Receptor Status Prediction in Breast Cancer Patients from DNA Methylation Data

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computational Natural Sciences by Research

by

Saksham Gupta
20161090
saksham.gupta@research.iiit.ac.in

International Institute of Information Technology, Hyderabad
(Deemed to be University)
Hyderabad - 500 032, INDIA

June 2023

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**Receptor Status Prediction in Breast Cancer Patients from DNA Methylation Data**" by **Saksham Gupta**, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Dr. Nita Parekh

To my family
and friends

Acknowledgments

I'd want to use this occasion to show my heartfelt gratitude to everyone who not only assisted me in finishing my research thesis but also helped me grow as a researcher.

First and foremost, I'd want to express my sincerest gratitude to my adviser, Prof. Nita Parekh, for her unwavering support over the years. She assisted me in honing my research skills, and I can see how far I've progressed ever since. This thesis was made possible by her unwavering commitment to high-quality work.

I would like to thank Prashanthi ma'am and Noor sir who helped me gain the necessary domain knowledge and were always willing to help me with whatever queries I had and brainstorm ideas. I'd also want to thank Dr. Tejas Dhamecha for his insightful comments and suggestions on my research.

I'd also like to express my gratitude to my family, who were a continual source of inspiration and encouragement throughout my research. I would not have been able to complete my research without their unwavering love and support. I would also like to thank all my friends who made my college life memorable and were a constant source of motivation.

Contents

Breast cancer continues to be a major worldwide health burden, and therapy and prognosis are greatly influenced by hormone receptor status. Immunohistochemistry (IHC) is currently the gold standard for evaluating the status of the estrogen receptor (ER), the progesterone receptor (PR), and the human epidermal growth factor receptor 2 (HER2). However, this approach has drawbacks, such as the potential for labelling errors and inconsistency with intrinsic subtypes. To increase the precision of predicting hormone receptor status, this thesis introduces a unique predictive modelling approach employing DNA methylation and gene expression data.

We created machine learning models that use data from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) to include DNA methylation profiles and gene expression data to forecast ER, PR, and HER2 status. Using MSigDB and STRINGDB, we discovered genes that had varying levels of methylation in relation to each receptor status and looked into the functional significance of these findings. Additionally, we used machine learning models based on noisy label training to address the issue of noisy labels that may be present as a result of potential mislabeling by IHC-based approaches.

The results of the study revealed that gene expression information and DNA methylation profiles are reliable indicators of hormone receptor status. Our models performed well as compared to traditional IHC techniques, indicating potential clinical value. Additionally, our method might predict brand-new biomarkers and offer deeper perceptions into the epigenetic pathways underlying breast cancer.

However, there are still certain drawbacks, such as class imbalance problems and the high-dimensionality of DNA methylation data, which may be resolved with the development of machine learning techniques and larger, more representative datasets. This thesis emphasises the potential of DNA methylation-based prediction models to increase the precision of determining hormone receptor status, provide useful information for individualised therapy approaches, and enhance patient prognosis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Breast Cancer

According to WHO, "Cancer is a large group of diseases that can start in almost any organ or tissue of the body when abnormal cells grow uncontrollably, go beyond their usual boundaries to invade adjoining parts of the body and/or spread to other organs". Cancer occurs when a single or a small group of cells produces faulty signals about how often the cells should multiply. These faulty signals may cause cells to grow uncontrollably and form a lump known as a tumour and are usually attributed to changes in the cell's DNA. While not all DNA changes are harmful, and cells quickly die off in case of DNA damage in most cases, some might lead to cancer. Breast Cancer (BC) is one of the most frequent cancers in women, with an incidence of one in every 29 Indian women. The chances of a woman dying from breast cancer currently is about 2.6%. Despite the fact that current statistics show a rise in breast cancer incidence, the mortality rate due to breast cancer has decreased by 1% every year from 2013 to 2018. Early detection and better therapies are thought to be responsible for the lower fatality rate. Breast cancer is currently seen as a collection of diseases that attack the same anatomical region, rather than a single disease, due to its nature. They respond to the same treatment in different ways, have varied clinical behaviours, and have diverse histological characteristics [35]. Molecular analysis of breast cancers by high-throughput methods have shown that these differences arise due to genetic and epigenetic changes at the molecular scale [36]. Such changes could lead to cancer in one or more of the following ways:

- High expression of oncogenes leads to the survival of cells that were otherwise designated for apoptosis and involve in cell growth and division instead.
- Repression of tumour suppressor genes prevents the regulation of cell multiplication.
- Production of abnormal proteins, which works differently than usual.

Morphologically, breast cancer can be split into two categories: In situ carcinoma and invasive carcinoma. The two subtypes are divided based on the fact that whether cancer has spread to other tissues from the site of origin or not. *In situ* carcinoma are those which have started in a milk duct or lobule but have not spread into the breast tissue; thus, cancer cannot metastasize beyond the breast to other body parts. On the other hand, *in situ* carcinoma can progress to invasive cancer in some cases, where they have spread into surrounding breast tissues. Invasive cancer could thus metastasize to

other parts of the body. Both of these subtypes (*in situ* carcinoma and invasive carcinoma) can further be classified into lobular or ductal carcinoma based on the fact whether cancer started in the milk-producing glands (lobules) or the cells that line a milk duct respectively. Based on their histological features, *in situ* ductal carcinoma is further divided into multiple subtypes such as solid, cribriform and comedo breast cancer. Similarly, invasive ductal carcinoma is also divided into subtypes such as medullary, tubular and inflammatory breast cancer. Figure 1.1 shows the normal duct and lobe cells and *in situ* carcinoma in both of them.

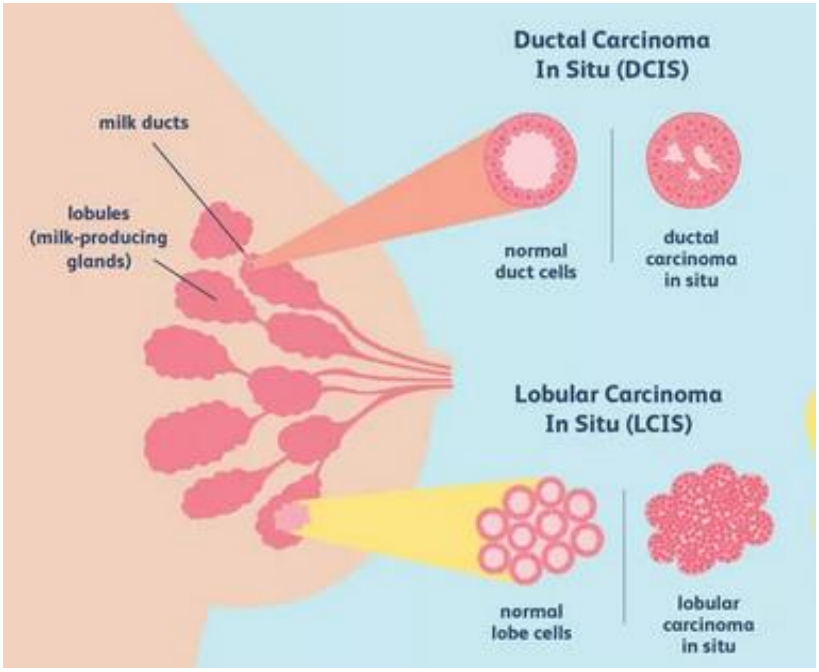
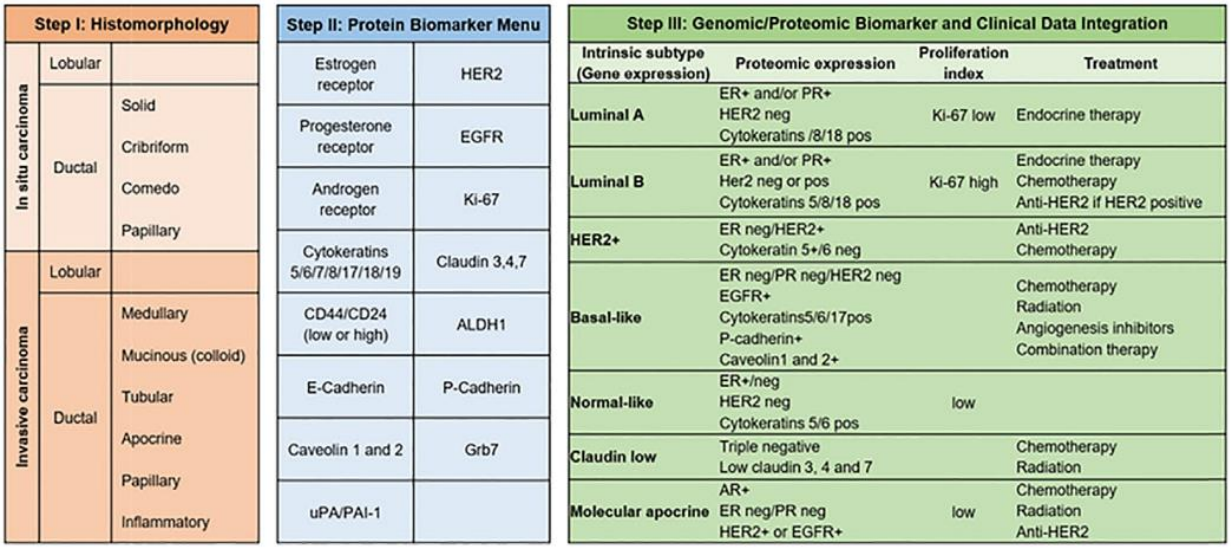


Figure 1.1 Normal duct and lobe cells along with their corresponding *in situ* carcinoma. ⟨Reproduced from: https://www.verywellhealth.com/breast-cancer-staging-stage-zero-429887⟩

Along with lifestyle-related factors, inherited DNA alterations can dramatically raise the risk of breast cancer. For example, a change in one of the tumour suppressor genes, BRCA1 and BRCA2, can be inherited by a child from a parent. Breast cancer is known to be a highly heterogeneous disease involving complex biological mechanisms. Breast cancer is classified into more than 20 major categories and 18 minor subtypes by the World Health Organization [53]. The histological classification of breast cancer, unlike the morphological classification (which is limited to the location of appearance of the tumour and whether the tumour has invaded the surrounding tissues), depends upon the cell type characteristics, the number of cells, type and location of secretion, immunohistochemical profile and architectural characteristics [54]. These features together define whether a tumour is ductal or lobular along with its sub-classification. The treatment plans and the survival rate of patients depend upon the accurate classification of breast cancer subtypes. Today, the histomorphological classification and clinical-pathological parameters are deemed insufficient to predict the real behaviour and treatment plan of cancer. Thus, molecular patterns of breast cancer are

analysed to get a better knowledge of the patient's breast cancer and thus determining the correct treatment for them. Traditionally, a patient's clinical subtype is determined by the expression status of three receptors [30] :

- Estrogen Receptor (ER)
- Progesterone Receptor (PR)
- Human Epidermal growth factor Receptor 2 (HER2)



Step I: Histomorphology		
In situ carcinoma	Lobular	
	Ductal	Solid
		Cribriform
		Comedo
		Papillary
Invasive carcinoma	Lobular	
	Ductal	Medullary
		Mucinous (colloid)
		Tubular
		Apocrine
		Papillary
		Inflammatory

Step II: Protein Biomarker Menu	
Estrogen receptor	HER2
Progesterone receptor	EGFR
Androgen receptor	Ki-67
Cytokeratins 5/6/7/8/17/18/19	Claudin 3,4,7
CD44/CD24 (low or high)	ALDH1
E-Cadherin	P-Cadherin
Caveolin 1 and 2	Grb7
uPA/PAI-1	

Step III: Genomic/Proteomic Biomarker and Clinical Data Integration			
Intrinsic subtype (Gene expression)	Proteomic expression	Proliferation index	Treatment
Luminal A	ER+ and/or PR+ HER2 neg Cytokeratins /8/18 pos	Ki-67 low	Endocrine therapy
Luminal B	ER+ and/or PR+ Her2 neg or pos Cytokeratins 5/8/18 pos	Ki-67 high	Endocrine therapy Chemotherapy Anti-HER2 if HER2 positive
HER2+	ER neg/HER2+ Cytokeratin 5+/6 neg		Anti-HER2 Chemotherapy
Basal-like	ER neg/PR neg/HER2 neg EGFR+ Cytokeratins5/6/17pos P-cadherin+ Caveolin1 and 2+		Chemotherapy Radiation Angiogenesis inhibitors Combination therapy
Normal-like	ER+/neg HER2 neg Cytokeratins 5/6 pos	low	
Claudin low	Triple negative Low claudin 3, 4 and 7		Chemotherapy Radiation
Molecular apocrine	AR+ ER neg/PR neg HER2+ or EGFR+	low	Chemotherapy Radiation Anti-HER2

Figure 1.2 Breast cancer subtypes are classified based on morphology, immunohistochemistry, and transcriptome data (Reproduced from [29]).

The molecular categorization of breast cancer may be split into four subgroups based on the expression of the above-mentioned receptors: Luminal A, Luminal B, Basal-like (Triple Negative) and HER2+. Other molecular subtypes of breast cancer are normal-like, claudin-low and molecular apocrine breast cancer. Figure 1.2 shows the breast cancer subtypes classified based on morphology, immunohistochemistry, and transcriptome data. The Luminal A subtype is the most prevalent, accounting for almost half of all breast cancer occurrences. The subtype is usually attributed to a highly favourable prognosis with a less painful clinical course. The subtype is known to have ER and/or PR positive status. Due to its positive hormone receptor status, patients suffering from this subtype benefit from hormone therapies. Hormone therapy is a technique that utilises hormones to delay or stop the development of cancer that is hormone-dependent. Luminal B subtype is observed in nearly 20-30% of all breast cancer cases. It is thought to be the most aggressive type of hormone-dependent breast cancer, requiring additional treatments such as chemotherapy in addition to hormonal therapy. HER2 subtype is observed in nearly 15-20% of all BC patients and is characterized by its strong HER2 expression along with low ER and PR expression. Anti-HER2 therapies like Trastuzumab and

Lapatinib work well on these tumours. TNBC (Triple-Negative Breast Cancer) is defined by a lack of expression for all three receptor statuses and occurs in roughly 10%-20% of all breast cancer cases. Patients with BRCA1 mutations, as well as young females, are more likely to develop these tumours. Owing to its morphological, genetic, and clinical heterogeneity, as well as a lack of targeted therapies, this subtype has proven to be difficult to be treated.

1.2 Epigenetics

Epigenetics is defined as the study of changes in gene function that are heritable at the mitotic and/or meiotic levels but do not involve a change in DNA sequence, as opposed to genetics [1]. The importance of epigenetics can be observed by the fact that different cells in our body serve different morphological and functional purposes, even though all of them have the same genetic information (DNA). This phenomenon of the same genetic material serving different purposes is possible because different body parts express different genes. Differential gene expression can occur throughout development and be maintained during mitosis [4]. Epigenetics has a major role in various biological functions such as tissue regeneration, genomic imprinting, X chromosome inactivation, transfer of information from one generation to the next, and ageing of organisms. Epigenetic alterations are known to be influenced by variables such as diet, behaviour, stress, physical exercise, working habits, drinking, etc [2]. Dietary factors vary a lot between people and between human populations and have shown to have a significant impact on the genome's epigenetic expression. Dietary changes can trigger not only epigenetic variations but can also transfix epigenetic changes. Environmental agents may also cause epigenome changes, which may lead to toxicity or carcinogenesis. Epigenetics has opened new doors in cancer research as studies have indicated a key role of epigenetic events in critical biological processes. DNA methylation, methyl-binding domain proteins, RNA-mediated gene silencing and histone modifications are some of the key processes which play a vital part in epigenetal control. Disruption of either DNA methylation or histone acetylation affects each other. For example, hypomethylation of CpG island in gene promoter regions leads to deacetylation of local histones and vice versa. However, the underlying mechanisms of the process are still unclear. Epigenetics have long been recognised as a factor in gene expression regulation and cancer. Epigenetic changes have been linked to a slew of different illnesses and diseases, including autoimmune disorders, obesity, hypertension, autism, and Alzheimer's disease.

Histone Modification: Histones are an important protein in providing a structural backbone to the DNA. Nucleosomes, the building blocks of chromatin, are formed by wrapping DNA around histone octamers i.e. two units each of H2A, H2B, H3 and H4. Each histone consists of a core and a histone tail. The extent to which DNA wraps around histones is altered when epigenetic factors attach to histone tails. Epigenetic alterations to the tails of histone proteins include methylation, acetylation, phosphorylation, and ubiquitination and may alter gene expression. The combination of these alterations affects the degree to which DNA is wrapped around histones, thus increasing or decreasing transcriptional accessibility. As shown in Figure 1.3, when a gene binds tightly around a histone, the gene remains inactive, but when an epigenetic factor adds to the histone tail, the DNA is loosely

wrapped around the histone, thus it is activated. Histone phosphorylation is important during mitosis and meiosis, and it interacts with other histone modifications that affect gene transcription. Phosphorylation can affect all histones at different sites at their tails. Histone methyltransferases (HMTs) bind to specific DNA sequences and interact with Trithorax group (Trx) proteins, Polycomb group (PcG) proteins, and RNA-interference (RNAi) to mediate histone methylation. Modifications to DNA methylation are also linked with modifications to histone marks; for example, in cell differentiation, the loss of active H3K4 tri-methylation and the retention of repressive H3K27 tri-methylation corresponds to an increase in DNA methylation. These findings show the presence of bidirectional crosstalk between DNA methylation and histone modifications.

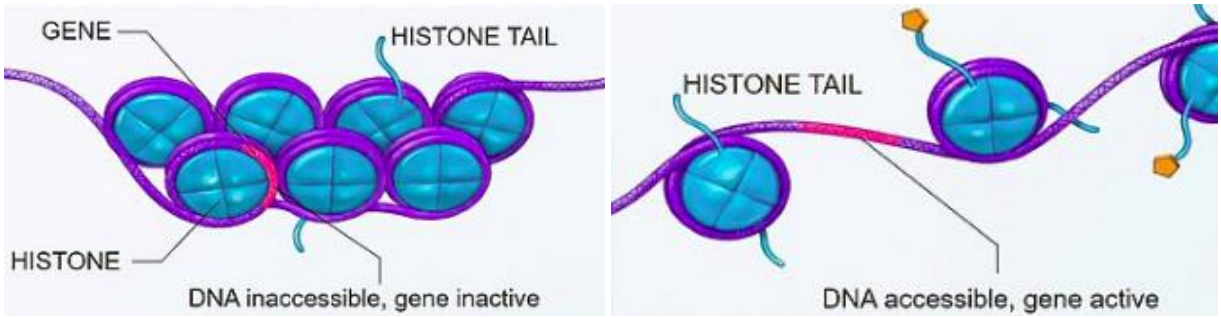


Figure 1.3 The role of histones in the activation and inactivation of genes. Left: When epigenetic factor is not attached to the histone tail, the DNA is tightly wound around the histones and is inaccessible, thus making the region inactive. Right: When epigenetic factor is attached to the histone tail, the DNA is loosely bound around the histones, making it accessible and the region activated. ⟨Reproduced from: https://en.wikipedia.org/wiki/Epigenetics#/media/File:Epigenetic_mechanisms.png⟩

DNA Methylation: DNA methylation is a key component of epigenetic mechanisms that control gene expression by chemically altering DNA without changing the underlying DNA sequence. Cytosine bases located 5’ to a guanine base gets a methyl group covalently attached to its 5-carbon position to convert cytosine to 5-methylcytosine. This cytosine alteration has no effect on the DNA sequence but may have an effect on its regulation. A CpG site is a position in the DNA sequence where a guanine nucleotide lies after cytosine along the 5’ to 3’ direction. An estimated 60 - 80% of the total 28 million CpG sites are methylated in the human genome [12]. Initiation and modification of DNA methylation are known to be caused by at least three DNA methyltransferases (DNMTs): DNMT1, DNMT3a and DNMT3b. DNMT1 has a high affinity to hemimethylated DNA and is required for correct embryonic development as it transfers DNA methylation patterns to a newly replicated DNA sequence. DNMT3a and DNMT3b have a high affinity for unmethylated DNA and can initiate de novo methylation. The exact mechanisms involved in DNA demethylation are unknown, but it is thought to occur via the ten-eleven-translocase (TET) enzymes oxidising 5-methylcytosine to 5-hydroxymethylcytosine, followed by a base-excision repair mechanism. Figure 1.4 shows the DNA methylation and demethylation processes. The precise role of DNA methylation throughout the genome is unclear, with various functions ascribed to distinct genomic locations. It is widely known that DNA methylation at gene promoters plays a function in transcriptional suppression. Hypermethylated CpG islands lead to tightly compacted chromatin, which prevents the initiation of

transcription. Thus, when gene promoters are hypermethylated, they are unable to bind with transcription factors that cause gene inactivation. DNA methylation is required for genomic imprinting, which occurs when one of the paternal or maternal alleles expresses a gene while the other allele is epigenetically repressed [74]. Female X chromosome silencing is linked with DNA hypermethylation and heterochromatin formation caused by antisense RNA.

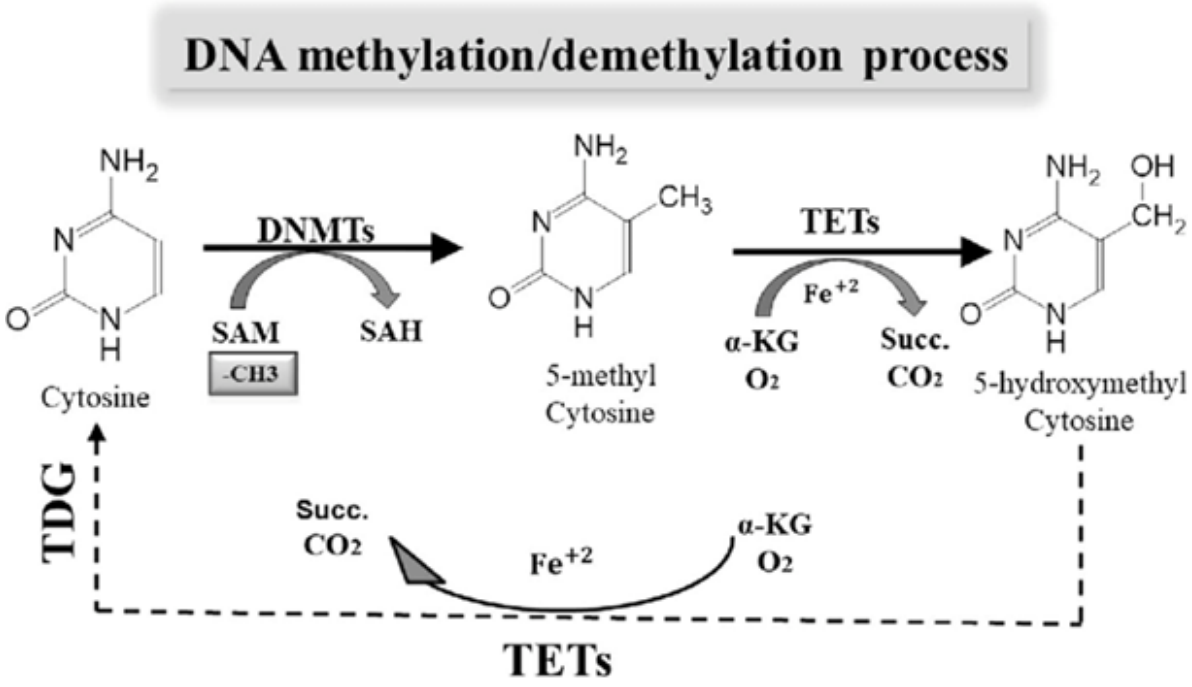


Figure 1.4 DNA methylation and demethylation processes occurring via DNMTs and TETs respectively (Reproduced from [73]).

1.2.1 Role of Epigenetics in Cancer

A lot of ongoing research shows the role of epigenetics in various human diseases, including cancer. The majority of epigenetic modifications occur during cell differentiation and are stably maintained over many cell division cycles, allowing cells to have distinct identities while retaining the same genetic material. If not maintained properly, these heritable epigenetic markers can result in improper activation or inhibition of different signalling pathways, leading to diseases like cancer. Traditionally, cancer was viewed as a purely genetic disease. But due to the development of epigenetics, it is now known that epigenetic alterations also play a crucial role in cancer and in addition to various genetic changes, human cancer cells have global epigenetic abnormalities. Two types of abnormal DNA methylation are found in human cancer: gene promoter-associated hypermethylation and global hypomethylation. Hypermethylation of promoter regions of tumour suppressors and other cancer-related genes has been implicated in the silencing of these genes in a number of studies. By misregulating chromatin structure and activity, abnormal histone-modifying factor activity may also

promote cancer development through deregulation of gene transcription and DNA repair. The fact that epigenetic alterations are reversible, makes epigenetic therapy a promising field for early detection, risk assessment and providing a better treatment plan for treating cancer patients by reversing or arresting the growth of cancer [3]. Cancer is a disease with a high degree of heterogeneity; thus, understanding the role of epigenetics in cancer could help us open doors to new treatment and personalised medications.

1.2.2 Role of Epigenetics in Breast Cancer

Breast cancer is caused by genetic and epigenetic changes, the latter happening during the early stages of the disease. Previous findings show a global decrease of DNA methylation in base-pair resolved primary breast cancer tissues [11]. This loss in DNA methylation leads to instability and activation of regulatory DNA sequences such as oncogenes, retrotransposons and genes that play a role in tumour cell development. High levels of methylation in tumour suppressor genes lead to the silencing of genes that block growth-promoting proteins. Many of the genes involved in tumour suppression, cell cycle control, apoptosis, angiogenesis, tissue invasion, and metastasis have been shown to be hypermethylated in primary breast tumours or BC cell lines [54]. Hypermethylation of the BRCA1 gene is one of the most researched epigenetic changes in BC, which results in the downregulation of tumour suppressor genes involved in DNA maintenance and repair. Hypermethylation of the promoter region of the PTEN gene, a tumour suppressor, activates the Akt pathway, suppresses apoptosis, and increases cell survival. Hypermethylation of the CpG islands in the promoter region has been linked to the progression of BC. Studies have shown that tumours that express hormone receptors and those that do not express hormone receptors have distinct epigenetic patterns [55]. In later research, it was shown that the differences in methylation patterns between hormone receptor-positive and hormone receptor-negative breast tumours affect the tumour response to hormonal treatment such as tamoxifen. Breast cancer subtypes have been linked to epigenetic patterns, and the methylation profiles of Basal-like, Luminal A, and Luminal B tumours differ. Methylation was found to be significantly higher in Luminal B samples than in Basal-like tumours. Cell cycle regulation, DNA repair, hormone regulation, cell adhesion, invasion, angiogenesis, and cellular growth-inhibitory signalling related genes are all impacted by abnormal epigenetic modifications in breast cancer.

1.3 DNA Methylation Profiling Methods

The first step in determining the methylation level of DNA is to transform the original DNA so that the methylated sequences can be differentiated from the unmethylated sequences. DNA methylation profiling can be mainly categorised into three different approaches [5, 6]:

1.3.1 Restriction Enzyme-Based Method

In this technique, either one of the two cytosines (methylated or unmethylated) are cleaved using restriction enzymes, and the other cytosines remain intact. The DNA fragments are then size-selected and sequenced (MRE-seq). Figure 1.5 shows the two commonly used Enzyme digestion based methods. The Differential Methylation Hybridization (DMH) method digests genomic DNA using MseI, which cuts the DNA into small fragments. The fragments are then ligated with linkers, and the unmethylated fragments are digested using either BstUI and/or HpaII restriction enzymes. The methylated fragments are then PCR amplified, which are then labelled and hybridised to arrays. This method is relatively more straightforward but has a major limitation: the restriction sites are mainly distributed in CpG islands; thus, it can identify a limited fraction of the total genomic CpG sites. To tackle this issue, the McrBC methodology uses the McrBC restriction enzyme that digests methylated fragments instead of unmethylated fragments, thus can digest densely methylated regions. Finally, high-throughput direct sequencing is used to determine the DNA methylation profile of the genome. The method is cost-effective and but the major disadvantage of McrBC is that it has relatively low resolution and is limited to regions lying near enzyme recognition sites.

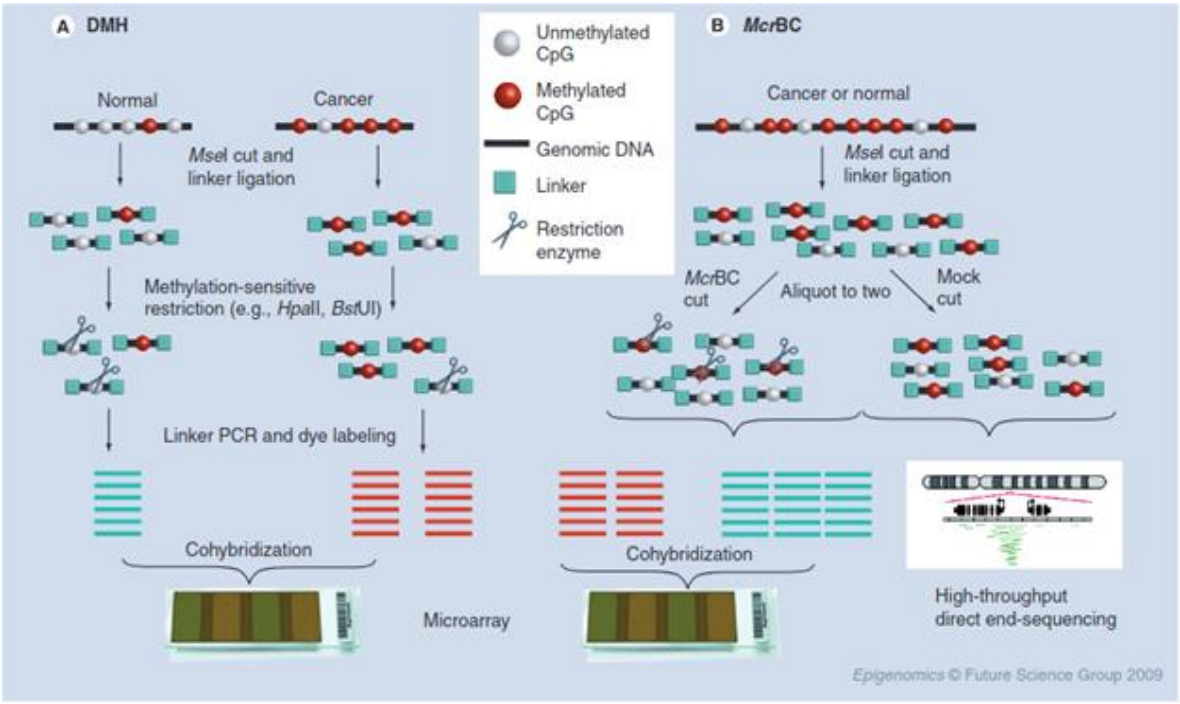


Figure 1.5 Restriction enzyme-based DNA methylation profiling methods. (A) DMH method; (B) McrBC method (Reproduced from [5]).

1.3.2 Bisulfite Conversion

When genomic DNA is treated with sodium bisulfite, unmethylated Cytosine (C) is deaminated to Uracil (U). In contrast, methylated Cytosine (C) residues remain unaltered from the effect of sodium bisulfite [7]. The Uracil gets converted to Thymine on PCR amplification. Figure 1.6 shows how bisulfite conversion and PCR amplification of a DNA fragment can lead to up to four different sequences. The methylation state of each position is inferred by aligning and comparing the bisulfite-converted sequences with the original sequence using tools such as Bismark [68]. The sequence reads are parallelly aligned against the bisulfite genome and the read with the best alignment is compared with the normal genomic sequence to infer the methylation states of the cytosines. Figure 1.7 shows two bisulfite conversion-based methods: Reduced Representation Bisulfite Sequencing (RRBS) and Padlock probes. The RRBS method fragments DNA using MspI or Bg/II digestion, which is then size-selected for DNA fragments between 40 to 220 base pair size. The DNA segments are treated with bisulfite conversion and PCR amplification before being analysed by next-generation sequencing methods. A padlock probe is made up of a standard linker sequence that connects two variable capture arms capable of annealing to two genomic DNA regions separated by hundreds of bases [66]. The padlock probe method uses padlock probes designed to target specific CpG sites, which can attach to both 5' and 3' ends of the target CpG sites. The probes attach to the bisulfite-converted target DNA sites. The extension and ligation process is done with the help of a mixture of dNTPs, polymerase and ligase. The leftover padlocks and genomic DNA is removed with the help of exonucleases, and finally, PCR amplification is performed before next-generation sequencing. There are also microarray-based bisulfite conversion methods used by the Infinium platform, which is discussed in the later section 1.4. This technique enables high-resolution analysis of genome-wide methylation patterns, but the drawback of these methods is their cost for whole-genome sequencing and substantial DNA degradation that leads to problems in designing hybridization array probes and mapping sequencing reads.

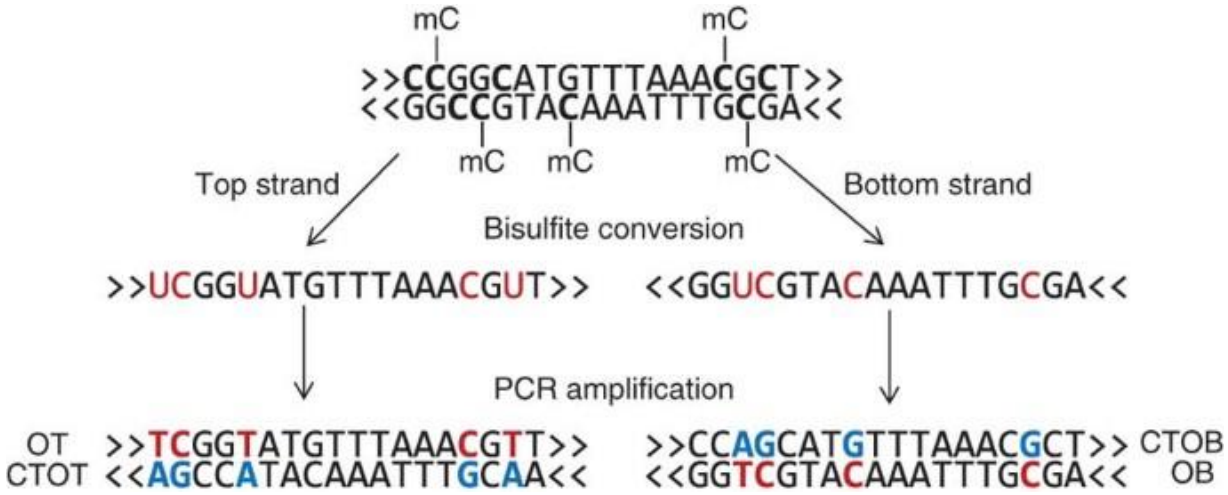


Figure 1.6 Bisulfite conversion of genomic DNA and subsequent PCR amplification (Reproduced from [67]). mC: 5-methylcytosine; OT: original top strand; CTOT: strand complementary to the original top strand; OB: original bottom strand; and CTOB: strand complementary to the original bottom strand.

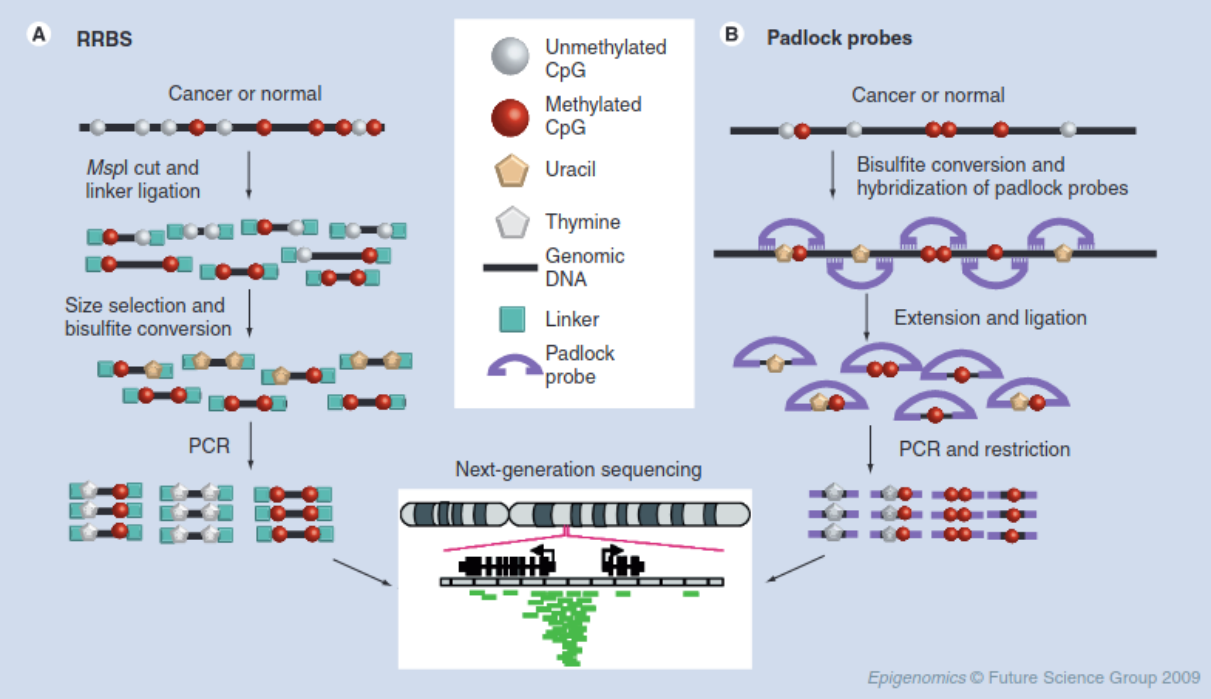


Figure 1.7 Bisulfite conversion-based DNA methylation profiling methods. (A) RRBS method; (B) Padlock probes method (Reproduced from [5]).

1.3.3 Affinity Enrichment

This technique uses the methyl-CpG-binding domain (MBD) proteins or 5mC-specific antibodies to capture methylated DNA sequences [6]. The antibody-based approach is known as methylated DNA immunoprecipitation (MeDIP) and the MBD proteins based approach is known as methylated CGI recovery assay (MIRA). Figure 1.8 shows the pipeline for MeDIP and MIRA approaches. In MeDIP, the DNA sequence is first fragmented using sonification and denaturation, following which anti-5mC antibodies that bind to methylated fragments are used to enrich those fragments. The enriched hypermethylated DNA fragments and the total input DNA are labelled with Cy5 and Cy3 fluorescent dyes respectively. The methylation status is finally determined by the ratio of fluorescent intensity of Cy5 to Cy3. MIRA method uses MBD proteins, unlike antibodies used by the MeDIP method. The method does not require denaturation of DNA and uses MBD to enrich methylated DNA. Affinity enrichment-based methods have certain advantages as they do not modify the DNA sequence like restriction enzyme and bisulfite conversion-based methods. The method is not restricted to the number of restriction sites present in the genome, unlike restriction enzyme-based methods; and there is no problem in designing hybridization array probes and mapping sequencing reads, unlike bisulfite conversion-based methods. However, both of these methods have certain limitations. The MeDIP method has a bias towards hypermethylated regions due to the antibodies' affinity binding; thus they are suitable for detecting hypermethylated CpG sites and has a major disadvantage in that it is rather inaccurate in hypomethylated regions. The MIRA method can detect hypomethylated regions; however, the use of MBD proteins causes preferential enrichment of CpG islands over other CpG regions [69].

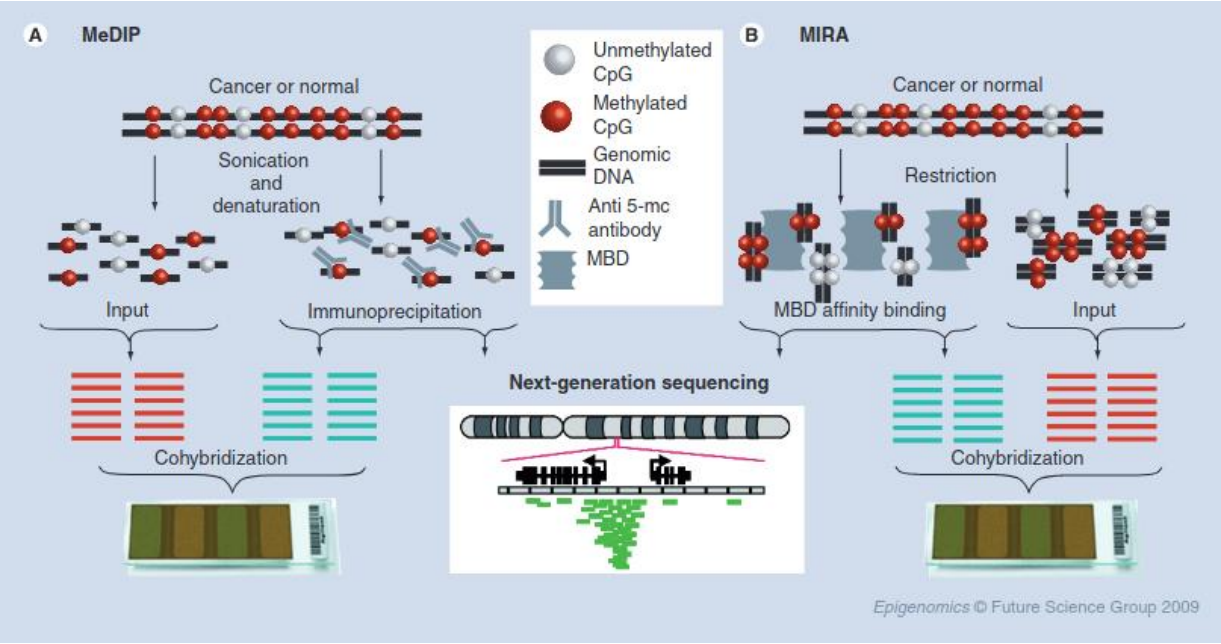


Figure 1.8 Affinity enrichment-based DNA methylation profiling methods. (A) MeDIP method; (B) MIRA method (Reproduced from [5]).

1.4 Platforms for DNA methylation

Once the DNA is transformed to distinguish a methylated sequence from an unmethylated sequence, the next step is the identification of the methylated and unmethylated sites. DNA hybridization array (also known as microarrays) and Next Generation Sequencing (NGS) are the two most often used ways of identification. NGS method allows millions of sequencing reactions in parallel. They do not require the reactions to be physically separated into different wells or tubes. Although different NGS platforms use different methods, the typical workflow is the same for all of them. The first step is library preparation which includes DNA fragmentation and adapter ligation. The next step is library amplification, followed by sequencing using different approaches [70]. Figure 1.9 shows the basic steps involved in NGS. The current NGS techniques are pyrosequencing, sequencing by ligation and sequencing by synthesis.

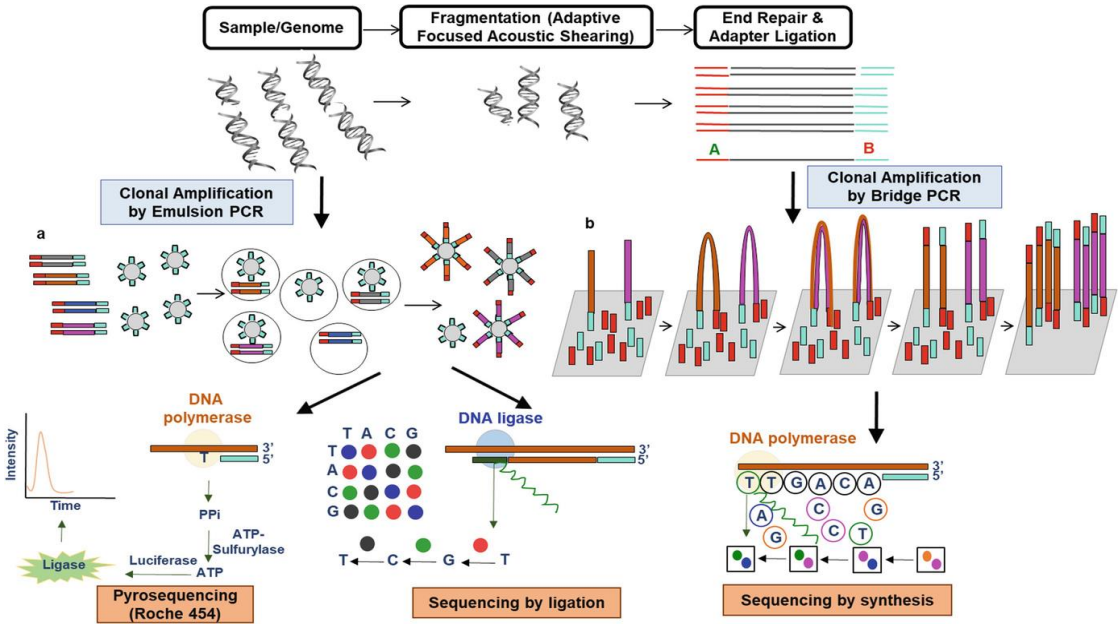


Figure 1.9 The fundamental stages involved in DNA sequencing using various NGS platforms (Reproduced from [70]).

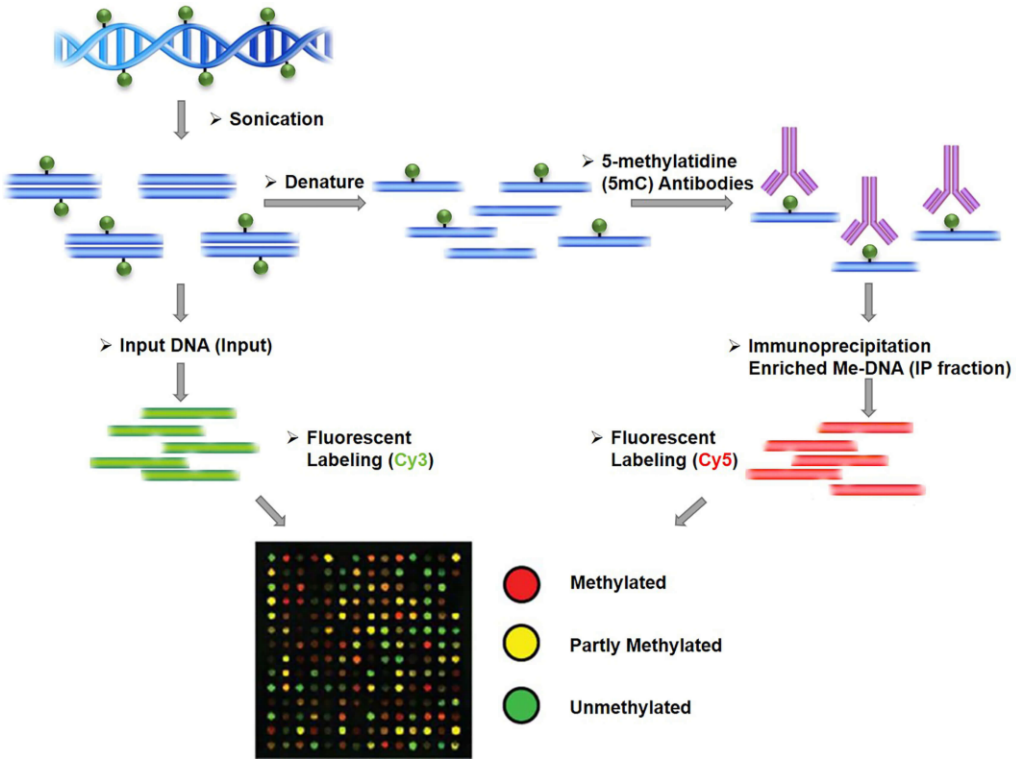


Figure 1.10 Steps involved in two-colour microarray-based methods (Reproduced from [71]).

One of the common approaches for DNA methylation profile analysis is using a two-color microarray method. Figure 1.10 shows the steps involved in two-colour microarray-based methods. A DNA microarray plate consists of multiple surface wells located on a glass or silicon slide. Each surface well has a unique DNA probe attached to them. The DNA probes of the sample are amplified and printed on the microarray slides. Two different fluorescent dyes that bind exclusively with methylated/unmethylated DNA are added to each probe and hybridize with the samples. As the dyes are distinct in color, it is possible to distinguish two different DNA states on the microarray plate. To do so, the chips are scanned and the fluorescent signal intensity of each probe is determined.

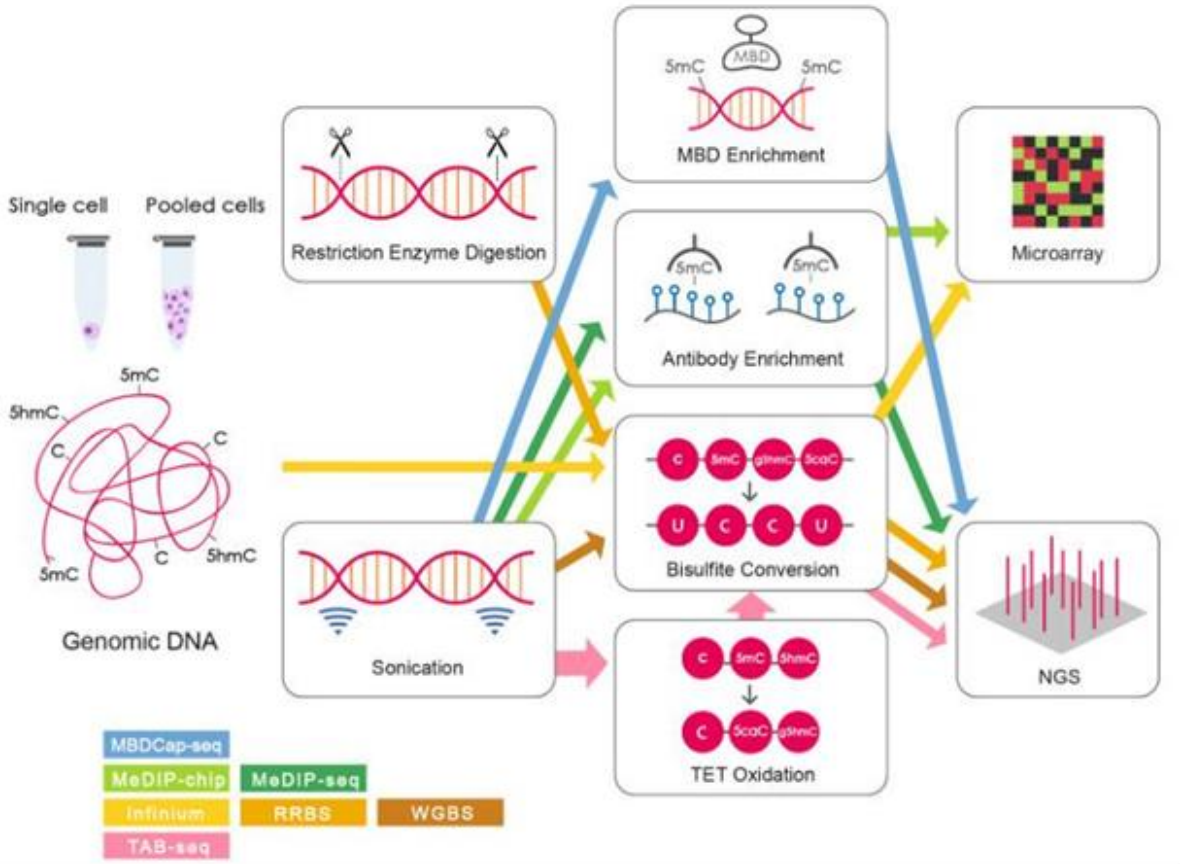


Figure 1.11 Commonly used DNA methylation profiling methods and their pipeline for methylation analysis (Reproduced from [6]).

Each profiling method can sequence the DNA by using either Microarray or NGS based techniques. Figure 1.11 shows a summary of commonly used methods and the profiling and sequencing method they follow. Either restriction enzyme digestion or sonification first fragments the genomic DNA, then the fragmented DNA is profiled by either antibody enrichment, MBD enrichment or bisulfite conversion.

The Illumina Infinium HumanMethylation platform uses bisulfite conversion-based methods for DNA transformation and microarray-based methods for methylation identification, and they are widely used for genome-wide DNA methylation-based studies. Illumina DNA methylation microarrays are currently one of the most frequently utilised techniques for studying DNA methylation in humans. One of the disadvantages of Illumina DNA methylation microarrays is that they are marketed as suitable for human samples only. Recent experiments on mice show that Illumina Human Infinium Methylation EPIC microarray can have limited use in differential methylation analysis in mice [38].

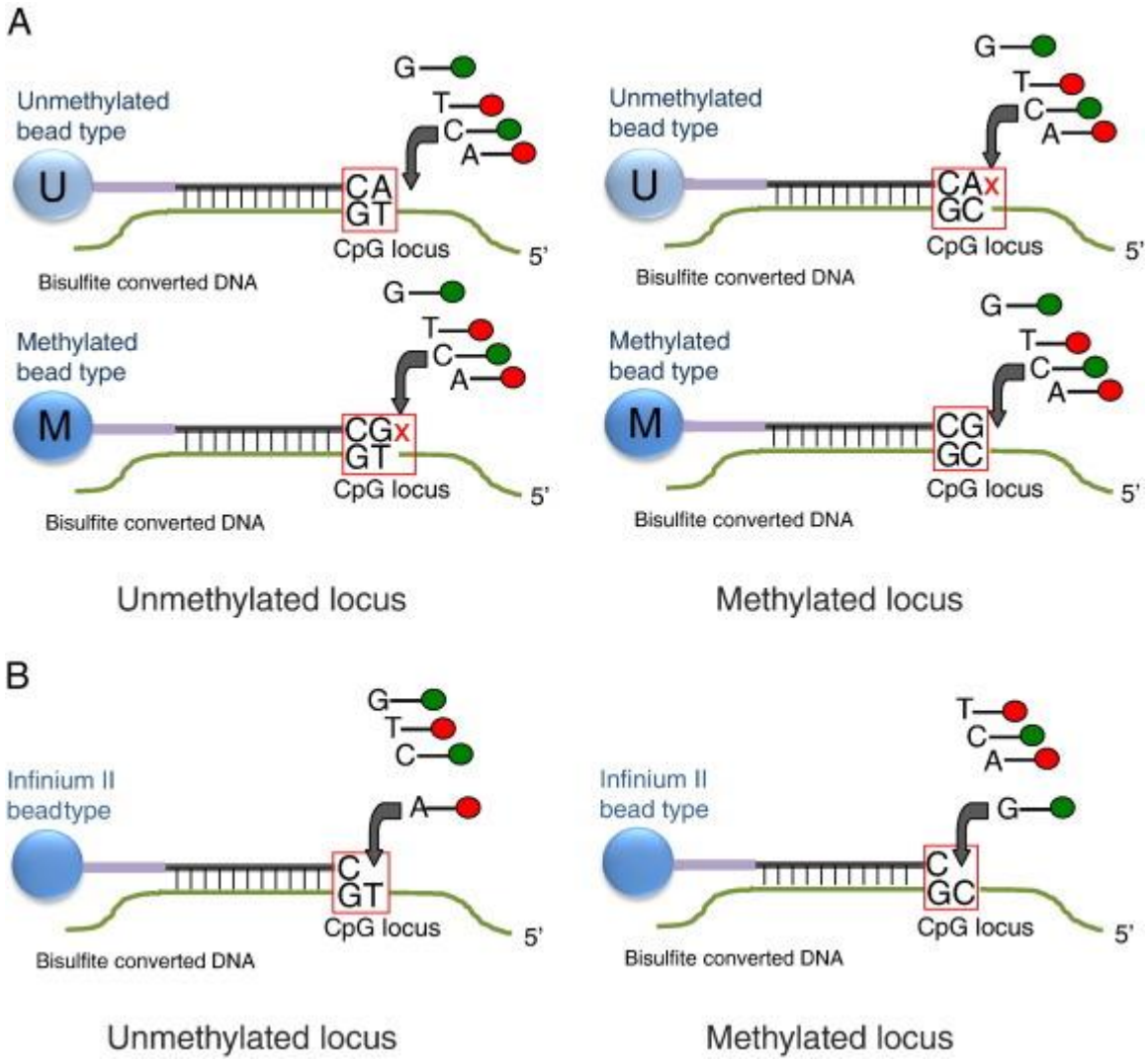


Figure 1.12 Comparison of the Infium I and Infium II assays used in the Illumina HumanMethylation 27K and 450K Beadchips**. A.** Infinium I assay involves using two types of probes. One probe is complementary to methylated cytosine. The second probe binds to the thymine that has appeared as a result of bisulfite conversion. **B.** Infinium II assay uses only one probe that can cover up to 3 CpG sites. During single-base extension, labelled adenine complements the thymine, while cytosine protected from the bisulfite conversion by the methyl group is complemented with the labelled guanine (Reproduced from [33]).

Illumina Methylation Beadchips are based on the Illumina Infinium assay. There are two types of Infinium assays - Infinium I and Infinium II. Figure 1.12 shows the comparision of the Infium I and Infium II assays used in the Illumina HumanMethylation 27K and 450K Beadchips. In Infinium I, two probes are used for each CpG locus: a “methylated” and an “unmethylated” probe. The methylated probe is designed to match the protected cytosine. The unmethylated probe matches the thymine base that appears as the result of bisulfite conversion. The Infinium II assay uses only one probe per locus.

Each probe can contain up to 3 underlying CpG sites. For each probe, its 3` terminus complements the base directly upstream of the query site. A single base extension results in the addition of a labelled guanine or adenine base, complementary to either the 'methylated' cytosine or 'unmethylated' thymine (unmethylated cytosine converted to thymine by bisulfite conversion and amplification), respectively. Depending on the approach utilized, Illumina Human Methylation Beadchips can detect a different number of CpG sites.

1.4.1 Illumina Humanmethylation 27K Beadchip

The Illumina Human Methylation 27K Beadchip platform targets 27,578 CpG sites. The CpG sites are found in 14,495 genes, which includes RefSeq genes from the NCBI CCDS Database. The Infinium I assay design used in the measurement of DNA methylation is done by two beads. Figure 1.12, part A shows how the Infinium I design uses two probes; one probe measures the Methylated (M) intensity of the CpG site, while the other probe measures the Unmethylated (U) intensity of the CpG site. The probes either bind to the protected cytosine (methylated design) or the thymine base formed during bisulfite conversion and amplification (unmethylated design). The methylation level can be then determined by the intensity ratio of the two probes. The major disadvantage of this platform is that it does not cover a lot of regions except the gene promoter regions in the genome.

1.4.2 Illumina Humanmethylation 450K Beadchip

The Illumina Humanmethylation 450K BeadChip targets 96% of the human genome's CpG islands, which provides methylation information of about 480K CpG sites [33]. It covers 99% of the RefSeq genes and has a 17-fold increase in coverage over the 27K Beadchip array. The 450K beadchip array covers all gene regions in addition to CpG islands and island shores, unlike the 27K beadchip which only covers gene promoter regions. Illumina 450K Beadchip combines both Infinium I and Infinium II approaches, unlike Illumina Human Methylation 27K BeadChip, which utilizes one type of probe. Due to application of both types of the coverage for the methylation analysis is increased. The two types of probe designs used by Illumina 450K beadchip are:

- Infinium I (for 135501 CpG sites)
- Infinium II (for 350076 CpG sites)

Therefore, the combination of both approaches can potentially detect 485K CpG sites in total. Considering that the detected sites can overlap, the beadchip can detect around 450K unique methylated sites or more, depending on their specific features. The principle of the Infinium I design is similar to the method used in Illumina Human Methylation 27K Beadchip described earlier. For Infinium II, the measurement of DNA methylation is achieved with the help of a single probe that measures the methylated (M) and unmethylated (U) intensities through red and green dye colours. As seen in Figure 1.12, B, Illumina II design uses a single bead to measure the methylation intensity of the CpG site. During single-base extension, the labeled probe either binds to the methylated cytosine or thymine (unmethylated cytosine converted to thymine by bisulfite conversion and amplification).

1.4.3 Illumina MethylationEPIC Beadchip (EPIC)

The Illumina MethylationEPIC Beadchip microarray is the most recent DNA methylation array available on the Illumina microarray platform. The EPIC and 450K microarray uses the same DNA methylation protocols and probe designs namely, Infinium I and Infinium II. The EPIC array provides methylation information for 866,836 cytosine sites on the human genome. It includes 450,161 CpG probes from the 450K beadchip. The EPIC beadchip omitted 32,260 probes present in 450K beadchip due to being flagged unreliable previously. EPIC includes 413,743 additional CpG probes, including over 350,000 CpG sites identified as possible enhancers in the FANTOM5 and ENCODE projects. The distribution of probes across various genome annotation categories (GENCODE19 genes, CpG islands, and regulatory areas) as defined by the ENCODE and FANTOM5 projects is shown in Figure 1.14. The probes are classified as those that are shared across EPIC and 450K and those that are exclusive to EPIC.

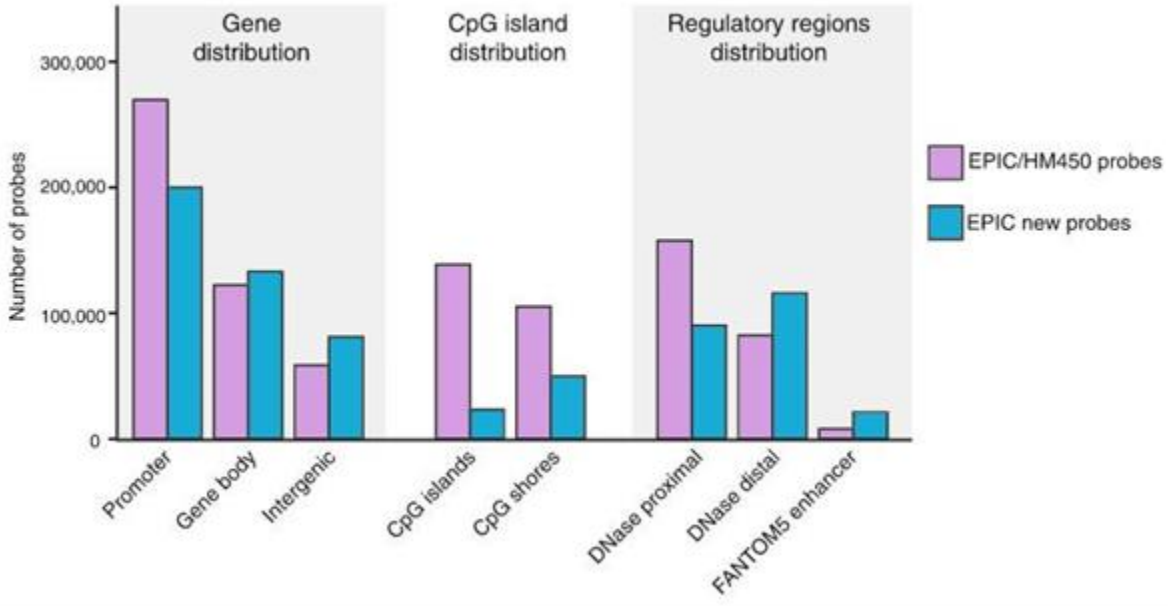


Figure 1.14 The distribution of probes across various categories of genome annotation based on publicly available catalogues (Reproduced from [72]).

1.5 Review of DNA methylation studies in Breast Cancer

Previously, several studies established a link between DNA methylation patterns and the risk of breast cancer. A study in 2019 [28] established the relationship between DNA methylation and gene expression in breast cancer patients. The study used the TCGA and METABRIC dataset for DNA methylation and gene expression data and derived the methylation profiles. They identified 368 differentially methylated CpG positions between tumour and normal breast tissue samples and showed their association with gene expression data. They found that 56% of hypermethylated CpGs were found

in upstream promoter regions, and 66% of hypomethylated CpG sites were found in the gene body. 209 of the 368 differentially methylated CpG sites, which were located in 169 genes, were differentially expressed between tumour and normal breast tissue. They observed that 70% of promoter CpG sites' methylation-expression was negatively correlated and 74% of gene body CpG sites' methylation-expression was positively correlated. They also identified novel DNA methylation markers that might be useful for diagnostic and prognostic roles in breast cancer.

1.6 Receptors and Their Role in Breast Cancer

Cancer biology is complex and therefore, predictive and prognostic markers play a crucial role in determining the treatment plan for breast cancer patients [31]. While predictive markers help predict a treatment plan's outcome, prognostic factors help determine the patient's clinical outcome in the absence of standard therapy. The most widely studied predictive markers whose expression status is associated with BC are: ER (Estrogen Receptor), PR (Progesterone Receptor), and the HER2 (Human Epidermal Growth factor Receptor 2). These have been used for intrinsic subtype classification of BC and are currently used in clinical setting. ER and PR bind their respective hormone ligands and regulate the expression of genes associated with cell growth and proliferation [99]. Similarly, HER2 also activates cell signaling pathways that drive cellular proliferation pathways [100].

In breast cancers, the hormone-receptor status is either positive or negative; for example, ER-positive breast cancer means that the cancer cells express the estrogen receptor. Establishing the hormone receptor status of breast cancer helps devise the appropriate treatment approach. Only three markers are primarily used to determine the molecular subtyping and subsequently the treatment plan for the patient: the Estrogen and Progesterone Receptors act as a predictive marker for response to hormone therapy [101], and the HER2 status acts as a predictive marker for HER2-targeted therapies [102]. Receptor status could also help to stage the cancer i.e., identify whether it is at its initial, receptor-positive and differentiated stage, or is at an advanced cancer stage, where the cells in a tumor mass are undifferentiated and receptor-negative [103]. However, the receptor-negative status alone cannot be used to stage a cancer as clinical studies have failed to establish its individual diagnostic value. Similarly, only ER expression alone poorly predicts outcome and further indicators of response or resistance are required. Compared to hormone-receptor positive BC, receptor-negative BC proliferate quickly, and may have metastasized before being diagnosed. In comparison to ER and PR-positive BC, other sub-types of breast cancers show increased risk of mortality over a 5-year survival period, ranging from a 1.5-fold increase in mortality in ER+/PR- cancers, 2-fold increase in ER-/PR+ cancers, and 2.6-fold increase in ER-/PR- cancers [40].

1.6.1 Estrogen Receptors

In 80% of breast malignancies, estrogen promotes tumour cell biology by activating estrogen receptors. Estrogen Receptors are a subtype of hormone receptors that are activated by the steroid hormone estrogen. 17-β-estradiol (E2) is the most common type of estrogen produced by the ovaries

and adrenal glands, and it exerts its function through activation of ER [104]. When activated, estrogen receptors stimulate normal breast epithelial cell growth and may also promote proliferation in invasive breast cancer cells [105]. For example, coactivators of E2, such as SRC-1 and CBP have histone acetyltransferase activity, and corepressors such as MTA1 and NCOR have histone deacetylase activity. These proteins play an epigenetic role by modifying gene expression. The multi-protein regulatory complex is dysregulated in ER-positive breast cancers [39].

Immunohistochemistry (IHC) is employed to prepare histological samples of tumor biopsies to determine the ER status in breast cancer cells. ER-negative breast cancer is a different disease entity, and has a poor clinical outcome when compared to ER-positive breast cancer [106]. If a breast cancer is ER-positive, it indicates an upregulation of ER-induced activation of gene expression, which in turn leads to growth and proliferation of breast cancer epithelial cells. Preventing estrogen from binding its receptor (ER) can reduce tumor growth and metastasis. This is referred to as hormone therapy. Through administering a drug such as tamoxifen, estrogen receptors are blocked, thereby downregulating the proliferation of breast cancer [97]. Hormone therapy aims to either decrease estrogen-stimulated growth (tamoxifen) or to reduce estrogen production (aromatase inhibitors) [41].

1.6.2 Progesterone Receptors

Progesterone Receptors are also a subtype of hormone receptor that are activated by the steroid hormone progesterone. When activated, PRs also stimulate the growth of tumour cells [52]. The PGR gene on chromosome 11 encodes progesterone receptors, and is an ER target gene- which means that when the ER is activated, PGR is one of the downstream genes expressed for growth and development of the mammary gland [47]. Because ER regulates PR expression, the presence of PR usually indicates that the ER pathway is intact and functioning. Around 70% of all breast cancers have PR-positive status. While PR and ER expression are highly correlated, the correlation is imperfect, resulting in four different phenotypes of combined expression. Both the ER and PR status need to be determined as each combination is associated with significantly varied rates of hormonal treatment response [45].

1.6.3 Human Epidermal Growth Factor Receptors 2

HER2 is a growth factor receptor protein that is encoded by the ERBB2 gene (Erb-B2 Receptor Tyrosine Kinase 2) on chromosome 17 in humans, and is a member of the ErbB family of transmembrane RTKs (receptor tyrosine kinase) proteins. This subclass of cell-surface growth factor receptors is known to influence cell differentiation, proliferation, and survival [50,51]. HER2 can undergo ligand-independent dimerization; when HER2 undergoes heterodimerization with HER3, it is a potent anti-apoptotic stimulator of the Phosphoinositide-3-kinase (PI3K)/Akt pathway- a cell signaling pathway that is most activated in oncogenesis [46].

HER2 status is commonly determined by either immunohistochemistry (IHC) method at the protein level, or by fluorescence in situ hybridization (FISH) or chromogenic in situ hybridization (CISH) method at the DNA level. In 15%–25% of breast cancers, HER2 is overexpressed [48,49]. The most

common signaling pathways upregulated in HER2 breast cancers are the mitogen-activated protein kinase pathway (MAPK), Phosphoinositide-3-kinase (PI3K)/Akt pathway, and protein kinase C (PKC) pathway. These signaling pathways are involved in cell proliferation, migration, apoptosis, and effect breast cancer tumor growth, proliferation and metastasis. It is observed that high HER2 amplification may result in poor outcome or even resistance to cancer treatments. HER2 amplified breast cancers show increased proliferation rates, more aneuploidy (presence of an abnormal number of chromosomes) [42], and tend to metastasize to the central nervous system (CNS) [43] and viscera [44]. The discovery that HER2-positive breast cancer patients had a poorer prognosis than HER2-negative breast cancer patients has laid the groundwork for establishing HER2-targeted treatment strategies in modern oncology [41]. Almost two decades after this observation, trastuzumab was established as a therapy known to benefit HER2-positive breast cancer patients. Another treatment designed to block the growth of HER2-positive cancers is lapatinib, which binds and blocks the receptor tyrosine kinase ATP-binding domain, thus inhibiting tumor growth [40].

1.7 Review of Studies for Receptor Status Prediction

The most common practice used by laboratories to determine the Receptor status of breast cancer patients is the ImmunoHistoChemistry (IHC) method. This method uses an antibody to target the extracellular or intramembranous domain of the receptor. The attached antibody is then visualised with the help of an antigen-antibody reaction that binds to a colour-emitting molecule. After that, the scoring is carried out using a light microscope. Although the IHC method is less accurate, it is faster and cheaper than the more accurate Fluorescence in situ hybridisation (FISH) method. Most pathologists initially use the IHC method and using the FISH method as a confirmatory technique or in intermediate cases where the results obtained by the IHC method are unsure. FISH test is widely used to identify the genetic material present in a person's cells and to visualize specific regions of chromosomes. In this cytogenetic technique, fluorescent labelled DNA probes are hybridized with denatured chromosomal DNA to visualise specific portions of chromosomes. For example, in HER2 testing FISH test is employed to see the number of copies of HER2 gene in each nucleus. Preferably 20-60 cells must be assessed from 3 different tumor fields to determine the HER2 to chromosome 17 ratio [37]. Both IHC and in situ hybridization (ISH) are FDA-approved techniques, but both methods have shown contradictory results. R Memon, (2021) studied HER2 status in patients of breast cancer between 2015 and 2020, and determined the discordance between the results of IHC and ISH. Along with their own study, they also performed pooled literature review analysis to see the results of relevant studies. They found that 1.6% cases had IHC-/ISH+ and 11.9% cases had IHC+/ISH- discordance in their own study. Also according to detailed literature analysis, they performed, the discordances (IHC+/ISH- and IHC-/ISH+) between two methods observed. with an IHC+/ISH− discordance was considerably higher than IHC−/ISH+ (13.8% vs. 3%, $P < .0001$) [92]. Due to the costly tests done at molecular-genetic level, immunohistochemical (IHC) markers are widely used as they are easily available and cost-effective. For this reason, subtyping of breast cancer by IHC is often practiced. Prior studies have shown that subtyping of breast cancer is crucial in prognosis and plays important role in endocrine therapy and chemotherapy. However, IHC-Luminal A was not significantly predictive of Pathologic Complete Response (pCR) [93]. Pathologic complete response

(pCR) is the complete absence of invasive or in situ cancer in the breast and axillary lymph nodes [32]. pCR is predictive of the outcomes of neoadjuvant treatment for breast cancer. Recently in 2020, there have been attempts to determine a patient's receptor status by using their Gene Expression Profiles [8]. The method identified predictor genes and performed receptor-status prediction using logistic regression based on those predictor genes. The method had a higher concordance with the intrinsic subtypes than the IHC-based method, with 5%-12% of the cases having predicted status different from the IHC-based status. It also showed that the patients with a mismatch between the predicted status and the IHC-based status had an overall lower survival rate and concluded that it provides a more reliable classification than the IHC method.

1.8 Need for Receptor Status Prediction Using DNA Methylation

Together with grade and stage, immunohistochemistry (IHC) subtypes are well-known independent prognostic markers of breast cancer in women. Testing of ER/PR/HER2 status of a tumour is usually done using a method known as IHC (immunohistochemistry). Immunohistochemistry (IHC) tumour markers such as estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and Ki67, are independent prognostic markers for breast cancer. Laboratory dependent factors such as choice of the specimen, choice of antibody and threshold for positivity can lead to different results for the same patient [9]. Through randomized controlled experiments it was observed that immunohistochemistry (IHC) resulted in discordant results when testing for hormone receptors (ER/PR) and HER2 expression in breast tumours. With HER2, 14.5% of cases were false-positive while 21.4% of false-negative cases were detected in the phase III trial [90]. This suggests that there are quality issues with IHC: (i) it requires stringent quality control; (ii) longer cold ischemic time and inappropriate time for fixation may result in false-negative results plausibly due to their detrimental effect on proteins; and (iii) HER2-positivity rate is affected by histological grade, age, HR and nodal status [91]. The accuracy of the results is essential because they can affect the patient's treatment plan and recurrence chance. Better accuracy of ER/PR status prediction can also help avoid side effects and potentially life-threatening toxicity that may occur due to treatments that might not even work for the patient. As explained in section 1.7, a recent study [8] has shown that Gene Expression Profiles might be better at predicting Hormone Receptor Status, which in turn showed that patients with different results from the IHC method had overall lower chances of survival. Other studies have indicated a strong correlation of DNA methylation profiles with breast cancer patients' ER and PR status [55]. Pyrosequencing methylation analysis was performed to investigate the methylation patterns of 12 tumour suppressor genes in 90 pairs of malignant and normal breast tissue. They found that methylation patterns of HIN-1, RASSF1A, RIL and CDH13 genes were strongly correlated to the ER and PR status. Building a machine learning-based ER/PR/HER2 prediction model can help in more accurate receptor status prediction, leading to better treatment and prognosis for patients.

1.9 Thesis overview

In this thesis, we have used a Machine Learning method to predict the Receptor Status (ER, PR and HER2) of Breast Cancer patients using DNA methylation data. In Chapter 2, we have briefly discussed the datasets used and data preprocessing, feature selection and prediction models that have been used for our study. We further discussed the pipeline used to predict the Receptor Status of Breast Cancer patients using DNA methylation data. In Chapter 3, we investigated the differences in DNA methylation profiles of breast cancer patients with different receptor statuses and the performance of our machine learning algorithm in predicting the receptor status, followed by a discussion. In Chapter 4, we have concluded the main points and contributions of our study and possible ideas that could be worked upon in the future.

Chapter 2

Methodology

2.1 Datasets used

We have used the TCGA-BRCA dataset [21] and four additional datasets from the NCBI GEO portal [22-24] consisting of DNA Methylation Data from the Infinium Hypermethylation 450K platform. The four other datasets were shortlisted from a list of more than 200 datasets based on the following criteria :

- The platform used should be Infinium Hypermethylation 450K platform.
- The data should be present in the IDAT format.
- The dataset is related to Breast Cancer.
- There should be at least ten samples in the dataset.
- The receptor status should be defined by IHC method.

The number of Breast Cancer samples total to 1514. All the data was downloaded in IDAT format due to the presence of control probes, as it helps in the preprocessing and Quality Control of the data. After downloading the data, it was preprocessed and normalised for further analysis, as explained in subsequent sections. Table 2.1 summarises the total number of samples and the count of patients with their ER positive/negative status, PR positive/negative status and HER2 positive/negative status. In the process of conducting 10-Fold Cross Validation, all available datasets were used. However, when it came to independent dataset testing, the GSE72251 was specifically used for that purpose, while the remaining datasets contributed to the training process. The table also mentions whether it contains information about the overall survival of the patients and PAM50 subtype classification.

The results from the IHC method depend on several laboratory dependent factors such as choice of the specimen, choice of the antibody, the threshold for positivity etc., as explained in Section 1.8 and could lead to different results from the same patient. This could mean that some of the samples might have misclassified receptor status. There has been a reported discordance rate of about 20%–50% between IHC-based clinical subtypes and intrinsic subtypes [25-27].

Thus, the Hormone Receptor status data labels were considered noisy labelled data for further downstream analysis.

Table 2.1 Summary of the datasets shortlisted for further analysis.

Dataset ID	Total case samples	ER +ve/-ve	PR +ve/-ve	HER2 +ve/-ve	Overall Survival	PAM50 subtype
GSE72245	118	64/53	0/0	30/88	Available	Available
GSE72251	119	70/49	54/65	25/94	Available	N/A
GSE84207	330	220/59	0/0	0/0	N/A	Available
GSE117439	52	36/16	0/0	0/0	N/A	N/A
TCGA-BRCA	895	650/188	699/344	164/564	Available	Available
Total	1514	1040/365	753/409	219/746	-	
Total number of samples after QC	1261	923/336	526/291	141/561	-	

A brief description of the datasets used is explained below.

- **GSE72245 and GSE72251:** The dataset consists of 118 and 119 samples, respectively, where fresh breast tumour tissues were collected immediately after the surgery. The DNA methylation was analysed on Infinium HumanMethylation450K bead arrays. The mean age of patients for the three datasets were 56 and 58, respectively. The histologic grade of the tumour was G1-G2 for 34 and 33 patients, respectively, while it was G3 for 84 and 85 patients, respectively. The IHC subtype for GSE72245 and GSE72251 datasets were HER2 for 30 and 25 patients, Luminal A for 25 and 27 patients, Luminal B for 32 and 31 patients and Basal-like for 31 and 34 patients, respectively.

- **GSE84207:** Using the Infinium HumanMethylation450K technology, the cohort was utilised to assess the DNA methylation levels of 330 patient tumours. The PAM50 subtype was Luminal A for 120 patients, Luminal B for 63 patients, Normal for 18 patients, HER2 for 37 patients and Basal-like for 34 patients. The PAM50 subtype was unknown for the rest of the patients.

- **GSE117439:** The dataset consists of 46-paired tumours. Primary tumours also referred to

as first tumours, were matched to a second tumour from the same woman. Twelve tumour pairs were from women with ER +ve first and second tumours. Five tumour pairs were from women with ER -ve first and second tumours. Six tumour pairs were from women with ER +ve first and ER -ve second tumours. Additionally, six ER +ve tumour samples were obtained from women who had no recurrence of BC during a seven-year follow-up period. The DNA methylation levels of the samples were determined using the Infinium HumanMethylation450K platform.

- **TCGA-BRCA:** In 2012, TCGA launched a Pan-Cancer analysis project aimed to gain knowledge about different cancer types. This included the BRCA project that was aimed at Breast tissue-related tumours. The dataset consists of 1098 cases and consists of various data such as DNA methylation data, copy number variation data, clinical data, sequencing reads, etc. Out of these 1098 cases, 895 had DNA methylation data of breast cancer tissue in IDAT format from the Illumina Humanmethylation 450K platform. The samples had to meet specific quality standards that include:
  - A primary untreated tumour along with normal tissue/blood sample from a matching source.
  - Frozen and sufficiently sized surgically removed samples.
  - A threshold of at least 80% tumour nuclei in the sample.

We also incorporated The Cancer Genome Atlas - Breast Invasive Carcinoma (TCGA-BRCA) RNA-Seq Gene Expression Data in our research to establish a link between DNA methylation and gene expression. The data, which was generated using the Illumina HiSeq platform, consists of a sizeable dataset from 1095 samples. Among these samples, we had information on ER status for 777 instances, with 599 classified as ER+ and 178 as ER-. Similarly, PR status was disclosed for 630 samples, which comprised 522 PR+ and 108 PR- instances. The HER2 status was provided for 767 samples, including 114 HER2+ and 653 HER2-. The processing of raw data for analysis was carried out using the TCGAbiolinks package [17, 107].

2.2 Preprocessing

Before using our data for any downstream analysis, all datasets were processed using the PyMethylProcess tool for quality control, preprocessing, and normalisation [61]. The PyMethylProcess tool helps to preprocess methylation IDAT files without the need to use R. The programme includes three distinct libraries for data quality control and normalisation, notably minfi, meffil, and ENmix [62-64]. We have used the meffil pipeline as it uses significantly less computation memory and processing time than the minfi pipeline and can handle larger datasets comprising thousands of samples. The meffil tool uses Functional Normalisation technique to normalize DNA methylation data. Functional Normalisation separates biological variation from technical variation using information from control probes [34]. The pipeline of the tool is mainly divided into five steps which are explained below:

- Quality Control
- Normalisation
- Data Imputation
- Non autosomal sites removal
- Feature selection

Microarray analysis is sensitive to variations in experimental conditions. It faces the issue of batch effects, especially in larger datasets, since it is not possible to process all the samples simultaneously and by the same technical personnel. DNA methylation is determined by treating the DNA with bisulfite, which deaminates and converts unmethylated cytosine to uracil. This process is carried out in batches and may introduce technical bias in one of two ways: either all unmethylated cytosines are not converted to uracil, or methylated cytosines are converted to uracil if the bisulfite conversion is not adequately regulated. This leads to unwanted variations in the data that can increase false positive and false negative rates. DNA collected from different samples such as white cells, blood spots, peripheral blood lymphocytes, and whole blood would also have slight differences in their methylation patterns.

To tackle such problems, quality control and normalisation of data are required. The quality control is performed based on sex outliers, methylated/unmethylated ratio, control probes mean, detection scores, etc. The meffil tool identifies sex outliers based on the difference between the total medial intensity between X and Y chromosome probes. Based on this, the meffil tool identified 7 sex detection outliers. 6 of them were from GSE117439 dataset, while one was from TCGA-BRCA dataset. Further quality control was performed by plotting the median methylation intensity against the median unmethylation intensity. The meffil tool considers samples as outliers when the median methylated signal is more than 3 standard deviations. There were a total of 16 samples, 15 of them were from TCGA-BRCA dataset and 1 was from GSE84207 dataset. The 450k array consists of control probes which can be used to judge the quality of processing steps such as staining, extension, hybridization, bisulfate conversion etc. These control probes are grouped in 42 categories of control types. There were 18 samples that didn't pass the control probe quality control, out of which 1 was from the GSE117439 dataset while the rest were from TCGA-BRCA dataset. Further quality control was performed by calculating the proportion of probes that did not pass the detection p-value and the meffil tool identified 34 samples with a high proportion of undetected probes. 1 of the sample was from GSE84207 dataset, 3 samples were from the GSE117439 dataset, while the rest 30 were from the TCGA-BRCA dataset. Further there were 0 samples with a high proportion of probes with low bead number and thus no further samples were removed. The meffil tool also identified 15755 probes with only background signal in high proportion of samples and thus they were removed from further analysis. The tool also removed 181 CpGs with low bead numbers in high proportion of samples.

Illumina Humanmethylation 450K and MethylationEPIC 850K arrays include control probes that vary due to technical variation and remain unaffected by biological variations. Functional Normalisation uses these control probes to distinguish between technical variations and biological

variations and is thus used by the meffil library.

The methylation level of a CpG site is represented by β-value, calculated by the following formula: $\beta=M/(M+U+\alpha)$. Here, M represents the methylated intensity of the CpG site, U represents the unmethylated intensity of the CpG site, and α is an offset that is generally set as 100 that helps stabilize beta values for small M and U values. CpG sites for which more than 50% of the samples had missing data were removed from further analysis. The missing β-values in the remaining data are imputed by the PyMethylProcess tool using the k-Nearest Neighbors (KNN) method.

The PyMethylProcess tool further removes sex probes to keep only autosomal sites for further analysis. This was done to remove any features dependent on sex so that the pipeline remains consistent irrescpective of the gender of the patient. CpG sites lying in the SNP regions were further removed. Finally, Feature Selection was performed using the PyMethylProcess tool to get the top 3,00,000 CpG sites with the highest amount of mean absolute deviation from the mean methylation from the set of CpGs. Figure 2.1 shows the t-SNE plot for the samples used in the final analysis after normalisation. Note the separate blue cluster of the GSE117439 dataset, this is due to the fact that GSE117439 dataset had both primary and secondary tumours.

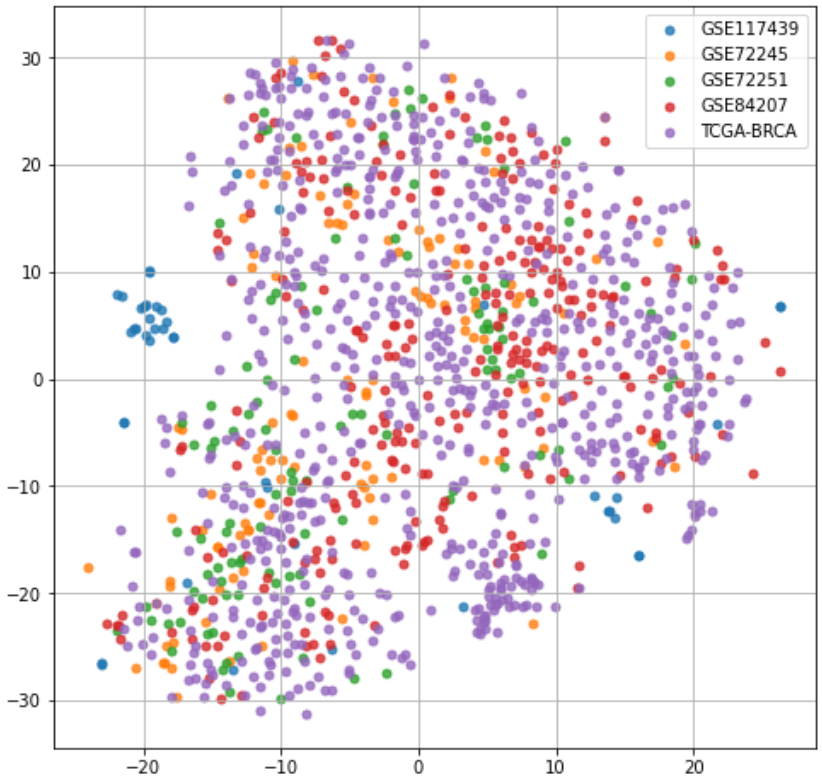


Figure 2.1 t-SNE plot for the cancer samples used in the final analysis after normalisation.

2.3 Dimensionality Reduction

A lot of data in biology, such as high-throughput data, consists of a high number of dimensions where the number of dimensions in the data is significantly greater than the number of samples available. Our study consists of more than 4,80,000 dimensions for CpG sites, while the number of patients in the study ranges from around 750-1300. This poses a lot of problems for any machine learning model. If not handled properly, the high dimensional data can significantly increase the time of the execution for the algorithm and negatively impact its performance. Increasing the number of features helps with the classifier's performance up to a certain limit, after which further increasing the features would decrease the classifier's performance. One reason is because increasing the number of features may result in an increase in noise, since not all characteristics are useful for categorization. Another issue is that increasing the number of features may result in overfitting the data, since the density of data points decreases exponentially as the number of features increases [65]. Due to this sparsity, it becomes much simpler to identify a separable hyperplane, since the probability of a training sample being on the incorrect side of the best hyperplane decreases as the number of features increases. This causes overfitting of the samples and leads to poorer results. The next section discusses some of the dimensionality reduction methods.

2.3.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is an unsupervised learning method that is used to reduce the dimension of high-dimensional data and to visualise it [13]. The primary objective of PCA is to minimise the number of dimensions by transforming the original dimensions to a new set of dimensions (known as a principal component) while maintaining as much information as possible from the original dataset. It tries to preserve the global structure of data while reducing the dimensions, which could lead to a loss of local structures. The first principal component indicates in the direction of the projected data with the greatest variation. The ith principal component may be defined as an orthogonal path to the first i – 1 principal component that maximises the variance of the projected data. Figure 2.2 explains the first two principal components of data.

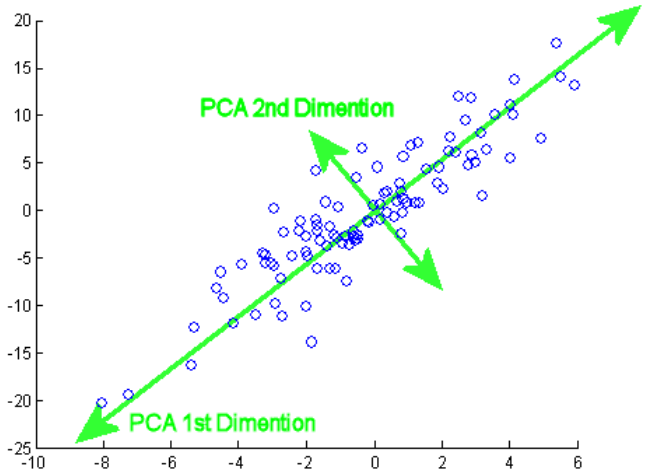


Figure 2.2 The first principal component of the data has the most variance while the second principal component of the data is orthogonal to the first component and has the most variance after the first component. ⟨Reproduced from: https://weigend.com/files/teaching/stanford/2008/stanford2008.wikispaces.com/file/view/pca_example.gif⟩

2.3.2 t-distributed stochastic neighbour embedding (t-SNE)

The t-distributed stochastic neighbour embedding (t-SNE) method is an unsupervised learning technique that is used to reduce the dimension of high-dimensional data and to visualise it [60]. The goal of t-SNE is to decrease data from a higher dimension to a lower dimension while preserving the data's neighbourhood. Unlike PCA, this method attempts to retain the data's local structure. Additional distinctions between t-SNE and PCA include the fact that it is a non-linear dimensionality reduction method and a non-deterministic algorithm (same inputs for the algorithm could lead to different outcomes on different runs).

2.4 Classification Algorithms in Machine Learning

Machine Learning is the study of how computers may learn to execute particular tasks from examples without being specifically programmed to do so. Classification algorithms try to predict unseen data’s class labels after learning discriminating patterns of different classes in the training phase. Some of the common classification algorithms in Machine Learning are briefly explained below.

2.4.1 Support Vector Machines (SVM)

SVM is a supervised machine learning algorithm that outputs an optimal hyperplane, using labelled training data to classify new samples [57]. In two-dimensional feature space, the hyperplane is a line; in three-dimensional feature space, the hyperplane is a plane, and so on. The method makes no assumptions about the data and instead attempts to fit an optimum hyperplane that separates the various groups with the greatest margin. As shown in Figure 2.3, the blue coloured samples are represented as label 1, and green coloured samples are represented as label -1.

The hyperplane can be written as the set of points $\vec{x}$ satisfying the following equation:

$$\vec{w} \,.\, \vec{x} = b \tag{2.1}$$

Here, $\vec{w}$ is a normal vector to the hyperplane and b is the negative of the intercept of the hyperplane. All the samples that lie on or above the following equation are classified into one class, represented by label 1.

$$\vec{w} \,.\, \vec{x} - b = 1 \tag{2.2}$$

All the samples that lie on or above the following equation are classified into the other class, represented by label -1.

$$\vec{w} \,.\, \vec{x} - b = -1 \tag{2.3}$$

The explained idea works fine if the class labels are linearly separable. If it's not the case, this linear boundary classifier can be converted into a non-linear boundary classifier with the help of the kernel trick. The main idea of kernels is that a non-linearly separable dataset might become linearly separable when it is projected in a higher dimension.

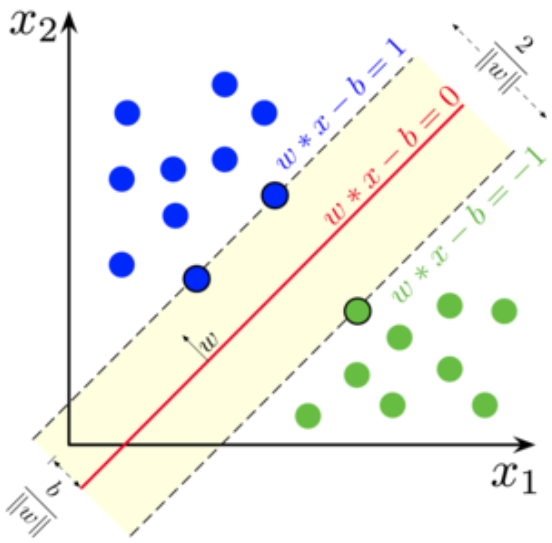


Figure 2.3 The SVM classifier outputs a solid line separating the two classes (red and blue) into two-dimensional space. The dotted lines are the functional margins, and samples lying on them are called support vectors. ⟨Reproduced from: https://en.wikipedia.org/wiki/Support-vector_machine⟩

2.4.2 Multi-Layer Perceptron (MLP)

A perceptron, also referred to as a neuron in Machine Learning is a linear classifier. A linear classifier is a classification algorithm that uses a linear predictor function to make predictions. A perceptron is the basis of all Artificial Neural Networks (ANN), which is somewhat inspired by a biological brain.

Figure 2.4 shows a model of the perceptron. The output is computed as:

$$Y = F\left(\sum_{i=1}^{m} w_i x_i + b\right) \tag{2.4}$$

Here F is a non-linear activation function, x is an m-dimensional input vector, w is the m-dimensional weight vector, b is the bias, and Y is the computed output.

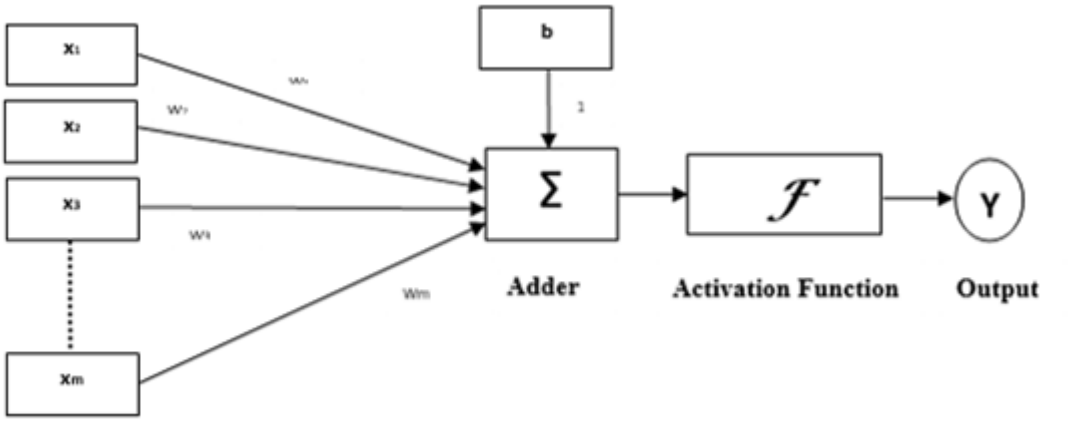


Figure 2.4 Model of a perceptron ⟨Reproduced from: https://www.tutorialspoint.com/artificial_neural_network/artificial_neural_network_supervised_learning.htm⟩

A perceptron has a limitation of only linear classification. This problem is overcome by multilayer perceptron (MLP), which stacks multiple perceptrons in a layer-wise fashion. Every MLP consists of three main parts:

- **Input layer:** This layer has the same number of perceptrons as the dataset's dimension. This layer's data is sent into the first hidden layer's neurons.
- **Hidden layers:** The capacity of an MLP to learn a complicated model is proportional to the number of hidden layers it contains [14]. Without the use of hidden layers, an MLP is capable of learning decision boundaries that are linearly separable. Any function that has a continuous mapping from one finite space to another may be approximated by a single hidden layer. It may represent any arbitrary decision boundary using two hidden layers. Three or more hidden layers are used to perform automatic feature engineering to learn complex representations.
- **Output layer:** This layer receives data from the hidden layer and predicts the input.

Figure 2.5 shows the architecture of an MLP. Generally, simple datasets don’t require more than two hidden layers. However, Deep Learning problems involving complex datasets such as in Computer Vision, NLP and time-series problems can have as many as 100 layers or more.

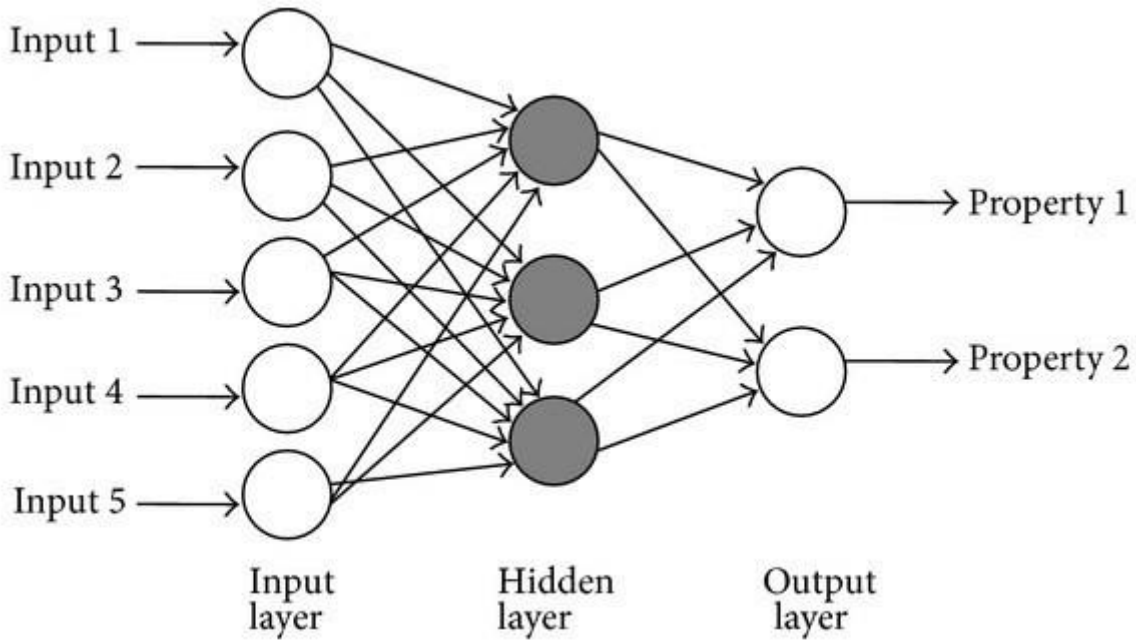


Figure 2.5 Architecture of a multilayer perceptron with a single hidden layer. ⟨Reproduced from: https://medium.com/pankajmathur/a-simple-multilayer-perceptron-with-tensorflow-3effe7bf3466⟩

2.4.3 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is one of the most straightforward supervised machine learning algorithms. KNN algorithms make the assumption that similar things lie in close proximity to each other. The algorithm is only useful when the assumption stated is true. The similarity between data points is usually determined by distance (such as Euclidean, Manhattan) between them. The main drawback of the KNN algorithm is that it's not scalable as it becomes significantly slower with the increase in data.

2.5 Noisy Label Training

Machine Learning has proved to be extremely helpful in providing meaningful information from a vast raw data set. It is known to have an impressive performance in various tasks such as language processing, image recognition and medical diagnosis. The success of these models is mainly dependent on a large amount of data that is correctly labelled. Always determining the correct label is sometimes infeasible and could also be expensive and time-consuming. In some cases, data labelling can be highly complex, even for Subject Matter Experts (SMEs) [18]. Such problems can lead to unreliable labels, also known as noisy labels, as they may differ from the ground truth. These noisy labels can affect the training of the machine learning model and can lead

to a drop in their prediction performance.

The receptor status of the samples are noisy labels as the labels determined in the clinical data could be wrong due to a variety of reasons, including laboratory dependent factors as explained in Section 1.8. To tackle this issue, we have used the Cleanlab tool [19] that uses the current state-of-the-art method for machine learning with noisy labels for binary classification [20]. The method, known as “Rank Pruning for Robust Classification with Noisy Labels” is the first time-efficient algorithm that achieves similar or better results than the previous state-of-the-art methods. The algorithm emphasizes the use of learning with confident examples than learning with more examples for training. Here confident examples mean the examples which have a very high probability of having the correct label. The algorithm uses a probabilistic classifier to estimate the number of samples that could be incorrectly labelled for each class. After removing samples with anticipated noisy labels from the training set, only the confident instances are utilised to train the machine learning model.

Previous research [15, 16] have used the Cleanlab technology effectively, demonstrating its potential usefulness in comparable settings. In one study [15], the Cleanlab tool was used to address the issue of noisy labelling in whole slide images when applying deep learning algorithms to classify brain tumours. The researchers discovered that integrating Cleanlab enhanced the classification model's accuracy, emphasising its utility in dealing with noisy labels in the context of histopathological investigation. Another important study [16] concentrated on using telemetry data to detect atrial fibrillation without considerable manual annotation. The researchers tested Cleanlab to other approaches and discovered that it performed similarly, suggesting its resilience and dependability in resolving the issues provided by noisy labels in biological datasets.

Given Cleanlab's effectiveness in prior studies involving varied biological datasets, its use in the current research on receptor status prediction in breast cancer is promising. It is hoped that by using Cleanlab's capabilities, the negative impact of noisy labels on prediction accuracy can be reduced, ultimately boosting the reliability and effectiveness of the receptor status prediction models for breast cancer.

2.6 Evaluation Metrics

Evaluation metrics are important for any machine learning project. It helps to quantify the quality of any machine learning or statistical model. Multiple evaluation metrics are necessary when evaluating a model, since a model may perform well on one evaluation metric but badly on another. In this thesis, we are studying receptor status classification; for all metrics, the classes ER/PR/HER-2 positive are defined as positive and classes ER/PR/HER2 negative are defined as negative. Figure 2.6 explains the terms True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) visually. Here, we take an example of an imaginary model used to predict an image as either a “Dog” or “Not Dog”.

There are four cases possible:

- The picture of a cat, correctly labelled as Not-dog, is a True Negative.
- The picture of a man with a dog filter incorrectly labelled as a Dog, is a False Positive.
- The picture of a dog wearing a helmet, incorrectly labelled as Not dog, is a False Negative.
- The picture of a dog, correctly labelled as Dog, is a True Positive.

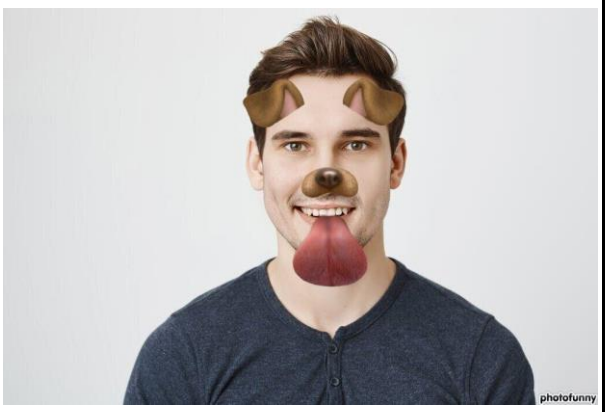
True Negative **Prediction: Not dog** 	**False Positive** **Prediction: Dog** 
False Negative **Prediction: Not dog** 	**True Positive** **Prediction: Dog** 

Figure 2.6 An example where a model classifies an image as either “Dog” or “Not Dog”. (This example has been designed using resources from Freepik.com).

Here are the Evaluation metrics we have used in our study:

2.6.1 Accuracy

The formula for accuracy is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.5}$$

But the Accuracy metric alone can be misleading. Consider an example of a model to identify terrorists with over 99.99% accuracy. But the model still might not be useful as it is an imbalanced classification problem, with the number of terrorists significantly lower than non-terrorists. The model could never identify a single terrorist and still could have good accuracy.

2.6.2 Precision

The formula for precision is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{2.6}$$

A model's precision indicates how many of the identified items are genuinely meaningful.

It is computed by dividing true positives by total positives. In the example shown in Figure 2.5, the precision percentage can be interpreted as the probability that an image which the model detected as a dog actually shows dog.

2.6.3 Recall

The formula for recall is as follows:

$$Recall = \frac{TP}{TP + FN} \tag{2.7}$$

Recall, also known as sensitivity, measures the proportion of actual positive instances correctly identified by a model. It is calculated by dividing true positives by the sum of true positives and false negatives. A higher recall value indicates a greater ability to capture all relevant items of a certain class. It is a crucial metric in scenarios where identifying all positives is important, such as medical diagnoses or quality control.

2.6.4 Matthews Correlation Coefficient (MCC)

The formula for MCC is as follows:

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.8}$$

An effective statistic for assessing binary classification models in machine learning is the Matthews Correlation Coefficient (MCC). When dealing with imbalanced datasets, the MCC gives a balanced measure even when the classes are of extremely different sizes, making it a more reliable measure. The MCC returns a value between -1 and +1 after accounting for both over- and under-predictions. A MCC of +1 denotes a perfect prediction, a value of 0 is equivalent to a random guess, and a value of -1 denotes complete discrepancy between the prediction and the observation. As a result, we selected the MCC to evaluate how well our models performed in predicting the hormone receptor status of breast cancer.

2.6.5 Confusion Matrix

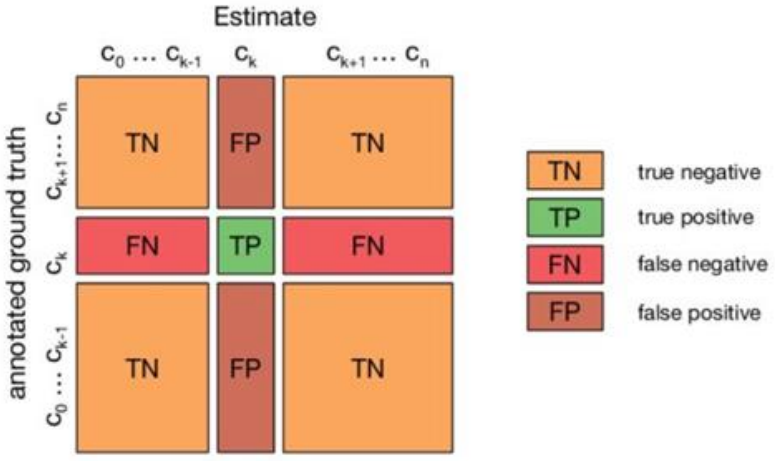


Figure 2.7 Confusion matrix for n-class classification. When considering the class k ($0 \leq k \leq n$), four distinct classification results may be achieved (Reproduced from [58]).

A confusion matrix is a table that is used to visualise and summarize the performance of a classification algorithm. It is an important evaluation metric that gives an insight into the type of errors made by the classification algorithm. In this thesis, each row of the confusion matrix represents the known label from the dataset, and each column represents the predicted label by the model. Figure 2.7 shows how a confusion matrix is divided into TP, FP, TN and FN.

2.7 Method pipeline for Receptor status prediction

Our study to predict the receptor status in breast cancer patients was performed individually for each receptor status. Figure 2.8 shows the overall workflow of the study. The pipeline is explained in detail in the following sections.

2.7.1 Preprocessing

The dataset used for predicting receptor status in breast cancer is compiled from several laboratories, which introduces background noise into the data. It is critical to do quality control, preprocessing, and data normalisation, as discussed in section 2.2, to ensure the reliability and correctness of the future analysis. A well-established pipeline based on PyMethylProcess's meffil module is used to do this. This pipeline, which is thoroughly described in section 2.2, includes a series of necessary actions to prepare the data for further analysis.

The data comprises of β-values for each CpG site. In cases where there are missing fields in the data, the PyMethylProcess programme imputes these values, delivering a complete dataset for enhanced downstream analysis. The noisy background inherent in data from different laboratories can be efficiently addressed by utilising PyMethylProcess's meffil-based pipeline. This preprocessing strategy is critical for optimising the dataset's quality and consistency, allowing for reliable and robust receptor status prediction models for breast cancer.

2.7.2 Prediction Using Differentially Methylated CpG sites

Given that the features in our dataset far outnumber the available samples, we needed to use a dimensionality reduction strategy, as described in section 2.3. To accomplish this, we identified CpG sites with a delta beta value greater than 0.2 and a multiple comparison corrected p-value less than 0.05 as characteristics for further investigation between malignant and normal samples. This method reduced the dataset to 19834 CpG locations. We further refined this selection by selecting only CpG sites in gene promoter regions, resulting in a total of 617 CpG sites. Following that, we estimated the delta beta values for each of these sites separately, taking into account the relevant subsets of ER+/ER-, PR+/PR-, and HER2+/HER2- samples.

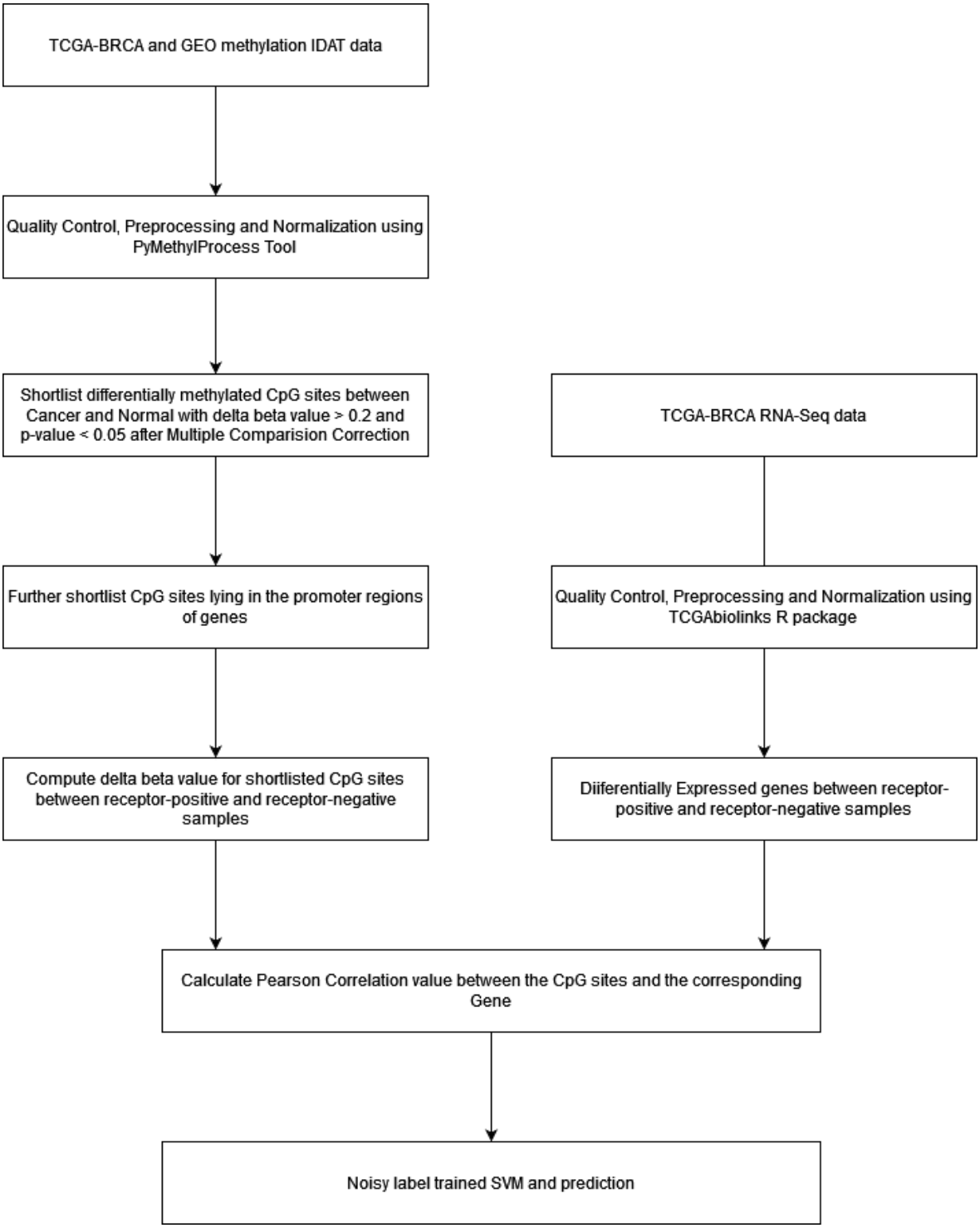


Figure 2.8 Workflow of CpG site selection, DEG selection, model training, receptor status prediction and analysis.

We used the TCGAanalyze_DEA function of the TCGAbiolinks tool to determine differentially expressed genes across receptor-positive and receptor-negative samples. This method generated a list of differentially expressed genes that included the mean expression values for both receptor-positive and receptor-negative samples. It also calculated $\log_2$(Fold Change), which is $\log_2$(mean(HR+)/mean(HR-)), and Delta, which is $\log_2$FC*(mean(ER+) - mean(ER-)). This approach found 2118 genes that differed between ER+ and ER- samples, 1271 genes that differed between PR+ and PR- samples, and 1066 genes that differed between HER2+ and HER2- samples.

Following that, we calculated the Pearson correlation between the selected CpG sites and their corresponding differentially expressed genes, if they were present. Our aim was to find the correlation values between the two. CpG sites with an absolute Pearson Correlation value greater than 0.4 were considered relevant for ER and PR. When multiple CpG sites were found within the promoter region of the same gene, the site with the highest correlation value was chosen. For HER2, a lower cutoff of 0.3 was used because a higher cutoff yielded insufficient CpG sites to accurately predict receptor status. As a result, we found 45 CpG sites for ER, 54 for PR, and 51 for HER2. Further discussion of these sites is done in section 3.2.1 and 3.3.

The final stage in our study was to use the CpG sites we had chosen as features in our Support Vector Machine (SVM) model. To partition the data, we used the Stratified 10-fold Cross-validation approach, with 90% put aside for training and 10% set away for validation. This stratification ensures that the class ratio remains constant over the K folds, replicating the ratio in the original dataset. In predicting ER status, for example, where we observed a class ratio of about 2.7:1, the stratified K-fold technique maintained this ratio throughout all folds.

We trained a Noisy-Label SVM for Receptor Status prediction, utilising the Cleanlab [19] tool as a noisy-label training wrapper and the scikit-learn [56,57] tool for the SVM model. The grid search approach was used to find the best hyperparameters while maximising the Matthews Correlation Coefficient (MCC) score. The kernel, degree, gamma and class weight was optimized using the grid search method for each receptor prediction model. The Radial Basis Function (RBF) Kernel was used in all receptors in our investigation.

Following modelling, we created Kaplan-Meier survival graphs [59] using the lifelines programme [10] to compare survival probabilities. The following chapter of the thesis elaborates on this procedure and the subsequent results.

Chapter 3

Results and Discussions

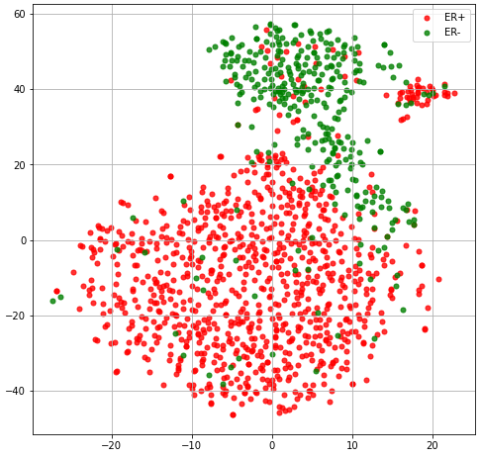
As explained previously in sections 1.7 and 1.8, the IHC method is commonly used to determine the receptor status of BC patients but suffers from lower accuracy due to a number of laboratory dependent factors. Determining the receptor status of BC patients with high accuracy is crucial as they help in determining the treatment plan and chance of recurrence for the patients. Recent studies have shown that Gene Expression-based Profiling can provide better accuracy for receptor status prediction. There have also been previous studies that show a strong correlation of differential methylation patterns with ER and PR status of patients [55]. Thus, it makes a strong case to study whether DNA methylation patterns can be used to predict the Receptor Status of BC patients and whether they would be beneficial over traditional IHC-based methods.

The pipeline to predict the Receptor status of patients from DNA methylation data is explained in detail in section 2.7. After preprocessing the data using the PyMethyProcess tool, we performed dimensionality reduction using a number of steps explained in section 2.7. We used these shortlisted CpG sites to train a noisy-label SVM model for receptor status prediction. In this chapter, we present our findings and observations from the research conducted for this thesis. In section 3.1, we reviewed the DNA methylation patterns of the data in relation to their Receptor status. We have further discussed the prediction results of our model with given label and intrinsic subtypes in section 3.2 and how the prognosis is affected for patients with misclassified labels between our ML model and the given label, as compared to the patients with the same classification. The differential methylation landscape between Receptor-positive and Receptor-negative samples is then discussed in section 3.3.

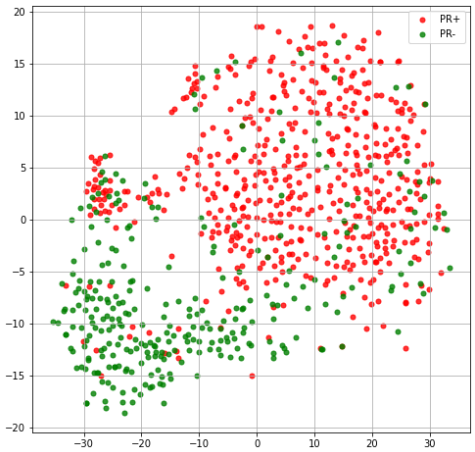
3.1 Macroscopic Landscape of DNA Methylation Patterns

To visualize the Receptor-positive and Receptor-negative groups' DNA methylation data in a lower dimension, we used t-SNE plots. t-SNE is an unsupervised dimensionality reduction algorithm that tries to reduce the dimensions of the data while preserving its neighbourhood. Figure 3.1 shows the t-SNE plots generated using the shortlisted methylated CpG positions. ER and PR show their positive and negative groups are clustered together without clear separation, but HER2 didn't have

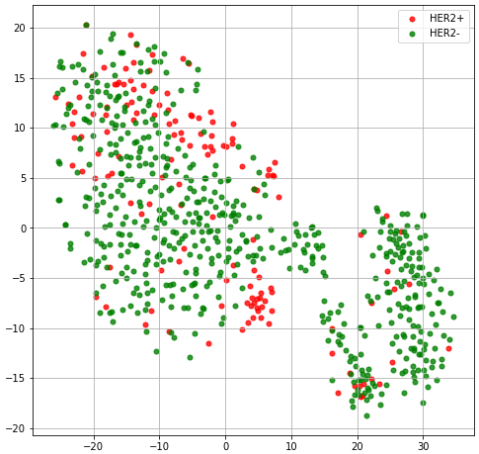
a very clear separation.



A



B



C

Figure 3.1 t-SNE plots between Receptor-positive and Receptor-negative samples (A) Estrogen Receptor, (B) Progesterone Receptor, (C) HER2 using the feature selected CpG positions.

However, there is an intermixing of some samples. Some of the intermixed samples could be the ones potentially misclassified by the traditional methods, while others could be due to the t-SNE's limited ability to preserve the neighborhood. The data is thus ideal for a machine-learning-based classification model since the groups are clustered together.

3.2 Analysis of Results for Receptor Status Prediction

3.2.1 Genes of CpG sites used as features for prediction

We evaluated the genes corresponding to the nominated CpG sites across all three receptors to gain a better understanding of the features used in our predictions. The Venn diagram, as shown in Figure 3.2, displays gene overlap between distinct receptors. We discovered 94 unique genes, with 12 of them shared by all three receptors. In addition, 21 genes were shared by PR and HER2, while 10 genes were shared by ER and PR. However, only one gene was discovered to be shared by ER and HER2. This study sheds light on the interconnections that exist between these receptors.

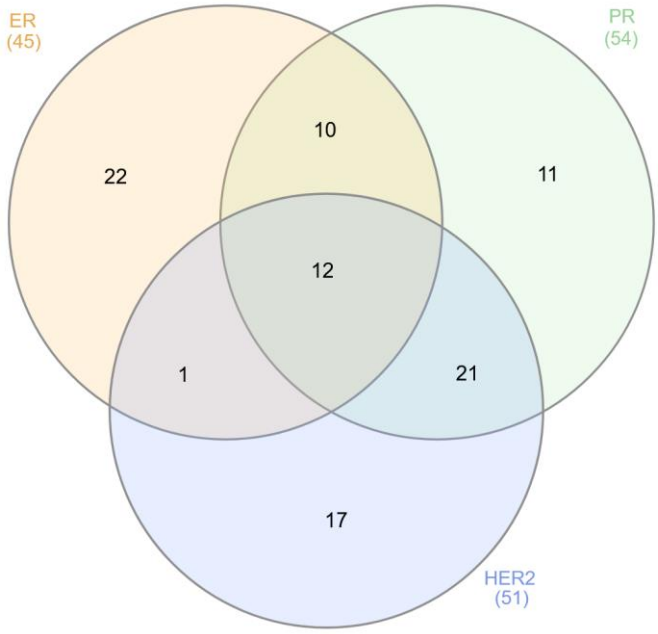


Figure 3.2 Venn Diagram Illustrating the Overlap of Genes Corresponding to Selected CpG Sites Across ER, PR, and HER2 Receptors.

We used OncoScore, a method that scores genes based on their connection with cancer in the scientific literature [76]. OncoScore is a text-mining algorithm that provides scores to genes based on the frequency with which they appear in cancer-related publications. Table A.3 in the appendix, shows the genes of CpG sites used for prediction, the OncoScore for each gene, whether a related CpG site was used for prediction for a certain receptor, the methylation status of the relevant CpG sites, and gene expression status when receptor-negative is considered normal. The bulk of the genes in the table have high OncoScores, indicating that they have previously been associated to cancer and show an inverse association between methylation and gene expression..

For example, the ERBB2 gene, which is hypomethylated and upregulated, is known to encode the HER2 receptor and has been linked to breast, ovarian, lung, and stomach cancer. Section 3.3 of the thesis conducts a more in-depth investigation of the methylation landscape.

3.2.2 Prediction Comparison Between the Given Label and ML Model

We examined a total of 1261 samples in our study. Five datasets, all of which contained DNA methylation data from breast cancer patients, were used to train and evaluate our machine learning model. We used stratified 10-fold cross-validation to assess the model. Table 3.1 shows the mean and standard deviation of the 10 folds' accuracy, precision, and recall. The table shows that the standard deviation between each fold for ER and PR is low, suggesting the model's robustness. However, for HER2, we found a significantly higher standard deviation as well as lower precision and recall scores. This is due to the relatively significant class imbalance and the low number of samples reporting HER2 status. Table 2.1 summarizes the datasets used in the research. Our evaluation metrics were accuracy, precision, and recall, with thorough explanations provided in section 2.6. Figure 3.3 shows the confusion matrix for the 10-fold cross validation method.

Table 3.2 presents the performance metrics - accuracy, precision, and recall - when employing GSE72251 as an independent testing dataset. This dataset was specifically selected as it encompasses comprehensive information about all three hormone receptors and contains 119 samples, approximately 10% of the entire dataset. For the validation on this dataset, we retrained our SVM model using the same features, however, the training dataset comprised all samples excluding those from GSE72251, whereas the testing dataset solely incorporated samples from the GSE72251 dataset. The predictions for ER, PR, and HER2 showcased commendable performance, with accuracies reaching 87%, 81%, and 93% respectively, accompanied by substantial precision and recall values, and minimal standard deviations. Figure 3.4 shows the confusion matrix for independent test dataset.

3.2.3 Kaplan Meier Survival Analysis

The treatment options of cancer patients have a significant impact on their survival probability, making it critical to examine the possible benefits of DNA-Methylation based Receptor status prediction for select patients. Using censored data, the Kaplan-Meier method, a survival analysis methodology, was used to compute the likelihood of mortality at a given period. This method uses individual survival data to estimate survival probabilities without assuming the form of the distribution. We can compare the survival probabilities of patients whose receptor statuses align with those whose statuses do not by analysing Kaplan-Meier plots. A Kaplan-Meier survival plot analysis is shown in Figure 3.5, comparing patients whose traditional and machine learning-based approach predictions agree with those whose projections differ.

Table 3.1: Performance Metrics Across 10-Fold Cross-Validation (A) ER, (B) PR, (C) HER2

(A)

	ER			
	Accuracy	Precision	Recall	MCC
Fold 1	89%	94%	90%	0.70
Fold 2	87%	90%	92%	0.64
Fold 3	83%	91%	86%	0.57
Fold 4	85%	90%	89%	0.61
Fold 5	82%	87%	88%	0.57
Fold 6	91%	95%	93%	0.74
Fold 7	85%	88%	91%	0.68
Fold 8	83%	87%	91%	0.61
Fold 9	87%	91%	91%	0.62
Fold 10	86%	92%	89%	0.64
Mean	86%	91%	90%	0.64
Std. Dev.	2.82%	2.71%	2.05%	0.05

(B)

	PR			
	Accuracy	Precision	Recall	MCC
Fold 1	88%	86%	96%	0.76
Fold 2	87%	88%	92%	0.76
Fold 3	85%	85%	94%	0.67
Fold 4	80%	80%	92%	0.58
Fold 5	89%	87%	98%	0.76
Fold 6	87%	86%	94%	0.64
Fold 7	88%	88%	94%	0.75
Fold 8	78%	78%	90%	0.61
Fold 9	88%	88%	94%	0.75
Fold 10	81%	84%	88%	0.62
Mean	85%	85%	93%	0.69
Std. Dev.	3.95%	3.46%	2.85%	0.07

(C)

	HER2			
	Accuracy	Precision	Recall	MCC
Fold 1	83%	64%	47%	0.47
Fold 2	87%	78%	50%	0.52
Fold 3	89%	80%	57%	0.61
Fold 4	90%	77%	71%	0.72
Fold 5	91%	100%	57%	0.70
Fold 6	87%	78%	50%	0.55
Fold 7	83%	60%	43%	0.43
Fold 8	87%	67%	71%	0.58
Fold 9	93%	91%	71%	0.73
Fold 10	79%	46%	43%	0.52
Mean	87%	74%	56%	0.58
Std. Dev.	4.22%	15.53%	11.39%	0.10

Table 3.2: Performance Metrics for Independent Test Dataset GSE72251

	ER	PR	HER2
Accuracy	87%	81%	93%
Precision	86%	74%	87%
Recall	94%	89%	80%
MCC	0.74	0.63	0.76

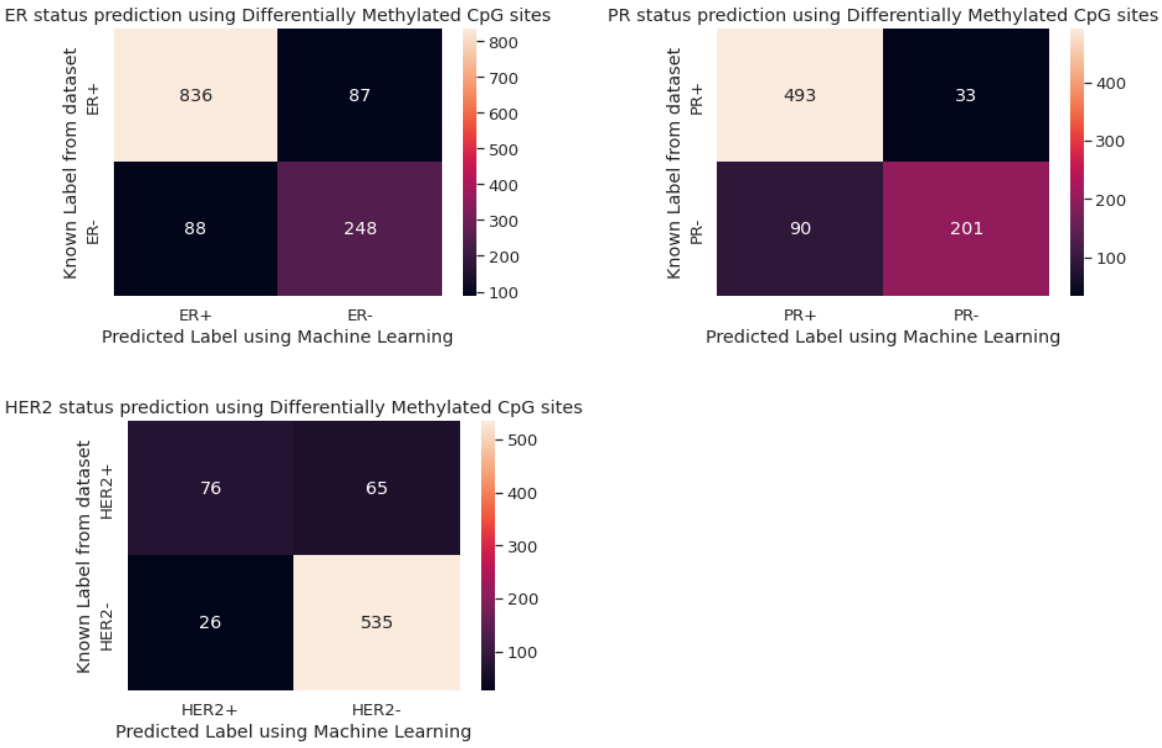


Figure 3.3: Cross-Validation Confusion Matrices for Hormone Receptor Status Prediction via 10-Fold Stratified Approach (a) ER, (b) PR, (c) HER2

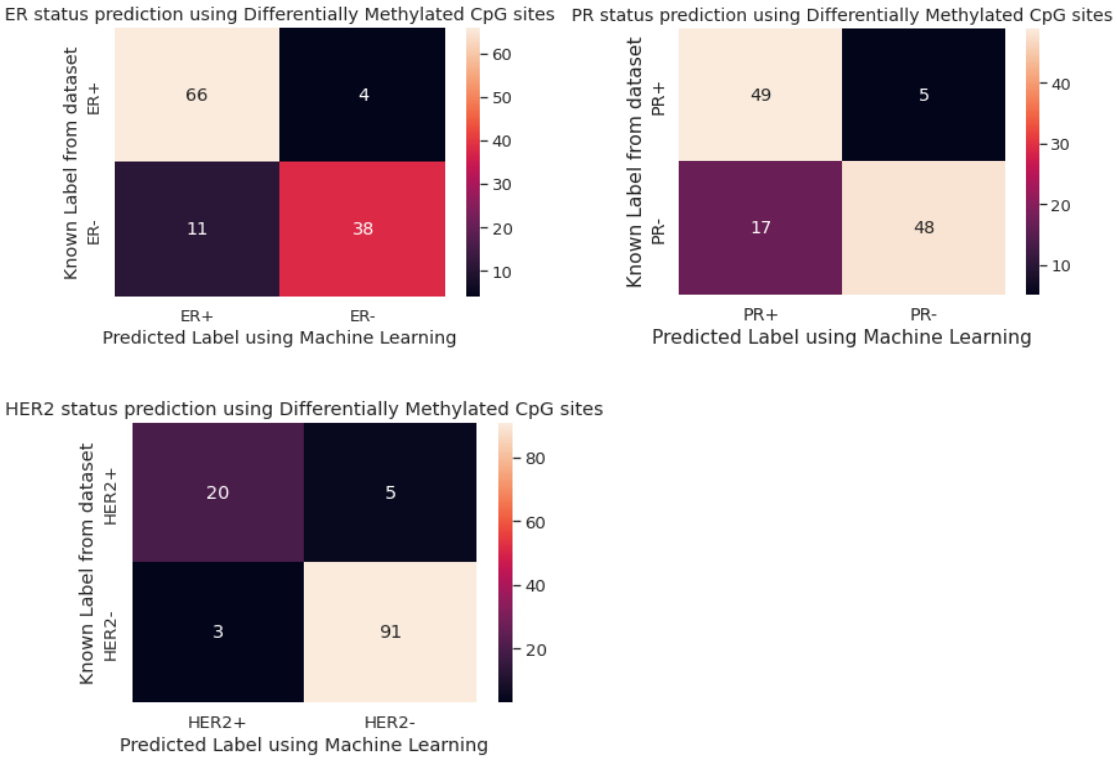
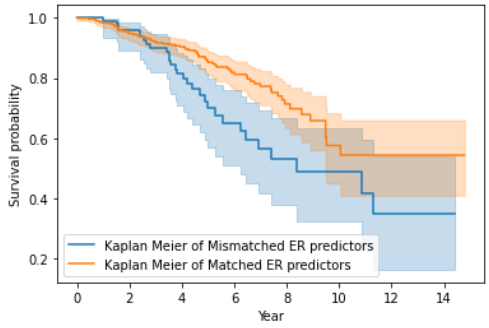
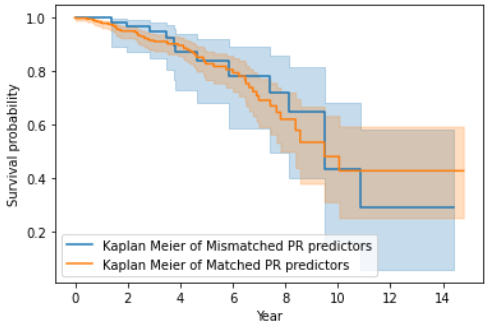


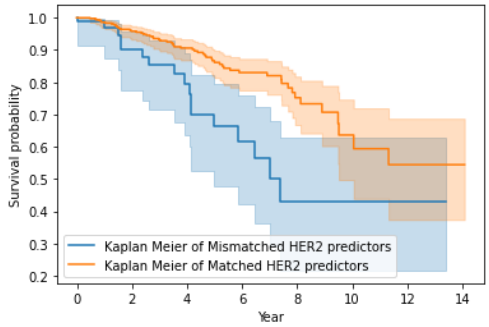
Figure 3.4: Cross-Validation Confusion Matrices for Hormone Receptor Status Prediction on independent dataset GSE72251 (a) ER, (b) PR, (c) HER2



A



B



C

Figure 3.5 Kaplan-Meier survival plots contrasting matching and non-matching Receptor status for (A) ER, (B) PR and (C) HER2.

We can see a significant difference in the survival probability of patients with mismatched ER (p-value < 0.005) and HER2 (p-value < 0.005) status up to the 6-year mark Patients with matching ER/HER2 receptor status predictions between the two approaches had a greater chance of survival than those with mismatched ER/HER2 receptor status predictions. Meanwhile, no significant change in survival probabilities was observed in PR status survival plots (p-value = 0.79 at the 6-year period).

Table 3.3 offers a detailed overview of the dataset, including the mean age, standard age deviation, and treatment types received by individuals with both matched and mismatched receptor statuses. The table consists of patients for whom clinical data was available. The columns are organised according to IHC labels. Table 3.3 shows that the mean ages of patients are remarkably stable across all receptor types, regardless of whether their receptor statuses are matched or mismatched. Out of the total patients around 5.5% had an age greater than 80 years. Despite this underlying homogeneity, when we look into the particular, surprising treatment patterns emerge. For example, in the DNA methylation-based prediction, patients with PR+ status are more likely than their counterparts to receive Hormone Therapy. Patients with PR- status, on the other hand, are more frequently exposed to Hormone Therapy in the IHC-Based Characterization group. In the case of

HER2 receptor status, a similar pattern can be observed. HER2- patients in the DNA methylation-based prediction group are more likely to receive both chemotherapy and hormone treatment. On the contrary, HER2+ individuals are more likely to receive these two types of therapy in the IHC-Based Characterization group. These findings highlight the need of proper receptor status identification because it directly determines therapeutic options used, potentially influencing patient outcomes.

Table 3.3: (A) Patient Demographics of matched and mismatched patients (B) Treatment Details for IHC-Based Characterization and DNA methylation-based prediction

(A)

	Matched						Mismatched					
	ER+	ER-	PR+	PR-	HER2+	HER2-	ER+	ER-	PR+	PR-	HER2+	HER2-
Mean age	57.94		57.65		57.65		56.80		60.33		58.70	
Std. Dev.	13.16		13.26		13.09		13.55		13.88		14.00	

(B)

	IHC-Based Characterization			DNA methylation-based prediction		
	ER+/-	PR+/-	HER2+/-	ER+/-	PR+/-	HER2+/-
Chemotherapy	281/112	245/146	51/222	262/131	261/130	25/248
Hormone Therapy	340/10	293/57	40/193	305/45	318/32	22/211
Others	44/15	34/18	16/32	44/15	40/12	23/25

3.2.4 Comparison with Intrinsic subtypes

Intrinsic subtypes of breast cancer are determined by their underlying biological traits rather than their behavioural elements, as the name implies. These intrinsic subtypes have received significant clinical attention throughout the years due to their ability to predict treatment responses and

prognosis outcomes. Positive ER and/or PR status and negative HER2 status define the Luminal A subtype. The Luminal B subtype is distinguished by positive ER and/or PR status, as well as HER2 positivity or negativity. Negative ER status and positive HER2 status distinguish the HER2-enriched subtype. The basal-like subtype is negative for ER, PR, and HER2. Finally, the negative HER2 status of the normal-like subtype defines it.

We used a comparison of the positive and negative occurrences of ER, PR, and HER2 statuses for each intrinsic subtype to assess the disparity between the intrinsic subtype and the clinical subtype, defined by ER, PR, and HER2 status. To accomplish this, we used labels from our dataset and predictions from our machine-learning model. Datasets GSE72245, GSE84207, and TCGA-BRCA provided PAM50 intrinsic subtype information, while the GSE72251 dataset provided IHC-based subtype information. We limited our analysis to samples from the TCGA-BRCA dataset for robust quality control and benchmarking.

Table 3.4 compares receptor status for each intrinsic subtype using two methods: (a) standard IHC-Based Characterization and (b) our proposed DNA methylation-based prediction. Table 3.4 shows that when employing the DNA methylation-based method rather than the traditional IHC-based approach, the discordance rates between the intrinsic and clinical subtypes (based on ER, PR, and HER2 status) are often lower. This is most noticeable in the intrinsic subtypes Luminal A, Luminal B, TNBC, and Normal-like. The DNA methylation-based prediction approach, for example, categorised samples as ER+ and PR+ in the Luminal A and Luminal B subcategories. In contrast, the conventional technique correctly recognised a higher proportion of these samples as ER- or PR-

For Luminal A, the IHC-based approach classified 23 samples as PR-, whereas the DNA methylation-based method classified only 5 samples as PR-. When comparing the two categorization systems, similar differences in the ER- status of Luminal A and PR- status of Luminal B samples were detected. The DNA methylation-based method recognised nearly all samples (with only three exceptions in ER) as negative for all three receptors (ER/PR/HER2) within the TNBC intrinsic subtype, in contrast to the old method, which classified some samples as receptor-positive. When it came to the Normal Breast-like subtype, the DNA methylation-based method predicted more samples as HER2- than the old method. Finally, the HER2-enriched subtype revealed considerable discordance between both approaches; however, given the limited sample size in comparison to other subtypes, drawing meaningful conclusions for this group is difficult.

Overall, these data indicate that DNA methylation-based prediction may provide more consistent receptor status classification based on intrinsic subtypes, thereby improving therapeutic strategy planning and patient prognosis.

Table 3.4 Receptor status for each intrinsic subtype by (a) IHC-Based Characterization and (b) DNA methylation-based prediction.

Subtype	(a) IHC-Based Characterization			(b) DNA methylation-based prediction		
	ER+/-	PR+/-	HER2+/-	ER+/-	PR+/-	HER2+/-
Luminal A	210/8	194/23	19/126	214/4	212/5	12/133
Luminal B	95/1	79/17	15/45	96/0	90/6	14/46
TNBC / Basal-like	6/61	3/64	1/45	3/64	0/67	0/46
Normal Breast-like	60/22	50/31	14/45	30/52	63/18	5/54
HER2 - enriched	4/19	4/19	13/6	21/2	2/21	12/7

3.2.5 Effects of Cleanlab Tool

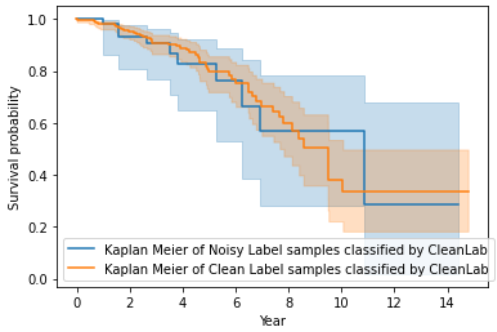
To evaluate the impact of the Cleanlab tool on our analysis, we compared the average Immunohistochemistry (IHC) scores for patients correctly and incorrectly labelled by the Cleanlab tool. Table 3.5 shows the mean IHC scores for each receptor state, along with the standard deviations. The IHC score normally varies from 0 to 4, with hormone receptor-negative samples gravitating to the lower end of the spectrum and hormone receptor-positive samples gravitating to the upper end. When we look at the table, we can see that there are significant disparities between samples with and without labelling difficulties, even when they have the same hormone receptor status as established by the IHC approach.

Consider the PR- samples: their IHC score should ideally be between 0 and 2. The mean score for samples with labelling errors, on the other hand, is 3.13 (with a low standard deviation), compared to 1.64 for samples without label discrepancies. In HER2+ samples, a similar pattern is observed: those with labelling abnormalities had a lower mean IHC score of 2.18, compared to 2.84 for those without labelling concerns. Such changes are also visible in ER+ and PR+ samples, but on a lower scale. This analysis emphasises the Cleanlab tool's possible impact on our study.

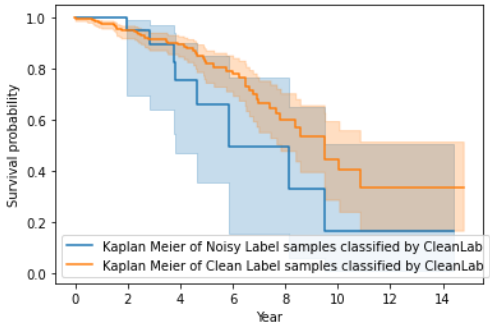
Figure 3.6 depicts Kaplan-Meier survival plots contrasting samples correctly and wrongly labelled for each receptor by Cleanlab. However, there are no significant survival differences between the two situations, making it difficult to draw solid inferences from these results.

Table 3.5 Mean and Std. Deviation (in brackets) of IHC scores for samples with label issues and samples without label issues.

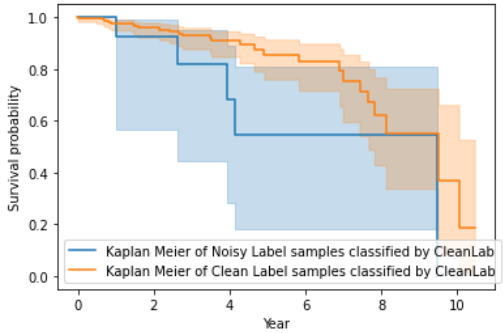
	Samples with label issues	Samples without label issues
ER+	1.87 (1.16)	2.38 (0.92)
ER-	No samples found	0.69 (0.82)
PR+	2.43(1.05)	2.83 (0.65)
PR-	3.13(0.33)	1.64 (1.11)
HER2+	2.18 (0.72)	2.84 (0.47)
HER2-	1.00 (0.00)	0.86 (0.44)



A



B



C

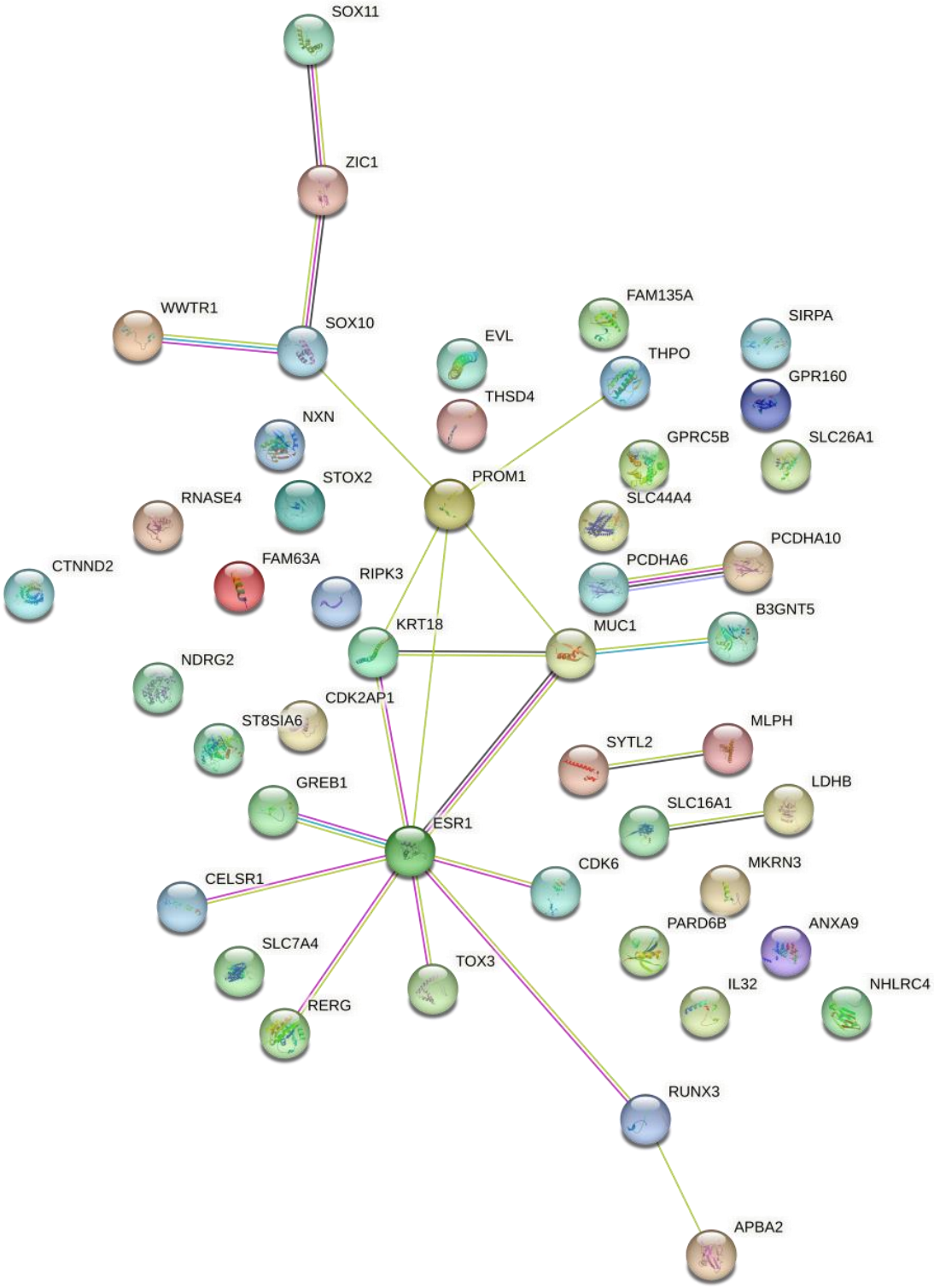
Figure 3.6 Kaplan-Meier survival plots contrasting samples labelled as correctly labelled and incorrectly labelled according to Cleanlab tool for (A) ER, (B) PR and (C) HER2.

3.3 Analysis of Genes Associated with CpG Sites Used for Prediction

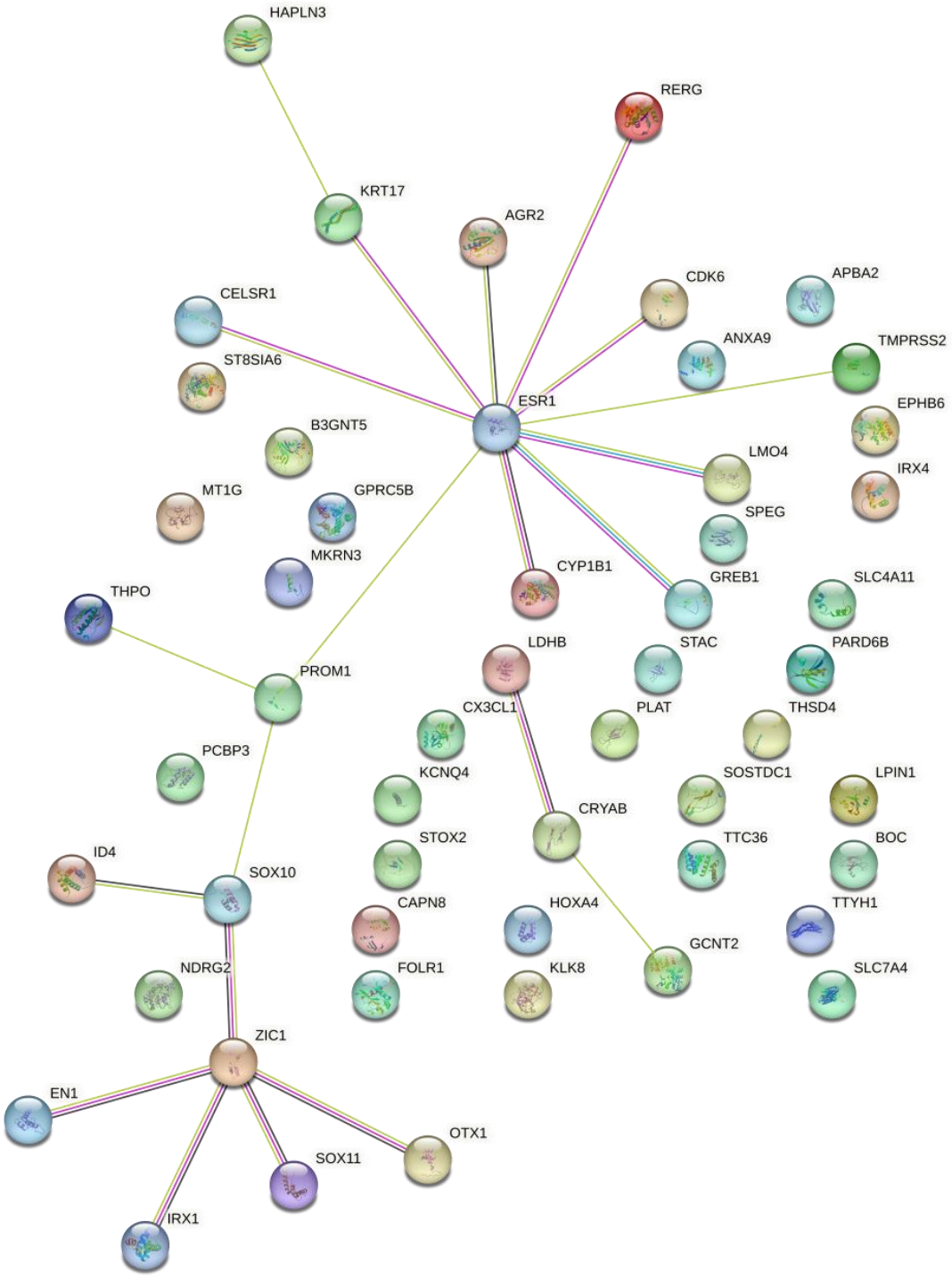
In continuation to the information provided in section 3.2.1, where we discussed the 94 unique genes across all three hormone receptors. We performed functional analysis for these genes individually for each hormone receptor using MSigDB[75]. Table A.1 in appendix, shows the list of genes and their relation to corresponding functional analysis in MSigDB. The description of the functional analysis keyword is attached in Appendix A.2. We also generated protein-protein interaction networks for these genes using STRINGDB[77]. Figure 3.7 shows the PPI networks for the genes associated with each hormone receptors.

ER consisted of 45 genes out of which 40 had an Oncoscore greater than zero. Similary, PR consisted of 54 genes out of which 50 had an Oncoscore greater than zero and HER2 had a total of 51 genes out of which 46 had an Oncoscore greater than zero. None of the genes which had a zero oncoscore were part of the biggest connected component in STRINGDB. They were also not involved in any of the pathways from the MSigDB functional analysis list. The functional analysis showed 6 genes related to pathways that cause Downregulation in ER+ cancers and were present in ER gene group. Our analysis confirmed the same with gene expression downregulated for all of them and methylation status hypermethylated for all of them. Out of these 5 genes were common in all three gene group, 3 of them (PROM1, SOX10, ZIC1) were part of largest connected component of STRINGDB in all receptors and 1 of them (SOX11) was part of the largest connected component in ER and PR. Similarly, we had 10 genes related to pathways that cause Upregulation in ER+ cancers and were present in ER gene group. Our analysis confirmed the same with gene expression upregulated for all of them and methylation status hypomethylated for all except one. There were 6 genes from them that were present in largest connected component of ER in STRINGDB out of which two of them (ESR1 and CELSR1) were also part of largest connected component of PR.

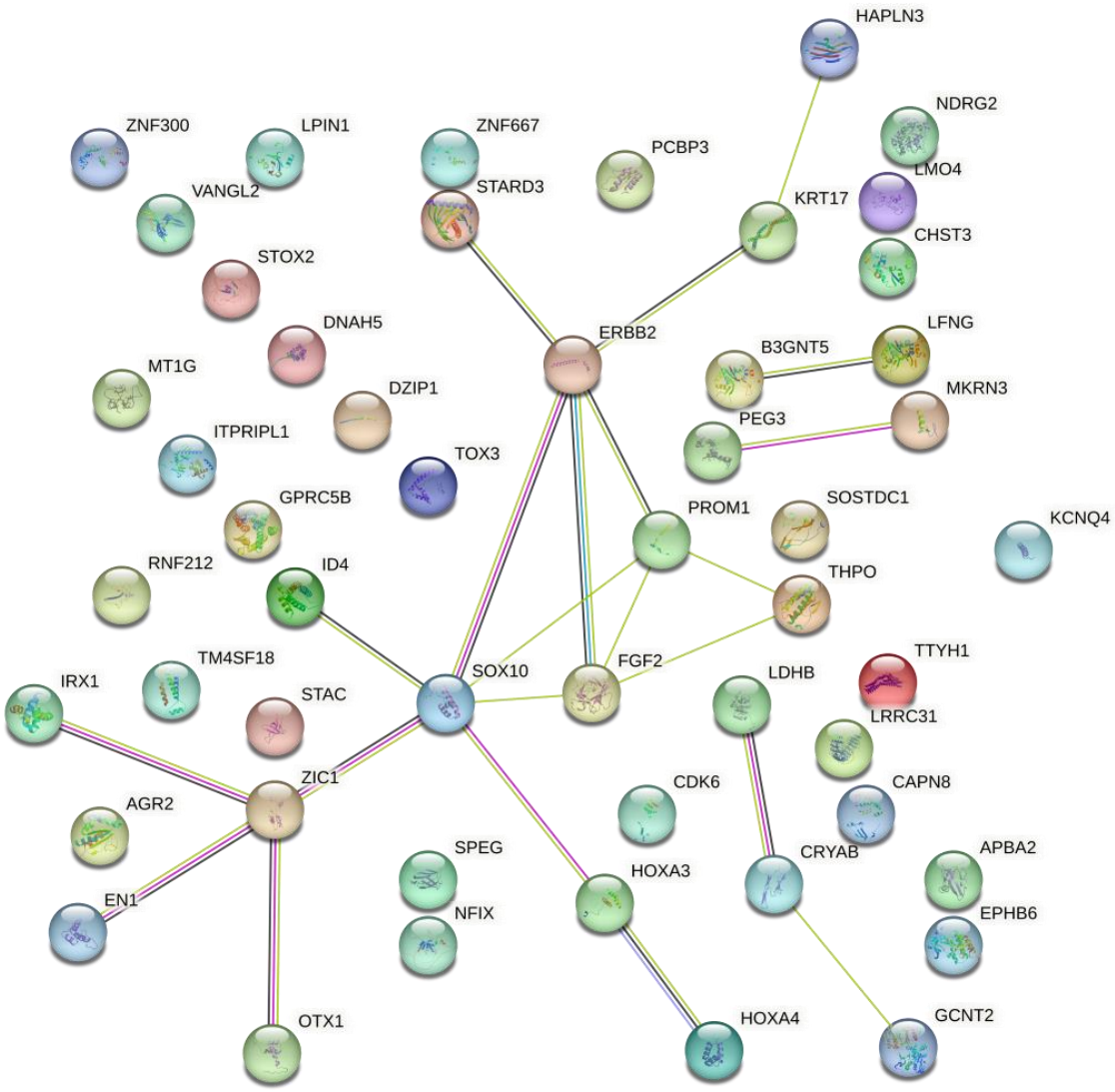
There were 28 genes from our list that were known to be Upregulated in TNBC by MSigDB. The same was found to be true in our analysis as well, where all genes were upregulated when ER, PR and HER2 status was negative. All were also Hypomethylated except APBA2, which was hypermethylated for ER and PR in this case and NXN which was hypermethylated for ER. There were 3 genes (PROM1, SOX10 and ZIC1) which were part of strongly connected component in all three receptors in STRINGDB, while there was SOX11 which was part of ER and PR, and there were ID4, KRT17 and EN1 which were part of both PR and HER2. The genes that form the largest connected component of STRINGDB are discussed in detail below.


SOX11
ZIC1
WWTR1
SOX10
EVL
FAM135A
SIRPA
THPO
GPR160
THSD4
NXN
GPRC5B
SLC26A1
STOX2
PROM1
RNASE4
SLC44A4
PCDHA10
CTNND2
FAM63A
RIPK3
PCDHA6
B3GNT5
KRT18
MUC1
NDRG2
CDK2AP1
ST8SIA6
MLPH
SYTL2
GREB1
LDHB
SLC16A1
ESR1
CDK6
MKRN3
CELSR1
PARD6B
ANXA9
SLC7A4
TOX3
IL32
NHLRC4
RERG
RUNX3
APBA2

A


HAPLN3
RERG
KRT17
AGR2
CDK6
APBA2
CELSR1
ANXA9
TMPRSS2
ST8SIA6
ESR1
EPHB6
B3GNT5
LMO4
IRX4
MT1G
GPRC5B
SPEG
MKRN3
CYP1B1
GREB1
SLC4A11
THPO
LDHB
STAC
PARD6B
PROM1
CX3CL1
PLAT
THSD4
PCBP3
KCNQ4
SOSTDC1
LPIN1
CRYAB
STOX2
TTC36
BOC
ID4
SOX10
CAPN8
HOXA4
TTYH1
GCNT2
NDRG2
FOLR1
KLK8
SLC7A4
ZIC1
EN1
OTX1
SOX11
IRX1

B



C

Figure 3.7: Protein-Protein Interaction Networks for (A) ER, (B) PR, (C) HER2.

GREB1 (growth regulation by estrogen in breast cancer 1) is an estrogen receptor (ER) gene target associated with proliferation and ER activity regulation in estrogen-responsive breast cancer cells [78]. GREB1 plays an important role in the estrogen-induced growth of breast cancer cells. Higher GREB1 expression in ER+ve breast cancer is associated with improved survival in response to the ER agonist tamoxifen [80]. Higher GREB1 expression was associated with both prolonged disease-free survival and sensitivity to tamoxifen treatment in a study that included only patients who received adjuvant tamoxifen monotherapy [81].

CDK6 belongs to the cyclin-dependent kinase (CDK) gene family, which is involved in breast cancer. CDK6 has been shown to be under-expressed in breast cancer cells, which may result in a poor prognosis for patients [79]. CDK4/6 inhibitors have changed the way advanced hormone receptor-positive, HER2-negative breast cancer is treated [82]. Inhibiting CDK6 has been shown to be a promising cancer treatment strategy [83].

MUC1 is a gene that codes for a transmembrane glycoprotein found on the surface of many epithelial cells. It's been found to be overexpressed in a variety of cancers, including TNBC breast cancer [84]. MUC1 overexpression has been linked to a poor prognosis in patients with HER2+ breast cancer [85]. MUC1 overexpression has also been linked to a more estrogen receptor (ER)-positive phenotype [86]. In our own analysis, we found MUC1 to be upregulated and hypermethylated in ER+ cancers.

TOX3 is a gene that has been identified as a susceptibility locus for breast cancer [87]. It binds to the BRCA1 promoter and suppresses BRCA1 expression2. In a mouse model of breast cancer, ectopic expression of TOX3 increased breast cancer cell proliferation, migration, and survival after apoptotic stimuli and was associated with tumor progression [88]. TOX3 upregulates a subset of ER target genes as well as genes involved in cell cycle, cancer progression, and metastasis in the MCF-7 breast cancer cell line [89].

RUNX3 is a transcription factor that suppresses tumor growth in breast cancer [94]. In human breast cancer cell lines and samples, it is frequently inactivated by hemizygous deletion of the Runx3 gene, hypermethylation of the Runx3 promoter, or cytoplasmic sequestration of RUNX3 protein [96]. RUNX3 inactivation has been linked to the onset and progression of breast cancer [96]. In severe combined immunodeficiency mice, RUNX3 inhibits the estrogen-dependent proliferation and transformation potential of ER-positive MCF-7 breast cancer cells and suppresses tumorigenicity of MCF-7 cells [94].

Prominin-1 (PROM1) is a glycoprotein found on the cell surface that has been shown to regulate PKA-induced gluconeogenesis, TGFβ-induced fibrosis, and IL-6-induced regeneration in the liver [98]. PROM1 is commonly reported as a neuronal and hematopoietic stem cell marker, but it is also expressed in cancer stem cells and cancer cells, including breast cancer [108]. PROM1 overexpression has been found in cancers of the brain, esophagus, leukemia, testis, ovary, and stomach, while PROM1 underexpression has been found in bladder, breast, and kidney cancers [108].

WWTR1, also known as TAZ, is a 14-3-3 binding protein with a PDZ binding motif that regulates the differentiation of mesenchymal stem cells [109]. It has been demonstrated that it plays an important role in the migration, invasion, and tumorigenesis of breast cancer cells [109]. TAZ is prominently expressed in human breast cancer cell lines, and its levels generally correlate with cancer cell invasiveness [109]. TAZ overexpression causes morphologic changes characteristic of cell transformation and promotes cell migration and invasion in low-expressing MCF10A cells. RNA interference-mediated knockdown of TAZ expression in MCF7 and Hs578T cells, on the

other hand, reduces cell migration and invasion [109].

SOX11, a transcription factor, has been linked to breast cancer. In breast cancer patients, high SOX11 expression is associated with poor overall survival and increased metastasis formation [110]. SOX11 has been shown in vivo to promote invasive transition, confirming its role in the progression of DCIS to invasive breast cancer [110].

RERG (RAS-related and estrogen-regulated growth inhibitor) is a one-of-a-kind RAS superfamily gene that has been linked to breast cancer [111]. High RERG expression has been found to correlate with a set of genes that define an estrogen receptor-positive breast tumour subtype, as well as a slow rate of tumour cell proliferation and a favorable prognosis for these cancer patients [111]. In breast cancer, RERG has been shown to be involved in the RAS pathway and ER-dependent transcription [112]. It was discovered to inhibit the Ras-activated pathway and to mediate RAS-driven biological effects [112]. RERG knockdown has been shown to increase the mobility of breast cancer cells and make them more resistant to SERM treatment [112].

The estrogen receptor protein (ESR1) is encoded by the ESR1 gene. It contributes to the pathogenesis of cancers like breast, endometrial, and prostate cancer. ESR1 mutations are a common cause of acquired resistance to estrogen deprivation by aromatase inhibition, which is the backbone of therapy in metastatic hormone receptor-positive breast cancer [113]. These mutations, which can affect tumor sensitivity to established and novel therapies, are a focus of current research [113].

KRT18 belongs to the keratin family of intermediate filament proteins. It is expressed in simple epithelial cells and is used to identify them [114]. KRT18 has been shown to be a useful breast cancer prognostic marker [114]. It has also been shown to be associated with malignant status and to function as an oncogene in the progression of colorectal cancer [115].

B3GNT5 (β1,3-N-acetylglucosaminyltransferase V) is a unique glycosyltransferase that has been linked to breast cancer [116]. Increased B3GNT5 transcription and glycosylation has been shown to promote breast cancer aggressiveness [116]. Through B3GNT5 overexpression and glycosylation-mediated protein stabilization, B3GNT5 promotes tumorigenesis by increasing SSEA-1 expression and cancer stem cell (CSC) properties in breast cancer cells [116].

The SOX10 gene encodes a transcription factor that is essential for neural crest-derived melanocytes and glia survival, maturation, and differentiation [117]. SOX10 has been linked to cancer progression and has been shown to significantly regulate tumour proliferation, migration, and apoptosis [118]. A Salk Institute team discovered that the gene SOX10 directly controls the growth and invasion of a significant proportion of difficult-to-treat triple-negative breast cancers [119]. SOX10 has also been linked to triple-negative breast cancer (TNBC) [119].

Zinc finger of the cerebellum 1 (ZIC1) is a gene that has been found to suppress breast cancer growth by targeting survivin[121]. Han et al. discovered that ZIC1 correlated negatively with

survivin in tumours and cells, and that higher ZIC1 RNA expression predicted better overall survival in breast cancer samples from The Cancer Genome Atlas (TCGA)[121]. ZIC1 overexpression inhibited cell proliferation, decreased mitochondrial membrane potential, and induced apoptosis in breast cancer cells in vitro by inactivating the Akt/mTOR/P70S6K pathway, suppressing survivin expression, modulating the cell cycle, releasing cytochrome c (Cyto-c) into the cytosol, and activating caspase proteins [121]. In vivo, increased ZIC1 expression inhibited the growth of implanted tumours and downregulated survivin expression in tumors [121]. On the whole, these findings demonstrate that ZIC1 plays a tumor suppressive role in breast cancer by targeting surviving and significantly downregulating its expression [121].

Thrombopoietin (THPO) is a protein that promotes cell growth and division (proliferation). It is well-known for its role in megakaryocyte proliferation and maturation as a megakaryocyte growth and development factor (MGDF)[122]. THPO has been linked to a poor prognosis in patients with gastric adenocarcinoma [122]. THPO expression was increased in tumour tissue and cells, and it was linked to a poor prognosis in patients with gastric adenocarcinoma [122].

SLC7A4 (solute carrier family 7 member 4) is a gene that codes for a protein in the solute carrier (SLC) family [123]. As glucose and glutamate transporters, the SLC family plays critical roles in cancer cell metabolism [124]. SLC7A4 has been linked to improved progression-free interval (PFI) and disease-specific survival (DSS) in breast cancer [124]. SLC7A4 was also associated with a favourable overall survival (OS), distant metastasis-free survival (DMFS), relapse-free survival (RFS), and post-progression survival (PPS) prognosis [124].

Approximately 30% of human breast cancers and many other cancer forms, including ovarian, stomach, bladder, salivary, and lung carcinomas, have the ERBB2 gene amplified or overexpressed. Numerous studies have found that ERBB2 amplification or overexpression disturbs normal cell-control processes, giving rise to aggressive cancer cells. Those with ERBB2-overexpressing breast cancer have significantly poorer overall survival rates and shorter disease-free intervals than those with ERBB2-negative breast cancer [125]. Breast tumours with ERBB2 amplification exhibit rapid tumour growth, a decreased survival rate, and accelerated disease progression. The molecular processes underpinning ERBB2's oncogenic action entail a complex signalling network that closely controls malignant cell migration and invasion, and thus metastatic potential [126].

TMPRSS2 is a transmembrane serine protease that is involved in the activation of the SARS-CoV-2 spike protein [127]. The gene TMPRSS2 has been linked to cancer susceptibility and severity [127]. High TMPRSS2 expression was associated with a shorter overall survival in breast invasive cancer (BRCA)[127][128].

Anterior gradient 2 (AGR2) is a protein that is involved in various signal transduction pathways that are crucial for cell survival and plays a critical function in oxidative protein folding in the endoplasmic reticulum [129]. AGR2 expression is related with oestrogen receptor (ER)-positive tumours in breast cancer, and its overexpression is a predictor of poor prognosis [130]. ER-alpha,

which is predominantly bound in tumours with poor outcomes, directly targets the AGR2 gene [130]. AGR2 has also been linked to the advancement of breast cancer via influencing the tumour immune microenvironment. Patients with low AGR2 expression may benefit from a combination of immune checkpoint inhibitors and TGF-β blockers [131].

ID4 is a protein in the ID (inhibitors of differentiation) family that has been linked to a stem-like phenotype and a poor prognosis in basal-like breast cancer [132]. It has been demonstrated that it promotes angiogenesis by increasing the expression of pro-angiogenic cytokines such as interleukin-8, CXCL1 and vascular endothelial growth factor [132]. ID4 can act as a tumour suppressor or oncogene in various tumour types, in addition to its role in angiogenesis. Its significance in breast cancer is unclear, as it has both an oncogenic and tumour suppressor function. ID4 is thought to act as both, however its involvement in breast cancer varies depending on the tumor's oestrogen receptor (ER) status [133].

The HOXA4 gene is a member of the HOX gene family. Several investigations have found HOXA4 to be involved with cancer. For example, HOXA4 expression has been observed to be down-regulated in lung cancer tissues when compared to non-cancerous tissues [134]. The expression of HOXA4 has been linked to tumour size, TNM stage, lymph node metastasis, and prognosis [134]. According to studies, HOXA4 expression is adversely linked with cell cycle, metastasis, and the Wnt signalling pathway. Furthermore, overexpression of HOXA4 in lung cancer cell lines inhibited cell proliferation, migration, and invasion [134]. These findings imply that HOXA4 is a possible diagnostic and prognostic marker in lung cancer, and that its overexpression may slow the course of the disease [134]. The involvement of HOXA4 has also been reported in colorectal cancer and epithelial ovarian cancer [135][136].

CYP1B1 is a gene that encodes an enzyme from the cytochrome P450 class. These enzymes have a role in a variety of bodily activities, including the breakdown of medications and the production of specific fats (lipids). The CYP1B1 enzyme is involved in biochemical reactions that involve the addition of an oxygen atom to other compounds. According to a study published in Cancer Letters, CYP1B1 catalyses the conversion of 17-β-estradiol (E2) to the catechol oestrogen metabolites 2-OH-E2 and 4-OH-E2, both of which have been implicated in breast carcinogenesis [137]. The study also claims that distinct single nucleotide polymorphisms (SNPs) in the CYP1B1 gene may explain the variation in enzymatic activity of the CYP1B1 protein and thus its ability to metabolise oestrogen between individuals [137].

KRT17 is a gene that encodes a keratin protein family member that is abundant in the skin's outer layer and protects epithelial cells from injury. KRT17 has been studied in a variety of cancers, including breast cancer. KRT17 expression was much lower in breast cancer tissues than in normal tissues, according to a study published in Biomolecules, particularly in the luminal-A, luminal-B, and human epidermal growth factor receptor-2 (HER2) + subtypes of breast cancer [138]. The study also discovered that lower KRT17 expression was substantially associated with a poor prognosis in breast cancer patients, particularly those with HER2 high and ER high tumors [138]. KRT17 was discovered to be engaged in antitumor immunity pathways, particularly the IL-

17 signaling pathway, and to be connected with a variety of immune cells, including natural killer (NK) and CD4 + T cells [138]. Finally, greater KRT17 expression indicated a better outcome in breast cancer patients with higher HER2 expression [138].

In breast cancer, the STARD3 (StAR related lipid transfer domain containing 3) gene has been found to be co-amplified with human epidermal growth factor receptor 2 (HER2) [139]. STARD3 is required in tumor cells for cholesterol transport and metabolism [139]. STARD3's potential role as a diagnostic and prognostic biomarker in breast cancer (BC) was investigated [139]. In a study of 112 patients with HER2-positive non-metastatic breast cancer treated with neoadjuvant systemic therapy (NST) and then surgery, STARD3 was found to be positive in 86.6% of cases and was significantly associated with pathological complete response (pCR) in univariate analysis ($p = 0.013$) and after adjustment for other known pathological parameters ($p = 0.044$) [140]. According to this study, assessing STARD3 overexpression status on initial biopsies of HER2-positive tumors adds benefit to the management of a subset of patients who have a high likelihood of no pathological response [140].

FGF2 has been demonstrated to promote breast cancer growth via hormone-independent activation and recruitment of oestrogen receptor alpha (ER) and PRB4 isoform to MYC regulatory sequences [141]. FGF2-induced effects were reversed by MYC inhibitors, antiestrogens, or antiprogestins [141]. A study on antiprogestin-resistant mammary carcinomas discovered that these tumours have lower levels of progesterone receptor A isoforms (PRA) than B isoforms (PRB)[142]. The study's goal was to determine the involvement of FGF2 isoforms in the advancement of breast cancer [142].

HAPLN3 is a novel link protein with hyaluronic acid binding and cell adhesion properties. It has been linked to cancer development and metastasis [143]. Transcriptome analysis confirmed MFGE8-HAPLN3 fusion as a new biomarker in triple-negative breast cancer (TNBC) in another study [144]. The study's goal was to find new fusion transcripts in TNBC [144]. TNBC had 189 fusion transcripts discovered using RNA sequencing data, 22 of which were recurrent fusions [144]. MFGE8-HAPLN3 was chosen as a new biomarker based on tumour selectivity and frameshift mutation and was verified in TNBC samples using PCR and Sanger sequencing [144].

IRX1 belongs to the Iroquois homeobox gene family and has been demonstrated to function as a tumour suppressor gene in different cancers. In patients with non-small cell lung cancer (NSCLC), for example, IRX1 promoter methylation may be a tumor-associated event and an independent predictor of survival advantage [145]. In addition, IRX1 has been identified as a tumour suppressor gene in gastric cancer [146].

The EN1 gene encodes the transcription factor EN1, which has been reported to be overexpressed in triple-negative breast tumors (TNBCs). Downregulation of EN1 has been demonstrated to diminish viability and tumorigenicity in TNBC cell lines preferentially and significantly [147]. Researchers identified genes involved in WNT and Hedgehog signaling, neurogenesis, and axonal guidance as direct EN1 transcriptional targets by combining gene expression changes after EN1

downregulation with EN1 chromatin binding patterns [147]. In patients with TNBC, high EN1 expression was associated with a shorter overall survival and an increased likelihood of acquiring brain metastases [147]. High EN1 expression has been linked to an increased risk of developing brain metastases in patients with triple-negative breast cancer (TNBC). These findings imply that the EN1 transcription factor controls neurogenesis-related genes and is linked to brain metastasis in TNBC. EN1 is thus a prognostic marker as well as a possible therapeutic target in TNBC [147].

OTX1 is a transcription factor that has been linked to cancer. OTX1, for example, has been shown to enhance the growth of colorectal cancer and hepatocellular carcinoma tumors [148]. In bladder cancer, OTX1 enhances cancer cell proliferation and motility by boosting cell cycle progression [148]. OTX1 expression is regulated by p53 in breast cancer, and its activity suggests a synergistic role with p53 in cancer stem cell (CSC) differentiation [149].

3.4 Discussion

This thesis investigated into how well hormone receptor status of patients with breast cancer (BC) may be predicted based on their DNA methylation patterns. Our strategy includes lowering the complexity of the DNA methylation data and constructing a machine-learning model to forecast receptor status. After applying this methodology to five datasets, it was possible to see some interesting trends and possible improvements over conventional Immunohistochemistry (IHC) techniques. Our t-SNE plots demonstrated that there was clustering, with some intermixing, between the receptor-positive and receptor-negative groups for ER, PR, and HER2. High OncoScore values supported the significant cancer relationship seen in gene analysis connected to CpG sites.

Satisfactory results were obtained from the model evaluation utilising stratified 10-fold cross-validation, particularly for the prediction of ER and PR status. The Kaplan-Meier technique of survival analysis showed a significant difference in the likelihood of survival of patients with mismatched ER (p-value 0.005) and HER2 (p-value 0.005) status up to the 6-year point. Notably, patients with correctly predicted ER/HER2 receptor status matching had higher survival rates. The significance of accurate receptor status identification in patient prognosis and treatment planning was highlighted by this.

When the discordance between intrinsic and clinical subtypes was analysed, it became clear that DNA methylation-based prediction frequently produced lower discordance rates, indicating more accurate classification of receptor status. Planning for therapeutic strategies and patient prognosis may be enhanced as a result. It was helpful to utilise the Cleanlab programme to find labelling mistakes because it showed that samples with and without labelling problems had significantly different IHC results. Although the Cleanlab did not significantly affect survival results, it did have a substantial impact in predicting receptor status and subsequent treatment plans.

The crucial function of the 94 distinct genes across the three hormone receptors- ER, PR, and HER2 takes centre stage in our discussion. As a result of our research, we can see that certain genes demonstrate a correlation with particular functional pathways as identified by MSigDB and are intricately linked in protein-protein interaction networks. Importantly, these genes exhibit patterns in their Oncoscore status, an indicator of their potential contribution to the development of cancer. Further supporting their significance in cancer dynamics is the fact that genes with a zero Oncoscore were neither included in any MSigDB-defined pathways nor the largest linked component in STRINGDB.

The discovery of particular genes, such as PROM1, SOX10, ZIC1, and SOX11, which span all three hormone receptors and function as essential nodes in the protein-protein interaction networks, is noteworthy. Under ER+ cancer or TNBC settings, these genes exhibited the expected expression and methylation patterns. Our research, reveals a group of genes that are known to be elevated in TNBC, a result that may provide some insight into the molecular complexity of this aggressive breast cancer subtype. The information obtained via this research sets the door for a more in-depth comprehension of the functional interactions between hormone receptors and cancer and the genetic components that affect the many phenotypes of cancer.

Prior efforts, as highlighted in a 2020 study [8], tried to predict the receptor status of breast cancer patients by utilising Gene Expression Profiling data. Logistic regression was used to find the predictor genes using data from the TCGA and METABRIC datasets. These widespread predictor genes were then utilised in a different Logistic Regression model to calculate the patients' receptor status. While this method outperformed the IHC-based method in terms of clinical significance and a reduction in incongruity with intrinsic subtypes, it fell short in resolving the problem of noisy labels, which results from potential mislabeling by the IHC-based method. In contrast, our analysis used gene expression data for feature selection and DNA methylation data for predicting receptor status. This strategy was chosen because it has been established that patient DNA methylation profiles and receptor status correlate, and because reversible epigenetic therapy has compelling therapeutic potential. Additionally, our approach addressed the problem of noisy labels in the dataset by implementing Noisy-Label training based machine learning models, which work to reduce the negative performance impact caused by inaccurate labels in the dataset.

Our machine-learning pipeline does have some limitations, though. Given the abundance of features in DNA methylation data, there is a chance that important characteristics for predicting Receptor Status may be accidentally overlooked during feature selection, and non-biologically relevant features may enter the prediction model. Due to the fact that the sample size currently available for training machine learning models is far smaller than the number of features per sample, a high-dimensional data environment has been created, which when combined with a small sample size leads to overfitting in machine learning models. In addition, class imbalance is a recurring problem in predictions of receptor state based on machine learning. Since 80% of all breast tumours are HER2-positive, there are difficulties with class imbalance during model training that could have an effect on performance. The potential inconsistencies in receptor status labels generated by laboratory procedures could hinder the machine learning models' anticipated

performance, even with our usage of a noisy label trained machine learning model. However, we remain enthusiastic about the potential to get over these restrictions in future research given the continual swift developments in data accessibility and machine learning.

Chapter 4

Conclusions

The proper assessment of receptor status is a crucial step in the effort to give breast cancer patients more specialised and personalised care. It immediately affects the course of treatment and aids in predicting the possibility of illness recurrence, thereby saving patients from unwanted side effects and potentially fatal toxicity from medicines that were administered without adequate knowledge. Immunohistochemistry (IHC), a technique known to be impacted by different laboratory-based factors and sensitive to discrepant results, is now the main method for predicting receptor status in breast cancer patients. By employing machine learning approaches to develop a DNA methylation-based receptor status prediction model for this thesis, we attempted to address these shortcomings. Given the well-established significance of DNA methylation in breast cancer and the evidence supporting a link between patient DNA methylation profiles and receptor status [55], this method presented a possible replacement for conventional IHC-based techniques.

Our findings showed that compared to the IHC-based method, receptor status predictions based on DNA methylation yielded a reduced discordance rate with intrinsic subtypes. Additionally, we discovered genes with variable levels of methylation in their promoter regions. Some of these genes have been linked to various breast cancer subtypes and cancer in general. In addition, we discovered genes about which little to no information had previously been known, opening up fresh research directions for potential diagnostic or prognostic indicators for breast cancer subtypes. This thesis, to our knowledge, is the first to use DNA methylation data to predict the ER, PR, and HER2 status of breast cancer patients. Our analysis sheds light on the potential value of DNA methylation information for predicting receptor status. This strategy might improve the selection of patients' appropriate treatment programmes, ultimately resulting in a better prognosis.

We acknowledge that this study still must be improved even though we were able to predict receptor status. Future studies should work to support the therapeutic applicability of our findings. To improve the performance of DNA methylation-based prediction models, larger datasets are required, especially those with more evenly distributed classes. To reduce the problem of noisy labels in training samples, receptor status labelling must be done with more care. It is also necessary to look more closely at the effects that distinct receptors' differential methylation has on

patients' gene expression profiles. Finally, clinical trials will be used as the final test for the applicability of our work.

We anticipate advances that will help overcome these issues as the availability of data and the field of machine learning continue to grow. By doing this, we want to advance the personalization and advancement of breast cancer treatment, ultimately improving patient outcomes.

Related Publications

Publications

1) **Receptor Status Prediction in Breast Cancer Patients Using Machine Learning Pipeline on DNA Methylation Data**, Saksham Gupta. In 2022 12th International Conference on Bioscience, Biochemistry and Bioinformatics (ICBBB '22). Association for Computing Machinery, New York, NY, USA, 38–43. https://doi.org/10.1145/3510427.3510433

Oral Presentations

1) **Receptor Status Prediction in Breast Cancer Patients Using Machine Learning Pipeline on DNA Methylation Data,** LanBix2021: Sri Lankan Conference on Bioinformatics 2021, Virtual, Sri Lanka, March 19-20, 2021.

Bibliography

[1] Wu Ct, Morris JR. Genes, genetics, and epigenetics: a correspondence. Science. 2001 Aug 10;293(5532):1103-5. doi: 10.1126/science.293.5532.1103. PMID: 11498582.

[2] Alegría-Torres JA, Baccarelli A, Bollati V. Epigenetics and lifestyle. Epigenomics. 2011;3(3):267-277. doi:10.2217/epi.11.22

[3] Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. Carcinogenesis. 2010;31(1):27-36. doi:10.1093/carcin/bgp220

[4] Jaenisch, R., Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet 33, 245–254 (2003). https://doi.org/10.1038/ng1089

[5] Zuo T, Tycko B, Liu TM, Lin JJ, Huang TH. Methods in DNA methylation profiling. Epigenomics. 2009 Dec;1(2):331-45. doi: 10.2217/epi.09.31. PMID: 20526417; PMCID: PMC2880494.

[6] Yong, W., Hsu, F. & Chen, P. Profiling genome-wide DNA methylation. Epigenetics & Chromatin 9, 26 (2016). https://doi.org/10.1186/s13072-016-0075-3

[7] Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A. 1992;89:1827–31.

[8] Yoon S, Won HS, Kang K, Qiu K, Park WJ, Ko YH. Hormone Receptor-Status Prediction in Breast Cancer Using Gene Expression Profiles and Their Macroscopic Landscape. Cancers (Basel). 2020;12(5):1165. Published 2020 May 5. doi:10.3390/cancers12051165

[9] Gown, A. Current issues in ER and HER2 testing by IHC in breast cancer. Mod Pathol 21, S8–S15 (2008). https://doi.org/10.1038/modpathol.2008.34

[10] Davidson-Pilon, (2019). lifelines: survival analysis in Python. Journal of Open Source Software, 4(40), 1317, https://doi.org/10.21105/joss.01317

[11] Hon G.C., et al. (2012). Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer.

[12] Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE, Gnirke A, Meissner A. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013 Aug 22;500(7463):477-81. doi: 10.1038/nature12433. Epub 2013 Aug 7. PMID: 23925113; PMCID: PMC3821869.

[13] Karl Pearson F.R.S. . (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11), 559–572. https://doi.org/10.1080/14786440109462720

[14] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning Representations by Back-Propagating Errors. Cambridge, MA, USA: MIT Press, 1988, pp. 696—699.

[15] Pei, L., Hsu, WW., Chiang, LA., Guo, JM., Iftekharuddin, K.M., Colen, R. (2021). A Hybrid Convolutional Neural Network Based-Method for Brain Tumor Classification Using mMRI and WSI. In: Crimi, A., Bakas, S. (eds) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2020. Lecture Notes in Computer Science(), vol 12659. Springer, Cham. https://doi.org/10.1007/978-3-030-72087-2_43

[16] Chen, B., Javadi, G., Jamzad, A., Hamilton, A., Sibley, S., Abolmaesumi, P., Maslove, D., & Mousavi, P. (2021). Detecting Atrial Fibrillation in ICU Telemetry data with Weak Labels. In Proceedings of the 6th Machine Learning for Healthcare Conference (pp. 176–195). PMLR.

[17] Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot T, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G and Noushmehr H. "TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data." Nucleic acids research (2015): gkv1507.

[18] R. V. Lloyd, L. A. Erickson, M. B. Casey, K. Y. Lam, C. M. Lohse,S. L. Asa, J. K. Chan, R. A. DeLellis, H. R. Harach, K. Kakudoet al.,"Observer variation in the diagnosis of follicular variant of papillarythyroid carcinoma,"The American Journal of Surgical Pathology,vol. 28, no. 10, pp. 1336–1340, 2004.

[19] Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021. Confident Learning: Estimating Uncertainty in Dataset Labels. *J. Artif. Int. Res.* 70 (May 2021), 1373–1411. DOI:https://doi.org/10.1613/jair.1.12125

[20] Northcutt CG, Wu T, Chuang IL. Learning with Confident Examples: Rank Pruning for Robust Classification with Noisy Labels. In: Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence UAI'17, AUAI Press; 2017. http://auai.org/uai2017/proceedings/papers/35.pdf.

[21] Network CGA, et al. Comprehensive molecular portraits of human breast tumours. Nature 2012;490(7418):61

[22] Jeschke J, Bizet M, Desmedt C, Calonne E et al. DNA methylation-based immune response signature improves patient diagnosis in multiple cancers. J Clin Invest 2017 Aug 1;127(8):3090-3102.

[23] Fleischer T, Tekpli X, Mathelier A, Wang S et al. DNA methylation at enhancers identifies distinct breast cancer lineages. Nat Commun 2017 Nov 9;8(1):1379

[24] Williams KE, Jawale RM, Schneider SS, Otis CN et al. DNA methylation in breast cancers: Differences based on estrogen receptor status and recurrence. J Cell Biochem 2019 Jan;120(1):738-755.

[25] Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009 Mar;27(8):1160–1167.

[26] Paquet ER, Hallett MT. Absolute assignment of breast cancer intrinsic molecular subtype. J Natl Cancer Inst 2015 Jan;107(1):357.

[27] Carey LA, Berry DA, Cirrincione CT, Barry WT, Pitcher BN, Harris LN, et al. Molecular Heterogeneity and Response to Neoadjuvant Human Epidermal Growth Factor Receptor 2 Targeting in CALGB 40601, a Randomized Phase III Trial of Paclitaxel PlusTrastuzumab With or Without Lapatinib. J Clin Oncol 2016 Feb;34(6):542–549.

[28] de Almeida, B.P., Apolónio, J.D., Binnie, A. et al. Roadmap of DNA methylation in breast cancer identifies novel prognostic biomarkers. BMC Cancer 19, 219 (2019). https://doi.org/10.1186/s12885-019-5403-0

[29] Fernandez-Rozadilla C,Simões AR,Lleonart ME,Carnero A andCarracedo A´(2021) Tumor Profiling at the Service of Cancer Therapy.Front. Oncol. 10:595613.doi: 10.3389/fonc.2020.595613

[30] Vallejos C.S., Gomez H.L., Cruz W.R., Pinto J.A., Dyer R.R., Velarde R., Suazo J.F., Neciosup S.P., León M., de la Cruz M.A., et al. Breast Cancer Classification According to Immunohistochemistry Markers: Subtypes and Association with Clinicopathologic Variables in a Peruvian Hospital Database. Clin. Breast Cancer. 2010;10:294–300. doi: 10.3816/CBC.2010.n.038.

[31] Chen, W., Colditz, G. Risk factors and hormone-receptor status: epidemiology, risk-prediction models and treatment implications for breast cancer. Nat Rev Clin Oncol 4, 415–423 (2007). https://doi.org/10.1038/ncponc0851

[32] von Minckwitz G, Untch M, Blohmer JU, Costa SD, Eidtmann H, Fasching PA, Gerber B, Eiermann W, Hilfrich J, Huober J, Jackisch C, Kaufmann M, Konecny GE, Denkert C, Nekljudova V, Mehta K, Loibl S. Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. J Clin Oncol. 2012 May 20;30(15):1796-804. doi: 10.1200/JCO.2011.38.8595. Epub 2012 Apr 16. PMID: 22508812.

[33] Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, Fan JB, Shen R. High density DNA methylation array with single CpG site resolution. Genomics. 2011 Oct;98(4):288-95. doi: 10.1016/j.ygeno.2011.07.007. Epub 2011 Aug 2. PMID: 21839163.

[34] Fortin JP, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome biology 2014 Dec;15(12):503–503. https://pubmed.ncbi.nlm.nih.gov/25599564, 25599564[pmid].

[35] Weigelt B, Reis-Filho JS. Histological and molecular types of breast cancer: is there a unifying taxonomy? Nat Rev Clin Oncol. 2009 Dec;6(12):718-30. doi: 10.1038/nrclinonc.2009.166. PMID: 19942925.

[36] Reis-Filho JS, Simpson PT, Gale T, Lakhani SR. The molecular genetics of breast cancer: the contribution of comparative genomic hybridization. Pathol Res Pract. 2005;201(11):713-25. doi: 10.1016/j.prp.2005.05.013. Epub 2005 Oct 3. PMID: 16325514.

[37] Payne, S. J. L., Bowen, R. L., Jones, J. L., & Wells, C. A. (2008). Predictive markers in breast cancer–the present. Histopathology, 52(1), 82-90.

[38] Gujar H, Liang JW, Wong NC, Mozhui K (2018) Profiling DNA methylation differences between inbred mouse strains on the Illumina Human Infinium MethylationEPIC microarray. PLoS ONE 13(3): e0193496. https://doi.org/10.1371/journal.pone.0193496.

[39] Saha Roy S, Vadlamudi RK. Role of estrogen receptor signaling in breast cancer metastasis. Int J Breast Cancer. 2012;2012:654698. doi: 10.1155/2012/654698. Epub 2011 Dec 19. PMID: 22295247; PMCID: PMC3262597.

[40] Dunnwald LK, Rossing MA, Li CI. Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. Breast Cancer Res. 2007;9(1):R6. doi: 10.1186/bcr1639. PMID: 17239243; PMCID: PMC1851385.

[41] Nagini S. Breast Cancer: Current Molecular Therapeutic Targets and New Players. Anticancer Agents Med Chem. 2017;17(2):152-163. doi: 10.2174/1871520616666160502122724. PMID: 27137076.

[42] Vranic S, Teruya B, Repertinger S, Ulmer P, Hagenkord J, Gatalica Z. Assessment of HER2 gene status in breast carcinomas with polysomy of chromosome 17. Cancer. 2011 Jan 1;117(1):48-53. doi: 10.1002/cncr.25580. Epub 2010 Aug 27. PMID: 20803611.

[43] Zimmer AS, Van Swearingen AED, Anders CK. HER2-positive breast cancer brain metastasis: A new and exciting landscape. Cancer Rep (Hoboken). 2020 Sep 3:e1274. doi: 10.1002/cnr2.1274. Epub ahead of print. PMID: 32881421.

[44] Kennecke H, Yerushalmi R, Woods R, Cheang MC, Voduc D, Speers CH, Nielsen TO, Gelmon K. Metastatic behavior of breast cancer subtypes. J Clin Oncol. 2010 Jul 10;28(20):3271-7. doi: 10.1200/JCO.2009.25.9820. Epub 2010 May 24. PMID: 20498394.

[45] Bardou VJ, Arpino G, Elledge RM, Osborne CK, Clark GM (2003). Progesterone receptor status significantly improves outcome prediction over estrogen receptor status alone for adjuvant endocrine therapy in two large breast cancer databases. J Clin Oncol 21:1973–1979.

[46] Way TD, Lin JK. Role of HER2/HER3 co-receptor in breast carcinogenesis. Future Oncol. 2005 Dec;1(6):841-9. doi: 10.2217/14796694.1.6.841. PMID: 16556064.

[47] Giulianelli S, Vaqué JP, Soldati R, Wargon V, Vanzulli SI, Martins R, Zeitlin E, Molinolo AA, Helguero LA, Lamb CA, Gutkind JS, Lanari C. Estrogen receptor alpha mediates progestin-induced mammary tumor growth by interacting with progesterone receptors at the cyclin D1/MYC promoters. Cancer Res. 2012 May 1;72(9):2416-27. doi: 10.1158/0008-5472.CAN-11-3290. Epub 2012 Mar 6. PMID: 22396492.

[48] Slamon DJ, Clark GM, Wong SG. Human breast cancer. Correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science. 1987;235:177-182.

[49] Sjögren S, Inganas M, Lindgren A. Prognostic and predictive value of c-erbB-2 overexpression in primary breast cancer, alone and in combination with other prognostic markers. J Clin Oncol. 1998.16:462-469.

[50] Yarden Y, Sliwkowski M. Untangling the ErbB signaling network. Nat Rev Mol Cell Biol. 2001;2:127-137.

[51] Gschwind A, Fischer OM, Ullrich A. The discovery of receptor tyrosine kinases: targets for cancer therapy. Nat Rev Cancer. 2004;4:361-370.

[52] Grimm SL, Hartig SM, Edwards DP. Progesterone Receptor Signaling Mechanisms. J Mol Biol. 2016 Sep 25;428(19):3831-49. doi: 10.1016/j.jmb.2016.06.020. Epub 2016 Jul 2. PMID: 27380738.

[53] Lakhani, S., International Agency for Research on Cancer, & World Health Organization (2012). WHO Classification of Tumours of the Breast. International Agency for Research on Cancer.

[54] Hinshelwood RA, Clark SJ (2008) Breast cancer epigenetics: normal human mammary epithelial cells as a model system. J Mol Med 86(12):1315–1328. doi: 10.1007/s00109-008-0386-3

[55] Feng, W., Shen, L., Wen, S. et al. Correlation between CpG methylation profiles and hormone receptor status in breast cancers. Breast Cancer Res 9, R57 (2007). https://doi.org/10.1186/bcr1762

[56] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[57] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

[58] Krüger, Frank. (2016). Activity, Context, and Plan Recognition with Computational Causal Behaviour Models.

[59] Kaplan, E., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association, 53(282), 457-481. doi:10.2307/2281868

[60] van der Maaten, L. & Hinton, G. (2008). Visualizing Data using t-SNE . Journal of Machine Learning Research, 9, 2579--2605.

[61] Levy JJ, Titus AJ, Salas LA, Christensen BC. PyMethylProcess-convenient high-throughput preprocessing workflow for DNA methylation data. Bioinformatics. 2019 Dec 15;35(24):5379-5381. doi: 10.1093/bioinformatics/btz594. PMID: 31368477; PMCID: PMC6954637.

[62] Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30(10):1363-1369. doi:10.1093/bioinformatics/btu049

[63] Min JL, Hemani G, Davey Smith G, Relton C, Suderman M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. Bioinformatics. 2018 Dec 1;34(23):3983-3989. doi: 10.1093/bioinformatics/bty476. PMID: 29931280; PMCID: PMC6247925.

[64] Xu Z, Niu L, Li L, Taylor JA. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. Nucleic Acids Res. 2016;44(3):e20. doi:10.1093/nar/gkv907

[65] Keogh E., Mueen A. (2017) Curse of Dimensionality. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7687-1_192

[66] Diep D, Plongthongkum N, Gore A et al (2012) Library-free methylation sequencing with bisulfite padlock probes. Nat Methods 9:270–272

[67] Krueger, F., Kreck, B., Franke, A. et al. DNA methylome analysis using short bisulfite sequencing data. Nat Methods 9, 145–151 (2012). https://doi.org/10.1038/nmeth.1828

[68] Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27(11):1571-1572. doi:10.1093/bioinformatics/btr167

[69] Rauch, T., Pfeifer, G. Methylated-CpG island recovery assay: a new technique for the rapid detection of methylated-CpG islands in cancer. Lab Invest 85, 1172–1180 (2005). https://doi.org/10.1038/labinvest.3700311

[70] Gupta N., Verma V.K. (2019) Next-Generation Sequencing and Its Application: Empowering in Public Health Beyond Reality. In: Arora P. (eds) Microbial Technology for the Welfare of Society. Microorganisms for Sustainability, vol 17. Springer, Singapore. https://doi.org/10.1007/978-981-13-8844-6_15

[71] T. R. Elvitigala et al., "High-Throughput Biological Data Analysis," in IEEE Control Systems Magazine, vol. 30, no. 6, pp. 81-100, Dec. 2010, doi: 10.1109/MCS.2010.938100.

[72] Pidsley, R., Zotenko, E., Peters, T.J. et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol 17, 208 (2016). https://doi.org/10.1186/s13059-016-1066-1

[73] Pop, Sevinci & Enciu, Ana-Maria & Tarcomnicu, Isabela & Gille, Elvira & Tanase, Cristiana. (2019). Phytochemicals in cancer prevention: modulating epigenetic alterations of DNA methylation. Phytochemistry Reviews. 18. 10.1007/s11101-019-09627-x.

[74] Ferguson-Smith, A. Genomic imprinting: the emergence of an epigenetic paradigm. Nat Rev Genet 12, 565–575 (2011). https://doi.org/10.1038/nrg3032

[75] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50. doi: 10.1073/pnas.0506580102. Epub 2005 Sep 30. PMID: 16199517; PMCID: PMC1239896.

[76] Rocco P, Daniele R, Roberta S, Alessandra P, Luca DS, Pierangelo F, et al. OncoScore: a novel, Internet-based tool to assess the oncogenic potential of genes. Sci Rep. 2017;7:46290.

[77] Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, Jensen LJ, von Mering C (2021). “The STRING database in 2021: customizable protein-protein

networks, and functional characterization of user-uploaded gene/measurement sets." Nucleic Acids Research (Database issue), 49.

[78] Haines CN, Braunreiter KM, Mo XM, Burd CJ. GREB1 isoforms regulate proliferation independent of ERα co-regulator activities in breast cancer. Endocr Relat Cancer. 2018 Jul;25(7):735-746. doi: 10.1530/ERC-17-0496. Epub 2018 Apr 25. PMID: 29695586; PMCID: PMC7158201.

[79] Zinia, J.A., Rahman, M.S. Evaluation of the prognostic significance of CDK6 in breast cancer. Netw Model Anal Health Inform Bioinforma 9, 40 (2020). https://doi.org/10.1007/s13721-020-00250-x

[80] Shin EM, Huynh VT, Neja SA, Liu CY, Raju A, Tan K, Tan NS, Gunaratne J, Bi X, Iyer LM, Aravind L, Tergaonkar V. GREB1: An evolutionarily conserved protein with a glycosyltransferase domain links ERα glycosylation and stability to cancer. Sci Adv. 2021 Mar 17;7(12):eabe2470. doi: 10.1126/sciadv.abe2470. PMID: 33731348; PMCID: PMC7968844.

[81] Corinne N Haines and others, GREB1 regulates PI3K/Akt signaling to control hormone-sensitive breast cancer proliferation, Carcinogenesis, Volume 41, Issue 12, December 2020, Pages 1660–1670, https://doi.org/10.1093/carcin/bgaa096

[82] Battisti NML, Ring A. CDK4/6 inhibition in HER2-positive breast cancer. Lancet Oncol. 2020 Jun;21(6):734-735. doi: 10.1016/S1470-2045(20)30164-9. Epub 2020 Apr 27. PMID: 32353343.

[83] Nebenfuehr S, Kollmann K, Sexl V. The role of CDK6 in cancer. Int J Cancer. 2020 Dec 1;147(11):2988-2995. doi: 10.1002/ijc.33054. Epub 2020 May 30. PMID: 32406095; PMCID: PMC7586846.

[84] Goode G, Gunda V, Chaika NV, Purohit V, Yu F, Singh PK. MUC1 facilitates metabolomic reprogramming in triple-negative breast cancer. PLoS One. 2017 May 2;12(5):e0176820. doi: 10.1371/journal.pone.0176820. Erratum in: PLoS One. 2017 Jun 1;12 (6):e0179098. PMID: 28464016; PMCID: PMC5413086.

[85] Pang Z, Dong X, Deng H, Wang C, Liao X, Liao C, Liao Y, Tian W, Cheng J, Chen G, Yi H, Huang L. MUC1 triggers lineage plasticity of Her2 positive mammary tumors. Oncogene. 2022 May;41(22):3064-3078. doi: 10.1038/s41388-022-02320-y. Epub 2022 Apr 23. PMID: 35461328.

[86] Rakha EA, Boyce RW, Abd El-Rehim D, Kurien T, Green AR, Paish EC, Robertson JF, Ellis IO. Expression of mucins (MUC1, MUC2, MUC3, MUC4, MUC5AC and MUC6) and their prognostic significance in human breast cancer. Mod Pathol. 2005 Oct;18(10):1295-304. doi: 10.1038/modpathol.3800445. PMID: 15976813.

[87] Seksenyan A, Kadavallore A, Walts AE, de la Torre B, Berel D, Strom SP, Aliahmad P, Funari VA, Kaye J. TOX3 is expressed in mammary ER(+) epithelial cells and regulates ER target genes in luminal breast cancer. BMC Cancer. 2015 Jan 30;15:22. doi: 10.1186/s12885-015-1018-2. PMID: 25632947; PMCID: PMC4324787.

[88] Han Y-J, Zhang J, Zheng Y, Huo D, Olopade OI (2016) Genetic and Epigenetic Regulation of TOX3 Expression in Breast Cancer. PLoS ONE 11(11): e0165559. https://doi.org/10.1371/journal.pone.0165559

[89] Seksenyan, A., Kadavallore, A., Walts, A.E. et al. TOX3 is expressed in mammary ER+ epithelial cells and regulates ER target genes in luminal breast cancer. BMC Cancer 15, 22 (2015). https://doi.org/10.1186/s12885-015-1018-2

[90] McCullough AE, Dell'orto P, Reinholz MM, Gelber RD, Dueck AC, Russo L, et al. Central pathology laboratory review of HER2 and ER in early breast cancer: an ALTTO trial [BIG 2-06/NCCTG N063D (Alliance)] ring study. Breast Cancer Res Treat 2014;143:485e92.

[91] Van Bockstal, M., Floris, G., Galant, C., Lambein, K., & Libbrecht, L. (2018). A plea for appraisal and appreciation of immunohistochemistry in the assessment of prognostic and predictive markers in invasive breast cancer. The Breast, 37, 52-55.

[92] Memon, R., Granada, C. N. P., Harada, S., Winokur, T., Reddy, V., Kahn, A. G., et al. (2021). Discordance between Immunohistochemistry and In Situ Hybridization to Detect HER2 Overexpression/Gene Amplification in Breast Cancer in the Modern Age A Single Institution Experience and Pooled Literature Review Study: Discordance between HER2 overexpression and gene amplification in breast cancer. Clinical Breast Cancer.

[93] Ohara, A. M., Naoi, Y., Shimazu, K., Kagara, N., Shimoda, M., Tanei, T., et al. (2019). PAM50 for prediction of response to neoadjuvant chemotherapy for ER-positive breast cancer. Breast cancer research and treatment, 173(3), 533-543.

[94] Huang, B., Qu, Z., Ong, C. et al. RUNX3 acts as a tumor suppressor in breast cancer by targeting estrogen receptor α. Oncogene 31, 527–534 (2012). https://doi.org/10.1038/onc.2011.252

[95] Lindqvist BM, Wingren S, Motlagh PB, Nilsson TK. Whole genome DNA methylation signature of HER2-positive breast cancer. Epigenetics. 2014;9(8):1149-1162. doi:10.4161/epi.29632

[96] Chen LF. Tumor suppressor function of RUNX3 in breast cancer. J Cell Biochem. 2012 May;113(5):1470-7. doi: 10.1002/jcb.24074. PMID: 22275124; PMCID: PMC3337355.

[97] Lumachi F, Brunello A, Maruzzo M, Basso U, Basso SM. Treatment of estrogen receptor-positive breast cancer. Curr Med Chem. 2013;20(5):596-604. doi: 10.2174/092986713804999303. PMID: 23278394.

[98] Bahn MS, Ko YG. PROM1-mediated cell signal transduction in cancer stem cells and hepatocytes. BMB Rep. 2023 Feb;56(2):65-70. doi: 10.5483/BMBRep.2022-0203. PMID: 36617467; PMCID: PMC9978360.

[99] Badowska-Kozakiewicz AM, Patera J, Sobol M, Przybylski J. The role of oestrogen and progesterone receptors in breast cancer - immunohistochemical evaluation of oestrogen and progesterone receptor expression in invasive breast cancer in women. Contemp Oncol (Pozn). 2015;19(3):220-5. doi: 10.5114/wo.2015.51826. Epub 2015 May 28. PMID: 26557763; PMCID: PMC4631285.

[100] Ménard, S., Pupa, S., Campiglio, M. et al. Biologic and therapeutic role of HER2 in cancer. Oncogene 22, 6570–6578 (2003). https://doi.org/10.1038/sj.onc.1206779

[101] Dembinski R, Prasath V, Bohnak C, Siotos C, Sebai ME, Psoter K, Gani F, Canner J, Camp MS, Azizi A, Jacobs L, Habibi M. Estrogen Receptor Positive and Progesterone Receptor Negative Breast Cancer: the Role of Hormone Therapy. Horm Cancer. 2020 Aug;11(3-4):148-154. doi: 10.1007/s12672-020-00387-1. Epub 2020 Jun 9. PMID: 32519274.

[102] Dunnwald LK, Rossing MA, Li CI. Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. Breast Cancer Res. 2007;9(1):R6. doi: 10.1186/bcr1639. PMID: 17239243; PMCID: PMC1851385.

[103] Largent, J.A., Ziogas, A. & Anton-Culver, H. Effect of reproductive factors on stage, grade and hormone receptor status in early-onset breast cancer. Breast Cancer Res 7, R541 (2005). https://doi.org/10.1186/bcr1198

[104] Eyster KM. The Estrogen Receptors: An Overview from Different Perspectives. Methods Mol Biol. 2016;1366:1-10. doi: 10.1007/978-1-4939-3127-9_1. PMID: 26585122.

[105] Chen GG, Zeng Q, Tse GM. Estrogen and its receptors in cancer. Med Res Rev. 2008 Nov;28(6):954-74. doi: 10.1002/med.20131. PMID: 18642351.

[106] Brufsky AM, Dickler MN. Estrogen Receptor-Positive Breast Cancer: Exploiting Signaling Pathways Implicated in Endocrine Resistance. Oncologist. 2018 May;23(5):528-539. doi: 10.1634/theoncologist.2017-0423. Epub 2018 Jan 19. PMID: 29352052; PMCID: PMC5947450.

[107] Mounir, Mohamed, Lucchetta, Marta, Silva, C T, Olsen, Catharina, Bontempi, Gianluca, Chen, Xi, Noushmehr, Houtan, Colaprico, Antonio, Papaleo, Elena (2019). “New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx.” PLoS computational biology, 15(3), e1006701.

[108] Saha, S.K., Islam, S.M.R., Kwak, KS. et al. PROM1 and PROM2 expression differentially modulates clinical prognosis of cancer: a multiomics analysis. Cancer Gene Ther 27, 147–167 (2020). https://doi.org/10.1038/s41417-019-0109-7

[109] Chan SW, Lim CJ, Guo K, Ng CP, Lee I, Hunziker W, Zeng Q, Hong W. A role for TAZ in migration, invasion, and tumorigenesis of breast cancer cells. Cancer Res. 2008 Apr 15;68(8):2592-8. doi: 10.1158/0008-5472.CAN-07-2696. PMID: 18413727.

[110] Sun Q, Du J, Dong J, Pan S, Jin H, Han X, Zhang J. Systematic Investigation of the Multifaceted Role of SOX11 in Cancer. Cancers. 2022; 14(24):6103. https://doi.org/10.3390/cancers14246103

[111] Finlin BS, Gau CL, Murphy GA, Shao H, Kimel T, Seitz RS, Chiu YF, Botstein D, Brown PO, Der CJ, Tamanoi F, Andres DA, Perou CM. RERG is a novel ras-related, estrogen-regulated and growth-inhibitory gene in breast cancer. J Biol Chem. 2001 Nov 9;276(45):42259-67. doi: 10.1074/jbc.M105888200. Epub 2001 Aug 31. PMID: 11533059.

[112] Hsu, Pei-Chen & Ho, Jar Yi. (2019). RERG involvement in the RAS pathway and ER-dependent transcription in breast cancer.. Journal of Clinical Oncology. 37. e14638-e14638. 10.1200/JCO.2019.37.15_suppl.e14638.

[113] Brett, J.O., Spring, L.M., Bardia, A. et al. ESR1 mutation as an emerging clinical biomarker in metastatic hormone receptor-positive breast cancer. Breast Cancer Res 23, 85 (2021). https://doi.org/10.1186/s13058-021-01462-3

[114] Iyer SV, Dange PP, Alam H, Sawant SS, Ingle AD, Borges AM, Shirsat NV, Dalal SN, Vaidya MM.

Understanding the role of keratins 8 and 18 in neoplastic potential of breast cancer derived cell lines. PLoS One. 2013;8(1):e53532. doi: 10.1371/journal.pone.0053532. Epub 2013 Jan 15. PMID: 23341946; PMCID: PMC3546083.

[115] Zhang J, Hu S, Li Y. KRT18 is correlated with the malignant status and acts as an oncogene in colorectal cancer. Biosci Rep. 2019 Aug 13;39(8):BSR20190884. doi: 10.1042/BSR20190884. PMID: 31345960; PMCID: PMC6692566.

[116] Miao, Z., Cao, Q., Liao, R. et al. Elevated transcription and glycosylation of B3GNT5 promotes breast cancer aggressiveness. J Exp Clin Cancer Res 41, 169 (2022). https://doi.org/10.1186/s13046-022-02375-5

[117] Qi J, Hu Z, Xiao H, Liu R, Guo W, Yang Z, Ma K, Su S, Tang P, Zhou X, Zhou J, Wang K. SOX10 – A Novel Marker for the Differential Diagnosis of Breast Metaplastic Squamous Cell Carcinoma. Cancer Manag Res. 2020;12:4039-404

[118] Yu, Liming & Peng, Fan & Dong, Xue & Chen, Ying & Sun, Dongdong & Jiang, Shuai & Deng, Chao. (2020). Sex-Determining Region Y Chromosome-Related High-Mobility-Group Box 10 in Cancer: A Potential Therapeutic Target. Frontiers in Cell and Developmental Biology. 8. 10.3389/fcell.2020.564740.

[119] Dravis C, Chung CY, Lytle NK, Herrera-Valdez J, Luna G, Trejo CL, Reya T, Wahl GM. Epigenetic and Transcriptomic Profiling of Mammary Gland Development and Tumor Models Disclose Regulators of Cell State Plasticity. Cancer Cell. 2018 Sep 10;34(3):466-482.e6. doi: 10.1016/j.ccell.2018.08.001. Epub 2018 Aug 30. PMID: 30174241; PMCID: PMC6152943.

[120] Al-Zahrani, K.N., Abou-Hamad, J., Pascoal, J. et al. AKT-mediated phosphorylation of Sox9 induces Sox10 transcription in a murine model of HER2-positive breast cancer. Breast Cancer Res 23, 55 (2021). https://doi.org/10.1186/s13058-021-01435-6

[121] Han W, Cao F, Gao XJ, Wang HB, Chen F, Cai SJ, Zhang C, Hu YW, Ma J, Gu X, Ding HZ. ZIC1 acts a tumor suppressor in breast cancer by targeting survivin. Int J Oncol. 2018 Sep;53(3):937-948. doi: 10.3892/ijo.2018.4450. Epub 2018 Jun 21. PMID: 29956756; PMCID: PMC6065452

[122] Zhou CL, Su HL, Dai HW. Thrombopoietin is associated with a prognosis of gastric adenocarcinoma. Rev Assoc Med Bras (1992). 2020 May;66(5):590-595. doi: 10.1590/1806-9282.66.5.590. Epub 2020 Jul 3. PMID: 32638965.

[123] Wang W, Zou W. Amino Acids and Their Transporters in T Cell Immunity and Cancer Therapy. Mol Cell. 2020 Nov 5;80(3):384-395. doi: 10.1016/j.molcel.2020.09.006. Epub 2020 Sep 29. PMID: 32997964; PMCID: PMC7655528.

[124] Yan L, He J, Liao X, Liang T, Zhu J, Wei W, He Y, Zhou X, Peng T. A comprehensive analysis of the diagnostic and prognostic value associated with the SLC7A family members in breast cancer. Gland Surg. 2022 Feb;11(2):389-411. doi: 10.21037/gs-21-909. PMID: 35284318; PMCID: PMC8899434.

[125] Tan M, Yu D. Molecular Mechanisms of ErbB2-Mediated Breast Cancer Chemoresistance. In: Madame Curie Bioscience Database [Internet]. Austin (TX): Landes Bioscience; 2000-2013.

[126] Camacho-Leal, M. del P., Sciortino, M., & Cabodi, S. (2017). ErbB2 Receptor in Breast Cancer: Implications in Cancer Cell Migration, Invasion and Resistance to Targeted Therapy. InTech. doi: 10.5772/66902

[127] Fu J, Liu S, Tan Q, Liu Z, Qian J, Li T, Du J, Song B, Li D, Zhang L, He J, Guo K, Zhou B, Chen H, Fu S, Liu X, Cheng J, He T, Fu J. Impact of TMPRSS2 Expression, Mutation Prognostics, and Small Molecule (CD, AD, TQ, and TQFL12) Inhibition on Pan-Cancer Tumors and Susceptibility to SARS-CoV-2. Molecules. 2022; 27(21):7413. https://doi.org/10.3390/molecules27217413

[128] Xiao X, Shan H, Niu Y, Wang P, Li D, Zhang Y, Wang J, Wu Y, Jiang H. TMPRSS2 Serves as a Prognostic Biomarker and Correlated With Immune Infiltrates in Breast Invasive Cancer and Lung Adenocarcinoma. Front Mol Biosci. 2022 Apr 26;9:647826. doi: 10.3389/fmolb.2022.647826. PMID: 35558557; PMCID: PMC9086397.

[129] Jach D, Cheng Y, Prica F, Dumartin L, Crnogorac-Jurcevic T. From development to cancer - an ever-increasing role of AGR2. Am J Cancer Res. 2021 Nov 15;11(11):5249-5262. PMID: 34873459; PMCID: PMC8640830.

[130] Salmans ML, Zhao F, Andersen B. The estrogen-regulated anterior gradient 2 (AGR2) protein in breast cancer: a potential drug target and biomarker. Breast Cancer Res. 2013 Apr 24;15(2):204. doi: 10.1186/bcr3408. PMID: 23635006; PMCID: PMC3672732.

[131] Zhang S, Liu Q, Wei Y, Xiong Y, Gu Y, Huang Y, Tang F, Ouyang Y. Anterior gradient-2 regulates cell communication by coordinating cytokine-chemokine signaling and immune infiltration in breast cancer. Cancer Sci. 2023 Feb 28. doi: 10.1111/cas.15775. Epub ahead of print. PMID: 36853166.

[132] Donzelli, S., Milano, E., Pruszko, M. et al. Expression of ID4 protein in breast cancer cells induces reprogramming of tumour-associated macrophages. Breast Cancer Res 20, 59 (2018). https://doi.org/10.1186/s13058-018-0990-2

[133] Nasif, D., Campoy, E., Laurito, S. et al. Epigenetic regulation of ID4 in breast cancer: tumor suppressor or oncogene?. Clin Epigenet 10, 111 (2018). https://doi.org/10.1186/s13148-018-0542-8

[134] Cheng, S., Qian, F., Huang, Q. et al. HOXA4, down-regulated in lung cancer, inhibits the growth, motility and invasion of lung cancer cells. Cell Death Dis 9, 465 (2018). https://doi.org/10.1038/s41419-018-0497-x

[135] Bhatlekar, S. et al. Identification of a developmental gene expression signature, including HOX genes, for the normal human colonic crypt stem cell niche: overexpression of the signature parallels stem cell overpopulation during colon tumorigenesis. Stem. Cells Dev. 23, 167–179 (2014).

[136] Yamashita, T. et al. Suppression of invasive characteristics by antisense introduction of overexpressed HOX genes in ovarian cancer cells. Int. J. Oncol. 28, 931–938 (2006).

[137] Gajjar K, Martin-Hirsch PL, Martin FL. CYP1B1 and hormone-induced cancer. Cancer Lett. 2012 Nov 1;324(1):13-30. doi: 10.1016/j.canlet.2012.04.021. Epub 2012 May 2. PMID: 22561558.

[138] Tang S, Liu W, Yong L, Liu D, Lin X, Huang Y, Wang H, Cai F. Reduced Expression of KRT17 Predicts Poor Prognosis in HER2high Breast Cancer. Biomolecules. 2022 Aug 25;12(9):1183. doi: 10.3390/biom12091183. PMID: 36139022; PMCID: PMC9496156.

[139] Fararjeh AFS, Al Khader A, Kaddumi E, Obeidat M, Al-Fawares O. Differential Expression and Prognostic Significance of STARD3 Gene in Breast Carcinoma. Int J Mol Cell Med. 2021 Winter;10(1):34-41. doi: 10.22088/IJMCM.BUMS.10.1.34. Epub 2021 May 22. PMID: 34268252; PMCID: PMC8256830.

[140] Lodi M, Voilquin L, Alpy F, Molière S, Reix N, Mathelin C, Chenard MP, Tomasetto CL. STARD3: A New Biomarker in HER2-Positive Breast Cancer. Cancers (Basel). 2023 Jan 5;15(2):362. doi: 10.3390/cancers15020362. PMID: 36672312; PMCID: PMC9856516.

[141] Giulianelli S, Riggio M, Guillardoy T, Pérez Piñero C, Gorostiaga MA, Sequeira G, Pataccini G, Abascal MF, Toledo MF, Jacobsen BM, Guerreiro AC, Barros A, Novaro V, Monteiro FL, Amado F, Gass H, Abba M, Helguero LA, Lanari C. FGF2 induces breast cancer growth through ligand-independent activation and recruitment of ERα and PRBΔ4 isoform to MYC regulatory sequences. Int J Cancer. 2019 Oct 1;145(7):1874-1888. doi: 10.1002/ijc.32252. Epub 2019 Mar 28. PMID: 30843188.

[142] Sahores, A., Figueroa, V., May, M. et al. Increased High Molecular Weight FGF2 in Endocrine-Resistant Breast Cancer. HORM CANC 9, 338–348 (2018). https://doi.org/10.1007/s12672-018-0339-4

[143] Kuo SJ, Chien SY, Lin C, Chan SE, Tsai HT, Chen DR. Significant elevation of CLDN16 and HAPLN3 gene expression in human breast cancer. Oncol Rep. 2010 Sep;24(3):759-66. doi: 10.3892/or_00000918. PMID: 20664984.

[144] Wang MY, Huang M, Wang CY, Tang XY, Wang JG, Yang YD, Xiong X, Gao CW. Transcriptome Analysis Reveals MFGE8-HAPLN3 Fusion as a Novel Biomarker in Triple-Negative Breast Cancer. Front Oncol. 2021 Jun 15;11:682021. doi: 10.3389/fonc.2021.682021. PMID: 34211850; PMCID: PMC8239224.

[145] Lee JY, Lee WK, Park JY, Kim DS. Prognostic value of Iroquois homeobox 1 methylation in non-small cell lung cancers. Genes Genomics. 2020 May;42(5):571-579. doi: 10.1007/s13258-020-00925-9. Epub 2020 Mar 21. PMID: 32200543.

[146] Jiang, J., Liu, W., Guo, X. et al. IRX1 influences peritoneal spreading and metastasis via inhibiting BDKRB2-dependent neovascularization on gastric cancer. Oncogene 30, 4498–4508 (2011). https://doi.org/10.1038/onc.2011.154

[147] Peluffo G, Subedee A, Harper NW, Kingston N, Jovanović B, Flores F, Stevens LE, Beca F, Trinh A, Chilamakuri CSR, Papachristou EK, Murphy K, Su Y, Marusyk A, D'Santos CS, Rueda OM, Beck AH, Caldas C, Carroll JS, Polyak K. EN1 Is a Transcriptional Dependency in Triple-Negative Breast Cancer Associated with Brain Metastasis. Cancer Res. 2019 Aug 15;79(16):4173-4183. doi: 10.1158/0008-5472.CAN-18-3264. Epub 2019 Jun 25. PMID: 31239270; PMCID: PMC6698222.

[148] Zhou L, Li H, Zhang D, Chen L, Dong H, Yuan Y, Wang T. OTX1 promotes tumorigenesis and progression of cervical cancer by regulating the Wnt signaling pathway. Oncol Rep. 2022 Nov;48(5):204. doi: 10.3892/or.2022.8419. Epub 2022 Sep 30. PMID: 36177903; PMCID: PMC9551656.

[149] Terrinoni, A., Pagani, I., Zucchi, I. et al. OTX1 expression in breast cancer is regulated by p53. Oncogene 30, 3096–3103 (2011). https://doi.org/10.1038/onc.2011.31

Table A.1: Shortlisted genes with their association with MSigDB functional analysis

Gene Symbol	CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_UP	SMID_BREAST_CANCER_LUMINAL_B_DN	SMID_BREAST_CANCER_LUMINAL_B_UP	FARMER_BREAST_CANCER_BASAL_VS_LULMINAL	DOANE_BREAST_CANCER_ESR1_DN	SMID_BREAST_CANCER_BASAL_DN	SMID_BREAST_CANCER_BASAL_UP	YANG_BREAST_CANCER_ESR1_UP	DOANE_BREAST_CANCER_ESR1_UP	VANTVEER_BREAST_CANCER_ESR1_DN
GREB1			✓	✓		✓		✓	✓	
MUC1						✓			✓	
AGR2				✓						
TOX3			✓			✓			✓	
RUNX3		✓					✓			
NDRG2		✓					✓			
PROM1		✓					✓			✓
WWTR1		✓					✓			
SOX11		✓			✓		✓			
ID4		✓					✓			
IL32		✓					✓			
CYP1B1		✓								
ESR1	✓		✓	✓		✓		✓	✓	
TTYH1		✓			✓		✓			✓
LMO4		✓		✓			✓			✓
FOLR1		✓					✓			
KRT18			✓	✓		✓		✓	✓	
KRT17		✓					✓			
LDHB		✓		✓			✓			✓
GPR160	✓									
KLK8							✓			
THSD4	✓									
PEG3							✓			
CX3CL1		✓		✓			✓			✓

Table A.1: Shortlisted genes with their association with MSigDB functional analysis

Gene Symbol	CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_UP	SMID_BREAST_CANCER_LUMINAL_B_DN	SMID_BREAST_CANCER_LUMINAL_B_UP	FARMER_BREAST_CANCER_BASAL_VS_LULMINAL	DOANE_BREAST_CANCER_ESR1_DN	SMID_BREAST_CANCER_BASAL_DN	SMID_BREAST_CANCER_BASAL_UP	YANG_BREAST_CANCER_ESR1_UP	DOANE_BREAST_CANCER_ESR1_UP	VANTVEER_BREAST_CANCER_ESR1_DN
CTNND2	✓									
SLC16A1		✓		✓			✓			
DZIP1		✓					✓			
SOX10		✓					✓			✓
NFIX							✓			
ANXA9	✓		✓	✓		✓		✓	✓	
SOSTDC1		✓					✓			
ZIC1		✓			✓		✓			
SLC44A4	✓			✓		✓			✓	
APBA2		✓					✓			
CHST3		✓			✓		✓			
GPRC5B							✓			✓
CRYAB		✓			✓		✓			
MLPH	✓		✓	✓		✓		✓	✓	
SYTL2						✓				
EN1		✓		✓	✓		✓			
IRX4							✓			
EVL	✓		✓	✓		✓			✓	
LPIN1										✓
CELSR1			✓	✓		✓			✓	
RNASE4						✓				
NXN							✓			

Table A.2: Functional analysis keyword description

CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL SAL_UP	Genes up-regulated in luminal-like breast cancer cell lines compared to the basal-like ones.
DOANE_BREAST_CANCER_ESR1_DN	Genes down-regulated in breast cancer samples positive for ESR1 [GeneID=2099] compared to the ESR1 negative tumors.
DOANE_BREAST_CANCER_ESR1_UP	Genes up-regulated in breast cancer samples positive for ESR1 [GeneID=2099] compared to the ESR1 negative tumors.
FARMER_BREAST_CANCER_BASAL_VS_LULMINAL	Genes which best discriminated between two groups of breast cancer according to the status of ESR1 and AR [GeneID=2099;367]: basal (ESR1- AR-) and luminal (ESR1+ AR+).
SMID_BREAST_CANCER_BASAL_DN	Genes down-regulated in basal subtype of breast cancer samles.
SMID_BREAST_CANCER_BASAL_UP	Genes up-regulated in basal subtype of breast cancer samles.
SMID_BREAST_CANCER_LUMINAL_B_DN	Genes down-regulated in the luminal B subtype of breast cancer.
SMID_BREAST_CANCER_LUMINAL_B_UP	Genes up-regulated in the luminal B subtype of breast cancer.
VANTVEER_BREAST_CANCER_ESR1_DN	Down-regulated genes from the optimal set of 550 markers discriminating breast cancer samples by ESR1 [GeneID=2099] expression: ER(+) vs ER(-) tumors.
YANG_BREAST_CANCER_ESR1_UP	Genes up-regulated in early primary breast tumors expressing ESR1 [GeneID=2099] vs the ESR1 negative ones.

Table A.3: Genes shortlisted for analysis and their methylation and gene expression status

		Gene shortlisted in receptor			Methylation Status			Gene Expression Status		
Gene Symbol	Oncoscore	HER2	PR	ER	HER2	PR	ER	HER2	PR	ER
ERBB2	88.61	Yes			Hypo			Up		
TMPRSS2	82.06		Yes			Hyper			Down	
GREB1	76.79		Yes	Yes		Hypo	Hypo		Up	Up
CDK6	75.3	Yes	Yes	Yes	Hyper	Hyper	Hyper	Down	Down	Down
MUC1	74.17			Yes			Hyper			Up
CDK2AP1	70.95			Yes			Hyper			Down
AGR2	70.83	Yes	Yes		Hypo	Hypo		Up	Up	
TOX3	70.04	Yes		Yes	Hypo		Hypo	Up		Up
RUNX3	69.43			Yes			Hyper			Down
NDRG2	68.7	Yes	Yes	Yes	Hyper	Hyper	Hyper	Down	Down	Down
PROM1	64.39	Yes	Yes	Yes	Hyper	Hyper	Hyper	Down	Down	Down
WWTR1	64.35			Yes			Hyper			Down
EPHB6	58.66	Yes	Yes		Hyper	Hypo		Down	Down	
SOX11	56.01		Yes	Yes		Hyper	Hyper		Down	Down
RERG	54.53		Yes	Yes		Hypo	Hypo		Up	Up
ID4	53.47	Yes	Yes		Hyper	Hyper		Down	Down	
RIPK3	49.62			Yes			Hypo			Up
HOXA4	49.47	Yes	Yes		Hyper	Hyper		Down	Down	
ZNF667	48.93	Yes			Hyper			Down		
IL32	48.45			Yes			Hyper			Down
ST8SIA6	48.07		Yes	Yes		Hypo	Hypo		Up	Up
CYP1B1	46.94		Yes			Hyper			Down	
ESR1	45.63		Yes	Yes		Hypo	Hypo		Up	Up
TTYH1	45.61	Yes	Yes		Hyper	Hyper		Down	Down	
PARD6B	44.13		Yes	Yes		Hypo	Hypo		Up	Up
LMO4	43.49	Yes	Yes		Hyper	Hyper		Down	Down	
FOLR1	43.33		Yes			Hyper			Down	
KRT18	39.92			Yes			Hypo			Up
KRT17	39.53	Yes	Yes		Hyper	Hyper		Down	Down	

Table A.3: Genes shortlisted for analysis and their methylation and gene expression status

		Gene shortlisted in receptor			Methylation Status			Gene Expression Status		
Gene Symbol	Oncoscore	HER2	PR	ER	HER2	PR	ER	HER2	PR	ER
LDHB	38.96	Yes	Yes	Yes	Hyper	Hyper	Hyper	Down	Down	Down
B3GNT5	36.22	Yes	Yes	Yes	Hyper	Hyper	Hyper	Down	Down	Down
GPR160	36.05			Yes			Hypo			Up
STARD3	34.54	Yes			Hypo			Up		
KLK8	32.71		Yes			Hyper			Down	
THSD4	31.1		Yes	Yes		Hypo	Hypo		Up	Up
PEG3	30.82	Yes			Hypo			Down		
CX3CL1	29.04		Yes			Hyper			Down	
SIRPA	28.93			Yes			Hyper			Down
CTNND2	28.69			Yes			Hyper			Up
SLC16A1	28.53			Yes			Hyper			Down
ZNF300	28.13	Yes			Hyper			Down		
DZIP1	27.69	Yes			Hyper			Down		
GCNT2	27.6	Yes	Yes		Hyper	Hyper		Down	Down	
FGF2	27.25	Yes			Hyper			Down		
SOX10	26.36	Yes	Yes	Yes	Hyper	Hyper	Hyper	Down	Down	Down
PCBP3	26.33	Yes	Yes		Hyper	Hypo		Down	Down	
HAPLN3	25.34	Yes	Yes		Hyper	Hyper		Down	Down	
NFIX	25.16	Yes			Hyper			Down		
CAPN8	24.8	Yes	Yes		Hyper	Hypo		Up	Up	
ANXA9	24.68		Yes	Yes		Hypo	Hypo		Up	Up
SOSTDC1	24.3	Yes	Yes		Hyper	Hyper		Down	Down	
ZIC1	24.16	Yes	Yes	Yes	Hyper	Hyper	Hyper	Down	Down	Down
SLC44A4	24.15			Yes			Hypo			Up
APBA2	24.01	Yes	Yes	Yes	Hyper	Hypo	Hypo	Down	Down	Down
IRX1	22.56	Yes	Yes		Hyper	Hyper		Down	Down	
THPO	22.54	Yes	Yes	Yes	Hyper	Hypo	Hypo	Up	Up	Up
CHST3	22.33	Yes			Hyper			Down		
.	21.3	Yes	Yes	Yes	Hyper	Hyper	Hyper	Down	Down	Down

Table A.3: Genes shortlisted for analysis and their methylation and gene expression status

		Gene shortlisted in receptor			Methylation Status			Gene Expression Status		
Gene Symbol	Oncoscore	HER2	PR	ER	HER2	PR	ER	HER2	PR	ER
HOXA3	20.81	Yes			Hyper			Down		
CRYAB	20.28	Yes	Yes		Hyper	Hyper		Down	Down	
OTX1	20.13	Yes	Yes		Hyper	Hyper		Down	Down	
SLC7A4	19.84		Yes	Yes		Hypo	Hypo		Up	Up
MLPH	18.89			Yes			Hypo			Up
VANGL2	18.44	Yes			Hyper			Down		
PLAT	17.91		Yes			Hypo			Up	
SYTL2	17.55			Yes			Hypo			Up
EN1	16.78	Yes	Yes		Hyper	Hyper		Down	Down	
IRX4	16.61		Yes			Hyper			Down	
RNF212	16.43	Yes			Hyper			Down		
STAC	16.01		Yes		Hyper	Hyper		Down	Down	
LFNG	14.72	Yes			Hypo			Up		
EVL	14.14			Yes			Hypo			Up
LPIN1	13.25	Yes	Yes		Hypo	Hypo		Down	Down	
MKRN3	12.74	Yes	Yes	Yes	Hyper	Hyper	Hyper	Down	Down	Down
CELSR1	12.42		Yes	Yes		Hypo	Hypo		Up	Up
RNASE4	11.53			Yes			Hypo			Up
DNAH5	9.92	Yes			Hypo			Up		
BOC	9.71		Yes			Hyper			Down	
NXN	6.62			Yes			Hypo			Down
SPEG	5.65	Yes	Yes		Hyper	Hyper		Down	Down	
KCNQ4	4.12	Yes	Yes		Hypo	Hypo		Down	Down	
SLC4A11	3.57		Yes			Hyper			Down	
SLC26A1	1.09			Yes			Hypo			Up
STOX2	0	Yes	Yes	Yes	Hypo	Hypo	Hypo	Down	Down	Down
PCDHA6	0			Yes			Hypo			Up
PCDHA10	0			Yes			Hyper			Up
FAM135A	0			Yes			Hyper			Down

Table A.3: Genes shortlisted for analysis and their methylation and gene expression status

		Gene shortlisted in receptor			Methylation Status			Gene Expression Status		
Gene Symbol	Oncoscore	HER2	PR	ER	HER2	PR	ER	HER2	PR	ER
NHLRC4	0			Yes			Hypo			Up
TTC36	0		Yes			Hypo			Up	
MT1L	0	Yes	Yes		Hyper	Hyper		Down	Down	
TCAM1P	0		Yes			Hypo			Down	
ITPRIPL1	0	Yes			Hyper			Down		
TM4SF18	0	Yes			Hyper			Down		
LRRC31	0	Yes			Hypo			Up		