Decoding the regulatory role and epiclonal dynamics of DNA methylation in 1482 breast tumours

Rajbir Nath Batra

Cancer Research UK Cambridge Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Darwin College November 2017

TO MY BELOVED GRANDFATHERS,

Lieutenant General **Raj**inder Nath Batra (1916-1995)

and

Major General **Bir**inder Singh Paintal (1928-2017)

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original, and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university.

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

This dissertation does not exceed the word limit of 60,000 words excluding figures, photographs, tables, appendices and bibliography, as specified by the Clinical Medicine Degree Committee, University of Cambridge.

Rajbir Nath Batra
November 2017

Acknowledgements

The work presented in this PhD thesis would never have come to fruition without the combined effort and support of many individuals, and it is with pleasure that I acknowledge their contributions.

Firstly, I would like to express my deepest gratitude to Professor Carlos Caldas for welcoming me into his dynamic and intellectually stimulating laboratory, and for the opportunity to experience (epi)genomics cancer research at the highest quality. I also thank him for his faith in my abilities to take on this fantastic project and entrusting me with the resources and the independence to see it through. His unrivalled scientific vision, strategic direction and mentorship have been instrumental in the progress of this project, as well as my professional development.

I would like to thank my research advisor, Dr Oscar M Rueda, for introducing me to the Caldas laboratory and for easing my transition from a mathematical statistician to a cancer scientist. I am also extremely grateful for our brainstorms and his excellent insights on the statistical and computational elements of this project.

I also thank Dr Suet-Feung Chin for adopting me as her student as well, and for taking on the unenviable role of DNA extraction for all the breast tissues and conducting the RRBS library preparation for a large proportion of the samples presented in this PhD thesis.

Both Oscar and Feung went far above and beyond their role as research advisors. They have given generously of their time, not only offering innovative insights into the project but also providing selfless support and advice in all matters during my time at Cambridge.

I would like to thank Dr Ana Tufegdzic Vidakovic for performing preliminary experiments and optimising the RRBS library preparation protocol that were vital for

initiating this project. Ana also conducted the RRBS library preparation for a large number of the breast samples presented in this thesis. I also thank her for her incisive biological interpretations.

I am very grateful to all members of the Caldas laboratory for providing an outstanding and multidisciplinary working environment. It was wonderful to belong to a group of colleagues where skill sharing and exchange of ideas was encouraged, both in the laboratory and at pubs. In particular, I acknowledge Professor Paul Edwards, Dr Alejandra Bruna and Dr Maurizio Callari for sharing their wisdom and for inspiring discussions. I also thank the Cancer Research UK Cambridge Institute, Genomics Core facility for their seamless execution of the large sequencing endeavour presented in this thesis. I thank Marion Karniely, Dr Ann Kaminski, Bethan Portlock, Elizabeth McIntyre and Joanne Heritage for all their advice and support through the administrative elements of my PhD.

I am grateful for the Wellcome Trust Mathematical Genomics and Medicine PhD programme for providing early career quantitative researchers with the academic and financial independence to contribute to modern medical research. In particular, I would like to thank Professor Simon Tavaré, Director of the Mathematical Genomics and Medicine Programme at Cambridge, and Director of the Cancer Research UK Cambridge Insitute, for his continuous guidance during my PhD. I recognise that this research would not have been possible without the financial assistance of the Wellcome Trust and the Cambridge Commonwealth Trust, and I express my gratitude to these funding bodies.

I am also very grateful for the opportunity to undertake two extremely fruitful travel fellowships during my PhD.

I thank Professor Ari Melnick and Assistant Professor Chris Mason for hosting me at their laboratories at Weill Cornell Medicine, New York, USA in Autumn 2016. This collaboration led to the investigation of epiclonal dynamics in the breast cancer cohort described in this thesis. I am also very grateful to Assistant Professor Francine Garrett-Bakelman for her guidance throughout my time at Weill Cornell. I would also like to thank the European Association for Cancer Research (EACR) for the EACR Travel Fellowship that funded this visit.

I also feel very privileged to have worked at the laboratory of Professor Amos Tanay at the Weizmann Institute of Science, Rehovot, Israel in Spring 2017. This ongoing collaboration has substantially advanced my understanding of the nuanced role of the epigenome in cancer, and has directly led to the investigation of epigenetic drift

described in this thesis. I would like to thank Aviezer Lifshitz for the many insightful discussions regarding epigenetics and computational algorithms during my time at Weizmann.

I would like to thank the breast cancer patients who enrolled in the study and contributed their tissue for cancer research. Without their voluntary participation, none of this work would have been made possible.

I would also like to thank Professor Nigel Arden, Associate Professor Kassim Javaid and Associate Professor Andrew Judge at the Nuffield Department of Orthopaedics, Rheumotology and Musculoskeletal Sciences, University of Oxford. They were instrumental in my initiation into medical research during my stint as a research statistician in Professor Arden's group after my Master's degree at Oxford.

I owe my utmost gratitude to my wife, Ankita. Her unconditional love, unwavering support and patience were indisputably the foundation upon which the past few years have been built. As a fellow laboratory member, she also volunteered to help with the RRBS library preparation which enabled the project to progress. But most importantly, I thank her for reminding me that there is much more to life than research.

I am endlessly indebted to my parents, Ranjit and Lalita Batra who have been a pillar of support and encouragement throughout my whole life. They encouraged me to have big dreams and gave me all the opportunities to chase after them. Without them, I would not be the person I am today. I also thank my sister Nayantara for being so loving, and providing a much-needed distraction during my PhD writing by visiting me in Cambridge and keeping me sane. I am also grateful to my parents in-law, Rakesh and Seema Sati, for their love, continuous counsel and support. I thank my grandmothers, Mrs Priyo Batra and Mrs Mira Paintal, for instilling in me the value of education and nurturing my interest in mathematics from a very young age.

Finally, I dedicate my thesis to my beloved grandfathers, the Late Lieutenant General Rajinder Nath Batra, and the Late Major General Birinder Singh Paintal who were exemplary men and my role models. I hope I have made you proud.

Abstract

Breast cancer is a clinically and molecularly heterogeneous disease displaying distinct therapeutic responses. Although recent studies have explored the genomic and transcriptomic landscapes of breast cancer, the epigenetic architecture has received less attention.

To address this, an optimised Reduced Representation Bisulfite Sequencing protocol was performed on 1482 primary breast tumours (and 237 matched adjacent normal tissues). This constitutes the largest breast cancer methylome yet, and this thesis describes the bioinformatics and statistical analysis of this study.

Noticeable epigenetic drift (both gain and loss of homogeneous DNA methylation patterns) was observed in breast tumours when compared to normal tissues, with markedly higher differences in late replicating genomic regions. The extent of epigenetic drift was also found to be highly heterogeneous between the breast tumours and was sharply correlated with the tumour's mitotic index, indicating that epigenetic drift is largely a consequence of the accumulation of passive cell division related errors.

A novel algorithm called DMARC (Directed Methylation Altered Regions in Cancer) was developed that utilised the tumour-specific drift rates to discriminate between methylation alterations attained as a consequence of stochastic cell division errors (*background*) and those reflecting a more instructive biological process (*directed*). Directed methylation alterations were significantly enriched for gene expression changes in breast cancer, compared to background alterations. Characterising these methylation aberrations with gene expression led to the identification of breast cancer subtype-specific epigenetic genes with consequences on transcription and prognosis.

Cancer genes may be deregulated by multiple mechanisms. By integrating with existing copy number and gene expression profiles for these tumours, DNA methylation

alterations were revealed as the predominant mechanism correlated with differentially expressed genes in breast cancer. The crucial role of DNA methylation as a mechanism to target the silencing of specific genes within copy number amplifications is also explored which led to the identification of a putative tumour suppressor gene, *THSZ2*.

Finally, the first genome-wide assessment of epigenomic evolution in breast cancer is conducted. Both, the level of intratumoural heterogeneity, and the extent of epiallelic burden were found to be prognostic, and revealed an extraordinary distinction in the role of epiclonal dynamics in different breast cancer subtypes.

Collectively, the results presented in this thesis have shed light on the somatic DNA methylation basis of inter-patient as well as intra-tumour heterogeneity in breast cancer. This complements our genetic knowledge of the disease, and will help move us towards tailoring treatments to the patient's molecular profile.

Table of contents

List of figures

List of tables

Abbreviations

5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
AML	Acute Myelogenous Leukaemia
ANOVA	ANalysis Of VAriance
ASCAT	Allele-Specific Copy Number Analysis of Tumours
AUC	Area Under the Curve
BAM	Binary Alignment Map
BCSS	Breast Cancer-Specific Survival
BMP	Bone Morphogenetic Protein
CIMP	CpG Island Methylator Phenotype
CIN	Chromosomal INstability
CLL	Chronic Lymphocytic Leukaemia
CNA	Copy Number Alteration
COSMIC	Catalogue Of Somatic Mutations In Cancer
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats

CRUK CI	Cancer Research United Kingdom Cambridge Institute
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag Sequencing
ChIP-Seq	Chromatin ImmunoPrecipitation Sequencing
DCIS	Ductal Carcinoma In Situ
DLBCL	Diffuse Large B Cell Lymphoma
DMARC	Directed Methylation Altered Regions in Cancer
DMR	Differentially Methylated Region
DNMT	DNA MethylTransferase
EGF	Epidermal Growth Factor
EMT	Epithelial to Mesenchymal Transition
EPIC	Infinium Methylation EPIC
EPM	Eloci Per Million
ER	Oestrogen Receptor
EWAS	Epigenome Wide Association Studies
FDR	False Discovery Rate
FFPE	Formalin-Fixed, Paraffin-Embedded
GATK	Genome Analysis ToolKit
GSEA	Gene Set Enrichment Analysis
HER2	Human Epidermal growth factor Receptor 2
HM27	Infinium Human Methylation 27
HM450	Infinium Human Methylation 450

HMEC	Human Mammary Epithelial Cell
HR	Hazards Ratio
ICGC	International Cancer Genome Consortium
IDC-NST	Invasive Ductal Carcinomas of No Special Type
IDH	Isocitrate DeHydrogenase
IHC	Immuno-HistoChemical
ITH	IntraTumour Heterogeneity
IntClust	Integrative Cluster
LFC	Log Fold Change
LOH	Loss Of Heterozygosity
MAR	Methylation Altered Region
MATH	Mutant-Allele Tumour Heterogeneity
MBD	Methyl-CpG-Binding Domain
MCF-7	Michigan Cancer Foundation-7
MEMo	Mutual Exclusivity MOdule
METABRIC	MolEcular TAxonomy of BReast cancer International Consortium
MSigDB	Molecular Signatures DataBase
MeDIP	Methyl-DNA ImmunoPrecipitation
NGS	Next-Generation Sequencing
OR	Odds Ratio
PAM	Prediction Analysis of Microarray

PARP	Poly (ADP-Ribose) Polymerase
PCR	Polymerase Chain Reaction
PDR	Proportion of Discordant Reads
PDTX	Patient Derived Tumour Xenograft
PE	Paired End
PFMR	Principal Functional Methylation Region
PRC	Polycomb-Repressed Chromatin
PR	Progesterone Receptor
Q-RRBS	Quantitative Reduced Representation Bisulphite Sequencing
RPMM	Recursively Partitioned Mixture Model
RRBS	Reduced Representation Bisulphite Sequencing
RefSeq	Reference Sequence
SCCRUB	Spatially Coordinated CpG-sites within the RRBS Universe in Breast cancer
SE	Single End
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SPRI	Solid Phase Reversible Immobilisation
TCGA	The Cancer Genome Atlas
TNBC	Triple Negative Breast Cancer
TOR	Time Of Replication

TSS	Transcription Start Site
UMI	Unique Molecular Identifier
VAF	Variant Allele Frequency
WGBS	Whole Genome Bisulphite Sequencing
bp	base pair
ctDNA	circulating tumour DNA
kbp	kilo base pair
oxBS-seq	oxidative BiSulphite sequencing
t-SNE	t-distributed Stochastic Neighbor Embedding

Chapter 1

Introduction

Contents

1.1 Cancer

Cancer has been notably described by Peter Nowell as an evolutionary process in which clones (groups of cells arising from a common descendent) that offer a selective advantage, are pushed towards rapid growth [Nowell, 2012]. Cancer progression, therefore, can be explained as a process of clonal diversification and adaptation under selective pressures [Greaves and Maley, 2012]. However, this inherently Darwinian-style evolution likely leads to the advancement of potent treatment resistant clones and may explain the failure of current cancer therapies [Aparicio and Caldas, 2013; Nowell, 2012]. Consideration of the evolutionary processes involved in cancer suggests that tumorigenesis is a multistep process [Foulds, 1958], determined by the acquisition of alterations leading to disruption of cell cycle regulation and attainment of self-sufficiency in cell proliferation. A linear evolution model associated with common initiating genetic mutations followed by a cumulative increase in alterations was proposed by Fearon and Volgestein [1990] in colorectal cancer. However, these findings have been challenged in breast cancer, with recent studies prescribing a more heterogeneous evolutionary model, where neighbouring clones within a tumour share common progenitor mutations but progressively acquire independent alterations in parallel [Shah et al., 2012].

The delineation of the evolutionary forces disrupted in cancer has also led to the classification of cancer driver genes into oncogenes and tumour suppressor genes. Genes whose alterations cause gain of function effects enabling autonomous and uncontrolled cell proliferation leading to the transformation of normal cells into cancer are known as *proto-oncogenes* [Adamson, 1987]. Proto-oncogenes encode proteins involved in normal cellular functions such as activating cell division and suppressing cell death. However, in the event of a gain of function alteration in these genes, they form mutated version of themselves known as oncogenes that are aberrantly expressed and drive tumorigenesis. The mechanisms associated with activating proto-oncogenes are varied. For instance, a chromosomal aberration involving reciprocal translocation between chromosome 9 and 22 leads to the formation of the BCR-ABL gene fusion [Rowley, 1973], popularly known as the Philadelphia chromosome [Nowell and Hungerford, 1960]. This oncogenic variant is associated with uncontrollable cell division and has been detected in various leukemias [Kurzrock et al., 2003]. Another example is the Ras proto-oncogene, involved in cell growth and division, which is activated by somatic point mutations in several adenocarcinomas including the pancreas, colon and the lung [Bos, 1989]. Additionally, copy number amplifications of the 17q12 amplicon harbouring *ERRB2*,

a growth factor receptor, causes overexpression of HER2 resulting in an aggressive form of breast cancer [Hynes and Stern, 1994].

In contrast, genes whose alterations cause loss of function effects contributing to the dysregulation in cell behaviour such as evasion of apoptosis are known as *tumour suppressor genes*. Tumour suppressor genes can be characterised as *gatekeepers* involved in regulating cell cycle checkpoints and inhibiting cell division in normal cells, or as *caretakers* involved in maintaining genome integrity [Kinzler and Vogelstein, 1997]. Alfred Knudson is credited with the first detection of a classical tumour suppressor gene, when he suggested that two independent mutational events within the same gene were required for the development of retinoblastoma [Knudson, 1971]. Subsequently, *RB1*, involved in cell cycle regulation, was identified as the target gene with the discovery that a somatic inactivating mutation often followed an inherited defective copy of the gene, ultimately leading to retinoblastoma in children [Cavenee et al., 1983; Comings, 1973]. Prominent examples of tumour suppressor genes in breast cancer include *TP53* [Ananiev et al., 2011] and *PTEN* [Li et al., 1997] that are also involved in cell cycle regulation. The breast cancer susceptibility genes, *BRCA1* and *BRCA2* with recognised roles in double-strand DNA break repair [Merajver et al., 1995], as well as *CDH1* which encodes E-cadherin, a protein involved in cell adhesion [Berx et al., 1998].

Historically, the main focus of cancer evolution has been on the acquisition of genetic alterations at oncogenes and tumour suppressor genes. However, in addition to genetic modifications such as somatic mutations and copy number alterations (CNAs), epigenetic aberrations (that are also heritable modifications) have been identified as molecular drivers of tumorigenesis [Baylin and Jones, 2011; Esteller, 2008; Jones and Baylin, 2002].

1.2 The role of DNA methylation in cancer

The term *epigenetics* was first coined by Conrad Waddington to describe how cells with the same genotype can differentiate into multiple phenotypes [Waddington, 1942, 1957]. Although, this description still holds true more than half a century later, the definition of epigenetic mechanisms has been updated to include mitotically heritable variations in transcription that do not modify the underlying DNA sequence. Several pioneering studies revealed that regulation of normal cellular processes is typically driven in a cell type-dependent manner, requiring a complex interplay between different layers of epigenetic information, including DNA methylation, nucleosome positions, histone modifications, and expression of noncoding RNA. These epigenetic mechanisms, together, help establish and consolidate the correct higher-order chromatin structures and gene-expression patterns during differentiation and development. Of these, DNA methylation remains the most studied epigenetic alteration in normal mammalian cells as well as in cancer. This is primarily due to two reasons: a) its strong and validated role in mammalian development, cellular differentiation and manifestation of some cancers; b) rapid technological advancements that have allowed systematic and high throughput DNA methylation analyses on a genome-scale..

1.2.1 DNA methylation is a well-balanced process in mammalian cells

DNA methylation was first discovered in 1948 [Hotchkiss, 1948] as a covalent modification of the cytosine base where a single methyl group is added to the carbon-5 position of the pyrimidine ring in post-replicative DNA. This reaction uses S-adenosyl-methionine as a methyl group donor, and is catalysed by the action of DNA methyltransferase enzymes, yielding 5-methylcytosine (5mC, DNA methylation) [Ehrlich et al., 1982], as depicted in Figure 1.1. Cytosine methylation is a widespread modification in the genome of mammalian cells that is imposed predominantly on CpG sites [Ziller et al., 2011], although non-CpG methylation is also found in certain tissues such as embryonic stem cells [Lister et al., 2009]. The human genome contains approximately 28 million CpG sites, but these are not evenly distributed [Deaton and Bird, 2011]. Rather, bulk of the genome is largely depleted of CpG sites with less than 25% of the expected frequency, due to increased mutability of methylated CpGs by spontaneous deamination [Coulondre et al., 1978]. By contrast, CpG sites can cluster in short (approximately 1 kb) CpG-rich regions, called CpG islands, which are located

in about 60% of transcription start sites of human genes (promoters), or in large repetitive DNA regions (such as centromeres and retro transposons) [Jones and Takai, 2001]. CpG islands are flanked by regions of comparatively lower CpG density, called CpG shores [Irizarry et al., 2009].

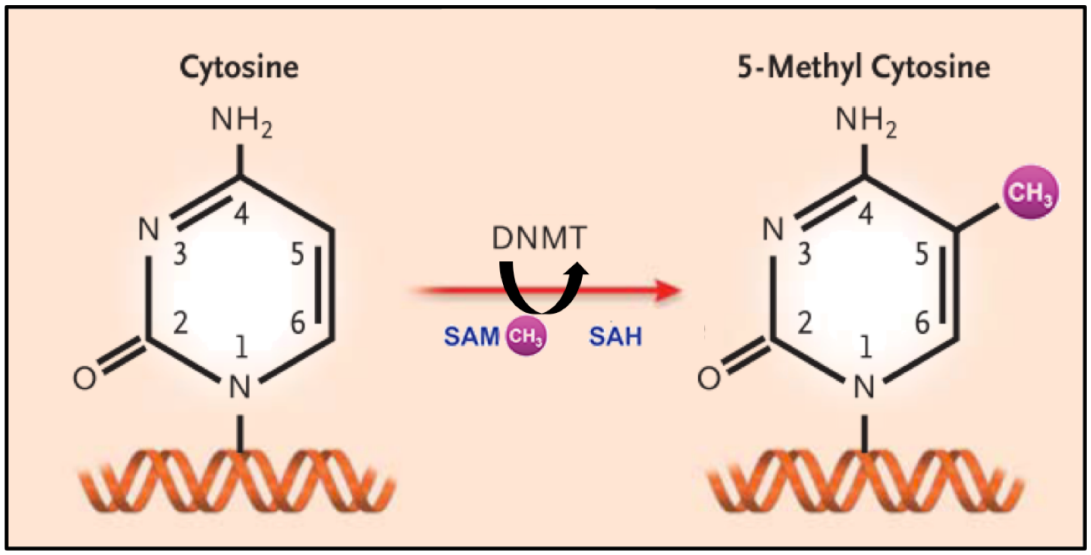


Figure 1.1: DNA methylation of cytosine in the mammalian genome. Figure adapted from Herman and Baylin [2003]. CH_3 = methyl group.

In 1975, two seminal papers [Holliday, R. & Pugh, 1975; Riggs, 1975] were the first to propose that DNA methylation could function as a potential epigenetic modification. They reported that precise DNA methylation patterns are established *de novo* during embryonic development and are mitotically heritable through multiple somatic cellular divisions in vertebrates. Since then the functional role of DNA methylation has been elucidated through a large number of studies that have shown that DNA methylation is necessary for cell development [Deaton and Bird, 2011; Suzuki and Bird, 2008], stem cell differentiation [Meissner, 2010], and control of some tissue-specific gene expression, with widespread effects on cellular growth and genomic stability [Jones and Takai, 2001]. DNA methylation is also involved in genomic imprinting, in which certain genes bypass epigenetic reprogramming resulting in their methylation states to be preserved in a parent-of-origin specific manner [Li et al., 1993].

In human genomes, the pattern of DNA methylation is cell-specific due to the tissue specific distribution of methylated and unmethylated CpG sites. However, several studies have revealed that promoter CpG islands are typically resistant to

methylation in normal cells and are associated with active gene expression during differentiation [Jones, 2012; Jones and Takai, 2001]. Unlike CpG island promoters, the role of DNA methylation within the gene body is yet to be understood. Although, extensive exonic methylation has been associated with active gene expression [Portela and Esteller, 2010], a recent report demonstrates that intragenic DNA methylation can protect the genebody from erroneous transcription initiation [Neri et al., 2017]. Beyond CpG islands, CpG-poor regions that are enriched in intergenic elements with the exception of enhancers, are typically methylated in normal cells. Similarly, CpG-poor promoters are silenced by DNA methylation and exhibit a closed chromatin structure unless gene expression is required [Stirzaker et al., 2014]. Repetitive sequences are usually methylated as well and involved in maintaining genome integrity [Jones, 2012].

As a result, DNA methylation has been increasingly recognised as a critical epigenetic alteration influencing gene regulation. Early studies established promoter DNA methylation as a key gene expression inhibitor after the discovery of two mechanisms. First, methyl-binding proteins that are recruited by the methyl group block the binding of transcriptional factors, thus acting as transcription inhibitors [Bird and Wolffe, 1999]. Second, methyl-binding proteins also recruit chromatin-remodelling factors possessing transcriptional repression domains that bind to these methylated promoters [Hendrich and Bird, 1998]. Recent evidence has also implicated non-promoter methylation such as intragenic methylation with increased transcription [Suzuki and Bird, 2008]. However, recent studies have indicated that the relationship between DNA methylation and transcription is much more nuanced than originally presumed. Which event -- DNA methylation or transcriptional regulation -- precedes the other, is still debatable. Although, some transcription factors have been demonstrated as methylation sensitive in the sense that they are inhibited from binding in the presence of DNA methylation (e.g. NRF), this mechanism is highly transcription factor and genome-context specific [Domcke et al., 2015; Feldmann et al., 2013]. Moreover, the binding (or absence) of other transcription factors (e.g. CTCF) have been shown to locally mediate altered DNA methylation levels [Lienert et al., 2011; Stadler et al., 2011]. Furthermore, experiments by Lock in the late 1980s [Lock et al., 1986, 1987] demonstrated that although DNA methylation of the *Hprt* gene followed inactivation of the X chromosome, methylation played a critical but secondary role in *locking* this inactive state. Therefore, depending on the context, DNA methylation may play an *initiating* or a *reinforcing* role in gene regulation. However, either way, DNA methylation provides an extremely informative readout of the epigenetic state of the tissue.

1.2.2 DNA methylation patterns are disrupted in cancer cells

Although, DNA methylation is a well-balanced process that exquisitely regulates gene expression in mammalian cells, the disruption of these normal epigenetic processes can have devastating consequences leading to developmental disorders as well as the initiation and progression of cancer [Gopalakrishnan et al., 2008; Robertson, 2005; Robertson and Wolffe, 2000]. Cancer was one of the first diseases in which the regulatory role of DNA methylation was determined, and altered DNA methylation patterns is a fundamental attribute observed in nearly all human cancers [Baylin and Jones, 2011; Esteller, 2008]. A schematic representation (adapted from [Stirzaker et al., 2014]) summarising DNA methylation patterns in normal cells and the deregulation observed in cancer cells is presented in Figure 1.2.

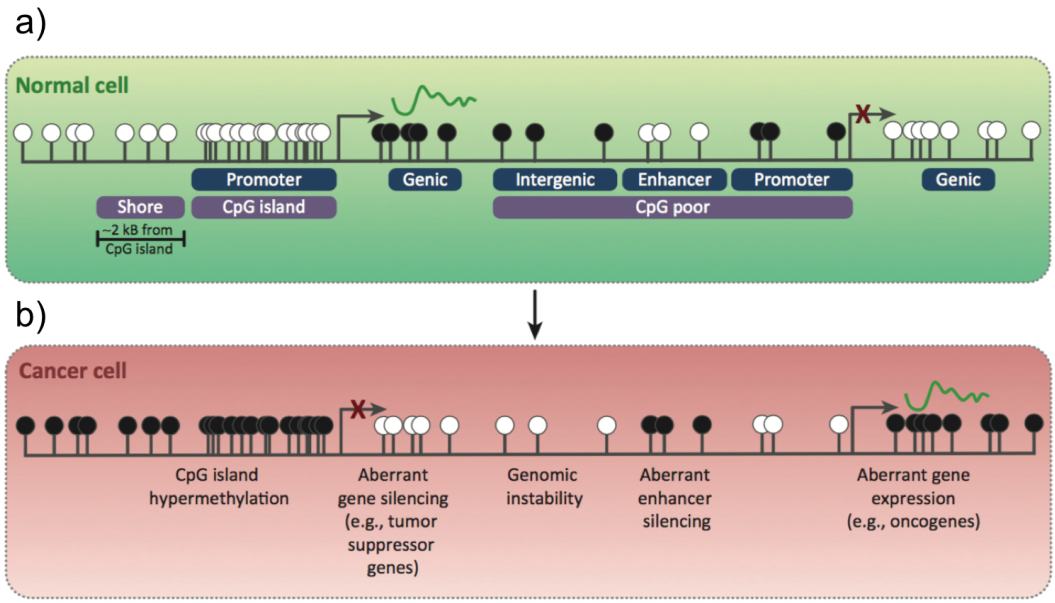


Figure 1.2: Summary of DNA methylation patterns in normal cells and the deregulation observed in cancer cells. Green squiggle = active gene expression. Red cross = Gene silencing. White circle = unmethylated CpG. Black circle = methylated CpG. Figure obtained from [Stirzaker et al., 2014].

In 1983, Feinberg and Vogelstein [1983a] observed that genomes of cancer cells have reduced levels of DNA methylation (hypomethylation) compared to normal tissues which represented the first link between DNA methylation and cancer. Promoters that were methylated and repressed in normal tissues were also found to be hypomethylated in human cancers leading to the reactivation of oncogenes including

the *ras* oncogenes [Feinberg and Vogelstein, 1983b] and *CCND2* [Oshimo et al., 2003]. Hypomethylation in tumour cells is also enriched in repetitive elements and has been shown to result in genomic instability through the expression and chromosomal rearrangements of repetitive DNA sequences, which is one of the hallmarks of cancer [Ehrlich, 2002, 2009]. In contrast, promoter CpG islands are prone to DNA hypermethylation in cancer, and often associated with the silencing of key tumour suppressor genes (such as *BRCA1* involved in DNA repair and *p16INK4a* involved in cell cycle control) [Esteller, 2000; Herman and Baylin, 2003; Jones and Baylin, 2007] which represents another hallmark of cancer. Promoter CpG shores also exhibit cancer-specific differential methylation and are associated with repression of genes including *HOX2A* [Rodríguez-Paredes and Esteller, 2011].

Although a majority of cancer studies have focused on promoter regions due to their established significance as an epigenetic transcriptional repressor, recently several reports have progressed to investigating the role of DNA methylation in other regions as well. DNA methylation in gene bodies has been shown to cause aberrant gene activation and may have an influence on deregulating alternative splicing [Kulis et al., 2012; Neri et al., 2017; Shukla et al., 2011]. Genome-scale analysis have also identified loss of methylation at enhancers in cancers presumably associated with increased transcription factor binding which can have varied transcriptional consequences for the target genes [Heyn et al., 2016; Kulis et al., 2012; Pellacani et al., 2016]. Additionally, polycomb repressive complexes (a group of repressive chromatin proteins), have also been found take part in epigenetic switching with DNA methylation; however, the exact role of DNA methylation in these regions in cancer is not clear [Baylin and Jones, 2011; Cedar and Bergman, 2009].

1.2.3 Epigenetic drift in cancer

It is well established that the DNA methylation landscape of normal cells gets altered gradually with age, and this phenomenon has been termed as epigenetic drift [Ahuja et al., 1998; Issa, 2014; Teschendorff et al., 2013]. High genome-coverage studies showed that these age-related methylation changes occurred genome-wide [Fraga et al., 2005], but are not randomly distributed [Day et al., 2013; Teschendorff et al., 2010]. A large component of this epigenetic drift has been shown to be a universal feature, independent of tissue type [Horvath et al., 2012; Teschendorff et al., 2013], and has been detected in cancer as well [Zheng et al., 2016]. But this imposes the question – what is the biological mechanism underpinning this epigenetic drift? Key experiments conducted by Yatabe et al. [2001] investigating DNA methylation patterns at neutral

loci in the stem cell population of colorectal cancer indicated an accumulation of errors in the maintenance of DNA methylation in line with the number of cell divisions. These neutral loci were specifically chosen to not have any functional consequences and were unlikely to be under selective pressures, which suggested that the accrual of DNA methylation replication errors largely contributed to the observed epigenetic drift [Kim et al., 2005; Siegmund et al., 2009].

Epigenetic drift in a tumour represents the divergence in its epigenome over time, and consequently, a higher level of epigenetic drift fuels epigenetic cellular heterogeneity which could enhance phenotypic plasticity enabling superior tumour evolution [Feinberg et al., 2006; Issa, 2011]. Consequently, this stochastic accumulation of methylation errors is also associated with increased proliferation rates and adverse clinical outcomes such as relapse or death [Teschendorff et al., 2016a].

In fact, the rate of DNA methylation replication errors (either gain or loss of methylation) has been estimated to be approximately 100 times higher than the rate of somatic mutations (Rate of DNA methylation errors within CpG dinucleotide = $\sim 10^{-5}$ per CpG site per cell division; Rate of somatic mutation within CpG dinucleotide = $\sim 10^{-7}$) [Siegmund et al., 2009]. Consequently, compared to somatic mutations, DNA methylation errors have been proposed as a superior molecular mark to serve as a mitotic clock to provide an estimate of the number of cell divisions in cancer [Issa, 2014; Kim et al., 2005]. However, although epigenetic drift is related to the number of cell divisions, it may also be acquired due to environmental exposures including inflammatory conditions and/or underlying genetic traits such as genome instability [Fraga et al., 2005; Jones et al., 2015; Teschendorff et al., 2013].

These observations collectively indicate that precancerous malignancies could carry DNA methylation signatures (called epigenetic field defects) which continue to intensify in line with the rate of mitotic divisions accompanying the initiation and progression of the cancer [Chai and Brown, 2009; Teschendorff et al., 2016a]. This has opened up the exciting possibility to use DNA methylation signatures in normal cells as a predictor of cancer risk, with promising results in cervical cancer [Teschendorff et al., 2012] and lung cancer [Teschendorff et al., 2015] among others. Recently, Yang et al. [2016] constructed a mitotic-index based epigenetic signature that was universally accelerated in cancer, in pre-invasive lesions and in normal cells at risk of neoplastic transformation, and thus could have a potential utility in diagnosis.

1.2.4 Deregulation of epigenetic modifiers in cancer

A significant discovery of genome-scale cancer studies has been the detection of mutations in epigenetic modifiers in multiple tumour types [Feinberg et al., 2016]. These include genes that are critical protagonists in the DNA methylation machinery, and therefore are associated with many downstream epigenetic alterations. For example, *DNMT3A*, which encodes a *de novo* DNA methyltransferase, is affected by somatic mutations in approximately 25% of acute myelogenous leukaemia (AML) cases [The Cancer Genome Atlas Research Network, 2013b]. These mutations have been shown to result in hypomethylation within the gene body, gene neighbourhood, and intergenic regions [Glass et al., 2017]. Moreover, mutations in ten-eleven translocation 2 (*TET2*), and isocitrate dehydrogenase 1 (*IDH1*) and 2 (*IDH2*) genes have been detected in gliomas and AML which resulted in a global hypermethylation phenotype [Abdel-Wahab et al., 2009; Figueroa et al., 2010; Noushmehr et al., 2010]. Remarkably, the mutations in *TET2* and *IDH1/ IDH2* were mutually exclusive in AML. These findings indicated involvement of these genes in a mechanism of active DNA methylation change. About the same time, seminal work discovered that the TET family of proteins are hydroxylating enzymes responsible for active DNA demethylation through enzymatic conversion of 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC) [Iyer et al., 2009; Kriaucionis and Heintz, 2009; Tahiliani et al., 2009]. TET enzymes and 5-hydroxymethylcytosine have since been established to play crucial roles in embryonic development, cellular differentiation and stem cell reprogramming [Hackett et al., 2013; Ito et al., 2010; Koh et al., 2011]. Loss of hydroxymethylcytosine is an epigenetic hallmark of cancers and both IDH and TET family of proteins have been implicated in the mechanism underlying this epigenetic deregulation [Figueroa et al., 2010; Lian et al., 2012].

1.2.5 Advancements in methylation profiling

Cancer studies focusing on methylation aberrations were initially restricted to relatively localised regions of the genome (predominantly promoter CpG rich regions). With the emergence of microarrays and next-generation sequencing (NGS) technologies, methylation-profiling strategies have been developed to perform systematic and high throughput DNA methylation analyses on a genomic-wide scale. Today, three broad genome-wide strategies for methylome profiling stand out as the most useful and popular. Each has different sample requirements, merits and challenges that are dictated by the research hypothesis, sample types and feasibility. These are discussed in depth below, and summarised in Table 1.1.

GENOME WIDE DNA METHYLATION PROFILING	Microarray-based	Immuno precipitation based	Bisulphite sequencing
Discrimination – C & 5mC	Bisulphite treatment	Antibody capture	Bisulphite treatment
NGS	Yes	No	Yes
Examples	Infinium HM27 Infinium HM450 Infinium EPIC	MeDip-seq MBDCap-Seq	WGBS RRBS PBAT (No PCR)
Compatible with low input DNA	Yes	No	Yes
Cost	£	£££	WGBS (£££££)/ RRBS (££)
Coverage	Infinium HM27 (0.1%) Infinium HM450 (~1.5%) Infinium EPIC (3%)	15-25%	WGBS (100% = 28M CpG sites) RRBS (5-10%)
Ease of bioinformatics analysis	Easy	Medium	Hard
Single nucleotide resolution	Yes	No	Yes
Quantification of DNA methylation	Absolute	Relative	Absolute
Cancer epiclonal analysis possible	No	No	Yes
Free from Copy number bias	Yes	No	Yes
Free from incomplete bisulphite sequencing bias	No	Yes	No (can be Quality Controlled)

Table 1.1: Comparison between the three broad types of DNA methylation profiling technologies. C = unmethylated cytosine. 5mC = methylated cytosine. NGS = Next Generation Sequencing.

1.2.5.1 Bisulphite microarrays

Array-based methods have been the popular choice for genome-wide DNA methylation analyses in a variety of cell types. These methylation assays commence with the bisulphite conversion of genomic DNA, followed by hybridisation to arrays that contain predesigned probes to distinguish methylated (cytosine) and unmethylated (converted to uracil) DNA fragments [Bibikova et al., 2009, 2006]. The Illumina GoldenGate Assay [Bibikova et al., 2006] and the Infinium Human Methylation 27 (HM27) BeadChip [Bibikova et al., 2009] were the first popular iterations of the microarray-based technologies that included 1536 and >27,000 cytosines respectively. However, they were quickly superseded by the widely used Infinium Human Methylation 450 (HM450) BeadChip that interrogates >480,000 cytosines across the human genome. This represents only approximately 1.7% of all CpG sites in the human genome, substantially less than other methods. These sites are enriched for CpG residues (99.3%) and located in 99% of (Reference Sequence) RefSeq gene promoters, gene bodies, and also some intergenic space [Bibikova et al., 2011]. However, the HM450 microarrays suffer from a lack of coverage at distal regulatory regions. In 2016, Illumina launched the Infinium Methylation EPIC (EPIC) BeadChip which interrogates >850,000 cytosines comprising of 90% of the HM450 probes, but significantly also including more than 350,000 probes representing potential enhancers [Pidsley et al., 2016] identified by FANTOM5 [Lizio et al., 2015] and ENCODE [Siggens and Ekwall, 2014].

With the advantage of requiring low amounts of input material, cost-effectiveness, suitability for high-throughput and ease of bioinformatics analysis, microarrays have been the most popular choice amongst researchers studying DNA methylation. However, these benefits come with the price of certain limitations. Firstly, the design is heavily biased due to preselection and inclusion of probes that interrogate only specific CpG sites that have been previously identified in methylation-based assays. Secondly, it is assumed that CpG sites located adjacent to those interrogated by the probes will be similarly methylated, which is known as the *co-methylation assumption* [Eckhardt et al., 2006]. Additionally, their use in cancer studies may also lead to biases since single nucleotide polymorphisms (SNPs) and mutations can affect the hybridisation step [Pidsley et al., 2013].

1.2.5.2 Immunoprecipitation-based technologies

Genome-wide immunoprecipitation-based methods rely on enrichment of a fraction of methylated DNA, followed by quantification by sequencing. Two common

immunoprecipitation-based approaches include using a monoclonal antibody specific for 5-methylcytosine (called methyl-DNA immunoprecipitation (MeDIP)) [Weber et al., 2005]; and affinity capture using Methyl-CpG-binding domain (MBD) proteins [Rauch and Pfeifer, 2005]. Due to biases in the different immunoprecipitation-based technologies, distinctive genomic regions are interrogated [Nair et al., 2011]. MeDIP is based on immunoprecipitation of single-stranded methylated DNA fragments and targets regions of low CpG density (e.g., intergenic regions). On the contrary, MBD isolates double-stranded fragments and favours enrichment of CpG-dense regions (e.g., CpG islands) [Robinson et al., 2010]. MBD-Seq performed on fully methylated DNA can yield approximately 18% coverage of the genome because it captures approximately 5 million methylated CpG sites [Stirzaker et al., 2014]. Approximately 30 million single-end reads are required for accurate interpretation of data.

The two key merits of immunoprecipitation-based technologies are a) achieving genome-wide coverage is relatively cheap; and b) enabling differentiation between types of DNA methylation (such as 5mC and 5hmC) by using antibodies that specifically detect one and not the other [Bock, 2012]. Conversely, a notable disadvantage of s that these techniques do not provide single-nucleotide resolution. Rather, they identify regions containing multiple methylated CpG sites typically at CpG-rich regions (for MeDIP) similar to chromatin immunoprecipitation sequencing (ChIP-Seq). Furthermore, these techniques are only marginally quantitative because the number of reads mapping to a particular region of the genome depends on the density of methylated CpG sites and is also highly susceptible to experimental biases [Stirzaker et al., 2014]. Consequently, it is more challenging to compare methylation levels between samples. Additionally, the immunoprecipitation-based techniques are prone to copy number biases, making them less suitable for cancer analyses, particularly breast and ovarian, that are dominated by CNAs.

1.2.5.3 Bisulfite sequencing techniques

Bisulphite-sequencing is considered the *gold standard* for DNA methylation analyses because it enables quantification of CpG methylation at single-base resolution [Frommer et al., 1992]. Briefly, the protocol involves treatment of the DNA with sodium bisulphite to convert cytosine to uracil, which is converted to thymine after polymerase chain reaction (PCR) amplification, whereas methylated cytosine residues are not converted and remain as cytosines due to a protective effect of their methyl group [Susan et al., 1994]. This is then followed by sequencing that enables the generation of genome-wide, single-base resolution DNA methylation maps from

bisulphite-converted DNA. After bisulphite conversion and sequencing, methylated and unmethylated cytosines will be distinguished as C and T, respectively [Frommer et al., 1992; Susan et al., 1994].

Whole Genome Bisulphite Sequencing (WGBS) is a well-established protocol that facilitates profiling of approximately 95% of all the CpG sites (28 million) in the human genome. WGBS has the advantage of whole-genome coverage, excellent reproducibility and of providing single-nucleotide resolution and thus high quantitative accuracy [Stirzaker et al., 2014]. Moreover, reads from WGBS data could be used to not only measure methylation at individual CpGs, but also to detect SNPs and mutations. Additionally, since multiple CpG sites are profiled for a large number of individual reads per sample, this sequencing method allows for estimation of intratumour heterogeneity in cancer studies [Landan et al., 2012]. Disadvantages of WGBS include the high cost and the inability to easily discriminate between 5mC and 5hmC [Booth et al., 2012]. Moreover, more than 500 million paired-end reads are required to achieve approximately 30-fold coverage [Stirzaker et al., 2014], and consequently conducting WGBS is extremely expensive. Bioinformatics analysis and accurate interpretation of bisulphite sequencing requires computational expertise [Stirzaker et al., 2014]. As a result, relatively few WGBS human cancer or related methylomes have been generated.

The increasing momentum of methylation studies, coupled with the considerable cost of whole epigenome technologies has sparked interest in the conception of more cost-effective methods. This has led to the development of genome-wide methylation analyses that are directed at specific genomic regions of interest (such as CpG rich regions) by using enrichment strategies. The archetype of this method is Reduced Representation Bisulphite Sequencing (RRBS), an efficient and high-throughput technique to analyse methylation profiles in which an enrichment strategy using restriction enzymes is combined with bisulphite sequencing to target a specific fraction of the genome, thereby reducing the per-sample cost of sequencing [Gu et al., 2010, 2011; Meissner et al., 2005]. A restriction enzyme such as *MspI* recognises and cuts the CCGG motifs that are overrepresented in high CpG content regions of the genome. Thus, RRBS is able to investigate methylation profiles at a majority of CpG islands and gene promoters, but does not interrogate intergenic or lowly methylated regions of the genome. This *reduced representation* of the genome is sequenced similarly to WGBS to generate a single-base pair resolution DNA methylation map. A minimum of 10 million sequencing reads are required for the downstream analysis of RRBS data sets, leading to approximately 3.7% actual coverage of CpG dinucleotides

genome-wide or approximately 1 million CpG sites [Gu et al., 2011; Stirzaker et al., 2014]. However, using enhanced RRBS protocols such as eRRBS described in Garrett-Bakelman et al. [2015] can yield approximately 3-4 million CpG sites.

RRBS offers most of the advantages of WGBS. Furthermore, it is more cost-effective since it targets bisulphite sequencing in an enriched population of the genome, while retaining single-nucleotide resolution. RRBS also requires significantly lower input DNA than WGBS, and can be extensively applied to model systems with limited DNA material availability [Bock, 2012]. Despite being biased towards CpG-rich regions, it has allowed discovery of novel biomarkers in addition to those identified using methylation arrays. Provided the coverage is sufficient, reads from RRBS data could also be used to detect single nucleotide variants (SNVs) [Gu et al., 2010], allowing simultaneous detection of genetic and epigenetic information in important regulatory regions of the genome that RRBS covers. Being a sequencing-based technology similar to WGBS, it can also capture epiallelic composition of the samples allowing the estimation of intratumour heterogeneity in cancer studies [Landan et al., 2012]. However, a lack of coverage at CpG-poor intergenic and distal regulatory elements is a potential disadvantage of the method. Similar to WGBS, it also suffers from a difficulty to differentiate between 5mC and 5hmC.

A limitation of bisulphite sequencing techniques is that the bisulphite treatment leads to considerable fragmentation of DNA, and therefore it typically requires relatively large quantities of DNA (1–5 µg for WGBS, lower for RRBS) followed by PCR amplification. Bisulphite conversion followed by PCR amplification also results in a highly skewed AT/GC composition which can introduce biases in the quantification of methylation [McInroy et al., 2016]. In RRBS, this bias is magnified due to the inability to discriminate between PCR-induced duplication artefacts or distinct molecular copies of fragments since the start and end sites of reads are largely driven by restriction enzyme digestion. However, two PCR-free protocols have since been established: post-bisulphite adaptor tagging (PBAT) [Miura et al., 2012] and recovery after bisulphite treatment (ReBuilT) [McInroy et al., 2016]. Consequently, these protocols result in lower levels of duplication as well as reduced biases in sequence context compared to bisulphite sequencing experiments that require PCR amplification. The PCR-duplicate predicament in RRBS can also be circumvented through the use of unique molecular identifiers (UMI) as demonstrated in the quantitative RRBS (Q-RRBS) method established by Wang et al. [2015]. These methods have also enabled the development of single-cell genome-wide bisulphite sequencing techniques [Farlik et al., 2015; Guo et al., 2013; Smallwood et al., 2014].

1.3 Heterogeneity in breast cancer

Breast cancer is an uncontrolled growth of cells that originates from the breast tissue. If left untreated, the tumour can invade surrounding tissues or metastasise to distant areas of the body leading to death. Breast cancer is one of the leading causes of mortality in women with greater than 450,000 deaths each year worldwide [Cancer Genome Atlas Network, 2012]. Most breast cancer patients (80%) are above the age of fifty, but younger women, and very rarely, men, are also diagnosed with breast cancer.

Over the past 25 years, there have been substantial improvements in overall survival outcomes, primarily due to developments in early diagnosis, early surgical interventions and efficacy of cytotoxic and targeted biological therapies [Peto et al., 2000]. Despite these advancements, a great challenge in the clinic is that breast cancer is not one single disease, but a group of heterogeneous entities with diverse variations in molecular and clinical outcomes [Dawson et al., 2013]. As a result, therapies that are not targeted may subject some patients to unnecessary toxicity and inferior efficacy. In fact, present clinical management causes overtreatment of 50% of breast cancer patients [Early Breast Cancer Trialists' Collaborative Group (EBCTCG), 2005], with implications on quality of life of patients as well as healthcare costs. At the same time, intrinsic or acquired tumour resistance to treatment leads to disease progression towards incurable metastatic disease in a significant proportion of patients.

There is a pressing need to expand our knowledge of the genetic and epigenetic aberrations underlying breast cancer, which would be vital in delineating *inter*-patient as well as *intra*-tumour heterogeneity. Successful translation of these findings into the clinic will move us closer towards the implementation of a more targeted, personalised approach to breast cancer medicine.

1.3.1 Histopathological evaluation of breast cancer

The current clinical management of breast cancer involves surgical excision of the tumour in combination with therapies administered either before (neoadjuvant) or after (adjuvant) the surgery. Therapies include radiotherapy, cytotoxic chemotherapy, endocrine therapy or molecular targeted therapies. Following surgery, therapeutic decisions are predominantly based on histopathological assessments of the surgical biopsy that involve a combination of morphological examinations alongside identifying immuno-histochemical (IHC) markers [Ali et al., 2014; Dawson et al., 2013]. These histopathological examinations have proven to have prognostic and predictive power, and hence remain the foundation of current clinical management of breast cancer.

1.3.1.1 Morphological and anatomical classification

Morphological features such as tumour grade and anatomical features such as tumour size and lymph node status, all form key tools as part of the histopathological arsenal. Tumour grade provides an assessment of its intrinsic characteristics including differentiation and proliferative activity [Elston et al., 1999]. Breast tumours are scored on a scale of 1-3, with 3 representing high-grade tumours. High-grade tumours have been shown to be correlated with high levels of the nuclear antigen Ki-67, a marker of cell proliferation and is associated with worse survival [Contesso et al., 1987]. Lymph node status and tumour size provide information related to the extent and stage of the disease. Lymph node status is typically defined as the number of lymph nodes in which the breast cancer has invaded, and is presently the most powerful prognostic factor in breast cancer [Carter et al., 1989]. Tumour size has also been shown to be a robust and independent prognostic factor in breast cancer [Carter et al., 1989]. Younger patient age at diagnosis has also been shown to be associated with worse prognosis for premenopausal women, independent of other histopathological parameters.

The morphological assessment also includes allocation of the histological subtype according to the tumour's architectural patterns and a semi-quantitative evaluation of its cytological characteristics [Elston et al., 1999]. A majority (75%) of breast cancers show no special characteristics and are defined as invasive carcinomas of no special type (NST) [Sinn and Kreipe, 2013]. There at least 17 special histological subtypes that are characterised by distinct microscopic appearances, histological growth patterns and diverse clinical courses [Wellings SR, Jensen HM, 1975]. They are relatively uncommon and include ductal carcinoma in situ; lobular carcinoma in situ; invasive lobular carcinoma; tubular carcinoma; medullary carcinoma; and mucinous carcinoma.

1.3.1.2 Histological molecular markers

The discovery of two key IHC markers, oestrogen receptor (ER) and human epidermal growth factor receptor 2 (HER2) provided a breakthrough in breast cancer research, and set the stage for the first targeted therapies in cancer.

Oestrogen receptor status

Oestrogen receptor (ER) expression status is a keystone of breast cancer management [Knight et al., 1977; Riggs and Hartmann, 2003], and is typically assessed using IHC [Harvey et al., 1999]. Most breast cancer patients have high levels of ER (75%, ER+ breast cancer), and are generally older patients with smaller

tumours [Anderson et al., 2002]. On the contrary, ER- breast cancers are more commonly diagnosed in younger patients. ER markers are widely applied to aid the selection of breast cancer patients for adjuvant hormonal based treatments such as tamoxifen, a non-steroid oestrogen antagonist in breast tissues [Fisher et al., 1998; Jensen et al., 1971]; and aromatase inhibitors that reduce the levels of oestrogen by interfering with its production [Bonneterre et al., 2000; Riggs and Hartmann, 2003]. These endocrine therapies have been associated with a striking improvement in survival predominantly in ER+ patients [Early Breast Cancer Trialists' Collaborative Group (EBCTCG), 1998, 2005]. ER+ patients generally experience a consistent risk of mortality over time, whereas ER- patients have a much higher risk of relapse and death within the first five years after diagnosis, after which the risk diminishes [Dent et al., 2007]. The rationale for this time-dependence is still unclear.

Progesterone receptor status

Progesterone receptor (PR) is an oestrogen-regulated hormone [Lee and Gorski, 1996] which has been shown to inhibit oestrogen-fuelled growth and tumour progression [Mohammed et al., 2015]. Most ER+ tumours are also PR+, although about 25% of all breast tumours are ER+/PR-, and is has been argued that this represents a distinct clinical and biological entity compared to ER-/ PR+ tumours [Rakha et al., 2007].

HER2 status

The overexpression and/ or amplification of the protein Human Epidermal growth factor Receptor 2 (HER2) is found in 12-15% of breast cancers with approximately half of these co-expressing hormone receptors [Curtis et al., 2012; King CR, Kraus MH, 1985; Konecny et al., 2003]. HER2 belongs to a family of transmembrane receptor tyrosine kinases that are activated by the binding of growth factor ligands such as the epidermal growth factor (EGF), which triggers receptor dimerization and activation of downstream cellular processes including proliferation [Baselga, 2010; Rowinsky, 2004]. Not only do *HER2* amplifications have great prognostic value in predicting survival and time to relapse [Dressman et al., 2003], this discovery also led to the development of anti-HER2 based therapies. The most prominent is trastuzumab (Herceptin), a monoclonal antibody that interferes with the HER2 receptor, which in combination with chemotherapy provides significant clinical benefit to HER2-positive breast cancer patients [Tan, 2003].

Breast cancers that are hormone receptor negative (ER- and PR-) as well as HER2 negative are classified as Triple Negative Breast Cancers (TNBCs) and account for

the remaining 12-17% of cancers [Foulkes et al., 2010]. TNBCs represent the least favourable IHC-based subtype of breast cancer in terms of prognosis due to the lack of therapeutic options available today [Ali et al., 2014; Dawson et al., 2013; Dent et al., 2007].

1.3.2 Molecular classification of breast cancer

The use of the histopathological assessments in conjunction with prognostic indices such as the Nottingham prognostic index (NPI) [Galea et al., 1992] and PREDICT [Wishart et al., 2010] are routinely performed in the clinic and provide great prognostic and predictive value in clinical management. However, despite these advancements, breast cancer patients with similar histopathological/ clinical features still exhibited significant variation in therapeutic response and survival [Dawson et al., 2013]. A probable explanation for this is that these assessments provide inadequate insight into the underlying molecular mechanisms and biological pathways that drives breast cancer heterogeneity.

1.3.2.1 Gene expression profiling - the Intrinsic subtypes

The emergence of microarray technology heralded the way for comprehensive molecular profiling of breast cancer towards the aim of revealing the unexplained heterogeneity. Pioneer gene expression profiling studies in breast cancer led to the identification of four distinct Intrinsic subtypes -- Luminal A, Luminal B, Basal-like and HER2-enriched -- and a normal breast-like group (called Normal-like) [Perou et al., 2000; Sorlie et al., 2001, 2003]. These molecular subtypes of breast cancer were discovered by focusing on an *intrinsic* gene set of 500 unique genes (initially 1753 genes), that had significantly higher variation between tumours from different patients than between successive samples from the same patient's tumour [Sorlie et al., 2003], and then performing unsupervised hierarchical clustering. These Intrinsic *subtypes* not only reflected pervasive transcriptomic differences underlying the cell biology of breast cancer, but importantly also showed significant variation in clinical outcomes including incidence, survival and therapeutic response [Sorlie et al., 2003]. A noteworthy finding of this study was that ER+ breast carcinomas (that arise from luminal epithelial cells), actually encompass two separate biological subtypes: Luminal A and Luminal B. They have distinctive gene expression profiles, with the Luminal A subtype having higher expression of *ESR1* and *GATA3*, whereas the Luminal B subtype had lower expression of luminal cell genes and exhibited higher levels of proliferation. Luminal B tumours also associated with higher grade and

appreciably worse prognosis [Sorlie et al., 2003; Sotiriou et al., 2003]. The other two Intrinsic Subtypes were characterised by ER-negativity with poor patient outcomes. HER2-enriched subtype is defined by dominant HER2 expression, whereas the Basal-like subtype shows characteristics of basal cells in the mammary epithelium. The Normal-like subtype is most similar to the normal mammary tissue with respect to its gene expression profile.

The evolution of gene expression studies has led to many similar analyses that have proposed additional subtypes and molecular signatures that could accurately predict disease outcome [Melisko, 2005; Teschendorff et al., 2007; van 't Veer et al., 2002]. Particularly, a sixth breast cancer subtype was discovered called Claudin^{low} that closely resembles mammary epithelial stem cells and is characterised by low expression levels of claudins and *CDH1* [Herschkowitz et al., 2007; Prat et al., 2010]. These tumours are predominantly ER, PR and HER2-negative, and reflect an intermediate prognosis between Luminal and Basal-like tumours. Parker et al. [2009] developed a single sample classifier called the Prediction Analysis of microarrays 50 (PAM50) to prospectively classify breast cancer tumours into the aforementioned Intrinsic subtypes. The PAM50 classifier uses a quantitative reverse transcriptase polymerase chain reaction (qRT-PCR) assay to measure the expression of 50 genes followed by the centroid-based Prediction Analysis of Microarrays (PAM) method [Tibshirani et al., 2002]. It's utility as a single sample classifier, the ability to apply it on formalin-fixed, paraffin-embedded (FFPE) tissue, as well as the reduced number of genes to consider has led to the widespread adoption of the Intrinsic subtype classification in breast cancer management.

These efforts have also prompted the development of multigene tests that employ distinct prognostic signatures for individual risk assessment in the clinic. In addition to PAM50, other prognostic tools include Mammaprint, OncotypeDX and Breast Cancer Index, that have been validated in clinical trials to establish their use in the clinic [Cardoso et al., 2008; Melisko, 2005; Sotiriou et al., 2006; Sparano and Paik, 2008; van de Vijver et al., 2002]. Although these attempts provided early insights into the molecular heterogeneity, they did not take into account the underlying molecular aberrations that might explain these transcriptomic variations, and ultimately subtype specificity. This prompted the development of multi-level genomic characterisation of breast cancer.

1.3.2.2 Profiling genomic alterations in breast cancer

Tumorigenesis as well as subtype-specific variations in gene expression profiles are heavily influenced by the underlying genomic architecture of breast cancers. Both inherited genetic variants and acquired genetic alterations contribute to the tumours' genetic diversity. However, it is the accumulation of somatic genomic changes (including somatic mutations and copy number aberrations) that enable the initiation and progression of breast carcinomas [Albertson et al., 2003]. Recurrent genomic aberrations can contribute to tumorigenesis by deregulating genes linked with the development of cancer pathophysiology such as evasion of apoptosis, sustained angiogenesis and limitless potential for proliferation [Aguirre-Ghiso et al., 2003; Hanahan and Weinberg, 2011]. Discovery of these somatic alterations, followed by differentiation between driver and passenger events amongst them would prove to be a key strategy in identifying these genes related to the initiation and progression of breast cancer.

The advent and rapid development of NGS technologies, combined with the continued use of older microarray technologies led to the realisation of a plethora of cancer genomics projects. These projects were designed to evaluate multi-level molecular characterisation of several cancer types, including breast cancer with the aim of fulfilling the aforementioned strategy of identifying driver genes, and are delineated below.

In breast cancer, the landscape of somatic alterations is dominated by copy number alterations; and well-defined genomic rearrangements have been characterised in breast cancer [Ciriello et al., 2013; Stephens et al., 2009]. Copy number profiling has also been linked with gene expression changes to improve the classification of the Intrinsic subtypes of breast cancer [Chin et al., 2006, 2007]. The emergence of low cost and deep whole-genome and whole-exome sequencing has also enabled the comprehensive characterisation of the mutational landscape of breast cancer, leading to the identification of genes previously implicated in breast cancer as well as novel frequently mutated genes [Banerji et al., 2012; Cancer Genome Atlas Network, 2012; Ellis et al., 2012; Nik-Zainal et al., 2016; Pereira et al., 2016; Shah et al., 2012; Stephens et al., 2012]. Several of these genes displayed Intrinsic subtype specific mutation patterns that could be linked into cellular pathways underlying tumour biology and clinical outcomes. Investigating the landscape of somatic alterations also enabled characterisation of the distribution of clonally dominant somatic mutations [Eirew et al., 2014; Nik-Zainal et al., 2012; Shah et al., 2012]. This allowed exploration into the dynamics of genomic clones and the role of intratumour heterogeneity in evolution.

1.3.2.3 Integrated analysis of copy number and gene expression - the Integrative clusters

The discovery efforts of the first-generation sequencing studies mentioned above signaled the move to identifying driver events in breast cancer as the optimum strategy for molecular classification. An extensive examination of 2000 breast cancers as part of the METABRIC (MolEcular TAxonomy of BReast cancer International Consortium) consortium involved the integration of the genomic (copy number) and transcriptomic landscapes of the disease [Curtis et al., 2012]. METABRIC represents the largest cohort of breast tumours on which molecular profiling has been undertaken, and this seminal study illuminated the profound impact that somatic CNAs have on gene expression in breast cancer. This strategy led to the identification of 1000 genes with *cis*-acting CNA (on gene expression) and used the integrative cluster method [Shen et al., 2009, iCluster] and internal validation to reveal 10 Integrative clusters (IntClusts 1-10). These clusters yielded a refinement of the molecular classification of the disease towards a driver-based taxonomy and demonstrated the extensive heterogeneity present within tumours classified according to IHC-subtypes and Intrinsic subtypes. Furthermore, each Integrative cluster was associated with different clinical features and survival outcomes. A complete description of the molecular and clinical characteristics of the Integrative clusters are provided in Dawson et al. [2013], and summarised in Table 1.2 (modified from [Dawson et al., 2013]). Since then, the tumours in IntClust 4, a genomically quiet cluster, have been further stratified based on ER status (IntClust 4ER+ and IntClust 4ER-), increasing the total number of clusters to 11. To allow for single-sample prediction of the Integrative clusters in external datasets, Ali et al. [2014] developed a 614 gene-based classifier to assign breast cancer IntClust classifications solely from gene expression data. Classifying the original METABRIC dataset (validation cohort: 983 patients) into the Integrative clusters using this gene expression-based approach produced a concordance of 98% when compared to the original integrative copy number and gene expression method described in Curtis et al. [2012].

IntClust	Frequency (*n*, %)	Defining molecular features	Expression (*n*, %)	PAM50 (*n*, %)	Clinical features	Prognosis (5-year, 10-year DSS)	Genomic instability
1	139 (7%)	17q23 amplification	ER+: 123 (88.49%) PR+: 60 (43.17%) HER2+: 20 (14.39%)	Basal: 9 (6.47%) HER2: 21 (15.11%) LumA: 11 (7.91%) LumB: 90 (64.75%) Normal: 8 (5.76%)	High grade	Intermediate 0.80, 0.69	High
2	72 (4%)	11q13/14 amplification	ER+: 69 (95.83%) PR+: 51 (70.83%) HER2+: 3 (4.17%)	Basal: 2 (2.78%) HER2: 6 (8.33%) LumA: 25 (34.72%) LumB: 36 (50%) Normal: 3 (4.17%)	No distinct clinical features	Poor 0.78, 0.51	High
3	290 (15%)	Paucity of copy number changes	ER+: 278 (95.86%) PR+: 211 (72.76%) HER2+: 1 (0.34%)	Basal: 4 (1.39%) HER2: 9 (3.14%) LumA: 195 (67.94%) LumB: 43 (14.98%) Normal: 36 (12.54%)	Low grade Low LN+	Good 0.93, 0.88	Low
4	343 (17%)	CNA devoid	ER+: 238 (69.39%) PR+: 155 (45.19%) HER2+: 20 (5.83%)	Basal: 64 (18.71%) HER2: 34 (9.94%) LumA: 106 (30.99%) LumB: 29 (8.48%) Normal: 109 (31.87%)	Low grade	Good 0.89, 0.76	Low
5	190 (10%)	ERBB2 amplification	ER+: 79 (41.58%) PR+: 40 (21.05%) HER2+: 181 (95.26%)	Basal: 21 (11.05%) HER2: 108 (56.84%) LumA: 18 (9.47%) LumB: 33 (17.37%) Normal: 10 (5.26%)	Younger age at diagnosis High grade High LN+	Poor 0.62, 0.45	Intermediate
6	85 (4%)	8p12 amplification	ER+: 85 (100%) PR+: 36 (45.88%) HER2+: 3 (3.53%)	Basal: 3 (3.53%) HER2: 10 (11.76%) LumA: 23 (27.06%) LumB: 43 (50.59%) Normal: 6 (7.06%)	No distinct clinical features	Intermediate 0.83, 0.59	High
7	190 (10%)	16p gain, 16q loss, 8q amplifcation	ER+: 187 (98.42%) PR+: 150 (78.95%) HER2+: 2 (1.05%)	Basal: 3 (1.59%) HER2: 9 (4.76%) LumA: 123 (65.08%) LumB: 41 (21.69%) Normal: 13 (6.88%)	Older age at diagnosis Low grade	Good 0.94, 0.81	Intermediate
8	299 (15%)	1q gain, 16q loss	ER+: 297 (99.3%) PR+: 236 (78.93%) HER2+: 1 (0.33%)	Basal: 1 (0.33%) HER2: 9 (3.01%) LumA: 192 (64.21%) LumB: 89 (29.77%) Normal: 8 (2.68%)	Older age at diagnosis Low grade	Good 0.88, 0.78	Intermediate
9	146 (7%)	8q gain, 20q amplification	ER+: 125 (85.62%) PR+: 79 (54.11%) HER2+: 10 (6.85%)	Basal: 20 (13.79%) HER2: 26 (17.93%) LumA: 24 (16.55%) LumB: 70 (48.28%) Normal: 5 (3.45%)	High grade	Intermediate 0.78, 0.62	High
10	226 (11%)	5q loss, 8q gain, 10p gain, 12p gain	ER+: 25 (11.06%) PR+: 19 (8.41%) HER2+: 6 (2.65%)	Basal: 202 (89.38%) HER2: 8 (3.54%) LumA: 1 (0.44%) LumB: 14 (6.19%) Normal: 1 (0.44%)	Younger age at diagnosis High grade Large tumours	Poor 0.71, 0.68	Intermediate

Table 1.2: Molecular and clinical features of the Integrative clusters. Table obtained from Dawson et al. [2013]. Frequencies and %s for the METABRIC dataset described in Curtis et al. [2012]. IntClust = Integrative cluster. DSS = disease-specific survival. LN+ = lymph-node involvement.

Subsequently, the contributions of microRNAs [Dvinge et al., 2013] and somatic point mutations [Dawson et al., 2013; Pereira et al., 2016] towards explaining the heterogeneity of breast cancer have also been explored in the METABRIC dataset. It has become quite clear that breast cancer subtypes possess different molecular wiring, but insights into the mechanism of tumorigenesis need to be extended beyond genomic alterations. The epigenome has long been understood as compelling regulators of tumorigenesis [Jones and Baylin, 2002, 2007]. However, the epigenetic landscape of breast cancer has received far less attention in comparison to its genomic counterparts.

1.3.3 Methylome studies in breast cancer

Epigenetic cancer studies prior to the recent technological advances in DNA methylation profiling focused on a small number of genes (such as *BRCA1*), and provided a limited representation of the breast cancer methylome [Berman et al., 2005; Dickinson et al., 2004; Esteller, 2000]. They demonstrated that methylation configurations are varied in different types of tumours, as well as between cancerous and stromal tissues. The advent of superior DNA methylation technologies, coupled with advances in bioinformatics, have enabled genome-wide investigation of DNA methylation patterns in breast cancer, which has further led to a quest for epigenetic regulators of tumorigenesis and prognostic/ predictive biomarkers.

Several studies have successfully examined the relationship between the breast cancer methylome and disease taxonomy and prognosis. Li et al. [2010] discovered that ER/ PR status was associated with the methylation levels of four genes. Fackler et al. [2011] further illustrated the potential of DNA methylation patterns to accurately predict the ER status of a tumour, besides identifying a CpG methylation signature that was significantly associated with breast cancer progression . Three studies using selected gene promoter panels revealed that the some of the Intrinsic subtypes (gene expression-based) of breast cancer harbour characteristic DNA methylation profiles [Bediaga et al., 2010; Holm et al., 2010; Rønneberg et al., 2011]. In particular, Luminal B tumours were associated with highest promoter methylation, Luminal A tumours with intermediate methylation, and Basal-like tumours harboured the lowest promoter methylation. HER2-enriched and Normal-like tumours did not exhibit a significant association with methylation derived subtypes in these studies. Eventually, these Intrinsic subtype-specific methylation patterns were largely verified in genome-wide studies with significantly higher sample sizes [Cancer Genome Atlas Network, 2012; Stefansson et al., 2015].

A study by Flanagan et al. [2010] determined that genome-wide methylation profiles could accurately predict the mutation status of breast cancer tumours, and identified a global hypomethylation signature that was associated with a novel subgroup of *BRCA1* mutated tumours. Conversely, *BRCA2* mutations were linked with promoter hypermethylation [Holm et al., 2010] suggesting that the BRCA family of proteins may participate in the regulation of DNA methylation.

Further DNA methylation profiling in breast cancers led to the discovery of new methylation-based subtypes with strong prognostic potential. Fang et al. [2011] identified a breast CpG island methylator phenotype (denoted by CIMP), characterised by coordinated methylation of a group of CpG islands, which was a strong determinant of breast cancer metastasis. The CIMP-positive phenotype was later found to be enriched in invasive lobular breast tumours [Roessler et al., 2015].

In 2011, Dedeurwaerder et al. [2011] examined genome-wide DNA methylation profiles in the largest breast cancer methylome study conducted up until that point (248 tumours), that revealed the existence of novel breast cancer subtypes not classified by current expression subtypes suggesting that methylation profiling might reflect the cellular origin of breast cancer cells. Furthermore, it highlighted an immune gene set associated with high prognostic value. A similar immune signature was revealed in a pathway-based integrated approach in breast cancer combining genome wide DNA methylation patterns with gene expression, microRNA and DNA copy number [Kristensen et al., 2012]. However, this study determined that the contribution of DNA methylation profiles was minimal, potentially due to the choice of the methylation platform which only profiled 1505 CpG sites.

The studies discussed above were able to expand the breast cancer classification to identify unique features that were not considered relevant through gene expression analysis, but they suffered from a lack of sufficient power. Most of the studies reported up until 2012 profiled less than 100 breast tissues (Dedeurwaerder et al. [2011] was the exception with 248 samples), and employed methylation profiling technologies that assayed only a limited number of CpG sites (up to $\sim$27,000) in the genome. However, in 2012, The Cancer Genome Atlas (TCGA) Network assayed 802 breast tumours using Infinium DNA HM450 methylation arrays, in addition to copy number, gene expression, exome-sequencing and microRNA profiling [Cancer Genome Atlas Network, 2012; Ciriello et al., 2015]. This represented the first large multiplatform analysis with epigenetic and genetic data in breast cancer, and was followed by an ER+ breast cancer study (n = 560) conducted by the International Cancer Genome Consortium (ICGC) in 2015 [Nik-Zainal et al., 2016]. Although, the

breast cancer methylome was not explored in depth in either of these two studies, the public availability of the TCGA data in particular, sparked an interest in integrating data across multiple molecular platforms to identify epigenetic drivers of tumorigenesis.

In 2015, Gao et al. [2015] performed a novel systems-level integrative analysis of the DNA methylation and transcriptomic landscapes using ER+ tumours from the TCGA breast cancer consortium dataset [Cancer Genome Atlas Network, 2012]. They identified nine epigenetic hotspots that were functionally deregulated in ER+ tumours including the W NT bone morphogenetic protein (BMP) signalling pathways. Although, the ER+ tumours were found to be remarkably homogeneous with respect to these epigenetic hotspots, Luminal B tumours exhibited higher deregulation. In the same year, investigation of the gene expression and methylation profiles in the TNBC sub-cohort of the TCGA dataset revealed three clusters with distinct prognosis [Stirzaker et al., 2015]. Aure et al. [2013] additionally examined microRNA expression in the TCGA breast cancer dataset, and revealed that DNA methylation along with copy number alterations influence mechanisms underlying microRNA dysregulation.

A shortcoming of the vast majority of the breast cancer methylome studies mentioned above, is that the focus has largely been retained on promoter regions due to their established significance as an epigenetic transcriptional repressor. Recent reports in breast cancer have progressed to investigating gene body methylation as well [Fleischer et al., 2014; Györffy et al., 2016]. These studies generally demonstrated a positive correlation between DNA methylation and gene expression in the gene body regions; however, they restricted their analysis to a selected panel of genes. Additionally, recent investigations have also explored the role of DNA methylation in enhancers in normal mammary tissues [Pellacani et al., 2016] and breast tumours [Fleischer et al., 2017; Heyn et al., 2016]. These studies demonstrated that reduced DNA methylation at enhancers was associated with increased transcription factor binding which can have transcriptional consequences for the target genes. Furthermore, Holm et al. [2016] identified seven novel breast cancer clusters with distinct methylation patterns that were correlated with the underlying chromatin states in normal mammary cells. This study suggested that DNA methylation profiles are informative indicators of the overall epigenetic state of breast tissues.

Despite multiple studies attempting to map breast cancer-associated epigenetic changes at the genome-wide level, not much evidence of the dynamics of methylation change and intratumour heterogeneity have been provided in breast cancer. Studies in chronic lymphocytic leukaemia (CLL) and diffuse large B-cell lymphomas (DLBCL) have emerged that have indicated that methylation patterns within cancer tissues are

highly heterogeneous and polymorphic [Landau et al., 2014; Pan et al., 2015]. This *epipolymorphism* has also been linked to adverse clinical outcomes such as relapse. Further, Li et al. [2014] developed a combinatorial entropy change calculation to identify loci that alter significantly in epiallelic compositions between two time points in the same sample. They applied this method to acute myeloid leukemia (AML) patients with paired observations at diagnosis and relapse, and observed that epiallelic compositions varied considerably during disease progression, and that a high degree of epigenetic allelic burden within a tumour was linked with adverse clinical outcomes [Li et al., 2016b]. However, due to the lack of NGS methylomes in epithelial tumours, the role of methylation disorder and intratumour heterogeneity in tumour evolution is poorly understood.

1.4 Scope of this thesis

Breast cancer is one of the leading causes of cancer death in women, and is a heterogeneous disease displaying distinct therapeutic responses. Accordingly, there has been a push to catalogue and characterise the molecular drivers in breast cancer to bring us closer towards the goal of personalised medicine. Recent efforts have focused on high-throughput genomic and transcriptomic profiling in large breast cancer datasets [Cancer Genome Atlas Network, 2012; Curtis et al., 2012; Dvinge et al., 2013; Nik-Zainal et al., 2016; Pereira et al., 2016], and revealed that breast cancer subtypes possess distinct molecular wiring. However, insights into the mechanism of tumorigenesis need to be extended beyond genomic alterations. Although, two large breast cancer consortiums, TCGA [Cancer Genome Atlas Network, 2012] and ICGC [Nik-Zainal et al., 2016], have included DNA methylation profiles as part of their multiplatform investigations, these breast cancer methylomes were not explored in-depth. Consequently, the *methylomic* basis of breast cancer heterogeneity is not fully understood, and this limitation in the literature can be explained by four fundamental reasons. Firstly, breast cancer is one of the most heterogeneous cancers at the molecular level, and in order to comprehensively reveal the contribution of DNA methylation in tumorigenesis and heterogeneity, a significantly large number of tumours needs to be investigated. Secondly, the abundant use of microarrays in which the selection of CpG probes is hypothesis driven has forbidden the discovery of novel methylation events in cancer, and has restricted focus primarily to promoter regions. Thirdly, the lack of long-term survival data in these studies, has stunted the investigation of the prognostic value of DNA methylation signatures in breast cancer. And finally, the absence of large NGS-based breast cancer methylomes has made it almost impossible to evaluate the role of intratumour heterogeneity in tumour evolution.

From this perspective, the METABRIC consortium dataset [Curtis et al., 2012] provides a vital resource for studying the epigenetic landscape of breast cancer. METABRIC represents the largest cohort of primary breast tumours available for observational study, and has excellent clinical annotation as well as long-term survival information. Additionally, the availability of multi-dimensional molecular profiles including copy number, gene expression, and somatic mutations in cancer driver genes enables the possibility to integrate DNA methylation into combined epigenomic and genomic pathways.

In order to expand our knowledge of the DNA methylation aberrations underlying human breast pathogenesis, the laboratory has conducted an NGS-based breast cancer methylome study of 1482 breast tumours and 237 adjacent matched normal tissues drawn from the METABRIC cohort. Such a large sample size guarantees adequate statistical power and appropriate representation of inter-patient heterogeneity to illuminate the full complexity of epigenetic aberrations in breast cancer. As described earlier, several genome-wide strategies for methylome profiling are currently available with different sample requirements, merits and challenges. RRBS was selected as the DNA methylation platform of choice since it allows quantification of DNA methylation at single nucleotide resolution and would also enable the exploration of intratumour methylation heterogeneity and epiclonal dynamics. Other advantages include the lower input DNA requirement and cost effectiveness compared to WGBS. The RRBS library preparation was performed by Dr Ana Tufegdzic Vidakovic, Dr Suet-Feung Chin, and Ankita Sati Batra; and sequencing was conducted by the Cancer Research UK Cambridge Institute (CRUK CI) Genomics Core.

This thesis largely delineates the *bioinformatics* and *statistical analysis* of the genome wide sequencing-based breast cancer methylomes of 1482 breast tumours and 237 adjacent matched normal tissues drawn from the METABRIC cohort. The current chapter (**Chapter 1**) provides a literature review illustrating the discovery of DNA methylation as an epigenetic mark, and its role in embryonic development and cancer (with a focus on breast cancer), along with an account of the significant efforts that have already been conducted to refine the taxonomy of breast cancer. The subsequent four chapters present four distinct fundamental features of the methylome analysis. Each of these four chapters commences with a focused introduction, reviewing the relevant literature and summarising the questions investigated in it; and concludes with a discussion of the results generated in the respective chapter.

Chapter 2 presents a detailed outline of the development of a robust RRBS pipeline that is not only suitable for high-throughput, but also maximises the information content yield while keeping feasibility in mind. An NGS-based cancer methylome study of this scale is unprecedented. Accordingly, several strategies have been implemented including optimising the sequencing parameters and bioinformatics methods with the ultimate aim of boosting the information yield. Furthermore, since methylation profiling has gained relative importance only recently, there are still unmet challenges in terms of the bioinformatics tools to analyse bisulphite sequencing data. In particular, quantitative methods for RRBS are currently undeveloped due to

the specific complexities of this protocol. Currently available statistical methods are reviewed and a novel algorithm called SCCRUB (Spatially Coordinated CpG-sites within RRBS Universe in Breast cancer) is constructed and implemented to define a functionally relevant RRBS universe of regions comprising of spatially coordinated CpG sites in breast cancer. Validation of the RRBS pipeline as well as a preliminary overview of the DNA methylation landscape generated from the 1482 breast tumours and 237 adjacent normal tissues is presented.

The primary objective of **Chapter 3** is to quantify the extent of DNA methylation alterations that have putative *functional* roles in a) tumorigenesis; b) and/or explaining inter-tumour heterogeneity in breast cancer. However, as described previously, cancers accumulate epigenetic drift over time. The presence of these methylation alterations, that are largely stochastic and not gene-specific in nature, can make the identification of epigenetic alterations truely associated with the initiation and progression of tumours quite challenging; an obstacle that is rarely considered in previous cancer methylome studies. Accordingly, in this chapter, background methylation differences in tumours (compared to normal tissues), that are a consequence of epigenetic drift, are first quantified. The genomic context and inter-tumour heterogeneity of these background methylation differences are also investigated, and are then utilised to feed the development of a novel algorithm called DMARC (Directed Methylation Altered regions in Cancer) to detect *directed* and *background* DNA methylation alterations in tumours. In addition, to illuminating the mechanism underlying an observed methylation difference in a tumour, this algorithm can also enrich for putative functional (expression-associated) methylation events. Finally, the heterogeneity in epigenetic programming in distinct breast cancer subtypes, and its consequences on transcriptional networks and survival is explored.

Chapter 4 extends the findings of the previous chapter, and compares the independent contributions of DNA methylation and copy number alterations in deregulating gene expression in breast cancer. The large number of samples in this study allows for the detection of mutually exclusive or co-occurring patterns between epigenetic (DNA methylation) and genomic (CNA) alterations, that can lead to the identification of genes with putative tumour suppressive or oncogenic roles. Interestingly, the crucial role of DNA methylation as a mechanism to target the silencing of specific genes within copy number amplifications is also investigated.

While the previous chapters investigated the inter-tumour DNA methylation heterogeneity in breast cancer, in **Chapter 5**, the 1482 RRBS breast cancer and 237 normal methylomes are reanalysed at the single read level to provide the first genome-

wide assessment of the role of epigenetic intra-tumour heterogeneity in breast cancer, and the largest for any single cancer type. Two previously established measures were implemented in the breast tumours. The Proportion of Discordant Reads (PDR) score represents the intratumour epigenetic diversity [Landau et al., 2014]; while the Eloci Per Milion (EPM) score represents the magnitude of dynamic epiallele shifting in a tumour compared to the normal tissue [Li et al., 2014]. These two orthogonal scores represent distinct properties of the epigenome and the relationship between them is explored for the first time. The link between intratumour heterogeneity measures inferred from epigenetic and genetic profiles is also examined, and finally, the prognostic potential of the DNA methylation intratumour diversity (PDR) and the magnitude of dynamic epiallele composition shifts (EPM) is explored.

The future directions of this project are also outlined in the final chapter (**Chapter 6**). With the wealth of data available, the possibilities are abundant. A comprehensive genome-wide profiling of hydroxymethylcytosine in breast cancer is warranted since it is a distinct epigenetic modification to methylcytosine with an alternate role in pathogenesis. Additionally, the investigation of DNA methylation signatures in i) a panel of breast Patient Derived Tumour Xenografts (PDTXs), representing one of the best preclinical models available today [Bruna et al., 2016]; and ii) in circulating tumour DNA, enabling the possibility of a *liquid biopsy*, is also proposed. These cohorts are already established in the Caldas laboratory and would enable the exploration of the potential of DNA methylation biomarkers to be used as tools for early predictors of breast cancer, monitoring, prognosis and stratification for different therapeutic approaches.

Chapter 2

DNA methylation profiling of a large breast cancer cohort

Contents

2.1 Introduction

Recent technological advances have allowed the characterisation of not only genetic aberrations such as somatic mutations and copy number alterations in cancer, but also allowed exploration of the epigenetic landscape including DNA methylation and its role in tumorigenesis. A majority of the previous studies of the breast cancer methylome have been limited by platforms that analyse only a very small fraction of the epigenome or have low resolution, and/ or insufficient power to unravel the epigenetic basis of breast cancer heterogeneity (Chapter 1). However, in 2012, 800 breast tumours were profiled with the Infinium HumanMethylation450 BeadChip (HM450) microarray by the TCGA consortium [Cancer Genome Atlas Network, 2012] and this represents the largest breast cancer methylome study conducted thus far. Multiple reports since then have used the HM450 platform in breast tumours leading to key findings including the identification of tumour subgroups with distinct DNA methylation patterns that were shown to have diverse molecular characteristics and prognosis [Györffy et al., 2016; Stefansson et al., 2015]. These studies delineated specific genomic regions that are epigenetically disrupted in breast cancer and had associations with other molecular events such as mRNA expression and copy number events [Gao et al., 2015; Holm et al., 2016; Teschendorff et al., 2016b].

The investigation of the DNA methylation landscape in the METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) cohort comprising of approximately 1500 breast tumours represents a significant development in the field and will enable the confirmation of previous findings as well as improve the understanding of the genome-wide patterns of DNA methylation in breast cancer. For the METABRIC study, Reduced Representation Bisulphite Sequencing (RRBS) has been chosen as the DNA methylation platform for a variety of reasons. Firstly, HM450 and RRBS represent complementary genome-wide techniques, profiling distinct sets of CpG sites in the genome, and this allows the discovery of novel methylation biomarkers. Moreover, being a next-generation sequencing (NGS) technique, it allows quantification of DNA methylation at single nucleotide resolution and also enables the exploration of intratumour methylation heterogeneity and epiclonal dynamics in breast cancer. Another advantage of RRBS is the possibility of DNA methylation profiling using very low input DNA amounts, since the need to characterise tumours at multiple levels puts tumour DNA at a premium.

While RRBS has been shown to be suitable for methylome profiling of cancer samples [Li et al., 2016b; Pan et al., 2015], most studies that have implemented this

strategy have relatively small sample sizes, probably due to the challenging nature of this protocol. The methylome study of approximately 1500 tumours presented in this thesis represents the largest bisulphite sequencing study in cancer. Accordingly, it was imperative to design a robust RRBS pipeline that is not only suitable for high-throughput, but also maximises the information content yield, while considering feasibility in terms of cost and input DNA material. This chapter presents several strategies that have been implemented including optimising the library preparation protocol, sequencing parameters and bioinformatics methods with the ultimate aim of boosting the information yield, which in the case of RRBS is the number of CpG sites detected with sufficient depth across all samples. A description of the filtering steps and quality control procedures are also provided.

Bisulphite sequencing technologies, in particular RRBS, present substantial challenges in the statistical analysis and biological interpretation of methylation differences between samples or groups of samples. In fact, only a few computational tools to analyse RRBS data exist. Since functional methylation alterations usually involve clusters of multiple CpG sites, an important consideration is the need to deal with methylation information at single nucleotide resolution. Currently available methods are explored, and a novel algorithm is developed and implemented to define a RRBS universe of regions comprising of spatially coordinated CpG sites in breast cancer.

Next, validation of the RRBS pipeline is performed. Specifically, the extent of technical variability in the methylome dataset is investigated. A comparison of methylation profiles in METABRIC samples mapped using two distinct platforms, RRBS and the universally used HM450 microarray, is also conducted to check whether the two technologies produce similar methylation estimates for the same tumour and same genomic region.

Finally, a preliminary overview of the DNA methylation landscape in 1482 breast tumours and 237 adjacent normal tissues is presented. An unsupervised analysis based on methylation profiles is also conducted in the METABRIC samples and compared with that obtained from the external TCGA breast cancer study [Cancer Genome Atlas Network, 2012].

2.1.1 Summary of aims

This chapter describes the generation of a sequencing-based genome-wide methylome breast cancer dataset comprising of 1482 tumours and 237 adjacent normal tissues drawn from the METABRIC cohort. This is achieved through the following steps:

1. An overview of the samples that are part of the METABRIC cohort, and a description of the multidimensional molecular and clinical data available for these samples.

2. Optimisation of the RRBS pipeline including determination of the sequencing parameters, and development of the bioinformatics pipeline. Quality assessment and filtering procedures are also detailed.

3. A review of existing methods to analyse single nucleotide methylation information as generated by bisulphite sequencing techniques. Construction and implementation of a novel algorithm to define an RRBS universe of regions comprising of spatially coordinated CpG sites in breast cancer.

4. Validation of the RRBS pipeline including investigating the extent of technical variability, and a comparison of RRBS and HM450 generated methylation profiles in the METABRIC study.

5. A preliminary overview of the DNA methylation landscape in the METABRIC samples, including comparison with the external TCGA breast cancer methylome.

2.2 Sample overview

METABRIC is a collection of over 2,000 primary fresh-frozen invasive breast cancer specimens and 473 matched normal tissues from tumour banks in the UK and Canada, associated with extensive clinical annotation and follow up [Curtis et al., 2012]. For DNA methylation profiling, a total of 1719 samples were characterised (following quality control procedures described in Section 2.3.4) including 1482 primary breast tumours and 237 adjacent normal samples. 1367 out of the 1980 tumours from the original METABRIC study [Curtis et al., 2012] were utilised, with insufficient DNA accounting for the missing samples. The additional 113 tumours were not published as part of the original METABRIC study despite belonging to the METABRIC cohort. These samples had either failed quality checks on the platforms used at the time, lacked corresponding gene expression data, or were processed after the initial publication was completed.

Multi-dimensional molecular data including gene expression (Illumina HT-12 v3 microarray), copy number aberrations (Affymetrix SNP 6.0 arrays), mutations in key cancer driver genes (targeted sequencing on 173 genes) as well as extensive clinical annotation were also available, and described below.

2.2.1 Gene expression data

Gene expression profiling of the METABRIC cohort was conducted using Illumina HT-12 v3 microarrays [Curtis et al., 2012]. Gene expression data was pre-processed and annotated using custom scripts relying on the *beadarray* package [Dunning et al., 2007] resulting in normalised relative $\log_2$ intensities for each probe. Details are described in, and the gene expression data was obtained from the original publication [Curtis et al., 2012]. Since Illumina microarrays use multiple probes localised at distinct loci to represent the expression of a single gene, low-quality probes were removed and the probe with the highest variation was chosen in order to determine gene-level expression data.

2.2.2 Copy number data

Copy number profiling of the METABRIC cohort was conducted using Affymetrix SNP 6.0 arrays [Curtis et al., 2012]. The raw data was pre-processed using aroma.affymetrix [Bengtsson et al., 2009] to obtain normalised relative $\log_2$ ratios for each probe, and B-allele fractions at SNP loci, as described in Curtis et al. [2012].

Two distinct analytical approaches were employed to generate continuous and allele-specific discrete estimates for copy number alterations respectively.

1. **Continuous approach:** The ratios were segmented using the circular binary segmentation (CBS) algorithm [Venkatraman and Olshen, 2007] implemented in the *DNAcopy* bioconductor package [Lipson and Liebert, 2006], as described in Curtis et al. [2012]. The segmented data was obtained from the original paper [Curtis et al., 2012]. Subsequently, the genome coordinates were converted from hg18 to hg19 using the *LiftOver* bioconductor package [Hinrichs, 2006]. Where continuous copy number estimates were required, the segmented mean $\log_2$ ratios were used. Overall, 1389 out of the 1482 tumour samples had associated *DNAcopy* number data available.

2. **Allele-specific approach:** The Allele-Specific Copy number Analysis of Tumours (ASCAT) algorithm [Van Loo et al., 2010] was implemented to obtain tumour-specific estimates of ploidy and aberrant cell fraction (tumour purity). Consequently, segmented allele-specific estimates of absolute copy number corrected for normal contamination were generated as described in Pereira et al. [2016]. The ASCAT data was obtained from Pereira et al. [2016]. Subsequently, the total copy number estimates were adjusted for ploidy as follows:

$$CNcorr = round(\frac{2 \times CNtot}{\pi}) \tag{2.1}$$

where *CNtot* and *CNcorr* refer to the total and corrected copy number calls respectively, and π is the estimated tumour ploidy. The corrected copy number was rounded to the nearest integer. Where allele-specific or discrete copy number data was required, these ASCAT estimates were used. Overall, 1345 out of the 1482 tumour samples had associated ASCAT copy number data available.

The downstream analysis described below was conducted for both DNAcopy-derived and ASCAT-derived copy number data. Germline copy number variants were removed using copy number data from the combination of HapMap populations and the adjacent matched normal tissues as described in Curtis et al. [2012]. Only somatic copy number alterations (CNA) were considered for which a diploid copy number state was assumed for the normal tissues. Gene annotation was conducted using coordinates obtained from RefSeqGene (Reference Sequence Gene) [Karolchik et al., 2014, hg19]. In order to summarise copy number at the gene-level, segmented copy number

alterations were overlapped with gene regions. Where multiple copy number segments were available for the same gene, the copy number state of the segment with maximal severity was assigned.

2.2.3 Mutation data

Targeted sequencing of the exons of 173 genes (panel of key cancer genes) was performed on the METABRIC cohort [Pereira et al., 2016]. The bioinformatics pipeline is described in Pereira et al. [2016] and summarised below. Variant calling was performed using MuTect [Cibulskis et al., 2013] for single nucleotide variants (SNVs), and Haplotype Caller for indels [McKenna et al., 2010]. Custom pipelines were utilised for filtering. Candidate driver genes in breast cancer were identified by looking for genes that harboured multiple recurrent or inactivating mutations, as these patterns are characteristic of oncogenes and tumour suppressors respectively as proposed by Vogelstein et al. [2013]. As described in Pereira et al. [2016], inactivating mutations included nonsense SNVs, frameshift substitutions, splice sites mutations; whereas recurrent mutations were defined as missense SNVs and in-frame substitutions that occur in the same codon. The proportions of recurrent mutations (oncogene score) or inactivating mutations (tumour suppressor score) observed for each gene were scored, and a threshold of 20% was used. Further details are described in, and the mutation data was obtained from Pereira et al. [2016].

2.2.4 Clinical data

Clinical data for the METABRIC cohort was obtained from published and in-preparation manuscripts [Curtis et al., 2012; Pereira et al., 2016; Rueda et al., 2017, in preparation]. Clinical data includes age at diagnosis; tumour-specific variables such as grade, size, number of lymph nodes, ER and HER2 status; as well as breast cancer classification into Intrinsic subtypes and Integrative clusters. Pathology-based mitotic indices for the tumours (scored by Dr Elena Provenzano) and digital lymphocytic infiltration scores (created by Dr Raza Ali) were also available. Long-term follow-up data including overall survival, breast cancer-specific survival, local and distant relapse was also available with a median follow-up time of 11 years.

2.3 Optimising the RRBS Pipeline

The preliminary goal of the project was to design a robust RRBS pipeline that was suitable for high-throughput, and that ensured accurate methylation calling as well as maximised the information content yield i.e. the number of CpG sites detected per sample. Previous work in the laboratory (conducted by Dr Ana Tufegdzic Vidakovic) led to the optimisation of the RRBS library preparation protocol (summarised in Vidakovic [2014], Section 2.3.1). Determination of the ideal sequencing parameters and optimisation of the RRBS bioinformatics pipeline was conducted as part of this thesis and described in Section 2.3.2 - 2.3.5.

2.3.1 RRBS library preparation protocol

Optimisations to the RRBS library preparation protocol implemented in this study are described in Vidakovic [2014], and are summarised briefly in this section.

A gel-free multiplexed RRBS method was adapted from Boyle et al. [2012] with modifications discussed below. Briefly, the original Boyle protocol commences with DNA digestion with the restriction endonuclease *MspI*, followed by end repair and A-tailing of the fragments. This is followed by ligation with methylated adapters. Subsequently, the DNA is bisulphite converted, after which the RRBS libraries are amplified. Finally, a purification step is conducted to size-select for fragments between 200-700 base pairs (bp). Typically multiplexing of samples is performed using barcoded DNA adaptors and pooled before the bisulphite conversion step.

2.3.1.1 Input DNA

The recommended DNA input for RRBS is 200ng of DNA [Boyle et al., 2012] . However, this amount can be considerably scaled down (to 5 ng), as determined by experiments performed in the laboratory [Vidakovic, 2014]. This is of particular significance for methylation profiling of the METABRIC samples, since the input DNA obtained from these tissues is low due to previous multiple molecular characterisations.

2.3.1.2 Unbalanced libraries

Typically, each RRBS library consists of 6-12 barcoded samples. In spite of starting from equal amounts of DNA, the original RRBS protocol suffers from an imbalance in the sequencing reads assigned to each sample within the pool [Boyle et al., 2012]. The imbalance appears to occur upstream of the Polymerase chain reaction (PCR)

amplification step where using equivalent volume did not represent equal input. A quantification step, pre-PCR and prior to sample pooling was introduced to normalise the input DNA which has positively affected library balancing and the final number of CpG loci detected [Vidakovic, 2014] .

2.3.1.3 RRBS library preparation workflow

The workflow of the modified RRBS library preparation protocol used in this study is shown in Figure 2.1. Further details are given in Vidakovic [2014].

Tissue processing was performed by Dr Suet-Feung Chin. In brief, DNA was extracted from 10x30micron sections using Qiagen QIAmp DNeasy Kits (Qiagen, Germany) and quantified fluorometrically by Qubit dsDNA High Sensitivity Quantification Reagent (ThermoFisher, USA). The DNA were normalised and plated into 96 well plates.

The RRBS libraries were generated by Dr Ana Tufegdzic Vidakovic, Dr Suet-Feung Chin, and Ankita Sati Batra.

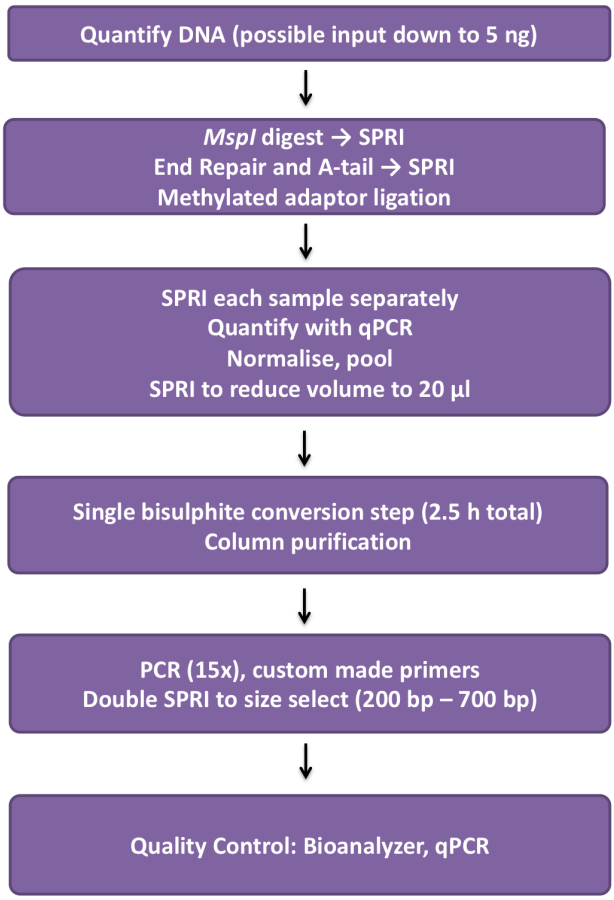


Figure 2.1: Schematic outline of the RRBS library preparation protocol. Figure is modified from Vidakovic [2014]. SPRI = Solid Phase Reversible Immobilisation.

2.3.2 Sequencing parameters

Multiple RRBS experiments were performed using DNA derived from MDA-MB-231 breast cancer cell lines (library preparation conducted by Dr Ana Tufegdzic Vidakovic) to determine the optimal sequencing parameters including the Illumina chemistry, read lengths, necessary sequencing depth, and comparison of single-end and paired-end sequencing. The aim was to maximize CpG site detection while considering feasibility and cost.

2.3.2.1 Illumina technology

The launch of v4 Illumina chemistry on the HiSeq 2500 machines (to replace the v3 chemistry on the HiSeq 2000 machines) in 2014 promised an increase in sequencing yield, and thus number of reads, as well as offering a reduction in run time and price. A 12 sample-multiplexed RRBS library sequenced with HiSeq 2500 yielded on average 13.2 million aligned reads per sample, whereas the same library on HiSeq 2000 yielded 8.8 million reads on average (*p-value* = 0.0018; t-test; Figure 2.2a). Importantly this also translated in an increase in CpG sites detected at 1×, 5×, and 10× coverage (Figure 2.2a).

2.3.2.2 Single-end vs. paired-end sequencing

Sequencing both ends of each read is a more efficient use of a library. Paired-end (PE) improves the ability to identify the relative positions of reads in the genome, making it much more effective than single-end (SE) in resolving structural rearrangements such as gene insertions, deletions, or inversions; and improving the assembly of repetitive regions. However, PE sequencing is approximately 50% more expensive and time-consuming to perform than SE [Genomics Core Cancer Research UK Cambridge Institute, 2015]. The key question is whether PE reads yield significantly more data than SE reads to justify this difference in price and cost.

PE reads are informative for the same DNA fragment (representing one DNA strand), but captures information from both sides of the fragment. In Whole Genome Bisulphite Sequencing (WGBS) experiments, the DNA fragments are randomly distributed across the genome, and can be selected to be large enough, and therefore PE sequencing yields considerably more unique CpG sites than SE. However, in RRBS, DNA digestion with *Msp1* results in fragments that start and end at the same sites of the genome, and the size-selection can often lead to fairly small fragments in contrast to a read length of 125bp [Babraham Bioinformatics, 2016b]. As a

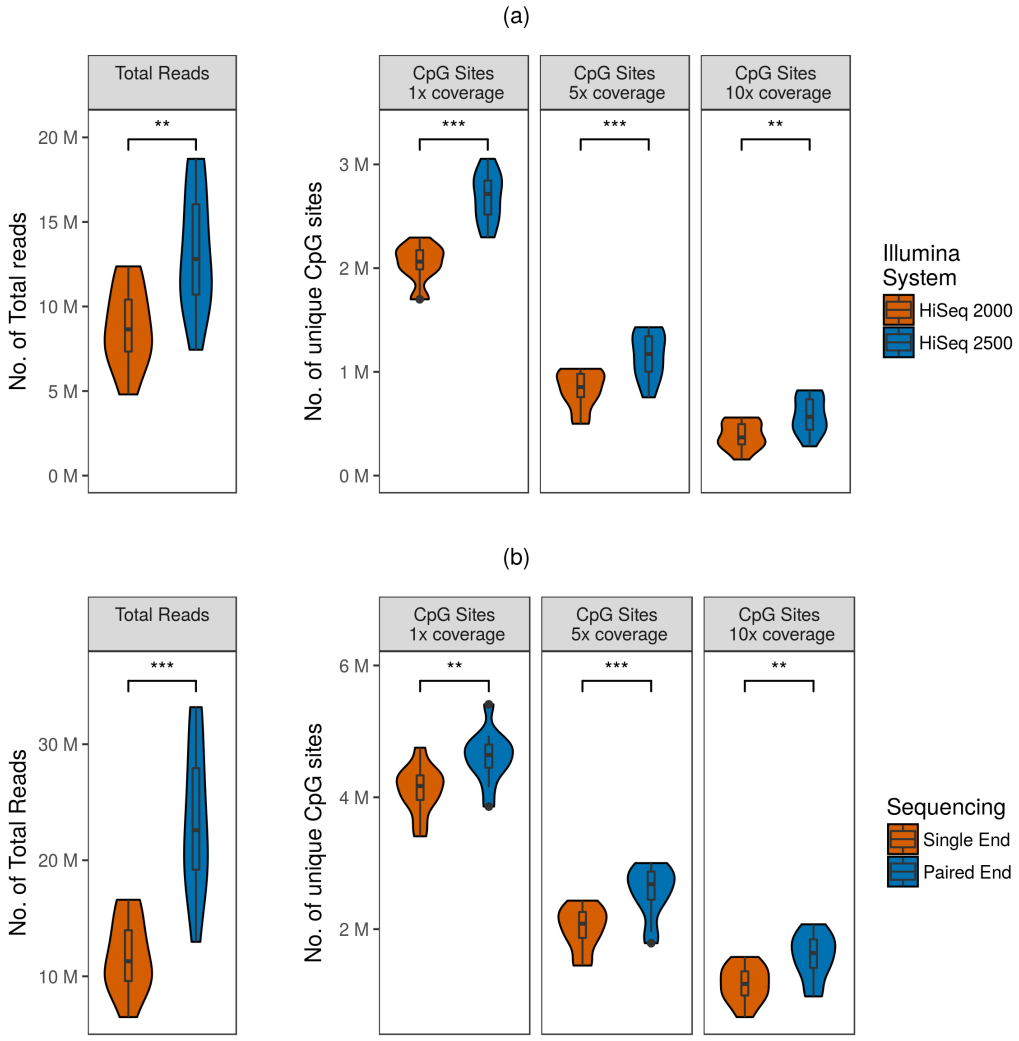


Figure 2.2: Single end sequencing was performed on the Illumina HiSeq 2500. (a) Comparison of yield between Illumina HiSeq 2000 and Illumina HiSeq 2500 within samples of a 12-sample library pool. **(b)** Comparison of yield between SE and PE within samples of a 12- sample library pool. In both comparisons, number of total reads detected, number of unique CpG sites detected at $1\times$, $5\times$ and $10\times$ coverage were compared. For both plots, boxplots indicate the interquartile range with 25^{th} percentile, median and 75^{th} percentile values illustrated. Whiskers indicate $1.5\times$ of the interquartile range. T-tests were used for statistical comparison. M=million. (. = *FDR p-value* < 0.1, $*$ = *FDR p-value* < 0.05, ** = *FDR p-value* < 0.01, *** = *FDR p-value* < 0.001, **** = *FDR p-value* < 0.0001).

consequence, PE reads overlap in the middle and yield redundant methylation information for the same strand and thus, do not yield twice the amount of methylation data as anticipated [Babraham Bioinformatics, 2016b]. In fact, SE sequencing with the same number of reads as a PE run is more likely to yield more methylation information in terms of unique CpG sites detected.

Notably, preliminary RRBS experiments revealed that only 27.6% more unique CpG sites at 5× coverage (*p-value* = .0040; t-test), and 36.6% more unique CpG sites at 10× coverage (*p-value* = 0.0009; t-test) were detected using PE vs. SE sequencing (Figure 2.2b). Keeping in mind the large number of samples in the METABRIC study, this finding strongly indicates that it is not cost-effective to use PE sequencing which is 50% more expensive than SE.

2.3.2.3 Read length

Most RRBS protocols have recommended 36-50 bp read runs for sequencing [Akalin et al., 2012a; Boyle et al., 2012] . The suggested reason for this is that the quality of base calling drops with an increase in read length and as a result, increasing the read length of RRBS sequences may not necessarily translate into a linear increase in CpG sites yield [Krueger et al., 2012]. Accordingly, many methylome studies use short reads sequencing for RRBS reads [Akalin et al., 2012a; Landau et al., 2014; Pan et al., 2015]. However, since RRBS library fragment sizes range from 200-700 bp, it is apparent that longer sequencing reads could increase both the number of CpGs detected and their coverage as long as sequencing quality does not drop. To investigate whether longer sequencing reads would improve high quality yield, a 12 sample-multiplexed RRBS library was sequenced with 125 bp read length. The reads were then trimmed *in silico* to 50 bp to obtain an *in silico* shorter sequencing output, and the quality of base calls and the yield in terms of number of CpGs detected were assessed. This experiment revealed that the quality of base calls remains high in longer sequenced reads (125 bp), substantiating the effectiveness of the enhanced RRBS protocol (Figure 2.3a). Moreover, increasing the read length from 50 bp to 125 bp allowed for detection of significantly more CpG sites (Figure 2.3b). This suggests a shift to longer read sequencing for RRBS libraries, contrary to what has been recommended in previous studies.

2.3.2.4 Sequencing depth

The success of the methylome study is contingent on each of the approximately 2000 samples yielding sufficient methylation information (unique CpG sites at adequate

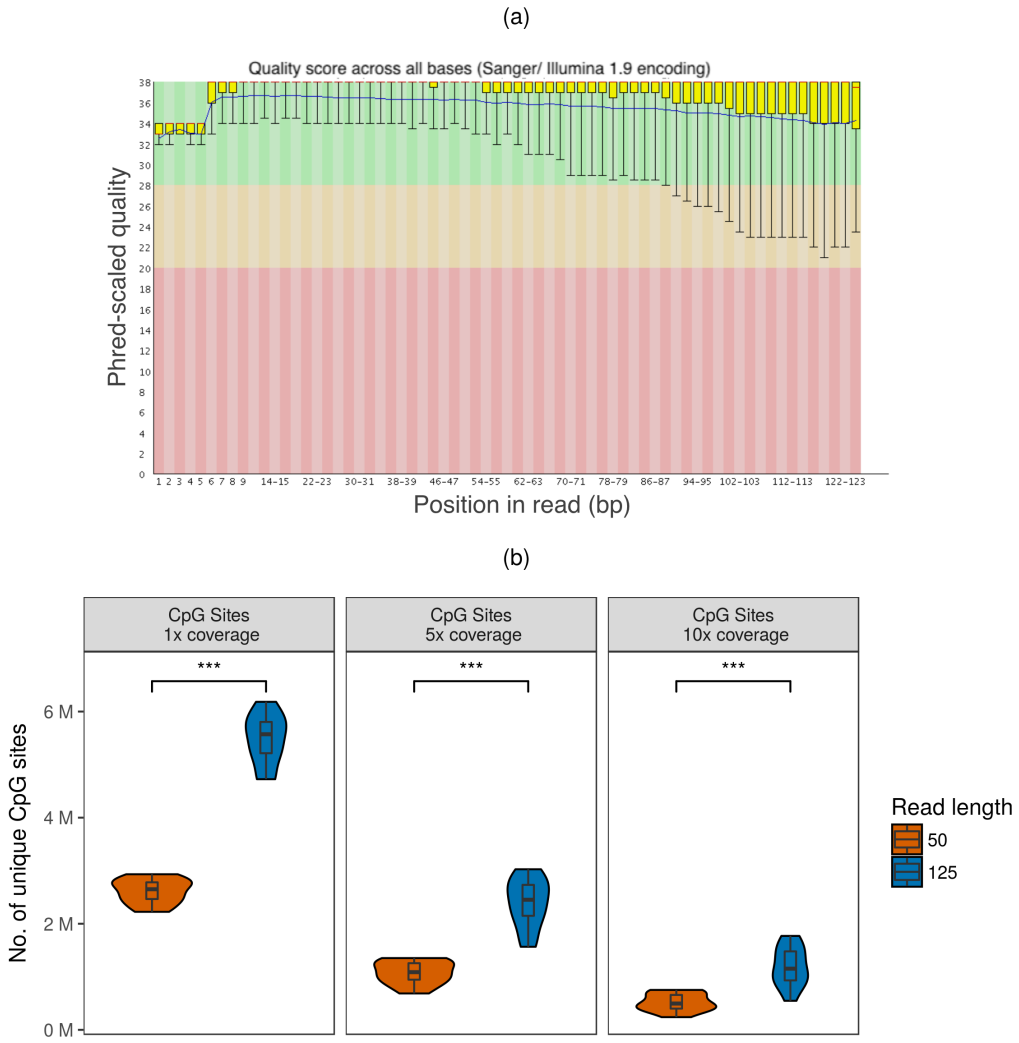


Figure 2.3: : Increased read length improves CpG detection and coverage. (a) Phred-scaled quality of base calls remains high for 125 bp reads. y-axis represent quality scores for a 12- sample library pool and x-axis represents the position in the read up to 125 bp. Phred-scaled quality > 20 constitute acceptable scores and Phred-scaled quality $\geq$ 30 represent good scores. **(b)** Comparison of yield between 50 bp and 125 bp reads within samples of a 12-library pool. Number of unique CpG sites detected with $1\times$, $5\times$ and $10\times$ coverage were compared. Boxplots indicate the interquartile range with 25^{th} percentile, median and 75^{th} percentile values illustrated. Whiskers indicate $1.5\times$ of the interquartile range. T-tests were used for statistical comparison. M=million. (. = *FDR p-value* < 0.1, $*$ = *FDR p-value* < 0.05, ** = *FDR p-value* < 0.01, *** = *FDR p-value* < 0.001, **** = *FDR p-value* < 0.0001).

depth). The target number of unique CpG sites to be profiled per sample was set to 2.5 million for the purpose of this study. A minimum of 5 read coverage is recommended to accurately estimate the methylation at each CpG site [Bock et al., 2010; Boyle et al., 2012; Meissner et al., 2008]. It is apparent that a higher number of reads would yield a higher number of CpG sites, however sequencing can quickly reach saturation depending on the required depth. Feasibility constraints necessitates the calculation of the optimal number of reads required to be sequenced per sample to reach the target number of CpG sites at the required depth, and consequently the determination of the extent of multiplexing. The extent of multiplexing has obvious consequence on the read yield and cost for each sample: pooling less samples leads to a higher read yield and higher costs while pooling more samples is more cost effective albeit with an associated compromise on read yield.

Data from preliminary RRBS experiments (multiplexed at 12 samples) was investigated, and *in silico* experimentation was performed to investigate the read yield for other levels multiplexing. The average sequencing yield per lane for the preliminary RRBS experiments was 160-180 million reads. Figure 2.4a reveals that the optimal extent of multiplexing is 8 samples, at which point the number of unique CpG sites detected on average per sample starts reaching saturation. Multiplexing at 8 samples results in approximately 20 million reads yield and results in detection of more than 2.5 million unique CpG sites at 5$\times$ coverage for each sample (interquartile range is higher than 25 million CpG sites, Figure 2.4a). Multiplexing less and increasing sequencing read depth is associated with higher costs, but with not a high reward in terms of additional unique CpG sites detected, whereas multiplexing more does not generate sufficient CpG sites. Interestingly, the number of unique CpG sites detected at 10$\times$ coverage does not reach saturation at 8-sample multiplexing, and a higher reward ratio is observed by multiplexing less (Figure 2.4b). However, the number of unique CpG sites detected at 1$\times$ coverage quickly approaches saturation (Figure 2.4c).

These findings reinforce the notion that RRBS is focused on a well-defined CpG-rich reduced representation of the genome which leads to relatively early saturation in detecting unique CpG sites with deeper sequencing. Profiled CpG sites are deemed beneficial for methylome studies only if they have sufficient coverage for accurate estimation of methylation levels (minimum 5$\times$ coverage). Deeper sequencing boosts the coverage of these detected CpG sites, but does not yield considerably more unique CpG sites.

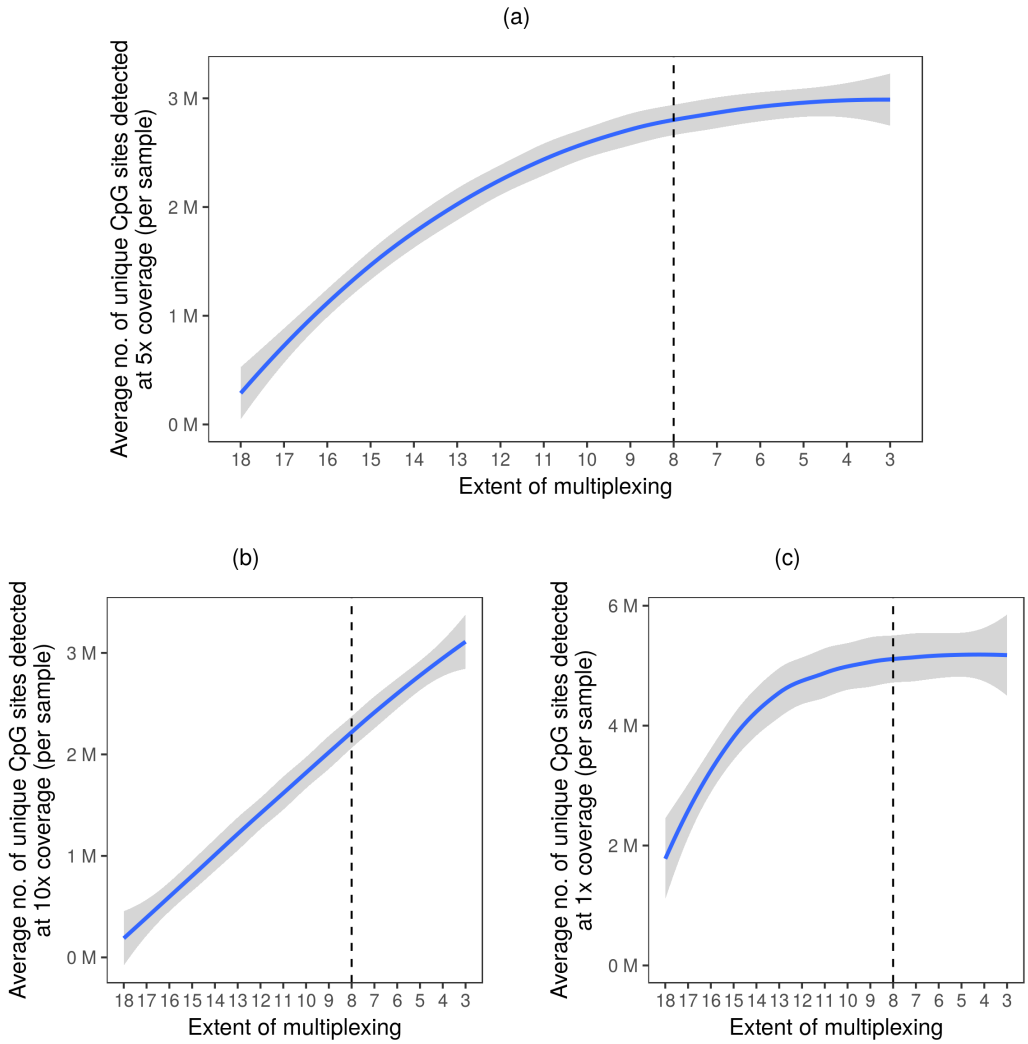


Figure 2.4: Multiplexing at the level of 8 samples provides a good balance between yield and feasibility. (a) The blue line represents the average number of CpG sites detected at 5× per sample as a function of the extent of multiplexing. Shaded area represents the interquartile range for the above relationship. Higher values on the y-axis represents lower multiplexing which yields a higher number of reads per sample and consequently a higher number of unique CpG sites detect at 5×. However, this starts reaching saturation at 8-sample multiplexing. **(b)** Average number of CpG sites detected at 10× per sample as a function of the extent of multiplexing. Does not reach saturation at 8-sample multiplexing. **(c)** Average number of CpG sites detected at 1× per sample as a function of the extent of multiplexing. Reaches saturation earlier than 8-sample multiplexing. In (a) (b) and (c) 8-sample multiplexing is indicated by the vertical dotted line. M=million.

2.3.2.5 Final Sequencing Parameters

For the METABRIC samples, sequencing was performed on the Illumina HiSeq 2500 (v4 chemistry), with single-end reads of 125 bp length. Multiplexing was conducted at the level of 8 samples per lane. Sequencing was performed by the Cancer Research Cambridge Institute (CRUK CI) Genomics Core and de-multiplexing by the CRUK CI Bioinformatics Core.

2.3.3 Bioinformatics pipeline

The RRBS bioinformatics pipeline for alignment and methylation calling has been well established [Babraham Bioinformatics, 2016a,b]. However, it was adapted to meet the specific challenges of the project. Moreover, as with all high throughput sequencing applications, it is critical to perform quality control (QC) to detect and account for sequencing and methylation calling errors. The modified RRBS bioinformatics pipeline and QC are detailed below. The schematic outline of the pipeline is illustrated in Figure 2.5.

2.3.3.1 Trimming

RRBS libraries with long read lengths (125 bp in this case) suffer from a variety of complications. As with all high throughput sequencing, long reads can be associated with poor qualities towards the 3' end. Moreover, DNA fragments generated from RRBS libraries are often shorter than 125 bp, and hence the sequencing reads may continue into the adapter sequence on the 3' end (Section 2.3.2). An additional characteristic of RRBS is that unmethylated cytosines (Cs) that are introduced during the enzymatic end repair step may also be sequenced if they are not trimmed appropriately. Trimming of the 3' ends was performed using Trim Galore! (version 0.3.7: powered by cutadapt) to i) remove bases with Phred-scaled quality score < 20, ii) remove adaptor contamination, and iii) remove the additional unmethylated Cs introduced during the end repair step.

2.3.3.2 Alignment

DNA treatment with the bisulphite chemical results in converting unmethylated cytosines (Cs) to thymines (Ts), whereas methylated Cs are largely protected from bisulphite-induced conversion. Aligning to the traditional human reference genome is problematic since after bisulphite conversion, a large number of Cs are not Cs any more, and consequently the DNA sequences will align with less concordance to the

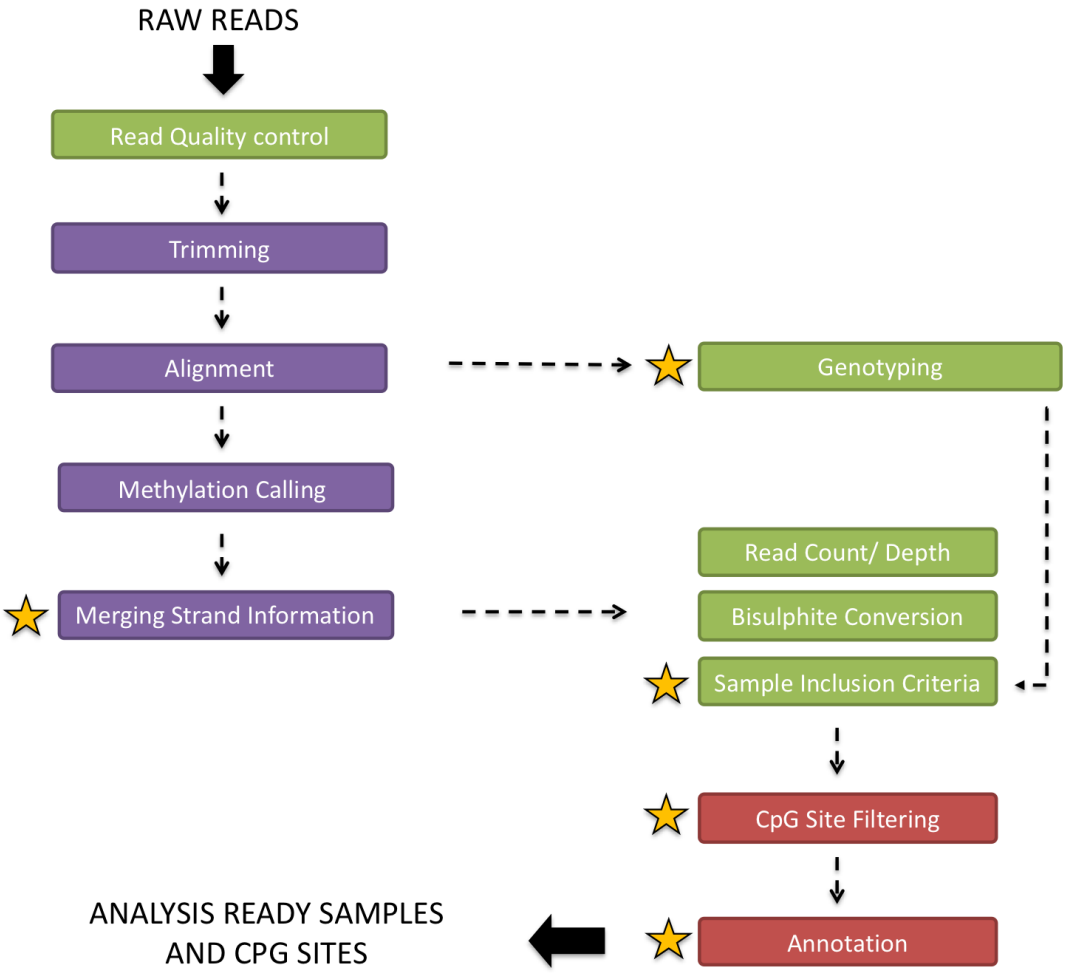


Figure 2.5: Schematic outline of the RRBS bioinformatics pipeline. Purple boxes indicate steps involved in the bioinformatics pipeline. Green boxes indicate steps involved in quality assessment and sample filtering. Red boxes indicate steps involved in CpG site filtering and annotation. Gold stars indicate steps included in addition to the RRBS bioinformatics pipeline established by Babraham Bioinformatics [2016a,b].

reference sequence. Additionally, the complexity of the bisulphite-treated library is also reduced [Bock et al., 2010]. Consequently, utilising standard aligners to align bisulphite converted reads to the human reference genome will result in a bias towards reads covering methylated Cs against unmethylated Cs. To complicate matters further, a single CpG site can have a different methylation state in different cells, and so building a reference human methylome is challenging [Krueger et al., 2012].

Algorithms have been proposed that avoid the bias towards methylation by removing the penalty associated with aligning a C or T in the read to a C in the reference genome. *Three-letter aligners* such as *Bismark* and *BS-Seeker* simplify bisulphite alignment by converting all Cs into Ts in the reads and for both strands of the genomic DNA sequence prior to alignment [Chen et al., 2010; Krueger and Andrews, 2011]. This way, they can carry out the alignment exclusively on a three-letter alphabet (namely, A, G and T) using standard aligners, such as *Bowtie* and *Bowtie2* [Langmead and Salzberg, 2012; Langmead et al., 2009]. As a result of the reduced sequencing complexity with only three letters remaining, a larger number of reads align to more than one position in the reference sequence and are discarded. Consequently, three letter aligners such as *Bismark* can be expected to achieve a lower genomic coverage than *wild-card aligners* such as *BSMAP* [Xi and Li, 2009] and *RRBSMAP* [Xi et al., 2012], but is free from the bias towards increased DNA methylation levels [Bock, 2012]. Additionally, *Bismark* is less likely to report non-unique alignments compared to the other popular three-card aligner, *BS-Seeker* [Chen et al., 2010; Krueger and Andrews, 2011].

Reads were aligned to the Human Genome Assembly GRCh37 [Lander et al., 2001, UCSC release hg19] using *Bismark (version 0.13.1)*. *Bowtie2 (version 2.2.4)* was chosen over *Bowtie* as the standard aligner due to improvements in speed, fraction of reads aligned and ability to perform gapped alignments [Langmead and Salzberg, 2012].

2.3.3.3 Methylation calling

To quantify absolute DNA methylation levels from bisulphite-sequencing data, the percentage of Cs and Ts are calculated among all reads aligned to each C in the genomic DNA sequence [Bock, 2012]. Single-base-pair methylation estimates were determined by quantifying evidence for methylated (unconverted) and unmethylated (converted) Cs at all CpG positions. Two approaches can be used to estimate methylation at a CpG site

1. **Beta-value**

A Beta-value is defined as the proportion of Cs (methylated CpGs) at a particular CpG site. Beta-value methylation estimates range from 0% to 100% where a value of 0% indicates that the CpG site was unmethylated in all assayed cells originating from the sample (i.e. no methylated molecules were measured) and a value of 100% indicates that every cell was methylated at that CpG site.

1. **M-value**

Beta-value methylation calls are largely bimodal at 0% and 100% (completely unmethylated and completely methylated respectively). The M-value is an alternate estimate that is calculated by applying the logistic transformation to Beta-values. This transformation results in a continuous methylation estimate from $-\infty$ to ∞ that expands the Beta-value distribution at 0% and 100% (Figure 2.6), and thus is more suitable for many quantitative statistical analyses [Irizarry et al., 2008]. Analogous to Beta-values, larger M-values also represent more evidence of methylation. The mathematical relationship between M-values and Beta-values is detailed in these two equations.

$$Beta = \frac{2^M}{2^M + 1} \tag{2.2}$$

$$M = log_2(\frac{Beta}{1 - Beta}) \tag{2.3}$$

Both Beta-values [Hovestadt et al., 2014; Kulis et al., 2012] and M-values [Stirzaker et al., 2014; Stone et al., 2015] have been commonly used as metrics to measure methylation levels. However, Beta-values have a more intuitive biological interpretation and their use is recommended for reporting results, while M-value methylation estimates are more statistically valid, and accordingly recommended for downstream statistical analyses such as differential methylation [Du et al., 2010]. Beta-value methylation estimates were called using *bismark_methylation_extractor* [Babraham Bioinformatics, 2016a], and M-values were generated in *R* using the Equation 2.3.

2.3.3.4 Merging strands

In mammals, the Cs at CpG dinucleotides on the two complementary DNA (+ and -) strands are symmetric with respect to methylation [Ehrlich et al., 1982]. If the C on one strand is methylated, so is the other. However, the 2 complementary Cs do not

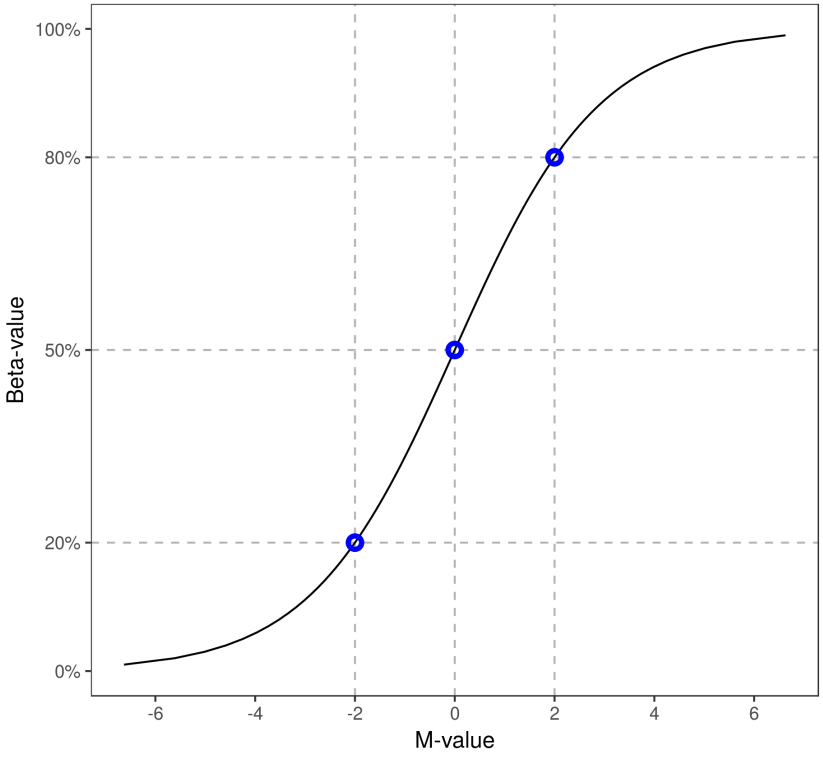


Figure 2.6: Relationship between M-value and Beta-value for methylation calling. Data points represent methylation Beta-values of 20%, 50% and 80%.

represent the exact same nucleotide location, but rather two consecutive nucleotides on separate strands. Consequently, this would result in methylation calls from the two consecutive bases, one from the + strand and one from the – strand. Standard bisulphite sequencing bioinformatics pipelines do not control for this and potentially different methylation estimates may be detected for the two bases since the random nature of directional RRBS libraries does not guarantee that reads originating from the + and – strands are represented equally, or even represent the same cells. However, due to the aforementioned symmetry in mammalian methylomes, it is necessary to merge the methylation information from the two complementary strands, since they truly represent the same CpG dinucleotide.

In the METABRIC dataset, methylation calls from complementary strands originating from the same CpG dinucleotide were highly correlated (median correlation = 0.89, 25th percentile = 0.86, 75th percentile = 0.91), and consequently the methylation information was merged using *coverage2cytosine* [Babraham Bioinformatics, 2016a].

2.3.4 Quality assessment and sample filtering

Systematic sequencing and base-calling errors that adversely affect downstream results are common and increasingly well characterised in high-throughput sequencing studies [Taub et al., 2010]. This is even more crucial for the METABRIC methylome study presented in this thesis, given the large number of samples involved and the unique features of RRBS libraries that make it susceptible to a variety of biases. While several of these biases (such as DNA fragments being shorter than the read length, have been described and controlled in Section 2.3.2), several other biases can arise and need to be detected. At each individual stage of the bioinformatics pipeline, a quality control (QC) step is introduced to guide the identification of the error-prone stage so that quick steps that can be taken to mitigate these errors. These steps are described below.

2.3.4.1 Read quality control

Quality control reports were generated for the raw sequence data using *FASTQC (version 0.11.2)* [Andrews Simon, 2015]. If base quality and base composition of the reads are not as expected, then trimming of the reads was performed as detailed in Section 2.3.3.1.

2.3.4.2 Genotyping

Single Nucleotide Polymorphism (SNP) data obtained from the Affymetrix SNP 6.0 arrays from the original METABRIC publication [Curtis et al., 2012] was used as the original genotyping data, and remapped from the genome assembly hg18 (on which genotyping was originally performed) to hg19 using the *LiftOver* package [Hinrichs, 2006]. Genotypes from the RRBS sequencing data was obtained by using GATK's Unified Genotyper with default parameters to call SNPs at the loci shared with the SNP 6.0 array. C/T and G/A SNPs were removed to prevent confounding with unmethylated cytosines. For each sample, the percentage of SNPs with the same genotype across the two platforms was used to confirm sample identity. Dr Harry Clifford performed genotyping on the samples in order to identify potential sampling or plating errors. The identities of 24 samples were corrected as a consequence of genotyping.

2.3.4.3 Read counts and depth of coverage

Multiplexing leads to a potential imbalance of samples within the library pools, which could result in decreased sequencing coverage in some samples. Although, the RBBS library preparation protocol has been optimised to reduce this imbalance (Section 2.3.1), some samples may not yield a sufficient number of unique CpG sites with adequate coverage due to insufficient input DNA. Figure 2.7a displays the number of aligned reads obtained for the METABRIC samples retained after filtering (median = 16.98 million reads, 25th percentile = 13.63 million reads, 75th percentile = 20.11 million reads). The mapping median mapping efficiency was 77.0% (25th percentile = 75.3%, 75th percentile = 78.1%; Figure 2.7b). As a reference, average mapping efficiency for an RRBS library is typically around 60%-65% [Garrett-Bakelman et al., 2015; Li et al., 2016b].

Only samples with more than 1.5 million unique CpGs at a minimum 5× coverage were retained. 100 samples (96 tumours and 4 normals) were discarded as a result of failing this criterion. After filtering, the number of unique CpGs with 1×, 5× and 10× coverage were computed for each sample (Figure 2.7d-f). The METABRIC samples had a median of 2.82 unique million CpGs at 5× (25th percentile = 2.52 million CpGs, 75th percentile = 3.05 M CpGs); median of 4.82 million unique CpGs at 1× (25th percentile = 4.54 million CpGs, 75th percentile = 5.20 M CpGs); and 1.88 million unique CpGs at 10× (25th percentile = 1.58 million CpGs, 75th percentile = 2.13 million CpGs).

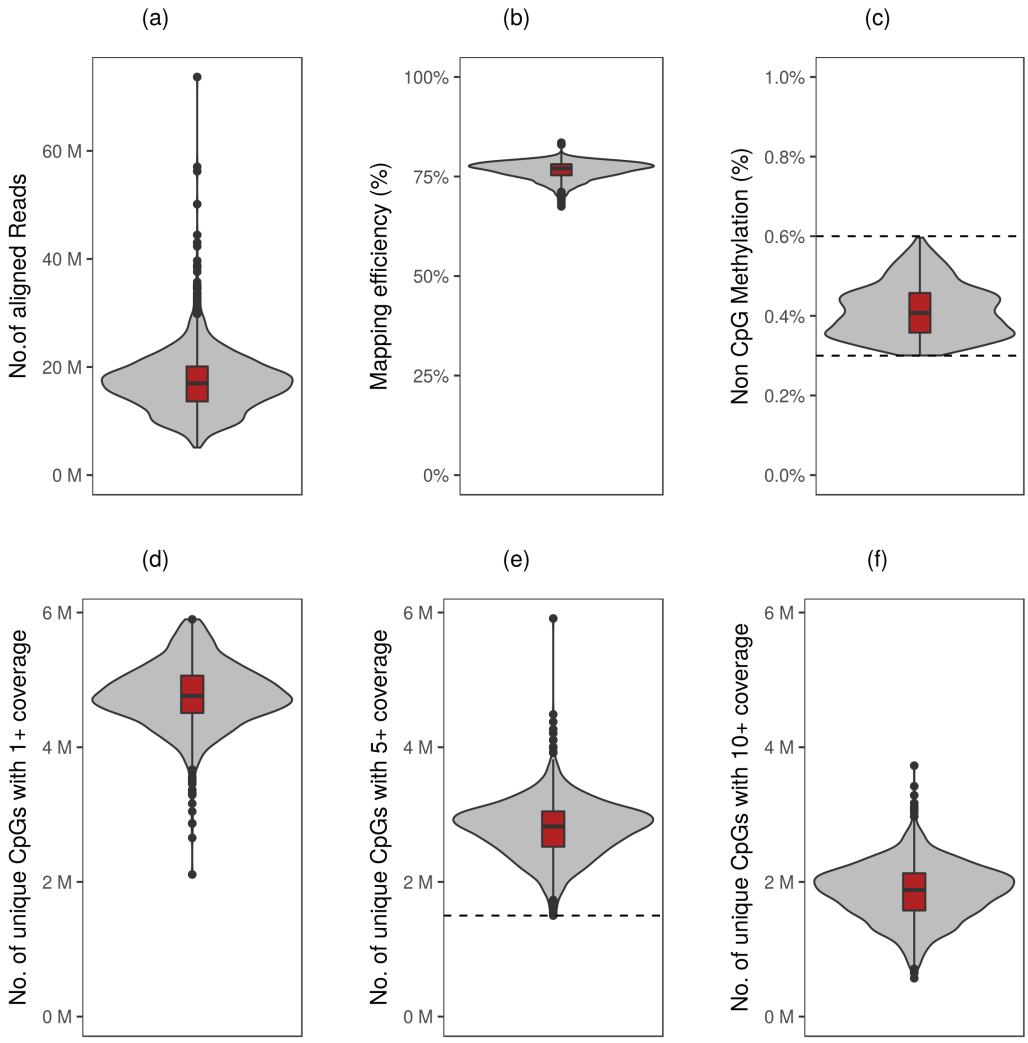


Figure 2.7: Sample filtering was based on the number of CpG sites detected at 5× coverage and non-CpG methylation %. **(a)** No. of aligned reads for all samples. **(b)** Mapping efficieny (%) for all samples. **(c)** Non-CpG methylation (%) for all samples. Dashed lines at 0.3% and 0.6% represent the thresholds used for sample filtering. **(d)** No. of CpG sites detected at 1× coverage. **(e)** No. of unique CpG sites detected at 5× coverage. Dashed lines at 1.5 M unique CpG sites represents the threshold used for sample filtering. **(f)** No. of CpG sites detected at 10× coverage. In (a) – (f) all METABRIC samples retained after filtering are plotted. Boxplots (red) indicate the interquartile range with 25th percentile, median and 75th percentile values illustrated. Whiskers indicate 1.5× of the interquartile range. M = million. No. = Number.

2.3.4.4 Bisulphite conversion

After methylation calling, the sensitivity and specificity of the bisulphite conversion is monitored. Elevated levels of observed non-CpG methylation can provide an indication of incomplete bisulphite conversion since non-CpG dinucleotides are rarely methylated in mammalian cells [Ziller et al., 2011].

Traditionally, non-CpG methylation levels of > 1% are used as a threshold to identify and discard samples that suffered from incomplete bisulphite conversion [Gu et al., 2010; Kulis et al., 2015]. However, one of the aims of the METABRIC project is to investigate and compare epiclonal compositions of the samples, and epiclonal estimates are extremely sensitive to variations in bisulphite conversion [Li et al., 2016b]. Consequently, a more stringent interval of [0.3%-0.6%] was used as a quality control criteria to retain samples. 122 samples (101 tumours and 21 normals) were discarded as a result of failing this criterion. Figure 2.7c portrays the global non-CpG methylation % for each retained sample.

2.3.4.5 Sample inclusion criteria

Only primary invasive breast tumours and adjacent normal breast tissues from female patients were retained. Ductal carcinoma in situ (DCIS, n=12), Non-invasive tumours (n=6), and recurrent tumours (n=13), and tumours from male patients (n=8) were removed. Where replicates or bilateral tumours exist, the sample with the highest number of CpGs at 5× coverage was picked to represent the patient. 7 replicates and 5 bilateral tumours were discarded as a consequence.

Overall 1719 METABRIC samples (1482 breast tumours and 237 adjacent normal tissues) were retained after quality control. Stratifying the breast tumours by three established taxonomies (see Chapter 1) – ER Status, Intrinsic subtypes and Integrative clusters – revealed that distinct breast cancer subtypes are also very well represented (Figure 2.8).

2.3.5 Determination of CpG sites

2.3.5.1 CpG site filtering

The starting sites of reads generated from RRBS are not random, but are dictated by *MspI* digestion, and hence this leads to an inability to distinguish PCR duplicates from two cancer cells having the same genotype and methylation information. Consequently, if experiments suffer from a high degree of PCR duplication, some reads will be

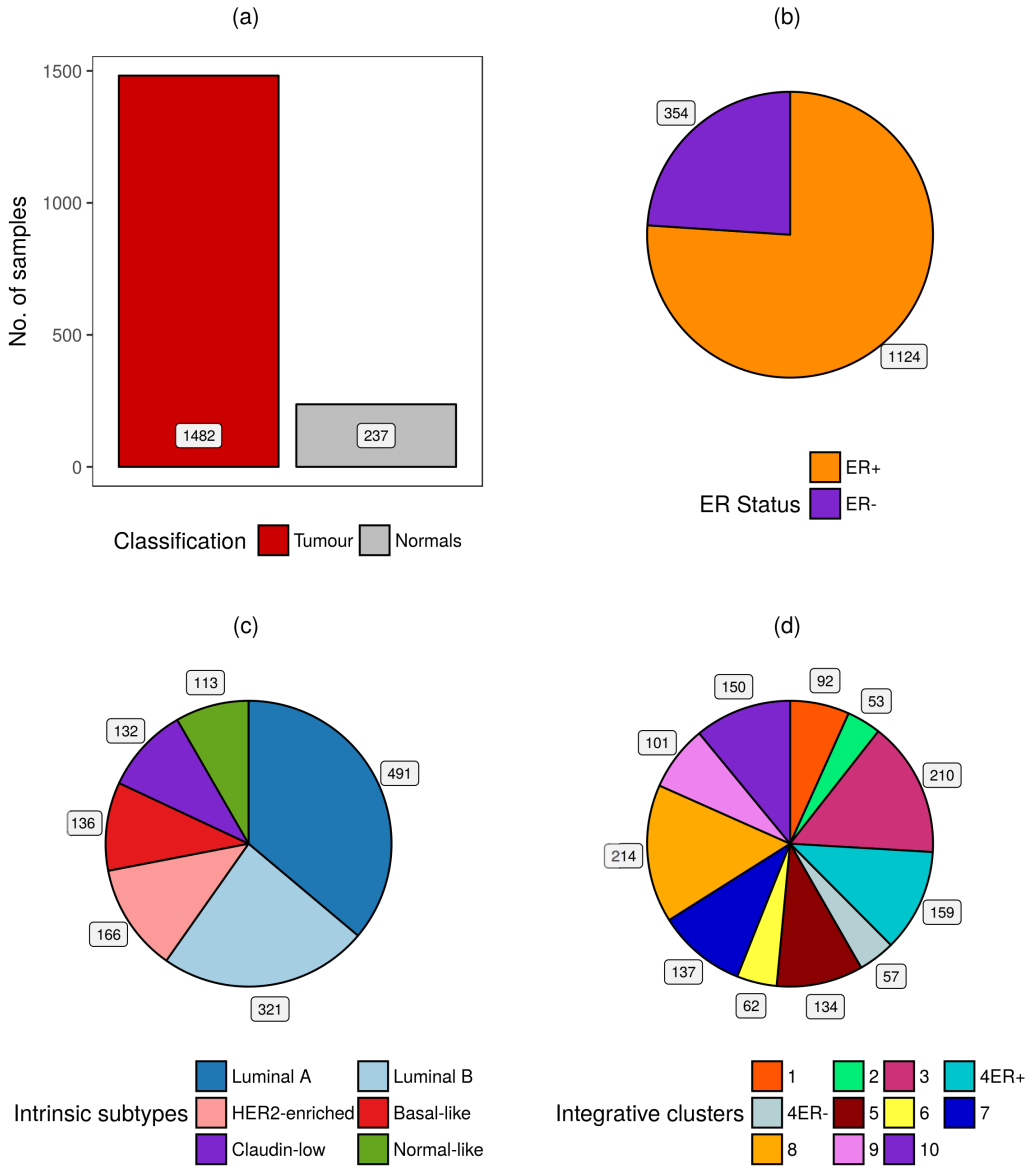


Figure 2.8: Breast cancer subtypes are well represented in the METABRIC methylome study. (**a**) Number of tumour and normal tissues retained in the study. **(b)** Distribution of tumours based on ER status. **(c)** Distribution of tumours based on Intrinsic subtype membership. **(d)** Distribution of tumours based on Integrative cluster membership.

preferentially over-amplified and will impair accurate determination of methylation estimates for these regions.

Read coverage per base distribution is an important metric that aids in filtering CpG sites that suffer from PCR duplication bias and consequently have very high read coverage [Akalin et al., 2012a]. CpG sites that have more than 99.99 percentile of coverage in each sample were discarded, as long as they are not located on known copy number amplifications (based on copy number estimates, see Section 2.2). This filtering step also avoids the inclusion of centromeric or telomeric repetitive regions [Kulis et al., 2015].

Furthermore, sufficient read coverage is required for providing adequate power of statistical tests, and accordingly CpG sites with low read coverage were also discarded. The number of CpG sites covered at various minimum depths and by different proportions of samples were also computed (Figure 2.9a). Increasing the minimum depth required, decreased the number of total number of CpG sites profiled, but increased the median coverage for the cohort. *1.25 CpG universe* represents the set of 5.41 million CpG sites with a minimum 1$\times$ coverage that are profiled in at least 25% of METABRIC samples. *5.50 CpG universe* represents the set of 2.72 million CpG sites with a minimum 5$\times$ coverage that are profiled in at least 50% of METABRIC samples. For this CpG universe, CpG sites covered by less than 5 reads were discarded due to insufficient confidence in their methylation estimates. 1506 METABRIC samples (87.6%) harboured more than 2 million CpG sites within the *5.50 CpG universe* (Figure 2.9b). *10.50 CpG universe* represents the set of 1.79 million CpG sites with a minimum 10$\times$ coverage that are profiled in at least 50% of METABRIC samples. Where single CpG information is analysed in this thesis, the *5.50 CpG universe* is used, with the exception of Chapter 5 in which a different CpG selection criterion is applied for the investigation of intratumour heterogeneity.

2.3.5.2 Annotation

RefSeq transcript annotation for hg19 was obtained from UCSC genome browser [Karolchik et al., 2014] to define TSS, exons and introns for each gene. Promoters were defined as 2 Kbp upstream and 500bp downstream of the transcription start site (TSS). One gene may be associated with more than one promoter. Intergenic regions were defined as regions complementing promoter and gene body regions. CpG island, CpG shore and open sea definitions for hg19 were also obtained from the UCSC genome browser [Karolchik et al., 2014].

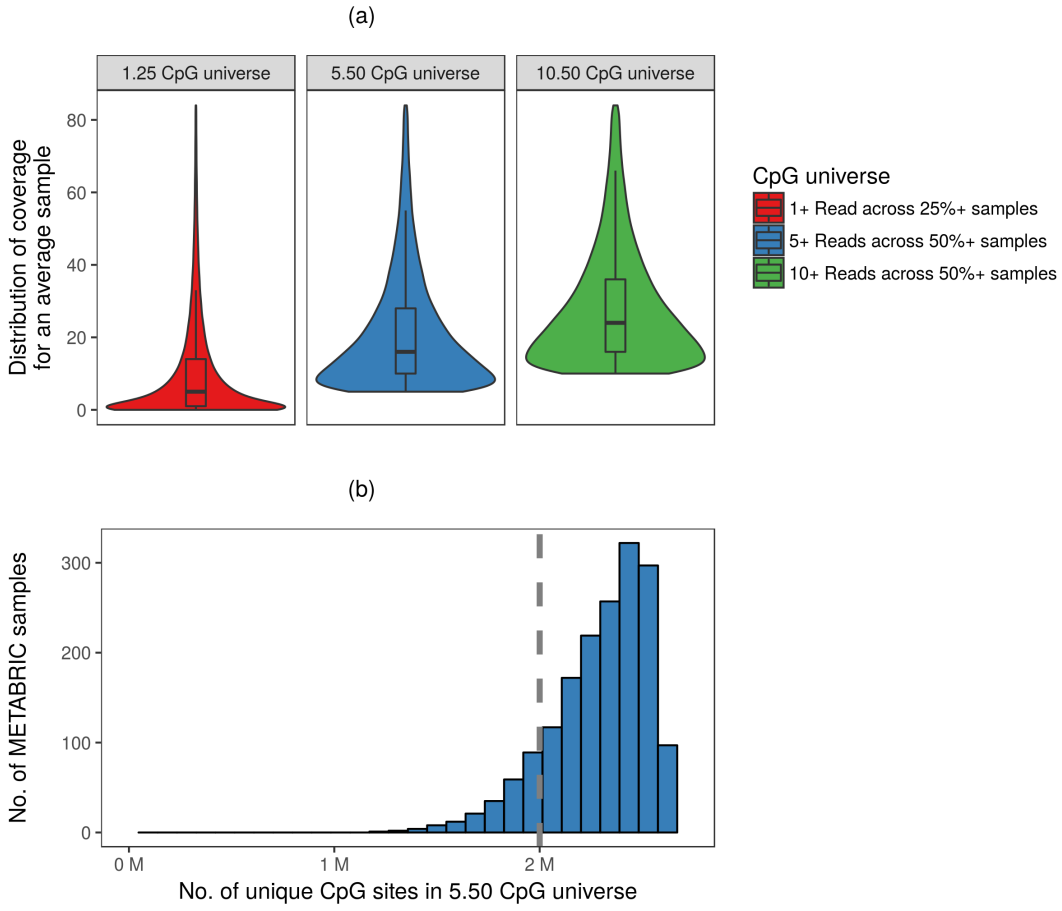


Figure 2.9: *5.50 CpG universe* represents the set of 2.7 M CpG sites at $\geq$5$\times$coverage profiled in $\geq$ 50% of METABRIC samples. (a) Comparison of coverage distribution for 3 CpG universes: (left) *1.25 CpG universe*, (middle) *5.50 CpG universe* (right) *10.50 CpG universe*. Definition of universes given in text. **(b)** Distribution of the number of unique CpG sites assayed as part of the 5.50 CpG universe in the METABRIC samples. Dashed vertical grey line represents 2 M unique CpG sites. M=million.

Human Mammary Epithelial Cell (HMEC) ChromHMM annotations for hg19 were downloaded from ENCODE to define the chromatin states based on genome-wide histone marks [Ernst et al., 2011]. The chromatin states summarised regions with coordinated chromatin modification patterns and included enhancers, polycomb-repressed chromatin (PRC) regions, insulators, heterochromatin and repetitive regions. Figure 2.10a illustrates the distribution of the CpG sites within the *5.50 CpG universe* according to RefSeq definitions, CpG content definitions and chromatin state definitions. Figure 2.10b compares the total number of CpG sites covered in the RRBS *5.50 CpG universe* with that covered in the Illumina HM450 microarray panel and the Illumina EPIC microarray panel. The RRBS *5.50 CpG universe* clearly interrogates a larger number of CpG sites in promoters, enhancers and PRC regions. Figure 2.10c illustrates the proportion of each genomic feature that is assayed as part of the *5.50 CpG universe*. For instance, approximately 75% of promoters present in the genome are covered by at least 1 CpG sites in the *5.50 CpG universe*. However, only 66% of promoters present in the genome are covered by at least 5 CpG sites in the *5.50 CpG universe*. Similarly, the gene bodies (comprising of all introns and exons of a gene) of 70% of all genes are covered by at least 5 CpG sites in the *5.50 CpG universe*.

Genes were also annotated by their known biological function (if any) by using gene-family annotation obtained from the Molecular Signatures Database [Subramanian et al., 2005, MSigDB]. Gene-family categories included tumour suppressors, oncogenes, translocated cancer genes, protein kinases, cell differentiation markers, homeodomain proteins, transcription factors and cytokines/ growth factors.

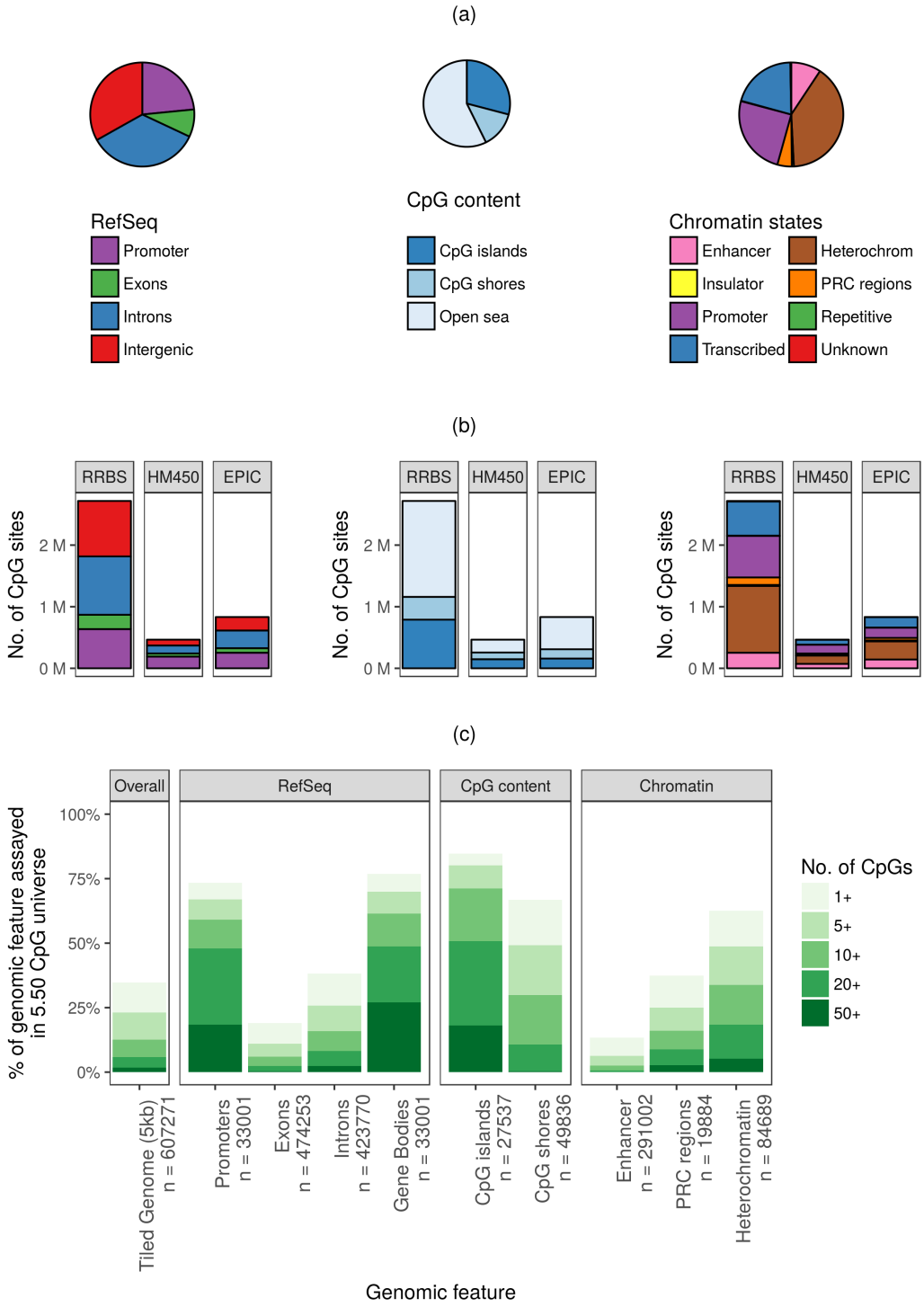


Figure 2.10: (Caption on next page.)

Figure 2.10: (Previous page.) **Proportion of promoters, exons, introns, enhancers and PRC regions interrogated by RRBS. (a)** illustrates the distribution of the CpG sites within the *5.50 CpG universe* according to RefSeq definitions, CpG content definitions and chromatin state definitions. **(b)** compares the total number of CpG sites covered in the *5.50 CpG universe* with that covered in the Illumina HM450 microarray panel and the Illumina EPIC microarray panel, according to RefSeq definitions, CpG content definitions and chromatin state definitions. **(c)** illustrates the proportion of each genomic feature that is assayed as part of the *5.50 CpG universe* at different overlap thresholds (CpG overlap at 1+, 5+, 10+, 20+, 50+ are denoted with different shades of green).

2.4 Analysing DNA methylation using RRBS

With recent advances in NGS technologies for methylation profiling such as WGBS and RRBS, there is an increasing demand for statistical tools to analyse bisulphite sequencing data. However, bisulphite sequencing technologies present substantial challenges in terms of data processing, statistical analysis and biological interpretation of observed differences, which are addressed below.

2.4.1 Existing approaches to analyse bisulphite sequencing data

There are a number of approaches that can be used to analyse methylation information obtained through RRBS profiling. Four distinct approaches can be used in bisulphite sequencing experiments and are detailed below.

2.4.1.1 Single CpG resolution

Previous publications have focused on estimating and comparing precise methylation levels at single-base resolution. For instance, Fisher's exact test [Lister et al., 2009] and logistic regression [Akalin et al., 2012b] have been used to identify single CpGs that are differentially methylated across two samples or groups of samples respectively. However, comparing methylation levels over single base resolution, as described in these studies, has a few disadvantages.

Firstly, methods such as Fisher's exact test and logistic regression either completely ignore biological variability within the conditions tested or underestimate it. Methylation levels between tumours are known to be highly variable [Györffy et al., 2016; Holm et al., 2016], and defining differentially methylated CpG sites in tumours versus normal tissues using these two approaches leads to overdispersion. As a result, the overall variability is underestimated resulting in overestimation of significantly differentially methylated CpG sites in tumours. The beta-binomial model is a more appropriate statistical method for identifying differentially methylated CpGs in bisulphite sequencing studies [Hebestreit et al., 2013].

Secondly, any statistical method that tests for differences in DNA methylation at a large number of genomic loci needs to correct for *multiple hypothesis testing*. This correction is usually done by controlling the ***false discovery rate (FDR)***[1], in which the

[1] **False discovery rate (FDR).** The false discovery rate is an estimate of the proportion of significant results (usually at *alpha = 0.05*) that are false positives. Originally developed by Benjamini and Hochberg [1995], FDR procedures essentially correct for this number of expected false discoveries, providing an estimate of the number of true results among those called significant.

distribution of uncorrected *p-values* is analysed and an FDR is inferred for each DMR. Because of the large number of CpGs in the genome, only the strongest single-CpG differences tend to remain significant after multiple testing corrections [Kuan and Chiang, 2012]. The result is often a high false-negative rate, especially when sample numbers and effect sizes are small.

Thirdly, the standard error of the single-CpG methylation estimate is inversely proportional to the number of reads covering the CpG. A high coverage of at least 30x per CpG site is required to minimise the standard error to an adequate level for accurate detection [Hansen et al., 2012]. However, such a high coverage design is unfeasible in particular for studies with high sample sizes.

2.4.1.2 Annotated region-level analysis

The regulatory role of DNA methylation at different genomic features has been discussed at length in Chapter 1. For instance, hypermethylated promoters in cancer have been associated with a strong repressive effect [Esteller, 2000], whereas loss of gene body methylation has been associated with both the suppresson and stimulation of genes [Kulis et al., 2012]. Therefore, delineating the role of DNA methylation at different genomic features should be of interest for DNA methylation based analyses. Furthermore, functionally relevant findings are generally associated with multiple CpG sites rather than a single CpG site, and methylation properties of the region as a whole determines its function [Eckhardt et al., 2006; Laird, 2010].

Therefore, an alternate option to single nucleotide analysis would be to carry out comparisons on predefined large genomic regions by averaging methylation levels over all CpG sites within individual genomic features, and then proceeding with statistical tests. Genomic features included in the analysis are i) CpG-density related regions such as CpG islands and CpG shores; ii) gene related regions such as promoters, gene bodies, and intergenic regions; iii) regulatory regions such as enhancers and PRC regions. This delivers an improvement over a single CpG analysis since neighbouring CpGs with similar differences in DNA methylation reinforce each other and improve the statistical power for detecting weak differences.

However, a limitation of using pre-defined genomic regions is that they may not exactly overlap functional differentially methylated regions. For instance, if a large fraction of a differentially methylated region extends beyond the annotation-defined region or is significantly smaller, then averaging methylation differences over all CpGs

within the annotation-defined region would effectively dilute this signal and potentially lead to a high false-negative rate (loss of detection).

2.4.1.3 Defining differentially methylated regions based on clustering of spatially correlated CpGs

The approach detailed in the previous subsection operates by first defining regions, and then conducting downstream analysis. Due to the aforementioned limitations of this approach, methods that detect differentially methylated regions unconstrained by a priori region definitions are more suitable. However, explicitly defining differentially methylated regions (DMRs) using single CpG information is difficult since controlling the FDR at the region-level whilst simultaneously defining the region poses a statistical challenge due to the spatial correlation of CpG sites [Bock, 2012; Robinson et al., 2014]. A number of methods have been developed that adjust for the correlation in neighbouring CpG sites such as the Stouffer-Liptak test [Dolzhenko and Smith, 2014; Li et al., 2013]. This additional step allows identification of precise differentially methylated regions with appropriate FDR control over single CpG sites. Other methods that attempt to control this FDR including a module for block finding for microarrays [Aryee et al., 2014]; cluster-wise weighted FDR strategy [Hebestreit et al., 2013]; and a tool called comb-p for DMR detection by combining spatially correlated *p-values* [Pedersen et al., 2012].

Although these approaches are statistically appropriate for RRBS analysis, they are only suitable for defining differentially methylated regions between two samples or between two groups of samples [Dolzhenko and Smith, 2014; Li et al., 2013]. Consequently, distinct region universes are constructed for each tumour and every comparison tested. This does not allow interrogation of methylation profiles over a common set of regions across all samples.

2.4.1.4 Smoothed gene unit analysis

The distribution of the epigenetic marks in the genome depends on distinct genomic features. Methylation that occurs directly on genes (promoter and gene bodies) is intuitively associated with regulation of the underlying gene itself and so it might be beneficial to focus the statistical analysis on genes separately. Furthermore, DNA methylation levels at neighbouring CpG sites have been shown to be strongly correlated [Zhang et al., 2015]. This implies that the probability that a CpG site is methylated can be assumed to vary smoothly along a gene, without distorting signal or losing

functional information [Hansen et al., 2012]. Precision can be improved by the use of modern statistical techniques such as local likelihood smoothing or smoothing splines.

A statistical method that combines the two concepts discussed above – a gene-centric analysis; and smoothing over CpG sites – has been developed in parallel to this thesis [Batra, 2015; Rueda, 2014, private communication] with the aim of increasing the statistical power as well as allowing gene-by-gene analyses. This method would enable testing of the role of individual genes between cancer and normal tissues (as well as between subtype-specific cancers), and also comparison of gene-by-gene differential patterns within the same tissue. However, this method is unsuitable for RRBS due to the disjointed nature of the profiled epigenome and is far more appropriate for WGBS and MBD-seq. Consequently, the smoothed gene unit analysis has not been utilised for quantifying methylation in this thesis.

2.4.2 Novel Method – Spatially Coordinated CpG-sites within the RRBS Universe in Breast cancer (SCCRUB)

One of the reasons that the microarray platform has been popular compared to bisulphite sequencing experiments is the relative ease in bioinformatics analysis. Each probe in a microarray analysis can be treated independently; conversely single CpG analysis in bisulphite experiments suffers from inadequate power and low functional relevance. Moreover, the unsuitability of the smoothed gene approach (previous subsection) in RRBS necessitates the need for a suitable method to delineate distinct regions for analysis. However, using predefined annotated regions for analysis results in the dilution of signal; and although explicitly defining differentially methylated regions overcomes this obstacle (discussed in Section 2.4.1.3), it does not yield a common set of regions for all tumours.

A novel algorithm called **S**patially **C**oordinated **C**pG-sites within the **R**RBS **U**niverse in **B**reast cancer (SCCRUB) is introduced that attempts to transform the RRBS data into distinct regions comprising of multiple CpG sites coordinated in their methylation behaviours[2]. The objective of this effort is to generate a universe of regions that is potentially functionally relevant and is smaller in size than the number of single CpG sites. This SCCRUB algorithm specifically identifies clusters of spatially correlated CpG sites across tumours and normal tissues by leveraging: a)

[2] This novel approach stemmed from discussions with Dr Cem Meydan during a travel fellowship to Professor Ari Melnick and Professor Chris Mason's laboratory at Weill Cornell Medicine, New York. The two principal investigators are incidentally the authors of Li et al. [2013], a method that defines DMRs based on clustering of spatially correlated CpG sites.

the large number of samples available (both tumours and normal tissues); and b) the single nucleotide resolution methylation estimates obtained from RRBS profiling. The algorithm consists of four components that are described below.

2.4.2.1 STEP 1: Determination of empirical region boundaries

In order to identify a set of regions comprising of neighbouring CpGs in close spatial proximity, an optimal threshold of genomic distance has to be estimated to determine the minimum gap between two distinct regions. This threshold is used to separate two adjacent CpG sites so that they are not considered in the same region. This is particularly vital for RRBS data since it is based on restriction enzyme digestion; and consequently, the profiled CpGs are not evenly distributed across the genome. The method and tool presented in Li et al. [2013] was used to determine this optimum boundary cut-off; however, the method was implemented on the *5.50 CpG universe* (which is a common universe of CpGs covered at $\geq 5\times$ in at least 50% of all samples; n = 2.7 million; Section 2.3.5), and not on a per sample basis as described in the original publication. Another key distinction is that method was originally used on differentially methylated CpG sites, whereas in this case it is implemented on all CpG sites. The distribution of the distance between consecutive CpG sites (using the *5.50 CpG universe*; n = 2.7 million) was examined. Log_2 transforming revealed a bimodal normal distribution (Figure 2.11a), with the assumption being that the first normal distribution represents distances between CpG sites belonging to the same functional region, whereas the second normal distribution represents the distance between boundary CpGs of distinct regions. An expectation maximisation (EM) algorithm was used to fit the bimodal normal distribution, and the weighted combined probability function was minimised to determine the appropriate boundary cut-off ($D = e^{8.2}$ = 294 bp) as described in Li et al. [2013]. This threshold represents the best separation point between the two distributions, and thus determines the minimum gap between consecutive functional region boundaries.

2.4.2.2 STEP 2: Defining regions

Once the boundary threshold (D=294 bp) was determined, regions comprising of neighbouring CpG sites were identified based on two criteria:

1. Distances between adjacent CpG sites were examined, and consecutive CpG sites with distances less than 294 bp were clustered into the same region. A distance of greater than 294 bp between consecutive CpG sites was used to mark the boundary point of two distinct regions.

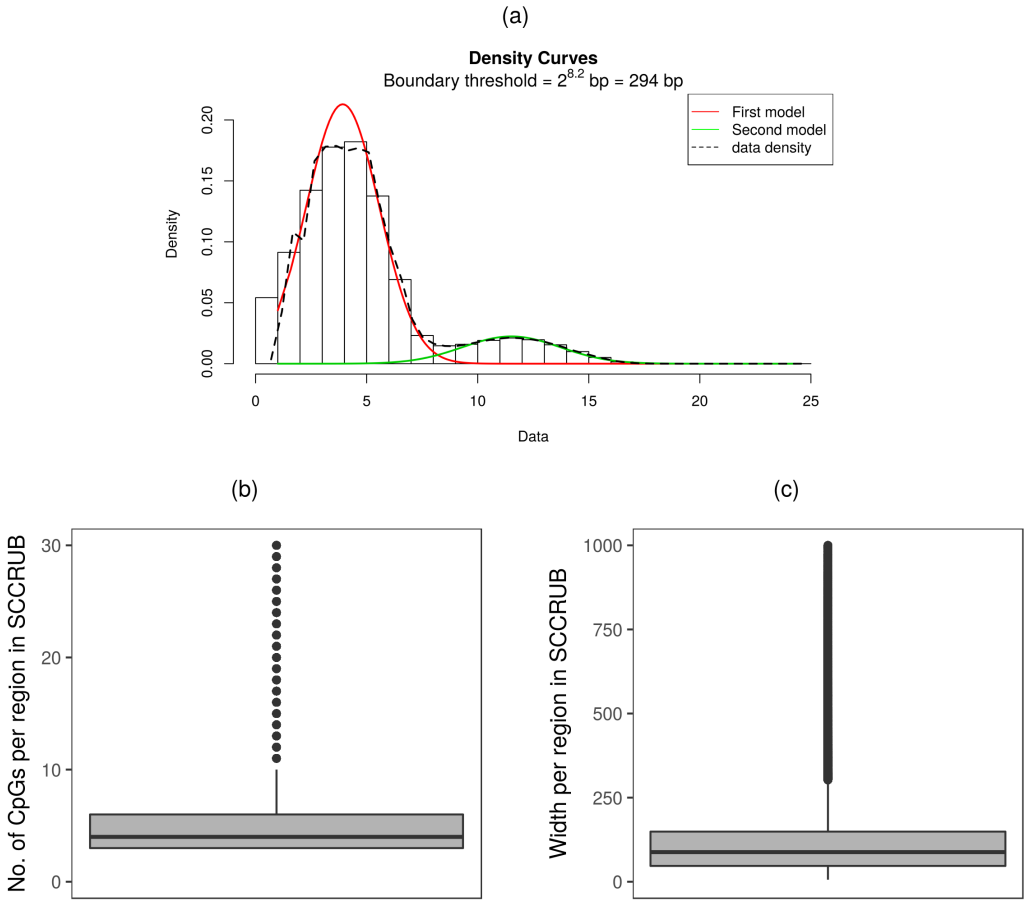


Figure 2.11: Generation of the novel SCCRUB universe. (a) The distribution of the distance between consecutive CpG sites within the *5.50 CpG universe*. x-axis is $\log_2$ transformed. The first normal distribution (red) represents distances between CpG sites belonging to the same functional region. The second normal distribution (green) represents the distance between boundary CpGs of distinct regions. The boundary cut-off is empirically estimated to be 294 bp and is used to determine the minimum gap between consecutive region boundaries. **(b)** Distribution of the number of CpG sites in each region in SCCRUB. **(c)** Distribution of the width of each region in SCCRUB. In (b) and (c), boxplots indicate the interquartile range with 25$^{\text{th}}$ percentile, median and 75$^{\text{th}}$ percentile values illustrated. Whiskers indicate 1.5$\times$ of the interquartile range.

2. Regions with at least 3 CpG sites were included, and the rest were discarded.

2.4.2.3 STEP 3: Filtering regions based on autocorrelation of CpG sites

CpG sites in the regions defined in STEP2 were further refined based on whether the CpG sites within a region were spatially correlated with each other. For each region, the following steps were implemented to obtain spatially correlated filtered sub-regions.

1. Spatial autocorrelations of methylation (M-values) between CpG sites were calculated using both tumour and normal samples. Two autocorrelations were calculated using lags of i) 1 CpG site, and ii) 2 CpG sites respectively, to account for inconsistencies in single CpG estimates. The maximum of the two was recorded for each CpG site.

2. A correlation threshold of 0.7 was used to mark a CpG site as PASS.

3. A forward sliding CpG approach (starting with the CpG with lowest base position and going higher) and backward sliding CpG approach (starting with CpG with highest position and going lower) were used to detect maximal sub-regions such that at least 60% of its CpG sites are marked as PASS. If the forward and backward approaches yield overlapping sub-regions, then the union is recorded; otherwise the sub-regions are recorded as two distinct regions.

4. Sub-regions with at least 3 CpG sites were included.

5. Step 2-4 were repeated with tiered correlation thresholds of 0.6, 0.5, 0.4, 0.3 and 0.2.

6. Sub-regions obtained from the 6 distinct correlation thresholds were compared. For overlapping sub-regions, a preference for the sub-region passing the most stringent (highest) correlation threshold (and consequently comprising the lowest number of CpG sites) was recorded. Non-overlapping sub-regions were recorded as distinct entities. Consequently, each region defined in STEP2 can yield more than one spatially correlated filtered sub-regions.

7. For each filtered sub-region, samples methylation estimates were obtained by aggregating estimates (Beta-values) over all CpG sites within the region. Methylation estimates were marked as unknown for samples that harboured less than 3 CpG sites for a sub-region since this constitutes insufficient data.

2.4.2.4 STEP 4: Annotation

Finally, the universe of spatially correlated filtered regions was comprehensively annotated using i) RefSeq gene coordinates: promoters, exons, introns and intergenic regions; ii) CpG content; and iii) chromatin based marks such as enhancers and PRC regions (see Section 2.3.5).

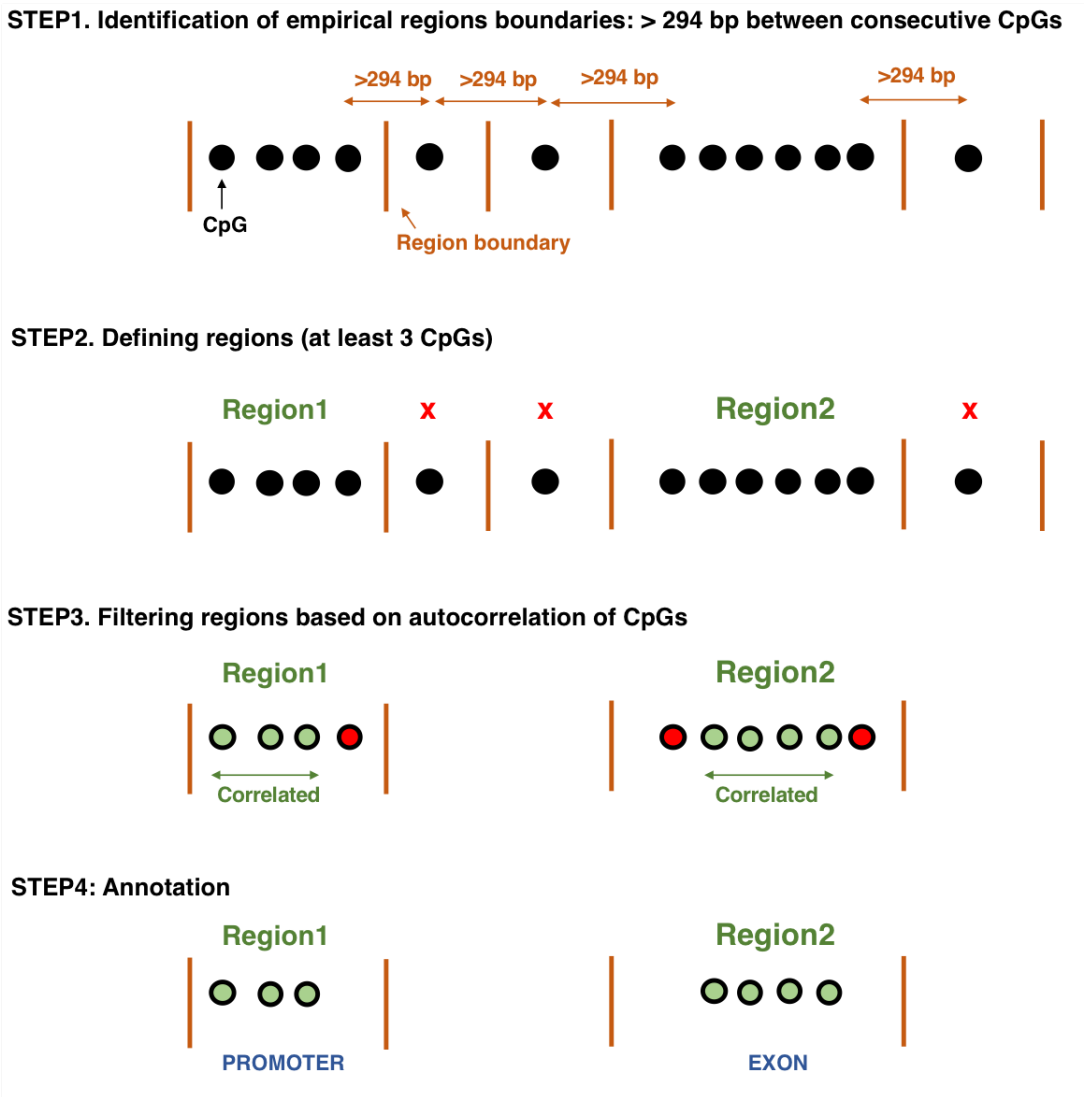


Figure 2.12: Schematic diagram of the SCCRUB algorithm.

A schematic diagram of the SCCRUB algorithm is shown in Figure 2.12. Implementing the SCCRUB algorithm on the *5.50 CpG universe* yielded 289,265 regions comprising of 4 CpG sites on average (median = 4, 25th percentile = 3, 75th percentile = 6; Figure 2.11b), and with an average width of 88bp (median = 88, 25th

percentile = 47, 75th percentile = 150; Figure 2.11c). The underlying premise of this universe of regions is that it potentially delineates all the regions within the RRBS-universe in which there is strong evidence that multiple CpGs (at least 3) are spatially coordinated in their methylation statuses. This has led to the identification of a novel set of regions that are potentially more functionally relevant than single CpG sites, but also represent a significantly smaller universe leading to lower false positives (SCCRUB: n = 289 K, *5.50 CpG universe*: n = 2,700 K = 2.7 M). Moreover, the SCCRUB algorithm is an unsupervised method unconstrained by any particular comparison, and consequently the resulting fixed set of SCCRUB regions can be used for any downstream analysis much like microarray probes. 289K distinct regions are assessed in SCCRUB, a manageable number for downstream statistical analysis, that can be compared to current and previous microarray technologies (HM27 = 27K probes; HM450 = 450K; EPIC = 850K probes). However, these regions comprise of methylation statuses of 1.7M spatially correlated CpG sites (and are constructed from the *5.50 CpG universe* of universe of 2.7M CpG sites). In this thesis, the SCCRUB universe of regions will be utilised exclusively in those downstream analyses that are involved in identifying regions of ordered (comprising of coordinated CpG sites) methylation alterations that have potential to be functional, such as the differential methylation analysis in Chapter 3. However, analyses in which stochastic methylation changes are examined will require single CpG investigations, and in these cases the *5.50 CpG universe* will be employed. Although, the SCCRUB universe provides many advantages (as described above), it is important to note that a large number of samples (both tumours and normal tissues) as well as large inter-tumour representation is necessary to accurately estimate CpG autocorrelations per region to feed accurate region definitions.

2.5 Pipeline validation

2.5.1 Tumour and normal methylation statuses

In order to conduct an unsupervised global assessment of the methylation profiles of the samples, the top 50% variably methylated SCCRUB regions (n=144632 regions) were chosen. t-Distributed Stochastic Neighbour Embedding (t-SNE), a technique for dimensionality reduction [Maaten and Hinton, 2008] was conducted on the pairwise Euclidean distance matrix obtained from the 1719 x 144632 matrix of methylation estimates over all 1719 samples and 144632 selected SCCRUB regions. Consequently, the methylation landscape of each sample (consisting of 144632 SCCRUB regions) could be largely represented by its two t-SNE components. Figure 2.13a illustrates the first two t-SNE components for each sample, and this allows easy visualisation and examination of the high dimensional methylation profiles for each sample. The normal tissues (grey crosses) exhibit a spectacular divergence from the tumours (red dots) implying distinct methylation profiles for the two groups.

Finally, tumours were classified according to ER status, as this classification encompasses significant differences in the fundamental biology of breast tumours; further, the normal tissues were also classified according to the ER status of the corresponding adjacent matched tumours (Figure 2.13b). The top two t-SNE components (representing sample-specific DNA methylation profiles) were tested for association with ER Status in the tumours and normal tissues separately (multinomial linear regression; Figure 2.13c). The DNA methylation profiles of the tumours are highly significantly associated with ER status which indicates that in addition to tumour-normal differences in breast cancer (Figure 2.13b, c) it would also be beneficial to test for subtype-specific differences in ER+ and ER- tumours separately. Crucially, normal samples have homogeneous methylation profiles with respect to the corresponding adjacent tumours' ER status (Figure 2.13b, c) which implies that unlike tumours, it is not necessary to divorce normal tissues by their respective matched subtypes. Accordingly, for all the analyses in this thesis, all normal tissues were pooled together, with the exception of intra-sample investigation in Chapter 5. Similar dichotomies between tumour and normal samples were noted for the Intrinsic subtype and Integrative cluster classifications (Figure 2.13c).

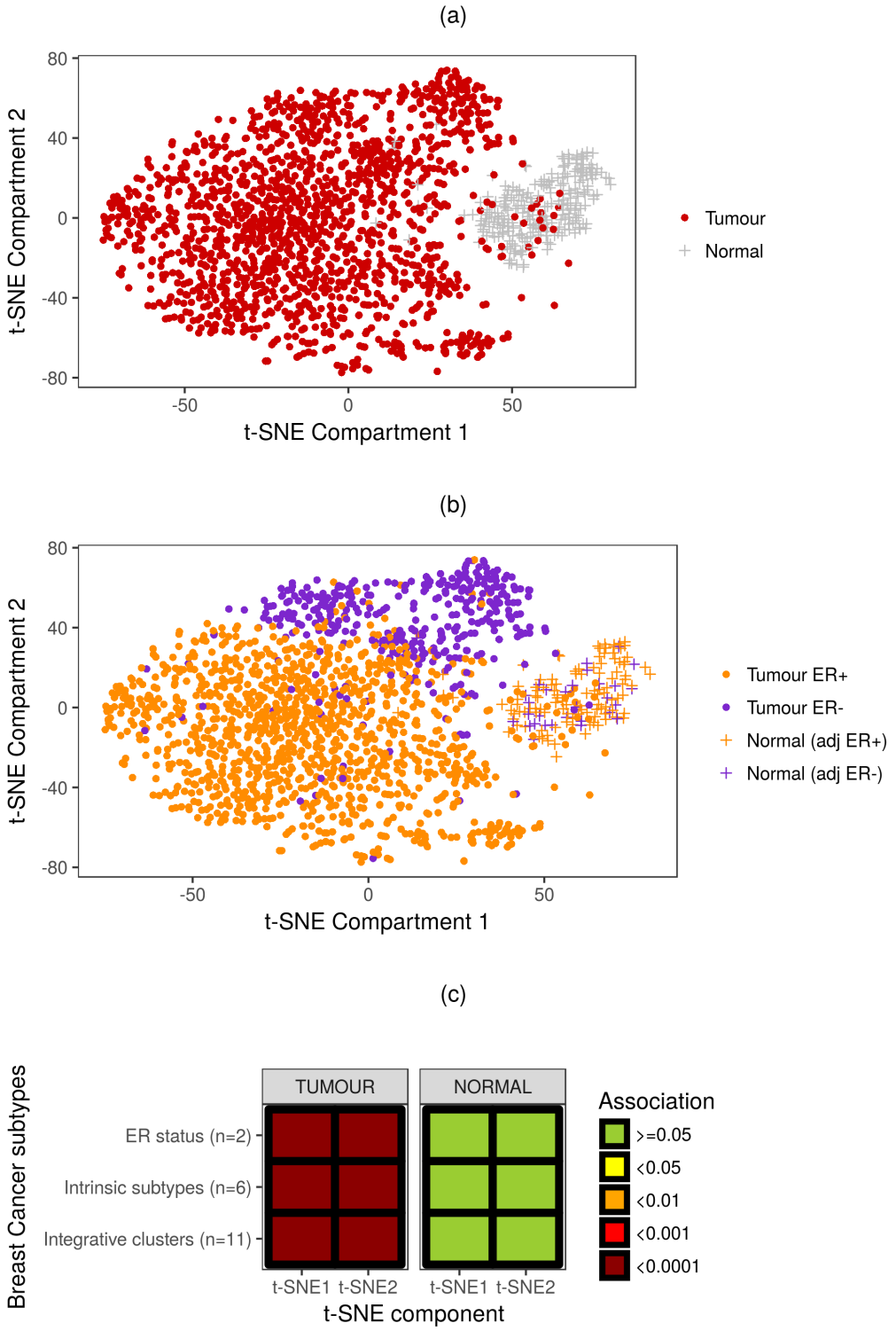


Figure 2.13: (Caption on next page.)

Figure 2.13: (Previous page.) **DNA methylation profiles of breast samples separate by tumour-normal classification and by tumour subtypes. (a)** The top 2 t-SNE components representing the sample specific methylation profile are plotted for each sample. Each point represents a breast tissue sample. Colour and shape of points represent tumour or normal classification. **(b)** The top 2 t-SNE components representing the sample specific methylation profile are plotted for each sample. Each point represents a breast tissue sample. Shape of points represent tumour or normal classification. Colour of points represent ER status in the case of tumours, and ER status of the adjacent matched tumour in the case of the normal tissues. **(c)** For three tumour subtype classifications (ER status, Intrinsic subtypes and Integrative clusters), the association between the tumour subtype and the 2 t-SNE components were investigated (multinomial linear regression). Similar associations were tested between the normal subtype (as defined by adjacent matched tumour subtype) and the 2 t-SNE components. Colour of the squares represent the level of significance of the association as detailed in the legend. adj = adjacent to tumour.

2.5.2 Technical variability and clinicopathological factors

In order to test whether technical effects and tumour-related clinicopathological factors were associated with DNA methylation profiles, a similar analysis (as above) was conducted for all 1482 tumours. The top two t-SNE components were tested with clinicopathological factors including age at diagnosis, tumour size, tumour grade and number of lymph nodes; and technical effects including METABRIC plate and RRBS batch (multiple plates were processed together and formed one batch). ER status was added as a confounder in the models. All clinicopathological factors were significantly associated with at least one t-SNE component suggesting that established clinical phenotypes have distinctive methylation profiles (linear regression; Figure 2.14). Conversely neither of the technical factors were associated with the 2 t-SNE components, implying that technical variability is not associated with the tumours' methylation profiles, and that a majority of the variation in methylation is explained by clinical/ biological variation.

2.5.3 Validation with HM450 data

To validate the RRBS pipeline, methylation calls from RRBS were compared to previous methylation estimates obtained previously using the HM450 microarray analysis of 12 METABRIC tumour samples (study conducted by Associate Professor Christina Curtis, Stanford University). Thirteen samples had data from both RRBS and HM450. Approximately 78.7K CpG sites were commonly covered in both HM450 and RRBS (read coverage ≥ 10). Sample-specific correlations were performed across these CpG sites yielding a median correlation of 0.93 (25th percentile = 0.90, 75th percentile

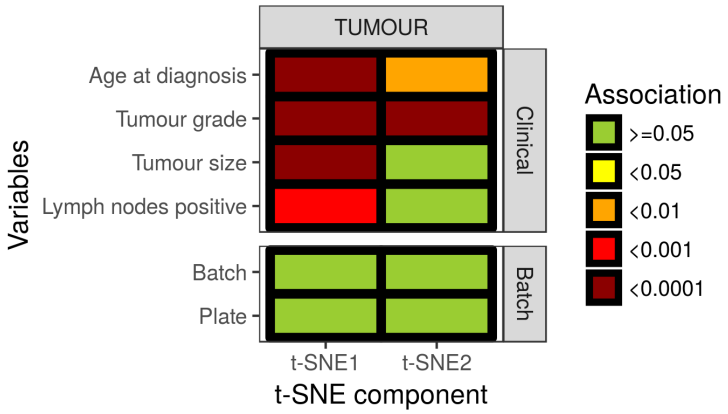


Figure 2.14: Clinicopathological variables but not technical variables are associated with DNA methylation profiles of the breast tumours. Associations between clinicopathological variables for the tumours and the 2 t-SNE components were investigated. Similarly, associations between technical variables and the 2 t-SNE components were investigated. Linear regression was used for continuous variables and multinomial regression was used for categorical variables. Colour of the squares represent the level of significance of the association as detailed in the legend.

= 0.94; Figure 2.15a). As expected, correlations between the two technologies increase with higher coverage in RRBS (Figure 2.15b). The high reproducibility between the RRBS platform and the HM450 platform indicates the robustness of these two technologies.

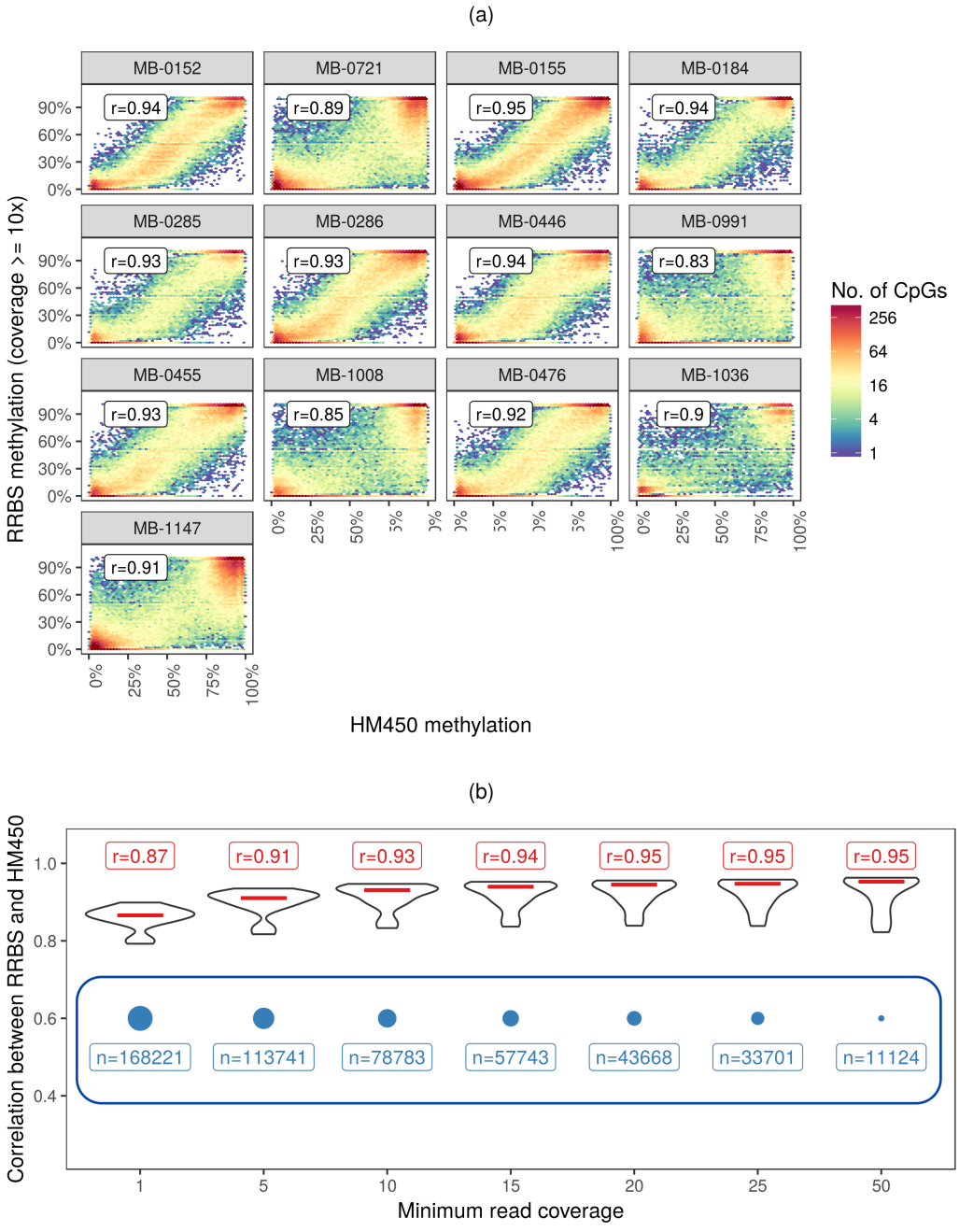


Figure 2.15: (Caption on next page.)

Figure 2.15: (Previous page.) **High reproducibility of DNA methylation calls between RRBS and HM450 platforms indicates the robustness of these procedures. (a)** Scatter plots of the RRBS and HM450 DNA methylation calls for thirteen breast tumours. Only CpG sites with RRBS read coverage $\geq 10\times$ and overlapping HM450 probes were plotted. Colours represent density of CpG sites. Median spearman correlation coefficients estimating the agreement between the RRBS and HM450 DNA methylation calls for each tumour were noted in the top left corner of the plots. **(b)** Distribution of the thirteen sample-specific median spearman correlation coefficients between the two methylation platforms, conducted for sets of CpG sites at different RRBS coverages that also overlap the HM450 probes. Median (red line) represents the average spearman correlation between the methylation calls from two platforms for the 13 breast tumours, and is noted above the boxplots. Size of the blue circle represents the number of CpG sites that were considered at the different RRBS read coverages, and the number is noted below the circle.

2.6 The breast cancer methylome in METABRIC

2.6.1 The global methylation landscape of breast cancer

The global methylation landscape is explored in tumour and normal samples. Investigating all CpG sites within the *5.50 CpG universe* revealed that on average, tumours have lower methylation levels genome-wide than normal tissues (Figure 2.16a, b). Stratifying the genome into relevant genomic features, also revealed differences (Figure 2.16a, b). An aggregate gain of methylation is observed in promoters (though difference was not significant in this study) and exons in breast tumours as observed in previous breast cancer studies [Fleischer et al., 2014; Györffy et al., 2016; Rønneberg et al., 2011]. On the contrary, a loss of methylation is observed in introns and heterochromatin in the tumours when compared to the normal tissue. In addition, enhancers and PRC regions are dramatically hypermethylated in breast tumours, agreeing with previous reports in other cancers [Ohm et al., 2007; Schlesinger et al., 2007; Widschwendter et al., 2007].

Comparing the aggregate methylation levels in the two predominant breast cancer subtypes, ER+ and ER-, tumours revealed that the global directionality of epigenetic change for a given genomic feature is the same in both subtypes (Figure 2.16c). Similar findings were observed when stratifying breast tumours by the Intrinsic subtypes and Integrative clusters (Appendix A.1). For example, all breast tumour subtypes have significantly increased global methylation in exons, enhancers and PRC regions. This indicates that the same fundamental mechanisms of epigenome control are breached in breast cancer resulting in the redistribution of DNA methylation across the genome. However, the extent of these changes vary across the different breast cancer subtypes, suggesting that the epigenome does contribute to defining these subtypes as distinct biological entities. A more formal investigation of tumour versus normal methylation alterations and subtype-specific methylation alterations is conducted in Chapter 3.

2.6.2 Comparison with the TCGA breast cancer methylome

The largest breast cancer methylome study preceding the METABRIC study described in this thesis was conducted by TCGA (The Cancer Genome Atlas) [Cancer Genome Atlas Network, 2012]. TCGA is a publically available collection of 1000 invasive breast cancer samples (and matched normal tissues), with multidimensional molecular characterisation (mRNA expression microarrays, DNA methylation microarrays, whole

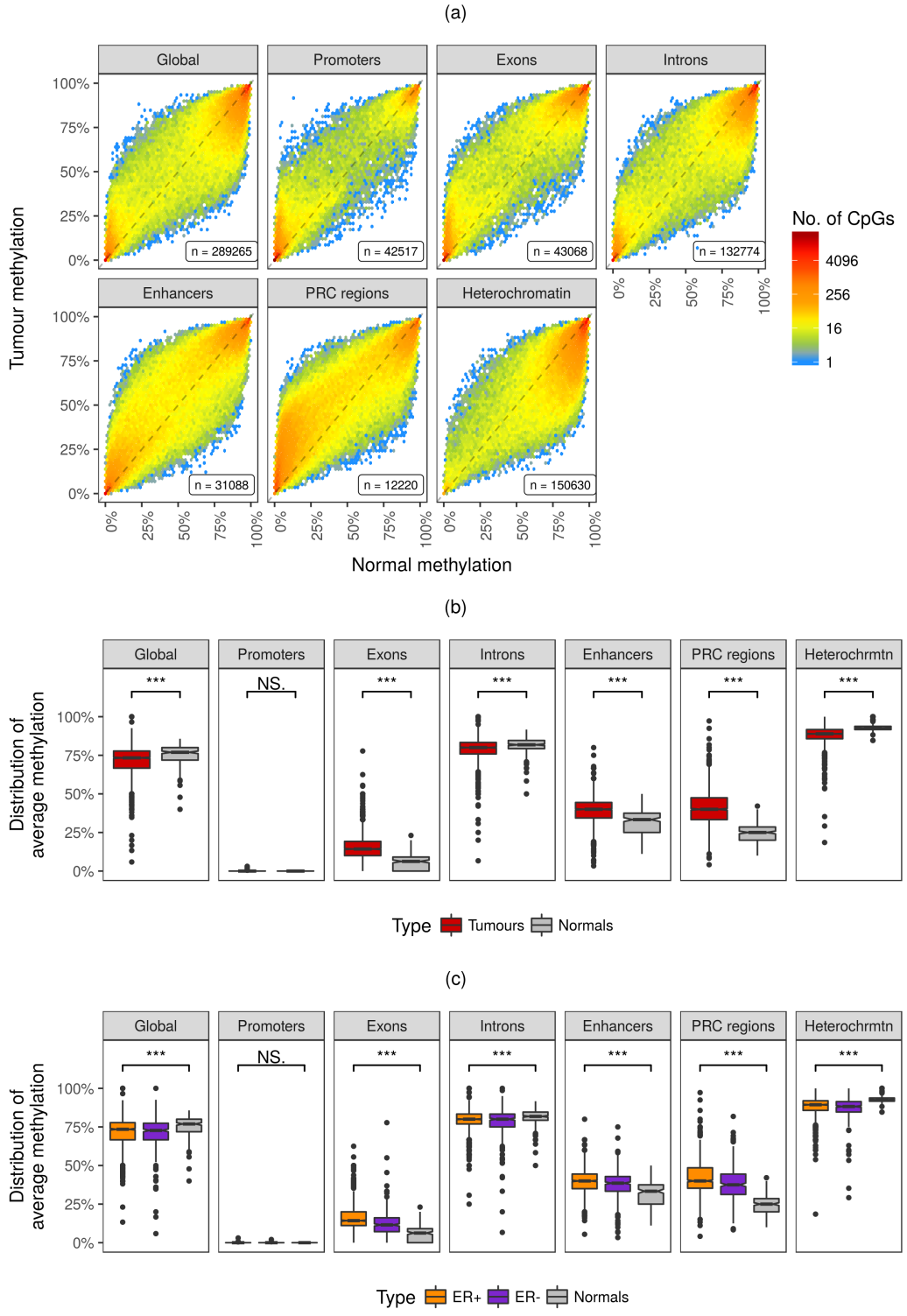


Figure 2.16: (Caption on next page.)

Figure 2.16: (Previous page.) **Supervised analysis of DNA methylation profiles reveal distinct epigenetic landscapes in tumours and normal tissues. (a)** Scatter plots of the average DNA methylation estimates for tumour and normal tissues for CpG sites within the *5.50 CpG universe* stratified by genomic feature. Colours represent density of CpG sites. Number of CpG sites within each genomic feature is noted in the bottom right corner of the plots. **(b)** Distribution of average methylation estimates for tumour and normal tissues stratified by genomic feature. For each sample, the median methylation level across each genomic feature, and the resulting distributions were plotted. For each genomic feature, methylation estimates between tumour and normal tissues were compared using the Wilcoxon rank-sum test. *FDR p-values* were denoted. **(c)** Distribution of average methylation estimates for ER+, ER- and normal tissues stratified by genomic feature. For each sample, the median methylation level across each genomic feature, and the resulting distributions were plotted. For each genomic feature, methylation estimates between these three categories were compared using the Kruskal Wallis test. *FDR p-values* were denoted. (N.S.= *FDR p-value* < 0.1, $*$ = *FDR p-value* < 0.05, ** = *FDR p-value* < 0.01, *** = *FDR p-value* < 0.001, **** = *FDR p-value* < 0.0001). Heterochrmtn = Heterochromatin.

exome sequencing, SNP arrays, microRNA sequencing and reverse-phase protein arrays (RPPA)), as well as clinical annotation.

In the original TCGA study, Infinium Human Methylation 27 (HM27) and Infinium Human Methylation 450 (HM450) arrays were used to profile methylation in 800 samples. Unsupervised analysis using a Recursively Partitioned Mixture Model (RPMM) was performed on 466 breast tumours using a set of 574 selected probes that were present in the HM27 microarray, as detailed in the original publication. This identified five clusters of tumours with distinct DNA methylation profiles [Cancer Genome Atlas Network, 2012].

In order to conduct a comparison of the METABRIC methylome with the TCGA methylome, a similar unsupervised analysis was conducted in the METABRIC samples. Since a vast majority of the probes in the HM27 microarray (used in the TCGA clustering analysis) localise to promoters (88%), only promoter SCCRUB regions were considered for this analysis. On aggregate, 42517 SCCRUB promoter regions were identified (out of which 4589 (11%) also overlapped with the HM27 microarray probes). For each region, z-scores were calculated for the 1482 tumours with respect to the methylation estimates of the 237 normal tissues. The top 50% variable regions were selected and a Partitioning Around Medoids (also known as K-medoids) clustering, which is a more robust version of K-means clustering [Kaufman and Rousseeuw, 1987] was conducted to cluster the tumours into 5 groups (same depth of clustering as conducted in the TCGA dataset). Table 2.1 (top panel) details the 5 methylation-groups identified in METABRIC, and they show remarkable concordance with the

5 methylation-clusters identified in the TCGA tumours. For instance, Groups 1 and 2 derived from both TCGA and METABRIC clustering efforts were revealed to be significantly enriched for the Luminal A subtype. Methylation group 3 derived from both TCGA and METABRIC clustering analyses demonstrated the strongest hypermethylation phenotype among all groups; they were depleted for *PIK3CA*, *MAP3K1* and *MAP2K4* mutations; and they largely overlapped with the Luminal B subtype. Group 5 showed the lowest aggregate methylation levels, highest frequency of *TP53* mutations and largely consisted of Basal-like tumours in both METABRIC and TCGA analyses. Moreover, the methylation subtypes derived from both METABRIC and TCGA analyses did not exhibit a significant association with the HER2-enriched breast cancer subtype. This remarkable agreement between the two sets of methylation-clusters that were independently derived from two completely distinct breast cancer datasets (TCGA and METABRIC tumours), and using distinct methylation platforms, strongly demonstrate the validity of the RRBS methylome data in METABRIC.

A 5-member unsupervised clustering solution derived from the METABRIC dataset was described above to conduct a like-for-like comparison with the TCGA cluster analysis (in which 5 methylation groups had been identified). In order to select the number of clusters that provides the most consistent clustering solution, the Dunn index was calculated for different number of clusters. The Dunn index aims to identify sets of clusters in which members of the same cluster have a small variance i.e. they are compact, while the means of distinct clusters are far apart i.e. they are separated. The 8-cluster solution had the highest Dunn index (compared with 3-9 member cluster solutions). Table 2.1 (bottom panel) details these 8 methylation-clusters identified in METABRIC. Remarkably many Intrinsic subtypes were split into 2 or more groups, and strong associations with the Integrative clusters were observed. For instance, the Luminal B subtype has been divorced into 2 methylation groups: i) Methylation group 3, which is enriched for IntClust 1, and has a dramatically hypermethylated phenotype, and ii) Methylation group 7, which is enriched for IntClust 7, and has a lesser hypermethylated phenotype. Similarly, the Basal-like subtype has also been split into three methylation groups: i) and ii) Methylation groups 4 and 8 which were both enriched for IntClust 10, had the lowest aggregate methylation levels, and highest frequencies of *TP53* mutations; and iii) Methylation group 5 which was enriched for IntClust 4ER- and the Claudin-low subtype, had moderate methylation levels and moderate prevalence of *TP53* mutations. Moreover, this clustering also identified a HER2-enriched subtype (Methylation group 6). This demonstrates how increased sample size provided by the METABRIC cohort enabled the refinement of the epigenetic taxonomy of breast cancer into 8 clusters, compared to the 5-cluster solution

obtained in the TCGA cohort [Cancer Genome Atlas Network, 2012]. In addition, these preliminary results substantiate the biological significance of the Integrative clusters and indicate that besides carrying distinct genetic rearrangements, these Integrative clusters might carry distinct epigenetic alterations as well.

No.	Average	Breast subtype enrichment			Avg. Mutation prevalence (%)				TCGA concordance
	methylation (%)	ER status	Intrinsic subtype	Integrative cluster	PIK3CA	MAP3K1	MAP2K4	TP53	
461	73.80	ER+	Luminal A	3	51.4	13.2	3.8	34.4	TCGA Methylation cluster 1
384	44.12	ER+	Luminal A	8	50.8	15.6	5.9	9.2	TCGA Methylation cluster 2
258	83.29	ER+	Luminal B	7	30.0	8.6	2.1	20.2	TCGA Methylation cluster 3
239	50.60	ER-	Claudin-low	10	32.7	5.7	2.4	61.1	None
140	0.30	ER-	Basal-like	10	9.6	2.4	0.0	84.0	TCGA Methylation cluster 5

No.	Average	Breast subtype enrichment			Avg. Mutation prevalence (%)			
	methylation (%)	ER status	Intrinsic subtype	Integrative cluster	PIK3CA	MAP3K1	MAP2K4	TP53
374	43.08	ER+	Luminal A	3	58.9	18.9	6.3	12.0
290	74.36	ER+	Luminal A	8	45.1	13.1	2.6	19.0
169	74.91	ER+	Luminal B	7	31.9	10.6	1.9	16.2
91	1.02	ER-	Basal-like	10	7.0	2.3	1.2	82.6
166	52.61	ER-	Claudin-low	4ER+ 4ER-	33.6	6.8	2.1	51.4
131	51.11	ER-	Her2-enriched	5	38.5	0.0	3.4	86.3
196	96.41	ER+	Luminal B	1	40.2	8.6	4.0	29.3
65	0.00	ER-	Basal-like	10	5.6	3.7	0.0	79.6

Table 2.1: Unsupervised clustering of DNA methylation profiles across the 1482 METABRIC breast tumours reveal subgroups with distinct clinical and molecular features. (top panel) 5 methylation-cluster solution which is extremely similar to that obtained in the TCGA breast cancer dataset [Cancer Genome Atlas Network, 2012]. Summaries of the clinical and molecular features of these clusters as well as most enriched established breast cancer subtype are detailed in the table. If an equivalent methylation cluster had been identified in TCGA, it is denoted in the TCGA concordance column. **(bottom panel)** 8 methylation-cluster solution based on the empirically derived Dunn index. Summaries of the clinical and molecular features of these clusters as well as the most enriched established breast cancer subtype are detailed in the table. For both panels, only promoter SCCRUB regions were considered to allow comparison with the TCGA breast cancer methylome clustering effort. Avg. methylation (%) represents the average methylation of the tumours in each subtype over the promoter SCCRUB regions considered.

2.7 Discussion

The investigation of the DNA methylation landscape in the METABRIC dataset comprising of approximately 1500 breast tumours constitutes the largest single cancer methylome cohort yet, and represents a significant development in the field of breast cancer. In addition to methylation profiling, the availability of gene expression, copy number and somatic mutation data allows a comprehensive investigation of the molecular multi-omic basis of breast cancer. Moreover, patients from the METABRIC cohort have rich clinical annotation and long follow-up times compared to other recent breast cancer studies, thus enabling the identification of prognostic biomarkers.

RRBS was selected as the DNA methylation platform of choice since it allows quantification of DNA methylation at single nucleotide resolution and enables the exploration of intratumour methylation heterogeneity and epiclonal dynamics. Other advantages include the lower input DNA requirement and cost effectiveness compared to WGBS. However, the RRBS technique described and implemented in this thesis requires PCR amplification, which can introduce biases in the uniformity of coverage as well as the confidence in the quantification of DNA methylation over the epigenome [McInroy et al., 2016] (see Chapter 1). Moreover, the inability to discriminate between PCR-induced duplication artefacts or distinct molecular copies of fragments in RRBS can distort DNA methylation estimates as well. Recently, advanced protocols have been established that can circumvent these limitations such as post-bisulphite adaptor tagging (PBAT) [Miura et al., 2012], quantitative RRBS (Q-RRBS) [Wang et al., 2015], and recovery after bisulphite treatment (ReBuilT) [McInroy et al., 2016]. One of the follow-up goals of this project is to perform a sensitivity analysis by implementing the PBAT technique on a subset of the breast tumours described in this thesis.

Despite numerous advancements in bisulphite sequencing chemistries and the bioinformatics pipeline over the last few years, considerable steps were implemented to maximise yield and ensure the accurate calling of methylation estimates in this massively parallel sequencing effort. In addition to improvements to the RRBS protocol conducted in the laboratory previously [Vidakovic, 2014], switching to Illumina v4 chemistry and increasing the length of sequencing reads improved the number of unique CpG sites detected from RRBS libraries. Although, data alignment and methylation calling procedures to analyse RRBS methylomes are well established [Babraham Bioinformatics, 2016b], the bioinformatics pipelines were also modified. For instance, methylation information was merged from the + and – strands since DNA methylation is symmetric in mammals. This not only provides a tremendous

boost in coverage, but consequently also results in a more accurate estimate of the DNA methylation status at CpG dinucleotides. Furthermore, following quality assessment procedures and CpG filtering, a universe of 2.72 million CpG sites with a minimum 5× coverage that are profiled in at least 50% of METABRIC samples was established (*5.50 CpG Universe*).

Statistical analysis of bisulphite sequencing data is challenging, predominantly due to the high false positives associated with analysing a large number of CpG sites profiled, and the difficulty of interpreting single CpG data when cancer specific methylation alterations usually involve multiple CpG sites [Bock, 2012]. In order to counter these limitations, a novel algorithm called **S**patially **C**oordinated **C**pG-sites within the **R**RBS **U**niverse in **B**reast cancer (SCCRUB) is developed and implemented that attempts to transform the RRBS data into distinct regions comprising of multiple CpG sites coordinated in their methylation behaviours. This led to the identification of a universe of regions that are functionally relevant, and also provides a significantly lower multiple testing problem than single CpG analysis due to the smaller size of the universe. In this thesis, the SCCRUB universe consisting of $\sim$ 289,000 regions was used exclusively for those analyses in which the predominant aim is to detect functional methylation alterations that are likely to involve multiple CpG sites that are altered in a coordinated manner. However, the SCCRUB universe cannot employed for analyses in which stochastic methylation changes at the single CpG site level are quantified. Two key investigations of this nature presented in this thesis include i) Section 3.1 (Chapter 3), in which single CpG specific information was utilised to describe the genomic context of epigenetic drift and estimate tumour-specific background methylation rates; and ii) Chapter 5 in which RRBS methylation information was reanalysed at the read level to investigate intratumour methylation heterogeneity.

An unsupervised clustering analysis over the 1482 breast tumours identified subgroups of tumours with distinct methylation profiles. The 5-group clustering solution demonstrated remarkable concordance with a clustering solution obtained in an external dataset, the TCGA breast cancer study [Cancer Genome Atlas Network, 2012]. The agreement between these Furthermore, most bioinformatics packages that are available for analysing bisulphite sequencing data are not optimised for large sample sizes. Consequently, a vast majority of the statistical methods and analyses presented in this thesis have been constructed using bespoke statistical algorithms, unless mentioned otherwise.

Next, validation of the RRBS pipeline was performed. An unsupervised t-SNE analysis established that the variation in DNA methylation profiles was predominantly

explained by the tumour/ normal classification, established breast cancer subtype taxonomies and clinicopathological variables, but notably not technical factors. The high reproducibility of DNA methylation calls between the RRBS platform and the universally used HM450 platform further established the robustness of the RRBS pipeline detailed in this chapter. two sets of methylation-clusters that were independently derived from two completely distinct breast cancer datasets (TCGA and METABRIC tumours), largely unique promoter regions (only 11% overlap between the SCCRUB promoter regions and HM27 probes), and two different methylation platforms (RRBS and HM27 microarrays) strongly substantiates the validity of the RRBS methylome data in METABRIC. Moreover, $\sim$ 240,000 SCCRUB regions lie outside promoters in this RRBS-derived methylome, and can be used to redefine and/ or refine the DNA methylation based epigenetic subtypes in breast cancer.

Finally, a supervised analysis demonstrated that tumours and normal tissues have distinct methylation profiles across various genomic features; and also revealed that the extent of methylation changes in tumours is subtype specific. A formal investigation delineating the precise locations of epigenetic deregulation (within the SCCRUB universe) between tumour and normal tissues as well as between breast tumour subtypes, and examination of their functional roles is conducted in Chapter 3.

Chapter 3

Identification of DNA methylation alterations in breast cancer

Contents

3.1 Introduction

The establishment and maintenance of precise epigenetic programs such as DNA methylation are essential for embryonic development and differentiation, and different tissues and cell types adopt characteristic methylation patterns that serve to determine and preserve specific biological processes (Chapter 1) [Holliday, R. & Pugh, 1975; Jones, 2012; Okano et al., 1999; Riggs, 1975]. While DNA methylation is a well-balanced process that regulates gene expression in mammalian cells, these normal DNA methylation patterns are largely disrupted during the initiation and progression of cancer. The first link between DNA methylation and cancer was reported by Feinberg and Vogelstein [1983a] who observed that genomes of cancer cells had reduced levels of DNA methylation (hypomethylation) compared to normal tissues. Repetitive elements were found to be enriched in hypomethylation and this was shown to result in genomic instability, which is one of the hallmarks of cancer [Ehrlich, 2002, 2009]. On the other hand, DNA hypermethylation of gene promoter has also been studied extensively and shown to favour tumorigenesis by repressing transcription in key tumour suppressor genes in many cancers [Esteller, 2000; Herman and Baylin, 2003; Kawaoi et al., 1992] which represents another hallmark of cancer. Consequently, there has been a drive to map these cancer-associated epigenetic changes with putative functional roles in gene transcriptional regulation. Recently, major international projects have identified key locations of divergent epigenetic changes between cancers and their normal counterparts [Cancer Genome Atlas Network, 2012; Hovestadt et al., 2014; Weisenberger, 2014], as well as between tumour subtypes [Figueroa et al., 2010; Kretzmer et al., 2015]. Moreover, these regions of differential methylation can serve as prognostic and predictive biomarkers [Gyparaki et al., 2013; Nikolaidis et al., 2012].

However, despite extensive profiling of cancer methylomes, there is insufficient evidence of the dynamics of the process driving methylation change in cancer tissues. For instance, if the promoter of a tumour suppressor gene is methylated, the key question is – how did this region gain methylation in tumour cells versus the normal tissue? Two distinct models have been put forward to explain this phenomenon: the *stochastic* model and the *instructive* model [Struhl, 2014; Tanay, 2017, private communication]. The fundamental distinction between these two models is whether this altered methylation is specifically targeted to the tumour suppressor gene (*instructive* model) or instead non-specifically over the whole genome (*stochastic* model). In the stochastic model, random changes in methylation occur independently and in parallel at different regions in different cells of the tumour as a consequence of DNA methylation replication errors (largely associated with cell division, see

Epigenetic drift, introduced in Chapter 1). One of these stochastic errors may result in the methylation increase of a specific tumour suppressor gene, and consequently give a cell a growth advantage (for instance, by silencing the tumour suppressor gene). For example, age-related DNA methylation linked with the epigenetic silencing of antagonists of the WNT-signalling pathway has been observed which was potentially implicated with stem-cell and progenitor expansion and survival [Baylin and Ohm, 2006; West et al., 2013; Zheng et al., 2016]. Conversely, in the instructive model, a genomic event such as a mutation in an oncogene triggers a transcriptional regulatory pathway that results in the DNA methylation of a tumour suppressor gene. For example, recent work in human colorectal cancer cell lines revealed that a *KRAS* mutation led to a surge in the concentration of the transcription factor, ZNF304, in the nucleus [Serra et al., 2014]. This transcription factor in turn recruits a co-repressor complex including DNA methylase leading to the concomitant DNA methylation and repression of specific loci including the *INK4-ARF* tumour suppressor. Although a few reports including those mentioned above have been able to identify the relevant mechanism culminating in tumour-specific methylation alterations in a few genes or pathways, this effort has been largely ignored in most high throughput cancer methylome studies.

The aim of genome-wide cancer methylome studies is to identify malignancy-associated epigenetic changes with putative functional roles in gene transcriptional regulation. Although instructive methylation changes are by definition linked with instructive regulatory pathways upstream of the methylation event, on the other hand stochastic methylation changes generally affect gene function, *only if* they hit a key gene maintaining healthy homeostasis (such as specific transcription factors or tumour suppressor genes). However, as discussed above, stochastic DNA methylation errors in tumours (collectively termed as epigenetic drift) are widespread all over the genome and do not target functional genes only; and thus, might be less likely to have relevance to tumorigenesis than instructive methylation differences. This imposes the question – can genome-wide cancer methylome studies accurately discriminate between methylation alterations with potential functional significance versus stochastic methylation changes that are not selected for? Unfortunately, this line of investigation has rarely been considered in cancer methylome studies and the relevant mechanism underlying observed methylation changes are usually not explicitly delineated. Indeed, with increasing sample sizes, larger genomic coverage and higher resolution profiling of cancer methylomes, it is extremely likely that a large contribution of the detected cancer-normal differentially methylated regions is a consequence of cell-division DNA replication errors. This is largely because the analytical approaches widely

available today employ a uniform methylation difference threshold (such as 20%) to identify differentially methylated regions irrespective of genomic context or cancer type [Akalin et al., 2012b; Hebestreit et al., 2013; Robinson et al., 2014]. However, this assumption is flawed. Firstly, it has been well established that methylation gain and loss rates are exceptionally dependent on CpG density [Horvath et al., 2012; Rakyan et al., 2010]. Regions that gain DNA methylation over time are enriched for CpG content, while low CpG regions tend to lose DNA methylation as a consequence of epigenetic drift. Subsequently key experiments investigating clonal dynamics of DNA methylation in somatic fibroblast populations revealed that higher methylation error rates are associated with late time of replication in the cell cycle and nuclear lamina interaction [Shipony et al., 2014]. In addition, stochastic epigenetic events result in increasing discordance in a tumour's epigenome and consequently, highly variable methylome profiles have been observed between different individuals and tissue types [Jones et al., 2015; Talens et al., 2012]. Based on this collective evidence, it is clear that the assumption of a uniform background methylation difference for all tumours and over the whole genome is erroneous, and can lead to widespread false positive findings that might overshadow those epigenetic events involved in transcriptional pathways.

One of the key objectives of this chapter is to quantify the extent of DNA methylation alterations with putative *functional* roles in tumorigenesis and/ or explaining inter-tumour heterogeneity in breast cancer. I aim to do this by first, linking detected methylation alterations with their presumed underlying mechanism, and secondly, identifying any relevant genes or pathways exhibiting concomitant modifications (as measured via gene expression). However, in order to shed light on the mechanisms underlying tumour-specific methylation alterations, it is necessary to first quantify the variation in Epigenetic drift in tumours (background DNA methylation alterations compared to normal tissues). The genomic context of these background methylation changes is investigated, as well as the inter-patient heterogeneity in what represents the first genome-wide characterisation of epigenetic drift in a large cohort of primary breast tumours. Finally, two epigenetic drift-related indices are introduced that allow evaluation of the extent and direction of epigenetic drift in the METABRIC dataset including their respective prognostic potentials.

The next section of this chapter involves the construction of a novel algorithm called DMARC (Directed Methylation Altered regions in Cancer) to detect *directed* and *background* DNA methylation alterations in tumours. Directed methylation comprise of all instructive methylation alterations as well as stochastic alterations that

are selected for and consequently observed more frequently in the tumour cell population. In contrast, background methylation alterations essentially comprise of all stochastic alterations that are not under selection. The algorithm incorporates the background methylation heterogeneity (estimated above) into the traditional differential methylation analysis resulting in the added functionality of being able to discern whether the identified methylation alterations are background or directed. The underlying premise of this discrimination is based on the assumption that directed methylation alterations will exhibit a larger methylation difference compared to that expected based on the estimated background differences. If developed effectively, DMARC would enable i) illumination of the mechanisms underlying these epigenetic events; and ii) enrichment of methylation events with potential functional roles.

The development of the novel algorithm DMARC is followed by its implementation in the 1482 breast cancer and 237 normal tissues to identify directed and background differentially methylated regions (DMRs) in the breast tumours compared to normal tissues. An analogous strategy is also used to detect methylation alterations at an individual tumour resolution to assemble a catalogue of tumour-specific directed and background methylation alterations. Finally, both directed and background methylation alterations were functionally characterised by linking them with concomitant modifications in gene expression and identifying the relevant biological pathways that are affected using gene set enrichment analysis. This line of analysis is conducted to achieve two goals. Firstly, it would determine whether the detected directed methylation alterations in breast cancer are more likely to have functional consequences than background methylation alterations as hypothesised. Secondly, it would validate the prominence of previously reported epigenetically altered genes in breast cancer, as well as revealing novel genes that have not yet been implicated.

Given the well-established distinct methylation landscapes between ER+ and ER- breast tumours (see Chapter 1), separate analyses on these two subtypes are also conducted. In addition, tumour vs. tumour methylation differences are identified to explicitly explore the distinct epigenetic profiles of ER+ and ER- tumours.

Finally, the functional differential methylation analysis described above was repeated in each of the previously defined breast cancer subtypes: 11 Integrative clusters (IntClusts) and 6 Intrinsic subtypes to examine the role of tumour-normal and tumour-tumour epigenetic differences in each subtype. Although a limited investigation has been conducted across the Intrinsic subtypes in previous reports [Bediaga et al., 2010; Holm et al., 2010], the large sample size of this study permits a

detailed genome-wide investigation that would enable the identification of subtype-specific epigenetic events involved in breast cancer transcriptional deregulation. This analysis would be crucial in assessing the contribution of DNA methylation alterations in explaining inter-tumour breast cancer heterogeneity.

3.1.1 Summary of aims

The previous chapter involved the generation of a large sequencing based breast cancer methylome comprising of 1482 breast tumours and 237 matched adjacent normal tissues. This chapter aims to identify DNA methylation alterations with putative *functional* roles in tumorigenesis and/or in explaining inter-tumour heterogeneity in breast cancer. This is achieved through the following steps:

1. Characterisation of epigenetic drift in breast tumours. Spatial-dependency and inter-tumour heterogeneity of background methylation alterations were investigated to elucidate the biological mechanisms underlying epigenetic drift.

2. Identification of class-specific and tumour-specific differential methylation alterations in breast cancer versus normal tissue.

3. Development and implementation of a novel algorithm that utilises background methylation heterogeneity to classify cancer methylation alterations as *background* or *directed*.

4. Assessment of the contribution of *directed* and *background* methylation alterations in regulating transcription.

5. Exploration of the heterogeneity in epigenetic programming in distinct breast cancer subtypes and its consequences on transcriptional networks and survival.

3.2 Epigenetic drift in breast cancer

As described in Chapter 1, epigenetic drift largely comprises of stochastic DNA methylation changes that lead to deregulation of normal methylation patterns. In order to elucidate the biological mechanisms underlying epigenetic drift, the spatial variation as well as the inter-patient heterogeneity in epigenetic drift was first quantified in the METABRIC breast cancer methylome cohort and then integrated with gene expression changes[1].

The extent of epigenetic drift in the breast tumours versus the normal tissues was examined. For the purpose of characterising epigenetic drift, only regions of the genome with methylation statuses unlikely to be under selection were considered [Yatabe et al., 2001]. Accordingly, regions of the genome were removed that have previously been implicated in neoplastic transformation through epigenetic processes. This strategy enriches for neutral methylation modifications by minimising confounding with alterations that are not stochastically or environmentally acquired. Specifically, a *background* set of CpG sites was established by eliminating CpG sites lying in promoter, enhancers or PRC region from the previously defined *5.50 CpG universe*. This newly constructed *Background CpG universe* comprised of 1.8 million CpG sites (*5.50 CpG universe* = 2.7 million CpG sites) that were covered in 50% of samples with at least 5 read coverage.

For each CpG site included in the *Background CpG universe*, average normal methylation levels were calculated by taking the mean across all 237 normal tissues. Subsequently, CpG site-specific methylation differences from the mean normal levels were calculated for each of the 1482 breast tumours separately. As discussed in Section 3.1, the extent and direction of methylation differences have been shown to be conditional on CpG density in cancer [Horvath et al., 2012], and on Time of Replication (TOR) and nuclear laminal interaction in fibroblast populations [Shipony et al., 2014]. Drift related methylation gains and losses have also been associated with PRC-marked promoter and enhancer regions [Day et al., 2013; Teschendorff et al., 2010]. However, as noted previously, these chromatin domains have been eliminated from the analysis to minimise bias attributed to functional methylation alterations that are likely to lie in these regions. For the sake of model parsimony, epigenetic drift related spatial dependency was assessed for only two genomic factors: CpG density and TOR. CpG density was estimated by quantifying the number of CpG dinucleotides

[1]The analysis described in Section 3.2 stemmed from discussions with Professor Amos Tanay and Aviezer Lifschitz during a fruitful travel fellowship to Professor Amos Tanay's laboratory at the Weizmann Institute of Science, Rehovot, Israel.

in the immediate 1Kbp surrounding each CpG. TOR status for the CpG sites was obtained from previously published Repli-seq experiments conducted on the Michigan Cancer Foundation-7 (MCF-7) breast cancer cell lines [Pope et al., 2014]. Although the TOR status observed in MCF-7 breast cancer cell lines may not be completely conserved across all breast cancer subtypes, it should give a reasonable indication of it.

The CpG sites were first stratified into 25 bins based on CpG density (1st 24 bins = 0 CpG/kbp to 120 CpG/kbp; bin size = 5 CpG/kbp; 25th bin = 120+ CpG/kbp) and then further stratified into 5 bins based on TOR (based on 20th percentiles) resulting in 125 bins with distinct CpG density and TOR profiles. The distribution of CpG sites across CpG density and TOR are illustrated in Figure 3.1a and 3.1b respectively. For each tumour, background methylation differences were aggregated across CpG sites lying within each bin. The mean background methylation difference between all tumours and all normals (across all CpG sites), and the number of CpG sites is calculated for each of the CpG density/ TOR bins and detailed in Appendix B.1.

3.2.1 Epigenetic drift is genomic-context dependent

In normal tissues, high background DNA methylation levels are observed at CpG-poor regions and low DNA methylation at CpG-rich regions (Figure 3.1c). It is likely that background methylation alterations in tumours will regress this strong bipolarity. As expected from previous reports [Jones and Baylin, 2002, 2007; Stirzaker et al., 2014; Teschendorff et al., 2013], breast tumours clearly gain methylation in CpG rich regions (methylation difference = 6.05%, *FDR p-value* < 1×10^{-300};Wilcoxon–Mann–Whitney test; CpGs with density > 100 CpGs/kbp) and lose methylation in CpG poor regions (methylation difference = -2.96%, *FDR p-value* < 1×10^{-300}; Wilcoxon–Mann–Whitney test; CpGs with density < 25 CpGs/kbp) when compared to normal tissues in the background epigenome (Figure 3.1c). Furthermore, tumours exhibit a marked increase in the variance of methylation statuses in CpG sites that replicate late/ very late in the cell cycle compared to CpG sites that replicate early/ very early (Ratio of variances in tumours = 3.86, *FDR p-value* = 1×10^{-300}; F-test), whereas this was not observed in normal tissues (Ratio of variances in normals = 1.07, *FDR p-value* = 0.4685; F-test) (Figure 3.1d).

Figure 3.2 (top panel) illustrates the tumour-normal methylation background differences at a tumour-specific resolution, and the association of epigenetic drift with CpG density and TOR (described above) is undeniable. Although the relationship of methylation gains with CpG-rich regions and losses with CpG-poor regions is already

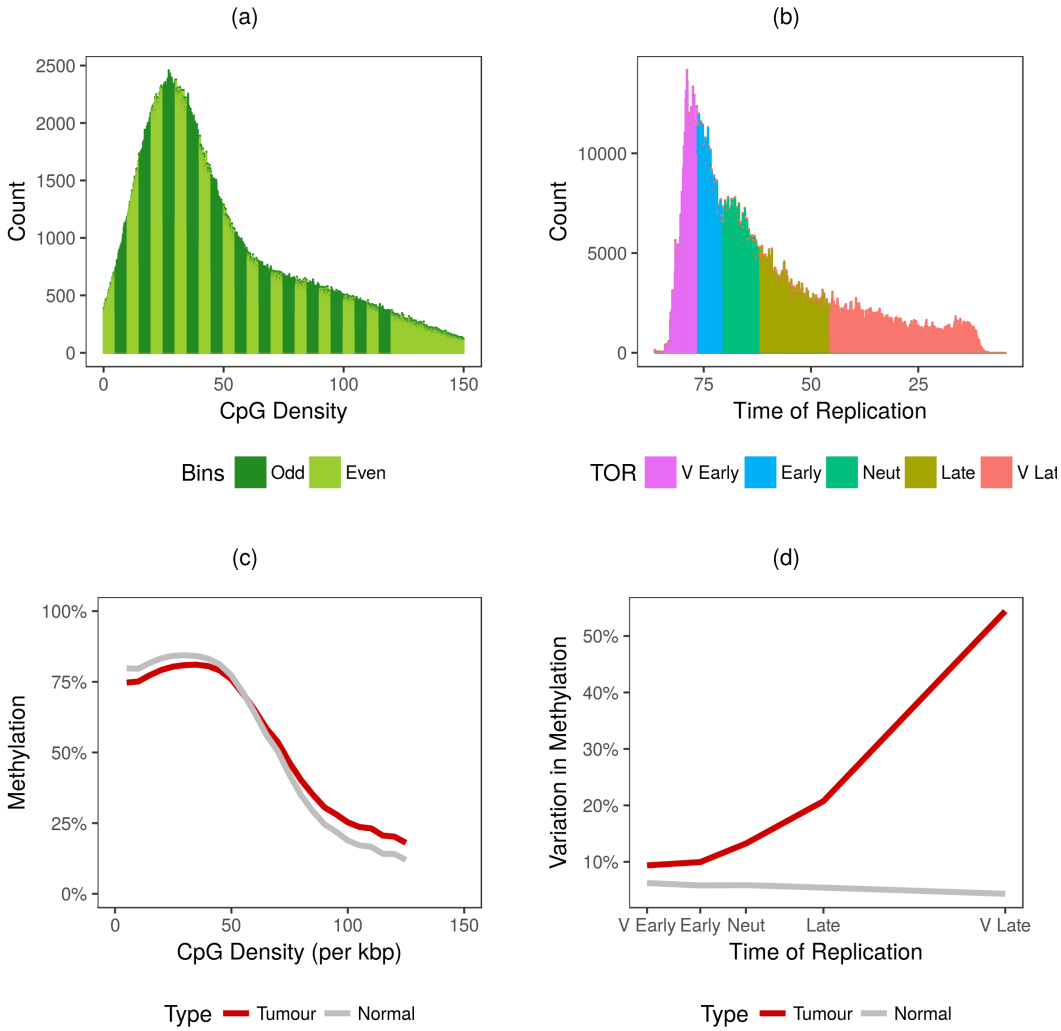


Figure 3.1: Background DNA methylation levels are dependent on CpG density and time of replication. Distribution of CpG sites in the background epigenome based on **(a)** 25 CpG density bins; and **(b)** 5 Time of Replication categories. **(c)** Mean methylation of tumour and normal tissues stratified by CpG density bins. Firstly, mean methylation levels of all background CpG sites per density bin were calculated for each sample. Next, for each CpG density bin, these sample-specific values were averaged over all normal tissues and tumour tissues respectively. **(d)** Variance in methylation of tumour and normal tissues stratified by time of replication. Firstly, mean methylation levels of all background CpG sites per TOR category were calculated for each sample. Next, for each TOR class, the variance of the sample-specific values was calculated over all normal tissues and tumour tissues respectively. V Early = Very Early. Neut= Neutral. V Late = Very Late.

well established in cancer, these findings demonstrate the remarkable predisposition of late TOR regions for accumulating both methylation changes (gains and losses) in breast cancer. A global loss of methylation has previously been hypothesised to be a consequence of replication-associated hypomethylation in late replicating regions in breast cancer cell lines [Hon et al., 2012] and prostate cancer [Berman et al., 2011]. However, the findings in this section strongly suggest that the direction of methylation change in late replicating regions is noticeably dependent on the underlying CpG density; nevertheless, an aggregate loss in these regions is observed since late replicating regions are generally CpG poor (as is the whole epigenome, Figure 3.2 – bottom panel). Late-replicating compartments of the genome have been functionally associated with heterochromatin and laminal associated domains [Hansen et al., 2010], which have been shown to be evolutionarily conserved [Pickersgill et al., 2006]. This suggests that mechanisms facilitating methylation errors in these regions involve the higher-order chromatin architecture within the cell nucleus as well as DNA replication timing.

3.2.2 Epigenetic drift is highly heterogeneous across breast tumours

Figure 3.2 also demonstrates that breast tumours exhibit extraordinary variations in their capacities to gain or lose methylation. Based on these estimated background methylation differences, two tumour-specific indices were developed to describe the extent and direction of epigenetic drift.

1. *Accumulation index*: Accumulation of epigenetic drift, represents the absolute extent of methylation alterations in the tumour compared to the normal. The accumulation index for each tumour was calculated by averaging absolute background methylation differences (irrespective of gain or loss) over all CpG sites. Higher values signify a larger extent of epigenetic divergence from the normal tissue with 0 being the lowest theoretical score.

2. *Direction index*: Direction of epigenetic drift, represents the tendency of a tumour to gain or lose methylation over the background epigenome. The direction index for each tumour calculated by averaging methylation differences across all CpG sites. Tumours tend to lose methylation globally due to higher proportion of CpG poor sites than CpG rich sites, and so the direction index would tend to be negative for most tumours. Accordingly, the index was normalised according to CpG density such that a value of 0 represents a tumour

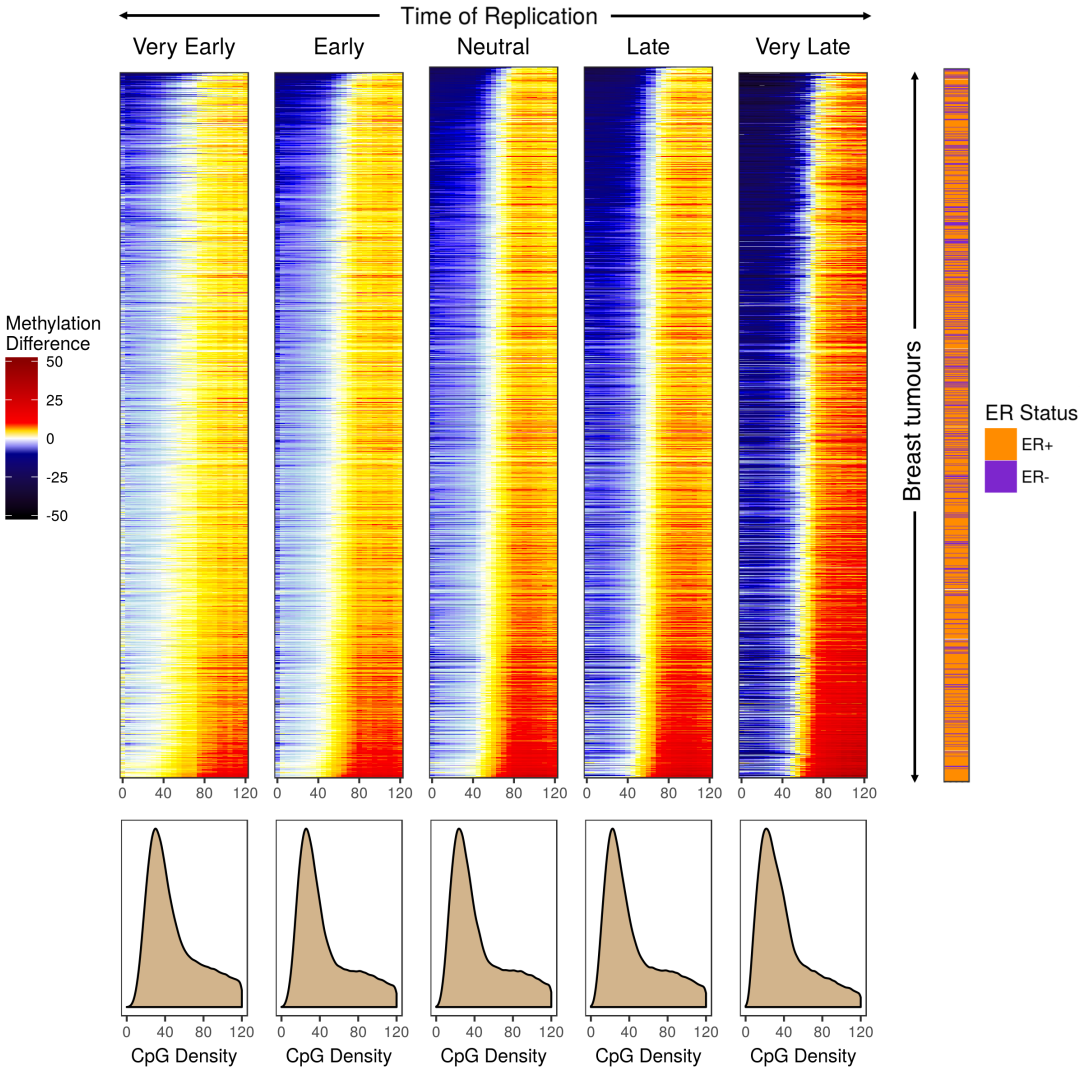


Figure 3.2: Epigenetic drift is genomic context-dependent and tumour-specific. (top panel) The five panels represent the TOR classes from Very Early to Very Late. In each panel, the x-axis represents CpG density bins. For each of the 125 CpG density/ TOR categories, mean tumour-normal methylation differences were calculated by averaging over background CpG sites and plotted. Colours represent the extent of tumour-normal methylation difference (gain denoted by red, loss denoted by blue; legend on far left). The y-axis represents the tumours. The tumours were ordered by average methylation difference aggregated over all 125 CpG density/ TOR bins, and not by breast cancer subtype. The legend on the far right indicates the ER status of the tumours. **(bottom panel)** Distribution of CpG sites in the background epigenome based on CpG density bins; stratified by the TOR classes. Figure conceptualised by Professor Amos Tanay, Weizmann Institute of Science, Rehovot, Israel.

that doesn't exhibit an aggregate increase or decrease in methylation (conditional on its underlying genomic context). Positive values represent an aggregate gain in methylation over the background genome while negative values represent an aggregate loss in methylation.

Examination of the accumulation and the direction indices of the 1482 breast tumours stratified by ER status showed that ER+ tumours on average tend to accumulate more methylation alterations than ER- tumours (mean accumulation: ER+ = 3.23, ER- = 2.64, *p-value* = 8.0×10^{-16}; t-test), implying that the former could have more divergent epigenomes (Figure 3.3a). However, since higher accumulation of epigenetic drift was associated with higher age (*p-value* = 3.1×10^{-16}, linear regression adjusted for tumour cellularity), this could also reflect the average younger ages of ER- patients (mean age: ER+ = 63 years, ER- = 54 years, *p-value* = 3.10×10^{-28}; t-test). Interestingly, ER+ tumours are also quite balanced in terms of the direction of these alterations, while the small fraction of ER- tumours that exhibit high accumulation also revealed a tendency for hypomethylation (mean Direction: ER+ = -0.17, ER- = -0.534, *p-value* = 1.4×10^{-6}; t-test).

Two epigenetic drift indices were formally examined across the previously established breast cancer molecular subtypes: 6 Intrinsic subtypes (Figure 3.3b) and 11 Integrative clusters (Figure 3.3c). The accumulation index was significantly associated with the Integrative cluster definition (*p-value* = 5.7×10^{-64}; Analysis of Variance) and the Intrinsic subtype definition (p = 7.7×10^{-71}; Analysis of Variance); while the direction index was less so (Direction vs. Integrative clusters: *p-value* = 1.8×10^{-7}; Direction vs. Intrinsic subtype: *p-value* = 3.2×10^{-6}; Analysis of Variance). The strong relationship of the two purely epigenetic based scores (accumulation and direction) with genetic and transcriptomic-defined subtypes of breast cancer confirms its biological significance as well as the relationship between genetic and epigenetic dysregulation in tumorigenesis. Moreover, further analysis also revealed tremendous variability in the epigenetic drift *within* ER+ and ER- tumours as well. For instance, tumours in IntClust 1 accumulates significantly more methylation alterations than IntClust 3 (mean accumulation: IntClust 1 = 4.35, IntClust 3 = 2.57, *p-value* = 1.2×10^{-20}; t-test) despite both clusters being predominantly composed of ER+ tumours. Similarly, two ER- enriched clusters, IntClust 4ER- and 10, exhibited distinct levels of epigenetic divergence (mean accumulation: IntClust 4ER- = 2.17, IntClust 10 = 2.63, *p-value* = 0.001; t-test).

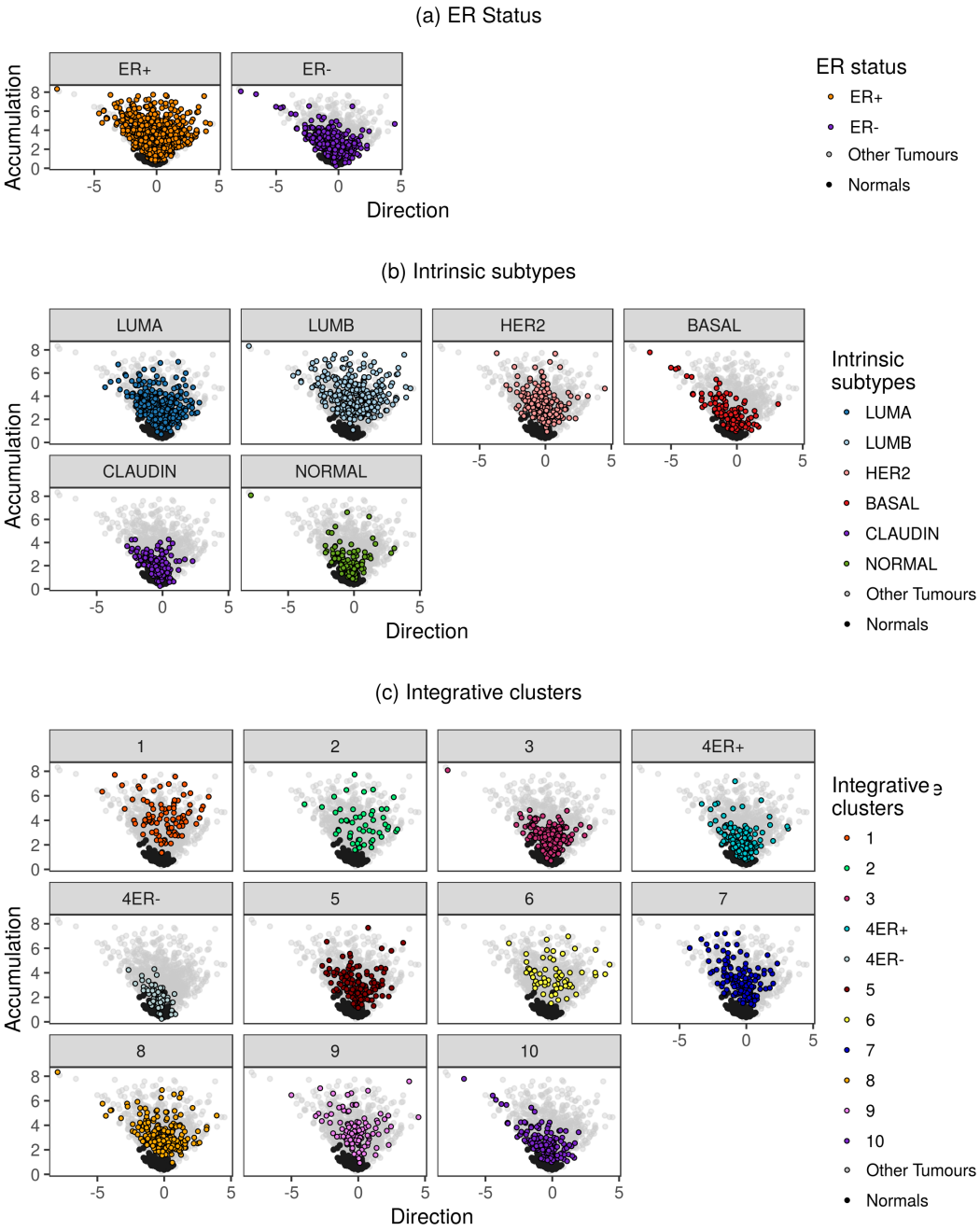


Figure 3.3: Epigenetic drift is highly heterogeneous across breast tumours. The *Accumulation* and *Direction* indices are plotted for each tumour (coloured points) stratified by breast cancer subtypes **(a)** ER Status; **(b)** Intrinsic subtypes; **(c)** Integrative clusters. Light grey points represent other tumours. Black points represent normal tissues. LUMA = Luminal A. LUMB = Luminal B. HER2 = HER2-enriched. BASAL = Basal-like. CLAUDIN = Claudin-low. NORMAL = Normal-like.

3.2.3 Accumulation of epigenetic drift is largely a consequence of mitotic errors

Epigenetic drift has largely been considered to comprise of alterations acquired as a consequence of stochastic DNA methylation maintenance errors [Kim et al., 2005; Teschendorff et al., 2013]. This hypothesis was explored in the METABRIC breast cancer methylome. Firstly, for each gene, a linear regression model was used to evaluate the association between a tumour's expression for the corresponding gene (dependent variable) and its accumulation index (independent variable). Tumour cellularity (as measured by ASCAT (Chapter 2) was significantly associated with the Accumulation index ($p\text{-}value = 7.6\times10^{-23}$), and therefore it was added as a confounder in the model. Next, the genes were ranked based on the relationship (β coefficients) between their expression and the accumulation index, and a Gene Set Enrichment Analysis (GSEA) was used to identify relevant mechanisms associated with a higher accumulation index using the *fgsea* package [Sergushichev, 2016]. The top 2 pathways associated with a tumour's capacity to accumulate methylation alterations were cell cycle (enrichment score = 2.18, *FDR p-value* = 0.0214; GSEA) and mitotic cell cycle (enrichment score = 2.06, *FDR p-value* = 0.0214; GSEA). Evidence for this relationship was reinforced with a strong association of a tumour's accumulation index with its mitotic index, a pathology-based score exclusively in ER+ tumours (ER+: $p\text{-}value = 5.1\times10^{-8}$, ER-: $p\text{-}value = 0.5820$; linear regression; Figure 3.4a). A sharp correlation was also revealed with the expression of the proliferation-related genes, Aurora Kinase A, *AURKA* (ER+: $FDR\ p\text{-}value = 2.0\times10^{-46}$, ER-: $FDR\ p\text{-}value = 0.0213$; linear regression; Figure 3.4b), and *MKI67* (ER+: $FDR\ p\text{-}value = 3.4\times10^{-28}$; ER-: $FDR\ p\text{-}value = 0.8591$; linear regression), in ER+ tumours. Although a higher accumulation index was also associated with higher age of the patients (Section 3.2.2), the relationship with mitotic index and proliferation was much sharper. These results collectively confirm the acquisition of background DNA methylation alterations in the epigenome is largely a consequence of the accumulation of passive replication related errors related to the number of cell divisions (mitotic clock), as initially hypothesised by Yatabe et al. [2001].

However, the relationship between epigenetic drift and mitotic clock (as measured by mitotic index and gene expression of *AURKA*) was much stronger in ER+ tumours than ER- tumours. In fact, in ER+ tumours, the accumulation index was significantly associated with Breast Cancer-Specific Survival (BCSS) indicating that ER+ tumours that accumulate more methylation alterations have lower survival times (univariable $p\text{-}value = 2.6\times10^{-5}$, multivariable $p\text{-}value = 0.0190$; Cox-proportional hazards model:

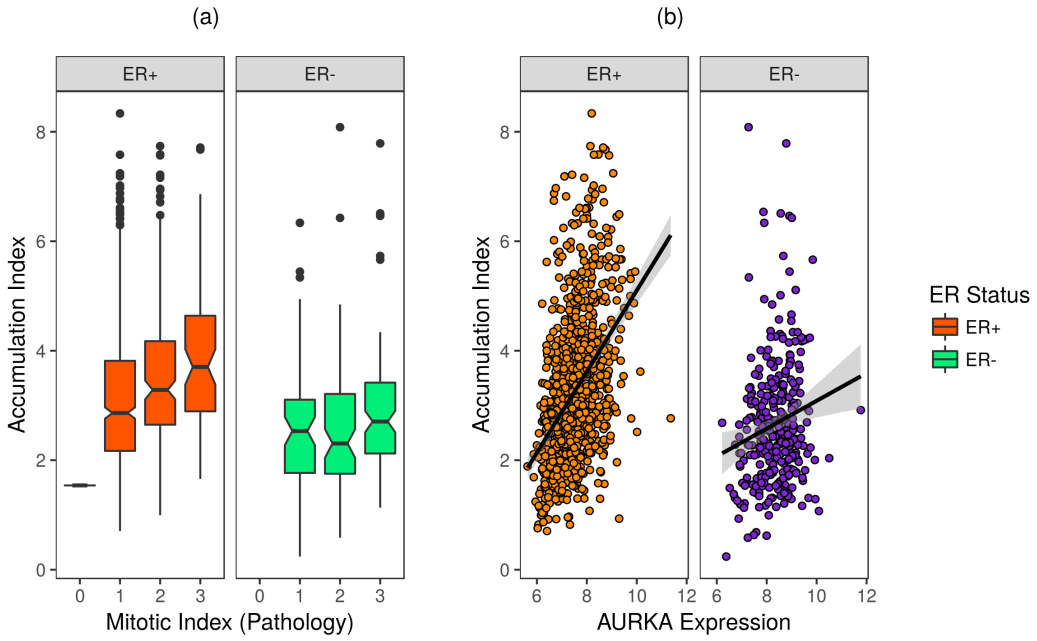


Figure 3.4: Accumulation of epigenetic drift is largely a consequence of mitotic errors. (a) Relationship between the Accumulation index and the mitotic index which is pathology-based score for ER+ and ER- tumours separately. **(b)** Relationship between the Accumulation index and expression of proliferation-related gene, *AURKA,* which is pathology-based score for ER+ and ER- tumours separately.

independent variable = BCSS, dependent variable = Accumulation index stratified into tertiles: with and without adjusting for clinical variables; Figure 3.5a). The dampened relationship in ER- patients may be explained by the fact that not all epigenetic drift related DNA methylation alterations are a consequence of stochastic cell division errors. Instead, background methylation alterations in ER- tumours may also be acquired due to environmental exposures including inflammatory conditions and/or underlying genetic traits such as genome instability [Jones et al., 2015; Teschendorff et al., 2013]. Moreover, deregulation of epigenetic modifiers have been shown to alter the global methylation landscape in ER- breast cancer (particularly in BRCA-like tumours) [Flanagan et al., 2010; Holm et al., 2010]. Although this is certainly an example of an instructive (and not stochastic) methylation event, these global changes would confound estimates of tumour background methylation differences, and therefore could be another explanation of the reduced association between the accumulation index and proliferation.

Remarkably, in ER- tumours, the Direction index was mildly but significantly associated with BCSS indicating that ER- tumours that undergo a global gain of background methylation have adverse clinical outcomes compared to those that lose methylation (univariable *p-value* = 0.1413, multivariable *p-value* = 0.0492; Cox-proportional hazards model: independent variable = BCSS, dependent variable = Direction index stratified into tertiles: with and without adjusting for clinical variables; Figure 3.5b). However, conducting a Gene Set Enrichment Analysis on the Direction index did not reveal any interesting pathways associated with a tumour's tendency to gain or lose methylation making interpretating this finding challenging. The global loss of background methylation in tumours could perhaps be explained by the contribution of lymphocytic infiltration in the tumour since they harbour lower DNA methylation estimates on average than epithelial cells [Dedeurwaerder et al., 2011]; however, examining a measure of the tumour's lymphocytic infiltration (digital pathology scores, see Chapter 2) revealed no such association with the Direction metric.

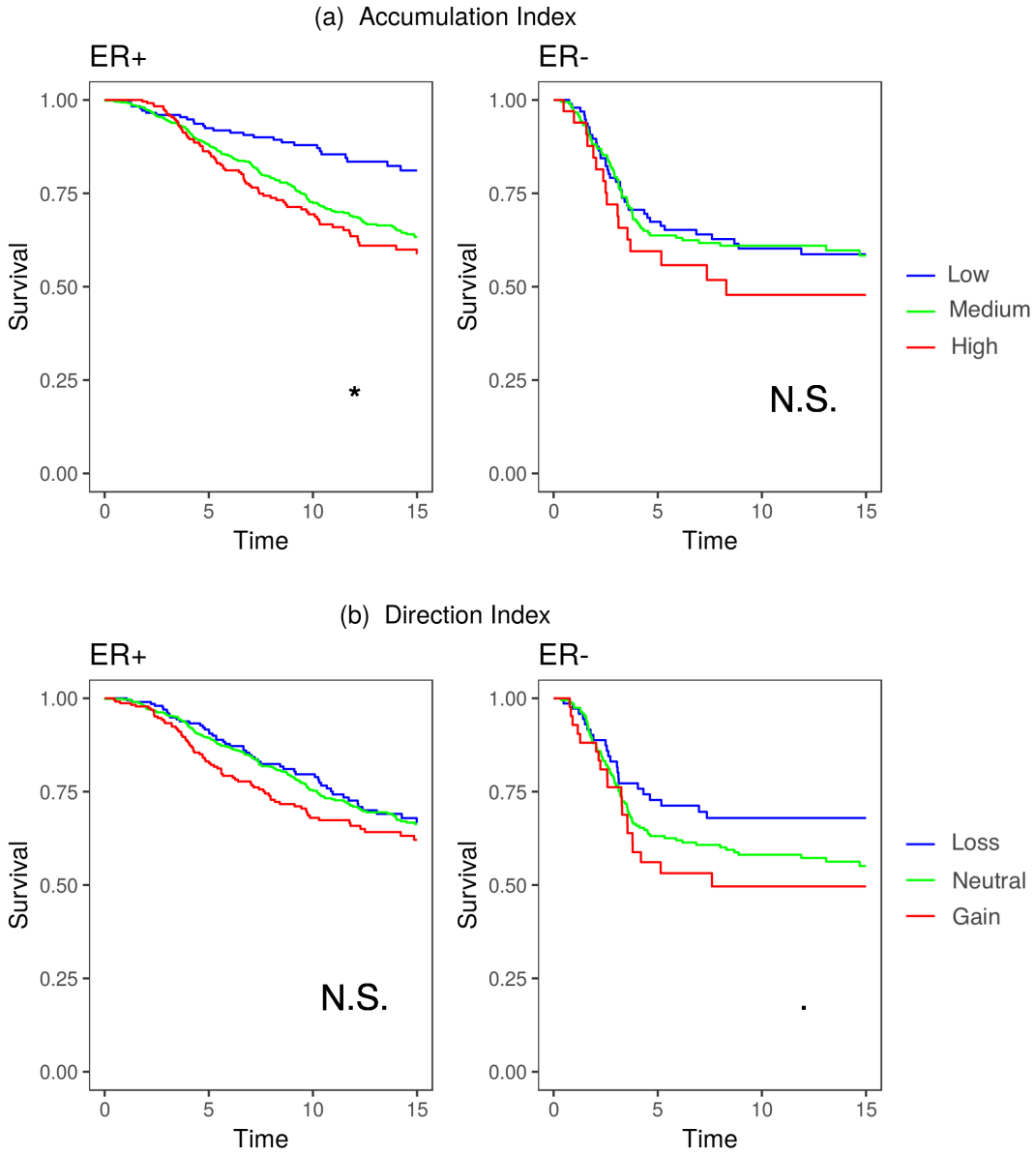


Figure 3.5: Accumulation index is prognostic in ER+ tumours and Direction index is prognostic in ER- tumours. (a) Kaplan-Meier survival estimates on BCSS stratified by Accumulation index (tertiles) for ER+ and ER- tumours. **(b)** Kaplan-Meier survival estimates on BCSS stratified by Direction index (tertiles) for ER+ and ER- tumours. Multivariable Cox proportional hazards models as described in the text. Significant analyses (alpha = 0.05) are labelled by an '*', and non-significant analyses by 'N.S.'.

3.3 Detecting class-specific and tumour-specific DNA methylation alterations

Genomic features such as promoter, enhancers and PRC region show increased DNA methylation in tumours whilst intron and intergenic regions show decreased methylation (Chapter 2). However, these global changes detected in tumours represent the aggregate of multiple specific hyper- and hypomethylation events, and it is essential to identify individual focal methylation events that may be associated with tumorigenesis. The analysis strategy presented below involves determining the genomic regions that exhibit statistically different methylation statuses when comparing all breast tumours with normal tissues (Section 3.3.1); followed by identifying individual tumour-specific methylation aberrations (Section 3.3.2).

3.3.1 Detecting Differential Methylation Regions (DMRs): Class-specific alterations

A Differentially Methylated Region (DMR) is a region of neighbouring CpG sites harbouring large methylation change in the same orientation (gain or loss) between two classes of multiple samples, such as between tumour and normal samples; or between two different tumour subtypes. These regions are likely to represent important locations of divergent epigenetic changes between these two classes with a potential regulatory role. To detect focal regions of methylation alterations, a differential methylation analysis was used to derive DMRs for the tumours compared to the normal tissue. Being a genome-wide method, RRBS enables the identification of methylation changes in a variety of different genomic features, thus providing an insight into epigenetic changes beyond the oft studied promoter.

SCCRUB is a set of coordinated regions defined based on correlations of proximal CpG sites (see Chapter 2). The underlying premise of SCCRUB is that it captures all regions within the RRBS-universe comprising of at least 3 CpGs that are spatially coordinated in their methylation behaviours. Therefore, SCCRUB represents a set of potentially functionally relevant regions over the genome. For the purpose of differential methylation analysis, only promoter, exon, intron, enhancer and PRC region were considered since they can be linked with 1 or more target genes to allow for investigation of the role of DMRs in explaining gene expression (Section 3.5). SCCRUB regions in other genomic features such as intergenic elements (excluding

enhancer and PRC region) and repetitive regions were excluded. This led to analysing a subset of the SCCRUB universe (187,095 out of 289,265 regions).

In this section, observed sample methylation estimates were used for the calculation of DMRs between the tumours and the normal tissues to obtain an overall perspective of the extent and precise locations of DMRs over the epigenome. The strategy of analysis is detailed below.

1. **Regression**

   For each region under consideration, a linear regression model was constructed using the observed methylation of the region as the independent variable and the 2-group classification as the predictor. The methylation values were transformed from Beta-values into M-values prior to fitting the model using a logistic transformation (see Chapter 2) leading to a distribution which is more appropriate for Gaussian-approaches.

2. **Adjusting for Multiple comparisons**

   An additional multiple-testing correction was applied to the *p-values* using FDR to account for the large number of genomic regions tested in parallel.

3. **Filtering**

   For the two classes being compared, mean methylation predictions from the linear regression models were transformed back to Beta-values. The average methylation difference between the two classes was calculated by subtracting these two estimates for each region. Subsequently, the genomic regions were filtered based on an absolute methylation difference exceeding 20% and a *FDR p-value* < 0.05. Hyper and hypo DMRs were identified separately depending on the orientation of the difference.

4. **Annotation**

   Regions were annotated into 5 genomic features (promoter, exon, intron, enhancer and PRC region) based on the definitions mentioned in Chapter 2.

Altogether, 13517 DMRs were identified in breast tumours compared to normal tissues (from within the 187,095 SCCRUB regions assessed). In addition to discriminating between the orientation of methylation differences (hyper and hypo), the DMRs were further stratified into 3 CpG density categories (low = CpG Density < 40 CpGs/kbp; high = CpG Density $\geq$ 80 CPGs/kbp; medium = remaining CpGs) to

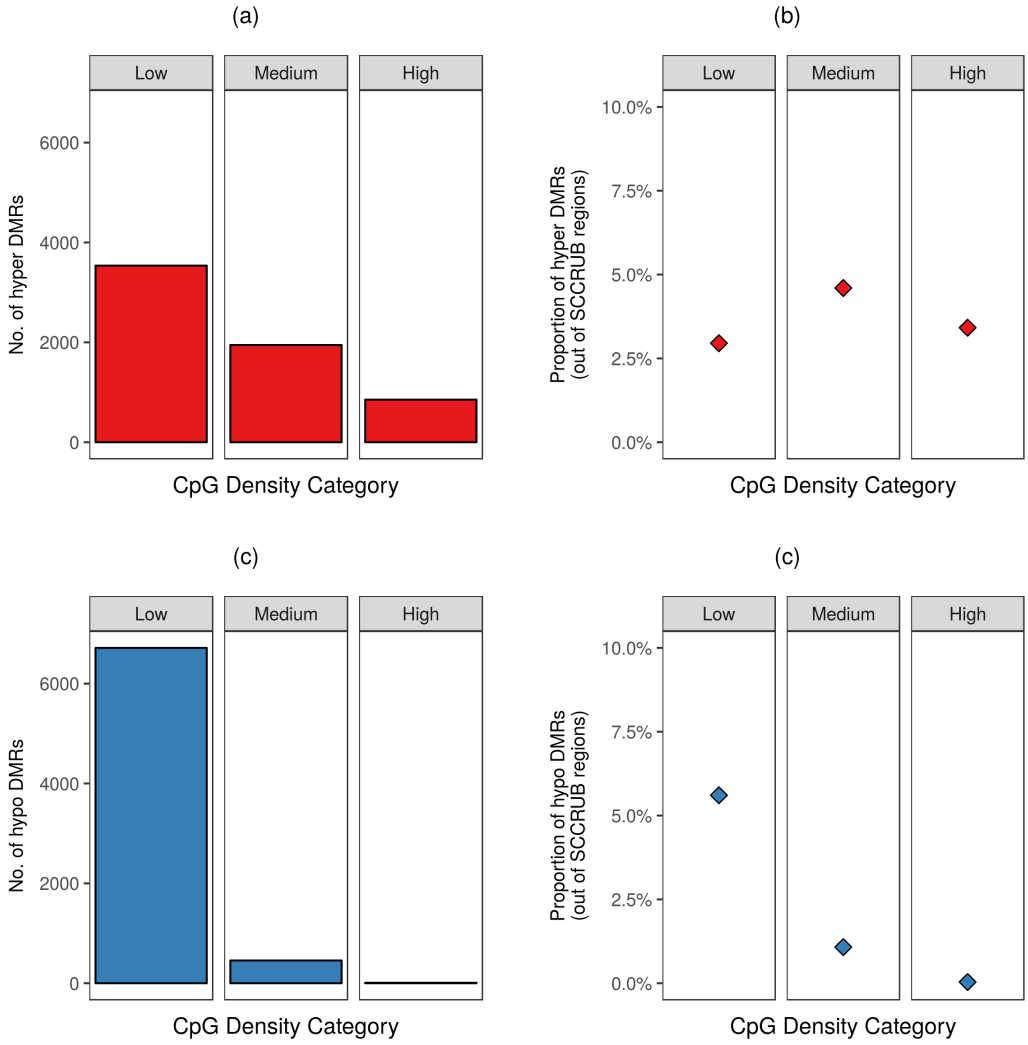


Figure 3.6: DMRs detected in breast tumours versus normal tissues. **(a)** Number of hyper DMRs. **(b)** Proportion of SCCRUB regions that are hyper DMRs. **(c)** Number of hypo DMRs detected. **(d)** Proportion of SCCRUB regions that are hypo DMRs. All plots are stratified by CpG density.

highlight context specific differences. The number of hyper and hypo DMRs identified per CpG classification are depicted in Figure 3.6a and 3.6c. whilst Figure 3.6b and 3.6d represent these numbers as a proportion of SCCRUB regions in each CpG classification separately. A higher number of hypo DMRs (n = 7180, 3.8% of all SCCRUB regions) were identified than hyper DMRs (n = 6337, 3.4% of all SCCRUB regions). A striking majority of hypo DMRs were identified in low CpG regions (6715 out of 7180, 93.5%; Figure 3.6c), suggesting a predisposition to lose methylation in low CpG regions (% of SCCRUB regions hypomethylated by CpG density: low = 5.60%, medium = 1.08%, high = 0.04%; Figure 3.6d). The CpG distribution of hyper DMRs also indicated that the majority were harboured in low CpG regions (3537 out of 6337, 55.8%; Figure 3.6a), although this was clearly due to the larger number of low CpG regions in the genome. Relatively speaking, medium and high CpG density regions had a greater propensity for gaining methylation (Percentage of SCCRUB regions hypermethylated stratified by CpG density: low = 2.95%, medium = 4.60%, high = 3.41%; Figure 3.6b).

The extent of hyper and hypo DMRs in breast cancer across different genomic features was quantified in Section 3.6. But before that analysis, tumour-specific methylation alterations are detected.

3.3.2 Detecting Methylation Altered Regions (MARs): Tumour-specific alterations

In the literature, DMRs have been used interchangeably as class-specific or tumour-specific methylation alterations. To avoid confusion between the two in this thesis, DMRs are used specifically to describe class-specific methylation alterations (as described in Section 3.3.1). For the purpose of identifying tumour-specific DNA methylation alterations versus the normal tissue, the concept of methylation altered regions (MARs) is introduced. A MAR is defined as a precise region of neighbouring CpGs that is significantly differentially methylated in an *individual* tumour compared to the normal tissue(s). Table 3.1 describes the difference between DMRs and MARs. The selection of regions used to define MARs can vary from i) universes such as the unsupervised-defined set of regions such as SCCRUB, or the set of all promoter; or ii) a supervised-comparison based filtered set of regions such as tumour DMRs.

SCCRUB is used as the universe of regions within which MARs were determined. In order to identify the set of tumours with a MAR for a particular region, the following two criteria were applied.

1. **Outlier test (Statistical significance).** This test attempts to address whether the methylation of the tumour is a significant outlier compared to the distribution of normal tissues for a specific region. A z-score is derived for each tumour by comparing its methylation estimate with the mean and standard deviation of methylation evaluated over all normal tissues. M-value methylation estimates were used for this calculation since they are more statistically valid than Beta-values in Gaussian analysis such as z-scores (Chapter 2). Specifically, if r labels the SCCRUB region under consideration; μ_r^M and σ_r^M represent the mean and standard deviation of methylation (M-value) of this region over the normal samples. If the methylation estimate (M-value) of the tumour t is X_{rt}^M, then the z-score of region r in tumour t is defined by $Z_{rt}^M = (X_{rt}^M - \mu_r^M)/\sigma_r^M$. Tumours with $|Z_{rt}^M| \geq 1.96$, corresponding to a *FDR p-value* of ≤ 0.05 were deemed to be statistical outlier with respect to the methylation of normal tissues for the region, r.

2. **Absolute difference test (Biological significance).** This test attempts to address whether the methylation difference between the tumour and the average normal tissue is biologically meaningful. The methylation difference for *each tumour* and the *mean of the normal tissues* was determined for this region. Beta-value methylation estimates were used for this calculation due to its intuitive biological interpretation (Chapter 2). Specifically, the mean normal methylation μ_r^M calculated above is transformed using the inverse-logit function to derive a Beta-converted methylation value μ_r^B. If the methylation estimate (Beta-value) of the tumour t for region r is labelled as X_{rt}^B, then the methylation difference is defined as $D_{rt}^B = (X_{rt}^B - \mu_r^B)$. Tumours that pass the 1st criteria (significant z-score, $|Z_{rt}^M| \geq 1.96$) and with $|D_{rt}^B| \geq 0.20$, were deemed to have a MAR at this region. If both the Z_{rt}^M (z-score) and D_{rt}^B (methylation difference) were positive, the MAR was classified as a hyper MAR, and if they were negative, the MAR was classified as a hypo MAR.

Using only the first criteria to determine methylation changes in relation to the normal tissue could generate large z-scores without necessarily a big change in absolute methylation value, particularly if the normal tissues are highly homogeneous with respect to methylation of that region. The combination of these two criteria ensures that tumours with MARs are not only statistical outliers with respect to the normal distribution, but the difference between the DNA methylation values are biologically relevant as well.

Differentially Methylated Region (DMR)	Methylation Altered Region (MAR)
A DMR is defined as a precise region of neighbouring CpGs that is on average significantly differentially methylated between two *classes* of multiple samples. It can be characterised as a hyper or hypo DMR depending on orientation.	A MAR is defined as a precise region of neighbouring CpGs that is significantly differentially methylated in an *individual* tumour compared to the matched normal tissue (or the pool of all normal tissues). It can be characterised as a hyper or hypo MAR depending on orientation.
A DMR represents a region of averaged methylation differences between a fixed group of tumours and a fixed group of normal tissues or between two fixed groups of tumours. DMRs can be subtype-specific.	A MAR and represents a region of acquired methylation alteration in a single tumour versus the normal tissue(s). MARs are tumour-specific.
At least one (usually more, but rarely all) tumour(s) used in the definition of a DMR must also harbour a MAR at the defined region.	A MAR for an individual tumour may occur within or outside a region defined by a DMR.

Table 3.1: Comparison between a DMR and MAR.

Tumour MARs can be detected in all SCCRUB regions of the epigenome. However, for the purpose of the following analysis, MARs were filtered such that only MARs within tumour DMR regions were considered. As explained previously, a DMR is a population-averaged methylation difference compared to normal tissue. Consequently, not *all* tumours necessarily harbour MARs within a DMR; and a tumour does not necessarily harbour MARs in *all* DMRs. For instance, if n DMRs are identified between breast tumour and normal tissues, each tumour could theoretically harbour between 0 and n MARs within these n regions. To illustrate this point, Figure 3.7a represents the proportion of hyper MARs detected in individual tumours within hyper DMRs stratified by CpG density; and Figure 3.7b represents the proportion of hypo MARs within hypo DMRs stratified by CpG density. Proportions of tumours were calculated instead of absolute numbers since the total number of tumours with methylation information varied depending on the region considered. This analysis revealed that there is considerable heterogeneity observed in the extent and location of MARs within the 1482 breast tumours. The recurrence of MARs in breast tumours is formally investigated in the next section.

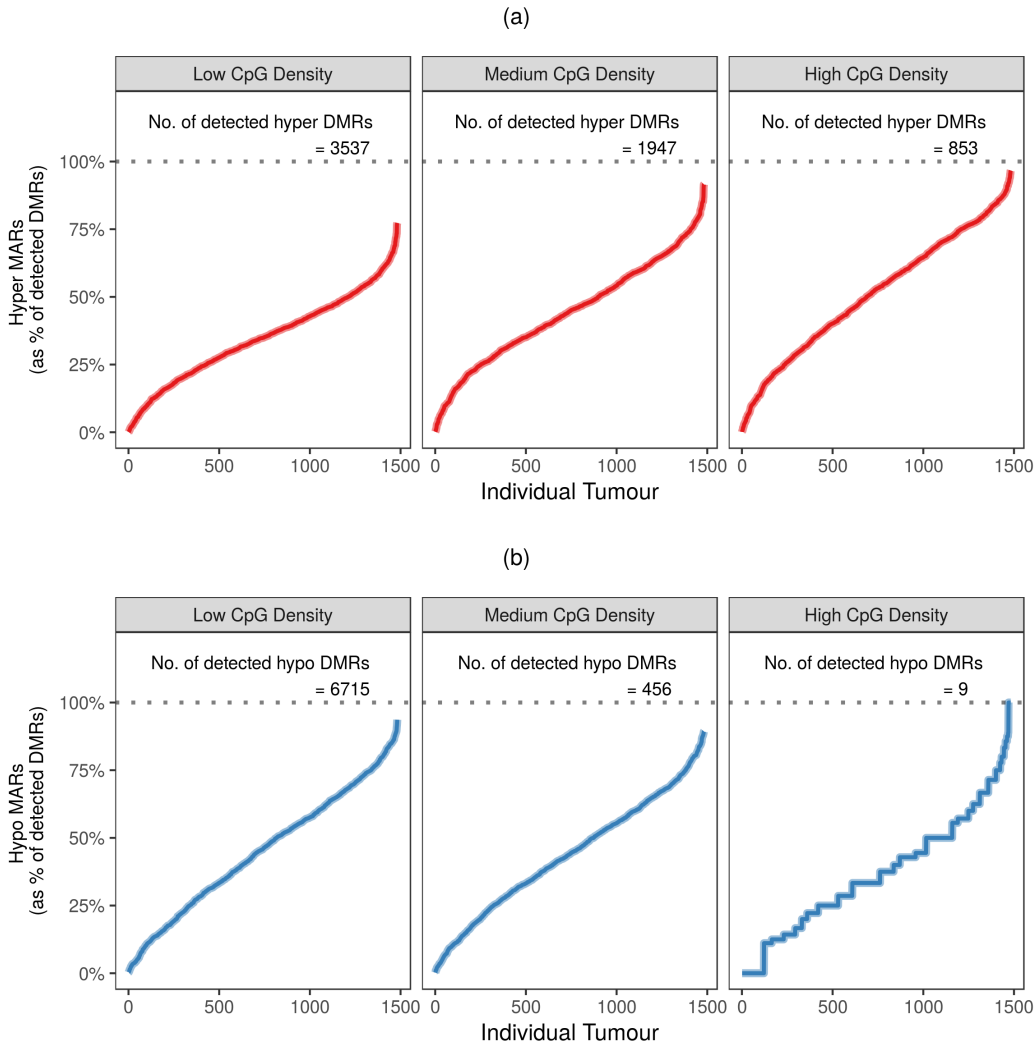


Figure 3.7: MARs detected within DMRs for all breast tumours. (a) Hyper MARs. Analysis conducted separately for 3 CpG density categories. The dotted line at 100% (y-axis) represents the total number of hyper DMRs detected in the CpG density category. The monotonic red lines represent the extent of hyper MARs as a proportion of the detected hyper DMRs (y-axis) for each of the 1482 tumours (x-axis). The tumours (x-axis) are ordered by increasing number of hyper MARs for each of the CpG density category separately. **(b)** Repeated for hypo MARs. The monotonic blue lines represent the extent of hypo MARs as a proportion of the detected hypo DMRs (y-axis).

3.4 DMARC – a novel algorithm for the identification of Directed Methylation Altered Regions in Cancer

The overarching goal of this chapter is to identify regions of functional methylation alterations over the breast cancer methylome. In Section 3.2 it was revealed that breast cancers have extraordinary variation in their background methylation differences that are not only tumour-dependent but also highly sequence-specific (CpG density and TOR). However, the detection of breast tumour DMRs and tumour-specific MARs (in Section 3.3) utilised uniform methylation difference thresholds on observed tumour methylation estimates (unadjusted for background differences) across the genome and for all tumours, as is standard practice in most current cancer methylome studies [Akalin et al., 2012b; Kulis et al., 2015; Robinson et al., 2014]. Consequently, it is highly likely that a considerable contribution of the DMRs and MARs are a consequence of background methylation differences ascertained in Section 3.2. It is necessary to deconvolute the contribution of background methylation differences and directed methylation differences. As defined earlier, directed methylation comprise of all instructive methylation alterations as well as stochastic alterations that are selected for and consequently observed more frequently in the tumour cell population. In contrast, background methylation alterations essentially comprise of all stochastic alterations that are not under selection. Discrimination between the two processes would shed light into methylation mechanisms with the hypothesis being that stochastic methylation differences are less likely to have functional consequences than directed methylation differences (Section 3.1).

A novel algorithm called DMARC (Directed Methylation Altered Regions in Cancer) is presented here. DMARC is able to characterise MARs and subsequently, DMRs into *directed* and *background* by incorporating the observed heterogeneity in background differences.

3.4.1 Directed Methylation Altered Regions

Given that MARs are tumour-specific as well as region-specific, they represent ideal candidates for adjustment using background methylation differences (which have already been established as tumour and context-specific), hence they were analysed first. As previously described, tumour-specific MARs within a region were determined if they passed the statistical outlier test as well as the absolute difference test (using a

uniform difference threshold of 20%) on comparison with the methylation distribution of normal tissues. While it is obvious that testing for an absolute methylation difference that is too low will lead to significant findings that are likely to be biologically irrelevant [Teschendorff et al., 2016b], it is less appreciated that failing to account for heterogeneity in the process of accumulating methylation differences can also lead to spurious results. A similar argument has been made in the case of mutations [Lawrence et al., 2013]. This point can be illustrated by the following example. Hyper MARs with an observed methylation difference D^B_{rt} of 25% were detected for a tumour in two distinct promoter regions. The first promoter region lies within the highest CpG density decile and highest TOR decile of the genome, and the second lies within median CpG density and TOR deciles. The critical question is – *are the mechanisms underlying the increase in promoter methylation the same for both regions?* Has methylation been purposely deposited to the promoter to fulfil some potential function that is under selection (*directed*), or is the hyper methylation merely a consequence of an accumulation of methylation errors that is occurring in regions with similar genomic frameworks over the genome (*background*)? Solely, by virtue of sequence context, a background methylation increase of 8% to 17% (inter-quartile range of background methylation difference for genomic context of the first region) can be expected in the first promoter region due to epigenetic drift, while a background difference of -2% to 1% (inter-quartile range of background methylation difference for genomic context of second region) is expected in the second promoter region. Therefore, if a constant methylation threshold of 20% is used, both promoter regions may be assumed to be associated with directed irregularity in the epigenetic machinery in that region. However, in this scenario it is apparent that the hyper MAR detected in the first region has a considerable contribution due to accumulation of cell-division associated methylation errors (linked to the background methylation difference); and the hyper MAR second region is more likely to be a consequence of directed deposition of methylation. Furthermore, heterogeneity in background methylation differences across different tumours can also lead to erroneous interpretations of methylation alterations. Tumours undergoing excessive cell proliferation will demonstrate additional accretion of methylation errors.

Thus, it is clear that background methylation differences need to be considered to make inferences on the methylation mechanism underlying a MAR. The background methylation differences (vs. normal tissues, M-values used), b^M_{gt} have already been approximated for different genomic contexts, g by using genomic covariates (such as CpG density and DNA replication timing) and tumours, t (see Section 3.2). In order to account for these background methylation differences, additional steps are proposed

post detection of MARs. Specifically, the genomic context, *g,* based on CpG density and DNA replication timing is identified for the region under consideration, *r* and a background adjusted methylation estimate is calculated as $\hat{X}^M_{rt} = X^M_{rt} - b^M_{gt}$, where X^M_{rt} represents the observed methylation estimate, as noted in Section 3.4.2. Beta-values, $\hat{X}^B_{rt}$ as well as M-values, $\hat{X}^M_{rt}$ for background adjusted methylation estimates can be calculated using the respective observed methylation estimates, X^B_{rt} or X^M_{rt}. Outlier tests ($|\hat{Z}^M_{rt}| \geq 1.96$); and the absolute difference test ($|\hat{D}^B_{rt}| \geq 0.20$) were calculated using the background adjusted methylation estimates ($\hat{X}^B_{rt}$ and $\hat{X}^M_{rt}$) instead of the observed methylation estimates (X^B_{rt} and X^M_{rt}, as described in Section 3.3.2), and if the MAR passed both these criteria it was characterised as a directed methylation altered region (directed-MAR), and if not then as a background methylation altered region (background-MAR). A schematic diagram of the DMARC algorithm is shown in Figure 3.17.

To illustrate the difference between directed-MARs and background-MARs, Figure 3.9a represents the proportion of hyper directed-MARs detected in individual tumours within hyper DMRs. This is similar to Figures 3.7a with the addition of grey dots to represent the proportion of directed-MARs. Background-MARs are defined as MARs that are not directed-MARs and are thus represented by the distance between total MARs and directed-MARs. Hypo directed-MARs and background-MARs are indicated in Figure 3.9b. This analysis revealed that most of the hyper background-MARs occur in CpG-rich regions (mean number of background-MARs per tumour stratified by CpG Density: low = 3, medium = 53, high = 116, *p-value* $< 1\times10^{-300}$; Analysis of Variance), while most of the hypo background-MARs occur in low CpG regions (mean number of hyper background-MARs per tumour by CpG Density: low = 1081, medium = 34, high = 0, *p-value* $< 1\times10^{-300}$; Analysis of Variance). However, due to the considerably larger number of CpG-poor DMRs (as a result of an average low CpG density globally), a vast majority of the identified background-MARs represent losses rather than gains in methylation (mean number of background-MARs per tumour by direction of methylation alterations: hyper = 170, hypo = 1115, *p-value* $= 1.2\times10^{-232}$; Wilcoxon signed rank test).

As explained previously, the vast majority of CpG sites in high CpG density and late TOR are unmethylated in normal tissues and will gradually gain methylation randomly in tumours due to DNA replication errors. Accordingly, the DMARC algorithm adjusts (increases) the threshold for directed methylation gains in these regions. However, since methylation losses in high CpG density sites in tumours represent a departure from what is expected based on its genomic context, all losses

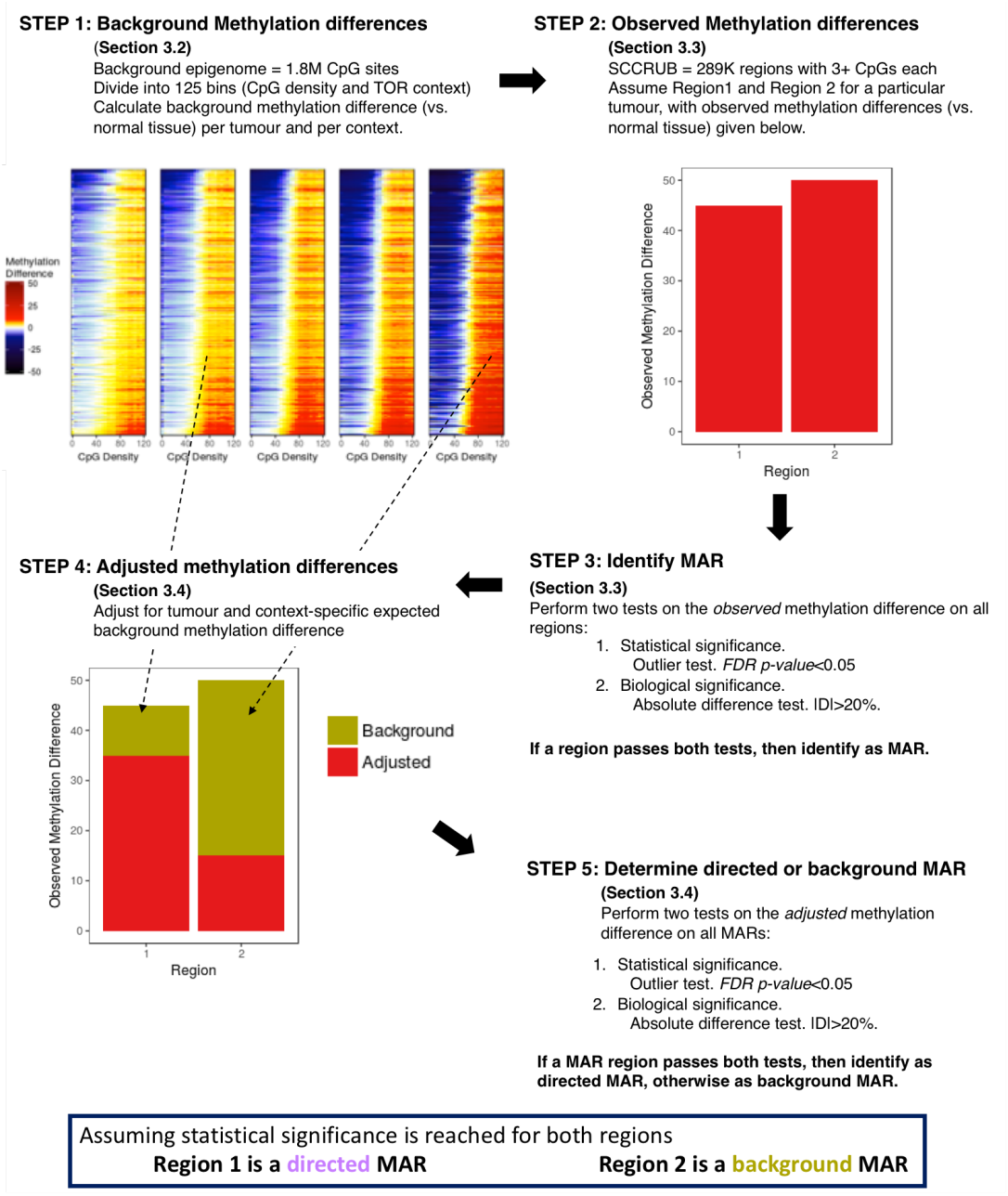


Figure 3.8: Schematic diagram of the DMARC algorithm. DMARC allows the discrimination between directed and background methylation altered regions in cancer.

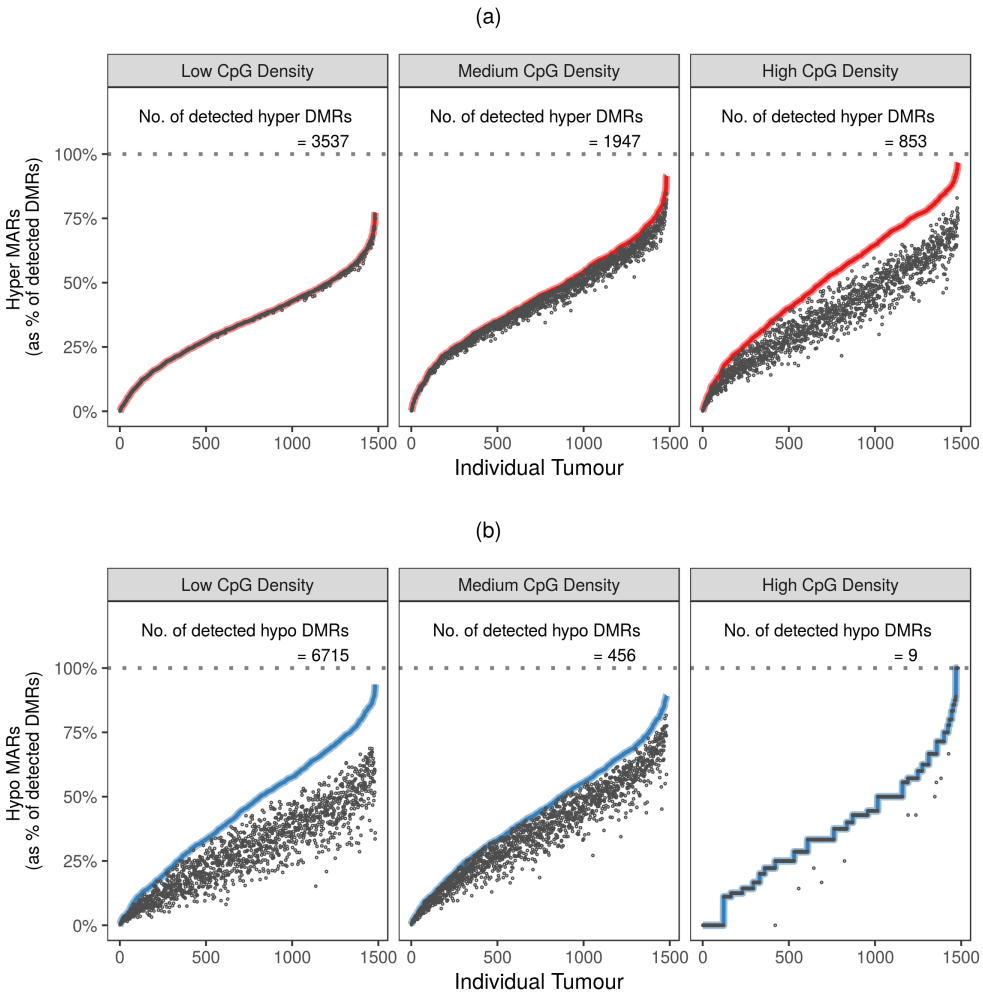


Figure 3.9: Directed-MARs detected within DMRs for all breast tumours. (a) Hyper MARs. Analysis conducted separately for 3 CpG density categories. The dotted line at 100% (y-axis) represents the total number of hyper DMRs detected in the CpG density category. The monotonic red lines represent the extent of hyper MARs as a proportion of the detected hyper DMRs (y-axis) for each of the 1482 tumours (x-axis). The black dots represent the proportion of tumour-specific hyper directed-MARs, and accordingly hyper background-MARs are represented by the distance between the hyper MARs (red line) and hyper directed-MARs (black dot). The tumours (x-axis) are ordered by increasing number of hyper MARs for each of the CpG density category separately. **(b)** Analysis repeated for hypo MARs. The monotonic blue lines represent the extent of hypo MARs as a proportion of the detected hypo DMRs (y-axis). The black dots represent the proportion of tumour-specific hypo directed-MARs.

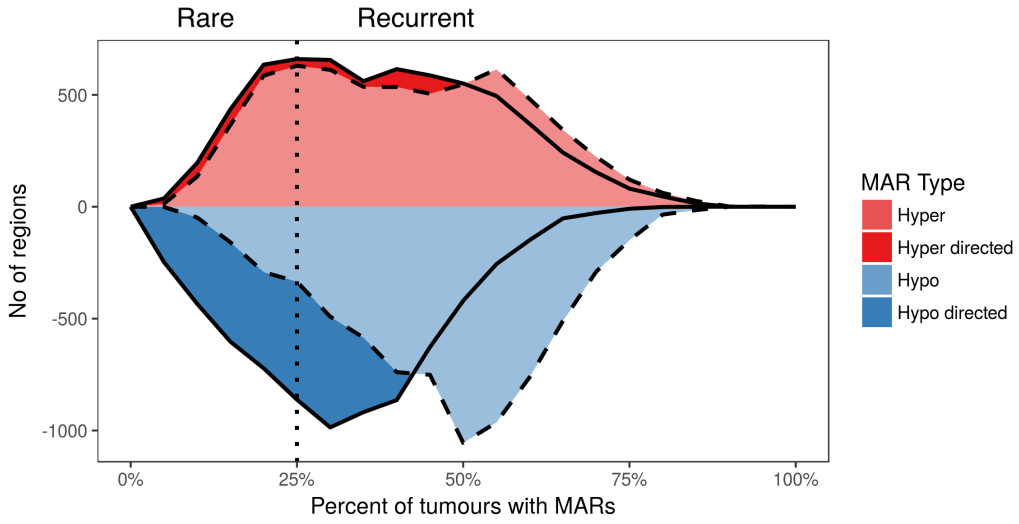


Figure 3.10: Recurrence of MARs. Regions that are altered in different proportions of tumours (x-axis) were enumerated as explained in the text and the resulting distribution was plotted for hyper and hypo regions separately. The solid lines and solid colours represent all MARs (background and directed), whereas the dashed lines and translucent colours represent directed-MARs. Red and blue curves represent the number of hyper and hypo regions respectively per frequency group of tumours. The vertical dotted line delimits the number of regions in which at least 25% of tumours harbour MARs and directed-MARs.

within high CpG density sites are assessed as directed MARs. This is why most hyper MARs in CpG-poor regions are classified as directed MARs, and most hypo MARS in CpG-rich regions are classified as directed MARs. This is a limitation of the DMARC algorithm, and a future version of the algorithm is being developed to address this.

Next, the recurrence of all MARs (background and directed) as well as specifically directed-MARs were investigated. For each region (all detected hyper and hypo DMRs are considered), the proportion of tumours that harboured a MAR (i.e. the proportion of tumours in which the region was aberrantly methylated) and the proportion of tumours that harboured directed-MARs were enumerated. This was done separately for hyper and hypo MARs. Proportions of tumours were calculated instead of absolute numbers since the total number of tumours with assayed methylation information varied depending on the region considered. The regions were ordered according to proportion of tumours with directed-MARs and the resulting distribution was plotted (Figure 3.10; solid colours, solid lines). A similarly constructed distribution for all MARs (directed and background) was plotted (Figure 3.10; translucent colours, dashed lines). Approximately 88% of the considered regions are altered (MARs) in at least 25% of tumours, and 76% of these regions have directed-MARs in the same fraction of tumours. Overall, these results indicate that there are thousands of recurrent methylation alterations in a substantial fraction of human breast cancers.

3.4.2 Directed Differentially Methylated Regions

The background estimates of methylation differences are sequence specific, and the corresponding CpG density and TOR are available for each DMR. However, there is also vast inter-tumour heterogeneity in the background methylation difference. DMRs are population-averaged difference (for instance, the identification of DMRs in tumours versus normal tissues involved comparison of 1482 breast tumours and 237 normal samples), and therefore, adjusting for tumour specific estimates directly is not trivial. However, once MARs are classified into directed-MARs and background-MARs it becomes straightforward to characterise DMRs into directed and background DMRs as well. For each DMR, the proportion of tumours with directed-MARs (out of total number tumours with MARs) was calculated. A threshold of 70% was used. If more than 70% of the MARs detected within the region were directed-MARs, then the DMR was denoted as a directed-DMR. The rest were classified as background-DMRs. Using this threshold, 67% of all DMRs were categorised as directed-DMRs, and 33% as background-DMRs. However, hypomethylation in low CpG density regions accounted

for a striking majority of background-DMRs (Figure 3.11) for reasons explained previously in Section 3.4.1.

Next, the potential regulatory roles of *directed* and *background* DNA methylation alteration in breast cancer pathogenesis was studied.

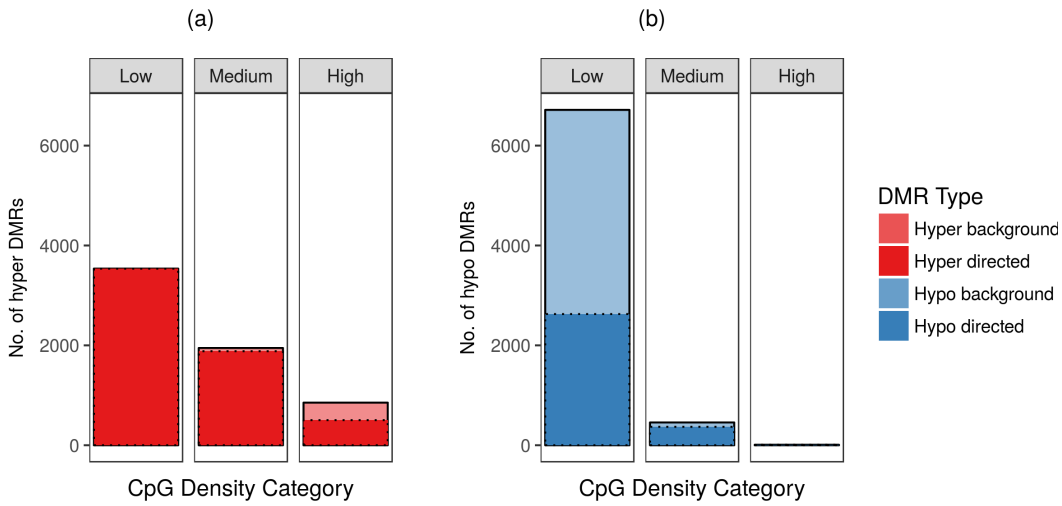


Figure 3.11: Directed and background DMRs detected in breast tumours versus normal tissues. (a) Number of directed and background hyper DMRs. **(b)** Number of directed and background hypo DMRs detected. Both plots are stratified by CpG density.

3.5 Altered DNA methylation is a regulatory mechanism in breast cancer

To explore the functional relevance of the identified DNA methylation alterations, the relationship between the expression and the methylation estimates of genes harbouring DMRs was evaluated. 11729 DMRs (both directed and background DMRs) were identified between breast tumours and normal tissues. But how many of these DMRs are potentially involved in gene silencing or activation?

3.5.1 Identification of expression-DMRs

Methylation that occurs directly on genes such as promoter, exon and intron is intuitively associated with potential regulation of the underlying gene itself and so DMRs within these 3 genomic features were examined to explore intragenic associations with expression, whereas, enhancer and PRC region DMRs were explored for distal cis-regulation. However, identification of the specific target gene or genes whose expression is modulated by distal regulatory elements can be challenging since the target gene is not necessarily the closest gene. In a Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) study to examine promoter-enhancer interactions in human cell lines including the MCF-7 breast cancer cells, Li et al. [2012] observed that almost half (40%) of enhancers bypass several genes to interact with their target promoter [Gao et al., 2015]. In fact, the median distance between the enhancer and promoter was 50kb, although a small proportion (less than 10%) of long distance enhancer-promoter interactions were observed up to millions of base pairs away from the enhancer. In order to identify candidate target genes modulated by a particular enhancer or PRC region DMR, nearby genes within a 400kb genomic window centred around the DMR were selected. Although target genes further than 200kb than the distal-regulatory DMRs would not be included, most regulatory DMR-gene pairs would be considered. DMRs for which a target gene could not be identified were excluded from further analysis.

To investigate the functional roles of both intragenic and distal methylation alterations, associations between methylation and gene expression was explored in the breast cancer samples for each *DMR-target gene* pair as defined above. A multivariable linear regression framework of gene expression (normalised $\log_2$ relative intensities) as the independent variable and using both DNA methylation (M-value) and CNA (segmented mean $\log_2$ ratios, see Chapter 2) as covariates was explored for each

DMR-gene pair. Gao and Teschendorff [2017] used a similar approach and noted that the inclusion of CNA in the model ensures that the relationship between gene expression and DNA methylation was due to concomitant alterations within the same set of tumours; and that the observed change in gene expression was not explained by concurrent copy number alterations in the tumours . For each DMR-target gene pair, two statistics were noted: i) the partial correlation ($rho_{meth\text{-}independent}$) between methylation of the DMR and expression of the target gene which is a measure of the independent strength and direction of this relationship whilst controlling for the effect of CNA; and ii) the regression *p-value* (t-test) of the methylation coefficient which indicates the strength of evidence for its association with gene expression. Only samples utilised in the identification of the DMRs were included in the analysis. For instance, for ER+ vs. normal DMRs, all ER+ and normal samples were used. However, not all 1719 METABRIC samples (1482 tumours and 237 normal tissues) had matched gene expression and CNA data. The 1342 tumours and 108 normal tissues that had matched gene expression (microarray; Chapter 2), CNA (array) and DNA methylation data were considered for this analysis.

A region-centric approach was desired to ensure a single DMR is represented only once in the analysis and the following two strategies were implemented to accomplish this.

1. For DMRs associated with more than one genomic feature, a biologically plausible priority ranking was employed: Promoter > Exon > Intron > Enhancer > PRC region. Using this approach, a DMR detected in an intragenic enhancer was considered simply as an intragenic (exon or intron) DMR and was only linked with the target gene that it lies within.

2. For distal regulatory DMRs that have more than one target gene, regression *p*-values were corrected for multiple testing (using the Benjamini-Hochberg procedure) for each of these DMRs separately. This controls the false discovery rate (FDR) for multiple genes tested per DMR. Subsequently, for each DMR, the target gene with the highest partial methylation-expression correlation was selected.

Although this filtering step ensured each DMR is represented only once, it is important to note that two distinct DMRs could be linked with the same gene. An additional round of FDR correction was conducted over all unique *DMR-target gene* pairs to control for the number of DMRs tested. *DMR-target gene* pairs with $|rho_{meth\text{-}independent}| \geq 0.40$ and *FDR p-value* < 0.05 were considered to be significantly

associated and denoted as expression-DMRs. The orientation of the DMR (hyper or hypo methylation) and the direction of concomitant change in expression (upregulation or repression) was utilised to classify the expression-DMRs into four categories: i) hyper expression-DMR associated with gene repression; ii) hyper expression-DMR associated with gene upregulation; iii) hypo expression-DMR associated with gene repression; iv) hypo expression-DMR associated with gene upregulation. The direction of $rho_{meth\text{-}independent}$ was used to characterise the direction of the methylation-expression relationship. A negative correlation denotes the well-studied view of increased methylation being associated with repression; however, positive correlations in which increased (decreased) methylation is associated with upregulation (downregulation) of the gene were also recorded.

3.5.2 Directed-DMRs are enriched for concomitant expression changes

Using this approach, a total of 1653 expression-DMRs were identified across the 5 genomic features assessed. The first line of analysis was to investigate the hypothesis that directed methylation alterations in breast cancer are more likely to have functional consequences (as measured by concomitant gene expression changes) than background methylation alterations. A Fisher's exact test was used to assess the relationship between directed-DMRs AND expression associated DMRs (expression-DMRs). 15.4% of all directed-DMRs were associated with expression changes. Reassuringly, directed-DMRs were significantly more likely to be associated with gene expression changes in breast cancer than background-DMRs ($\log_2$ odds of expression-DMR being directed = 0.155, background = -0.358; Odds Ratio = 1.43; *p-value* = 1.2×10^{-9}; Figure 3.12a).

Gene Set Enrichment Analysis (GSEA: pathways tested included Hallmark gene sets obtained from the Molecular Signatures Database [Subramanian et al., 2005, MSigDB]) were performed on genes that were significantly up and down regulated (expression-DMRs) to identify relevant mechanisms that are disrupted due to methylation alterations. Enrichment scores were calculated and significantly enriched pathways (enrichment = observed/ expected > 5; *FDR p-value* < 0.05; hypergeometric test) were identified. Pathways that were simultaneously significantly enriched for both up and down regulated genes were also noted and their enrichment scores were added to reflect this dual role. The analysis was conducted separately on directed and background expression-DMRs, and significantly enriched pathways were listed for the two methylation mechanisms (Figure 3.12b). Interestingly, gene expression

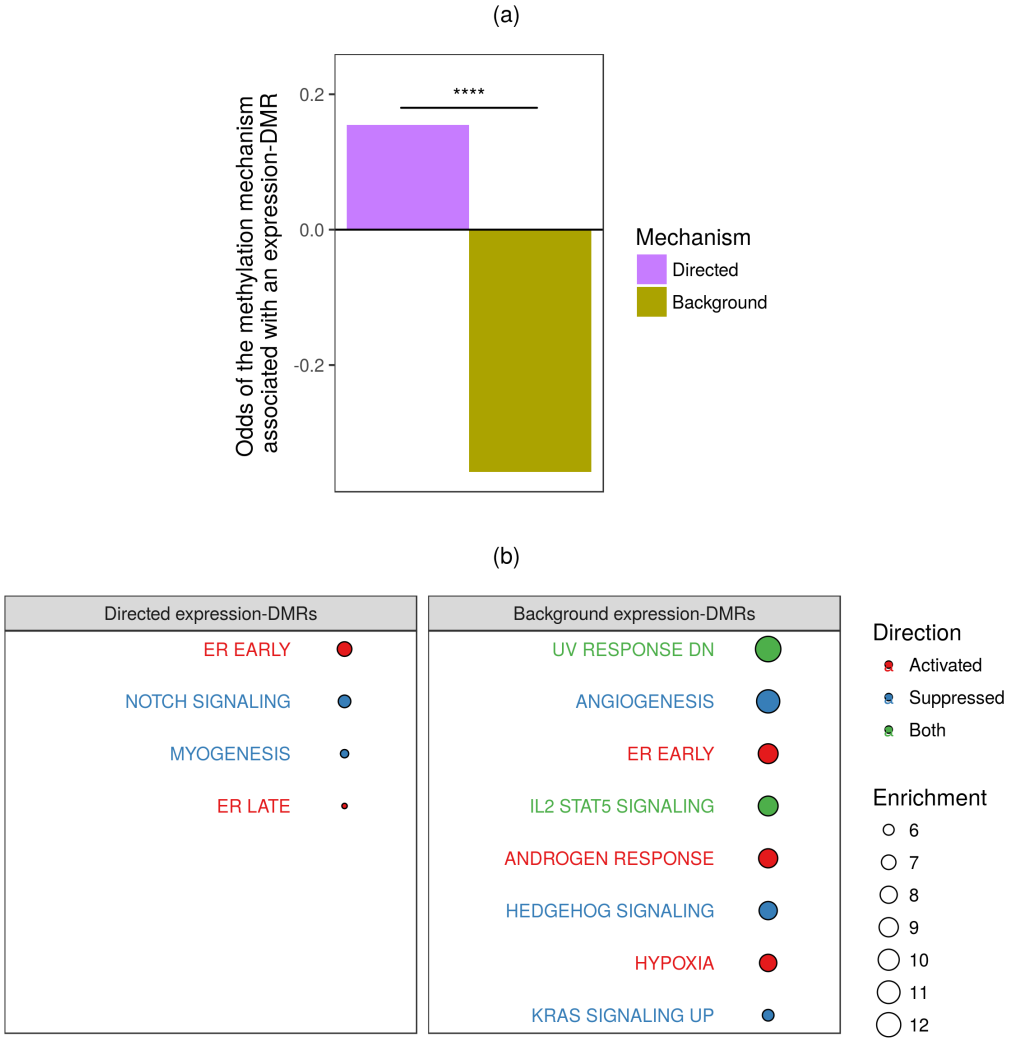


Figure 3.12: Directed DMRs are enriched for concomitant expression changes. (a) Odds of an expression-DMR being directed or background. Fisher's exact test used for calculation of *p-value*. **(b)** Significantly enriched gene sets comprising of at least 4 genes that are regulated by directed and background DMRs (Enrichment (observed/ expected) > 5, *FDR p-value* < 0.05; hypergeometric test, as explained in text). Colour of points represent direction of regulation. Size of points represent level of enrichment.

programmes disturbed by background expression-DMRs included many pathways elicited by specific cell-extrinsic stresses. Particularly response to ultraviolet radiation as well as interleukin-2 stimulation, a type of cytokine signalling in the immune system, were associated with *both* up and downregulated genes. Background DMRs were also implicated in the aberrant upregulation of genes in response to low oxygen levels (hypoxia). Conversely, directed DMRs were more likely to be linked with cell-intrinsic biological signalling pathways such as ER and NOTCH signalling, further substantiating the conclusion that the mechanisms underlying background and directed DMRs are distinct.

Therefore, discerning between directed DMRs and background DMRs has not only enabled enrichment for expression-related consequences (directed-DMRs, Figure 3.12a), but also separation of methylation-related functional modifications that are involved in cellular stress pathways in response to extrinsic exposures (background-DMRs, Figure 3.12b). Moreover, a component of background DMRs are also likely to be a consequence of DNA methylation maintenance errors as a result of mitotic division that disrupt the epigenome in a non-targeted manner (Section 3.2 and DMARC model). Based on this evidence, it is tempting to speculate that alterations associated with epigenetic drift (background DMRs) are not functionally relevant. However, these stochastic alterations may occasionally hit key genes in cancer-relevant pathways such as ER-signalling, androgen-signalling, hedgehog-signalling and KRAS signalling pathways. Consequently, although discrimination between directed and background DMRs (and expression-DMRs) has proved extremely valuable for a better understanding of methylation processes, it would be erroneous to ignore background DMRs entirely. Therefore, for the remainder of this thesis, both background and directed DMRs (and expression-DMRs) are considered, and in some investigations, the distinction between the two mechanisms is highlighted.

The regulatory role of hyper and hypo DMRs across different genomic features was investigated separately in ER+ and ER- tumours in Section 3.6.

3.6 Subtype-specific epigenetic programming in breast cancer

3.6.1 Tumour-normal differences in ER+ and ER- breast cancer

The extent of aggregate DNA methylation changes varies greatly across the breast cancer subtypes (Figure 2.6c, Appendix A.1), suggesting that breast tumours harbour subtype-specific methylation alterations as well. Hyper and hypo DMRs (versus the normal tissues; methylation difference $\geq 20\%$; *FDR p-value* ≤ 0.05) were identified in ER+ and ER- tumours separately as this classification encompasses significant differences in the fundamental biology of breast tumours. The extent of hyper and hypo DMRs (versus normal tissues) in each of ER+ and ER- breast tumours across 5 different genomic features: promoter, exon, intron, enhancers and PRC region was quantified. Both directed and background DMRs were considered. Multiple 2-class comparisons are possible within the same classification. For example, ER status divides the samples into three categories (ER+ tumours, ER- tumours and normal tissues), and thus three 2-class comparisons are possible: ER+ vs. normals, ER- vs. normals and ER+ vs. ER-). To account for this, standard errors for the classes (and hence *p-values*) were adjusted following a multiple comparison procedure for simultaneous inference using the changepoint contrast [Hothorn et al., 2008].

Figure 3.13 (top panel) displays the proportion of hyper and hypo DMRs within each genomic feature (relative to the total number of SCCRUB regions within the respective genomic feature) detected across both tumour subtypes. Since the total number of SCCRUB regions assessed within a genomic feature are constant for ER+ and ER- tumours, this allows comparison between the 2 subtypes, and consequently, identification of the one that has stronger tendencies towards directing their epigenetic changes to specific regions of the genome. Figure 3.13 (middle panel) displays the ratio between the number of hyper and hypo DMRs for each genomic feature. This allows determination of the inclination of a tumour subtype to significantly gain or lose methylation in a specific genomic feature. Finally, an enrichment analysis (enrichment = observed/ expected, hypergeometric test) was conducted across hyper and hypo DMRs separately, to investigate whether any of these 5 genomic features are preferentially targeted within that particular subtype. For example, are *promoter* more likely to be hyper methylated (compared with other genomic features) within ER+ tumours? For this purpose, a hypergeometric test was conducted independently for each of the 5 genomic features: promoter, exon, intron, enhancers and PRC region;

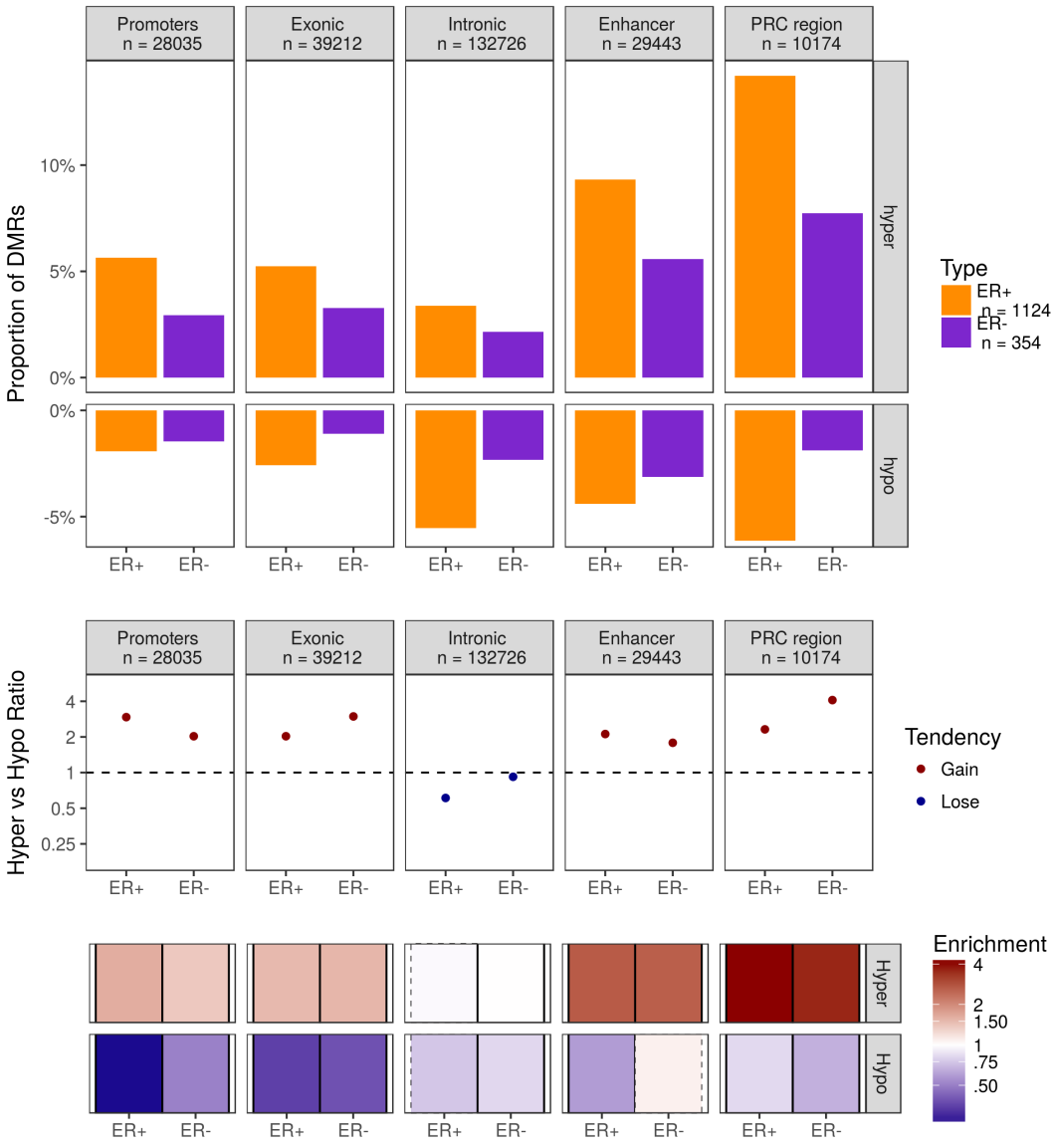


Figure 3.13: (Caption on next page.)

Figure 3.13: (Previous page.) **Hyper and hypo DMRs detected in ER+ and ER- tumours. (top panel)** The proportion of hyper and hypo DMRs within each genomic feature, relative to the total number of SCCRUB regions within the respective genomic feature (the total number is noted in the boxes at the top), detected across ER+ and ER- tumours. x-axis represents the tumour subtype (orange is ER+, purple is ER-) and the five panels represent the individual genomic features. Analyses were conducted separately for the 2 tumour subtypes and 5 genomic features resulting in 10 bars. Positive bars on y-axis represents hyper DMRs and negative bars represents hypo DMRs. **(middle panel)** The ratio between the number of hyper and hypo DMRs detected for each genomic feature which represents the inclination of a tumour subtype to significantly gain or lose methylation in a specific genomic feature. **(bottom panel)** Enrichment analysis of hyper and hypo DMRs across the 5 genomic features conducted separately for each tumour subtype, as explained in the text (hypergeometric test). Top squares represent enrichment of hyper DMRs and bottom represent enrichment of hypo DMRs. Colour of the squares represent level of enrichment (observed/ expected). Red represents enriched (enrichment > 1), blue represents depleted (enrichment < 1) and white represents no enrichment (enrichment = 1). Square boundaries represent whether the enrichment was significant (solid lines = *FDR p-value* < 0.05; dotted lines = *FDR p-value* > 0.05).

and across the 2 DMR orientation classifications: hyper and hypo. This was repeated across the 2 tumour subtypes separately: ER+ and ER- tumours.

This analysis revealed that the most enriched methylation events across all breast tumours in general (both ER+ and ER- tumours) are hyper methylation of PRC region with 14.2% of PRC region in ER+ being differentially hypermethylated (enrichment = 4.08, *FDR p-value* < 1.0×10^{-255}) and 7.7% in ER- tumours (enrichment = 3.61, *FDR p-value* < 1.0×10^{-255}). Previous studies have noted hypermethylation of PRC region in cancer [Ohm et al., 2007; Schlesinger et al., 2007; Widschwendter et al., 2007], and more recently in chronic lymphocytic leukemia (CLL) [Kulis et al., 2012]. Epigenetic switching from polycomb-repressive complex mediated gene silencing to a more stable silencing conveyed by DNA methylation has been proposed as a likely explanation [Baylin and Jones, 2011]. Nevertheless, the observation that hyper DMRs are most enriched in PRC region is a first for breast cancer.

Hypermethylation of enhancer regions (9.3% in ER+; 5.6% in ER-), promoter regions (5.6% in ER+; 2.9% in ER-) and exonic regions (5.2% in ER+; 3.2% in ER-) were also significantly enriched (enrichment > 1.3, *FDR p-value* < 0.0001; all comparisons mentioned). Enrichment of hypermethylation in these genomic features concurs with findings from the previous section since these regions lie largely in medium/ high CpG dense regions and thus are more likely to undergo hypermethylation in tumours. Promoter and exonic hypermethylation has been reported previously in ER+ and ER- tumours [Fleischer et al., 2014; Györffy et al., 2016; Rønneberg et al.,

2011], however, methylation alterations in enhancer regions have been understudied in breast cancer. In addition, the most underrepresented events in breast cancer (both ER+ and ER- tumours) were hypomethylation of promoter and exon, suggesting that that these regions hardly undergo demethylation, possibly because they are unmethylated in normal tissue (see Chapter 2). Conversely, intron undergo both hyper and hypo methylation to a similar extent in both tumour subtypes. In fact, regions within intron were neither significantly enriched/ depleted for hyper DMRs nor hypo DMRs.

On aggregate, ER+ tumours amass more hyper as well as hypo DMRs than ER- tumours (Hyper DMRs: ER+ = 12310, ER- = 7392; Hypo DMRs: ER+ = 10814, ER- = 5047). Although this may be in part due to smaller sample size for ER- tumours, this observation is consistent with the analysis of epigenetic drift in Section 3.2 where ER+ tumours were found to accumulate more background methylation differences (both gains and losses) than ER- tumours. This suggests that the epigenetic machinery is deregulated to a higher degree in ER+ tumours, although this could also be a consequence of ER+ patients being older than ER- patients.

This raises the interesting question – how many *breast cancer DMRs* (vs. normal tissue) are shared between ER+ and ER- tumours, and how many are subtype-specific? Jaccard indices (size of intersection over size of union), which is a statistic used for comparing similarity and diversity of sample sets was used to measure the extent of shared DMRs between ER+ and ER- tumours. A higher value indicates that a larger proportion of DMRs are shared between the two subtypes, while a lower value indicates subtype-specificity. 3577 hyper DMRs (Jaccard index = 0.43) and 2387 hypo DMRs (Jaccard index = 0.22) were commonly detected across the two breast subtypes (Figure 3.14). For these DMRs, the same fundamental mechanisms of epigenome control are hijacked in breast cancer irrespective of subtype. This was prevalent to a higher degree for gains rather than losses in methylation, indicating that losses in DNA methylation are more subtype-specific. The analysis was repeated for each genomic feature separately. Enhancer DMRs had the lowest Jaccard index among all genomic features indicating that that hypo and hypermethylated regions in enhancers are largely distinct in ER+ and ER- tumours.

Using the differential methylation and enrichment strategies applied above, DMRs were identified for the 6 Intrinsic subtypes (Appendix A.2) and the 11 Integrative clusters as well (Appendix A.3).

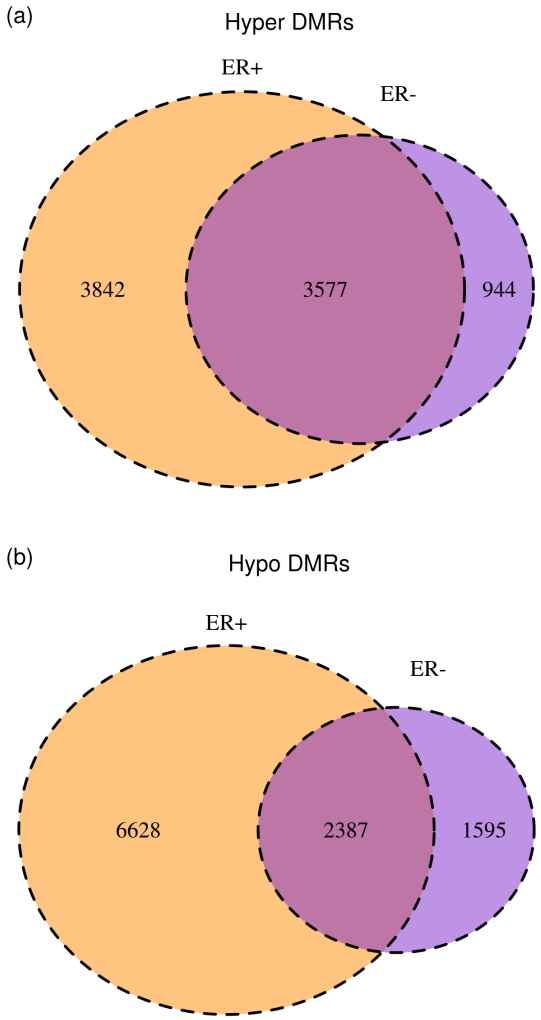


Figure 3.14: Subtype-specific DMRs detected in ER+ and ER- tumours. (a) Venn diagram of hyper DMRs detected in ER+ and ER- tumours. ER+ specific hyper DMRs are represented in orange. ER- specific hyper DMRs are represented in purple. Hyper DMRs shared by the two subtypes are represented in orangish-purple. The number of DMRs belonging to each of the three categories are denoted. **(b)** Same as (a) but for hypo DMRs.

3.6.2 Associations with gene expression

The distribution of the cancer-specific disrupted methylation marks in the genome, such as DMRs, is clearly dependent on the underlying genomic feature as noted in the previous section. However, a key question is whether these DMRs identified in breast cancer are potentially involved in gene silencing or upregulation, and do they differ by genomic feature? The regulatory role of hypermethylation in promoter has been discussed at length in the literature, but recent studies have expanded this investigation to explore the function of altered methylation in the gene body [Kulis et al., 2012] and enhancers [Heyn et al., 2016] in cancer. The detection of both hyper and hypo DMRs, across a variety of genomic features, encompassing both gene-associated and intergenic elements (Section 3.3) has allowed a comprehensive investigation of the regulatory role of altered methylation in ER+ and ER- breast tumours (methods described in Section 3.5.1).

In ER+ tumours, 799 (5.6%) of all expression-DMRS were associated with upregulation of the respective genes and 1599 (11.2%) with repression suggesting that DNA methylation plays a stronger role in silencing genes in this subtype (Figure 3.15a – top panel). Interestingly, stratifying DMRs by orientation of methylation change revealed that both hyper and hypo DMRs were more likely to be associated with repression of gene expression (hyper DMRs: upregulated = 5.4%, downregulated = 11.5%; hypo DMRs: upregulated = 5.6%, downregulated = 10.9%) in ER+ tumours. However, the consequences of hyper or hypo methylation on gene expression were markedly dependent on which genomic feature they occurred. Twenty independent enrichment analyses (enrichment = observed/ expected) were conducted to test these associations formally across hyper and hypo DMRs in the 5 genomic features (Figure 3.15a - bottom panel). Specifically, hypergeometric tests were performed to investigate whether, hyper promoter DMRs, for instance, had a preferential association with upregulation or downregulation of the respective target genes compared to other DMRs. Promoter DMRs exhibit the most glaring directional associations with gene expression. 13.4% hypermethylated promoter were significantly associated with gene repression, and this coordinated pattern was highly enriched (enrichment = 1.20; *FDR p-value* = 5.1×10^{-3}) while only 2.5% hypermethylated promoter were associated with gene upregulation (enrichment = 0.44; *FDR p-value* = 2.7×10^{-8}). Conversely, hypomethylated ones were strongly enriched for gene upregulation (17.1%; enrichment = 3.06; *FDR p-value* = 1.1×10^{-15}). These results demonstrate the repressive role of methylation in promoter in cancer, and confirms conclusions from multiple cancer epigenomics studies [Hovestadt et al., 2014; Teschendorff et al., 2016b].

Hyper or hypo DMRs in exon and intron were not enriched for significant gene expression changes in ER+ tumours. However, filtering for gene body regions that localise with intragenic enhancers led to the noteworthy discovery that these hypomethylated intragenic enhancers were significantly enriched for upregulation of genes (14.4%; enrichment = 2.58; *FDR p-value* = 8.9×10^{-16}). This result is consistent with a report in CLL in which gene-body hypomethylation related with enhancers was revealed as a major cause of aberrant gene stimulation [Kulis et al., 2012]. Further investigations in distal-regulatory elements such as enhancers and PRC region in ER+ tumours, revealed that the associations between methylation and expression were not exclusively neagtive. A considerable fraction of distal-regulatory (both enhancer and PRC region) hyper DMRs and hypo DMRs were significantly associated with gene upregulation as well as repression (Figure 3.15a). The higher proportion of distal regulatory expression-DMRs (15.1%) than intragenic expression-DMRs (7.5%) is partially due to the fact that each enhancer or PRC region is tested with more than one target gene increasing the likelihood of finding a significant correlation. On aggregate, significantly fewer hypo DMRs were detected in enhancers compared to hyper DMRs (Figure 3.13); nevertheless, hypo DMRs were significantly more likely to be associated with gene expression changes (Figure 3.15a). And hypo DMRs were significantly more likely to be associated with upregulation (enrichment = 2.83; *FDR p-value* < 1×10^{-255}), than down regulation (enrichment = 1.65; *FDR p-value* = 4.4×10^{-7}). Collectively these results indicate a strong gene regulatory role for hypomethylated enhancers in ER+ tumours, both within the gene body and further away. A likely mechanism is that reduced methylation at enhancers is associated with increased transcription factor binding which has transcriptional consequences for the target genes [Heyn et al., 2016; Pellacani et al., 2016]. Hyper DMRs in PRC region were the most enriched methylation events observed across the genome in ER+ tumours (Figure 3.15a). Surprisingly, hypermethylated PRC region were not associated with gene silencing, suggesting that these regions reflected the pre-existing repressed chromatin state at silenced genes in the normal tissue from which the cancer originates.

In ER- tumours, approximately equal proportions of all expression-DMRS were associated with upregulation (8.4%) and repression (8.9%) of the respective target genes (Figure 3.15b). This was in sharp juxtaposition to ER+ tumours in which there was a stronger tendency for methylation-associated silencing (gene repression = 11.2%, gene activation = 5.6%). This discrepancy can largely be attributed to the contrasting roles of enhancer DMRs in ER+ and ER- tumours, that have already been confirmed to be largely subtype specific (Section 3.6.1). A significantly higher proportion of hypo

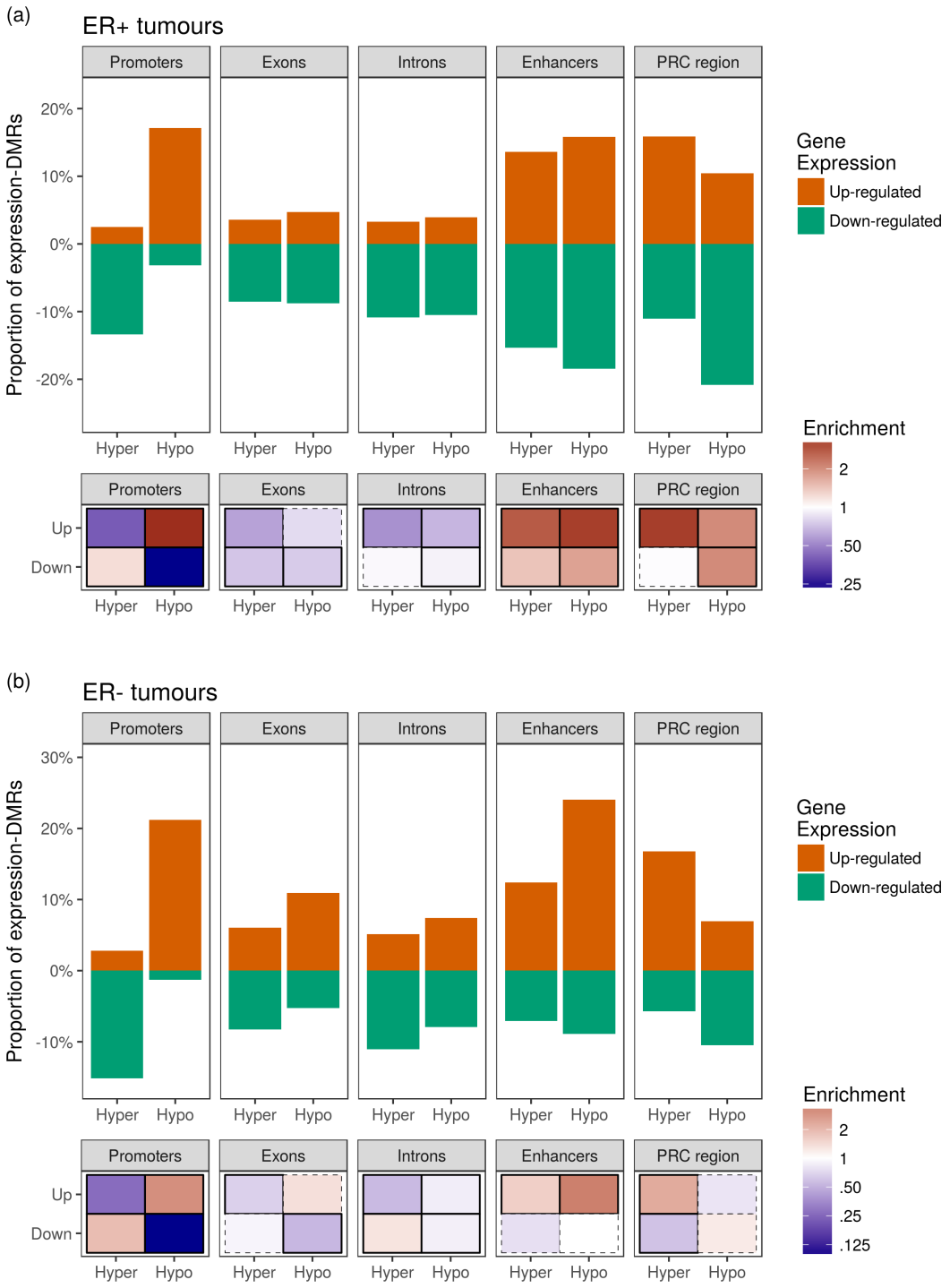


Figure 3.15: (Caption on next page.)

Figure 3.15: (Previous page.) **Expression-DMRs detected in ER+ and ER- tumours. (a) ER+ tumours.** (top panel) The proportion of expression-DMRs detected relative to the total number of direction-specific (hyper or hypo) DMRs identified within a genomic feature. x-axis represents the methylation direction (hyper and hypo) and the five panels represent the individual genomic features. Analyses were conducted separately for the 2 methylation directions and 5 genomic features resulting in 10 bars. Positive bars (orange) on the y-axis represents upregulating expression-DMRs and negative bars (green) represents downregulating expression-DMRs. (bottom panel) Enrichment analysis of up and downregulating expression-DMRs across the 5 genomic features conducted separately for each methylation direction, as explained in the text (hypergeometric test). Top squares represent enrichment of upregulating expression-DMRs and bottom represent enrichment of downregulating expression-DMRs. Colour of the squares represent level of enrichment (observed/ expected). Red represents enriched (enrichment > 1), blue represents depleted (enrichment < 1) and white represents no enrichment (enrichment = 1). Square boundaries represent whether the enrichment was significant (solid lines = *FDR p-value* < 0.05; dotted lines = *FDR p-value* > 0.05). **(b) ER- tumours.** Same as (a) but for ER- tumours.

DMRs in enhancers were correlated with gene upregulation in ER- tumours (24.0%) compared to ER+ tumours (15.8%); and far fewer enhancer DMRs were associated with gene silencing in ER- tumours (7.9%) than ER+ tumours (16.5%). Transcriptional consequences of hyper or hypo methylation localising in other genomic features such as promoter and the gene bodies were largely similar between the two breast cancer subtypes (Figure 3.15).

Next GSEA (pathways tested included Hallmark MSigDB gene sets) were conducted on genes that were significantly dysregulated and had concomitant methylation alterations (expression-DMRs). Pathways that were significantly enriched (enrichment = observed/ expected > 1, *FDR p-value* < 0.05; hypergeometric test) in upregulating and downregulating expression-DMRs represent pathways whose components are potentially epigenetically regulated (upregulated and suppressed respectively). The analysis was conducted separately in ER+ and ER- tumours to classify epigenetically mediated pathways as subtype-specific and shared. Specifically, if a pathway was enriched only in one subtype then it was classified as subtype-specific. If a pathway was enriched in both breast cancer subtypes but the ratio of the enrichment scores for the two subtypes was greater than 1.5, then the pathway was classified as subtype-specific for the subtype with the higher enrichment score. Finally, if a pathway was enriched in both breast cancer subtypes but the ratio of the enrichment scores for the two subtypes was less than 1.5, then the pathway was classified as shared between ER+ and ER- tumours. Enriched pathways that were

upregulated or suppressed and classified as subtype-specific or shared were listed in Figure 3.16.

Figure 3.16: Cancer pathways are epigenetically regulated in a subtype-specific manner. Significantly enriched gene sets comprising of at least 4 genes that are regulated by DMRs (Enrichment (observed/ expected) > 5, *FDR p-value* < 0.05; hypergeometric test, as explained in text) were identified. This was conducted for upregulating expression-DMRs (activated gene sets - top) and downregulating expression-DMRs (suppressed gene sets - bottom). Separate analysis for ER+ and ER- tumours. Enriched gene sets were classified as subtype-specific or shared as described in the text. Colour of points represent direction of regulation. Size of points represent level of enrichment.

G2M checkpoint genes and E2F targets are upregulated and associated with expression-DMRs in both ER+ and ER- tumours, suggesting that DNA methylation is involved in cell cycle activity. For instance, promoter hypomethylation is associated with activation of *WEE1* in ER+ tumours (correlation = 0.42, *FDR p-value* = 3.0×10^{-8}), a nuclear kinase that acts as a mitotic gatekeeper [Nurse, 2004]. Moreover, *WEE1* has also been to shown play a key role in controlling histone synthesis [Mahajan et al., 2012]. This provides further evidence for a direct link between epigenetics and cell cycle progression. Conversely, genes defining epithelial-mesenchymal transition

(EMT) as well as those involved in NOTCH signalling were downregulated in both ER+ and ER- tissues. This is an intriguing discovery since both EMT and NOTCH signalling are associated with breast cancer progression [Reedijk, 2012; Tomaskovic-Crook et al., 2009], and additional work is required to tease out the context in which these pathways are downregulated.

Both early and late oestrogen response stand out as significantly activated pathways in ER+ tumours, and conversely are suppressed in ER- tumours (only early ER response), implying that the most evident phenotypic differences in the two breast cancer subtypes (ER signalling) are strongly linked with epigenetic deregulation. Androgen response was also upregulated in ER+ tumours (associated with only promoter and genebody expression-DMRs; Pathways enriched via expression-DMRs in specific genomic features for ER+ and ER- tumours were listed in Appendix A.4 and Appendix A.5 respectively), which is a fascinating finding since oestrogen gene expression programmes have been shown to be recapitulated by androgen signalling [Robinson et al., 2012]. However, it is uncertain whether ER signalling is a consequence of the aberrant DNA methylation, or if ER signalling converges on and actively shapes the epigenome.

Furthermore, genes belonging to the p53 pathway are downregulated exclusively in ER- tumours. Interestingly, *TP53*, a tumour suppressor involved in cell cycle regulation, DNA damage response and apoptosis [Gasco et al., 2002], has been shown to be the most frequently mutated gene in ER- tumours [Cancer Genome Atlas Network, 2012; Pereira et al., 2016]. This imposes the question whether aberrant methylation in components of the p53 pathway is an alternative pathway to tumorigenesis where there is no p53 gene mutation, and is an avenue for future investigations. Another possibility is that methylation of downstream genes ensued as a consequence of genetically-directed pathway inactivity. In contrast, predominantly ER+ tumours (and not ER- tumours) harbour expression-DMRs associated with downregulation of genes in the TGFβ pathway among others. This corroborates previous findings in the lab that TGFβ had context specific effects on breast tumour initiating cells that were subtype dependent [Bruna et al., 2012] with further work indicating that the methylome plays a central role in moulding the context-specific effect of TGFβ in breast cancer [Tufegdzic Vidakovic et al., 2015]. Collectively, these results strongly suggest that ER+ and ER- tumours require different degrees of epigenomic programming to achieve a breast cancer phenotype.

3.6.3 Subtype specific epigenetic regulators in breast cancer

The previous section revealed specific cellular pathways that are epigenetically regulated in ER+ and ER- tumours. This theme is explored further by explicitly identifying genes with sub-type specific expression-DMRs in ER+ and ER- tumours. Specifically, regions that are differentially methylated compared to the normal tissues *as well as* other tumours (in the same direction) are identified; and if concomitant alterations in expression of the target gene are observed in the tumour subtype compared to normal tissue *as well as* other tumours (in the same direction), then these regions are denoted as subtype specific epigenetic regulators in breast cancer. This strategy led to the identification of potential oncogene and tumour suppressor gene candidates that are specific to either ER+ (Table 3.2, Table 3.3) or ER- breast cancer (Table 3.4, Table 3.5). Survival analysis was also conducted for these genes using a Cox-proportional hazards model adjusted for clinical variables including grade, size, lymph node status, and age at diagnosis. Methylation status of regions that were independently and significantly predictive of BCSS were identified.

A striking example of regional hypomethylation across 10 consecutive CpG sites was identified in one of the intron of *BMPR1B* in ER+ tumours (methylation difference: ER+ vs. normal tissue = -34.8%, ER+ vs. ER- = -35.8%) is shown in Figure 3.17a. The methylation status of the tumours was also significantly associated with expression resulting in higher expression of *BMPR1B* in ER+ tumours (correlation = 0.49, *FDR p-value* = 3.3×10^{-51}). Interestingly, *promoter* associated epigenetic upregulation of *BMPR1B* has been noted in a report integrating the methylomic and transcriptomic architectures in ER+ breast cancer [Gao et al., 2015]. Although in the METABRIC dataset, a ER+ specific expression-DMR was also detected in the promoter of *BMPR1B*, the methylation status of the intron (and not the promoter) was significantly associated with BCSS. Therefore, data presented in this thesis not only confirms the epigenetic upregulation of *BMPR1B*, it also provides evidence of the subtype specificity of this feature. And its value as a prognostic biomarker is revealed for the first time in breast cancer. Furthermore, luminal tumours (that are largely ER+) have recently been proposed to arise from an amplified *BMP2/ BMPR1B* mediated normal response [Chapellier et al., 2015]. This indicates that this methylation event might play an oncogenic role in ER+ tumours. Other examples of subtype specific epigenetic associated upregulation include *SIAH2*, a transcription factor involved in hypoxia and hippo signalling [Adam et al., 2015]; *AFF3*, a mediator of the oncogenic effects of β-catenin signalling [Von Bergh et al., 2002]; *SPDEF* (Sam-pointed domain containing Ets transcription factor), a known transcription factor with a proposed role

Gene	Feature	DMR		Avg. Meth. Difference		Expression		Prognostic
		Direction	Mechanism	vs. N	vs. T	$\|rho\|$	FDR p	
ZNF552	intron	Hypo	Background	-53.8	-47.5	0.45	<0.0001	Strong (Bad)
AGR3	intron	Hypo	Directed	-47.2	-45.6	0.65	<0.0001	No
TBC1D9	intron (Enh)	Hypo	Background	-34.7	-41.4	0.47	<0.0001	Strong (Bad)
CHCHD5	exon	Hypo	Background	-21.5	-39.0	0.50	<0.0001	Weak (Bad)
BAIAP2	intron (Enh)	Hypo	Directed	-24.4	-37.7	0.48	<0.0001	Weak (Bad)
GRTP1	exon	Hypo	Directed	-30.0	-37.6	0.56	<0.0001	No
C1orf64	promoter	Hypo	Directed	-34.9	-36.9	0.59	<0.0001	No
MUC1	exon	Hypo	Directed	-42.6	-36.2	0.50	<0.0001	Weak (Bad)
C3orf52	intron	Hypo	Background	-35.8	-36.2	0.56	<0.0001	Strong (Bad)
BMPR1B	intron	Hypo	Directed	-34.8	-35.8	0.49	<0.0001	Strong (Bad)
SPDEF	intron	Hypo	Directed	-37.0	-34.4	0.67	<0.0001	Strong (Bad)
GATA3	intron	Hypo	Directed	-35.3	-32.8	0.71	<0.0001	No
MSI2	intron (Enh)	Hypo	Background	-32.4	-32.5	0.57	<0.0001	No
EEF1A2	exon	Hyper	Directed	47.5	32.5	0.49	<0.0001	Strong (Bad)
ASB13	intron	Hypo	Directed	-30.8	-31.7	0.49	<0.0001	No
SIAH2	intron	Hypo	Directed	-38.7	-31.0	0.58	<0.0001	Weak (Bad)
MLPH	intron (Enh)	Hypo	Directed	-39.9	-30.2	0.61	<0.0001	Strong (Bad)
AFF3	intron	Hypo	Directed	-31.8	-29.8	0.58	<0.0001	No
NSMCE1	exon	Hypo	Background	-26.7	-28.4	0.47	<0.0001	Strong (Bad)
TFF3	promoter	Hypo	Directed	-24.7	-28.4	0.45	<0.0001	Strong (Bad)
ENTPD5	intron	Hypo	Directed	-26.0	-27.2	0.46	<0.0001	No
TPRN	intron	Hypo	Directed	-29.2	-26.9	0.65	<0.0001	No
ESR1	intron	Hypo	Directed	-27.7	-26.4	0.65	<0.0001	No
C19orf33	promoter	Hypo	Directed	-26.3	-26.0	0.71	<0.0001	No
SCNN1A	intron	Hypo	Directed	-34.0	-25.8	0.44	<0.0001	No
MYO6	intron	Hypo	Directed	-32.2	-25.5	0.45	<0.0001	Strong (Bad)
DEGS2	intron	Hypo	Directed	-36.6	-24.5	0.72	<0.0001	No
PVRL2	intron	Hypo	Directed	-35.7	-24.0	0.56	<0.0001	No
STARD10	intron	Hypo	Directed	-25.5	-23.9	0.56	<0.0001	Strong (Bad)
TFF1	promoter	Hypo	Directed	-26.1	-23.5	0.47	<0.0001	No
RBM47	intron	Hypo	Directed	-31.9	-23.2	0.44	<0.0001	Strong (Bad)
EVL	intron	Hypo	Directed	-24.7	-22.9	0.43	<0.0001	Strong (Bad)
TUBA3D	exon	Hypo	Directed	-31.6	-22.8	0.43	<0.0001	No
SLC2A10	intron	Hypo	Background	-23.1	-22.8	0.43	<0.0001	No
EXOC6	enhancer	Hyper	Directed	33.1	22.7	0.47	<0.0001	Weak (Bad)
SREBF1	intron (Enh)	Hypo	Directed	-33.2	-22.5	0.42	<0.0001	No
KRT8	intron (Enh)	Hypo	Directed	-43.7	-21.6	0.60	<0.0001	Strong (Bad)
HPN	exon	Hypo	Directed	-38.5	-20.4	0.56	<0.0001	No
NAT1	PRC region	Hyper	Directed	26.7	20.2	0.41	<0.0001	Strong (Bad)

Table 3.2: Upregulated genes with subtype-specific expression-DMRs in ER+ tumours. ER+ expression-DMRs that fulfilled three criteria were listed: i) they had a methylation difference of at least 20% vs. normal tissue (*vs. N*); ii) they had a methylation difference of at least 20% versus ER- tumours (*vs. T*); and iii) they were associated with upregulation of the gene: $|rho_{meth\text{-}independent}| > 0.40$ ($|rho|$ in table), and *FDR p-value* < 0.05 (FDR p in table). Analyses were performed as described in the text. For genes with 2 or more expression-DMRs fulfilling the criteria, the region with the largest ER+ vs ER- methylation difference was recorded. Expression-DMRs that were independently predictive of BCSS (multivariable Cox-proportional hazards model as described in the text; *FDR p-value* < 0.1) were identified as strong prognostic biomarkers. DMRs with *FDR p-value* < 0.2 were marked as weak prognostic biomarkers. DMRs that were prognostic were marked as '(Bad)'if they were associated with worse BCSS, and '(Good)'otherwise. Genes listed in decreasing order of ER+ vs. ER- methylation difference. Exonic and intronic DMRs overlapping enhancers are marked with '(Enh)'.

Gene	Feature	DMR		Avg. Meth. Difference		Expression		Prognostic
		Direction	Mechanism	vs. N	vs. T	$\|rho\|$	FDR p	
TBC1D4	intron	Hypo	Directed	-47.7	-41.9	0.58	<0.0001	Weak (Bad)
RSU1	intron	Hypo	Background	-52.2	-41.1	0.65	<0.0001	Strong (Bad)
SIRPA	intron	Hypo	Directed	-46.7	-37.3	0.65	<0.0001	No
PDXK	exon	Hyper	Directed	34.8	36.5	0.47	0.0003	No
PGM1	enhancer	Hyper	Directed	26.3	36.4	0.44	0.0002	No
NCK2	enhancer	Hypo	Background	-50.3	-30.9	0.43	<0.0001	Strong (Bad)
EPHB1	intron	Hypo	Background	-41.7	-27.8	0.53	<0.0001	No
FAM171A1	intron	Hypo	Directed	-39.8	-25.8	0.70	<0.0001	Strong (Bad)
TCF7L1	intron	Hyper	Directed	39.2	25.4	0.58	<0.0001	Strong (Bad)
EGFR	enhancer	Hypo	Background	-38.3	-25.2	0.66	<0.0001	No
FZD9	promoter	Hyper	Directed	27.2	25.1	0.57	<0.0001	Strong (Bad)
MID1	intron	Hypo	Background	-33.9	-25.1	0.68	<0.0001	No
GPM6B	intron	Hypo	Directed	-33.4	-24.4	0.66	<0.0001	No
S100A9	enhancer	Hypo	Directed	-21.4	-24.0	0.43	<0.0001	No
PRKCA	intron	Hypo	Directed	-30.3	-22.6	0.57	<0.0001	Strong (Bad)
CX3CL1	promoter	Hyper	Directed	39.3	21.6	0.68	<0.0001	Strong (Bad)
GPT2	promoter	Hyper	Directed	22.3	20.7	0.55	<0.0001	Strong (Bad)
SFRP1	enhancer	Hypo	Directed	-34.4	-20.2	0.63	<0.0001	No

Table 3.3: Downregulated genes with subtype-specific expression-DMRs in ER+ tumours. ER+ expression-DMRs that fulfilled three criteria were listed: i) they had a methylation difference of at least 20% vs. normal tissue (*vs. N*); ii) they had a methylation difference of at least 20% versus ER- tumours (*vs. T*); and iii) they were associated with downregulation of the gene: $|rho_{meth\text{-}independent}| > 0.40$ ($|rho|$ in table), and *FDR p-value* < 0.05 (FDR p in table). Analyses were performed as described in the text. For genes with 2 or more expression-DMRs fulfilling the criteria, the region with the largest ER+ vs ER- methylation difference was recorded. Expression-DMRs that were independently predictive of BCSS (multivariable Cox-proportional hazards model as described in the text; *FDR p-value* < 0.1) were identified as strong prognostic biomarkers. DMRs with *FDR p-value* < 0.2 were marked as weak prognostic biomarkers. DMRs that were prognostic were marked as '(Bad)'if they were associated with worse BCSS, and '(Good)'otherwise. Genes listed in decreasing order of ER+ vs. ER- methylation difference. Exonic and intronic DMRs overlapping enhancers are marked with '(Enh)'.

Gene	Feature	DMR		Avg. Meth. Difference		Expression		Prognostic
		Direction	Mechanism	vs. N	vs. T	$\|rho\|$	FDR p	
ENO1	intron	Hypo	Background	-31.2	-43.2	0.64	0.0025	No
IDH2	enhancer	Hypo	Directed	-36.7	-41.3	0.69	<0.0001	No
NMI	intron	Hypo	Background	-33.9	-34.0	0.59	<0.0001	No
CLIC3	exon (Enh)	Hypo	Directed	-27.2	-31.1	0.54	0.0051	No
CARD9	promoter	Hypo	Directed	-25.6	-29.2	0.58	0.0021	Weak (Bad)
STIL	intron	Hypo	Directed	-32.2	-28.0	0.71	<0.0001	No
P2RY8	intron	Hypo	Directed	-32.7	-27.9	0.46	0.0006	No
AFG3L2	intron	Hypo	Directed	-24.5	-27.3	0.62	<0.0001	No
SNX8	intron	Hypo	Directed	-24.7	-26.8	0.61	0.0257	No
GPSM2	enhancer	Hypo	Directed	-28.4	-25.8	0.65	0.0069	No
SLC7A5	intron	Hypo	Directed	-47.8	-25.6	0.79	<0.0001	Weak (Bad)
TYMP	enhancer	Hypo	Background	-27.5	-24.3	0.75	<0.0001	No
LAMP3	PRC region	Hyper	Directed	40.4	24.2	0.64	<0.0001	No
PLK1	intron (Enh)	Hypo	Background	-31.3	-24.2	0.69	<0.0001	No
TMC6	exon	Hypo	Directed	-22.1	-23.9	0.54	0.0196	No
HSPA14	enhancer	Hypo	Directed	-20.8	-23.1	0.61	<0.0001	No
TRIP13	PRC region	Hypo	Directed	-27.2	-22.6	0.76	<0.0001	No
NFE2L3	intron (Enh)	Hypo	Directed	-30.6	-22.3	0.76	<0.0001	No
XPO5	promoter	Hypo	Directed	-25.1	-21.9	0.70	<0.0001	No
CSK	promoter	Hypo	Directed	-22.6	-21.9	0.69	0.0236	No
C1orf106	promoter	Hypo	Background	-24.0	-21.6	0.77	<0.0001	No
FAM83D	intron	Hypo	Directed	-24.1	-21.6	0.78	<0.0001	No
GSDMC	exon (Enh)	Hypo	Directed	-22.4	-21.5	0.66	<0.0001	No
TOMM5	intron	Hypo	Directed	-21.0	-21.4	0.40	0.0096	No
EVI2B	intron	Hypo	Directed	-30.6	-20.6	0.65	<0.0001	No
PTPRCAP	promoter	Hypo	Directed	-20.9	-20.2	0.69	<0.0001	No

Table 3.4: Upregulated genes with subtype-specific expression-DMRs in ER- tumours. ER- expression-DMRs that fulfilled three criteria were listed: i) they had a methylation difference of at least 20% vs. normal tissue (*vs. N*); ii) they had a methylation difference of at least 20% versus ER+ tumours (*vs. T*); and iii) they were associated with upregulation of the gene: $|rho_{meth\text{-}independent}| > 0.40$ ($|rho|$ in table), and *FDR p-value* < 0.05 (FDR p in table). Analyses were performed as described in the text. For genes with 2 or more expression-DMRs fulfilling the criteria, the region with the largest ER+ vs ER- methylation difference was recorded. Expression-DMRs that were independently predictive of BCSS (multivariable Cox-proportional hazards model as described in the text; *FDR p-value* < 0.1) were identified as strong prognostic biomarkers. DMRs with *FDR p-value* < 0.2 were marked as weak prognostic biomarkers. DMRs that were prognostic were marked as '(Bad)'if they were associated with worse BCSS, and '(Good)'otherwise. Genes listed in decreasing order of ER+ vs. ER- methylation difference. Exonic and intronic DMRs overlapping enhancers are marked with '(Enh)'.

Gene	Feature	DMR		Avg. Meth. Difference		Expression		Prognostic
		Direction	Mechanism	vs. N	vs. T	|*rho*|	FDR p	
DNALI1	promoter	Hyper	Directed	51.5	45.3	0.75	<0.0001	No
GATA3	exon (Enh)	Hyper	Directed	33.7	37.2	0.47	<0.0001	No
MAST4	promoter	Hyper	Directed	30.5	29.7	0.55	0.0113	No
IRS1	exon	Hyper	Directed	26.4	28.4	0.78	<0.0001	No
PTPRT	intron	Hypo	Directed	-28.3	-25.2	0.52	<0.0001	No
WFS1	intron (Enh)	Hyper	Directed	30.3	23.9	0.59	<0.0001	No
APBB3	enhancer	Hypo	Directed	-28.5	-20.8	0.73	0.0078	Weak (Bad)
ARRB1	intron	Hyper	Directed	22.7	20.4	0.65	0.0003	No

Table 3.5: Downregulated genes with subtype-specific expression-DMRs in ER- tumours. ER- expression-DMRs that fulfilled three criteria were listed: i) they had a methylation difference of at least 20% vs. normal tissue (*vs. N*); ii) they had a methylation difference of at least 20% versus ER+ tumours (*vs. T*); and iii) they were associated with downregulation of the gene: $|rho_{meth\text{-}independent}| > 0.40$ ($|rho|$ in table), and *FDR p-value* < 0.05 (FDR p in table). Analyses were performed as described in the text. For genes with 2 or more expression-DMRs fulfilling the criteria, the region with the largest ER+ vs ER- methylation difference was recorded. Expression-DMRs that were independently predictive of BCSS (multivariable Cox-proportional hazards model as described in the text; *FDR p-value* < 0.1) were identified a strong prognostic biomarkers. DMRs with *FDR p-value* < 0.2 were marked as weak prognostic biomarkers. DMRs that were prognostic were marked as '(Bad)'if they were associated with worse BCSS, and '(Good)'otherwise. Genes listed in decreasing order of ER+ vs. ER- methylation difference. Exonic and intronic DMRs overlapping enhancers are marked with '(Enh)'.

in breast tumour progression [Sood et al., 2009]; and *TFF3* which has been shown to stimulate invasion and angiogenesis in advanced cancer [Ahmed et al., 2012]. The methylation status of the latter two genes also demonstrated significant prognostic potential in breast tumours (Table 3.2).

Promoter hypermethylation and epigenetic inactivation of *SFRP1* and *FZD9* (Figure 3.17b) was also observed in ER+ tumours, but not in ER- tumours. These genes are antagonists of the WNT signalling pathway which has been recognised as an oncogenic regulator of cancer development [Polakis, 2000] . Therefore, the epigenetic silencing of these genes could result in higher activity of the WNT pathway leading to cancer progression [Ohm et al., 2007]. Although epigenetic silencing of these genes has recently been implicated in breast cancer [Gao et al., 2015; Györffy et al., 2016], the results here also demonstrate that this is specific to only ER+ tumours. Furthermore, promoter methylation of *FZD9* was also revealed as a novel prognostic biomarker with potential value in clinical management. Interestingly, hypermethylation of a WNT-inhibitor, *DKK3,* has previously been shown to be associated with the patient's age [Veeck et al., 2009], and Gao et al. [2015] hypothesised that epigenetic silencing of WNT-antagonists including *SFRP1* and *FZD9* might also be a consequence of

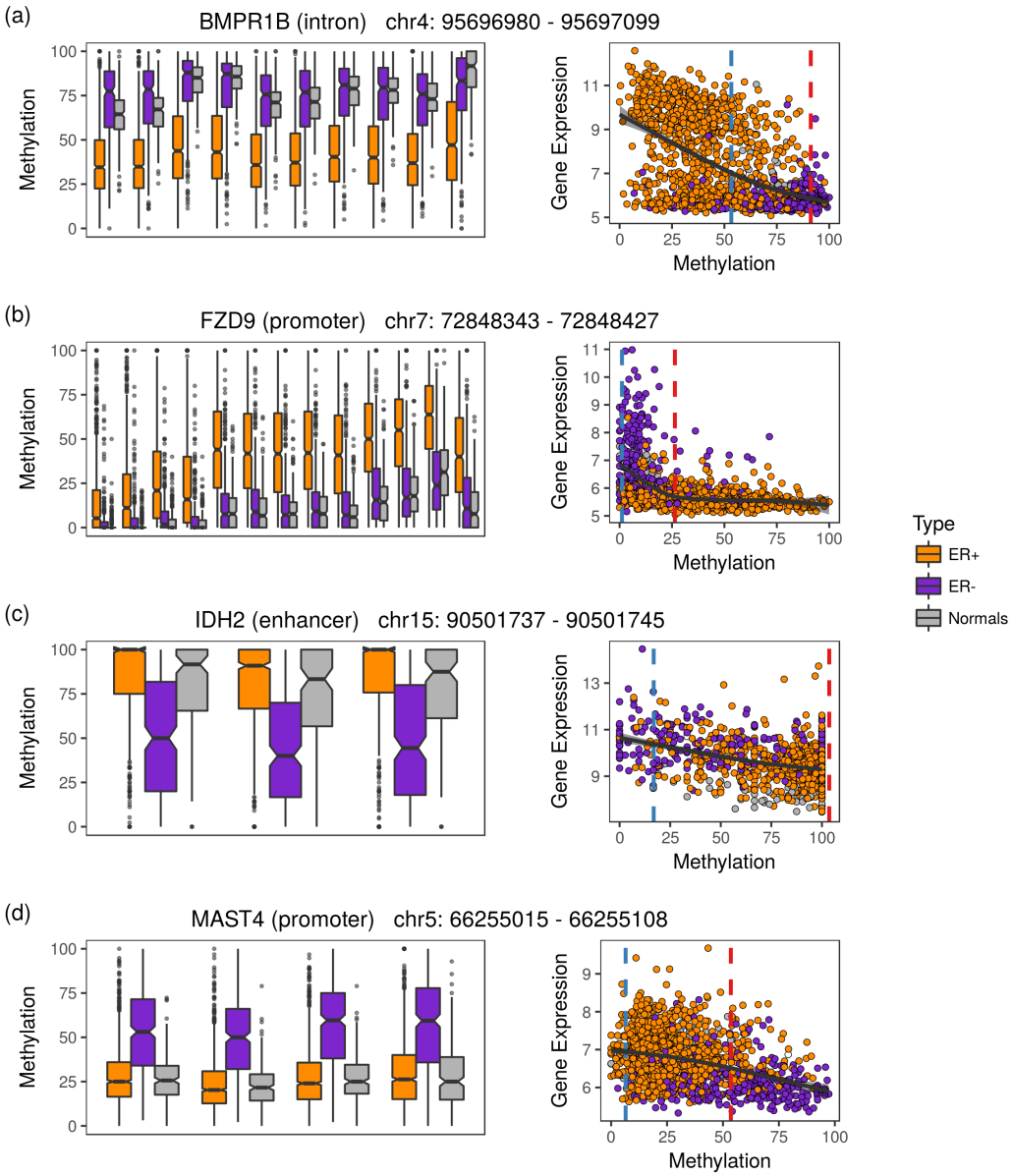


Figure 3.17: Examples of genes with subtype-specific expression-DMRs in ER+ and ER- tumours. (a) ***BMPR1B*** (left) Average methylation levels for ER+ tumours, ER- tumours, and normal tissues across all CpGs at this locus. (right) Scatter plot between DNA methylation and expression levels at this locus across all samples. Black line represents the loess smoothed relationship between methylation and expression. Blue vertical dashed line denotes boundary for hypo MAR definition. Red vertical dashed line denotes boundary for hyper MAR definition. Repeated for **(b)** ***FZD9*** **(c)** ***IDH2*** and **(d)** ***MAST4.***

age-related epigenetic drift initiated early in the pathogenesis of cancer. However, the promoter of both *SFRP1* and *FZD9* exhibited larger methylation departures than expected based on the background model (both were directed alterations; Table 3.3), indicating that although these regions may have accumulated background methylation differences, it is likely that they were selected for.

Remarkably, in ER- tumours, enhancer hypo methylation of *IDH2* was associated with its upregulation (Figure 3.17c). *IDH2* is one the key epigenetic modifiers involved in the TET-mediated DNA demethylation in cancer. Recurrent hotspot *IDH2* mutations have also been found in a rare subtype of breast cancer (Solid papillary carcinoma with reverse polarity) that are largely ER- [Chiang et al., 2016]. Other than this rare subtype, mutations in *IDH2* are rarely detected in breast tumours [Chiang et al., 2016] suggesting that the functional disruption of epigenetic enzymes such as *IDH2* in breast cancer may be caused by a mechanism other than mutations, such as epigenetic events. This suggests that the aberrant methylation observed in ER- cancers may be a consequence of activation of *IDH2* which indeed could be a result of an epigenetic alteration. This potentially may also confound the estimation of background methylation differences in ER- tumours which might explain the dampened association between the accumulation index and proliferation in ER- tumours.

The subtype-specific epigenetic regulation of *GATA3*, a known cofactor in the ER transcription complex, which is activated in ER+ tumours and silenced in ER- tumours offers a fascinating insight into context-dependent behaviour of DNA methylation and is explored in detail in Chapter 4. Promoter hypermethylation associated epigenetic silencing of *MAST4*, a protein kinase, was also observed specifically in ER- tumours (Figure 3.17d).

Subtype specific expression-DMRs were also identified for the 6 Intrinsic subtypes (Appendix B.2) and the 11 Integrative clusters (Appendix B.3).

3.7 Discussion

The aim of genome-wide cancer methylome studies is to identify malignancy-associated epigenetic changes with putative functional roles in gene transcriptional regulation. This would help in better understanding of the mechanisms underlying tumorigenesis, and also in developing rational approaches to therapy stratification. However, a fundamental problem facing cancer methylome studies today and in the future, is that widespread false positive findings might overshadow those epigenetic events involved in transcriptional pathways. This is particularly true for WGBS methylomes, since theoretically every single CpG is profiled. However, despite the fact that methylome profiling technologies such as microarrays (27K, 450K and EPIC) are enriched for functional regions, not all detected methylation changes are directed or selected for. It is likely that a large contribution of the detected cancer-normal differential methylation regions is a consequence of cell-division DNA replication errors that may not contribute to cancer biology. However, this issue is rarely considered, and consequently the relevant mechanism underlying observed methylation changes are not delineated explicitly in genome-wide cancer methylome studies.

In order to shed light on the mechanisms underlying tumour-specific methylation alterations, first, the extent of background or neutral DNA methylation changes in breast tumours (compared to the normal tissues) was quantified. This represents the first characterisation of genome-wide epigenetic drift in a large cohort of primary breast tumours, and was found to be highly context specific with methylation gains largely observed in CpG rich regions and losses in low CpG regions. Moreover, late TOR regions demonstrated a remarkable predisposition for accumulating both methylation changes (gains and losses) in breast cancer. This corroborates similar findings in mutational studies where a marked increase in mutation rates has been observed in late replicating domains of the human genome [Lawrence et al., 2013; Stamatoyannopoulos et al., 2009], suggesting a generalised mechanism involving replication time-dependent damage in both genetic and epigenetic compartments. A likely explanation for these observations is that modifications in late replicating regions lack sufficient time to undergo repair or that the mismatch repair systems might have eroded by this time [Stamatoyannopoulos et al., 2009].

Two indices – *Accumulation* and *Direction* - were introduced to assess the extent and direction of epigenetic drift respectively in the METABRIC cohort. Both indices were highly heterogeneous across all breast tumours and significantly associated with

ER status, as well as molecular subtypes of breast cancer, thus indicating the relationship between genetic, epigenetic and transcriptomic dysregulation in tumorigenesis. Integrating the tumour's accumulation index with its mitotic index as well as gene expression changes confirmed that epigenetic drift is largely a consequence of the accumulation of passive replication related errors related to the number of cell divisions (mitotic clock), as initially hypothesised by Yatabe et al. [2001]. Although this relationship between accumulation of epigenetic drift and mitotic clock was observed in both ER+ and ER- tumours, it was much stronger in ER+ tumours in which it was significantly prognostic of lower BCSS. The diminished relationship in ER- patients raises the interesting postulation that a considerable proportion of epigenetic drift accumulated in ER- tumours is acquired due to environmental exposures and not cell-division errors, and will be examined in future work.

Next, functional regions of the methylome were assessed to identify focal differential methylation regions (DMRs) and methylation altered regions (MARs) by comparing tumours and normal tissues. The key distinction between these two alterations is that DMRs are regions of averaged methylation differences between a class of tumours and a group of normal tissues, while MARs are tumour-specific. DMRs can also represent regions of altered methylation between two fixed groups of tumours, and this is discussed in Section 3.6.

A crucial enhancement to the detection of both DMRs and MARs was the implementation of the novel DMARC algorithm that has been developed as part of this thesis. Traditional cancer methylation analyses utilise a uniform methylation difference threshold across the genome and for all tumours. However, the epigenetic drift analysis revealed extraordinary variation in background methylation differences within breast tumours; and in background methylation differences across the genome, which was strongly correlated with CpG density and DNA replication timing. By incorporating background methylation heterogeneity into the methylation analyses, DMARC is able to identify *directed* methylation alteration regions that are a consequence of instructive regulatory pathways versus *background* methylation alteration that are untargeted and largely associated with cell division errors. The underlying premise of DMARC is similar to the MutSigCV algorithm for the identification of mutation driver genes in cancer which is based on the assumption that cancer driver genes harbour more mutations to that expected for an observed background rate [Lawrence et al., 2013]. Accordingly, DMARC enabled the explicit

delineation of the mechanism underlying methylation changes, be it tumour-specific MARs or class-specific DMRs.

As expected, directed-DMRs were significantly enriched for gene expression changes in breast cancer compared to background-DMRs. This strongly suggests that background-DMRs, that have been largely demonstrated to be a consequence of a stochastic accumulation of cell-division errors are likely to have fewer functional consequences than directed-DMRs. More stringent thresholds (based on the background estimates) were used to detect directed-DMRs, and this partially contributes to the observed enrichment. However, this cannot explain the noteworthy separation of the biological pathways disrupted at the transcriptional level by the two distinct methylation mechanisms. Specifically, genes that harboured background-DMRs as well as concomitant expression changes were more likely to be involved in pathways elicited by specific cell-extrinsic stresses such as immune system engagement and hypoxia. Conversely, genes with directed-DMRs were enriched for altering cell-intrinsic biological signalling pathways such as ER and NOTCH signalling. Thus, DMARC enables a better understanding of methylation processes and its role in disease aetiology. However, it is important to note that bioinformatic approaches go only so far in highlighting important genes and pathways that might be drivers in pathogenesis. Functional studies are required to validate the relevance and mechanism underlying the epigenetic events described here.

Although discrimination between directed and background DMRs has proved extremely valuable for biological interpretation, it would be erroneous to ignore stochastic DMRs and assume they are functionally irrelevant for two reasons. Firstly, these stochastic methylation alterations may occasionally hit key genes in cancer-relevant pathways and be under selection and be identified as directed. Examining methylation calls aggregated over all cells in the tumour does not allow discrimination based on selection-related methylation dynamics. However, evaluating methylation content based on epialleles inferred from single reads can be used to distinguish methylation alterations that occur in a noisy and stochastic fashion versus those that are occur in a deterministic manner. This paves the way for a comprehensive analysis of the role of intratumour methylation heterogeneity in tumour evolution which is conducted in Chapter 5. Secondly, although the extent of background methylation alterations in a tumour is undoubtedly associated with CpG density and DNA replication timing, these two genomic variables are not absolute predictors. Observed methylation differences (larger than 20%) would always be detected as MARs, however the DMARC algorithm may suffer from type 1 or type 2 errors in classifying directed from background

methylation alterations. Shipony et al. [2014] demonstrated that gain and loss of methylation in somatic cells is also associated with nucleosome occupancy and nuclear laminal interaction in addition to replication timing. It is likely that integrating these additional genomic variables into the DMARC algorithm will improve performance.

A caveat in the tumour-normal differential analyses (both DNA methylation and gene expression) is that the comparisons are conducted against the normal tissue that is found adjacent to the breast tumour tissue. Teschendorff et al. [2016a] established that the normal-adjacent breast tissue harbour epigenetic field defects (largely stochastic age-associated methylation alterations), and consequently, this may reduce the sensitivity to detect potential regulatory methylation events. However, the authors also demonstrated that breast tumours exhibit larger methylation differences on comparison with normal-adjacent tissues at the loci harbouring field defects, and so this strategy is still appropriate. Moreover, this also raises another advantage of DMARC since it can discriminate between stochastic age-related methylation alterations and putative directed events.

Another key benefit of the DMARC algorithm is that it provides a novel way to account for tumour purity. A critical issue in cancer methylome studies is that tumour tissues are highly heterogeneous that suffer from contamination from adjacent normal cells. Consequently, methylation profiles obtained from these tumour tissues are in fact mixed signals from tumour and normal components depending on tumour purity [Jaffe and Irizarry, 2014; Zheng et al., 2014]. Although tackling intra-sample cell heterogeneity has been studied in the context of epigenome-wide association studies (EWAS) [Houseman et al., 2014; Jaffe and Irizarry, 2014; Rahmani et al., 2016; Teschendorff et al., 2017; Zheng et al., 2017a], differential methylation analysis between tumour and normal tissues with the consideration of tumour purity represents a related but fundamentally distinct challenge and is understudied [Wang et al., 2016]. Consequently, statistical methods for differential methylation analysis implemented in a vast majority of cancer methylome studies do not account for tumour purity [Akalin et al., 2012b; Aryee et al., 2014; Assenov et al., 2014] which leads to biased and often erroneous results [Wang et al., 2016; Zheng et al., 2017b]. For instance, let us assume two tumours – A and B – with Tumour A having a higher tumour purity than tumour B. The epigenetic drift analysis revealed that the accumulation index (magnitude of background methylation alterations) in a tumour was significantly associated with ASCAT-defined tumour purity estimates ($p\text{-}value = 7.6 \times 10^{-23}$). Consequently, for the same genomic loci, tumour A would also harbour a higher absolute background methylation difference (compared to normal tissues) than tumour B on account of its

higher tumour purity. Using uniform methylation difference thresholds as per traditional differential methylation analysis would erroneously result in the detection of a larger number of MARs for Tumour A. Conversely, by incorporating background estimates into constructing tumour-specific methylation difference thresholds, DMARC is able to account for purity-based biases. Although there a few methods available to estimate tumour purity from DNA methylation content, only a few recent reports have been able to use this information to deconvolute the methylation profiles of the tumour analysis for downstream differential methylation analysis [Barrett et al., 2017; Zheng et al., 2017b, 2014]. Interestingly, DMARC does not require explicit estimates of tumour purity as an input, although the accumulation index is a good indicator of it, nor does it utilise the epiallelic content obtained from bisulphite sequencing techniques and thus can be used for microarray techniques as well.

DMARC is suitable for all high-throughput genome-wide cancer methylome studies, and in principle, it can be applied across all popular methylation profiling techniques that provide single CpG resolution such as microarrays and bisulphite sequencing techniques. The benefits of utilising DMARC in cancer methylome analysis are clearly outlined above, however, it is important to note, that it relies extensively on accurate estimations of tumour-specific background methylation differences in distinct genomic contexts. Since theoretically every single CpG is profiled in WGBS, it is obvious that the implementation of DMARC in WGBS cancer methylomes is not only highly recommended to deconvolute directed and background methylation alterations, but the algorithm will also benefit from superior resolution at all genomic contexts. Although microarray technologies are biased towards functionally relevant and *interesting* CpG sites, approximately 50% of the probes localise in the background or neutral compartment of the epigenome (as defined in Section 3.2) which points to their suitability in the implementation of DMARC. Specifically, HM450 and EPIC include 215K CpG sites and 452 K CpG sites respectively as part of the *background CpG universe*. RRBS is even more suitable as demonstrated in this study, with approximately 1816 K background CpG sites assayed. Moreover, background CpG sites across all above-mentioned methylation platforms have adequate distributions of DNA replication timing that are similar to the whole epigenome. However, a limitation of microarray technologies is that both HM450 and EPIC suffer from inadequate representation of background CpG sites at CpG-rich regions (HM450 = 13 K, EPIC = 13 K, EPIC = 478 K; CpG Density $\geq$ 80 CPGs/kbp). Additionally, despite being depleted for CpG-low regions, the RRBS epigenome contains a larger number of background CpG sites at CpG-poor regions compared to microarray technologies (HM450 = 141 K, EPIC = 376 K, RRBS = 849 K; CpG

Density < 40 CpGs/kbp). Consequently, DMARC may not perform as accurately in microarray technologies, and future work will assess their utility in pan-cancer microarray-based methylomes publically available from TCGA [Bass et al., 2014; Cancer Genome Atlas Network, 2012; Cancer Genome Atlas Research Network, 2016; Collisson et al., 2014; Hammerman et al., 2012; Muzny et al., 2012; The Cancer Genome Atlas Research Network, 2013a; Weinstein et al., 2014].

Despite multiple studies attempting to map cancer-associated epigenetic changes at the genome-wide level, one of the shortcomings of previous breast cancer methylome analyses [Cancer Genome Atlas Network, 2012; Dedeurwaerder et al., 2011; Gao et al., 2015; Roessler et al., 2015] is that the focus has largely been retained on the promoter regions due to their established significance as an epigenetic transcriptional repressor. Although recent reports in breast cancer have progressed to investigating gene body methylation as well [Fleischer et al., 2014; Györffy et al., 2016], these studies restricted their analysis to a selected panel of genes. There is also noteworthy absence in the examination of the epigenetic roles of intergenic distal regulatory regions such as enhancers and PRC regions in breast cancer. In the current project, the extent of DNA methylation alterations in breast cancer was assessed in a variety of genomic features, encompassing both gene-associated elements as well as distal-regulatory elements. Hyper as well as hypo DMRs in all the genomic features were detected in this breast cancer methylome study. Notably, these epigenetic differences were not only incriminated with gene silencing, but in fact a large number of DMRs were implicated with the upregulation of potential oncogenes as well. Collectively, the results in this chapter strongly indicate that the activity of major signalling pathways in breast cancer are at least partly regulated by the epigenome, and more so, by epigenetic deregulation of not only promoter but also other genomic elements.

The importance of breast cancer stratification is highlighted in Section 3.6, where ER status was initially used for classification. Some genes such as *GATA3* and *ESR1* are specifically linked to oestrogen signalling and so may be relevant only in an ER+ context. Conversely, various genes including *IDH2* and *MAST4* were identified as subtype specific epigenetic regulators in ER- cancers, and the methylation statuses were also associated with worse prognosis. GSEA also revealed distinct pathways that were epigenetically disrupted in ER+ and ER tumours, such as oestrogen-signalling in ER+ tumours and the p53 pathway in ER- tumours. Interestingly, DMRs within enhancers were largely identified as subtype specific, and hypomethylated enhancers (both within the gene body and further away) in particular, were demonstrated to have a strong gene regulatory role. A likely mechanism is that reduced methylation at enhancers

is associated with increased transcription factor binding which has transcriptional consequences for the target genes. This has been demonstrated in recent studies across different cancers [Heyn et al., 2016] and in normal mammary epigenomes [Pellacani et al., 2016]. ER binding sites are also available from ChIP-Seq experiments conducted on these tumours [Ross-Innes et al., 2012], and one of the follow-up goals of this project is to correlate enhancer DMRs in ER+ tumours with differential ER binding, as was performed in a recent report studying ER+ tumours [Stone et al., 2015]. Furthermore, motif analysis can be implemented to map putative binding sites of various transcription factors to activating hypo enhancer DMRs in ER+ and ER- breast tumours. This would enable the construction of subtype-specific transcription factor regulatory networks in breast cancer.

The results presented in this chapter also strongly suggested that the epigenome contributes to defining these subtypes as distinct biological entities that affect the clinical evolution of the disease. However, these subtype-specific methylation patterns might be associated with the epigenetic imprints of presumed cellular origins. Recent reports have explored epigenomic profiles in different cell populations from normal mammary glands [Gascard et al., 2015; Pellacani et al., 2016] and revealed striking differences in the degree of epigenomic reprogramming between them. Future investigations post this thesis are proposed to integrate the METABRIC breast cancer methylomes with those from the normal mammary subpopulations obtained from the Roadmap Epigenomics Project [Kundaje et al., 2015] to shed light on epigenomic reprogramming in normal mammary development and leading to the initiation of breast cancer.

The DMR and MAR analysis are implemented on SCCRUB – the universe of spatially coordinated regions defined in Chapter 2. SCCRUB consists of approximately 289K focal regions comprising of 4 CpG sites on average with an average width of 88bp (see Chapter 2), and this is largely driven by *Msp1* sites, the restriction enzyme used in RRBS. Although differential methylation analyses on SCCRUB regions can identify focal hypomethylated regions, an alternate approach needs to be employed to detect long-range methylation alterations such as the recently identified hypomethylated blocks in various cancers [Berman et al., 2011; Hansen et al., 2011; Hovestadt et al., 2014]. Future work will identify whether multiple hypo DMRs (from within the SCCRUB universe) cluster together in specific regions of the genome. This will be followed by an investigation into whether these epigenetic domains correspond to large chromatin regions (LOCKs), nuclear organisation (LADs), or DNA repeats and subtelomeric regions, and whether they have consequences on genome instability.

Genetic events might have a dominant role in shaping the epigenome by affecting cellular components that participate in epigenetic pathways (see Chapter 1). Although the widely recognised epigenetic modifiers such as *DNMT3A*, *IDH1* and *TET1* are rarely mutated in breast cancer, *BRCA1* mutations have been associated with global DNA hypomethylation [Flanagan et al., 2010]. Conversely, *BRCA2* mutations were linked with promoter hypermethylation [Holm et al., 2010]. While the reasons for this remain unknown, this suggests that besides their recognised functions in DNA repair [Joosse, 2012], BRCA family of proteins also participate in regulation of DNA methylation. A recent computational report using the TCGA consortium breast cancer data identified three epigenetic modifier genes – *UHRF1*, *WHSC1* and *CBX7* – that were universally deregulated across different cancers including breast cancer and associated with concomitant disruption in global DNA methylation levels [Yang et al., 2015]. One of the follow-up goals of this project is to integrate the mutational, transcriptomic and epigenomic data in METABRIC to confirm these findings and determine additional epigenetic enzymes in breast cancer.

The results in this chapter support the view that altered DNA methylation is associated with the disruption of key transcriptional pathways in breast cancer. Although it is possible that the observed methylation alterations instigate concomitant silencing or upregulation of the gene in question, recent studies have also demonstrated how the presence or absence of an upstream transcription factor (through genomic events) can mediate active DNA methylation changes at promoter and distal regulatory elements [Domcke et al., 2015; Feldmann et al., 2013; Yin et al., 2017]. This study provides a snapshot of the expression and methylation status of the breast tumour at the time of surgery. And consequently, the associative statistical models presented here cannot distinguish between a causative role for DNA methylation and an effects model, in which the observed methylation modification is a consequence of an upstream gene deregulation event [Teschendorff et al., 2016b]. However, even in the scenario where DNA methylation alterations follow gene deregulation, it has been shown to play a critical role in *locking* this transcriptional state [Bird, 2002; Lock et al., 1987; Siegfried and Cedar, 1997]. Therefore, differential DNA methylation marks do not only reflect altered transcription factor activity [Fleischer et al., 2017; Schübeler, 2015], but as demonstrated in this chapter, can also serve as powerful biomarkers for breast cancer subtype-specific diagnosis and prognosis. Collectively, these results have confirmed that DNA methylation alterations are an extremely informative indicator of the epigenetic disruption of transcriptional pathways in cancer, and they contribute significantly to breast cancer heterogeneity.

Chapter 4

Integration of DNA methylation alterations with genomic events

Contents

4.1 Introduction

Genetic alterations such as mutations, copy number changes and epigenetic alterations such as DNA methylation events represent distinct mechanisms by which gene function in cancer can be deregulated [Cancer Genome Atlas Network, 2012; Ciriello et al., 2013; Curtis et al., 2012; Vogelstein et al., 2013]. Genomic and epigenomic profiling of these alterations have led to the identification of a number of genes that contribute to tumour progression, but they have mostly been studied in isolation from each other. Since all these types of events possess gene-regulatory function, it is clear that there must be an interaction between them during malignant transformation [Cancer Genome Atlas Network, 2012; Shen and Laird, 2013]. The development of integrative approaches has allowed discovery of joint patterns across multiple data types such as DNA methylation, expression and copy number events (iCluster [Shen et al., 2009], moCluster [Meng et al., 2016], consensus clustering (cluster of clusters) [Lancichinetti et al., 2009]). However, these methods, involve dimensionality reduction across multiple molecular landscapes with the primary objective of understanding the taxonomy of disease; and are not concerned with the identification of the distinct mechanism (or synergy of multiple mechanisms) associated with deregulation of cancer. Curtis et al. [2012] successfully bridged the gap between these two objectives by first identifying 1000 genes in which CNAs influenced expression in *cis* in breast cancer; and subsequently conducting integrative clustering on this set of genes. This analysis revealed 10 (now 11) novel tumour subgroups (the Integrative clusters; IntClusts) with distinct clinical features and prognosis. However, the DNA methylation landscape has not been profiled in this METABRIC dataset until now.

In the same year as the METABRIC publication, The Cancer Genome Atlas (TCGA) presented results for approximately 500 patients with data available on DNA copy number, DNA methylation, exome, mRNA, microRNA and protein profiles [Cancer Genome Atlas Network, 2012]. Approximately, 300 breast tumours were added to the consortium in 2015 resulting in a total of ∼800 breast tumours [Ciriello et al., 2015]. This represented the first large multiplatform analysis with epigenetic and genetic data in breast cancer, and was followed by an ER+ breast cancer study (n = 560) conducted by the International Cancer Genome Consortium (ICGC) in 2015 [Nik-Zainal et al., 2016]. Although, the breast cancer methylome was not explored in depth in either of these two studies, the public availability of this data, in particular TCGA, has sparked an interest in integrating data across multiple molecular frameworks to identify epigenetic drivers of tumorigenesis. Teschendorff et al. [2016b] conducted a detailed investigation of the multi-omic basis of transcription factor dysregulation

in cancer using pan-cancer data obtained from TCGA. Excitingly, they revealed that promoter hypermethylation is a more frequent event than copy number losses or inactivating mutations, and is the more likely mechanism associated with silencing of transcription factors in cancer. However, only methylation of the gene promoter was evaluated, and mechanisms underlying the activation of genes was not investigated; and notably, breast cancer samples were not included in the study. About the same time, another pan-cancer report using TCGA data (this time including breast cancer) revealed that DNA methylation alterations prefer to target genes in the extracellular and transmembrane domains (enriched for cytokines and growth factors, cell differentiation markers) rather than intracellular domains (enriched for transcription factors) [Gao and Teschendorff, 2017]. However, this study too, only focused on promoter methylation events negatively correlated with expression; and there is a pressing need to expand cancer methylome analyses beyond this traditional outlook.

The breast cancer landscape has previously been shown to be dominated by copy number events [Ciriello et al., 2013]. Although, the landmark study [Curtis et al., 2012] mentioned above also revealed that the breast cancer transcriptome is largely influenced by acquired somatic CNAs with the identification of 1000 *cis*-acting CNA genes, the variation in expression for approximately 60% of genes across the genome was unaccounted for by CNA. Could a significant contribution of this unaccounted variation in the transcriptional deregulation be explained by DNA methylation? A recent demonstration that this was in fact the case in other cancers [Teschendorff et al., 2016b] suggests that this is a promising hypothesis; although, as mentioned above, only inactivation of transcription factors was evaluated. In this chapter, the independent contributions of *cis*-association of CNA and of DNA methylation in explaining gene expression were compared in order to identify genes that were regulated by CNA, DNA methylation or both alterations. Moreover, the contributions of CNA and DNA methylation alterations in explaining inter patient breast cancer heterogeneity (such as between ER+ and ER- tumours) were also compared.

Many cancer genes exhibit contradicting expression profiles to what is expected given the underlying copy number alterations. For instance, tumours with copy number gains do not show an anticipated upregulation of the underlying gene. The large number of genes showcasing this abnormality suggests that this is not a technical artefact. And so, the hypothesis that an alternate mechanism such as DNA methylation could be used to target specific genes in order to modulate the regulatory role of *cis* CNA events is proposed and investigated.

Finally, the prevalence of methylation altered regions (MARs) and of CNAs were compared within genes that were differentially expressed in tumours compared to normal tissues to identify which alteration accounts for a higher fraction of tumours in these potentially functional genes. This section of analysis helps to support results from the above comparison of the regulatory roles of DNA methylation and CNA, but critically also allows investigation of patterns of co-occurrence and mutual exclusivity between these two distinct mechanisms for the same gene. A biologically plausible rationale for finding two independent co-occurring events is that they specifically augment each other's tumour suppressive or oncogenic roles [Knudson, 1971]. In this case, the second alteration provides a selective advantage for the cancer cell(s) and will consequently be observed more frequently than expected by chance. A case in point is loss of heterozygosity (LOH), a common event in tumour suppressor genes in cancer (e.g. *RB1* in retinoblastoma, *BRCA1* in breast cancer), in which a mutation or an epigenetic hit is accompanied by hemizygous deletion of the non-mutant allele ensuring that only the function of the mutant allele is observed [Cavenee et al., 1983; Esteller, 2000; Kawaoi et al., 1992; Merajver et al., 1995]. On the other hand, a pattern of two mutually exclusive events on the same gene could be an indication that i) both alterations have the same functional consequence. For instance, loss of an allele would give the same effect as a dominant negative mutation; ii) the second alteration offers no further selective advantage than the first hit; or iii) the second alteration leads to a disadvantage for the cell, eventually leading to cell death. Consequently, they will be observed less frequently than expected by chance. Identification of methylation alterations among such patterns of independent cancer mechanisms on the same gene helps to delineate the functional role of DNA methylation in tumorigenesis, but also supports the discovery of novel tumour suppressor genes and oncogenes and validation of previously established ones in breast cancer.

4.1.1 Summary of aims

In the previous chapter, potential functional methylation alterations were identified using a combination of strategies including accounting for tumour-specific and context-specific background rates; and subsequently characterising these events using concomitant changes in gene expression. This chapter extends these findings, and aims to characterise the regulatory role of methylation patterns and explore the molecular multi-omic landscape of *cis* gene deregulation in breast cancer. This is achieved through the following steps:

1. Quantification and comparison of the independent contributions of DNA methylation and CNA alterations to identify the predominant mechanism driving deregulation of gene expression in breast cancer.

2. Investigation of whether methylation events can modify the effect of copy number amplifications and copy number losses in a gene.

3. Calculation of the comparative prevalence of MARs and CNAs in differentially expressed genes in breast cancers. Identification of MARs that co-occur or are mutually exclusive with CNA events within the same gene, and explore the regulatory consequences of these configurations.

4.2 Identification of the principal functional methylation region (PFMR) of a gene

The purpose of this chapter is to identify the relevant mechanism – Copy number, mutations, or methylation - that is predominantly associated with *cis* activation or silencing of genes in breast cancer. Although definite estimates of CNA, gene expression and mutation (only 173 genes) information are available per gene (see Chapter 2), DNA methylation changes are widespread and variable over multiple regions within the gene body as well as distal *cis*-regulatory elements (Chapter 3). For example, *GATA3* has 3 promoter regions, 4 exons, 3 introns, 9 introns, 3 enhancers and 9 polycomb-repressed complex marked (PRC) regions assayed as part of the SCCRUB universe. For the purpose of identifying the principal functional methylation region (PFMR) of the genes, the SCCRUB region associated with the gene having the highest correlation with expression was selected.

To serve the dual objectives of i) evaluating the independent roles of DNA methylation and of CNA in explaining gene expression; and ii) making an informed choice for the PFMR per gene, a multivariable regression framework was constructed. This is similar to expression analysis described in Chapter 3, with the fundamental difference being that all regions within the SCCRUB universe were considered, and not just those identified as DMRs. Each gene may be associated with multiple methylation regions but only single CNA and gene expression estimates are available, and accordingly multiple tests were performed for each gene using distinct methylation inputs against the same CNA and expression inputs. For each region-gene combination, gene expression (normalised $\log_2$ intensities, see Chapter 2) was treated as the independent variable and both DNA methylation (M-value, see Chapter 2) and CNA (segmented mean $\log_2$ ratios, see Chapter 2) as continuous covariates. The inclusion of both DNA methylation and CNA in the model ensures that their independent contribution to gene expression can be determined for each gene [Gao and Teschendorff, 2017]. Tumours that had matched gene expression (microarray, see Chapter 2), CNA (SNP 6.0 array, see Chapter 2) and DNA methylation data were considered for this analysis. A total of 15968 genes with gene expression (microarray; Chapter 2), CNA (array) and DNA methylation data available were used.

The analysis in Chapter 3 examined associations from a DMR and genomic feature (such as promoters, exons etc) standpoint. In this section, a gene-central perspective is considered. On average, each gene was associated with 24.6 (SD = 23.2) distinct regions of coordinated clusters of CpG sites as defined by the SCCRUB

universe (2.3 regions within promoters, 1.4 region within exon, 5.9 regions within introns, 11.5 regions within enhancers and 3.5 regions within PRC regions; Figure 4.1a - left). For every gene individually, the partial correlations between methylation and expression were recorded for each of its multiple regions ($r_{meth\text{-}independent}$, which is a measure of the independent strength and direction of this relationship whilst controlling for the effect of CNA). A key question is whether methylation on intragenic regions (including promoters) are more likely to be associated with the expression of the underlying gene than distal regulatory elements? To test this hypothesis, the region with the highest $r_{meth\text{-}independent}$ was noted for each gene, and a hypergeometric test was conducted to compare observed vs. expected proportions for each feature. All three intragenic features (promoters, exons and introns) were significantly more likely to contain the region in which methylation most correlated with expression; whereas enhancers and PRC regions were significantly depleted for the strongest methylation correlation for the gene (Figure 4.1a – middle and right). It is true, however, that these regions were selected from within the SCCRUB universe and not the entire genome, which means that the formal comparison of the regulatory roles of different genomic features cannot be performed. Nevertheless, promoters displayed the highest enrichment (hypergeometric test: enrichment = 1.69; *FDR p-value* $< 1\times10^{-255}$) which strongly confirms previous findings that promoters have the most impact on gene expression [Jiao et al., 2014].

Although Chapter 3 undoubtedly illustrates that distal *cis*-regulatory features such as enhancers and PRC regions are epigenetically altered in breast cancers and that they also have significant impact on gene expression, it is challenging to prove these associations without using chromosome conformation capture assays such as Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) [Fullwood et al., 2009] or Hi-C [Lieberman-Aiden et al., 2009] to identify distal *cis*-interactions with promoters. This, along with the observation that promoter and gene body regions are more likely to play the strongest methylation-specific regulatory roles for the underlying gene (Figure 4.1), led to the decision to focus further investigation in this chapter exclusively on promoters, exons and introns. However, this does not diminish the importance of understanding the behaviour of distal regulatory methylation alterations in the presence or absence of other genomic events; they have been ignored purely for practical purposes.

For each gene, the region (within the three intragenic features) with the strongest regulatory role (highest $r_{meth\text{-}independent}$) was selected as a functional region, and its methylation estimate was used as the representative methylation value for the gene.

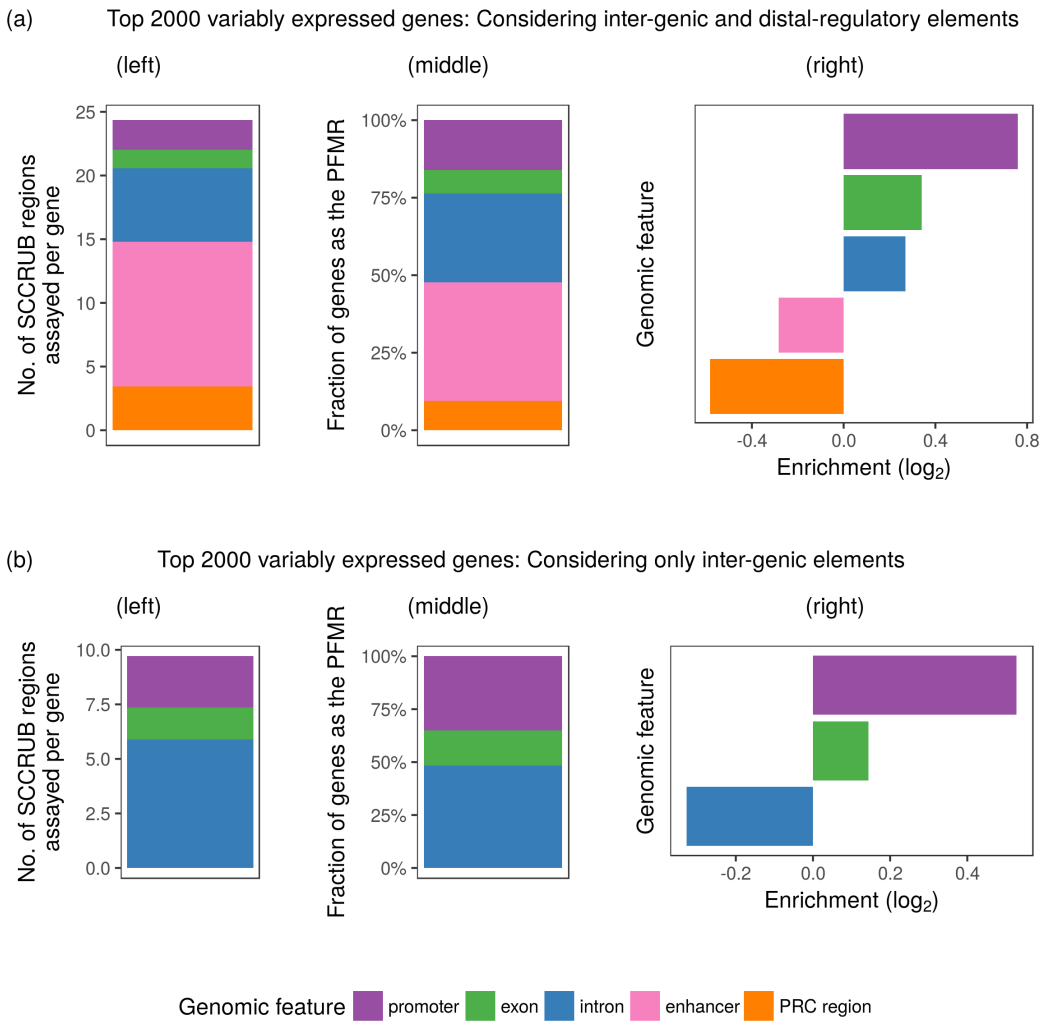


Figure 4.1: Promoters are likely to harbour the PFMR of a gene. (a) Considering intergenic and distal regulatory elements. (left) Average number of SCCRUB regions detected per gene per genomic feature. (middle) Fraction of genes that have a specific genomic feature as the principal functional methylation region. (right) Enrichment ($\log_2$) of a genomic feature to be detected as the principal functional methylation region of a gene. **(b)** Considering only inter-genic elements. same as described for (a). PFMR = principal functional methylation region. The top 2000 variably expressed genes (VAR2000) were considered.

Promoters represent the strongest methylation-specific regulatory region for 35% of the genes; exons for 17% of the genes; and introns for 48% of the genes (Figure 4.1b). Given that there are twice as many intron regions covered in SCCRUB than promoter regions (1.2 million vs. 0.6 million), promoter regions are the most enriched (enrichment = 1.44, hypergeometric test: $p\text{-}value < 1 \times 10^{-255}$; Figure 4.1f).

4.3 *Cis*-acting DNA methylation and CNA regulate the breast cancer transcriptome

4.3.1 Inter-patient heterogeneity in breast cancer

Multiple reports have shown that the breast cancer transcriptome is largely influenced by *cis*-acting CNAs [Ciriello et al., 2013; Curtis et al., 2012]. In this section, the contribution of *cis*-acting DNA methylation in explaining gene expression is quantified and compared to that for *cis*-acting CNA. Specifically, for each gene the partial correlation between methylation and expression ($r_{meth\text{-}independent}$, a measure of the strength of the independent contributions of methylation on expression) and between CNA and expression ($r_{cna\text{-}independent}$, a measure of the strength of the independent contributions of methylation on expression) are compared. At the outset, focus was set on the 2000 most variably expressed genes in breast cancer in this cohort (this gene set will be henceforth denoted as VAR2000), since this represents a straightforward strategy to enrich for genes likely to explain the considerably inter-tumour heterogeneity in breast cancer.

Figure 4.2a demonstrates the comparison of the *cis*-regulatory roles of methylation and CNA in the VAR2000 gene set. A stringent threshold of 0.40 was applied on the partial correlation estimates ($r_{meth\text{-}independent}$ and $r_{cna\text{-}independent}$) to select genes in which CNA (blue points), or DNA methylation (purple points), or both (green points), were highly correlated with the expression of each gene. This high threshold ensured that the selected genes had a strong relationship between the respective mechanism and gene expression. The proportion of genes (out of the 2000 genes in VAR2000) in each of these 3 categories was plotted (Figure 4.2b – left panel). Similarly, the proportions were also plotted for the set of all genes (15968 genes) in Figure 4.2b (right panel). 21.7% of VAR2000 genes were methylation-regulated, whereas only 4.9% were CNA-regulated. A clear preference for methylation control of gene expression was observed in VAR2000 when compared to the set of all genes (All genes: CNA = 14.2% vs. Methylation = 5.5%). This corresponds to a substantial increase in the number of methylation-modulated genes with a simultaneous decrease in the number of CNA-modulated genes in VAR2000 compared to what is expected. Accordingly, Fisher's exact tests were used to formally assess three hypotheses: i) is the proportion of methylated-regulated genes significantly enriched or depleted in VAR2000 compared to what is expected based on all genes?; ii) is the proportion of CNA-regulated genes significantly enriched or depleted in VAR2000 compared to what is expected based on

Gene Set	N (total)	N (available)	Disease-specific	Description
VAR2000	2000	2000	Breast cancer	Genes with highest variation in expression in the METABRIC dataset [Curtis et al., 2012, this study]
SORLIE	437	392	Breast cancer	List of variably expressed genes used to classify breast tumour into Intrinsic subtypes [Perou et al., 2000; Sorlie et al., 2003]
PAM50	50	47	Breast cancer	Reduced list of variably expressed genes used to classify breast tumours into Intrinsic subtypes [Parker et al., 2009]
IC10	614	528	Breast cancer	Genes with high correlation between gene expression and cis-acting CNA. Used to classify breast tumours into Integrative clusters [Curtis et al., 2012]
COSMIC	602	549	All cancers	Catalogue of Somatic Mutations in Cancer [Forbes et al., 2015, cancer.sanger.ac.uk]

Table 4.1: Definition of 5 gene sets. N (total) represents the number of genes defined in the gene set. N (available) represents the number of genes profiled in the METABRIC dataset described in this thesis.

all genes?; and iii) Is the enrichment or depletion in methylation-regulated significantly different from CNA-regulated genes?

Figure 4.2c illustrates that methylation-modulated genes are significantly enriched in variably expressed (VAR2000) breast cancer genes (OR = 8.15, *FDR p-value* = 2.0×10^{-164}) while CNA-driven genes are significantly depleted (OR = 0.28; *FDR p-value* = 2.2×10^{-45}). Identical analyses were conducted for four additional gene sets that are defined in Table 1, comprising of breast cancer specific genes (Sorlie, PAM50 and IC10), and the Catalogue of Somatic Mutations in Cancer (COSMIC). The PAM50 gene set of 50 genes (47 genes available in this cohort) which has been shown to robustly classify breast tumours into the Intrinsic subtypes showed a similar enrichment in methylation-regulated genes (16 out of 47 genes, OR = 8.88, *FDR p-value* = 2.2×10^{-9}) and depletion in CNA-regulated genes (4 out of 47 genes, OR = 0.56, *p-value* = 0.401). The Sorlie gene set (previous iteration of the PAM50 gene set) is comprised of approximately equal proportions of methylated-modulated (17.86%, OR = 3.91, *p-value* = 2.1×10^{-18}) and CNA-modulated (19.64%, OR = 1.48, *p-value* = 0.0034) genes, however, the methylation-altered genes are enriched to a considerably

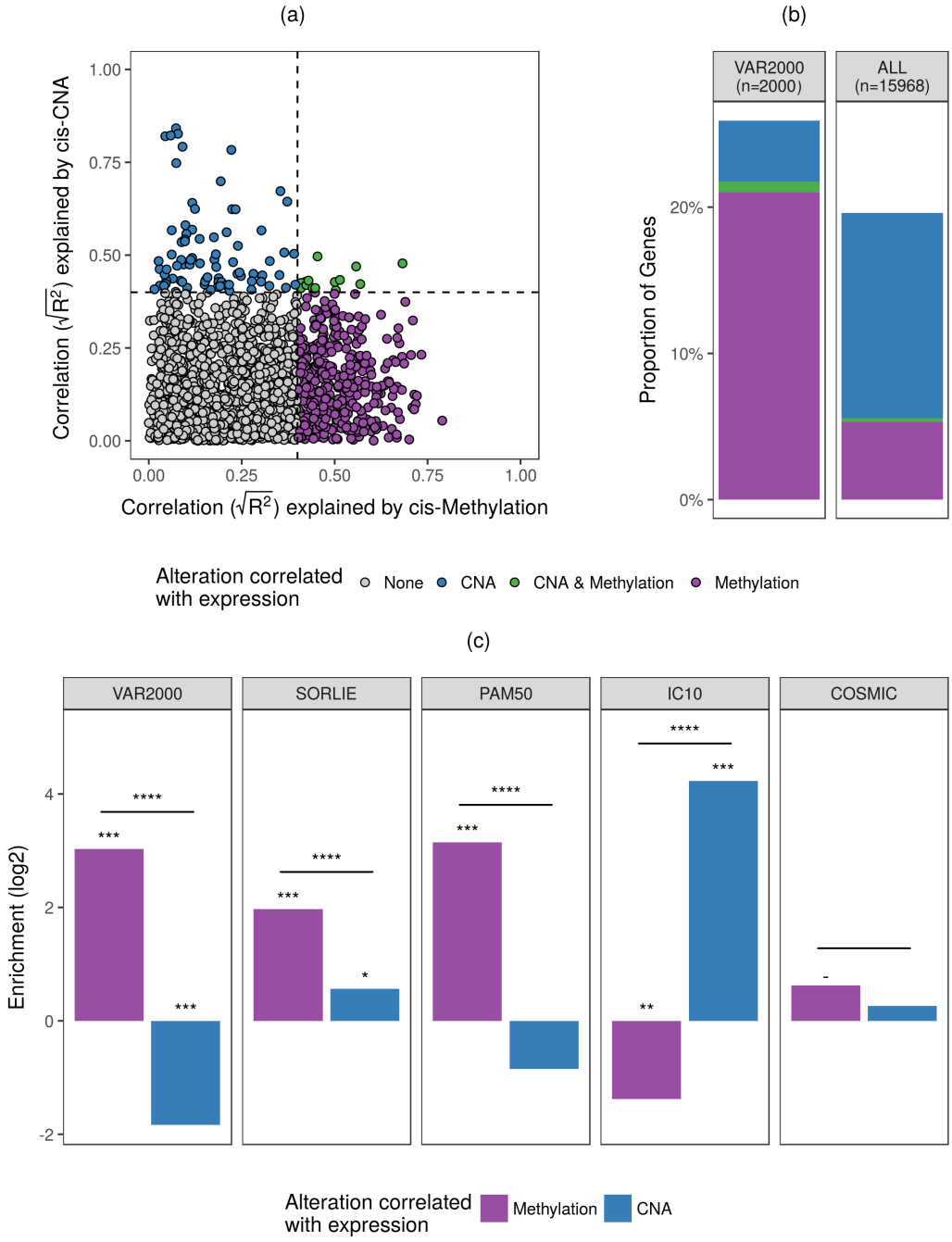


Figure 4.2: (Caption on next page.)

Figure 4.2: (Previous page.) ***Cis*-acting DNA methylation events are the predominant mechanism regulating variably expressed genes in breast cancer. (a)** Scatter plot of the partial correlations between *cis*-acting DNA methylation and gene expression (x-axis) and *cis*-acting CNA and gene expression (y-axis) for each of the 2000 variably expressed genes in breast cancer (VAR2000). Each point represents one gene. Colour of the points represents whether the gene can be described as methylation-regulated (purple), CNA-regulated (blue), both (green), or none (white) using a partial correlation threshold of 0.40. **(b)** Proportion of genes that are characterised as methylation-regulated (purple), CNA-regulated (blue) or both (green) within the VAR2000 gene set (left); and within the set of all genes (right, ALL gene set). **(c)** Enrichment analysis to investigate whether the proportion of methylated-regulated genes and CNA-regulated genes are significantly enriched or depleted in the gene set of interest compared to what is expected based on all genes. x-axis represents methylated-regulated genes (left bar - purple) and CNA-regulated genes (right bar - blue) and the five panels represents the 5 gene sets considered as defined in Table 4.1. Height of bars on the y-axis represents enrichment ($\log_2$(observed/ expected odds)). Asterisks on the two bars per panel represents *FDR p-values* (hypergeometric test) to test if the methylation (left) or CNA (right) regulation is enriched or depleted. Asterisks between the two bars represents *FDR p-values* (hypergeometric test) to explicitly test the difference in enrichment between methylation and CNA regulation. (. = *FDR p-values* < 0.1, $*$ = *FDR p-values* < 0.05, ** = *FDR p-values* < 0.01, *** = *FDR p-values* < 0.001, **** = *FDR p-values* < 0.0001).

higher extent than CNA-altered genes (*p-value* = 2.6×10^{-7}). These results provide substantial evidence that inter-tumour heterogeneity in the breast cancer transcriptome is largely associated with *cis* methylation differences, however, as argued in Chapter 3 (and later in Section), correlations between DNA methylation and expression may not be indicative of causation. Remarkably, COSMIC (the set of commonly mutated genes in all cancer types) also showed a higher enrichment for methylation-regulated genes than CNA-regulated genes (OR = 1.54 vs. OR = 1.20, *p-value* = 0.305), although a higher absolute frequency of CNA-driven genes were present (8.20% vs. 16.57%). This indicates that DNA methylation and CNA mechanisms interact with mutations in cancer and warrants the integrative analysis of DNA methylation and CNA complementing driver mutation genes in breast cancer.

The only gene set that displayed an exception to this pattern was IC10, in which the majority of genes were CNA-modulated genes (72.54%, OR = 18.80, p = 8.9×10-*212*), and showed significant depletion for methylation-modulated genes (2.27%, OR = 0.38, *p-value* = 3.1×10^{-4}). However, this observation is undoubtedly driven by the fact that the IC10 gene selection was purely based on high CNA-expression correlation for the same cohort [Curtis et al., 2012], and thus does not provide a fair assessment of the comparative regulatory roles of DNA methylation and CNA.

4.3.2 Tumour-normal and tumour-tumour differences

Next, the extent of the contributions of CNA and DNA methylation in explaining tumour-normal differences (in particular ER+ vs. normal and ER- vs. normal) or tumour-tumour differences (ER+ vs. ER-) were examined. Differential expression analysis was used to identify genes that were upregulated and genes downregulated in each of these comparisons separately (using *limma*: an absolute fold change > 1.5 and *FDR p-value* < 0.05 were used as thresholds for significance). A statistical framework to identify partial-correlations of DNA methylation and CNA ($r_{meth\text{-}independent}$ and $r_{cna\text{-}independent}$, similar to Section 4.2 and 4.3) was conducted. However, the regression models were adapted using appropriate sample weights (weighted least squares regression) to ensure that the two groups being compared contributed equally to the correlations. For example, for ER+ vs. normal comparisons, the 1124 ER+ tumours were weighted lower than 237 normal samples to prevent the ER+ tumours from dominating the results.

Genes downregulated in ER+ tumours vs. normal are largely influenced by methylation (7.65%) than CNA (5.92%) as illustrated in Figure 4.3a (left). In fact, on comparison with all genes, methylation-regulated genes were enriched (OR = 1.95, *FDR p-value* = 1.6×10^{-8}) and CNA-regulated genes were depleted (OR = 0.42, *FDR p-value* = 1.7×10^{-16}) in those silenced in ER+ tumours vs. normals (Figure 4.3b - left). On the contrary, genes upregulated in ER+ tumours vs. normal are dictated largely by copy number changes, with 20.51% of the overexpressed genes correlated with CNA, while only 11.2% correlated with methylation. Nonetheless, upregulated genes showed a higher enrichment for methylation-regulated genes than CNA-regulated genes (OR = 3.23 vs. OR = 1.93, *FDR p-value* = 6.22×10^{-6}).

Gene Set Enrichment Analysis (GSEA: pathways tested included Hallmark, REACTOME, oncogenic gene sets obtained from the Molecular Signatures Database [Subramanian et al., 2005, MSigDB]; hypergeometric test) were performed on the CNA and methylation-regulated genes to identify whether distinct pathways were disrupted by CNA and methylation alterations. CNA drives increased expression of key protagonists of cell cycle progression such as *AURKA* which encodes Aurora kinase A, a marker of proliferation that facilitates entry into M-phase of the cell cycle [Marumoto et al., 2005]; *CDK4* (Cyclin-dependent kinase 4) which is involved in G1/S phase cell cycle progression [Dai et al., 2012; Massagué, 2004]; and *NEK2*, a a cell cycle-regulated kinase involved in the control of centrosome separation [Fry, 2002]. This corresponded to significant enrichments of cell cycle progression pathways such as mitotic cell cycle (enrichment = 3.23, *FDR p-value* = 1.1×10^{-5}, REACTOME), targets

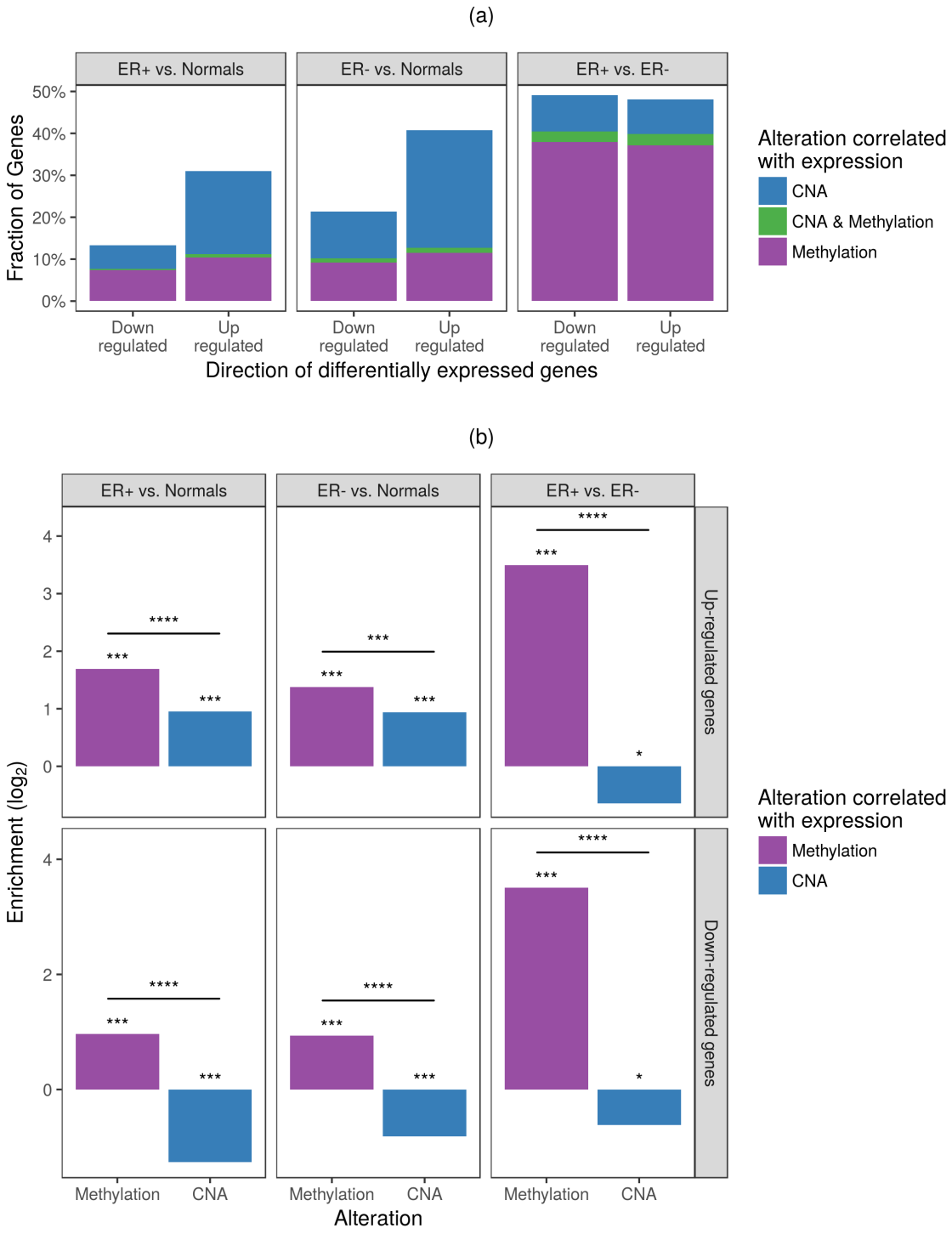


Figure 4.3: (Caption on next page.)

Figure 4.3: (Previous page.) ***Cis*-acting DNA methylation events are the predominant mechanism associated with silenced genes in breast cancer. (a)** Proportion of genes that are characterised as methylation-regulated (purple), CNA-regulated (blue) or both (green) in the differentially expressed genes between ER+ tumours and normals (left), between ER- tumours and normals (middle) and between ER+ and ER- tumours (right). x-axis represents the direction of regulation. **(b)** Enrichment analysis to investigate whether the proportion of methylated-regulated genes and CNA-regulated genes are significantly enriched or depleted in the gene set of interest compared to what is expected based on the set of all genes (ALL gene set). x-axis represents methylated-regulated genes (purple) and CNA-regulated genes (blue) and the six panels represents the 6 gene sets – 3 differentially expressed gene sets (left, middle, right) x 2 directions of regulation (top, bottom). Height of bars on the y-axis represents enrichment ($\log_2$(observed/ expected odds)). Asterisks on the two bars per panel represents *FDR p-values* (hypergeometric test) to test if the methylation (purple) or CNA (blue) regulation is enriched or depleted. Asterisks between the two bars represents *FDR p-values* (hypergeometric test) to explicitly test the difference in enrichment between methylation and CNA regulation. (. = *FDR p-values* < 0.1, $*$ = *FDR p-values* < 0.05, ** = *FDR p-values* < 0.01, *** = *FDR p-values* < 0.001, **** = *FDR p-values* < 0.0001).

of E2F transcription (enrichment = 5.24, *FDR p-value* = 4.4×10^{-9}, Hallmark), MYC targets (enrichment = 5.14, *FDR p-value* = 6.1×10^{-9}, Hallmark), G2/M checkpoint (enrichment = 3.91, *FDR p-value* = 6.6×10^{-6}, Hallmark), S phase (enrichment = 5.05, *FDR p-value* = 1.2×10^{-5}, REACTOME). Conversely, methylation was associated with activation of key regulators of the oestrogen signalling pathway such as *ESR1* and *GATA3* [Stone et al., 2015], as well as genes defining early (enrichment = 11.30, *FDR p-value* = 1×10^{-16}, Hallmark) and late response to oestrogen (enrichment = 8.94, *FDR p-value* = 2.6×10^{-12}, Hallmark). Genes silenced by methylation alterations were also associated with down-stream modulation of the AKT pathway (enrichment = 9.48, *FDR p-value* = 8.1×10^{-9}, Oncogenic signature), and mTOR pathway (enrichment = 5.72; *FDR p-value* = 8.6×10^{-5}, Oncogenic signature).

Differential expression analysis of ER- vs. normal showed similar results to ER+ vs. normals, with CNA dominating upregulated genes (CNA = 29.23% vs. Methylation = 12.74%; Figure 4.3a - middle). However, even though methylation regulated only 12.74% of overexpressed genes, this was significantly higher than expected (Methylation OR = 2.60 vs. CNA OR = 1.92, *FDR p-value* = 1.1×10^{-4}; Figure 4.3b - middle). CNA-regulated genes were also observed marginally more frequently in downregulated genes than methylation-regulated genes (12.14% vs. 10.24%) in concordance with the reported high chromosomal instability associated with ER- tumours [Curtis et al., 2012; Hu et al., 2009]. Nonetheless, methylation-regulation was observed significantly higher than expected (OR = 1.91, *FDR p-value* = 6.5×10^{-11}) in striking contrast to CNA (OR = 0.57, *FDR p-value* = 4.0×10^{-13})

which further substantiates the methylation-associated silencing role in cancer [Baylin, 2005; Herman and Baylin, 2003]. Analogous to the ER+ vs. normal analysis, CNA directed keys pathways associated with cell-cycle progression including mitotic cell cycle (enrichment = 5.02, *FDR p-value* = 1×10^{-16}, REACTOME), mitotic M/G1 phases (enrichment = 5.97, *FDR p-value* = 2.7×10^{-15}, REACTOME). Conversely, DNA methylation was involved in modulation of ER- specific pathways including downregulation of both oestrogen-response and TP53 signalling (enrichment = 6.26, *FDR p-value* = 6.0×10^{-7}, Oncogenic signature).

Analysing genes differentially expressed between the two breast cancer subtypes (ER+ vs. ER-) revealed that methylation-associated epigenetic control of gene expression (both up and downregulation) dominated the landscape (upregulated genes: OR = 11.25, *FDR p-value* = 2.29×10^{-101}, downregulated genes: OR = 11.36, *FDR p-value* = 2.9×10^{-94}; Figure 4.3b - right). All together, these results showed that whereas CNA plays a strong role leading to the divergence of tumours from normal tissues, DNA methylation, on the other hand, was the preferred mechanism for refining subtype-specific differences leading to inter-tumour heterogeneity.

Interestingly, in all comparisons conducted, the scatter plots of partial correlations of CNA and DNA methylation revealed an 'L' type shape (not shown, but similar to Figure 4.2a) which implies that expression of genes highly correlated with DNA methylation are not likely to be correlated with CNA and vice versa (*FDR p-value* for all comparisons < 0.05; Fisher's exact test). Thus, these two gene regulatory mechanisms are largely complementary, i.e. they do not commonly target the same genes.

4.4 DNA methylation as the CNA-modifier in gene expression

Although, Curtis et al. identified 2000 *cis*-acting CNA genes influencing the breast cancer transcriptome, CNAs did not explain the variation in expression for a majority of genes across the genome [Curtis et al., 2012]. In fact, results from the previous section indicated that a large contribution of inter-tumour heterogeneity in the breast cancer transcriptome can be explained by *cis* methylation differences. In order to assess whether DNA methylation adds significant gene regulation in addition to CNA, a nested regression framework was constructed. In the first iteration (the crude or unadjusted model), gene expression was treated as the independent variable and CNA as the dependent variable. For each gene, three statistics were recorded: i) the effect estimate of CNA (β_{cna}), which indicates the extent and direction of this relationship; ii) the FDR-corrected regression *p-value* of the CNA ($FDR\ p\text{-}value_{cna}$), which indicates the strength of evidence for its association with gene expression; and iii) a measure of the variation in expression explained by CNA (R^2_{cna}). In the second iteration of the regression model (confounder or adjusted model), DNA methylation (M-value) was added as the second predictor to the regression model. For each gene, the three statistics mentioned were recorded. i) the independent effect estimate of CNA ($\beta_{cna\text{-}independent}$),; ii) the FDR-corrected regression *p-value* of the independent effect of CNA *(FDR p-value* ($FDR\ p\text{-}value_{cna\text{-}independent}$); iii) and the total variation in expression explained by the combined CNA and DNA methylation model (R^2_{total}). was noted. The additional variation in expression explained by DNA methylation is ($R^2_{addmeth}$). with

$$R^2_{total} = R^2_{cna} + R^2_{addmeth} \tag{4.1}$$

It has been established that CNAs do not explain the variation in expression for a majority of genes (only 14.2% of all genes, only 4.2% of VAR2000 genes, Section 4.3.1) across the breast cancer genome. However, a key question is – are these non CNA-regulated genes largely diploid or is this a consequence of acquired copy number aberrations not inducing expected expression changes at the RNA level? As mentioned earlier, expression of 4.9% (n = 97) of VAR2000 genes in the genome were significantly regulated by copy number losses or copy number gains/ amplifications ($r_{meth\text{-}independent} \geq 0.40$, $FDR\ p\text{-}value_{cna} < 0.05$; linear regression), henceforth known as CNA-cis genes. 61% (n = 1220) of VAR2000 genes were weakly but positively significantly associated with CNA ($r_{meth\text{-}independent} < 0.40$, $\beta_{cna} > 0$, $FDR\ p\text{-}value_{cna}$

Figure 4.4: DNA methylation explains significant variation in transcriptional activity in the top 2000 variably expressed genes in breast cancer. Variation in expression (y-axis) explained by just CNA (R^2_{cna}) and CNA + DNA methylation (R^2_{total}) per gene. The slope of the lines connecting the two dots represent the additional variation in expression explained by DNA methylation. Each line represents a gene. A threshold of 10% was used to signify genes with significant added contribution (gold). This was analysed separately across the 4 categories of genes as defined in the text – CNA-*contradictory* (left), CNA-*independent* (middle left), CNA-*weakregulated* (middle right) and CNA-*cis* (right). % of genes with significant methylation contribution is shown in gold. Total number of genes in each category are shown in brackets in the panel legend. Only the 2000 variably expressed genes (VAR2000) were considered. CNA-weakregltd = CNA-*weakregulated*.

< 0.05; linear regression), henceforth known as CNA-*weakregulated* genes. These represent genes where copy number gains/ amplifications and copy number losses show mild consequences in transcription. Surprisingly, 4.2% (n = 83) of VAR2000 genes showed contradictory response to CNA (CNA-*contradictory* genes; $\beta_{cna} < 0$, *FDR p-value*$_{cna}$ < 0.05; linear regression) including genes such as *GATA3*, *CCND2* and *TSHZ2* with known tumorigenic roles. The remaining, 28.4% (n = 567) of VAR2000 genes were largely diploid or those in which CNA did not exhibit a significant (expected or contradictory) association with CNA (CNA-*independent* genes; *FDR p-value*$_{cna} \geq$ 0.05; linear regression).

Figure 4.4 illustrates the additional variation in expression explained by DNA methylation across these 4 categories of genes, and a threshold of 10% was used to signify genes with significant added contribution. CNA-*cis* genes did not exhibit significant additional methylation regulation (only 18.6% genes), which was expected since CNA and DNA methylation do not commonly target the same genes. 32.6% of CNA-*independent* genes showed significant contributions by DNA methylation, suggesting that a large extent of the unaccounted variation in expression reported in the literature could be explained by epigenetic control. Remarkably, higher proportions of CNA-*weakregulated* (39.6%) and CNA-contradictory genes (42.2%) had significant methylation contributions (*p-value* = 0.0021; chi-square test). This observation raises the hypothesis that DNA methylation could be used as a mechanism to modulate the effect of CNA on the breast cancer transcriptome (deactivate in the case of CNA-contradictory; enhance in the case of CNA-*weakregulated* genes).

4.4.1 DNA methylation alterations target potential tumour suppressor genes in genomic amplifications: *TSHZ2*

Breast cancer has largely been considered to be a copy number driven disease, where the normal diploid state of the genome is altered in large chromosomal regions. The function of copy number gains is exerted through amplifying the number of copies and thus increasing the expression of the underlying oncogene(s). Identifying the *driver* oncogenes within these large genomic regions remains a challenge, as any given amplification region might harbour several tens or hundreds of genes. Besides containing an oncogene, the large amplification event might also span genes whose overexpression is not beneficial or is even toxic for the cancer. Thus, to fully benefit from an amplification event, cancer cells need to selectively inhibit the expression of unwanted genes within amplified genomic regions. Is it possible that DNA methylation alterations could be used to target such genes and silence them within copy number

alterations? This postulation was explored in frequent amplicons in breast cancer, and one example is detailed below.

A set of 62 breast tumours have 4 or more copies of the chromosome 20q13 cytoband which consists of several genes (Figure 4.5a). *ZNF217*, an oncogene, which has been associated with promotion of Epithelial-to-Mesenchymal Transition (EMT) as well as the development of metastasis in mice *in vivo*, is the presumed target for this amplification [Krig et al., 2010; Vendrell et al., 2012]. In fact, breast tumours with the amplification exhibited significantly higher expression for *ZNF217* (Figure 4.5b). However, the immediate adjacent gene is *TSHZ2* (Teashirt zinc finger homeobox 2), which has been identified as a CNA-contradictory gene. The 62 tumours have an amplification for this gene as well, but they do not report higher expression. Why is this so? Examining the promoter of *TSHZ2* revealed that breast tumours with the amplification had higher average promoter methylation (Figure 4.5c). In fact, about 64.5% of tumours with amplifications had significant promoter hypermethylation which is an enrichment of almost 4 times as expected (expected = 33.1%, odds ratio = 3.67, *p-value* = 1.4×10^{-6}; Fisher's exact test; Figure 4.5d). Exploring the regulatory role of *TSHZ2* promoter methylation revealed that hypermethylation was associated with down regulation of the gene in tumours with amplifications (*p-value* = 0.0024; linear regression; Figure 4.5e). *TSHZ2* has been proposed as a potential tumour suppressor in breast cancer via regulation of *GLI1*, the downstream transcription factor of Hedgehog signalling [Habib and O'Shaughnessy, 2016; Yamamoto et al., 2011]. Collectively, this evidence strongly suggests that by virtue of being located on the same cytoband as *ZNF217, TSHZ2* has been amplified in 62 breast tumours. However, due to its tumour suppressive properties, promoter hypermethylation has been used as a mechanism to selectively silence this gene.

The lack of overexpression normally associated with copy number amplification in *TSHZ2* is attributed to the presence of altered DNA methylation, and will be henceforth termed as *a diminishing effect* of DNA methylation. The diminishing effects of DNA methylation manifests in the suppression of the anticipated effects of CNA in transcription, and therefore can explain the detection of CNA-contradictory genes such as *TSHZ2*.

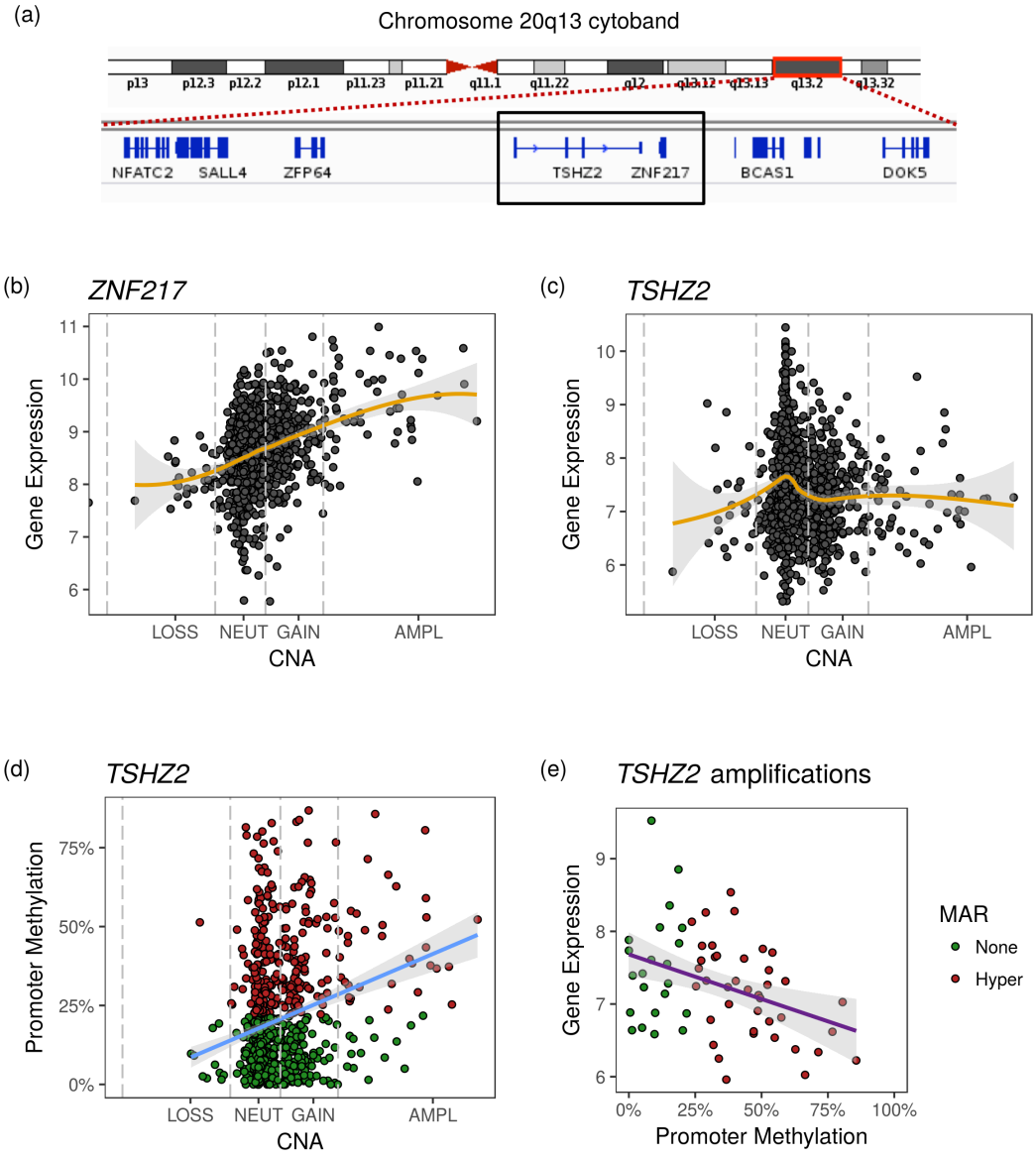


Figure 4.5: (Caption on next page.)

Figure 4.5: (Previous page.) **Anticipated effect of copy number amplification in *TSHZ2* is buffered by promoter hypermethylation (a)** Genes in chromosome 20q13 cytoband. **(b)** Relationship between *ZNF217* gene expression and CNA for all breast tumours. Gold line represents the loess smoothed curve (95% confidence interval shaded grey). **(c)** Relationship between *TSHZ2* gene expression and promoter methylation for all breast tumours. Gold line represents the loess smoothed curve (95% confidence interval shaded grey). **(d)** Relationship between *TSHZ2* promoter methylation and CNA for all breast tumours. Blue line represents the linear regression slope (95% confidence interval shaded grey). **(e)** Relationship between *TSHZ2* gene expression and promoter methylation for tumours with *TSHZ2* amplification (copies ≥ 4). Purple line represents the linear regression slope (95% confidence interval shaded grey). In (b) (c) and (d) Dashed vertical lines represents the discrete CNA boundaries as denoted on the x-axis. In (d) and (e) points are coloured by presence of hyper MAR (red) or not (green) in the tumour. LOSS = loss of copy number. NEUT = Neutral copy number. GAIN = gain of copy number. AMPL = Amplification of copy number.

4.4.2 DNA methylation can diminish or enhance the role of CNA in a subtype specific manner

ER+ and ER- tumours encompass significant differences in the fundamental biology of breast tumours. Considerable subtype-specific CNAs and DNA methylation alterations have been associated with differentially genes between ER+ and ER- tumours (Section 4.3). Could a mechanism of DNA methylation alterations modulating the effect of CNA in targeted genes, as introduced above, contribute to phenotypic differences (at the level of mRNA) between ER+ and ER- tumours? In order to formally evaluate this postulation, the extent of DNA methylation modifying the effect of CNA on gene expression was examined for all genes that were differentially expressed between ER+ and ER- tumours using a nested regression framework, similar to that described in Section 4.4. A key adaptation to this regression framework was that appropriate sample weights were used to ensure that the two ER subtypes contributed equally to the models (as described previously in Section 4.3.2). The crude model (first iteration: $mRNA \sim CNA$; Section 4.4) delineates the effect of CNA on gene expression without adjusting for DNA methylation as a confounder. This model represents the *observed effect* of CNA on gene expression (β_{cna}), and has been used in previous studies [Curtis et al., 2012]. The adjusted model (second iteration: $mRNA \sim CNA + Methylation$; Section 4.4) includes DNA methylation as a confounder. This enables the deconvolution of the independent role of the CNA ($\beta_{cna-independent}$). The effect estimates of CNA in the observed model (β_{cna}) and adjusted model ($\beta_{cna-independent}$) were compared as follows.

$$\Delta = \beta_{cna} - \beta_{cna-independent} \tag{4.2}$$

Delta modification (Δ) represents the direction and the magnitude of the methylation-assisted modification in the role of CNA. *GATA3*, a gene that is significantly upregulated in ER+ tumours, exhibited a large negative modification (Δ) in the *observed role* of CNA, and was explored in detail below.

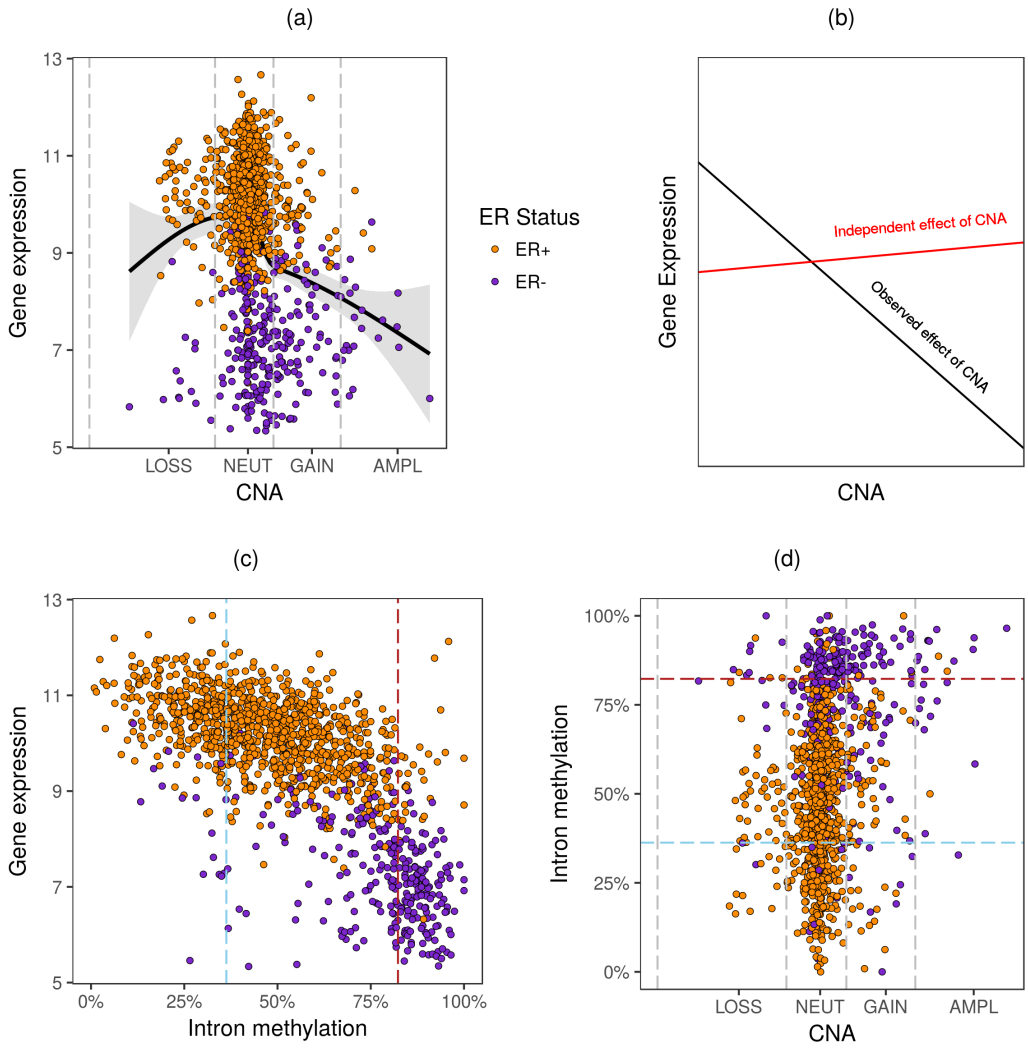


Figure 4.6: (Caption on next page.)

Figure 4.6: (Previous page.) **DNA methylation at the *GATA3* intron produces subtype-specific consequences in breast cancer. (a)** Relationship between *GATA3* gene expression and CNA for all breast tumours. Black line represents the loess smoothed curve (95% confidence interval shaded grey). **(b)** Comparison of observed effect of CNA on gene expression (black) and independent adjusted effect of CNA on gene expression (red) shown as distinct slopes. The reduced *observed* CNA role of *GATA3* is attributed to the diminishing influence of DNA methylation. **(c)** Relationship between *GATA3* gene expression and intron methylation for all breast tumours. **(d)** Relationship between *GATA3* intron methylation and CNA for all breast tumours. Blue line represents the linear regression slope (95% confidence interval shaded grey). In (a) (c) and (d) points are coloured by ER status. In (a) (c) and (d) Dashed vertical lines represents the discrete CNA boundaries as denoted on the x-axis. In (c) the blue vertical dashed line denotes boundary for hypo MAR definition, and red vertical dashed line denotes boundary for hyper MAR definition. In (d) the blue horizontal dashed line denotes boundary for hypo MAR definition, and red horizontal dashed line denotes boundary for hyper MAR definition. LOSS = loss of copy number. NEUT = Neutral copy number. GAIN = gain of copy number. AMPL = Amplification of copy number.

4.4.2.1 DNA methylation at the *GATA3* intron produces subtype-specific consequences in breast cancer

GATA3 is a known cofactor in the ER transcription complex and a marker of ER+/luminal breast cancer [Takaku et al., 2015]. However, ER+ tumours with copy number deletions in *GATA3* do not present with reduced expression (Figure 4.6a). Conversely, *GATA3* amplification were more common in ER- tumours but the affected tumours did not reflect an increase in gene expression. Consequently, on fitting the crude unadjusted model, a negative effect estimate of CNA with gene expression was observed for *GATA3* (β_{cna} = -2.3, as defined in Section 4.4; Figure 4.6b – black line). Remarkably, examining the methylation status of the *GATA3* gene body (intron) revealed that tumours with amplifications were far more likely to have a higher methylation status than tumours (Figure 4.6c), and this hyper methylation was significantly associated with reduced *GATA3* expression (Figure 4.6d). This strongly suggests that hypermethylation in the *GATA3* gene body was suppressing the effect of *GATA3* amplifications in the ER- tumours. Adjusting for the suppressing role of DNA methylation as a confounder, revealed that the *independent* effect of CNA on *GATA3* expression ($\beta_{cna\text{-}independent}$ = 0.2, as defined in Section 4.4; Figure 4.6b – red line) was much higher than *observed* effect (β_{cna} = -2.3, as defined in Section 4.4; Figure 4.6d – black line). The reduced *observed* role of CNA (compared to the methylation-adjusted effect estimate of CNA, Δ = -2.5) is clearly attributed to the diminishing influence of DNA methylation. Specifically, in the presence of a *GATA3* intron DNA methylation, ER- tumours with amplifications did not reflect a

concomitant increase in gene expression. *GATA3* has been shown to be involved in upregulating *ESR1* in ER+ cancer, and is not required to be highly expressed in ER- tumours. Based on this collective evidence, it is possible that the *GATA3* amplicon in ER− tumours encompasses other driver oncogenes, and that suppression of *GATA3* expression is mediated by DNA methylation.

4.4.2.2 DNA methylation diminishing CNA function

A substantial decrease in the *observed* effect estimate ($\Delta \leq$ -0.1) of CNA was used to identify other potential DNA methylation diminishing agents in the context of differentially expressed genes between ER+ and ER-. For these genes, in presence of a DNA methylation diminishing agent, tumours with amplifications do not reflect a concomitant increase in gene expression, and/ or tumours with copy number losses do not exhibit an expected reduction. This analysis revealed that 112 genes were upregulated and 95 genes were downregulated in ER+ tumours potentially as a consequence of a methylation-associated CNA-diminishing mechanism (Figure 4.7a – blue bars). Of note, this mechanism was identified in ER+ *upregulated genes* such as *GATA3* (detailed above), *DNALI1* and *SPDEF* (identified in Chapter 3), a known transcription factor involved in ER+ breast cancer pathogenesis [Sood et al., 2009]; and in ER+ *repressed genes* such as *IL32*, *MID1*, and the *SFRP1* (identified as a gene with ER+ specific tumour suppressive role, see Chapter 3) (Figure 4.7b – left panels). Collectively, these results strongly indicate that DNA methylation is being used as a CNA buffer to target critical regulators, which subsequently leads to subtype-specific breast cancer pathogenesis.

4.4.2.3 DNA methylation enhancing CNA function

Similarly, a substantial increase in the *observed* effect estimate ($\Delta \geq 0.1$) of CNA implies an enhancing effect of CNA on gene expression, bearing in mind that this could result in either hyper-activating genes in *tumours with* amplifications, or further deactivating gene function in the presence of copy number losses. In ER+ tumours, 139 upregulated genes and 180 downregulated genes were identified with CNA-enhancing roles (Figure 4.7a – orange bars). For instance, *FOXA1*, a known pioneer factor that is a fundamental determinant of ER activity in breast cancer [Hurtado et al., 2011] was shown to be significantly overexpressed in ER+ tumours as a consequence of a methylation-associated CNA-enhancing mechanism (Figure 4.7b – right panels). Conversely, Sry-related HMg-Box genes, *SOX10* and *SOX11* were overexpressed in ER- tumours due to CNA-enhancing configurations (Figure 4.7b – right panels). Fascinatingly, these genes have been shown be transcription factors and

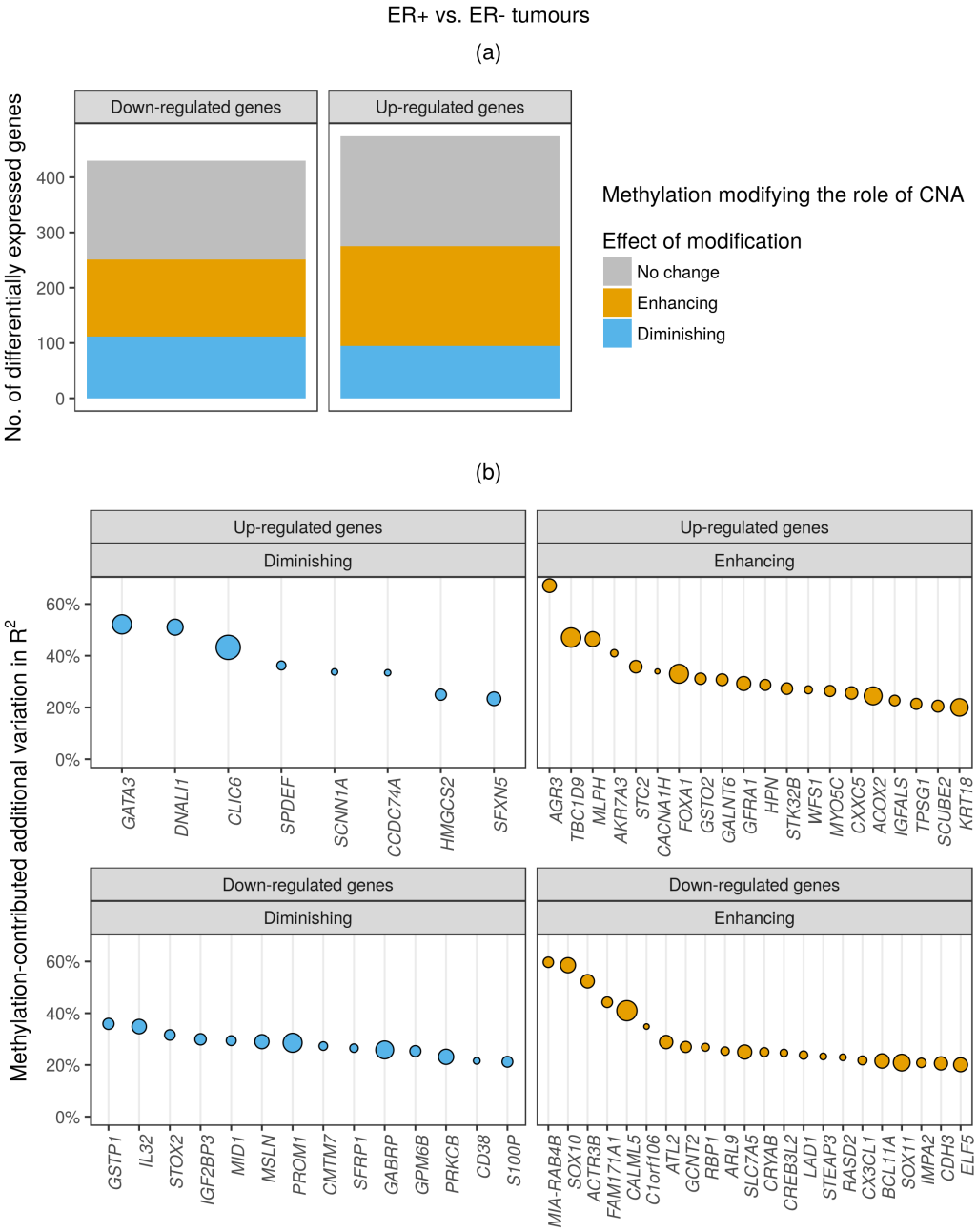


Figure 4.7: (Caption on next page.)

Figure 4.7: (Previous page.) **DNA methylation modifies the role of CNA in differentially expressed genes between ER+ and ER- tumours. (a)** Number of genes with evidence of methylation-associated copy number diminishing (blue bars) and number of genes with evidence of methylation-associated copy number enhancing (orange bars) were quantified among genes differentially expressed in ER+ vs. ER- tumours (downregulated – left panel) and (upregulated – right panel). **(b)** Genes with evidence of methylation-associated copy number diminishing were listed in the x-axis on the left panels (blue) and genes with evidence of methylation-associated copy number enhancing were listed on the x-axis in the right panels (orange). Additional variation in expression (R^2) contributed by DNA methylation was illustrated on the y-axis, and genes were ordered by decreasing values of this metric. Only genes with $R^2 \geq 20\%$ were illustrated. The shape of points represented the absolute magnitude of delta (Δ) of the modification in the role of CNA. Only genes that were differentially expressed between ER+ and ER- tumours were considered. Upregulated genes are represented in the top two panels and downregulated genes in the lower two panels.

critical regulators exclusively of ER- tumours [Cimino-Mathews et al., 2013; Shepherd et al., 2016]. These results indicate that *both* DNA methylation and CNA are necessary for the regulation of key cancer genes.

This line of analysis raises a closely related question -- how often do DNA methylation and CNA events co-occur in the same tumour, and what are the regulatory consequences? This is investigated in detail in the next section.

4.5 DNA methylation and CNA are complementary mechanisms in cancer

Conclusions made in the previous sections suggest that DNA methylation events may act as complementary/ or additive mechanisms to CNA within the same gene in regulating expression. To formally test this hypothesis, the prevalence of functional methylation alterations and CNAs were compared. The analysis was focused on genes that were differentially expressed in tumours compared to normal tissues (separately conducted for ER+ and ER- tumours, see Section 4.3). Functional methylation events for a specific tumour were defined as detected methylation altered regions (MARs, see Chapter 3) in the corresponding principal functional methylation region (PFMR, as described in Section 4.2) for the differentially expressed genes under consideration. Functional methylation events could be hyper or hypo MARs. A bipartite strategy was used for the definition of functional copy number events: for genes that were downregulated in tumours, homozygous deletions (complete loss of genomic material, *loss*) as well as hemizygous deletions (loss of one allele in a diploid gene, irrespective of number of copies in the other allele, *LOH*) were defined as functional events. On the other hand, amplifications (threshold of 4 or more copies) were defined as functional copy number events for genes upregulated in tumours.

4.5.1 Identification of potential tumour suppressors

Firstly, significantly silenced genes in ER+ and ER- tumours (vs. normal tissues, separate analysis for ER+ and ER-) were explored in order to identify potential tumour suppressor candidates. The proportion of tumours with DNA functional methylation (x-axis) and CNA losses (y-axis) were plotted for downregulated genes in ER+ and ER- tumours (Figure 4.8). Genes that were recurrently (in more than 25% tumours) altered due to MARs or CNAs or both were identified. 27.3% of downregulated genes in ER+ tumours have recurrent MARs (Figure 4.8, purple and green points) compared to only 15.3% of genes in ER- tumours (Figure 4.8 - purple and green). The opposite trend is observed for copy number losses with a higher number of ER- downregulated exhibiting recurrent copy number losses (ER+ = 16.4%, ER- = 56.6%; Figure 4.8, blue and green points). This parallels findings from Section 4.3.2 where DNA methylation was revealed to play a stronger role in silencing genes in ER+ tumours compared to ER- tumours.

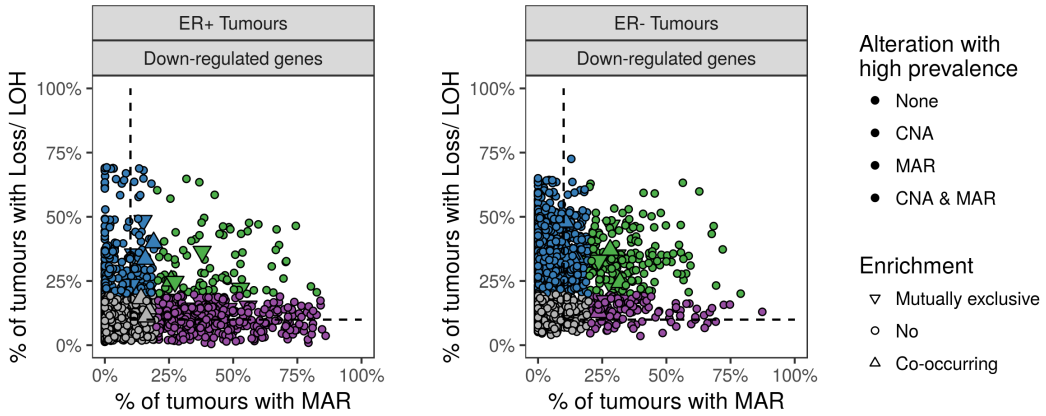


Figure 4.8: Silenced genes in breast cancer exhibit different propensities to DNA methylation or copy number loss. Scatter plots of the prevalence of MAR events (x-axis) against copy number loss (y-axis) in differentially downregulated genes in ER+ tumours (left panel) and ER- tumours (right panel). Each point represents a silenced gene. Colours represent the propensity of the gene for DNA for either MAR events (purple), CNA events (blue), both (green) or none (grey) based on a recurrence threshold of 25%. Genes enclosed by the dotted black lines are considered for the co-occurrence and mutually exclusive analyses as described in the text. Upward triangles represent genes with co-occurring MAR and CNA events, while downward triangles represent genes with mutually exclusive MAR and CNA events.

Interestingly, many genes were detected that were intrinsically prone to genomic loss (Figure 4.8, blue points) i.e. they exhibited recurrent copy number losses in breast cancer ($\geq$ 25% tumours) but were not commonly altered at the level of DNA methylation (i.e. $<$ 25% tumours harboured MARs for these genes). Genes exhibiting this pattern in both ER+ and ER- tumours include *CDH5* and *CDH13* involved in cytoskeletal organisation [Berx and van Roy, 2009; van Roy, 2014]. However, ER- tumours harboured considerably more genes with a propensity for copy number loss such as *NR2F1* and *SETBP1*. Interestingly these 2 genes also exhibited this pattern in lung squamous cell carcinoma and colon adenoma carcinoma [Teschendorff et al., 2016b]. Conversely, genes with a propensity for DNA methylation alterations and not copy number losses (Figure 4.8, purple points) such as *HOXA4* and *HOXA5* were detected in ER+ tumours. These *HOXA* genes encode transcription factors that have critical roles in embryogenesis and tissue differentiation and evidence for epigenetic silencing has also been demonstrated in acute myeloid leukaemia [Musialik et al., 2015] and lung cancers [Teschendorff et al., 2016b]. Interestingly, methylation-regulated silencing of *CDH13* was commonly observed in ER- tumours (compared

to CNA-regulated silencing in ER+ tumours) indicating the preference of distinct mechanisms to downregulate this gene in ER+ and ER- tumours.

Several genes undergoing *both* copy number losses and MARs in tumours were also identified. A lower threshold of 10% prevalence was used, but for *both* types of alteration (Figure 4.8, points enclosed by dotted black lines). These genes were also investigated for patterns of co-occurrence and mutual exclusivity for the two mechanisms -- CNA and MARs within the same gene, using a hypergeometric test. Furthermore, associations with gene expression were also examined for these genes. One-way ANOVA tests were used to explore expression changes across 4 categories of tumours: i) tumours that were wild-type (i.e. had no MAR or copy number loss); ii) tumours that only had MARs; iii) tumours that only had copy number losses; and iv) tumours in which concomitant MARs and copy number losses co-occurred for the same gene. Gene expression was z-transformed using the mean and SD of the wild-type tumours (1^{st} group of samples) such that the wild-type tumours are given an expression z-score value of 0. This allows straightforward comparisons between the groups across different genes. Tumour categories with less than 5 samples were not considered, and ANOVA *p-values* were corrected for multiple testing (FDR). This analysis was conducted separately for i) genes exhibiting co-occurrence; and ii) genes exhibiting mutual exclusivity, and is detailed in the following two sections.

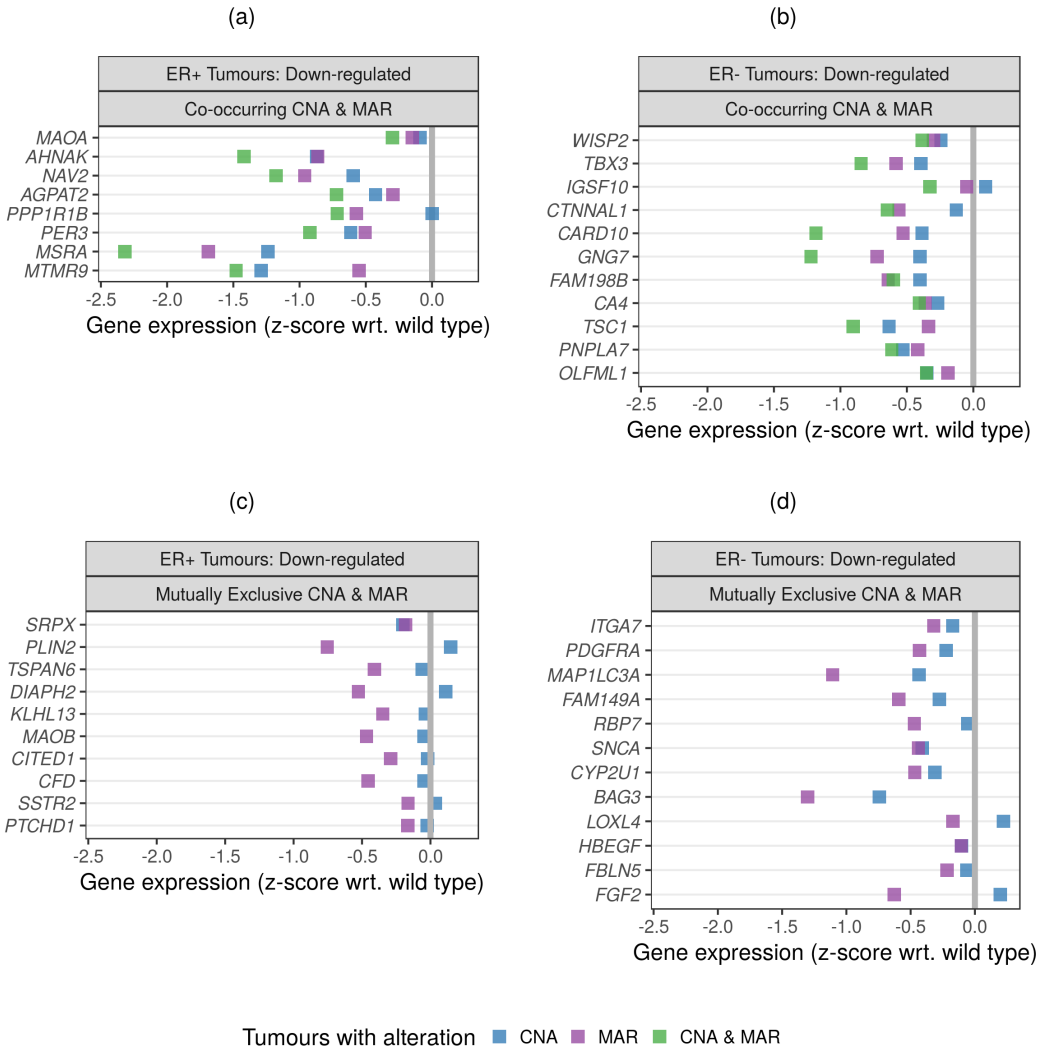


Figure 4.9: (Caption on next page.)

Figure 4.9: (Previous page.) **Silenced genes in breast cancer exhibit co-occurring or mutually exclusive patterns of DNA methylation and copy number loss. (a)** All silenced genes with significantly co-occurring MAR and CNA events within the same tumour, detected in ER+ tumours. **(b)** same as (a) but for ER- tumours. **(c)** All silenced genes with significantly mutually exclusive MAR and CNA events within the same tumour, detected in ER+ tumours. **(d)** same as (c) but for ER- tumours. In (a), (b), (c) and (d), each row represents a gene. The coloured squares represent the mean z-scores of gene expression data for three categories: (blue) tumours that only had copy number losses; (purple) tumours that only had MARs; (green) tumours in which concomitant MARs and copy number losses co-occurred for the same gene. The z-scores were constructed using the mean and SD of wild type tumours as detailed in the text, and the grey vertical line at 0 represents the mean z-scores of gene expression for wild type tumours. Tumour categories with less than 5 samples were not considered. In (c) and (d), tumours with concomitant MARs and copy number losses events are not shown since this constitutes a rare or significantly depleted category. Tumours with *AHNAK* MARs (pink) and those with *AHNAK* copy number losses (blue) have similar expression on average which is why the CNA (blue) data point is not visible.

4.5.1.1 Downregulated genes with co-occurring CNA and DNA methylation profiles

Genes in which *tumours harboured MARs as well as CNA losses simultaneously* were observed more than expected by chance (Enrichment > 1.5, *FDR p-value* < 0.05; hypergeometric test) were identified and were termed as *co-occurring genes*. Only significantly silenced genes in ER+ and ER- tumours, that had a prevalence of 10% or more for both CNA and MARs were considered (No. of such silenced genes: ER+ = 402, ER- = 501). Co-occurring silenced genes were identified separately in ER+ (n = 8) and ER- tumours (n = 11), and are indicated with upward triangles in Figure 4.8; and detailed in Table 4.2a and Table 4.2b. These genes might have tumour-suppressive roles in the respective subtypes and accordingly associations with gene expression were examined as described above. Figure 4.9a and Figure 4.9b displays the average expression z-scores across the 4 categories of tumours defined above for the co-occurring silenced genes. In 7 out of 8 (87.5%) co-occurring silenced genes detected in the ER+ subtype, the tumours with concomitant MARs and copy number losses (Figure 4.9a, b – green points) suffered the highest loss in expression compared to other tumour categories (ANOVA *FDR p-value* < 0.05). Together with the observation that this co-occurrence was a significantly enriched event, these finding strongly imply that the two independent mechanisms augment each other's silencing roles and their combination provides a selective advantage for the tumour. These associations identified for concomitant MARs and CNA events support the classical definition for loss of heterozygosity (LOH). Similarly, 8 out of 11 (72.7%) co-occurring silenced

genes detected in ER- tumours demonstrated evidence of loss of heterozygosity. These 15 genes (ER+ = 7, ER- = 8) are put forward as candidate tumour suppressors. Of note are the transcription factors, *PER3* (ER+ tumours, Figure 4.10a) and *TBX3* (ER- tumours, Figure 4.10b).

Gene	Feature	Methylation		CNA		Enrichment			Expression	
		Direction	(%)	CNA	(%)	Type	Extent	FDR p*	Direction	FDR p**
(a) Co-occurring genes: ER+ tumours										
MAOA	promoter	Hypo	16.7	Loss/ LOH	14.2	Co-occurring	1.70	0.0005	Down	0.1246
AHNAK	intron	Hyper	10.2	Loss/ LOH	16.2	Co-occurring	1.70	0.0061	Down	<0.0001
NAV2	intron	Hyper	16.2	Loss/ LOH	11.1	Co-occurring	1.67	0.0060	Down	<0.0001
AGPAT2	intron	Hyper	10.4	Loss/ LOH	14.5	Co-occurring	1.63	0.0205	Down	<0.0001
PPP1R1B	promoter	Hyper	14.1	Loss/ LOH	18.2	Co-occurring	1.55	0.0059	Down	<0.0001
PER3	promoter	Hyper	25.8	Loss/ LOH	25.1	Co-occurring	1.51	<0.0001	Down	<0.0001
MSRA	intron	Hypo	19.0	Loss/ LOH	40.0	Co-occurring	1.51	<0.0001	Down	<0.0001
MTMR9	intron	Hypo	15.8	Loss/ LOH	33.5	Co-occurring	1.50	0.0003	Down	<0.0001
(b) Co-occurring genes: ER- tumours										
WISP2	intron	Hypo	12.7	Loss/ LOH	13.5	Co-occurring	2.08	0.0375	Down	0.2706
TBX3	intron	Hyper	11.2	Loss/ LOH	32.7	Co-occurring	1.95	0.0044	Down	0.0026
IGSF10	promoter	Hypo	23.6	Loss/ LOH	12.8	Co-occurring	1.85	0.0098	Down	0.6289
CTNNAL1	intron	Hypo	10.2	Loss/ LOH	35.5	Co-occurring	1.69	0.0121	Down	0.0258
CARD10	exon	Hypo	31.9	Loss/ LOH	24.5	Co-occurring	1.68	0.0001	Down	<0.0001
GNG7	intron	Hypo	13.0	Loss/ LOH	34.1	Co-occurring	1.62	0.0323	Down	<0.0001
FAM198B	exon	Hyper	28.1	Loss/ LOH	37.1	Co-occurring	1.58	<0.0001	Down	0.0001
CA4	exon	Hypo	10.6	Loss/ LOH	48.0	Co-occurring	1.56	0.0036	Down	0.0203
TSC1	exon	Hypo	13.2	Loss/ LOH	35.5	Co-occurring	1.55	0.0213	Down	<0.0001
PNPLA7	exon	Hypo	14.8	Loss/ LOH	34.3	Co-occurring	1.53	0.0268	Down	0.0011
OLFML1	promoter	Hypo	10.4	Loss/ LOH	38.5	Co-occurring	1.52	0.0469	Down	0.0421
(c) Mutually exclusive genes: ER+ tumours										
SRPX	intron	Hypo	19.6	Loss/ LOH	13.9	Exclusive	0.25	<0.0001	Down	0.0468
PLIN2	intron	Hyper	10.3	Loss/ LOH	23.6	Exclusive	0.28	<0.0001	Down	<0.0001
TSPAN6	intron	Hypo	18.1	Loss/ LOH	17.6	Exclusive	0.30	<0.0001	Down	0.0001
DIAPH2	intron	Hypo	15.2	Loss/ LOH	18.1	Exclusive	0.32	<0.0001	Down	<0.0001
KLHL13	intron	Hypo	20.1	Loss/ LOH	18.8	Exclusive	0.36	<0.0001	Down	0.0002
MAOB	intron	Hypo	16.8	Loss/ LOH	13.5	Exclusive	0.38	0.0028	Down	<0.0001
CITED1	promoter	Hyper	13.0	Loss/ LOH	28.3	Exclusive	0.40	<0.0001	Down	0.0120
CFD	promoter	Hyper	10.5	Loss/ LOH	17.8	Exclusive	0.42	0.0189	Down	0.0001
SSTR2	exon	Hypo	13.7	Loss/ LOH	14.4	Exclusive	0.44	0.0118	Down	0.2785
PTCHD1	intron	Hypo	17.9	Loss/ LOH	13.4	Exclusive	0.48	0.0087	Down	0.2785
(d) Mutually exclusive genes: ER- tumours										
ITGA7	promoter	Hyper	15.5	Loss/ LOH	30.4	Exclusive	0.07	<0.0001	Down	0.1149
PDGFRA	intron	Hyper	12.6	Loss/ LOH	21.5	Exclusive	0.24	0.0379	Down	0.0348
FAM149A	promoter	Hyper	16.6	Loss/ LOH	36.2	Exclusive	0.29	0.0002	Down	0.0006
MAP1LC3A	promoter	Hyper	26.2	Loss/ LOH	17.2	Exclusive	0.29	0.0043	Down	<0.0001
RBP7	intron	Hyper	16.7	Loss/ LOH	34.4	Exclusive	0.37	0.0041	Down	0.0416
SNCA	promoter	Hyper	17.0	Loss/ LOH	28.4	Exclusive	0.44	0.0379	Down	0.0045
CYP2U1	intron	Hyper	18.9	Loss/ LOH	29.6	Exclusive	0.46	0.0323	Down	0.0045
BAG3	intron	Hyper	20.1	Loss/ LOH	32.6	Exclusive	0.46	0.0152	Down	<0.0001
HBEGF	promoter	Hyper	28.5	Loss/ LOH	46.6	Exclusive	0.48	<0.0001	Down	0.7440
FBLN5	exon	Hypo	12.3	Loss/ LOH	44.7	Exclusive	0.48	0.0479	Down	0.2528
LOXL4	intron	Hyper	15.4	Loss/ LOH	36.1	Exclusive	0.48	0.0255	Down	0.2394
FGF2	promoter	Hyper	27.7	Loss/ LOH	24.7	Exclusive	0.50	0.0369	Down	0.0006

Table 4.2: (Caption on next page.)

Table 4.2: (Previous page.) **Silenced genes in breast cancer exhibit co-occurring or mutually exclusive patterns of DNA methylation and copy number loss. (a)** All silenced genes with significantly co-occurring MAR and CNA events within the same tumour, detected in ER+ tumours. **(b)** same as (a) but for ER- tumours. **(c)** All silenced genes with significantly mutually exclusive MAR and CNA events within the same tumour, detected in ER+ tumours. **(d)** same as (c) but for ER- tumours. The direction and prevalence (%) of the MAR event and the direction and prevalence (%) of the CNA event (only copy number losses considered) were detailed for each gene. The direction and extent of enrichment (observed/ expected) and *FDR p-value* and are indicated based on the hypergeometric test as defined in the text. For (a) and (b), the co-occurring genes are ordered by decreasing enrichment (Odds Ratio; OR). The direction of expression change and the *FDR p-value* based on the one-way ANOVA of the mean expression z-scores across the 4 categories of tumours are indicated. For (c) and (d), the mutually exclusive genes are ordered by increasing enrichment (OR, decreasing depletion). The direction of expression change and the *FDR p-value* based on the one-way ANOVA of the mean expression z-scores across the 3 categories of tumours are indicated. FDR p* = *FDR p-value* from the hypergeometric test; FDR p** = *FDR p-value* from the ANOVA.

4.5.1.2 Downregulated genes with mutually exclusive CNA and DNA methylation profiles

Genes in which *tumours harboured MARs as well as copy number losses in a mutually exclusive fashion* (enrichment < 0.5, *FDR p-value* < 0.05; hypergeometric test) were identified. These genes exhibited higher levels of MARs in tumours that had no copy number losses for the given gene, compared with tumours harbouring copy number losses for the given gene, and vice versa, and were termed as *mutually exclusive genes*. As mentioned above, only significantly silenced genes in ER+ and ER- tumours, that had a prevalence of 10% or more for both CNA and MARs were considered (No. of such silenced genes: ER+ = 402, ER- = 501). Mutually exclusive silenced genes were identified separately in ER+ (n = 10) and ER- tumours (n = 12), and are indicated with downward triangles in Figure 4.8; and detailed in Table 4.2c and Table 4.2d. For these genes, a pattern of two mutually exclusive events is an indication that the second alteration offers no further selective advantage than the first hit, or in fact leads to a disadvantage for the cell leading to cell death, which explains why tumours with concomitant MAR and copy number losses were observed less frequently than expected by chance. Therefore, in these genes, MAR and copy number losses represent two alternative mechanisms for silencing. Figure 4.9c and Figure 4.9d displays the average expression z-scores across 3 categories of tumours (tumours with concomitant MARs and copy number losses are not considered since this constitutes a rare or significantly depleted category). In 7 out of 10 (70.0%) mutually exclusive silenced genes detected in the ER+ subtype (8 out of 12 i.e. 66.7% in ER-), tumours with MAR

events (Figure 4.9c, d – green points) suffered the highest loss in expression compared to other 2 tumour categories (ANOVA *FDR p-value* < 0.05) suggesting that DNA methylation is the predominant mechanism in downregulating these potential tumour suppressors. These 15 genes (ER+ = 7, ER- = 8) are also put forward as candidate methylation-regulated tumour suppressors. Of note are *CITED1* (ER+ tumours, Figure 4.10c), a transcription factor which is a regulator of oestrogen signalling; and the fibroblast growth factor, *FGF2* (ER- tumours, Figure 4.10d). Expression of both these genes have been shown to be correlated with good prognosis in breast cancer [McBryan et al., 2007; Yiangou et al., 1997] thus revealing their potential function as tumour suppressor genes.

4.5.1.3 *BRCA1* demonstrates classical tumour suppressor behaviour

BRCA1 (Breast cancer susceptibility gene 1) is a canonical tumour suppressor gene that plays a critical role in homologous recombination to repair DNA damage and prevent tumour development [Easton et al., 1993; Hall et al., 1990; Miki et al., 1994; Moynahan, 2002; Moynahan et al., 1999]. Germline inactivating mutations in *BRCA1* were established to have a strong role in hereditary cases of breast and ovarian cancers [Castilla et al., 1994; Friedman et al., 1994]; and this *first hit* was shown to be accompanied by a *second hit* through the somatically acquired loss of chromosomal material on the wild type allele i.e. LOH [Merajver et al., 1995]. Collectively, these observations supported the identification of *BRCA1* as a classical tumour suppressor gene, based on Knudson's seminal *two hit model* for tumorigenesis [Knudson, 1971]. Since then several studies have revealed that other alterations such as somatic mutations and promoter hypermethylation in *BRCA1* can also drive the initiation and progression of breast cancer [Cancer Genome Atlas Network, 2012; Esteller, 2000; Pereira et al., 2016; Polak et al., 2017].

The availability of a large cohort of breast tumours with expression, copy number, mutational and DNA methylation information on *BRCA1* (approximately 1344 samples: 1241 breast tumours and 103 adjacent normal breast tissues from the METABRIC dataset had data on all four platforms) allowed a comprehensive examination of the different configurations that non-hereditary genetic and epigenetic events combine with each other to silence *BRCA1* in breast cancer. Firstly, *BRCA1* promoter methylation was estimated for all 1241 breast tumours and 103 normal tissues. A 16 CpG SCCRUB region was detected in the promoter of *BRCA1* (chr17:41276641 – 41277275). 35 tumours were identified with a significant hyper MAR at this locus in the *BRCA1* promoter (Figure 4.11a). All hyper MARs were

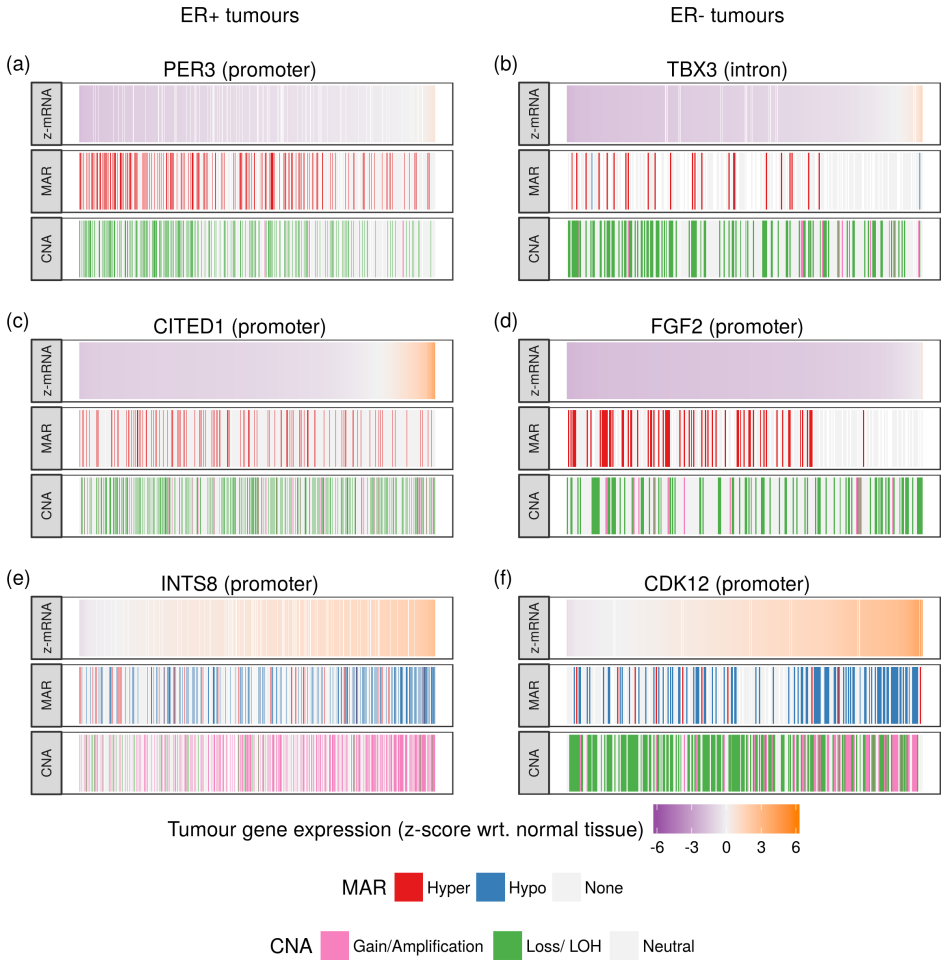


Figure 4.10: Examples of key genes in breast cancer exhibiting co-occurring or mutually exclusive patterns of DNA methylation or copy number loss. **(a)** *PER3*, a significantly downregulated gene in ER+ tumours, exhibiting concomitant genomic losses and promoter hyper MARs. **(b)** *TBX3*, a significantly downregulated gene in ER- tumours, exhibiting concomitant genomic losses and intron hyper MARs **(c)** *CITED1*, a significantly downregulated gene in ER+ tumours, exhibiting mutually exclusive genomic losses and promoter hyper MARs. **(d)** *FGF2*, a significantly downregulated gene in ER- tumours exhibiting mutually exclusive genomic losses and promoter hyper MARs. **(e)** *INTS8*, a significantly upregulated gene in ER+ tumours, exhibiting concomitant genomic amplifications and promoter hypo MARs. **(f)** *CDK12*, a significantly upregulated gene in ER- tumours exhibiting concomitant genomic amplifications and promoter hypo MARs. For each gene, a heatmap representation of the (top) z-score mRNA change vs. normal, (middle) MAR status – hyper, hypo or none, and (bottom) CNA status – Loss/ LOH, Gain/ amplification or neutral, is plotted for each tumour. Tumours are sorted by increasing gene expression.

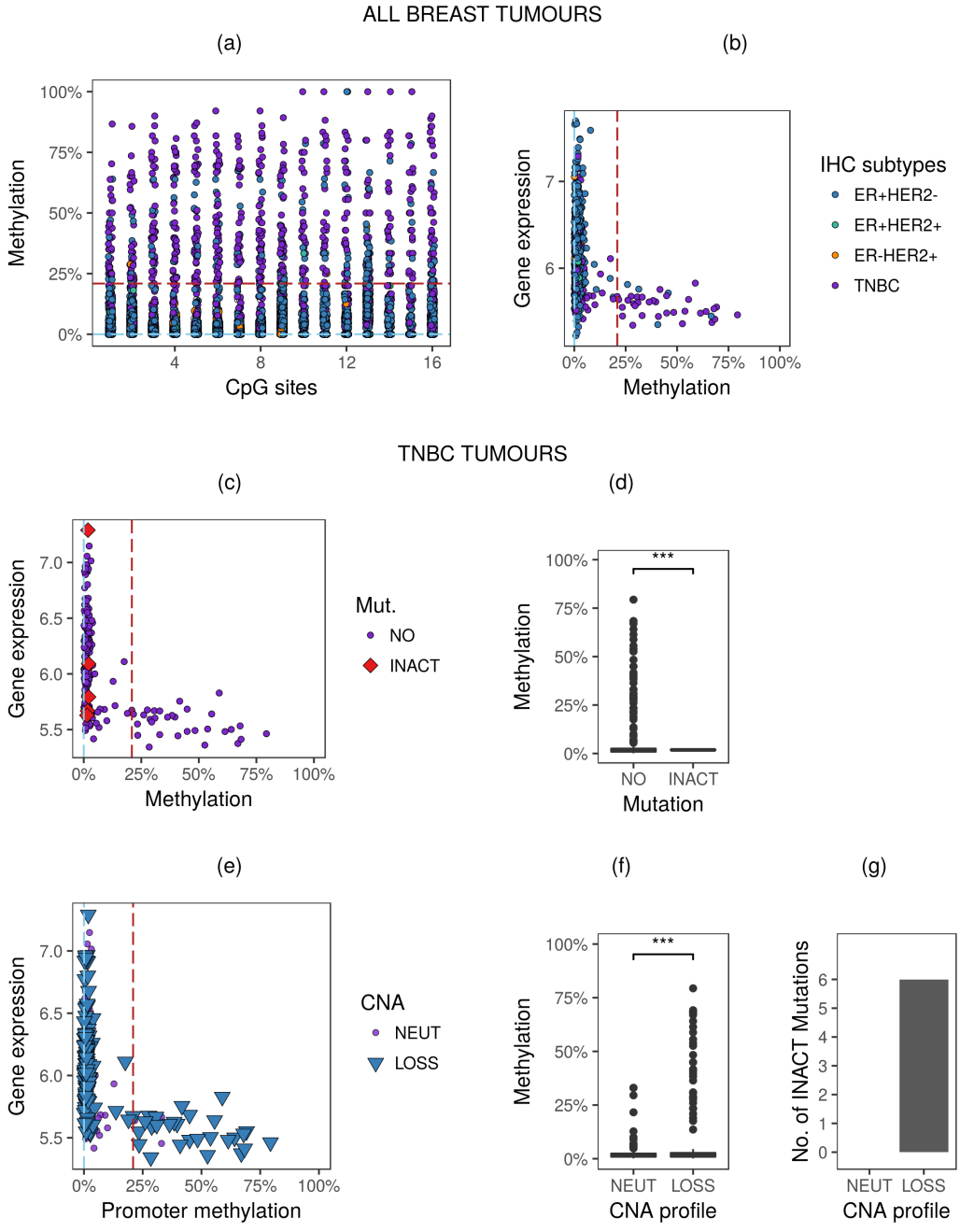


Figure 4.11: (Caption on next page.)

Figure 4.11: (Previous page.) ***BRCA1* demonstrates classical tumour suppressor behaviour.** **(a)** DNA methylation estimates for all tumours across the 16 CpG sites at the *BRCA1* promoter region defined using SCCRUB. Each point represents a tumour for a specific CpG site, coloured by IHC status. **(b)** Relationship between *BRCA1* gene expression and promoter methylation (mean for the promoter SCRRUB region) for all breast tumours. Each point represents a tumour, coloured by IHC status. **(c)** Relationship between *BRCA1* gene expression and promoter methylation for all TNBC tumours. Each point represents a tumour, coloured by mutation status. **(d)** *BCRA1* promoter methylation estimates compared between TNBC tumours with and without an inactivating *BRCA1* mutation. *P-value* based on t-test as described in text. **(e)** Relationship between *BRCA1* gene expression and promoter methylation for all TNBC tumours. Each point represents a tumour, coloured by copy number status. **(f)** *BCRA1* promoter methylation estimates compared between TNBC tumours with and without a copy number loss. *P-value* based on t-test as described in text. **(g)** No. of inactivating *BCRA1* mutation for TNBC tumours with and without a copy number loss. *P-value* based on Fisher's exact test as described in text. In (a) the blue horizontal dashed line denotes boundary for hypo MAR definition, and red horizontal dashed line denotes boundary for hyper MAR definition. In (b) (c) and (e) the blue vertical dashed line denotes boundary for hypo MAR definition, and red vertical dashed line denotes boundary for hyper MAR definition. *BRCA1* = Breast cancer susceptibility gene 1. TNBC = Triple negative breast cancer. Mut. = Mutation. INACT = Inactivating mutation. NEUT = Neutral copy number. LOSS = loss of copy number. (. = *FDR p-values* < 0.1, $*$ = *FDR p-values* < 0.05, ** = *FDR p-values* < 0.01, *** = *FDR p-values* < 0.001, **** = *FDR p-values* < 0.0001).

confirmed as *directed*-MARs (as defined in Chapter 3), and the methylation status was significantly associated with silencing of the gene (*p-value* < 1×10^{-16}; linear regression; Figure 4.11b). Triple negative breast tumours (TNBC, breast tumours that are ER-, PR- and HER2-, see Chapter 1) were significantly enriched for the promoter hypermethylation event corroborating several previous reports [Xu et al., 2013; Yamashita et al., 2015; Zhang et al., 2015; Zhu et al., 2015], with 30 out of 35 breast tumours harbouring a *BRCA1* promoter hyper MAR, also having TNBC characteristics (enrichment = 29.7, *p-value* = 3.8×10^{-18}; hypergeometric test). Subsequent analysis was focused on the 232 TNBC tumours. Next, the somatic mutational profiles for these tumours were examined and 6 TNBC tumours with inactivating *BRCA1* mutations were detected (Figure 4.11c). Remarkably, none of these 6 tumours harboured hyper MARs at the *BRCA1* promoter indicating that inactivating mutations and promoter hypermethylation events are mutually exclusive in *BRCA1* (*p-value* = 1.4×10^{-07}; t-test; Figure 4.11d). Examining the *BRCA1* copy number landscape revealed that a vast majority of the TNBC tumours (61.2%) harboured LOH (Figure 4.11e); and that the TNBC tumours with LOH were significantly enriched for promoter hypermethylation (*p-value* = 6.9×10^{-5}; t-test; Figure 4.11f) strongly suggesting a preference for these two *BRCA1* events to

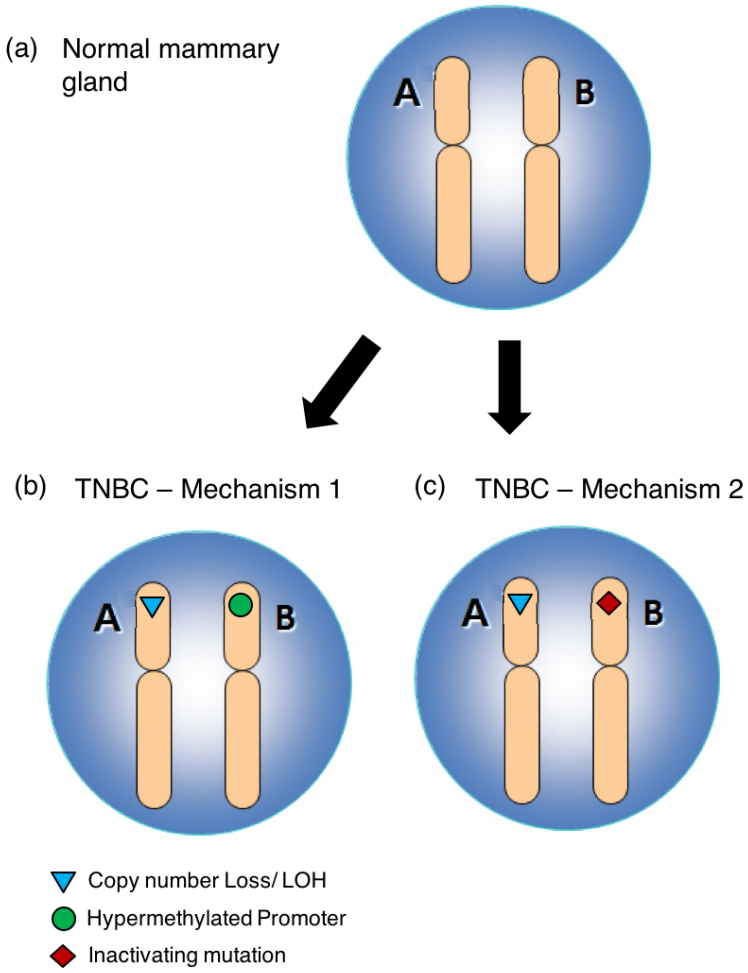


Figure 4.12: Two non-hereditary mechanisms illustrated for *BRCA1* silencing in TNBCs. (a) Status of the 2 BRCA1 alleles in the normal mammary gland. **(b)** Mechanism 1 for silencing *BRCA1* in TNBCs: LOH is accompanied by *BRCA1* promoter hypermethylation. **(c)** Mechanism 2 for silencing *BRCA1* in TNBCs: LOH is accompanied by inactivating *BRCA1* mutation. *BRCA1* = Breast cancer susceptibility gene 1. TNBC = Triple negative breast cancer.

co-occur in the same tumour. Integrating the mutational and copy number profiles of *BRCA1* for these tumours revealed the third piece of the puzzle, that is, all 6 tumours harbouring inactivating BRCA1 mutations also suffered LOH (*p-value* = 0.0853; hypergeometric test; Figure 4.11g). Although this was marginally significant in TNBCs (at alpha = 0.05), expanding this association to all breast tumours revealed a strong statistical significance (*p-value* = 0.0003; hypergeometric test) again indicating a tendency for these two somatic genetic events to occur on the two alleles of *BRCA1*.

Collectively, this data has revealed two distinct non-hereditary mechanisms that lead to both alleles of *BRCA1* being hit, and supporting its role as a tumour suppressor gene (Figure 4.12). Firstly, tumours with *BRCA1* promoter hypermethylation almost always harboured LOH (26 out of 30 tumours with hyper MARs also had LOH) ensuring that only the function of the hypermethylated allele is observed; and secondly, the 6 inactivating somatic *BRCA1* mutations were always accompanied by LOH. Interestingly, the mechanism involving DNA methylation (n = 26) occurs more than 4 times as frequently as the mechanism involving somatic mutations (n = 6), which strongly substantiates the crucial role of promoter hypermethylation as a *BRCA1* epigenetic silencing event in TNBCs. Fascinatingly, the inactivating *BRCA1* mutations were completely mutually exclusive with promoter hyper MARs suggesting that a mechanism utilising these two events as distinct hits is not preferred.

4.5.2 Identification of potential oncogenes

Next, in order to identify potential oncogene candidates in breast cancer, genes significantly upregulated in ER+ and ER- tumours (vs. normal tissues, separate analysis for ER+ and ER-) were explored. The proportion of tumours with DNA functional methylation (x-axis) and copy number amplification events (y-axis) were plotted for upregulated genes in ER+ and ER- tumours (Figure 4.10). Genes that were recurrently (in more than 25% tumours) altered due to MARs or CNAs or both were identified. However, although, amplifications induce the activation of a larger number genes compared to DNA methylation in both ER+ and ER- tumours, (Section 4.3), a higher number of upregulated genes had recurrent MARs (ER+ = 13.9%; ER- = 11.5%; Figure 4.13, purple and green points) than copy number amplifications (ER+ = 1.6%; ER- = 1.1%; Figure 4.13, blue dots and green points). This contradictory observation is likely explained by the fact that i) copy number amplifications (Total: 4+ copies) are sufficient to significantly increase the expression of genes, while DNA methylation plays a weaker/ fine-tuning role; and ii) a gain of 1 copy (total: 3 copies) in a gene could still increase gene expression, but is ignored in the prevalence analysis

resulting in the reporting of a lower number of tumours harbouring activating copy number alterations. Genes intrinsically prone to copy number amplifications (Figure 4.13, blue points) include protein kinases such as *MAPKAPK2* and *NUAK2* in ER+ tumours; and downstream components of the PI3K/ AKT/ mTOR pathway such as *SQLE*, *GSDMC* and *MAL2* in ER- tumours. Conversely, genes with a propensity for DNA methylation alterations and not copy number losses (Figure 4.13, purple points) such as *ESR1* and *TFF1* were detected in ER+ tumours. Both these genes are involved in oestrogen signalling [Ariazi et al., 2006; Prest et al., 2002].

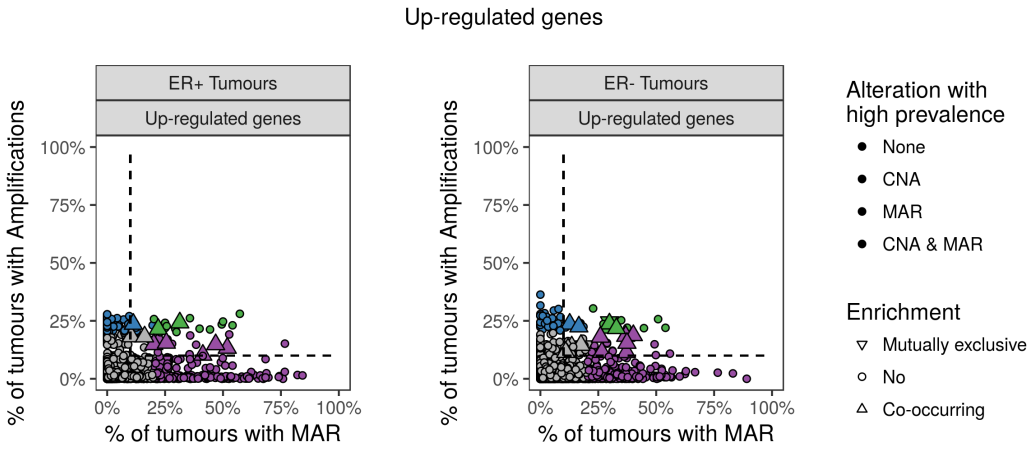


Figure 4.13: Upregulated genes in breast cancer exhibit different propensities to DNA methylation or copy number amplification. Scatter plots of the prevalence of MAR events (x-axis) against copy number amplifications (y-axis) in differentially upregulated genes in ER+ tumours (left panel) and ER- tumours (right panel). Each point represents an upregulated gene. Colours represent the propensity of the gene for DNA for either MAR events (purple), CNA events (blue), both (green) or none (grey) based on a recurrence threshold of 25%. Genes enclosed by the dotted black lines are considered for the co-occurrence and mutually exclusive analyses as described in the text. Upward triangles represent genes with co-occurring MAR and CNA events, while downward triangles represent genes with mutually exclusive MAR and CNA events. Tumours with *MED1* copy number amplifications (blue) and those with both CNA and MAR (green) have similar expression on average which is why the CNA (blue) data point is not visible.

Several genes undergoing *both* CNA amplifications and MARs in tumours were also identified. A lower threshold of 10% prevalence was used, but for *both* types of alteration (Figure 4.13, points enclosed by dotted black lines). These genes were also investigated for patterns of co-occurrence and mutual exclusivity for the two mechanisms -- CNA and MARs within the same gene, using a hypergeometric test. One-way ANOVA tests (see Section 4.5.1) were used to explore expression changes across 4 categories of tumours: i) tumours that were wild-type (i.e. had no MAR or

copy number loss); ii) tumours that only had MARs; iii) tumours that only had copy number losses; and iv) tumours in which concomitant MARs and copy number losses co-occurred for the same gene. This analysis was conducted separately for i) genes exhibiting co-occurrence; and ii) genes exhibiting mutual exclusivity, and is detailed in the following two sections.

4.5.2.1 Upregulated genes with co-occurring CNA and DNA methylation profiles

A similar analysis as described in Section 4.5.1.1 was conducted to identify upregulated genes with co-occurring CNA and DNA methylation profiles, separately in ER+ (n = 12) and ER- tumours (n = 16) (indicated with upward triangles in Figure 4.13; and detailed in Table 4.3a, b). These genes might have oncogenic roles in the respective subtypes and accordingly associations with gene expression were examined as described above. Figure 4.14a and Figure 4.14b display the average expression z-scores across the 4 categories of tumours defined above for the co-occurring upregulated genes. In all of the 12 co-occurring upregulated genes detected in the ER+ subtype (13 out of 16 in the ER- subtype), tumours with concomitant MARs and copy number amplification events (Figure 4.13a, b – green points) exhibited the highest gain in expression compared to other tumour categories (ANOVA *FDR p-value* < 0.05). This strongly implies that the two independent mechanisms augment each other's activating roles and their combination provides a selective advantage for the tumour. These 25 genes (ER+ = 12, ER- = 13) are put forward as candidate oncogenes. Of note are *INTS8* (ER+ tumours, Figure 4.10e), a key constituent of the RNA polymerase II mediated transcription machinery which has been shown to be highly mutated in several cancers [Federico et al., 2017]; and *CDK12* (ER- tumours, Figure 4.10f), a regulatory kinase which has been shown to promote breast cancer cell invasion [Tien et al., 2017].

4.5.2.2 Upregulated genes with mutually exclusive CNA and DNA methylation profiles

A similar analysis as described in Section 4.5.1.2 was implemented to identify genes in which tumours harbour MARs as well as CNA amplifications in a mutually exclusive fashion. Only two such genes, *ARHGAP30* and *SLA* were identified in ER- tumours, and none in ER+ tumours (indicated with downward triangles in Figure 4.13; and detailed in Table 4.3c, d).

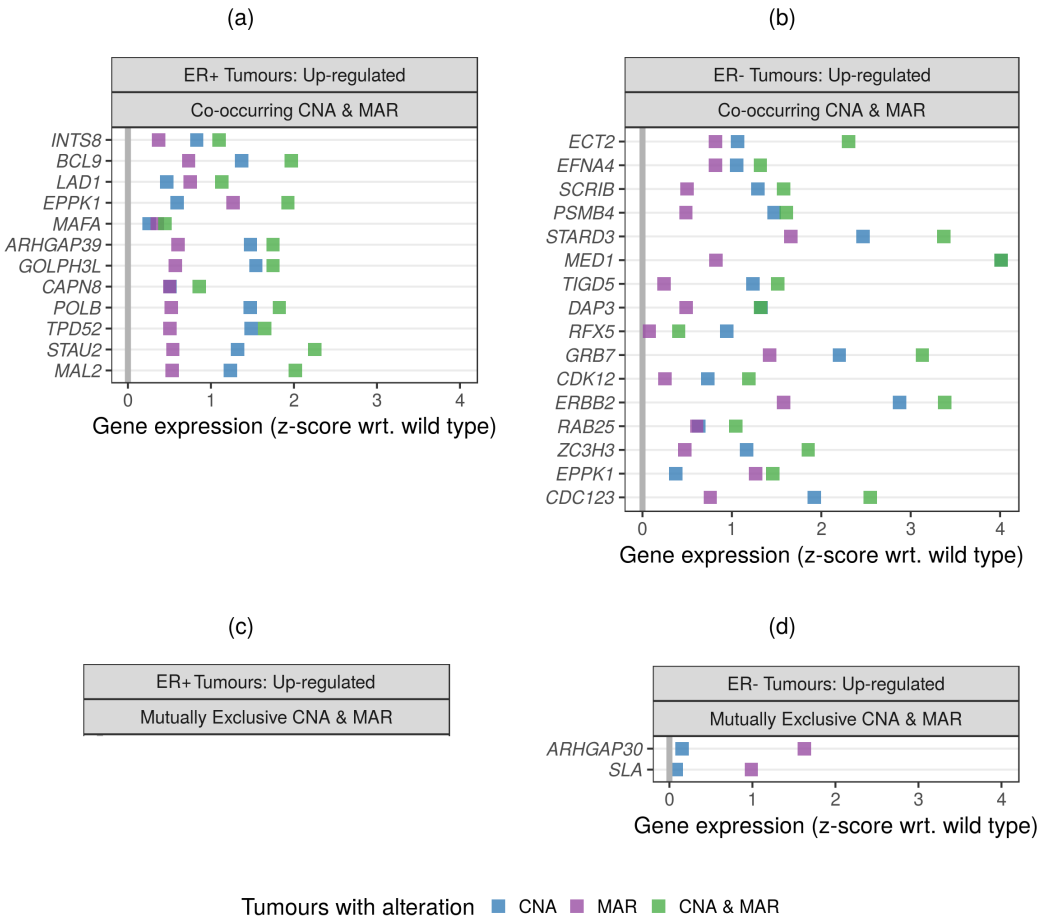


Figure 4.14: Upregulated genes in breast cancer exhibit co-occurring or mutually exclusive patterns of DNA methylation and copy number amplification. (a) All upregulated genes with significantly co-occurring MAR and CNA events within the same tumour, detected in ER+ tumours. **(b)** same as (a) but for ER- tumours. **(c)** All upregulated genes with significantly mutually exclusive MAR and CNA events within the same tumour, detected in ER+ tumours. **(d)** same as (c) but for ER- tumours. In (a) – (d), each row represents a gene. The coloured squares represent the mean z-scores of gene expression data for three categories: (blue) tumours that only had copy number amplifications; (purple) tumours that only had MARs; (green) tumours in which concomitant MARs and copy number amplifications co-occurred for the same gene. The z-scores were constructed using the mean and SD of wild type tumours as detailed in the text, and the grey vertical line at 0 represents the mean z-scores of gene expression for wild type tumours. Tumour categories with less than 5 samples were not considered. In (c) and (d), tumours with concomitant MARs and copy number amplifications are not shown since this constitutes a rare or significantly depleted category.

Gene	Feature	Methylation		CNA		Enrichment			Expression	
		Direction	(%)	CNA	(%)	Type	Extent	FDR p*	Direction	FDR p**
(a) Co-occurring genes: ER+ tumours										
INTS8	promoter	Hypo	22.3	Amplification	16.2	Co-occurring	2.21	<0.0001	Up	<0.0001
BCL9	intron	Hyper	16.2	Amplification	18.2	Co-occurring	2.16	<0.0001	Up	<0.0001
LAD1	promoter	Hypo	11.3	Amplification	23.6	Co-occurring	1.95	<0.0001	Up	<0.0001
EPPK1	intron	Hypo	25.5	Amplification	15.4	Co-occurring	1.91	<0.0001	Up	<0.0001
MAFA	promoter	Hyper	21.2	Amplification	16.6	Co-occurring	1.73	<0.0001	Up	0.0001
ARHGAP39	exon	Hyper	20.1	Amplification	14.8	Co-occurring	1.73	<0.0001	Up	<0.0001
GOLPH3L	intron	Hypo	22.0	Amplification	21.4	Co-occurring	1.72	<0.0001	Up	<0.0001
CAPN8	exon	Hypo	31.3	Amplification	24.2	Co-occurring	1.65	<0.0001	Up	<0.0001
POLB	intron	Hypo	41.3	Amplification	10.3	Co-occurring	1.59	<0.0001	Up	<0.0001
TPD52	intron	Hypo	46.8	Amplification	14.8	Co-occurring	1.58	<0.0001	Up	<0.0001
STAU2	exon	Hypo	51.8	Amplification	13.3	Co-occurring	1.56	<0.0001	Up	<0.0001
MAL2	intron	Hypo	10.4	Amplification	18.2	Co-occurring	1.54	0.0217	Up	<0.0001
(b) Co-occurring genes: ER- tumours										
ECT2	exon	Hypo	10.9	Amplification	11.5	Co-occurring	3.05	0.0055	Up	<0.0001
EFNA4	promoter	Hypo	11.9	Amplification	12.9	Co-occurring	2.80	0.0003	Up	<0.0001
SCRIB	exon	Hypo	16.6	Amplification	22.5	Co-occurring	2.50	<0.0001	Up	<0.0001
PSMB4	promoter	Hypo	17.5	Amplification	14.5	Co-occurring	2.35	0.0011	Up	<0.0001
STARD3	promoter	Hyper	25.9	Amplification	18.0	Co-occurring	2.32	<0.0001	Up	<0.0001
MED1	intron	Hypo	37.4	Amplification	15.4	Co-occurring	2.22	<0.0001	Up	<0.0001
TIGD5	exon	Hypo	12.7	Amplification	23.7	Co-occurring	2.18	0.0009	Up	<0.0001
DAP3	intron	Hypo	10.5	Amplification	11.8	Co-occurring	2.12	0.0492	Up	<0.0001
RFX5	exon	Hyper	13.5	Amplification	13.9	Co-occurring	2.11	0.0136	Up	<0.0001
GRB7	promoter	Hypo	32.8	Amplification	21.8	Co-occurring	2.08	<0.0001	Up	<0.0001
CDK12	promoter	Hypo	36.8	Amplification	18.4	Co-occurring	2.05	<0.0001	Up	<0.0001
ERBB2	exon	Hypo	40.1	Amplification	18.8	Co-occurring	1.92	<0.0001	Up	<0.0001
RAB25	intron	Hypo	26.6	Amplification	11.5	Co-occurring	1.83	0.0096	Up	<0.0001
ZC3H3	exon	Hyper	29.8	Amplification	23.8	Co-occurring	1.65	0.0008	Up	<0.0001
EPPK1	intron	Hypo	29.8	Amplification	22.0	Co-occurring	1.60	0.0018	Up	<0.0001
CDC123	intron	Hypo	37.1	Amplification	11.4	Co-occurring	1.51	0.0449	Up	<0.0001
(c) Mutually exclusive genes: ER+ tumours										
(d) Mutually exclusive genes: ER- tumours										
ARHGAP30	promoter	Hypo	19.0	Amplification	14.1	Exclusive	0.24	0.0390	Up	<0.0001
SLA	promoter	Hypo	29.9	Amplification	24.5	Exclusive	0.28	<0.0001	Up	<0.0001

Table 4.3: Upregulated genes in breast cancer exhibit co-occurring or mutually exclusive patterns of DNA methylation and copy number amplification. (a) All upregulated genes with significantly co-occurring MAR and CNA events within the same tumour, detected in ER+ tumours. **(b)** same as (a) but for ER- tumours. **(c)** All upregulated genes with significantly mutually exclusive MAR and CNA events within the same tumour, detected in ER+ tumours. **(d)** same as (c) but for ER- tumours. The direction and prevalence (%) of the MAR event and the direction and prevalence (%) of the CNA event (only copy number amplifications considered) were detailed for each gene. The direction and extent of enrichment (observed/ expected) and *FDR p-value* are indicated based on the hypergeometric test. For (a) and (b), the co-occurring genes are ordered by decreasing enrichment. The direction of expression change and the *FDR p-value* based on the one-way ANOVA of the mean expression z-scores across the 4 categories of tumours are indicated. For (c) and (d), the mutually exclusive genes are ordered by increasing enrichment (decreasing depletion). The direction of expression change and the *FDR p-value* based on the one-way ANOVA of the mean expression z-scores across the 4 categories of tumours are indicated. FDR p* = *FDR p-value* from the hypergeometric test; FDR p** = *FDR p-value* from the ANOVA.

4.6 Discussion

The copy number and mutational landscapes have been studied extensively in breast cancer [Cancer Genome Atlas Network, 2012; Ciriello et al., 2013; Curtis et al., 2012; Nik-Zainal et al., 2016; Pereira et al., 2016], but relatively little is known about the contribution of DNA methylation events and how they interact with genomic events. In order to reveal the molecular multi-omic basis of breast cancer, a considerably large number of tumours needs to be analysed due to the substantial heterogeneity of the disease. The availability of copy number, gene expression, mutations and now DNA methylation data across ∼1200 primary breast tumours in METABRIC has now enabled a large systematic genome-wide exploration into the distinctive roles of epigenetic and genetic events in breast cancer.

Comparing the contributions of CNA and DNA methylation events breast cancer revealed a poignant juxtaposition in the regulatory roles of CNA and DNA methylation. CNA played a stronger role in driving the expression of key protagonists of cell cycle progression leading to the divergence of tumours from normal tissues. Conversely, silenced genes in breast tumours were significantly enriched for DNA methylation alterations, and not *cis* genomic loss corroborating a previous report investigating multiple cancers [Teschendorff et al., 2016b]. Moreover, focusing on the top 2000 variably expressed genes in breast cancer as well as the set of genes differentially expressed genes between the two ER subtypes established that DNA methylation was the preferred mechanism for refining subtype-specific differences. For instance, expression of genes specifically linked to the oestrogen signalling (overexpressed specifically in ER+ tumours) as well as TP53 signalling (downregulated in ER- tumours) were demonstrated to be modulated by DNA methylation alterations. Conversely, the role of gene amplifications or deletions appears to be much more modest in driving inter-tumour heterogeneity.

A key finding in this chapter was that not all CNA events have the anticipated consequences in gene transcription, and other mechanisms such as DNA methylation could be used as a mechanism to modulate (diminish or enhance) the effect of CNA on the breast cancer transcriptome. Although the detection of copy number amplifications has become routine in breast cancer [Curtis et al., 2012], overexpression of some genes within these large amplification domains might not be beneficial for the cancer cells. DNA methylation was demonstrated as a potential genomic amplification diminishing agent to selectively inhibit the expression of these genes. This can lead to the identification of putative tumour suppressors (such as *THSZ2)* as well as

genes that are toxic for breast cancer. Multiple genes were identified in which DNA methylation events modulated the role of CNA. Striking examples include gene body methylation acting as a diminishing agent in *GATA3* resulting in the subtype-specific downregulation of the gene in ER- tumours; and the CNA enhancing effect of DNA methylation in *FOXA1* leading to its upregulation in ER+ tumours.

The comparison of the prevalence of functional methylation and copy number events in differentially expressed genes in breast tumours (versus the normal tissue) demonstrated that probabilistically, DNA methylation is a more likely mechanism than CNA in deregulation of key cancer genes. Silenced and overexpressed genes exhibiting a higher propensity for DNA methylation than CNA events, and vice versa, were identified in ER+ and ER- tumours. Fascinatingly, this also led to the identification of genes that were recurrently silenced ($\geq$25% tumours) in both ER+ and ER- tumours, but using distinct mechanisms. For instance, *STAT5A* and *HLF* (both genes were MAR-silenced in ER+ tumours, but CNA-silenced in ER- tumours); and *CDH13* (CNA-silenced in ER+ tumours, MAR-silenced in ER- tumours).

However, it is important to note that the associative statistical models presented here can only demonstrate *inferred function* of these epigenomic and genomic events in silencing or over-expressing key genes (as discussed in Chapter 3). Functional and/ or clinical investigation is required for providing definite evidence of the *causal function* of individual alterations in genes, and for labelling the targeted genes as tumour suppressors or oncogenes [Stricker et al., 2016]. Nevertheless, the analysis presented here provides convincing evidence that DNA methylation is a superior indicator of the underlying regulatory activity of the gene than CNA.

The principal objective of quantifying the prevalence of MAR and CNA events was to identify genes that demonstrated significant configurations of mutual exclusivity or co-occurrence in these two events for the same gene and in the same tumour. Only 2 genes exhibiting CNA and MAR events in a mutually exclusive fashion were identified in the context of upregulated genes. Along with the previously noted strong role of CNA in upregulating genes, this indicates that the occurrence of only MAR events in the breast tissue may not be dominant enough to sufficiently overexpress the gene since in the absence of copy number gains (or amplifications), DNA methylation can only act on a maximum of two alleles, This is why mutually exclusive patterns are rarely observed in the context of upregulated genes in breast cancer. Conversely, in the context of silenced genes, 15 genes (ER+ = 7, ER- = 8) with mutually exclusive patterns for MAR and copy number amplifications were detected, suggesting that MAR and copy number losses represent two alternative mechanisms for silencing.

This also indicates that for these genes, one alteration (either MAR or copy number loss) is sufficient for silencing, or the second alteration is a disadvantage for the cell. Consequently, tumours with both hits will be observed less frequently than expected by chance. A previous effort to detecting silenced transcription factors with mutually exclusive promoter hypermethylation and copy number loss events in multiple cancers only revealed one gene each in 4 out of the 6 cancer datasets investigated [Teschendorff et al., 2016b]. The identification of 15 subtype-specific genes in the METABRIC dataset, demonstrates the benefit of i) evaluating gene body methylation in addition to promoter methylation and of ii) having high power in integrative molecular analyses.

Detection of co-occurrence patterns strongly imply that the two independent mechanisms, CNA and DNA methylation, augment each other's deregulating roles providing a selective advantage for the tumour. Genes demonstrating classical tumour suppressor behaviour as described by Knudson [1971] with LOH (and a MAR event on the present allele(s)) included *PER3* in ER+ tumours and *TBX3* in ER- tumours. Interestingly, investigating the mutational profile of *TBX3,* revealed that tumours with LOH and inactivating mutations were also associated with a lower expression than tumours with only inactivating mutations. *BRCA1* also demonstrated patterns of co-occurrence (between CNA and MAR events, and between CNA and mutations), as well as mutual exclusivity (between MAR and mutations) in TNBC tumours. The co-occurrence analysis also revealed genes such as *INTS8* (ER+ tumours) and *CDK12* (ER- tumours) that showcased concomitant MAR and amplification events associated with high expression. Both genes have previously shown to have oncogenic behaviour in breast cancer [Federico et al., 2017; Tien et al., 2017]. Therefore, the detection of non-random patterns of CNA and MAR events in tumours in the same gene has enabled the identification of potential tumour suppressors and oncogenes in breast cancer in which DNA methylation has a functional role.

However, a caveat to the mutually exclusivity and co-occurrence analysis is that mutation events were ignored in this analysis. Consequently, genes in a tumour which seemed to harbour only one event (among DNA methylation or copy number alterations), may also be affected by somatic (or germline) mutations. As a result, many LOH events that occur due to alternate combinations of genetic and epigenetic are not captured in this analysis. Accordingly, one of the follow-up goals of this project is to conduct a similar analysis on the prevalence of DNA methylation aberrations, copy number alterations and somatic mutational events across all genes available in the METABRIC dataset [Pereira et al., 2016], as well as in the TCGA breast cancer dataset [Cancer Genome Atlas Network, 2012].

In a similar vein, mutual exclusivity analysis can also be conducted between *different genes* (by any of the genetic or epigenetic events profiled) rather than on *the same gene*. This strategy can be used to identify sets of genes belonging to the same molecular pathway that are recurrently altered in a mutually exclusive fashion. For instance, the Mutual Exclusivity Modules (MEMo) algorithm [Ciriello et al., 2012] was implemented in the TCGA breast cancer dataset [Cancer Genome Atlas Network, 2012] to identify gene modules exhibiting mutually exclusive alterations across the mutational and copy number landscape. This effort revealed several gene modules including the receptor Tyrosine kinase (RTK)–PI(3)K signalling pathway; the p38-Mitogen-Activated Protein Kinase (MAPK) signalling pathway; and the TP53 signalling pathway. Including DNA methylation alterations into such analysis is one of the future objectives of this project, and this promises to provide new insights into the pathogenic mechanisms of breast cancer.

Chapter 5

The role of epiclonal dynamics in tumour evolution

Contents

5.1 Introduction

Historically, the main focus of cancer evolution has been on genetic heterogeneity. Numerous studies have revealed the contribution of genetic intratumour heterogeneity in clonal evolution in breast cancer [Nik-Zainal et al., 2012; Shah et al., 2012; Wang et al., 2010], and other cancers [Cooper et al., 2015; Gerlinger et al., 2012; Navin et al., 2011], as well as its association with poor clinical outcomes [Diaz Jr et al., 2012; Merlo et al., 2010]. For example, genetic clonal diversity measures were shown to predict neoplastic progression to oesophageal adenocarcinoma [Maley et al., 2006], and more recently, deep sequencing analysis revealed that the presence of subclonal driver mutations was an independent risk factor for rapid disease progression in chronic lymphocytic leukaemia (CLL) [Landau et al., 2013].

However, in addition to genetic alterations such as somatic mutations, epigenetic aberrations are also heritable modifications that have been shown to be drivers of tissue-specific phenotypic differences [Baylin and Jones, 2011; Esteller, 2008, Chapter 3 and 4]. In 2006, Feinberg et al. [2006] proposed that epigenetic modifications can disrupt stem and progenitor cells through the aberrant expression of *tumour progenitor genes*. These early stage epigenetic cancer alterations were hypothesised to drive epigenetic plasticity, increase the probability of genetic mutations and fuel tumour progression. Darryl Shibata's group were one of the first to analyse epigenetic intratumour heterogeneity (ITH) by investigating DNA methylation patterns at neutral loci in colorectal cancer [Kim et al., 2005; Siegmund et al., 2009]. These loci were assumed to have no functional consequences and unlikely to be under selective pressures, and thus could serve as a molecular mitotic clock to provide an estimate of the cell divisions (see Chapter 1). A follow-up study utilised deep methylation sequencing to examine the molecular heterogeneity between and within individual colorectal cancer glands and simulate the evolutionary history of the tumour [Sottoriva et al., 2013]. This analysis revealed high ITH at the DNA methylation level, and advocated the use of neutral genomic features in understanding the evolution of cancers [Mazor et al., 2016].

Moreover, genetically similar tissues have exhibited variations in phenotypic outcomes such as tumour propagation potential, survival and response to therapy, likely indicating the influence of the epigenome to tumour evolution [Kreso et al., 2013; Shaffer et al., 2017; Sharma et al., 2010], and the necessity for the comprehensive assessment of ITH at the epigenetic level.

Recent studies have used genome-wide DNA methylation microarrays to examine multiple samples of primary tumour and metastatic sites in prostate [Aryee et al., 2013; Brocks et al., 2014] and brain cancers [Mazor et al., 2015]. These reports revealed extensive spatial DNA methylation heterogeneity in the solid tumours, and successfully constructed phylo-epigenetic trees using distance-based clustering of the epigenetic signatures. Interestingly, in all three studies, evolutionary histories inferred from genome-wide DNA methylation signatures were highly similar to those constructed from CNA [Aryee et al., 2013; Brocks et al., 2014] and somatic mutations [Mazor et al., 2015]. Multiple spatial sample analysis conducted in breast cancer also showed substantial clonal epigenetic heterogeneity within tumours, although this study was not genome wide, and only focused on promoter hypermethylation of 24 established tumour suppressor genes [Moelans et al., 2014].

Sampling multiple regions within a tumour is a powerful approach that has been used to determine its evolutionary history as described above. However, inferring clonal dynamics of the epigenome using single samples is still in a formative stage. One reason for this is that conventional DNA methylation analysis such as microarrays estimate aggregate methylation values at each CpG site from the population of cancer cells over a tumour sample. While these technologies easily allow comparison between multiple samples of a tumour, the evaluation of heterogeneity of individual single cells within a tumour is not possible. However, the advent of bisulfite sequencing technologies such as WGBS and RRBS, has enabled the assessment of epigenetic allele (or epiallele) compositions at single molecules (reads) derived from the population of cancer cells in a tumour. In a landmark paper, Landan et al. [2012] quite elegantly introduced the concept of epigenetic polymorphism or epipolymorphism of a given locus as the "probability that two randomly sampled DNA molecules from the cell population differ in their methylation pattern". They used this epipolymorphism scores to explore the dynamics of the process driving methylation change in cancer tissues, to distinguish methylation changes that occur in a noisy and stochastic fashion versus those that are occur in a deterministic manner. Moreover, the epiallelic diversity within the tumour also provides a measure of overall ITH, and Pan et al. [2015] leveraged the aggregate epipolymorphism scores in an RRBS analysis of diffuse large B cell lymphoma (DLBCL) to demonstrate that patients who relapse within 5 years have higher epigenetic ITH at diagnosis than those who do not relapse. Subsequently, Landau et al. [2014] developed the notion of proportion of discordant reads (PDR), as an alternate measure of intratumour DNA methylation heterogeneity, where increased PDR levels were reflective of stochastic heterogeneity in tumour samples [Landau et al., 2014]. Through RRBS analysis of CLL, they established that patients with

higher ITH were associated with lower remission times. Furthermore, the high PDR levels within an individual lymphoma were highly correlated with the number of subclonal mutations, further demonstrating the link between epigenetic and genetic ITH.

Although these reports have attempted to track the evolution of DNA methylation in tumours by matching DNA methylation ITH scores (such as epipolymorphism and PDR) in tumours at distinct time points such as diagnosis and relapse [Pan et al., 2015], Li et al. [2014] argued that a disparity (or similarity) in epipolymorphism measures between two time points does not necessarily reveal the underlying shift in epigenetic clonality. They developed an algorithm called *methclone* to explicitly compare the epiallelic distributions of a tumour at two distinct disease stages using a combinatorial entropy change calculation to identify loci that alter significantly in epiallelic compositions. In a follow-up study by the same group, this method was applied to a cohort of 138 acute myeloid leukaemia (AML) patients with paired observations at diagnosis and relapse, and they observed that epiallelic compositions varied considerably during disease progression, and that a high degree of epigenetic allelic burden within a tumour was linked with adverse clinical outcomes [Li et al., 2016b].

While these initial results in leukaemias and lymphomas are promising, it is yet unclear whether DNA methylation ITH measures are associated with adverse outcomes in solid tumours [Mazor et al., 2016]. A very recent report utilising RRBS on 140 Ewing sarcomas indicated substantial epigenetic heterogeneity (PDR) within tumours, however, the PDR levels were not significantly associated with metastatic disease [Sheffield et al., 2017]. To the best of our knowledge, the above-mentioned study in Ewing Sarcoma is the only comprehensive (> 50 tumours), genome-wide assessment of epigenetic ITH in solid tumours, though RRBS (or adapted methods) methylomes in prostate cancer [Lin et al., 2013] and chondrosarcoma [Lu et al., 2013] have been made available with approximately 20 tumours each in the two cohorts. Consequently, the role of methylation disorder in tumour evolution in epithelial tumours including breast cancer is extremely poorly understood. This is in part due to the fact of high cost and complexity of genome-wide bisulphite sequencing technologies (such as RRBS, WGBS), but also due to the relative difficulty in procuring longitudinal solid tumour samples compared to blood-based cancers.

5.1.1 Summary of aims

Chapters 3 and 4 investigated the inter-tumour DNA methylation heterogeneity in breast cancer. In this chapter, the 1482 next-generation sequencing(NGS) breast cancer and 237 normal methylomes (part of the METABRIC cohort) were reanalysed to provide the first genome-wide assessment of the role of epigenetic *intra*-tumour heterogeneity in breast cancer, and the largest for any single cancer. This is achieved through the following steps:

1. Characterisation of the epigenetic intratumour heterogeneity within breast tumours to explain the dynamics underlying methylation changes in breast cancer tissues.

2. Identification of loci with significant epiallelic compositional changes involved in the initiation and progression of breast tumours from normal tissues. Investigation of the association of these epigenetically shifted loci with gene expression alterations in tumours.

3. Examination of the link between ITH measures inferred from genetic and epigenetic profiles in breast cancer.

4. Investigation of the prognostic potential of DNA methylation ITH and dynamic epiallelic composition shifts in breast cancer.

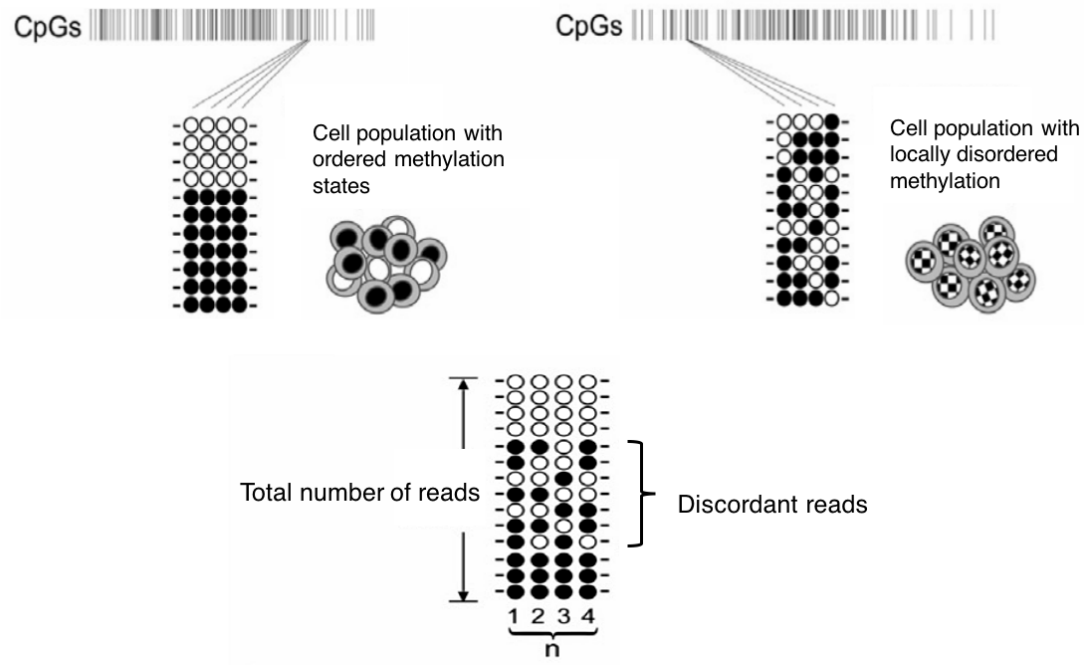
5.2 Intratumour DNA methylation heterogeneity

The previously established PDR score [Landau et al., 2014] was used to estimate intratumour heterogeneity of each sample (tumours and normal) from their DNA methylation profiles.

5.2.1 Calculation of the PDR score

RRBS, being a NGS-based technology, is a powerful tool for examining intratumour heterogeneity, and each sequencing read captures the DNA methylation information of one allele of a single cell derived from the population of cells in the tumour. If there are 4 consecutive CpG sites that are close enough to be captured by a single read, then given that DNA methylation is a binary mark, there are 2^4 or 16 possible methylation patterns, henceforth known as epialleles. Groups of 4 adjacent CpGs that were frequently covered (at least 20$\times$ coverage) by the same read are identified as loci of interest. The nature of RRBS data makes it an optimal technique for the identification of such loci since the sequencing reads naturally tend to start and end at the same DNA coordinates. By reanalysing RRBS data at the read level, the pattern of epialleles at a locus can be detected which gives a snapshot of the distinct cellular subpopulations within the tumour. For instance, two 4-CpG loci that have approximately identical methylation estimates (calculated as the proportion of methylated CpGs – black circles), conversely, showed distinct patterns of methylation ITH. The first locus is compatible with a mixture of cell populations with clear but distinct methylation states, and lower epigenetic ITH (Figure 5.1a). In contrast, the second locus, is compatible with a mixture of cell populations with locally disordered methylation resulting in higher ITH (Figure 5.1a).

The PDR (proportion of discordant reads, Figure 5.1b) score has been proposed as a measure of DNA methylation ITH, where high PDR levels were reflective of locally disordered DNA methylation in a CpG locus, and low values represent homogeneous DNA methylation patterns indicative of a selective process. By reanalysing bisulfite-sequencing (RRBS) data at the read level on all 4-CpG loci (using the criteria defined above) per tumour, sequencing reads were characterised into two categories: a) discordant reads, defined as those that contain both methylated and unmethylated CpG sites; and b) concordant reads, those that contain only methylated or unmethylated CpG sites. The proportion of discordant reads was calculated for each CpG locus as described in Landau et al. [2014]. Since the PDR levels are dependent on DNA methylation levels [Landan et al., 2012; Landau et al.,



$$\textbf{DNA methylation for 4 CpG locus} = \frac{\textbf{Number of methylated CpG } \textit{sites}}{\textbf{Total number of CpG } \textit{sites}}$$

$$\textbf{PDR score for 4 CpG locus} = \frac{\textbf{Number of discordant } \textit{reads}}{\textbf{Total number of } \textit{reads}}$$

Figure 5.1: Proportion of discordant reads is a measure of intra-sample methylation heterogeneity. (a) Methylation patterns from two distinct 4-CpG loci. The locus on the left is compatible with a mixture of cell populations with clear but distinct methylation states, and lower epigenetic ITH. In contrast, the locus on the right is compatible with a mixture of cell populations with locally disordered methylation resulting in higher ITH. **(b)** DNA methylation and PDR scores for a 4-CpG locus were calculated using the given formulas. In both panels, each row denotes methylation information for the 4-CpG locus captured by a single read representing one cell from the sample. Black circles = methylated CpG sites. White circles = unmethylated CpG sites. PDR = Proportion of Discordant reads. ITH = Intratumour heterogeneity. Figures modified from Landau et al. [2014].

2014], all loci were binned based on average methylation level into 21 bins spanning 0–100% methylation levels. The PDR for each 4-CpG loci (black dot) is displayed as a function of methylation for a representative ER+ tumour (Figure 5.2). The PDR levels are clearly dependent on the underlying DNA methylation levels. Loci with methylation proportions at the two extremes (0% and 100%) are characterised by low PDR scores. Conversely, loci with intermediate methylation can be achieved via several methylation patterns, and consequently have higher PDR values. For each sample, median PDR values were calculated across the 21 bins spanning DNA methylation levels from 0% to 100% (red line). A sample-specific PDR score was calculated by determining the area under median PDR line.

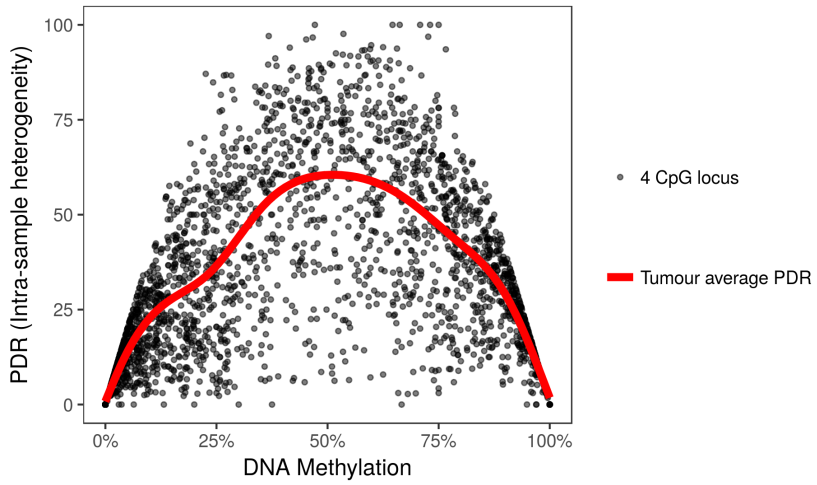


Figure 5.2: PDR scores are dependent on underlying DNA methylation levels. PDR scores and DNA methylation estimates are plotted for all 4-CpG loci obtained from a representative ER+ tumour. The PDR for each 4-CpG loci (black dot) is displayed as a function of DNA methylation. Tumour average (median) PDR values (red line) were calculated across the 21 bins spanning DNA methylation levels from 0% to 100% as explained in the text. Only 4-CpG loci with coverage greater than 30× are plotted, although 4-CpG loci with coverage greater than 20× were analysed.

5.2.2 Breast tumours have lower epigenetic intratumour heterogeneity than normal tissues

The average PDR for each sample was calculated as described above globally (averaging over all 4-CpG loci in the RRBS universe), and separately across different genomic features (such as promoters, exons, introns, enhancers and

polycomb-repressed chromatin (PRC) regions). Next, the tumours were stratified by ER status and the distribution of PDR scores is illustrated for ER+ tumours, ER- tumours and normal tissues (Figure 5.3). Notably, the breast tumours were considerably heterogeneous with respect to the PDR scores (compared to the normal tissues) suggesting there is potential to link the epigenetic state of a tumour with the underlying clonal diversity as implemented in previous studies [Landan et al., 2012; Landau et al., 2014]. On average, breast tumours had a lower average ITH than the normal tissue (Global median PDR: Tumour = 40.21, Normal = 42.76, *FDR p-value* = 8.6×10^{-33}; Wilcoxon test) which was also observed in DLBCL [Pan et al., 2015], and is consistent with a epiclonal selection process underlying tumorigenesis. This could be explained by the fact the normal mammary epithelial have diverse cell types with distinct methylation profiles. Furthermore, ER- tumours exhibited lower ITH than ER+ tumours (Global median PDR: ER+ = 40.63, ER- = 38.78, *FDR p-value* = 5.9×10^{-21}; Wilcoxon test). ER- tumours display a higher tumour grade than ER+ tumours, and therefore the lower epigenetic ITH could be interpreted as reflecting a series of recent clonal outgrowths from more diverse cell populations [Mazor et al., 2016].

Promoters, exons and enhancer regions exhibited the lowest ITH in tumours compared to other genomic features (Median Tumour PDR: promoters = 37.32, exons = 37.01, enhancers = 37.61, introns = 39.69, PRC regions = 40.98, global = 40.21), as was also observed in CLL [Landau et al., 2014]. Since exons are regions that are under a strong selective constraint, a low epigenetic ITH need not indicate a special function for CpG methylation [Cohen et al., 2011]. However, the promoters displayed a far larger ITH loss between tumours and normal tissues than exons or enhancers (median PDR loss between tumours and normal: promoters = -3.79, exons = -2.86, enhancers = -1.61, global = 2.55), and also showed the highest divergence in PDR between ER+, ER- and normal tissues (Kruskal-Wallis test). This indicates that a tumour evolves under a regime selecting for specific methylation patterns in promoters, thus implying a functional role of promoter DNA methylation in tumorigenesis.

A subset of ER+ and ER- tumours, with PDR scores that are higher than the normal tissue, are also clearly identified (outliers with high PDR scores in Figure 5.3). Interestingly, tumours with high epigenetic ITH in one genomic feature were also likely to have high epigenetic ITH (pairwise correlations of PDR scores between all genomic features > 0.85). Further characterisation of this group of tumours with disordered methylation levels (high PDR scores) is conducted in Section 5.4 and 5.5.

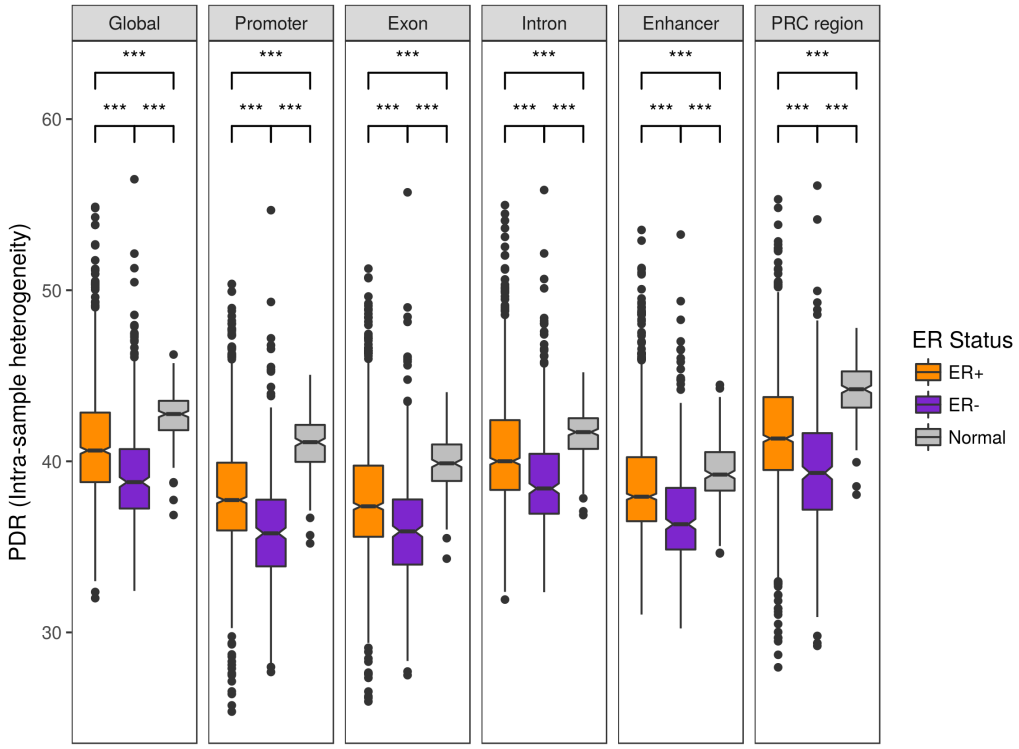


Figure 5.3: Breast tumours have lower epigenetic intratumour heterogeneity than normal tissues. Distribution of average PDR scores for ER+, ER- and normal tissues stratified by genomic feature. Separate panels represent different genomic features, and only loci that overlapped a given feature were considered (the first panel, Global, represents all loci). For each genomic feature, pairwise Wilcoxon rank-sum tests were used to compare PDR scores between these three categories. *FDR p-values* were denoted. Only 4-CpG loci at 20× coverage were considered. (. = *FDR p-value* < 0.1, ∗ = *FDR p-value* < 0.05, ** = *FDR p-value* < 0.01, *** = *FDR p-value* < 0.001, **** = *FDR p-value* < 0.0001).

5.2.3 Late replicating regions are associated with disordered methylation

Intratumour heterogeneity scores such as PDR for specific regions can also help to delineate the dynamics of the process driving methylation changes in cancer tissues from normal tissues. PDR values within CpG loci can distinguish a controlled increase in the frequency of a specific epiallele (low PDR), from multiple stochastic changes in the frequencies of many epialleles (high PDR). For each 4-CpG locus, PDR values were averaged (mean) across all 1482 tumours to give a loci-specific score for breast cancer, and these regions were stratified by time of replication (TOR) in the cell cycle (8 groups from early to late: $\geq$ 80, 70-80, 60-70, 50-60, ..., 20-30, < 20). As mentioned in Chapter 3, TOR status for the CpG sites was obtained from previously published Repli-seq experiments conducted on the Michigan Cancer Foundation-7 (MCF-7) breast cancer cell line [Pope et al., 2014]. Significantly higher PDR was observed in regions with later TOR (median PDR scores: earliest TOR group = 26.4, latest TOR group = 36.1, *p-value* $\leq 2\times10^{-16}$, Jonckheere-Terpstra test for trend, Figure 5.4). An equally strong PDR-TOR relationship (*p-value* $\leq 2\times10^{-16}$, Jonckheere-Terpstra test for trend) was detected in the background epigenome (as defined in Chapter 3), indicating that regions that replicate later in the cell cycle were prone to higher stochastic variation in methylation, an observation also reported for promoter regions in CLL [Landau et al., 2014]. This was in strong concordance with the findings in Chapter 3, in which regions with later TOR were shown to accumulate significantly higher DNA methylation related drift. However, much weaker PDR-TOR trends were observed for promoter, exonic and enhancer regions in this breast cancer methylome (data not shown), suggesting that highly conserved genes (that replicate earlier in the cell cycle) may also have high levels of ITH that coud be associated with tumorigenesis.

5.2.4 Directed-DMRs associated with concomitant expression changes harbour ordered methylation patterns.

Next, the 4-CpG loci were stratified based on whether they overlapped differentially methylated regions (DMRs) between tumour and normal tissues as defined in Chapter 3 ($\geq$ 20% absolute difference in average methylation, *FDR p-value* < 0.05). A lower degree of PDR was observed at DMRs regardless of whether they were hyper and hypo methylated compared to loci that were did not undergo methylation change in tumours (*p-value* < 2.2×10^{-16}; Kruskal-Wallis test; Figure 5.5a). This implies that

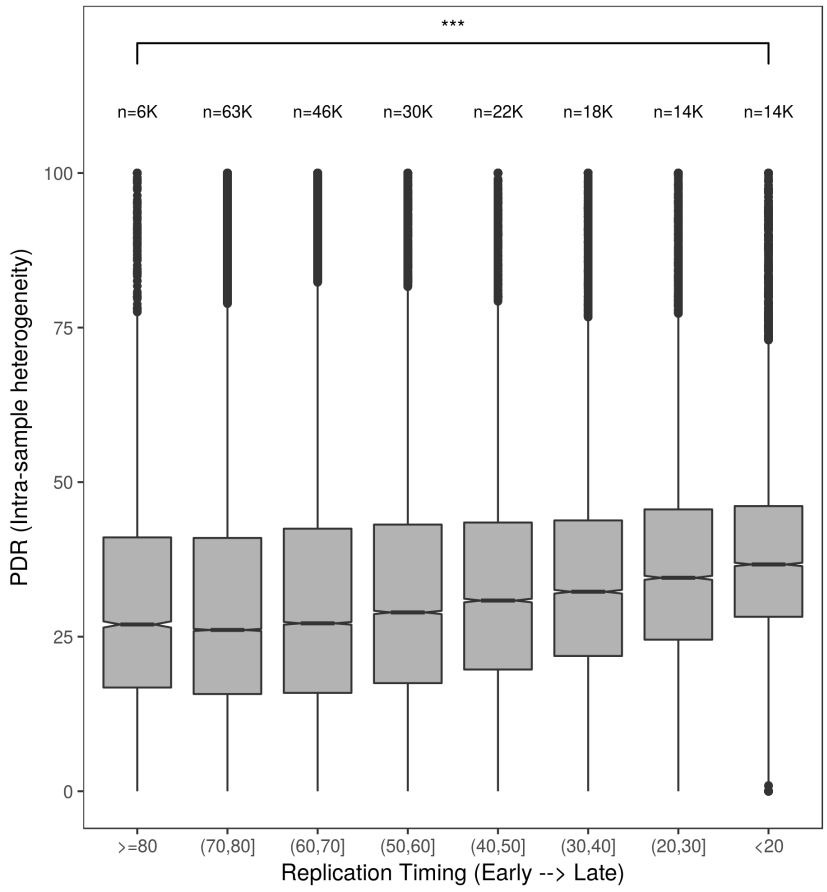


Figure 5.4: Late replicating regions are associated with disordered methylation. Distribution of PDR scores for loci within 7 time of replication (TOR) categories from early to late. The number of loci considered within each TOR category is denoted. Average PDR scores for each locus is calculated by averaging over all tumours with available information. Only 4-CpG loci at 20$\times$ coverage were considered. Jonckheere-Terpstra test for trend was used to compare the PDR scores between the seven categories. *FDR p-values* were denoted. (. = *FDR p-value* < 0.1, $*$ = *FDR p-value* < 0.05, ** = *FDR p-value* < 0.01, *** = *FDR p-value* < 0.001, **** = *FDR p-value* < 0.0001).

loci with significant deviations in tumour-averaged methylation underwent ordered methylation changes indicating positive selection, a finding that was observed in CLL as well [Landau et al., 2014].

Nevertheless, is it possible to identify DMRs that are a culmination of stochastic accumulation of methylation changes? Considerable progress in this regard was made in Chapter 3, in which aggregate (not intratumour) methylation estimates were used to characterise DMRs into i) directed-DMRs, that were hypothesised to be the result of targeted methylation differences; and ii) background-DMRs, hypothesised to be the consequence of accumulation of largely stochastic cell-division related methylation errors. In this section, intratumour PDR scores were used to further characterise regions within these two DMR categories (Figure 5.5b). B*ackground* DMRs showed a significantly stronger stochastic component than directed DMRs (median PDR: directed DMRs = 38.9, background DMRs = 42.1, $p\text{-}value < 2.2\times10^{-16}$; Wilcoxon test) which further supports this DMR classification. Furthermore, directed-DMRs associated with a concomitant alteration in gene expression (directed ex*pression*-DMRs) had relatively homogeneous methylation patterns compared to those not associated with expression changes (Figure 5.5c) (median PDR: directed expression DMRs = 35.3, directed DMRs not associated with expression = 39.1, $p\text{-}value = 8.3\times10^{-16}$; Wilcoxon test). This suggests that methylation changes with an apparent functional role in tumorigenesis (measured at the mRNA level) are more likely to be selected for and reflect an ordered methylation state across the tumour. A similar trend of lower PDR scores for directed expression DMRs was also noted when focusing on loci within promoters (median PDR: directed expression DMRs = 35.1, directed DMRs not associated with expression = 39.1, $p\text{-}value = 0.017$; Wilcoxon test).

Notably, intra-patient DNA methylation homogeneity was also observed for promoter methylation differences associated with gene expression in a multiple region analysis of metastatic prostate cancer [Aryee et al., 2013]. In conclusion, lower PDR scores within a locus are a good indicator that the methylation status of the corresponding locus has a potential functional role and has undergone positive selection.

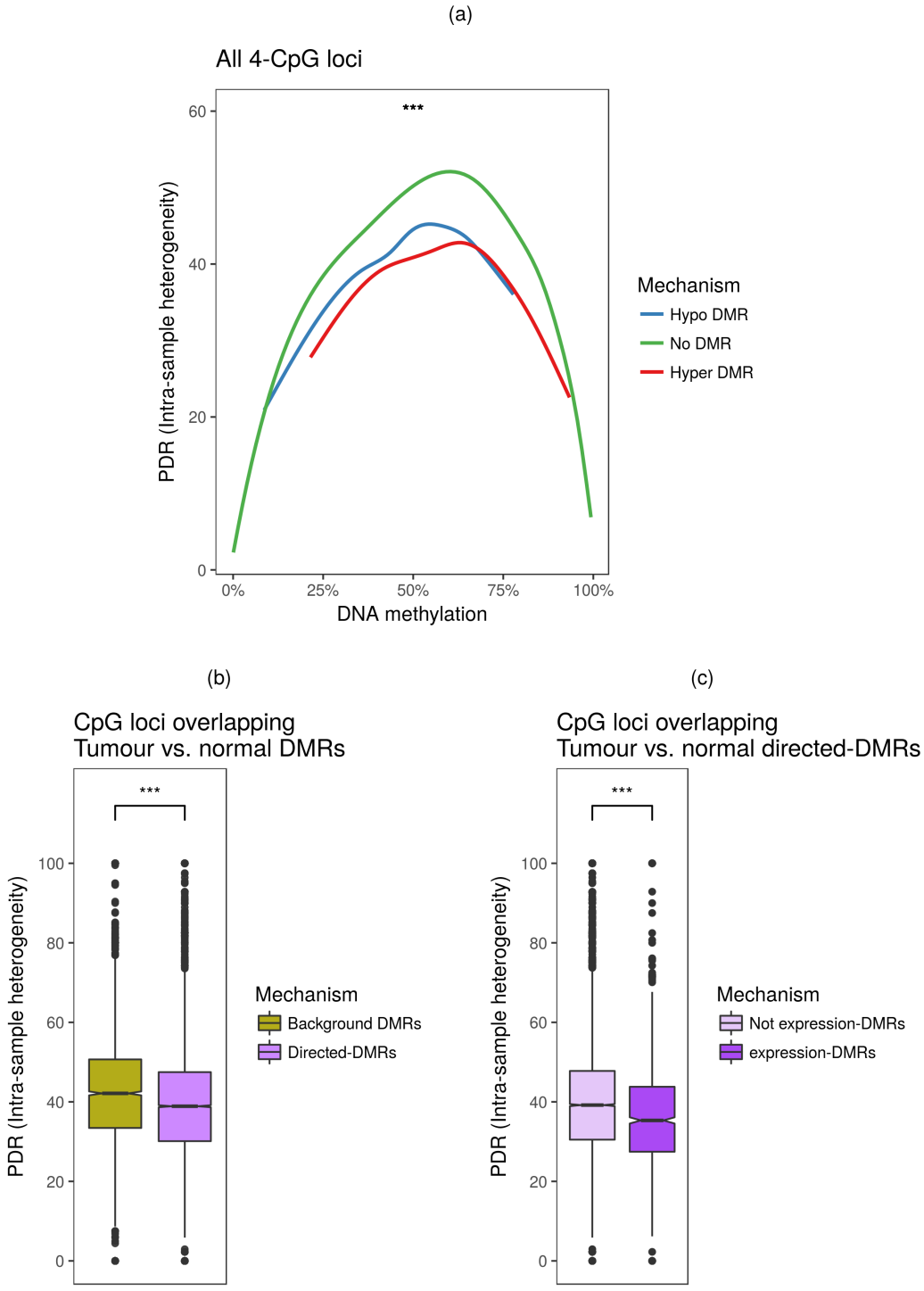


Figure 5.5: (Caption on next page.)

Figure 5.5: (Previous page.) **Directed-DMRs associated with concomitant expression changes have lower intratumour methylation heterogeneity. (a)** The relationship between average PDR scores and DNA methylation was compared for loci within hyper DMRs, hypo DMRs and for loci not overlapping DMRs. Loess curves are illustrated for each of these three categories. **(b)** Distribution of PDR scores for loci within directed-DMRs and background-DMRs. **(c)** Distribution of PDR scores for loci within directed-DMRs that are associated with expression changes and directed-DMRs not associated with expression changes. In (a) (b) and (c) average PDR scores for each locus is calculated by averaging over all tumours with available information. Only 4-CpG loci at 20$\times$ coverage were considered. In (a) Kruskal-Wallis test was used to compare PDR scores between the categories. In (b) and (c) Wilcoxon rank-sum tests were used to compare PDR scores between the two categories. *FDR p-values* were denoted. (. = *FDR p-value* < 0.1, $*$ = *FDR p-value* < 0.05, ** = *FDR p-value* < 0.01, *** = *FDR p-value* < 0.001, **** = *FDR p-value* < 0.0001).

5.3 Evolutionary dynamics of epigenetic changes in breast cancer

In order to unravel the evolutionary dynamics involved in tumorigenesis from an epigenetic perspective, an alternate method called *methclone* [Li et al., 2014] was used to measure the shifting of epiallelic compositions from the normal tissue to the breast tumour. This represents an orthogonal approach to the PDR score (which quantified the epiallelic diversity within individual samples), since it examines the transformation of epiallelic states between two samples.

5.3.1 Detection of significant epiallelic composition shifts in breast tumours relative to normal samples

The epiallelic composition shift analysis was performed using *methclone*, as described in Li et al. [2014]. Briefly, for each 4-CpG locus (minimum 20 read coverage in both the tumour and the corresponding normal tissue), the epiallele patterns of compositional changes between the tumour and the normal tissue was evaluated to calculate the combinatorial entropy change (ΔS) of epialleles (Figure 5.6). Loci with combinatorial entropy change (ΔS) < -90 were identified as undergoing significant epiallelic composition shifts (henceforth known as eloci). For each tumour, the number of eloci was normalised to the total number of CpG loci considered in the analysis, giving rise to an estimate of **E**loci **P**er **M**illion CpGs (henceforth known as EPM).

The EPM score thus represents a global estimate of epigenetic allelic burden per tumour. It is important to note that the epiallele composition shift analysis between a tumour and a normal tissue, can not only detect loci with organised changes in methylation leading to a substantial difference in mean methylation (such as those detected using a DMR analysis), but in addition can also detect differential variability in methylation. Early DNA methylation changes in normal cells at risk of neoplastic transformation and in early cancers (also called epigenetic field defects) have been shown to reflect a stochastic nature [Teschendorff et al., 2016a, 2012; Teschendorff and Widschwendter, 2012]. Therefore, in contrast to a traditional DMR analysis, the EPM analysis can also detect such stochastic and heterogeneous DNA methylation pattern changes.

As discussed in Chapter 3, the original normal tissue from which the tumours arose was not available, and instead the matched adjacent normal tissue was used

which may contain epigenetic field defects [Johnson et al., 2017; Teschendorff et al., 2016a] as well as tumour contamination, and thus reduce sensitivity in detecting eloci in the tumours. Furthermore, not all breast tumours had matched normal tissue available (only 143 tumours), and so a matched 1-to-1 *methclone* analysis as described in the original publication [Li et al., 2014] was not possible. A modified *methclone* strategy was developed[1] and implemented in the breast cancer cohort as described below. A panel of 50 reference normal breast samples (out of 237 available) was randomly selected, and each tumour was linked individually with each of these 50 normal samples leading to 74100 pairwise *methclone* analyses (1482 tumours x 50 reference normal samples). For each analysis, the list of eloci (significantly shifting epialleles) was identified and the global EPM was calculated as described earlier. Accordingly, each tumour had 50 lists of eloci and 50 EPM scores (resulting from individual comparisons with the 50 reference normal samples). For each tumour, a consensus list of eloci was determined by identifying loci that were covered by a minimum of 20 normal samples and consistently met the threshold criteria ($\Delta S < -90$ in $\geq$10% of the normal comparisons; minimum of 5 normal comparisons). Methylation profiles of normal tissues have been shown to be relatively homogeneous (Chapter 2), thus supporting the suitability of this approach. The global tumour-specific EPM score was calculated by taking the geometric mean of the 50 individual EPM scores based on the separate normal comparisons. Similarly, genomic feature specific EPMs were also calculated for each tumour by focusing the analysis on individual genomic features including promoters, exons, introns, enhancers and PRC regions.

To further, contextualise EPM shifts in breast tumours with respect to normal tissues, EPM scores were similarly calculated for the normal tissues. The remaining 187 normal samples, not part of the reference normal panel, were linked with the 50 reference normal samples leading to a further 9350 pairwise *methclone* analyses (187 normal tissues x 50 reference normal tissues). A similar consensus approach was utilised to calculate the EPM scores for the 187 normal tissues.

An alternate *methclone* strategy was also tested in which the ER status of the tumour and adjacent normal tissue was considered. Specifically, ER+ tumours were compared only with those reference normal tissues that were adjacent to ER+ tumours, rather than all 50 reference normal samples, and the same was repeated for ER- tumours. However, since methylation profiles of normal tissues have been shown to be homogeneous with respect to ER status of the adjacent tumour (Chapter 2), this

[1] This modified *methclone* strategy was developed by me as part of the PhD thesis, and discussed with the authors of the original *methclone* publication [Li et al., 2014] during a visit to their laboratories at Weill Cornell Medicine, New York.

alternate strategy did not yield dissimilar results (data not shown) and was not utilised for further analysis.

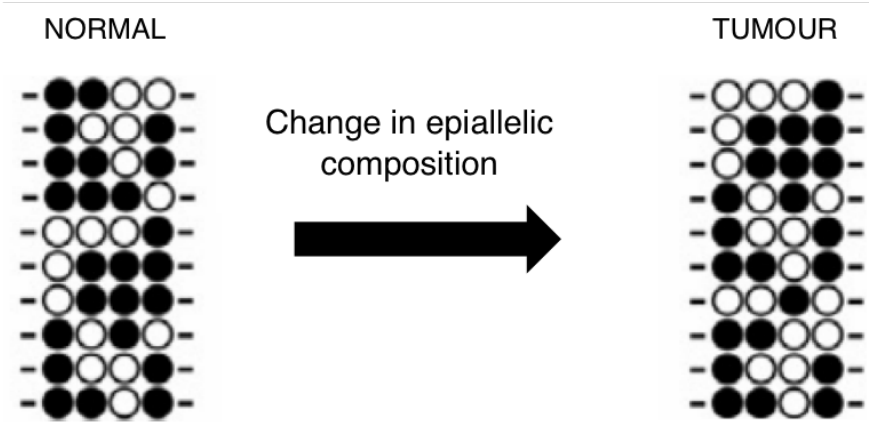


Figure 5.6: Epiallelic composition shift detection by *methclone*. Schematic plot of epiallelic composition of cells for the same 4-CpG locus in the normal tissue and in the matched tumour tissue. Each row denotes methylation information for the 4-CpG locus captured by a single read representing one cell from the sample. Black circles = methylated CpG sites. White circles = unmethylated CpG sites.

5.3.2 Breast tumours undergo subtype-specific and genome feature-specific epiallelic composition shifts

The breast tumours were stratified by ER status and the distribution of EPM scores is illustrated for ER+ tumours, ER- tumours and normal tissues globally and for specific genomic features (Figure 5.7). Whereas, EPM scores for the tumours largely represent the *degree of epiallelic burden* associated with tumorigenesis from the normal tissue, conversely, EPM scores for the normal tissues likely represents the inter-patient variation in the degree of epigenetic field defects in normal breast tissues [Li et al., 2016b]. Globally, breast tumours exhibited higher EPM scores than normal tissues (Global median EPM ($\log_{10}$): Tumour = 4.08, Normal = 3.78, *FDR p-value* = 3.0×10^{-135}; Wilcoxon rank-sum test) indicating that the epigenetic alterations associated with tumorigenesis are much more dynamic than epigenetic field defects. The number of eloci captured between the different normal samples is also very stable (less heterogeneous) suggesting that inter-patient epigenetic variation in normal breast tissues is low, an observation also noted in an independent study in AML patients [Li et al., 2014]. This further substantiates the findings in Chapter 2, that adjacent normal breast tissues have relatively homogeneous methylation profiles.

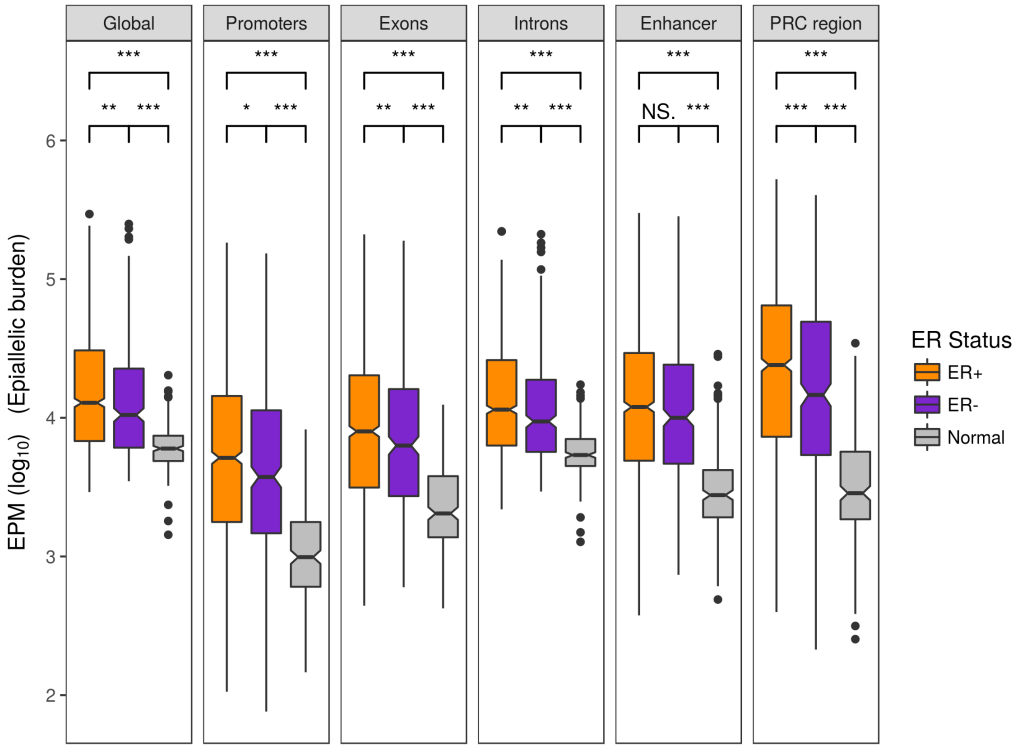


Figure 5.7: Breast tumours exhibit high epiallelic composition shifts compared to normal tissues. Distribution of average EPM scores ($\log_{10}$) for ER+, ER- and normal tissues stratified by genomic feature. Separate panels represent different genomic features, and only loci that overlapped a given feature were considered (the first panel, Global, represents all loci). For each genomic feature, pairwise Wilcoxon rank-sum tests were used to compare EPM scores between these three categories. *FDR p-values* were denoted. Only 4-CpG loci at 20× were considered. (. = *FDR p-value* < 0.1, $*$ = *FDR p-value* < 0.05, ** = *FDR p-value* < 0.01, *** = *FDR p-value* < 0.001, **** = *FDR p-value* < 0.0001).

Higher tumour EPM scores were observed across all individual genomic features as well (*FDR p-value* < 1.0×10^{-113} for all comparisons). Interestingly, higher EPMs were detected in intergenic regulatory regions such as enhancers and PRC regions compared to genic regions, with the promoters containing the lowest proportion of average number of significantly deregulated loci (Median Tumour $\log_{10}$ EPM: promoters = 3.68, exons = 3.85, introns = 4.03, enhancer = 4.05, PRC regions = 4.31). However, in tumours, promoter as well as PRC loci displayed the highest variability in EPMs (measured by interquartile range), as well as displaying large departures from the normal tissues. This suggests that there is potential to link promoter epiallele shifts in a tumour with its underlying evolutionary dynamics.

5.3.3 High epiallelic composition shifts at promoters are linked with tumour-specific gene expression changes

Given the high tumour variability in promoter EPM scores (described above) and the well-recognised and reported role of promoter methylation in regulating transcription in tumours, the presence of epiallelic composition shifts (eloci) in the promoter of a gene was investigated for concomitant alterations in its mRNA expression. For each tumour, genes were classified into whether their promoters harboured eloci or not. For all genes, log fold changes (LFC) in gene expression were calculated between the tumour and the 50 reference normal samples. The standard deviation (SD) in LFC, and the percentage of genes differentially expressed (LFC > 1) were noted for the two genes categories: i) with promoter eloci; and ii) without promoter eloci. Figure 5.8a illustrates that on average genes containing substantial shifts in epiallelic compositions at their promoters showed significantly higher variance in gene expression alterations in the tumour compared to the normal tissue (*p-value* = 1.0×10^{-14}; Wilcoxon signed rank test). In addition, genes with promoters harbouring eloci were also significantly more likely to be differentially expressed (*p-value* = 2.9×10^{-6}, adjusted for concomitant copy number alterations; Wilcoxon signed rank test, Figure 5.8b), confirming the contribution of DNA methylation dynamics at promoters in deregulating the transcriptional landscape in tumours.

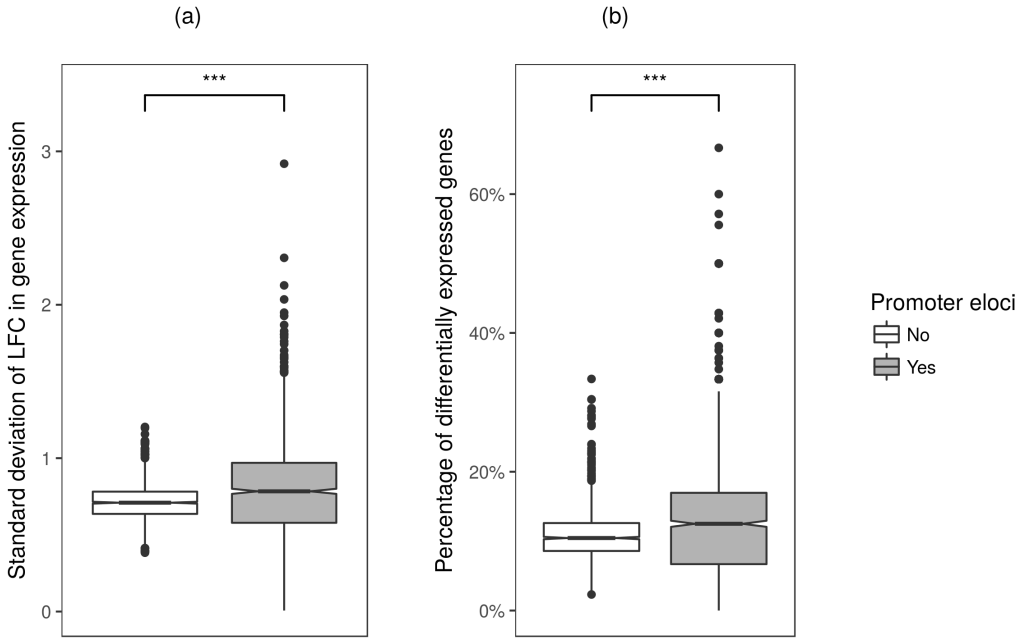


Figure 5.8: High epiallelic composition shifts at promoters are linked with tumour-specific gene expression changes. For each tumour, genes were stratified into whether they harboured an eloci in the promoter or not. **(a)** Boxplot of standard deviation of LFC in transcript expression in each tumour ($\log_2$ fold change versus 50 reference normal tissues). For each tumour this is calculated for genes with (right) or without eloci (left) in their promoters. **(b)** Boxplot of percentage of genes in each tumour that are differentially expressed (compared to 50 reference normal tissues). For each tumour, this is calculated for genes with (right) or without eloci (left) in their promoters. In (a) and (b), Wilcoxon signed rank tests (paired) were used to compare the scores between the two categories. *p-values* were denoted. Only 4-CpG loci at 20$\times$ within promoters were considered. (. = *p-value* < 0.1, $*$ = *p-value* < 0.05, ** = *p-value* < 0.01, *** = *p-value* < 0.001, **** = *p-value* < 0.0001).

5.4 PDR and EPM represent distinct properties of the epigenome

A tumour's PDR score and EPM score represent two orthogonal measures of its epigenome. The PDR score represents intra-tumour epigenetic diversity, while the EPM score represents the magnitude of dynamic epiallelic composition shifting in the tumour compared to the normal tissue. The relationship of these scores are compared with the established genetic ITH scores and the epigenetic drift derived *Accumulation index* (Chapter 3).

5.4.1 Patterns of genetic and epigenetic intratumour heterogeneity

The Integrative cluster (IntClust) classification divides breast tumours into 11 tumour subtypes based on integrating genomic (CNA) profiles with gene expression. The two epigenetic scores defined above, PDR and EPM are investigated within these clusters. Since promoter loci displayed the highest divergence in PDR scores between tumour subtypes and normal tissues (Section 5.2), further investigation of intratumour DNA methylation dynamics in this chapter is largely focused on the promoter regions. The distribution of tumour-specific promoter PDR scores across the Integrative clusters was very heterogeneous (Kruskal Wallis $p\text{-}value < 2.2\times10^{-16}$; Figure 5.9a), as were the promoter EPM scores (Kruskal Wallis $p\text{-}value < 2.2\times10^{-16}$; Figure 5.9b), suggesting that in addition to carrying distinct genetic rearrangements, these tumour subtypes were also diverse with respect to the dynamics driving methylation changes. Similar heterogeneity across the subtypes was observed using PDR scores and EPM scores inferred from 4-CpG loci in other genomic features (data not shown).

But is the level of epigenetic ITH correlated with the level of genetic ITH? The mutant-allele tumour heterogeneity (MATH) score [Mroz and Rocco, 2013], which is a tumour-specific score based on the variation in variant allele frequency (VAF) of all mutations in the tumour, was used to quantify genetic ITH. Only 173 genes were profiled for mutations, so only these genes were used for calculation of the MATH score (MATH scores for METABRIC tumours obtained from Pereira et al. [2016]). Remarkably, a very strong (but negative) correlation was observed between PDR scores and the MATH scores (rho = -0.82, $p\text{-}value = 0.0010$; spearman correlation over Integrative clusters weighted by number of tumours in each subgroup; Figure 5.9c). Stratifying by ER status also revealed significant negative correlations (ER+:

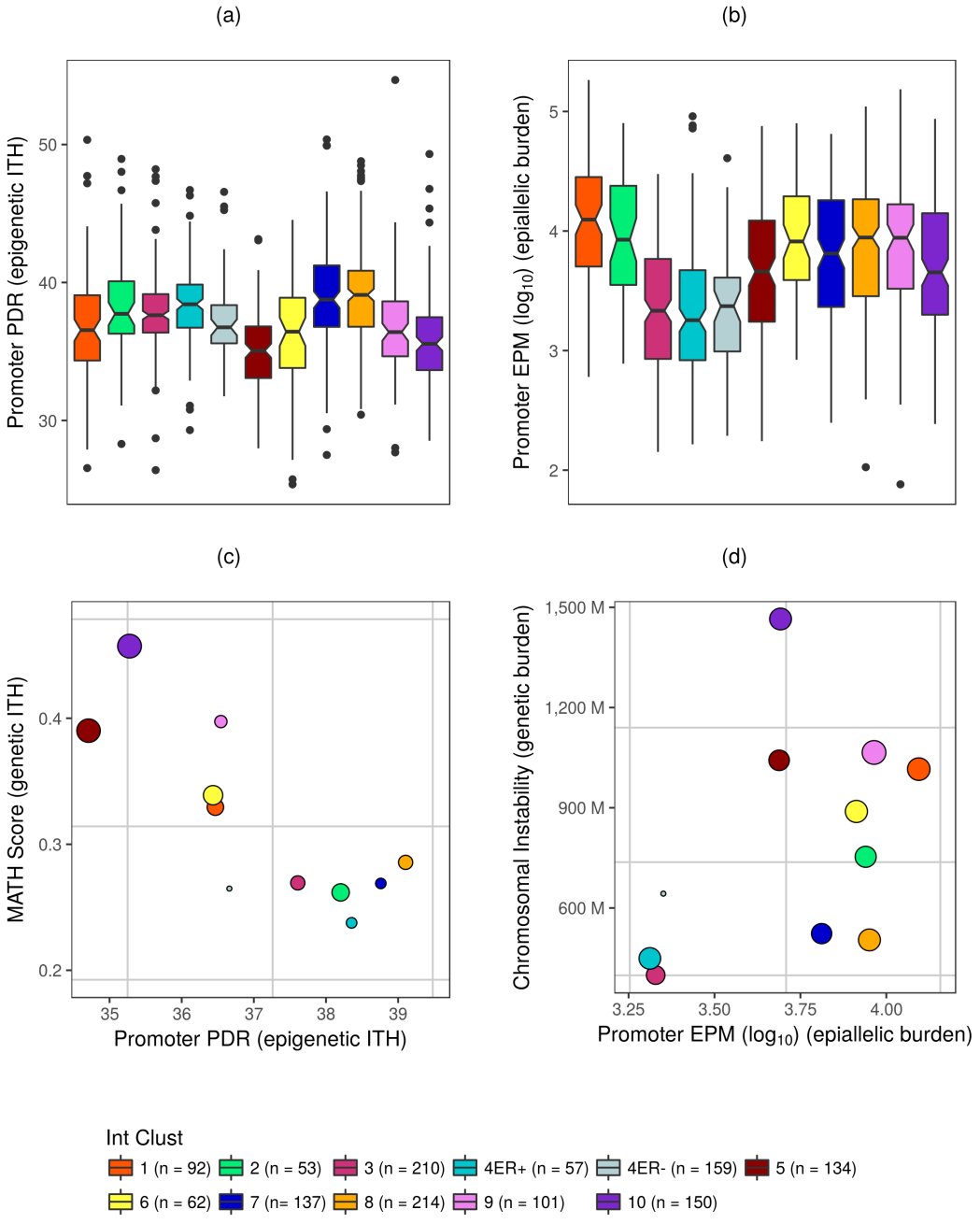


Figure 5.9: (Caption on next page.)

Figure 5.9: (Previous page.) **Patterns of genetic and epigenetic intratumour heterogeneity. (a)** Distribution of average promoter PDR scores for the 11 Integrative clusters. **(b)** Distribution of average promoter EPM scores ($\log_{10}$) for the 11 Integrative clusters. **(c)** Bubble plot of median promoter PDR scores and MATH scores for each Integrative cluster. Grey lines depict the quartiles for both scores (vertical lines, PDR quartiles; horizontal lines, MATH score quartiles) in the cohort as a whole. The areas of the circles are proportional to number of samples in each Integrative cluster. **(d)** Bubble plot of median promoter EPM scores ($\log_{10}$) and CIN scores for each Integrative cluster. In (a) (b) (c) and (d), only 4-CpG loci at 20× within promoters were considered. In (c) and (d), grey lines depict the quartiles for the two scores in the cohort as a whole. The areas of the circles are proportional to the covariance between the scores plotted in each Integrative cluster.

rho = -0.65, *p-value* = 0.0287; ER-: rho = -0.73, *p-value* = 0.0159). IntClust 10 that are predominantly Basal-like tumours and IntClust 5 tumours that are predominantly HER2+ tumours have relatively high mutation ITH [Pereira et al., 2016], and high chromosal instability [Curtis et al., 2012]. Conversely, these tumours exhibited the least ITH at the epigenetic level (Median PDR). This suggests a linked relationship between genetic and epigenetic heterogeneity, with tumours displaying high levels of genetic diversity conversely being associated with highly ordered promoter methylation patterns indicative of selection. The basis behind this relationship can be partially explained by presence of epigenetic field defects in genetically homogenous tumours.

Chromosomal instability (CIN), is also a tumour-specific score obtained by calculating the fraction of the genome altered by CNAs was also examined (CIN scores for METABRIC tumours obtained from Pereira et al. [2016]). Since, breast cancer is a copy number driven disease, CIN can be used to represent the extent of the genetic burden in the tumour. Mutational burden cannot be used since the mutational profiles of only 173 key cancer driver genes are available for these tumours. The relationship between the EPM score and the CIN score was also investigated since they represent the epigenetic and genetic burden of the tumour respectively. A weak positive correlation was observed with the EPM score (rho = 0.36, *p-value* = 0.2472; spearman correlation over Integrative clusters weighted by number of tumours in each subgroup; Figure 5.9d). However, stratifying by ER status revealed mildly significant positive correlations between epigenetic and genetic burden (ER+: rho = 0.55, *p-value* = 0.0773; ER-: rho = 0.69, *p-value* = 0.0274).

5.4.2 Epiallelic burden is correlated with epigenetic drift

Given the association of background methylation changes (background DMRs) with a higher epigenetic ITH (PDR score; Figure 5.5) denoting a stochastic-oriented process at a region-level, it might be tempting to theorise that a tumour's global PDR score is associated with accumulation metric that quantifies the extent of epigenetic drift in the tumours (calculated in Chapter 3). However, no such correlation was observed when testing the relationship between two aggregated measures -- PDR score and the Accumulation index – across all tumours (rho = 0.01, *p-value* = 0.735, partial correlation adjusted for ER status; Figure 5.10a). Similar results were obtained for PDR scores inferred only from promoter loci (rho = 0.02, *p-value* = 0.470, partial correlation adjusted for ER status), as well as PDR scores inferred from loci within the background or neutral epigenome defined in Chapter 3 (rho = 0.04, *p-value* = 0.107, partial correlation adjusted for ER status). This strongly implies that a tumour's tendency to accumulate stochastic related methylation errors is not associated with the magnitude of functional methylation changes in the promoter that have undergone positive selection, and that these two independent processes affect distinct loci within the genome.

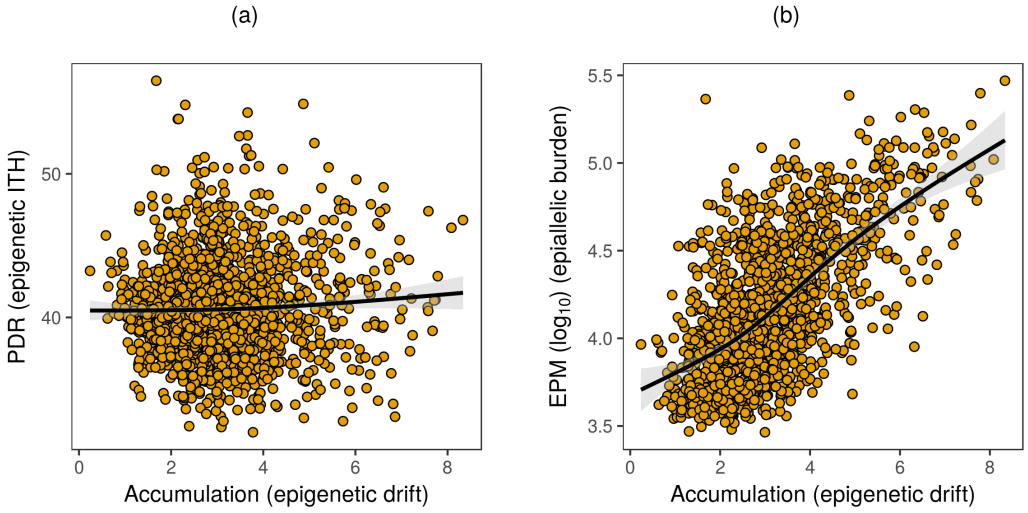


Figure 5.10: The degree of epiallelic composition shifting is correlated with epigenetic drift. (a) Scatter plot of global PDR score and Accumulation metric for the 1482 tumours. **(b)** Scatter plot of global EPM score ($\log_{10}$) and Accumulation metric for the 1482 tumours. Each point represents a tumour. In both panels, the black line represents the loess curve (95% confidence interval in grey) for the relationship. Only 4-CpG loci at 20× within promoters were considered.

However, both EPM scores and the Accumulation index represent departures in the tumour tissue compared to normal tissues. Moreover, as mentioned earlier, the EPM analysis can detect stochastic and heterogeneous DNA methylation pattern changes that also largely contribute to epigenetic drift. Consequently, a significantly positive correlation was detected between these two scores (rho = 0.62; *p-value* = 2.3×10^{-171}; partial correlation adjusted for ER status. Log_{10} EPM scores were used for statistical testing; Figure 5.10b). Similar results were obtained for EPM scores inferred only from promoter loci (rho = 0.53, *p-value* = 1.07×10^{-112}, partial correlation adjusted for ER status. Log_{10} EPM scores were used for statistical testing), as well as PDR scores inferred from loci within the background or neutral epigenome defined in Chapter 3 (rho = 0.64, *p-value* = 3.4×10^{-187}, partial correlation adjusted for ER status. Log_{10} EPM scores were used for statistical testing). This suggests that a large contribution of epiallelic composition shifts in tumours compared to normal tissues are the consequence of epigenetic drift related methylation errors.

5.4.3 Relationship between PDR scores and EPM scores

The relationship between the two epiclonal measures of the tumour's epigenome - PDR score and EPM score - was also of considerable interest, particularly since they represent orthogonal scores that have never been linked previously. The association between the scores was investigated in promoters across the 1482 breast tumours (Figure 5.11a). Given the largely non-linear relationship between the two scores, the tumours were classified into 4 quadrants based on mean promoter PDR and promoter EPM (log_{10}) scores for the whole cohort to allow for easier interpretation. However, an exception to the non-linear relationship was observed within tumours in quadrant 1 (green; Figure 5.11a), that showed a strong correlation between PDR and EPM scores (rho = 0.47; *p-value* = 1.3×10^{-25}; partial correlation adjusted for ER status. Log_{10} EPM scores were used for statistical testing; n = 437). In this subgroup, tumours that harboured frequent epiallelic composition shifts were also associated with a higher degree of locally disordered methylation in promoters. This implies that although these tumours have a large number of epiclonal shifts compared to the normal tissue, these methylation changes are more likely to be stochastic and heterogeneous, rather than ordered methylation changes reflective of functional alterations under selection.

Similar correlations between the PDR score and the EPM score were not noted in the other quadrants. Tumours with a propensity for selection-related methylation dynamics (lower PDR) were not associated with epiallelic burden and were found to be similarly distributed across the tumours with a lower epiallelic burden (low PDR

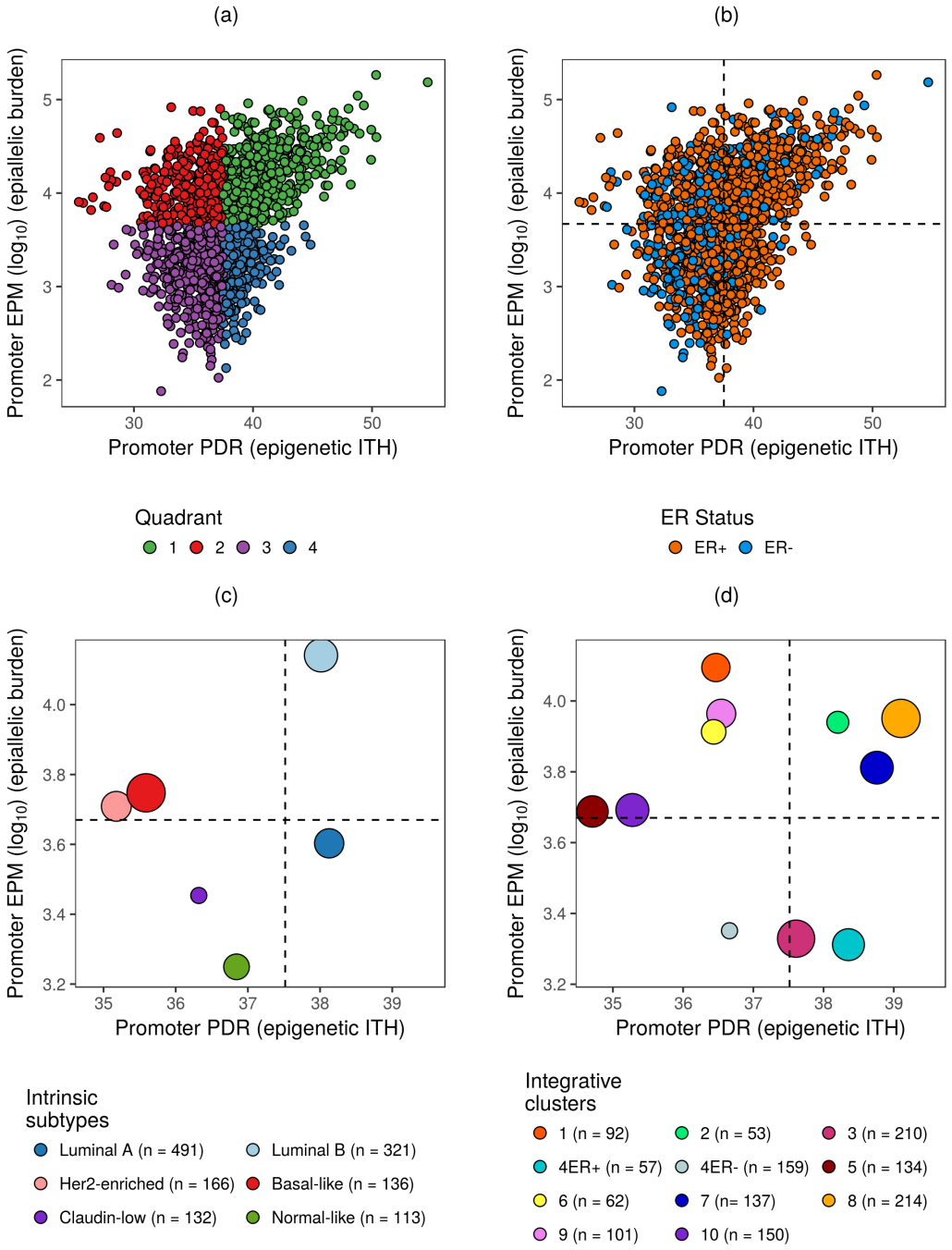


Figure 5.11: (Caption on next page.)

Figure 5.11: (Previous page.) **Breast cancer subtypes are associated with the 4-group methylation classifier. (a)** Scatter plot of promoter PDR score and promoter EPM ($\log_{10}$) scores for the 1482 tumours. A 4-group methylation classier was constructed using mean values (based on the whole cohort) for the two scores. Colours represent subgroup classification. The black line represents the loess curve (95% confidence interval in grey) for the relationship. **(b)** Scatter plot of promoter PDR score and promoter EPM ($\log_{10}$) scores for the 1482 tumours. Colours represent ER status. **(c)** Bubble plot of median promoter PDR scores and median promoter EPM ($\log_{10}$) scores for each Intrinsic subtype. **(d)** Bubble plot of median promoter PDR scores and median promoter EPM ($\log_{10}$) scores for each Integrative cluster. In (a) (b) (c) and (d), only 4-CpG loci at 20× within promoters were considered. Grey dotted lines depict the means for both scores (vertical line, mean promoter PDR; horizontal lines, mean promoter EPM ($\log_{10}$)) in the cohort as a whole. In (c) and (d) the areas of the circles are proportional to the covariance between the PDR and EPM ($\log_{10}$) scores in each subtype.

and low EPM, quadrant 3, purple; Figure 5.11a) or a higher burden (low PDR and high EPM, quadrant 2, red; Figure 5.11a). This strongly indicates that a tumour's propensity for ordered methylation changes (likely reflecting a selection process) in the promoter is independent of the number of epiallelic composition shifts that it harbours.

This confirms that the PDR score and the EPM score of tumours represent distinct properties of the epigenome, and the combination of the two scores into 4 categories (based on mean values of the whole cohort) can be used to discriminate between tumours into those where these shifts are a consequence of stochastic drift or selection, as well as between tumours with a low or high epiallelic burden. A description of this 4-group classifier is detailed in Table 5.1. ER status was significantly associated with this 4-category epigenetic based classification (*p-value* = 3.3×10^{-19}; chi-square test; Figure 5.11b), with ER+ tumours being strongly predisposed to disordered methylation patterns with enrichments in quadrant 1 (OR = 2.0, *FDR p-value* = 1.8×10^{-6}; Fisher's exact test) and more so in quadrant 4 (OR = 3.5; *FDR p-value* = 4.8×10^{-10}; Fisher's exact test). Conversely, ER- tumours were enriched in quadrant 3 (OR = 2.6, *FDR p-value* = 3.1×10^{-13}; Fisher's exact test) had a tendency for selection-related methylation dynamics (lower PDR) but with a lower total epiallelic burden in promoters. This classification was also significantly associated with the Intrinsic subtype definition (*p-value* = 2.6×10^{-58}; chi-square test; Figure 5.11c) and the Integrative cluster definition (*p-value* = 7.2×10^{-65}; chi-square test; Figure 5.11d). The strong relationship of this purely epigenetic based classifier with a *genetic and transcriptomic* defined classification of breast cancer such as Integrative clusters strongly confirms its biological significance, and also the linked nature of genetic and epigenetic dysregulation in tumorigenesis. Moreover, the IntClusts 1,5,6,9 and

Intrinsic clusters - Luminal B, HER2 and Basal-like, are all associated with poor prognosis and were also all enriched in quadrant 2 [Curtis et al., 2012; Parker et al., 2009]. Consequently, the utility of this methylation-based classifier as a prognostic indicator in breast cancer was formally investigated in Section 5.5.

Description			Epigenetic scores		Clinical variables		Enriched subtypes			Genetic scores	
Quadrant	Colour	N	PDR	EPM	Age	Grade	ER Status	Integrative	Intrinsic	CIN	MATH
1	Green	437	High ITH	High burden	63.2	2.4	ER+	7 8	Luminal B	767	0.33
2	Red	314	Low ITH	High burden	62.7	2.6		1 5 6 9	Luminal B, Her2, Basal	1061	0.40
3	Purple	471	Low ITH	Low burden	58.9	2.6	ER-	4ER- 5 10	Her2, Basal, Claudin, Normal	931	0.37
4	Blue	260	High ITH	Low burden	57.9	2.0	ER+	3 4ER+	Luminal A, Claudin, Normal	450	0.28

Table 5.1: Description of the 4-group methylation classifier. Quadrant and colour denotes the quadrant number and colour of the respective subgroup as illustrated in Figure 5.11. N denotes the number of tumours. PDR denotes the level of epigenetic intratumour heterogeneity. EPM denotes the level of epigenetic burden. Age and Grade represents the average age at diagnosis and average tumour grade. ER Status, Intrinsic (subtype) and Integrative (cluster) represent the established breast cancer subtype(s) that is (are) enriched (Odds Ratio > 1.5, *FDR p-value* < 0.05; Fisher's Exact test). CIN denotes the average chromosomal instability. MATH denotes the average mutant-allele tumour heterogeneity. Analysis as described in text.

5.5 PDR and EPM scores are prognostic in breast cancer

5.5.1 Construction of survival models

Cox proportional hazard models [Cox, 1972] were used to explore the prognostic potential of the two epigenetic measures generated in this chapter -- promoter PDR score and the promoter EPM score. The prognostic value of the CIN score was also investigated for comparison. Breast Cancer-Specific Survival (BCSS) was explored for these measures, and accordingly, the endpoint of interest was defined as death due to breast cancer. Patients with deaths due to other or unknown causes were censored at the times of those deaths, and all other patients were censored at the time of last contact. The three continuous predictors (PDR; EPM; CIN scores) were scaled using their respective standard deviations (SD), such that the hazard ratio (HR) is interpreted as the ratio of hazard rates corresponding to an increase of one 1 SD of the corresponding predictor. Higher hazard ratios indicate that tumours with higher values of the predictor are associated with lower survival times i.e. higher risk of death associated with breast cancer [Zwiener et al., 2011].

All survival models were adjusted for confounding clinicopathological variables including ER status, grade, size, lymph node status, and age at diagnosis. Age, Lymph node status and size of the tumour were treated as a continuous variable; while ER status (ER+ vs. ER-) and grade (1/2 vs. 3) were treated as binary variables. The effect of treatment is also important to consider as it may act as a confounding factor in survival analyses. However, since treatment decisions for early breast cancer were highly dependent on the standard clinical parameters mentioned above, treatment was not used as an additional variable in these analyses. As described for the original METABRIC study [Curtis et al., 2012], nearly all ER+ patients that were lymph node-negative did not receive adjuvant chemotherapy status; conversely, all ER− patients that were lymph node-positive patients did. Moreover, the patients with ER+ disease who did receive chemotherapy were also more likely to have high-grade tumours. None of the HER2+ patients in the original METABRIC cohort received adjuvant trastuzumab since they were diagnosed prior to its use in clinic.

5.5.2 PDR and EPM scores collectively, but not individually are highly significant predictors of BCSS

Firstly, survival models were constructed to assess the individual contributions of the two promoter methylation-based scores: PDR and EPM scores. Given the dominance of copy number events in breast cancer [Ciriello et al., 2013], the prognostic value of the CIN score was also considered in a third model. Next, to evaluate the combined contributions of these measures, three pairwise multivariable models, i) CIN +PDR; ii) CIN + EPM; iii) EPM+PDR; and a final multivariable model with all three predictors, CIN +PDR + EPM. These seven survival models were also adjusted for the clinical confounding variables (enumerated above). Further, a likelihood ratio (LR) test was conducted for each model, to assess whether the described model provides a better fit than the *null-clinical* model (comprising of solely the 5 clinicopathological parameters). The results of the seven models are presented in Table 5.2.

Lower promoter PDR scores were predictive of lower BCSS after adjustment of clinical parameters (HR = 0.9016, *p-value* = 0.0497; adjusted Cox proportional hazards model; Table 5.2) indicating that tumours associated with lower epigenetic heterogeneity (selection-related methylation patterns) at promoters were associated with worse outcomes. Conversely, higher promoter EPM scores were not predictive of lower BCSS (HR = 1.0534, *p-value* = 0.2934, EPM scores were $\log_{10}$ transformed; adjusted Cox proportional hazards model; Table 5.2). However, combining these two scores gave rise to a model where both promoter PDR scores and promoter EPM scores demonstrated evidence of being prognostic (PDR *p-value* = 0.0111, EPM *p-value* = 0.0587, Model 6 in Table 5.2). Moreover, the likelihood ratio test indicated that amongst all seven models tested, the combined EPM + PDR model provides the best *goodness of fit* (lowest LR test *p-value*), and was significantly informative when compared to the *null-clinical* model (LR test *p-value* = 0.0236). The partial hazard ratios (adjusted for each other and clinicopathological variables, Model 6 in Table 5.2) for promoter EPM and PDR scores are illustrated in Figure 5.12, again confirming that tumours with higher epiallelic burden and tumours with evidence of ordered methylation patterns (indicative of a selection process) in promoters were associated with adverse prognosis (higher hazard ratio; lower BCSS).

Higher CIN scores were mildly predictive individually (adjusted Cox proportional hazards model: HR = 1.1029, *p-value* = 0.0764); however, inclusion of CIN with either of the methylation scores did not yield models with a better fit than the *null-clinical*

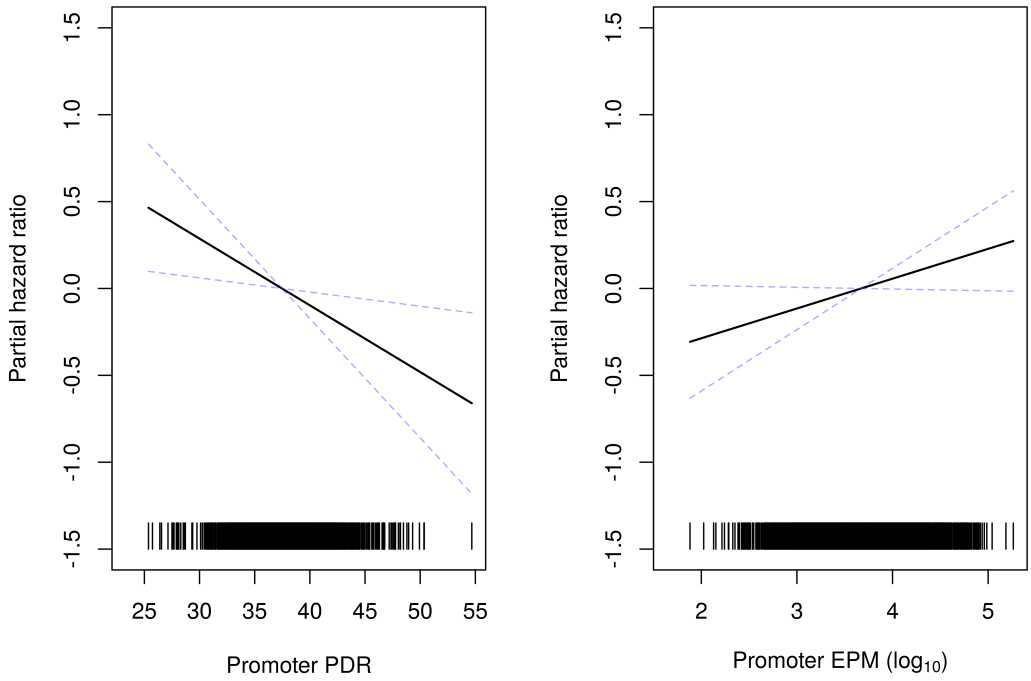


Figure 5.12: PDR and EPM scores are both prognostic in breast cancer when combined. (left panel) Hazard ratio for promoter PDR score (partial i.e. adjusted) from Model 6 in Table 5.2. **(right panel)** Hazard ratio for promoter EPM ($\log_{10}$) score (partial i.e. adjusted) from Model 6 in Table 5.2. Survival models constructed as described in text and Table 5.2. of the circles are proportional to number of samples in each subtype.

Model	Variable	HR (scaled)	Wald's test	LR Test
Model1: EPM	EPM	1.05	0.2934	**0.2931**
Model2: PDR	PDR	0.90	0.0497	**0.0482**
Model3: CIN	CIN	1.10	0.0764	**0.0798**
Model4: CIN+PDR	CIN	1.10	0.0920	
	PDR	0.95	0.3698	**0.1435**
Model5: CIN+EPM	CIN	1.10	0.0930	
	EPM	1.02	0.6925	**0.1993**
Model6: EPM+PDR	EPM	1.11	0.0587	
	PDR	0.87	0.0111	**0.0236**
Model7: EPM+PDR+CIN	EPM	1.05	0.4005	
	PDR	0.93	0.2423	
	CIN	1.09	0.1434	**0.2042**

Table 5.2: PDR and EPM scores are both prognostic in breast cancer when combined. The first column represents the epigenetic or genetic variables included in the survival model fitted using Cox proportional hazards, as described in the text. All survival models were adjusted for clinicopathological variables including ER Status, Age at diagnosis, No of lymph nodes, Size and Grade of tumour. HR represents the scaled hazard ratio, and Wald's test represents the Wald's test *p-value*. Both statistical estimates were adjusted for other variables in the model and calculated as described in the text. LR test represents the *p-value* from the likelihood ratio (chi-square) test of the model versus the *null-clinical* model.

model, implying that the intratumour promoter methylation measures – PDR and EPM – have a greater prognostic potential than chromosomal instability.

5.5.3 PDR + EPM methylation classifier is prognostic

Given that PDR and EPM collectively are highly prognostic of BCSS, the utility of the 4-group classifier based on these two scores (defined in Section 5.4) as a prognostic indicator in breast cancer was investigated. Kaplan-Meier survival curves (endpoints: BCSS) were constructed to test the prognostic ability of the promoter methylation-based classifier in all tumours and this model was highly prognostic (Log-likelihood test *p-value* = 1.2×10^{-6}, Figure 5.13a). As expected from the findings in the previous section, breast tumours with *both* high epiallelic burden in promoters and low epigenetic ITH (quadrant 2, red; HR = 2.27; *p-value* = 1.4×10^{-6}, reference = quadrant 4; Cox proportional hazards model) had the shortest survival times. Conversely tumours with a low epiallelic burden and high epigenetic ITH in promoters (quadrant

4, blue (reference): HR = 1) were linked with the best prognosis. The survival curves were also significantly prognostic in ER+ tumours with quadrant 2 tumours exhibiting the shortest BCSS times (Figure 5.13b), but not in ER- tumours (Figure 5.13c). However, crucially the 4-group promoter methylation-based classifier was significantly prognostic within three Integrative clusters (Log-likelihood test *p-values* for all < 0.05; Figure 5.13d-f). For instance, IntClust 8 (largely ER+; low grade) tumours that harbour a high epiallelic burden in the promoters had the worst prognosis (quadrants 1 (green) and 2 (red)), while IntClust 9 (largely ER+; high grade) tumours with low epigenetic burden in the promoters were likely to have the worst outcomes (quadrants 3 (purple) and 4 (blue)). This finding in the IntClust 9 subtype can explained due to the identified driver role of 8q cis-acting copy number alterations and 20q amplification in these tumours, indicating that role of epigenetics may not be necessary. Remarkably, IntClust 10 (predominantly ER-; high grade) tumours with high epigenetic burden as well as high epigenetic diversity in the promoters were likely to have the worst outcomes (quadrants 1 (green)). This reflects the extraordinary heterogeneity observed in the role of epiallelic dynamics in breast tumours, and also indicates the ability to refine the molecular taxonomy of breast cancer by investigating the epigenome.

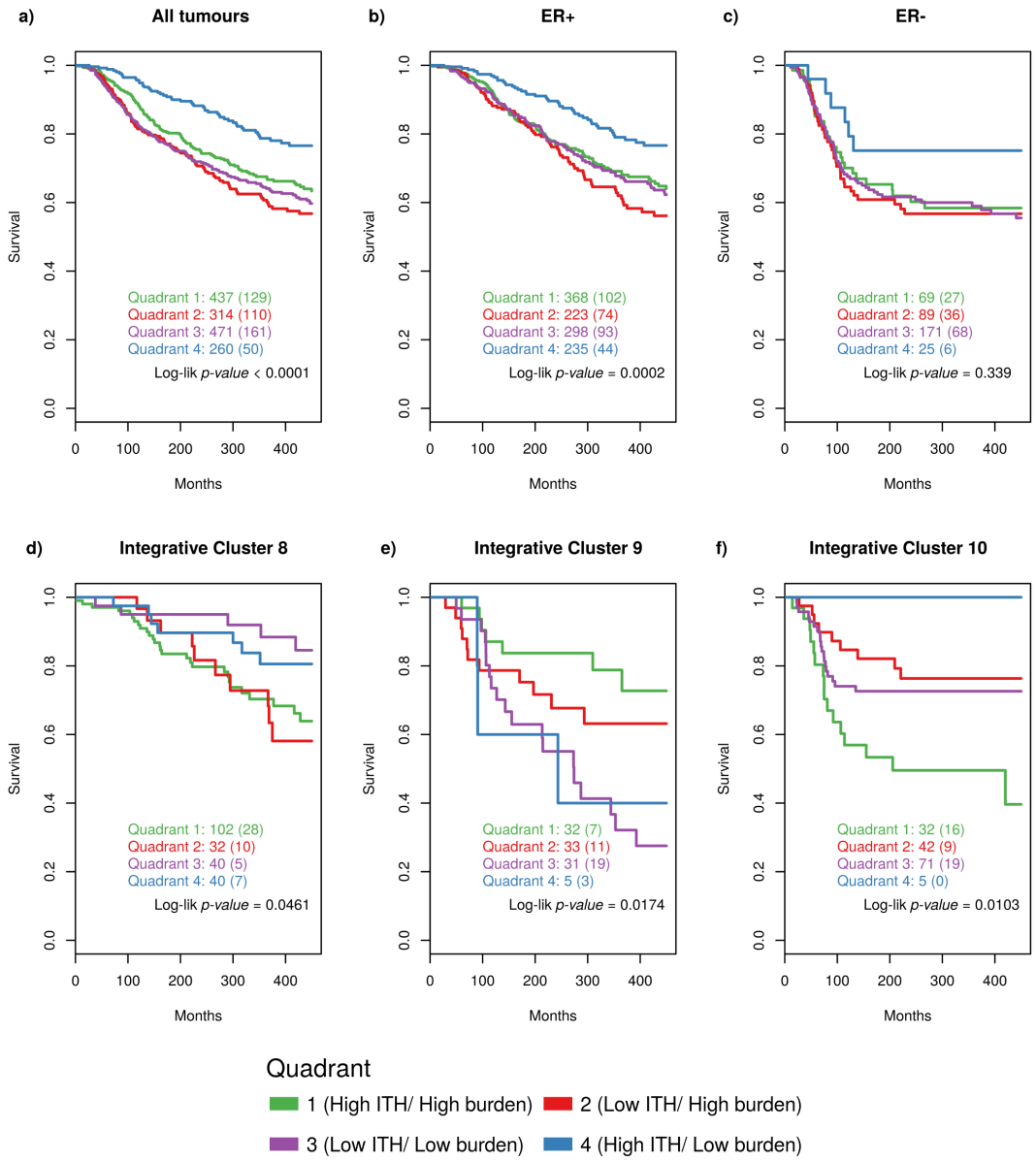


Figure 5.13: The 4-group methylation classifier is prognostic in breast cancer. Kaplan-Meier survival curves (endpoints: BCSS) were constructed to test the prognostic ability of the 4-group methylation-based classifier in **(a)** all tumours, **(b)** ER+ tumours, **(c)** ER- tumours, **(d)** IntClust 8 tumours, **(e)** IntClust 9 tumours, and **(f)** IntClust 10 tumours. The numbers of samples under consideration in each methylation subgroup are indicated, and the numbers in brackets represent the deaths occurring in the cohort. Log-lik *p-value* represents the significance of the survival curves based on a log-likelihood test.

5.6 Discussion

Extensive profiling of cancer methylomes have delineated specific genomic regions that on average are differentially methylated between cancer and normal tissues. However, there is insufficient evidence of the dynamics of the process driving methylation change in cancer tissues. Are these differences a consequence of ordered and regulated changes in methylation patterns, or rather a result of stochastic and heterogeneous process occurring independently and non-specifically over the epigenome? This question was briefly addressed using tumour and region-specific estimates of the background methylation differences in Chapter 3. However, RRBS, which is an NGS based methylation profiling platform also allows the capture of the epiallelic composition from contiguous CpG sites on sequencing reads. Leveraging this information can provide deeper insights into the role of the epigenome in cancer.

Utilising the proportion of discordant reads (PDR) method [Landau et al., 2014] enabled the enumeration of epigenetic ITH for the 1482 breast tumours at different regions of the genome, with lower PDR values within a CpG locus representing a controlled increase in the frequency of a specific epiallele while higher PDR scores represent multiple stochastic changes in the frequencies of many epialleles. Therefore, PDR scores leverage the snapshot of the epigenetic state in a tumour to unravel the journey that has been traversed ultimately culminating in this state. Late replicating regions were associated with disordered methylation patterns compared with early replicating regions, corroborating the findings in Chapter 3 in which these regions were shown to accumulate significantly higher drift (stochastic) related methylation errors. Furthermore, directed-DMRs associated with concomitant expression changes also harboured lower epigenetic ITH than other loci, indicating that these functionally-marked regions potentially underwent an ordered change in methylation patterns likely reflecting a positive selection mechanism.

A similar loss in ITH (epipolymorphism scores) was observed in hyper methylated loci in a study of colorectal tumour-normal pairs, albeit only a select panel of putative DMRs were assayed [Landan et al., 2012]. However, they demonstrated that the subpopulation of cells in the tumour with increased methylation levels actually exhibited high epipolymorphism indicating a strong stochastic component, and the loss in polymorphism at such loci is actually attributed to a *co-existing but distinct* subpopulation of cells that remains completely resistant to methylation. Consequently, the authors [Landan et al., 2012] made an interesting argument that although deterministic epigenetic patterns are observed in cancer cells that gain

resistance to methylation, once this protection is removed, methylation accumulates in a locally disordered manner leading to an increase in methylation levels (and consequently identified as a hyper DMR). However, this postulation remains to be confirmed in other published genome-wide cancer methylome studies, and will be examined in our breast cancer cohort in the future.

Next, the shifting of epiallel compositions from the normal tissue to the breast tumour was assessed using *methclone* [Li et al., 2014]. The *methclone* algorithm can not only detect differentially methylated loci, but in addition can also detect differential variability in methylation that might be a consequence of stochastic processes. Promoters with evidence of epiallelic composition shifting (eloci) in the tumour (compared to the normal tissue) were associated with a higher variability in transcriptional change (compared to the normal tissues), as well as a higher probability of being differential expressed in the tumour, indicating the utility of this approach. Interestingly, tumours with a high number of aggregate eloci (calculated as eloci per million or EPM) also demonstrated evidence of high epigenetic drift (measured by the *Accumulation* index, see Chapter 3). This may not be surprising, given that both EPM and epigenetic drift give an account of the accumulation of stochastic heterogeneous methylation patterns in tumours compared to the normal tissue (though EPM does not exclusively detect stochastic changes). However, what is notable is the fact that methods underlying the calculation of these two scores, *Accumulation* index (developed in Chapter 3) and *methclone* [Li et al., 2014] are vastly distinct. The former utilises information obtained from average methylation differences in various genomic contexts while the latter leverages single read information to detect changes in epiclonal composition.

Epiallelic composition shifting represents an orthogonal approach to quantifying epigenetic ITH within individual samples, since it examines the transformation of epiallelic states between two samples. While tumour-specific aggregate EPM scores represent the overall epiallelic burden of the tumour including ordered methylation changes as well as stochastic and heterogeneous DNA methylation pattern changes, the tumour-specific PDR scores contextualises this burden by characterising the level of epigenetic ITH in these regions.

However, a caveat in the estimation of the two epiallelic (PDR and EPM) scores is that measurements from tumour samples may include a mixture of tumour and adjacent normal tissue depending on tumour purity [Jaffe and Irizarry, 2014; Zheng et al., 2014, also discussed in Chapter 3]. Consequently, this would lead to biased PDR and EPM scores. Although, a few methods to account for normal tissue contamination

in cancer methylome studies are available for microarrays and other platforms using aggregate tumour methylation profiles [Zheng et al., 2017b, 2014, DMARC developed in Chapter 3], true decontamination at the epiallele level is challenging. A recent report details a novel Bayesian approach to infer tumour sample purity from bisulphite sequencing data and incorporate these estimates to deconvolute the tumour epiallele profiles [Barrett et al., 2017]. Moreover, this method also controls for experimental noise due to variation in bisulphite conversion. Future work will utilise this protocol in the RRBS dataset described in this thesis, to allow a more accurate quantification of the EPM and PDR scores in the 1482 breast tumours.

A limitation of the RRBS protocol (see Chapter 1, 2) is the inability to discriminate between PCR-induced duplication artefacts or distinct molecular copies of fragments since the start and end sites of reads are largely driven by restriction enzyme digestion. This can also distort estimates of intratumour methylation heterogeneity estimates. This predicament in RRBS can be circumvented through the use of unique molecular identifiers (UMI) as demonstrated in the quantitative RRBS (Q-RRBS) method established by Wang et al. [2015]. Future studies using singe-cell genome-wide bisulphite sequencing techniques [Farlik et al., 2015; Guo et al., 2013; Smallwood et al., 2014] in breast cancer will be able to generate exquisite measures of ITH and provide valuable insights into their role in tumour evolution.

Several studies have reported substantial agreement between ITH inferred from DNA methylation compared with CNAs and somatic mutations [Aryee et al., 2013; Brocks et al., 2014; Mazor et al., 2015]. However, these studies examined spatial heterogeneity via overall DNA methylation measurements at CpG sites based on microarrays, and hence do not completely address the issue of epiclonal heterogeneity. A recent report in CLL documented high epigenetic heterogeneity that was associated with a higher number of subclonal mutations [Landau et al., 2014]. However, this genetic score is not truly reflective of genetic ITH, but rather the degree of mutational burden. The link between genetic and epigenetic ITH remains to be fully understood. In this chapter, measurements of genetic and epigenetic intratumour heterogeneity (MATH score and PDR score respectively) across more than 1000 tumours were compared and remarkably, a significant negative correlation between the two was observed. Higher genetic ITH has been associated with adverse clinical outcomes in breast cancer [Pereira et al., 2016] and other cancers [Landau et al., 2013; Merlo et al., 2010] while lower epigenetic ITH was associated with shorter survival times in the breast cancer dataset described in this chapter. This may reflect that epigenetic and genetic routes to malignancy in breast tissues are related but longitudinal experiments

examining *de novo* genetic and epigenetic alterations in cancer cells would be required to shed further light into this interesting relationship.

This chapter presents the first investigation of the interplay between the two intratumour DNA methylation scores – PDR (epigenetic ITH) and EPM (epigenetic burden) -- in cancer, as well as the first examination of the prognostic value of these indices in breast cancer (or in any epithelial malignancy). Interestingly, EPM scores were not prognostic individually (in addition to clinicopathological variables), but when combined with PDR scores, both scores were prognostic (in addition to clinicopathological variables) with tumours demonstrating evidence of higher epigenetic burden and lower epigenetic heterogeneity associated with worse outcomes. The resulting model using the two orthogonal epigenetic scores was the most significantly prognostic model when compared to models with the individual scores as well as models with CIN (chromosomal instability index). Given that breast cancer is dominated by copy number alterations, the superior prognostic potential of the epigenetic scores indicates the strong role of the epigenome in breast cancer.

The prognostic potential of the PDR and EPM scores supported the construction of a simple methylation-based classifier based on the mean values of these scores in the whole cohort. This classification was not only highly prognostic in all breast tumours, but crucially also explained different phenotypes within established breast cancer subtypes. Tumours with high epigenetic burden within IntClust 8 and tumours with low epigenetic ITH in IntClust 9 were associated with worse outcomes compared to other tumours within the respective subtypes. These tumours were largely ER+, and consequently when considering all ER+ tumours, the combination of higher epiallelic burden and lower ITH in promoter regions (quadrant 2, red) were demonstrated to have the shorter breast cancer-specific survival times. Although, the higher epiallelic burden (EPM score) in promoters have also been previously linked with adverse outcomes in AML [Li et al., 2016b], the findings that *lower* DNA methylation ITH in promoters is associated with *worse* prognosis is novel. Lower DNA methylation ITH is an indicator of ordered methylation patterns that may reflect a selection process of methylation alterations which enhance tumour fitness through the aberrant expression of tumour suppressors and oncogenes (directed ex-DMRs have lower PDRs, Section 5.2). Tumours with lower ITH also have a higher grade, and so it could be argued that they represent recent epiclonal expansions in which the fittest subclone emerged resulting in lower ITH. In fact, the study in DLBCL [Pan et al., 2015] also reported decreased methylation ITH at diagnosis compared to the normal cell population from which they arose, and a further decrease in methylation ITH at relapse corroborating

the postulation that as a cancer progresses, the epiclonal state rapidly becomes homogeneous.

In sharp contrast, to the ER+ subtypes (IntClust 8 and 9), tumours within IntClust 10 that had high epigenetic burden as well as high epigenetic diversity were associated with worse prognosis. These tumours are predominantly ER-/ Basal-like, thus reflecting the extraordinary distinction in the role of epiallelic dynamics in different breast cancer subtypes. The observation that higher epigenetic heterogeneity is associated with worse outcomes corroborates findings from the seminal study that established PDR scores and applied them in CLL [Landau et al., 2014]. They also reported that tumours with *higher* DNA methylation ITH at diagnosis correlated with shorter progression-free survival in CLL suggesting that higher epigenetic diversity could enhance phenotypic plasticity, thus enabling superior tumour evolution.

This 4-group methylation classifier was constructed using two tumour-specific epigenetic indices that were calculated by aggregating scores over all 4-CpG promoter loci (at 20×) within the RRBS universe. However, the epiallele dynamics measured by the PDR and EPM scores can be retained at the single locus level allowing the discovery and prioritisation of epiclonal selection and epiallele composition shifting in key cancer genes. It is anticipated that this effort will substantially improve the understanding of the role of stochastic drift, phenotypic plasticity and natural selection in evolution of tumours, as well as result in the development of an enhanced methylation classifier with superior prognostic ability.

Chapter 6

Summary and Perspective

The investigation of the DNA methylation landscape in the METABRIC cohort comprising of 1482 breast tumours and 237 matched adjacent normal tissues using RRBS is described in this thesis. This constitutes the largest single cancer methylome study yet, and being a next-generation sequencing (NGS) technique, represents a significant development in the field of breast cancer.

In **Chapter 2**, a robust RRBS bioinformatics pipeline that is not only suitable for high-throughput, but also maximises the information content yield was developed and implemented. This pipeline was also validated by comparing with microarray profiling performed on some of the samples, as well as with an external dataset [Cancer Genome Atlas Network, 2012]. Quantitative methods for RRBS are currently undeveloped due to the specific complexities this protocol, and to address this, a novel algorithm called SCCRUB (Spatially Coordinated CpG-sites within the RRBS Universe in Breast cancer) was constructed and implemented to define a functionally relevant RRBS universe of regions comprising of spatially coordinated CpG sites in breast cancer. The dataset generated in this thesis represents the methylation profiles of a large number of breast tumours, and consequently will be of great value to the breast cancer community. It has already been used as a validation dataset in a recent report investigating germline and somatic alterations underlying deficient homologous recombination repair in breast cancer [Polak et al., 2017].

In **Chapter 3**, the epigenetic drift defined as the extent of background or neutral DNA methylation changes in breast tumours (compared to the normal tissues) was quantified. This represents the first characterisation of genome-wide epigenetic drift in a large cohort of primary breast tumours, and was found to be highly context specific with methylation gains largely observed in CpG rich regions and losses in low CpG regions. Moreover, late time of replication regions demonstrated a remarkable predisposition for accumulating both methylation changes (gains and losses) in breast cancer. The extent of epigenetic drift was also found to be highly heterogeneous between the breast tumours and was sharply correlated with the tumour's mitotic index. This confirmed that epigenetic drift is largely a consequence of the accumulation of passive replication related errors related to the number of cell divisions as postulated by Yatabe et al. [2001]. The presence of these methylation alterations, that are largely stochastic and non-specific in nature, can make the identification of epigenetic alterations, that are truly associated with the initiation and progression of tumours, quite challenging; an obstacle that is rarely considered in previous cancer methylome

studies. Accordingly, the background methylation differences in tumours were utilised to feed the development of a novel algorithm called DMARC (Directed Methylation Altered Regions in Cancer) to detect directed and background DNA methylation alterations in tumours. In addition, to illuminating the mechanism underlying an observed methylation difference in a tumour, directed methylation alterations identified by the algorithm were significantly enriched for putative functional (gene expression) changes in breast cancer compared to background alterations.

Another key benefit of the DMARC algorithm is that it provides a novel way to account for tumour purity. Moreover, DMARC is suitable for all high-throughput genome-wide cancer methylome studies, and in principle, it can be applied across all popular methylation profiling techniques that provide single CpG resolution such as microarrays and bisulphite sequencing techniques. Future work will assess their utility in pan-cancer microarray-based methylomes publically available from TCGA [Bass et al., 2014; Cancer Genome Atlas Network, 2012; Cancer Genome Atlas Research Network, 2016; Collisson et al., 2014; Hammerman et al., 2012; Muzny et al., 2012; The Cancer Genome Atlas Research Network, 2013a; Weinstein et al., 2014].

Chapter 3 concluded with the implementation of the DMARC algorithm in the large cohort of breast tumours and normal tissues. Collectively, the results strongly indicated that the activity of major signalling pathways in breast cancer are at least partly regulated by the epigenome, and more so, by epigenetic deregulation of not only promoters but also other gene-associated elements as well as distal-regulatory elements. Functional characterisation of differentially methylated regions in ER+ and ER- tumours, also led to identification of subtype-specific candidate targets that were not only incriminated with gene silencing, but also implicated with the upregulation of genes. For instance, *SPDEF* and *TFF3* both harboured directed-DMRs specifically linked to differential regulation in ER+ tumours and were also associated with worse prognosis. On the other hand, genes including *IDH2* and *MAST4* were identified as subtype specific epigenetic regulators (both harboured directed-DMRs) in ER- breast cancers and were also associated with worse prognosis. Furthermore, gene set enrichment analysis revealed distinct pathways that were epigenetically disrupted in ER+ and ER- tumours. For instance, oestrogen-signalling was epigenetically disrupted in ER+ tumours and the p53 pathway in ER- tumours, thus revealing the heterogeneity in epigenetic programming in distinct breast cancer subtypes.

Comparing the contributions of CNAs and DNA methylation events breast cancer in **Chapter 4**, revealed a poignant juxtaposition in the regulatory roles of CNA and DNA methylation. CNA played a stronger role in driving the expression of key

protagonists of cell cycle progression leading to the divergence of tumours from normal tissues. Conversely, silenced genes in breast tumours were significantly enriched for DNA methylation alterations, and not in *cis* genomic loss. Moreover, focusing on the top 2000 variably expressed genes in breast cancer as well as the set of genes differentially expressed genes between the two ER subtypes established that DNA methylation was the preferred mechanism for refining subtype-specific differences.

The crucial role of DNA methylation as a mechanism to target the silencing of specific genes within copy number amplifications is also explored. This led to the identification of a putative tumour suppressor gene, *THSZ2*, which lies within the amplified 20q13 cytoband but is silenced potentially as a consequence of promoter hypermethylation. Several genes were identified in which DNA methylation events modulated the role of CNA in a similar manner. Striking examples include gene body methylation acting as a diminishing agent in *GATA3* resulting in the subtype specific downregulation of the gene in ER- tumours; and the CNA enhancing effect of DNA methylation in *FOXA1* leading to its upregulation in ER+ tumours. The large number of samples also allowed for the detection of mutually exclusive or co-occurring patterns between epigenetic (DNA methylation) and genetic (CNA) alterations, that led to the identification of genes with putative tumour suppressive roles (such as *CITED1* in ER+ tumours and *FGF2* in ER- tumours) or oncogenic roles (such as *INTS8* in ER+ tumours and *CDK12* in ER- tumours).

Although, it is possible that the DNA methylation alterations detected in this thesis may actually instigate the silencing or overexpression of a gene that is also observed, recent studies have also shown that the concomitant deregulation in expression and methylation patterns can be a consequence of an upstream transcription factor inactivation event [Domcke et al., 2015; Feldmann et al., 2013; Yin et al., 2017], or due to an alteration in the chromatin structure [Ohm et al., 2007; Schlesinger et al., 2007; Widschwendter et al., 2007]. The associative statistical models presented in the thesis cannot identify which is the causal modification and can only demonstrate *inferred function* of DNA methylation events in silencing or over-expressing key genes. Functional investigation such as epigenome editing is required for providing definite evidence of the *causal function* of individual alterations in genes, and for labelling the targeted genes as tumour suppressors or oncogenes [Stricker et al., 2016]. However, irrespective of whether DNA methylation plays an *initiating* or a *reinforcing* role in gene regulation, the detection of this modification provides an extremely informative readout of the underlying epigenetic state of the tissue. Differential DNA methylation marks do not only reflect altered transcription

factor activity [Fleischer et al., 2017; Schübeler, 2015], but as demonstrated in this thesis, they can also serve as powerful biomarkers for breast cancer subtype-specific diagnosis and prognosis.

The results presented in **Chapter 4** underscore the multiplicity of mechanisms by which key cancer genes may be deregulated. However, it is also evident that a large component of the variation in the breast cancer transcriptome is still unaccounted for (only ∼25% of the top 2000 variably expressed genes were significantly regulated by DNA methylation or CNA). This implies that a majority of gene expression variation is still to be explained by other events such as microRNA [Dvinge et al., 2013], somatic mutations [Nik-Zainal et al., 2016; Pereira et al., 2016] that have been shown to play an important role in breast cancer. Other epigenetic events such as histone marks [Kondo et al., 2008; Seligson et al., 2005] and chromosomal interactions [Dryden et al., 2014; Rousseau et al., 2014; Zeitz et al., 2013] are also altered in cancer with devastating transcriptional consequences. A complete comprehension of the mechanisms underlying breast pathogenesis and tumour heterogeneity can only be achieved when all events have been characterised. Information generated from different molecular interrogations can be used to feed an integrative network analysis to identify specific molecular pathways that are disrupted in breast cancer that are indistinguishable if studied at a single gene level. For instance, integrating the DNA methylation, copy number, mRNA and microRNA expression using PARADIGM (Pathway recognition algorithm using data integration on genomic models) algorithm [Vaske et al., 2010] emphasised the importance of FOXM1 and ERBB4 signalling in breast cancer [Kristensen et al., 2012]. Furthermore, combining the genomic and transcriptomic landscapes using the METABRIC dataset has already revealed 11 novel molecular subgroups with distinct clinical features and prognosis [Curtis et al., 2012]. It will be interesting to see whether reconciling the tumour methylomes generated in this thesis with mRNA expression, microRNA expression, somatic mutation and copy number landscapes would identify therapeutically tractable signatures and improve classification towards a driver-based taxonomy.

In **Chapter 5**, the 1482 RRBS breast cancer and 237 normal tissue methylomes were reanalysed at the single read level to provide the first genome-wide assessment of the role of epigenetic intratumour heterogeneity in breast cancer, and the largest for any single cancer type. In ER+ tumours, a *higher epigenetic burden* but *lower DNA methylation* intratumour *heterogeneity* was associated with worse prognosis. This indicates that ordered methylation patterns that may reflect a selection process of methylation alterations in these tumours which enhances tumour fitness through the

aberrant expression of tumour suppressors and oncogenes. Conversely, in IntClust 10 tumours (that are largely ER-/ Basal-like tumours), a *higher epigenetic burden* but *higher DNA methylation intratumour heterogeneity* was associated with lower survival, implying that in these tumours, higher epigenetic diversity could enhance phenotypic plasticity, thus enabling superior tumour evolution. These findings reflect the extraordinary distinction in the role of epiallelic dynamics in different breast cancer subtypes, and provide deep insights into the role of the epigenome in cancer, that cannot be captured by other methylation platforms such as immunoprecipitation-based sequencing or microarray technologies.

The existence of another cytosine modification, 5-hydroxymethylcytosine (5hmC), was discovered in nuclear DNA in the brain [Kriaucionis and Heintz, 2009]. Subsequently, the TET family of proteins were revealed as hydroxylating enzymes that were responsible for the enzymatic conversion of 5-methylcytosine (5mC) to 5hmC resulting in DNA demethylation [Iyer et al., 2009; Tahiliani et al., 2009]. Hydroxymethylcytosine has since been proposed as a predominantly stable epigenetic modification with crucial roles in embryonic development, cellular differentiation and stem cell reprogramming [Bachman et al., 2014; Hackett et al., 2013]. Several studies have also demonstrated that hydroxymethylcytosine is positively associated with transcriptional activity [Ito et al., 2010] and is enriched at regulatory elements and within gene bodies of actively expressed genes [Madzo et al., 2014; Stroud et al., 2011]. 5-hydroxymethylcytosine levels are dramatically reduced compared to the normal tissue [Haffner et al., 2011; Song et al., 2011; Yang et al., 2013]. The global loss of hydroxymethylcytosine is an epigenetic hallmark of cancer mediated by the tumour suppressive role of TET2, and this phenotype has been observed in various malignancies including myeloid leukemias [Delhommeau et al., 2009; Langemeijer et al., 2009], melanoma [Lian et al., 2012] and glioblastomas [Johnson et al., 2016; Raiber et al., 2017], but remains understudied in breast tumours.

Although RRBS has been utilised to comprehensively map DNA methylation alterations in this breast tumour dataset, this methylation technology (along with other bisulphite conversion methods such as WGBS and Infinium microarrays) suffer from the inability to distinguish between 5-methylcytosine and 5-hydroxymethylcytosine. Consequently, the biological interpretations provided for DNA methylation might be confounded by DNA hydroxymethylation levels. However, this is not likely to alter major conclusions drastically, since hydroxymethylation levels are much rarer in the genome than methylation levels in cancer [Li et al., 2016a]. Nevertheless, a comprehensive genome-wide profiling of hydroxymethylcytosine in breast cancer

is warranted since it is a distinct epigenetic modification with an alternate role in pathogenesis. Fortunately, a new technique called oxidative bisulphite sequencing (oxBS-seq) has been developed with an additional oxidation step prior to bisulfite conversion which allows for 5-methylcytosine and 5-hydroxymethylcytosine to be distinguished [Booth et al., 2012, 2013]. One of the immediate future goals of this project is to utilise whole genome oxBS-seq in breast tumours from the METABRIC dataset presented here to fulfil two principal objectives. Firstly, since the whole epigenome will be profiled rather than a reduced fraction, this investigation will enable us to confirm, refine and expand the role of DNA methylation in breast cancer. Secondly, this will provide a comprehensive account of the DNA hydroxymethylation landscape at different genomic features and specific loci in breast cancer, and expand the view of this epigenetic modification beyond the well-reported global reduction of its levels observed in cancer.

DNA methylation signatures potentially offer key advantages over transcriptomic profiles as biomarkers in cancer. While gene expression profiles provide a snapshot of the transcriptional activity at a particular time, DNA methylation depicts a more robust characterisation of the long-term epigenetic and transcriptional state [Szyf, 2012]. Moreover, DNA methylation is a much more robust and stable marker than mRNA and therefore could serve as excellent molecular biomarkers for prediction, prognosis, monitoring and stratification of breast cancer. Spatially and temporally repeated biopsies are unfeasible for monitoring of patients' response to therapy, particularly for metastatic lesions. An emerging alternative is the detection of circulating tumour DNA (ctDNA), enabling the possibility of a liquid biopsy for systemic non-invasive monitoring of the disease [Dawson et al., 2013; Esposito et al., 2014; Murtaza et al., 2013]. Information collected by mining methylomes of the METABRIC dataset presented here will be instrumental to predict the best collection of methylation signatures to be assessed in ctDNA. Two clinical studies (DETECT, metastatic breast cancer patients; and NEO-TANGO, early breast cancer patients) have been established in collaboration with the Caldas laboratory, which could potentially be used to investigate the feasibility of the implementing methylation-based liquid biopsies.

A lack of good pre-clinical models, faithfully reflecting clinical response to therapies has brought about a high attrition rate in both drug and biomarker development. A large bio-bank of PDTX models obtained by subcutaneous implantation of surgical tumour sample cores from the Breast Cancer Unit at the Addenbrooke's Hospital into immune-compromised mice has been generated by the Caldas laboratory [Bruna et al., 2012]. These models represent breast cancer

inter-patient and *intra*-tumour heterogeneity, constituting the best preclinical model available today. They are deeply characterised with multi-dimensional molecular data including methylation profiling using RRBS, and *ex vivo* high-throughput screening of clinically approved treatments, new anti-cancer drugs and new drug-drug combinations. Work done in parallel to this thesis and in collaboration with the PDTX consortium [Bruna et al., 2012] has led to the development and optimisation of the bioinformatics RRBS pipeline with the crucial added capability of deconvoluting mouse stromal contamination in the PDTXs [Callari et al., 2017, in preparation].

The well-known paradigm of promoter hypermethylation of the *BRCA1* gene conferring sensitivity to poly (ADP-ribose) polymerase family (PARP) inhibitors [Veeck et al., 2010] has already been detected in one of the PDTX models. This indicates the potential of epigenetic biomarkers as powerful targets for therapeutic intervention in the clinical setting. It is anticipated that further integration of methylation profiles and drug sensitivity data in the panel of PDTXs can lead to the identification of new predictive epigenetic biomarkers in breast cancer, which will move us closer towards tailoring treatments to the patient's molecular profile. Moreover, comparison of drug sensitivity in matched primary and metastatic tumours, as well as serial passaging of the same PDTX tumour model, conjugated with epiclonal architecture modelling, can give insights into mechanisms of intrinsic and acquired resistance. Preliminary analysis of the epigenetic intratumour heterogeneity (PDR score) in the PDTXs have revealed that that higher epigenetic heterogeneity can drive variable response to drug therapy (PDR score vs. standard deviation of area under the curve response across all drugs tested: correlation = 0.70, *p-value* = 0.0037).

Given, the important role of DNA methylation changes in cancer, there has been much excitement regarding potential therapeutic interventions based on reversing these epigenetic abnormalities. The predominant approach for targeting DNA methylation is the utilisation of DNA methyltransferase (DNMT) inhibitors like 5-azanucleosides in order to obtain tumour suppressive effects [Barletta et al., 1997; Daskalakis et al., 2002; Jones and Taylor, 1980]. However, the unspecific nature of the DNA methyltransferase (DNMT) inhibition induced by 5-azanucleosides proved to be too toxic for clinical therapy initially. Seminal clinical work in adjusting the dose of the 5-azacytidine treatment regimen demonstrated that its use in conjunction with standard chemotherapy had significant improvements in survival [Fenaux et al., 2009; Silverman and Mufti, 2005] and led to its adoption in the clinic, although only for leukaemia [Kaminskas, 2005]. Identification of breast cancer subtype-specific methylation signatures in this thesis can pave the way for further research into incorporating the utility of the

CRISPR-Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats-Cas9) system towards the development of gene-specific and patient-targeted demethylating agents as a potential cancer therapy in breast cancer [Choudhury et al., 2016; Liu et al., 2016; Xu et al., 2016].

References

Abdel-Wahab, O., Mullally, A., Hedvat, C., Garcia-Manero, G., Patel, J., Wadleigh, M., Malinge, S., Yao, J., Kilpivaara, O., Bhat, R., Huberman, K., Thomas, S., Dolgalev, I., Heguy, A., Paietta, E., Le Beau, M. M., Beran, M., Tallman, M. S., Ebert, B. L., Kantarjian, H. M., Stone, R. M., Gilliland, D. G., Crispino, J. D., and Levine, R. L. (2009). Genetic characterization of TET1, TET2, and TET3 alterations in myeloid malignancies. *Blood*, 114(1):144–147.

Adam, M. G., Matt, S., Christian, S., Hess-Stumpp, H., Haegebarth, A., Hofmann, T. G., and Algire, C. (2015). SIAH ubiquitin ligases regulate breast cancer cell migration and invasion independent of the oxygen status. *Cell Cycle*, 14(23):3734–3747.

Adamson, E. D. (1987). Oncogenes in development. *Development (Cambridge, England)*, 99(4):449–71.

Aguirre-Ghiso, J. A., Estrada, Y., Liu, D., and Ossowski, L. (2003). ERKMAPK activity as a determinant of tumor growth and dormancy; regulation by p38SAPK. *Cancer Research*, 63(7):1684–1695.

Ahmed, A. R., Griffiths, A. B., Tilby, M. T., Westley, B. R., and May, F. E. (2012). TFF3 is a normal breast epithelial protein and is associated with differentiated phenotype in early breast cancer but predisposes to invasion and metastasis in advanced disease. *American Journal of Pathology*, 180(3):904–916.

Ahuja, N., Li, Q., Mohan, A. L., Baylin, S. B., and Issa, J. P. J. (1998). Aging and DNA methylation in colorectal mucosa and cancer. *Cancer Research*, 58(23):5489–5494.

Akalin, A., Garrett-Bakelman, F. E., Kormaksson, M., Busuttil, J., Zhang, L., Khrebtukova, I., Milne, T. A., Huang, Y., Biswas, D., Hess, J. L., Allis, C. D., Roeder, R. G., Valk, P. J. M., Löwenberg, B., Delwel, R., Fernandez, H. F., Paietta, E., Tallman, M. S., Schroth, G. P., Mason, C. E., Melnick, A., and Figueroa, M. E. (2012a). Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. *PLoS Genetics*, 8(6).

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., and Mason, C. E. (2012b). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology*, 13(10):R87.

Albertson, D. G., Collins, C., McCormick, F., and Gray, J. W. (2003). Chromosome aberrations in solid tumors. *Nature Genetics*, 34(4):369–376.

Ali, H. R., Rueda, O. M., Chin, S.-F., Curtis, C., Dunning, M. J., Aparicio, S. A., and Caldas, C. (2014). Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biology*, 15(8):431.

Ananiev, J., Tchernev, G., Patterson, J. W., Gulubova, M., and Ganchev, G. (2011). p53 - "The Guardian of Genome". *Acta Medica Bulgarica*, 38(2):72–82.

Anderson, W. F., Chatterjee, N., Ershler, W. B., and Brawley, O. W. (2002). Estrogen receptor breast cancer phenotypes in the Surveillance, Epidemiology, and End Results database. *Breast Cancer Research and Treatment*, 76(1):27–36.

Andrews Simon (2015). FastQC: a quality control tool for high throughput sequence data. *http://www.bioinformatics.babraham.ac.uk/projects/fastqc*.

Aparicio, S. and Caldas, C. (2013). The Implications of Clonal Genome Evolution for Cancer Medicine. *New England Journal of Medicine*, 368(9):842–851.

Ariazi, E. A., Ariazi, J. L., Cordera, F., and Jordan, V. C. (2006). Estrogen receptors as therapeutic targets in breast cancer. *Current topics in medicinal chemistry*, 6(May 2016):181–202.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., and Irizarry, R. A. (2014). Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30(10):1363–1369.

Aryee, M. J., Liu, W., Engelmann, J. C., Nuhn, P., Gurel, M., Haffner, M. C., Esopi, D., Irizarry, R. A., Getzenberg, R. H., Nelson, W. G., Luo, J., Xu, J., Isaacs, W. B., Bova, G. S., and Yegnasubramanian, S. (2013). DNA Methylation Alterations Exhibit Intraindividual Stability and Interindividual Heterogeneity in Prostate Cancer Metastases. *Science Translational Medicine*, 5(169):169ra10–169ra10.

Assenov, Y., Müller, F., Lutsik, P., Walter, J., Lengauer, T., and Bock, C. (2014). Comprehensive analysis of DNA methylation data with RnBeads. *Nature Methods*, 11(11):1138–1140.

Aure, M., Leivonen, S.-K., Fleischer, T., Zhu, Q., Overgaard, J., Alsner, J., Tramm, T., Louhimo, R., Alnæs, G. I., Perälä, M., Busato, F., Touleimat, N., Tost, J., Børresen-Dale, A.-L., Hautaniemi, S., Troyanskaya, O. G., Lingjærde, O., Sahlberg, K., and Kristensen, V. N. (2013). Individual and combined effects of DNA methylation and copy number alterations on miRNA expression in breast tumors. *Genome Biology*, 14(11):R126.

Babraham Bioinformatics (2016a). Bismark Bisulfite Mapper User Guide. Technical report, Babraham Bioinformatics.

Babraham Bioinformatics (2016b). Reduced Representation Bisulfite-Seq – A Brief Guide to RRBS. Technical report, Babraham Bioinformatics.

Bachman, M., Uribe-Lewis, S., Yang, X., Williams, M., Murrell, A., and Balasubramanian, S. (2014). 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nature Chemistry*, 6(12):1049–1055.

Banerji, S., Cibulskis, K., Rangel-Escareno, C., Brown, K. K., Carter, S. L., Frederick, A. M., Lawrence, M. S., Sivachenko, A. Y., Sougnez, C., Zou, L., Cortes, M. L., Fernandez-Lopez, J. C., Peng, S., Ardlie, K. G., Auclair, D., Bautista-Piña, V., Duke, F., Francis, J., Jung, J., Maffuz-Aziz, A., Onofrio, R. C., Parkin, M., Pho, N. H., Quintanar-Jurado, V., Ramos, A. H., Rebollar-Vega, R., Rodriguez-Cuevas, S., Romero-Cordoba, S. L., Schumacher, S. E., Stransky, N., Thompson, K. M., Uribe-Figueroa, L., Baselga, J., Beroukhim, R., Polyak, K., Sgroi, D. C., Richardson, A. L., Jimenez-Sanchez, G., Lander, E. S., Gabriel, S. B., Garraway, L. A., Golub, T. R., Melendez-Zajgla, J., Toker, A., Getz, G., Hidalgo-Miranda, A., and Meyerson, M. (2012). Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, 486(7403):405–409.

Barletta, J. M., Rainier, S., and Feinberg, A. P. (1997). Reversal of loss of imprinting in tumor cells by 5-aza-2'-deoxycytidine. *Cancer research*, 57(1):48–50.

Barrett, J. E., Feber, A., Herrero, J., Tanic, M., Wilson, G., Swanton, C., and Beck, S. (2017). Quantification of tumour evolution and heterogeneity via Bayesian epiallele detection. *BMC Bioinformatics*, 18 VN - r(1):354.

Baselga, J. (2010). Treatment of HER2-overexpressing breast cancer. *Annals of Oncology*, 21(suppl_7):vii36–vii40.

Bass, A. J., Thorsson, V., Shmulevich, I., Reynolds, S. M., Miller, M., Bernard, B., Hinoue, T., Laird, P. W., Curtis, C., Shen, H., Weisenberger, D. J., Schultz, N., Shen, R., Weinhold, N., Kelsen, D. P., Bowlby, R., Chu, A., Kasaian, K., Mungall, A. J., Gordon Robertson, A., Sipahimalani, P., Cherniack, A., Getz, G., Liu, Y., Noble, M. S., Pedamallu, C., Sougnez, C., Taylor-Weiner, A., Akbani, R., Lee, J.-S., Liu, W., Mills, G. B., Yang, D., Zhang, W., Pantazi, A., Parfenov, M., Gulley, M., Blanca Piazuelo, M., Schneider, B. G., Kim, J., Boussioutas, A., Sheth, M., Demchok, J. A., Rabkin, C. S., Willis, J. E., Ng, S., Garman, K., Beer, D. G., Pennathur, A., Raphael, B. J., Wu, H.-T., Odze, R., Kim, H. K., Bowen, J., Leraas, K. M., Lichtenberg, T. M., Weaver, S., McLellan, M., Wiznerowicz, M., Sakai, R., Getz, G., Sougnez, C., Lawrence, M. S., Cibulskis, K., Lichtenstein, L., Fisher, S., Gabriel, S. B., Lander, E. S., Ding, L., Niu, B., Ally, A., Balasundaram, M., Birol, I., Bowlby, R., Brooks, D., Butterfield, Y. S. N., Carlsen, R., Chu, A., Chu, J., Chuah, E., Chun, H.-J. E., Clarke, A., Dhalla, N., Guin, R., Holt, R. A., Jones, S. J. M., Kasaian, K., Lee, D., Li, H. A., Lim, E., Ma, Y., Marra, M. A., Mayo, M., Moore, R. A., Mungall, A. J., Mungall, K. L., Ming Nip, K., Gordon Robertson, A., Schein, J. E., Sipahimalani, P., Tam, A., Thiessen, N., Beroukhim, R., Carter, S. L., Cherniack, A. D., Cho, J., Cibulskis, K., DiCara, D., Frazer, S., Fisher, S., Gabriel, S. B., Gehlenborg, N., Heiman, D. I., Jung, J., Kim, J., Lander, E. S., Lawrence, M. S., Lichtenstein, L., Lin, P., Meyerson, M., Ojesina, A. I., Sekhar Pedamallu, C., Saksena, G., Schumacher, S. E., Sougnez, C., Stojanov, P., Tabak, B., Taylor-Weiner, A., Voet, D., Rosenberg, M., Zack, T. I., Zhang, H., Zou, L., Protopopov, A., Santoso, N., Parfenov, M., Lee, S., Zhang, J., Mahadeshwar, H. S., Tang, J., Ren, X., Seth, S., Yang, L., Xu, A. W., Song, X., Pantazi, A., Xi, R., Bristow, C. A., Hadjipanayis, A., Seidman, J., Chin, L., Park, P. J., Kucherlapati, R., Akbani, R., Ling, S., Liu, W., Rao, A.,

Weinstein, J. N., Kim, S.-B., Lee, J.-S., Lu, Y., Mills, G., Laird, P. W., Hinoue, T., Weisenberger, D. J., Bootwalla, M. S., Lai, P. H., Shen, H., Triche Jr, T., Van Den Berg, D. J., Baylin, S. B., Herman, J. G., Getz, G., Chin, L., Liu, Y., Murray, B. A., Noble, M. S., Arman Askoy, r. B., Ciriello, G., Dresdner, G., Gao, J., Gross, B., Jacobsen, A., Lee, W., Ramirez, R., Sander, C., Schultz, N., Senbabaoglu, Y., Sinha, R., Onur Sumer, S., Sun, Y., Weinhold, N., Thorsson, V., Bernard, B., Iype, L., Kramer, R. W., Kreisberg, R., Miller, M., Reynolds, S. M., Rovira, H., Tasman, N., Shmulevich, I., Ng, S., Haussler, D., Stuart, J. M., Akbani, R., Ling, S., Liu, W., Rao, A., Weinstein, J. N., Verhaak, R. G. W., Mills, G. B., Leiserson, M. D. M., Raphael, B. J., Wu, H.-T., Taylor, B. S., Black, A. D., Bowen, J., Ann Carney, J., Gastier-Foster, J. M., Helsel, C., Leraas, K. M., Lichtenberg, T. M., McAllister, C., Ramirez, N. C., Tabler, T. R., Wise, L., Zmuda, E., Penny, R., Crain, D., Gardner, J., Lau, K., Curely, E., Mallery, D., Morris, S., Paulauskis, J., Shelton, T., Shelton, C., Sherman, M., Benz, C., Lee, J.-H., Fedosenko, K., Manikhas, G., Potapova, O., Voronina, O., Belyaev, D., Dolzhansky, O., Kimryn Rathmell, W., Brzezinski, J., Ibbs, M., Korski, K., Kycler, W., Łaźniak, R., Leporowska, E., Mackiewicz, A., Murawa, D., Murawa, P., Spychała, A., Suchorska, W. M., Tatka, H., Teresiak, M., Wiznerowicz, M., Abdel-Misih, R., Bennett, J., Brown, J., Iacocca, M., Rabeno, B., Kwon, S.-Y., Penny, R., Gardner, J., Kemkes, A., Mallery, D., Morris, S., Shelton, T., Shelton, C., Curley, E., Alexopoulou, I., Engel, J., Bartlett, J., Albert, M., Park, D.-Y., Dhir, R., Luketich, J., Landreneau, R., Janjigian, Y. Y., Kelsen, D. P., Cho, E., Ladanyi, M., Tang, L., McCall, S. J., Park, Y. S., Cheong, J.-H., Ajani, J., Constanza Camargo, M., Alonso, S., Ayala, B., Jensen, M. A., Pihl, T., Raman, R., Walton, J., Wan, Y., Demchok, J. A., Eley, G., Mills Shaw, K. R., Sheth, M., Tarnuzzer, R., Wang, Z., Yang, L., Claude Zenklusen, J., Davidsen, T., Hutter, C. M., Sofia, H. J., Burton, R., Chudamani, S., and Liu, J. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, 513(7517):202–209.

Batra, R. (2015). Characterising the epigenetic landscape of breast cancer. Technical report, University of Cambridge.

Baylin, S. B. (2005). DNA methylation and gene silencing in cancer. *Nature Clinical Practice Oncology*, 2:S4–S11.

Baylin, S. B. and Jones, P. A. (2011). A decade of exploring the cancer epigenome — biological and translational implications. *Nature Reviews Cancer*, 11(10):726–734.

Baylin, S. B. and Ohm, J. E. (2006). Epigenetic gene silencing in cancer – a mechanism for early oncogenic pathway addiction? *Nature Reviews Cancer*, 6(2):107–116.

Bediaga, N. G., Acha-Sagredo, A., Guerra, I., Viguri, A., Albaina, C., Ruiz Diaz, I., Rezola, R., Alberdi, M. J., Dopazo, J., Montaner, D., de Renobales, M., Fernández, A. F., Field, J. K., Fraga, M. F., Liloglou, T., and de Pancorbo, M. M. (2010). DNA methylation epigenotypes in breast cancer molecular subtypes. *Breast Cancer Research*, 12(5):R77.

Bengtsson, H., Ray, A., Spellman, P., and Speed, T. P. (2009). A single-sample method for normalizing and combining full-resolution copy numbers from multiple platforms, labs and analysis methods. *Bioinformatics*, 25(7):861–867.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300.

Berman, B. P., Weisenberger, D. J., Aman, J. F., Hinoue, T., Ramjan, Z., Liu, Y., Noushmehr, H., Lange, C. P. E., van Dijk, C. M., Tollenaar, R. A. E. M., Van Den Berg, D., and Laird, P. W. (2011). Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina–associated domains. *Nature Genetics*, 44(1):40–46.

Berman, H., Zhang, J., Crawford, Y. G., Gauthier, M. L., Fordyce, C. A., McDermott, K. M., Sigaroudinia, M., Kozakiewicz, K., and Tlsty, T. D. (2005). Genetic and epigenetic changes in mammary epithelial cells identify a subpopulation of cells involved in early carcinogenesis. *Cold Spring Harbor Symposia on Quantitative Biology*, 70:317–327.

Berx, G., Becker, K. F., Höfler, H., and Van Roy, F. (1998). Mutations of the human E-cadherin (CDH1) gene. *Human Mutation*, 12(4):226–237.

Berx, G. and van Roy, F. (2009). Involvement of members of the cadherin superfamily in cancer.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., Fan, J. B., and Shen, R. (2011). High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4):288–295.

Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R., and Gunderson, K. L. (2009). Genome-wide DNA methylation profiling using Infinium ® assay. *Epigenomics*, 1(1):177–200.

Bibikova, M., Lin, Z., Zhou, L., Chudin, E., Garcia, E. W., Wu, B., Doucet, D., Thomas, N. J., Wang, Y., Vollmer, E., Goldmann, T., Seifart, C., Jiang, W., Barker, D. L., Chee, M. S., Floros, J., and Fan, J. B. (2006). High-throughput DNA methylation profiling using universal bead arrays. *Genome Research*, 16(3):383–393.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes and Development*, 16(1):6–21.

Bird, A. P. and Wolffe, A. P. (1999). Methylation-induced repression-belts, braces, and chromatin. *Cell*, 99(5):451–454.

Bock, C. (2012). Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*, 13(10):705–719.

Bock, C., Tomazou, E. M., Brinkman, A. B., Müller, F., Simmer, F., Gu, H., Jäger, N., Gnirke, A., Stunnenberg, H. G., and Meissner, A. (2010). Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nature Biotechnology*, 28(10):1106–1114.

Bonneterre, J., Thürlimann, B., Robertson, J. F., Krzakowski, M., Mauriac, L., Koralewski, P., Vergote, I., Webster, A., Steinberg, M., and Von Euler, M. (2000). Anastrozole versus tamoxifen as first-line therapy for advanced breast cancer in 668 postmenopausal women: Results of the tamoxifen or arimidex randomized group efficacy and tolerability study. *Journal of Clinical Oncology*, 18(22):3748–3757.

Booth, M. J., Branco, M. R., Ficz, G., Oxley, D., Krueger, F., Reik, W., and Balasubramanian, S. (2012). Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution. *Science*, 336(6083):934–937.

Booth, M. J., Ost, T. W. B., Beraldi, D., Bell, N. M., Branco, M. R., Reik, W., and Balasubramanian, S. (2013). Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nature Protocols*, 8(10):1841–1851.

Bos, J. L. (1989). ras Oncogenes in Human Cancer : A Review ras Oncogenes in Human Cancer : A Review. *Cancer Research*, 49(17):4682–4689.

Boyle, P., Clement, K., Gu, H., Smith, Z. D., Ziller, M., Fostel, J. L., Holmes, L., Meldrim, J., Kelley, F., Gnirke, A., and Meissner, A. (2012). Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. *Genome Biology*, 13(10):R92.

Brocks, D., Assenov, Y., Minner, S., Bogatyrova, O., Simon, R., Koop, C., Oakes, C., Zucknick, M., Lipka, D. B., Weischenfeldt, J., Feuerbach, L., Cowper-Sallari, R., Lupien, M., Brors, B., Korbel, J., Schlomm, T., Tanay, A., Sauter, G., Gerhäuser, C., and Plass, C. (2014). Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell Reports*, 8(3):798–806.

Bruna, A., Greenwood, W., Le Quesne, J., Teschendorff, A., Miranda-Saavedra, D., Rueda, O. M., Sandoval, J. L., Vidakovic, A. T., Saadi, A., Pharoah, P., Stingl, J., and Caldas, C. (2012). TGFβ induces the formation of tumour-initiating cells in claudinlow breast cancer. *Nature Communications*, 3:1055.

Bruna, A., Rueda, O. M., Greenwood, W., Batra, A. S., Callari, M., Batra, R. N., Pogrebniak, K., Sandoval, J., Cassidy, J. W., Tufegdzic-Vidakovic, A., Sammut, S. J., Jones, L., Provenzano, E., Baird, R., Eirew, P., Hadfield, J., Eldridge, M., McLaren-Douglas, A., Barthorpe, A., Lightfoot, H., O'Connor, M. J., Gray, J., Cortes, J., Baselga, J., Marangoni, E., Welm, A. L., Aparicio, S., Serra, V., Garnett, M. J., and Caldas, C. (2016). A Biobank of Breast Cancer Explants with Preserved Intra-tumor Heterogeneity to Screen Anticancer Compounds. *Cell*, 167(1):260–274.e22.

Callari, M., Batra, A. S., Batra, R. N., Sammut, S.-J., Greenwood, W., Clifford, H., Hercus, C., Chin, S.-F., Bruna, A., Rueda, O. M., and Caldas, C. (2017). Computational approach to discriminate human and mouse sequences in patient-derived tumour xenografts. *in preparation*.

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70.

Cancer Genome Atlas Research Network (2016). Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *New England Journal of Medicine*, 374(2):135–145.

Cardoso, F., Van't Veer, L., Rutgers, E., Loi, S., Mook, S., and Piccart-Gebhart, M. J. (2008). Clinical application of the 70-gene profile: The MINDACT trial. *Journal of Clinical Oncology*, 26(5):729–735.

Carter, C. L., Allen, C., and Henson, D. E. (1989). Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer*, 63(1):181–187.

Castilla, L. H., Couch, F. J., Erdos, M. R., Hoskins, K. F., Calzone, K., Garber, J. E., Boyd, J., Lubin, M. B., Deshano, M. L., Brody, L. C., Collins, F. S., and Weber, B. L. (1994). Mutations in the BRCA1 gene in families with early-onset breast and ovarian cancer. *Nature Genetics*, 8(4):387–391.

Cavenee, W. K., Dryja, T. P., Phillips, R. a., Benedict, W. F., Godbout, R., Gallie, B. L., Murphree, a. L., Strong, L. C., and White, R. L. (1983). Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. *Nature*, 305(5937):779–784.

Cedar, H. and Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics*, 10(5):295–304.

Chai, H. and Brown, R. E. (2009). Review: Field effect in cancer-an update. *Annals of Clinical and Laboratory Science*, 39(4):331–338.

Chapellier, M., Bachelard-Cascales, E., Schmidt, X., Clément, F., Treilleux, I., Delay, E., Jammot, A., Ménétrier-Caux, C., Pochon, G., Besançon, R., Voeltzel, T., Caron De Fromentel, C., Caux, C., Blay, J. Y., Iggo, R., and Maguer-Satta, V. (2015). Disequilibrium of BMP2 levels in the breast stem cell niche launches epithelial transformation by overamplifying BMPR1B cell response. *Stem Cell Reports*, 4(2):239–254.

Chen, P.-Y., Cokus, S. J., and Pellegrini, M. (2010). BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics*, 11(1):203.

Chiang, S., Weigelt, B., Wen, H. C., Pareja, F., Raghavendra, A., Martelotto, L. G., Burke, K. A., Basili, T., Li, A., Geyer, F. C., Piscuoglio, S., Ng, C. K., Jungbluth, A. A., Balss, J., Pusch, S., Baker, G. M., Cole, K. S., Von Deimling, A., Batten, J. M., Marotti, J. D., Soh, H. C., McCalip, B. L., Serrano, J., Lim, R. S., Siziopikou, K. P., Lu, S., Liu, X., Hammour, T., Brogi, E., Snuderl, M., Iafrate, A. J., Reis-Filho, J. S., and Schnitt, S. J. (2016). IDH2 mutations define a unique subtype of breast cancer with altered nuclear polarity. *Cancer Research*, 76(24):7118–7129.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W. L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T., Chen, F., Feiler, H., Tokuyasu, T., Kingsley, C., Dairkee, S., Meng, Z., Chew, K., Pinkel, D., Jain, A., Ljung, B. M., Esserman, L., Albertson, D. G., Waldman, F. M., and Gray, J. W. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, 10(6):529–541.

Chin, S. F., Teschendorff, A. E., Marioni, J. C., Wang, Y., Barbosa-Morais, N. L., Thorne, N. P., Costa, J. L., Pinder, S. E., van de Wiel, M. A., Green, A. R., Ellis, I. O., Porter, P. L., Tavaré, S., Brenton, J. D., Ylstra, B., and Caldas, C. (2007). High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biology*, 8(10):R215.

Choudhury, S. R., Cui, Y., Lubecka, K., and Stefanska, B. (2016). CRISPR-dCas9 mediated TET1 targeting for selective DNA demethylation at BRCA1 promoter. *Oncotarget*, 7(Ii):1–12.

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., and Getz, G. (2013). Sensitive detection

of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3):213–219.

Cimino-Mathews, A., Subhawong, A. P., Elwood, H., Warzecha, H. N., Sharma, R., Park, B. H., Taube, J. M., Illei, P. B., and Argani, P. (2013). Neural crest transcription factor Sox10 is preferentially expressed in triple-negative and metaplastic breast carcinomas. *Human Pathology*, 44(6):959–965.

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Research*, 22(2):398–406.

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., Bowlby, R., Shen, H., Hayat, S., Fieldhouse, R., Lester, S. C., Tse, G. M., Factor, R. E., Collins, L. C., Allison, K. H., Chen, Y. Y., Jensen, K., Johnson, N. B., Oesterreich, S., Mills, G. B., Cherniack, A. D., Robertson, G., Benz, C., Sander, C., Laird, P. W., Hoadley, K. A., King, T. A., Akbani, R., Auman, J. T., Balasundaram, M., Balu, S., Barr, T., Benz, S., Berrios, M., Beroukhim, R., Bodenheimer, T., Boice, L., Bootwalla, M. S., Bowen, J., Brooks, D., Chin, L., Cho, J., Chudamani, S., Davidsen, T., Demchok, J. A., Dennison, J. B., Ding, L., Felau, I., Ferguson, M. L., Frazer, S., Gabriel, S. B., Gao, J. J., Gastier-Foster, J. M., Gehlenborg, N., Gerken, M., Getz, G., Gibson, W. J., Hayes, D. N., Heiman, D. I., Holbrook, A., Holt, R. A., Hoyle, A. P., Hu, H., Huang, M., Hutter, C. M., Hwang, E. S., Jefferys, S. R., Jones, S. J., Ju, Z., Kim, J., Lai, P. H., Lawrence, M. S., Leraas, K. M., Lichtenberg, T. M., Lin, P., Ling, S., Liu, J., Liu, W., Lolla, L., Lu, Y., Ma, Y., Maglinte, D. T., Mardis, E., Marks, J., Marra, M. A., McAllister, C., Meng, S., Meyerson, M., Moore, R. A., Mose, L. E., Mungall, A. J., Murray, B. A., Naresh, R., Noble, M. S., Olopade, O., Parker, J. S., Pihl, T., Saksena, G., Schumacher, S. E., Shaw, K. R., Ramirez, N. C., Rathmell, W. K., Roach, J., Robertson, A. G., Schein, J. E., Schultz, N., Sheth, M., Shi, Y., Shih, J., Shelley, C. S., Shriver, C., Simons, J. V., Sofia, H. J., Soloway, M. G., Sougnez, C., Sun, C., Tarnuzzer, R., Tiezzi, D. G., Van Den Berg, D. J., Voet, D., Wan, Y., Wang, Z., Weinstein, J. N., Weisenberger, D. J., Wilson, R., Wise, L., Wiznerowicz, M., Wu, J., Wu, Y., Yang, L., Zack, T. I., Zenklusen, J. C., Zhang, J., Zmuda, E., and Perou, C. M. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell*, 163(2):506–519.

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, 45(10):1127–1133.

Cohen, N. M., Kenigsberg, E., and Tanay, A. (2011). Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell*, 145(5):773–786.

Collisson, E. A., Campbell, J. D., Brooks, A. N., Berger, A. H., Lee, W., Chmielecki, J., Beer, D. G., Cope, L., Creighton, C. J., Danilova, L., Ding, L., Getz, G., Hammerman, P. S., Neil Hayes, D., Hernandez, B., Herman, J. G., Heymach, J. V., Jurisica, I., Kucherlapati, R., Kwiatkowski, D., Ladanyi, M., Robertson, G., Schultz, N., Shen, R., Sinha, R., Sougnez, C., Tsao, M.-S., Travis, W. D., Weinstein, J. N., Wigle, D. A., Wilkerson, M. D., Chu, A., Cherniack, A. D., Hadjipanayis, A., Rosenberg, M., Weisenberger, D. J., Laird, P. W., Radenbaugh, A., Ma, S., Stuart,

J. M., Averett Byers, L., Baylin, S. B., Govindan, R., Meyerson, M., Rosenberg, M., Gabriel, S. B., Cibulskis, K., Sougnez, C., Kim, J., Stewart, C., Lichtenstein, L., Lander, E. S., Lawrence, M. S., Getz, G., Kandoth, C., Fulton, R., Fulton, L. L., McLellan, M. D., Wilson, R. K., Ye, K., Fronick, C. C., Maher, C. A., Miller, C. A., Wendl, M. C., Cabanski, C., Ding, L., Mardis, E., Govindan, R., Creighton, C. J., Wheeler, D., Balasundaram, M., Butterfield, Y. S. N., Carlsen, R., Chu, A., Chuah, E., Dhalla, N., Guin, R., Hirst, C., Lee, D., Li, H. I., Mayo, M., Moore, R. A., Mungall, A. J., Schein, J. E., Sipahimalani, P., Tam, A., Varhol, R., Gordon Robertson, A., Wye, N., Thiessen, N., Holt, R. A., Jones, S. J. M., Marra, M. A., Campbell, J. D., Brooks, A. N., Chmielecki, J., Imielinski, M., Onofrio, R. C., Hodis, E., Zack, T., Sougnez, C., Helman, E., Sekhar Pedamallu, C., Mesirov, J., Cherniack, A. D., Saksena, G., Schumacher, S. E., Carter, S. L., Hernandez, B., Garraway, L., Beroukhim, R., Gabriel, S. B., Getz, G., Meyerson, M., Hadjipanayis, A., Lee, S., Mahadeshwar, H. S., Pantazi, A., Protopopov, A., Ren, X., Seth, S., Song, X., Tang, J., Yang, L., Zhang, J., Chen, P.-C., Parfenov, M., Wei Xu, A., Santoso, N., Chin, L., Park, P. J., Kucherlapati, R., Hoadley, K. A., Todd Auman, J., Meng, S., Shi, Y., Buda, E., Waring, S., Veluvolu, U., Tan, D., Mieczkowski, P. A., Jones, C. D., Simons, J. V., Soloway, M. G., Bodenheimer, T., Jefferys, S. R., Roach, J., Hoyle, A. P., Wu, J., Balu, S., Singh, D., Prins, J. F., Marron, J., Parker, J. S., Neil Hayes, D., Perou, C. M., Liu, J., Cope, L., Danilova, L., Weisenberger, D. J., Maglinte, D. T., Lai, P. H., Bootwalla, M. S., Van Den Berg, D. J., Triche Jr, T., Baylin, S. B., Laird, P. W., Rosenberg, M., Chin, L., Zhang, J., Cho, J., DiCara, D., Heiman, D., Lin, P., Mallard, W., Voet, D., Zhang, H., Zou, L., Noble, M. S., Lawrence, M. S., Saksena, G., Gehlenborg, N., Thorvaldsdottir, H., Mesirov, J., Nazaire, M.-D., Robinson, J., Getz, G., Lee, W., Arman Aksoy, B., Ciriello, G., Taylor, B. S., Dresdner, G., Gao, J., Gross, B., Seshan, V. E., Ladanyi, M., Reva, B., Sinha, R., Onur Sumer, S., Weinhold, N., Schultz, N., Shen, R., Sander, C., Ng, S., Ma, S., Zhu, J., Radenbaugh, A., Stuart, J. M., Benz, C. C., Yau, C., Haussler, D., Spellman, P. T., Wilkerson, M. D., Parker, J. S., Hoadley, K. A., Kimes, P. K., Neil Hayes, D., Perou, C. M., Broom, B. M., Wang, J., Lu, Y., Kwok Shing Ng, P., Diao, L., Averett Byers, L., Liu, W., Heymach, J. V., Amos, C. I., Weinstein, J. N., Akbani, R., Mills, G. B., Curley, E., Paulauskis, J., Lau, K., Morris, S., Shelton, T., Mallery, D., Gardner, J., Penny, R., Saller, C., Tarvin, K., Richards, W. G., Cerfolio, R., Bryant, A., Raymond, D. P., Pennell, N. A., Farver, C., Czerwinski, C., Huelsenbeck-Dill, L., Iacocca, M., Petrelli, N., Rabeno, B., Brown, J., Bauer, T., Dolzhanskiy, O., Potapova, O., Rotin, D., Voronina, O., Nemirovich-Danchenko, E., Fedosenko, K. V., Gal, A., Behera, M., Ramalingam, S. S., Sica, G., Flieder, D., Boyd, J., Weaver, J., Kohl, B., Huy Quoc Thinh, D., Sandusky, G., Juhl, H., Duhig, E., Illei, P., Gabrielson, E., Shin, J., Lee, B., Rogers, K., Trusty, D., Brock, M. V., Williamson, C., Burks, E., Rieger-Christ, K., Holway, A., Sullivan, T., Wigle, D. A., Asiedu, M. K., Kosari, F., Travis, W. D., Rekhtman, N., Zakowski, M., Rusch, V. W., Zippile, P., Suh, J., Pass, H., Goparaju, C., Owusu-Sarpong, Y., Bartlett, J. M. S., Kodeeswaran, S., Parfitt, J., Sekhon, H., Albert, M., Eckman, J., Myers, J. B., Cheney, R., Morrison, C., Gaudioso, C., Borgia, J. A., Bonomi, P., Pool, M., Liptay, M. J., Moiseenko, F., Zaytseva, I., Dienemann, H., Meister, M., Schnabel, P. A., Muley, T. R., Peifer, M., Gomez-Fernandez, C., Herbert, L., Egea, S., Huang, M., Thorne, L. B., Boice, L., Hill Salazar, A., Funkhouser, W. K., Kimryn Rathmell, W., Dhir, R., Yousem, S. A., Dacic, S., Schneider, F., Siegfried, J. M., Hajek, R., Watson, M. A., McDonald, S., Meyers, B., Clarke, B., Yang, I. A., Fong, K. M.,

Hunter, L., Windsor, M., Bowman, R. V., Peters, S., Letovanec, I., Khan, K. Z., Jensen, M. A., Snyder, E. E., Srinivasan, D., Kahn, A. B., Baboud, J., Pot, D. A., Mills Shaw, K. R., Sheth, M., Davidsen, T., Demchok, J. A., Yang, L., Wang, Z., Tarnuzzer, R., Claude Zenklusen, J., Ozenberger, B. A., Sofia, H. J., Travis, W. D., Cheney, R., Clarke, B., Dacic, S., Duhig, E., Funkhouser, W. K., Illei, P., Farver, C., Rekhtman, N., Sica, G., Suh, J., and Tsao, M.-S. (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550.

Comings, D. E. (1973). A general theory of carcinogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 70(12):3324–3328.

Contesso, G., Mouriesse, H., Friedman, S., Genin, J., Sarrazin, D., and Rouesse, J. (1987). The importance of histologic grade in long-term prognosis of breast cancer: a study of 1,010 patients, uniformly treated at the Institut Gustave-Roussy. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 5(9):1378–1386.

Cooper, C. S., Eeles, R., Wedge, D. C., Van Loo, P., Gundem, G., Alexandrov, L. B., Kremeyer, B., Butler, A., Lynch, A. G., Camacho, N., Massie, C. E., Kay, J., Luxton, H. J., Edwards, S., Kote-Jarai, Z., Dennis, N., Merson, S., Leongamornlert, D., Zamora, J., Corbishley, C., Thomas, S., Nik-Zainal, S., O'Meara, S., Matthews, L., Clark, J., Hurst, R., Mithen, R., Bristow, R. G., Boutros, P. C., Fraser, M., Cooke, S., Raine, K., Jones, D., Menzies, A., Stebbings, L., Hinton, J., Teague, J., McLaren, S., Mudie, L., Hardy, C., Anderson, E., Joseph, O., Goody, V., Robinson, B., Maddison, M., Gamble, S., Greenman, C., Berney, D., Hazell, S., Livni, N., Fisher, C., Ogden, C., Kumar, P., Thompson, A., Woodhouse, C., Nicol, D., Mayer, E., Dudderidge, T., Shah, N. C., Gnanapragasam, V., Voet, T., Campbell, P., Futreal, A., Easton, D., Warren, A. Y., Foster, C. S., Stratton, M. R., Whitaker, H. C., McDermott, U., Brewer, D. S., and Neal, D. E. (2015). Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nature Genetics*, 47(4):367–372.

Coulondre, C., Miller, J. H., Farabaugh, P. J., and Gilbert, W. (1978). Molecular basis of base substitution hotspots in Escherichia coli. *Nature*, 274(5673):775–780.

Cox, D. (1972). Regression models and life tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Caldas, C., Aparicio, S., Curtis†, C., Shah, S. P., Caldas, C., Aparicio, S., Brenton, J. D., Ellis, I., Huntsman, D., Pinder, S., Purushotham, A., Murphy, L., Caldas, C., Aparicio, S., Caldas, C., Bardwell, H., Chin, S.-F., Curtis, C., Ding, Z., Gräf, S., Jones, L., Liu, B., Lynch, A. G., Papatheodorou, I., Sammut, S. J., Wishart, G., Aparicio, S., Chia, S., Gelmon, K., Huntsman, D., McKinney, S., Speers, C., Turashvili, G., Watson, P., Ellis, I., Blamey, R., Green, A., Macmillan, D., Rakha, E., Purushotham, A., Gillett, C., Grigoriadis, A., Pinder, S., di Rinaldis, E., Tutt, A., Murphy, L., Parisien, M., Troup, S., Caldas, C., Chin, S.-F., Chan, D., Fielding, C., Maia, A.-T., McGuire, S., Osborne, M., Sayalero, S. M., Spiteri, I., Hadfield, J., Aparicio, S., Turashvili, G., Bell, L., Chow,

K., Gale, N., Huntsman, D., Kovalik, M., Ng, Y., Prentice, L., Caldas, C., Tavaré, S., Curtis, C., Dunning, M. J., Gräf, S., Lynch, A. G., Rueda, O. M., Russell, R., Samarajiwa, S., Speed, D., Markowetz, F., Yuan, Y., Brenton, J. D., Aparicio, S., Shah, S. P., Bashashati, A., Ha, G., Haffari, G., McKinney, S., Langerød, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Børresen-Dale, A.-L., Brenton, J. D., Tavaré, S., Caldas, C., and Aparicio, S. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486 VN -(7403):346–352.

Dai, M., Al-Odaini, A. A., Arakelian, A., Rabbani, S. A., Ali, S., and Lebrun, J.-J. (2012). A novel function for p21Cip1 and acetyltransferase p/CAF as critical transcriptional regulators of TGFβ-mediated breast cancer cell migration and invasion. *Breast Cancer Research*, 14(5):3055.

Daskalakis, M., Nguyen, T. T., Nguyen, C., Guldberg, P., Köhler, G., Wijermans, P., Jones, P. A., and Lübbert, M. (2002). Demethylation of a hypermethylated P15/INK4B gene in patients with myelodysplastic syndrome by 5-Aza-2'-deoxycytidine (decitabine) treatment. *Blood*, 100(8):2957–64.

Dawson, S.-J., Rueda, O. M., Aparicio, S., and Caldas, C. (2013). A new genome-driven integrated classification of breast cancer and its implications. *The EMBO Journal*, 32(5):617–628.

Day, K., Waite, L. L., Thalacker-Mercer, A., West, A., Bamman, M. M., Brooks, J. D., Myers, R. M., and Absher, D. (2013). Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. *Genome Biology*, 14(9):R102.

Deaton, A. M. and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes and Development*, 25(10):1010–1022.

Dedeurwaerder, S., Desmedt, C., Calonne, E., Singhal, S. K., Haibe-Kains, B., Defrance, M., Michiels, S., Volkmar, M., Deplus, R., Luciani, J., Lallemand, F., Larsimont, D., Toussaint, J., Haussy, S., Rothé, F., Rouas, G., Metzger, O., Majjaj, S., Saini, K., Putmans, P., Hames, G., van Baren, N., Coulie, P. G., Piccart, M., Sotiriou, C., and Fuks, F. (2011). DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO Molecular Medicine*, 3(12):726–741.

Delhommeau, F., Dupont, S., Valle, V. D., James, C., Trannoy, S., Massé, A., Kosmider, O., Le Couedic, J.-P., Robert, F., Alberdi, A., Lécluse, Y., Plo, I., Dreyfus, F. J., Marzac, C., Casadevall, N., Lacombe, C., Romana, S. P., Dessen, P., Soulier, J., Viguié, F., Fontenay, M., Vainchenker, W., and Bernard, O. A. (2009). Mutation in TET2 in Myeloid Cancers. *New England Journal of Medicine*, 360(22):2289–2301.

Dent, R., Trudeau, M., Pritchard, K. I., Hanna, W. M., Kahn, H. K., Sawka, C. A., Lickley, L. A., Rawlinson, E., Sun, P., and Narod, S. A. (2007). Triple-negative breast cancer: Clinical features and patterns of recurrence. *Clinical Cancer Research*, 13(15):4429–4434.

Diaz Jr, L. A., Williams, R. T., Wu, J., Kinde, I., Hecht, J. R., Berlin, J., Allen, B., Bozic, I., Reiter, J. G., Nowak, M. A., Kinzler, K. W., Oliner, K. S., and Vogelstein, B. (2012). The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature*.

Dickinson, R. E., Dallol, A., Bieche, I., Krex, D., Morton, D., Maher, E. R., and Latif, F. (2004). Epigenetic inactivation of SLIT3 and SLIT1 genes in human cancers. *British Journal of Cancer*, 91(12):2071–2078.

Dolzhenko, E. and Smith, A. D. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics*, 15(1):215.

Domcke, S., Bardet, A. F., Adrian Ginno, P., Hartl, D., Burger, L., and Schübeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature*, 528(7583):575–579.

Dressman, M. A., Baras, A., Malinowski, R., Alvis, L. B., Kwon, I., Walz, T. M., and Polymeropoulos, M. H. (2003). Gene expression profiling detects gene amplification and differentiates tumor types in breast cancer. *Cancer Research*, 63(9):2194–2199.

Dryden, N. H., Broome, L. R., Dudbridge, F., Johnson, N., Orr, N., Schoenfelder, S., Nagano, T., Andrews, S., Wingett, S., Kozarewa, I., Assiotis, I., Fenwick, K., Maguire, S. L., Campbell, J., Natrajan, R., Lambros, M., Perrakis, E., Ashworth, A., Fraser, P., and Fletcher, O. (2014). Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Research*, 24(11):1854–1868.

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., and Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11(1):587.

Dunning, M. J., Smith, M. L., Ritchie, M. E., and Tavaré, S. (2007). Beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics*, 23(16):2183–2184.

Dvinge, H., Git, A., Gräf, S., Salmon-Divon, M., Curtis, C., Sottoriva, A., Zhao, Y., Hirst, M., Armisen, J., Miska, E. A., Chin, S.-F., Provenzano, E., Turashvili, G., Green, A., Ellis, I., Aparicio, S., and Caldas, C. (2013). The shaping and functional consequences of the microRNA landscape in breast cancer. *Nature*, 497(7449):378–382.

Early Breast Cancer Trialists' Collaborative Group (EBCTCG) (1998). Tamoxifen for early breast cancer: an overview of the randomised trials. *The Lancet*, 351(9114):1451–1467.

Early Breast Cancer Trialists' Collaborative Group (EBCTCG) (2005). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: An overview of the randomised trials. *Lancet*, 365(9472):1687–1717.

Easton, D. F., Bishop, D. T., Ford, D., and Crockford, G. P. (1993). Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium. *American journal of human genetics*, 52(4):678–701.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., Haefliger, C., Horton, R., Howe, K., Jackson, D. K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K., and Beck, S. (2006).

DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics*, 38(12):1378–1385.

Ehrlich, M. (2002). DNA methylation in cancer: too much, but also too little. *Oncogene*, 21(35):5400–5413.

Ehrlich, M. (2009). DNA hypomethylation in cancer cells. *Epigenomics*, 1(2):239–259.

Ehrlich, M., Gama-Sosa, M. A., Huang, L. H., Midgett, R. M., Kuo, K. C., Mccune, R. A., and Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucleic Acids Research*, 10(8):2709–2721.

Eirew, P., Steif, A., Khattra, J., Ha, G., Yap, D., Farahani, H., Gelmon, K., Chia, S., Mar, C., Wan, A., Laks, E., Biele, J., Shumansky, K., Rosner, J., McPherson, A., Nielsen, C., Roth, A. J. L., Lefebvre, C., Bashashati, A., de Souza, C., Siu, C., Aniba, R., Brimhall, J., Oloumi, A., Osako, T., Bruna, A., Sandoval, J. L., Algara, T., Greenwood, W., Leung, K., Cheng, H., Xue, H., Wang, Y., Lin, D., Mungall, A. J., Moore, R., Zhao, Y., Lorette, J., Nguyen, L., Huntsman, D., Eaves, C. J., Hansen, C., Marra, M. A., Caldas, C., Shah, S. P., and Aparicio, S. (2014). Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, 518(7539):422–426.

Ellis, M. J., Ding, L., Shen, D., Luo, J., Suman, V. J., Wallis, J. W., Van Tine, B. A., Hoog, J., Goiffon, R. J., Goldstein, T. C., Ng, S., Lin, L., Crowder, R., Snider, J., Ballman, K., Weber, J., Chen, K., Koboldt, D. C., Kandoth, C., Schierding, W. S., McMichael, J. F., Miller, C. A., Lu, C., Harris, C. C., McLellan, M. D., Wendl, M. C., DeSchryver, K., Allred, D. C., Esserman, L., Unzeitig, G., Margenthaler, J., Babiera, G. V., Marcom, P. K., Guenther, J. M., Leitch, M., Hunt, K., Olson, J., Tao, Y., Maher, C. A., Fulton, L. L., Fulton, R. S., Harrison, M., Oberkfell, B., Du, F., Demeter, R., Vickery, T. L., Elhammali, A., Piwnica-Worms, H., McDonald, S., Watson, M., Dooling, D. J., Ota, D., Chang, L.-W., Bose, R., Ley, T. J., Piwnica-Worms, D., Stuart, J. M., Wilson, R. K., and Mardis, E. R. (2012). Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature*, 486 VN -(7403):353–360.

Elston, C. W., Ellis, I. O., and Pinder, S. E. (1999). Pathological prognostic factors in breast cancer. *Critical Reviews in Oncology/Hematology*, 31(3):209–223.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49.

Esposito, A., Bardelli, A., Criscitiello, C., Colombo, N., Gelao, L., Fumagalli, L., Minchella, I., Locatelli, M., Goldhirsch, A., and Curigliano, G. (2014). Monitoring tumor-derived cell-free DNA in patients with solid tumors: Clinical perspectives and research opportunities. *Cancer Treatment Reviews*, 40(5):648–655.

Esteller, M. (2000). Promoter Hypermethylation and BRCA1 Inactivation in Sporadic Breast and Ovarian Tumors. *Journal of the National Cancer Institute*, 92(7):564–569.

Esteller, M. (2008). Epigenetics in Cancer. *New England Journal of Medicine*, 358(11):1148–1159.

Fackler, M. J., Umbricht, C. B., Williams, D., Argani, P., Cruz, L. A., Merino, V. F., Teo, W. W., Zhang, Z., Huang, P., Visvananthan, K., Marks, J., Ethier, S., Gray, J. W., Wolff, A. C., Cope, L. M., and Sukumar, S. (2011). Genome-wide methylation analysis identifies genes specific to breast cancer hormone receptor status and risk of recurrence. *Cancer Research*, 71(19):6195–6207.

Fang, F., Turcan, S., Rimner, A., Kaufman, A., Giri, D., Morris, L. G. T., Shen, R., Seshan, V., Mo, Q., Heguy, A., Baylin, S. B., Ahuja, N., Viale, A., Massague, J., Norton, L., Vahdat, L. T., Moynahan, M. E., and Chan, T. A. (2011). Breast Cancer Methylomes Establish an Epigenomic Foundation for Metastasis. *Science Translational Medicine*, 3(75):75ra25–75ra25.

Farlik, M., Sheffield, N. C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., and Bock, C. (2015). Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics. *Cell Reports*, 10(8):1386–1397.

Fearon, E. R. and Volgestein, B. (1990). A Genetic Model for Colorectal tumorogenesis. *Cell*, 61(5):759–767.

Federico, A., Rienzo, M., Abbondanza, C., Costa, V., Ciccodicola, A., and Casamassimi, A. (2017). Pan-cancer mutational and transcriptional analysis of the integrator complex. *International Journal of Molecular Sciences*, 18(5).

Feinberg, A. P., Koldobskiy, M. A., and Göndör, A. (2016). Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nature Reviews Genetics*, 17(5):284–299.

Feinberg, A. P., Ohlsson, R., and Henikoff, S. (2006). The epigenetic progenitor origin of human cancer. *Nature Reviews Genetics*, 7(1):21–33.

Feinberg, A. P. and Vogelstein, B. (1983a). Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, 301(5895):89–92.

Feinberg, A. P. and Vogelstein, B. (1983b). Hypomethylation of ras oncogenes in primary human cancers. *Biochemical and Biophysical Research Communications*, 111(1):47–54.

Feldmann, A., Ivanek, R., Murr, R., Gaidatzis, D., Burger, L., and Sch??beler, D. (2013). Transcription Factor Occupancy Can Mediate Active Turnover of DNA Methylation at Regulatory Regions. *PLoS Genetics*, 9(12).

Fenaux, P., Mufti, G. J., Hellstrom-Lindberg, E., Santini, V., Finelli, C., Giagounidis, A., Schoch, R., Gattermann, N., Sanz, G., List, A., Gore, S. D., Seymour, J. F., Bennett, J. M., Byrd, J., Backstrom, J., Zimmerman, L., McKenzie, D., Beach, C. L., and Silverman, L. R. (2009). Efficacy of azacitidine compared with that of conventional care regimens in the treatment of higher-risk myelodysplastic syndromes: a randomised, open-label, phase III study. *The Lancet Oncology*, 10(3):223–232.

Figueroa, M. E., Lugthart, S., Li, Y., Erpelinck-Verschueren, C., Deng, X., Christos, P. J., Schifano, E., Booth, J., van Putten, W., Skrabanek, L., Campagne, F., Mazumdar, M., Greally, J. M., Valk, P. J., Löwenberg, B., Delwel, R., and Melnick, A. (2010). DNA Methylation Signatures Identify Biologically Distinct Subtypes in Acute Myeloid Leukemia. *Cancer Cell*, 17(1):13–27.

Fisher, B., Costantino, J. P., Wickerham, D. L., Redmond, C. K., Kavanah, M., Cronin, W. M., Vogel, V., Robidoux, A., Dimitrov, N., Atkins, J., Daly, M., Wieand, S., Tan-Chiu, E., Ford, L., and Wolmark, N. (1998). Tamoxifen for Prevention of Breast Cancer: Report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study and other National Surgical Adjuvant Breast and Bowel Project Investigators. *Journal of the National Cancer Institute*, 90(18):1371–1388.

Flanagan, J. M., Cocciardi, S., Waddell, N., Johnstone, C. N., Marsh, A., Henderson, S., Simpson, P., da Silva, L., Khanna, K., Lakhani, S., Boshoff, C., and Chenevix-Trench, G. (2010). DNA Methylome of Familial Breast Cancer Identifies Distinct Profiles Defined by Mutation Status. *American Journal of Human Genetics*, 86(3):420–433.

Fleischer, T., Edvardsen, H., Solvang, H. K., Daviaud, C., Naume, B., Børresen-Dale, A. L., Kristensen, V. N., and Tost, J. (2014). Integrated analysis of high-resolution DNA methylation profiles, gene expression, germline genotypes and clinical end points in breast cancer patients. *International Journal of Cancer*, 134(11):2615–2625.

Fleischer, T., Tekpli, X., Mathelier, A., Wang, S., Nebdal, D., Dhakal, H. P., Sahlberg, K. K., Schlichting, E., Børresen-Dale, A.-L., Borgen, E., Naume, B., Eskeland, R., Frigessi, A., Tost, J., Hurtado, A., and Kristensen, V. N. (2017). DNA methylation at enhancers identifies distinct breast cancer lineages. *Nature Communications*, 8(1):1379.

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., Kok, C. Y., Jia, M., De, T., Teague, J. W., Stratton, M. R., McDermott, U., and Campbell, P. J. (2015). COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, 43(D1):D805–D811.

Foulds, L. (1958). The natural history of cancer. *Journal of Chronic Diseases*, 8(1):2–37.

Foulkes, W. D., Smith, I. E., and Reis-Filho, J. S. (2010). Triple-Negative Breast Cancer. *New England Journal of Medicine*, 363(20):1938–1948.

Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ballestar, M. L., Heine-Suner, D., Cigudosa, J. C., Urioste, M., Benitez, J., Boix-Chornet, M., Sanchez-Aguilera, A., Ling, C., Carlsson, E., Poulsen, P., Vaag, A., Stephan, Z., Spector, T. D., Wu, Y.-Z., Plass, C., and Esteller, M. (2005). From The Cover: Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences*, 102(30):10604–10609.

Friedman, L. S., Ostermeyer, E. A., Szabo, C. I., Dowd, P., Lynch, E. D., Rowell, S. E., and King, M.-C. (1994). Confirmation of BRCA1 by analysis of germline mutations linked to breast and ovarian cancer in ten families. *Nature Genetics*, 8(4):399–404.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., and Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences*, 89(5):1827–1831.

Fry, A. M. (2002). The Nek2 protein kinase: a novel regulator of centrosome structure. *Oncogene*, 21(40):6184–6194.

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., Chew, E. G. Y., Huang, P. Y. H., Welboren, W.-J., Han, Y., Ooi, H. S., Ariyaratne, P. N., Vega, V. B., Luo, Y., Tan, P. Y., Choy, P. Y., Wansa, K. D. S. A., Zhao, B., Lim, K. S., Leow, S. C., Yow, J. S., Joseph, R., Li, H., Desai, K. V., Thomsen, J. S., Lee, Y. K., Karuturi, R. K. M., Herve, T., Bourque, G., Stunnenberg, H. G., Ruan, X., Cacheux-Rataboul, V., Sung, W.-K., Liu, E. T., Wei, C.-L., Cheung, E., and Ruan, Y. (2009). An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, 462(7269):58–64.

Galea, M. H., Blamey, R. W., Elston, C. E., and Ellis, I. O. (1992). The Nottingham prognostic index in primary breast cancer. *Breast Cancer Research and Treatment*, 22(3):207–219.

Gao, Y., Jones, A., Fasching, P. A., Ruebner, M., Beckmann, M. W., Widschwendter, M., and Teschendorff, A. E. (2015). The integrative epigenomic-transcriptomic landscape of ER positive breast cancer. *Clinical Epigenetics*, 7(1):126.

Gao, Y. and Teschendorff, A. E. (2017). Epigenetic and genetic deregulation in cancer target distinct signaling pathway domains. *Nucleic Acids Research*, 45(2):583–596.

Garrett-Bakelman, F. E., Sheridan, C. K., Kacmarczyk, T. J., Ishii, J., Betel, D., Alonso, A., Mason, C. E., Figueroa, M. E., and Melnick, A. M. (2015). Enhanced Reduced Representation Bisulfite Sequencing for Assessment of DNA Methylation at Base Pair Resolution. *Journal of Visualized Experiments*, (96):52246.

Gascard, P., Bilenky, M., Sigaroudinia, M., Zhao, J., Li, L., Carles, A., Delaney, A., Tam, A., Kamoh, B., Cho, S., Griffith, M., Chu, A., Robertson, G., Cheung, D., Li, I., Heravi-Moussavi, A., Moksa, M., Mingay, M., Hussainkhel, A., Davis, B., Nagarajan, R. P., Hong, C., Echipare, L., O'Geen, H., Hangauer, M. J., Cheng, J. B., Neel, D., Hu, D., McManus, M. T., Moore, R., Mungall, A., Ma, Y., Plettner, P., Ziv, E., Wang, T., Farnham, P. J., Jones, S. J., Marra, M. A., Tlsty, T. D., Costello, J. F., and Hirst, M. (2015). Epigenetic and transcriptional determinants of the human breast. *Nature Communications*, 6:6351.

Gasco, M., Shami, S., and Crook, T. (2002). The p53 pathway in breast cancer. *Breast Cancer Research*, 4(2):70.

Genomics Core Cancer Research UK Cambridge Institute (2015). Communication. Technical report, Cancer Research UK Cambridge Institute.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N. Q., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C. R., Nohadani, M., Eklund, A. C., Spencer-Dene, B., Clark, G., Pickering, L.,

Stamp, G., Gore, M., Szallasi, Z., Downward, J., Futreal, P. A., and Swanton, C. (2012). Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine*, 366(10):883–892.

Glass, J. L., Hassane, D., Wouters, B. J., Kunimoto, H., Avellino, R., Garrett-Bakelman, F. E., Guryanova, O. A., Bowman, R., Redlich, S., Intlekofer, A. M., Meydan, C., Qin, T., Fall, M., Alonso, A., Guzman, M. L., Valk, P. J., Thompson, C. B., Levine, R., Elemento, O., Delwel, R., Melnick, A., and Figueroa, M. E. (2017). Epigenetic identity in AML depends on disruption of nonpromoter regulatory elements and is affected by antagonistic effects of mutations in epigenetic modifiers. *Cancer Discovery*, 7(8):868–883.

Gopalakrishnan, S., Van Emburgh, B. O., and Robertson, K. D. (2008). DNA methylation in development and human disease. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 647(1-2):30–38.

Greaves, M. and Maley, C. C. (2012). Clonal evolution in cancer. *Nature*, 481(7381):306–313.

Gu, H., Bock, C., Mikkelsen, T. S., Jäger, N., Smith, Z. D., Tomazou, E., Gnirke, A., Lander, E. S., and Meissner, A. (2010). Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nature Methods*, 7(2):133–136.

Gu, H., Smith, Z. D., Bock, C., Boyle, P., Gnirke, A., and Meissner, A. (2011). Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nature Protocols*, 6(4):468–481.

Guo, H., Zhu, P., Wu, X., Li, X., Wen, L., and Tang, F. (2013). Single-Cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Research*, 23(12):2126–2135.

Györffy, B., Bottai, G., Fleischer, T., Munkácsy, G., Budczies, J., Paladini, L., Børresen-Dale, A. L., Kristensen, V. N., and Santarpia, L. (2016). Aberrant DNA methylation impacts gene expression and prognosis in breast cancer subtypes. *International Journal of Cancer*, 138(1):87–97.

Gyparaki, M. T., Basdra, E. K., and Papavassiliou, A. G. (2013). DNA methylation biomarkers as diagnostic and prognostic tools in colorectal cancer. *Journal of Molecular Medicine*, 91(11):1249–1256.

Habib, J. G. and O'Shaughnessy, J. A. (2016). The hedgehog pathway in triple-negative breast cancer. *Cancer Medicine*, 5(10):2989–3006.

Hackett, J. A., Sengupta, R., Zylicz, J. J., Murakami, K., Lee, C., Down, T. A., and Surani, M. A. (2013). Germline DNA Demethylation Dynamics and Imprint Erasure Through 5-Hydroxymethylcytosine. *Science*, 339(6118):448–452.

Haffner, M. C., Chaux, A., Meeker, A. K., Esopi, D., Gerber, J., Pellakuru, L. G., Toubaji, A., Argani, P., Iacobuzio-Donahue, C., Nelson, W. G., Netto, G. J., DeMarzo, A., and Yegnasubramanian, S. (2011). Global 5-hydroxymethylcytosine content is significantly reduced in tissue stem/progenitor cell compartments and in human cancers. *Oncotarget*, 2(8):627–637.

Hall, J., Lee, M., Newman, B., Morrow, J., Anderson, L., Huey, B., and King, M. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250(4988):1684–1689.

Hammerman, P. S., Lawrence, M. S., Voet, D., Jing, R., Cibulskis, K., Sivachenko, A., Stojanov, P., McKenna, A., Lander, E. S., Gabriel, S., Getz, G., Sougnez, C., Imielinski, M., Helman, E., Hernandez, B., Pho, N. H., Meyerson, M., Chu, A., Chun, H.-J. E., Mungall, A. J., Pleasance, E., Gordon Robertson, A., Sipahimalani, P., Stoll, D., Balasundaram, M., Birol, I., Butterfield, Y. S. N., Chuah, E., Coope, R. J. N., Corbett, R., Dhalla, N., Guin, R., He, A., Hirst, C., Hirst, M., Holt, R. A., Lee, D., Li, H. I., Mayo, M., Moore, R. A., Mungall, K., Ming Nip, K., Olshen, A., Schein, J. E., Slobodan, J. R., Tam, A., Thiessen, N., Varhol, R., Zeng, T., Zhao, Y., Jones, S. J. M., Marra, M. A., Saksena, G., Cherniack, A. D., Schumacher, S. E., Tabak, B., Carter, S. L., Pho, N. H., Nguyen, H., Onofrio, R. C., Crenshaw, A., Ardlie, K., Beroukhim, R., Winckler, W., Hammerman, P. S., Getz, G., Meyerson, M., Protopopov, A., Zhang, J., Hadjipanayis, A., Lee, S., Xi, R., Yang, L., Ren, X., Zhang, H., Shukla, S., Chen, P.-C., Haseley, P., Lee, E., Chin, L., Park, P. J., Kucherlapati, R., Socci, N. D., Liang, Y., Schultz, N., Borsu, L., Lash, A. E., Viale, A., Sander, C., Ladanyi, M., Todd Auman, J., Hoadley, K. A., Wilkerson, M. D., Shi, Y., Liquori, C., Meng, S., Li, L., Turman, Y. J., Topal, M. D., Tan, D., Waring, S., Buda, E., Walsh, J., Jones, C. D., Mieczkowski, P. A., Singh, D., Wu, J., Gulabani, A., Dolina, P., Bodenheimer, T., Hoyle, A. P., Simons, J. V., Soloway, M. G., Mose, L. E., Jefferys, S. R., Balu, S., O'Connor, B. D., Prins, J. F., Liu, J., Chiang, D. Y., Neil Hayes, D., Perou, C. M., Cope, L., Danilova, L., Weisenberger, D. J., Maglinte, D. T., Pan, F., Van Den Berg, D. J., Triche Jr, T., Herman, J. G., Baylin, S. B., Laird, P. W., Getz, G., Noble, M., Voet, D., Saksena, G., Gehlenborg, N., DiCara, D., Zhang, J., Zhang, H., Wu, C.-J., Yingchun Liu, S., Lawrence, M. S., Zou, L., Sivachenko, A., Lin, P., Stojanov, P., Jing, R., Cho, J., Nazaire, M.-D., Robinson, J., Thorvaldsdottir, H., Mesirov, J., Park, P. J., Chin, L., Schultz, N., Sinha, R., Ciriello, G., Cerami, E., Gross, B., Jacobsen, A., Gao, J., Arman Aksoy, B., Weinhold, N., Ramirez, R., Taylor, B. S., Antipin, Y., Reva, B., Shen, R., Mo, Q., Seshan, V., Paik, P. K., Ladanyi, M., Sander, C., Akbani, R., Zhang, N., Broom, B. M., Casasent, T., Unruh, A., Wakefield, C., Craig Cason, R., Baggerly, K. A., Weinstein, J. N., Haussler, D., Benz, C. C., Stuart, J. M., Zhu, J., Szeto, C., Scott, G. K., Yau, C., Ng, S., Goldstein, T., Waltman, P., Sokolov, A., Ellrott, K., Collisson, E. A., Zerbino, D., Wilks, C., Ma, S., Craft, B., Wilkerson, M. D., Todd Auman, J., Hoadley, K. A., Du, Y., Cabanski, C., Walter, V., Singh, D., Wu, J., Gulabani, A., Bodenheimer, T., Hoyle, A. P., Simons, J. V., Soloway, M. G., Mose, L. E., Jefferys, S. R., Balu, S., Marron, J. S., Liu, Y., Wang, K., Liu, J., Prins, J. F., Neil Hayes, D., Perou, C. M., Creighton, C. J., Zhang, Y., Travis, W. D., Rekhtman, N., Yi, J., Aubry, M. C., Cheney, R., Dacic, S., Flieder, D., Funkhouser, W., Illei, P., Myers, J., Tsao, M.-S., Penny, R., Mallery, D., Shelton, T., Hatfield, M., Morris, S., Yena, P., Shelton, C., Sherman, M., Paulauskis, J., Meyerson, M., Baylin, S. B., Govindan, R., Akbani, R., Azodo, I., Beer, D., Bose, R., Byers, L. A., Carbone, D., Chang, L.-W., Chiang, D., Chu, A., Chun, E., Collisson, E., Cope, L., Creighton, C. J., Danilova, L., Ding, L., Getz, G., Hammerman, P. S., Neil Hayes, D., Hernandez, B., Herman, J. G., Heymach, J., Ida, C., Imielinski, M., Johnson, B., Jurisica, I., Kaufman, J., Kosari, F., Kucherlapati, R., Kwiatkowski, D., Ladanyi, M., Lawrence, M. S., Maher, C. A., Mungall, A., Ng, S., Pao, W., Peifer, M., Penny, R., Robertson, G., Rusch, V., Sander, C., Schultz, N., Shen, R., Siegfried, J., Sinha, R., Sivachenko, A., Sougnez,

C., Stoll, D., Stuart, J., Thomas, R. K., Tomaszek, S., Tsao, M.-S., Travis, W. D., Vaske, C., Weinstein, J. N., Weisenberger, D., Wigle, D. A., Wilkerson, M. D., Wilks, C., Yang, P., John Zhang, J., Jensen, M. A., Sfeir, R., Kahn, A. B., Chu, A. L., Kothiyal, P., Wang, Z., Snyder, E. E., Pontius, J., Pihl, T. D., Ayala, B., Backus, M., Walton, J., Baboud, J., Berton, D. L., Nicholls, M. C., Srinivasan, D., Raman, R., Girshik, S., Kigonya, P. A., Alonso, S., Sanbhadti, R. N., Barletta, S. P., Greene, J. M., Pot, D. A., Tsao, M.-S., Bandarchi-Chamkhaleh, B., Boyd, J., Weaver, J., Wigle, D. A., Azodo, I. A., Tomaszek, S. C., Christine Aubry, M., Ida, C. M., Yang, P., Kosari, F., Brock, M. V., Rogers, K., Rutledge, M., Brown, T., Lee, B., Shin, J., Trusty, D., Dhir, R., Siegfried, J. M., Potapova, O., Fedosenko, K. V., Nemirovich-Danchenko, E., Rusch, V., Zakowski, M., Iacocca, M. V., Brown, J., Rabeno, B., Czerwinski, C., Petrelli, N., Fan, Z., Todaro, N., Eckman, J., Myers, J., Kimryn Rathmell, W., Thorne, L. B., Huang, M., Boice, L., Hill, A., Penny, R., Mallery, D., Curley, E., Shelton, C., Yena, P., Morrison, C., Gaudioso, C., Bartlett, J. M. S., Kodeeswaran, S., Zanke, B., Sekhon, H., David, K., Juhl, H., Van Le, X., Kohl, B., Thorp, R., Viet Tien, N., Van Bang, N., Sussman, H., Duc Phu, B., Hajek, R., Phi Hung, N., Khan, K. Z., Muley, T., Mills Shaw, K. R., Sheth, M., Yang, L., Buetow, K., Davidsen, T., Demchok, J. A., Eley, G., Ferguson, M., Dillon, L. A. L., Schaefer, C., Guyer, M. S., Ozenberger, B. A., Palchik, J. D., Peterson, J., Sofia, H. J., Thomson, E., Hammerman, P. S., Neil Hayes, D., Wilkerson, M. D., Schultz, N., Bose, R., Chu, A., Collisson, E. A., Cope, L., Creighton, C. J., Getz, G., Herman, J. G., Johnson, B. E., Kucherlapati, R., Ladanyi, M., Maher, C. A., Robertson, G., Sander, C., Shen, R., Sinha, R., Sivachenko, A., Thomas, R. K., Travis, W. D., Tsao, M.-S., Weinstein, J. N., Wigle, D. A., Baylin, S. B., Govindan, R., and Meyerson, M. (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674.

Hansen, K. D., Langmead, B., and Irizarry, R. A. (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13(10):R83.

Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., Wen, B., Wu, H., Liu, Y., Diep, D., Briem, E., Zhang, K., Irizarry, R. A., and Feinberg, A. P. (2011). Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, 43(8):768–775.

Hansen, R. S., Thomas, S., Sandstrom, R., Canfield, T. K., Thurman, R. E., Weaver, M., Dorschner, M. O., Gartler, S. M., and Stamatoyannopoulos, J. A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences*, 107(1):139–144.

Harvey, J. M., Clark, G. M., Osborne, C. K., and Allred, D. C. (1999). Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer. *Journal of Clinical Oncology*, 17(5):1474–1481.

Hebestreit, K., Dugas, M., and Klein, H. U. (2013). Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, 29(13):1647–1653.

Hendrich, B. and Bird, A. (1998). Identification and Characterization of a Family of Mammalian Methyl-CpG Binding Proteins. *Molecular and Cellular Biology*, 18(11):6538–6547.

Herman, J. G. and Baylin, S. B. (2003). Gene Silencing in Cancer in Association with Promoter Hypermethylation. *New England Journal of Medicine*, 349(21):2042–2054.

Herschkowitz, J. I., Simin, K., Weigman, V. J., Mikaelian, I., Usary, J., Hu, Z., Rasmussen, K. E., Jones, L. P., Assefnia, S., Chandrasekharan, S., Backlund, M. G., Yin, Y., Khramtsov, A. I., Bastein, R., Quackenbush, J., Glazer, R. I., Brown, P. H., Green, J. E., Kopelovich, L., Furth, P. A., Palazzo, J. P., Olopade, O. I., Bernard, P. S., Churchill, G. A., Van Dyke, T., and Perou, C. M. (2007). Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biology*, 8(5):R76.

Heyn, H., Vidal, E., Ferreira, H. J., Vizoso, M., Sayols, S., Gomez, A., Moran, S., Boque-Sastre, R., Guil, S., Martinez-Cardus, A., Lin, C. Y., Royo, R., Sanchez-Mut, J. V., Martinez, R., Gut, M., Torrents, D., Orozco, M., Gut, I., Young, R. A., and Esteller, M. (2016). Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. *Genome Biology*, 17(1):11.

Hinrichs, A. S. (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, 34(90001):D590–D598.

Holliday, R. & Pugh, J. E. (1975). DNA modification mechanisms and gene activity during development. *Science*, 187(4173):226–232.

Holm, K., Hegardt, C., Staaf, J., Vallon-Christersson, J., Jönsson, G., Olsson, H., Borg, Å., and Ringnér, M. (2010). Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns. *Breast Cancer Research*, 12(3):R36.

Holm, K., Staaf, J., Lauss, M., Aine, M., Lindgren, D., Bendahl, P.-O., Vallon-Christersson, J., Barkardottir, R. B., Höglund, M., Borg, Å., Jönsson, G., and Ringnér, M. (2016). An integrated genomics analysis of epigenetic subtypes in human breast tumors links DNA methylation patterns to chromatin states in normal mammary cells. *Breast Cancer Research*, 18(1):27.

Hon, G. C., Hawkins, R. D., Caballero, O. L., Lo, C., Lister, R., Pelizzola, M., Valsesia, A., Ye, Z., Kuan, S., Edsall, L. E., Camargo, A. A., Stevenson, B. J., Ecker, J. R., Bafna, V., Strausberg, R. L., Simpson, A. J., and Ren, B. (2012). Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Research*, 22(2):246–258.

Horvath, S., Zhang, Y., Langfelder, P., Kahn, R. S., Boks, M. P., van Eijk, K., van den Berg, L. H., and Ophoff, R. A. (2012). Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biology*, 13(10):R97.

Hotchkiss, D. (1948). The quantitative serparation of purines, pyrimidines, and nucleosides by paper chromatography. *Journal of Biological Chemistry*, 175(1):315–332.

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363.

Houseman, E. A., Molitor, J., and Marsit, C. J. (2014). Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics*, 30(10):1431–1439.

Hovestadt, V., Jones, D. T. W., Picelli, S., Wang, W., Kool, M., Northcott, P. A., Sultan, M., Stachurski, K., Ryzhova, M., Warnatz, H.-J., Ralser, M., Brun, S., Bunt, J., Jäger, N., Kleinheinz, K., Erkek, S., Weber, U. D., Bartholomae, C. C., von Kalle, C., Lawerenz, C., Eils, J., Koster, J., Versteeg, R., Milde, T., Witt, O., Schmidt, S., Wolf, S., Pietsch, T., Rutkowski, S., Scheurlen, W., Taylor, M. D., Brors, B., Felsberg, J., Reifenberger, G., Borkhardt, A., Lehrach, H., Wechsler-Reya, R. J., Eils, R., Yaspo, M.-L., Landgraf, P., Korshunov, A., Zapatka, M., Radlwimmer, B., Pfister, S. M., and Lichter, P. (2014). Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature*, 510(7506):537–541.

Hu, X., Stern, H. M., Ge, L., O'Brien, C., Haydu, L., Honchell, C. D., Haverty, P. M., Peters, B. A., Wu, T. D., Amler, L. C., Chant, J., Stokoe, D., Lackner, M. R., and Cavet, G. (2009). Genetic Alterations and Oncogenic Pathways Associated with Breast Cancer Subtypes. *Molecular Cancer Research*, 7(4):511–522.

Hurtado, A., Holmes, K. A., Ross-Innes, C. S., Schmidt, D., and Carroll, J. S. (2011). FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nature Genetics*, 43(1):27–33.

Hynes, N. E. and Stern, D. F. (1994). The biology of erbB-2/neu/HER-2 and its role in cancer. *Biochim Biophys Acta*, 1198(2-3):165–84.

Irizarry, R. A., Ladd-Acosta, C., Carvalho, B., Wu, H., Brandenburg, S. A., Jeddeloh, J. A., Wen, B., and Feinberg, A. P. (2008). Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Research*, 18(5):780–790.

Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., Ji, H., Potash, J. B., Sabunciyan, S., and Feinberg, A. P. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genetics*, 41(2):178–186.

Issa, J.-P. (2011). Epigenetic variation and cellular Darwinism. *Nature Genetics*, 43(8):724–726.

Issa, J. P. (2014). Aging and epigenetic drift: A vicious cycle. *Journal of Clinical Investigation*, 124(1):24–29.

Ito, S., D'Alessio, A. C., Taranova, O. V., Hong, K., Sowers, L. C., and Zhang, Y. (2010). Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*, 466(7310):1129–1133.

Iyer, L. M., Tahiliani, M., Rao, A., and Aravind, L. (2009). Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. *Cell Cycle*, 8(11):1698–1710.

Jaffe, A. E. and Irizarry, R. A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*, 15(2):R31.

Jensen, E. V., Block, G. E., Smith, S., Kyser, K., and DeSombre, E. R. (1971). Estrogen receptors and breast cancer response to adrenalectomy. *National Cancer Institute monograph*, 34:55–70.

Jiao, Y., Widschwendter, M., and Teschendorff, A. E. (2014). A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics*, 30(16):2360–2366.

Johnson, K. C., Houseman, E. A., King, J. E., and Christensen, B. C. (2017). Normal breast tissue DNA methylation differences at regulatory elements are associated with the cancer risk factor age. *Breast Cancer Research*, 19(1):81.

Johnson, K. C., Houseman, E. A., King, J. E., von Herrmann, K. M., Fadul, C. E., and Christensen, B. C. (2016). 5-Hydroxymethylcytosine localizes to enhancer elements and is associated with survival in glioblastoma patients. *Nature Communications*, 7:13177.

Jones, M. J., Goodman, S. J., and Kobor, M. S. (2015). DNA methylation and healthy human aging. *Aging Cell*, 14(6):924–932.

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492.

Jones, P. a. and Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. *Nature reviews. Genetics*, 3(6):415–28.

Jones, P. A. and Baylin, S. B. (2007). The Epigenomics of Cancer. *Cell*, 128(4):683–692.

Jones, P. A. and Takai, D. (2001). The role of DNA methylation in mammalian epigenetics. *Science (New York, N.Y.)*, 293(5532):1068–70.

Jones, P. A. and Taylor, S. M. (1980). Cellular differentiation, cytidine analogs and DNA methylation. *Cell*, 20(1):85–93.

Joosse, S. A. (2012). BRCA1 and BRCA2: a common pathway of genome protection but different breast cancer subtypes. *Nature Reviews Cancer*, 12(5):372–372.

Kaminskas, E. (2005). FDA Drug Approval Summary: Azacitidine (5-azacytidine, VidazaTM) for Injectable Suspension. *The Oncologist*, 10(3):176–182.

Karolchik, D., Barber, G. P., Casper, J., Clawson, H., Cline, M. S., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., Harte, R. A., Heitner, S., Hinrichs, A. S., Learned, K., Lee, B. T., Li, C. H., Raney, B. J., Rhead, B., Rosenbloom, K. R., Sloan, C. A., Speir, M. L., Zweig, A. S., Haussler, D., Kuhn, R. M., and Kent, W. J. (2014). The UCSC Genome Browser database: 2014 update. *Nucleic Acids Research*, 42(D1):70.

Kaufman, L. and Rousseeuw, P. J. (1987). Clustering by means of medoids. *Statistical Data Analysis Based on the L 1-Norm and Related Methods. First International Conference*, pages 405–416416.

Kawaoi, A., Matsumoto, H., Suzuki, K., and Moriyama, S. (1992). Histogenesis of diisopropanolnitrosamine (DIPN)-induced tumors of the rat thyroid gland. *Virchows Archiv B Cell Pathology Including Molecular Pathology*, 61(1):49–56.

Kim, J. Y., Tavare, S., and Shibata, D. (2005). Counting human somatic cell replications: Methylation mirrors endometrial stem cell divisions. *Proceedings of the National Academy of Sciences*, 102(49):17739–17744.

King CR, Kraus MH, A. S. (1985). Amplification of a novel v-erB-related gene in a human mammary carcinoma. *Science*, 229:974-97(4717):974–976.

Kinzler, K. W. and Vogelstein, B. (1997). Cancer-susceptibility genes. Gatekeepers and caretakers. *Nature*, 386(6627):761, 763.

Knight, W. A., Livingston, R. B., Gregory, E. J., and McGuire, W. L. (1977). Estrogen receptor as an independent prognostic factor for early recurrence in breast cancer. *Cancer Res*, 37(12):4669–4671.

Knudson, A. G. (1971). Mutation and Cancer: Statistical Study of Retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4):820–823.

Koh, K. P., Yabuuchi, A., Rao, S., Huang, Y., Cunniff, K., Nardone, J., Laiho, A., Tahiliani, M., Sommer, C. A., Mostoslavsky, G., Lahesmaa, R., Orkin, S. H., Rodig, S. J., Daley, G. Q., and Rao, A. (2011). Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. *Cell Stem Cell*, 8(2):200–213.

Kondo, Y., Shen, L., Cheng, A. S., Ahmed, S., Boumber, Y., Charo, C., Yamochi, T., Urano, T., Furukawa, K., Kwabi-Addo, B., Gold, D. L., Sekido, Y., Huang, T. H.-M., and Issa, J.-P. J. (2008). Gene silencing in cancer by histone H3 lysine 27 trimethylation independent of promoter DNA methylation. *Nature Genetics*, 40(6):741–750.

Konecny, G., Pauletti, G., Pegram, M., Untch, M., Aguilar, Z., Wilson, C., Rong, H.-m., Bauerfeind, I., Felber, M., Wang, H.-j., Beryt, M., Seshadri, R., Hepp, H., and Slamon, D. J. (2003). Quantitative Association Between HER-2 / neu and Steroid Hormone Receptors in Hormone Receptor-Positive Primary Breast Cancer. *Journal of the National Cancer Institute*, 95(2):142–153.

Kreso, A., O'Brien, C. A., van Galen, P., Gan, O. I., Notta, F., Brown, A. M. K., Ng, K., Ma, J., Wienholds, E., Dunant, C., Pollett, A., Gallinger, S., McPherson, J., Mullighan, C. G., Shibata, D., and Dick, J. E. (2013). Variable Clonal Repopulation Dynamics Influence Chemotherapy Response in Colorectal Cancer. *Science*, 339(6119):543–548.

Kretzmer, H., Bernhart, S. H., Wang, W., Haake, A., Weniger, M. A., Bergmann, A. K., Betts, M. J., Carrillo-de Santa-Pau, E., Doose, G., Gutwein, J., Richter, J., Hovestadt, V., Huang, B., Rico, D., Jühling, F., Kolarova, J., Lu, Q., Otto, C., Wagener, R.,

Arnolds, J., Burkhardt, B., Claviez, A., Drexler, H. G., Eberth, S., Eils, R., Flicek, P., Haas, S., Hummel, M., Karsch, D., Kerstens, H. H. D., Klapper, W., Kreuz, M., Lawerenz, C., Lenze, D., Loeffler, M., López, C., MacLeod, R. A. F., Martens, J. H. A., Kulis, M., Martín-Subero, J. I., Möller, P., Nagel, I., Picelli, S., Vater, I., Rohde, M., Rosenstiel, P., Rosolowski, M., Russell, R. B., Schilhabel, M., Schlesner, M., Stadler, P. F., Szczepanowski, M., Trümper, L., Stunnenberg, H. G., Küppers, R., Ammerpohl, O., Lichter, P., Siebert, R., Hoffmann, S., and Radlwimmer, B. (2015). DNA methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control. *Nature Genetics*, 47(11):1316–1325.

Kriaucionis, S. and Heintz, N. (2009). The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. *Science*, 324(5929):929–930.

Krig, S. R., Miller, J. K., Frietze, S., Beckett, L. A., Neve, R. M., Farnham, P. J., Yaswen, P. I., and Sweeney, C. A. (2010). ZNF217, a candidate breast cancer oncogene amplified at 20q13, regulates expression of the ErbB3 receptor tyrosine kinase in breast cancer cells. *Oncogene*, 29(40):5500–5510.

Kristensen, V. N., Vaske, C. J., Ursini-Siegel, J., Van Loo, P., Nordgard, S. H., Sachidanandam, R., Sorlie, T., Warnberg, F., Haakensen, V. D., Helland, A., Naume, B., Perou, C. M., Haussler, D., Troyanskaya, O. G., and Borresen-Dale, A.-L. (2012). Integrated molecular profiles of invasive breast tumors and ductal carcinoma in situ (DCIS) reveal differential vascular and interleukin signaling. *Proceedings of the National Academy of Sciences*, 109(8):2802–2807.

Krueger, F. and Andrews, S. R. (2011). Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11):1571–1572.

Krueger, F., Kreck, B., Franke, A., and Andrews, S. R. (2012). DNA methylome analysis using short bisulfite sequencing data. *Nature Methods*, 9(2):145–151.

Kuan, P. F. and Chiang, D. Y. (2012). Integrating Prior Knowledge in Multiple Testing under Dependence with Applications to Detecting Differential DNA Methylation. *Biometrics*, 68(3):774–783.

Kulis, M., Heath, S., Bibikova, M., Queirós, A. C., Navarro, A., Clot, G., Martínez-Trillos, A., Castellano, G., Brun-Heath, I., Pinyol, M., Barberán-Soler, S., Papasaikas, P., Jares, P., Beà, S., Rico, D., Ecker, S., Rubio, M., Royo, R., Ho, V., Klotzle, B., Hernández, L., Conde, L., López-Guerra, M., Colomer, D., Villamor, N., Aymerich, M., Rozman, M., Bayes, M., Gut, M., Gelpí, J. L., Orozco, M., Fan, J.-B., Quesada, V., Puente, X. S., Pisano, D. G., Valencia, A., López-Guillermo, A., Gut, I., López-Otín, C., Campo, E., and Martín-Subero, J. I. (2012). Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nature Genetics*, 44(11):1236–1242.

Kulis, M., Merkel, A., Heath, S., Queirós, A. C., Schuyler, R. P., Castellano, G., Beekman, R., Raineri, E., Esteve, A., Clot, G., Verdaguer-Dot, N., Duran-Ferrer, M., Russiñol, N., Vilarrasa-Blasi, R., Ecker, S., Pancaldi, V., Rico, D., Agueda, L., Blanc, J., Richardson, D., Clarke, L., Datta, A., Pascual, M., Agirre, X., Prosper, F.,

Alignani, D., Paiva, B., Caron, G., Fest, T., Muench, M. O., Fomin, M. E., Lee, S.-T., Wiemels, J. L., Valencia, A., Gut, M., Flicek, P., Stunnenberg, H. G., Siebert, R., Küppers, R., Gut, I. G., Campo, E., and Martín-Subero, J. I. (2015). Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nature Genetics*, 47(7):746–756.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A., Wang, X., ClaussnitzerYaping Liu, M., Coarfa, C., Alan Harris, R., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., David Hawkins, R., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Scott Hansen, R., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Abdennur, N., Adli, M., Akerman, M., Barrera, L., Antosiewicz-Bourget, J., Ballinger, T., Barnes, M. J., Bates, D., Bell, R. J. A., Bennett, D. A., Bianco, K., Bock, C., Boyle, P., Brinchmann, J., Caballero-Campo, P., Camahort, R., Carrasco-Alfonso, M. J., Charnecki, T., Chen, H., Chen, Z., Cheng, J. B., Cho, S., Chu, A., Chung, W.-Y., Cowan, C., Athena Deng, Q., Deshpande, V., Diegel, M., Ding, B., Durham, T., Echipare, L., Edsall, L., Flowers, D., Genbacev-Krtolica, O., Gifford, C., Gillespie, S., Giste, E., Glass, I. A., Gnirke, A., Gormley, M., Gu, H., Gu, J., Hafler, D. A., Hangauer, M. J., Hariharan, M., Hatan, M., Haugen, E., He, Y., Heimfeld, S., Herlofsen, S., Hou, Z., Humbert, R., Issner, R., Jackson, A. R., Jia, H., Jiang, P., Johnson, A. K., Kadlecek, T., Kamoh, B., Kapidzic, M., Kent, J., Kim, A., Kleinewietfeld, M., Klugman, S., Krishnan, J., Kuan, S., Kutyavin, T., Lee, A.-Y., Lee, K., Li, J., Li, N., Li, Y., Ligon, K. L., Lin, S., Lin, Y., Liu, J., Liu, Y., Luckey, C. J., Ma, Y. P., Maire, C., Marson, A., Mattick, J. S., Mayo, M., McMaster, M., Metsky, H., Mikkelsen, T., Miller, D., Miri, M., Mukame, E., Nagarajan, R. P., Neri, F., Nery, J., Nguyen, T., O'Geen, H., Paithankar, S., Papayannopoulou, T., Pelizzola, M., Plettner, P., Propson, N. E., Raghuraman, S., Raney, B. J., Raubitschek, A., Reynolds, A. P., Richards, H., Riehle, K., Rinaudo, P., Robinson, J. F., Rockweiler, N. B., Rosen, E., Rynes, E., Schein, J., Sears, R., Sejnowski, T., Shafer, A., Shen, L., Shoemaker, R., Sigaroudinia, M., Slukvin, I., Stehling-Sun, S., Stewart, R., Subramanian, S. L., Suknuntha, K., Swanson, S., Tian, S., Tilden, H., Tsai, L., Urich, M., Vaughn, I., Vierstra, J., Vong, S., Wagner, U., Wang, H., Wang, T., Wang, Y., Weiss, A., Whitton, H., Wildberg, A., Witt, H., Won, K.-J., Xie, M., Xing, X., Xu, I., Xuan, Z., Ye, Z., Yen, C.-a., Yu, P., Zhang, X., Zhang, X., Zhao, J., Zhou, Y., Zhu, J., Zhu, Y., Ziegler, S., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., Kellis, M., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard,

P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., and Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.

Kurzrock, R., Kantarjian, H. M., Druker, B. J., and Talpaz, M. (2003). Philadelphia Chromosome – Positive Leukemias : From basic mechanisms to molecular therapeutics. *Ann Intern Med*, 138(10):819–31.

Laird, P. W. (2010). Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, 11(3):191.

Lancichinetti, A., Fortunato, S., and Kertész, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11.

Landan, G., Cohen, N. M., Mukamel, Z., Bar, A., Molchadsky, A., Brosh, R., Horn-Saban, S., Zalcenstein, D. A., Goldfinger, N., Zundelevich, A., Gal-Yam, E. N., Rotter, V., and Tanay, A. (2012). Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nature Genetics*, 44(11):1207–1214.

Landau, D. A., Carter, S. L., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M. S., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L., Wan, Y., Zhang, W., Shukla, S. A., Vartanov, A., Fernandes, S. M., Saksena, G., Cibulskis, K., Tesar, B., Gabriel, S., Hacohen, N., Meyerson, M., Lander, E. S., Neuberg, D., Brown, J. R., Getz, G., and Wu, C. J. (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*, 152(4):714–726.

Landau, D. A., Clement, K., Ziller, M. J., Boyle, P., Fan, J., Gu, H., Stevenson, K., Sougnez, C., Wang, L., Li, S., Kotliar, D., Zhang, W., Ghandi, M., Garraway, L., Fernandes, S. M., Livak, K. J., Gabriel, S., Gnirke, A., Lander, E. S., Brown, J. R., Neuberg, D., Kharchenko, P. V., Hacohen, N., Getz, G., Meissner, A., and Wu, C. J. (2014). Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia. *Cancer Cell*, 26(6):813–825.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D.,

Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G. R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F. A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S.-P., Yeh, R.-F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Patrinos, A., and Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Langemeijer, S., Kuiper, R., Berends, M., Knops, R., Aslanyan, M., Massop, M., van Hoogen, P., van Kessel, A. G., Raymakers, R., Verburgh, E., Hagemeijer, A., Vandenberghe, P., de Witte, T., Van der Reijden, B., and Jansen, J. (2009). P052 Acquired mutations in TET2 are common in myelodysplastic syndromes. *Leukemia Research*, 33(SUPPL. 1):838–842.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–9.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky,

N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D.-A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. A., Mora, J., Lee, R. S., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W. M., Biegel, J. A., Stegmaier, K., Bass, A. J., Garraway, L. A., Meyerson, M., Golub, T. R., Gordenin, D. A., Sunyaev, S., Lander, E. S., and Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218.

Lee, Y. J. and Gorski, J. (1996). Estrogen-induced transcription of the progesterone receptor gene does not parallel estrogen receptor occupancy. *Proc Natl Acad Sci U S A*, 93(26):15180–15184.

Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. *Nature*, 366(6453):362–365.

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., Poh, H. M., Goh, Y., Lim, J., Zhang, J., Sim, H. S., Peh, S. Q., Mulawadi, F. H., Ong, C. T., Orlov, Y. L., Hong, S., Zhang, Z., Landt, S., Raha, D., Euskirchen, G., Wei, C. L., Ge, W., Wang, H., Davis, C., Fisher-Aylor, K. I., Mortazavi, A., Gerstein, M., Gingeras, T., Wold, B., Sun, Y., Fullwood, M. J., Cheung, E., Liu, E., Sung, W. K., Snyder, M., and Ruan, Y. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1-2):84–98.

Li, J., Yen, C., Liaw, D., Podsypanina, K., Bose, S., Wang, S. I., Puc, J., Miliaresis, C., Rodgers, L., McCombie, R., Bigner, S. H., Giovanella, B. C., Ittmann, M., Tycko, B., Hibshoosh, H., Wigler, M. H., and Parsons, R. (1997). PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science*, 275(5308):1943–7.

Li, L., Lee, K. M., Han, W., Choi, J. Y., Lee, J. Y., Kang, G. H., Park, S. K., Noh, D. Y., Yoo, K. Y., and Kang, D. (2010). Estrogen and progesterone receptor status affect genome-wide DNA methylation profile in breast cancer. *Human Molecular Genetics*, 19(21):4273–4277.

Li, M., Gao, F., Xia, Y., Tang, Y., Zhao, W., Jin, C., Luo, H., Wang, J., Li, Q., and Wang, Y. (2016a). Filtrating colorectal cancer associated genes by integrated analyses of global DNA methylation and hydroxymethylation in cancer and normal tissue. *Scientific Reports*, 6(1):31826.

Li, S., Garrett-Bakelman, F., Perl, A. E., Luger, S. M., Zhang, C., To, B. L., Lewis, I. D., Brown, A. L., D'Andrea, R. J., Ross, M. E., Levine, R., Carroll, M., Melnick, A., and Mason, C. E. (2014). Dynamic evolution of clonal epialleles revealed by methclone. *Genome Biology*, 15(9):472.

Li, S., Garrett-Bakelman, F. E., Akalin, A., Zumbo, P., Levine, R., To, B. L., Lewis, I. D., Brown, A. L., D'Andrea, R. J., Melnick, A., and Mason, C. E. (2013). An optimized algorithm for detecting and annotating regional differential methylation. *BMC Bioinformatics*, 14(Suppl 5):S10.

Li, S., Garrett-Bakelman, F. E., Chung, S. S., Sanders, M. A., Hricik, T., Rapaport, F., Patel, J., Dillon, R., Vijay, P., Brown, A. L., Perl, A. E., Cannon, J., Bullinger, L., Luger, S., Becker, M., Lewis, I. D., To, L. B., Delwel, R., Löwenberg, B., Döhner, H., Döhner, K., Guzman, M. L., Hassane, D. C., Roboz, G. J., Grimwade, D., Valk, P. J. M., D'Andrea, R. J., Carroll, M., Park, C. Y., Neuberg, D., Levine, R., Melnick, A. M., and Mason, C. E. (2016b). Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nature Medicine*, 22(7):792–799.

Lian, C. G., Xu, Y., Ceol, C., Wu, F., Larson, A., Dresser, K., Xu, W., Tan, L., Hu, Y., Zhan, Q., Lee, C. W., Hu, D., Lian, B. Q., Kleffel, S., Yang, Y., Neiswender, J., Khorasani, A. J., Fang, R., Lezcano, C., Duncan, L. M., Scolyer, R. A., Thompson, J. F., Kakavand, H., Houvras, Y., Zon, L. I., Mihm, M. C., Kaiser, U. B., Schatton, T., Woda, B. A., Murphy, G. F., and Shi, Y. G. (2012). Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of Melanoma. *Cell*, 150(6):1135–1146.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326(5950):289–293.

Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F., and Schübeler, D. (2011). Identification of genetic elements that autonomously determine DNA methylation states. *Nature Genetics*, 43(11):1091–1097.

Lin, P.-C., Giannopoulou, E. G., Park, K., Mosquera, J. M., Sboner, A., Tewari, A. K., Garraway, L. A., Beltran, H., Rubin, M. A., and Elemento, O. (2013). Epigenomic Alterations in Localized and Advanced Prostate Cancer. *Neoplasia*, 15(4):373–IN5.

Lipson, D. and Liebert, M. A. (2006). DNA Copy Number Data Analysis. *R package version 1.50.1.*, 13(2):215–228.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., and Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322.

Liu, X. S., Wu, H., Ji, X., Stelzer, Y., Wu, X., Czauderna, S., Shu, J., Dadon, D., Young, R. A., and Jaenisch, R. (2016). Editing DNA Methylation in the Mammalian Genome. *Cell*, 167(1):233–247.e17.

Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., Mungall, C. J., Arner, E., Baillie, J., Bertin, N., Bono, H., de Hoon, M., Diehl, A. D., Dimont, E., Freeman, T. C., Fujieda, K., Hide, W., Kaliyaperumal, R., Katayama, T., Lassmann, T., Meehan, T. F., Nishikata, K., Ono, H., Rehli, M., Sandelin, A., Schultes, E. A., 't Hoen, P., Tatum, Z., Thompson, M., Toyoda, T., Wright, D. W., Daub, C. O., Itoh, M., Carninci, P., Hayashizaki, Y., Forrest, A., and Kawaji, H. (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biology*, 16(1):22.

Lock, L. F., Melton, D. W., Caskey, C. T., and Martin, G. R. (1986). Methylation of the mouse hprt gene differs on the active and inactive X chromosomes. *Molecular and cellular biology*, 6(3):914–24.

Lock, L. F., Takagi, N., and Martin, G. R. (1987). Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. *Cell*, 48(1):39–46.

Lu, C., Venneti, S., Akalin, A., Fang, F., Ward, P. S., DeMatteo, R. G., Intlekofer, A. M., Chen, C., Ye, J., Hameed, M., Nafa, K., Agaram, N. P., Cross, J. R., Khanin, R., Mason, C. E., Healey, J. H., Lowe, S. W., Schwartz, G. K., Melnick, A., and Thompson, C. B. (2013). Induction of sarcomas by mutant IDH2. *Genes and Development*, 27(18):1986–1998.

Maaten, L. V. D. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research 1*, 620(1):267–84.

Madzo, J., Liu, H., Rodriguez, A., Vasanthakumar, A., Sundaravel, S., Caces, D. B. D., Looney, T. J., Zhang, L., Lepore, J. B., Macrae, T., Duszynski, R., Shih, A. H., Song, C. X., Yu, M., Yu, Y., Grossman, R., Raumann, B., Verma, A., He, C., Levine, R. L., Lavelle, D., Lahn, B. T., Wickrema, A., and Godley, L. A. (2014). Hydroxymethylation at gene regulatory regions directs stem/early progenitor cell commitment during erythropoiesis. *Cell Reports*, 6(1):231–244.

Mahajan, K., Fang, B., Koomen, J. M., and Mahajan, N. P. (2012). H2B Tyr37 phosphorylation suppresses expression of replication-dependent core histone genes. *Nature Structural & Molecular Biology*, 19(9):930–937.

Maley, C. C., Galipeau, P. C., Finley, J. C., Wongsurawat, V. J., Li, X., Sanchez, C. A., Paulson, T. G., Blount, P. L., Risques, R.-A., Rabinovitch, P. S., and Reid, B. J. (2006). Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature Genetics*, 38(4):468–473.

Marumoto, T., Zhang, D., and Saya, H. (2005). Aurora-A — A guardian of poles. *Nature Reviews Cancer*, 5(1):42–50.

Massagué, J. (2004). G1 cell-cycle control and cancer. *Nature*, 432(7015):298–306.

Mazor, T., Pankov, A., Johnson, B. E., Hong, C., Hamilton, E. G., Bell, R. J. A., Smirnov, I. V., Reis, G. F., Phillips, J. J., Barnes, M. J., Idbaih, A., Alentorn, A., Kloezeman, J. J., Lamfers, M. L. M., Bollen, A. W., Taylor, B. S., Molinaro, A. M., Olshen, A. B., Chang, S. M., Song, J. S., and Costello, J. F. (2015). DNA Methylation and Somatic Mutations Converge on the Cell Cycle and Define Similar Evolutionary Histories in Brain Tumors. *Cancer Cell*, 28(3):307–317.

Mazor, T., Pankov, A., Song, J. S., and Costello, J. F. (2016). Intratumoral Heterogeneity of the Epigenome. *Cancer Cell*, 29(4):440–451.

McBryan, J., Howlin, J., Kenny, P. a., Shioda, T., and Martin, F. (2007). ERalpha-CITED1 co-regulated genes expressed during pubertal mammary gland development: implications for breast cancer prognosis. *Oncogene*, 26(44):6406–6419.

McInroy, G. R., Beraldi, D., Raiber, E. A., Modrzynska, K., Van Delft, P., Billker, O., and Balasubramanian, S. (2016). Enhanced methylation analysis by recovery of unsequenceable fragments. *PLoS ONE*, 11(3).

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303.

Meissner, A. (2010). Epigenetic modifications in pluripotent and differentiated cells. *Nature Biotechnology*, 28(10):1079–1088.

Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, 33(18):5868–5877.

Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C., Jaffe, D. B., Gnirke, A., Jaenisch, R., and Lander, E. S. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454 VN -(7205):766–770.

Melisko, M. (2005). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *Women's Oncology Review*, 5(1):45–47.

Meng, C., Helm, D., Frejno, M., and Kuster, B. (2016). MoCluster: Identifying Joint Patterns Across Multiple Omics Data Sets. *Journal of Proteome Research*, 15(3):755–765.

Merajver, S. D., Frank, T. S., Xu, J., Pham, T. M., Calzone, K. A., Bennett-Baker, P., Chamberlain, J., Boyd, J., Garber, J. E., Collins, F. S., and Weber, B. L. (1995). Germline BRCA1 Mutations and Loss of the Wild-Type Allele in Tumors from Families with Early Onset Breast and Ovarian Cancer. *Clinical Cancer Research*, 1(5):539–544.

Merlo, L. M. F., Shah, N. A., Li, X., Blount, P. L., Vaughan, T. L., Reid, B. J., and Maley, C. C. (2010). A comprehensive survey of clonal diversity measures in Barrett's esophagus as biomarkers of progression to esophageal adenocarcinoma. *Cancer Prevention Research*, 3(11):1388–1397.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L., Ding, W., and al. Et (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*, 266(5182):66–71.

Miura, F., Enomoto, Y., Dairiki, R., and Ito, T. (2012). Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Research*, 40(17).

Moelans, C. B., de Groot, J. S., Pan, X., van der Wall, E., and van Diest, P. J. (2014). Clonal intratumor heterogeneity of promoter hypermethylation in breast cancer by MS-MLPA. *Modern Pathology*, 27(6):869–874.

Mohammed, H., Russell, I. A., Stark, R., Rueda, O. M., Hickey, T. E., Tarulli, G. A., Serandour, A. A. A., Birrell, S. N., Bruna, A., Saadi, A., Menon, S., Hadfield, J., Pugh, M., Raj, G. V., Brown, G. D., D'Santos, C., Robinson, J. L. L., Silva, G., Launchbury, R., Perou, C. M., Stingl, J., Caldas, C., Tilley, W. D., and Carroll, J. S. (2015). Progesterone receptor modulates ERα action in breast cancer. *Nature*, 523(7560):313–317.

Moynahan, M. E. (2002). The cancer connection: BRCA1 and BRCA2 tumor suppression in mice and humans. *Oncogene*, 21(58):8994–9007.

Moynahan, M. E., Chiu, J. W., Koller, B. H., and Jasint, M. (1999). Brca1 controls homology-directed DNA repair. *Molecular Cell*, 4(4):511–518.

Mroz, E. A. and Rocco, J. W. (2013). MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. *Oral Oncology*, 49(3):211–215.

Murtaza, M., Dawson, S.-J., Tsui, D. W. Y., Gale, D., Forshew, T., Piskorz, A. M., Parkinson, C., Chin, S.-F., Kingsbury, Z., Wong, A. S. C., Marass, F., Humphray, S., Hadfield, J., Bentley, D., Chin, T. M., Brenton, J. D., Caldas, C., and Rosenfeld, N. (2013). Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature*, 497(7447):108–112.

Musialik, E., Bujko, M., Kober, P., Grygorowicz, M. A., Libura, M., Przestrzelska, M., Juszczyński, P., Borg, K., Florek, I., Jakóbczyk, M., and Siedlecki, J. A. (2015). Promoter DNA methylation and expression levels of HOXA4, HOXA5 and MEIS1 in acute myeloid leukemia. *Molecular Medicine Reports*, 11(5):3948–3954.

Muzny, D. M., Bainbridge, M. N., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., Kovar, C. L., Lewis, L. R., Morgan, M. B., Newsham, I. F., Reid, J. G., Santibanez, J., Shinbrot, E., Trevino, L. R., Wu, Y.-Q., Wang, M., Gunaratne, P., Donehower, L. A., Creighton, C. J., Wheeler, D. A., Gibbs, R. A., Lawrence, M. S., Voet, D., Jing, R., Cibulskis, K., Sivachenko, A., Stojanov, P., McKenna, A., Lander, E. S., Gabriel, S., Getz, G., Ding, L., Fulton, R. S., Koboldt, D. C., Wylie, T., Walker, J., Dooling, D. J., Fulton, L., Delehaunty, K. D., Fronick, C. C., Demeter, R., Mardis, E. R., Wilson, R. K., Chu, A., Chun, H.-J. E., Mungall, A. J., Pleasance, E., Gordon Robertson, A., Stoll, D., Balasundaram, M., Birol, I., Butterfield, Y. S. N., Chuah, E., Coope, R. J. N., Dhalla, N., Guin, R., Hirst, C., Hirst, M., Holt, R. A., Lee, D., Li, H. I., Mayo, M., Moore, R. A., Schein, J. E., Slobodan, J. R., Tam, A., Thiessen, N., Varhol, R., Zeng, T., Zhao, Y., Jones, S. J. M., Marra, M. A., Bass, A. J., Ramos, A. H., Saksena, G., Cherniack, A. D., Schumacher, S. E., Tabak, B., Carter, S. L., Pho, N. H., Nguyen, H., Onofrio, R. C., Crenshaw, A., Ardlie, K., Beroukhim, R., Winckler, W., Getz, G., Meyerson, M., Protopopov, A., Zhang, J., Hadjipanayis, A., Lee, E., Xi, R., Yang, L., Ren, X., Zhang, H., Sathiamoorthy, N., Shukla, S., Chen, P.-C., Haseley, P., Xiao, Y., Lee, S., Seidman, J., Chin, L., Park, P. J., Kucherlapati, R., Todd Auman, J., Hoadley, K. A., Du, Y., Wilkerson, M. D., Shi, Y., Liquori, C., Meng, S., Li, L., Turman, Y. J., Topal, M. D., Tan, D., Waring, S., Buda, E., Walsh, J., Jones, C. D., Mieczkowski, P. A., Singh, D., Wu, J., Gulabani, A., Dolina, P., Bodenheimer, T., Hoyle, A. P., Simons, J. V., Soloway, M., Mose, L. E., Jefferys, S. R., Balu, S., O'Connor, B. D., Prins, J. F., Chiang, D. Y., Neil Hayes, D., Perou, C. M., Hinoue, T., Weisenberger, D. J., Maglinte, D. T.,

Pan, F., Berman, B. P., Van Den Berg, D. J., Shen, H., Triche Jr, T., Baylin, S. B., Laird, P. W., Getz, G., Noble, M., Voet, D., Saksena, G., Gehlenborg, N., DiCara, D., Zhang, J., Zhang, H., Wu, C.-J., Yingchun Liu, S., Shukla, S., Lawrence, M. S., Zhou, L., Sivachenko, A., Lin, P., Stojanov, P., Jing, R., Park, R. W., Nazaire, M.-D., Robinson, J., Thorvaldsdottir, H., Mesirov, J., Park, P. J., Chin, L., Thorsson, V., Reynolds, S. M., Bernard, B., Kreisberg, R., Lin, J., Iype, L., Bressler, R., Erkkilä, T., Gundapuneni, M., Liu, Y., Norberg, A., Robinson, T., Yang, D., Zhang, W., Shmulevich, I., de Ronde, J. J., Schultz, N., Cerami, E., Ciriello, G., Goldberg, A. P., Gross, B., Jacobsen, A., Gao, J., Kaczkowski, B., Sinha, R., Arman Aksoy, B., Antipin, Y., Reva, B., Shen, R., Taylor, B. S., Chan, T. A., Ladanyi, M., Sander, C., Akbani, R., Zhang, N., Broom, B. M., Casasent, T., Unruh, A., Wakefield, C., Hamilton, S. R., Craig Cason, R., Baggerly, K. A., Weinstein, J. N., Haussler, D., Benz, C. C., Stuart, J. M., Benz, S. C., Zachary Sanborn, J., Vaske, C. J., Zhu, J., Szeto, C., Scott, G. K., Yau, C., Ng, S., Goldstein, T., Ellrott, K., Collisson, E., Cozen, A. E., Zerbino, D., Wilks, C., Craft, B., Spellman, P., Penny, R., Shelton, T., Hatfield, M., Morris, S., Yena, P., Shelton, C., Sherman, M., Paulauskis, J., Gastier-Foster, J. M., Bowen, J., Ramirez, N. C., Black, A., Pyatt, R., Wise, L., White, P., Bertagnolli, M., Brown, J., Chan, T. A., Chu, G. C., Czerwinski, C., Denstman, F., Dhir, R., Dörner, A., Fuchs, C. S., Guillem, J. G., Iacocca, M., Juhl, H., Kaufman, A., Kohl III, B., Van Le, X., Mariano, M. C., Medina, E. N., Meyers, M., Nash, G. M., Paty, P. B., Petrelli, N., Rabeno, B., Richards, W. G., Solit, D., Swanson, P., Temple, L., Tepper, J. E., Thorp, R., Vakiani, E., Weiser, M. R., Willis, J. E., Witkin, G., Zeng, Z., Zinner, M. J., Zornig, C., Jensen, M. A., Sfeir, R., Kahn, A. B., Chu, A. L., Kothiyal, P., Wang, Z., Snyder, E. E., Pontius, J., Pihl, T. D., Ayala, B., Backus, M., Walton, J., Whitmore, J., Baboud, J., Berton, D. L., Nicholls, M. C., Srinivasan, D., Raman, R., Girshik, S., Kigonya, P. A., Alonso, S., Sanbhadti, R. N., Barletta, S. P., Greene, J. M., Pot, D. A., Mills Shaw, K. R., Dillon, L. A. L., Buetow, K., Davidsen, T., Demchok, J. A., Eley, G., Ferguson, M., Fielding, P., Schaefer, C., Sheth, M., Yang, L., Guyer, M. S., Ozenberger, B. A., Palchik, J. D., Peterson, J., Sofia, H. J., and Thomson., E. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337.

Nair, S. S., Coolen, M. W., Stirzaker, C., Song, J. Z., Statham, A. L., Strbenac, D., Robinson, M. D., and Clark, S. J. (2011). Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpG sequence coverage bias. *Epigenetics*, 6(1):34–44.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W. R., Hicks, J., and Wigler, M. (2011). Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94.

Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., Maldotti, M., Anselmi, F., and Oliviero, S. (2017). Intragenic DNA methylation prevents spurious transcription initiation. *Nature*, 543(7643):72–77.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L. B., Martin, S., Wedge, D. C., Van Loo, P., Ju, Y. S., Smid, M., Brinkman, A. B., Morganella, S., Aure, M. R., Lingjærde, O. C., Langerød, A., Ringnér, M., Ahn, S.-M., Boyault, S., Brock, J. E., Broeks, A., Butler,

A., Desmedt, C., Dirix, L., Dronov, S., Fatima, A., Foekens, J. A., Gerstung, M., Hooijer, G. K. J., Jang, S. J., Jones, D. R., Kim, H.-Y., King, T. A., Krishnamurthy, S., Lee, H. J., Lee, J.-Y., Li, Y., McLaren, S., Menzies, A., Mustonen, V., O'Meara, S., Pauporté, I., Pivot, X., Purdie, C. A., Raine, K., Ramakrishnan, K., Rodríguez-González, F. G., Romieu, G., Sieuwerts, A. M., Simpson, P. T., Shepherd, R., Stebbings, L., Stefansson, O. A., Teague, J., Tommasi, S., Treilleux, I., Van den Eynden, G. G., Vermeulen, P., Vincent-Salomon, A., Yates, L., Caldas, C., van't Veer, L., Tutt, A., Knappskog, S., Tan, B. K. T., Jonkers, J., Borg, Å., Ueno, N. T., Sotiriou, C., Viari, A., Futreal, P. A., Campbell, P. J., Span, P. N., Van Laere, S., Lakhani, S. R., Eyfjord, J. E., Thompson, A. M., Birney, E., Stunnenberg, H. G., van de Vijver, M. J., Martens, J. W. M., Børresen-Dale, A.-L., Richardson, A. L., Kong, G., Thomas, G., and Stratton, M. R. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54.

Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., Shlien, A., Cooke, S. L., Hinton, J., Menzies, A., Stebbings, L. A., Leroy, C., Jia, M., Rance, R., Mudie, L. J., Gamble, S. J., Stephens, P. J., McLaren, S., Tarpey, P. S., Papaemmanuil, E., Davies, H. R., Varela, I., McBride, D. J., Bignell, G. R., Leung, K., Butler, A. P., Teague, J. W., Martin, S., Jönsson, G., Mariani, O., Boyault, S., Miron, P., Fatima, A., Langerod, A., Aparicio, S. A., Tutt, A., Sieuwerts, A. M., Borg, Å., Thomas, G., Salomon, A. V., Richardson, A. L., Borresen-Dale, A. L., Futreal, P. A., Stratton, M. R., and Campbell, P. J. (2012). The life history of 21 breast cancers. *Cell*, 149(5):994–1007.

Nikolaidis, G., Raji, O. Y., Markopoulou, S., Gosney, J. R., Bryan, J., Warburton, C., Walshaw, M., Sheard, J., Field, J. K., and Liloglou, T. (2012). DNA methylation biomarkers offer improved diagnostic efficiency in lung cancer. *Cancer Research*, 72(22):5692–5701.

Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P., Pan, F., Pelloski, C. E., Sulman, E. P., Bhat, K. P., Verhaak, R. G. W., Hoadley, K. A., Hayes, D. N., Perou, C. M., Schmidt, H. K., Ding, L., Wilson, R. K., Van Den Berg, D., Shen, H., Bengtsson, H., Neuvial, P., Cope, L. M., Buckley, J., Herman, J. G., Baylin, S. B., Laird, P. W., and Aldape, K. (2010). Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma. *Cancer Cell*, 17(5):510–522.

Nowell, P. and Hungerford, D. (1960). A minute chromosome in human chronic granulocytic leukemia. *Science*, 132(3438):1488–1501.

Nowell, P. C. (2012). The Clonal Evolution of Tumor Cell Populations. *Science (New York, N.Y.)*, 194(4260):23–28.

Nurse, P. (2004). Wee beasties. *Nature*, 432(7017):557–557.

Ohm, J. E., McGarvey, K. M., Yu, X., Cheng, L., Schuebel, K. E., Cope, L., Mohammad, H. P., Chen, W., Daniel, V. C., Yu, W., Berman, D. M., Jenuwein, T., Pruitt, K., Sharkis, S. J., Watkins, D. N., Herman, J. G., and Baylin, S. B. (2007). A stem cell–like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nature Genetics*, 39(2):237–242.

Okano, M., Bell, D. W., Haber, D. A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257.

Oshimo, Y., Nakayama, H., Ito, R., Kitadai, Y., Yoshida, K., Chayama, K., and Yasui, W. (2003). Promoter methylation of cyclin D2 gene in gastric carcinoma. *International journal of oncology*, 23(6):1663–1670.

Pan, H., Jiang, Y., Boi, M., Tabbò, F., Redmond, D., Nie, K., Ladetto, M., Chiappella, A., Cerchietti, L., Shaknovich, R., Melnick, A. M., Inghirami, G. G., Tam, W., and Elemento, O. (2015). Epigenomic evolution in diffuse large B-cell lymphomas. *Nature Communications*, 6:6921.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., and Bernard, P. S. (2009). Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167.

Pedersen, B. S., Schwartz, D. A., Yang, I. V., and Kechris, K. J. (2012). Comb-p: Software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics*, 28(22):2986–2988.

Pellacani, D., Bilenky, M., Kannan, N., Heravi-Moussavi, A., Knapp, D. J., Gakkhar, S., Moksa, M., Carles, A., Moore, R., Mungall, A. J., Marra, M. A., Jones, S. J., Aparicio, S., Hirst, M., and Eaves, C. J. (2016). Analysis of Normal Human Mammary Epigenomes Reveals Cell-Specific Active Enhancer States and Associated Transcription Factor Networks. *Cell Reports*, 17(8):2060–2074.

Pereira, B., Chin, S.-F., Rueda, O. M., Vollan, H.-K. M., Provenzano, E., Bardwell, H. A., Pugh, M., Jones, L., Russell, R., Sammut, S.-J., Tsui, D. W. Y., Liu, B., Dawson, S.-J., Abraham, J., Northen, H., Peden, J. F., Mukherjee, A., Turashvili, G., Green, A. R., McKinney, S., Oloumi, A., Shah, S., Rosenfeld, N., Murphy, L., Bentley, D. R., Ellis, I. O., Purushotham, A., Pinder, S. E., Børresen-Dale, A.-L., Earl, H. M., Pharoah, P. D., Ross, M. T., Aparicio, S., and Caldas, C. (2016). The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nature Communications*, 7:11479.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. a., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. a., Fluge, O., Pergamenschikov, a., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, a. L., Brown, P. O., and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752.

Peto, R., Boreham, J., Clarke, M., Davies, C., and Beral, V. (2000). UK and USA breast cancer deaths down 25% in year 2000 at ages 20-69 years. *Lancet*, 355(9217):1822.

Pickersgill, H., Kalverda, B., de Wit, E., Talhout, W., Fornerod, M., and van Steensel, B. (2006). Characterization of the Drosophila melanogaster genome at the nuclear lamina. *Nature Genetics*, 38(9):1005–1014.

Pidsley, R., Y Wong, C. C., Volta, M., Lunnon, K., Mill, J., and Schalkwyk, L. C. (2013). A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*, 14(1):293.

Pidsley, R., Zotenko, E., Peters, T. J., Lawrence, M. G., Risbridger, G. P., Molloy, P., Van Djik, S., Muhlhausler, B., Stirzaker, C., and Clark, S. J. (2016). Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, 17(1):208.

Polak, P., Kim, J., Braunstein, L. Z., Karlic, R., Haradhavala, N. J., Tiao, G., Rosebrock, D., Livitz, D., Kübler, K., Mouw, K. W., Kamburov, A., Maruvka, Y. E., Leshchiner, I., Lander, E. S., Golub, T. R., Zick, A., Orthwein, A., Lawrence, M. S., Batra, R. N., Caldas, C., Haber, D. A., Laird, P. W., Shen, H., Ellisen, L. W., D'Andrea, A. D., Chanock, S. J., Foulkes, W. D., and Getz, G. (2017). A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nature Genetics*.

Polakis, P. (2000). Wnt signaling and cancer Wnt signaling and cancer. *Genes Dev.*, 14(650):1837–1851.

Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D. L., Wang, Y., Hansen, R. S., Canfield, T. K., Thurman, R. E., Cheng, Y., Gülsoy, G., Dennis, J. H., Snyder, M. P., Stamatoyannopoulos, J. A., Taylor, J., Hardison, R. C., Kahveci, T., Ren, B., and Gilbert, D. M. (2014). Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515(7527):402–405.

Portela, A. and Esteller, M. (2010). Epigenetic modifications and human disease. *Nature Biotechnology*, 28(10):1057–1068.

Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., He, X., and Perou, C. M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Research*, 12(5):R68.

Prest, S. J., May, F. E. B., and Westley, B. R. (2002). The estrogen-regulated protein, TFF1, stimulates migration of human breast cancer cells. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 16(6):592–594.

Rahmani, E., Zaitlen, N., Baran, Y., Eng, C., Hu, D., Galanter, J., Oh, S., Burchard, E. G., Eskin, E., Zou, J., and Halperin, E. (2016). Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nature Methods*, 13(5):443–445.

Raiber, E. A., Beraldi, D., Martínez Cuesta, S., R, M. G., Kingsbury, Z., Becq, J., James, T., Lopes, M., Allinson, K., Field, S., Humphray, S., Santarius, T., Watts, C., Bentley, D., and Balasubramanian, S. (2017). Base resolution maps reveal the importance of 5-hydroxymethylcytosine in a human glioblastoma. *NPJ Genom Med*, 2(1):6.

Rakha, E. A., El-Sayed, M. E., Green, A. R., Paish, E. C., Powe, D. G., Gee, J., Nicholson, R. I., Lee, A. H. S., Robertson, J. F. R., and Ellis, I. O. (2007). Biologic and clinical characteristics of breast cancer with single hormone receptor-positive phenotype. *Journal of Clinical Oncology*, 25(30):4772–4778.

Rakyan, V. K., Down, T. A., Maslau, S., Andrew, T., Yang, T. P., Beyan, H., Whittaker, P., McCann, O. T., Finer, S., Valdes, A. M., Leslie, R. D., Deloukas, P., and Spector, T. D. (2010). Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Research*, 20(4):434–439.

Rauch, T. and Pfeifer, G. P. (2005). Methylated-CpG island recovery assay: a new technique for the rapid detection of methylated-CpG islands in cancer. *Laboratory Investigation*, 85(9):1172–1180.

Reedijk, M. (2012). Notch signaling and breast cancer. *Advances in Experimental Medicine and Biology*, 727:241–257.

Riggs, A. D. (1975). X inactivation, differentiation, and DNA methylation. *Cytogenetic and Genome Research*, 14(1):9–25.

Riggs, B. L. and Hartmann, L. C. (2003). Selective Estrogen-Receptor Modulators — Mechanisms of Action and Application to Clinical Practice. *New England Journal of Medicine*, 348(7):618–629.

Robertson, K. D. (2005). DNA methylation and human disease. *Nature Reviews Genetics*, 6(8):597–610.

Robertson, K. D. and Wolffe, A. P. (2000). DNA methylation in health and disease. *Nature Reviews Genetics*, 1(1):11–19.

Robinson, J. L. L., MacArthur, S., Ross-Innes, C. S., Tilley, W. D., Neal, D. E., Mills, I. G., and Carroll, J. S. (2012). Androgen receptor driven transcription in molecular apocrine breast cancer is mediated by FoxA1. *The EMBO Journal*, 31(6):1617–1617.

Robinson, M. D., Kahraman, A., Law, C. W., Lindsay, H., Nowicka, M., Weber, L. M., and Zhou, X. (2014). Statistical methods for detecting differentially methylated loci and regions. *Frontiers in Genetics*, 5(SEP):324.

Robinson, M. D., Stirzaker, C., Statham, A. L., Coolen, M. W., Song, J. Z., Nair, S. S., Strbenac, D., Speed, T. P., and Clark, S. J. (2010). Evaluation of affinity-based genome-wide DNA methylation data: Effects of CpG density, amplification bias, and copy number variation. *Genome Research*, 20(12):1719–1729.

Rodríguez-Paredes, M. and Esteller, M. (2011). A combined epigenetic therapy equals the efficacy of conventional chemotherapy in refractory advanced non-small cell lung cancer. *Cancer Discovery*, 1(7):557–559.

Roessler, J., Ammerpohl, O., Gutwein, J., Steinemann, D., Schlegelberger, B., Weyer, V., Sariyar, M., Geffers, R., Arnold, N., Schmutzler, R., Bartram, C. R., Heinrich, T., Abbas, M., Antonopoulos, W., Schipper, E., Hasemeier, B., Kreipe, H., and Lehmann, U. (2015). The CpG island methylator phenotype in breast cancer is associated with the lobular subtype. *Epigenomics*, 7(2):187–199.

Rønneberg, J. A., Fleischer, T., Solvang, H. K., Nordgard, S. H., Edvardsen, H., Potapenko, I., Nebdal, D., Daviaud, C., Gut, I., Bukholm, I., Naume, B., Børresen-Dale, A. L., Tost, J., and Kristensen, V. (2011). Methylation profiling with a panel of cancer related genes: Association with estrogen receptor, TP53 mutation status and expression subtypes in sporadic breast cancer. *Molecular Oncology*, 5(1):61–76.

Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., Dunning, M. J., Brown, G. D., Gojis, O., Ellis, I. O., Green, A. R., Ali, S., Chin, S.-F., Palmieri, C., Caldas, C., and Carroll, J. S. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, 481 VN -(7381):389–393.

Rousseau, M., Ferraiuolo, M. A., Crutchley, J. L., Wang, X. Q., Miura, H., Blanchette, M., and Dostie, J. (2014). Classifying leukemia types with chromatin conformation data. *Genome Biology*, 15(4):R60.

Rowinsky, E. K. (2004). The erbB family: targets for therapeutic development against cancer and therapeutic strategies using monoclonal antibodies and tyrosine kinase inhibitors. *Annual Review of Medicine*, 55:433–457.

Rowley, J. D. (1973). A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining. *Nature*, 243(5405):290–293.

Rueda, O. (2014). Lab communication. Technical report, Cancer Research UK Cambridge Institute.

Rueda, O. M., Sammut, S.-J., Chin, S.-F., Caswell, J. L., Ali, H. R., Pereira, B., Seoane, J. A., Batra, R. N., Bruna, A., Callari, M., Provenzano, E., Liu, B., Wright, K., Parisien, M., Gillett, C., McKinney, S., Green, A. R., Murphy, L., Purushotham, A., Ellis, I. O., Pharoah, P. D., Rueda, C., Aparicio, S. A., Curtis, C., and Caldas, C. (2017). The (spatio-temporal) dynamics of breast cancer relapse. *in preparation*.

Schlesinger, Y., Straussman, R., Keshet, I., Farkash, S., Hecht, M., Zimmerman, J., Eden, E., Yakhini, Z., Ben-Shushan, E., Reubinoff, B. E., Bergman, Y., Simon, I., and Cedar, H. (2007). Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nature Genetics*, 39(2):232–236.

Schübeler, D. (2015). Function and information content of DNA methylation. *Nature*, 517(7534):321–326.

Seligson, D. B., Horvath, S., Shi, T., Yu, H., Tze, S., Grunstein, M., and Kurdistani, S. K. (2005). Global histone modification patterns predict risk of prostate cancer recurrence. *Nature*, 435(7046):1262–1266.

Sergushichev, A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv*, page 060012.

Serra, R. W., Fang, M., Park, S. M., Hutchinson, L., and Green, M. R. (2014). A KRAS-directed transcriptional silencing pathway that mediates the CpG island methylator phenotype. *eLife*, 2014(3).

Shaffer, S. M., Dunagin, M. C., Torborg, S. R., Torre, E. A., Emert, B., Krepler, C., Beqiri, M., Sproesser, K., Brafford, P. A., Xiao, M., Eggan, E., Anastopoulos, I. N., Vargas-Garcia, C. A., Singh, A., Nathanson, K. L., Herlyn, M., and Raj, A. (2017). Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature*, 546(7658):431–435.

Shah, S. P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G., Bashashati, A., Prentice, L. M., Khattra, J., Burleigh, A., Yap, D., Bernard, V., McPherson, A., Shumansky, K., Crisan, A., Giuliany, R., Heravi-Moussavi, A., Rosner, J., Lai, D., Birol, I., Varhol, R., Tam, A., Dhalla, N., Zeng, T., Ma, K., Chan, S. K., Griffith, M., Moradian, A., Cheng, S.-W. G., Morin, G. B., Watson, P., Gelmon, K., Chia, S., Chin, S.-F., Curtis, C., Rueda, O. M., Pharoah, P. D., Damaraju, S., Mackey, J., Hoon, K., Harkins, T., Tadigotla, V., Sigaroudinia, M., Gascard, P., Tlsty, T., Costello, J. F., Meyer, I. M., Eaves, C. J., Wasserman, W. W., Jones, S., Huntsman, D., Hirst, M., Caldas, C., Marra, M. A., and Aparicio, S. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 486 VN -(7403):395–399.

Sharma, S. V., Lee, D. Y., Li, B., Quinlan, M. P., Takahashi, F., Maheswaran, S., McDermott, U., Azizian, N., Zou, L., Fischbach, M. A., Wong, K. K., Brandstetter, K., Wittner, B., Ramaswamy, S., Classon, M., and Settleman, J. (2010). A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations. *Cell*, 141(1):69–80.

Sheffield, N. C., Pierron, G., Klughammer, J., Datlinger, P., Schönegger, A., Schuster, M., Hadler, J., Surdez, D., Guillemot, D., Lapouble, E., Freneaux, P., Champigneulle, J., Bouvier, R., Walder, D., Ambros, I. M., Hutter, C., Sorz, E., Amaral, A. T., de Álava, E., Schallmoser, K., Strunk, D., Rinner, B., Liegl-Atzwanger, B., Huppertz, B., Leithner, A., de Pinieux, G., Terrier, P., Laurence, V., Michon, J., Ladenstein, R., Holter, W., Windhager, R., Dirksen, U., Ambros, P. F., Delattre, O., Kovar, H., Bock, C., and Tomazou, E. M. (2017). DNA methylation heterogeneity defines a disease spectrum in Ewing sarcoma. *Nature Medicine*, 23(3):386–395.

Shen, H. and Laird, P. W. (2013). Interplay between the cancer genome and epigenome.

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912.

Shepherd, J. H., Uray, I. P., Mazumdar, A., Tsimelzon, A., Savage, M., Hilsenbeck, S. G., and Brown, P. H. (2016). The SOX11 transcription factor is a critical regulator of basal-like breast cancer growth, invasion, and basal-like gene expression. *Oncotarget*, 7(11):13106–13121.

Shipony, Z., Mukamel, Z., Cohen, N. M., Landan, G., Chomsky, E., Zeliger, S. R., Fried, Y. C., Ainbinder, E., Friedman, N., and Tanay, A. (2014). Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature*, 513(7516):115–119.

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, 479(7371):74–79.

Siegfried, Z. and Cedar, H. (1997). DNA methylation: a molecular lock. *Current biology : CB*, 7(5):R305–R307.

Siegmund, K. D., Marjoram, P., Woo, Y.-J., Tavare, S., and Shibata, D. (2009). Inferring clonal expansion and cancer stem cell dynamics from DNA methylation

patterns in colorectal cancers. *Proceedings of the National Academy of Sciences*, 106(12):4828–4833.

Siggens, L. and Ekwall, K. (2014). Epigenetics, chromatin and genome organization: Recent advances from the ENCODE project. *Journal of Internal Medicine*, 276(3):201–214.

Silverman, L. R. and Mufti, G. J. (2005). Methylation inhibitor therapy in the treatment of myelodysplastic syndrome. *Nature clinical practice. Oncology*, 2 Suppl 1(December):S12–23.

Sinn, H. P. and Kreipe, H. (2013). A brief overview of the WHO classification of breast tumors, 4th edition, focusing on issues and updates from the 3rd edition.

Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*, 11(8):817–820.

Song, C.-X., Szulwach, K. E., Fu, Y., Dai, Q., Yi, C., Li, X., Li, Y., Chen, C.-H., Zhang, W., Jian, X., Wang, J., Zhang, L., Looney, T. J., Zhang, B., Godley, L. A., Hicks, L. M., Lahn, B. T., Jin, P., and He, C. (2011). Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nature Biotechnology*, 29(1):68–72.

Sood, A. K., Wang, J., Mhawech-Fauceglia, P., Jana, B., Liang, P., and Geradts, J. (2009). Sam-pointed domain containing Ets transcription factor in luminal breast cancer pathogenesis. *Cancer Epidemiology Biomarkers and Prevention*, 18(6):1899–1903.

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lonning, P. E., and Borresen-Dale, A.-L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874.

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lonning, P. E., Brown, P. O., Borresen-Dale, A.-L., and Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, 100(14):8418–8423.

Sotiriou, C., Neo, S.-Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., Fox, S. B., Harris, A. L., and Liu, E. T. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences*, 100(18):10393–10398.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Van de Vijver, M. J., Bergh, J., Piccart, M., and Delorenzi, M. (2006). Gene expression profiling in breast cancer: Understanding the molecular

basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, 98(4):262–272.

Sottoriva, A., Spiteri, I., Shibata, D., Curtis, C., and Tavaré, S. (2013). Single-molecule genomic data delineate patient-specific tumor profiles and cancer stem cell organization. *Cancer Research*, 73(1):41–49.

Sparano, J. A. and Paik, S. (2008). Development of the 21-gene assay and its application in clinical practice and clinical trials. *Journal of Clinical Oncology*, 26(5):721–728.

Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., Wirbelauer, C., Oakeley, E. J., Gaidatzis, D., Tiwari, V. K., and Schübeler, D. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480 VN -(7378):490–495.

Stamatoyannopoulos, J. A., Adzhubei, I., Thurman, R. E., Kryukov, G. V., Mirkin, S. M., and Sunyaev, S. R. (2009). Human mutation rate associated with DNA replication timing. *Nature Genetics*, 41(4):393–395.

Stefansson, O. A., Moran, S., Gomez, A., Sayols, S., Arribas-Jorba, C., Sandoval, J., Hilmarsdottir, H., Olafsdottir, E., Tryggvadottir, L., Jonasson, J. G., Eyfjord, J., and Esteller, M. (2015). A DNA methylation-based definition of biologically distinct breast cancer subtypes. *Molecular Oncology*, 9(3):555–568.

Stephens, P. J., McBride, D. J., Lin, M.-L., Varela, I., Pleasance, E. D., Simpson, J. T., Stebbings, L. A., Leroy, C., Edkins, S., Mudie, L. J., Greenman, C. D., Jia, M., Latimer, C., Teague, J. W., Lau, K. W., Burton, J., Quail, M. A., Swerdlow, H., Churcher, C., Natrajan, R., Sieuwerts, A. M., Martens, J. W. M., Silver, D. P., Langerød, A., Russnes, H. E. G., Foekens, J. A., Reis-Filho, J. S., van 't Veer, L., Richardson, A. L., Børresen-Dale, A.-L., Campbell, P. J., Futreal, P. A., and Stratton, M. R. (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, 462(7276):1005–1010.

Stephens, P. J., Tarpey, P. S., Davies, H., Van Loo, P., Greenman, C., Wedge, D. C., Zainal, S. N., Martin, S., Varela, I., Bignell, G. R., Yates, L. R., Papaemmanuil, E., Beare, D., Butler, A., Cheverton, A., Gamble, J., Hinton, J., Jia, M., Jayakumar, A., Jones, D., Latimer, C., Lau, K. W., McLaren, S., McBride, D. J., Menzies, A., Mudie, L., Raine, K., Rad, R., Spencer Chapman, M., Teague, J., Easton, D., Langerød, A., OSBREAC, Lee, M. T. M., Shen, C.-Y., Tee, B. T. K., Huimin, B. W., Broeks, A., Vargas, A. C., Turashvili, G., Martens, J., Fatima, A., Miron, P., Chin, S.-F., Thomas, G., Boyault, S., Mariani, O., Lakhani, S. R., van de Vijver, M., van 't Veer, L., Foekens, J., Desmedt, C., Sotiriou, C., Tutt, A., Caldas, C., Reis-Filho, J. S., Aparicio, S. A. J. R., Salomon, A. V., Børresen-Dale, A.-L., Richardson, A., Campbell, P. J., Futreal, P. A., Stratton, M. R., Karesen, R., Schlichting, E., Naume, B., Sauer, T., and Ottestad, L. (2012). The landscape of cancer genes and mutational processes in breast cancer. *Nature*, 486 VN -(7403):400–404.

Stirzaker, C., Taberlay, P. C., Statham, A. L., and Clark, S. J. (2014). Mining cancer methylomes: Prospects and challenges. *Trends in Genetics*, 30(2):75–84.

Stirzaker, C., Zotenko, E., Song, J. Z., Qu, W., Nair, S. S., Locke, W. J., Stone, A., Armstong, N. J., Robinson, M. D., Dobrovic, A., Avery-Kiejda, K. A., Peters, K. M., French, J. D., Stein, S., Korbie, D. J., Trau, M., Forbes, J. F., Scott, R. J., Brown, M. A., Francis, G. D., and Clark, S. J. (2015). Methylome sequencing in triple-negative breast cancer reveals distinct methylation clusters with prognostic value. *Nature Communications*, 6:5899.

Stone, A., Zotenko, E., Locke, W. J., Korbie, D., Millar, E. K. A., Pidsley, R., Stirzaker, C., Graham, P., Trau, M., Musgrove, E. A., Nicholson, R. I., Gee, J. M. W., and Clark, S. J. (2015). DNA methylation of oestrogen-regulated enhancers defines endocrine sensitivity in breast cancer. *Nature Communications*, 6:7758.

Stricker, S. H., Köferle, A., and Beck, S. (2016). From profiles to function in epigenomics. *Nature Reviews Genetics*, 18(1):51–66.

Stroud, H., Feng, S., Morey Kinney, S., Pradhan, S., and Jacobsen, S. E. (2011). 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biology*, 12(6):R54.

Struhl, K. (2014). Is DNA methylation of tumour suppressor genes epigenetic? *eLife*, 2014(3).

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

Susan, J. C., Harrison, J., Paul, C. L., and Frommer, M. (1994). High sensitivity mapping of methylated cytosines. *Nucleic Acids Research*, 22(15):2990–2997.

Suzuki, M. M. and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, 9(6):465–476.

Szyf, M. (2012). DNA methylation signatures for breast cancer classification and prognosis. *Genome Medicine*, 4(3):26.

Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L. M., Liu, D. R., Aravind, L., and Rao, A. (2009). Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science*, 324(5929):930–935.

Takaku, M., Grimm, S. A., and Wade, P. A. (2015). GATA3 in breast cancer: Tumor suppressor or oncogene?

Talens, R. P., Christensen, K., Putter, H., Willemsen, G., Christiansen, L., Kremer, D., Suchiman, H. E. D., Slagboom, P. E., Boomsma, D. I., and Heijmans, B. T. (2012). Epigenetic variation during the adult lifespan: Cross-sectional and longitudinal data on monozygotic twin pairs. *Aging Cell*, 11(4):694–703.

Tan, A. (2003). Ongoing adjuvant trials with trastuzumab in breast cancer. *Seminars in Oncology*, 30(5 Suppl 16):54–64.

Tanay, A. (2017). Lab communication. Technical report, Weizmann Institute of Science.

Taub, M. A., Corrada Bravo, H., and Irizarry, R. A. (2010). Overcoming bias and systematic errors in next generation sequencing data. *Genome Medicine*, 2(12):87.

Teschendorff, A. E., Breeze, C. E., Zheng, S. C., and Beck, S. (2017). A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics*, 18(1):105.

Teschendorff, A. E., Gao, Y., Jones, A., Ruebner, M., Beckmann, M. W., Wachter, D. L., Fasching, P. A., and Widschwendter, M. (2016a). DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. *Nature Communications*, 7:10478.

Teschendorff, A. E., Jones, A., Fiegl, H., Sargent, A., Zhuang, J. J., Kitchener, H. C., and Widschwendter, M. (2012). Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. *Genome Medicine*, 4(3):24.

Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Weisenberger, D. J., Shen, H., Campan, M., Noushmehr, H., Bell, C. G., Maxwell, A. P., Savage, D. A., Mueller-Holzner, E., Marth, C., Kocjan, G., Gayther, S. A., Jones, A., Beck, S., Wagner, W., Laird, P. W., Jacobs, I. J., and Widschwendter, M. (2010). Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Research*, 20(4):440–446.

Teschendorff, A. E., Miremadi, A., Pinder, S. E., Ellis, I. O., and Caldas, C. (2007). An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome Biology*, 8(8):R157.

Teschendorff, A. E., West, J., and Beck, S. (2013). Age-associated epigenetic drift: Implications, and a case of epigenetic thrift? *Human Molecular Genetics*, 22(R1).

Teschendorff, A. E. and Widschwendter, M. (2012). Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics*, 28(11):1487–1494.

Teschendorff, A. E., Yang, Z., Wong, A., Pipinikas, C. P., Jiao, Y., Jones, A., Anjum, S., Hardy, R., Salvesen, H. B., Thirlwell, C., Janes, S. M., Kuh, D., and Widschwendter, M. (2015). Correlation of Smoking-Associated DNA Methylation Changes in Buccal Cells With DNA Methylation Changes in Epithelial Cancer. *JAMA Oncology*, 1(4):476.

Teschendorff, A. E., Zheng, S. C., Feber, A., Yang, Z., Beck, S., and Widschwendter, M. (2016b). The multi-omic landscape of transcription factor inactivation in cancer. *Genome Medicine*, 8(1):89.

The Cancer Genome Atlas Research Network (2013a). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456):43–49.

The Cancer Genome Atlas Research Network (2013b). Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *New England Journal of Medicine*, 368(22):2059–2074.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572.

Tien, J. F., Mazloomian, A., Cheng, S. W., Hughes, C. S., Chow, C. C., Canapi, L. T., Oloumi, A., Trigo-Gonzalez, G., Bashashati, A., Xu, J., Chang, V. C., Shah, S. P., Aparicio, S., and Morin, G. B. (2017). CDK12 regulates alternative last exon mRNA splicing and promotes breast cancer cell invasion. *Nucleic Acids Research*, 45(11):6698–6716.

Tomaskovic-Crook, E., Thompson, E. W., and Thiery, J. P. (2009). Epithelial to mesenchymal transition and breast cancer. *Breast Cancer Research*, 11(6):213.

Tufegdzic Vidakovic, A., Rueda, O. M., Vervoort, S. J., Sati Batra, A., Goldgraben, M. A., Uribe-Lewis, S., Greenwood, W., Coffer, P. J., Bruna, A., and Caldas, C. (2015). Context-Specific Effects of TGF-β/SMAD3 in Cancer Are Modulated by the Epigenome. *Cell Reports*, 13(11):2480–2490.

van de Vijver, M. J., He, Y. D., van 't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. (2002). A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. *New England Journal of Medicine*, 347(25):1999–2009.

Van Loo, P., Nordgard, S. H., Lingjaerde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B., Perou, C. M., Borresen-Dale, A.-L., and Kristensen, V. N. (2010). Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915.

van Roy, F. (2014). Beyond E-cadherin: roles of other cadherin superfamily members in cancer. *Nature Reviews Cancer*, 14(2):121–134.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, 26(12).

Veeck, J., Ropero, S., Setien, F., Gonzalez-Suarez, E., Osorio, A., Benitez, J., Herman, J. G., and Esteller, M. (2010). BRCA1 CpG island hypermethylation predicts sensitivity to poly(adenosine diphosphate)-ribose polymerase inhibitors.

Veeck, J., Wild, P. J., Fuchs, T., Schüffler, P. J., Hartmann, A., Knüchel, R., and Dahl, E. (2009). Prognostic relevance of Wnt-inhibitory factor-1 (WIF1) and Dickkopf-3 (DKK3) promoter methylation in human breast cancer. *BMC Cancer*, 9(1):217.

Vendrell, J. A., Thollet, A., Nguyen, N. T., Ghayad, S. E., Vinot, S., Bièche, I., Grisard, E., Josserand, V., Coll, J. L., Roux, P., Corbo, L., Treilleux, I., Rimokh, R., and Cohen, P. A. (2012). ZNF217 is a marker of poor prognosis in breast cancer that drives epithelial-mesenchymal transition and invasion. *Cancer Research*, 72(14):3593–3606.

Venkatraman, E. S. and Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6):657–663.

Vidakovic, A. T. (2014). *Epigenetic determinants of context specificity in breast cancer*. PhD thesis, University of Cambridge.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer Genome Landscapes. *Science*, 339(6127):1546–1558.

Von Bergh, A. R., Beverloo, H. B., Rombout, P., Van Wering, E. R., Van Weel, M. H., Beverstock, G. C., Kluin, P. M., Slater, R. M., and Schuuring, E. (2002). LAF4, an AF4-related gene, is fused to MLL in infant acute lymphoblastic leukemia. *Genes Chromosomes and Cancer*, 35(1):92–96.

Waddington, C. H. (1942). Canalization of Development and the Inheritance of Acquired Characters. *Nature*, 150(3811):563–565.

Waddington, C. H. (1957). The strategy of the genes. A discussion of some aspects of theoretical biology. *The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser.*, page 262 p.

Wang, F., Zhang, N., Wang, J., Wu, H., and Zheng, X. (2016). Tumor purity and differential methylation in cancer epigenomics. *Briefings in Functional Genomics*, 15(6):408–419.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16).

Wang, K., Li, X., Dong, S., Liang, J., Mao, F., Zeng, C., Wu, H., Wu, J., Cai, W., and Sun, Z. S. (2015). Q-RRBS: A quantitative reduced representation bisulfite sequencing method for single-cell methylome analyses. *Epigenetics*, 10(9):775–783.

Weber, M., Davies, J. J., Wittig, D., Oakeley, E. J., Haase, M., Lam, W. L., and Schübeler, D. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genetics*, 37(8):853–862.

Weinstein, J. N., Akbani, R., Broom, B. M., Wang, W., Verhaak, R. G. W., McConkey, D., Lerner, S., Morgan, M., Creighton, C. J., Smith, C., Kwiatkowski, D. J., Cherniack, A. D., Kim, J., Sekhar Pedamallu, C., Noble, M. S., Al-Ahmadie, H. A., Reuter, V. E., Rosenberg, J. E., Bajorin, D. F., Bochner, B. H., Solit, D. B.,

Koppie, T., Robinson, B., Gordenin, D. A., Fargo, D., Klimczak, L. J., Roberts, S. A., Au, J., Laird, P. W., Hinoue, T., Schultz, N., Ramirez, R., Hansel, D., Hoadley, K. A., Kim, W. Y., Damrauer, J. S., Baylin, S. B., Mungall, A. J., Gordon Robertson, A., Chu., A., Kwiatkowski, D. J., Sougnez, C., Cibulskis, K., Lichtenstein, L., Sivachenko, A., Stewart, C., Lawrence, M. S., Getz, G., Lander, E., Gabriel., S. B., Creighton, C. J., Donehower, L., Cherniack, A. D., Kim, J., Carter, S. L., Saksena, G., Schumacher, S. E., Sougnez, C., Freeman, S. S., Jung, J., Sekhar Pedamallu, C., Bhatt, A. S., Pugh, T., Getz, G., Beroukhim, R., Gabriel, S. B., Meyerson, M., Mungall, A. J., Gordon Robertson, A., Chu, A., Ally, A., Balasundaram, M., Butterfield, Y. S. N., Dhalla, N., Hirst, C., Holt, R. A., Jones, S. J. M., Lee, D., Li, H. I., Marra, M. A., Mayo, M., Moore, R. A., Schein, J. E., Sipahimalani, P., Tam, A., Thiessen, N., Wong, T., Wye, N., Bowlby, R., Chuah, E., Guin, R., Jones, S. J. M., Marra, M. A., Hinoue, T., Shen, H., Bootwalla, M. S., Triche Jr, T., Lai, P. H., Van Den Berg, D. J., Weisenberger, D. J., Laird, P. W., Hansel, D., Hoadley, K. A., Balu, S., Bodenheimer, T., Damrauer Alan P. Hoyle, J. S., Jefferys, S. R., Meng, S., Mose, L. E., Simons, J. V., Soloway, M. G., Wu, J., Kim, W. Y., Parker, J. S., Neil Hayes, D., Roach, J., Buda, E., Jones, C. D., Mieczkowski, P. A., Tan, D., Veluvolu, U., Waring, S., Todd Auman, J., Perou, C. M., Wilkerson, M. D., Santoso, N., Parfenov, M., Ren, X., Pantazi, A., Hadjipanayis, A., Seidman, J., Kucherlapati, R., Lee, S., Yang, L., Park, P. J., Baylin, S. B., Wei Xu, A., Protopopov, A., Zhang, J., Bristow, C., Mahadeshwar, H. S., Seth, S., Song, X., Tang, J., Zeng, D., Chin, L., Guo, C., Weinstein, J. N., Akbani, R., Broom, B. M., McConkey, D., Casasent, T. D., Liu, W., Ju, Z., Motter, T., Peng, B., Ryan, M., Wang, W., Verhaak, R. G. W., Su, X., Yang, J.-Y., Lorenzi, P. L., Yao, H., Zhang, N., Zhang, J., Mills, G. B., Kim, J., Noble, M. S., Cho, J., DiCara, D., Frazer, S., Gehlenborg, N., Heiman, D. I., Lin, P., Liu, Y., Stojanov, P., Voet, D., Zhang, H., Zou, L., Chin, L., Getz, G., Bernard, B., Kreisberg, D., Reynolds, S., Rovira, H., Shmulevich, I., Ramirez, R., Schultz, N., Gao, J., Jacobsen, A., Arman Aksoy, B., Antipin, Y., Ciriello, G., Dresdner, G., Gross, B., Lee, W., Reva, B., Shen, R., Sinha, R., Onur Sumer, S., Weinhold, N., Ladanyi, M., Sander, C., Benz, C., Carlin, D., Haussler, D., Ng, S., Paull, E. O., Stuart, J., Zhu, J., Liu, Y., Zhang, W., Taylor, B. S., Lichtenberg, T. M., Zmuda, E., Barr, T., Black, A. D., George, M., Hanf, B., Helsel, C., McAllister, C., Ramirez, N. C., Tabler, T. R., Weaver, S., Wise, L., Bowen, J., Gastier-Foster, J. M., Weinstein, J. N., Lerner, S., Jian, W., Tello, S., Ittman, M., Castro, P., McClenden, W. D., Morgan, M., Gibbs, R., Liu, Y., Saller, C., Tarvin, K., DiPiero, J. M., Owens, J., Bollag, R., Li, Q., Weinberger, P., Czerwinski, C., Huelsenbeck-Dill, L., Iacocca, M., Petrelli, N., Rabeno, B., Swanson, P., Shelton, T., Curley, E., Gardner, J., Mallery, D., Penny, R., Van Bang, N., Thi Hanh, P., Kohl, B., Van Le, X., Duc Phu, B., Thorp, R., Viet Tien, N., Quang Vinh, L., Sandusky, G., Burks, E., Christ, K., Gee, J., Holway, A., Moinzadeh, A., Sorcini, A., Sullivan, T., Al-Ahmadie, H. A., Bajorin, D. F., Bochner, B. H., Garcia-Grossman, I. R., Regazzi, A. M., Solit, D. B., Rosenberg, J. E., Reuter, V. E., Koppie, T., Boice, L., Kimryn Rathmell, W., Thorne, L., Bastacky, S., Davies, B., Dhir, R., Gingrich, J., Hrebinko, R., Maranchie, J., Nelson, J., Parwani, A., Bshara, W., Gaudioso, C., Morrison, C., Alexopoulou, V., Bartlett, J., Engel, J., Kodeeswaran, S., Antic, T., O'Donnell, P. H., Smith, N. D., Steinberg, G. D., Egea, S., Gomez-Fernandez, C., Herbert, L., Jorda, M., Soloway, M., Beaver, A., Carter, S., Kapur, P., Lewis, C., Lotan, Y., Robinson, B., Hansel, D., Guo, C., Bondaruk, J., Czerniak, B., Akbani, R., Broom, B. M., Liu, Y., Zhang, W., Weinstein, J. N., Lerner, S., Morgan, M., Kim, J., Cherniack, A. D., Freeman,

S. S., Sekhar Pedamallu, C., Noble, M. S., Kwiatkowski, D. J., Al-Ahmadie, H. A., Bajorin, D. F., Bochner, B. H., Solit, D. B., Rosenberg, J. E., Reuter, V. E., Koppie, T., Robinson, B., Skinner, E., Ramirez, R., Schultz, N., Hansel, D., Kim, W. Y., Guo, C., Bondaruk, J., Aldape, K., Czerniak, B., Jensen, M. A., Kahn, A. B., Pihl, T. D., Pot, D. A., Srinivasan, D., Wan, Y., Ferguson, M. L., Claude Zenklusen, J., Davidsen, T., Demchok, J. A., Mills Shaw, K. R., Sheth, M., Tarnuzzer, R., Wang, Z., Yang, L., Hutter, C., Ozenberger, B. A., Sofia, H. J., and Eley, G. (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492):315–322.

Weisenberger, D. J. (2014). Characterizing DNA methylation alterations from the cancer genome atlas. *Journal of Clinical Investigation*, 124(1):17–23.

Wellings SR, Jensen HM, M. R. (1975). An Atlas of Subgross Pathology of the Human Breast With Special Reference to Possible Precancerous Lesions. J Natl Cancer Inst 55: 231-273. *Journal of the National Cancer Institute*, 55(2):231–273.

West, J., Beck, S., Wang, X., and Teschendorff, A. E. (2013). An integrative network algorithm identifies age-associated differential methylation interactome hotspots targeting stem-cell differentiation pathways. *Scientific Reports*, 3(1):1630.

Widschwendter, M., Fiegl, H., Egle, D., Mueller-Holzner, E., Spizzo, G., Marth, C., Weisenberger, D. J., Campan, M., Young, J., Jacobs, I., and Laird, P. W. (2007). Epigenetic stem cell signature in cancer. *Nature Genetics*, 39(2):157–158.

Wishart, G. C., Azzato, E. M., Greenberg, D. C., Rashbass, J., Kearins, O., Lawrence, G., Caldas, C., Pharoah, P. D., Haybittle, J., Blamey, R., Elston, C., Johnson, J., Doyle, P., Campbell, F., Todd, J., Dowle, C., Williams, M., Elston, C., Ellis, I., Hinton, C., Blamey, R., Haybittle, J., D'Eredita, G., Giardina, C., Martellotta, M., Natale, T., Ferrarese, F., Blamey, R., Ellis, I., Pinder, S., Lee, A., Macmillan, R., Morgan, D., Robertson, J., Mitchell, M., Ball, G., Haybitle, J., Elston, C., Blamey, R., Pinder, S., Ball, G., Ellis, I., Elston, C., Mitchell, M., Haybittle, J., Ravdin, P., Siminoff, L., Davis, G., Mercer, M., Hewlett, J., Gerson, N., Parker, H., Olivoto, I., Bajdik, C., Ravdin, P., Speers, C., Coldman, A., Norris, B., Davis, G., Chia, S., Gelmon, K., Campbell, H., Taylor, M., Harris, A., Gray, A., Azzato, E., Greenberg, D., Shah, M., Blows, F., Driver, K., Caporaso, N., Pharoah, P., May, S., Hosmer, D., Joensuu, H., Lehtimaki, T., Holli, K., Elomaa, L., Turpeenniemi-Hujanen, T., Kataja, V., Anttila, A., Lundin, M., Isola, J., Lundin, J., Shen, Y., Yang, Y., Inoue, L., Munsell, M., Miller, A., Berry, D., Wishart, G., Greenberg, D., Britton, P., Chou, P., Brown, C., Purushotham, A., and Duffy, S. (2010). PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Research*, 12(1):R1.

Xi, Y., Bock, C., Müller, F., Sun, D., Meissner, A., and Li, W. (2012). RRBSMAP: A fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing. *Bioinformatics*, 28(3):430–432.

Xi, Y. and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, 10(1):232.

Xu, X., Tao, Y., Gao, X., Zhang, L., Li, X., Zou, W., Ruan, K., Wang, F., Xu, G.-l., and Hu, R. (2016). A CRISPR-based approach for targeted DNA demethylation. *Cell Discovery*, 2:16009.

Xu, Y., Diao, L., Chen, Y., Liu, Y., Wang, C., Ouyang, T., Li, J., Wang, T., Fan, Z., Fan, T., Lin, B., Deng, D., Narod, S. A., and Xie, Y. (2013). Promoter methylation of BRCA1 in triple-negative breast cancer predicts sensitivity to adjuvant chemotherapy. *Annals of Oncology*, 24(6):1498–1505.

Yamamoto, M., Cid, E., Bru, S., and Yamamoto, F. (2011). Rare and frequent promoter methylation, respectively, of TSHZ2 and 3 genes that are both downregulated in expression in breast and prostate cancers. *PLoS ONE*, 6(3).

Yamashita, N., Tokunaga, E., Kitao, H., Hitchins, M., Inoue, Y., Tanaka, K., Hisamatsu, Y., Taketani, K., Akiyoshi, S., Okada, S., Oda, Y., Saeki, H., Oki, E., and Maehara, Y. (2015). Epigenetic Inactivation of BRCA1 Through Promoter Hypermethylation and Its Clinical Importance in Triple-Negative Breast Cancer. *Clinical Breast Cancer*, 15(6):498–504.

Yang, H., Liu, Y., Bai, F., Zhang, J.-Y., Ma, S.-H., Liu, J., Xu, Z.-D., Zhu, H.-G., Ling, Z.-Q., Ye, D., Guan, K.-L., and Xiong, Y. (2013). Tumor development is associated with decrease of TET gene expression and 5-methylcytosine hydroxylation. *Oncogene*, 32(5):663–669.

Yang, Z., Jones, A., Widschwendter, M., and Teschendorff, A. E. (2015). An integrative pan-cancer-wide analysis of epigenetic enzymes reveals universal patterns of epigenomic deregulation in cancer. *Genome Biology*, 16(1):140.

Yang, Z., Wong, A., Kuh, D., Paul, D. S., Rakyan, V. K., Leslie, R. D., Zheng, S. C., Widschwendter, M., Beck, S., and Teschendorff, A. E. (2016). Correlation of an epigenetic mitotic clock with cancer risk. *Genome Biology*, 17(1):205.

Yatabe, Y., Tavare, S., and Shibata, D. (2001). Investigating stem cells in human colon by using methylation patterns. *Proceedings of the National Academy of Sciences*, 98(19):10839–10844.

Yiangou, C., Gomm, J. J., Coope, R. C., Law, M., Luqmani, Y. A., Shousha, S., Coombes, R. C., and Johnston, C. L. (1997). Fibroblast growth factor 2 in breast cancer: occurrence and prognostic significance. *British journal of cancer*, 75(1):28–33.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Syed, K., Das, P. K., Kivioja, T., Dave, K., Zhong, F., Nitta, K. R., Taipale, M., Popov, A., Ginno, P. A., Domcke, S., Yan, J., Schübeler, D., Vinson, C., and Taipale, J. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, 356(6337):eaaj2239.

Zeitz, M. J., Ay, F., Heidmann, J. D., Lerner, P. L., Noble, W. S., Steelman, B. N., and Hoffman, A. R. (2013). Genomic Interaction Profiles in Breast Cancer Reveal Altered Chromatin Architecture. *PLoS ONE*, 8(9).

Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T., and Engelhardt, B. E. (2015). Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biology*, 16(1):14.

Zheng, S. C., Beck, S., Jaffe, A. E., Koestler, D. C., Hansen, K. D., Houseman, A. E., Irizarry, R. A., and Teschendorff, A. E. (2017a). Correcting for cell-type heterogeneity in epigenome-wide association studies: revisiting previous analyses. *Nature Methods*, 14(3):216–217.

Zheng, S. C., Widschwendter, M., and Teschendorff, A. E. (2016). Epigenetic drift, epigenetic clocks and cancer risk. *Epigenomics*, 8(5):705–719.

Zheng, X., Zhang, N., Wu, H.-J., and Wu, H. (2017b). Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. *Genome Biology*, 18(1):17.

Zheng, X., Zhao, Q., Wu, H.-J., Li, W., Wang, H., Meyer, C. A., Qin, Q. A., Xu, H., Zang, C., Jiang, P., Li, F., Hou, Y., He, J., Wang, J., Wang, J., Zhang, P., Zhang, Y., and Liu, X. S. (2014). MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes. *Genome Biology*, 15(7):419.

Zhu, X., Shan, L., Wang, F., Wang, J., Wang, F., Shen, G., Liu, X., Wang, B., Yuan, Y., Ying, J., and Yang, H. (2015). Hypermethylation of BRCA1 gene: implication for prognostic biomarker and therapeutic target in sporadic primary triple-negative breast cancer. *Breast Cancer Research and Treatment*, 150(3):479–486.

Ziller, M. J., Müller, F., Liao, J., Zhang, Y., Gu, H., Bock, C., Boyle, P., Epstein, C. B., Bernstein, B. E., Lengauer, T., Gnirke, A., and Meissner, A. (2011). Genomic distribution and Inter-Sample variation of Non-CpG methylation across human cell types. *PLoS Genetics*, 7(12).

Zwiener, I., Blettner, M., and Hommel, G. (2011). Survival analysis: part 15 of a series on evaluation of scientific publications. *Dtsch Arztebl Int*, 108(10):163–169.

Appendix A

Supplementary figures

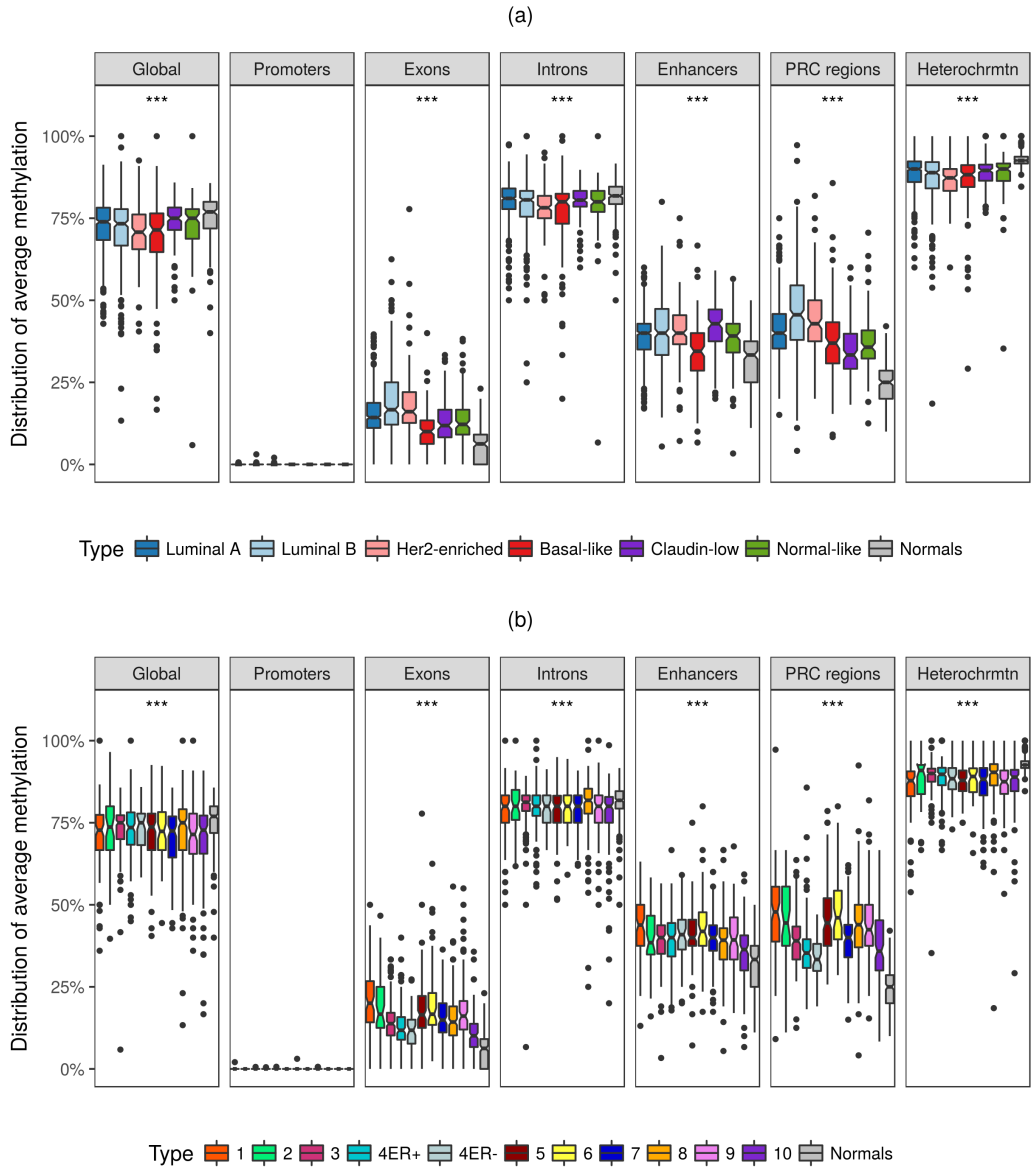


Figure A.1: Supervised analysis of DNA methylation profiles reveal distinct epigenetic landscapes in breast cancer subtypes. (a) Distribution of average methylation estimates for the breast cancer Intrinsic subtypes and normal tissues stratified by genomic feature. For each sample, the median methylation level across each genomic feature, and the resulting distributions were plotted. For each genomic feature, methylation estimates between these three categories were compared using the Kruskal Wallis test. *FDR p-values* were denoted. **(b)** Same as (a) but for the breast cancer Integrative clusters. (N.S.= *FDR p-value* < 0.1, $*$ = *FDR p-value* < 0.05, ** = *FDR p-value* < 0.01, *** = *FDR p-value* < 0.001, **** = *FDR p-value* < 0.0001). Heterochrmtn = Heterochromatin.

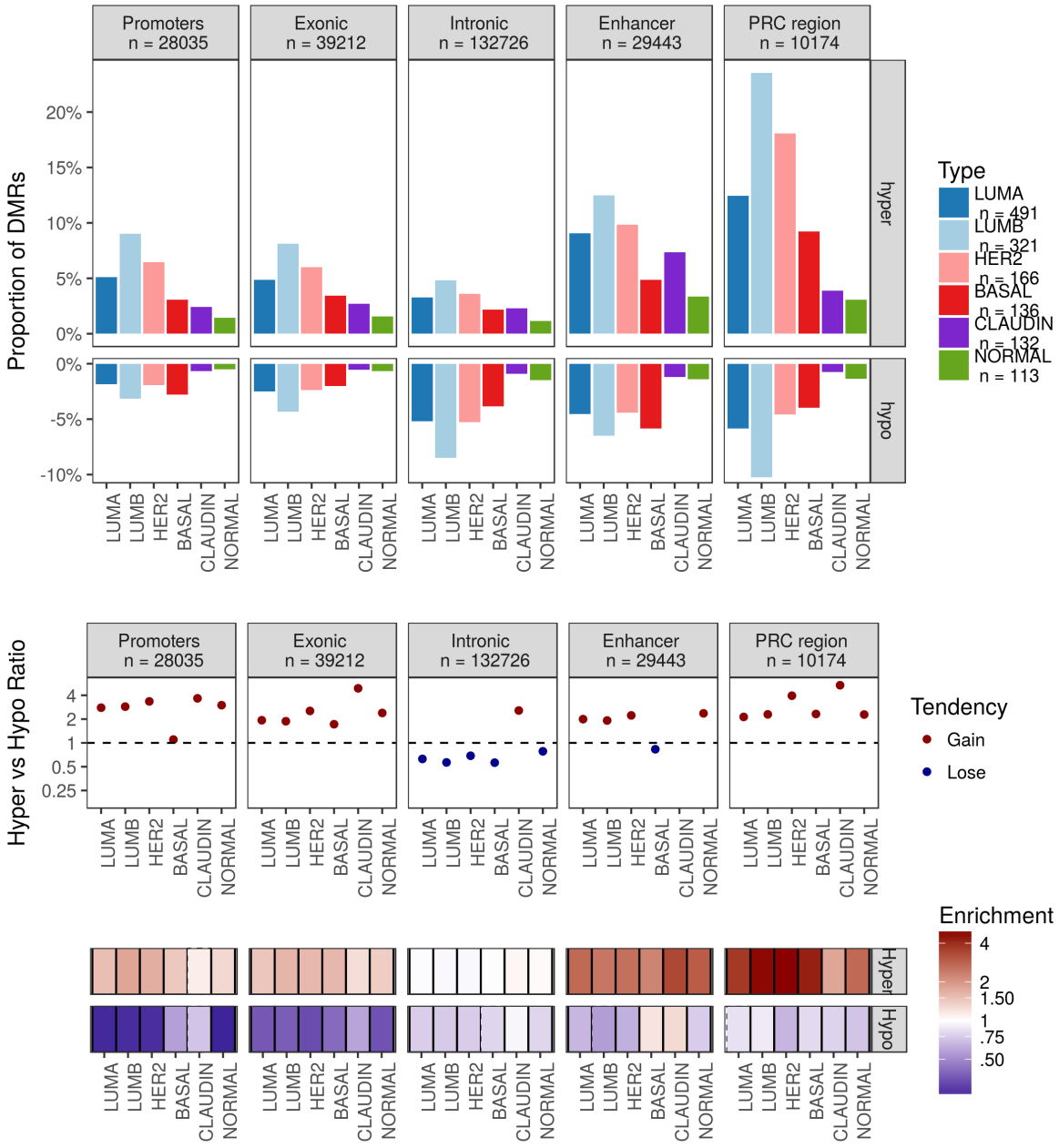


Figure A.2: (Caption on next page.)

Figure A.2: (Previous page.) **Hyper and hypo DMRs detected in the Intrinsic subtypes. (top panel)** The proportion of hyper and hypo DMRs within each genomic feature, relative to the total number of SCCRUB regions within the respective genomic feature (the total number is noted in the boxes at the top), detected across the Intrinsic subtypes. x-axis represents the tumour subtype and the five panels represent the individual genomic features. Analyses were conducted separately for the 6 tumour subtypes and 5 genomic features resulting in 30 bars. Positive bars on y-axis represents hyper DMRs and negative bars represents hypo DMRs. **(middle panel)** The ratio between the number of hyper and hypo DMRs detected for each genomic feature which represents the inclination of a tumour subtype to significantly gain or lose methylation in a specific genomic feature. **(bottom panel)** Enrichment analysis of hyper and hypo DMRs across the 5 genomic features conducted separately for each tumour subtype, as explained in the text (hypergeometric test). Top squares represent enrichment of hyper DMRs and bottom represent enrichment of hypo DMRs. Colour of the squares represent level of enrichment (observed/ expected). Red represents enriched (enrichment > 1), blue represents depleted (enrichment < 1) and white represents no enrichment (enrichment = 1). Square boundaries represent whether the enrichment was significant (solid lines = *FDR p-value* < 0.05; dotted lines = *FDR p-value* > 0.05). LUMA = Luminal A. LUMB = Luminal B. HER2 = HER2-enriched. BASAL = Basal-like. CLAUDIN = Claudin-low. NORMAL = Normal-like.

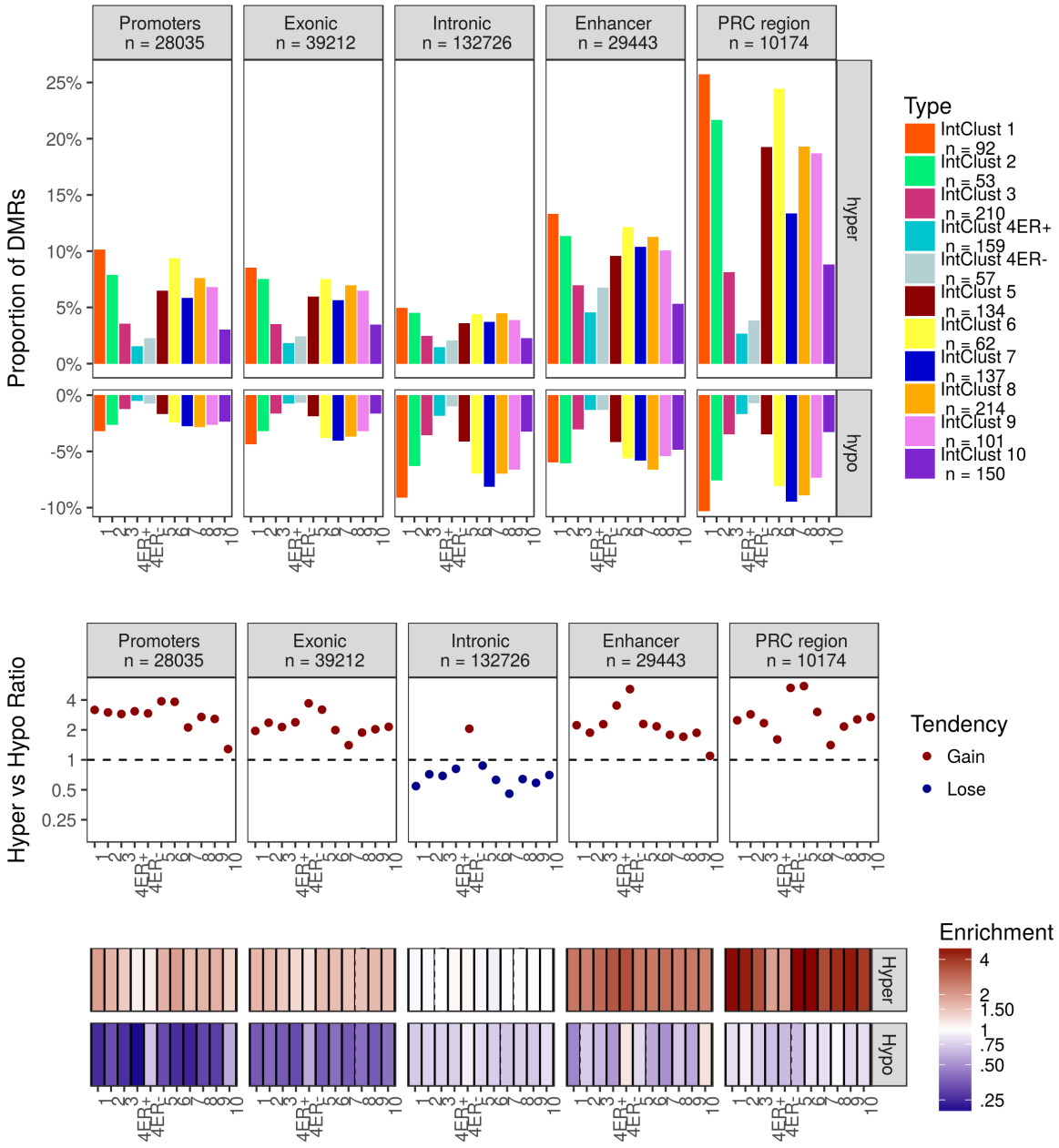


Figure A.3: (Caption on next page.)

Figure A.3: (Previous page.) **Hyper and hypo DMRs detected in the Integrative clusters. (top panel)** The proportion of hyper and hypo DMRs within each genomic feature, relative to the total number of SCCRUB regions within the respective genomic feature (the total number is noted in the boxes at the top), detected across the Integrative clusters. x-axis represents the tumour subtype and the five panels represent the individual genomic features. Analyses were conducted separately for the 11 tumour subtypes and 5 genomic features resulting in 55 bars. Positive bars on y-axis represents hyper DMRs and negative bars represents hypo DMRs. **(middle panel)** The ratio between the number of hyper and hypo DMRs detected for each genomic feature which represents the inclination of a tumour subtype to significantly gain or lose methylation in a specific genomic feature. **(bottom panel)** Enrichment analysis of hyper and hypo DMRs across the 5 genomic features conducted separately for each tumour subtype, as explained in the text (hypergeometric test). Top squares represent enrichment of hyper DMRs and bottom represent enrichment of hypo DMRs. Colour of the squares represent level of enrichment (observed/ expected). Red represents enriched (enrichment > 1), blue represents depleted (enrichment < 1) and white represents no enrichment (enrichment = 1). Square boundaries represent whether the enrichment was significant (solid lines = *FDR p-value* < 0.05; dotted lines = *FDR p-value* > 0.05).

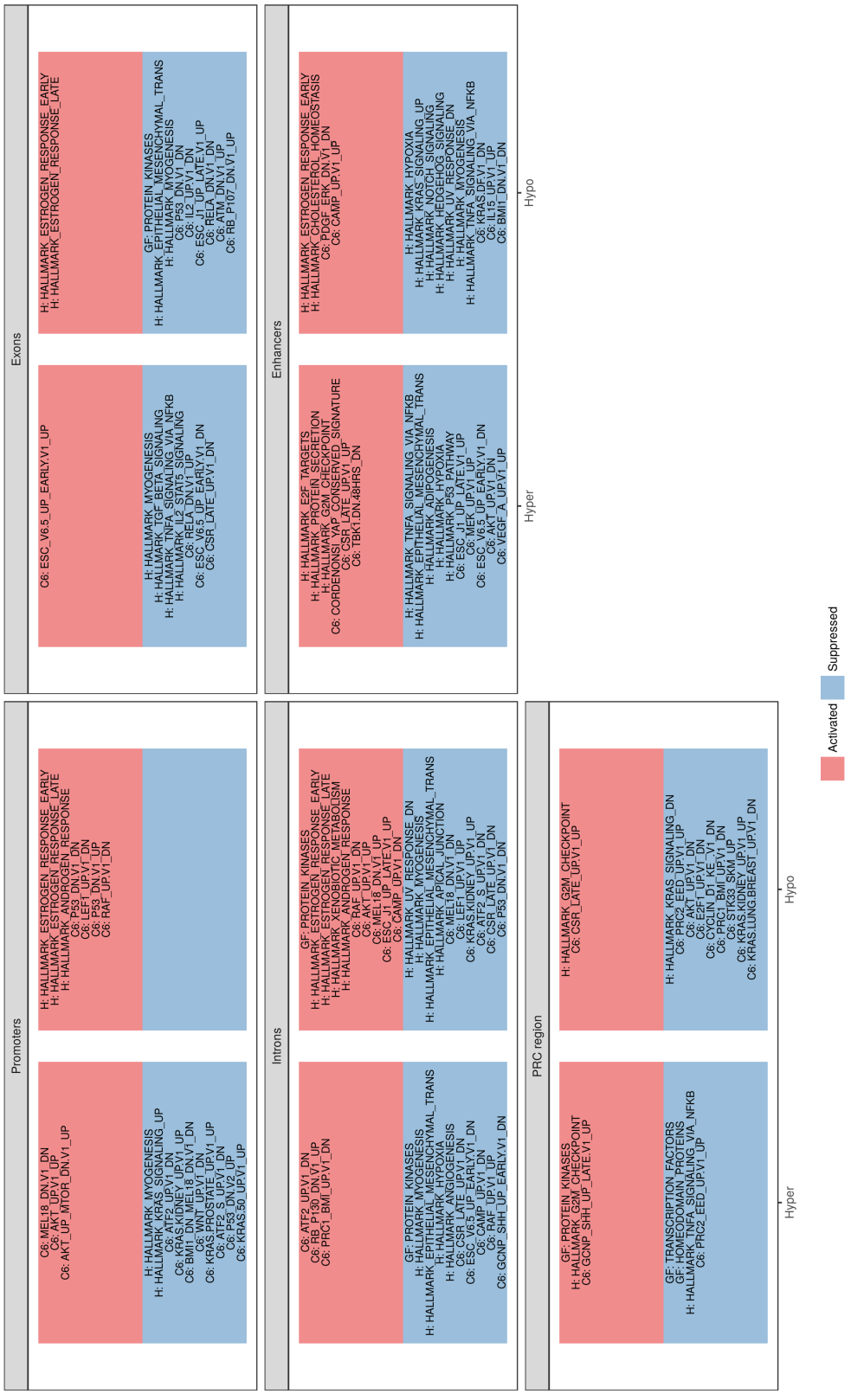


Figure A.4: (Caption on next page.)

Figure A.4: (Previous page.) **Cancer pathways are epigenetically regulated in a genomic feature-specific manner in ER+ tumours.** Significantly enriched gene sets comprising of at least 4 genes that are regulated by DMRs (Enrichment (observed/ expected) > 1.5, *FDR p-value* < 0.05; hypergeometric test, as explained in text) were identified. This was conducted separately for hyper (left) and hypo (right) DMRs; and for upregulating expression-DMRs (Activated: red) and downregulating expression-DMRs (Suppressed: blue). Separate analyses were conducted for expression-DMRs within 5 disinct genomic features. Directed and background DMRs (versus normal tissues) considered. Gene Set Enrichment Analysis (GSEA: pathways tested included Gene Families, Hallmark and oncogenic gene sets obtained from Molecular Signatures Database [Subramanian et al., 2005, MSigDB]). Top 10 enriched gene sets included per analysis.

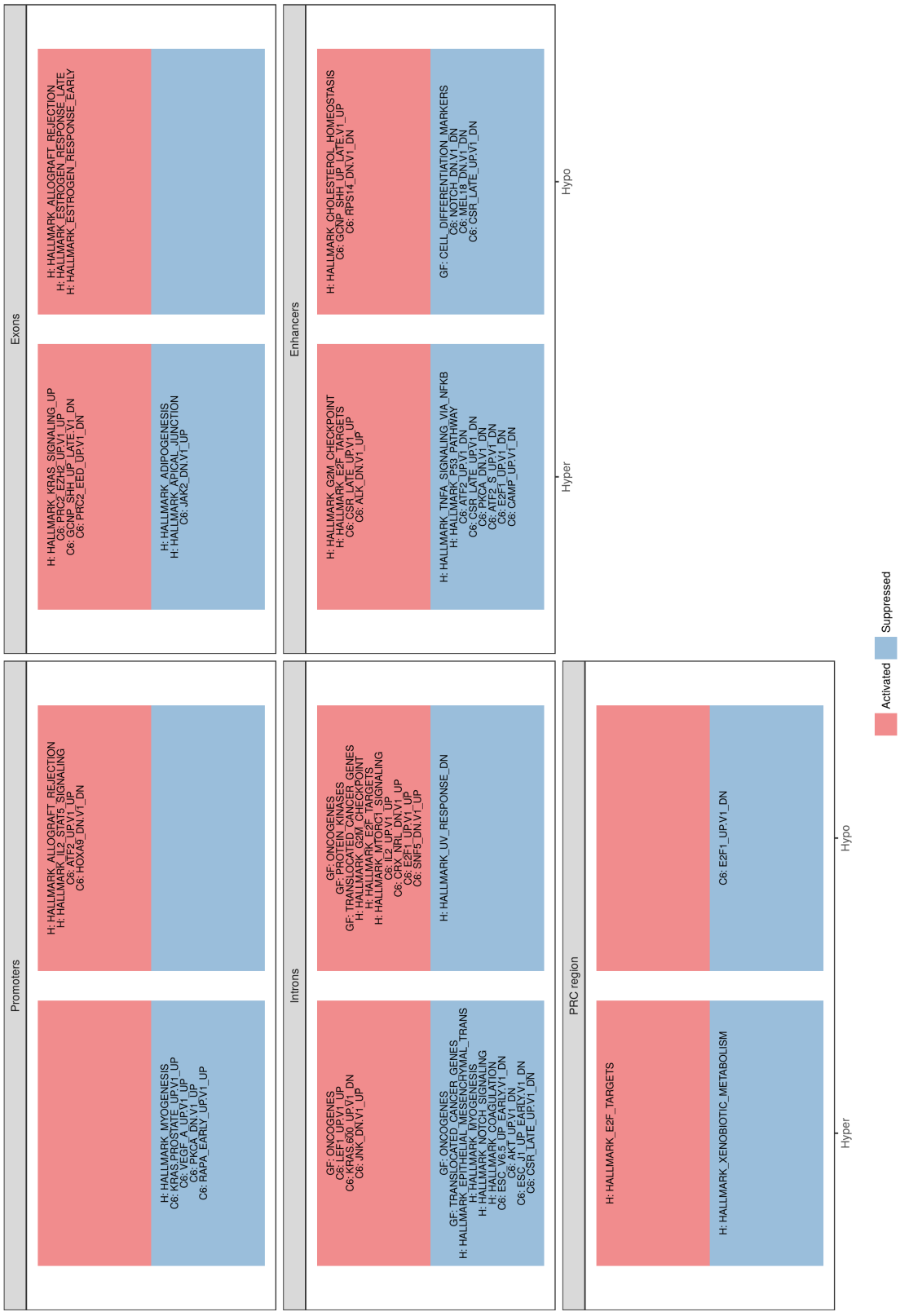


Figure A.5: (Caption on next page.)

Figure A.5: (Previous page.) **Cancer pathways are epigenetically regulated in a genomic feature-specific manner in ER- tumours**. Significantly enriched gene sets comprising of at least 4 genes that are regulated by DMRs (Enrichment (observed/ expected) > 1.5, *FDR p-value* < 0.05; hypergeometric test, as explained in text) were identified. This was conducted separately for hyper (left) and hypo (right) DMRs; and for upregulating expression-DMRs (Activated: red) and downregulating expression-DMRs (Suppressed: blue). Separate analyses were conducted for expression-DMRs within 5 disinct genomic features. Directed and background DMRs (versus normal tissues) considered. Gene Set Enrichment Analysis (GSEA: pathways tested included Gene Families, Hallmark and oncogenic gene sets obtained from Molecular Signatures Database [Subramanian et al., 2005, MSigDB]). Top 10 enriched gene sets included per analysis.

Appendix B

Supplementary tables

TOR ⟶	Very Early		Early		Neutral		Late		Very Late	
CpG density ↓	Diff	N	Diff	N	Diff	N	Diff	N	Diff	N
0 - 5	-9.49	1016	-7.70	610	-6.71	337	-3.02	104	-14.95	1849
5 - 10	-9.07	14077	-6.94	9906	-4.77	6532	-4.00	2760	-14.76	19712
10 - 15	-9.17	21826	-6.63	18700	-4.85	14370	-3.79	7468	-15.05	23989
15 - 20	-9.13	48224	-6.34	44653	-4.73	41368	-3.59	26598	-15.06	46273
20 - 25	-8.74	35667	-6.07	35414	-4.44	35726	-3.38	27410	-14.61	32595
25 - 30	-8.15	47105	-5.68	49684	-4.02	53150	-3.05	48707	-13.88	44795
30 - 35	-7.67	25091	-5.38	27303	-3.79	29757	-2.68	30991	-12.67	25849
35 - 40	-7.08	29444	-4.91	30075	-3.44	33591	-2.38	39613	-11.80	32544
40 - 45	-6.32	14150	-4.70	14049	-2.87	15503	-1.61	20257	-10.93	16318
45 - 50	-5.21	14926	-3.50	15041	-2.31	15764	-1.07	22816	-9.44	15960
50 - 55	-3.49	6885	-2.22	6870	-1.44	7329	-0.52	11049	-7.36	7117
55 - 60	-1.57	7810	-0.75	7914	-0.11	8032	0.23	12337	-4.77	7782
60 - 65	0.32	4562	0.70	4322	0.86	4616	1.06	6370	-2.09	4397
65 - 70	1.49	5590	2.32	5621	1.92	5769	1.27	7645	0.41	5298
70 - 75	4.15	3316	3.59	3435	3.35	3475	1.61	4350	3.64	2858
75 - 80	5.67	4781	5.29	4605	4.50	4838	3.11	5483	6.86	3539
80 - 85	6.59	2650	7.00	2942	5.72	2859	3.64	3238	7.88	1888
85 - 90	7.65	3686	7.25	4224	5.48	3865	4.07	4415	9.30	2744
90 - 95	8.04	2140	7.79	2495	5.36	2236	4.23	2670	10.06	1637
95 - 100	8.13	2693	7.90	3285	6.21	3007	5.10	3614	10.34	2093
100 - 105	8.29	1536	7.36	2015	6.12	1715	5.19	2242	11.79	1243
105 - 110	8.34	2105	7.27	2516	6.22	2423	4.71	2777	12.01	1611
110 - 115	7.77	1054	6.83	1422	6.54	1303	5.59	1522	12.19	846
115 - 120	7.34	1515	6.33	2032	5.95	1679	5.14	1920	13.04	1071
120 - 250	6.87	4332	6.66	6343	5.38	5590	5.67	7477	12.98	3077

Table B.1: Details of the 125 CpG density/ time of replication bins. The mean background methylation difference between all tumours and the normal tissues across all CpG sites (Diff), and the number of CpG sites (N) is calculated for each of the 125 CpG density/ Time of Replication bins. CpG density (25 rows) is measured as number of CpGs/ kbp. TOR = Time of Replication. TOR is stratified based on 20th percentiles (5 columns).

Subtype	Direction of regulation	No. of genes		Genes
		All	Directed	
Luminal A	Up	4	4	*C1orf64*[1], *AGR3*[3], *FGD3*[3], *TMEM101*[1]
Luminal A	Down	0	0	
Luminal B	Up	16	12	*NEK2*[4], *AFF3*[3], *C3orf52*[3], *BMPR1B*[3], *TBC1D9*[3]*, *SPDEF*[3], *ESR1*[3], *AGR3*[3], *WWP1*[3], *ASB13*[3], *MSI2*[4], *ZNF552*[3]*, *PARD6B*[3], *PARD6B*[3], *EEF1A2*[2], *EEF1A2*[2]
Luminal B	Down	2	2	*CX3CL1*[1], *SIRPA*[3]
HER2-enriched	Up	5	5	*MTFR1*[3], *IDH2*[4], *ERBB2*[5], *FAM110A*[2], *NCOA3*[3]
HER2-enriched	Down	0	0	
Basal-like	Up	72	53	*STIL*[3], *LRP8*[3], *USP1*[3], *LRRC8D*[3], *C1orf106*[1]*, *LIN9*[4], *HEATR1*[1]*, *CAD*[4], *ATL2*[4]*, *ATL2*[4]*, *MSH6*[4]*, *MCM6*[3], *RPUSD3*[1], *ACTL6A*[4], *LAMP3*[5], *ECE2*[4], *IQCG*[2], *TRIP13*[5], *PITX1*[1]*, *FOXC1*[1], *GCNT2*[13], *GCNT2*[31]*, *NUDT3*[3], *XPO5*[1], *CNKSR3*[3], *NFE2L3*[3]*, *NFE2L3*[3], *NFE2L3*[3], *MGC72080*[3], *ACTR3B*[2], *EIF4EBP1*[32], *EIF4EBP1*[23], *EIF4EBP1*[32]*, *GSDMC*[2], *TOP1MT*[3], *TOMM5*[3], *CALML5*[2], *HSPA14*[4], *HSPA14*[4], *MASTL*[4]*, *RPS24*[4], *CEP55*[3], *ETV6*[3], *DDX11*[3], *RPL6*[5], *TFDP1*[5], *RPP25*[4], *IDH2*[4], *MSLN*[1], *PLK1*[3]*, *CENPN*[2], *SLC7A5*[3], *SLC7A5*[3], *FAM64A*[4], *AFG3L2*[3], *SLMO1*[3]*, *SLMO1*[3], *SEH1L*[3], *SMARCA4*[2], *ASF1B*[4], *DDX39A*[4], *DDX39A*[4], *PDCD2L*[3], *SUPT5H*[1], *CDC25B*[3], *RNF24*[3], *FAM83D*[3], *RBM38*[2]*, *GNAS*[3]*, *GNAS*[3]*, *CHAF1B*[2], *ZBED4*[4]
Basal-like	Down	15	11	*DNALI1*[1], *AFF3*[3], *IRS1*[2], *IRS1*[2], *WFS1*[3], *OLFM1*[3], *GATA3*[2], *C16orf45*[3], *C16orf45*[3]*, *PALM*[3], *PTPRT*[3]*, *PTPRT*[3], *PTPRT*[3]*, *PTPRT*[3]*, *CLIC6*[1]
Claudin-low	Up	34	26	*LAPTM5*[3], *ARHGAP30*[1], *TRAF3IP3*[1]*, *NMI*[3]*, *INPP5D*[23], *INPP5D*[32]*, *INPP5D*[32]*, *STK10*[3], *STK10*[3], *STK10*[3], *STK10*[3], *FAM65B*[13], *FAM65B*[31]*, *FAM65B*[13], *MYO1G*[2], *FERMT3*[1], *TBC1D10C*[2], *PTPRCAP*[1], *P2RY6*[3], *PTPN6*[13], *PTPN6*[31], *LCP1*[3], *PLCB2*[2], *LAT*[1], *PLCG2*[3]*, *ACAP1*[1], *EVI2B*[3], *S1PR4*[12], *S1PR4*[21], *S1PR4*[21], *MCM5*[4], *CYTH4*[3], *PARVG*[3], *P2RY8*[3]
Claudin-low	Down	5	5	*DNALI1*[1], *IRS1*[2], *CMYA5*[1], *MYO5C*[3], *IGF1R*[3]
Normal-like	Up	0	0	
Normal-like	Down	0	0	

Table B.2: Genes harbouring subtype-specific expression-DMRs in the Intrinsic subtypes. For each subtype, genes harbouring expression-DMRs that fulfilled three criteria were identified as follows: i) DMR with a methylation difference of at least 20% vs. normal tissue; ii) DMR with a methylation difference of at least 20% versus other tumours; and iii) DMR is associated with altered expression of the gene: Partial correlation between methylation and expression, $|rho_{meth\text{-}independent} > 0.40$, and *FDR p-value* < 0.05. The superscript details the genomic feature(s) in which the expression-DMR was found. 1 = promoter; 2 = exon; 3 = intron; 4 = enhancer; 5 = PRC region. If more than expression-DMR was found in one gene, then all genomic features were detailed in order of decreasing methylation difference (vs. other tumours). '*' represents Background DMR. Absence of '*' represents Directed-DMR.

Subtype	Direction of Regulation	No. of genes		Genes
		All	Directed	
IntClust 1	Up	5	4	*AFF3*[3], *CCND1*[4], *TMEM104*[2], *ICT1*[2], *TBC1D16*[2]
IntClust 1	Down	8	7	*CNN3*[1], *FAT1*[2], *KHDRBS3*[5]*, *ASS1*[3], *CX3CL1*[1], *SOX9*[2], *NFIX*[3], *NFIX*[3]
IntClust 2	Up	2	2	*PPFIA1*[1], *SLC16A6*[3]
IntClust 2	Down	1	1	*SFRP1*[1]
IntClust 3	Up	0	0	
IntClust 3	Down	0	0	
IntClust 4ER-	Up	31	31	*ENO1*[3], *LAPTM5*[3], *ARHGAP30*[1], *INPP5D*[23], *INPP5D*[32], *INPP5D*[32], *STK10*[3], *STK10*[3], *STK10*[3], *STK10*[3], *FAM65B*[1], *FAM65B*[1], *TBC1D10C*[12], *TBC1D10C*[21], *PTPRCAP*[1], *PTPRCAP*[1], *P2RY6*[3], *PTPN6*[13], *PTPN6*[31], *ARHGAP9*[3], *LCP1*[3], *IGSF6*[2], *LAT*[1], *ACAP1*[1], *EVI2B*[3], *S1PR4*[12], *S1PR4*[21], *S1PR4*[21], *ITGB2*[3], *PARVG*[3], *P2RY8*[3]
IntClust 4ER-	Down	5	5	*IRS1*[2], *CMYA5*[1], *GATA3*[2], *GATA3*[2], *IGF1R*[3]
IntClust 4ER+	Up	0	0	
IntClust 4ER+	Down	0	0	
IntClust 5	Up	8	7	*ENO1*[3]*, *MTFR1*[3], *IDH2*[4], *PSMB3*[4], *ERBB2*[5], *ERBB2*[5], *ERBB2*[5], *FAM110A*[2]
IntClust 5	Down	0	0	
IntClust 6	Up	9	8	*CELSR2*[2], *AFF3*[3], *ESR1*[3], *RAB11FIP1*[2], *WHSC1L1*[1], *CCND1*[4], *RERG*[3], *ZFHX3*[3], *SLC2A4RG*[2]
IntClust 6	Down	1	1	*NFIX*[3]
IntClust 7	Up	1	1	*TMEM101*[1]
IntClust 7	Down	0	0	
IntClust 8	Up	19	18	*FMO5*[3], *MUC1*[2], *MAPKAPK2*[3], *CAPN8*[2], *CAPN9*[3], *SIAH2*[3], *BMPR1B*[3], *TBC1D9*[3]*, *AGR3*[3], *PSD3*[3], *FGD3*[3], *ANKRD30A*[3], *GRTP1*[2], *SREBF1*[3], *TMEM101*[1], *KDM4B*[3], *KDM4B*[3], *ZNF552*[3], *TFF1*[1]
IntClust 8	Down	0	0	
IntClust 9	Up	1	1	*RAMP1*[3]
IntClust 9	Down	0	0	
IntClust 10	Up	55	42	*LAPTM5*[3], *STIL*[3], *LRP8*[3], *USP1*[3], *LRRC8D*[3], *C1orf106*[1]*, *FBXO28*[4], *CAD*[4], *ATL2*[4]*, *ATL2*[4]*, *OTX1*[3], *MCM6*[3], *NMI*[3]*, *RPUSD3*[1], *LAMP3*[5], *IQCG*[2], *TRIP13*[5], *PITX1*[1]*, *FOXC1*[1], *GCNT2*[1], *XPO5*[1], *MB21D1*[3], *SNX8*[3], *NFE2L3*[3]*, *NFE2L3*[3], *MGC72080*[3], *ACTR3B*[2], *EIF4EBP1*[3], *GSDMC*[2], *TOP1MT*[3], *CALML5*[2], *HSPA14*[4], *RPP38*[4], *MASTL*[4]*, *RPS24*[4], *CEP55*[3], *DDX11*[3], *TFDP1*[5], *PLK1*[3]*, *CENPN*[2], *SLC7A5*[3], *FAM64A*[4], *MRPL12*[4], *AFG3L2*[3], *SLMO1*[3]*, *SLMO1*[3], *DDX39A*[4], *DDX39A*[4], *DDX39A*[4], *PDCD2L*[3], *CDC25B*[3], *FAM83D*[3], *RBM38*[2]*, *TYMP*[4]*, *TYMP*[4]*
IntClust 10	Down	15	13	*DNALI1*[1], *RHOB*[4], *AFF3*[3], *IRS1*[2], *IRS1*[2], *WFS1*[3], *TFAP2B*[2], *GATA3*[2], *ARRB1*[3], *C16orf45*[3], *PALM*[3], *PTPRT*[3]*, *PTPRT*[3], *PTPRT*[3]*, *CLIC6*[1]

Table B.3: (Caption on next page.)

Table B.3: (Previous page.) **Genes harbouring subtype-specific expression-DMRs in the Integrative clusters.** For each subtype, genes harbouring expression-DMRs that fulfilled three criteria were identified as follows: i) DMR with a methylation difference of at least 20% vs. normal tissue; ii) DMR with a methylation difference of at least 20% versus other tumours; and iii) DMR is associated with altered expression of the gene: Partial correlation between methylation and expression, $|rho_{meth\text{-}independent}| > 0.40$, and *FDR p-value* < 0.05. The superscript details the genomic feature(s) in which the expression-DMR was found. 1 = promoter; 2 = exon; 3 = intron; 4 = enhancer; 5 = PRC region. If more than expression-DMR was found in one gene, then all genomic features were detailed in order of decreasing methylation difference (vs. other tumours). '*' represents Background DMR. Absence of '*' represents Directed-DMR.