# Copy-number aware methylation deconvolution analysis of cancers

*Elizabeth Larose Cadieux*

University College London

and

The Francis Crick Institute

A thesis submitted for the degree of

Doctor of Philosophy

University College London

PhD Supervisors: Dr. Peter Van Loo and Prof. Stephan Beck

March 28, 2021

# Declaration

I, Elizabeth Larose Cadieux, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

# Abstract

DNA methylation has long been known to play a role in tumourigenesis. To this day, interpretation of bulk tumour bisulphite sequencing data has been hampered by normal contamination levels and tumour copy number. To address this issue, we develop two computational tools: (1) ASCAT.m, which allows Allele-Specific Copy number Analysis of Tumour methylation data directly from bulk tumour reduced representation bisulphite sequencing (RRBS) data and (2) CAMDAC, a method for Copy Number-Aware Methylation Deconvolution Analysis of Cancer, from bulk tumour and adjacent normal RRBS data.

We describe a set of rules to compute allelic imbalance independently of bisulphite conversion and correct normalised read coverage estimates for sequencing biases. We apply ASCAT.m to non-small cell lung cancers from the epiTRACERx study with multi-region bulk tumour RRBS and adjacent normal. ASCAT.m genotypes, allele-specific copy numbers and tumour purity and ploidy estimates are in excellent agreement with those obtained from matched whole-exome and a subset of whole-genome sequencing of the same samples. We observe a correlation between whole-genome doubling and relapse-free survival in lung squamous cell carcinoma but not in adenocarcinoma. We see widespread genomic instability across both histological subtypes.

We model bulk tumour methylation rates as a mixture of tumour and normal signals weighed for tumour purity and copy number and formalise this relationship into CAMDAC equations. The errors between predicted and observed methylation rates were low. Normal infiltrates Fluorescence-activated cell sorting (FACS)-purified from the bulk tumour were similar in composition to the adjacent matched

normal lung, suggesting the latter is a suitable proxy for deconvolution. Single nucleotide variant (SNV)- and FACS-purified tumour methylation rates are in good agreement with CAMDAC deconvoluted estimates. Purification successfully removes shared normal signal, decreasing correlations between patients and to normal after purification. Samples with shared ancestry remain highly correlated.

Purified methylation rates yield accurate tumour-normal and tumour-tumour differential methylation calls independent of tumour purity and copy number. We find hundreds of ubiquitously early clonal gene promoter epimutations across the epiTRACERx cohort, showcasing the potential of DNA methylation markers for early cancer detection. CAMDAC purified profiles reveal both phylogenetic and inter-tumour relationships as well as provide insight in tumour evolutionary history.

Quantifying allele-specific methylation on chromosome X in females, we uncover extraction biases against the Barr body. X inactivation is random at the scale of our normal lung cancer samples. Phasing of methylation rates with polymorphisms confirms the presence of allele-specific methylation at the *H19/IGF2* locus. Loss of imprinting is observed in 5 tumours, all involving demethylation of the maternal allele. We attempt to quantify the ratio of clonal allele-specific to bi-allelic epimutations in tumours in regions of $1+1$, which we define as regulatory and stochastic methylation changes, respectively. Utilising this ratio, we try to extract the number of stochastic epimutations in regions of $2+0$ with copy numbers 1 and 2 and time those copy number gains.

We find that SNVs at gene promoters often lead to hypermethylation of neighbouring CpGs on the same copy or allele, suggesting the ablation of a transcription factor binding site. Non-expressed neo-antigen are enriched for promoter hypermethylation, indicating methylation plays a role in immune escape.

To conclude, CAMDAC purified methylation rates are key to unlock insights into comparative cancer epigenomics and intra-tumour epigenetic heterogeneity.

# Impact statement

In this work, we present a cohort of 38 non-small cell lung cancer patients with multi-sample reduced representation bisulphite sequencing, totalling 122 tumour and 37 adjacent normal samples. Lung cancer causes the largest proportion of cancer-related death and this dataset could provide unprecedented insights into the non-small cell lung cancer methylome. However, bulk tumour methylation sequencing data is convoluted by normal cell contamination, tumour purity and copy number. This is particularly relevant in lung cancers, which have lower purities than most other cancer types on average.

To address this issue, we first develop a computational method for obtaining allele-specific copy numbers and tumour purity estimates directly from RRBS data, ASCAT.m. For the first time, we formalise the relationship between methylation rates, tumour purity and copy number into the core CAMDAC equations.

Crucially, only CAMDAC purified methylation profiles enable accurate differential methylation analysis, and as such, we find hundreds of early clonal promoter epimutations present in virtually all non-small cell lung cancer samples, showcasing the immense potential DNA methylation sequencing data for diagnostic purposes. Early detection has significant implications on patient outcome in non-small cell lung cancer, with the 5-year survival decreasing from 70% to below 15% between cases diagnosed in stage I *versus* II and above, respectively. We discuss plans to apply CAMDAC to more samples from the TRACERx study recently sent for sequencing to gain deeper insight into methylation biomarkers of non-small cell lung cancer.

Throughout this work, we showcase the possibilities unlocked by tumour RRBS data. We demonstrate that CAMDAC purified profiles reveal both phylogenetic and inter-tumour relationships as well as provide insight in tumour evolutionary history. We show that it is possible to study the interplay between somatic mutations and epimutations and namely that DNA methylation plays a role in suppressing neo-antigen presentation.

# Acknowledgements

I would like to begin by thanking everyone who has contributed to the work presented in this thesis. Firstly, I would like to thank my primary and secondary supervisors, Peter and Stephan, who gave me not one but two labs to call a home. Jonas has, in many ways, acted as a third supervisor providing not only scientific advise but also moral support. I am very thankful for all your help and for your time! Working with you three, I discovered my passion for cancer epigenetics and I will be forever grateful for that. Thank you the Van Loo and Beck lab members, especially Jonas, Clem, Kerstin, Dr. Matt, Max, Annelien, Tom, Haixi, Toby, Ismail, Simone, Olga, Amy, Allison, Iben, Ric, Ron and Teresa with whom I shared most of my PhD. Every single one of you helped to make my PhD a truly special experience.

I am incredibly grateful to Gareth, Andy, Miljana, Stephan and Charlie for getting Peter, Jonas and myself involved in the epiTRACERx study. Specifically, I would like to thank you all for entrusting me with the development of methods for the analysis of these data. Thank you to all members of the TRACERx lung cancer consortium for providing whole-exome sequencing analyses as well as detailed clinical information. This dataset is the result of an enormous collaborative effort between hundreds of researchers, clinicians and patients across the UK. Special thanks to Nicky and Rachel for a fruitful collaboration, investigating the role of DNA methylation in anti-tumour immune response.

Finally, I would like to take the time to acknowledge my friends, Tom, Tea, Zuzana, Flavia, Julia, Harveen, Isabel, Kelsey, Philippa and Teddy, for their continued support throughout my PhD and for always believing in me, you guys are incredible and I could not ask for better friends. I would also like to thank my fam-

ily in Canada and my in-laws, here in the UK, for their continued support. Special thanks to my Mum and Dad for managing to take care of me even at a distance, through daily video chats and messages. Thanks to my brother, Phil, my sister-in-law, Katherine, and my niece Charlotte, for bringing a smile to my face every time you video called. Louis, you are my rock, thank you for taking care of everything during these past few weeks enabling me to focus completely on my thesis. Loki the cat, thanks for being you, stay cheeky!

# Contents

# List of Figures

# List of Tables

# List of abbreviations

| | |
|---|---|
| **ASCAT** | Allele-Specific Copy Number Analysis of Tumour |
| **ASCAT.m** | Allele-Specific Copy Number Analysis of Tumour Methylation data |
| **ASM** | Allele-Specific Methylation |
| **CAMDAC** | Copy Number-Aware Methylation Deconvolution Analysis of Cancer |
| **DNA** | DeoxyriboNucleicAcid |
| **DMP** | Differentially Methylated Position |
| **DMR** | Differentially Methylated Region |
| **FACS** | Fluorescent-Activated Cell Sorting |
| **ICR** | Imprinting Control Region |
| **LOH** | Loss of Heterozygosity |
| **LUAD** | LUng ADenocarcinoma |
| **LUSC** | LUng Squamous cell Carcinoma |
| **NSCLC** | Non-Small Cell Lung Cancer |
| **RRBS** | Reduced Representation Bisulphite Sequencing |
| **SNP** | Single Nucleotide Polymorphism |
| **SNV** | Single Nucleotide Variant |
| **TFBS** | Transcription Factor Binding Site |
| **WES** | Whole-Exome Sequencing |
| **WGD** | Whole-Genome Doubling |
| **WGS** | Whole-Genome Sequencing |

# Chapter 1

# Introduction to the cancer methylome

## 1.1 DNA methylation and transcription

Epigenetics is the study of inheritable DNA modifications that allow cells to acquire specialised phenotypes without altering their DNA sequence. There are two types of epigenetic inheritance: mitotic inheritance, which is well studied and applies to this thesis, and meiotic inheritance, which is less well studied and controversial in humans [1]. DNA methylation is an important epigenetic mark. Six different methylation modifications are currently known of which covalent modification at carbon 5 of cytosine (C) is the most common, resulting in 5-methylcytosine (5mC) [2]. Cytosine methylation was first reported in the tuberculosis bacterium [3]. Decades later, interest in DNA methylation spiked after 5mC was observed in mammalian cells [4] and evidence of its role in gene regulation was uncovered, namely thanks to the observation of methylation-driven chromosome X inactivation in females [5–7].

Indeed, the importance of methylation in mammalian gene regulation is particularly striking when looking at X inactivation, the process by which gene dosage compensation is achieved in females involving widespread methylation of one chromosome X copy [8]. On a smaller scale, genomic imprinting also leads to allele-specific expression, by methylation of one parental allele at a number of genes loci across autosomes [9, 10]. Loss of imprinting is associated with several diseases, including cancer [11].

Specifically, methylation regulates gene expression by either (1) recruiting methyl-CpG binding domain proteins (*MBD*s) which themselves promote histone deacetylases (*HDAC*s) activity leading to a closed chromatin state [12] and late replication timing [13, 14], or, (2) through increased steric hindrance at transcription factor binding sites (TFBSs, [15], **Figure 1.1**). The former mechanism is widespread in the human genome despite requiring several epigenetic modalities working in synergy. Early experiments showed that genes methylated *in vitro* were not immediately silenced by methylation but rather indirectly inhibited after a short period of time during which the genomic region adopted a closed chromatin structure [16, 17]. The second mechanism is thought to be more dynamic and reportedly gives rise to tissue-specific, cell state- and, perhaps unsurprisingly, disease-specific methylation.

While methylation usually leads to gene silencing, it is worth noting that demethylation of promoters does not necessarily equate to increased expression of associated genes. Other factors must also be met to induce gene expression such as availability of the necessary transcription factors, nucleosome depleted TFBSs and active transcription marks on neighbouring histones such as H3 and H4 acylation and H3K4 (tri)methylation [18]. Linker histone H1 (not depicted) is less well-studied but is thought to play an important role in modulating chromatin accessibility [19].

## 1.2 Measuring genome-wide CpG methylation

In adult mammalian cells, DNA methylation occurs almost exclusively in CpG context [20]. The first approach to measure DNA methylation levels took advantage of CpG methylation-sensitive restriction enzymes [21]. This technique enabled researchers to study genome-wide methylation levels, albeit limited to CpG loci at the enzyme recognition sequences. With the advent of next generation sequencing, whole genome bisulphite sequencing (WGBS) was developed and allowed assessment of methylation levels at every CpG and at single base pair resolution [22]. In WGBS, input DNA is typically treated with sodium bisulphite prior to sequenc-

ing, oxidising unmethylated Cs to uracil (U) with high conversion rate (under-conversion rate $r_{uc}$ <1%) and leaving 5mC virtually unscathed (over-conversion rate $r_{oc}$ ∼4%, **Figure 1.2**) [23–25].

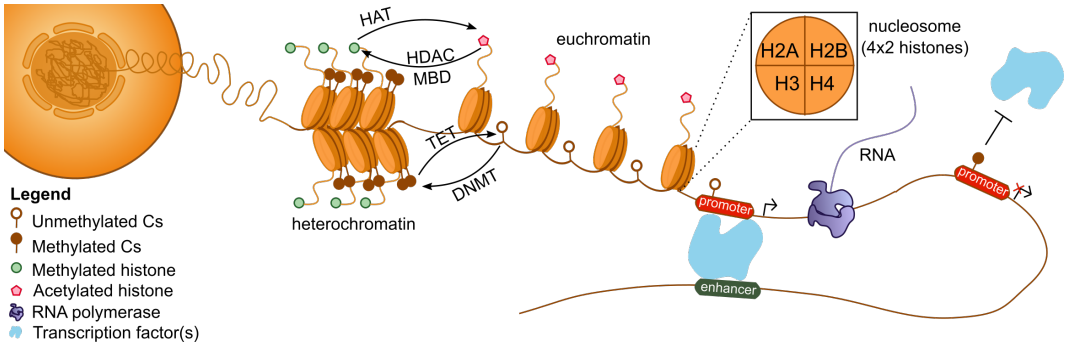


**Figure 1.1: Epigenetic marks determine chromatin architecture and regulate transcription.**
Eukaryotic DNA resides in the nucleus and exists in the form of chromatin. In the human genome, DNA is condensed into 23 diploid chromosomes, 22 autosomes and one of each X and Y sex chromosomes in males while females have 2 copies of X and none of Y. Within chromosomes, DNA is wrapped around nucleosomes, each made up of 8 histone sub-units, 2 of each H2A, H2B, H3 and H4. Methylation of cytosines and of certain histone residues, usually tri-methylation of H3K9 or H3K27, promote a repressive heterochromatin state. In comparison, acetylation of these same residues combined with DNA demethylation enables binding of transcription factors and, if the latter are available, may result in RNA polymerase activity. Transcription may be stimulated by cis-regulatory elements such as nearby enhancers.

After bisulphite conversion of unmethylated Cs into Us, their guanine (G) double-strand partners are left unchanged, which means the forward and reverse strands are no longer complementary. Post-polymerase chain reaction (PCR), four possible products are therefore obtained, one for each of the original strands and their new complements, with Us converted into thymines (T). At this stage, an unmethylated CpG locus could generate four different dinucleotides read-outs from either the original top, original bottom or their new complements, TG(+), GT(-), AC(-) and CA(+), respectively (**Figure 1.2**). Base read-outs are reported in terms of the forward strand and so GT(-) is catalogued as CA(-). At methylated CpG sites, cytosines are unaltered by bisulphite conversion and so only CG(+) and CG(-) are generated. Methylated read adaptors are often used to generate directional libraries as they enable selective sequencing of both the original strands while

discarding the complements. Dinucleotide outputs from directional bisulphite sequencing at unmethylated or methylated CpGs will be TG(+) and CA(-) or CG on both strands respectively. Assuming the absence of heterozygous single nucleotide polymorphisms or variants, the methylation rate is easily computed by compiling dinucelotide counts, dividing CG counts by the total number of reads supporting any of the four methylation informative dinucleotides (i.e. TG(+), CA(-), CG, [26]).

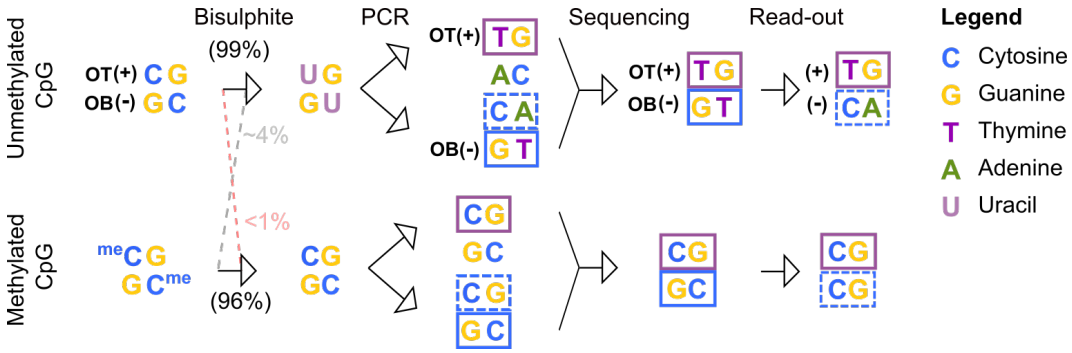


**Figure 1.2: Directional bisulphite sequencing dinucleotide read-out at CpGs.** Expected dinucleotide products from directional bisulphite sequencing data at unmethylated (top) and methylated (bottom) CpGs. Bisulphite conversion oxidises unmethylated Cs into Us leaving its methylated counterpart unaltered. Over- and under-conversion rates may lead to erroneous read counts and are depicted as a grey or red dashed line, respectively, each accompanied by reported estimates. At unmethylated CpGs, bisulphite conversion leads to base pair mismatch which results in four different PCR products, the original top (OT+), original bottom (OB-), and their complements. Thanks to directional sequencing adapters, only the original strands are sequenced. OT+ dinucleotides are reported as is whilst the OB- bases are catalogued as their reverse complement, (i.e. GT- becomes CA- and GC- becomes CG+).

Developed a decade or so later, methylation microarrays quickly rose in popularity as a considerably cheaper alternative to WGBS. As part of this protocol, bisulphite treated and PCR-amplified DNA is hybridised onto arrayed oligonucleotides probes each marking a regions of interest for a subset of CpGs, leveraging fluorescent labels that can discriminate between C and 5mC alleles [27, 28]. Bisulphite sequencing approaches have become the gold standard, but methylation arrays are still in use to this day. Despite plummeting sequencing costs in recent years, the shift from microarrays to sequencing has been slow, namely due to lack of availability of bioinformatics analysis tools compatible with bisulphite sequencing data.

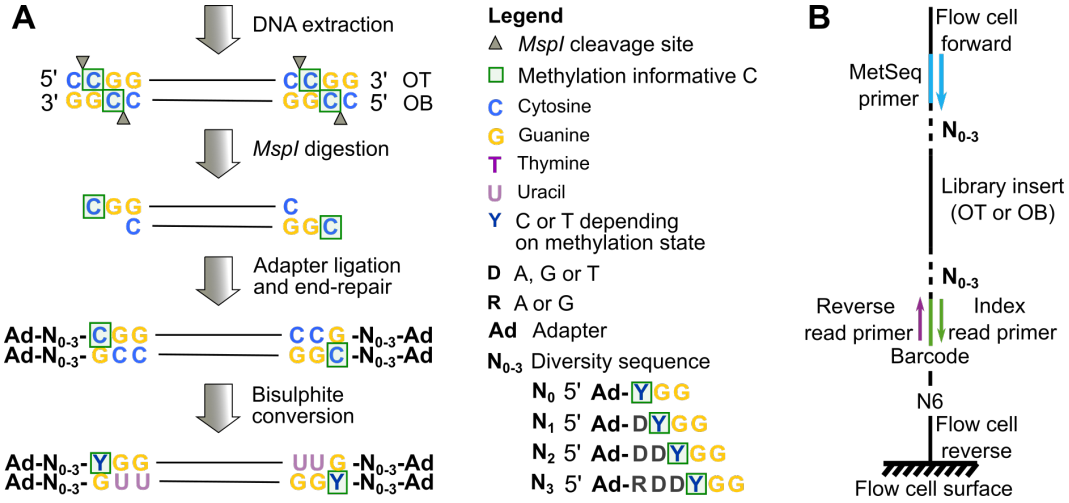


**Figure 1.3: NuGEN Ovation RRBS protocol.**
(A) DNA extraction is followed by *MspI* digestion leaving a 2bp overhang. Only the 5′ CCGG remains methylation informative after fragment end-repair. Note that a sequence of 0 to 3 bases is added to each library molecule in order to avoid clustering issues during sequencing. Adapter ligation and bisulphite conversion follow. Library amplification and sequencing is implied but not shown here. (B) NuGEN Ovation libraries include an integrated molecular tag (N6) that enables removal of duplicate reads from the dataset after PCR and barcode enables multiplexing up to 16 different samples per flow cell, both of which are 6 nucleotide in length. A custom forward primer enables directional sequencing of the original top and bottom strands while standard Illumina sequencing primers are used for the reverse and index reads.

Reduced representation bisulphite sequencing (RRBS) [29–32] is raising in popularity compared with WGBS since it requires smaller input quantities (10-300ng versus 5$\mu$g) and has lower sequencing costs ($\geq$10 million versus $>$500 million reads) [31, 33]. RRBS gives a read out of $\sim$2 million CpGs, more than twice the number covered by the most recent Illumina EPIC array (Reviewed in [34]). RRBS relies on restriction enzymes such as *MspI* (CˆCGG recognition motif) to digest DNA into fragments that are enriched for CpG dinucleotides (**Figure 1.3**). Contrary to whole-genome bisulphite sequencing (WGBS) [35], every fragment produced by *MspI* digestion contains information for at least one CpG per single end read. RRBS has been used to identify novel parental imprinting loci [36], to study embryogenesis [37, 38], cell differentiation [39] or tumour biology [40]. In addition, reference DNA methylation profiles for a variety of cell types have been constructed from RRBS data [41].

## 1.3 CpG Island methylation and gene regulation

Bisulphite sequencing and microarray technologies have enabled researchers to gain a deeper understanding of genome-wide DNA methylation patterns. In humans, 70-80% of CpGs motifs are methylated, most of which are scattered across the repetitive genome [42, 43]. A large fraction of the remaining CpGs are found in high density clusters called CpG Islands (CGIs), shores ($\pm$2kb around islands) and shelves ($\pm$3kb around shores). CGIs often overlap gene promoters and enhancers and are usually unmethylated irrespective of gene expression levels [44]. As previously mentioned, several other factors must be met, in addition to demethylation, to activate gene expression. However, methylation of promoter- or enhancer-associated CGIs, either on its own or in concert with repressive histone modifications, is correlated with gene silencing. Interestingly, housekeeping gene promoters are enriched for CGIs [45]. Because *MspI* digestion is biased for these CpG-rich regulatory regions [29, 46] and also targets a number of CGI shores, exons, $3'$ and $5'$ untranslated regions (UTRs) and repetitive elements [47], RRBS is particularly well suited to study methylation with respect to gene regulation.

## 1.4 DNA methylation machinery: introducing key enzymes

Whilst the methylome of a healthy differentiated cell is relatively stable, it undergoes complete reprogramming in early embryogenesis [37, 38, 48, 49]. After genome-wide chromatin activation, DNA methylation and histone repressive marks both need to be re-established quickly, particularly in the repetitive genome to keep transposable elements (TEs) in check and protect cells against hijacking by endogenous viral DNA [50, 51]. This is crucial as 45-50% of the human genome encodes TEs such as endogenous retroviruses (ERVs), long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) [52,53]. Mutations affecting the DNA methylation machinery have been observed in cancer and are known to disrupts normal epigenetic programmes.

DNA methyltransferases (DNMTs) are a class of enzymes responsible for DNA methylation maintenance post-replication as well as during *de novo* methylation. *DNMT1* is the main enzyme tasked with re-establishing symmetrical methylation at hemimethylated CpGs following DNA replication in S-Phase [54]. *DNMT3C* protects sperm DNA from retrotransposons activity by methylating repeats irrespective of the germline or offspring methylation state [55]. *DNMT3A* and *B* are best known for their role in *de novo* methylation namely during embryonic development. They are the principal enzymes regulating CpG (re)methylation after widespread genome activation.

Epigenetic reprogramming is a tightly regulated multilayered process involving methylation. DNA methylation is implicated in repression of transposable element in later stages of embryogenesis whilst repressive histone post-translational modifications are necessary for early silencing as shown *in vitro* in *DNMT3A* and *B* double knockouts embryonic stem cells [56]. In the newly fertilised mammalian embryo, retrotransposons are silenced by KRAB-Zinc Finger Proteins (KZFP) that bind to methylated DNA thereby recruiting a repressive chromatin complex via *KAP1/TRIM28* interaction which ultimately results in both cytosine and histone H3K9 trimethylation [57]. Loss of methylation is prevented at imprinted loci through the same KZFP-dependent mechanism. In contrast, variably methylated Intracisternal A Particle Long Terminal Repeats (IAP LTRs) are not perpetuated across generations but rather through context dependent *de novo* methylation by *DNMTA/B*, such as CTCF binding site proximity [58]. The methylation level of IAP LTR harbouring promoter regions can influence the expression of nearby genes. For example, IAP methylation leads to $A_{vy}$ (agouti viable-yellow) ectopic gene expression in the mouse and visible phenotypic consequences ensue, in the form of variable coat colour.

*In vitro* experiments have shown that *DNMT1* has high affinity for hemimethylated DNA successfully converting $\sim$99.7% of hemimethylated loci. Nevertheless, this means roughly 3 in 1000 methylated CpG dinucleotides will lose methylation after each cell division [59]. *DNMT1* errors rarely result in gain of methylation, but

when they do, a preference for methylation at CCGG motifs is observed, at least in healthy cells. The error rate as measured between generations in healthy plant cells showed a 3-fold higher loss than gain of methylation with respect to the founder generation, $2.56 \times 10^{-4}$ and $6.30 \times 10^{-4}$ respectively [60]. Gain of methylation was 25 times more likely than demethylation at transposable elements, suggesting either selection forces at play or an alternative methylation maintenance mechanism for the repetitive genome.

Ten-eleven translocation (*TET*) family of dioxygenases (*TET1*, *2* and *3*) play a role opposite to that of *DNMT*s (**Figure 1.4**), catalysing the step-wise oxidation of 5mC into 5-hydroxymethylcytosine (5hmC) [61], 5-formylcytosine (5fC) and subsequently 5-carboxylcytosine (5caC) [62, 63]. These oxidised 5mC derivatives are no longer recognised by *DNMT1* leading to loss of methylation upon DNA replication [64]. In some cases, 5mC sites are missed by the methylase creating a transient hemi-methylated loci which, in the absence of further errors, should regain symmetrical methylation following replication in the next cell cycle. Demethylation may also occur through thymine DNA glycosylase (*TDG*)-mediated base excision repair [65]. *In vitro* experiments suggest that *DNMT*s can catalyse the dehydroxymethylation 5hmC and decarboxylation of 5caC into C [66].

## 1.5 Cancer, a disease of the (epi)genome

Epigenetic mutations or 'epimutations', including aberrant DNA methylation and chromatin architecture alterations, are now acknowledged as a universal feature of cancer development [67–69]. Healthy cells accumulate (epi)mutations throughout their lifetime [70–72] and while most have no effect, a small subset may provide a selective advantage for the cell [73]. A cell may eventually acquire a fully malignant phenotype following successive gains of hallmark cancer cellular capabilities [74], while continuing to evolve in response to environmental pressures [75].

The cancer methylome will display characteristic of its cell of origin, with varying degree of somatic DNA methylation changes [76, 77]. Global hypomethylation has long been known to occur in cancer cells [78], destabilising the genome

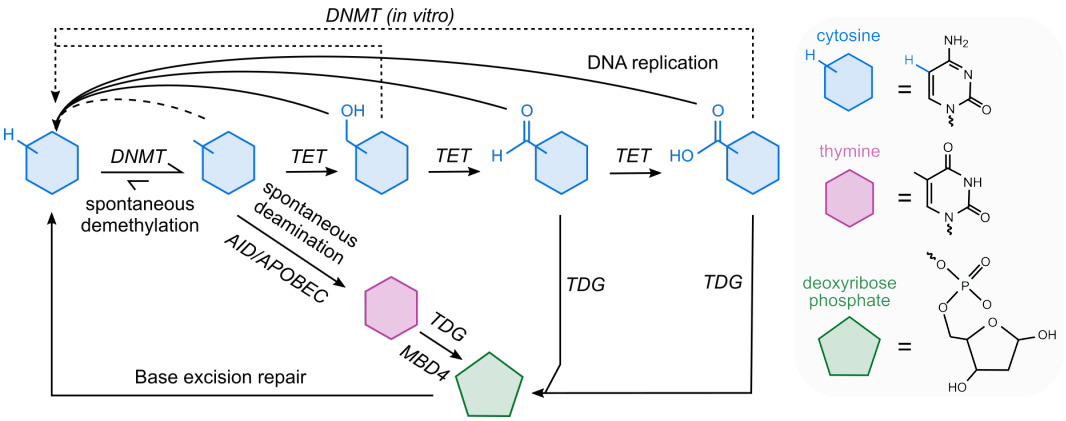


**Figure 1.4: An overview of cytosine modifications in mammals.**
Cytosine nucleotides can be methylated at the 5′ carbon position on the pyrimidine ring. This modification is catalysed by *DNMT*s. Passive demethylation of methylated cytosines is a slow process but is known to accumulate with ageing and can result in loss of methylation. In addition, *TET*-assisted oxidation of 5mC into 5hmC, 5fC or 5caC leads to cytosine demethylation upon replication, *DNMT*-driven decarboxylation or through *TDG*-mediated base excision repair.

namely by reactivation of the repetitive genome [79]. Hypomethylation at centromeres specifically favours aneuploidy [80]. Recent reports suggest these effects are the results of long range hypomethylation blocks rather than individual hypomethylated CpGs, as described in colon cancer [81]. Aberrant gain of methylation usually operates on a smaller scale, silencing individual promoter- or enhancer-associated CGIs. Gene promoter hypermethylation-driven tumour suppressor deactivation was first reported in cancer cells at the *VHL* and *CDKN2/p16/MST1* locus [82–84] and has since been reported at a number of genes across cancer types. Hence, DNA methylation profiles can provide useful information on (disease) cell states and could become powerful biomarkers [85–87].

The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have revealed recurrent somatic aberrations and their clonality for the majority of known cancer types [88, 89]. In comparison, the cancer methylome is considerably less well charted. This is likely due to complexity of interpreting bulk tumour methylation data, which is confounded by somatic copy number alterations (CNAs) and admixed normal cells, and the lack of computational methods to correct for these effects in downstream tumour differential methylation analyses.

Tumour-normal differential methylation can arise due to (1) error-prone DNMTs activity, (2) aberrant transcription factor activity, (3) spontaneous deamination of methylated CpGs and (4) coupled TET- and TDG-mediated demethylation of 5mC oxidation products into apyrimidinic sites [65] in cancer cells with deficient base excision repair [90]. Stochastic differentially methylated positions (DMPs) and CG-destroying/-forming single nucleotide variants (SNVs) should be inherited unless back-mutated. Altering the tumour genetic sequence is non-reversible under the infinite sites assumptions, which states that the human genome can be considered infinite and thus the probability of a given base being mutated twice is zero. Infinite site assumptions violation should be rare, albeit more common than for point mutations due to relatively higher (epi)mutation rate. Regulatory-driven differentially methylated regions (DMRs) are likely to be dynamically methylated in response to signalling, also violating the assumption and further complicating tumour phylogeny reconstruction.

Popular DMP and DMR calling methods for bisulphite sequencing data have been reviewed in recent articles by either Hebestreit and Klein [91] and Robinson *et al.* [92]. Methylation rates at C>T mutations in CpG context are confounded by the variant allele, which is indistinguishable from the bisulphite-converted unmethylated base. We note that C>T SNVs are enriched at CpGs, both in cancer and normal cells [93].

## 1.6 Sources of intermediate methylation in bulk tumour data

Cancer cells are constantly evolving in response to selective pressures from the environment in which they exist, fuelled by the activity of various mutational processes [94]. Darwinian selection and clonal expansion of the fittest cells creates subclones and moulds the landscape of intra-tumour heterogeneity (ITH, **Figure 1.5**, left). Subclonal epimutations can be detected given they are present in a large enough cancer cell fraction (CCF) and depending on tumour purity and copy number. Cell-type heterogeneity within the admixed normal cell populations can

also give rise intermediate tumour DNA methylation levels due to cell type specific methylation (**Figure 1.5**, middle left [95–97]), although this effect is presumably limited at sufficiently high tumour DNA content. Potential sources of intermediate methylation also include allele-specific methylation at germline imprinted loci or allele-specific somatic DNA methylation alterations (**Figure 1.5**, middle right). Lastly, DNA methylation erosion can lead to intermediate methylation (**Figure 1.5**, right). This occurs as part of healthy ageing, but the process is accelerated in rapidly replicating cancer cells.

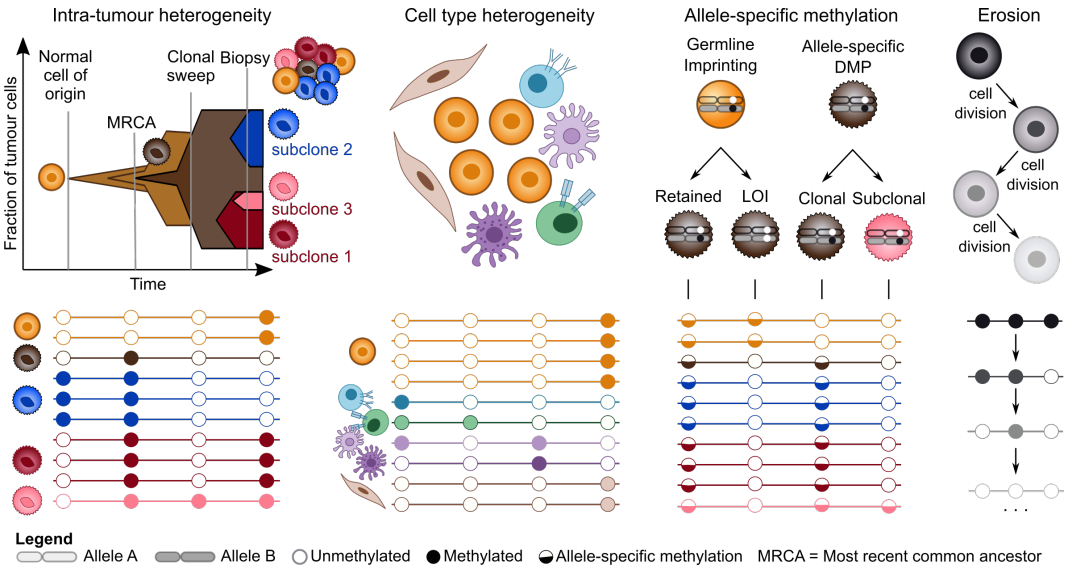


**Figure 1.5: Potential sources of intermediate DNA methylation levels.**
Sources of intermediate methylation (top) and example bulk methylation rate distribution (bottom). Circles indicate the methylation state at a given loci formed of one or more neighbouring CpGs. This could represent an individual intragenic CpG or a CGI.

## 1.7 Capturing DNA methylation intra-tumour heterogeneity

DNA methylation ITH has been reported by independent research laboratories [98–100], but its role in tumour evolution and its impact on patient outcome both remain unclear. At present, methods to quantify ITH from bisulphite sequencing data [40, 101–103] assume low normal contamination levels and/or the absence of copy number alterations.

Heterogeneity metrics are reviewed in Scherer *et al.* [104] including the popular proportion of discordant reads (PDR [40]), a measure of intra-molecular heterogeneity. Individual reads are classified as either concordant (i.e. fully methylated or unmethylated) or discordant (i.e. partially methylated) and the PDR is calculated as $PDR = \frac{countsdiscordant}{countsdiscordant + countsconcordant}$. This method relies on the assumption that neighbouring CpGs normally display concordant methylation which is lost either due to DMPs or DNA methylation erosion in tumours. Alternatively, one can compute the methylated haplotype load (MHL, [105]), a weighted mean of the fraction of fully methylated haplotypes and its substrings of 2 of more CpG loci. Subclonal differential methylation need not be locally discordant to be captured by the MHL from bulk data. In any case, both scores are affected by variations in tumour purity and copy number.

Although less well-known, the Bayesian epiallele detection (BED) approach seems a promising method to evaluate methylation ITH [106]. An epiallele is defined as a sequence of $n$ CpG sites on one read molecule for which there are $2^n$ possible methylation patterns. Epialleles may be compared between overlapping reads and epiallele frequencies can be calculated. BED estimates the underlying number of epialleles using the Bayes information criterion (BIC), accounting for experimental noise and preventing the model from inferring too many epialleles unless the evidence is sufficiently strong. Epialleles obtained from tissue-matched normal can be used as a proxy for the normal contaminating cells and based on this the authors determine tumour purity and extract purified tumour epialleles. A more accurate method would require consideration of both tumour purity and copy number.

## 1.8 Reconstructing tumour phylogenies

When reconstructing tumour evolutionary histories, accurate estimates of tumour purity and copy number are critical [107]. First, the ratio of cancer cells carrying a mutation, referred to as the cancer cell fraction (CCF), is extracted from the variant allele frequencies (VAF) taking care to correct for local copy number

and the aberrant cell fraction. In turn, we can use the distribution of mutations in CCF space in a given tumour sample to infer the underlying subclonal architecture. This distribution usually has a peak at $CCF = 1$, which corresponds to the most recent common ancestor shared by all tumour cells while peaks at lower CCF are attributed to subclonal populations of cells. Clustering of subclonal mutations is possible with various algorithms. Individual mutations are assigned a probabilistic cluster assignments from which phylogenetic trees can be inferred. Subclobal reconstruction algorithms include but are not limited to DPClust [107], PhylogicNDT [108], PhyloWGS [109], PyClone [110]. These and 7 other methods are described and compared in Dentro *et al.* [111]. Whilst these fantastic tools enable researchers to uncover the genetic evolutionary history of tumours from subclonal copy number and single nucleotide variant information, they offer no epigenomic information and are not compatible with methylation data. Bisulphite sequencing experiments yield genomic and epigenomic data and therefore have the potential to deepen our understand of the interplay between these two components throughout tumour evolution.

## 1.9 Multi-sample studies

While tumour subclones present at large enough cancer cell fractions can be detected within single biopsies as above-described, the illusion of clonality can occur when spatially segregated tumour subclones do not overlap with the single sampled region. Multi-region sequencing dataset address this caveat and are therefore particularly powerful for ITH investigations. For example, analyses of the first 100 patients from the Tracking Non–Small-Cell Lung Cancer Evolution through Therapy (TRACERx) prospective cohort study (Funded by Cancer Research UK and others; TRACERx, ClinicalTrials.gov number, NCT01888601), also called the TRACERx100 cohort, suggest that without the use of multi-region WES, 65% of branched subclone clusters would have been (erroneously) classified as clonal [112]. They were able to identify that, whilst most driver mutations were clonal (EGFR, MET, BRAF and TP53), others like PIK3CA and NF1 were often

subclonally aberrated, a relevant observation that could influence treatment choice. The study involving various hospitals, universities and research institutes across the UK began recruitment back in 2014 and aims to enrol around 850 NSCLC patients in stages IA through IIIA and perform high-depth, multi-region WES for each surgically resected tumour [113]. Multi-omics data, including RNA sequencing and RRBS data has been obtained for a subset of samples and more sequencing is under way. The main objective of the study is to investigate potential correlation between measures of intra-tumour heterogeneity and clinical outcome. One of the major reported findings was that subclonal copy number heterogeneity is negatively correlated with disease-free survival, while SNV clonality does not correlated with outcome. Whether or not DMP clonality is prognostic in NSCLC is unknown. A method to infer DMP clonality from multi-region bisulphite sequencing data is needed to answer this question.

## 1.10 Thesis summary

To summarise, while it is well-established that DNA methylation plays a role in tumourigenesis, there is a clear need for computational methods to facilitate the interpretation of bulk tumour bisulphite sequencing data.

To address this, we developed ASCAT.m, our tool for allele-specific copy number profiling and purity estimation from tumour RRBS data (see Chapter 2). We generated multi-region RRBS (range 2-7) of the primary tumour for 38 NSCLC patients from the TRACERx study and applied ASCAT.m on these samples, referred to as the epiTRACERx cohort. We validated our approach by comparing ASCAT.m outputs with those obtained from matched whole-exome sequencing (WES) [112] and 7 newly generated whole-genome sequencing (WGS) tumour samples from 3 patients as well as each patient-matched adjacent normal, showing high concordance. ASCAT.m copy number profiles reveal recurrent alterations in NSCLC and highlights differences between lung adenocarcinoma and squamous cell carcinoma. Whole genome doubling was often observed and was correlated with worse prognosis in lung squamous cell carcinoma, but not in adenocarcinoma.

With accurate tumour copy number and purity estimates in hand, we formalise the relationship between methylation rates, copy number and tumour purity. This relationship is the guiding principle of our algorithm for Copy number-Aware Methylation Deconvolution Analysis of Cancers (CAMDAC) from bulk tumour and tissue-matched normal RRBS data (see Chapter 3). Inter-sample distances between methylation profiles shows that CAMDAC efficiently removes shared normal signals from bulk and highlight differences between patients while retaining high correlations between samples of shared clonal ancestry. CAMDAC purified tumour methylation rates are in agreement with SNV purified estimates across the NSCLC cohort. Simulated and real data show that CAMDAC deconvoluted tumour methylation rates improve differential methylation calls both between tumour and normal cells and between different tumours or sampled regions.

In chapter 4, we use CAMDAC purified methylomes and DMPs to obtain DMRs calls and gain deeper insight into NSCLC methylomes (Chapter 4). DMR ubiquity analysis reveals that methylation heterogeneity is correlated with relapse-free survival. We identify hundreds of recurrently early clonal epimutations across the epiTRACERx cohort, supporting the use of DNA methylation data to improve early detection of NSCLC. In contrast to the bulk, CAMDAC pure tumour methylation profiles reveal intra-tumour subclonal relationships. We obtain DMR calls on SNV-deconvoluted methylation rates, and found DMRs were usually in-*cis* with respect to somatic mutations. DMRs in-*cis* were usually hypermethylation, implying that SNVs possibly lead to the ablation on TBFSs and, in the absence of transcription factor binding, enables methylation of neighbouring CpGs by DNMTs. Paying particular attention to expression levels at genes harbouring neo-antigen mutations, we uncover hypermethylation as a mechanism for immune evasion and published this finding [114]). Building on this work, we show the effect to be even stronger after applying CAMDAC to deconvolve the bulk tumour data.

Finally, we set out to investigate copy- and allele-specific methylation in NSCLC harnessing CAMDAC purified tumour methylation profiles and ASCAT.m copy numbers (Chapter 5). The mode of allele-specific methylation on chromo-

some X in females is below 0.4 on average, indicating the presence of extraction biases against the condensed inactive X chromosome copy. There were no corresponding allelic bias at heterozygous SNPs, suggesting that the inactive X copy is of random parental origin in a given cell, at least within our normal lung samples. The presence of heterozygous SNPs at the germline imprinted locus *H19/IGF2* enables SNP-phasing of methylation rates in 30/37 normal samples, each case validating the presence of allele-specific methylation. We saw loss of imprinting in 5 tumours, each with demethylation of the maternal allele. We define two types of DMPs, stochastic and regulatory, and use regions of $1+1$ to estimate their relative abundance. We find that regulatory DMPs dominate the epimutational landscape of NSCLC, at least in the epiTRACERx cohort. We leverage epimutation with copy numbers 1 and 2 in $2+0$ to time these copy number gains and find they usually occur late in epimutational time.

# Chapter 2

# Allele-specific copy number analysis of cancers from bisulphite sequencing data

## 2.1 Introduction

### 2.1.1 Bulk tumour methylation rates are confounded by tumour purity and copy number

Solid tumour samples are often highly heterogeneous and contain a mixture of tumour and normal cells, the ratio of which is commonly referred to as the tumour purity ($\rho$) or aberrant cell fraction. The relative amounts of tumour and normal DNA depends not only on tumour purity, but also on copy number [115–117]. If the methylation rate is different between the tumour and normal cells, the bulk tumour methylation rate will be affected by the ratio of tumour to normal DNA and thus by both tumour purity and copy number (**Figure 2.1A**).

For example, take a hypothetical bulk tumour sample of purity $\rho = 0.4$ and a CpG locus that is completely unmethylated in the normal contaminating cells, $m_n = 0$, fully methylated in the tumour, $m_t = 1$, and located on a tumour copy number segment with a total of 3 copies, $n_t = 3$. In this bulk mixture, we have 6 unmethylated normal CpGs from 3 diploid cells for every 6 methylated loci with CpG copy number 3 from 2 cancer cells. The bulk tumour methylation rate should therefore fall near: $m_b = \frac{6}{6+6} = 0.5$. Variation in tumour purity and copy number will clearly impact the value of $m_b$ (**Figure 2.1B**). Indeed, several studies have demonstrated that variation in copy number [118–120] and purity [121] lead to in-

creased false positives and negatives in differential methylation analysis, scientists are yet to correct bulk cancer methylomes for both of these confounders simultaneously, probably due to the lack of computational tools for obtaining purity and methylation estimates directly from methylation data.

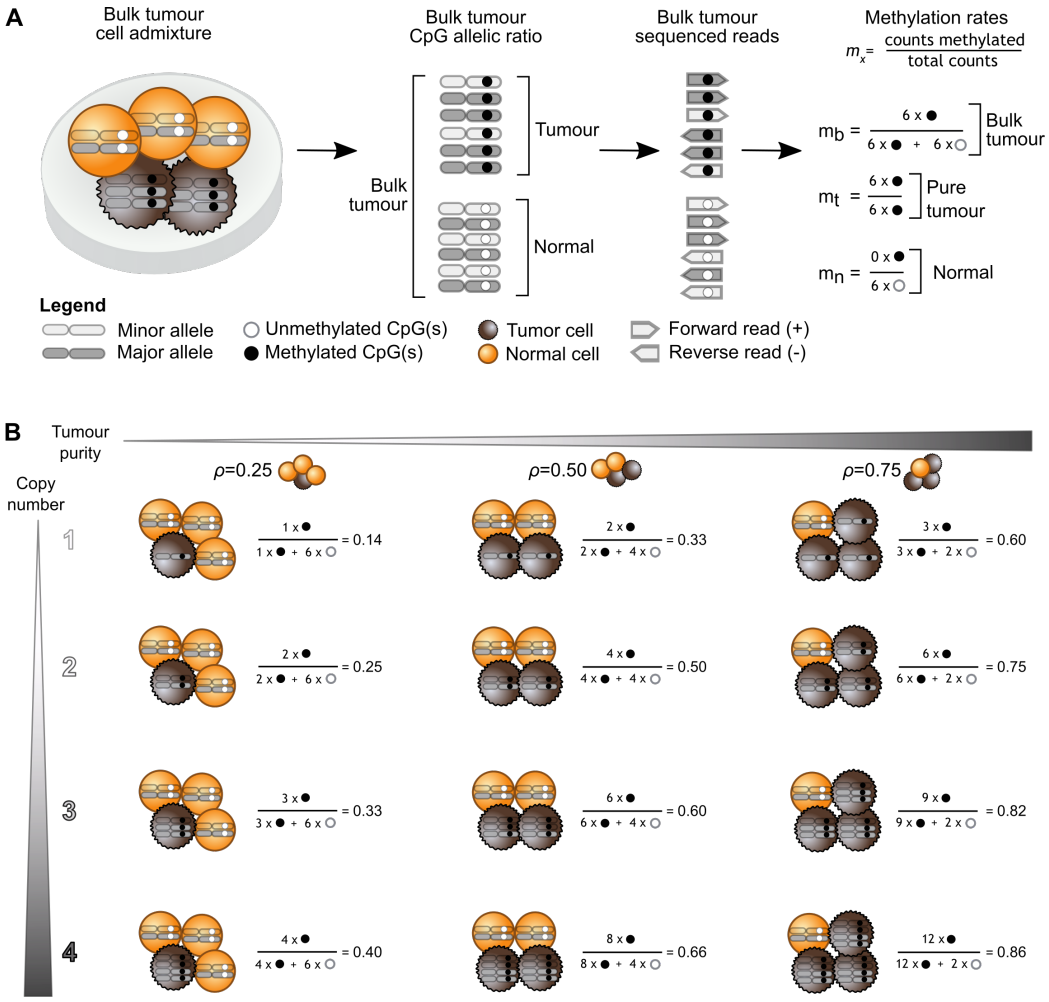


**Figure 2.1: Tumour purity and copy number affect bulk methylation rates.**
(A) Example bulk tumour ($m_b$), pure tumour ($m_t$) and normal ($m_n$) methylation rates at a tumour-normal differentially methylated CpG with total tumour copy number $n_t = 3$ and $\rho = 0.4$. (B) Bulk methylation rates for a CpG locus which is unmethylated in the normal contaminating cells and methylated in the pure tumour cells stratified by a range of purity and copy number.

### 2.1.2 Somatic copy number alterations are universal features of cancer genomes

In addition to their above-outlined effect on bulk tumour methylation rates, somatic copy number alterations also play an important role in tumourigenesis. As copy number modulates expression, amplification at oncogenes and deletions at tumour suppressors will provide cells with a selective advantage. In LUAD and LUSC, focal amplification of EGFR, MYC and deletions of tumour suppressor gene CDKN2A/B are reported driver events [122]. Whole genome doubling (WGD) is a genome-wide copy number alteration involving duplication of all chromosome copies. In non-WGD cancer cells with large proportion of loss of heterozygosity (LOH) where a single chromosome copy remains, mutations may be disadvantageous if not lethal [123]. While WGD is energetically costly, it also mitigates this phenomena known as Muller's ratchet. Genome doubling is therefore correlated with poor prognosis. Timing of chromosomal gains in tumour evolution including genome doubling from whole-genome sequencing analysis of 2,658 cancers as part of the Pan-Cancer Analysis of Whole Genomes (PCAWG) revealed that copy number alterations can be early drivers of tumourigenesis, such as gains of chromosome 7, 19 and 20 which virtually always occur early in glioblastoma development [124]. WGD was found to occur as early as 20 years prior to diagnosis in WGD ovarian cancers, a finding highlighting incredible potential for early diagnosis of this disease.

### 2.1.3 Purity and copy number estimation from bulk tumour data

Experimentally determining the cancer purity and overall ploidy of bulk tumour tissue is possible through cell sorting-based technology such as Fluorescent-Activated Cell Sorting (FACS) [125]. FACS of bulk tumour data is time-consuming, labour intensive and as such may not be suitable for large sequencing efforts.

*In silico* purity estimation from copy number profiling algorithms is a highly scalable and accurate alternative. Indeed, these estimates show good correlations

with ESTIMATE and LUMP purity scores based on RNA-Seq data and leukocyte infiltration, respectively [126], all of which are more accurate than those obtained by histological examination of stained frozen tissue slides [115]. ASCAT is a popular *in silico* method for clonal allele-specific copy number (ASCN) profiling of tumours [117]. While originally designed for SNP array data, ASCAT has been adapted to a range of platforms including WGS [127] and WES [112]. The ABSOLUTE [115] copy number profiling tool is comparable to ASCAT in terms of outputs, but requires computationally expensive statistical modelling. Both approaches derive tumour purity, ploidy and allele-specific copy numbers from the read depth (LogR), corrected for GC content biases, and the allelic imbalance (B-Allele Frequency, BAF) at heterozygous single nucleotide polymorphisms (SNPs). In ASCAT, BAF and corrected LogR values are segmented simultaneously using allele-specific piece-wise constant fitting (ASPCF). The optimal partitioning is fed into a function which computes the goodness of fit score for all possible allele-specific copy number profiles given the input BAF and LogR segments, and a grid of possible values for both purity, $\rho$, and ploidy, $\psi$. The goodness of fit score for each purity and ploidy solution is defined by the distance between raw copy number segments and non-negative whole numbers weighed for segment size. The assumption is that most of the genome is clonal, and therefore the optimal solution is one which minimises the distance metric. SNP-based copy number profiling approaches like ASCAT are yet to be adapted for bisulphite sequencing data.

The published algorithms for copy number analysis of tumour methylation data are fewer in numbers and only generate log read depth profiles [128, 129] as opposed to absolute copy numbers and/or do not provide simultaneous tumour purity estimates. Standalone purity estimation is possible from bulk tumour methylation array data using InfiniumPurify [121]. InfiniumPurify purity values show good correlation with both LUMP and ABSOLUTE and outperforms immunohistochemistry estimates across samples from The Cancer Genome Atlas (TCGA [130]). MethylPurify enables purity estimation method that is designed for bisulphite sequencing and not array data. The approach models purity from intermedi-

ate methylation rate distribution at tumour-normal differentially methylated positions (DMPs), assuming monoclonal tumour cells and homogeneous methylation amongst normal contaminants and ignoring copy number variation. The relative success of MethylPurify in estimating tumour purity directly from methylation data suggests that bulk tumour DNA methylation rates are predictably affected by normal contaminants. Moreover, it shows that for most cancer types, intermediate methylation at DMPs stems from somatic evolution and not from normal cell type heterogeneity given the assumption that normal infiltrating cells are homogeneous. However, the main drawback of these tools for tumour purity estimation from methylation data is that they do not provide absolute copy number profiles. As such, purity estimates from methylation rates are not corrected for copy number which can lead to erroneous purity estimates, for example in whole genome doubled tumours.

In the past, researchers have sometimes chosen to obtain additional SNP array data to compute copy number profiles from established pipelines instead of extracting the information directly from the methylation data [120, 131]. This approach incurs additional costs in terms of tumour material, time and the arrays themselves. In a clinical setting, it is particularly important to avoid wasting precious tumour material and so obtaining additional array or sequencing data may not be feasible. In some cancer types, such as non-small cell lung cancer, operable tumours are surgically resected as first line of treatment which means plenty of material is available (depending on the tumour size), but this is not standard practice for all solid tumours. It is often the case that only limited amounts of material can be collected via tumour biopsies. A novel method to generate both accurate tumour purity estimates and copy number profiles directly from bisulphite sequencing data is clearly needed.

### 2.1.4 Chapter summary

To address this issue, we introduce ASCAT.m, a new tool that enables both allele-specific copy number and purity inference directly from bulk tumour RRBS data. We begin this chapter with a detailed explanation of the methodology behind ASCAT.m. We then apply ASCAT.m to the epiTRACERx pilot RRBS dataset

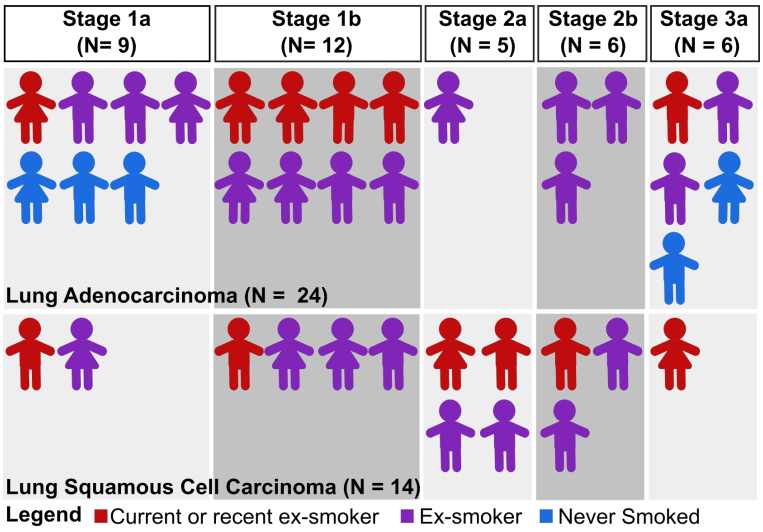


**Figure 2.2: EpiTRACERx cohort.**
Depiction of the histological subtypes, gender, smoking status and tumour stage distribution across the epiTRACERx cohort.

(**Figure 2.2, Table B.1**), which is part of the TRACERx prospective cohort study [113]. The epiTRACERx cohort consists of multi-region RRBS data for a subset of 38 non-small cell lung cancer patients (122 tumour samples, 3.3 samples per patient on average, range 2-7) chosen from the first 100 patients of the TRACERx study [113]. Copy number analyses of multi-region whole-exome sequencing from the TRACERx study revealed high variability in both tumour purity and copy number [112]. The epiTRACERx RRBS cohort is therefore highly amenable in its role as a pilot dataset for ASCAT.m. We observed high correlation between ASCAT.m tumour purity and copy number estimates and values obtained from matched WES data [112] and, from a subset of unpublished WGS samples recently generated by collaborators. Copy number profiles revealed a high prevalence of WGD in the epiTRACERx NSCLC cohort. Whole genome doubling is negatively correlated with patient outcome in LUSC likely due to high levels of loss of heterozygosity.

## 2.2 Results

### 2.2.1 Computing BAF and LogR from bisulphite sequencing data

Inspired from existing methods copy number and purity estimation from array or sequencing data [115–117], ASCAT.m requires coverage (LogR) and allelic imbalance (BAF) information at a sufficient number of (heterozygous) SNPs in order to compute tumour purity, ploidy and allele specific copy numbers (ASCN, **Figure 2.3**). The epiTRACERx RRBS dataset was predicted to be particularly well-suited for ASCAT.m because *MspI* digestion enriches for CpG Islands and the latter are known hotspots for C>T polymorphisms [132].

We began by compiling read counts at all 1000 genome SNP positions [133] from bulk tumour and patient-matched normal RRBS data (**Figure 2.3**). The LogR at the $i^{th}$ SNP ($SNP_i$) was then calculated by taking the total read coverage of $SNP_i$ in the tumour, $cov_{t,i}$, normalised by the coverage in the normal, $cov_{n,i}$. Leveraging findings from different studies, we show that RRBS-derived raw LogR values suffer from at least 3 sequencing coverage biases as a result of the RRBS protocol (**Figure 2.4**). (1) Enzymatic digestion with *MspI* yields libraries with heterogeneous insert size distributions reflecting variability in the distance between any two CCGG cleavage sites which can be as little as a few base pairs to hundreds [134]. (2) Extreme GC contents are known to bias polymerase chain reaction amplification and coverage [135] and bisulphite converted sequences have decreased GC content, depending on CpG methylation status and density. (3) Coverage is inflated in genomic regions that tend to replicate earlier during S-phase compared with those that replicate later. Replication timing differs not only across the genome but also between cell types [136]. ASCAT.m corrects LogR estimates for each of these three sources of bias (See Methods Section 2.4.3.1, **Figure 2.4**).

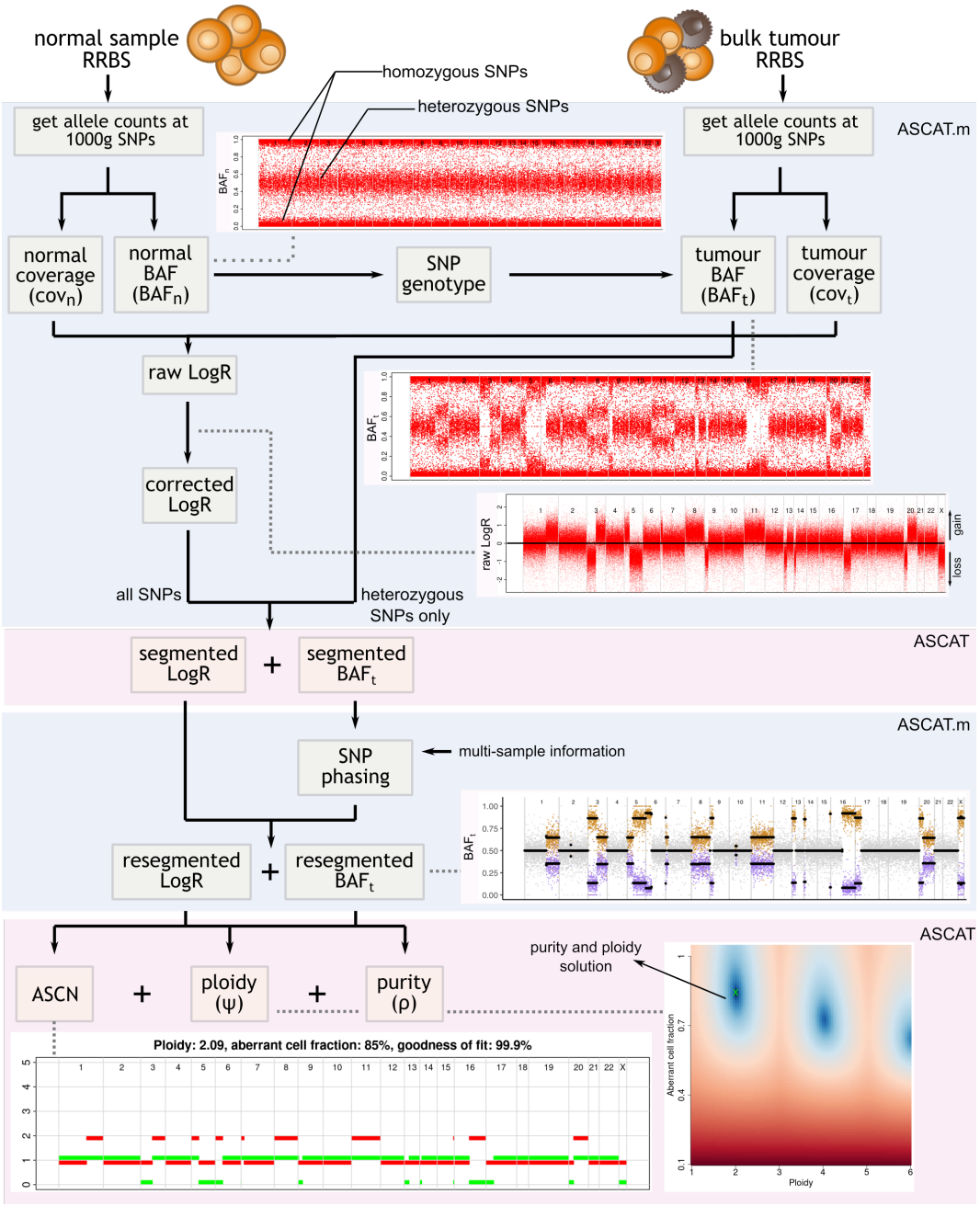


**Figure 2.3: ASCAT.m workflow.**
Allele counts are compiled at all 1000 genome SNP positions [133] for bulk tumour and patient-matched normal RRBS data. SNP genotyping is carried out on the normal sample and tumour BAF is obtained at heterozygous SNPs. LogR is computed from the matched normal and tumour coverage at each SNP. BAF and corrected LogR values are fed into ASCAT standard ASPCF function. For multi-region datasets, re-segmentation of BAF and LogR tracks is performed with multi-sample phasing of BAF segments with allelic imbalance in at least one sample. The best copy number solution is found using a purity and ploidy grid search approach. The solution which maximises the goodness of fit to integer copy number is selected.

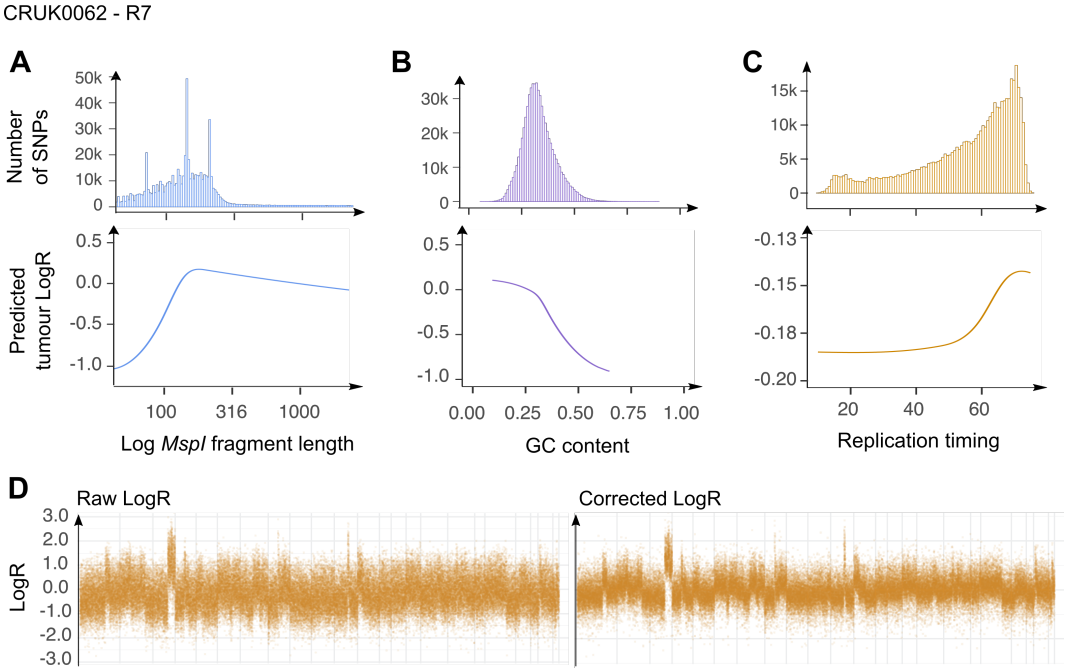


**Figure 2.4: Sources of bias in LogR values.**
(A-C) *MspI* fragment length (A), GC content (B) and replication timing (C) affect tumour LogR. The observed peaks in the fragment length distribution originate from *MspI*-containing micro-satellite repeats of distinct lengths and is characteristic of human RRBS libraries. (D) A linear combination of three natural splines, modelling the effect of each of the three biases described in (A), (B) and (C) with respect to tumour LogR is used to correct the raw LogR and yield the corrected values.

Next, ASCAT.m requires normal ($BAF_n$) and tumour BAF ($BAF_t$) values at enough heterozygous SNPs to establish germline genotypes and identify regions of allelic imbalance respectively (**Figure 2.3**). Computing BAF values from RRBS is challenging compared with genome sequencing and array data because bisulphite conversion leads to unexpected reference and alternate allele read counts at SNP loci with a G and/or C allele. Indeed, unmethylated Cs are converted to Ts during library preparation, yielding four possible bisulphite DNA strands: (complementary to) original top and (complementary to) original bottom with roughly the same likelihood (**Figure 2.5A**). Most bisulphite sequencing protocols are directional, including the RRBS NuGEN assay used for this work, meaning only bases from original top (+) or complementary to original bottom strand (-) are sequenced. Methylation further complicates BAF calculations since roughly half of the SNPs reported by RRBS are located at CpGs. Across the epiTRACERx cohort, polymorphisms at

CpGs account for 49.9% of SNPs, 83.3% of which are CpG>TpG polymorphisms.

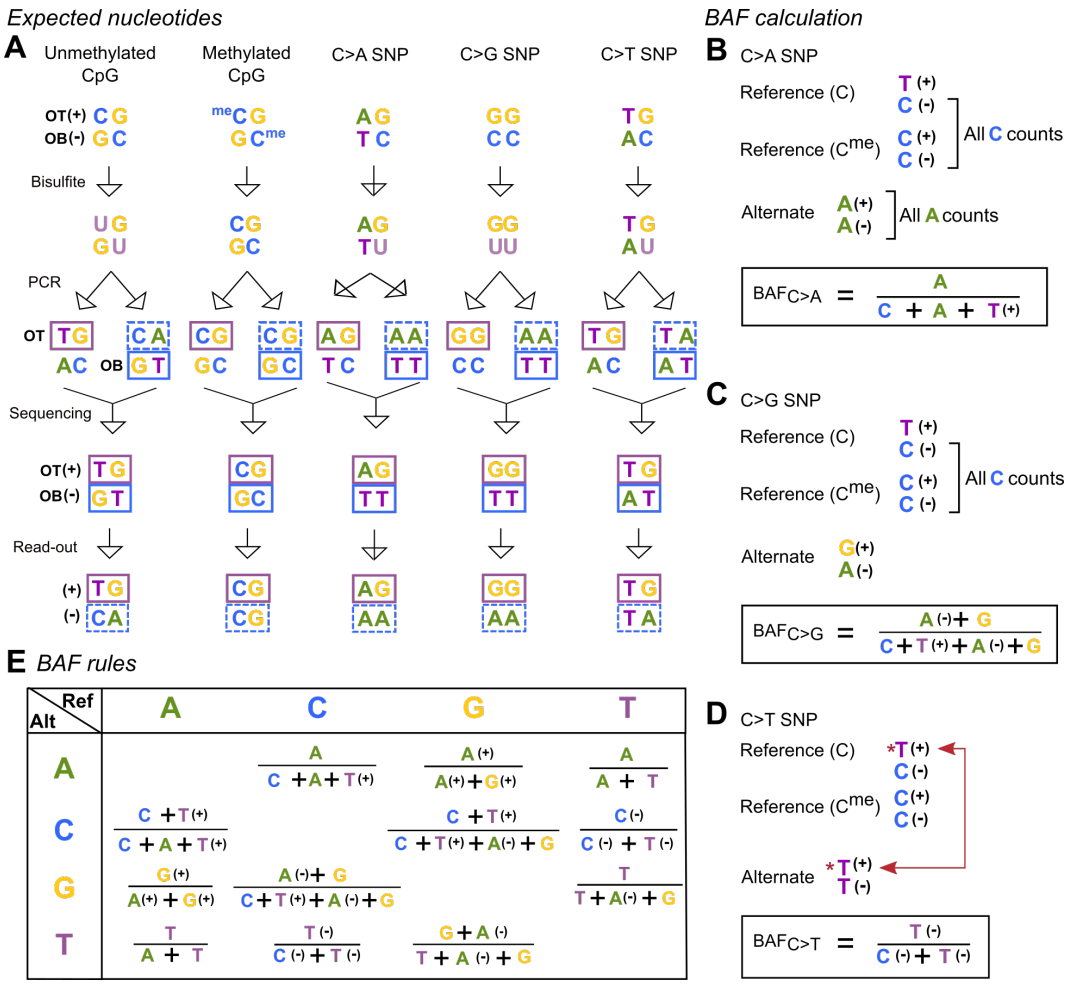


**Figure 2.5: ASCAT.m BAF calculation rules.**
(A) Transformation of (un)methylated reference CpG, and alternate ApG, GpG and TpG alleles during bisulphite sequencing. (B-D) Derivation of BAF rules from strand specific base pile ups C>A (B), C>G (C) and C>T (D) SNP loci. (E) BAF formulae for all SNP types.

To address these issues, we propose a set of strand specific allele counting and BAF calculation rules for all SNP types for ASCAT.m. For reference, the derivation for C>A, C>G and C>T SNPs is depicted in (**Figure 2.5A-D**). For a C>A SNP at a CpG, all reads supporting A on the + or - strand, i.e. A(+) and A(-), can be assigned to the alternate allele unambiguously. Reads reporting C(+) and T(+) can be uniquely attributed to the methylated and unmethylated C, respectively, while C(-) counts could arise from either methylation state. The al-

lelic imbalance can be calculated as $BAF_{C>A} = \frac{A}{A+C+T(+)}$ (**Figure 2.5B**). At C>G SNPs, the alternate allele is represented by G(+) and A(-) nucleotides with $BAF_{C>G} = \frac{G(+)+A(-)}{G(+)+A(-)+C+T(+)}$. The situation is more complex for C>T SNPs because the bisulphite converted unmethylated reference cytosine base is indistinguishable from the alternate allele on the + strand with both alleles leading to a T(+) read-out (**Figure 2.5A,D**). However, base counts taken from the - strand are unique to either the alternate allele, T(-), or the (un)methylated reference, C(-). Therefore, it is still possible to quantify allelic imbalance at C>T SNPs: $BAF_{C>T} = \frac{T(-)}{C(-)+T(-)}$. Using only the - strand implies that, on average, only half of the total number of reads can be used to measure allelic imbalance at C>T SNPs. Following this line of reasoning, we outline a set of rules to compute BAF values at all types of SNPs (**Figure 2.5E**). The same principles above can be generalised to non-CG cytosine methylation, known to occur at low percentages in the mammalian genome [137]. ASCAT.m BAF rules are robust to both CpG and non-CpG cytosine methylation.

### 2.2.2 Comparing genotypes derived from RRBS and WGS data

To test our approach, we leveraged unpublished WGS data generated by colleagues for 3 NSCLC patients also part of the epiTRACERx cohort, including normal samples (see section 2.4.2.2). We compared genotyping outputs generated at overlapping SNPs between the 2 platforms for each of the normal samples. We calculated the false positive rates (FPR) and false negative rates (FNR) for RRBS data using the WGS-derived genotypes as ground truth. The average FPR across all SNP types and samples was 0.3% whilst the mean FNR was 25% (**Figure 2.6**). The average FNR was highly context dependent ($\chi^2$ test, p-val $< 2.2 \times 10^{-16}$, $FNR_{CCGG} = 83\%$, $FNR_{CG} = 20\%$, $FNR_{other} = 15\%$). Increasing sequencing depth only lead to a small decrease in FNR: by requiring a minimum SNP coverage of 30 instead of 10 reads, the average FNR dropped to roughly 1 in 5 heterozygous SNPs. Polymorphic CCGGs perturbing or creating a CCGG *MspI* recognition sequence, can lead to allele-specific fragments during RRBS library preparation skewing allelic coverage and were the largest source of false negatives (49%). We observed a

bias towards SNPs erroneously assigned as homozygous reference (72%) compared with homozygous alternate (28%) at false negatives (**Figure 2.6B**). This suggest an alignment bias, likely due to limited mappability of short *MspI* fragments with alternate alleles. In line with this hypothesis, SNPs with low reference allele mapping quality scores (MAPQ $< 40$) were enriched for homozygous reference false negatives (Wilcoxon test, p-val $< 2.2 \times 10^{-16}$), especially outside CCGG context.

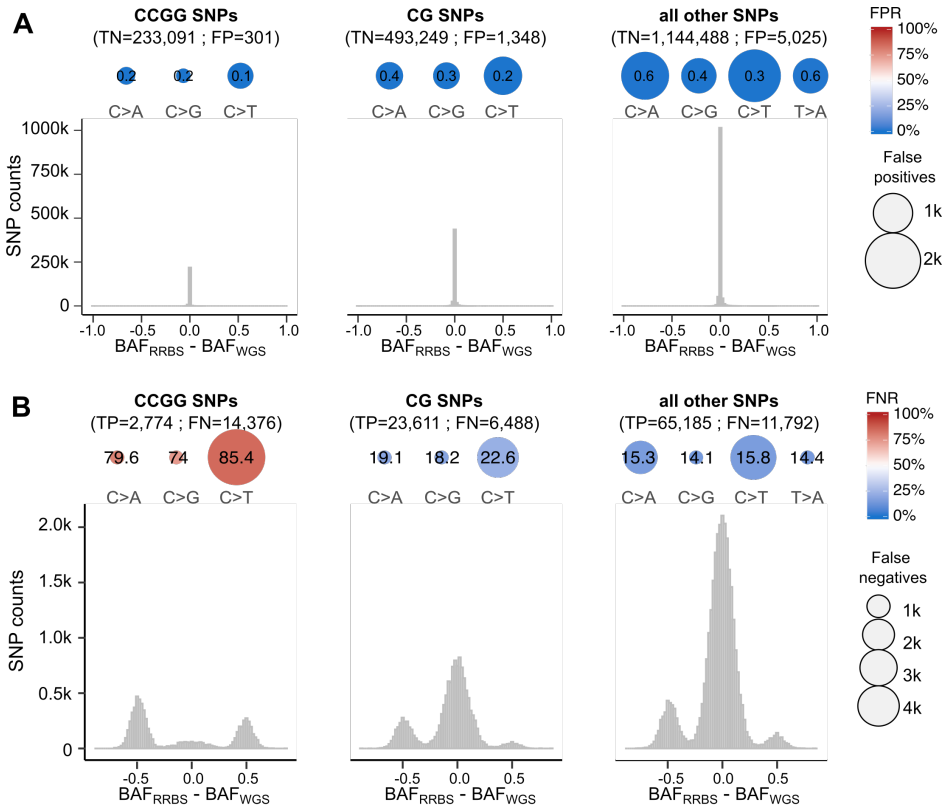


**Figure 2.6: Comparing ASCAT.m and ASCAT on WGS data.**
(A-B) False positive (A) and false negative (B) heterozygous call rates across SNP types (top) and ASCAT.m BAF estimate error distribution histograms of the distribution of CAMDAC BAF estimate errors and noise (bottom). SNPs are considered heterozygous when $0.1 \geq \text{BAF} \geq 0.9$ and a gold standard alleleCounter pipeline on WGS of the normal sample subset to RRBS-covered regions is taken as ground truth.

### 2.2.3 Multi-sample SNP phasing improves segmentation

Next, we compute BAF and corrected LogR values for all tumour samples in the epiTRACERx cohort. Given sufficient tumour purity and sequencing coverage, copy number segments with clonal allelic imbalance will generate two distinct BAF

bands. This can be used to assign heterozygous SNPs to the gained or lost alleles and to phase all SNP alleles in the copy number segment. SNP phasing information is relayed between different samples from the same patient. For example, consensus phasing of the 5 tumour regions of CRUK0069 revealed over half a dozen mirrored subclonal allelic imbalance (MSAI) events (**Figure 2.7**).

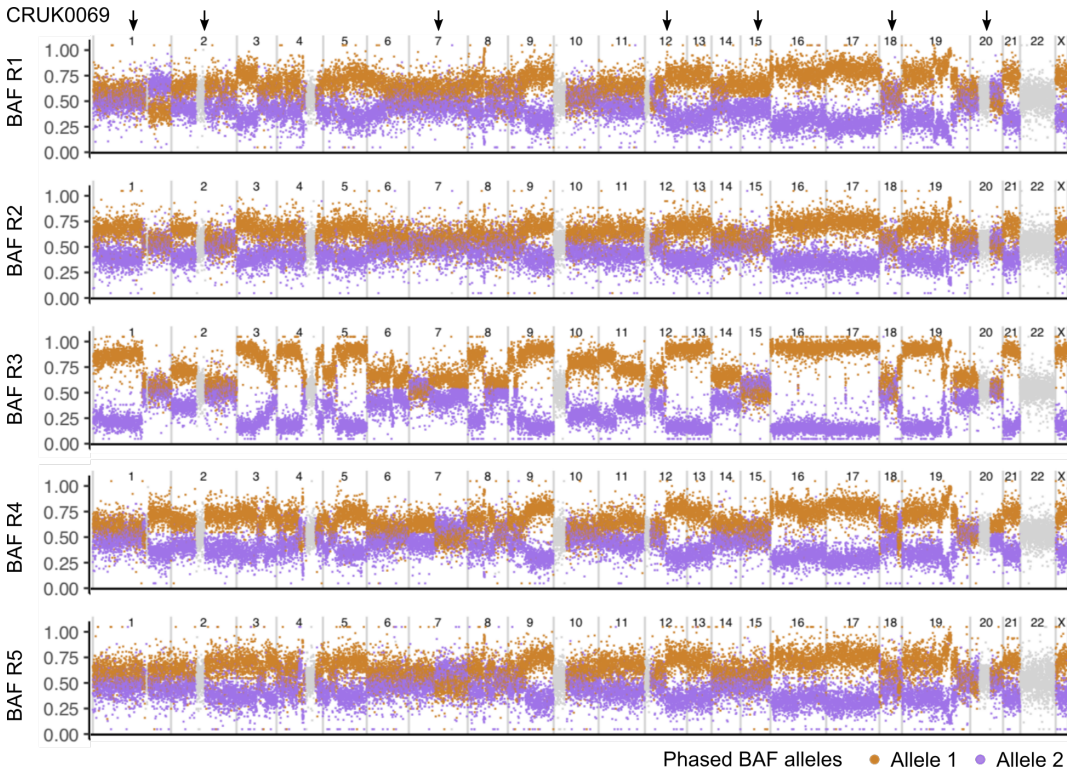


**Figure 2.7: Creating haplotypes from multi-sample BAF estimates.**
(A) The phased BAF profile for regions 1 to 5 is shown with phased segments being randomly assigned alleles 1 (orange) or 2 (purple). The same labelling is used for the $i^{th}$ SNP across all 5 tumour samples enabling visual identification of MSAI events.

ASCAT.m phased BAF and LogR profiles are then fed into the ASCAT piecewise constant fitting function modified to account for multi-sample haplotyping (**Figure 2.3**). The output segmented BAF and LogR profiles are fed into the final copy number fitting function, where a grid search approach is used to identify the optimal purity and ploidy solution, that is to say the solution where, on average, allele-specific copy number segments have the shortest distance to integer copy number states. Accurate segmentation is key to generate high quality copy number

profiles. For example, take a clonal copy number segment with *major allele* + *minor allele* $= 2 + 1$. If allelic imbalance is not detected from the BAF track by ASCAT's allele specific piece-wise constant fitting (ASPCF), the segment is misclassified as a balanced copy number segment. At the true underlying tumour purity, the total copy number would be unaffected by this misclassification and would appear near copy number 3, but the minor allele segment would be mistakenly positioned halfway, at 1.5 copies. Because the minor allele is 0.5 away from the nearest integer copy number, this may penalise the score of the true solution such that it is no longer a local minimum of the grid search. This is especially true if large and/or multiple segments are misinterpreted as balanced. Multi-sample haplotyping successfully rescues allelic imbalance signal in lower purity and/or coverage samples, for example in CRUK0069-R2 ($\rho = 0.32$, **Figure 2.8**). We note that BAF bands for the epiTRACERx RRBS data are wider than for matched high coverage WGS and WES data, further increasing the importance of haplotyping.

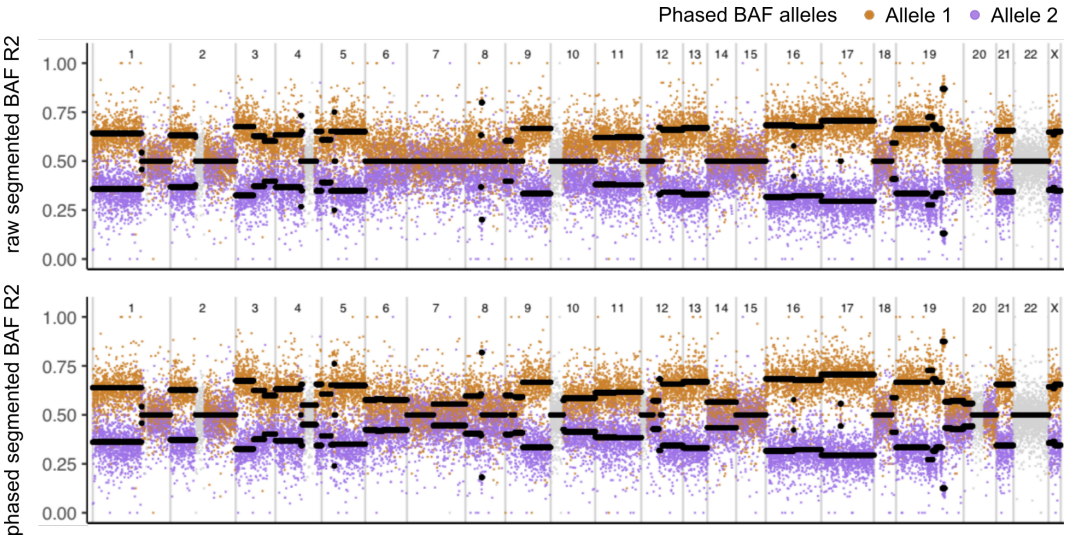


**Figure 2.8: BAF segmentation is improved by multi-sample haplotyping.**
Raw (top) and phased (bottom) segmented BAF for low purity ($\rho = 0.32$) and ploidy ($\psi = 2.67$) sample CRUK0069-R2 leveraging multi-sample information.

### 2.2.4 Comparing copy number profiles, purity and ploidy derived from RRBS versus matched WES and WGS data

We compared ASCAT(.m) BAF, LogR and allele-specific copy number segments for 7 tumour samples taken from the above-mentioned 3 patients with matched WGS and RRBS data (See Methods Section 2.4.2.2). Despite the higher number of false negative homozygous SNPs in our RRBS data, BAF estimates generated by ASCAT.m enable correct identification of BAF bands separation at genomic regions with allelic imbalance as determined by SNP phasing while these are sometimes missed by the matched WGS data (**Figure 2.9** and Supplementary Figures SA.1-6). The LogR tracks are in good agreement between the two platforms across all samples indicating that biases in coverage introduced by the RRBS protocol are adequately modelled and removed by ASCAT.m. We compared final allele-specific copy number calls to data obtained from both platforms and, unsurprisingly, the two are virtually identical (**Figure 2.9** and Supplementary Figures SA.1-6).

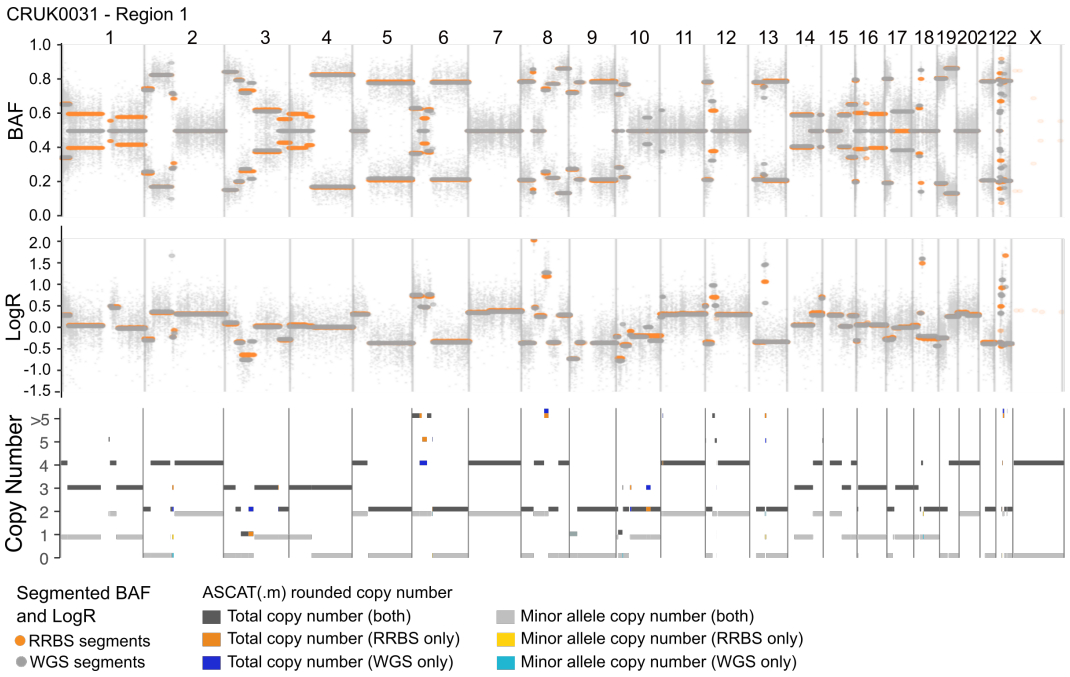


**Figure 2.9: Comparing WGS- and RRBS-derived ASCAT(.m) BAF, LogR and copy number segments for a representative tumour sample.**
Direct comparison of BAF (top), LogR (middle) and allele-specific copy number (bottom) profiles derived by ASCAT(.m) from matched RRBS and WGS for CRUK0031-R1. Segmented BAF and LogR values are plotted at heterozygous SNPs only. Raw BAF and LogR values at every heterozygous SNPs is shown in for the WGS data only. This sample is male which explains why virtually no data points are plotted on chromosome X.

ASCAT(.m) ploidy ($\psi$) and purity ($\rho$) validation was performed by comparing RRBS estimates with those inferred from WGS (mean coverage 67x) and high-coverage WES (mean coverage 464x) performed on the same samples (Table 3, [112]). Tumour purity and ploidy values derived from RRBS showed excellent agreement (**Figure 2.10A-B**) with both WGS ($corr_\rho = 0.996$, $corr_\psi = 0.991$) and WES estimates ($corr_\rho = 0.984$, $corr_\psi = 0.978$). We note that tumour purity variability is relatively high whilst ploidy is typically more homogeneous within tumours (**Figure 2.10C**).

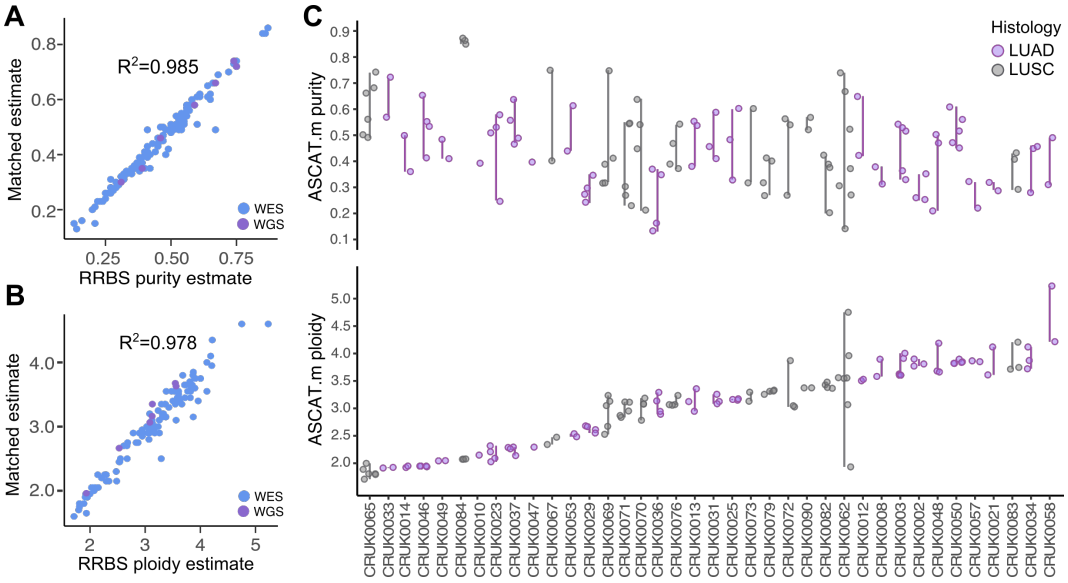


**Figure 2.10: ASCAT(.m) purity and ploidy estimates across epiTRACERx and sequencing platforms.**
(A-B) Direct comparison of ASCAT(.m) RRBS-derived purity (A) and ploidy (B) estimates with matched WES (blue) and, where available, WGS data (purple). (C) Patient-wise ASCAT.m tumour ploidy (top panel) and purity (bottom panel) distribution, as inferred by ASCAT.m.

### 2.2.5 The somatic copy number variation landscape of NSCLC

Whole genome doubling (WGD) status is defined by the fraction of the genome with loss of heterozygosity (LOH) and tumour ploidy (**Figure 2.11A**). ASCAT.m results show that WGD is widespread across lung adenocarcinoma (LUAD, 71%) and squamous cell lung carcinoma (LUSC, 86%) patients, while LOH is enriched in LUSC compared with LUAD, with 94% versus 26% of samples harbouring LOH

in ≥30% of the covered genome respectively (Wilcoxon p-val $= 1.48 \times 10^{-11}$). We note that WGD is negatively associated with relapse-free survival in LUSC but not in LUAD (**Figure 2.11B**). Having a higher fraction of the genome with LOH, genome doubling likely provides a strong selective advantage in LUSC decreasing the likelihood of acquiring lethal mutations on the sole copy of an essential gene [123]. No differences in overall purity or ploidy were noted between the two lung cancer subtypes ($\rho$ Wilcoxon p-val $= 0.411$, $\psi$ Wilcoxon p-val $= 0.276$). However, LUAD samples tend to be of higher ploidy than LUSC when comparing samples of the same WGD status between histologies (WGD Wilcoxon p-val $= 9.83 \times 10^{-3}$, non-WGD Wilcoxon p-val $= 1.85 \times 10^{-2}$).

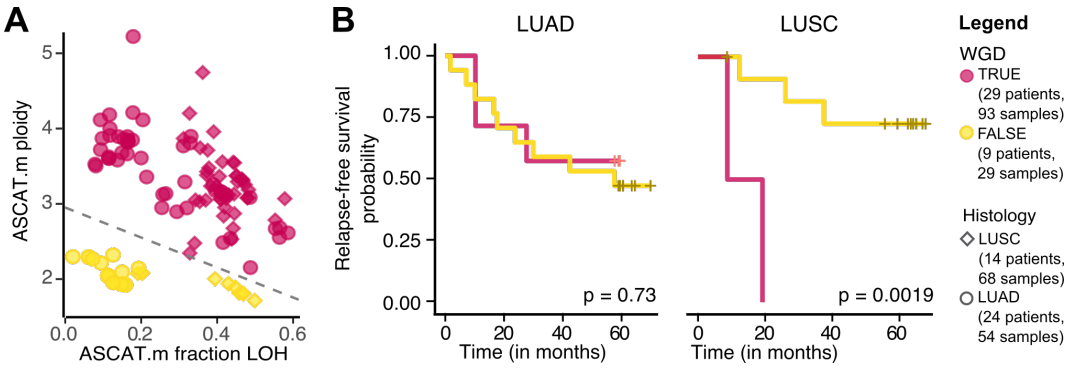


**Figure 2.11: Whole genome doubling predicts outcome in LUSC but not LUAD.** (A) Whole genome doubling status is a function of both tumour ploidy and the fraction of the genome with loss of heterozygosity. (B) Relapse-free survival probability for LUAD (left) and LUSC (right) cases stratified by WGD status from ASCAT.m.

Genomic instability is a defining feature of both cancer types (**Figure 2.12**). Certain copy number gains and losses are observed at a high frequency across the two subtypes, such as +8q and -8p, while other patterns are strikingly different between subgroups. For example, +3q is found in virtually all LUAD samples and is absent from LUSC. Gains and losses are defined based on allele-specific copy number and with respect to tumour ploidy (2.4). This is important as most LUAD and LUSC are tetraploid and will appear to have high levels of gains and fewer losses if not corrected for baseline ploidy. Indeed, reports by others based on GISTIC2.0 data in LUAD only identified losses of 8p, and 18 with frequencies greater than 0.5 [138].

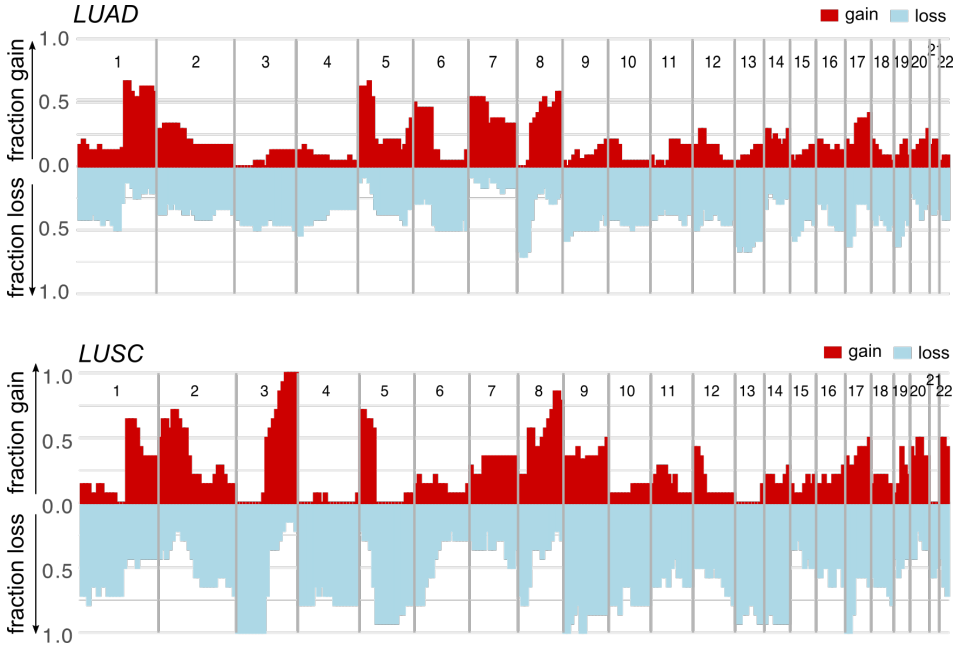


**Figure 2.12: NSCLC somatic copy number variation landscape.**
Fraction of samples with copy number gains (red) and losses (blue) in 10Mb bins across the genome for LUAD (top) and LUSC (bottom).

## 2.3 Discussion

To summarise, we have shown that ASCAT.m enables accurate allele-specific copy number profiling and simultaneous tumour purity estimation from RRBS data, obviating the need to perform separate copy number profiling experiments, reducing cost and saving time. We demonstrated that ASCAT.m BAF calculation rules are successful in identifying heterozygous SNPs outside CCGG context. We showed that multi-sample BAF phasing improves detection of allelic imbalance from BAF, especially in samples of low tumour purity. ASCAT(.m) output allele specific copy number profiles for tumour samples with matched RRBS and WGS showed good agreement and overall tumour purity and ploidy estimates were highly correlated between WES, WGS and RRBS platforms. WGD was found to be widespread in NSCLC and was significantly associated with increased probability of relapse in LUSC but not LUAD and link this result to LOH abundance. Finally, the copy number variation landscape of LUAD and LUSC was obtained and compared.

We note that patient-matched tumour adjacent normal samples may not always be available. In this study, we build reference RRBS coverage profiles which can be used to compute LogR in the absence of patient-matched normal data. Given sufficiently low tumour purity or the absence of LOH, approximate heterozygous SNPs can assignments may be obtained directly from the tumour BAF. We ran ASCAT.m with tissue-matched normal for one tumour sample where no patient-matched normal data was available. Allele-specific copy number and purity estimates can therefore be generated by ASCAT.m without patient-matched normal using the reference coverage panel generated in this study. Although this was not tested, we hypothesise that patient-matched blood normal RRBS data may be used for ASCAT.m as an alternative to tissue-matched normal. We therefore conclude that ASCAT.m will be highly valuable for tumour copy number and purity estimation from bisulphite sequencing data irrespective of cancer types and matched normal availability.

While evaluating ASCAT.m for genotyping purposes, we noted allelic skewage at heterozygous SNP in CCGG tetranucleotides context due to SNP-biased enzymatic digestion and outside CCGG motifs, likely due to poor alignment of short *MspI* fragments with alternate alleles and low mappability of short single end bisulphite sequencing reads. It is worth underscoring that these allelic biases are largely specific to the single-end RRBS protocol. Paired-end sequencing is usually recommended to increase mapping quality but is not suitable in combination with RRBS due to many fragments being shorter than twice the read length leading to duplicate sequencing of bases from the same DNA molecule. It is likely that the above biases could be avoided altogether by using an alternative DNA methylation profiling method.

Paired-end whole-genome bisulphite sequencing (WGBS), for example, would not suffer from enzymatic digestion biases potentially reducing the number of false negatives and increasing the mappability of low degeneracy bisulphite treated reads thanks to paired-end sequencing. While genotyping bias may be reduced, sequencing costs would be much greater for paired-end WGBS than for single-end RRBS. Unlike RRBS, WGBS is less cost-effective due to uneven CpG concentration out-

side CpG islands meaning many reads will not contain any methylation information [31]. WGBS usually requires large quantity of input DNA material (at least 10-fold higher than NuGEN RRBS protocol). For large scale analyses, it may not be feasible to carry out WGBS as per the reasons outlined above. Standalone Nanopore sequencing technology or in combination with enzymatic conversion of modified cytosine bases is a promising new alternative to provide accurate methylation and SNP information as well as valuable long range phasing information with minimal input DNA requirements [139].

Despite high number of false negatives, ASCAT.m BAF rules were shown to be accurate at called heterozygous SNPs and could be used beyond genotyping purposes. In future, we speculate that these rules may form the basis of a new module for ASCAT.m enabling *de novo* identification of single nucleotide variants (SNVs). It would be incredibly valuable to add SNV calls to the output one can obtain directly from bisulphite sequencing data. While RRBS data is suitable for genotyping purposes and copy number profiling with ASCAT.m, we suspect WGBS or Nanopore sequencing may be more accurate for genotyping and by extension for *de novo* identification of single nucleotide variants.

We note that several *in silico* methods for tumour purity and copy number inference have been developed since ASCAT was first published in 2010. The Battenberg method [116] for instance was inspired from ASCAT but designed for WGS. Battenberg exploits haplotype information to generate subclonal allele-specific copy number profiles. WGBS should in theory enable haplotyping and therefore allow subclonal copy number profiling directly from methylation data. In multi-sample studies like epiTRACERx, we can evaluate clonality to some extent by comparing tumour regions from the same primary. However, subclonal copy numbers and DMP calls would be useful in reconstructing phylogenies from single sample biopsies and plan to extend ASCAT.m with haplotyping with WGS data.

## 2.4 Methods

The methods described below were recently published as part of our bioRxiv preprint [140].

### 2.4.1 epiTRACERx methylation study design

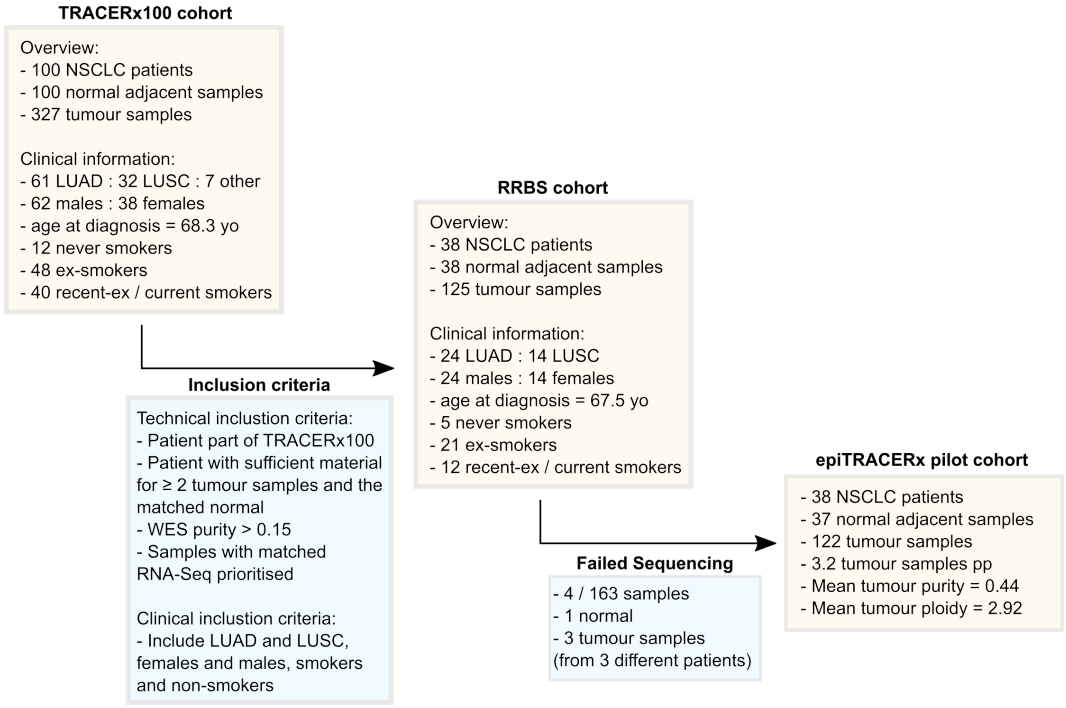


**Figure 2.13: EpiTRACERx patient inclusion criteria and cohort clinical features.** This flow chart shows how patients from the TRACERx100 cohort were selected for inclusion in the epiTRACERx methylation study. The RRBS cohort is the same as the final epiTRACERx cohort minus 4 samples which failed sequencing.

Samples from the first 100 patients of the TRACERx lung cancer cohort were selected for multi-region RRBS (**Figure 2.13**). Patients with data for samples from 2 or more tumour regions and the adjacent matched normal, all with sufficient material remaining, were considered for bisulphite sequencing. Tumour samples with purity below 15% were discarded with the exception of CRUK0062-R6 which was included for comparison with the other 6 sampled regions from this patient's primary tumour. Patients with tumour samples of high purity were prioritised as well as those with matched RNA-Seq data [114]. Samples were obtained across both lung adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC) subtypes, genders

and smoking status. The clinical and technical inclusion criteria are summarised in **Figure 2.13** and the resulting cohort is depicted in **Figure 2.2** and **Table B.1**.

## 2.4.2 Sequencing methods

### *2.4.2.1 RRBS sequencing protocols*

RRBS sequencing was performed by Miljana Tanic and Pawan Dhami at University College London prior to my involvement in the epiTRACERx study.

Multi-region RRBS data was generated for about 1 in 3 NSCLC patients from the TRACERx 100 cohort (122/327 tumour regions from 38/100 patients, each with matched normal). The NuGEN Ovation RRBS Methyl-Seq System was adapted by the manufacturer for automation on Agilent Bravo liquid handling robot. This set up was then used to prepare libraries by enzymatically digesting 100ng of gDNA with *MspI*, an methylation insensitive enzyme that cleaves DNA at 5′-CCGG-3′ motifs. The enzyme breaks the phosphodiester bonds upstream of CpG dinucleotides, leaving a 2bp overhang suitable for adaptor ligation and then a final end repair step. Qiagen's EpiTect Fast DNA Bisulfite Kit was used for bisulphite conversion of the resulting libraries.

Bisulphite converted libraries were then amplified by polymerase chain reaction using 12 cycles and purified using Agencourt RNAClean XP magnetic beads (suitable for DNA purification). Library quantification was performed by Qubit dsDNA HS Assay (Invitrogen) and quality control was carried out using Agilent Bioanalyzer High Sensitivity DNA Assay (Agilent Technologies). We multiplexed 8 samples per flow cell and sequenced on HiSeq2500 system using HiSeq SBS Kit v4 in 100bp paired-end runs for our pilot patient CRUK0062 and in single-end for the rest of the cohort, yielding an average of 150M raw sequencing reads per sample. Paired-end sequencing was not applied to the larger cohort because our pilot data revealed a high number of mates with negative insert sizes due to short *MspI* fragments and outweigh any increase in mapping efficiency.

Sequencing outputs were quality checked with FastQC (Babraham Institute, https://www.babraham.ac.uk/), adapter sequences trimming was performed with Cutadapt [141] and reads were aligned to the UCSC hg19 refer-

ence assembly using Bismark v0.14.4 [26]. Read deduplication was carried out using NuDup, leveraging NuGEN's molecular tagging technology (NuGEN, https://github.com/nugentechnologies/nudup). On average $1.88 \times 10^8$ reads per sample remained post-processing and alignment (**Table B.2**), resulting in an average of 4.5 million CpGs being supported by at least 1 read in any one sample. A subset of samples from the epiTRACERx RRBS dataset were deposited at the European Genome-phenome Archive (EGA) under accession number EGAS00001003484 as part of Rosenthal *et al.* [114]. The full cohort may be accessed through the Cancer Research UK & University College London Cancer Trials Centre (ctc.tracerx@ucl.ac.uk) for academic non-commercial research purposes upon reasonable request, and subject to review of a project proposal that will be evaluated by a TRACERx data access committee, entering into an appropriate data access agreement and any applicable ethical approvals.

#### *2.4.2.2 WGS sequencing protocols*

Whole genome sequencing was performed on 7 samples from 3 patients included in the TRACERx100 cohort and the epiTRACERx cohort. The WGS data was generated by Edinburgh Genomics. Samples were sequenced on Illumina HiSeq in paired-end 100bp runs. Sequencing outputs were processed by Michelle Dietzen. Reads were quality checked with FastQC v0.11.5 (Babraham Institute, https://www.babraham.ac.uk/) and aligned to the UCSC hg19 reference assembly using BWA-MEM v0.7.15 ( [142], http://bio-bwa.sourceforge.net). Alignments were saved as Binary Alignment Map files (BAM), sorted and indexed with SAMtools v0.1.19 (http://github.com/samtools/).

### 2.4.3 Computational method development and analyses

#### *2.4.3.1 ASCAT.m*

ASCAT.m stands for Allele-Specific Copy number Analysis of Tumours from Methylation data. Like ASCAT [117], ASCAT.m requires both BAF and LogR information at germline SNPs to compute tumour purity and copy number. To obtain these variables, base counts are compiled at all 1000 genome SNP positions [133]

that overlap with RRBS data. Briefly, the LogR at the $i^{th}$ SNP is taken as the base-2 logarithm of the read depth ratio of the tumour coverage, $cov_{t,i}$, to the normal reference, $cov_{n,i}$, divided by the average of this ratio. LogR values are easily computed from RRBS data, as per Equation (1):

$$LogR = log_2\left(\frac{cov_{t,i}/cov_{n,i}}{\frac{\sum_{j=1}^{k}(cov_{t,j}/cov_{n,j})}{k}}\right) \quad (1)$$

For the majority of the epiTRACERx cohort, normalisation is carried out with patient-matched tumour-adjacent normal samples, but LogR values can also be obtained from blood normal RRBS data. Patient-matched normal reference samples remove noise from germline copy number variants. For one female patient, CRUK0047, we do not have patient-matched adjacent normal. We generated male and female reference coverage profiles by taking the median depth at every SNP position across gender-matched RRBS data from the epiTRACERx cohort. This female reference profile was used to compute LogR for CRUK0047.

Next, normalised LogR values are corrected for sequencing coverage biases due to variation in *MspI* fragment length, GC content and replication timing across the genome (**Figure 2.4**). Technical and biological biases affecting sequencing coverage can differ between the normal and bulk tumour data and so normalisation is not sufficient to fully remove these confounders.

In order to obtain the corresponding fragment length for each single end read, we reconstruct the underlying *MspI* fragment distribution for each patient using the matched normal. We note the absence of CCGG motifs within RRBS reads, confirming complete enzymatic digestion. Thus, covered CCGG motifs correspond to *MspI* fragment ends. To identify these boundaries, we create a list of all CCGG motifs from the reference genome accounting for SNP forming/destroying cleavage sites and evaluate CCGG coverage at each of these by compiling read counts in support of the following trinucleotides at the 5' end: CGG(+), TGG(+), CCG(-) or CCA(-). Having identified the fragments supported by reads (i.e. both ends have total CCGG counts greater than 0), we annotate their respective lengths.

Next, for replication timing correction, we leverage publicly available cell line Repli-seq data from the ENCODE project (Dataset GEO Accession: GSE34399). We generate 15 reference profiles, one for each cell line, by calculating the replication timing at every 1000 genome SNPs. The replication timing profile which is the most correlated with the observed coverage biases is selected for each tumour sample and used to compute the replication timing at every SNP position.

GC content values are generated per fragments using the bulk tumour reads and adjusting the reference GC content for bisulphite conversion and methylation rate. When fragments are longer than 2x the read length (i.e. $\geq$200bp), we make the following assumptions at bases that are not covered by any sequencing reads: (1) Cs outside CpG dinucleotides are converted to thymines and (2) CpGs have methylation rates equal to the mean of all CpG loci supported by reads in the same fragment.

Finally, we fit the observed LogR to a linear combination of the natural splines (df = 5) of *MspI* fragment length, replication timing and GC content. The model residuals provide corrected LogR values.

Calculating the BAF at SNP $i$ requires the number of reads supporting the alternate allele, $r_{alt,i}$, and the reference allele, $r_{ref,i}$, $BAF = \frac{r_{alt,i}}{r_{alt,i} + r_{ref,i}}$. Strand-specific BAF rules allow to distinguish between the methylated or unmethylated reference and the alternate allele and 1000 genome SNPs. We explain these calculations in detail for all SNP types in the results section (**Section 2.2**). Unexpected alleles can arise due to rare polymorphisms, SNVs or misalignments. We exclude SNPs from downstream analyses when the unexpected allele makes up more 5% of the total allele counts at a given SNP.

Genotyping is performed on the patient-matched normal and conservative heterozygosity boundaries ($0.3 \geq BAF_n \geq 0.7$). At heterozygous SNPs, given high enough tumour purity and sequencing coverage, copy number segments with clonal allelic imbalance will generate two distinct BAF bands. ASCAT.m can assign heterozygous SNPs to the gained or lost alleles and phases all SNP alleles in the copy number segment. If SNP phasing is available for any one sampled tumour region,

this information is relayed to all of the other samples from the same patient. These haplotypes can be used to identify mirrored subclonal allelic imbalance [112] and to rescue signal in tumours where there is allelic imbalance but no clear separation. NSCLC tumour purity is usually low and RRBS data has wider BAF bands than matched higher coverage WGS and WES data, further increasing the importance of haplotyping to rescue signal.

ASCAT.m partially phased BAF and corrected LogR estimates are fed into ASCAT [117] piece-wise constant segmentation function (penalty = 200) leveraging germline heterozygous SNPs and copy number fitting functions (gamma = 1) to obtain allele-specific copy number profiles and purity estimates for each tumour region. For CRUK0047, heterozygous SNPs are identified as all SNPs with tumour BAF above 0.15 and below 0.85. Note that we set the minimum germline coverage for SNP inclusion to 10 whilst one read is deemed sufficient in the tumour. In the tumour, we identify SNPs with coverage below 10, but only exclude them if their nearest neighbours within a 10kb moving window display coverage above this threshold. In this manner, we only remove isolated low coverage singletons. This prevents creating bias against homozygous deletions.

Ploidy was QC'ed by leveraging intra-tumour information. 6 patients were flagged by QC for having large intra-tumour differences in their ploidy estimates. In each of those cases, we looked for evidence supporting an alternative ploidy solution in the purity and ploidy solution matrix. If such a better suited solution existed in line with the the overall tumour ploidy profile, the ASCAT copy number fitting step was re-run forcing the solution to the ploidy towards a diploid or tetraploid solution (setting MINPLOIDY and MAXPLOIDY boundary variables). We refit 7 tumour copy number profiles in total. In comparison, 17 samples were manually curated and re-run in the matched exome sequencing data [112].

#### *2.4.3.2 Tumour copy-number profiling from WGS data*

Base counts were extracted at 1000 genome SNPs subset for having coverage in the matched RRBS data using alleleCount (http://cancerit.github.io/alleleCount/) on the 10 WGS samples from patients CRUK0031, CRUK0062 and CRUK0069

(7 multi-region tumour and 3 adjacent normal samples). ASCAT GC content and replication timing LogR corrections were performed for each tumour sample. The normal genotype, tumour BAF and corrected normalised LogR values were fed into ASCAT piece-wise constant segmentation (penalty = 200) and obtained copy number and purity (gamma = 1). Segmented BAF, LogR and final allele specific copy number profiles generated from WGS and RRBS data were compared.

#### 2.4.3.3 Comparing RRBS and WGS-derived SNP genotypes

Base counts obtained by ASCAT(.m) at 1000 genome SNPs were recycled for genotyping purposes. Genotypes derived from WGS and RRBS of normal samples from patients CRUK0031, CRUK0062 and CRUK0069 were compared. Because short read WGS data is well-established for use in genotyping, the positive heterozygous SNPs ($0.1 \geq BAF_n \geq 0.9$) calls generated from the WGS data of the 3 patient's normal samples are considered to provide ground truth. A minimum SNP read depth of 10 required on both platforms for inclusion. The mean ASCAT.m false positive rate (FPR) is the number of erroneous heterozygous SNP assignments. It is computed as the sum of all false positives over the sum of false positive and true negative across samples. Similarly, the mean false negative rate (FNR) is taken as the sum of all false negatives divided by the sum of false negatives and true positives. Three different contexts are evaluated: (i) SNPs at CCGG, the recognition sequence of the *MspI* enzyme used to digest DNA and enrich for CpGs during library preparation, (ii) SNPs at CpGs (excluding CCGG motifs), and (iii) all other SNPs. A chi-square analysis is performed to test whether or not the FNR is context dependent. Next, we also evaluate the impact of heterozygosity boundaries and minimum coverage thresholds on both FNR and FPR. Finally, we take a deep look into the various sources of false negative homozygous SNPs.

#### 2.4.3.4 Determining regions of copy number gains and losses

First, the genome was divided into 10 Mb bins. For each chromosome, the first bin was set to the genomic position with the lowest index that generated copy number information in at least one sample. Next, each bin was classified as a gain, loss, both or neither for each sample. After compiling pan-cohort information, the

fraction of samples with gains and/or losses for each 10 Mb genomic bin was calculated. This fraction was divided by the number of samples with copy number information overlapping this genomic region to get the fraction samples with losses or gains at that loci. Regions were deemed neutral if the allele-specific copy number was *major allele* + *minor allele* $= 1 + 1$ in diploid (2n) or $2 + 2$ in whole-genome doubled (4n) samples. A bin was considered to harbour a copy number gain when the copy number of the major allele was $\geq$2 in 2n and $\geq$3 in 4n samples. Loss of heterozygosity events were classified as losses, irrespective of the major allele copy number. In WGD tumours, segments of copy number *major allele* $+1$ have lost one copy of the minor allele and are therefore classified as losses. For example, $2+1$ is a gain in 2n tumours but a loss in 4n tumours while $3+1$ is classed as a gain in both 2n and 4n samples and also as a loss in 4n tumours. Indeed, certain allele-specific copy number states are achieved by both gains and losses and should be counted as such.

# Chapter 3

# Copy-number aware methylation deconvolution and analysis of cancer

## 3.1 Introduction

### 3.1.1 Bulk tumour methylation deconvolution methods

In the previous chapter, we presented a simple two-component model which defined the bulk tumour as a mixture of aberrant and normal cells. We described the relationship between tumour DNA content and methylation rates at tumour-normal differentially methylated CpGs (**Figure 2.1**) and concluded that tumour purity and copy number must be accounted for correct interpretation of bulk cancer methylomes, in line with published reports [118–121].

Separation of aneuploid cancer cells from diploid normal populations is possible by FACS based on nuclei staining and scatter angles [125]. Crucially, prior knowledge of tumour-specific protein markers is additionally required to distinguish near diploid tumour populations from admixed normal cells since nuclei staining is uninformative. Moreover, FACS is low throughput and not usually scalable to large cohorts due to time and costs and, to our knowledge, this technique has never been used in combination with bisulphite sequencing. This is potentially because sorting puts added stress on nuclei, which combined with bisulphite-driven DNA degradation [143], may considerably reduce DNA quality. Large amounts of input material are likely needed to compensate for this. A computational tumour purification alternative would therefore save researchers time, costs and precious tumour material.

A handful of tools have been developed to computationally interrogate normal contamination levels from bulk tumour methylation sequencing and array data [105, 106, 144]. As introduced in the first chapter (section 1.7), the BED algorithm identifies the number of unique methylation sequences of *n* CpGs present at a given genomic locus, termed epialleles, and quantifies methylation ITH from the Shannon entropy [106]. In addition, epialleles from a patient-matched non-aberrant lung tissue sample can be used as proxy for the tumour contaminating normal cells and removed from the bulk weighed for tumour purity yielding the pure tumour epiallele distribution. Despite having only been tested on a single squamous cell carcinoma primary tumour with multi-region RRBS data, BED could be modified to account for copy number and is a potentially powerful approach to extract purified tumour methylomes and study copy- and allele-specific methylation. InfiniumPurify and MethylPurify do not extract purified tumour methylation rates but do account for tumour purity in differential methylation analysis, a key part of cancer research, and do so in a reference free fashion. All approaches discussed ignore tumour copy number profiles which leads to increased false positives and negatives rates [144].

### 3.1.2 Differential methylation analysis from bisulphite sequencing data

Differential methylation analysis is key to understanding the cancer methylome. DMP and DMR calling methods from bisulphite sequencing data are reviewed by Hebestreita and Klein [91]. Bisulphite sequencing yields read count data and thus DMP calling from RRBS and WGBS is often based on Beta or Beta-Binomial regression models. DMR calling is reportedly most successful when it builds on DMP calls as opposed to *de novo* differential methylation analysis of regions. There is however consensus as to the best way to bin CpGs (and hence DMPs) into regions for DMR calling. Some researchers prefer binning CpGs by genomic features while others opt for a feature-agnostic approach.

Spatially separated bins are easily determined from RRBS data due to sparsity. RRBS is also particularly well-suited to assay (potentially) disease-causing changes in methylation and dynamically regulated CpGs, which are usually concentrated in

CGI shores and at TFBSs [33, 145]. Compared with WGBS, RRBS data is a cost-effective way to query aberant methylation at regulatory CpGs.

### 3.1.3 Chapter summary

In this chapter, we present Copy Number-Aware Methylation Deconvolution Analysis of Cancer (CAMDAC), a novel computation tool for deconvolution of pure tumour signals and allele-specific methylation analysis, and apply it to multi-region bulk tumour RRBS from the epiTRACERx cohort totalling 122 tumour and 37 tumour-adjacent normal samples from 38 NSCLC patients (section 2.4.1).

To begin with, we show that bulk tumour and normal methylation rates are affected by polymorphisms and develop SNP-independent methylation rate calculation rules based on strand-specific dinucleotides counts. The output methylation rate is the average per CpG allele and importantly is bound by [0,1], regardless of heterozygous SNPs.

Next, we define the bulk tumour as a two-component mixture of tumour and normal cell populations. Assuming the normal is a reasonable proxy for the normal contaminating cells, we visualise tumour-normal DMPs populations by plotting bulk methylation rates distributions thresholding on the normal methylation rate and stratifying CpGs by sample purity and ASCN. We formalise the relationship between methylation rates and tumour DNA content into CAMDAC equation 2 and validate this model by comparing the observed and predicted DMP population peak position across samples and copy number states. We demonstrate that predictions fit the observed data with noise proportional to tumour purity. CAMDAC purified tumour methylomes successfully remove shared normal signal, decreasing correlations between patients and to normal after purification, while sampled regions from the same tumour remain closely connected. Where possible, we obtain SNV deconvoluted tumour methylation rates and compare them with CAMDAC $m_t$. We observe good agreement between the two approaches, validating our tool.

Subsequently, we describe our Bayesian DMP calling method. We demonstrate from simulated and real data that differential methylation analysis from purified methylomes considerably reduces false positive and negative rates. Finally,

we present FACS tumour RRBS data for a subset of epiTRACERx bulk tumour samples, separating tumour and normal contaminating cells. We compare the composition of the FACS and bulk normal and conclude the adjacent normal is a suitable proxy for the tumour infiltrating normal cells. Because the adjacent normal has a significant epithelial component, it is an appropriate substitute for the NSCLC cell of origin in differential methylation analyses. Finally, we compare tumour-normal DMP calls based on bulk and CAMDAC pure tumour methylation rate with those obtained from FACS-purified tumour methylomes and see better overlap post-deconvolution.

## 3.2 Results

### 3.2.1 SNP-independent methylation rate estimation

In the previous chapter, we showed that the allele distribution at SNPs overlapping with CpG loci are affected by methylation. Conversely, we posit that polymorphisms confound methylation rates. Across the epiTRACERx RRBS dataset, heterozygous CpGs account for 11,026 (0.31%) of CpG sites, 83.3% of which are CpG>TpG polymorphisms and methylation rates at these positions using default approaches show markedly different distributions (**Figure 3.1**A-C). In tumour samples, methylation at heterozygous CpG>TpG shows copy number dependence (**Figure 3.1D-F**).

Addressing this, we propose an new set of rules to compute methylation rates accounting for SNP status (**Methods, 3.4.2**). (1) We define methylation rates as the average methylation per CpG allele ensuring the methylation rate at a polymorphic CpG can take any value between 0 and 1, rather than, for example, between 0 and 0.5 in a diploid region. (2) In directional bisulphite sequencing protocols, the original top and bottom strands encode methylation information for the first and second position of a CpG, respectively. Therefore, we can differentiate (un)methylated CpGs based on TG(+) and CA(-) versus CG(+) and CG(-) dinucleotide counts (**Figure 3.2A**). (3) At CpG>TpG SNPs, reads supporting the unmethylated CpG and the CpG-destroying alleles are only separable on the reverse strand (CA(-) *ver-*

*sus* TA(-), **Figure 3.2B**). Likewise, at CpG>CpA SNPs, only the top strand may be used to unambiguously distinguish alleles (**Figure 3.2C**).

Excluding counts supporting the CpG-destroying allele, we compiled strand-specific dinucleotides and calculated methylation rates per CpG allele as: $m = \frac{CG}{CG+TG(+)+CA(-)}$, except at CpG>TpG and CpG>CpA SNPs, where only reads from the bottom strand and top strand, respectively, inform the estimates (**Figure 3.2B-C**). When accounting for these confounders, intermediate methylation signals are removed (**Figure 3.1A-F**), confirming that CAMDAC methylation rates are robust to polymorphisms.

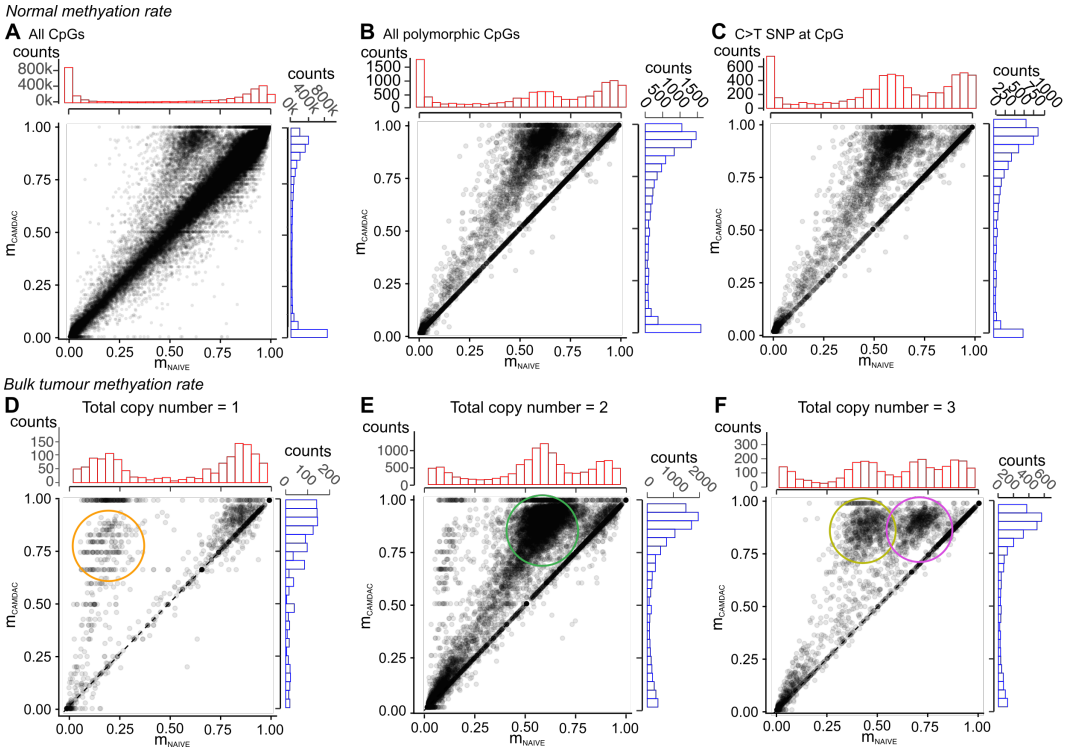


**Figure 3.1: Naïve *versus* CAMDAC polymorphism independent methylation rates.** (A-C) CAMDAC compared with naïve normal methylation rate estimates at (A) all CpGs, (B) polymorphic CpGs and (C) CpG>TpG SNPs, selected from a random sample of 3,000,000 CpGs from this cohorts 37 normal lung samples. (D-F) CAMDAC versus naïve bulk tumour methylation rate estimates for CpG>TpG SNPs in segments with total copy number (D) 1, (E) 2 and (F) 3, pooling the data from all 3 sampled regions from patient CRUK0084 of near-equal and high tumour purity (range 0.85-0.87). The data points highlighted by the orange, green and yellow circles indicate heterozygous C>T SNPs with CpG allele copy number 1 and the pink circle CpGs with copy number 2.

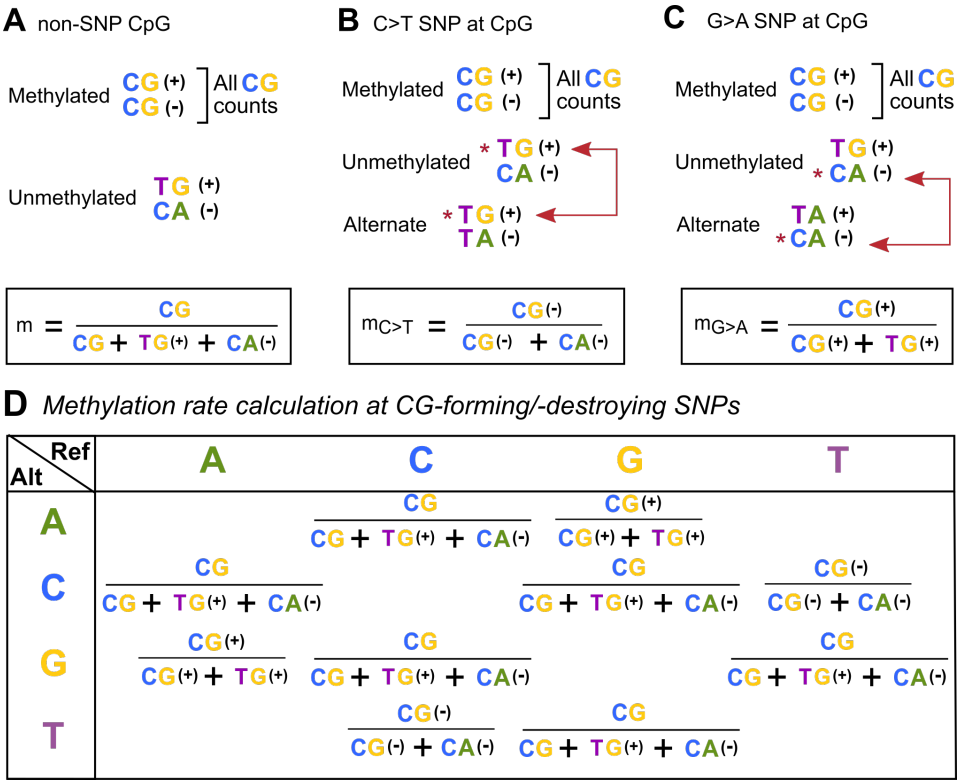


**Figure 3.2: Rules for calculating polymorphism independent methylation rates with CAMDAC.**
(A) Derivation of methylation rate estimates at non-polymorphic CpGs. (B-C) Derivation of the CpG-forming allele-specific methylation rate at a (B) CpG>TpG and (C) CpG>CpA SNP. (B) Methylation rate formulae for all possible polymorphic CpGs.

### 3.2.2 Bulk tumour methylation rates are affected by tumour purity and copy number alterations

Analysis of the epiTRACERx cohort with ASCAT.m revealed substantial variability in tumour copy number and purity, in line with previous work using WES of the same samples [112]. Here, we take advantage of this to assess the effect of both these confounders on bulk methylation rates at differentially methylated positions.

In the absence of a robust method to identify differential methylation independently of normal contamination and copy number, we first developed a method to visualise candidate DMPs. Selecting CpG loci that were confidently unmethylated in the normal contaminating cells, using the patient-matched adjacent normal as a proxy ($m_n$ posterior 99% highest density interval ($\text{HDI}^{99}$) $\subseteq$ [0, 0.2], **Figure 3.3A-B**), and plotted the bulk tumour methylation rates at those positions, unpolluted

by signals from non-aberrant cells. Most of these sites were also unmethylated in the bulk tumour (**Figure 3.3C**), exhibiting methylation rates close to 0 (88% with $HDI^{99} \subseteq [0, 0.2]$). This is as expected and suggests that most sites are not differentially methylated between tumour and normal cells. Zooming in, we detect gain of methylation in a subset of CpGs (**Figure 3.3D**). Leveraging copy number from ASCAT.m, we see the modal peak of differential methylation shift with tumour copy number (**Figure 3.3E-F**).

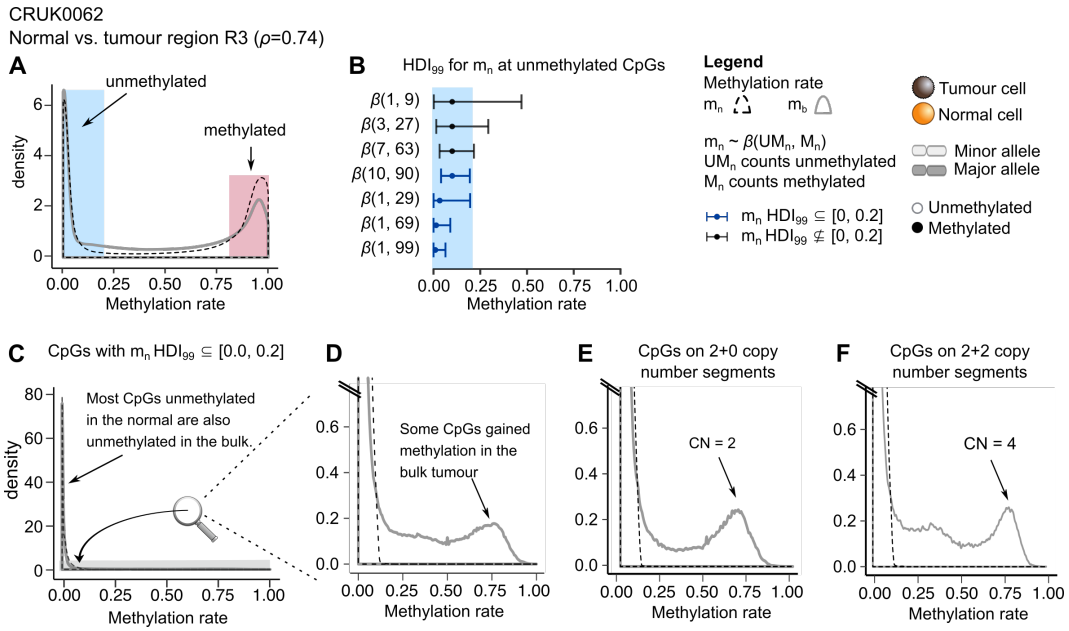


**Figure 3.3: Bulk tumour methylation rates at CpGs confidently unmethylated in the matched normal.**
(A) Normal and bulk tumour methylation rate density distributions for sample CRUK0062-R3 and adjacent patient-matched normal showing most CpGs are either fully methylated or unmethylated. A slight increase in the number of CpGs with intermediate methylation levels is found in the bulk tumour. (B) Simulated 99% Highest Density Interval ($HDI^{99}$) at example unmethylated CpGs with variable number of methylated and unmethylated read counts and total CpG coverage. (C-F) Normal and bulk tumour methylation rate ($m_n$ and $m_b$, respectively) histogram for CpGs labelled as confidently unmethylated in the normal ($HDI^{99} \subseteq [0,0.2]$). The majority of these loci are also unmethylated in the bulk tumour. (D-F) Zoom-in, highlighting CpGs with gained methylation in the cancer cells across all copy numbers (D), and in regions of copy number states 2+0 (E) and 2+2 (F).

To appreciate the combined effect of normal contamination and tumour copy number, we again selected CpGs that were confidently unmethylated in the normal ($HDI^{99} \subseteq [0, 0.2]$) and further stratified them by ASCAT.m allele-specific copy number state for each of three tumour regions with different purity and visualised their bulk tumour methylation rates (**Figure 3.4**). A bi-allelic DMP population is apparent across all histograms and its modal methylation rate varies along with tumour purity and total copy number. Similar observations can be made from profiles which we generated by selecting CpGs that were confidently methylated in the normal ($HDI^{99} \subseteq [0.8, 1]$, **Figure S7**).

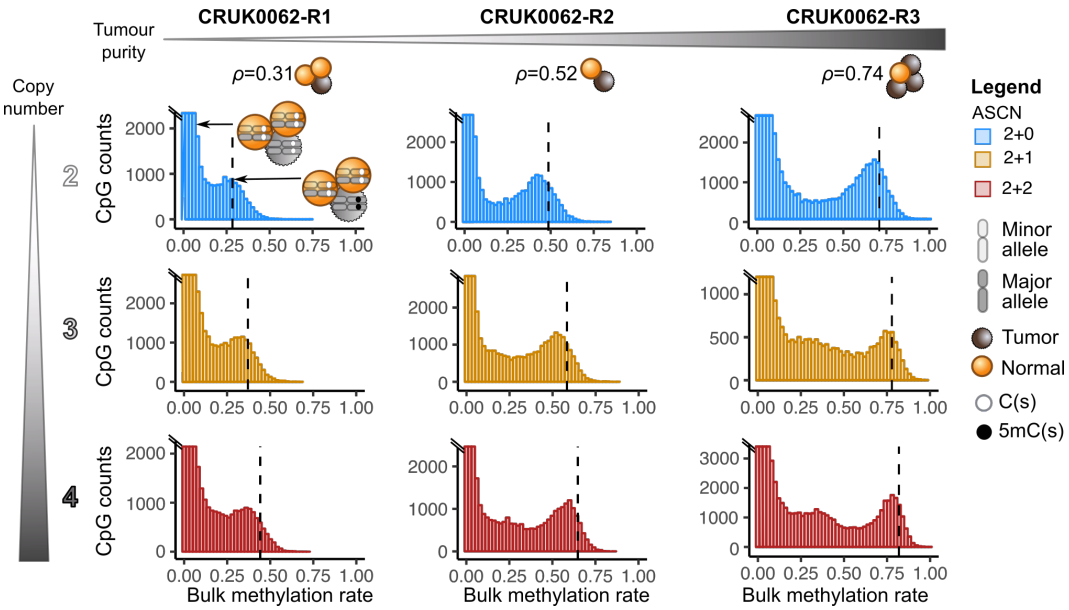


**Figure 3.4: Tumour purity and copy number affect methylation rates.**
Bulk methylation rate histograms for tumour regions 1-3 of patient CRUK0062, for CpGs which are confidently unmethylated in the adjacent normal sample. CpGs are stratified by allele-specific copy number. A dashed line indicates the expected mode of the methylation rate peak corresponding to clonal differentially methylated CpGs on all copies ($m_t = 1$).

These results suggest that correct interpretation of bulk tumour intermediate methylation signals requires consideration of both tumour purity and copy number. To address this, we present CAMDAC, a computation tool enabling correction of bulk tumour methylation rates for these confounders to yield the pure tumour methylomes, leveraging tumour purity and copy number estimates from ASCAT.m.

### 3.2.3 Modelling bulk tumour methylation rates

In the introduction to the previous section, we briefly showed that when a tumour clone is methylated at CpGs that are unmethylated in the normal contaminating cells, the ratio of methylated reads at that locus obtained from bulk bisulphite sequencing will rise with tumour purity and copy number (**Figures 2.1B, 3.4**). Conversely, if the tumour clone is hypomethylated with respect to the admixed normal populations, we record a decrease in methylation rate with increasing purity and copy number (**Figure S7**). We formally define this relationship modelling the bulk tumour methylation rate ($m_b$) as the sum of a tumour component (with methylation rate $m_t$) and a normal component (with methylation rate $m_n$), weighted for their relative DNA contributions, which can be calculated as the product of purity and copy numbers ($\rho$ and $1-\rho$, and $n_t$ and $n_n$ for tumour and normal, respectively):

$$m_b = \frac{\rho n_t m_t + n_n m_n (1-\rho)}{\rho n_t + n_n (1-\rho)} \tag{2}$$

Applying this equation, we compute the expected bulk tumour methylation mode for clonal bi-allelic DMPs across our lung cancer RRBS data (represented as a dashed line in **Figures 2.1B, 3.4**). We then compare our prediction with the observed $m_b$ distribution peak stratifying CpGs by sample purity, copy number and normal contaminant methylation state, using the adjacent normal samples as a proxy for the admixed normal cells. For each subset of loci, we carry out separate beta regression to extract the mode of the observed $m_b$ distribution peak generated by DMPs present on all tumour copies. Setting a minimum effect size threshold is necessary to obtain a solution at low tumour purities, although this may artificially increase the distance between DMPs and non-differentially methylated CpGs. The estimated mode of the $m_b$ peaks representing clonal bi-allelic tumour-normal DMPs align closely with those theoretically predicted under equation 2, for different values of tumour purity and copy number, and holds across all tumour samples (**Figure 3.5**). We conclude that the two-component model portrayed in equation 2 explains the observed bulk tumour methylation rate distribution in our lung cancer cohort.

The median absolute differences between the expected and the observed fit-

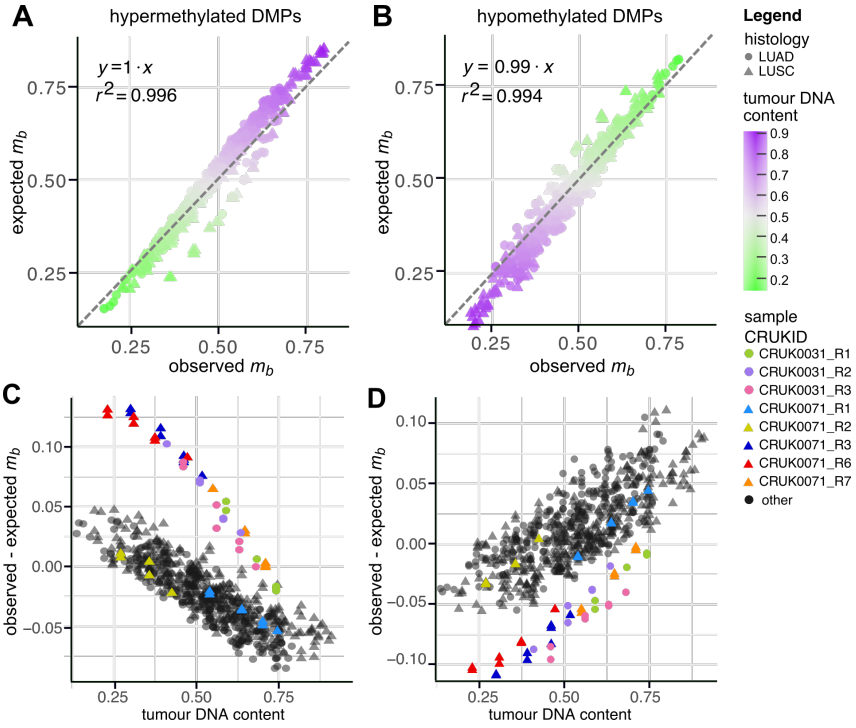


**Figure 3.5: Validation of CAMDAC equations across sample purities and copy number states.**
(A) Comparison of the predicted methylation rate under the CAMDAC equations as a function of purity and copy number state, and the observed peak position for clonal bi-allelic tumour-normal hypermethylated CpGs that are unmethylated in the adjacent normal sample. (B) As in A, but for unmethylated CpGs that are methylated in the adjacent normal sample. (C-D) Observed minus expected $m_b$ peak position in relation to tumour DNA content at clonal hyper- (C) and hypomethylated (D) CpGs across copy number states and tumour samples. Samples from patients CRUK0031 and CRUK0071 are highlighted.

ted values was 0.0254 (range -0.131 to 0.109) and is correlated with tumour DNA content (**Figure 3.5C,D**), calculated as $\frac{\rho n_t}{\rho n_t + n_n(1-\rho)}$, at both hyper- (Pearson correlation = 0.703, p-val = $5.43\times10^{-96}$) and hypomethylated DMPs (Pearson correlation = -0.692, p-val = $7.42\times10^{-92}$). At low tumour fraction, there is significant bleeding of non-differentially methylated CpGs into the DMP peaks, and thus setting a minimum effect size threshold to fit the observed data can bias the fitted estimates removing signal (**Figures 3.4, S7**). At high tumour fraction, DMP populations form well-defined clonal bi-allelic peaks in the $m_b$ distribution. The observed is systematically shifted away from the expected and towards non-differentially methylated CpGs. We hypothesise that methylation erosion in rapidly dividing tumour cells may play a role in this effect. Nevertheless, the reported errors are small

in comparison to absolute methylation changes at clonal epimutations and thus unlikely to affect downstream analyses. We note that samples from CRUK0031 and a subset from CRUK0071 show noticeably larger errors than the rest of the cohort. It is apparent from the bulk tumour distribution that the observed modal value computed by beta regression is not a good a fit for the observed data in those samples, explaining the discrepancy with CAMDAC equations.

Overall, the predicted modal methylation rate under CAMDAC equations at clonal bi-allelic DMPs was in agreement with the observed data. This suggest that differential methylation is the most important driver of intermediate methylation, at least at the scale of our non-small cell lung cancer cohort.

### 3.2.4 Deconvolution of bulk into pure tumour methylomes

CAMDAC equation 2 formalises the relationship between the bulk, purified tumour and normal methylation rates in relation to tumour DNA content and is a good CAMDAC model for our NSCLC dataset. Leveraging estimates of tumour purity and copy numbers directly from RRBS obtained with ASCAT.m and assuming the adjacent normal is a suitable substitute for the methylation rate of the tumour polluting normal cells, we have all the necessary variables to solve equation 2 for $m_t$ and obtain deconvolved tumour methylation rate estimates:

$$m_t = \frac{m_b(\rho n_t + n_n(1-\rho)) - n_n m_n(1-\rho)}{\rho n_t} \tag{3}$$

We applied CAMDAC to the epiTRACERx cohort and assessed the output purified methylomes. After deconvolution, CpGs that have become clonally methylated on all copies in tumour cells are expected to have purified tumour methylation rates close to $m_t = 1$. *Vice versa*, tumour-normal hypomethylated loci should approach $m_t = 0$. We see a high correlation between the expected and the observed purified tumour methylation rates at these clonal differentially methylated loci present on all copies (**Figure 3.6A**). The error on the CAMDAC predictions is proportional to tumour DNA content (**Figure 3.6B-C**), with the noise decreasing along with tumour sample purity (**Figure 3.6D**) causing the error to plateau at $\pm 0.1$.

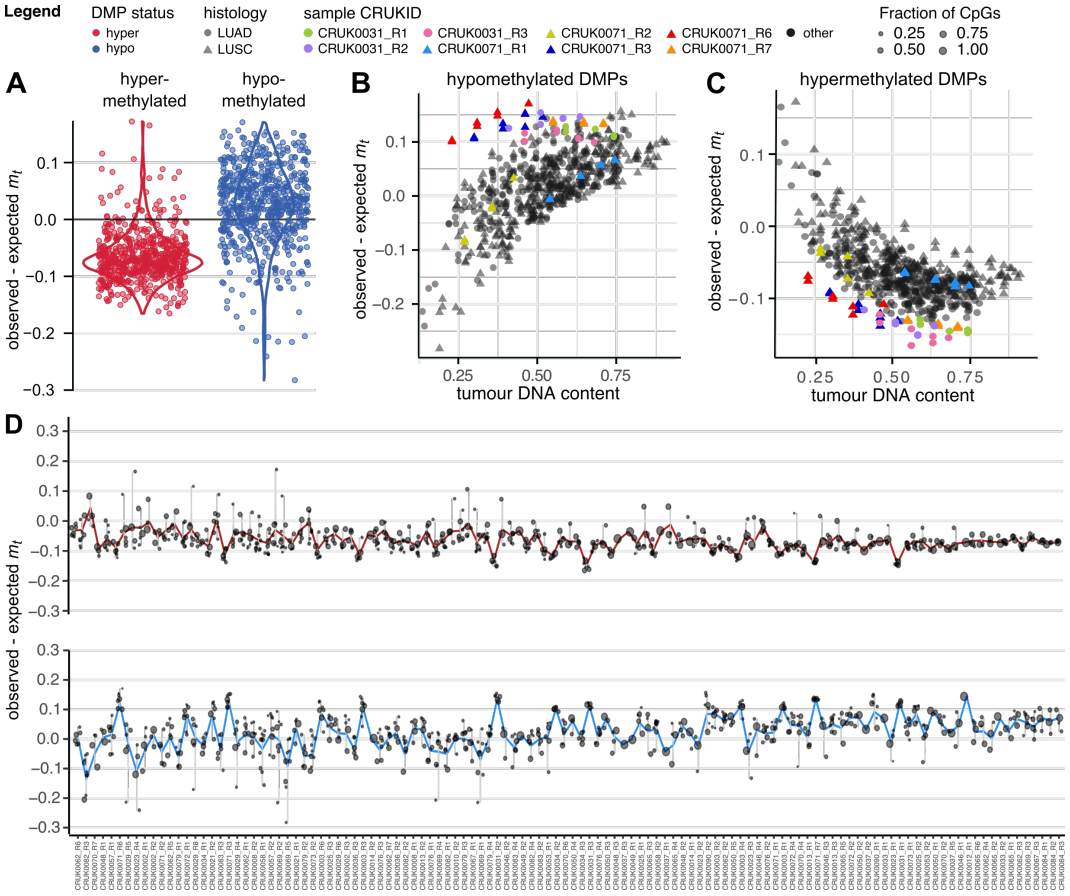


**Figure 3.6: Validation of CAMDAC purified $m_t$ at DMPs.**
(A) Observed deconvolved tumour methylation rate at clonal hypo- (blue) and hypermethylated (red) CpGs on all tumour chromosome copies compared with the expected, taken as the mean of the adjacent normal beta distribution peak for unmethylated and methylated CpGs respectively. (B-C) Observed minus expected $m_t$ peak position in relation to tumour DNA content at clonal hyper- (B) and hypomethylated (C) CpGs across copy number states and for all tumour samples. Samples that were outliers in $m_b$ are highlighted. (D) Observed minus expected deconvolved tumour methylation rate at clonal hyper- (top) and hypomethylated (bottom) per tumour sample.

We showed that our CAMDAC model can account for copy number and tumour purity to correctly predict the mode of clonal bi-allelic DMPs. However, we often noted the presence of distinct DMP populations forming peaks between the bi-allelic and non-differentially methylated clusters, suggesting that these CpGs were aberrated on a subset of tumour copies. For example, this signal is visible in the methylation rate distributions of DMPs stratified by allele-specific copy number and normal methylation rate in samples with sufficiently high purity (**Figures 3.4, S7**). Allele-specific methylation has been reported in healthy cells and is known

to play an important role in chromosome X inactivation in females, at germline imprinted genes and at polymorphic regulatory sequences [146, 147]. DNA methylation ITH could also contribute to intermediate methylation signals, but published work on the same samples showed little subclonal signal within single tumour biopsies, with most clustered mutations appearing clonal in individual samples and only found to be subclonal after multi-sample analyses [112]. We therefore assume that subclonal signal will not substantially alter methylation rates and ignore this source of intermediate methylation going forward.

We sought to assess CAMDAC purified methylomes by comparing Pearson correlations between sample pairs. Sample pairs from the same tumour and patient have shared clonal origins and so were expected to have lower distances than between patient or tumour-normal pairs. In line with predictions, deconvolved tumour methylomes showed increased distances to matched normals and between tumour samples from different patients compared to bulk signals, while samples from the same patient remained correlated (**Figure 3.7**). From this, we conclude that

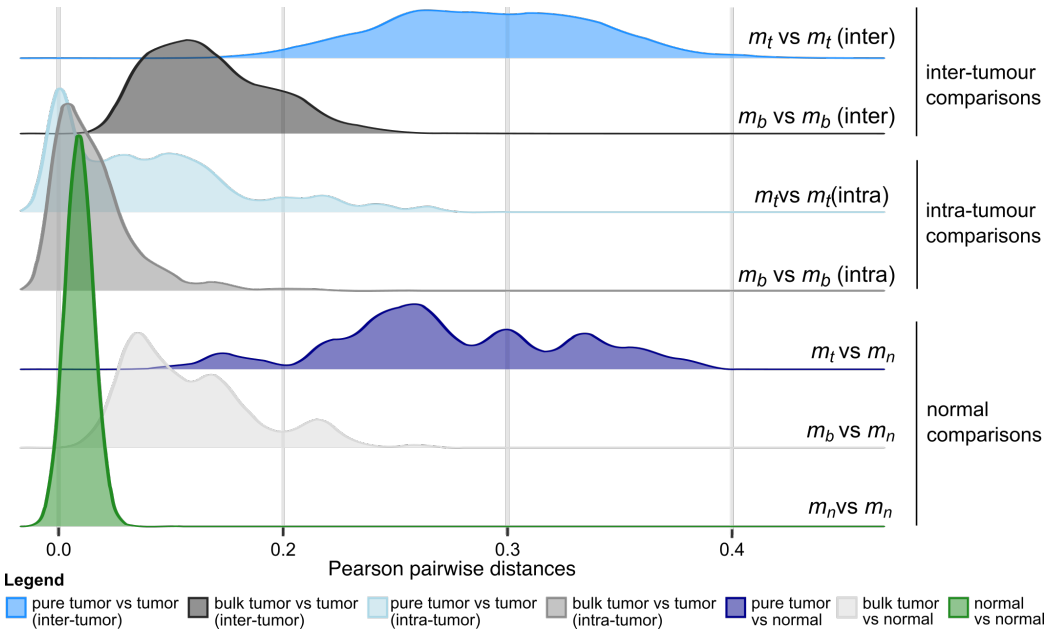


**Figure 3.7: Comparing normal, bulk tumour and CAMDAC purified methylomes.** Correlation between bulk tumour, CAMDAC deconvolved tumour, and adjacent normal methylation profiles, separating samples from the same (intra-tumour) *versus* different patients (inter-tumour).

CAMDAC successfully deconvolves tumour-specific signals, removing the shared normal component from the bulk profiles.

### 3.2.5 Comparing CAMDAC and SNV purified methylomes

We validate CAMDAC pure tumour estimates by phasing CpG methylation to clonal SNVs present on the same read and on all copies in regions with loss of heterozygosity called from newly obtained WGS data and previously published WES [112]. At these sites, tumour reads must report the variant allele, and methylation rates obtained from this subset of reads can be used as an unbiased estimate of the pure tumour methylation rate. The variant allele frequencies of somatic genetic alterations were computed from bisulphite sequencing data using the same rules devised for BAF calculation at germline SNPs (**Figure 2.5**). Overall, RRBS- and matched WES/WGS-derived VAF estimates were highly correlated (Pearson correlation = 0.86, **Figure 3.8A**). We obtained phased methylation estimates for a total of 4,485 CpGs across our dataset and observed concordance between these SNV purified $m_t$ values and CAMDAC estimates (Pearson correlation = 0.97, **Figure 3.8B**). We thus conclude that deconvolution of bulk profiles with CAMDAC leads to accurate pure tumour methylation rate estimates.

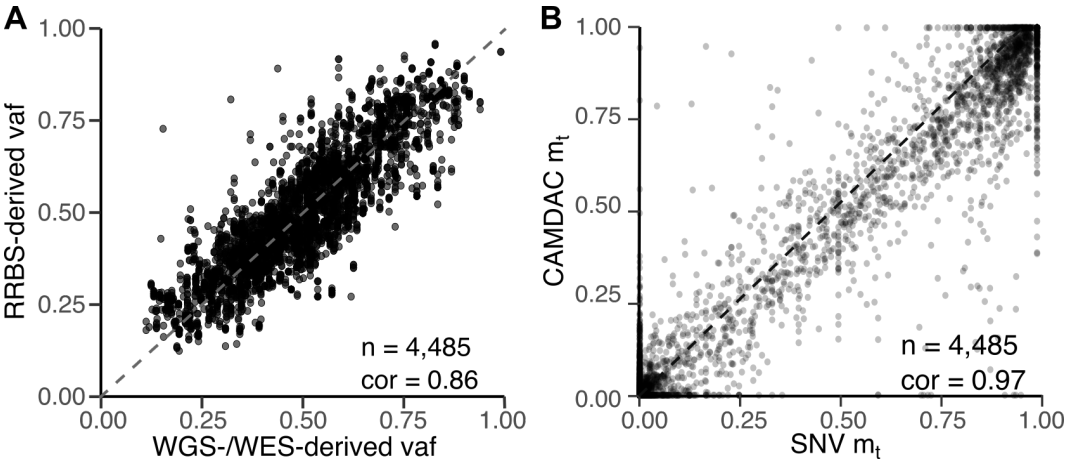


**Figure 3.8: SNV deconvoluted versus CAMDAC $m_t$ estimates.**
(A) Comparison of variant allele frequencies of single-nucleotide variants derived from RRBS and WES/WGS in regions of loss of heterozygosity. (B) Scatter plot comparing CAMDAC $m_t$ with SNV purified methylation rates. Pearson correlation is displayed.

### 3.2.6 Inferring differential methylation from CAMDAC purified methylation rates

So far, we assessed methylation rates at DMPs by setting thresholds on the matched normal, stratifying CpGs by copy number state and sample purity, and modelled peaks of differential methylation. Here, we develop a formal approach for tumour-normal DMP calling from CAMDAC purified tumour profiles and normal lung samples (as proxy for the normal cell of origin). We compute the probabilities of a CpG having gained or lost methylation during tumour evolution, $P(m_t > m_n)$ and $P(m_t < m_n)$, respectively (**Methods, section 3.4.5**). Replacing $m_t$ with equation 3, these probabilities can be expressed in terms $m_b$ and $m_n$, hence enabling use of a beta distribution Bayesian model based on bulk tumour and normal (un)methylated read counts ($UM_b$, $M_b$, $UM_n$ and $M_n$). The resulting probability density is a scaled difference of two beta posteriors ($m_n$ and $m_b$ $Beta(M_x, UM_x)$, where $x \in [n, b]$), calculated as (**Methods, section 3.4.5**):

$$P(m_t > m_n) = P(C(m_n - m_b) < 0) \qquad \text{where} \quad C = \frac{\rho n_t}{\rho n_t + n_n(1-\rho)} \tag{4}$$

CAMDAC equations 3 and 4 reveal the relationship between tumour purity, copy number and read depth. From this, we can deduce the effect of these variables on tumour–normal DMP detection power. Specifically, the standard deviation around the observed $m_n$ and $m_b$ increases with decreasing normal and tumour coverage, respectively. In other words, a reduction in coverage flattens the two Beta distributions. Int turn, higher tumour copy number leads to increases in local read depth reducing the variance of $m_b$ but also, together with increasing purity, widens the gap between bulk tumour and normal methylation rates at DMPs, further increasing power.

For tumour–tumour DMP calling, we calculated the probabilities $P(m_{t1} > m_{t2})$ and $P(m_{t1} < m_{t2})$, substituting $m_{ti}$ by equation 3 and resampling $m_t$ from the posterior distributions of $m_n$ and $m_b$ weighed for tumour DNA content (**Methods, section 3.4.6**).

In addition to the statistical test, a minimum methylation rate difference is usually required by differential methylation analysis tools to tease out biologically significant from statistically significant signals. We set this threshold to $|m_t - m_n| > 0.2$, which is relatively low compared with other cancer studies (e.g. 0.25 [148], 0.3 [149, 150]).

### 3.2.7 Evaluating CAMDAC performance on simulated DMPs

We discuss our tumour-normal and tumour-tumour DMP simulation outputs, including false positive and false negative rates and evaluate the impact of tumour purity, copy number and read depth on CAMDAC DMP calling power (**Methods, section 3.4.7**).

First, we created two sample sets of differing tumour DNA content by taking the 20 lowest and highest purity samples in our cohort, each group with purities $\rho < 0.3$ and $\rho > 0.58$, respectively. For all those samples, we compiled the copy number and coverage information at every autosomal CpG and noted the associated sample purity. We then sampled CpGs from this list, taking the bulk tumour coverages, copy numbers and sample purities to calculate the number of reads coming from tumour and normal cells. Next, we compiled vectors of confidently unmethylated ($HDI^{99} \in [0, 0.2]$) and methylated ($HDI^{99} \in [0.8, 1]$) CpGs from the adjacent normals. A normal methylation prior was randomly selected from these two vectors and the tumour methylation prior was either sampled from the same prior as the normal, to simulate non-differentially methylated loci, or from the opposite, to generate our true positive DMP set. We used these methylation priors in combination with the sampled coverages to generate the counts methylated in the tumour and normal. The bulk is finally computed as the sum of the normal and tumour read counts.

We generated equal numbers of inter- and intra-tumour sample pairs, from the same or across purity groups and copy number states $\in$ 1,2,3,4,5,$\geq$6, totalling 168 possible combinations each with 10,000 simulated CpGs. Every CpG pair is included in the tumour-tumour DMP analysis while individual CpGs are fed into tumour-normal DMP calling, setting the probability and minimum effect size thresholds to 0.01 and 0.2, respectively. Our results revealed that, in contrast to $m_b$,

thresholding based on $m_t$ removed dependence on purity and copy number, with true positive bi-allelic and mono-allelic DMPs (in balanced regions) showing absolute $m_t - m_n$ methylation rate difference near 1 and 0.5 respectively (**Figure 3.9**).

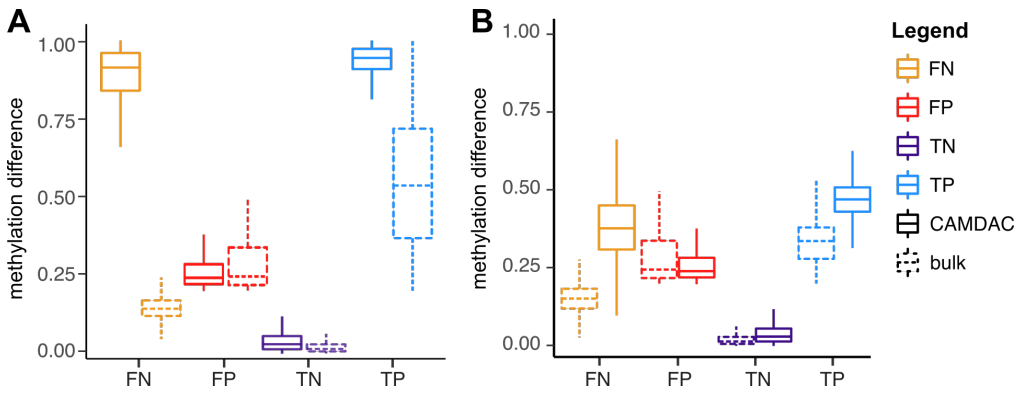


**Figure 3.9: Absolute tumour-normal methylation difference in simulated data.** Average absolute normal minus $m_t$ (solid line) and $m_b$ (dashed line) methylation difference at simulated bi-allelic (left) and mono-allelic (right) false negatives (FN), false positives (FP), true negatives (TN) and true positives (TP).

In comparison to $m_b$, setting the effect size on $m_t$ lowered the false negative rates in tumours with high normal contamination levels and at CpGs with low copy number (**Figure 3.10**), while retaining low numbers of false positives at all copy numbers ($FPR < 3.0 \times 10^{-3}$). Likewise, tumour-tumour DMP calls based on $m_b$ were highly polluted with false positive DMPs, while those derived from purity and copy number adjusted $m_t$ estimates were not. We conclude that use of CAMDAC $m_t$ values considerably reduces false positives in tumour-tumour comparisons while retaining a similarly low rate of false negatives (**Figures 3.11 and S8**).

To evaluate CAMDAC performance on real data, we compared intra-tumour DMPs called using $m_b$ or $m_t$ for patient CRUK0062, which we selected for its large number of samples with varying tumour purity (**Figure 3.12A**). In this setting, most CAMDAC $m_t$-based calls are also identified using $m_b$ and the effect of tumour purity on power can readily be seen as an increase in the number of DMPs identified with sample purity. Note however that, using $m_t$, more DMPs are called when both samples are high purity, i.e. statistical power is the highest. In contrast, when using $m_b$, more DMPs are called when two samples differ more in purity. These findings

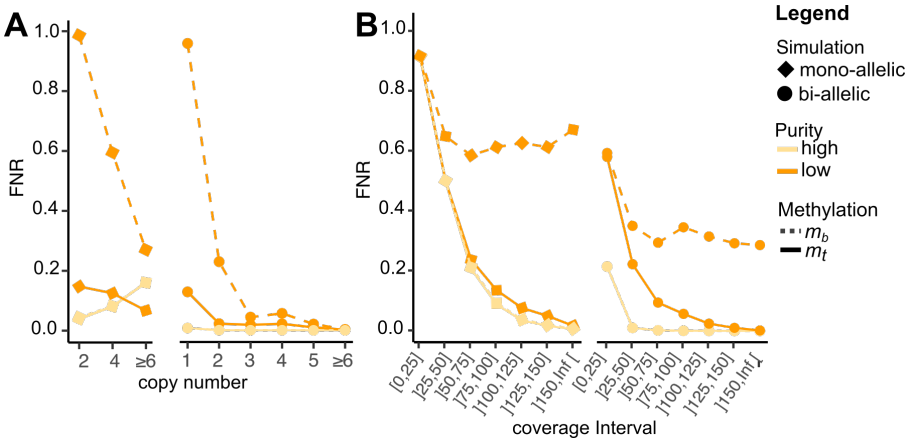


**Figure 3.10: Tumour-normal DMP simulation results.**
False negative rates at mono- (diamond) and bi-allelic (circle) DMPs for samples falling in the low (orange) and high (yellow) tumour purity categories using CAMDAC $m_t$ (solid line) and $m_b$ (dashed line) with tumour copy nuber (A) or by coverage interval (B).

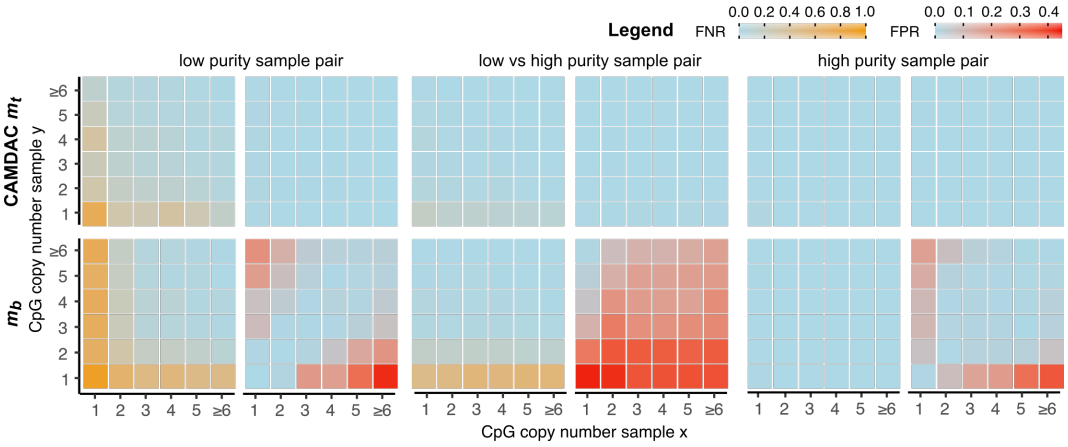


**Figure 3.11: Bi-allelic tumour-tumour DMP simulation results.**
False negative and false positive rates as a function of the copy number state, averaged across simulated sample pairs of low (left panel), low versus high (middle panel) and high (right panel) tumour purities with CAMDAC deconvolved (top row) and bulk (bottom row) tumour methylation rates.

are in line with our simulation results and suggest that also on real data, controlling methylation rates for tumour purity and copy number greatly reduces the number of false positive DMP calls, while maintaining low false negative rates.

To get a global overview of the performance of CAMDAC $m_t$ and bulk $m_b$ for tumour-tumour differential methylation analysis, we analyse simulated pairwise loci selected CpG loci from samples of low or high tumour purities both within and

between patients and obtained DMP calls (Methods). As expected, results showed a greater number of DMPs for inter-patient comparisons than between samples of shared clonal origins (**Figure 3.12B**). Furthermore, DMP calls unique to the bulk tumour were frequent between samples of differing purities taken from the same patient, suggesting a high false-positive rate of DMP calling without deconvolution.

Taken together, analyses of both simulated and observed data show that CAMDAC enables accurate calling of both tumour-normal and tumour-tumour differential methylation from RRBS data, accounting for both purity and copy number.

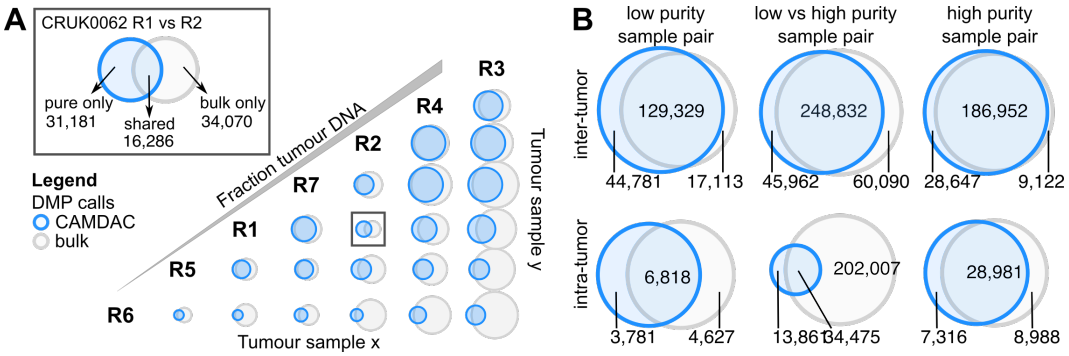


**Figure 3.12: DMP calling on real data.**
(A) Comparing observed tumour-tumour differential methylation calls between CRUK0062 regions from bulk (grey) and CAMDAC (blue) approaches. Samples are ordered by tumour DNA content. (B) Venn diagrams showing the overlap of tumour-tumour DMPs between bulk and CAMDAC call sets for intra- and inter-tumour DMPs.

### 3.2.8 Validation of normal lung as reference for CAMDAC and differential methylation

CAMDAC purification and differential methylation analysis modules require a proxy for the methylation rates of the tumour-infiltrating normal cells and the tumour-initiating cell, $m_{n,infil}$ and $m_{n,init}$, respectively. In this work, we substitute the tumour-adjacent patient-matched lung normal as an estimate of both these variables and thus refer to them interchangeably as $m_n$.

Deconvolution of cellular subtypes in LUAD and LUSC from bulk tumour RNA-Seq data suggests variable immune infiltration levels between sampled region of the same patient and across the cohort [114]. We posit that immune, fibroblast

and epithelial populations in the adjacent normal will roughly match the tumour composition if sampled within sufficient proximity. Studies in mouse and human have shown that the lung epithelium hosts at least 32 different cellular subtypes [151, 152]. While most of these components represent a small fraction of cells, lung tissue is enriched for two epithelial cellular subtypes, both of which happen to be a cell of origin of lung adeno- and/or squamous cell carcinoma. A surfactant-producing population in the alveolar space referred to as alveolar type II cells (AT2 cells) could be the cell of origin of lung adenocarcinoma [153]. The bronchiolar club (Clara) cells can presumably initiate both LUAD and LUSC [154, 155]. The bulk adjacent normal should therefore be densely populated by the NSCLC cells of origin and thus an adequate proxy for tumour-normal differential methylation analysis.

To test our hypothesis, we performed RRBS of diploid, $pop_D$, and aneuploid, $pop_A$, populations separated by FACS from 5 different patients selected for having sufficient material available and high tumour ploidy (**Methods, section 3.4.8**). Assuming diploid cells are all normal, we compared the cellular composition obtained from EpiDISH [96] on the sorted $pop_D$ and bulk normal data (**Table 3.1**). EpiDISH is a reference-based cell-type deconvolution approach designed for array data leveraging the methylation rates of epithelial, immune and fibroblast cells at 716 probes. On average, RRBS data overlaps with 27.23% (195) of these reference probes. The mean absolute difference between the adjacent normal and sorted populations from the same patient were 0.130, 0.167 and 0.178 for epithelial, fibroblast and immune components, respectively. The tumour-adjacent lung tissue had consistently higher epithelial cell fraction than the tumour-infiltrating normal cells separated by FACS suggesting that, of these two normals, it is the best substitute for the cell of origin of NSCLC (Wilcoxon paired test p-val = 0.0556).

Next, we obtained DMP calls between the tumour-adjacent normal samples and the patient-matched diploid population that were isolated by FACS to gauge their similarity. Overall, the mean absolute difference ($\Delta m$) between paired profiles was $\Delta m = 0.0672$ and on average 95.28% of CpGs were classified as non-

**Table 3.1: Comparison of the cellular composition of bulk and sorted normals.**

| CRUK ids | sample | design | epithelial | fibroblast | immune |
|---|---|---|---|---|---|
| CRUK0090 | N | bulk | 0.388 | 0.208 | 0.404 |
| CRUK0090 | R1 | sorted | 0.362 | 0.016 | 0.621 |
| CRUK0062 | N | bulk | 0.316 | 0.302 | 0.382 |
| CRUK0062 | R5 | sorted | 0.199 | 0.338 | 0.462 |
| CRUK0079 | N | bulk | 0.335 | 0.231 | 0.434 |
| CRUK0079 | R1 | sorted | 0.224 | 0.155 | 0.622 |
| CRUK0070 | N | bulk | 0.349 | 0.176 | 0.475 |
| CRUK0070 | R2 | sorted | 0.221 | 0.301 | 0.478 |
| CRUK0050 | N | bulk | 0.491 | 0.159 | 0.349 |
| CRUK0050 | R4 | sorted | 0.222 | 0.0263 | 0.752 |

differentially methylated. This suggests that cell-type heterogeneity between sorted and bulk normals generates few DMPs.

There were a total of 2,557,754 CpGs that were covered in all samples, 6.86% and 10.94% of which were hyper- and hypomethylated in at least one sorted sample with respect to the patient-matched bulk tissue. Of these DMPs, 28.81% were shared between at least 2 patients with most showing consistent direction of methylation difference across pairs (98.1%). Additionally, 92.73% of CpG sites that fell within the 150bp region spanning either side of an EpiDISH cellular deconvolution reference probes were recurrently differentially methylated across samples. Given these observations, we posit that most sorted versus bulk normal DMPs arise from cell-type specific methylation and variable cell type composition, and that these loci may be used to assess normal infiltrate composition.

Finally, we compared tumour-normal DMPs calls based on $m_b$ and CAMDAC $m_t$ with those obtained from the sample-matched tumour cells purified by FACS. In both cases, we used the adjacent normal as a proxy for the normal cell of origin. As expected, we observed good overlap between all DMP call sets for tumour samples with high tumour DNA content and the number of DMPs was correlated with power (**Figure 3.13**). At lower tumour copy number and purity, CAMDAC deconvolved methylation rates enabled identification of a large fraction of DMPs also identified by FACS, while bulk tumour methylomes recalled considerably fewer. Comparison

of non-overlapping calls is difficult due to the FACS data being of lower coverage (mean$_{FACSm_t}$ = 24, mean$_{CAMDACm_t}$ = 41, mean$_{m_b}$ = 78) and, potentially, cell-type biases during nuclei sorting.

To sum up, the adjacent normal is seemingly more suitable for differential methylation analyses than the sorted non-tumour cells due to having a larger epithelial component and, despite variation in composition, the adjacent normal is an adequate proxy for the normal tumour contaminating cells. Tumour-normal DMP calls based on CAMDAC $m_t$ have greater overlap with the FACS sorted data than the bulk, further advocating use of CAMDAC over bulk methylation rates for accurate identification of epimutations.

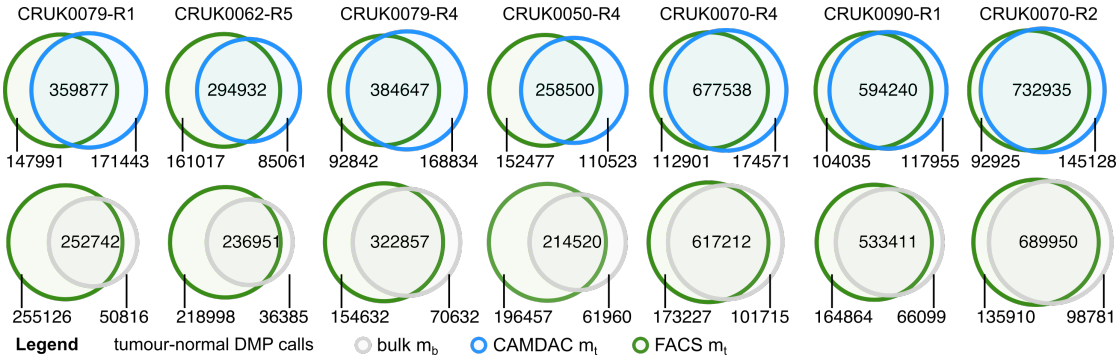


**Figure 3.13: Comparing tumour-normal differential methylation based on CAMDAC $m_t$, $m_b$ and FACS purified tumour methylation rates.**
Venn diagrams showing the overlap between tumour-normal DMP calls performed on the adjacent matched normal and CAMDAC $m_t$ versus FACS $m_t$ (top row) and $m_b$ versus FACS $m_t$ (bottom row). Individual Venn diagrams are to scale. Samples are ordered by the tumour sample tumour DNA content.

## 3.3 Discussion

Data presented in the previous chapter (**Section 2.2.4**) showed that LUAD and LUSC both have high levels of normal contamination. In line with this observation, they each ranked in 4th and 7th lowest position in terms of mean purity when compared with 30 other cancer types [144]. Tumour-normal deconvolution is therefore particularly important to correctly interpret NSCLC methylomes. To overcome this issue, we introduced our tool for Copy Number-Aware Methylation Deconvolution

Analysis of Cancer, CAMDAC and applied it to multi-region RRBS data performed on the primary tumour of NSCLC patients from the epiTRACERx cohort.

We showed that SNPs affect methylation levels in both normal and tumour samples and thus computed SNP-independent methylation rates. This ensured that CpG methylation could take any values between 0 and 1, as it should. By thresholding on the adjacent matched normal and stratifying CpGs by ASCAT.m copy numbers and tumour purity, we could visualise DMP populations. Following the same approach, we demonstrated that bulk methylation rates at DMPs depend on tumour purity, copy number and both tumour and admixed normal methylation rates and formalised this relationship as CAMDAC equation 2. To test this model, we extracted the modal methylation rate of clonal bi-allelic DMP populations from the observed $m_b$ distribution and compared them with the expected, showing good agreement across tumour purities and copy number states for both hyper- and hypomethylation positions. We concluded that tumour purity and copy number explains the majority of the observed intermediate methylation values in the $m_b$ distribution. With ASCAT.m purity and copy numbers and polymorphism-independent methylation rates all obtained directly from tumour and adjacent matched normal RRBS, we could compute the CAMDAC $m_t$ estimates as per equation 3.

Analysis of inter-sample Pearson distances confirmed that CAMDAC effectively removes the shared normal component from the bulk, increasing distances between different patients and normal and tumour sample pairs while retaining high intra-tumour correlations, as was expected due to shared clonal origins. Where possible, we phased SNVs from newly obtained WGS and published WES [112] in regions of LOH with the mutant allele on all tumour copies, enabling unbiased separation of tumour and normal reads. Despite the alignment bias against reads containing alternate alleles in RRBS data, which we discussed in the previous chapter (**Section 2.2.2**), the RRBS- and WES/WGS-derived VAFs were highly correlated. This suggests that, when the alternate allele is mapped successfully, the VAF is reliable and thus *de novo* SNV calls from RRBS using CAMDAC calculation rules would be possible. SNV purified tumour methylation rates validate CAMDAC de-

convolved $m_t$ estimates. Next, we performed tumour-normal and tumour-tumour DMP calls of both simulated and real data, revealing that only $m_t$ values enable accurate differential methylation analyses.

Finally, we presented additional RRBS data acquired for both tumour and admixed normal cell populations separated by ploidy with FACS from five different patients. For two of the patients, we sequenced an additional aneuploid population from a different tumour region. Results suggest that the tumour-adjacent normals have greater epithelial content then the non-tumour infiltrating cells, making the former the most suitable proxy for the NSCLC cell of origin and thus for tumour-normal differential methylation analyses. Despite differences in cell type composition, normal methylation rates computed from the diploid populations which were sorted by FACS and those obtained from the adjacent normal were in agreement with over 95% of CpGs classified as non-DMPs at overlap probability $< 0.01$ and effect size $> 0.2$. Where present, DMPs tended to be recurrent across normal pairs and were likely due to cell-specific methylation. Tumour-normal DMPs derived from the adjacent matched normal and either the $m_b$, CAMDAC $m_t$ or FACS $m_t$ revealed greater overlap between calls based on CAMDAC- and FACS-purified methylation rates than between $m_b$ and FACS $m_t$.

Overall, CAMDAC allows for correct interpretation of bulk tumour intermediate methylation levels as well as extraction of purified profiles unpolluted by admixed normal cells.

The data presented in this chapter showed that, at least in NSCLC, the methylation rate distribution at CpGs having lost methylation in tumour cells is noisier than hypermethylated loci, likely due to (i) ongoing methylation erosion, especially in rapidly dividing tumour cells, (ii) bisulphite over-conversion and (iii) cell-type specific methylation in normal contaminating cells.

Although ignored in our model thus far, ongoing tumour evolution gives rise to intra-tumour heterogeneity. Diverging evolutionary trajectories from the most recent common ancestor (MRCA) will lead to coexisting subgroups of cells called subclones. If these populations harbour substantially different (epi)genetic profiles

and are present in a sufficiently large fraction of cells, this signal could further complicate interpretation of bulk tumour profiles [126]. Work from the PCAWG consortium showed that, on average less than 5 and 10% of SNVs detected from single LUAD and LUSC bulk tumour biopsies were subclonal, respectively [126]. Clonal bi-allelic DMP signal is unlikely to be affected by this, but clonal allele-specific epimutations could be confounded by subclonal bi-allelic aberrations in high purity samples with sufficiently large co-existing subclones and sufficiently high cancer cell fraction.

For example, take a hypothetical bulk tumour admixture with $n_t = 3$ and $\rho = 0.4$, and a large set of CpGs that were unmethylated in the normal and the MRCA, but became fully methylated in 50% of tumour cells. These sites would form a potentially detectable subclonal DMP cluster around $m_b = \frac{1}{4}$ or $m_t = \frac{1}{2}$ in pure tumour methylation space (**Figure 3.14**. In reality, most tumours will have a dominant subclone, and the small population of tumour cells that are heterogeneous at a subset of CpGs are unlikely to interfere with global differential methylation detection [144].

Cell type heterogeneity between normal cell populations can also contribute to intermediate methylation signal. In samples of low tumour purity, a normal cell population present in a large fraction of the normal contaminating cells with cell-type specific methylation will generate intermediate bulk tumour methylation levels. For example, if one or more CpGs are methylated in $\frac{1}{3}$ of normal contaminants and otherwise unmethylated in a sample of purity $\rho = 0.4$, tumour copy number $n_t = 3$ and normal copy number $n_n = 2$, this would lead to intermediate methylation centred around $m_b = \frac{1}{6}$ (**Figure 3.14**). However, we show that the adjacent normal cellular composition is sufficiently close to the normal contaminating cells to serve as a proxy to remove this signal from the bulk tumour looking at correlations pre- and post-deconvolution and comparing adjacent normal methylation rates directly with that of the admixed normal cells isolated by FACS for a subset of samples.

To conclude, our tool for copy number-aware methylation deconvolution analysis of cancers, CAMDAC, allows users to fully exploit the wealth of infor-

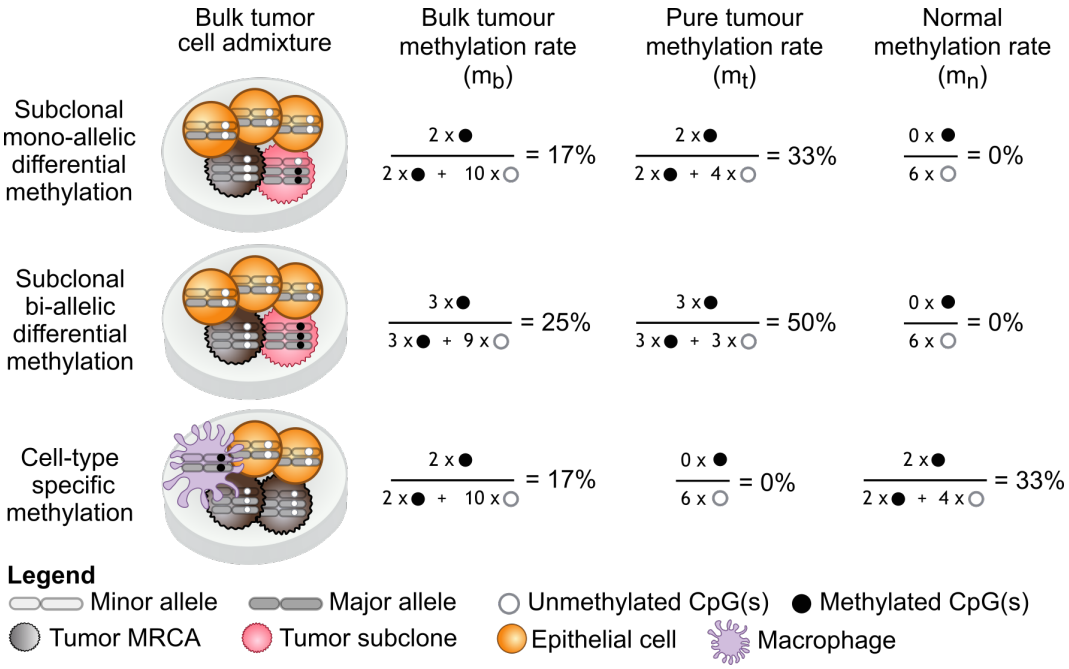


**Figure 3.14: Intra-tumour and cell type heterogeneity may also generate intermediate methylation signal in bulk tumour data.**
Example bulk tumour ($m_b$), pure tumour ($m_t$) and normal ($m_n$) methylation rates in a bulk tumour admixture of purity $\rho = 0.4$ for an individual or group of CpG(s) of clonal tumour copy number $n_t = 3$ and with either mono- or bi-allelic subclonal tumour-normal differential methylation or cell type specific methylation.

mation present within RRBS, enabling combined copy number profiling and purity estimation with accurate DMP analysis derived from CAMDAC $m_t$. CAMDAC has the potential to unveil unique insights into cancer biology and taxonomy as well as methylation ITH, directly from bulk bisulphite sequencing of solid tumours. CAMDAC is expected to further our understanding of the interplay between epigenetic and genetic mutations as well as their roles throughout tumor evolution.

## 3.4 Methods

The methods described below were recently published as part of our bioRxiv preprint [140]. The epiTRACERx cohort selection, characteristics and sequencing methods can be found in the previous chapter (**Methods 2.4**).

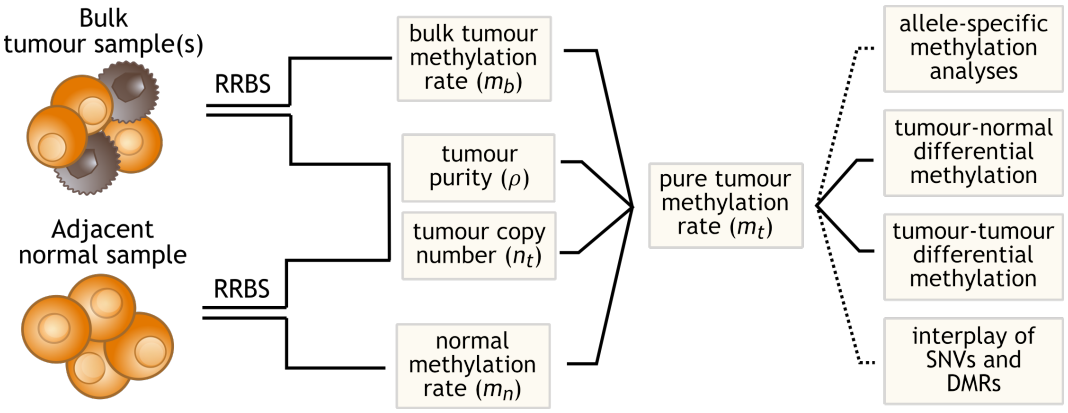


**Figure 3.15: CAMDAC pipeline overview.**
CAMDAC takes bulk tumour and a normal proxy for the tumour contaminating cells as input. The default is to used the same normal sample as a representative methylome for the cell of origin in differential methylation analyses or a different normal may be provided. CAMDAC extracts SNP-independent $m_b$ and $m_n$ and leverages copy number profiles and tumour purity estimates from ASCAT.m to compute $m_t$ and perform DMP and DMR calling. Allele-specific methylation analysis and phasing to SNVs / heterozygous SNP can be done applying CAMDAC principles, but is yet to be implemented as part of the package.

### 3.4.1 Copy-number aware methylation deconvolution analysis of cancers (CAMDAC)

As depicted below, CAMDAC requires RRBS data prepared from bulk tumour and matched adjacent normal samples. At this moment, CAMDAC is only compatible with human (directional) RRBS data. The input must be quality and adapter trimmed with PCR duplicates removed and subsequently aligned to hg19 (hg38, GRCH37 and GRHCH38 formats also compatible). The tumour and matched normal data must be provided in .bam file format. The files must be sorted and indexed using SAMtools (http://github.com/samtools/). The input tumour and matched normal sequencing data is used to compute tumour purity and allele-specific copy numbers with ASCAT.m as well as SNP-independent bulk tumour and normal methylation rates. From these data, we obtain CAMDAC purified $m_t$ values and perform tumour-normal and tumour-tumour differential methylation analysis. Allele-specific methylation analyses and phasing of differentially methylated loci to SNVs or heterozygous SNPs can both be achieved by applying CAMDAC principles, but this is yet to be implemented as part of the package.

### 3.4.2 SNP-independent methylation rate calculation

The bulk tumour and matched normal methylation rate is readily computed by taking the ratio of methylated CpG read counts to the sum of methylated ($M$) and unmethylated ($UM$) read counts, $m = \frac{M}{M+UM}$. CAMDAC uses strand-specific dinucleotide counts to distinguish between methylated and unmethylated CpGs, $m = \frac{CG}{CG+TG(+)+CA(-)}$. At CpG>TpG (TpT>CpG) and CpG>CpA (CpA>CpG) SNPs, only reads from the bottom strand and top strand, respectively, contribute to methylation rates. Moreover, only the CpG-forming allele contributes to the methylation rate at polymorphic CpGs. This enables the methylation rate at a heterozygous CpG to vary between 0 and 1, instead of between 0 and 0.5 in a diploid example, and ensures further independence between methylation rate and copy number estimates.

We compiled bulk tumour and normal methylation rates for all CpGs which fell within the above-mentioned reference RRBS genomic regions list. For each patient, we discarded all CpGs that failed to reach a minimum coverage of 10 in the matched normal RRBS data. CpGs that had less than 3 reads in a given tumour sample were also filtered out from that sample.

### 3.4.3 CAMDAC-purified tumour methylation rates from RRBS data

Bulk tumour methylation rate ($m_b$) could be expressed as a function of the methylation rate in the tumour cells ($m_t$) and contaminating normal cells ($m_n$), scaled for the purity of the sample ($\rho$) and the local copy number state. This normal copy number is set to $n_n = 2$ for autosomal CpGs and 1 copy at polymorphic CpGs or outside pseudoautosomal regions regions on chromosome X in males in the normal. Similarly, $n_t$ is the tumour total copy number obtained from ASCAT.m, with the exception being polymorphic CpGs where the CpG allele-specific copy number is used. For CpG-destroying SNPs with BAF<0.5 or CpG-forming SNPs with BAF>0.5, the major allele copy is used. *Vice versa*, if the BAF>0.5 and the SNP is CpG-destroying or the BAF<0.5 and the SNP is CpG-forming, only the minor

allele informs the methylation rate.

We obtained total copy number at each CpG and tumour purity directly from RRBS with ASCAT.m and thus, assuming the matched normal is a reasonable proxy for the contaminating normal, we can derive $m_t$.

Finally, $m_t$ is a rate and so must be bound by the 0 and 1. But, due to biological and technical noise, values can fall outside of these boundaries. For downstream analyses, we round negative $m_t$ estimates up to 0 and those above the upper boundary are capped at 1. The $m_t$ 99% HDI virtually always overlaps with allowed values.

We validate our model by comparing observed and expected clonal bi-allelic DMPs modal peak position for CpGs that were confidently unmethylated ($HDI^{99} \subseteq [0,0.2]$) of methylated ($HDI^{99} \subseteq [0.8,1]$) in the adjacent patient-matched normal. For this subset of CpGs, we compiled allele-specific copy number (ASCN) segments are retained ASCN states with $\geq$ 10,000 loci (inter-quartile range 180,125-807,774). Sex chromosomes were excluded due to sequencing biases against the packed Barr body shifting methylation estimates. Segments meeting these criteria had total copy number ranging from 1-8. Beta regression was used to estimate the mode of the peak generated by hyper- and hypomethylated DMP populations in the observed bulk tumour and CAMDAC methylation rates distribution stratified by tumour purity, copy number and matched normal methylation rate. The modal methylation rate of the peaks at 0 and 1 in the patient matched normal were estimated by beta regression and these unmethylated ($m_0$) and methylated ($m_1$) values were used to derive the expected values as opposed to exactly 0 and 1. The bulk tumour expected at hypermethylated loci is computed by feeding sample purity, segment copy number, $n_n$=2 and $m_n = m_0$ and $m_t = m_1$ equation 2. *Vice versa*, $m_n$ and $m_t$ are substituted with $m_1$ and $m_0$ to calculate the expectation at hypomethylated loci. The predicted values are set to exactly $m_0$ and $m_1$ in the purified profiles for CpGs having lost and gained methylation respectively.

In the bulk comparisons, the mean observed and expected methylation rates were 0.49 compared with 0.52 for hypermethylated loci and 0.46 for both at hypomethylated CpG sites. This further supports that intermediate methylation gives

rise to intermediate methylation signal. The mean absolute errors were 0.032 and 0.029 at CpGs that gained and lost methylation, respectively. As for the bulk, the error between modelled and expected pure tumour methylation rates is correlated with tumour DNA content, but after purification, the error plateaus to -0.084 at tumour fractions greater than 0.53 for hypermethylated loci. Similarly, at hypomethylated DMPs, it plateaued from 0.51 at a value of 0.052.

### 3.4.4 SNV-phased methylation rate estimates

Leveraging SNV calls from newly obtained WGS data and previously published WES [112], we phased CpG methylation to all SNVs, excluding loci with $VAF \leq 0.1$ in a tumour sample or $VAF > 0$ in the patient-matched adjacent normal tissue. Allele-specific methylation counts were compiled for all reads that could be phased to exactly one SNV. Phased methylation rates were obtained for 32,874 CpGs and 6,529 SNVs across samples (14,514 and 2,984 unique CpGs and SNVs respectively across patients). The VAF was derived from the mutant (*mut*) and wild type ($WT$) reads counts: $VAF = \frac{counts_{mut}}{counts_{mut} + counts_{WT}}$. RRBS-derived VAF estimates were compared with those obtained from the WES/WGS (Pearson correlation = 0.86, **Figure 3.8A**). The mutation copy number, $n_{mut}$, was then computed as a function of the variant allele frequency, tumour purity and copy number: $n_{mut} = \frac{1}{\rho} \times VAF \times \rho n_t + n_n(1-\rho)$. The wild type allele copy number, $n_{WT}$, is obtained by subtracting $n_{mut}$ from $n_t$: $n_{WT} = n_t - n_{mut}$. The mutant allele methylation rate, $m_{mut}$, is extracted by taking the counts methylated ($M_{mut}$) and unmethylated ($UM_{mut}$) divided by all counts phased to the variant allele: $m_{mut} = \frac{M_{mut}}{M_{mut} + UM_{mut}}$. The wild type allele methylation rate, $m_{WT}$, is confounded by signal from normal contaminating cells and must be deconvolved. For this, we use a modified version of the CAMDAC equations 2 and 3, where the tumour methylation rate and copy number are expressed in terms of the mutant and wild type alleles.

$$m_b = \frac{\rho(n_{mut}m_{mut} + n_{WT}m_{WT}) + n_n m_n(1-\rho)}{\rho(n_{mut} + n_{WT}) + n_n(1-\rho)} \tag{5}$$

$$m_{WT} = \frac{m_b(\rho(n_{mut} + n_{WT}) + n_n(1-\rho)) - n_n m_n(1-\rho) - \rho n_{mut} m_{mut}}{\rho n_{WT}} \tag{6}$$

We validate CAMDAC $m_t$ by comparison with methylation estimates phased to clonal SNVs present on all copies in regions with loss of heterozygosity across our cohort ($n_{mut} = n_t$). At these sites, all reads reporting the variant allele can directly be assigned to the tumour cells, and methylation rates obtained from this subset of reads should be an unbiased estimate of the purified tumour methylation rate (i.e. $n_{mut} = 0$). Overall, 4,485 CpG loci met these criteria. A high correlation was observed between these SNV deconvoluted $m_t$ values and CAMDAC estimates (Pearson correlation = 0.97, **Figure 3.8B**).

### 3.4.5 Identifying tumour-normal DMPs

We develop a statistical test for DMP calling between tumour and normal. The number of methylated reads at a CpG can be modelled as Beta-Binomial distribution with mean equal to the methylation rate and following a Beta distribution. The bulk tumour and matched normal methylation rate at the $i^{th}$ CpG, $m_{x,i}$, can be expressed in terms of the observed methylated and unmethylated read counts $M_{x,i}$ and $UM_{x,i}$, respectively, where $x$ is $b$ for the bulk tumour and $n$ for the normal:

$$m_{x,i} = Beta(M_{x,i}, UM_{x,i}) \tag{7}$$

For tumour-normal DMP calling, we test whether or not the CAMDAC purified tumour methylation rate at the $i^{th}$ locus, $m_{t,i}$, is different from the normal methylation rate, $m_{n,i}$. In other words, we evaluate whether $m_{t,i}$ is statistically different from $m_{n,i}$.

$$\Delta B = m_{t,i} - m_{n,i} \tag{8}$$

As per CAMDAC equation 3, we defined $m_{t,i}$ as a difference between two Beta distributions, $m_{b,i}$ and $m_{n,i}$, scale for tumour purity and copy number. Substituting expression 2 into $\Delta B$, we obtain the following and simplify:

$$\Delta B = \frac{m_{b,i}(\rho n_{t,i} + n_{n,i}(1-\rho)) - n_{n,i} m_{n,i}(1-\rho)}{\rho n_{t,i}} - m_{n,i} \tag{9}$$

$$\Delta B = C \times (m_{b,i} - m_{n,i}) \quad \text{where} \quad C = \frac{\rho n_{t,i}}{\rho n_{t,i} + n_{n,i}(1-\rho)} \tag{10}$$

Expression 10 suggests that our model is independent of copy number and tumour purity. However, the power to call DMPs will intrinsically depend on purity, local copy number and CpG coverage. The former two will alter the magnitude of $\Delta B$ and the latter will impact the variance of each $m_{b,i}$ and $m_{n,i}$. Since there is a closed-form solution for testing $P(m_{b,i} > m_{n,i})$, but not for $P(m_{b,i} = m_{n,i})$ [156], the null and alternate hypotheses are written as follows:

$H_0 : m_{b,i} = m_{n,i}$ The methylation rate of the ith CpG is identical in normal and tumour

$H_1 : m_{b,i} > m_{n,i}$ The tumour is hypermethylated at this locus

$H_1 : m_{b,i} < m_{n,i}$ The tumour is hypomethylated at this locus

where,

$$P(m_{b,i} > m_{n,i}) = \sum_{j=0}^{M_{n,i}-1} \frac{B(M_{b,i}+j, UM_{b,i}+UM_{n,i})}{(UM_{b,i}+j)+B(1+j, UM_{n,i})+B(M_{b,i}, UM_{b,i})} \tag{11}$$

$$P(m_{b,i} < m_{n,i}) = \sum_{j=0}^{M_{b,i}-1} \frac{B(M_{n,i}+j, UM_{n,i}+UM_{b,i})}{(UM_{n,i}+j)+B(1+j, UM_{b,i})+B(M_{n,i}, UM_{n,i})} \tag{12}$$

$$= 1 - P(m_{b,i} > m_{n,i}) \tag{13}$$

Equation 11 can be rewritten as follows, incorporating our $B(0.5, 0.5)$ prior to each methylation count variable. The prior informs on the underlying methylation rate distribution and ensures finite logarithms:

$$\begin{aligned} P(m_{b,i} > m_{n,i}) = \sum_{j=0}^{M_{n,i}-1} & logB(M_{b,i}+j+0.5, UM_{b,i}+UM_{n,i}+0.5) - \\ & logB(UM_{b,i}+j+0.5) - \\ & logB(1+j+0.5, UM_{n,i}+0.5) - \\ & logB(M_{b,i}+0.5, UM_{b,i}+0.5) \end{aligned} \tag{14}$$

We compute the probability that a CpG site is hypo- or hypermethylated and for easier interpretation, we express these probabilities as their complement (C = 1-P), which is a measure of the overlap between $m_{b,i}$ and $m_{n,i}$. If $C(m_{b,i} > m_{n,i})$ or $C(m_{b,i} > m_{n,i}) \leq 0.01$ we accept $H_1$ or $H_2$ respectively. Differential methylation analyses incorporating a beta distribution model reportedly show higher true positive and lower false discovery rate are obtained compared with Fisher's and z-score methods [157]. Nevertheless, given high enough coverage, even a small difference in methylation can become statistically significant. In order to focus our analysis on biologically significant methylation changes, we require a minimum effect size of 0.2 between the purified tumour methylation rate, $m_{t,i}$, and the matched normal, $m_{n,i}$, for DMP calling. In theory, this allows for subclonal or allele specific changes to be picked up whilst removing spurious signal. We obtained a second set of DMPs by applying the minimum effect size threshold on the difference between the bulk tumour methylation rate, $m_{b,i}$, and the normal methylation rate, $m_{n,i}$ as is customary [91,92,158–160], and compared the output with CAMDAC calls. The threshold of 0.2 was deemed sufficiently low to capture most mono-allelic aberrations whilst being high enough to remove false positives and filter noise from the heterogeneous normal contaminating cells.

### 3.4.6 Identifying tumour-tumour DMPs

The purified tumour methylation rate is not a beta distribution but rather the difference between two betas, the bulk tumour and normal methylation rate, scaled for tumour purity and copy number. As such, there is no exact solution to compute the highest density interval for $m_{t,i}$. To address this, we simulate a credible 99% HDI for $m_{t,i}$ at every CpG. We use the tumour purity and CpG copy number and simulate 2000 data points for the bulk tumour ($m_{b,i}$) and matched normal methylation rate ($m_{n,i}$), given $m_{x,i} \sim B(M_{x,i}, UM_{x,i})$. Substituting these into Eq(2), we obtain a vector of values for $m_{t,i}$ and readily extract the 99% HDI. If the purified tumour methylation rate HDI does not overlap between any two tumour regions at a given CpG and the minimum effect size, 0.2, is reached, a tumour-tumour DMP is identified.

### 3.4.7 Simulating tumour-normal and tumour-tumour DMPs

To appreciate the effect of tumour purity on both tumour-normal and tumour-tumour differential methylation analysis, we extracted the 20 lowest and highest purity samples in our cohort, $\rho < 0.3$ and $\rho > 0.58$, respectively. We combined the methylation information at all overlapping autosomal CpGs from these samples including the patient-matched normal of selected samples. The normal methylation rate at confidently unmethylated and methylated CpGs was extracted. Confidently unmethylated and methylated CpGs are respectively defined as having their methylation rate 99% HDI boundaries in the 0.0-0.2 and 0.8-1.0 intervals. This vector of values incorporates information for both distributions such as mean methylation rate and deviation as well as their respective contribution to the overall bimodal normal methylation rate profile. Sampling from this vector will yield our simulation priors.

Next, we randomly selected intra- and inter-tumour CpG pairs from samples within or across purity categories and of equal or differing total copy numbers (range 1-6). For all 168 possible combination of these simulation parameters, we sampled 10,000 loci from each tumour samples randomly assigned as x and y as well as their matched normal. For each locus, we begin by obtaining the coverage information for the $i^th$ CpG selected from sample $s = x \quad or \quad y$. We know the bulk tumour coverage ($cov_{b,s,i}$), the tumour copy number ($n_{t,s,i}$), the normal copy number ($n_{n,s,i} = 2$) and the global tumour purity ($\rho_{s,i}$) and so the tumour DNA fraction ($f_{t,s,i}$) is $f_{t,s,i} = (n_{t,s,i} \times \rho_{s,i})/(n_{t,s,i} \times \rho_{s,i} + (n_{n,s,i} \times (1 - \rho_{s,i}))$. We can deduce the purified tumour coverage as $cov_{t,s,i} = cov_{b,s,i} \times f_{t,s,i}$ where $cov_{t,s,i} \sim Binom(cov_{b,s,i}, f_{t,s,i})$. The matched normal coverage ($cov_{n,s,i}$) is therefore $cov_{n,s,i} = cov_{b,s,i} - cov_{t,s,i}$.

We then sample a normal methylation rate prior ($p_{n,i}$) from the confidently unmethylated ($v_{unmeth}$) and methylated CpGs ($v_{meth}$) from the matched normal data which is used to simulate the normal methylation rates of normal contaminating cells from both tumour samples x and y. For each sample, we obtain the counts methylated ($M_{n,s,i}$) and unmethylated ($UM_{n,s,i}$): $M_{n,s,i} \sim Binom(cov_{n,s,i}, p_{n,i})$ and $UM_{n,s,i} = cov_{n,s,i} - M_{n,s,i}$.

The purified tumour methylation rate is obtained by randomly selecting from the same vector as the matched normal ($p_{n,i}, p_{t,i} \in v_{unmeth}$ or $p_{n,i}, p_{t,i} \in v_{meth}$) or from opposite vector states ($p_{t,i} \in v_{unmeth}$ and $p_{n,i} \in v_{meth}$ or $p_{t,i} \in v_{meth}$ and $p_{n,i} \in v_{unmeth}$). We obtain the counts methylated ($M_{t,s,i}$) and unmethylated ($UM_{t,s,i}$) as $M_{t,s,i} \sim Binom(cov_{t,s,i}, p_{t,i})$ and $UM_{t,s,i} = cov_{t,s,i} - M_{t,s,i}$. The bulk methylation counts for both samples are easily calculated by adding the tumour and normal counts: $M_{b,s,i} = M_{t,s,i} - M_{n,s,i}$ and $UM_{b,s,i} = UM_{t,s,i} - UM_{n,s,i}$. In balanced copy number regions, we also simulate allele-specific DMPs whereby one allele is in the normal ground state and the other is differentially methylated. We obtain the counts methylated from the minor allele, allele A, and for the major allele, allele B, and combine them to obtain total counts methylated and unmethylated.

The expected absolute tumour-normal methylation difference at simulated bi-allelic DMPs is $|m_t - m_n| \approx 0.95$. In the bulk, the magnitude of the difference depends on sample purity and CpG copy number. The purified methylation rate at mono-allelic DMPs usually depends on the copy number of the mutated allele, however, in balanced copy number regions, where $n_A = n_B$ and given that one copy clonal differentially methylated CpG(s) and the other is in the ground state, $m_t = \frac{M_{t,A} + M_{t,B}}{M_{t,A} + UM_{t,A} + M_{t,B} + UM_{t,B}} = 0.5$. The expected tumour-normal difference at simulated mono-allelic DMPs is thus $|m_t - m_n| \approx 0.5$. Finally, tumour-normal and tumour-tumour differential methylation calls were made using CAMDAC differential methylation analysis and the output compared with the ground truth. False negative and positive rates are obtained for the bulk and deconvolved tumour simulated data.

## 3.4.8 RRBS sequencing of FACS populations

### *3.4.8.1 Nuclei extraction and FACS*

Nuclei extraction and FACS were performed by Annelien Verfaillie. Briefly, Frankenstein protocol (dx.doi.org/10.17504/protocols.io.3fkgjkw) was used for nuclei isolation from fresh frozen tissue samples of seven epiTRACERx samples: CRUK0050-R4, CRUK0070-R2, CRUK0070-R4, CRUK0079-R1, CRUK0079-R4, CRUK0062-R5, CRUK0090-R1. The tumour is minced and placed into a lysis

buffer. The lysate is homogenised using a pestle and the homogenate is filtered out using a 70 µm filter. The extracted nuclei were then washed and suspended into a buffer containing Pi followed by another filtering step, this time with a smaller 35 µm filter. Note that the Pi chromatin staining buffer concentration is 70 µg/mL of Pi, 1% BSA, 1×PBS.

Nuclei were run through the FACS machine, which separated nuclei by Pi staining intensity, reflecting nuclei ploidy, and by side scatter, reflecting cell morphology. We use the first few cells to adjust the nuclei sorting parameters so as to eliminate debris and select near-diploid, $pop_D$, and the most prevalent aneuploid population, $pop_A$. For example, replicating cell populations were small but detectable, and were excluded. Nuclei were sorted and populations were collected directly into tubes containing the 200µL of PBS and 2% FCS, yielding between 100,000-300,0000 events per population. Finally, DNA extraction was prepared using the Zymo Research Quick DNA-microprep plus kit (D4074) following the indications from manufacturer. DNA was eluted in 12µL of elution buffer.

#### *3.4.8.2 RRBS*

The NuGEN Ovation RRBS Methyl-Seq System protocol was used to prepare libraries and for bisulphite sequencing using the same approach as described for the bulk tumour RRBS dataset, with the exception that no automation was required (**Methods, section 2.4.2.1**). Library preparation was carried out by Alex McLatchie and Cristina Cotobal-Martin. Libraries were prepared by enzymatically digesting ∼100ng of gDNA with *MspI*. Qiagen's EpiTect Fast DNA Bisulfite Kit was used for bisulphite conversion of the resulting DNA fragments. Bisulphite converted libraries were then amplified by PCR using 12 cycles and purified using Agencourt® RNAClean® XP magnetic beads. Library quantification was performed by Qubit dsDNA HS Assay (Invitrogen) and quality control was carried out using Agilent Bioanalyzer High Sensitivity DNA Assay (Agilent Technologies). In cases with two samples from the same patient, both $pop_A$ and only one $pop_D$ were made into libraries and sequenced to save on costs. We therefore sequenced 7 tumour aneuploid and 5 (presumably) normal diploid populations.

RRBS was performed by at the Francis Crick Institute sequencing facility under the supervision of Robert Goldstone. The 12 samples were multiplexed across 4 lanes on HiSeq 4000 using the HiSeq® 3000/4000 SBS Kit. As for the bulk tumour samples, 100bp SE and 10bp reads were generated for the NuGEN RRBS library insert and unique molecular identifiers, respectively. We aimed to sequence 120,000,000 reads per sample.

Processing of the sequencing reads was performed by the author of this report. Sequencing reads were QC'ed, adapter trimmed, aligned to hg19, PCR-deduplicated and output binary alignment map were sorted and indexed following the same procedure as described for the epiTRACERx bulk tumour RRBS dataset (**Methods, section 2.4.2.1**).

On average, 109,864,415 raw sequencing reads were obtained per samples. Mapping efficiency averaged around 70.5%, as expected for bisulphite sequencing data aligned with Bismark. However, samples had very high duplication rates (average 57.18%), leaving only 31,533,497 reads per sample post-processing. This is potentially due to combined bisulphite- and FACS-driven DNA degradation. This effect was not observed on the previous cohort generated using the same protocol, but without FACS.

### 3.4.9 WGS SNV calls

WGS samples were obtained for 7 samples from 3 patients included in the TRACERx100 cohort and the epiTRACERx cohort and processed as described in **section 2.4.2.2**. Somatic mutations were identified with MuTECT v1.1.7 [161] using the same pipeline as for the published TRACERx100 WES-derived SNV calls [112] and was run by Michelle Dietzen from the McGranahan Lab at University College London. SNVs present in the tumour-adjacent normal tissue were removed from downstream analyses.

# Chapter 4

# Initials insights into the NSCLC methylome

## 4.1 Introduction

### 4.1.1 Rationale for studying NSCLC

In 2018, around 1,700 Americans died of cancer each day in the United States only according to reports by the American Cancer Society [162]. Of these, roughly 1 in 4 deaths were attributable to tumours originating in the lung or bronchus, totalling around 149,000 yearly casualties. NSCLC accounts for about 80% of all lung cancer diagnoses in the US, with LUAD and LUSC being the two most common histological subtypes. Unfortunately, the disease is often diagnosed late, with one study reporting that 68-80% of cases between 2005 and 2010 were classed as stage II and above [163]. Due to late diagnosis and lack of adequate treatment, the 5-year survival rate is below 15% [164]. For all of the above reasons, large scale sequencing efforts have been deployed to provide insights into the NSCLC genome namely from The Cancer Genome Atlas [165, 166].

In recent years, research focus has shifted away from simply charting genetic alterations and towards analysing intra-tumour heterogeneity as well as reconstructing tumour evolutionary histories. The TRACERx lung cancer study was designed to answer both these important questions and involves sampling of NSCLC tumours in space, taking a multi-sample sequencing approach, and in time, collecting tissue from the primary as well as any recurrence and metastases [113]. Preliminary results from the TRACERx first 100 patients revealed widespread somatic mutation and copy number ITH. Driver mutations in key genes such as *EGFR*, *TP53*, *BRAF*

and *MET* were virtually always clonal, while others, including *PIK3CA* and *NF1*, were heterogeneous [112]. The presence of subclonal driver mutations has important clinical implications, presenting a potential mechanism of resistance to targeted therapy. Crucially, several mutations which appeared clonal from single tumour biopsies were re-classified as subclonal based on multi-region data. This demonstrates that the illusion of clonality can be an issue when estimating the mutational profile of tumours from individual samples and making clinical decisions in relation to these findings.

### 4.1.2 Current understanding of the NSCLC methylome

The NSCLC methylome is considerably less well charted than its somatic sequence and copy number alteration landscape. Promoter hypermethylation of both *CDKN2A* and *CDH13* is known to drastically increase risk of recurrence in surgically resected stage I lung tumours (odds ratio 15.5, [167]). To name a few, hypermethylation of *CDH1* [168], *DAPK* [169], *DLEC1* [170], *MLH1* [170] or *RASSF1A* [171] in NSCLC is also linked with poor patient outcome. A handful of other events have been reported, most of which are included either in reviews by Tsou *et al.* [172], Belinsky [173] or, more recently, Langevin, Kratzke and Kelsey [174]. Importantly, studies often disagree as to the prevalence of each hypermethylation events (**Table 4.1**). This discrepancy can represent skewed patient selection and sampling biases, especially in smaller cohorts, as well as the use of different techniques to probe DNA methylation levels. Additionally, we posit that the lack of a method to faithfully purify bulk tumour methylation estimates and obtain DMP calls that are robust to variations in tumour purity and copy number contributes significantly to the observed variation. Correctly predicting the prevalence of epimutated gene promoters is important for use as biomarkers and in early detection.

The study of DNA methylation ITH has been mostly limited to high tumour purity samples and cancer types with little to no aneuploidy such as haematological malignancies [40, 148, 189, 190], Ewing sarcoma [103] and cancers of the central nervous system [191]. One recent article [150] investigated DNA methylation ITH

**Table 4.1: Commonly hypermethylated gene promoters in NSCLC.**

| Pathway | Gene | Histological subtype | Prevalence (in %) | Sample type | Technique | Source |
|---|---|---|---|---|---|---|
| Apoptosis | *DAPK* | LUAD | 24-48 | TU | ML | [169, 175] |
| | | LUSC | 25-31 | TU | ML | [169, 175] |
| Cell cycle | *CDKN2A* | LUAD | 13-67 | TU | MSP | [171, 176, 177] |
| | | LUSC | 37-70 | TU | MSP | [171, 176, 177] |
| | *PAX5* | LUAD | 52-64 | TU | MSP | [178] |
| | | LUSC | 61-74 | TU | MSP | [178] |
| | *CHFR* | NSCLC | 10-19 | TU | MSP | [179, 180] |
| Differen- | *RARβ* | NSCLC | 26 | TU | MSP | [181] |
| tiation | | LUAD | 61 | TU | MSP | [170] |
| | | LUSC | 51 | TU | MSP | [170] |
| DNA Repair | *MGMT* | LUAD | 27-47 | TU | MSP | [176, 182] |
| | | LUSC | 19 | TU | MSP | [176] |
| Invasion | *CDH1* | LUAD | 16 | TU | MSP | [176] |
| | | LUSC | 19 | TU | MSP | [176] |
| | *CDH13* | LUAD | 16 | TU | MSP | [176] |
| | | NSCLC | 43 | TU | MSP | [183] |
| | *TIMP3* | LUAD | 24 | TU | MSP | [176] |
| | | LUSC | 23 | TU | MSP | [176] |
| | *LAMA3* | LUAD | 58 | CL | MSP | [184] |
| | | LUSC | 27 | CL | MSP | [184] |
| | *LAMB3* | LUAD | 32 | CL | MSP | [184] |
| | | LUSC | 20 | CL | MSP | [184] |
| | *LAMC3* | LUAD | 32 | CL | MSP | [184] |
| | | LUSC | 13 | CL | MSP | [184] |
| MSI | *MHL1* | LUAD | 23 | TU | MSP | [170] |
| | | LUSC | 46 | TU | MSP | [170] |
| NF$-\kappa\beta$ | *DLEC1* | LUAD | 32 | TU | MSP | [170] |
| | | LUSC | 48 | TU | MSP | [170] |
| | *UBE2N* | LUAD | 46 | TU | MSP | [170] |
| | | LUSC | 27 | TU | MSP | [170] |
| RAS | *RASSF1* | NSCLC | 30-38 | TU | BS,MSP | [185, 186] |
| | | LUAD | 27 | TU | MSP | [170] |
| | | LUSC | 25 | TU | MSP | [170] |
| | *RASSF5* | NSCLC | 18-24 | TU,CL | MSP | [187] |

*MSI = microsatellite instability, TU = patient primary tumour, CL = cell line, MSP = methylation-specific PCR, ML = MethyLight assay [188], BS = Targetted bisulphite sequencing

and tumour evolutionary trajectories in LUAD, leveraging multi-sample methylation array data obtained for 205 and 75 primary tumours and matched adjacent normal tissue samples, respectively, from 68 patients part of the EAGLE study [192]. Methylomes covered on average 338,730 CpG probes and profiles were corrected for tumour purity, but not ploidy. The study found that heterogeneity between samples from the same primary tumour was lower than between patients, with $\sim 90\%$

of tumour regions from the same patient clustering together. Correcting for tumour copy number would likely increase this percentage. Nevertheless, results suggest that DNA methylation ITH, particularly at CpG Islands, predicts not only survival but also time to metastasis.

### 4.1.3 Chapter summary

First, we performed DMR calling on CAMDAC purified methylomes and identified thousands of DMRs in NSCLC that were covered by RRBS and present at a detectable CCF. We annotated DMRs, revealing that RRBS data covered more intragenic regions than originally expected. As a result, most DMRs were hypomethylated, but promoter-associated CGIs were usually hypermethylated. We surveyed recurrently hypo- and hypermethylated DMRs and identified several early events as well as histological subtype-specific epimutations. We measured intra-tumour DMR ubiquity was correlated with patient outcome. Our results indicate at least 3 samples are needed to adequately survey intra-tumour heterogeneity. As a multi-sample RRBS dataset, the epiTRACERx cohort is particularly well-suited to study both intra- and inter-tumour epigenetic heterogeneity in NSCLC. Clustering of CAMDAC pure tumour methylation rates at promoter DMRs separated samples by sex, histology and patient while DMP clustering revealed intra-tumour subclonal relationships. In summary, CAMDAC offers unique insights into the NSCLC methylome.

## 4.2 Results

### 4.2.1 The epiTRACERx cohort epimutational landscape

#### *4.2.1.1 General overview*

In an attempt to identify disease-causing epimutations, it is common to search for differentially methylated regions (DMRs) instead of individual CpGs [92]. CAMDAC builds on its tumour-normal DMP calls to uncover DMRs. CpGs are binned into clusters setting a threshold on inter-CpG distance and tumour-normal DMRs are called if a given neighbourhood harbours a hotspots of $n$ consecutive

and *m* total DMPs (Methods, **section 4.4.1**). On average, we found 22,578 DMRs per sample and a combined total of 30,233 unique aberrated loci per patient, corresponding to 13.5% and 16.8% of covered loci, respectively (**Figure 4.1, first and second panel**). The DMR rate was weakly correlated with the number of samples per patient (Pearson correlation = 0.424, p-val = 0.00794), but the relationship was no longer significant after removing patients with two or less sampled tumour regions (Pearson correlation = 0.286, p-val = 0.174). This result advocates that at least 3 separate tumour biopsies are needed to chart the (detectable) epimutational profile of a patient's tumours.

Overall, 62.9% of tumour-normal DMRs across patients were hypo- as opposed to hypermethylated (**Figure 4.1, third panel**). This makes sense since more than 66.2% of bins covered by RRBS data were intragenic, which are expected to be methylated in the normal cells and could undergo demethylation during tumourigenesis. The fraction of hypo- and hyper-methylated DMRs varied greatly across patients and genetic features. On average, 26.7% of DMRs recorded at CGIs were hypomethylated, meaning they were usually hypermethylated. On the contrary, frequent loss of methylation was observed at CGI shelves (70.9% of DMRs).

#### *4.2.1.2 DMR ubiquity levels could inform NSCLC prognosis*

Next, we evaluated the ubiquity of epimutations as a proxy measure of tumour heterogeneity. Methylation clusters which were consistently epimutated across all sampled tumour regions from a given patient were termed ubiquitous. In cases where at least one sample was confidently epimutated, but others were not, we assessed whether the non-differentially methylated allele was confidently in the normal ground state. If at least one normal and mutant state were observed at a given methylation locus across samples from the same patient, the bin was classified as non-ubiquitous. In other words, the DMR was not present in all samples from the same tumour. Otherwise, if we could not distinguish whether the non-differentially methylated allele was in the normal ground state in at least one samples, the methylation bin could not be confirmed as non-ubiquitous and was instead categorised as 'undetermined' (**Figure 4.1, fourth panel**).

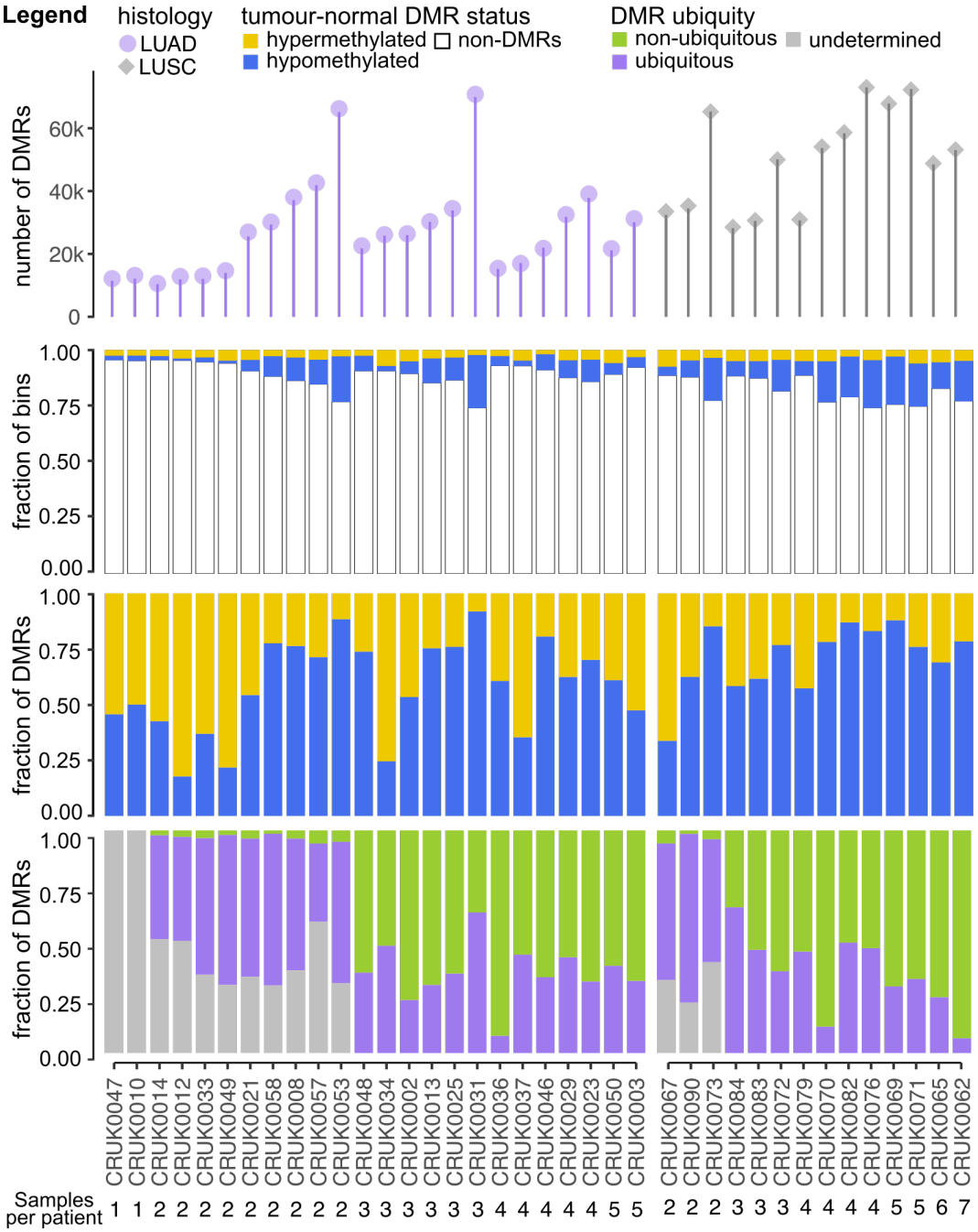


**Figure 4.1: Evaluating DMR calls across the epiTRACERx cohort.**
Total number of DMRs recorded per patient (top panel), fraction of loci classified as DMRs or non-differentially methylated (second row), fraction of DMRs with hypo- and hypermethylation (third row), DMR ubiquity breakdown per patient (fourth panel). Patients are ordered by the number of sampled tumour regions.

Overall, 36.1% of DMRs were ubiquitous across the 24 tumours for which we sampled 3 or more regions and showed little to no variation across genetic features for a given patient. DMR ubiquity levels did not affect overall survival (Pearson correlation = 0.102, p-val = 0.262) but were significantly correlated with relapse-free survival (Pearson correlation = 0.336, p-val= $1.49 \times 10^{-4}$). However, we cannot exclude that this effect may be driven by the the anti-correlation between ubiquity estimates and the number of samples per patient, which itself reflects tumour size as per the TRACERx study design [112].

#### 4.2.1.3 Promoter-associated differentially methylated CpG Islands

To identify key methylation events in NSCLC progression, we evaluated 16,906 distinct methylation bins that were covered by RRBS, overlapped with CGIs and spanned gene promoters (totalling 17,781 different genes). Regions were selected for having coverage in all tumour and normal samples for a total of 22,173 different annotated CpG clusters per sample, with some duplicate bins due to overlapping genes. Almost half of these methylation bins were not aberrated in any samples (49.6%, 10,988) and a further 6.07% (1,345) were only altered in a single patient or sample. Of the remaining 9,840 loci that were epimutated in at least 2 patients, 71.5% were hypermethylated in the tumour with respect to the normal (median CAMDAC pure methylation difference = 0.437). This result is in line with the expectation that promoter-associated CGIs are usually hypermethylated in cancer [193]. In total, 609 gene promoter-associated methylation bins were differentially methylated in more than 90% of patients. Comparative analysis of matched RNA-Seq and RRBS data could help narrow down key alterations that result in differential expression, but this is beyond the scope of this work.

#### 4.2.1.4 Recurrently hypomethylated gene promoters

The top 5 most recurrently hypomethylated loci were oncogenes *SH2B1* (106 samples, 35 patients), *SFN* (109 samples, 37 patients), *NELFCD* (120 samples, 37 patients), *FAM83H-AS1* (115 samples, 38 patients) and *TUBA1C* (118 samples, 38 patients). Suppression of *SH2B1* expression with micro-RNAs has been shown to reduce NSCLC cell proliferating potential and metastasis formation [194]. Our

results suggest that, in the absence of silencing micro-RNAs, promoter hypomethylation could enable expression of this oncogene. *SFN* over-expression has been reported in LUAD and is known to be regulated by aberrant hypomethylation [195]. We posit that the same mechanism is at play in LUAD and LUSC. *NELFCD* is known to be over-expressed in colorectal cancer and this is usually linked to focal copy number amplification [196]. Our data suggest DNA hypomethylation may also contribute to the observed *NELFCD* upregulation. Over-expression of *TUBA1C* is a predictor of poor prognosis in pancreatic ductal adenocarcinoma [197]. *FAM83H-AS1* is a long non-coding RNA (LncRNA) and is also reportedly up-regulated in LUAD with high expression levels associated with worse outcome [198]. Despite hypomethylation being only weakly correlated with gene expression, our results imply that these five recurrent altered genes likely play an important role in lung cancer.

#### *4.2.1.5 Frequently hypermethylated gene promoters*

The role of promoter hypermethylation in cancer, and more specifically in NSCLC, is well understood. We take a look at known hypermethylated gene promoters and members of the same family, selecting methylation bins that had coverage in all samples (**Figure 4.2**) and compared epimutation rates with published estimates (reviewed in [173]).

We report *PAX5* promoter DMRs in 81.6% of epiTRACERx patients (91 samples, 31 patients) a higher percentage than the $\sim 52-75\%$ estimates published by others [178]. This is likely explained by increased sensitivity in CAMDAC deconvoluted profiles as opposed to bulk tumour tissues. This epimutation was ubiquitous in 74.1% of patients after excluding the 2 patients with only one tumour region and a further 2 cases for which ubiquity was undetermined, suggesting *PAX5* alteration is an early even in NSCLC evolution. Highlighting the importance of paired-box (*PAX*) transcription factors in cell cycle deregulation during lung cancer tumourigenesis, several other members of the PAX family were repeatedly altered, including *PAX* genes 1-3 and 6-9 which were hypermethylated in 73.7-100% of NSCLC in this cohort (78-120 samples, 28-38 patients). *PAX6* hypermethylation has been re-


**Figure 4.2: Prevalent promoter DMRs at genes and their families across the epiTRACERx cohort.**
CAMDAC DMR status at gene promoters reported as frequently hypermethylated in the literature (**Table 4.1**) and members of the same family across samples and divided by histological subtypes. If a given gene promoter comprises of more than one spatially segregated DMR, one being hyper- and the other being hypo-methylated, the bin DMR status is labelled as 'mixed'. Otherwise, if all regions are in the same direction of methylation, the methylation bin can be categorised as either having gained (hyper) or lost (hypo) methylation. Gene promoter-associated methylation bins that did not have coverage in all samples were excluded. Samples are ordered by patient ids and genes are ordered by family and prevalence, except for the bottom panel, where genes that showed distinct DMR status or prevalence between LUAD and LUSC are listed.

ported previously with a prevalence of 85% in a single biopsy cohort of 20 LUSC patients [199] while it was detected in 94.4% and 100% of epiTRACERx LUSC samples based on CAMDAC bulk and purified tumour methylation profiles, respectively. Interestingly, LncRNA PAX8-AS1 is regularly hypomethylated at its promoter. We hypothesise PAX8-AS1 activation further inhibits transcription of the overlapping *PAX8* gene. In summary, hypermethylation of *PAX* genes is omnipresent and usually ubiquitous, suggesting they occur early in tumour development, and thus that they may present an opportunity for early detection.

The promoter of tumour suppressor gene *RASSF1* was hypermethylated in 84.2% of patients (32 patient, 94 samples) after combining methylation bins associated with the locus, a much higher percentage than the previously published 30-38% range, showcasing the increased DMR detection power from CAMDAC purified methylomes. *RASSF2* was the second most commonly differentially methylated *RASS* family member and was observed in 44 samples (36.1%) from 15 patients (39.5%, 9 LUAD, 6 LUSC) followed by *RASSF5* (*NORE1A*) and *RASSF10* which were detected in 8 (21.2%) and 3 (7.9%) cases, respectively. Contrary to other family member, the *RASSF7* promoter often lost methylation in the tumour (15/122 samples) while *RASSF3* and *4* were never altered. *RASS* genes were not always clonally aberrated and were present in a lower percentage in tumours and thus likely occur later in tumour development, compared with *PAX* loci. Sporadically mutated genes are less useful for diagnostic purposes, but can be powerful biomarkers of disease prognosis. A larger cohort of *RASS*-negative patients would be needed to evaluate the prognostic value of *RASSF* promoter alterations.

Methylation of cadherins (*CDH*) and protocadherins (*PCDH*) is a known feature of several cancers, including NSCLC [176, 183]. Gain of methylation at the *CDH13* gene promoter is reportedly present in 16% of LUAD patients [176], while we detected it in 42% of samples (51 samples) representing 53% of NSCLC patients (20/38 patients, 16/24 LUAD and 4/14 LUSC). *CDH13* mutation was ubiquitous in 80% of cases. The most commonly epimutated *(P)CDH* family member were *PDCH8*, a known breast cancer tumour suppressor gene [200]. At least part of the

promoter region associated with this gene was hypermethylated in every single patient of the epiTRACERx NSCLC cohort (121/122 samples). Prior to CAMDAC, this alteration would have been missed from at least 6 samples. Because CAMDAC tumour-normal DMP calling has a very low false positive rate, even from bulk, we can use a minimum methylation difference threshold of 0.2, which is low compared with that used in other cancer studies (e.g. 0.25 [148], 0.3 [149, 150]). The number of false negatives would increase with higher effect size. The mean methylation difference across PAX6 DMRs was 0.384 in the bulk, compared with 0.629 post-deconvolution, stressing the importance of deconvolution.

*DAPK* gain of promoter methylation was observed in 27 samples from 9 patients, 8 of which were LUSC and therefore the epimutation rates were 57.1% in LUSC and 4.17% in LUAD. Both estimates differed widely from the published values in both 35-33% and 24-48%, respectively.

Genes of the laminin family, specifically *LAMA3*, *LAMB3* and *LAMC2* are known to the hypermethylated in both lung adenocarcinoma and squamous cell lung cancer cell lines. In the epiTRACERx data, *LAMB1* was most often hypermethylated, while *LAMA3* and *LAMC3* were LUSC- and LUAD-specific, respectively. *LAMP3* was often hypomethylated in LUSC and less often in LUAD. These observation suggests that *LAM* genes could be used as biomarkers of the two histologies, as well as *MLPH*, *CDH3* and *SAFTA3*, which were virtually always hypermethylated in LUSC. *CHFR* gain of methylation only occurred in LUSC and was found it 5 samples (9.10%) from 3 different patients (21.4%) which is comparable to published data. Two more genes showed bias between histologies, *CHMP4BP1* and *HMHA1*, showing increased methylation in one and loss of methylation in the other. This was possible as these loci exhibited intermediate normal methylation rates across all normal samples.

To summarise, we evaluated known hypermethylated genes and their families revealing numerous new loci that were also consistently hypermethylated as expected, while others were surprisingly hypomethylated. Epimutation prevalence was often higher in the epiTRACERx cohort then previously reported, thanks to de-

convolution with CAMDAC leading to lower false negative rates. Strikingly, a small number of DMRs were specific to LUAD or LUSC. We identify a number of genes with potential as biomarkers for early NSCLC detection and lung cancer subtype diagnosis. On a wider cohort, CAMDAC DMR calls may enable the identification of prognostic markers, for example to predict recurrence or outcome.

### 4.2.2 Intra- and inter-tumour sample relationships

In order to investigate sample relationships, we performed clustering of all tumour and normal samples. First, we extracted promoter methylation bins that were covered in all samples and that were tumour-normal DMRs in at least one sample pair. Promoters are enriched for CpGs that can modulate gene expression and are thus likely correlated with phenotype. Using the average methylation values, we carried out uniform manifold approximation and projection (UMAP) using the average normal or CAMDAC purified methylation estimate in normal and tumour samples respectively for each of these bins (**Figure 4.3A,B**).

We observe four main clusters, two of each cancer and normal samples. The two normal lung subgroups corresponded to either male or female samples (**Figure 4.3A**), presumably dominated by differences in chromosome X methylation across sexes. Tumour samples did not separate strongly by sex, likely due to the majority of female samples having at least partial (52%) if not complete (34%) chromosome X LOH. The main dividing feature between tumour samples was histological subtypes, potentially indicative of a different cell type of origin. This is in line with reports that a LUAD usually develops at the bronchioalveolar duct junction from the alveolar type I/II and Clara progenitor cells while LUSC are often located in the trachea and are thought to originate from aberrant basal cells [155]. Regions sampled from the same tumour were found to cluster in close proximity, reflecting ancestral relationships, while samples from different patients showed greater inter-sample distances (**Figure 4.3C,D**).

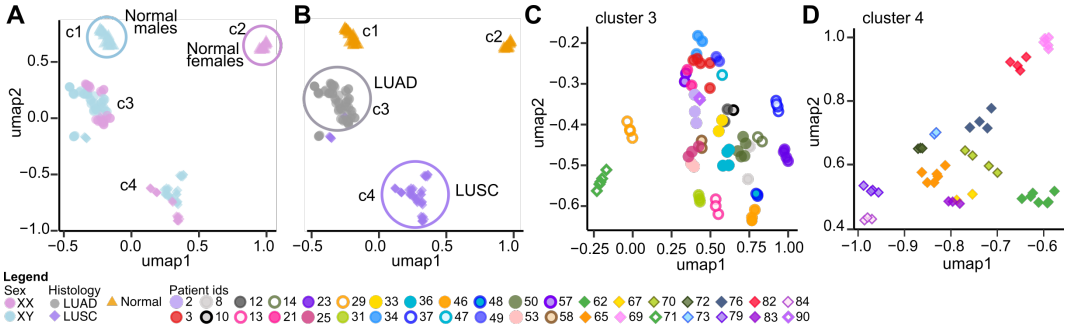


**Figure 4.3: Relationships between tumour and normal methylation profiles.** UMAP of the average deconvolved tumour and normal methylation rates across promoters-associated methylation bins that are DMRs in at least one sample, highlighted sample sex (A), histology (B) and patient of origin, zooming in the LUAD (C) and LUSC (D) clusters.

### 4.2.3 CAMDAC deconvoluted methylomes reflect phylogenetic relationships

To assess clonal relationships between samples, we used DMPs as opposed to DMRs and did not limit our analysis to promoter-associated CpG positions. In theory, this allows for the inclusion of potentially isolated stochastic methylation changes which can contain valuable information. For each patient, we sub-selected all CpG loci which were tumour-normal DMPs in at least one sample. Pearson correlations between tumour and normal methylation values at clonal bi-allelic DMPs, which make up the majority of DMPs, should be anti-correlated. Leveraging patient CRUK0062, selected for having the most tumour regions and highest level of methylation ITH as well as a wide range of purity and ploidy values, we demonstrate that CAMDAC deconvoluted methylation rates captured this effect, while on the contrary, the bulk tumour methylation profiles were increasingly correlated with the normal lung epithelium methylome with decreasing tumour DNA fraction (**Figure 4.4A**). This further supports that CAMDAC removes shared normal signals from bulk tumour data, enabling accurate comparison of pure tumour signals.

Genetic mutations identified in a given tumour are phylogenetically related, having evolved from a single founder cell. In recent years, several methods have been developed to obtain phylogenies from exome and genome sequencing experiments [107, 116]. We therefore make the reasonable assumption that DMPs should

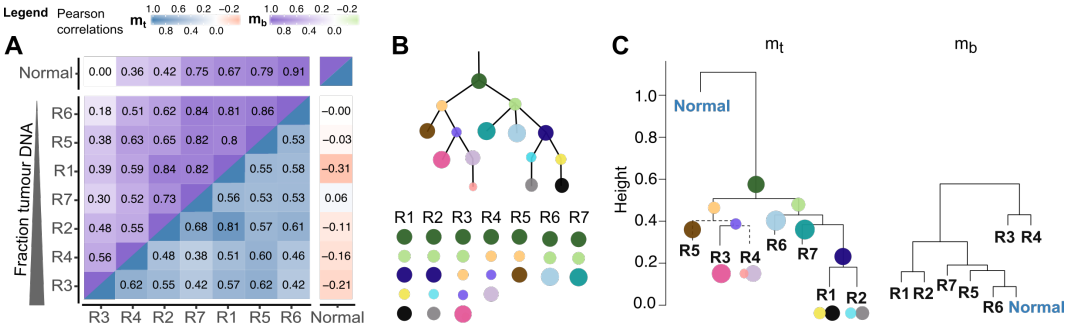


**Figure 4.4: CAMDAC deconvoluted methylomes are free from normal contamination and reflect phylogenetic relationships.**
(A) Pearson correlations between CRUK0062 tumour and adjacent normal samples in the bulk (left triangle) and purified tumour (right triangle). Tumour samples are ordered by tumour DNA content. (B) CRUK0062 phylogenetic tree inferred by SNV clustering and reproduced from published WES data [112]. (C) Hierarchical clustering of the same samples based on the purified (left) and bulk tumour (right) methylation rates at all CpGs that were differentially methylated in at least one tumour region, overlaying the SNV clusters from (C). Patient CRUK0062 was selected as an example for having the most tumour regions.

equally inform phylogenetic relationships between tumour samples. Using the same sites as in **Figure 4.4A**, we performed hierarchical clustering of CAMDAC pure tumour methylation rates and found that resulting sample clusters accurately mirrored evolutionary relationships between samples derived from SNVs (**Figure 4.4B,C**). In contrast, normal contaminated bulk tumour samples did not reproduce inferred sample relationships, and clustered by tumour DNA content. From this, we conclude that purified tumour methylation profiles present a unique opportunity to study ITH in solid cancers, unconfounded by signals from normal contaminating cells.

## 4.3 Discussion

We analysed CAMDAC tumour-normal DMR calls across the epiTRACERx lung cancer cohort. We suggest that a minimum of 3 tumour regions are needed to capture the epimutational landscape of a given tumour. The average DMR ubiquity levels were correlated with relapse-free survival, but not overall survival. We extracted promoter-associated differentially methylated CpG Islands and surveyed the five most commonly hypomethylated loci, all of which were known oncogenes. Next, we investigated published hypermethylated DMRs and genes of the same

families, revealing widespread hypermethylation across samples for several genes. Crucially, some alterations were specific to LUSC or LUAD. Prevalence as determined by CAMDAC tended to be higher than previously reported, likely due to increased sensitivity post-deconvolution. Unsurprisingly, clustering of promoter DMRs separated tumour samples by histology. Normal samples were separated by sex, as expected since we included chromosome X in our analyses. Tumour samples were not strictly segregated by sex, possibly due to most females having chromosome X LOH and the abundance of tumour-specific epimutations across the remainder of the genome. Finally, we perform hierarchichal clustering of individual DMPs and overlay SNV-derived phylogenetic trees, revealing that epigenetic and somatic mutations follow the same evolutionary trajectory.

In NSCLC, the 5-year survival for patients diagnosed in stage I is greater than 70%, whilst, alarmingly, the survival drops to about 15% in later stages. Widespread promoter hypermethylation events therefore present a unique opportunity for early cancer detection. In long-time smokers, *CDKN2A* hypermethylation in sputum was indicative of a 2-fold increase in the risk of lung cancer development [201]. While this and other tested genes did not perfectly predict NSCLC, the study suggests that targeted methylation analysis of bronchoalveolar fluid has potential in detection of early drivers. A routine minimally invasive blood or sputum test for NSCLC in high risk individuals could allow for early diagnosis and drastically improve patient outcome. CAMDAC DMR calling outputs for the epiTRACERx cohorts revealed several epimutations which could be included in such a targeted panel, including *PAX1, 3* and *5-9*, *RASSF1*, *PCDH8, 17* and *19*, *PCDHGA11* and *12* and *LAMB1*.

Hypomethylation is only weakly correlated with gene expression because several other elements must be in place to allow transcription. Nevertheless, recurrently hypomethylated loci could be valuable for diagnostic purposes and worthy of inclusion in a diagnostic panel given they occurred early in tumourigenesis. Although they are less frequent than gain of methylation events, our observations suggest such epimutations do exist, namely at the promoters of *TUBA1C*, *FAM83H-AS1*, *NELFCD*, *SFN*, *SH2B1*, and that these alterations play a role in lung cancer.

*LAMA3*, *CDH3*, *SFTA3*, *MLPH*, *HMHA1* and *CHMP4BP1* are histology-specific and would make a great addition to any NSCLC early detection panel as they could help distinguish between LUAD and LUSC upon diagnosis.

Lastly, epimutations that appear in a subset of samples could potentially serve as prognostic biomarkers. A larger cohort could give insight into the outcomes of LUAD and LUSC patients with and without certain alterations. Indeed, the upcoming release of the entire TRACERx cohort will provide unprecedented insight into spatial and longitudinal intra-tumour heterogeneity as well as lung cancer evolution in relation to survival and response to therapy and across ethnicity, smoking status, sexes and histological subtypes.

We show that CAMDAC purified methylomes contain phylogenetic information and that clustering of these profiles reveals inter-sample relationships. In order to build phylogenetic trees directly from epimutations, in a similar fashion to somatic mutations, subclonal reconstruction is required as described in Dentro *et al.* [111]. Developing such a tool is beyond the scope of this work. Alternatively, single cell bisulphite sequencing could enable phylogenetic reconstruction without the need for deconvolution as done in one study in chronic lymphocytic leukemia [149]. While the approach provides unique insight into DNA methylation heterogeneity and evolutionary trajectories, it comes with obvious drawback. For example, the depth of single cell data is typically very low (less then 1X), thus requiring binning of both copy number, mutational and methylation values, thereby losing single base pair resolution. Moreover, single cell experiments are technically challenging, lower throughput and more costly. Unless thousands of cells are sequenced per tumour, the output is unlikely to capture intra-tumour heterogeneity nuances as well as CAMDAC deconvoluted tumour data. Our results indicate that CAMDAC purified profiles captures epigenetic ITH given at least 3 tumour regions are sequenced, in line with observations from genome analyses of the same samples [112].

## 4.4 Methods

The methods described below were recently published as part of our bioRxiv preprint [140].

### 4.4.1 Tumour-normal DMR calling

First, CpGs are grouped into bins whereby neighbouring loci with inter-CpG distance below or equal to 100bp are grouped together. Bins with less than 3 CpGs with coverage equal or greater than 10 and 3 in the normal and purified tumour, respectively, were discarded. On average, 259,312 methylation bins were covered by all sampled tumour regions in a given patient (**Figure 4.5**) and 180,444 bins met the inclusion criteria in all 122 tumour samples and their matched normal.

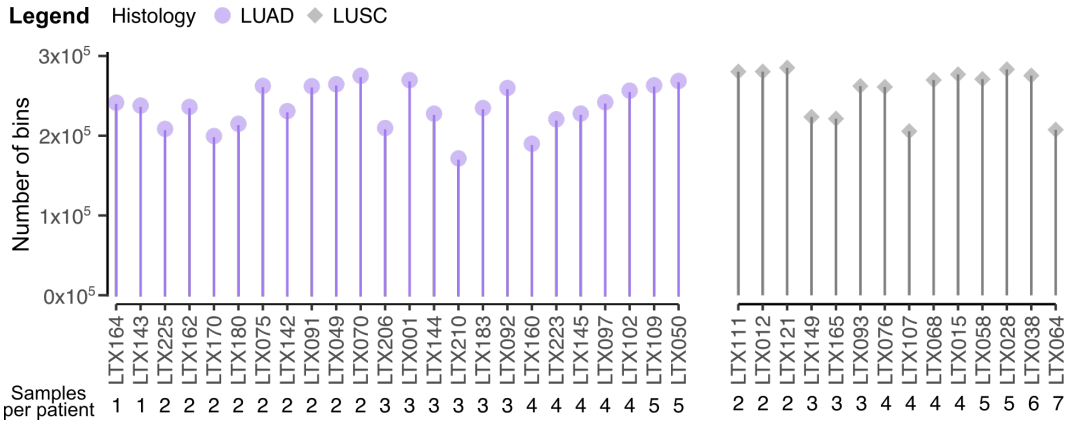


**Figure 4.5: Distribution of annotated methylation bins.**
Intragenic CpGs from exons, introns and UTRs are concatenated under the category intragenic to simplify the Venn diagram.

For each bin, we compute the total number of DMPs as well as the maximum number of consecutive DMPs with effect size 0.2 and $p < 0.01$ per bin. Genomic regions with at least $m = 5$ DMPs in total and $n = 4$ or more consecutive DMPs are deemed differentially methylated. The average number of CpGs and DMPs per bin were 15 and 2.5, respectively, and within DMRs, these values were 22 and 12. Limiting ourselves to regions covered in all tumour samples, we find on average 30,233 bins were classified as differentially methylated in at least one sample per patient, corresponding to 16.8% of methylation bins covered. This number goes up

to 35,591 when looking at genomic regions covered by all samples from individual patients, which represents 13.5% of patient-specific covered bins. Only 28,337 DMRs could be identified from bulk prior to CAMDAC (10.8% of patient-specific covered bins).

Each bin is annotated for downstream analyses. Annotated gene features include CpG Islands, shores and shelves, exons, introns, 5′UTR, 3′UTR, promoters and enhancers. CpG islands, intragenic features and enhancers genomic coordinates are extracted from Ensembl via biomaRt. Genomic coordinates of CpG island shores and shelves are defined as the regions 0-2kb and 2kb-5kb either side of a CpG island, respectively. When islands, shores and shelves annotations overlap due to neighbouring CGIs, the closest annotation to the island is chosen. Intragenic annotations are simplified for each gene to the GENCODE basic transcript set. We note that CGIs covered by RRBS are more often intragenic than intergenic (**Figure 4.6A**). Each gene transcript promoter is defined as starting 2.5kb upstream and ending 250bp downstream of the transcription start site. Promoters overlap significantly with CpG Islands, shores and shelves (**Figure 4.6B**). Enhancer regions are annotated along with the associated GeneHancer ids [202]. Enhancers show less overlap with CpG dense regions ((**Figure 4.6C**). For this work, we used CAMDAC with hg19 annotation set, but hg38 is also available. Note that a given CpG cluster can be associated to several features, although inter-/intragenic annotations are mutually exclusive and so are the three CGI annotations. The number of annotated bins covered for each category is consistent across the epiTRACERx cohort (**Figure 4.6D**).

To validate our DMR calls, we leverage SNV purified tumour methylation rates. First, we perform tumour-normal DMP calling following the same logic as in equation 14, but this time computing the probability that $P(m_{mut} > m_n)$, $P(m_{WT} > m_n)$ and their complements. We feed the DMP calls into CAMDAC DMR calling as described above. We then classified DMRs as either occurring on both alleles, only the mutant allele (in-*cis*) or only the wild type allele (in-*trans*). DMRs in regions with loss of wild type allele are categorised separately as it is not

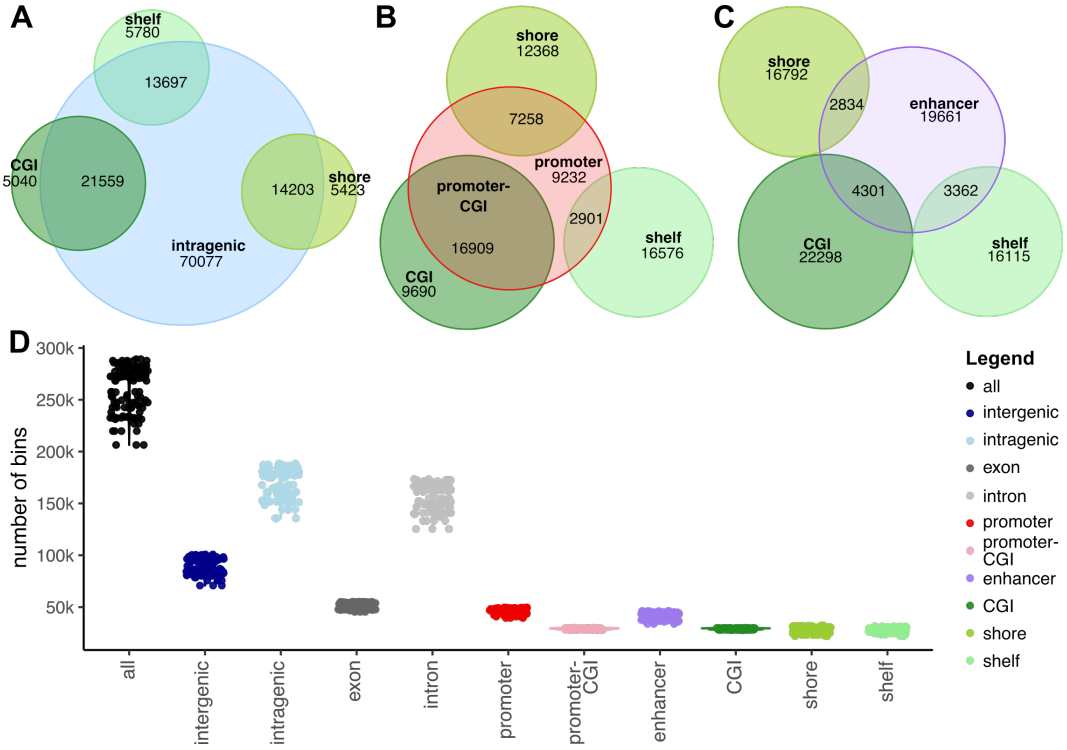


**Figure 4.6: Overview of genomic features covered by RRBS and annotated by CAMDAC.**
(A) Venn diagram showing the proportion of bins at CpG Islands, shores and shelves which are intra- *vs.* intergenic. Intragenic CpGs from exons, introns and UTRs are concatenated under the category intragenic to simplify the Venn diagram. (B-C) Comparing the overlap between CGI-related annotations and either promoters (B) or enhancers (C). (D) Breakdown of the number of bins assigned to each annotation. Each dot represent a sample.

possible to determine whether the DMR was in-*cis*, in-*trans* or on both alleles prior to the loss.

### 4.4.2 Clustering tumour and normal methylation profiles

We performed hierarchical clustering at the single patient level across 15 patients with 4 or more sampled tumour regions (average = 5 samples, range 4-7) to investigate inter-sample clonal relationships. Clustering was carried out on the purified tumour methylation rate at all CpGs that were DMPs in at least one of the tumour samples. The clustering output nicely fit the SNV phylogenetic trees derived from patient-matched multi-region WES data. This analysis was repeated on the bulk tumour methylation rate and the outputs were com-

pared. Bootstrap hierarchical clustering was performed using the R package pvclust (http://www.sigmath.es.osaka-u.ac.jp/shimo-lab/prog/pvclust/) with the hierarchical clustering method set to the average and using a Pearson distance matrix. For each clustering, we ran 100 bootstrap iterations with r-values going from 0.5 to 1.5.

Next, we compiled a list of all CpGs which fell within a promoter-associated tumour-normal DMR in one or more samples based on CAMDAC purified methylomes (totaling 8,570 gene promoters). We then calculated the mean methylation rate for each of those genomic regions across this cohort's 159 samples (122 tumour and 37 normal lung samples) and performed UMAP clustering analyses. We repeated this analysis selecting promoter regions based on the bulk tumour-normal DMR calls (totaling 8,387 gene promoters) feeding the mean bulk tumour methylation rates into the UMAP.

# Chapter 5

# Quantifying allele- and copy-specific methylation in NSCLC

## 5.1 Introduction

### 5.1.1 Allele-specific methylation in normal and tumour cells

Allele-specific methylation (ASM) is reported at a number of genomic loci and serves to modulate gene expression herein maintaining normal and possibly disease cellular functions. For example, germline imprinting involves methylation of one parental allele set during gamete formation and is maintained throughout epigenetic reprogramming, resulting in parent-of-origin expression [203]. Loss of imprinting (LOI) is associated with various diseases [204], and thus recent efforts have culminated in comprehensive mapping of imprint loci across the human genome [205]. Chromosome X inactivation in females is another well-characterised occurrence of ASM in normal tissues and involves methylation of one chromosome copy in females to allow for gene dosage compensation between the sexes [8].

### 5.1.2 Measuring allele-specific methylation

Allele-specific methylation rates can be obtained from bisulphite sequencing data by phasing CpG to heterozygous SNPs on the same read or read pair from single- and paired-end reads, respectively. In normal tissues, reports suggest that $\sim$5-8% of SNPs across the human genome lead to ASM [206]. Targeted bisulphite sequencing experiments enriched for CpG-dense DNA fragments revealed that 38%-88% of the observed intermediate CpG methylation signals across 16 normal

cell lines [207], were dependent on heterozygous loci, supposedly via the CpG-destroying allele itself or indirectly, for example, through TFBS ablation. One obvious drawback of this approach is that it requires neighbouring CpGs to identify ASM sites.

To overcome this limitation, a method was developed to infer ASM status directly from methylation data [208]. The authors designed this method for the analysis of normal diploid mammalian cell lines where ASM is the most likely source of intermediate methylation. They make the reasonable assumption that CpGs are either methylated on one allele or equally (un)methylated on both copies in all cells. They then compute the likelihood of each scenario at a given locus from the overlapping read distribution. The Bayesian information criterion is used to call regions of ASM. This model succeeds in identifying ASM in cell lines but breaks down when it comes to bulk tissue analyses where additional source of intermediate methylation are at play.

To our knowledge, the BED algorithm [106] is the only published method that enables to detect $k$ epialleles where $k$ can take values greater than 2. None of these tools can be directly applied to impure bulk tumour bisulphite sequencing data, which requires adjustment for both tumour purity and copy number to correctly identify allele-specific methylation. CAMDAC purified methylation rates therefore present a unique opportunity to gain insights into ASM in NSCLC.

### 5.1.3 Chapter summary

First, we show that CAMDAC SNP-independent $m_n$ enable visualisation of ASM. Evaluating chromosome X inactivation in females, we identify biases against the inactive X copy reducing its coverage and thus decreasing methylation levels across this chromosome. The BAF values at heterozygous SNPs were not skewed, X inactivation is random at the scale of our normal samples (i.e. it can silence either parental allele). Next, we assess methylation rates at the *H19/IGF2* germline imprinted locus. We validate the presence of ASM at this locus by phasing to heterozygous SNPs in 30/37 normal lung samples. Modifying CAMDAC equations to obtain allele-specific pure tumour methylation rates, we see LOI (without LOH) in

five patients. In every instance of LOI, we observe loss of methylation.

Next, we evaluate ASM levels in tumours in a more general manner. We assign epimutation copy numbers to hypermethylated DMPs identified by CAMDAC DMP calling. We quantify the extent of allele- and copy-specific methylation in regions of $1+1$ and $2+0$, using regions of $1+0$ as negative controls. We hypothesise that DMPs on one copy in $1+1$ reflect mainly stochastic methylation changes while those on both copies possibly reflect aberrant gene regulatory signalling. Regulatory DMPs constitute the (slim) majority of loci, at least in our dataset. In regions of $2+0$, stochastic DMPs can exist on both copies if they were acquired prior to the copy number gain. We filter out regulatory signals to extract stochastic DMPs, and, by comparing the number of clonal stochastic epimutations on 1 *versus* 2 copies, we obtained timings estimates for the copy number gain in epimutational time. Our results suggest copy number gains, at least in $2+0$, usually occur late in tumour evolution. Whether or not this holds for all copy number gains remains to be determined.

We perform DMR calling on SNV-phased CpG methylation estimates and find that allele-specific DMRs at these loci usually occurred on the same copy as the mutation (in-*cis*). In-*cis* DMRs were commonly hypermethylated, implying that the genetic mutations often lead to the ablation of a neighbouring TFBS. We also investigate the relationship between neo-antigen mutations and DNA methylation. We uncover that promoter hypermethylation can help suppress neo-antigen presentation and published this observation [114].

## 5.2 Results

### 5.2.1 Modelling allele- and copy-specific methylation rates

Previously, we introduced a simple model where the bulk tumour consists of two homogeneous cellular components: normal and tumour cells (**Figure 2.1**). Building on this model, we investigate the impact of ASM on $m_b$. To illustrate this effect, we recycle our previous example taking an unmethylated autosomal locus of tumour copy number $n_t = 3$ existing in a bulk tumour mixture of purity $\rho = 0.4$.

As before, we note that for most genomic CpGs, the tumour and normal populations will not be differentially methylated and thus, the bulk tumour, $m_b$, pure tumour, $m_t$, and normal, $m_n$, methylation rates will be equal and near 0 (**Figure 5.1**, top row). If this locus were to gain methylation on a single, two or all three copies in the tumour, the resulting bulk tumour methylation rate would fall near $m_b = \frac{1}{6}, \frac{1}{3}$ or $\frac{1}{2}$, respectively (**Figure 5.1**, rows 2-4). If present in a sufficiently large population of cells, allele-specific methylation may be detectable in the bulk tumour methylation rate distribution. We therefore speculate that intermediate CAMDAC purified $m_t$ values can, for the most part, be attributed to copy- or allele-specific differential methylation signal and attempt to quantify this.

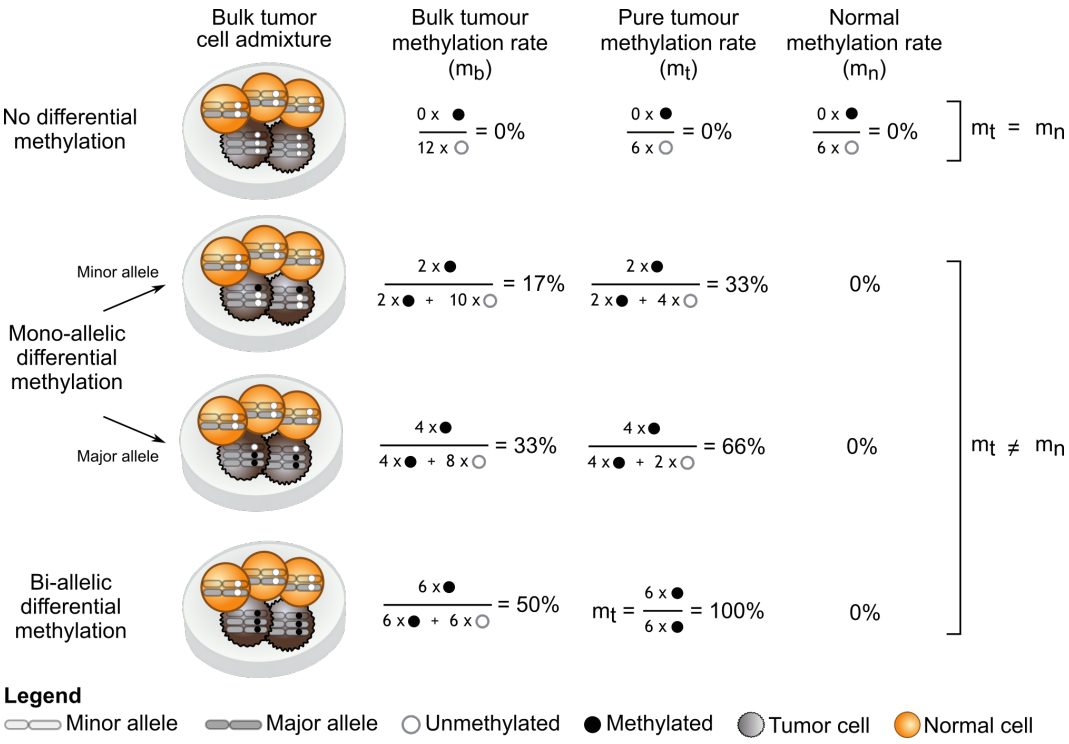


**Figure 5.1: Differential methylation between tumour and normal populations of cells leads to intermediate bulk methylation levels.**
Example bulk tumour ($m_b$), pure tumour ($m_t$) and normal ($m_n$) methylation rates in a bulk tumour mixture of purity $\rho = 0.4$ for an individual CpG (or a group of CpGs) of clonal tumour copy number $n_t = 3$ and with either no differential methylation (top row) or DMPs on the minor (second row), major (third row) or both alleles (fourth row). Note that a similar relationship is observed for loci that were methylated in the normal and underwent demethylation in the tumour, but with mirrored methylation rates.

### 5.2.2 Recapitulating known germline ASM events

We begin by evaluating the detection sensitivity of RRBS-derived CAMDAC methylation rates for known germline ASM signals, starting with chromosome X inactivation. In the female normal lung samples from the epiTRACERx cohort, we expect to find widespread ASM of one chromosome X copy. Interestingly, chromosome X modal intermediate methylation estimates were below 0.5 for all of the 13 normal female samples in this cohort (**Figure 5.2A**). Additionally, sequencing coverage on X in females was lower in comparison with autosomes (**Figure 5.2B**). We hypothesise that DNA extraction biases against the condensed Barr body (inactive chromosome X copy) lead to a reduction in coverage. There was no correlation between the mode of the methylation rate distribution and either smoking (Pearson correlation = -0.204, p-value = 0.502) or age (Pearson correlation = 0.155, p-value = 0.614). Observation of the BAF distribution at heterozygous SNPs did not reveal any deviation from the expected (**Figure 5.2C**), indicating that chromosome X inactivation is random in our albeit small female normal lung dataset.

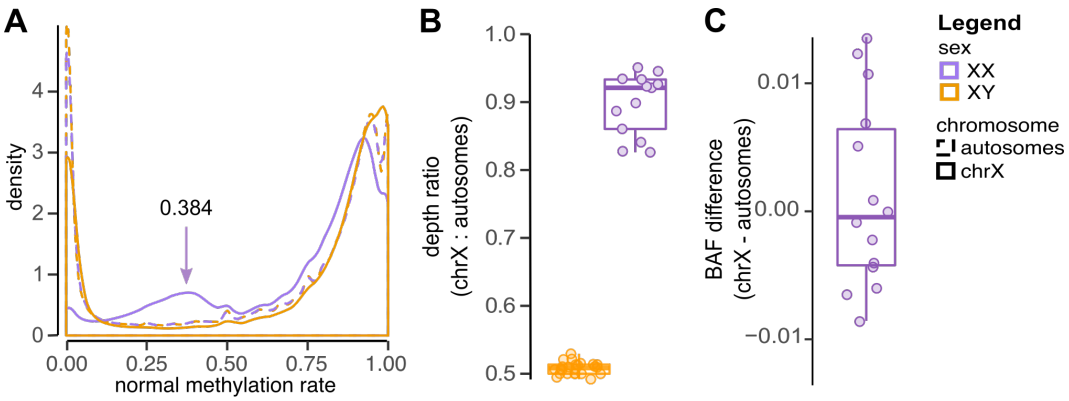


**Figure 5.2: Chromosome X modal methylation suggest RRBS coverage is skewed against the inactive copy.**
(A) Normal lung methylation rate density distribution across autosomes (dashed) and chromosome X (solid line) for male (orange) compared with female (purple) samples. (B) Ratio of the depth of coverage at autosomes versus chromosome X in males and females. (C) Boxplot showing the distribution of the difference between the median BAF at heterozygous SNPs across chromosome X minus autosomes in female samples.

Next, we investigated the imprinting control region (ICR) of the *IGF2/H19* locus [209] and assessed whether we could detect intermediate methylation at this well-known germline imprinted loci. The ICR is covered by RRBS data and the CpGs overlapping with the region had methylation rates around $m_n$=0.5 in all normal lung samples. For example, 86 CpGs spanning the ICR and neighbouring H19 promoter in sample CRUK0073-N appear confidently methylated on one of the two alleles as inferred from the bulk normal data ($m_n$ HDI$^{99} \subseteq [0.25, 0.75]$, **Figure 5.3A**, top panel). To validate the occurrence of ASM, we phased CpG methylation rate estimates to germline heterozygous SNPs, where possible. In 31/37 (84%) normal samples, we found at least one heterozygous SNPs overlapping with the ICR and, on average, 15 imprinted CpGs could be validated by phasing ($|m_{alleleA} - m_{alleleB}| > 0.7$, **Figure 5.3A** (top panel)). In contrast, phased methylation estimates at the H19 exon, where available, showed no allele-specific methylation.

LOI is linked to various diseases. Specifically, LOI has been reported at the *IGF2/H19* locus in colorectal cancer [210]. We therefore set out to investigate this phenomenon across our lung cancer cohort. We found LOI in CRUK0073-R2 (**Figure 5.3A**, bottom panel) as well as a further 4 tumours, all of which involved demethylation of the maternal allele rendering the ICR fully unmethylated (**Figure 5.3B**). Loss of heterozygosity also led to loss of imprinting in 8 additional tumours, but demonstrated no evidence of epiallelic bias.

### 5.2.3 Quantifying allele- and copy-specific methylation

CpGs outside of either imprinted regions and the inactive X chromosome copy in females are usually presumed to be symmetrically methylated on both alleles, at least in normal cells. Heterozygous SNPs at CpGs can also lead to intermediate methylation values, but, by design, methylation at polymorphic CpGs is computed by CAMDAC as the average per CpG allele (**Figure 3.1**). As a result, clonal allele- (and copy-)specific DMPs are likely the principal contributor of intermediate methylation signals in purified tumour methylomes.

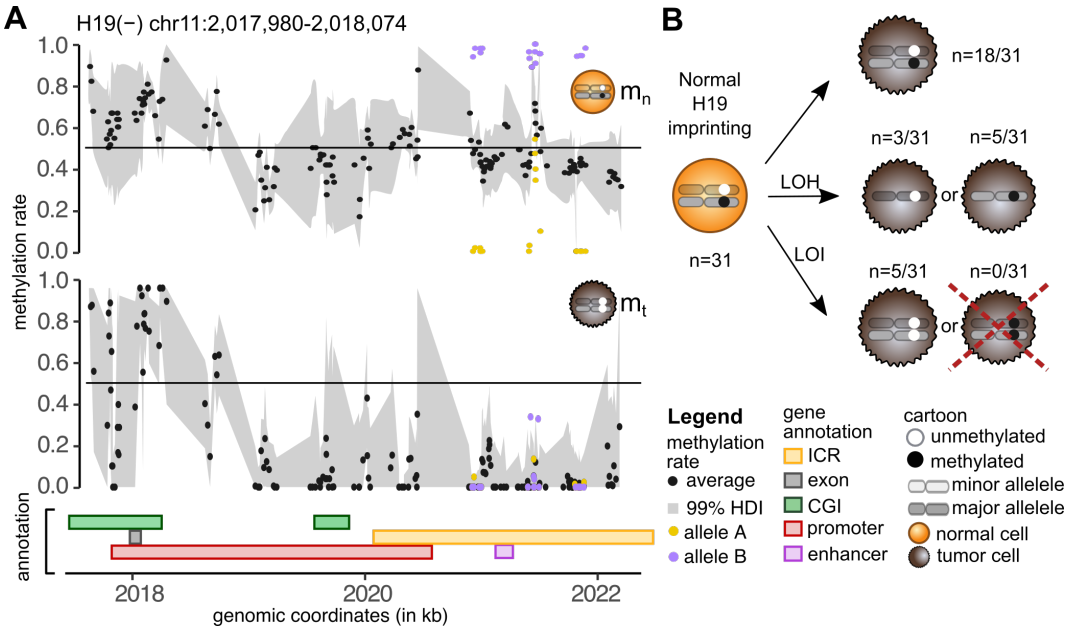


**Figure 5.3: Evaluating normal and tumour imprinting status at the *IGF2/H19* imprinted locus.**
(A) Normal lung (CRUK0073-N, top) and CAMDAC pure tumour (CRUK0073-R2, middle) methylation rate across the *IGF2/H19* imprinted locus. Black dots represent the average methylation point estimates per CpG allele and the grey ribbon is the $HDI^{99}$ around each of them. CpG methylation point estimates phased to either parental alleles are also displayed (purple and yellow). Genomic annotations of the GENCODE basic *H19* transcript (grey) and its promoter (red) defined as the region 250 and 2500bp downstream and upstream of the transcript start site, respectively. Neighbouring CpG islands (green), enhancers (magenta) and the *IGF2/H19* imprinting control region (yellow) are also shown (bottom panel). (E) Summary of the somatic changes observed at the *IGF2/H19* ICR.

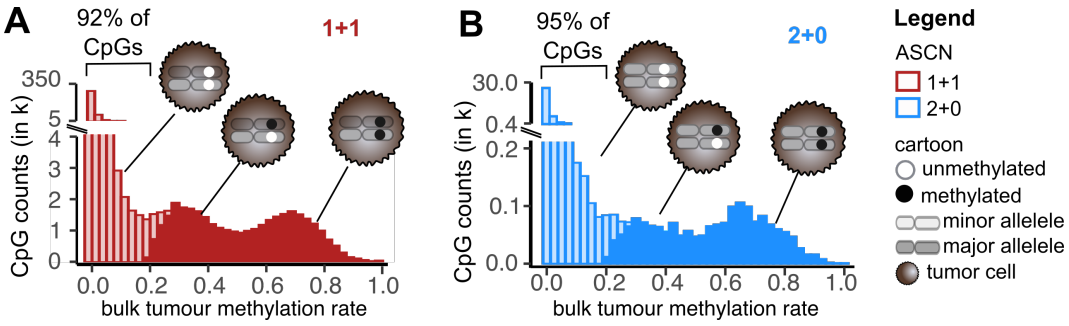


**Figure 5.4: Allele- and copy-specific hypermethylation.**
Tumour methylation rate distribution zooming in on tumour normal hypermethylated CpGs that were confidently unmethylated in the adjacent normal in regions of ASCN (A) $1+1$ and (B) $2+0$ in sample CRUK0062-R3.

To illustrate this effect, we selected loci that were fully unmethylated in the cell of origin, using the tumour-adjacent normal lung as proxy ($m_n$ HDI$^{99} \subseteq [0, 0.2]$) and focused on regions of total copy number $n_t = 2$. CpGs were then stratified according to ASCN (i.e. either $1+1$ or $2+0$). As there are only two copies, the clonal allele- or copy-specific DMPs clusters located around $m_t = 0.5$ should be easily distinguishable from peaks on either side coming from non-differentially methylated CpGs ($m_t = 0$) or DMPs on both copies ($m_t = 1$). Taking CRUK0062-R3 as an example, two peaks emerge in the tumour methylation rate histogram each corresponding to hypermethylated DMP populations with epimutation copy numbers 1 and 2 (**Figure 5.4**). We set out to quantify the relative size of these two methylation clusters in the epiTRACERx cohort. We subset our analysis to 58 bulk and 2 FACS sorted tumour RRBS samples that harboured both $1+1$ and $2+0$ ASCN segments each containing at least 200 tumour-normal DMPs as well as $1+0$ segments for use as proxy to calculate subclonal contamination (**Methods, section 5.4.3**).

Observations from the TRACERx100 cohort indicate that there is usually one major detectable genetic subclone per sampled tumour region [112]. Our findings suggest that epigenetic and genetic changes follow the same somatic evolutionary trajectories (**Figure 4.4B,C**), in line with recently published data from independent research laboratories [150]. In light of these facts, we hypothesise that one major clone exists per sampled tumour region and thus that most DMPs in $1+0$ should be present on one copy in all cells after CAMDAC. Consequently, the majority of DMPs should have pure tumour methylation rates near 1, assuming $m_t$ adequately reflects the variant allele frequencies of epimutations. We note that CAMDAC DMP calling requires a minimum tumour-normal absolute methylation difference of $|m_t - m_n| > 0.2$, and thus we cannot detect subclones with very low CCFs. However, we can measure the number of DMPs in subclones with intermediate CCFs values near 0.5 and compare this with the clonal signal near 1, thresholding on $m_t$. In $1+1$ and $2+0$, DMPs in $\sim$50% of cells present on both copies will appear at the same methylation rates as clonal DMPs present on one copy. We leveraged the observed

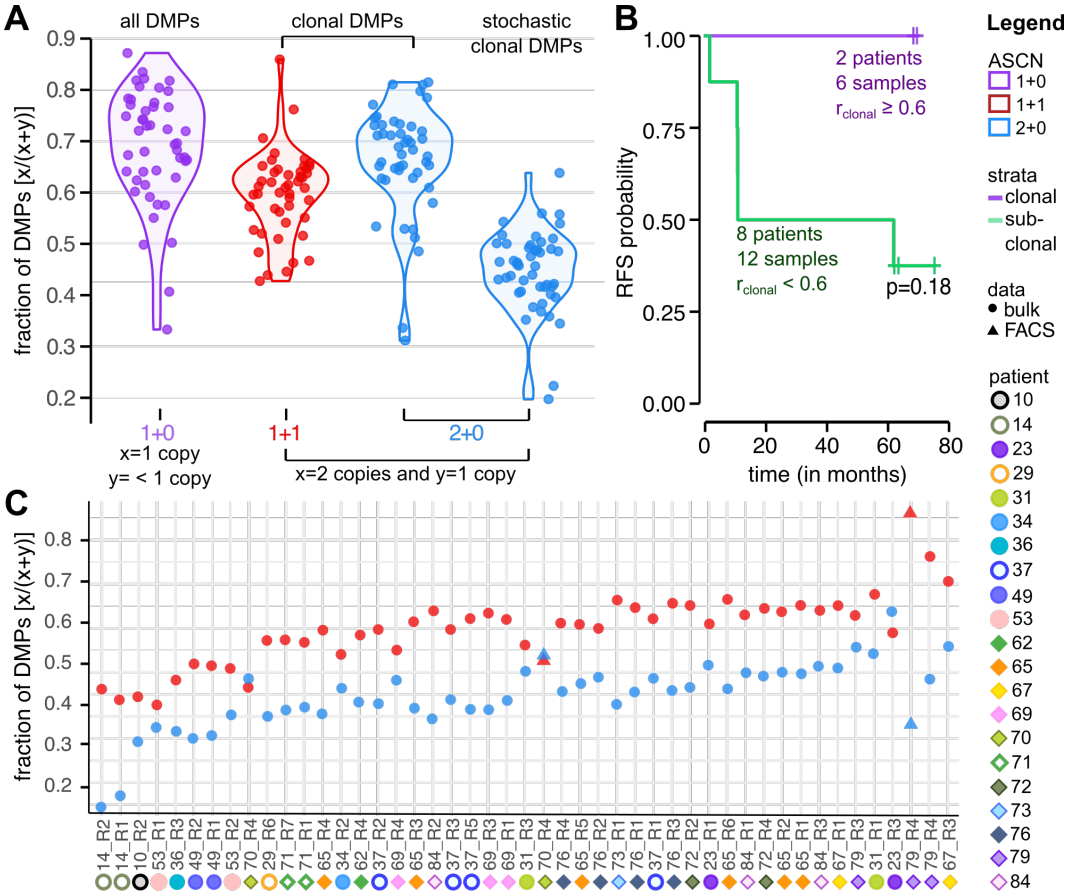


**Figure 5.5: Epimutation copy numbers give insights into subclonal, copy- and allele-specific methylation.**
(A) The first violin shows the fraction of clonal copy number $n_t = 1$ *versus* subclonal ($n_t < 1$) epimutations in $1+0$ (purple). The second and third depict the fraction of clonal DMPs with copy numbers 2 and 1 in $1+1$ (red) and $2+0$ (blue), after attempting to remove subclonal contamination. The fourth distribution is a variation of the third, but limited to what we estimate to be stochastic DMPs. (B) Relapse-free survival (RFS) in LUAD appears worse for highly subclonal tumours, but the effect is not statistically significant. Samples with subclonal scores falling below the $25^{th}$ quantile ($q^{25} = 0.6$) are deemed subclonal whilst the rest are classed as predominantly clonal. (C) Showing the same fractions as depicted in the second and fourth violin from (A), but per sample.

ratio of DMPs at high and intermediate deconvoluted methylation rates in $1+0$ to estimate the level of subclonal contamination in $1+1$ and $2+0$.

We calculated the fraction of clonal copy number 1 ($m_t \in [0.6, 1]$) *versus* subclonal ($n_t < 1$, $m_t \in [0.2, 0.6[$) epimutations in regions of $1+0$ allele-specific copy number and found that, on average, 69.8% of DNA methylation alterations were clonal (**Figure 5.5A**). The presence of subclonal DMPs as measured in regions

of $1+0$ was not significantly correlated relapse-free survival in LUAD (**Figure 5.5B**). Leveraging this clonality estimate, we corrected DMP counts in an attempt to reduce the impact of subclonal contamination of the clonal ASM peaks in $1+1$ and $2+0$. Indeed, subclonal DMPs present on both copies in a large subclone with say a CFF of 50% would overlap with the methylation rates of DMPs on one copy and pressent in all sells in regions with total tumour copy number $n_t = 2$. By measuring We obtained the number of DMPs with copy numbers 1 and 2, which we expressed as the fraction of epimutation on all copies *versus* that on a single copy. In $1+1$ and $2+0$ ASCN segments, the mean values were 0.597 (inter-quartile range [0.556, 0.643]), and 0.664 (inter-quartile range [0.634, 0.731], **Figure 5.5A**).

Speculating on the origin of DMPs, we propose that DNA methylation changes can be classified into two categories: (1) stochastic and (2) gene regulatory epimutations. We suggest that epimutations found on a single copy in segments of total copy number $n_t = 2$ are the result of stochastic methylation changes. These alterations can reflect errors in DNA methylation maintenance during replication, abnormal *TET* activity or the impact of neighbouring somatic mutation truncating or creating TFBSs [59, 65, 93]. Following clonal expansion, stochastic DMPs would be on one copy in all tumour cells assuming faithful replication. On the other hand, DNA methylation alterations in response to somatic regulatory changes should affect all copies equally. For instance, tumour specific silencing of a transcription factor may lead to hypermethylation of some of the associated TFBSs.

We propose that the infinite sites model [211] used in tumour evolutionary analyses [107] also applies to DMPs. As per this model, we assume that stochastic methylation changes are unlikely to occur twice at the same CpG position. This means that DMPs present on all copies in regions without loss of heterozygosity, such as $1+1$, are unlikely to be caused by stochastic alterations and are for the most part attributed to regulatory differential methylation. In $2+0$ segments, DMPs on both copies of the gained allele could be either regulatory or stochastic while those on one copy are likely to be stochastic.

We then computed the ratio of DMPs with copy numbers 1 and 2 in $1+1$

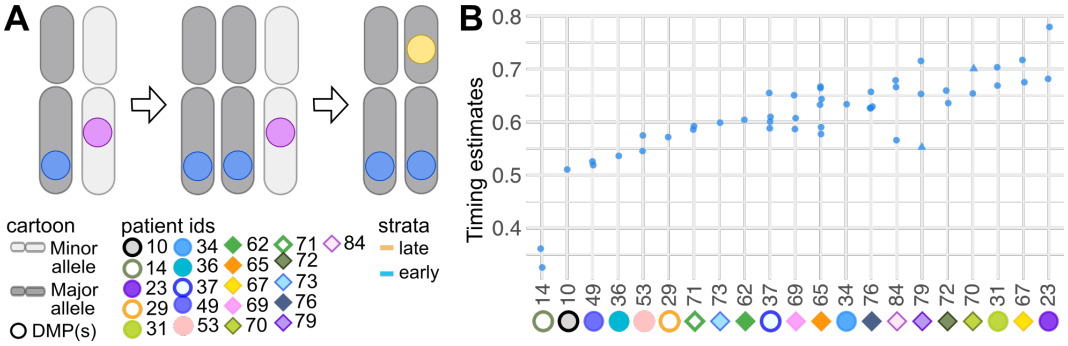


**Figure 5.6: Timing copy number gains in epimutational time.**
(A) Timing schematic. Stochastic epimutations appearing before the gain on the gained allele in $2+0$ are found on two copies while, in principle, those having occurred after the gain are present on a single copy. (B) Copy number gain timing estimates in $2+0$ across patients.

as proxy for the relative contribution of stochastic and regulatory epimutations to the overall number of DMPs. We leverage this fraction to extract the number of stochastic DMPs on all copies $2+0$. We then compute the fraction of stochastic epimutations only in $2+0$ with copy numbers 1 and 2. The mean ratio was 0.446 (inter-quartile range [0.410, 0.494], **Figure 5.5A**). In regions of $1+1$ and $2+0$, As the mutation rate makes sense because looking at large enough bins mut rate in constant that the values obtained for are correlated within samples **Figure 5.5C**). Samples from the same patients showed similar values and so did estimates from matched bulk and FACS sorted RRBS data.

In regions of $2+0$, we posit that stochastic epimutations on both copies were acquired prior to the copy number gain, whilst those on one copy occurred after (**Figure 5.6A** [124]). We utilise the stochastic clonal DMP counts on 1 ($counts_{1,stochastic}$) and 2 ($counts_{2,stochastic}$) copies in $2+0$ to derive timing estimates for this copy number gain as: $timing = \frac{2 \times counts_{2,stochastic}}{counts_{1,clonal} + 2 \times counts_{2,stochastic}}$ (**Methods, section 5.4.3**). Results suggest that gains usually occurred in the second half of epimutational time (**Figure 5.6B**), with the mean value sitting slightly higher in LUSC than in LUAD, 0.635 and 0.577 respectively. We found good agreement between timing values across samples from the same patient.

To conclude, we showed that allele-specific epimutations gives rise to intermediate methylation in CAMDAC purified profiles. We verified that detectable

epimutations were usually clonal, as measured by the ratio of DMPs on one or less than one copy in $1+0$. We then defined two classes of epimutations, stochastic and regulatory DMPs, and estimated the rate of each from epimutations in $1+1$ copy number segments. Leveraging clonal and stochastic DMP rates inferred from $1+0$ and $1+1$ segments we extracted the number of stochastic DMPs on a single or on both copies of the gained allele in $2+0$ and used these to get timing estimates for the copy number gain in epimutational time. Note that this result is based only on the timing of gains in $2+0$ in a small subset of epiTRACERx samples. Our observations and work by others [126] both suggest that, in lung cancers, copy number gains tend to occur late, at least in $2+0$. We note that we cannot evaluate whether our approach enables accurate timing of early gains. Nevertheless, our results suggest that methylation data does contain timing information, at least in $2+0$, and is a first step in developping a tool to harness this.

### 5.2.4 The interplay between mutations and somatic mutations

We briefly mentioned that genetic mutations can alter the methylation levels of neighbouring CpGs, for example through the ablation or creation of TFBSs. We set out to formally investigate the relationship between somatic mutation and differential methylation. We obtained DMR calls on the SNV deconvoluted CAMDAC pure tumour methylation rates already computed as part of analyses included in the previous chapters (**Methods, sections 3.4.4, 4.4.1**). Across all samples, we extracted a combined total of 5,727 phased methylation bins, 603 of which were DMRs (**Figure 5.7A**). This SNV deconvoluted DMR rate, 10.52%, is in line with the per sample average obtained from CAMDAC purified DMR calling which was 12.51%. We note that SNV deconvoluted methylation rates may have lower coverage and thus greater uncertainty which will influence DMR calling. The overlap between RRBS and WES is low, as opposed to WGS. We obtained 5 and 21 DMRs out of 36 and 247 SNV-phased methylation bins per sample on average, respectively for the two platforms. Given the 3 patients with WGS have on average 23,348 SNVs and that RRBS covers $< 2\%$ of the genome, the total number of SNV-phaseable methylation bins is sensible.

We then classified DMRs as either bi-allelic or mono-allelic with or without loss of the wild type (WT) allele. We further divided the latter category into epimutations that were in-*cis* with the mutant allele ($n_{cis} = 258$) and those which were specifically located on the WT allele (in-*trans*, $n_{trans} = 43$). The observed enrichment for DMRs in-*cis* to SNVs (Binomial test, $p = 1.67 \times 10^{-33}$) implies that a causal relationship exists between certain somatic genetic sequence alterations and aberrant DNA methylation.

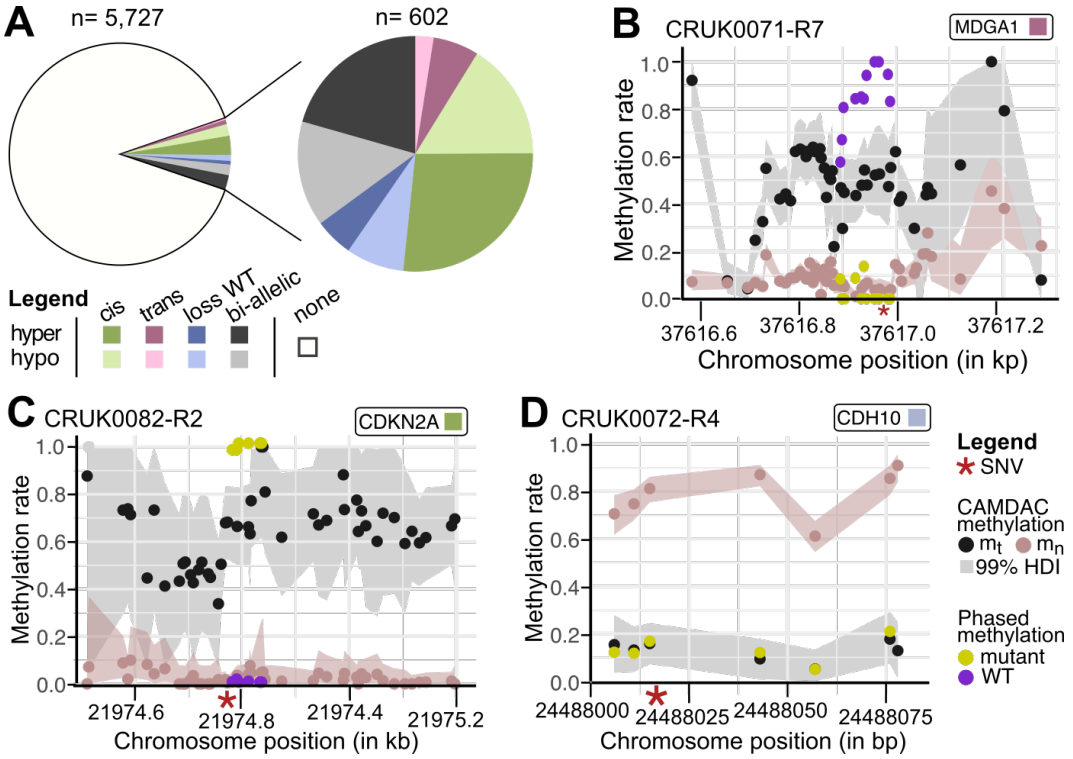


**Figure 5.7: Interplay between somatic mutations and methylation changes.**
(A) Pie chart of the methylation status of SNV-phaseable regions. Categories include bi-allelic, WT loss, in-*cis*, in-*trans* and non-DMRs. Epimutations were further divided into into hypo- and hyper-methylated loci. (B-D) Examples of (B) in-*trans*, (C) in-*cis* and (D) loss of WT phased DMRs.

Promoter hypermethylation, or possibly intragenic hypomethylation, of the wild type allele combined with a deleterious mutant allele likely results in stronger downregulation of tumour suppressor genes than one modality alone, offering a possible explanation for the occurrence of DMRs in-*trans* to SNVs. Alternatively, activating mutations at oncogenes together with together with hypomethylation of

the wild type allele could also favour tumour progression. The theory is that the combined effect of a somatic mutation and DNA methylation change on separate copies leads to greater impact on expression than each alteration alone.

Phased epimutations in-*trans* relative to SNVs were usually hypermethylated ($p = 2.19 \times 10^{-3}$, Binomial test). Enhancer GH06F037648, which is intragenic to *MDGA1*, was clonally hypermethylated on at least one wild type allele copy in all CRUK0071 regions (**Figure 5.7B**). Findings from the ENCODE project predict that this enhancer also regulates *RNF8* expression, a reported tumour suppressor gene in breast cancer thanks to its role in DNA double strand break repair and Notch signalling downregulation [212]. Other studies in both LUAD and breast cancers suggest that, on the contrary, activation of this ubiquitin ligase is indicative of tumour progression because it can stabilise transcription factors *RXR*$\alpha$ and *Twist* promoting epithelial–mesenchymal transition and increased proliferation [213]. The various roles on *RNF8* in tumourigenesis are discussed in a review by Zhou *et al.* [214].

DMRs in-*cis* with respect to the mutant allele showed no bias for either gain or loss of methylation. The occurrence of either hyper- or hypomethylated DMRs in-*cis* may depend on whether the SNV results in the deletion or, although less likely, the formation of TFBSs, respectively. In the case of hypomethylated regions in-*cis*, a more plausible explanation is that the loci is in fact hydroxymethylated, an intermediate state in tumour demethylation pathway which is unaffected by bisulphite conversion and thus indistinguishable from methylated CpG in RRBS data. Patient CRUK0082 harbours a clonal promoter mutation at the tumour suppressor gene CDKN2A where only the mutant allele hypermethylated (**Figure 5.7C**).

In cases where the mutant allele was differentially methylated and the WT allele was lost in the tumour cells, we could not establish whether the DMR originally occurred on both alleles or in-*cis*. There was WT loss at the *CDH10* gene locus in patient CRUK0072 (**Figure 5.7D**). The remaining copies harboured a missense mutation and hypomethylation of the surrounding CpGs. Specifically, this example SNV-phased DMR was found in a region of both LOH and WT allele loss and thus the SNV was present on all remaining copies. As such, CAMDAC purified

methylation rates were in excellent agreement with SNV deconvoluted estimates as previously detailed (**Chapter 3, sections 3.2.5, 3.4.4**).

Taken together, CAMDAC deconvolution and phasing enables deeper understanding of the interplay between genetic mutations and aberrant DNA methylation. In our lung cancer cohort, we observed frequent phasing of the hypermethylated allele to the mutant SNV allele, potentially through ablation of adjacent transcription factor binding sites.

### 5.2.5 DNA methylation and immune escape

Somatic sequence alterations can create tumour-specific neo-antigens that can be recognised by competent immune cells initiating an anti-tumour immune response. This not only requires effective antigen presentation to the cell surface, but also the presence of relevant immune effectors nearby to detect these molecules [215]. In Rosenthal *et al.* [114], we show that, in NSCLC, tumour cells can suppress antigen presentation via several different mechanisms, including loss of heterozygosity at the human leukocyte antigen locus, loss of function mutation of the major histocompatibility complex, alterations at the enhanceosome of these loci, aberration of the peptide generation pathway and epigenetic silencing of genes harbouring neo-antigens. The analysis linking DNA methylation with immune escape in Rosenthal *et al.* was carried out by the author of this thesis.

Leveraging a list of expressed (mutant read counts $> 30$) and non-expressed (mutant read counts $= 0$) neo-antigenic transcripts derived by colleagues, we set out to evaluate whether promoter methylation could play a role in modulating neo-antigen transcription in NSCLC (**Methods, section 5.4.4**). Neo-antigens, expression levels and methylation rates were respectively determined from matched WES, RNA sequencing and RRBS data of 79 TRACERx samples from 28 different patients.

We obtained CAMDAC tumour-normal DMP calls for all CpGs contained in the region spanning 2kb down- and upstream of the most upstream transcription start site at genes harbouring neo-antigens. Note that methylation rates at the promoters could not be phased to distant intragenic neo-antigen mutations. For exam-

ple, the *LAMB1* locus has two distinct intragenic neo-antigen mutations in sample CRUK0057-R1. These mutations are not expressed and the *LAMB1* gene promoter is hypermethylated (**Figure 5.8**, left). In comparison, a randomly sampled unmutated control sample was not hypermethylated (**Figure 5.8**, right).

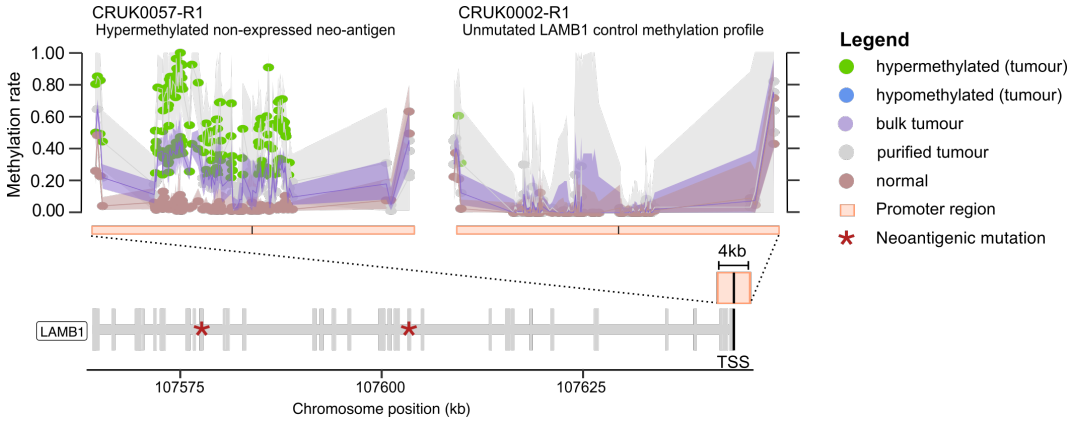


**Figure 5.8: Hypermethylated non-expressed neo-antigen example.**
Methylation rates at CpGs overlapping the most upstream transcription start site of the *LAMB1* gene locus. Under the methylation profiles the two neo-antigenic mutation are shown relative to all exonic regions across *LAMB1* transcripts. The most upstream transcription start site is highlighted. Methylation rates are shown for the normal (rose), bulk tumour (purple) and pure tumour (grey) in sample CRUK0057-R1, which harbours two non-expressed neo-antigens and is hypermethylated (green, left) and for an un-mutated control sample, CRUK0002-R1, which is neither mutated nor differentially methylated (right).

Strikingly, we saw a 11.4-fold increase in the number of gene promoters harbouring hypermethylated DMPs for non-expressed compared with expressed neo-antigens ($\chi^2$ test, p-val = $1.6 \times 10^{-4}$, **Figure 5.9A**). To ensure gene hypermethylation was correlated with the presence of neo-antigen mutations and mutant expression levels, as opposed to the gene loci themselves, we randomly sampled an un-mutated controls for each gene in **Figure 5.9A**. As predicted, controls were usually non-differentially methylated, irrespective of the expression level of the matched mutant transcript (**Figure 5.9B-C**). Overall, this analysis implies that promoter hypermethylation plays a role in silencing the expression of tumour neo-antigens.

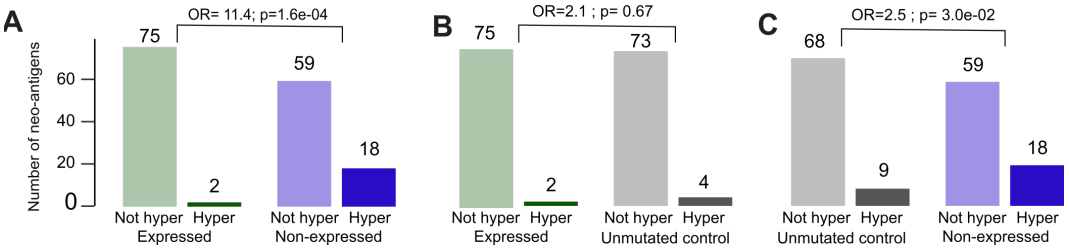


**Figure 5.9: Hypermethylation across expressed *versus* non-expressed genes harbouring neo-antigenic mutations in (epi)TRACERx.**
Comparing promoter hypermethylation status at (A) non-expressed *versus* expressed neo-antigens, (B) expressed neo-antigens *versus* unmutated controls of the same genes and (C) non-expressed neo-antigens *versus* unmutated controls of the same genes.

## 5.3 Discussion

In summary, we showed that allele (and copy-)specific methylation is a common feature of normal lung tissue and NSCLC. Surprisingly, we found that the modal methylation on chromosome X in females was lower than the expected value. We saw reduced coverage at the inactive X copy due to extraction biases against the Barr body, at least in RRBS data. In light of this finding, we advise researchers to take caution when interpreting methylation rates on X in females. Next, we evaluated allele-specific methylation at the *H19/IGF2* germline imprinted locus. We detected ASM in all normal samples and were able to validate this by phasing to heterozygous SNPs in 30/37 samples. We identified loss of imprinting in five patients, all of which lead to complete loss of methylation at CpGs spanning the imprinting control region.

We went on to evaluate allele- and copy-specific methylation in a more general setting across tumour samples. From this, we defined two classes of DMPs, stochastic and regulatory DMPs and assessed the relative quantities of each based on the ratio of epimutations with copy numbers 1 and 2 in $1+1$ regions. Our results show that regulatory DMPs tend to dominate the mutational landscape of our NSCLC samples, although we cannot exclude that this is in part due to RRBS data being skewed for CGIs.

Leveraging the estimated number of stochastic alterations on 1 versus 2 copies, we attempt to extract timing values for copy number gains in $2+0$ copy number seg-

ments only. An obvious drawback of our method is that we were limited to timing gains in regions of $2+0$ and in samples which also harboured a sufficient number of DMPs in regions of $1+0$ and $1+1$, which were used for subclonal and regulatory DMP decontamination, respectively. Nevertheless, timing estimates in regions of $2+0$ suggest that virtually all of these gains take place in the second half of epimutational time. This result is in line with published timing of LUAD samples derived from whole-genome sequencing data, the current state-of-the-art for timing purposes [124]. We note that a different cancer type cohort with both late and early gains would be better suited for the development of a methylation-based copy number timing method. Nevertheless, our data indicates that methylation sequencing data may have the potential to inform on tumour evolutionary trajectories.

Indeed, studies of whole genomes have enabled timing of somatic mutation in tumour evolution. To do this accurately and for all gains - as opposed to limiting ourselves to $2+0$ - one is required to perform clustering of (sub)clonal mutations. Although beyond the scope of this work, a probabilistic approach to assign epimutation copy numbers, cluster clonal DMPs and from this compute timing estimates may enable timing of higher copy number gains. Identifying recurrently early clonal, late clonal or late subclonal epigenetic alterations has obvious clinical implications, in early diagnosis, disease monitoring and for developping new treatments.

A similar approach has been utilised to situate copy number events in mutational time, as detailed in Jolly and Van Loo [216] and Gerstung *et al.* [124]. However, this approach is not directly applicable for use with DNA methylation data, where both subclonal and regulatory contamination must be taken into account. Moreover, bleeding between peaks of non-differentially methylated CpGs and subclonal and clonal mutations is likely much more significant in methylation data than for SNVs, posing an additional challenges to clonal reconstruction.

For example, normal methylation levels will vary according to cellular composition and may be susceptible to tumour field effects, while in comparison, genetic sequence alterations rarely reach a detectable CCF in healthy tissues [217]. At high

tumour DNA content, normal methylation noise is unlikely to affect clonal DMP calls, but at low tumour DNA fraction, it could have a non-negligible effect, especially for subclonal DMPs with low effect sizes. Furthermore, epimutations are reversible, meaning that clonal epimutations can revert to the ground state in a subset of cells, decreasing the methylation rate difference between non-differentially methylated CpGs and clonal epimutations.

We note that whole-genome bisulphite sequencing would be particularly well-suited to methylation timing analysis since the entire genome contains many individual isolated CpGs unlikely to be affected by regulatory changes and thus able to reflect accurate timing, unlike RRBS which is enriched for regulatory elements.

Subsequently, we showcased the power of RRBS as a tool to study the interplay between epigenetic and somatic changes. We classified SNV-phased DMRs as bi-allelic, in-*cis*, in-*trans*, on both alleles or on the mutant allele in regions of WT loss. Overall, we identified very few recurrent SNV-phased DNA methylation changes between patients. This is likely due to fact that SNVs need to occur at specific bases in order to cause in DNA methylation changes. Most SNV-phased DMRs were in-*cis* and hypermethylated, suggesting that somatic mutations can result in TFBS ablation thereby preventing transcription factor binding and allowing *DNMT* activity.

The rarity of loci harbouring SNV-phased DMRs in more than one patient could also be explained in part by the small cohort size, the low mappability of RRBS reads with alternate alleles (**see Chapter 2, section 2.2.2**) and the weak overlap between RRBS and WES data from which DMR and SNV calls were made, respectively. In future, trading two sequencing platforms for either Nanopore sequencing or matched paired-end WGBS and WGS data would help address the latter two potential sources of bias, but would significantly increase experimental costs.

It is important to note that despite totalling 122 samples, the epiTRACERx study comprises of only 38 different patients, 14 LUSC and 24 LUAD, at the time of writing. In the next epiTRACERx RRBS dataset, additional samples from both

current and new patients will be available and thus it may be possible to identify genes with recurrent SNV-phased DMRs. Specifically, frequently in-*trans* alterations would suggest strong selective forces for aberrant gene regulation. In the absence of recurrent individual hits, one could look at enrichment for in-*trans* aberrations within gene sets. Similarly, a larger cohort would be needed to fully appreciate the impact of LOI and subclonal methylation on clinical outcome in NSCLC.

Finally, we describe work completed as part of Rosenthal *et al.* [114]. We leverage mulit-omics data to unveal the relationship between immune escape and DNA methylation. We find that promoter hypermethylation is enriched at genes containing non-expressed as opposed to expressed neo-antigenic mutations. By comparison with unmutated controls of the same genes, we confirm that hypermethylation is specific to non-expressed neo-antigens. This finding has significant clinical implications, suggesting that combining a *DNMT* inhibitor with immune checkpoint blockade therapy could improve anti-tumour response and patient outcome in NSCLC.

## 5.4 Methods

The methods described below were recently published as part of our bioRxiv preprint [140].

### 5.4.1 Chromosome X allele-specific methylation analysis

First, we compute the median coverage on autosomes ($med_{autosomes}$) and on chromosomes X ($med_X$) for each sample, excluding pseudoautosomal regions (PAR1 and PAR2) in males. To correct for spatial correlation between CpGs on the same read molecule, we only include the coverage values of 1 CpG per read ($cov_i$). We obtain the log normalised depth ($LogD_i$) estimates for the $i^{th}$ CpG: $LogD_i = log_2(\frac{cov_i}{med_{chr}})$, where $med_{chr}$ is the median depth across all chromosomes, excluding Y in males. We centre the LogD distribution around 0 by subtracting the median ($med_{LogD}$): $LogD_{i,corrected} = LogD_{i,corrected} - med_{LogD})$. Finally, we get the median $LogD_{i,corrected}$ values on X and convert it to the base 2 exponent to obtain the coverage ratio, dividing samples by sex. The mean coverage ratio for males and

females on X *versus* the rest of the genome were 0.523 and 0.902, respectively.

The modal methylation rate on X in females was computed as the local maximum of the methylation rate density falling in the [0.3,0.7] interval taking 1 CpG per read on chromosome X in all 13 normal female lung samples. Since one copy of X is unmethylated and the other is methylated, we expected a value near 0.5, but the observed mode of methylation was 0.384. The observed reduction in coverage is thus likely biased against the inactive X copy, leading to a decrease in global methylation.

To see whether there was a corresponding allelic bias, we compiled the ASCAT.m normal BAF values for all heterozygous SNPs (**see Methods, section 2.4.3.1**) and compared the means on autosomes and on chromosome X in females, 0.454 and 0.453, respectively. We did not observe allelic bias, suggesting X-inactivation is random in our normal lung dataset.

### 5.4.2 Allele-specific methylation at the *H19/IGF2* imprinting control region

We obtained allele-specific methylation rate estimates at the ICR of the *H19/IGF2* by phasing the germline heterozygous SNPs. The normal methylation rate for the reference allele, $m_{n,ref}$, is computed from the methylated ($M_{n,ref}$) and unmethylated ($UM_{n,ref}$) reads phased to the reference allele: $m_{n,ref} = M_{n,ref}/(M_{n,ref} + UM_{n,ref})$. *Vice versa*, the alternate allele normal methylation rate, $m_{n,alt}$, is calculated as: $m_{n,alt} = M_{n,alt}/(M_{n,alt} + UM_{n,alt})$. Overall, 31/37 patients had polymorphisms at the *H19/IGF2* ICR. On average, we could phase 15 CpGs with clear allele-specific methylation at the ICR ($m_{alleleA}$ - $m_{alleleB} > 0.7$) per sample.

Obtaining bulk tumour methylation rates at the reference and alternate allele, $m_{b,ref}$ and $m_{b,alt}$, respectively, follows the same principle. To obtain the purified tumour methylation rate, we first assign clonal copy numbers to each allele using the BAF at the heterozygous and the allele-specific copy number from ASCAT.m. If BAF $< 0.5$, then major allele, $n_A$, is the reference allele copy number, $n_{ref}$=$n_A$ and the minor allele, $n_B$, is the alternate allele, $n_{alt}$=$n_B$. *Vice versa*, if BAF $> 0.5$,

$n_{alt}$=$n_A$ and $n_{ref}$=$n_B$. Both the reference and alternate allele methylation rates are confounded by signal from normal contaminating cells and must be deconvolved. We modified CAMDAC equations 2 and 3 for this purpose, where x is either the reference or alternate allele:

$$m_{b,x} = \frac{\rho n_{t,x} m_{t,x} + n_{n,x} m_{n,x}(1-\rho)}{\rho n_{t,x} + n_{n,x}(1-\rho)} \tag{15}$$

$$m_{t,x} = \frac{m_{b,x}(\rho n_{t,x} + n_{n,x}(1-\rho)) - n_{n,x} m_{n,x}(1-\rho)}{\rho n_{t,x}} \tag{16}$$

Using $m_{t,x}$, we classified tumour samples has having retained imprinting when absolute differences between the reference and alternate allele phased methylation rates at 1 or more CpGs across the ICR were larger than 0.7 (i.e. $|m_{t,ref} - m_{t,alt}| > 0.7$). Assignments were then manually validated. Tumours with loss of heterozygosity were classified as having lost either the inactive or active copies by taking the average methylation rate across CpGs within the ICR which were phased to the remaining allele. Similarly, in tumours with loss of imprinting, we took the average methylation level across CpGs at the ICR phased to either alleles to determine whether gain or loss of methylation had occurred.

### 5.4.3 Quantifying epimutation copy numbers

We set out to quantify allele-specific *versus* bi-allelic DMPs. Unlike SNVs, DMPs can occur as a results of regulatory changes and so can take any copy number and DMP populations are larger in number than genetic mutational clusters. The combined effect of a wide range of multiplicity and large cluster size lead to significant overlap between copy- and allele-specific and bi-allelic DMP populations with increasing total copy number. We therefore limited our analysis to $1+1$ and $2+0$ copy number segments. We used $1+0$ as a negative control as there should be no ASM in 1+0 region. We subset our analysis to 58 bulk and 2 FACS sorted tumour RRBS samples that harboured both $1+1$ and $2+0$ ASCN segments each containing at least 200 DMPs as well as $1+0$ segments for use as ASM negative control.

In those samples, we selected CpG loci which were confidently unmethylated in the adjacent normal ($m_n$ $\text{HDI}^{99} \subseteq [0,0.2]$) and hypermethylated in tumour cells ($P(m_t > m_n) > 0.995$ and $m_t - m_n > 0.2$). The methylation rate distribution at hypomethylated DMPs are less informative than at hypermethylated DMPs because of increased noise (**Figure 3.6D**). We divide the pure tumour methylation space into discrete bins: $m_t \in [0.2,0.6[$ or $[0.6, 1]$.

In $1+0$, hypermethylated DMPs falling in the second interval are likely clonal with copy number 1, while those below are presumable subclonal. At very low purity, CAMDAC estimates have high uncertainty and DMP calls are biased against subclonal and mono-allelic DMPs. We trimmed 13 samples with low tumour DNA fraction ($f_t \leq 0.20$) to remove any correlation between tumour purity and the ratio of clonal to subclonal DMPs counts ($r_{clonal}$). After trimming, the correlation between the mean sample purity and $r_{clonal}$ per patient was -0.117 (p-val $= 0.614$). On average, $r_{clonal}$ counts ratio was 0.698 across all samples and 0.660 in LUAD and 0.707 in LUSC.

Leveraging $r_{clonal}$ values computed above for each sample, we then corrected DMP counts of copy number 1 ($counts_1$) for subclonal contamination: $counts_{1,clonal} = counts_1 \times r_{clonal}$. We could then compute the ratio of clonal copy number 1 and 2 ($counts_2$) DMPs for each sample and for ASCNs $1+1$ and $2+0$. In $1+1$, we used this fraction as a proxy for the ratio of stochastic to regulatory DMPs, $r_{stochastic} = \frac{counts_{1,clonal}}{counts_{1,clonal} + counts_2}$. From this, we extracted the number of DMPs having occurred from stochastic methylation changes only at copy number 2 in $2+0$ $counts_{2,stochastic} = counts_2 \times r_{stochastic}$. Finally, we obtain the ratio of stochastic epimutation with copy numbers 1 and 2 in $2+0$. DMPs on one copy must have occurred after the gain took place while those on 2 copies were already present prior to the copy number alteration. The ratio of late to early epimutations is calculated as: $\frac{counts_{2,stochastic}}{counts_{1,clonal} + counts_{2,stochastic}}$. Finally, we obtained copy number gain timing estimates leveraging stochastic DMP counts on 1 and 2 copies in $2+0$: $timing = \frac{2 \times counts_{2,stochastic}}{counts_{1,clonal} + 2 \times counts_{2,stochastic}}$.

### 5.4.4 Evaluating promoter hypermethylation at genes harbouring neo-antigens

These methods were previously published in Rosenthal *et al.* [114] and in the results of a collaboration between the author of this thesis, the authors of the paper and the wider TRACERx consortium.

Matched RNA sequencing data was available for 28 out of the 38 patients included in the epiTRACERx with multi-region sequencing totalling 79 out of 122 tumour samples. Matched normal RNA sequencing data was not available. To exclude genes that are not expressed in lung cancer, Rachel Rosenthal compiled a reference lung cancer transcriptome comprised of genes that were ubiquitously expressed in the TCGA non-small cell lung cancer dataset.

Using this reference panel, Rachel extracted neo-antigenic mutations from matched WES data and classified genes harbouring neo-antigens into two groups: expressed neo-antigenic transcripts, where the mutant allele was present in at least 30 reads, and non-expressed locus, where no reads supported the mutant allele. She found 883 and 375 expressed and non-expressed neo-antigenic transcripts in samples with matched RRBS data, respectively. Of those 407 and 77 were unique, while others were duplicates from different sampled tumour regions of the same patient. Correlations between samples of the same patient could skew downstream analyses, we therefore only used on sample per patient, taking the sample with the largest variant allele frequency for each neo-antigenic loci. The higher the VAF, the more likely we are to detect methylation-driven silencing of neo-antigens, if present.

Expressed neo-antigens were slightly biased for higher RNA sequencing coverage and VAF. To even out the expression and mutational profiles between expressed and non-expressed neo-antigens, we downsampled the expressed neo-antigenic loci to match as closely as possible the gene expression and the variant allele frequency distributions observed for the non-expressed neo-antigens. We performed CAMDAC DMP calling from bulk and normal methylation rates at promoters (2kb up- and downstream of TSS). Hochberg family-wise error rate correction is then

applied and promoters are flagged as hypermethylated when 3 or more CpGs were significantly hypermethylated ($q < 0.05$). Promoter counts are tested in a 2x2 contingency table (methylation status vs expression status or mutation status) using a $\chi^2$-test. We re-ran this analysis with CAMDAC purified tumour methylation rates and saw mo difference in the outcome of the $\chi^2$-test. However, we saw greater tumour-normal methylation differences at hypermethylated neo-antigenic promoters post-deconvolution.

# Chapter 6

# Discussion

## 6.1 Summary

This work began with the proposition that bulk tumour methylation rates are affected by tumour copy number and purity. As such, these two confounders present a barrier to our understanding of the cancer methylome. Addressing this issue, we describe a robust framework for the analysis of bulk tumour RRBS data, unpolluted by normal signals.

First, we discuss ASCAT.m, our method for allele-specific copy number profiling directly from methylation sequencing data. Importantly, this tool obviates the need for additional SNP array or WGS data. To develop this approach, we use the epiTRACERx pilot cohort, comprised of 38 NSCLC patients selected from the TRACERx first 100 patients each with 2-7 tumour regions for which we collected RRBS data, totalling of 122 tumour and 37 adjacent normal lung samples. Previously published WES [112] of the same samples as well as newly obtained WGS data for a subset of samples were available for validation, making the epiTRACERx particularly well-suited for ASCAT.m method development.

Comparing RRBS- and WGS-derived genotypes from 3 NSCLC patients, we find that genotypes obtained directly from bisulphite sequencing data are reliable, except for polymorphic CCGGs. For multi-region sequencing studies, multi-sample segmentation greatly improves copy number profiles. Resulting segmented LogR and BAF profiles and allele-specific copy numbers were comparable between ASCAT(.m) performed on RRBS and gold standard WGS. Overall, purity

and ploidy values were in good agreement between RRBS data and WGS/WES of the same samples, including one sample without patient-matched normal.

Indeed, ASCAT.m performs well in one case without tumour-adjacent matched normal, which we substitute by a panel of sex-matched normals. Analysing ASCAT.m copy number profiles across the epiTRACERx cohort, we find commonalities and differences between LUAD and LUSC. Genome doubling is a frequent feature of NSCLC according to our and previously published results [123]. WGD was significantly associated with increased probability of relapse in LUSC but not LUAD likely due to increase prevalence of LOH in the former.

Subsequently, we introduce our method for copy number-aware methylation deconvolution analysis of cancer, CAMDAC. We first obtain SNP-independent bulk tumour and normal methylation rates for all samples in the epiTRACERx cohort. Next, we visualised DMP populations in the $m_b$ distribution by thresholding on $m_n$, assuming the adjacent normal cell composition is a suitable match for the normal contaminants and stratifying CpGs by ASCAT.m copy numbers and tumour purity. From this, we formalise the relationship between methylation rates and tumour copy number and purity as CAMDAC equations 2 and 3.

We model the position of the clonal bi-allelic DMP population as predicted by CAMDAC both pre- ($m_b$) and post-deconvolution ($m_t$) and find that both are in agreement with the observed. The error on the CAMDAC predictions is proportional to tumour DNA content, decreasing with increasing copy number and tumour purity. Although the errors are small, we note that the observed methylation rates is systematically shifted away from the expected and towards non-differentially methylated CpGs and suggest that this effect is caused by methylation erosion in fast replicating tumour cells.

Next, we demonstrate that by removing shared signal from normal contaminating cells, CAMDAC purified profiles have increased pairwise distances to normals and between samples from different patients, while samples from the same patients remain correlated, reflecting their shared ancestry. Leveraging SNV calls from WGS and WES of the same samples, we perform SNV deconvolution. SNV

deconvoluted methylation rates in regions of LOH where the SNV is on all copies should be an unbiased estimate of the pure tumour methylation rate. We report good agreement between SNV- and CAMDAC-purified methylation rates at these loci.

We develop a method for tumour-normal and tumour-tumour DMP calling from CAMDAC deconvoluted methylation rates and compared its performance with the bulk tumour on simulated and real data. DMP calls were obtained for simulated mono- (balanced regions only) and bi-allelic DMPs, revealing that only $m_t$ values enable accurate differential methylation analyses. Applying the effect size threshold to the bulk, as is customary, leads to false negatives. Indeed, false negative rates based on the bulk increased with decreasing copy number and purity, especially at mono-allelic DMPs. False positive rates at tumour-tumour DMPs derived from bulk were highest in low *versus* high purity sample pairs and simulated CpGs of differing copy number. A similar trend is observed from real data.

We perform RRBS of sorted diploid and aneuploid populations separated by FACS for 7 bulk tumour fresh frozen tissue samples taken from 5 different epiTRACERx patients. Diploid cells extracted from the bulk tumour are assumed to represent normal infiltrates while the higher ploidy population is purely tumour cells. We first use these data to validate the use of tumour-adjacent normal lung methylation rates as substitutes for that of the normal contaminating cells in CAMDAC equation 3. We perform cell-type deconvolution using EpiDISH on the normal infiltrates isolated by FACS and on the tumour-adjacent normals, comparing results between samples from the same patient. The bulk normal lung tissue samples are shown to be of similar cellular composition to the sorted normal contaminants, suggesting that it is a suitable proxy for bulk tumour deconvolution, at least in NSCLC.

We also investigate the suitability of the bulk normal lung samples as substitutes for the NSCLC cell of origin in differential methylation analysis. We find that the bulk normals contain a large percentage of epithelial cells, more so than the sorted normal infiltrates, suggesting that they are the best available proxy for the NSCLC cell of origin. Using the FACS-purified aneuploid tumour cell populations,

we also demonstrate that the tumour-normal DMP calls based on CAMDAC $m_t$ had greater overlap with the FACS sorted data than the bulk, further advocating use of CAMDAC over bulk methylation rates for accurate identification of epimutations.

We subsequently obtain DMR calls building on CAMDAC purified methylomes and DMPs. Roughly 12.5% of methylation bins covered by RRBS reads in all epiTRACERx samples were DMRs. We measure intra-tumour DMR ubiquity and observe a positive correlation with patient outcome. Importantly, results indicate that a minimum of 3 samples is required to adequately sample intra-tumour heterogeneity. Clustering of CAMDAC pure tumour methylation rates at promoter DMRs separated samples by sex, histology and patient while DMP clustering revealed intra-tumour subclonal relationships.

Investigating the most frequent epigenetic alterations across the epiTRACERx cohort, we note that recurrently hypomethylated promoters are few in numbers. This is likely because most promoter CGIs are not methylated to begin with. Nevertheless, recurrent loss of methylation is observed at 5 oncogenes, *H2B1*, *SFN*, *NELFCD*, *FAM83H-AS1* and *TUBA1C*, suggesting that demethylation may play a role in transcription regulation at those loci. In contrast, promoter hypermethylation was abundant, with over 600 gene promoters having gained methylation in $\geq$ 90% of patients in our NSCLC cohort. Focusing on known hypermethylated genes and their families, we usually find epimutations to occur at a higher rate than previously reported in the literature. We propose that use of CAMDAC purified tumour methylomes for DMR calling leads to the observed reduction in the number of false negatives. Hundreds of DMRs are ubiquitous within patients, suggesting they occurred early in tumour evolution, and across our cohort, demonstrating the potential of methylation data for early diagnosis of NSCLC. Interestingly, a handful of DMRs were histological subtype-specific: *MLPH*, *CDH3*, *SAFTA3*, *CHMP4BP1* and *HMHA1*.

Finally, we leverage CAMDAC pure tumour methylation rates and ASCAT.m allele-specific copy numbers to gain insight into copy- and allele-specific methylation in NSCLC. We find that allele (and copy)-specific methylation is a universal

feature of both normal lung tissue and NSCLCs. To being with, we illustrate the presence of allele-specific methylation in tumour-adjacent normal lung tissue by looking at X inactivation and incidentally uncover the presence of extraction biases against the Barr body. We also note that X inactivation is random, at the scale of our normal samples. Next, we use SNP-phased methylation rates to confirm the presence of allele-specific methylation at the *H19/IGF2* germline imprinting control region. Looking at this locus in tumour samples, we saw loss of imprinting in 5 patients, leading to demethylation of CpGs spanning the ICR in all instances.

Next, we assess allele- and copy-specific methylation in tumours, measuring the number of DMPs with epimutation copy numbers 1 and 2 in regions of $1+1$ and $2+0$. As proxy for the ratio of clonal to subclonal DMPs, we also compute the number of DMPs on 1 or less than 1 copy in regions of $1+0$. We then use this ratio as a proxy to remove subclonal contamination from allele- and copy-specific DMPs in regions of $1+1$ and $2+0$. Speculating on the origins of DMPs, we define two classes of DNA methylation changes, stochastic and regulatory alterations. We estimate their relative frequency based on the ratio of epimutations with copy numbers 1 and 2 in $1+1$ regions, assuming all DMPs with epimutation copy number 1 are stochastic and those with copy number 2 are regulatory. Our results show that regulatory DMPs tend to dominate the mutational landscape of our NSCLC samples, although we cannot exclude that this is in part due to RRBS data being skewed for CGIs.

Leveraging the estimated ratio of regulatory to stochastic DMPs computed from $1+1$ copy number segments, we attempt to extract the number of stochastic alterations on 2 copies in $2+0$. We take the estimated number of clonal stochastic DMPs with epimutation copy numbers 1 and 2 in regions of $2+0$ and attempt to estimate the timing of the copy number gain. Our method is currently limited to timing gains in regions of $2+0$ and in samples that also harbour sufficiently many DMPs in regions of $1+0$ and $1+1$, which are needed as proxy for subclonal and regulatory DMP decontamination, respectively. Nevertheless, timing estimates in regions of $2+0$ suggest that gains take place in the second half of epimutational

time, in line with published timing of LUAD samples derived from whole-genome sequencing data [124].

We investigate the relationship between somatic alterations and DNA methylation changes. We often find a causal relationship between the two, with SNV-phased DMRs commonly appearing in-*cis* with respect to the mutation. DMRs in-*cis* are usually hypermethylated, suggesting that SNVs lead to the deletion of TFBSs and consequently enable *DNMT* activity. In the same line of idea, we describe the relationship between methylation and gene expression at tumour neo-antigens across epiTRACERx. We find that non-expressed tumour neo-antigens were significantly enriched for hypermethylation compared with expressed neo-antigens and unmutated controls of the same genes. Our results implies that DNA methylation plays a role in immune escape, suppressing neo-antigen presentation. This analysis has been published as part of Rosenthal *et al.* [114].

Overall, this work sheds light on the wealth of genetic and epigenetic information contained in bisulphite sequencing data. We show that ASCAT.m copy numbers and CAMDAC deconvoluted methylomes are a first step towards gaining a deeper understanding of the cancer methylome.

## 6.2 Strengths of this work

For the first time, we present a simple model for bulk tumour data, formalising the relationship between tumour copy number, purity and methylation rates in bulk tumour bisulphite sequencing data. We formalise this relationship into CAMDAC equations 2 and 3. Importantly, use of CAMDAC pure tumour methylomes greatly reduces false negatives in differential methylation analyses while keeping false positives low. As a result, we see increased prevalence of recurrently epimutated gene promoters compared to levels reported in the literature. CAMDAC purified DNA methylation estimates could prove to be an even more powerful cancer biomarker than previously thought based on bulk data, which is subject to high false negative rates at low tumour purity and copy number. Moreover, our results suggest that intra-tumour heterogeneity analysis by clustering of methylation rates at DMPs

from bulk simply reflects tumour DNA content, and only $m_t$ reflects sample relationships. We speculate that CAMDAC $m_t$ can also provide deeper insights in tumour heterogeneity in other solid cancers.

CAMDAC relies on ASCAT.m, which enables copy number profiling directly from RRBS data, saving researchers time and money by eliminating the need for matched SNP array or WGS, at least for copy number profiling and genotyping purposes. Adding to its usefulness, ASCAT.m does not require a patient-matched normal. Indeed, we showed in one example case without tumour-adjacent patient-matched normal lung RRBS data that a panel of sex-matched normal is suitable to obtain LogR estimates. Although affecting only a small portion of the genome, it is worth noting that germline copy number variants may be mistaken as somatic copy number alterations with this approach. Heterozygous SNPs can be inferred directly from the tumour BAF profile if samples are of sufficiently low tumour purity, which is the case for most NSCLC samples.

We also show that CAMDAC deconvolution and differential methylation analysis can be run on tissue- and sex-matched normals. This is important as it is not always possible to obtain patient-matched normal tissue. In cases where the cell of origin is unknown, the tissue-matched normal may not be readily obtainable. CAMDAC provides provides $HDI^{99}$ estimates on its $m_t$ values, which can be used to identify tumour-tumour DMPs. While early clonal DMPs present in all samples will be missed by this approach, it can be used to identify methylation differences between for example histological subtypes, primary tumours and metastases, or groups of samples with different prognosis. One can obtain valuable information from ASCAT.m and CAMDAC simply with a cohort of tumour samples with RRBS data.

Furthermore, we show that RRBS data allows combined analysis of SNVs and CAMDAC DMPs, providing insights into the interplay between somatic genetic sequence alterations and DNA methylation changes. Most SNV-phased DMRs were hypermethylated and in-*cis* with respect to SNVs, suggesting a causal relationship between genetic mutations and epimutations, likely through the ablation of TFBSs,

enabling *DNMT* activity. In-*trans* DMRs also exist, reflecting two independent epigenetic and genetic events affecting different copies of the same locus. Such alterations were rare, and we speculate that they could indicate the presence of tumour suppressor genes where strong selection forces are at play.

Lastly, the TRACERx consortium is a powerful collaborative effort between research laboratories with different field of expertise and the resulting multi-omics dataset collected and analysed by each lab is immensely valuable. We were able to combine WES, RNA sequencing and RRBS data of the same samples to investigate the relationship between DNA methylation and immune escape. As part of published work [114], we uncover that silenced neo-antigens are enriched for hypermethylated CpGs compared with expressed neo-antigens and unmutated controls. Methylation is reversible by *DNMT* inhibitors and therefore we posit that administering *DNMT* inhibitors in combination with immune checkpoint blockade in patients with otherwise competent immune presentation and recognition pathways could greatly improve treatment response and patient survival.

## 6.3 Limitations

### 6.3.1 Comparing reduced-representation with whole-genome bisulphite sequencing methylation data

We discussed the advantages of RRBS data throughout this manuscript. Here, we point out some of the limitations. Briefly, while the NuGEN Ovation RRBS Methyl-Seq System protocol enables cost-effective analysis of DNA methylation in key regulatory regions such as CGIs, promoters and enhancers, it cannot probe CpGs outside *MspI* fragments. As a result, differentially methylated regions detected from CpG-dense RRBS reads should mainly involve regulatory alterations, while stochastic DNA methylation changes are likely fewer in number.

In this work, we suggest that, like SNVs, random stochastic epimutations contain timing information and attempt to extract these loci from RRBS data. In fact, WGBS (and Nanopore WGS) potentially covers a large number of isolated loci in low CpG density regions covered that are unlikely to be affected by aberrant gene

regulation and possibly only prone to stochastic alterations. Compared with RRBS, WGBS and Nanopore sequencing would probably allow for more accurate timing analysis, obviating the need for decontamination of potentially dynamic regulatory aberrations by focusing on loci in low CpG density areas of the genome. In this work, we find that RRBS is well-suited to identify differential methylation at regulatory loci, which are likely to play a role in tumourigenesis, but we speculate that stochastic DNA methylation are better represented in WGBS data, which could refine timing of copy number gains and shed light on tumour evolutionary trajectories.

We also note that RRBS data is not particularly well-suited to the investigation of methylation rates at repetitive elements due to the low mappability of bisulphite converted reads. This mapping bias is worsened for reads with alternate alleles or from short *MspI* fragments (i.e. shorter than 100bp). For example, RRBS is unlikely to provide coverage at repetitive endogenous retroviral elements. Recent publications showed that expression of endogenous retroviruses in cancer cells can stimulate an immune response resulting in apoptosis [218, 219]. The authors of both papers go on to demonstrate that transcription of viral RNA is silenced due to aberrant *DNMT* activity. These findings have significant clinical implications, with results showing that *DNMT* inhibitors, through endogenous retroviruses upregulation, can (re-)instigate cell death. Indeed, administration of a *DNMT* inhibitor in low dosage was shown to slow tumour growth in colorectal cancer [218]. These reports offer a possible explanation for the anti-tumour activity of *DNMT* inhibitors and highlight the importance of repetitive elements in tumourigenesis. Nanopore data yields long reads with high mappability and may be more adequate to probe methylation levels at repetitive elements than RRBS.

### 6.3.2 Challenges in timing copy number gains from methylation data

Our method relies on the infinite sites model, implying that no CpG may be mutated twice given the size of the human genome. Violations to the infinite sites model for genetic mutations are rare. The mutation rate for NSCLC is one of the highest, near that of melanoma, reportedly $\sim 1 \times 10^{-5}$ mutation per base pair [94].

Nevertheless, the likelihood of a given nucleotide being mutated twice, either on the same or the opposite allele or copy is low.

In contrast, the epimutation rate at CpGs is much higher, with roughly 1 in 10 CpGs identified as epimutated by CAMDAC, at least in RRBS data. It is therefore probable that random DNA methylation changes affecting a given copy or allele are subsequently acquired on a second, in two distinct steps. In such cases, our assumption that all loci on 2 copies in $1 + 1$ are of regulatory origin breaks down. DNA methylation is reversible, meaning that stochastic DMPs can be erased. These caveats will complicate the development of both timing and phylogenetic reconstruction tools from methylation data.

DNA methylation erosion is known to occur in ageing cells and rapidly dividing tumour cells, increasing noise at methylated CpGs in both normal contaminants and tumour cells. Erosion is likely to be transient at regulatory DMRs and unlikely to reverse methylation without corresponding changes in signalling. Methylated stochastic DMPs are most likely to be erased by this process, as they are probably passenger epimutations with no evolutionary and/or regulatory pressure to remain in the altered state.

NSCLC samples often harbour late copy number gains while other cancer types, such as glioblastoma, have well-defined early gains [126]. Testing whether our methylation-based copy number gain timing approach can not only detect late gains as implied in this work, but also known early gains could validate our hypothesis that methylation sequencing data has the potential to inform on tumour evolutionary trajectories.

In summary, we hypothesise that bisulphite sequencing data harbours timing information, but highlight several challenges in developping methods to robustly investigate tumour epigenetic evolutionary trajectories and the need for samples.

## 6.4 Future perspectives

### 6.4.1 The epiTRACERx study: next steps

Having now developed tools to analyse tumour RRBS data, unpolluted by signals from normal infiltrating cells, we plan to expand the epiTRACERx cohort. The new cohort will include longitudinal data (i.e. metastasis), providing insights into the metastatic seeding potential of clones with specific alterations and enabling comparison of the methylome of patient-matched primary tumour and metastasis samples.

We note that the ASCAT.m and CAMDAC methods described in this work can be directly applied to analysis of bulk tumour RRBS data from local lung metastases, using the patient matched normal for copy number profiling with ASCAT.m and for both deconvolution and tumour-normal differential methylation analyses. Samples taken from metastases can be combined with primary tumour regions to improve multi-sample segmentation and downstream copy number profiling by ASCAT.m. However, for lymph nodes or distant metastases, adjacent normal tissue will be required for used as proxy for the normal contaminants in CAMDAC.

### 6.4.2 CAMDAC beyond NSCLC

To increase the value and accessibility of CAMDAC to other cancer types and cohort designs, it would be useful to build an extensive normal reference panel with (1) a collection of normal RRBS samples to obviate the need for germline samples to get LogR estimates in ASCAT.m and (2) different bulk tissues and cell types so matched normal is not needed for CAMDAC deconvolution and differential methylation analysis. Read coverage data extracted from the normal lung samples presented in this work provide an adequate normal reference for ASCAT.m. Sequencing coverage is not controlled data and could be made publicly available. For deconvolution, a variety of cell types with RRBS data have already been compiled [41], and could be utilised in combination with reference profiles generated by our lab in future. It would be interesting to compare differential methylation analysis on the epiTRACERx cohort performed using patient-matched adjacent normal sam-

ples *versus* normal lung cell lines.

Knowing that Small Intestine Neuro-Endocrine Tumours (SINETs) harbour very few genetic aberrations, we plan to mine the SINET methylome to better understand the evolution of this disease. So far, we applied CAMDAC to a small pilot cohort of 7 SINETs with low coverage multi-region RRBS (range 2-5) of the primary tumour (see Chapter 6). We find that multi-region copy number segmentation is key to obtain good quality profiles with ASCAT.m at low coverage. We find that tumour-adjacent healthy intestinal tissue is an adequate proxy for the normal contaminating cells and thus for tumour-normal deconvolution with CAMDAC. However, neuro-endocrine cells only make up a small percentage of the small intestine niche and thus methylation data from an endocrine cell line will be needed for accurate tumour-normal differential methylation analysis. Hopefully, after procuring the adequate normal reference, this work will shed light on the origin of SINETs and offer potential therapeutic targets as well as biomarkers.

### 6.4.3 Combining normal cell-type and bulk tumour deconvolution

We show that, in lung, the adjacent normal tissue is a relatively good fit for the tumour infiltrating normal cells. This may however not be the case in all bulk tumour samples. In theory, we posit that it would be possible to adapt cell-type deconvolution algorithms, such as EpiDISH [96], to determine the cellular composition of normal infiltrates. In NSCLC, for example, given tumour purity and ploidy, the epithelial fraction would be split between tumour and normal, while all immune and fibroblast components would be attributed to normal infiltrates. With the normal composition in hand, we could use reference normal RRBS profiles of each cell types to create a proxy for the normal contaminating cells. This approach would be useful in cancer types where the adjacent normal cannot be obtained.

### 6.4.4 Investigating the interplay between genetic and epigenetic modalities

Our SNV-phased deconvolution of methylation rates revealed the causal relationship between DMRs in-*cis* to SNVs. Our analysis incidentally showed that the variant allele frequencies of mutations derived from RRBS, building on ASCAT.m BAF calculation rules, were comparable between RRBS and WGS/WES. While our analysis leverages a list of SNVs pre-curated from matched WES and WGS, it indicates that *de novo* SNV calls could be made directly from RRBS, in addition to allele-specific copy numbers, tumour purity, pure tumour methylation rates and differential methylation analysis.

We are yet to investigate the interplay between epigenetic markers, mainly DNA methylation, histone post-translational modifications and nucleosome positioning, which work together to modulate chromatin structure and transcription factor binding, thereby regulating gene expression. In future, combining CAMDAC with an analogous method to deconvolute bulk tumour RNA and chromatin immunoprecipitation sequencing data has the potential to provide unprecedented insight in the interplay of these modalities and their deregulation in cancer.

# Appendix A

## Supplementary Figures

**Figures S1-6: Comparing WGS- and RRBS-derived ASCAT(.m) BAF, LogR and copy number segments for a representative tumour sample.** Direct comparison of BAF (top) LogR (middle), and allele-specific copy number (bottom) profiles derived by ASCAT(.m) from matched RRBS and WGS. We show results for the 7 tumour samples for which matched RRBS and WGS data was available but they were not included as examples in the main body. The samples are CRUK0031-R3 (Figure S1), CRUK0062-R1 (Figure S2), CRUK0062-R3 (Figure S3), CRUK0062-R4 (Figure S4), CRUK0069-R3 (Figure S5), CRUK0069-R4 (Figure S6). BAF and LogR values are plotted at heterozygous SNPs only and this sample is male.

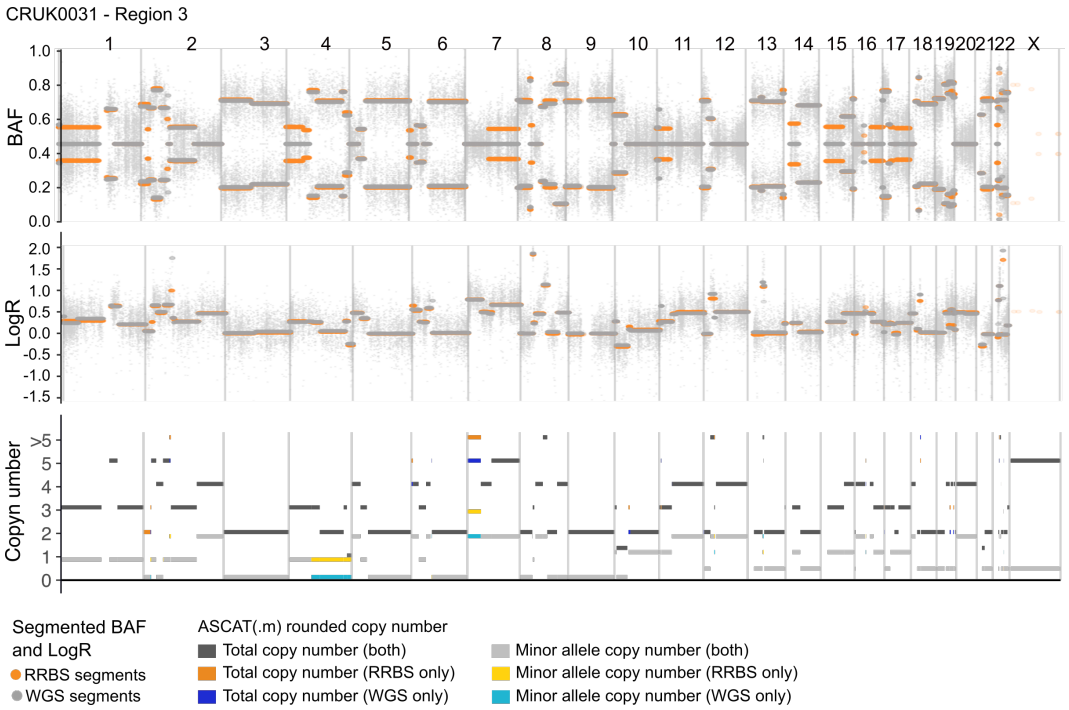


**Figure S1: CRUK0031-R3.**

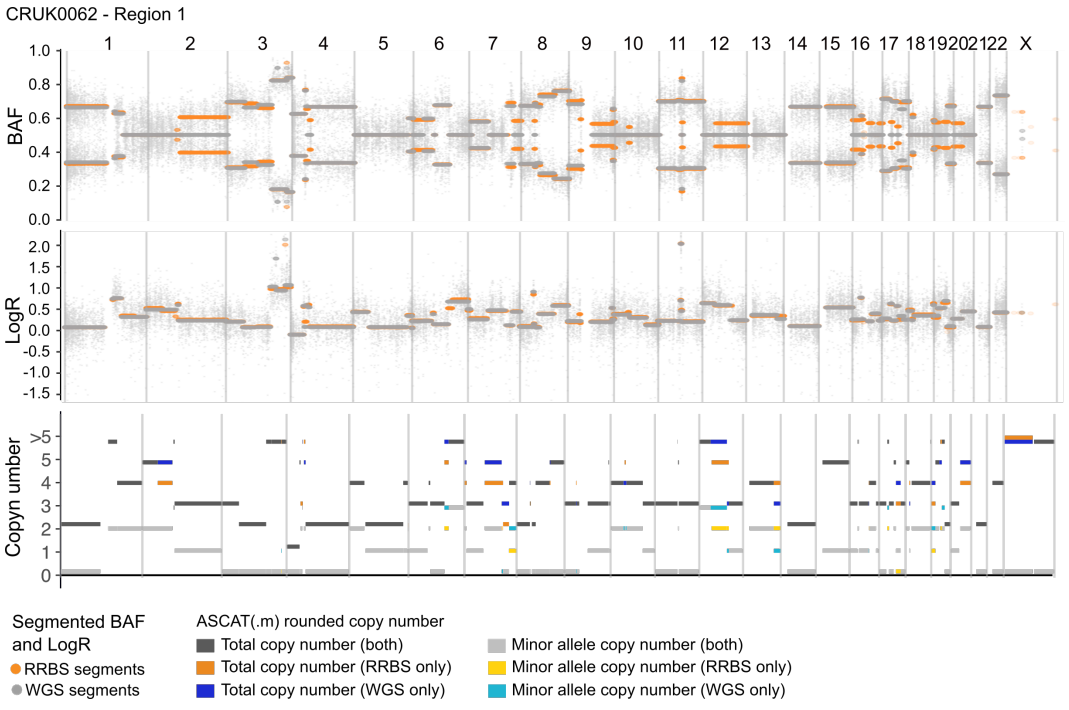


**Figure S2: CRUK0062-R1.**

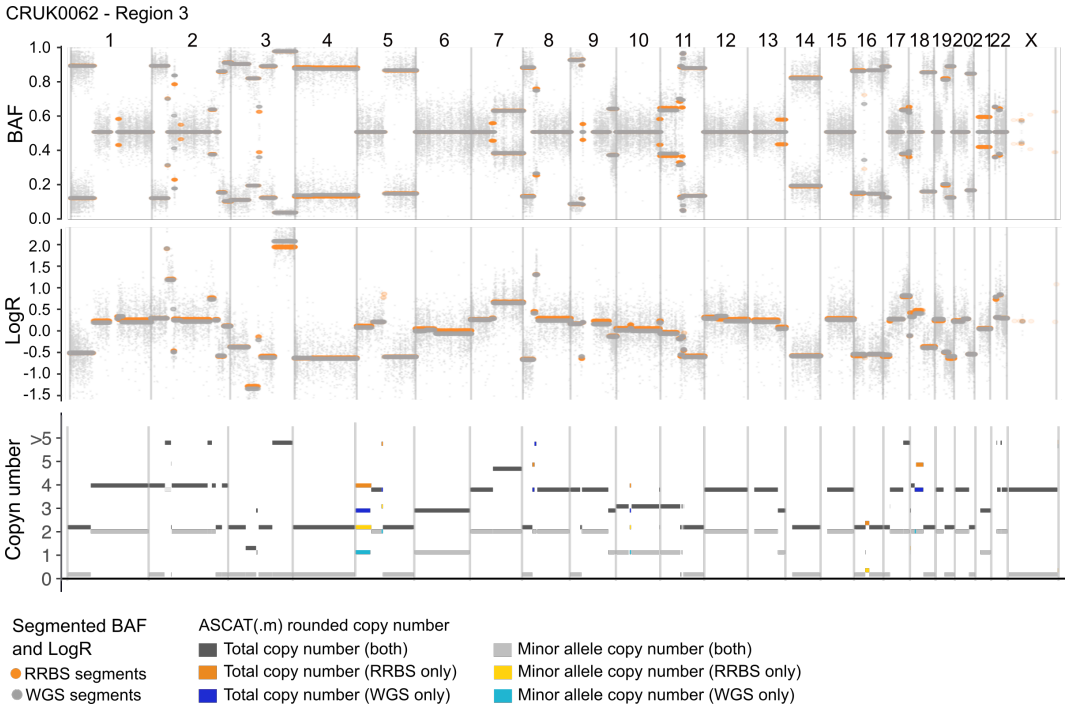


**Figure S3: CRUK0062-R3.**

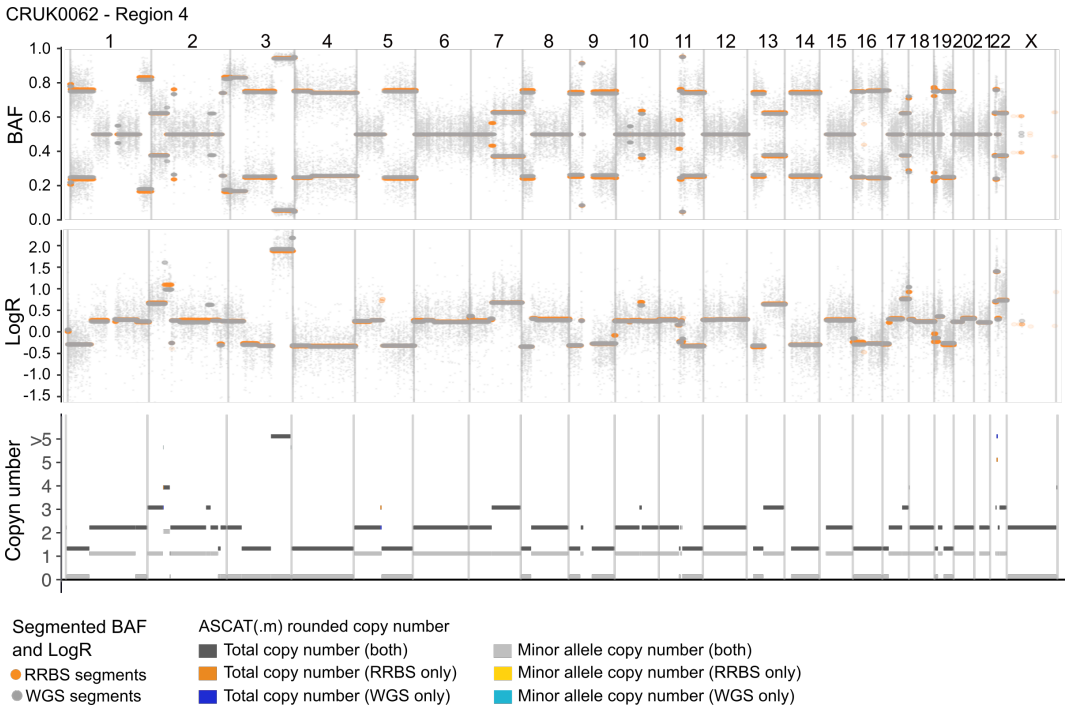


**Figure S4: CRUK0062-R4.**

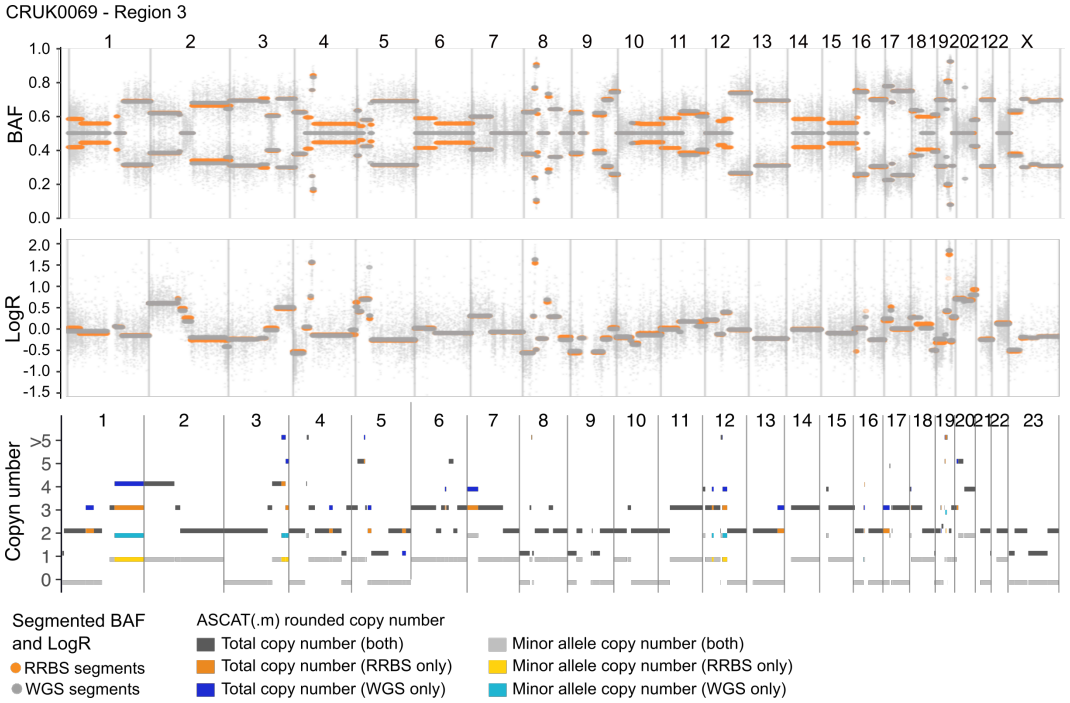


**Figure S5: CRUK0069-R3.**

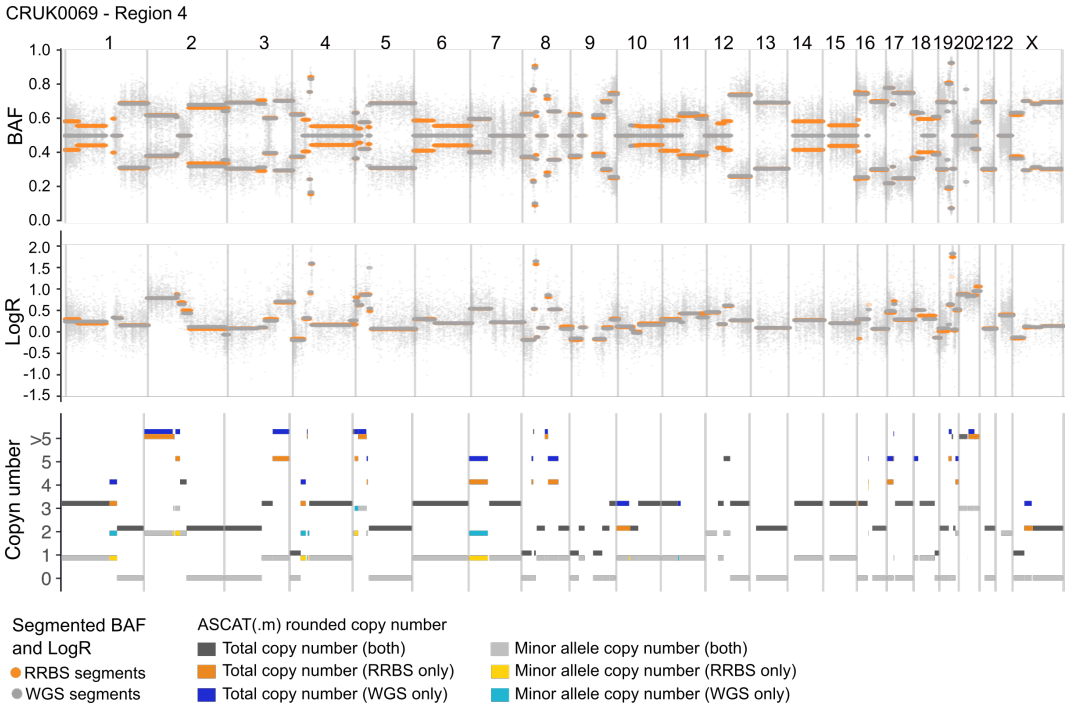


**Figure S6: CRUK0069-R4.**

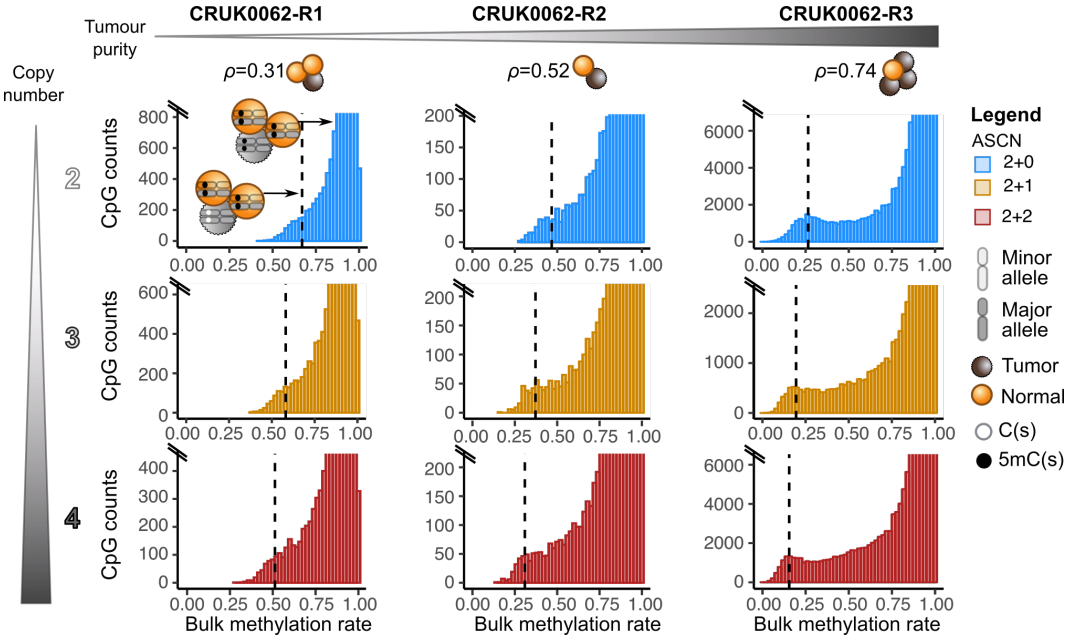


**Figure S7: Tumour purity and copy number affect methylation rates.**
Bulk methylation rate histograms for tumour regions 1-3 of patient CRUK0062, for CpGs which are confidently methylated in the adjacent normal sample. CpGs are stratified by copy number. A dashed line indicates the expected mode of the methylation rate peak corresponding to clonal differentially methylated CpGs on all copies ($m_t = 0$).

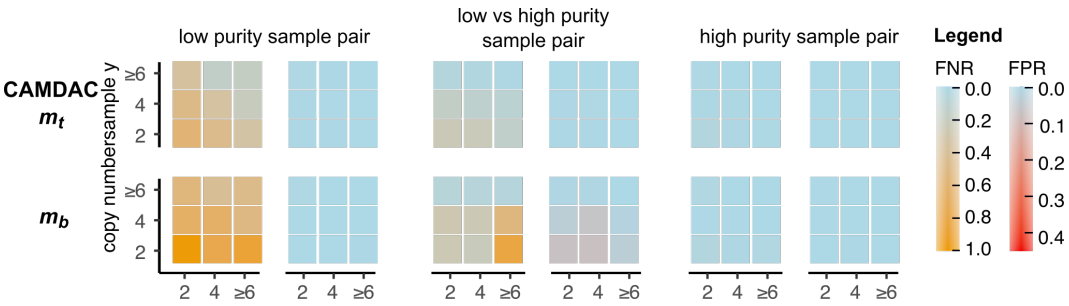


**Figure S8: Mono-allelic tumour-tumour DMP simulation.**
Results of mono-allelic tumour-tumour DMP simulations. False negative and false positives rates as a function of tumour copy number for low (left panel), low versus high (middle panel) and high purity (right panel) sample pairs at mono-allelic epimutations.

# Appendix B

# Supplementary Tables

**Table B.1: Cohort clinical information for the TRACERx methylation study.**

| Publication ID | Region IDs | Age group | Gender | Diagnosis | Stage | Smoking Status |
|---|---|---|---|---|---|---|
| CRUK0002 | R1,R2,R3 | 80-89 | Male | LUAD | 1b | Ex-Smoker |
| CRUK0003 | R1,R2,R3,R4,R6 | 70-79 | Female | LUAD | 3a | Never Smoked |
| CRUK0008 | R1,R2 | 70-79 | Male | LUAD | 1a | Ex-Smoker |
| CRUK0010 | R2 | 60-69 | Male | LUAD | 3a | Never Smoked |
| CRUK0012 | R1,R2 | 60-69 | Male | LUAD | 1a | Ex-Smoker |
| CRUK0013 | R1,R2,R3 | 60-69 | Male | LUAD | 3a | Ex-Smoker |
| CRUK0014 | R1,R2 | 60-69 | Male | LUAD | 1a | Never Smoked |
| CRUK0021 | R1,R2 | 80-89 | Female | LUAD | 1a | Never Smoked |
| CRUK0023 | R1,R2,R3,R4 | 60-69 | Male | LUAD | 2b | Ex-Smoker |
| CRUK0025 | R1,R2,R3 | 50-59 | Male | LUAD | 1b | Recent Ex-Smoker |
| CRUK0029 | R4,R5,R6,R8 | 50-59 | Male | LUAD | 3a | Ex-Smoker |
| CRUK0031 | R1,R2,R3 | 50-59 | Male | LUAD | 1b | Recent Ex-Smoker |
| CRUK0033 | R1,R2 | 60-69 | Male | LUAD | 1a | Never Smoked |
| CRUK0034 | R1,R2,R3 | 60-69 | Female | LUAD | 1b | Ex-Smoker |
| CRUK0036 | R1,R2,R3,R4 | 60-69 | Female | LUAD | 1b | Recent Ex-Smoker |
| CRUK0037 | R1,R2,R3,R5 | 80-89 | Male | LUAD | 2b | Ex-Smoker |
| CRUK0046 | R1,R2,R3,R4 | 60-69 | Female | LUAD | 2a | Ex-Smoker |
| CRUK0047 | R2 | 70-79 | Female | LUAD | 1a | Ex-Smoker |
| CRUK0048 | R1,R2,R3 | 70-79 | Female | LUAD | 1b | Recent Ex-Smoker |
| CRUK0049 | R1,R2 | 60-69 | Female | LUAD | 1a | Recent Ex-Smoker |
| CRUK0050 | R1,R2,R3,R4,R5 | 50-59 | Male | LUAD | 3a | Current Smoker |
| CRUK0053 | R1,R2 | 60-69 | Male | LUAD | 1b | Ex-Smoker |
| CRUK0057 | R1,R2 | 60-69 | Female | LUAD | 1b | Recent Ex-Smoker |
| CRUK0058 | R1,R2 | 70-79 | Male | LUAD | 2b | Ex-Smoker |
| CRUK0062 | R1,R2,R3,R4,R5,R6,R7 | 50-59 | Male | LUSC | 2b | Recent Ex-Smoker |
| CRUK0065 | R1,R2,R3,R4,R5,R6 | 70-79 | Male | LUSC | 2b | Ex-Smoker |
| CRUK0067 | R1,R3 | 60-69 | Male | LUSC | 1a | Recent Ex-Smoker |
| CRUK0069 | R1,R2,R3,R4,R5 | 70-79 | Female | LUSC | 1b | Ex-Smoker |
| CRUK0070 | R2,R4,R6,R7 | 50-59 | Male | LUSC | 2a | Recent Ex-Smoker |
| CRUK0071 | R1,R2,R3,R6,R7 | 60-69 | Male | LUSC | 2a | Ex-Smoker |
| CRUK0072 | R1,R2,R4 | 60-69 | Male | LUSC | 1b | Ex-Smoker |
| CRUK0073 | R1,R2 | 70-79 | Female | LUSC | 1a | Ex-Smoker |
| CRUK0076 | R1,R2,R3,R4 | 60-69 | Male | LUSC | 1b | Recent Ex-Smoker |
| CRUK0079 | R1,R2,R3,R4 | 60-69 | Female | LUSC | 2a | Recent Ex-Smoker |
| CRUK0082 | R1,R2,R3,R4 | 60-69 | Female | LUSC | 3a | Recent Ex-Smoker |
| CRUK0083 | R2,R3,R4 | 70-79 | Male | LUSC | 2b | Ex-Smoker |
| CRUK0084 | R1,R2,R3 | 60-69 | Female | LUSC | 1b | Ex-Smoker |
| CRUK0090 | R1,R2 | 60-69 | Male | LUSC | 2a | Ex-Smoker |

**Table B.2: Reduced representation bisulphite sequencing experiment statistics.**

| Publication ID | Region IDs | Sequencing strategy | Number of reads | | | | |
|---|---|---|---|---|---|---|---|
| | | | pre-alignment | post-alignment | duplicated | unique | first mate only |
| CRUK0002 | N | Single-End | 176,577,580 | 118,293,212 | 32,510,856 | 85,782,356 | |
| CRUK0002 | R1 | Single-End | 109,780,858 | 72,681,231 | 12,385,656 | 60,295,575 | |
| CRUK0002 | R2 | Single-End | 193,374,886 | 125,803,430 | 29,118,753 | 96,684,677 | |
| CRUK0002 | R3 | Single-End | 187,715,188 | 130,136,542 | 24,230,929 | 105,905,613 | |
| CRUK0003 | N | Single-End | 160,473,093 | 108,733,935 | 8,599,836 | 100,134,099 | |
| CRUK0003 | R1 | Single-End | 200,334,726 | 135,600,635 | 8,448,022 | 127,152,613 | |
| CRUK0003 | R2 | Single-End | 310,128,838 | 209,781,946 | 12,292,698 | 197,489,248 | |
| CRUK0003 | R3 | Single-End | 217,245,738 | 144,526,174 | 8,878,308 | 135,647,866 | |
| CRUK0003 | R4 | Single-End | 203,844,235 | 139,022,618 | 9,700,560 | 129,322,058 | |
| CRUK0003 | R6 | Single-End | 165,730,526 | 114,446,343 | 7,666,303 | 106,780,040 | |
| CRUK0008 | N | Single-End | 225,831,238 | 150,658,818 | 13,512,196 | 137,146,622 | |
| CRUK0008 | R1 | Single-End | 351,041,032 | 231,256,287 | 16,967,399 | 214,288,888 | |
| CRUK0008 | R2 | Single-End | 245,593,118 | 163,941,546 | 10,211,390 | 153,730,156 | |
| CRUK0010 | N | Single-End | 192,819,487 | 130,866,815 | 33,095,290 | 97,771,525 | |
| CRUK0010 | R2 | Single-End | 177,333,787 | 117,565,424 | 41,977,384 | 75,588,040 | |
| CRUK0012 | N | Single-End | 257,363,758 | 168,123,698 | 58,873,685 | 109,250,013 | |
| CRUK0012 | R1 | Single-End | 138,437,889 | 90,844,187 | 33,140,371 | 57,703,816 | |
| CRUK0012 | R2 | Single-End | 111,784,902 | 73,251,187 | 28,048,382 | 45,202,805 | |
| CRUK0013 | N | Single-End | 141,737,697 | 90,657,162 | 52,718,241 | 37,938,921 | |
| CRUK0013 | R1 | Single-End | 174,260,865 | 115,347,510 | 26,830,800 | 88,516,710 | |
| CRUK0013 | R2 | Single-End | 145,760,227 | 96,451,089 | 14,385,105 | 82,065,984 | |
| CRUK0013 | R3 | Single-End | 170,067,074 | 111,015,955 | 22,924,015 | 88,091,940 | |
| CRUK0014 | N | Single-End | 143,562,839 | 94,518,013 | 29,965,967 | 64,552,046 | |
| CRUK0014 | R1 | Single-End | 157,023,226 | 104,907,557 | 45,150,892 | 59,756,665 | |
| CRUK0014 | R2 | Single-End | 88,937,506 | 59,369,190 | 23,225,005 | 36,144,185 | |
| CRUK0021 | N | Single-End | 259,152,036 | 170,980,696 | 13,771,917 | 157,208,779 | |
| CRUK0021 | R1 | Single-End | 239,708,830 | 159,256,899 | 10,505,082 | 148,751,817 | |
| CRUK0021 | R2 | Single-End | 244,113,091 | 164,207,547 | 10,811,883 | 153,395,664 | |
| CRUK0023 | N | Single-End | 189,845,885 | 131,686,853 | 16,764,595 | 114,922,258 | |
| CRUK0023 | R1 | Single-End | 223,695,274 | 153,450,622 | 12,538,940 | 140,911,682 | |
| CRUK0023 | R2 | Single-End | 121,951,818 | 84,639,798 | 11,344,714 | 73,295,084 | |
| CRUK0023 | R3 | Single-End | 191,421,238 | 129,896,640 | 10,232,953 | 119,663,687 | |
| CRUK0023 | R4 | Single-End | 220,881,732 | 146,794,041 | 8,628,254 | 138,165,787 | |
| CRUK0025 | N | Single-End | 316,379,753 | 203,776,340 | 79,039,614 | 124,736,726 | |
| CRUK0025 | R1 | Single-End | 196,148,346 | 125,412,731 | 28,357,239 | 97,055,492 | |
| CRUK0025 | R2 | Single-End | 133,249,133 | 87,735,212 | 30,669,699 | 57,065,513 | |
| CRUK0025 | R3 | Single-End | 193,583,657 | 126,774,857 | 26,115,835 | 100,659,022 | |
| CRUK0029 | N | Single-End | 228,332,002 | 146,820,145 | 9,599,976 | 137,220,169 | |
| CRUK0029 | R4 | Single-End | 186,393,652 | 125,064,146 | 6,128,547 | 118,935,599 | |
| CRUK0029 | R5 | Single-End | 197,670,506 | 134,538,849 | 7,333,553 | 127,205,296 | |
| CRUK0029 | R6 | Single-End | 129,733,337 | 87,515,820 | 2,969,394 | 84,546,426 | |
| CRUK0029 | R8 | Single-End | 142,505,442 | 94,901,517 | 3,629,347 | 91,272,170 | |
| CRUK0031 | N | Single-End | 150,705,771 | 102,682,345 | 5,549,114 | 97,133,231 | |
| CRUK0031 | R1 | Single-End | 229,952,348 | 150,893,313 | 7,305,813 | 143,587,500 | |
| CRUK0031 | R2 | Single-End | 163,541,715 | 109,035,678 | 3,996,871 | 105,038,807 | |
| CRUK0031 | R3 | Single-End | 267,701,825 | 185,717,904 | 10,507,904 | 175,210,000 | |
| CRUK0033 | N | Single-End | 134,119,007 | 88,615,119 | 41,563,766 | 47,051,353 | |
| CRUK0033 | R1 | Single-End | 148,509,494 | 99,092,456 | 27,240,292 | 71,852,164 | |
| CRUK0033 | R2 | Single-End | 197,893,842 | 132,661,250 | 42,020,038 | 90,641,212 | |
| CRUK0034 | N | Single-End | 134,752,760 | 91,306,117 | 3,124,934 | 88,181,183 | |
| CRUK0034 | R1 | Single-End | 111,641,432 | 77,116,402 | 3,265,592 | 73,850,810 | |
| CRUK0034 | R2 | Single-End | 127,878,414 | 88,022,832 | 2,494,256 | 85,528,576 | |
| CRUK0034 | R3 | Single-End | 169,027,791 | 115,243,277 | 4,403,058 | 110,840,219 | |
| CRUK0036 | N | Single-End | 122,727,492 | 74,669,574 | 12,483,692 | 62,185,882 | |
| CRUK0036 | R1 | Single-End | 140,786,079 | 95,284,773 | 32,450,832 | 62,833,941 | |
| CRUK0036 | R2 | Single-End | 208,496,792 | 130,223,874 | 26,734,423 | 103,489,451 | |
| CRUK0036 | R3 | Single-End | 121,973,872 | 82,881,569 | 11,885,771 | 70,995,798 | |
| CRUK0036 | R4 | Single-End | 198,610,895 | 132,284,379 | 31,799,765 | 100,484,614 | |
| CRUK0037 | N | Single-End | 96,569,632 | 65,127,608 | 14,984,730 | 50,142,878 | |
| CRUK0037 | R1 | Single-End | 134,467,907 | 92,404,439 | 18,971,383 | 73,433,056 | |
| CRUK0037 | R2 | Single-End | 145,587,783 | 89,796,131 | 21,099,936 | 68,696,195 | |
| CRUK0037 | R3 | Single-End | 155,760,117 | 99,202,061 | 16,553,245 | 82,648,816 | |
| CRUK0037 | R5 | Single-End | 164,886,388 | 113,050,996 | 22,133,460 | 90,917,536 | |
| CRUK0046 | N | Single-End | 99,313,402 | 67,614,858 | 8,490,388 | 59,124,470 | |
| CRUK0046 | R1 | Single-End | 122,394,806 | 82,688,578 | 7,377,427 | 75,311,151 | |
| CRUK0046 | R2 | Single-End | 148,515,781 | 101,509,093 | 13,749,080 | 87,760,013 | |
| CRUK0046 | R3 | Single-End | 136,471,123 | 93,594,253 | 9,365,580 | 84,228,673 | |
| CRUK0046 | R4 | Single-End | 160,510,970 | 109,089,471 | 11,796,092 | 97,293,379 | |

| Publication ID | Region IDs | Sequencing strategy | Number of reads | | | | |
|---|---|---|---|---|---|---|---|
| | | | pre-alignment | post-alignment | duplicated | unique | first mate only |
| CRUK0047 | R2 | Single-End | 117,741,255 | 80,012,445 | 16,458,310 | 63,554,135 | |
| CRUK0048 | N | Single-End | 105,661,152 | 69,292,850 | 12,506,126 | 56,786,724 | |
| CRUK0048 | R1 | Single-End | 205,194,953 | 129,045,020 | 29,926,683 | 99,118,337 | |
| CRUK0048 | R2 | Single-End | 111,854,131 | 72,608,853 | 13,001,310 | 59,607,543 | |
| CRUK0048 | R3 | Single-End | 237,534,761 | 159,600,549 | 39,436,988 | 120,163,561 | |
| CRUK0049 | N | Single-End | 169,817,224 | 111,010,675 | 48,624,188 | 62,386,487 | |
| CRUK0049 | R1 | Single-End | 92,626,321 | 64,693,260 | 6,189,508 | 58,503,752 | |
| CRUK0049 | R2 | Single-End | 79,492,970 | 54,702,441 | 4,749,321 | 49,953,120 | |
| CRUK0050 | N | Single-End | 153,897,300 | 106,299,535 | 7,652,691 | 98,646,844 | |
| CRUK0050 | R1 | Single-End | 214,681,573 | 149,811,170 | 11,012,964 | 138,798,206 | |
| CRUK0050 | R2 | Single-End | 132,925,239 | 91,903,081 | 5,160,116 | 86,742,965 | |
| CRUK0050 | R3 | Single-End | 159,614,536 | 109,177,226 | 6,664,450 | 102,512,776 | |
| CRUK0050 | R4 | Single-End | 121,863,483 | 86,079,159 | 7,954,686 | 78,124,473 | |
| CRUK0050 | R5 | Single-End | 210,626,040 | 145,025,505 | 16,812,769 | 128,212,736 | |
| CRUK0053 | N | Single-End | 156,483,904 | 105,654,211 | 7,793,328 | 97,860,883 | |
| CRUK0053 | R1 | Single-End | 223,988,888 | 153,046,780 | 8,316,822 | 144,729,958 | |
| CRUK0053 | R2 | Single-End | 160,224,975 | 110,565,270 | 6,084,544 | 104,480,726 | |
| CRUK0057 | N | Single-End | 172,406,307 | 117,509,826 | 9,815,269 | 107,694,557 | |
| CRUK0057 | R1 | Single-End | 194,314,384 | 135,623,834 | 8,720,607 | 126,903,227 | |
| CRUK0057 | R2 | Single-End | 247,608,714 | 171,005,391 | 11,872,436 | 159,132,955 | |
| CRUK0058 | N | Single-End | 200,733,687 | 117,631,593 | 41,832,873 | 75,798,720 | |
| CRUK0058 | R1 | Single-End | 101,922,040 | 68,656,066 | 9,455,847 | 59,200,219 | |
| CRUK0058 | R2 | Single-End | 125,484,040 | 86,263,751 | 16,529,467 | 69,734,284 | |
| CRUK0062 | N | Paired-End | 152,333,066 | 98,254,494 | 2,526,892 | 95,727,602 | 46,600,355 |
| CRUK0062 | R1 | Paired-End | 256,504,028 | 166,734,428 | 4,233,859 | 162,500,569 | 79,133,355 |
| CRUK0062 | R2 | Paired-End | 231,783,292 | 154,534,238 | 3,099,023 | 151,435,215 | 74,168,096 |
| CRUK0062 | R3 | Paired-End | 220,294,066 | 145,162,762 | 3,184,253 | 141,978,509 | 69,397,128 |
| CRUK0062 | R4 | Paired-End | 175,406,104 | 117,793,172 | 2,629,235 | 115,163,937 | 56,267,351 |
| CRUK0062 | R5 | Paired-End | 174,208,958 | 116,200,478 | 2,908,975 | 113,291,503 | 55,191,264 |
| CRUK0062 | R6 | Paired-End | 162,725,478 | 108,272,938 | 2,380,023 | 105,892,915 | 51,756,446 |
| CRUK0062 | R7 | Paired-End | 138,575,046 | 93,359,494 | 1,788,742 | 91,570,752 | 44,891,005 |
| CRUK0065 | N | Single-End | 232,906,183 | 156,803,508 | 14,037,230 | 142,766,278 | |
| CRUK0065 | R1 | Single-End | 214,524,438 | 145,263,862 | 10,811,974 | 134,451,888 | |
| CRUK0065 | R2 | Single-End | 280,314,392 | 188,035,758 | 16,120,426 | 171,915,332 | |
| CRUK0065 | R3 | Single-End | 152,548,102 | 98,985,662 | 6,491,985 | 92,493,677 | |
| CRUK0065 | R4 | Single-End | 146,745,671 | 100,847,498 | 8,009,293 | 92,838,205 | |
| CRUK0065 | R5 | Single-End | 272,179,018 | 179,814,561 | 12,692,904 | 167,121,657 | |
| CRUK0065 | R6 | Single-End | 145,256,217 | 99,628,402 | 6,825,782 | 92,802,620 | |
| CRUK0067 | N | Single-End | 272,802,297 | 186,029,212 | 10,571,322 | 175,457,890 | |
| CRUK0067 | R1 | Single-End | 296,787,548 | 200,950,994 | 20,576,429 | 180,374,565 | |
| CRUK0067 | R3 | Single-End | 218,523,836 | 148,855,611 | 12,560,009 | 136,295,602 | |
| CRUK0069 | N | Single-End | 130,876,889 | 90,869,087 | 6,812,915 | 84,056,172 | |
| CRUK0069 | R1 | Single-End | 161,318,796 | 115,629,840 | 7,668,007 | 107,961,833 | |
| CRUK0069 | R2 | Single-End | 228,915,136 | 156,846,230 | 10,315,263 | 146,530,967 | |
| CRUK0069 | R3 | Single-End | 344,081,324 | 241,298,905 | 17,536,408 | 223,762,497 | |
| CRUK0069 | R4 | Single-End | 229,496,537 | 163,062,306 | 9,137,095 | 153,925,211 | |
| CRUK0069 | R5 | Single-End | 168,660,408 | 116,034,662 | 7,989,767 | 108,044,895 | |
| CRUK0070 | N | Single-End | 187,675,744 | 130,320,639 | 11,425,957 | 118,894,682 | |
| CRUK0070 | R2 | Single-End | 161,053,122 | 108,991,413 | 5,274,874 | 103,716,539 | |
| CRUK0070 | R4 | Single-End | 202,225,100 | 126,820,203 | 15,401,020 | 111,419,183 | |
| CRUK0070 | R6 | Single-End | 208,425,163 | 141,639,227 | 10,240,961 | 131,398,266 | |
| CRUK0070 | R7 | Single-End | 46,509,567 | 31,834,584 | 1,642,240 | 30,192,344 | |
| CRUK0071 | N | Single-End | 368,748,102 | 251,429,251 | 22,637,538 | 228,791,713 | |
| CRUK0071 | R1 | Single-End | 187,799,234 | 125,506,331 | 9,970,498 | 115,535,833 | |
| CRUK0071 | R2 | Single-End | 337,178,117 | 229,262,086 | 21,594,182 | 207,667,904 | |
| CRUK0071 | R3 | Single-End | 203,958,067 | 138,962,176 | 7,675,737 | 131,286,439 | |
| CRUK0071 | R6 | Single-End | 179,970,484 | 123,067,101 | 9,989,923 | 113,077,178 | |
| CRUK0071 | R7 | Single-End | 161,296,724 | 113,926,280 | 8,423,611 | 105,502,669 | |
| CRUK0072 | N | Single-End | 229,103,940 | 151,482,236 | 8,206,440 | 143,275,796 | |
| CRUK0072 | R1 | Single-End | 246,881,836 | 166,081,071 | 14,798,115 | 151,282,956 | |
| CRUK0072 | R2 | Single-End | 253,822,733 | 167,270,558 | 12,801,062 | 154,469,496 | |
| CRUK0072 | R4 | Single-End | 241,197,792 | 161,386,223 | 10,056,498 | 151,329,725 | |

| Publication ID | Region IDs | Sequencing strategy | Number of reads | | | | |
|---|---|---|---|---|---|---|---|
| | | | pre-alignment | post-alignment | duplicated | unique | first mate only |
| CRUK0073 | N | Single-End | 191,064,007 | 132,610,014 | 15,254,641 | 117,355,373 | |
| CRUK0073 | R1 | Single-End | 220,753,580 | 152,995,691 | 12,552,849 | 140,442,842 | |
| CRUK0073 | R2 | Single-End | 180,471,575 | 125,045,179 | 12,802,061 | 112,243,118 | |
| CRUK0076 | N | Single-End | 245,069,612 | 162,515,246 | 9,940,017 | 152,575,229 | |
| CRUK0076 | R1 | Single-End | 177,400,379 | 123,907,729 | 6,183,967 | 117,723,762 | |
| CRUK0076 | R2 | Single-End | 212,061,856 | 151,153,527 | 8,953,063 | 142,200,464 | |
| CRUK0076 | R3 | Single-End | 270,645,020 | 185,800,099 | 9,901,404 | 175,898,695 | |
| CRUK0076 | R4 | Single-End | 312,695,624 | 210,144,508 | 11,947,501 | 198,197,007 | |
| CRUK0079 | N | Single-End | 211,013,183 | 144,791,682 | 9,459,070 | 135,332,612 | |
| CRUK0079 | R1 | Single-End | 264,775,778 | 179,371,207 | 11,553,046 | 167,818,161 | |
| CRUK0079 | R2 | Single-End | 311,163,392 | 212,087,133 | 15,161,716 | 196,925,417 | |
| CRUK0079 | R3 | Single-End | 164,893,649 | 113,108,705 | 6,593,671 | 106,515,034 | |
| CRUK0079 | R4 | Single-End | 210,247,585 | 143,565,398 | 6,565,688 | 136,999,710 | |
| CRUK0082 | N | Single-End | 181,865,369 | 119,474,117 | 15,882,569 | 103,591,548 | |
| CRUK0082 | R1 | Single-End | 184,541,943 | 127,914,804 | 8,990,451 | 118,924,353 | |
| CRUK0082 | R2 | Single-End | 325,893,467 | 220,294,206 | 16,984,725 | 203,309,481 | |
| CRUK0082 | R3 | Single-End | 247,704,764 | 168,251,377 | 11,369,164 | 156,882,213 | |
| CRUK0082 | R4 | Single-End | 214,372,257 | 146,195,961 | 9,650,198 | 136,545,763 | |
| CRUK0083 | N | Single-End | 170,475,446 | 113,710,069 | 36,699,024 | 77,011,045 | |
| CRUK0083 | R2 | Single-End | 119,129,680 | 80,639,195 | 14,607,043 | 66,032,152 | |
| CRUK0083 | R3 | Single-End | 119,782,004 | 79,040,704 | 20,899,790 | 58,140,914 | |
| CRUK0083 | R4 | Single-End | 146,709,576 | 99,124,226 | 33,500,031 | 65,624,195 | |
| CRUK0084 | N | Single-End | 178,454,733 | 113,640,247 | 18,520,306 | 95,119,941 | |
| CRUK0084 | R1 | Single-End | 159,224,808 | 103,948,419 | 14,106,917 | 89,841,502 | |
| CRUK0084 | R2 | Single-End | 149,184,870 | 99,969,238 | 18,019,522 | 81,949,716 | |
| CRUK0084 | R3 | Single-End | 142,176,976 | 92,272,775 | 21,221,270 | 71,051,505 | |
| CRUK0090 | N | Single-End | 282,352,055 | 191,099,891 | 11,248,006 | 179,851,885 | |
| CRUK0090 | R1 | Single-End | 172,996,407 | 119,044,613 | 6,685,291 | 112,359,322 | |
| CRUK0090 | R2 | Single-End | 162,405,212 | 113,581,149 | 5,905,199 | 107,675,950 | |

# Appendix C

## Running CAMDAC

The CAMDAC package which can be downloaded from github (https://github.com/elarosecadieux/CAMDAC). ASCAT.m in integrated within CAMDAC. After cloning the repository to your location of choice and running the install.R script, you simply need to set the path_to_CAMDAC variable to your install path in all CAMDAC function calls. This path should point inside the CAMDAC folder. At time of writing, CAMDAC is only compatible with human (directional) RRBS data, but we plan to extend the algorithm to support further platforms, namely WGBS. The input must be quality and adapter trimmed with PCR duplicates removed. Binary alignment map files aligned to hg19, hg38, GRCH37 and GRHCH38 genome builds are accepted formats.

Single nucleotide allele counts are obtained for all CpG and 1000g SNP positions [133] that overlap with RRBS data. To speed up the computation, we built a reference RRBS genome listing all genomic regions supported by $\geq$5 reads in $\geq$1 of the 37 epiTRACERx normals. The get_allele_counts() function will only query 1000g SNP and CpG positions which fall within the above-mentioned genomic regions. This step greatly reduces the number of loci to investigate since RRBS data only covers about 2% of the human genome [31]. Reference loci files are divided into 25 smaller files to reduce memory requirements. The get_allele_counts() function is run within a loop whereby the $i^{th}$ iteration of the function pulls the corresponding $i^{th}$ reference file.

As ASCAT.m can be applied to multi-region data, both (anonymised) patient and sample identification string variables are required, patient_id and sample_id.

The sex follows the ASCAT format, ”XX” for females and ”XY” for males. This flag is important for copy number estimates on chromosome X as well as global tumour purity and ploidy values. Users are ask to provide the full BAM file path and name. By default, no mapping quality threshold is set to avoid creating a bias against the alternate allele at SNPs. Instead, reads are discarded when they do not align with *MspI* CCGG recognition sites. With allele counts in hand, we can calculate the BAF and LogR.

If you are unsure of the reference genome build of your BAM file, but know it is either hg19, hg38, GRCH37 or GRHCH38, set build=NULL and let CAMDAC determine the build version for you. We recommend that users create a directory to store CAMDAC outputs which we will call parent_dir. When submitting get_allele_counts() via the CAMDAC wrapper, the path is set automatically to the below. Users running CAMDAC outside the wrapper should set this same path. CAMDAC efficiency is improved by setting the number of cores to either 8 or 12. This may differ depending on your local machine or remote server compute power. When running get_allele_counts() for the first time, it is recommended that users set the test flag to TRUE.

```
# Run function iterating over each reference file
for(a in 1:25){
    get_allele_counts(i=a,patient_id, sample_id, sex,
   bam_file, mq=0, path_to_CAMDAC, build=NULL, path, n_cores,
   test=FALSE)
}
```

Next, we concatenate the 25 allele counts ’.fst’ files and convert these data to GRanges objects from the GenomicRanges R library [220]. *MspI* fragment reference profiles are created from the normal at this point.

```
# format allele counts output
format_output(patient_id, sample_id, sex, is_normal,
   is_patient_matched_normal, path, path_to_CAMDAC, build,
   txt_output=FALSE)
```

Note that bulk tumour and normal methylation rates are computed prior to BAF and LogR because they are needed to inform the GC content of each *MspI* fragment, which itself is required for LogR correction and copy number profiling. CAMDAC's polymorphism-independent methylation rate calculation is covered in **Chapter 3, Section 3.2.1**. If a patient-matched normal proxy for the tumour cell of origin is not available for differential methylation analysis, a reference panel can be used. The reference sample(s) should be at the very least sex-matched to avoid contamination of tumour-normal and tumour-tumour DMPs with female-to-male differential methylation signal.

```
# run bulk tumour and normal methylation processing
run_methylation_data_processing(patient_id, sample_id,
    normal_infiltrates_proxy_id, normal_origin_proxy_id, path,
    min_tumour=3, min_normal=10, n_cores, reference_panel=NULL)
```

Next, tumour and normal BAF and LogR values are combined to obtain allele-specific copy number profiles. Methylation rates are used to correct LogR for *MspI* fragment GC content. Note that the code for phasing BAF from multi-region data and improve copy number profiles is not included in the current version of CAMDAC but is available to share upon request.

```
# Get copy number (includes LogR bias correction)
get_copy_number(patient_id, sample_id, normal_id, build, path,
    min_tumour=1, min_normal=10, sex, path_to_CAMDAC,
    chr_names=c(1:22,"X"), n_cores, reference_panel=NULL,
    fragments_reference=NULL)
```

With methylation rates in hand for both the bulk tumour and normal infiltrates proxy as well as tumour copy number and purity estimates, we obtain the pure tumour methylation rates as detailed in **Chapter 3, Section 3.2.4**.

```
# Get pure tumour methylation rates
get_pure_tumour_methylation(patient_id, sample_id,
    normal_infiltrates_proxy_id, sex, path, path_to_CAMDAC, build,
    n_cores, reseg=FALSE)
```

With purified tumour methylation rates in hand, one can accurately examine distances between normal and tumour and between tumour methylomes themselves in R. Clustering analyses can also easily be performed by the user using well-established R packages. Leveraging CAMDAC purified methylomes, we then obtain differentially methylated positions and regions.

```
# Carry out differential methylation analysis
get_differential_methylation(patient_id, sample_id,
    normal_origin_proxy_id, sex, path, path_to_CAMDAC, build,
    effect_size=0.2, min_DMP_counts_in_DMR=5,
    min_consec_DMP_in_DMR=4, n_cores, reseg=FALSE)
```

Finally, users may choose to look for recurrently aberrated loci across their cohort. We recommend any gene-set enrichment analysis to be limited to hyper-methylated promoter-associated CpG Islands given that methylation at these loci is most correlated with expression. Users may leverage normal, deconvoluted tumour methylation rates and tumour-normal DMP calls to identify clonal bi-allelic and allele-specific methylation changes to shed light into tumour evolutionary histories.

# Bibliography


[1] Ksenia Skvortsova, Nicola Iovino, and Ozren Bogdanović. Functions and mechanisms of epigenetic inheritance in animals. *Nature Reviews Molecular Cell Biology*, 19(12):774–790, 2018.

[2] Stefan H. Stricker, Anna Köferle, and Stephan Beck. From profiles to function in epigenomics. *Nature Reviews Genetics*, 18(1):51–66, 2017.

[3] Treat B. Johnson and Robert D. Coghill. Researches on pyrimidines. C111. The discovery of 5-methyl-cytosine in tuberculinic acid, the nucleic acid of the tubercle bacillus. *Journal of the American Chemical Society*, 47(11):2838–2844, 1925.

[4] Rollin D. Hotchkiss. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *The Journal of biological chemistry*, 175(1):315–32, 1948.

[5] Susumu Ohno, William D. Kaplan, and Riojun Kinosita. Formation of the sex chromatin by a single X-chromosome in liver cells of Rattus norvegicus. *Experimental Cell Research*, 18(2):415–418, 1959.

[6] Ruth Holliday and James E. Pugh. DNA modification mechanisms and gene activity during development. *Science*, 187(4173):226–232, 1975.

[7] Arthur D. Riggs. X inactivation, differentiation, and DNA methylation. *Cytogenetic and Genome Research*, 14(1):9–25, 1975.

[8] Mary F. Lyon. Sex chromatin and gene action in the mammalian X-chromosome. *American journal of human genetics*, 14(2):135–48, 1962.

[9] James McGrath and Davor Solter. Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell*, 37(1):179–183, 1984.

[10] M. Azim Surani and Sheila C. Barton. Development of gynogenetic eggs in the mouse: Implications for parthenogenetic embryos. *Science*, 222(4627):1034–1036, 1983.

[11] Claudine Junien and Isabelle Henry. Genetics of Wilms' tumor: A blend of aberrant development and genomic imprinting. *Kidney International*, 46(5):1264–1279, 1994.

[12] Alan P. Wolffe and Marjori A. Matzke. Epigenetics: Regulation Through Repression. *Science*, 286(5439):481–486, 1999.

[13] Laura Lande-Diner, Jianmin Zhang, Ittai Ben-Porath, Ninette Amariglio, Ilana Keshet, Merav Hecht, Veronique Azuara, Amanda G. Fisher, Gideon Rechavi, and Howard Cedar. Role of DNA Methylation in Stable Gene Repression. *Journal of Biological Chemistry*, 282(16):12194–12200, 2007.

[14] Jianmin Zhang, Feng Xu, Tamar Hashimshony, Ilana Keshet, and Howard Cedar. Establishment of transcriptional competence in early and late S phase. *Nature*, 420(6912):198–202, 2002.

[15] Peri H. Tate and Adrian P. Bird. Effects of DNA methylation on DNA-binding proteins and gene expression. *Current Opinion in Genetics and Development*, 3(2):226–231, 1993.

[16] G. Buschhausen, Burghardt Wittig, Monica Graessmann, and Adolf Graessmann. Chromatin structure is required to block transcription of the methylated herpes simplex virus thymidine kinase gene. *Proceedings of the National Academy of Sciences of the United States of America*, 84(5):1177–1181, 1987.

[17] Stefan U. Kass, Nicoletta Landsberger, and Alan P. Wolffe. DNA methylation directs a time-dependent repression of transcription initiation. *Current Biology*, 7(3):157–165, 1997.

[18] Nathaniel D. Heintzman, Rhona K. Stuart, Gary Hon, Yutao Fu, Christina W Ching, R. David Hawkins, Leah O. Barrera, Sara Van Calcar, Chunxu Qu, Keith A. Ching, Wei Wang, Zhiping Weng, Roland D. Green, Gregory E. Crawford, and Bing Ren. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39(3):311–318, 2007.

[19] Dmitry V. Fyodorov, Bing-Rui Zhou, Arthur I. Skoultchi, and Yawen Bai. Emerging roles of linker histones in regulating chromatin structure and function. *Nature Reviews Molecular Cell Biology*, 19(3):192–206, 2018.

[20] J. Doskočil and F. Šorm. Distribution of 5-methylcytosine in pyrimidine sequences of deoxyribonucleic acids. *Biochimica et Biophysica Acta*, 55(6):953–959, 1962.

[21] Adrian P. Bird. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, 8(7):1499–1504, 1980.

[22] Ryan Lister, Mattia Pelizzola, Robert H. Dowen, R. David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R. Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, Lee Edsall, Jessica Antosiewicz-Bourget, Ron Stewart, Victor Ruotti, A. Harvey Millar, James A. Thomson, Bing Ren, and Joseph R. Ecker. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, 2009.

[23] M. Frommer, L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, and C. L. Paul. A genomic sequencing protocol that yields a positive display of 5- methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, 89:1827–1831, 1992.

[24] Susan J. Clark, Janet Harrison, Cheryl L. Paul, and Marianne Frommer. High sensitivity mapping of methylated cytosines. *Nucleic Acids Research*, 22(15):2990–2997, 1994.

[25] Junhua Zhou, Minqiong Zhao, Zefang Sun, Feilong Wu, Yucong Liu, Xianghua Liu, Zuping He, Quanze He, and Quanyuan He. BCREval: A computational method to estimate the bisulfite conversion ratio in WGBS. *BMC Bioinformatics*, 21, 2020.

[26] Felix Krueger and Simon R. Andrews. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11):1571–1572, 2011.

[27] Cora Mund, Verena Beier, Peter Bewerunge, Michael Dahms, Frank Lyko, and Jörg D. Hoheisel. Array-based analysis of genomic DNA methylation patterns of the tumour suppressor gene p16INK4A promoter in colon carcinoma cell lines. *Nucleic Acids Research*, 33(8), 2005.

[28] Huidong Shi, Sabine Maier, Inko Nimmrich, Pearlly S. Yan, Charles W. Caldwell, Alexander Olek, and Tim Hui Ming Huang. Oligonucleotide-based microarray for DNA methylation analysis: Principles and applications, 2003.

[29] Patrick Boyle, Kendell Clement, Hongcang Gu, Zachary D Smith, Michael Ziller, Jennifer L Fostel, Laurie Holmes, Jim Meldrim, Fontina Kelley, Andreas Gnirke, and Alexander Meissner. Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. *Genome Biology*, 13(10):R92, 2012.

[30] Alexander Meissner, Tarjei S. Mikkelsen, Hongcang Gu, Marius Wernig, Jacob Hanna, Andrey Sivachenko, Xiaolan Zhang, Bradley E. Bernstein, Chad Nusbaum, David B. Jaffe, Andreas Gnirke, Rudolf Jaenisch, and Eric S. Lander. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770, 2008.

[31] Hongcang Gu, Zachary D Smith, Christoph Bock, Patrick Boyle, Andreas Gnirke, and Alexander Meissner. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nature Protocols*, 6(4):468–481, 2011.

[32] A. Meissner, Andreas Gnirke, George W. Bell, Bernard Ramsahoye, Eric S. Lander, and Rudolf Jaenisch. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, 33(18):5868–5877, 2005.

[33] M J Ziller, K D Hansen, A Meissner, and M J Aryee. Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nat Methods*, 12(3):230–4, 2015.

[34] Ruth Pidsley, Elena Zotenko, Timothy J. Peters, Mitchell G. Lawrence, Gail P. Risbridger, Peter Molloy, Susan Van Djik, Beverly Muhlhausler, Clare Stirzaker, and Susan J. Clark. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, 17, 2016.

[35] Shawn J. Cokus, Suhua Feng, Xiaoyu Zhang, Zugen Chen, Barry Merriman, Christian D. Haudenschild, Sriharsa Pradhan, Stanley F. Nelson, Matteo Pellegrini, and Steven E. Jacobsen. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, 452:215–219, 2008.

[36] Yonatan Stelzer, Daniel Ronen, Christoph Bock, Patrick Boyle, Alexander Meissner, and Nissim Benvenisty. Identification of novel imprinted differentially methylated regions by global analysis of human-parthenogenetic-induced pluripotent stem cells. *Stem Cell Reports*, 1(1):79–89, 2013.

[37] Zachary D. Smith, Michelle M. Chan, Tarjei S. Mikkelsen, Hongcang Gu, Andreas Gnirke, Aviv Regev, and Alexander Meissner. A unique regula-

tory phase of DNA methylation in the early mammalian embryo. *Nature*, 484:339–344, 2012.

[38] Zachary D. Smith, Michelle M. Chan, Kathryn C. Humm, Rahul Karnik, Shila Mekhoubad, Aviv Regev, Kevin Eggan, and Alexander Meissner. DNA methylation dynamics of the human preimplantation embryo. *Nature*, 511:611–615, 2014.

[39] Melanie Ehrlich and Michelle Lacey. DNA methylation and differentiation: silencing, upregulation and modulation of gene expression. *Epigenomics*, 5(5):553–568, 2013.

[40] Dan A. Landau, Kendell Clement, Michael J. Ziller, Patrick Boyle, Jean Fan, Hongcang Gu, Kristen Stevenson, Carrie Sougnez, Lili Wang, Shuqiang Li, Dylan Kotliar, Wandi Zhang, Mahmoud Ghandi, Levi Garraway, Stacey M. Fernandes, Kenneth J. Livak, Stacey Gabriel, Andreas Gnirke, Eric S. Lander, Jennifer R. Brown, Donna Neuberg, Peter V. Kharchenko, Nir Hacohen, Gad Getz, Alexander Meissner, and Catherine J. Wu. Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia. *Cancer Cell*, 26(6):813–825, 2014.

[41] Katherine E. Varley, Jason Gertz, Kevin M. Bowling, Stephanie L. Parker, Timothy E. Reddy, Florencia Pauli-Behn, Marie K. Cross, Brian A. Williams, John A. Stamatoyannopoulos, Gregory E. Crawford, Devin M. Absher, Barbara J. Wold, and Richard M. Myers. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Research*, 23(3):555–567, 2013.

[42] Adrian Bird. DNA methylation patterns and epigenetic memory. *Genes & Development*, 16(1):6–21, 2002.

[43] Rudolf Jaenisch and Adrian Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33(S3):245–254, 2003.

[44] Adrian P. Bird. CpG-rich islands and the function of DNA methylation. *Nature*, 321(6067):209–213, 1986.

[45] Margaret Gardiner-Garden and Marianne Frommer. CpG Islands in vertebrate genomes. *Journal of Molecular Biology*, 196(2):261–282, 1987.

[46] Zachary D. Smith, Hongcang Gu, Christoph Bock, Andreas Gnirke, and Alexander Meissner. High-throughput bisulfite sequencing in mammalian genomes. *Methods*, 48(3):226–232, 2009.

[47] Hongcang Gu, Christoph Bock, Tarjei S Mikkelsen, Natalie Jäger, Zachary D Smith, Eleni Tomazou, Andreas Gnirke, Eric S Lander, and Alexander Meissner. Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nature Methods*, 7(2):133–136, 2010.

[48] Natasha Lane, Wendy Dean, Sylvia Erhardt, Petra Hajkova, Azim Surani, Jörn Walter, and Wolf Reik. Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. *Genesis*, 35(2):88–93, 2003.

[49] Sébastien A. Smallwood, Shin-ichi Tomizawa, Felix Krueger, Nico Ruf, Natasha Carli, Anne Segonds-Pichon, Shun Sato, Kenichiro Hata, Simon R. Andrews, and Gavin Kelsey. Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nature Genetics*, 43(8):811–814, 2011.

[50] Timothy H. Bestor. DNA methylation: evolution of a bacterial immune function into a regulator of gene expression and genome structure in higher eukaryotes. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 326(1235):179–187, 1990.

[51] Detlev Jähner, Heidi Stuhlmann, Colin L. Stewart, Klaus Harbers, Jürgen Löhler, Iva Simon, and Rudolf Jaenisch. De novo methylation and expression of retroviral genomes during mouse embryogenesis. *Nature*, 298(5875):623–628, 1982.

[52] Helen M. Rowe and Didier Trono. Dynamic control of endogenous retroviruses during development. *Virology*, 411(2):273–287, 2011.

[53] Wanding Zhou, Gangning Liang, Peter L. Molloy, and Peter A. Jones. DNA methylation enables transposable element-driven genome expansion. *Proceedings of the National Academy of Sciences*, 117(32):19359–19366, 2020.

[54] Timothy H. Bestor. The DNA methyltransferases of mammals. *Human Molecular Genetics*, 9(16):2395–2402, 2000.

[55] Joan Barau, Aurélie Teissandier, Natasha Zamudio, Stéphanie Roy, Valérie Nalesso, Yann Hérault, Florian Guillou, and Déborah Bourc'his. The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. *Science*, 354(6314):909–912, 2016.

[56] Dylan Pannell. Retrovirus vector silencing is de novo methylase independent and marked by a repressive histone code. *The EMBO Journal*, 19(21):5884–5894, 2000.

[57] Michaël Imbeault, Pierre-Yves Helleboid, and Didier Trono. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*, 543(7646):550–554, 2017.

[58] Anastasiya Kazachenka, Tessa M. Bertozzi, Marcela K. Sjoberg-Herrera, Nic Walker, Joseph Gardner, Richard Gunning, Elena Pahita, Sarah Adams, David Adams, and Anne C. Ferguson-Smith. Identification, Characterization, and Heritability of Murine Metastable Epialleles: Implications for Non-genetic Inheritance. *Cell*, 175(5):1259–1271.e13, 2018.

[59] Rachna Goyal. Accuracy of DNA methylation pattern preservation by the Dnmt1 methyltransferase. *Nucleic Acids Research*, 34(4):1182–1188, 2006.

[60] Adriaan van der Graaf, René Wardenaar, Drexel A. Neumann, Aaron Taudt, Ruth G. Shaw, Ritsert C. Jansen, Robert J. Schmitz, Maria Colomé-Tatché,

and Frank Johannes. Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. *Proceedings of the National Academy of Sciences*, 112(21):6676–6681, 2015.

[61] Shinsuke Ito, Ana C. D'Alessio, Olena V. Taranova, Kwonho Hong, Lawrence C. Sowers, and Yi Zhang. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*, 466(7310):1129–1133, 2010.

[62] Yu-Fei He, Bin-Zhong Li, Zheng Li, Peng Liu, Yang Wang, Qingyu Tang, Jianping Ding, Yingying Jia, Zhangcheng Chen, Lin Li, Yan Sun, Xiuxue Li, Qing Dai, Chun-Xiao Song, Kangling Zhang, Chuan He, and Guo-Liang Xu. Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. *Science*, 333(6047):1303–1307, 2011.

[63] Shinsuke Ito, Li Shen, Qing Dai, Susan C. Wu, Leonard B. Collins, James A. Swenberg, Chuan He, and Yi Zhang. Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. *Science*, 333(6047):1300–1303, 2011.

[64] Matthew T. Bennett, M. T. Rodgers, Alexander S. Hebert, Lindsay E. Ruslander, Leslie Eisele, and Alexander C. Drohat. Specificity of Human Thymine DNA Glycosylase Depends on N -Glycosidic Bond Stability. *Journal of the American Chemical Society*, 128(38):12510–12519, 2006.

[65] Rahul M. Kohli and Yi Zhang. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature*, 502(7472):472–479, 2013.

[66] Zita Liutkevičiūtė, Edita Kriukienė, Janina Ličytė, Milda Rudytė, Giedrė Urbanavičiūtė, and Saulius Klimašauskas. Direct Decarboxylation of 5-Carboxylcytosine by DNA C5-Methyltransferases. *Journal of the American Chemical Society*, 136(16):5884–5887, 2014.

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

[68] William A. Flavahan, Elizabeth Gaskell, and Bradley E. Bernstein. Epigenetic plasticity and the hallmarks of cancer. *Science*, 357(6348):eaal2380, 2017.

[69] Hui Shen and Peter W. Laird. Interplay between the Cancer Genome and Epigenome. *Cell*, 153(1):38–55, 2013.

[70] Gregory Hannum, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, Srini-Vas Sadda, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, Rob Deconde, Menzies Chen, Indika Rajapakse, Stephen Friend, Trey Ideker, and Kang Zhang. Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Molecular Cell*, 49(2):359–367, 2013.

[71] Jia-Jie Hao, De-Chen Lin, Huy Q Dinh, Anand Mayakonda, Yan-Yi Jiang, Chen Chang, Ye Jiang, Chen-Chen Lu, Zhi-Zhou Shi, Xin Xu, Yu Zhang, Yan Cai, Jin-Wu Wang, Qi-Min Zhan, Wen-Qiang Wei, Benjamin P Berman, Ming-Rong Wang, and H Phillip Koeffler. Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. *Nature Genetics*, 48(12):1500–1507, 2016.

[72] Steve Horvath. DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10):R115, 2013.

[73] Andrew P. Feinberg, Rolf Ohlsson, and Steven Henikoff. The epigenetic progenitor origin of human cancer. *Nature Reviews Genetics*, 7(1):21–33, 2006.

[74] Douglas Hanahan and Robert A. Weinberg. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674, 2011.

[75] Daniela F. Quail and Johanna A. Joyce. Microenvironmental regulation of tumor progression and metastasis. *Nature Medicine*, 19(11):1423–1437, 2013.

[76] Valerie Greger, Eberhard Passarge, Wolfgang Höpping, Elmar Messmer, and Bernhard Horsthemke. Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. *Human Genetics*, 83(2):155–158, 1989.

[77] Rafael A Irizarry, Christine Ladd-Acosta, Bo Wen, Zhijin Wu, Carolina Montano, Patrick Onyango, Hengmi Cui, Kevin Gabo, Michael Rongione, Maree Webster, Hong Ji, James B Potash, Sarven Sabunciyan, and Andrew P Feinberg. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genetics*, 41(2):178–186, 2009.

[78] Andrew P. Feinberg and Bert Vogelstein. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, 301(5895):89–92, 1983.

[79] Colum P. Walsh, J. Richard Chaillet, and Timothy H. Bestor. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature Genetics*, 20(2):116–117, 1998.

[80] Guo-Liang Xu, Timothy H. Bestor, Déborah Bourc'his, Chih-Lin Hsieh, Niels Tommerup, Merete Bugge, Maj Hulten, Xiaoyan Qu, James J. Russo, and Evani Viegas-Péquignot. Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature*, 402(6758):187–191, 1999.

[81] Benjamin P. Berman, Daniel J. Weisenberger, Joseph F. Aman, Toshinori Hinoue, Zachary Ramjan, Yaping Liu, Houtan Noushmehr, Christopher P E Lange, Cornelis M van Dijk, Rob A E M Tollenaar, David Van Den Berg, and Peter W. Laird. Regions of focal DNA hypermethylation and long-range hy-

pomethylation in colorectal cancer coincide with nuclear lamina–associated domains. *Nature Genetics*, 44(1):40–46, 2012.

[82] James G. Herman, Farida Latif, Yongkai Weng, Michael I. Lerman, Berton Zbar, Sue Liu, Dvorit Samid, D. S. Duan, James R. Gnarra, and W. Marston Linehan. Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. *Proceedings of the National Academy of Sciences*, 91(21):9700–9704, 1994.

[83] James G. Herman, Adrian Merlo, Li Mao, Rena G. Lapidus, J P Issa, Nancy E. Davidson, David Sidransky, and Stephen B. Baylin. Inactivation of the CDKN2/p16/MTS1 gene is frequently associated with aberrant DNA methylation in all common human cancers. *Cancer research*, 55(20):4525–30, 1995.

[84] Mirella Gonzalez-Zulueta, Christina M. Bender, Allen S. Yang, T Nguyen, Robert W. Beart, J M Van Tornout, and Peter A. Jones. Methylation of the 5′ CpG island of the p16/CDKN2 tumor suppressor gene in normal and transformed human tissues correlates with gene silencing. *Cancer research*, 55(20):4531–5, 1995.

[85] Andrew Feber, Pawan Dhami, Liqin Dong, Patricia de Winter, Wei Shen Tan, Mónica Martínez-Fernández, Dirk S. Paul, Antony Hynes-Allen, Sheida Rezaee, Pratik Gurung, Simon Rodney, Ahmed Mehmood, Felipe Villacampa, Federico de la Rosa, Charles Jameson, Kar Keung Cheng, Maurice P. Zeegers, Richard T. Bryan, Nicholas D. James, Jesus M. Paramio, Alex Freeman, Stephan Beck, and John D. Kelly. UroMark—a urinary biomarker assay for the detection of bladder cancer. *Clinical Epigenetics*, 9(1):8, 2017.

[86] Holger Heyn and Manel Esteller. DNA methylation profiling in the clinic: applications and challenges. *Nature Reviews Genetics*, 13(10):679–692, 2012.

[87] Alexander Koch, Sophie C. Joosten, Zheng Feng, Tim C. de Ruijter, Muriel X. Draht, Veerle Melotte, Kim M. Smits, Jurgen Veeck, James G. Her-

man, Leander Van Neste, Wim Van Criekinge, Tim De Meyer, and Manon van Engeland. Analysis of DNA methylation in cancer: location revisited. *Nature Reviews Clinical Oncology*, 15(7):459–466, 2018.

[88] ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93, 2020.

[89] Carolyn Hutter and Jean Claude Zenklusen. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell*, 173(2):283–285, 2018.

[90] Nicola J. Curtin. DNA repair dysregulation from cancer driver to therapeutic target. *Nature Reviews Cancer*, 12(12):801–817, 2012.

[91] Hans Ulrich Klein and Katja Hebestreit. An evaluation of methods to test predefined genomic regions for differential methylation in bisulfite sequencing data. *Briefings in Bioinformatics*, 17(5):796–807, 2016.

[92] Mark D. Robinson, Abdullah Kahraman, Charity W. Law, Helen Lindsay, Malgorzata Nowicka, Lukas M. Weber, and Xiaobei Zhou. Statistical methods for detecting differentially methylated loci and regions. *Frontiers in Genetics*, 5:1–7, 2014.

[93] Christine Coulondre, Jeffrey H. Miller, Philip J. Farabaugh, and Walter Gilbert. Molecular basis of base substitution hotspots in Escherichia coli. *Nature*, 274(5673):775–780, 1978.

[94] Ludmil B. Alexandrov, Serena Nik-Zainal, David C. Wedge, Samuel A. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P. Butler, Carlos Caldas, Helen R. Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörd, John A. Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinski, Natalie Jäger, David T. W. Jones, David Jones, Stian Knappskog, Marcel Kool, Sunil R. Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C.

Munshi, Hiromi Nakamura, Paul A. Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V. Pearson, Xose S. Puente, Keiran Raine, Manasa Ramakrishna, Andrea L. Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N. Schumacher, Paul N. Span, Jon W. Teague, Yasushi Totoki, Andrew N. J. Tutt, Rafael Valdés-Mas, Marit M. van Buuren, Laura van 't Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R. Yates, Jessica Zucman-Rossi, P. Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M. Grimmond, Reiner Siebert, Elías Campo, Tatsuhiro Shibata, Stefan M. Pfister, Peter J. Campbell, and Michael R. Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013.

[95] Eugene Houseman, William P Accomando, Devin C Koestler, Brock C Christensen, Carmen J Marsit, Heather H Nelson, John K Wiencke, and Karl T Kelsey. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13(1):86, 2012.

[96] Andrew E. Teschendorff, Charles E. Breeze, Shijie C. Zheng, and Stephan Beck. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics*, 18(1):105, 2017.

[97] Ginell Elliott, Chibo Hong, Xiaoyun Xing, Xin Zhou, Daofeng Li, Cristian Coarfa, Robert J.A. Bell, Cecile L. Maire, Keith L. Ligon, Mahvash Sigaroudinia, Philippe Gascard, Thea D. Tlsty, R. Alan Harris, Leonard C. Schalkwyk, Misha Bilenky, Jonathan Mill, Peggy J. Farnham, Manolis Kellis, Marco A. Marra, Aleksandar Milosavljevic, Martin Hirst, Gary D. Stormo, Ting Wang, and Joseph F. Costello. Intermediate DNA methylation is a conserved signature of genome regulation. *Nature Communications*, 6(1):6363, 2015.

[98] David Brocks, Yassen Assenov, Sarah Minner, Olga Bogatyrova, Ronald Simon, Christina Koop, Christopher Oakes, Manuela Zucknick, Daniel Bern-

hard Lipka, Joachim Weischenfeldt, Lars Feuerbach, Richard Cowper-Sal·lari, Mathieu Lupien, Benedikt Brors, Jan Korbel, Thorsten Schlomm, Amos Tanay, Guido Sauter, Clarissa Gerhäuser, and Christoph Plass. Intratumor DNA Methylation Heterogeneity Reflects Clonal Evolution in Aggressive Prostate Cancer. *Cell Reports*, 8(3):798–806, 2014.

[99] Johanna Klughammer, Barbara Kiesel, Thomas Roetzer, Nikolaus Fortelny, Amelie Nemc, Karl-Heinz Nenning, Julia Furtner, Nathan C. Sheffield, Paul Datlinger, Nadine Peter, Martha Nowosielski, Marco Augustin, Mario Mischkulnig, Thomas Ströbel, Donat Alpar, Bekir Ergüner, Martin Senekowitsch, Patrizia Moser, Christian F. Freyschlag, Johannes Kerschbaumer, Claudius Thomé, Astrid E. Grams, Günther Stockhammer, Melitta Kitzwoegerer, Stefan Oberndorfer, Franz Marhold, Serge Weis, Johannes Trenkler, Johanna Buchroithner, Josef Pichler, Johannes Haybaeck, Stefanie Krassnig, Kariem Mahdy Ali, Gord von Campe, Franz Payer, Camillo Sherif, Julius Preiser, Thomas Hauser, Peter A. Winkler, Waltraud Kleindienst, Franz Würtz, Tanisa Brandner-Kokalj, Martin Stultschnig, Stefan Schweiger, Karin Dieckmann, Matthias Preusser, Georg Langs, Bernhard Baumann, Engelbert Knosp, Georg Widhalm, Christine Marosi, Johannes A. Hainfellner, Adelheid Woehrer, and Christoph Bock. The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space. *Nature Medicine*, 24(10):1611–1624, 2018.

[100] Tali Mazor, Aleksandr Pankov, Brett E Johnson, Chibo Hong, Emily G Hamilton, Robert J A Bell, Ivan V Smirnov, Gerald F Reis, Joanna J Phillips, Michael J Barnes, Ahmed Idbaih, Agusti Alentorn, Jenneke J Kloezeman, Martine L M Lamfers, Andrew W Bollen, Barry S Taylor, Annette M Molinaro, Adam B Olshen, Susan M Chang, Jun S Song, and Joseph F Costello. DNA Methylation and Somatic Mutations Converge on the Cell Cycle and Define Similar Evolutionary Histories in Brain Tumors. *Cancer cell*, 28(3):307–317, 2015.

[101] Gilad Landan, Netta Mendelson Cohen, Zohar Mukamel, Amir Bar, Alina Molchadsky, Ran Brosh, Shirley Horn-Saban, Daniela Amann Zalcenstein, Naomi Goldfinger, Adi Zundelevich, Einav Nili Gal-Yam, Varda Rotter, and Amos Tanay. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nature Genetics*, 2012.

[102] Sheng Li, Francine E Garrett-Bakelman, Stephen S Chung, Mathijs A Sanders, Todd Hricik, Franck Rapaport, Jay Patel, Richard Dillon, Priyanka Vijay, Anna L Brown, Alexander E Perl, Joy Cannon, Lars Bullinger, Selina Luger, Michael Becker, Ian D Lewis, Luen Bik To, Ruud Delwel, Bob Löwenberg, Hartmut Döhner, Konstanze Döhner, Monica L Guzman, Duane C Hassane, Gail J Roboz, David Grimwade, Peter J M Valk, Richard J D'Andrea, Martin Carroll, Christopher Y Park, Donna Neuberg, Ross Levine, Ari M Melnick, and Christopher E Mason. Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nature Medicine*, 22(7):792–799, 2016.

[103] Nathan C Sheffield, Gaelle Pierron, Johanna Klughammer, Paul Datlinger, Andreas Schönegger, Michael Schuster, Johanna Hadler, Didier Surdez, Delphine Guillemot, Eve Lapouble, Paul Freneaux, Jacqueline Champigneulle, Raymonde Bouvier, Diana Walder, Ingeborg M Ambros, Caroline Hutter, Eva Sorz, Ana T Amaral, Enrique de Álava, Katharina Schallmoser, Dirk Strunk, Beate Rinner, Bernadette Liegl-Atzwanger, Berthold Huppertz, Andreas Leithner, Gonzague de Pinieux, Philippe Terrier, Valérie Laurence, Jean Michon, Ruth Ladenstein, Wolfgang Holter, Reinhard Windhager, Uta Dirksen, Peter F Ambros, Olivier Delattre, Heinrich Kovar, Christoph Bock, and Eleni M Tomazou. DNA methylation heterogeneity defines a disease spectrum in Ewing sarcoma. *Nature Medicine*, 23(3):386–395, 2017.

[104] Michael Scherer, Almut Nebel, Andre Franke, Jörn Walter, Thomas Lengauer, Christoph Bock, Fabian Müller, and Markus List. Quantita-

tive comparison of within-sample heterogeneity scores for DNA methylation data. *Nucleic Acids Research*, 48(8):e46–e46, 2020.

[105] Shicheng Guo, Dinh Diep, Nongluk Plongthongkum, Ho-Lim Fung, Kang Zhang, and Kun Zhang. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nature Genetics*, 49(4):635–642, 2017.

[106] James E Barrett, Andrew Feber, Javier Herrero, Miljana Tanic, Gareth A Wilson, Charles Swanton, and Stephan Beck. Quantification of tumour evolution and heterogeneity via Bayesian epiallele detection. *BMC Bioinformatics*, 18(1):354, 2017.

[107] Stefan C. Dentro, David C. Wedge, and Peter Van Loo. Principles of Reconstructing the Subclonal Architecture of Cancers. *Cold Spring Harbor Perspectives in Medicine*, 7(8):a026625, 2017.

[108] Ignaty Leshchiner, Dimitri Livitz, Justin Gainor, Daniel Rosebrock, Oliver Spiro, Aina Martinez, Edmund Mroz, Jessica Lin, Chip Stewart, Jaegil Kim, Liudmila Elagina, Ivana Bozic, Mari Mino-Kenudson, Marguerite Rooney, Sai-Hong Ignatius Ou, Catherine Wu, James Rocco, Jeffrey Engelman, Alice Shaw, and Gad Getz. Comprehensive analysis of tumour initiation, spatial and temporal progression under multiple lines of treatment. *bioRxiv*, 2018.

[109] Amit G. Deshwar, Shankar Vembu, Christina K. Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):35, 2015.

[110] Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P. Shah. PyClone: statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4):396–398, 2014.

[111] Stefan C. Dentro, Ignaty Leshchiner, Kerstin Haase, Maxime Tarabichi, Jeff Wintersinger, Amit G. Deshwar, Kaixian Yu, Yulia Rubanova, Geoff Macintyre, Ignacio Vázquez-García, Kortine Kleinheinz, Dimitri G. Livitz, Salem Malikic, Nilgun Donmez, Subhajit Sengupta, Jonas Demeulemeester, Pavana Anur, Clemency Jolly, Marek Cmero, Daniel Rosebrock, Steven Schumacher, Yu Fan, Matthew Fittall, Ruben M. Drews, Xiaotong Yao, Juhee Lee, Matthias Schlesner, Hongtu Zhu, David J. Adams, Gad Getz, Paul C. Boutros, Marcin Imielinski, Rameen Beroukhim, S. Cenk Sahinalp, Yuan Ji, Martin Peifer, Inigo Martincorena, Florian Markowetz, Ville Mustonen, Ke Yuan, Moritz Gerstung, Paul T. Spellman, Wenyi Wang, Quaid D. Morris, David C. Wedge, Peter Van Loo, on behalf of the PCAWG Evolution Group, Heterogeneity Working, and the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network. Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types. *bioRxiv*, page 312041, 2018.

[112] Mariam Jamal-Hanjani, Gareth A. Wilson, Nicholas McGranahan, Nicolai J. Birkbak, Thomas B.K. Watkins, Selvaraju Veeriah, Seema Shafi, Diana H Johnson, Richard Mitter, Rachel Rosenthal, Max Salm, Stuart Horswell, Mickael Escudero, Nik Matthews, Andrew Rowan, Tim Chambers, David A Moore, Samra Turajlic, Hang Xu, Siow-Ming Lee, Martin D Forster, Tanya Ahmad, Crispin T Hiley, Christopher Abbosh, Mary Falzon, Elaine Borg, Teresa Marafioti, David Lawrence, Martin Hayward, Shyam Kolvekar, Nikolaos Panagiotopoulos, Sam M Janes, Ricky Thakrar, Asia Ahmed, Fiona Blackhall, Yvonne Summers, Rajesh Shah, Leena Joseph, Anne M Quinn, Phil A Crosbie, Babu Naidu, Gary Middleton, Gerald Langman, Simon Trotter, Marianne Nicolson, Hardy Remmen, Keith Kerr, Mahendran Chetty, Lesley Gomersall, Dean A. Fennell, Apostolos Nakas, Sridhar Rathinam, Girija Anand, Sajid Khan, Peter Russell, Veni Ezhil, Babikir Ismail, Melanie Irvin-Sellers, Vineet Prakash, Jason F Lester, Malgorzata Kornaszewska, Richard Attanoos, Haydn Adams, Helen Davies, Stefan Dentro, Philippe Taniere, Brendan O'Sullivan, Helen L Lowe, John A Hartley, Natasha Iles,

Harriet Bell, Yenting Ngai, Jacqui A Shaw, Javier Herrero, Zoltan Szallasi, Roland F Schwarz, Aengus Stewart, Sergio A Quezada, John Le Quesne, Peter Van Loo, Caroline Dive, Allan Hackshaw, and Charles Swanton. Tracking the Evolution of Non–Small-Cell Lung Cancer. *New England Journal of Medicine*, 376(22):2109–2121, 2017.

[113] Mariam Jamal-Hanjani, Alan Hackshaw, Yenting Ngai, Jacqueline Shaw, Caroline Dive, Sergio Quezada, Gary Middleton, Elza de Bruin, John Le Quesne, Seema Shafi, Mary Falzon, Stuart Horswell, Fiona Blackhall, Iftekhar Khan, Sam Janes, Marianne Nicolson, David Lawrence, Martin Forster, Dean Fennell, Siow Ming Lee, Jason Lester, Keith Kerr, Salli Muller, Natasha Iles, Sean Smith, Nirupa Murugaesu, Richard Mitter, Max Salm, Aengus Stuart, Nik Matthews, Haydn Adams, Tanya Ahmad, Richard Attanoos, Jonathan Bennett, Nicolai Juul Birkbak, Richard Booton, Ged Brady, Keith Buchan, Arrigo Capitano, Mahendran Chetty, Mark Cobbold, Philip Crosbie, Helen Davies, Alan Denison, Madhav Djearman, Jacki Goldman, Tom Haswell, Leena Joseph, Malgorzata Kornaszewska, Matthew Krebs, Gerald Langman, Mairead MacKenzie, Joy Millar, Bruno Morgan, Babu Naidu, Daisuke Nonaka, Karl Peggs, Catrin Pritchard, Hardy Remmen, Andrew Rowan, Rajesh Shah, Elaine Smith, Yvonne Summers, Magali Taylor, Selvaraju Veeriah, David Waller, Ben Wilcox, Maggie Wilcox, Ian Woolhouse, Nicholas McGranahan, and Charles Swanton. Tracking Genomic Cancer Evolution for Precision Medicine: The Lung TRACERx Study. *PLoS Biology*, 2014.

[114] Rachel Rosenthal, Elizabeth Larose Cadieux, Roberto Salgado, Maise Al Bakir, David A. Moore, Crispin T. Hiley, Tom Lund, Miljana Tanić, James L. Reading, Kroopa Joshi, Jake Y. Henry, Ehsan Ghorani, Gareth A. Wilson, Nicolai J. Birkbak, Mariam Jamal-Hanjani, Selvaraju Veeriah, Zoltan Szallasi, Sherene Loi, Matthew D. Hellmann, Andrew Feber, Benny Chain, Javier Herrero, Sergio A. Quezada, Jonas Demeulemeester, Peter Van Loo, Stephan

Beck, Nicholas McGranahan, and Charles Swanton. Neoantigen-directed immune escape in lung cancer evolution. *Nature*, 567(7749):479–485, 2019.

[115] Scott L Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W Laird, Robert C Onofrio, Wendy Winckler, Barbara A Weir, Rameen Beroukhim, David Pellman, Douglas A Levine, Eric S Lander, Matthew Meyerson, and Gad Getz. Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, 30(5):413–421, 2012.

[116] Serena Nik-Zainal, Peter Van Loo, David C. Wedge, Ludmil B. Alexandrov, Christopher D. Greenman, King Wai Lau, Keiran Raine, David Jones, John Marshall, Manasa Ramakrishna, Adam Shlien, Susanna L. Cooke, Jonathan Hinton, Andrew Menzies, Lucy A. Stebbings, Catherine Leroy, Mingming Jia, Richard Rance, Laura J. Mudie, Stephen J. Gamble, Philip J. Stephens, Stuart McLaren, Patrick S. Tarpey, Elli Papaemmanuil, Helen R. Davies, Ignacio Varela, David J. McBride, Graham R. Bignell, Kenric Leung, Adam P. Butler, Jon W. Teague, Sancha Martin, Goran Jönsson, Odette Mariani, Sandrine Boyault, Penelope Miron, Aquila Fatima, Anita Langerod, Samuel A J R Aparicio, Andrew Tutt, Anieta M. Sieuwerts, Åke Borg, Gilles Thomas, Anne Vincent Salomon, Andrea L. Richardson, Anne Lise Borresen-Dale, P. Andrew Futreal, Michael R. Stratton, and Peter J. Campbell. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, 2012.

[117] P. Van Loo, Silje H. Nordgard, O. C. Lingjaerde, Hege G. Russnes, Inga H. Rye, Wei Sun, Victor J. Weigman, Peter Marynen, Anders Zetterberg, Bjørn Naume, Charles M. Perou, A.-L. Borresen-Dale, and Vessela N. Kristensen. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915, 2010.

[118] Jung H. Kim, Saravana M. Dhanasekaran, John R. Prensner, Xuhong Cao, Daniel Robinson, Shanker Kalyana-Sundaram, Christina Huang, Sunita Shankar, Xiaojun Jing, Matthew Iyer, Ming Hu, Lee Sam, Catherine Grasso,

Christopher A. Maher, Nallasivam Palanisamy, Rohit Mehra, Hal D. Kominsky, Javed Siddiqui, Jindan Yu, Zhaohui S. Qin, and Arul M. Chinnaiyan. Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. *Genome Research*, 21(7):1028–1041, 2011.

[119] Yoshinao Ruike, Yukako Imanaka, Fumiaki Sato, Kazuharu Shimizu, and Gozoh Tsujimoto. Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. *BMC Genomics*, 11(1):137, 2010.

[120] Mark D. Robinson, Dario Strbenac, Clare Stirzaker, Aaron L. Statham, Jenny Song, Terence P. Speed, and Susan J. Clark. Copy-number-aware differential analysis of quantitative DNA sequencing data. *Genome Research*, 22(12):2489–2496, 2012.

[121] Xiaoqi Zheng, Naiqian Zhang, Hua-Jun Wu, and Hao Wu. Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. *Genome Biology*, 18(1):17, 2017.

[122] Travis I. Zack, Steven E. Schumacher, Scott L. Carter, Andrew D. Cherniack, Gordon Saksena, Barbara Tabak, Michael S. Lawrence, Cheng-Zhong Zhang, Jeremiah Wala, Craig H. Mermel, Carrie Sougnez, Stacey B. Gabriel, Bryan Hernandez, Hui Shen, Peter W. Laird, Gad Getz, Matthew Meyerson, and Rameen Beroukhim. Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, 45(10):1134–1140, 2013.

[123] Saioa López, Emilia L. Lim, Stuart Horswell, Kerstin Haase, Ariana Huebner, Michelle Dietzen, Thanos P. Mourikis, Thomas B. K. Watkins, Andrew Rowan, Sally M. Dewhurst, Nicolai J. Birkbak, Gareth A. Wilson, Peter Van Loo, Mariam Jamal-Hanjani, Charles Swanton, and Nicholas McGranahan. Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. *Nature Genetics*, 52(3):283–293, 2020.

[124] Moritz Gerstung, Clemency Jolly, Ignaty Leshchiner, Stefan C. Dentro, Santiago Gonzalez, Daniel Rosebrock, Thomas J. Mitchell, Yulia Rubanova, Pavana Anur, Kaixian Yu, Maxime Tarabichi, Amit Deshwar, Jeff Wintersinger, Kortine Kleinheinz, Ignacio Vázquez-García, Kerstin Haase, Lara Jerman, Subhajit Sengupta, Geoff Macintyre, Salem Malikic, Nilgun Donmez, Dimitri G. Livitz, Marek Cmero, Jonas Demeulemeester, Steven Schumacher, Yu Fan, Xiaotong Yao, Juhee Lee, Matthias Schlesner, Paul C. Boutros, David D. Bowtell, Hongtu Zhu, Gad Getz, Marcin Imielinski, Rameen Beroukhim, S. Cenk Sahinalp, Yuan Ji, Martin Peifer, Florian Markowetz, Ville Mustonen, Ke Yuan, Wenyi Wang, Quaid D. Morris, Paul T. Spellman, David C. Wedge, and Peter Van Loo. The evolutionary history of 2,658 cancers. *Nature*, 578(7793):122–128, 2020.

[125] Sreemanti Basu, Hope M. Campbell, Bonnie N. Dittel, and Avijit Ray. Purification of specific cell population by fluorescence activated cell sorting (FACS). *Journal of Visualized Experiments*, 2010.

[126] Stefan C. Dentro, Ignaty Leshchiner, Kerstin Haase, Maxime Tarabichi, Jeff Wintersinger, Amit G. Deshwar, Kaixian Yu, Yulia Rubanova, Geoff Macintyre, Jonas Demeulemeester, Ignacio Vázquez-García, Kortine Kleinheinz, Dimitri G. Livitz, Salem Malikic, Nilgun Donmez, Subhajit Sengupta, Pavana Anur, Clemency Jolly, Marek Cmero, Daniel Rosebrock, Steven Schumacher, Yu Fan, Matthew Fittall, Ruben M. Drews, Xiaotong Yao, Juhee Lee, Matthias Schlesner, Hongtu Zhu, David J. Adams, Gad Getz, Paul C. Boutros, Marcin Imielinski, Rameen Beroukhim, S. Cenk Sahinalp, Yuan Ji, Martin Peifer, Inigo Martincorena, Florian Markowetz, Ville Mustonen, Ke Yuan, Moritz Gerstung, Paul T. Spellman, Wenyi Wang, Quaid D. Morris, David C. Wedge, Peter Van Loo, and on behalf of the PCAWG Evolution and Heterogeneity Working Groupthe PCAWG consortium., the PCAWG consortium. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *bioRxiv*, 2020.

[127] Keiran M. Raine, Peter Van Loo, David C. Wedge, David Jones, Andrew Menzies, Adam P. Butler, Jon W. Teague, Patrick Tarpey, Serena Nik-Zainal, and Peter J. Campbell. ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Current Protocols in Bioinformatics*, 56(1), 2016.

[128] Andrew Feber, Paul Guilhamon, Matthias Lechner, Tim Fenton, Gareth A. Wilson, Christina Thirlwell, Tiffany J. Morris, Adrienne M. Flanagan, Andrew E. Teschendorff, John D. Kelly, and Stephan Beck. Using high-density DNA methylation arrays to profile copy number alterations. *Genome Biology*, 15(2):R30, 2014.

[129] Wen Wei Liao, Ming Ren Yen, Evaline Ju, Fei Man Hsu, Larry Lam, and Pao Yang Chen. MethGo: A comprehensive tool for analyzing whole-genome bisulfite sequencing data. *BMC Genomics*, 16(12):S11, 2015.

[130] The Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

[131] Andrew Feber, Gareth A. Wilson, Lu Zhang, Nadege Presneau, Bernadine Idowu, Thomas A. Down, Vardhman K. Rakyan, Luke A. Noon, Alison C. Lloyd, Elia Stupka, Vassia Schiza, Andrew E. Teschendorff, Gary P. Schroth, Adrienne Flanagan, and Stephan Beck. Comparative methylome analysis of benign and malignant peripheral nerve sheath tumors. *Genome Research*, 21(4):515–524, 2011.

[132] Dan Zhou, Zhenli Li, Dan Yu, Ledong Wan, Yimin Zhu, Maode Lai, and Dandan Zhang. Polymorphisms involving gain or loss of CpG sites are significantly enriched in trait-associated SNPs. *Oncotarget*, 6(37):39995–40004, 2015.

[133] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[134] Zhifu Sun, Julie Cunningham, Susan Slager, and Jean-Pierre Kocher. Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. *Epigenomics*, 7(5):813–828, 2015.

[135] Yuval Benjamini and Terence P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10):e72–e72, 2012.

[136] Tyrone Ryba, Ichiro Hiratani, Junjie Lu, Mari Itoh, Michael Kulik, Jinfeng Zhang, Thomas C. Schulz, Allan J. Robins, Stephen Dalton, and David M. Gilbert. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Research*, 20(6):761–770, 2010.

[137] Matthew D. Schultz, Yupeng He, John W. Whitaker, Manoj Hariharan, Eran A. Mukamel, Danny Leung, Nisha Rajagopal, Joseph R. Nery, Mark A. Urich, Huaming Chen, Shin Lin, Yiing Lin, Inkyung Jung, Anthony D. Schmitt, Siddarth Selvaraj, Bing Ren, Terrence J. Sejnowski, Wei Wang, and Joseph R. Ecker. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*, 523(7559):212–216, 2015.

[138] Xiaohong Han, Qiaoyun Tan, Sheng Yang, Junling Li, Jianping Xu, Xuezhi Hao, Xingsheng Hu, Puyuan Xing, Yutao Liu, Lin Lin, Lin Gui, Yan Qin, Jianliang Yang, Peng Liu, Xingyuan Wang, Wumin Dai, Dongmei Lin, Hua Lin, and Yuankai Shi. Comprehensive Profiling of Gene Copy Number Alterations Predicts Patient Prognosis in Resected Stages I–III Lung Adenocarcinoma. *Frontiers in Oncology*, 9:556, 2019.

[139] Jared T. Simpson, Rachael E. Workman, P. C. Zuzarte, Matei David, L. J. Dursi, and Winston Timp. Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, 14(4):407–410, 2017.

[140] Elizabeth Larose Cadieux, Miljana Tanić, Gareth A Wilson, Toby Baker, Michelle Dietzen, Pawan Dhami, Heli Vaikkinen, Thomas B K Watkins,

Nnennaya Kanu, Selvaraju Veeriah, Mariam Jamal-Hanjani, Nicholas McGranahan, Andrew Feber, Charles Swanton, Stephan Beck, Jonas Demeulemeester, and Peter Van Loo. Copy number-aware deconvolution of tumor-normal DNA methylation profiles. *bioRxiv*, 2020.

[141] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10, 2011.

[142] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[143] Nelly Olova, Felix Krueger, Simon Andrews, David Oxley, Rebecca V. Berrens, Miguel R. Branco, and Wolf Reik. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biology*, 19(1):33, 2018.

[144] Xiaoqi Zheng, Qian Zhao, Hua-Jun Wu, Wei Li, Haiyun Wang, Clifford A Meyer, Qian Alvin Qin, Han Xu, Chongzhi Zang, Peng Jiang, Fuqiang Li, Yong Hou, Jianxing He, Jun Wang, Jun Wang, Peng Zhang, Yong Zhang, and Xiaole Shirley Liu. MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes. *Genome Biology*, 15(7):419, 2014.

[145] Andrew E. Teschendorff and Martin Widschwendter. Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics*, 28(11):1487–1494, 2012.

[146] Kristi Kerkel, Alexandra Spadola, Eric Yuan, Jolanta Kosek, Le Jiang, Eldad Hod, Kerry Li, Vundavalli V. Murty, Nicole Schupf, Eric Vilain, Mitzi Morris, Fatemeh Haghighi, and Benjamin Tycko. Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nature Genetics*, 40(7):904–908, 2008.

[147] Wolf Reik and Jörn Walter. Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics*, 2(1):21–32, 2001.

[148] Marta Kulis, Simon Heath, Marina Bibikova, Ana C Queirós, Alba Navarro, Guillem Clot, Alejandra Martínez-Trillos, Giancarlo Castellano, Isabelle Brun-Heath, Magda Pinyol, Sergio Barberán-Soler, Panagiotis Papasaikas, Pedro Jares, Sílvia Beà, Daniel Rico, Simone Ecker, Miriam Rubio, Romina Royo, Vincent Ho, Brandy Klotzle, Lluis Hernández, Laura Conde, Mónica López-Guerra, Dolors Colomer, Neus Villamor, Marta Aymerich, María Rozman, Mónica Bayes, Marta Gut, Josep L Gelpí, Modesto Orozco, Jian-Bing Fan, Víctor Quesada, Xose S Puente, David G Pisano, Alfonso Valencia, Armando López-Guillermo, Ivo Gut, Carlos López-Otín, Elías Campo, and José I Martín-Subero. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nature genetics*, 44(11):1236–42, 2012.

[149] Federico Gaiti, Ronan Chaligne, Hongcang Gu, Ryan M. Brand, Steven Kothen-Hill, Rafael C. Schulman, Kirill Grigorev, Davide Risso, Kyu-Tae Kim, Alessandro Pastore, Kevin Y. Huang, Alicia Alonso, Caroline Sheridan, Nathaniel D. Omans, Evan Biederstedt, Kendell Clement, Lili Wang, Joshua A. Felsenfeld, Erica B. Bhavsar, Martin J. Aryee, John N. Allan, Richard Furman, Andreas Gnirke, Catherine J. Wu, Alexander Meissner, and Dan A. Landau. Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature*, 569(7757):576–580, 2019.

[150] Xing Hua, Wei Zhao, Angela C Pesatori, Dario Consonni, Neil E Caporaso, Tongwu Zhang, Bin Zhu, Mingyi Wang, Kristine Jones, Belynda Hicks, Lei Song, Joshua Sampson, David C Wedge, Jianxin Shi, and Maria Teresa Landi. Genetic and epigenetic intratumor heterogeneity impacts prognosis of lung adenocarcinoma. *Nature Communications*, 11(1):2459, 2020.

[151] Yuan Yuan, Yilu Zhou, Yali Li, Charlotte Hill, Rob M. Ewing, Mark G. Jones, Donna E. Davies, Zhenglin Jiang, and Yihua Wang. Deconvolution

of RNA-Seq Analysis of Hyperbaric Oxygen-Treated Mice Lungs Reveals Mesenchymal Cell Subtype Changes. *International Journal of Molecular Sciences*, 21(4):1371, 2020.

[152] Douglas Arneson, Xia Yang, and Kai Wang. MethylResolver—a method for deconvoluting bulk DNA methylation profiles into known and unknown cell contents. *Communications Biology*, 3(1):422, 2020.

[153] Kate D. Sutherland, Ji-Ying Song, Min Chul Kwon, Natalie Proost, John Zevenhoven, and Anton Berns. Multiple cells-of-origin of mutant K-Ras–induced mouse lung adenocarcinoma. *Proceedings of the National Academy of Sciences*, 111(13):4952–4957, 2014.

[154] Carla F. Bender Kim, Erica L. Jackson, Amber E. Woolfenden, Sharon Lawrence, Imran Babar, Sinae Vogel, Denise Crowley, Roderick T. Bronson, and Tyler Jacks. Identification of Bronchioalveolar Stem Cells in Normal Lung and Lung Cancer. *Cell*, 121(6):823–835, 2005.

[155] Kate D. Sutherland and Anton Berns. Cell of origin of lung cancer. *Molecular Oncology*, 4(5):397–403, 2010.

[156] David Robinson. *Introduction to Empirical Bayes: Examples from Baseball Statistics*. Kindle edi edition, 2017.

[157] Emanuele Raineri, Marc Dabad, and Simon Heath. A Note on Exact Differences between Beta Distributions in Genomic (Methylation) Studies. *PLoS ONE*, 9(5):e97349, 2014.

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

[159] Kasper D Hansen, Benjamin Langmead, and Rafael A Irizarry. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13(10):R83, 2012.

[160] Hao Wu, Tianlei Xu, Hao Feng, Li Chen, Ben Li, Bing Yao, Zhaohui Qin, Peng Jin, and Karen N Conneely. Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Research*, 43:141, 2015.

[161] Kristian Cibulskis, Michael S. Lawrence, Scott L. Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S. Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3):213–219, 2013.

[162] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. Cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*, 69(1):7–34, 2019.

[163] Rafael Meza, Clare Meernik, Jihyoun Jeon, and Michele L. Cote. Lung Cancer Incidence Trends by Gender, Race and Histology in the United States, 1973–2010. *PLOS ONE*, 10(3):e0121323, 2015.

[164] Angela Risch and Christoph Plass. Lung cancer epigenetics and genetics. *International Journal of Cancer*, 123(1):1–7, 2008.

[165] The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550, 2014.

[166] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525, 2012.

[167] Malcolm V. Brock, Craig M. Hooker, Emi Ota-Machida, Yu Han, Mingzhou Guo, Stephen Ames, Sabine Glöckner, Steven Piantadosi, Edward Gabrielson, Genevieve Pridham, Kristen Pelosky, Steven A. Belinsky, Stephen C. Yang, Stephen B. Baylin, and James G. Herman. DNA Methylation Markers and Early Recurrence in Stage I Lung Cancer. *New England Journal of Medicine*, 2008.

[168] Dong Sun Kim, Mi Jin Kim, Ji Yun Lee, Young Zoo Kim, Eun Jin Kim, and Jae Yong Park. Aberrant methylation of E-cadherin and H-cadherin genes in nonsmall cell lung cancer and its relation to clinicopathologic features. *Cancer*, 110(12):2785–2792, 2007.

[169] Ximing Tang. Hypermethylation of the Death-Associated Protein (DAP) Kinase Promoter and Aggressiveness in Stage I Non-Small-Cell Lung Cancer. *Journal of the National Cancer Institute*, 92(18):1511–1516, 2000.

[170] T. J. Seng, N. Currey, W. A. Cooper, C-S Lee, C. Chan, L. Horvath, R. L. Sutherland, C. Kennedy, B. McCaughan, and M R J Kohonen-Corish. DLEC1 and MLH1 promoter methylation are associated with poor prognosis in non-small cell lung carcinoma. *British Journal of Cancer*, 99(2):375–382, 2008.

[171] Duk Hwan Kim, Heather H. Nelson, John K. Wiencke, Shichun Zheng, David C. Christiani, John C. Wain, Eugene J. Mark, and Karl T. Kelsey. p16(INK4a) and histology-specific methylation of CpG islands by exposure to tobacco smoke in non-small cell lung cancer. *Cancer research*, 61(8):3419–24, 2001.

[172] Jeffrey A. Tsou, Jeffrey A. Hagen, Catherine L. Carpenter, and Ite A. Laird-Offringa. DNA methylation analysis: a powerful new tool for lung cancer diagnosis. *Oncogene*, 21(35):5450–5461, 2002.

[173] Steven A. Belinsky. Gene-promoter hypermethylation as a biomarker in lung cancer. *Nature Reviews Cancer*, 4(9):707–717, 2004.

[174] Scott M. Langevin, Robert A. Kratzke, and Karl T. Kelsey. Epigenetics of lung cancer. *Translational Research*, 165(1):74–90, 2015.

[175] Duk-Hwan Kim, Heather H. Nelson, John K. Wiencke, David C. Christiani, John C. Wain, Eugene J. Mark, and Karl T. Kelsey. Promoter methylation of DAP-kinase: association with advanced stage in non-small cell lung cancer. *Oncogene*, 20(14):1765–1770, 2001.

[176] Sabine Zöchbauer-Müller, Kwun M. Fong, Arvind K. Virmani, Joseph Geradts, Adi F. Gazdar, and John D. Minna. Aberrant promoter methylation of multiple genes in non-small cell lung cancers. *Cancer Research*, 2001.

[177] Alessandra Bearzatto, Davide Conte, Milo Frattini, Nadia Zaffaroni, Francesca Andriani, Debora Balestra, Luca Tavecchio, Maria Grazia Daidone, and Gabriella Sozzi. p16(INK4A) Hypermethylation detected by fluorescent methylation-specific PCR in plasmas from non-small cell lung cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 8(12):3782–7, 2002.

[178] William A. Palmisano, Kevin P. Crume, Marcie J. Grimes, Sally A. Winters, Minoru Toyota, Manel Esteller, Nancy Joste, Stephen B. Baylin, and Steven A. Belinsky. Aberrant promoter methylation of the transcription factor genes PAX5 alpha and beta in human cancers. *Cancer research*, 63(15):4620–5, 2003.

[179] Kotaro Mizuno, Hirotaka Osada, Hiroyuki Konishi, Yoshio Tatematsu, Yasushi Yatabe, Tetsuya Mitsudomi, Yoshitaka Fujii, and Takashi Takahashi. Aberrant hypermethylation of the CHFR prophase checkpoint gene in human lung cancers. *Oncogene*, 21(15):2328–2333, 2002.

[180] Paul G. Corn. Frequent hypermethylation of the 5′ CpG island of the mitotic stress checkpoint gene Chfr in colorectal and non-small cell lung cancer. *Carcinogenesis*, 24(1):47–51, 2003.

[181] Yoshio Tomizawa, Hironobu Iijima, Taisuke Nomoto, Yasuki Iwasaki, Yoshimi Otani, Satoshi Tsuchiya, Ryusei Saito, Kunio Dobashi, Takashi Nakajima, and Masatomo Mori. Clinicopathological significance of aberrant methylation of RARbeta2 at 3p24, RASSF1A at 3p21.3, and FHIT at 3p14.2 in patients with non-small cell lung cancer. *Lung Cancer*, 46(3):305–312, 2004.

[182] Manel Esteller, Stanley R. Hamilton, Peter C. Burger, Stephen B. Baylin, and James G. Herman. Inactivation of the DNA repair gene O6-methylguanine-DNA methyltransferase by promoter hypermethylation is a common event in primary human neoplasia. *Cancer research*, 59(4):793–7, 1999.

[183] K O Toyooka, Shinichi Toyooka, Arvind K. Virmani, Ubaradka G. Sathyanarayana, David M. Euhus, Michael Gilcrease, John D. Minna, and Adi F. Gazdar. Loss of expression and aberrant methylation of the CDH13 (H-cadherin) gene in breast and lung carcinomas. *Cancer research*, 61(11):4556–60, 2001.

[184] Ubaradka G. Sathyanarayana, Shinichi Toyooka, Asha Padar, Takashi Takahashi, Elizabeth Brambilla, John D. Minna, and Adi F. Gazdar. Epigenetic inactivation of laminin-5-encoding genes in lung cancers. *Clinical cancer research*, 9(7):2665–2672, 2003.

[185] Reinhard Dammann, Chun Li, Jung-Hoon Yoon, Philip L. Chin, Steven Bates, and Gerd P. Pfeifer. Epigenetic inactivation of a RAS association domain family protein from the lung tumour suppressor locus 3p21.3. *Nature Genetics*, 25(3):315–319, 2000.

[186] David G. Burbee, Eva Forgacs, Sabine Zöchbauer-Müller, Latha Shivakumar, Kwun Fong, Boning Gao, Dwight Randle, Masashi Kondo, Arvind Virmani, Scott Bader, Yoshitaka Sekido, Farida Latif, Sara Milchgrub, Shinichi Toyooka, Adi F. Gazdar, Michael I. Lerman, Eugene Zabarovsky, Michael White, and John D. Minna. Epigenetic inactivation of RASSF1A in lung and breast cancers and malignant phenotype suppression. *Journal of the National Cancer Institute*, 93(9):691–9, 2001.

[187] Luke Hesson, Ashraf Dallol, John D. Minna, Eamonn R. Maher, and Farida Latif. NORE1A, a homologue of RASSF1A tumour suppressor gene is inactivated in human cancers. *Oncogene*, 22(6):947–954, 2003.

[188] Cindy A. Eads, K D Danenberg, K Kawakami, L B Saltz, C Blake, D Shibata, P V Danenberg, and P W Laird. MethyLight: a high-throughput assay to measure DNA methylation. *Nucleic acids research*, 28(8):E32, 2000.

[189] Jessica Nordlund, Christofer L. Bäcklin, Per Wahlberg, Stephan Busche, Eva C. Berglund, Maija-Leena Eloranta, Trond Flaegstad, Erik Forestier, Britt-Marie Frost, Arja Harila-Saari, Mats Heyman, Ólafur G. Jónsson, Rolf Larsson, Josefine Palle, Lars Rönnblom, Kjeld Schmiegelow, Daniel Sinnett, Stefan Söderhäll, Tomi Pastinen, Mats G. Gustafsson, Gudmar Lönnerholm, and Ann-Christine Syvänen. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biology*, 14(9):r105, 2013.

[190] Christopher C. Oakes, Marc Seifert, Yassen Assenov, Lei Gu, Martina Przekopowitz, Amy S. Ruppert, Qi Wang, Charles D. Imbusch, Andrius Serva, Sandra D. Koser, David Brocks, Daniel B. Lipka, Olga Bogatyrova, Dieter Weichenhan, Benedikt Brors, Laura Rassenti, Thomas J. Kipps, Daniel Mertens, Marc Zapatka, Peter Lichter, Hartmut Döhner, Ralf Küppers, Thorsten Zenz, Stephan Stilgenbauer, John C. Byrd, and Christoph Plass. DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nature Genetics*, 48(3):253–264, 2016.

[191] David Capper, David T. W. Jones, and Martin Sill *et al.* DNA methylation-based classification of central nervous system tumours. *Nature*, 555(7697):469–474, 2018.

[192] Maria Teresa Landi, Dario Consonni, Melissa Rotunno, Andrew W. Bergen, Alisa M. Goldstein, Jay H. Lubin, Lynn Goldin, Michael Alavanja, Glen Morgan, Amy F. Subar, Ilona Linnoila, Fabrizio Previdi, Massimo Corno, Maurizia Rubagotti, Barbara Marinelli, Benedetta Albetti, Antonio Colombi, Margaret Tucker, Sholom Wacholder, Angela C. Pesatori, Neil E. Caporaso, and Pier Alberto Bertazzi. Environment and Genetics in Lung cancer Etiol-

ogy (EAGLE) study: An integrative population-based case-control study of lung cancer. *BMC Public Health*, 8(1):203, 2008.

[193] Stephen B Baylin and James G Herman. DNA hypermethylation in tumorigenesis: epigenetics joins genetics. *Trends in Genetics*, 16(4):168–174, 2000.

[194] Wei Chen, Jun Wang, Sulai Liu, Shaoqiang Wang, Yuanda Cheng, Wolong Zhou, Chaojun Duan, and Chunfang Zhang. MicroRNA-361-3p suppresses tumor cell proliferation and metastasis by directly targeting SH2B1 in NSCLC. *Journal of Experimental & Clinical Cancer Research*, 35(1):76, 2016.

[195] Aya Shiba-Ishii and Masayuki Noguchi. Aberrant stratifin overexpression is regulated by tumor-associated CpG demethylation in lung adenocarcinoma. *The American Journal of Pathology*, 180(4):1653–1662, 2012.

[196] B. Carvalho, C. Postma, S. Mongera, E. Hopmans, S. Diskin, M. A. Van De Wiel, W. Van Criekinge, O. Thas, A. Matthäi, M. A. Cuesta, J. S.Terhaar Sive Droste, M. Craanen, E. Schröck, B. Ylstra, and G. A. Meijer. Multiple putative oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression. *Gut*, 58(1):79–89, 2009.

[197] Mugahed Abdullah Hasan Albahde, Piao Zhang, Qiuqiang Zhang, Guoqi Li, and Weilin Wang. Upregulated Expression of TUBA1C Predicts Poor Prognosis and Promotes Oncogenesis in Pancreatic Ductal Adenocarcinoma via Regulating the Cell Cycle. *Frontiers in Oncology*, 2020.

[198] Jie Zhang, Shumei Feng, Wenmei Su, Shengbin Bai, Lei Xiao, Lihui Wang, Dafydd G. Thomas, Jules Lin, Rishindra M. Reddy, Philip W. Carrott, William R. Lynch, Andrew C. Chang, David G. Beer, You Min Guo, and Guoan Chen. Overexpression of FAM83H-AS1 indicates poor patient survival and knockdown impairs cell proliferation and invasion via MET/EGFR signaling in lung cancer. *Scientific Reports*, 2017.

[199] Tibor A Rauch, Xueyan Zhong, Xiwei Wu, Melody Wang, Kemp H Kernstine, Zunde Wang, Arthur D Riggs, and Gerd P Pfeifer. High-resolution mapping of DNA hypermethylation and hypomethylation in lung cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 105(1):252–7, 2008.

[200] J. S. Yu, S. Koujak, S. Nagase, C-M Li, T. Su, X. Wang, M. Keniry, L. Memeo, A. Rojtman, M. Mansukhani, H. Hibshoosh, B. Tycko, and R. Parsons. PCDH8, the human homolog of PAPC, is a candidate tumor suppressor of breast cancer. *Oncogene*, 27(34):4657–4665, 2008.

[201] Steven A. Belinsky, Kieu C. Liechty, Frederick D. Gentry, Holly J. Wolf, Justin Rogers, Kieu Vu, Jerry Haney, Tim C. Kennedy, Fred R. Hirsch, York Miller, Wilbur A. Franklin, James G. Herman, Stephen B. Baylin, Paul A. Bunn, and Tim Byers. Promoter hypermethylation of multiple genes in sputum precedes lung cancer incidence in a high-risk cohort. *Cancer research*, 66(6):3338–44, 2006.

[202] Simon Fishilevich, Ron Nudel, Noa Rappaport, Rotem Hadar, Inbar Plaschkes, Tsippi Iny Stein, Naomi Rosen, Asher Kohn, Michal Twik, Marilyn Safran, Doron Lancet, and Dana Cohen. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database : the journal of biological databases and curation*, 2017.

[203] Anne C. Ferguson-Smith. Genomic imprinting: the emergence of an epigenetic paradigm. *Nature Reviews Genetics*, 12(8):565–575, 2011.

[204] Marisa S. Bartolomei and Anne C. Ferguson-Smith. Mammalian Genomic Imprinting. *Cold Spring Harbor Perspectives in Biology*, 3(7):a002592–a002592, 2011.

[205] Katherine S. Pollard, David Serre, Xu Wang, Heng Tao, Elin Grundberg, Thomas J. Hudson, Andrew G. Clark, and Kelly Frazer. A genome-wide ap-

proach to identifying novel-imprinted genes. *Human Genetics*, 122(6):625–634, 2008.

[206] Vitor Onuchic, Eugene Lurie, Ivenise Carrero, Piotr Pawliczek, Ronak Y. Patel, Joel Rozowsky, Timur Galeev, Zhuoyi Huang, Robert C. Altshuler, Zhizhuo Zhang, R. Alan Harris, Cristian Coarfa, Lillian Ashmore, Jessica W. Bertol, Walid D. Fakhouri, Fuli Yu, Manolis Kellis, Mark Gerstein, and Aleksandar Milosavljevic. Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci. *Science*, 361(6409):eaar3146, 2018.

[207] Robert Shoemaker, Jie Deng, Wei Wang, and Kun Zhang. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Research*, 20(7):883–889, 2010.

[208] Fang Fang, Emily Hodges, Antoine Molaro, Matthew Dean, Gregory J. Hannon, and Andrew D. Smith. Genomic landscape of human allele-specific DNA methylation. *Proceedings of the National Academy of Sciences*, 109(19):7332–7337, 2012.

[209] Mathias A. E. Frevel, Stephen J. Sowerby, George B. Petersen, and Anthony E. Reeve. Methylation Sequencing Analysis Refines the Region of H19 Epimutation in Wilms Tumor. *Journal of Biological Chemistry*, 274(41):29331–29340, 1999.

[210] Hengmi Cui, Patrick Onyango, Sheri Brandenburg, Yiqian Wu, Chih Lin Hsieh, and Andrew P. Feinberg. Loss of imprinting in colorectal cancer linked to hypomethylation of H19 and IGF2. *Cancer Research*, 62(22):6442–6446, 2002.

[211] M. Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893–903, 1969.

[212] Li Li, Kiran Kumar Naidu Guturi, Brandon Gautreau, Parasvi S. Patel, Amine Saad, Mayako Morii, Francesca Mateo, Luis Palomero, Haithem Barbour, Antonio Gomez, Deborah Ng, Max Kotlyar, Chiara Pastrello, Hartland W. Jackson, Rama Khokha, Igor Jurisica, El Bachir Affar, Brian Raught, Otto Sanchez, Moulay Alaoui-Jamali, Miguel A. Pujana, Anne Hakem, and Razq Hakem. Ubiquitin ligase RNF8 suppresses Notch signaling to regulate mammary development and tumorigenesis. *Journal of Clinical Investigation*, 128(10):4525–4542, 2018.

[213] Jingyu Kuang, Lu Min, Chuanyang Liu, Si Chen, Changsong Gao, Jiaxin Ma, Xiaomin Wu, Wenying Li, Lei Wu, and Lingyun Zhu. RNF8 Promotes Epithelial–Mesenchymal Transition in Lung Cancer Cells via Stabilization of Slug. *Molecular Cancer Research*, 18(11), 2020.

[214] Tingting Zhou, Fei Yi, Zhuo Wang, Qiqiang Guo, Jingwei Liu, Ning Bai, Xiaoman Li, Xiang Dong, Ling Ren, Liu Cao, and Xiaoyu Song. The Functions of DNA Damage Factor RNF8 in the Pathogenesis and Progression of Cancer. *International Journal of Biological Sciences*, 15(5):909–918, 2019.

[215] Jérôme Galon, Anne Costes, Fatima Sanchez-Cabo, Amos Kirilovsky, Bernhard Mlecnik, Christine Lagorce-Pagès, Marie Tosolini, Matthieu Camus, Anne Berger, Philippe Wind, Franck Zinzindohoué, Patrick Bruneval, Paul Henri Cugnenc, Zlatko Trajanoski, Wolf Herman Fridman, and Franck Pagès. Type, Density, and Location of Immune Cells Within Human Colorectal Tumors Predict Clinical Outcome. *Science*, 313(5795):1960–1964, 2006.

[216] Clemency Jolly and Peter Van Loo. Timing somatic events in the evolution of cancer. *Genome Biology*, 19(1):95, 2018.

[217] Iñigo Martincorena, Joanna C. Fowler, Agnieszka Wabik, Andrew R. J. Lawson, Federico Abascal, Michael W. J. Hall, Alex Cagan, Kasumi Murai, Krishnaa Mahbubani, Michael R. Stratton, Rebecca C. Fitzgerald, Penny A.

Handford, Peter J. Campbell, Kourosh Saeb-Parsy, and Philip H. Jones. Somatic mutant clones colonize the human esophagus with age. *Science*, 362(6417):911–917, 2018.

[218] Katherine B. Chiappinelli, Pamela L. Strissel, Alexis Desrichard, Huili Li, Christine Henke, Benjamin Akman, Alexander Hein, Neal S. Rote, Leslie M. Cope, Alexandra Snyder, Vladimir Makarov, Sadna Buhu, Dennis J. Slamon, Jedd D. Wolchok, Drew M. Pardoll, Matthias W. Beckmann, Cynthia A. Zahnow, Taha Merghoub, Timothy A. Chan, Stephen B. Baylin, and Reiner Strick. Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. *Cell*, 162(5):974–986, 2015.

[219] David Roulois, Helen Loo Yau, Rajat Singhania, Yadong Wang, Arnavaz Danesh, Shu Yi Shen, Han Han, Gangning Liang, Peter A. Jones, Trevor J. Pugh, Catherine O'Brien, and Daniel D. De Carvalho. DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts. *Cell*, 162(5):961–973, 2015.

[220] Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, 9(8):e1003118, 2013.