Development and application of computational methods to study DNA modifications



Gergana Venkova Velikova

Kellogg College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2020

Abstract

The epigenetic modifications of DNA shape cell fate in development, differentiation, and disease. The existing gold standard sequencing technologies for epigenetic DNA modifications are based on sodium bisulfite, which is a harsh chemical treatment resulting in DNA degradation. A novel bisulfite-free and base-resolution sequencing method, TET Assisted Pyridine-borane Sequencing (TAPS) was developed to detect the most abundant DNA modifications. In comparison to bisulfite sequencing (BS), TAPS relies on mild reactions for the detection of modified bases. From a bioinformatics perspective, sodium bisulfite substantially reduces information content, complicating data processing and the detection of genetic variation. In fact, most existing modification calling tools for bisulfite-treated data do not distinguish between modifications and genetic variants, which results in false positives. A computational tool, asTair, was created to process DNA modification sequencing data. It was designed primarily for handling TAPS sequencing output, but also contains functions that are useful for bisulfite sequencing data analyses. It was shown that TAPS has more even coverage than BS while having a comparable conversion rate over CpGs, and is applicable to low input samples. A Deep Neural Network (DNN) model that detects single-nucleotide variants in TAPS- and BS-converted sequencing data was created to enable sensitive modification and variant calling. The algorithm showed precision and recall above 0.9 for classifying variants, modifications and reference positions. The model outperformed available variant callers for whole-genome sequencing and BS data. Applying such a model on real datasets could improve the accuracy of identifying real DNA modifications masked by genetic variation and errors, as around a sixth of all SNPs could be misclassified as modifications.

Contents

List of Abbreviations **vii**

1 Introduction **2**

1.1 Analysis of epigenomic data 4

1.1.1 The chemistry and biological roles of DNA modifications 5

1.1.2 DNA sequencing 14

1.1.3 Introduction to sodium bisulfite treatment 18

1.1.4 Introduction of TAPS chemistry 20

1.2 How mutations arise 23

1.3 Variant calling on DNA 25

1.3.1 Introduction to variant calling methods on genomic data . 27

1.3.2 Introduction to variant calling methods on epigenomic data 28

1.4 Common machine learning methods with applications in the biomedical sciences 28

1.4.1 Historical notes and importance 29

1.4.2 Characteristics of selected ML algorithms 29

1.4.3 Application of ML in the biomedical sciences 31

1.5 Aims . . . 33

2 Computational methods **35**

2.1 Computational environments . . . 35

2.1.1 Hardware environments . . . 35

2.1.2 Software environments . . . 36

2.2 File formats . . . 36

2.3 Reference genomes . . . 37

2.3.1 Human . . . 37

2.3.2 Mouse . . . 37

2.3.3 Spike-ins . . . 38

2.4 In-house sequencing data generation . . . 40

2.5 Genomic annotations . . . 41

2.5.1 Human hg38 . . . 41

2.5.2 Mouse mm9 . . . 41

2.6 Additional data . . . 42

2.6.1 Synthetic data . . . 42

2.6.2 Public datasets . . . 44

2.7 Alignment and pre-processing of sequencing data . . . 44

2.7.1 Whole-genome sequencing data . . . 44

2.7.2 Whole-genome bisulfite sequencing data . . . 45

2.7.3 TET-assisted pyridine borane sequencing data . . . 46

2.8 Modification calling . . . 47

2.9 Variant calling . . . 48

2.9.1 Germline variants . . . 48

2.9.2 Somatic variants 49
2.9.3 Mutational signatures 50
2.9.4 Copy number and structural variation 50
2.10 Computation and statistics 51
2.11 Machine Learning 51
2.12 Downsampling 51
2.13 Quality metrics 52
2.13.1 FastQC 52
2.13.2 Read quality metrics 53
2.14 Visualisation 54
2.14.1 Sequencing reads and genomic tracks 54
2.14.2 Figures generation 54

3 Development of the TAPS modification calling method 55

3.1 Data-driven chemical development of TAPS 56
3.1.1 First whole-genome run 61
3.1.2 Second whole-genome run 64
3.1.3 Third whole-genome run 66
3.1.4 Fourth whole-genome run 69
3.1.5 First test run 74
3.1.6 Second test run 74
3.1.7 Fifth whole-genome run 78
3.1.8 Sequencing runs comparison 81
3.2 Development of the modification calling computational pipeline . 86
3.2.1 Alignment of TAPS data 87

3.2.2 Modification calling in TAPS data . . . 96
3.3 TAPS data analysis of mESC E14 line . . . 97
3.4 Discussion and future directions . . . 105

4 Model development for variant calling on WGBS and TAPS data 112

4.1 Description of the used WGS, TAPS and WGBS datasets . . . 113
4.2 Demodification . . . 126
4.2.1 Feature exploration . . . 128
4.2.2 Machine learning model . . . 141
4.2.3 Neural networks model . . . 145
4.3 Demodification variant discovery . . . 148
4.4 Germline variant discovery . . . 153
4.5 Error analysis . . . 156
4.6 Discussion and future directions . . . 162

5 Somatic variant calling on TAPS and WGBS data 168

5.1 Somatic calling comparisons on WGS data . . . 169
5.2 Base error estimation comparisons on WGS data . . . 172
5.3 Direct somatic calling on TAPS and WGBS data . . . 176
5.4 Deterministic demodification of TAPS and WGBS data . . . 183
5.5 Probabilistic demodification of TAPS and WGBS data . . . 187
5.6 Discussion and future directions . . . 199

6 Germline variant calling on TAPS and WGBS data 204

6.1 Germline variant calling comparison on WGS data . . . 205

6.2 Deterministic demodification of TAPS and WGBS data 207
6.3 Probabilistic demodification of TAPS and WGBS data 212
6.4 Benchmarking against other tools 217
6.4.1 Benchmarking on synthetic data 217
6.4.2 Benchmarking on real data 219
6.5 Discussion and future directions 221

7 Discussion **223**

8 Conclusion **229**

Appendices **232**

I Additional results 233
I Development of the NN model 233
II Additional figures and tables 235
III cfDNA projects 263
I Feature exploration studies 263
II Classification of HCC 268
IV Enrichment-bisulfite method development project 270
V Publications and contributions 276

References **279**

List of Abbreviations

5-caC 5-carboxycytosine

5-fC 5-formylcytosine

5-hmC 5-hydroxymethylcytosine

5-mC 5-methylcytosine

bp base pair

BS bisulfite sequencing

bWGS whole-genome sequencing when matched with WGBS

cfDNA cell-free DNA

CGI CpG island

CNV copy number variation

DHU dihydrouracil

DNA deoxyribonucleic acid

HPLC high-performance liquid chromatography

iid independent and identically distributed

IUPAC International Union of Pure and Applied Chemistry

kb kilobase

LR logistic regression

Mb megabase

ML machine learning

MS/MS tandem mass spectroscopy

NN neural network

NGS next-generation sequencing

PCR polymerase chain reaction

ppm parts per million

RF random forest

RNA ribonucleic acid

SNP single-nucleotide polymorphism

SNV single-nucleotide variant

TAPS TET assisted pyridine-borane sequencing

TET ten-eleven translocation enzymes

tWGS whole-genome sequencing when matched with TAPS

WGBS whole-genome bisulfite sequencing

WGS whole-genome sequencing

Acknowledgements

First of all, I would like to express my gratitude to my supervisor Dr Benjamin Schuster-Böckler, who sincerely guided me through difficult times to become a better researcher. I own to him my initiation in the field of Bioinformatics and my independence during the DPhil. Also, I am grateful for his assistance in writing this thesis.

I also want to thank my other supervisors that have changed during the time of the program—Dr Chun-Xiao Song and Dr Francesco Boccellato, for their help and advice. I am also grateful to Dr Andrew Roth for sharing expertise for the part of my work that included computational models.

I take this opportunity to express how much I am indebted to Dr Robert Gilbert and Dr Sarah De Val for their help and support through my DPhil degree. Without their help, I do not think that would have gotten where I am now—a step from submission of the doctoral thesis.

I am immensely grateful to Dr Markèta Tomkovà, Paulina Siejka-Zielińska, Olena Yavorska, Dr Ying Bi for their help through discussions and problem-solving during these four years.

Often the words of wisdom and understanding come from the outer academic shell and the deep space of the non-academic world only to bring light by acting as cloud dispersers. This is why I would like to thank Dr Carlos Ruiz Puig, Paolo Spingardi, and Georgina Berridge for the exciting professional and personal discussions through these four years. Furthermore, I want to thank other alumni and active members of the Ludwig Institute for Cancer Research, and the Oxford University Language Centre for sharing thoughts and experience.

I also thank my family for their support and overall low expectations towards me (this is a DPhil, not an Sc.D). My mother, Snezhanka, sparked my

love for science, my sister, Stanislava, taught me to write and to multiply before primary school, and my father, Venko, who dared once to throw away a homework of mine. Thank you again for the decades of acceptance and care. Finally, I can hardly describe how much I owe to Stanko, who understood my thesis by diffusion and never got too bored or scared of the characteristics of a typical DPhil student in terms of caring for the insignificant real word outside of the thesis.

Following my thesis defence, I would like to express my gratitude to my assessors Prof Jim Hughes and Dr Tobias Warnecke for their time and effort in reviewing my dissertation, and for the stimulating discussion and advice regarding my doctoral work.

"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk"

attributed to von Neumann

"Some people are so gifted in the art of pie charts and ppt slide shows they could prove Riemann's hypothesis with just three pie charts spread on four slides."

Unknown

"Instruments register only those things they're designed to register. Space still contains infinite unknowns."

Mr Spock, Star Trek, TOS

Chapter 1

Introduction

The rich library collection of life on Earth contains works written in the languages of two nucleic acids: 2-deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) that were first discovered in the nuclei of human leukocytes in 1869 [236], [85]. DNA encodes the information about the building blocks of life for the majority of known organisms (excluding RNA viruses), creating structural and functional diversity that enables its carriers to adapt and survive in dynamically changing environments. The DNA alphabet is a simple one consisting of four important letters (called nucleotides or bases)—adenine (A), cytosine (C), thymine (T), guanine (G), and their metabolites[1]. These bases comprise strings that organise to double helixes and further higher dimensional structures following rules studied though the 20th century. Eucaryotic DNA sequences organise spatially on different levels adhering to local and distal spatial rules of

[1] Throughout the thesis, nucleotides are labelled using the notation recommended by the International Union of Pure and Applied Chemistry (IUPAC). As an addition to the four deoxyribonucleotides and uracil (U), IUPAC labels can show whether the base is a purine (R), pyrimidine (Y), or is unknown (N). Also, the notation can suggest that an ambiguous nucleotide is any of the main three bases. Such an example is marking a nucleotide as H meaning any base except guanine [82].

A-T and C-G base-pairing [62], double-strandedness [365], [359], the nucleosome [253], [182], the 30-nm fibre [103] and higher chromatin organisation [271]. DNA organisation in cells serves not only the purpose of compacting and preserving genetic material but also enables highly controlled modulation and exact timing of cellular processes through epigenetic mechanisms [333]. Epigenetics, as a scientific domain, studies stable changes in gene expression that are not a result of an alteration in the primary alphabet of DNA and can be heritable although very dynamically changing [159], [330]. Epigenetic mechanisms orchestrate complex interdependent processes in organisms such as demethylation of the paternal genome in the zygote [223], X-chromosome inactivation [218], [282], genomic imprinting [225], [328], organismal development [24] and cellular differentiation [173], protection from intracellular parasites [58], [258], [265], repression of gene expression [278], genome stabilisation through silencing of repetitive elements [240], [120], diverse protein interactions and regulation of transcription [237], inflammation [277], and ageing [255]. Epigenetic modulation takes the form of direct DNA modification, or through post-translational changes in histone proteins which help the higher dimensional DNA organisation and packaging within the confined space of the nucleus, nucleosomal positioning, and higher-order chromatin organisation, or through non-coding RNAs [24], [59]. Epigenetics is also thought to be one of the driving forces for multicellularity [115]. Multicellular organisms start their development from one single cell called a zygote that instead of forming a colony of identical cells that contain the same starting genetic material after multiple cycles of cell division, they have hundreds of cell types having the same DNA but different morphology and function. The ability of cells to specialise using sophisticated means of communication, where their form and function are modulated on different levels of signalling, enabled evolutionary complexity which was impossible before.

Epigenetic changes accumulate over the life span hallmarking ageing [52] and encompassing the initiation and progression of malignancies in complex diseases where an interplay between the genome and the environment exists [94]. The list of diseases with links to epigenetics include neurological and neurodegenerative conditions, cardiovascular diseases, autoimmune disorders metabolic disease and obesity and cancer [346], [198], [134], [268], [344], [97]. Epigenetic changes in cancer, for example, are rapid and pervasive—DNA and histone modifications may change within 48 hours of an external stimulus in a cancer cell due to selective pressure [94].

Epigenetic changes occur early in the development of malignancies [167] which makes them prospective early diagnostics tools [297], and due to their treatment potential because of their reversibility [94] and their importance in immunity [90], epigenetic biomarkers, and drug targets are sought [379], [134]. The diverse and crucial roles of epigenetic mechanisms navigating normal and pathological development signify the need for basic and applied research in epigenetics. Epigenomics studies epigenetic changes in their entirety in the organismal genome through biological experimentation and computational methods [96]. A wide range of approaches to study epigenetic modifications relies on sequencing technologies and the availability of computational tools for data analysis [138], [288].

This thesis focuses on methods to study DNA modifications and their potential application on clinical data.

1.1 Analysis of epigenomic data

The section presents the discovery of DNA modifications, their generalised chemistry, and their abundance and functions within genomes with a focus

on mammals. Also, a critical review of laboratory techniques and data analysis approaches to study DNA modifications are discussed. The development of reliable methods to study DNA modifications is pivotal because the information obtained may shed light on complex physiological processes. Furthermore, methods with applicability in the clinic could boost early diagnostics and personalised medicine given that epigenetic changes in DNA occur significantly early in the course of the disease development and may be used as potential treatment targets.

1.1.1 The chemistry and biological roles of DNA modifications

DNA modifications are ubiquitous and show a high level of diversity in the tree of life. Up to date, there are over 50 types of verified DNA modifications (17 on adenines, 11 on cytosines, 11 on guanines, and 15 on thymines) with $\approx$ 90% of these modifications occurring naturally where several DNA modifications are observed only in bacteria and viruses. In contrast, over half of all curated modifications are indicative of DNA damage and malignancies [322]. The search for new modifications, particularly in bacteria and viruses continues, as these modifications can serve as potential drug targets. Technologies such as single-molecule real-time (SMRT) sequencing can be used to study novel modifications [212].

DNA modifications differ significantly in their rarity and function, with the most common DNA modification in eukaryotes being 5-methyl cytosine (5-mC), while in prokaryotes the most abundant modified base is N6-methyladenine (6-mA) [376]. The transfer of a methyl group to cytosine at its 5' position in DNA by DNA methyltransferases using S-adenosyl methionine as a substrate, results in the formation of 5-mC. This process is known as DNA methylation and plays

an essential role in the dynamic control of cellular functions. Methylcytosine was first synthesised in 1904 [125] and discovered by hydrolysis of tuberculinic acid in 1925 [162]. Still, its existence was under debate until later confirmatory studies: from 1948 [144] using paper chromatography and DNA extracted from calf thymus suggesting that a compound chemically similar to cytosine could be 5-methyl cytosine, and from 1950 [373] which showed that small quantities of 5-mC were found in different samples (tested vertebrate tissues had 5-8%, locusts 1% and wheat 33% proportion of methylated to all cytosines). From a phylogenetic perspective, DNA methylation is common in diverse domains of life such as viruses [36], bacteria and archaea [296], plants [135], fungi [29], mammals [39], and other animals [44], [88], and it can occur at other positions in cytosines or in adenines. It was suggested that epigenetic changes might provide plasticity for organismal development given the environment [159]. Also, in a comparative study of methylation roles in vertebrates, the methylation levels appear to be influenced by the environmental conditions, as well as to correlate with body temperature [347].

Although the occurrence of methylated cytosines in DNA was known for almost a century, the interest in its roles and distribution have increased since it was showed that 5-mC plays a regulatory role in mammalian cells silencing particular genes [142], [282]. In a large proportion of eukaryotes, the methylation occurs in cytosine-guanine pairs known as CpGs; however, non-CpG methylation has been observed in mammals, and other vertebrates [106] and is the dominating methylation pattern in plants and fungi [135],[210],[55]. Only around 55% of the methylation events in *Arabidopsis* are in CpG context, and 22-23% are in CHH and CHG contexts [210] where H stands for C, T, or A. There are also differences in terms of methylation levels distribution in cytosine contexts as well as different variability within and between cell types. Also, the event of cy-

tosine methylation in non-CpG contexts appears to decrease the probability for methylation of neighbouring cytosines which is in contrast to CpG methylation patterns where the methylation values of CpGs in close proximity are correlated [360].

The functional and structural importance of CpGs has been a subject of interest since the 1960s when it was found that these dinucleotides occur at a much lower frequency than what is statistically expected by GC content [164][2]. A possible explanation of why the CpG dinucleotides were depleted from the genome except for specific GC-rich stretches of DNA known as CpG islands (CGI) [109] was that 5-mC were mutational hotspots where methylated cytosines were spontaneously deaminated to thymines [83]—an error that cannot be excised by DNA-uracil glycosidase. The initial description of CGIs defined as being of a certain length ($\geq$ 200 bp) and having GC content of $\geq$ 50% and a ratio of observed versus expected CpGs of 0.6 [109]. Although there have been changes to the definition of what is considered a CGI, novel computational approaches still use some characteristics of the original description such as the high GC and CpG density together with certain length expectations [302].

Methylation shows tissue-specific patterns [106] occurring in different regions in terms of CpG density. These regions other than CGIs, which have the highest concentration of CpGs, include CpG shores (located at up to 2 kb distance from the CGI), CpG shelves (located at up to 2 kb distance from the shores), and CpGs more than 4 kb away from a CGI are labelled as open sea [92]. CGIs playing a role in the regulation of transcription are commonly observed in the 5' ends of genes. CGIs are presented in regulatory sequences spanning promoters, enhancers, and insulators. By definition, promoters are genomic regions preceding the transcription start site (TSS) that contain specific sequences

[2]This article suggested that 5-mC may cover a high proportion of the observed CpG dinucleotides.

that are recognised by regulatory proteins, while enhancers are distal sequences that upregulate transcription and insulators block interactions between other regulatory sequences [281]. In the human genome, over 70% of all promoters contain CGIs [302]. CGIs are usually unmethylated (but are methylated during X-chromosome inactivation and imprinting), and it was observed that promoters of housekeeping genes contain CGIs, while CGIs were less common in the promoters of cell-type-specific genes. The lack of CGIs in the promoters of tissue-specific genes allows for modulating and silencing of gene expression in cell- and tissue type-dependent manner [43]. Moreover, CGIs are negatively associated with nucleosome positioning [274], which might be a result of the need for recognition of the promoter by the transcription initiation complex or the destabilisation of the nucleosome [88]. Moreover, CGIs can be grouped into clusters based on their function and distribution [384]. DNA methylation does not occur only at regulatory sequences such as promoters, enhancers, and insulators, but also may occur within gene bodies where it does not act as a repressive mark [163], and in highly repetitive DNA sequences where such sequences are repressed [119]. Higher diversity in the methylation patterns of CpG shores and shelves were observed in comparison to CGIs [92].

DNA methylation is a stable, heritable modification passed down to the daughter cells during mitotic cell division [43]. It is maintained during replication in a semi-conservative manner where the newly synthesised unmethylated DNA strand is modified by a maintenance DNA methyltransferase enzyme (DNMT1) [136]. DNMT1 shows high specificity towards hemimethylated CG positions which are then methylated following the template orientation. Also, new methylation marks can be introduced to DNA sequence with the *de novo* methyltransferases DNMT3a and DNMT3b [65]. The *de novo* methyltransferases play important roles during embryonic development [252], maintenance

of genome methylation [163], repression of repetitive sequences [170], and the transition to a malignant state in tumours [260]. In certain cases of rapid cell division such as early embryonic and neoplastic development, DNA may lose 5-mC due to the passive demethylation pathway when the maintenance DNA methyltransferase cannot mark the novel synthesised strand [251]. Another way passive demethylation can occur is through excision repair of methylated cytosine bases [169]. However, the passive demethylation pathway is not the only possible mechanism to explain the loss of methylation. The other mechanism is the active demethylation pathway where DNA methylation is removed enzymatically. In plants a group of DNA glycosylases such as Demeter (DME), Repressor of silencing 1 (ROS1), Demeter-like 2 (DML2), and Demeter-like 3 (DML3) catalyse the direct removal of methylated cytosines which plays a role in development [205]. In animals the ten-eleven translocation (TET) enzymes catalyse the oxidation of 5-methylcytosine in a stepwise manner to cytosine through 5-hydroxymethylcytosine (5-hmC), then 5-formilcytosine (5-fC), and finally to 5-carboxycytosine (5-caC) [331]. 5-fC and 5-caC are substrates to the thymine DNA glycosylase (TDG) and are removed by the base excision repair (BER) (see Figure 1.1) [157]. The TET enzymes exist in three different forms Tet1, Tet2, Tet3 that differ in their tissue of origin, and substrate specificity [233]. The TET enzymes are α-ketoglutarate (α-KG) and Fe(II)-dependent thus cofactors and substrates such as vitamin C may induce their activity [46], [233]. TET enzymes show substrate specificity oxidising 5-hmC and 5-fC at 4.9-7.6 times lower rates than 5-mC [157].

The function and distribution of 5-mC have been studied extensively for decades, but less is known about 5-mC active demethylation reaction products. 5-hmC was first detected in T2, and T6 bacteriophages in 1952 [374] and later discovered in adult rat tissues in 1972 [263] where an unrealistically high

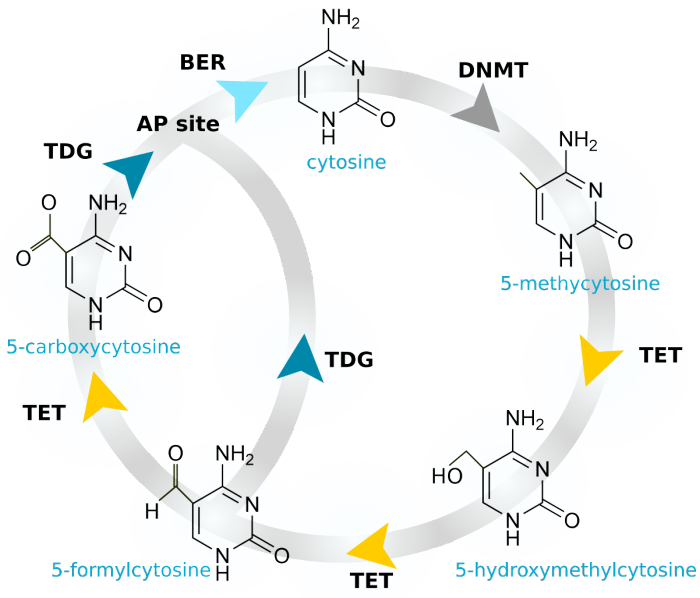


Figure 1.1: The active demethylation pathway. Cytosine is modified at the C5 position by DNA methyltransferases (DNMT) to 5-methylcytosine (5-mC), which is then consecutively oxidised by the ten-eleven translocation (TET) enzymes to cytosine passing through 5-hydroxymethylcytosine (5-hmC), 5-formylcytosine (5-fC), and 5-carboxycytosine (5-caC). Then 5-fC and 5-caC are removed by thymine DNA glycosylase (TDG) leaving an apyrimidinic (AP) site. Finally, the base at the lesion is restored by base excision repair (BER). Based on an illustration from EA Raiber et al., 2017 [272].

abundance of 5-hmC was suggested (15% of all cytosines) but made a later confirmed observation that 5-hmC has a higher concentration in the brain compared to the liver. Almost 40 years later, 5-hmC was rediscovered and confirmed in significant quantities in embryonic and brain cells [331], [184]. 5-hmC has a tissue-specific abundance and has been a subject of major scientific interest in the decades after its rediscovery. A difference between 5-mC and 5-hmC is that the largest differences between tissues in terms of 5-mC are due to differential methylation patters, while for 5-hmC the largest observed difference is its relative abundance [114], [206]. 5-hmC is more common in neural tissues ($\approx 0.6\%$ of cytosines in Purkinje neurons) [184] and embryonic stem cells (ESC) [331]. The

tissue-specific distribution of 5-hmC, together with its stability, has suggested that it is not only an intermediate from the active demethylation pathway of 5-mC but is a regulatory epigenetic mark [114]. 5-hmC was found to play a role in the maintenance of pluripotency [180]. It is enriched at gene bodies and regions surrounding the TSS where it is associated with gene expression [318] and in some repetitive sequences such as short interspersed nuclear elements (SINEs) and long terminal repeats (LTRs) [380]. Furthermore, 5-hmC could inform of the tissue of origin [320], [204].

5-fC and 5-caC have lower abundancies ($\leq$ 20 ppm) and were confirmed as products from enzymatic oxidation of TET [264], [157]. Despite the rarity of 5-fC, it is considered to play a regulatory role in the genome [33] and is found in enhancers and other regulatory regions [319]. The rarest cytosine modification resulting from the active demethylation pathway, 5-caC, is enriched at regulatory sequences such as promoters and enhancers [310] and may interact with RNA Pol II elongation complex [355]. Both 5-fC and 5-caC have been detected in satellite repeats [310]. A high proportion of studies up to date that focused on 5-caC and 5-fC used localisation and correlation of function and presence of the DNA modification. However, the binding activity of transcription factors to symmetric and hemi-modified 5-caC and 5-fC [321], together with the physical interaction between 5-fC and nucleosome-bound histones [273], can be used as a further proof that these rare products of the TET-catalysed active demethylation pathway have regulatory roles in cells.

DNA modifications in cancer

Epigenetic modifications play a double role in cancer development and progression. On the one hand, abnormal epigenetic regulation (particularly DNA modifications) can trigger mutational processes, but on the other hand, DNA

modifications may have a preventive role against cancer development [97], [338]. This section focuses on known DNA modifications' genomic distribution and patterns seen in cancer, and the potential applications of epigenetics for early diagnostics and prognostication of cancer.

The transformation of normal to neoplastic cells includes pervasive changes on multiple levels of regulation of cellular processes. Cancer cells differ from the cells that they were derived from in their abnormal methylation and hydroxylation patterns such as decreased global levels, abnormal repression or permissive marking, and hypervariability [340], [130], aggregation of mutations, chromatin, and genome instability resulting in altered gene expression and cellular signalling [104]. The first interest in the role of DNA epigenetic changes in cancer was after the discovery in 1983 that tumours have abnormal methylation patterns compared to their healthy tissues of origin. The transformation from a pre-malignant to the metastatic state shows a progressive reduction in the global level of methylation [98]. The abnormal DNA modification patterns observed in tumours include lower global methylation levels [206] combined with lower 5-hmC levels [188], and specific hypermethylation of regulatory sequences [97]. The hypermethylation of the promoters of tumour suppressor genes is thought to play a role in cancer initiation [60], and it was shown that changes in DNA modifications precede the occurrence of mutations in cancers [167]. It is thought that epigenetic changes give a selective advantage to tumours to survive in the hostile stromal environment through increased plasticity [338], [104].

The most common single nucleotide substitution of C>T occurs in the cells because of either spontaneous deamination of modified cytosines (5-hmC is more readily deaminated than 5-mC [122]) to thymines or 5-hmUs (as demodification products of 5-mC and 5-hmC respectively) or through the limited availability of S-adenosine methionine suggesting that DNA modifications and particularly

5-mC levels should be positively correlated with the rate of introduction of mutations [389]. Such association may be linked with the observation that malignancies have higher prevalence within older populations while one of the age-related changes in DNA is hypomethylation in specific genomic regions [341]. Usually, the higher methylation occurring in cancer is linked to repressing tumour suppressor genes. Thus, the TET enzymes can be used to restore the unmethylated state [275]. An example of the importance of TET enzymes for normal cellular regulations is that the Tet2 gene is frequently mutated in hematopoietic malignancies [275].

Given the importance of DNA modifications in cancer development, there is an increased interest in epigenetics-based anticancer therapies. There are two approaches for the epigenetic treatment of cancers: through the direct effect of anticancer drugs [87] or modulation of interactions of epigenetically regulated processes with the immune system [140]. The first group of methods includes DNA methyltransferases inhibitors (DNMTi) and inhibitors of histone deacetylases (HDACi) [87]. The application of DNMT inhibitors may lead to apoptosis and repression of cell proliferation and is attributed to the demethylation of regulatory elements [75]. Given the importance of the active demethylation pathway in cancer, the TET enzymes can be used as a therapeutic approach to remove hypermethylation [275]. The interplay of DNA modifications with the immune system may elicit antitumor responses, and aid immunotherapy has sparked interest in the recent years [140]. Immunotherapy of cancers includes therapeutic approaches that increase the natural immune reaction of different classes of innate and adaptive immune cells (primarily T-cells) to recognise and destroy malignant cells [234]. Recently, a link between Tet1 mutations and response to immune checkpoint inhibitors in solid tumours (significantly enriched in non-melanoma skin cancers) was discovered and was suggested as a potential

biomarker for the success of the treatment [370].

1.1.2 DNA sequencing

Although the structure and physical properties of DNA have been known since the 1950s, no methods were available to researchers to study the nucleotide sequence. Chromatographic methods available at the time and used for analytical problems were unable to provide positional information of individual nucleotides [350]. The first effort to decode a nucleic acid code started on RNA because of its shorter sequences, known highly specific RNases, and general lack of complementarity [371]. The first nucleic acid sequence which was read through complete digestion and consecutive search for larger sequence fragments (this method is known as fractionation and has been used before to study amino acid composition) was alanine transfer RNA from *Saccharomyces cerevisiae* and was sequenced in 1965 [141]. In the same year, Sanger et al. used fractionation of RNA radiolabeled with ^{32}P nucleotides combined with electrophoretic methods which enabled the better separation of fragments by molecular weight [298]. The reading of individual nucleotides in the sequence was enabled by 1) the enzyme spleen phosphodiesterase which removes bases consecutively from the 5' terminal end of nucleic acid sequence [279], and 2) electrophoretic properties of labelled nucleotides. After the development of methods to study nucleotide sequences in RNA, they were expanded and applied to DNA. A role for the DNA experimentation was played by studying simpler sequences from bacteriophages T7 [362], λ [372], which had cohesive ends (12 bp GC-rich single-strand overhangs), where initial trials to add radiolabeled nucleotides and deduce sequences were made [371]. Later Gilbert and Maxam sequenced a short DNA sequence that was not at phage ends but was the 27 bases of the lac operator of *E.coli* [113]. Gilbert and Maxam's method was based on B.G. Barrell's fractionation method

described in [32]. The discovery of class II restriction enzymes [316] that cut the DNA at specific 4-6 bp sequences, and the development and applicability of gel electrophoretic methods for separation of molecules [299] helped the progress of the first generation of sequencing methods. The first generation of sequencing methods comprises the *chemical method*, where the DNA sequence is deduced by radiolabeling and cleavage, and the *chain termination method*, which reads DNA sequence by the synthesis of labelled nucleotides. The chemical method by cleavage was created by Maxam and Gilbert [222]. Also, this technique had high sensitivity and was able to produce fragments with lengths over 100 bp. The chain termination method by Sanger, which steps on principles from earlier RNA sequencing method, uses analogues of triphosphates (arabinoside triphosphates and 2,3-dideoxy nucleoside triphosphates). These 2,3-dideoxy nucleoside triphosphates were terminating the synthesis of DNA by blocking the DNA polymerase I [300]. Maxam and Gilbert's method had less applicability than Sanger's method because of its toxicity and speed. At the same time, the less sensitive Sanger sequencing is still used today as a quantitative measure and validation of laboratory methods and results.

One of the main shortcomings of the first generation sequencing methods is their scale which limited their utility for the sequencing of large whole genomes. The ability to clone and amplify DNA fragments before sequencing enabled parallel and high-yield sequencing, which had the potential to scale up to full eucaryotic genomes. These later methods are known as next-generation sequencing (NGS) and are currently dominating genomics and epigenomic research. The dominating NGS technology is Illumina where genomes are fragmented to short sequences that are amplified in a flow cell by cluster formation of molecules generated through sequencing by reversible terminator chemistry where fluorescent labels are used by the sequencing machine to deduce the base call [151], [248].

One of the limitations of the NGS methods is the short read lengths that are complicating *de novo* assembly of genomes, the alignment of lowly mappable repetitive sequences, and the discovery of large structural variation [343]. Furthermore, the need for PCR and cluster amplification of reads introduces biases in the representation of fragments, as well as background errors [30]. Finally, the base calling of fluorescent dyes by the sequencer can be perturbed by genomes with unbalanced compositions [152].

The third generation of sequencing methods includes Single-molecule real-time sequencing (SMRT) and Nanopore sequencing, which can read DNA directly without PCR amplification and to output long reads [343]. The emerging fourth generation of sequencing is expected to preserve the spatial properties of nucleic acids under study [172].

Introduction of sequencing methods in epigenomics

Given the importance of DNA modifications regulating processes in healthy development and ageing, as well as in disease progression, there is a multitude of methods for studying them (see Figure 1.2). These methods differ in their ability to quantify modifications, in their sensitivity, resolution, application, and costs. This subsection gives an overview of existing DNA modification sequencing methods and describes their primary applicability and shortcomings. This section presents only the most essential techniques for biomedical applications.

Laboratory-based quantifications of DNA modifications and their reactions use different types of assays employing the physical, chemical, optical, and spatial properties of the modified bases [266]. The most common method to study DNA methylation is through methylation arrays due to their affordability and coverage of CpGs in regulatory regions [42].

With the advent of NGS, epigenome and particularly methylome analyses

have become accessible, and multiple methods for studying DNA modifications have been developed in the past 20 years. NGS enables both targeted, and whole-genome approaches for studying DNA modifications, and the existing methods are quantitative in nature. The fully quantitative methods provide base-resolution (per nucleotide) information. In contrast, semi-quantitative methods provide information about the position of DNA modifications (at a larger scale), but not about the modification level of individual positions. The first group can be divided into sodium-bisulfite based methods and bisulfite-free methods. Sodium bisulfite deamination of unmethylated cytosines is known since the 1970s [309], [133] when the main interest was in the mutagenic properties of sodium bisulfite. During that time, the requirement for acidic conditions during cytosine deamination was known [308] and was applied in the following studies. In the next decades, the interest in DNA methylation grew, and the focus changed to development and disease.

Sodium bisulfite has been applied for studying DNA with Sanger sequencing two decades later [106]. It was observed in the 1970s that sodium bisulfite is able to deaminate unmodified cytosines in single-stranded DNA [307], which is a property used in sodium bisulfite sequencing methods. The NGS bisulfite method known as methylC-seq was first applied in 2008 on *Arabidopsis thaliana* [210] and mouse [78] and showed promising results in terms of genomic coverage and conversion rates (as high as 99.14%). The main scientific interest in methylation for biomedical application concerns epigenetic modifications on CpGs. For this reason, as a selective approach covering about 2-3% of the whole genome containing high CpG densities and decreasing costs, reduced representation bisulfite sequencing (RRBS), was developed [232]. Due to the inability of MethylC-seq to distinguish between 5-hmC and 5-mC, novel sequencing methods were developed: oxBS that was able to read exclusively 5-mC [49] and

TAB-Seq that distinguishes only 5-hmC [380].

Apart from the bisulfite-based base resolution methods, there are bisulfite-free sequencing technologies. A large group of them are chemical and antibody enrichment methods that selectively amplify regions rich with an epigenetic mark of interest [360], [102], [318], [259]. Although such methods may obtain full genome information at lower costs, they do not provide positional information of the modified cytosines and do not inform of the exact modification level. The second group contains novel bisulfite-free methods that provide an alternative way to study cytosine modifications at base resolution [306], [383], [216], [290].

Finally, the third generation sequencing technologies SMRT and Nanopore can directly detect DNA modifications [105], [181], [377].

1.1.3 Introduction to sodium bisulfite treatment

The treatment of modified cytosine bases with sodium bisulfite is the most common approach to study DNA modifications. Sodium bisulfite does not show preferential activity for a particular methylation pattern [106]. From a chemical point of view, sodium bisulfite reacts with unmethylated cytosines, 5-fC, and 5-caC (but not 5-mC and 5-hmC), deaminating them to uracil, which is read as thymine during PCR [146], [118]. A small proportion of methylated cytosines are known to be deaminated by sodium bisulfite [106]. Due to the harsh chemical conditions of high temperature and acidic pH for an incubation lasting hours, DNA damage occurs in bisulfite-treated DNA primary because of depyrimidination under acidic conditions [335].

There have been significant developments in bisulfite sequencing quality and input due to technological advancements. The recommendation by Illumina is to add ϕX as a calibration control because of its well-studied genome and balanced nucleotide composition [152]. Bisulfite sequencing traditionally required

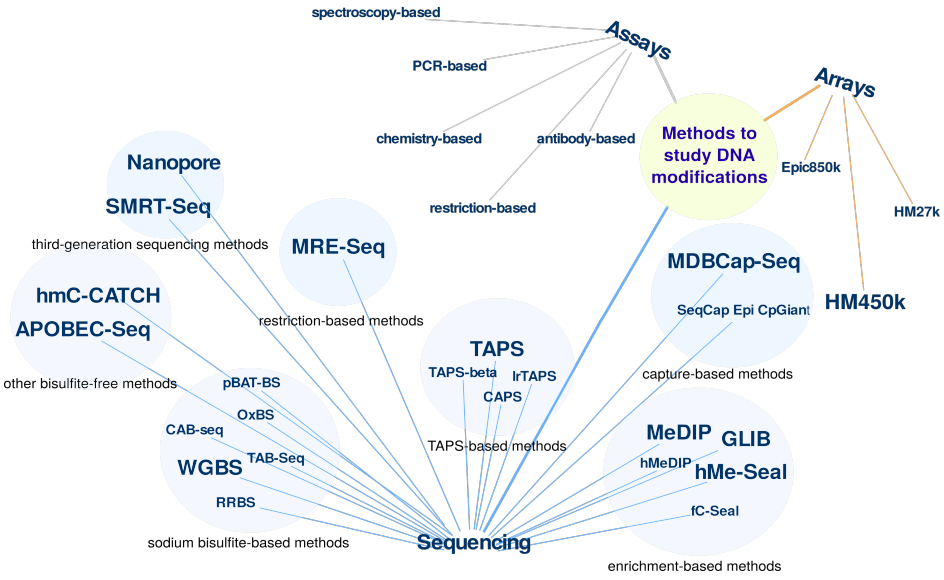


Figure 1.2: A map of major qualitative and quantitative methods for studying DNA modifications (based on Plongthongkum et al, 2014 [266], with additional information from [306], [383], [216], and [214]). DNA modifications can be studied qualitatively with a variety of assays based on the physical properties, chemistry, and affinity of modifications (particularly methylated CpGs), and their methylation levels can be assessed with arrays and sequencing methods. Arrays cover a subset of cytosine positions (usually CpGs) and provide site-specific information about the modification status of a cytosine. Sequencing methods are versatile in terms of targets, chemistry, coverage, and application. Only a selected sample of NGS and third generation sequencing methods are showed.

10-25% unmethylated DNA [16] to be added to sequence due to the insensitivity of earlier sequencing machines that were unable to handle unbalanced DNA compositions. In comparison, newer Illumina machines with an updated base quality score table can sequence bisulfite libraries with only 5% spike-in of ϕX [388]. Furthermore, the choice of an extraction kit can influence the recovery of bisulfite fragments, which reflects on base quality and coverage where higher sequencing depth in regions with lower GC content can be achieved [388]. Also, the improvement of the bisulfite protocol in terms of spike-in [329] and lower input quantities [388] suggest that the method can be used for sensitive

application after modifying the original MethylC-Seq protocol.

Regardless of the advances made in the sequencing of bisulfite-treated data, many challenges remain. From a bioinformatics perspective, sodium bisulfite substantially reduces the information content of the underlying DNA sequences by conversion of over 95% of all cytosines to thymines (see Figure 1.3 and Table 1.1). The loss of sequence complexity complicates the alignment of sequences to the genome, which may result in incorrect positioning of fragments and discarding of unmappable reads. The poor alignment would influence the downstream analyses and the correctness of calling modifications. There are two main approaches to bisulfite alignment. The first one uses a *wild-card* substitution of all cytosines and thymines to the IUPAC symbol Y, thus enabling the permissive alignment of modified DNA sequences [190]. The second approach called the *three-letter* method modifies the reference genome effectively aligning to two genomes (T- and A-converted) depending on the origin and conversion status of the read [186]. The two methods of alignment of bisulfite data suffer from limitations such as a possible increase in the incorrect alignment of reads and can be computationally intensive and time-consuming.

1.1.4 Introduction of TAPS chemistry

TET-assisted pyridine borane sequencing (TAPS) is a novel base-resolution bisulfite-free method to study cytosine modifications that converts modified cytosines to thymines in a multiple-step process [216]. In TAPS cytosine modifications (methylation and all intermediates of the active demethylation pathway) are oxidised to 5-caC by the TET enzyme (see Figure 1.1), followed by a conversion to 5,6-dihydrouracil (DHU or UH_2) by specific borane compounds (pyridine or 2-picoline borane) and DNA fragment amplification where DHU is read as thymine during PCR [211]. Both 2-picoline and pyridine borane can reduce 5-

fC and 5-caC, but leaving the complete demethylation of 5-mC and 5-hmC to 5-caC has a higher efficiency due to the complete oxidation of the TET substrate. DHU is a naturally occurring base resulting from either the reductive pathway of uracil in healthy and neoplastic tissues [247], [345] where the activity of the main reductive enzyme dihydropyrimidine dehydrogenase (DPD) can be elevated or decreased compared to normal in certain malignancies giving DHU the potential to be used as as a biomarker [364], [325], [132], or from DNA damage on cytosines by external factors such as radiation [211]. Also, the ratio of uracil to DHU can be used to evaluate pre-treatment the possible toxicity of 5-fluorouracil, which is used as standard chemotherapeutic drug [18], [378], [235]. Two additional TAPS alternatives TAPS-β (enabling the sequencing only of 5-mC due to the β-glucosyltransferase glucose transfer to 5-hmC positions protecting them from TET oxidation) and chemical-assisted pyridine borane sequencing (CAPS, which allows for the specific reading of 5-hmC after its chemical oxidation to 5-fC by potassium perruthenate $KRuO_4$) are under development. CAPS is methodologically similar to hmC-CATCH [383]. Given the relative abundance of the modified cytosines to all cytosines, it is expected that approximately 1-5% of all cytosines will be converted [43], [272], [245], which is contrary to sodium bisulfite based methods, where conversion occurs on the majority of the cytosine in the genome (≈95-99%) (see Figure 1.3 and Table 1.1). Thus, the library complexity and the DNA alphabet complexity of TAPS is closer to WGS, the proxy of which can be seen as the base quality that remains consistently high between bases. Furthermore, TAPS-based methods have a potential clinical utility given that they convert modified cytosines through milder chemical reactions and incubation conditions (lower temperatures and pH around 5), which leads to less DNA degradation. The input preserving property of TAPS makes it a suitable candidate for studies relying on low-input sam-

ples such as single-cell sequencing (scDNA) and sequencing of cell-free DNA (cfDNA). TAPS showed no bias against longer DNA fragments and was successfully applied to Oxford Nanopore and PacBio Single-Molecule Real-Time (SMRT) sequencing [214]. Furthermore, reduced representation TAPS (rrTAPS) and cfDNA TAPS are currently under development.

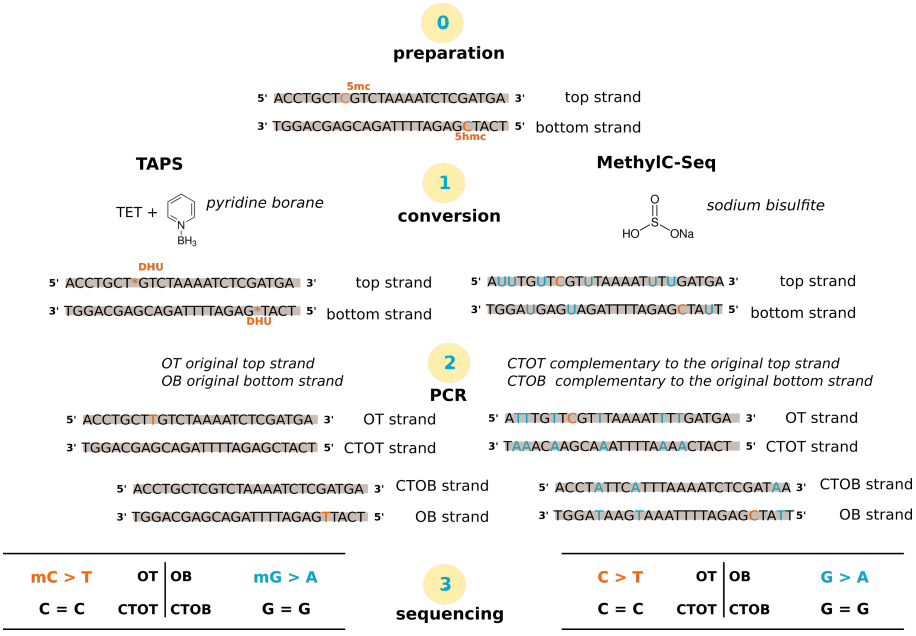


Figure 1.3: This simplified scheme shows the difference between TAPS and WGBS in terms of changes to the reference and the base expectation at cytosine positions. The two methods require three general steps following DNA extraction and library preparation: (1) conversion (here TAPS-treated DNA goes through an enzymatic reaction with TET followed by reduction with pyridine or 2-picoline borane, and WGBS needs only sodium bisulfite), (2) PCR amplification of DNA fragments (this step is generally similar between both methods but can differ in its parameters) and (3) sequencing where WGBS requires the addition of 10-20% non-methylated DNA to sequence the sample. During the conversion step, in TAPS terminally oxidised modified cytosines, 5-caC, are converted to dihydrouridine (DHU), which after PCR is read as thymine. In contrast, sodium bisulfite converts non-modified cytosine to uracil, which is amplified as thymine during PCR.

Technology	5-mC	5-hmC	5-fC	5-caC	C
WGS	C	C	C	C	C
WGBS	C	C	T	T	T
oxBS	C	T	T	T	T
TAB-seq	T	C	T	T	T
TAPS	T	T	T	T	C
TAPS-β	T	C	C	C	C
CAPS	C	T	C	C	C

Table 1.1: Summary of base-resolution high throughput sequencing methods for studying cytosine modifications in DNA when compared to whole-genome sequencing (WGS). Bisulfite-based methods (WGBS, oxBS, and TAB-seq) convert unmodified and unprotected modified cytosines (in oxBS these are 5-hmC, 5-fC, 5-caC, and in TAB-seq—5-mC, 5-fC, and 5-caC) to thymines, while TAPS-based methods (TAPS, TAPS-β, and CAPS) convert only the modified cytosines of interest to thymines. The C>T change of unmodified cytosines that are the majority of cytosines is the reason for the decrease in the proportion of cytosines in bisulfite-treated DNA, which complicates sequencing and data analysis.

1.2 How mutations arise

Cancer is a complex disease resulting from cellular dysregulation on multiple levels as a result of epigenomic and genomic changes accumulated over time. Tumour cells differ from their normal counterparts in biological characteristics known as the *hallmarks* of cancer, that are linked to evading growth suppressors and cell death, increasing proliferation and metastasis, and switching on angiogenesis and proliferative signalling [128]. Cancer, as a disease, has been described in antiquity [79]. Still, the first observation between the link between chromosomal aberrations and abnormal cell division and malignancies was of Boveri in 1914, while the suggestion of the mutational origins of cancer was first recorded in the 1920s [93], followed by a demonstration of the connection between mutations and the development of cancer in the 1980s (radiation and chemical mutagens were known since earlier) [313]. The first and most influential theories describing models of carcinogenesis were the multi-stage model

[31] and the two-stage model [242]. These models explain the transition from a normal to a cancerous cell through a different number of transformative steps, with proliferation and survival being the first key event for cancer development.

Mutations can be divided into driver and passenger (neutral) according to their role in promoting cancer development. Driver mutations trigger the oncogenic transformation by changing the balance of cell proliferation and death. In contrast, passenger mutations accumulate during tumour development but do not act like switches of proliferation and malignancy [351]. Instead, passenger mutations can be used as 'molecular clocks' to study cancer evolution [50]. The driver genes can be further divided into epi-drivers and mut-drivers where the first class is aberrantly expressed in cancer but not frequently mutated in cancer, and the second group has an accumulation of cancer-specific genomic differences [351]. Given the importance of cancer driver genes for the neoplastic development and potential treatment, efforts are made to discover and categorise candidate driver genes in cancers [35].

Given that mutations accumulate over time, there are differences in the number of driver and passenger mutations in pediatric versus adult malignancies [351]. Also, the rate accumulation correlates with the number of cell divisions [351]. In certain situations, the high number of cancer-specific mutations may have delivered a survival benefit to the patient, such as during immunotherapy owing to the expression of neoantigens (antigens specific to the cancer cell) which are recognised by the immune system resulting in the elimination of cancer cells [121], [117], [294].

Cancer can also be described as a result of evolutionary processes on the cellular level where the high diversity of cancer cells ensures their selection and survival [67]. Tumours are composed of both different cell types, and cells with their clones, increasing the diversity in the accumulated genomic differences.

The high level of heterogeneity may complicate treatment and lead to poor patient outcomes [230], while the level of heterogeneity varies considerably between cancer types and stages [199]. Furthermore, there is a link between the development of specific mutational patterns and tumorigenesis. The discovery that the mutational patterns of cancer are a combination of different processes that can be identified and be represented as mutational signatures have enabled a more in-depth understanding of oncogenesis [23]. The mutational processes affecting cells and inducing irreversible changes in the sequence and structure of DNA can be divided into exogenous and endogenous. The first group includes different carcinogenic factors of the environment, such as UV light, ionising radiation, certain chemicals, tobacco smoking, asbestos, and some viruses. At the same time, endogenous mutations are driven by replication and DNA repair, nucleotide deamination, oxidative damage, and DNA modifications [63]. Furthermore, the genetic mutations can be classified into being germline or somatic, and following the Weismann barrier rule, the somatic mutations occur in the body tissues and are mitotically heritable, but in contrast to the germline, somatic variants are not passed by vertical transfer [361].

1.3 Variant calling on DNA

This section presents an overview of mutational processes with a focus on cancer and the computational framework of variant calling. It explains why the existing pipelines created for WGS and exome sequencing data are not suitable for direct application on TAPS and bisulfite converted data.

Variant calling entails the discovery of differences between individual genomes that could be a result of either natural species variation or disease inception and progression (*e.g.* tumorigenesis). The most common approach to variant calling requires whole-genome or targeted sequencing of DNA, followed by alignment

and comparison to a universal (reference) genome. Reference genomes are complete genomic sequences of a certain species usually represented by a single sequence from an individual genome, or a consensus genetic information from different members of the species [219]. However, a single reference genome does not capture natural intraspecies variation which may be present in the form of additional genes in bacteria, or highly repetitive or divergent (*e.g.* the major histocompatibility complex) genomic regions in eukaryotes [311]. In prokaryote genomes, it was showed that the choice of the reference genome may influence coverage and be linked to short read misalignment [342]. This is why efforts to build pan-genomic[3] references are made for prokaryotes [348], plants [38], and humans [311]. The general requirement for the existence of a reference complicates the analyses of a novel or less extensively studied species where such a universal genome is not established. In such cases, variant calling is enabled by *de novo* assembly and comparison of genomes [276]. For samples with a known reference, the raw sequencing data need to be properly reconstructed through alignment to the available reference prior to variant calling. The DNA fragment mapping is a crucial step which may incorporate errors due to misalignment. Depending on the rarity of the variant, the library complexity (and the observed PCR-generated duplicates of single DNA fragments), and the intrinsic error profile of the sequencing machine different approaches can be used for the analysis and the improvement of variant calling [293], [246], [179]. The short sequencing reads generated through NGS, require to be aligned with high precision but their lengths complicate analysis at highly repetitive regions [179]. Strategies to reduce the weight of dubious reads when evaluating a possible variant include removal of PCR duplicates, local realignment, recalibration of base qualities, and filtering of candidate variants [179].

[3]Pan-genome is the collection of genomes that represents the species genetic variation [219],[348].

1.3.1 Introduction to variant calling methods on genomic data

The complete sequencing of the human genome in the early 2000s paved the way for studies of the genetic variation within populations and its meaning for disease development and treatment [84]. The advent of NGS technologies has enabled the cost-effective study of genetic variation using targeted and whole-genome approaches. The most common type of a genetic variant that is considered a germline mutation is the single nucleotide polymorphism (SNP) with a frequency of 1 in 1200-1500 bp [84], and its identification is used to study individual and population differences in disease predisposition and physiology [195]. The discovery of SNPs is highly dependent on the quality of the sequencing data, where either higher sample depths (>20x) or larger cohorts are required [249].

Somatic variant calling concerns the discovery of cell and tissue-specific variants that occur in healthy tissues and malignancies alike, but the primary applications concerns diseases such as cancer [201]. While the prevalent type of mutations in the majority of tumours are single nucleotide substitutions [351], [34], large copy number variation and structural aberrations also mark cancer genomes. These larger scale forms of genetic variation require different kinds of algorithms and models for variant discovery [20]. Usually, the discovery of cancer-specific variants involves the use of a paired tumour and healthy sample, which enables the separation between germline and somatic variants.

A considerable fraction of variant calling algorithms used today uses a probabilistic framework for the discovery and filtering of variants [249]. Standard requirements for data quality and fidelity of alignment of sequencing reads are observed, with a common step of a local realignment of haplotypes, scoring and filtering of the found variants [213], [283].

1.3.2 Introduction to variant calling methods on epigenomic data

In contrast to the genomic NGS data, the most common types of base-resolution epigenomic techniques modify DNA (usually converting cytosines or its modified variants into thymines). These methods introduce another level of complexity for variant calling where the C>T base substitution could be recognised as either a modified base (TAPS) or unmodified base (WGBS). Another complication in epigenomic variant calling is the assumption made by SNP (germline) variant callers of strand complementarity, that is not preserved after chemical conversion in directional libraries (presenting sequencing information only about the original top, OT and the original bottom, OB) [215]. These difficulties explain the existence of a limited number of germline and somatic variant callers for WGBS and RRBS data at present [215], [14], [108], [123]. Furthermore, the application of WGS-specific variant callers on epigenomic data may be suboptimal due to their inability to recognise C>T variants from modifications or reference positions, together with the directionality and the chemical biases introduced during treatment. Also, the filtering algorithms of WGS variant callers may have assumptions about error patterns that differ from the ones observed in WGBS and TAPS.

1.4 Common machine learning methods with applications in the biomedical sciences

The section presents brief historical notes on the development of machine learning (ML) methods, their classification and application with a focus on the biomedical sciences.

1.4.1 Historical notes and importance

Machine learning includes a vast array of techniques for pattern discovery in data and can trace back its origins to 1950s [291]. A bottleneck for the application of ML learning approaches in the 1950-1960s was the lack of powerful computational environments. In the later decades, changes in industrial interests have brought up the application of statistical methods instead of ML. Thus, the popularity of machine learning varied throughout the century. The highest peaks of interest in ML was the 1990s and after 2000s with the rediscovery of deep learning [1]. Deep learning (DL) which as a type of ML tracing its origins back to computational neuroscience in 1943 as a result of the experimentation on artificial neurons used to study how the brain works [224]. Deep learning has been growing in popularity because of the flexibility of its models and the ability to discover patterns in extensive datasets.

1.4.2 Characteristics of selected ML algorithms

There are three main machine learning paradigms to date: supervised, unsupervised, and reinforcement learning. The supervised learning method includes tasks such as classification and regression and requires as input data labels. The unsupervised methods, such as clustering, recognise patterns in the data without additional input information. In the reinforcement learning paradigm, the algorithm learns by trial and error through reward maximisation. A typical machine learning pipeline requires the cleaning and pre-processing of data, which is then split into training and testing set (commonly a validation set is used as well). One of the most critical steps during data pre-processing and model building is feature selection where specific properties of the data are used to solve the problem at hand (*e.g.* improve class separability in classification tasks). These datasets play different roles in the model development, where the model learns

specific characteristics of the data on the training set, followed by optimisation and tuning of parameters tested on the validation set. The performance of the complete model is then evaluated against on testing set. A common problem is the insufficient generalisation ability of the model that may be a result of overfitting to the noise of the training data, or the unrepresentativeness of the testing data. Overfitting can be prevented through direct simplification the model used and regularisation which penalises complex models [45].

Logistic regression Logistic regression is among the most commonly used ML algorithms for binary classification due to its simplicity, and certain alterations to the model can enable classification of multiple classes through multinomial regression [143].

Random forests Random forests (RF) are an example of an ensemble method where the final performance of the algorithm depends on the voting of multiple decision trees classifiers or regressors. The decision trees as algorithms represent hierarchical models of conditions and decisions and are commonly used in predictive modelling applications [289]. The predictive power of random forests lies in the building of decision forests through multiple decision trees where the improved performance is explained by the pseudorandom selection of features [139]. This step increases the variability of trees and improves performance. Due to their interpretability based on the ranking of features by importance RF are a commonly used approach for biological applications [269].

Neural networks Due to their principal similarity to biological neurons, artificial neural networks (ANN or NN), organise in layers with individual connections between the units and have activation functions that are the output of neurons after computation which are then input into the next layer. There are three functionally distinct categories of layers that are the input layer, the

hidden layers and the output layer. The input layer takes in the training data, while the output layer gives the result of classification or regression. The hidden layers are composed of connected units that participate in the learning process and frequently the larger number of hidden layers (resulting in a deep neural network, DNN) can improve model performance. At the same time, caution should be taken to avoid overfitting of models with multiple hidden layers. The learning process is manifested as a forward propagation to evaluate the value of a target function, and backpropagation to minimise the error of learning by modifying the weights of neural connections. Neural networks differ in their architectures with the simplest being the fully connected feed-forward network where the layers are linearly connected [45].

1.4.3 Application of ML in the biomedical sciences

Machine learning approaches have been successfully used for a diverse range of problems in the natural sciences and medicine due to their ability to discover patterns in data. Common problems with a clinical application that are addressed successfully with machine learning algorithms include drug discovery and repurposing [56], stratification of patients [382], diagnosis and prognostication [89]. Genomic and epigenomic variant and modification calling algorithms may output data that can be used for training of machine learning models. Such models can be applied for diagnosis and prognostication of disease. Furthermore, machine learning approaches are applied successfully for variant calling showing high accuracy of performance [267], [73].

However, common limitations of the application of machine learning methods for biomedical problems are data availability, the lack of correct labelling for supervised learning tasks, and the quality of data. Often in biological studies, the sample size can consist of fewer samples than the number of features while

the recommended ratio for training a stable non-overfitting model is five to ten times fewer features than the data samples [45].

1.5 Aims

DNA modifications play essential roles in regulating the normal and pathological physiological state of organisms. In cancer, the DNA modification and mutational processes are intertwined where particular methylation patterns may correlate with higher or lower mutation rates, as well as the highly mutated genomes show aberrant DNA modifications. The most commonly used methods to study DNA modifications at present suffer from at least one type of limitation from the following: 1) limited coverage, 2) not providing base resolution information, 3) not applicable to 'precious' samples where the available DNA is of a very low quantity or is of rare origin (patient samples, ancient samples). Until now, the most extensive datasets with DNA modification information used array data that does not provide complete genomic information. For complex diseases like cancer, the ability to mine full genomes (CpG and non-CpG modification and genetic variants) could provide insights into disease mechanisms and treatment which were not available until now given the limitations of the methods to study DNA modifications. The current practical limitations of studying DNA modifications could hypothetically be resolved by a sensitive base-resolution method that can be applied on low input samples and allows for simultaneous variant calling. The novel chemical method to study DNA modifications developed and tested as part of this thesis is TAPS which was presented in the introduction.

The main objectives of this thesis were studying novel epigenomic data and developing computational tools and pipelines for the analysis of the generated data. The present work aimed at exploring potential issues and biases of the new sequencing DNA method TAPS, and studying the usability of the tools on real data and discusses its applicability for biomedical research and clinical studies. By creating novel computational methods to study DNA modifications and iden-

tifying potential biases while analysing data, contributions were made towards the bioinformatic community who would be able to use these methods. The motivation behind this project was that being able to call both modifications and mutations from the same dataset could not only decrease the application costs, but could provide a specific 2D code, where both mutational and modification patterns are distinguishable, and this information can be used for better tissue classification and for studying diseases. The first aim to develop computational methods to analyse TAPS data was focused on developing a modification calling tool and creating a TAPS data analysis pipeline (chapter 3), and on building a classifier which is able to distinguish between cytosine modifications after TAPS conversion and germline variants (SNP) (chapter 4). In addition to the main aim, TAPS-specific biases in comparison with the gold standard method for studying cytosine modifications WGBS and standard WGS were explored with a focus on applications in chapters 3, 5, and 6.

Chapter 2

Computational methods

This chapter gives an overview of the main tools, environments, and datasets used for the analyses included in this thesis.

2.1 Computational environments

2.1.1 Hardware environments

The TAPS method development analyses and historical data analyses were conducted on the Ludwig Institute for Cancer Research high-performance computing cluster (Cortex). All of the variant calling experiments and the larger part of the final TAPS development analysis were run on the Biomedical Research Computing (BMRC) high-performance cluster that is part of the Big Data Institute. Part of the machine learning training sessions and experiments were performed using NVidia GPU-accelerated servers (with Tesla P100 GPUs) for their acceleration and reliability [11].

2.1.2 Software environments

The primary environment used for the installation of specific Python versions and bioinformatic tools was Conda [5] because of its ease of use and a wide range of packages. The main reason for using Conda and Python environments was the encapsulation of the required packages and their dependencies enabling running them in isolation, thus preventing possible clashes due to changes between versions of dependencies outside of the environment.

2.2 File formats

This section gives a short overview of the data file formats used and mentioned throughout the thesis without providing an in-depth introduction. The most common formats used in the bioinformatic analyses were: 1) annotation formats usually in the form of text tab-delimited files containing genomic locations and other features, 2) alignment formats such as uncompressed and BGZIP compressed tab-delimited files containing aligned sequencing reads, and 3) raw data formats including files containing single or multi-column strings per line [238]. In the first group, the files have zero-based coordinates (BED, bedGraph, mods, cov) or 1-based chromosome location (GTF, VCF, BCF). The VCF or BCF format [295] is commonly used to store detailed genetic variant information. It is the default output of many variant calling tools, while BED, bedGraph and GTF files contain genomic positions and their annotations. The second group describes the SAM format [203] with its compressed alternatives BAM and CRAM (columnar), which have complete information about the sequencing reads and their alignment and allow for rapid random access of data. The third group includes raw sequencing reads in FASTQ format and genomic references in FASTA format.

2.3 Reference genomes

The genomic sequence was combined with the spike-in sequences in a reference FASTA file used for alignment to ensure better alignability of reads. During the TAPS method development phase, a spike-in only reference was used separately before the mapping to the combined full genome and spike-ins reference to speed up the development cycle.

2.3.1 Human

For the few cfDNA samples prepared during the TAPS method development, the canonical hg19 was used as a reference [196]. During the TAPS variant calling development, the human reference genome that was used for alignment was hg38 [196]. The initial batch used for TAPS variant calling development was aligned to a reference prepared by Dr Benjamin Schuster-Böckler, where ALT sequences were removed, and chromosomal sequences were organised alphanumerically. For the rest of the analysis, the release hg38.p13 was used for TAPS, WGBS, and WGS. The original sequence contained spaces within the sequence names that were replaced with underscores and slight changes to the naming: 1) renaming of the main chromosomes to 'chr' and a number or an uppercase letter, 2) removing repetitive descriptions such as 'unlocalised genomic scaffold, GRCh38.p13 Primary Assembly', 'Homo sapiens unplaced genomic scaffold, GRCh38.p13 Primary Assembly' and 'genomic patch of type FIX/Novel, GRCh38.p13 PATCHES', before 3) sorting the sequences (so that the canonical chromosomes would come first) and outputting the genomic reference.

2.3.2 Mouse

For TAPS method development and modification calling the canonical mm9 reference was used [358].

2.3.3 Spike-ins

Enterobacteria phage λ

The DNA sequence of Enterobacteria phage λ (J02459.1) was downloaded from GenBank [303] and was used as a proxy for estimating the conversion and failed conversion rates in TAPS and WGBS data. The approach for measuring the incomplete conversion of bisulfite-treated DNA on CpH positions (that are expected to be unmodified) is commonly used and can be traced since the early 2000s [314]. The method may give poor estimates in sequences with real modification on CpH positions such as DNA from plants or ES cells. In bisulfite-treated DNA sequences, the incomplete conversion can inflate the rate of false positives, by leaving out unconverted cytosines which will be detected as modified. For TAPS the λ phage DNA was methylated on CpG positions with M.SssI methyltransferase, and conversion was assessed on them. In TAPS the incomplete conversion inflates the false-negative rate where modified cytosines that should be converted to thymines are not. A difference between TAPS and WGBS in using bacteriophage λ to assess the success of conversion is that in TAPS, the observed conversion rate also depends on the enzymatic methylation preceding conversion.

ϕX174 bacteriophage

ϕX174 bacteriophage (shortened as ϕX) is used as quality control by Illumina because of its small genome and balanced base composition [152]. The balanced composition is essential for cluster formation and error evaluation during sequencing, and it is recommended that balanced libraries include 1% ϕX (quality control v3), whereas highly C-G or A-T imbalanced libraries such as WGBS and the genomes of some prokaryotes require up to $\geq$ 40% spike-in [152], [16]

(newer Illumina sequencing machines can handle better unbalanced libraries and require less ϕX input [388]). Given the ubiquity of ϕX174 bacteriophage DNA, its sequence was downloaded from Illumina iGENOMES and included in the alignment reference FASTA.

Other viral genomes

The reference sequences for common human viruses: Human papillomavirus (HPV, GenBank accession number NC_027779.1), Hepatitis B virus (HBV, GenBank accession number LT717702.1), Hepatitis C virus (HCV, GenBank accession number JX227964.1) and Epstein-Barr virus (EBV, GenBank accession number AJ507799.2) were included to the human reference to increase the chance that dubious reads will not align to the human genome, and possible real viral DNA would be able to align. No microbiome-specific analyses were conducted during TAPS modification calling and variant discovery methods.

In-house generated reference sequences

All of the presented spike-ins in this section were created and tested in Dr Chun-Xiao Song's lab (more details can be found in the supplementary notes of the TAPS method paper [216]).

120-mers

There were two short spike-ins—120mer-1 and 120mer-2 made by primer extension. These spike-ins had methylated and hydroxymethylated positions and were used to evaluate the bias of the method towards specific cytosine modifications.

160-mer

This NCNN spike-in was generated with the annealing of two oligonucleotide sequences carrying different cytosine modifications in equal quantities and ex-

tension with Klenow Polymerase and dNTPs and 5 U. The sequence is short at 160 bp but has two highly diverse positions: 51 (N5-mCNN, methylated in 5'-3' direction) and 111 (N5-hmCNN, hydroxymethylated in 5'-3' direction). The spike-in was used to evaluate TAPS efficiency on CpH contexts and compare methylation and hydroxymethylation.

2 kb-mers

During the development of the TAPS method, three distinct 2 kb spike-ins were used. Initially, the sequences labelled 1 (methylated at GpC positions by M.CviPI methyltransferase, 2005 bp), 2 (hydroxymethylated at CpGs in sequencing runs 1 and 2, later methylated at CpG, 2026 bp), 3 (unmodified 2018 bp) were used to evaluate different modification contexts, but after the NCNN was used for this purpose, and only spike-in 3 in its unmodified state was used to measure the false-positive conversion from the TAPS method. They were prepared by PCR amplification from a pNIC28-Bsa4 plasmid.

2.4 In-house sequencing data generation

The TAPS and WGBS data generated for the TAPS development, and the TAPS and WGS of clinical samples of patients with oesophageal adenocarcinoma (OAC) and Barrett's oesophagus were prepared by Paulina Siejka-Zielińska and other lab members from Dr Chun-Xiao Song's lab. The WGBS and its matching WGS samples of Barrett's oesophagus patients were prepared earlier (2015-2017) within the Ludwig Institute for Cancer Research. The preliminary validation step of the applicability of TAPS on clinical samples used two tissue samples, which were from Barrett's oesophagus patient and a fetal oesophagus tissue. In addition to them, two WGS samples from a different biopsy from the same

Barrett's oesophagus patient and a matched duodenum were sequenced. These pilot WGS and TAPS samples were sequenced at ≈ 20x mean depth, while later TAPS and WGBS samples had sequencing coverages of ≈ 30x on healthy patient-matched tissues and ≈ 45-60x for malignant tissues. The sequencing machines used for the in-house data analysed in this thesis are described in Table 2.1.

Study	Sequencing type	Sequencing machine	Sequencing depth	Read length	Adapters	Organisms
TAPS method development	TAPS, WGBS	NextSeq	1-20x	75-80 bp	Nextera	mouse, human
Barrett's oesophagus study	WGS, WGBS	HiSeq	35-60x (WGS), 65-75x (WGBS)	125 bp	Illumina	human
Barrett's oesophagus and OAC studies	WGS, TAPS	NovaSeq	20x (pilot), 30x, 60x	151 bp	Illumina	human

Table 2.1: Summary of the sequencing runs produced in-house that were used in the thesis. The samples were presented in greater detail within chapters 3, 4, 5, and 6.

2.5 Genomic annotations

2.5.1 Human hg38

Information about functional regions in the human genome (gene annotations, known variation, regulatory sequences) was downloaded from UCSC genomic browser [189] and GENCODE [131]. As mappability estimates for the WGS and WGBS data, UMAP and BISMAP tables were used [168].

2.5.2 Mouse mm9

Similarly to the hg38 genome information, the functional regions and mappability tracks for mouse were downloaded from the same sources listed above. Additionally, the germline variants (SNPs) for the mESC cell line E14 was downloaded from [153], and information about the mESC chromatin states was ob-

tained from [48]. Genomic regions with having higher rates of mapping artefacts were excluded from the analysis [25] (downloaded from https://sites.google.com/site/anshulkundaje/projects/blacklists).

2.6 Additional data

2.6.1 Synthetic data

Alignment experiments For the study of TAPS alignment, synthetic pair-end sequencing reads were created through ART [145] using the parameters -p -ss NS50 –errfree –minQ 15 -k 0 -nf 0 -l 75 -c 270000 -m 240 -s 0 -ir 0 -ir2 0 -dr 0 -dr2 0 -sam -rs 10 for a sequence in mm9 chromosome 4 (genomic coordinates chr4:41000000-43000000). The reads without sequencing errors and indels were then used directly by conversion of ART's SAM output to BAM. These synthetic reads were subsequently TAPS, or WGBS modified at CpG positions (with proportions of fully modified positions of 0.0, 0.2, 0.5, 0.8, 1.0) using asTair simulate. Their modification levels were compared to the expected by modification calling and visualised. The reads were then converted to FASTQ using bedtools bamtofastq [270] after sorting the BAM files by read names and realigned with different WGS- and WGBS-specific aligners.

The alignment correctness was assessed by the proximity and exact matching to the read coordinates after alignment to the read coordinates in the original SAM file output by ART, and the modification level correctness was assessed as being within a margin of error of the expected modification at the position. The assessment of the error in the position and modification level after realignment were done through a custom Python script. The unmodified realigned reads we used to measure the alignment error without methylation due to randomness and the sequence composition of the genomic region.

Performance experiments For testing of asTair v2 call performance versus Bismark, synthetic reads were generated over the bacteriophage λ (with parameters as previously described with the exception of the coverage where -c 1000000 reads were created). In this experiment, the proportion of CpGs that were used to convert fully in WGBS- and TAPS-like manner was 50%. The original processing pipeline described was used to prepare the sequencing reads for modification calling. The performance comparison of modification calling with asTair and Bismark was measured with Linux command *time*, and the analysis steps were conducted using a single thread on one Intel Xeon CPU with 250 GB of memory (as described in the methods of [216]).

SNP calling validation Synthetic data were simulated with ART to resemble high-quality Illumina TruSeq data with 150 bp pair-end reads and mean coverage of 30x using chr21 from hg38 genome build. Then one million SNPs and indels were introduced with VarSim [244]. After that asTair simulate was used to modify the data in TAPS and WGBS manner at a subset of the CpG positions. Reads were then realigned with astair align –method CtoT for the WGBS-like data, astair align –method mCtoT for the TAPS-like data and BWA mem for the unmodified synthetic data. Strelka2 germline, Octopus germline, GATK HaploCaller, and the NN model for SNP discovery with selected features were used on the TAPS and WGS data. For WGBS, bisulfite-specific germline callers: BS-SNPer [108], Biscuit [14], CGmapTools [123] and Bis-SNP [215] were applied. These tools were used together with the described previously WGS germline callers, and the NN selected model for SNP calling on WGBS data. Samtools mpileup was applied on the data to evaluate any possible differences from the list of variants given from VarSim due to alignment. The list of positions with evidence for alternative alleles was intersected with the truth set for any of the datasets.

2.6.2 Public datasets

For the TAPS development experiments, E14 WGBS raw data from Habibi et al. 2013 paper [126] (NCBI GEO SuperSeries accession number GSE41923), and high-depth sequencing coverage WGS data [27] with (NCBI Sequence Read Archive accession number SRS635099) were downloaded and processed for comparison to TAPS.

2.7 Alignment and pre-processing of sequencing data

The short reads obtained by sequencing were quality and adapter trimmed using TrimGalore! (version 0.4.4) [186] which is a wrapper for cutadapt (version 2.3) [220] that simplifies trimming and enables specific bisulfite sequencing removal of low-quality bases and short reads, with parameters –gzip –length 35 –paired to do Illumina sequencing specific adapter trimming and to keep a gzip-ped version of pair-end reads with sequencing quality of at least 20 and length of at least 35 bp.

2.7.1 Whole-genome sequencing data

The WGS data used for variant calling was processed using standard approaches summarised in Figure 2.1. The quality and adapter trimmed sequencing reads were aligned to the reference genome using the default pair-end settings of BWA (versions 0.7.12 and 0.7.17) [202]. The sorting per genomic position and indexing of the BAM files was done using samtools (version 1.3.1 or 1.8) [203] as a preliminary step to the removal of PCR and optical duplicates using the MarkDuplicates tool of the GATK suite (version 4.1.4.0) [227]. In order to do

haplotyping and variant calling on WGS, the deduplicated reads have to be pre-processed further using the GATK suite (version 4.1.4.1). Base recalibration was conducted with the BaseRecalibrator function and providing dbsnp_146 information as –known_sites, and then the base recalibration table is applied using ApplyBQSR. The base recalibration procedure empirically evaluates the systematic bias of the sequencing machine using a probabilistic model, and the available sequencing reads data. The primary assumption of the method is that all variation outside of the provided known SNP positions can be treated as an error, and the base quality scores should be modified accordingly.

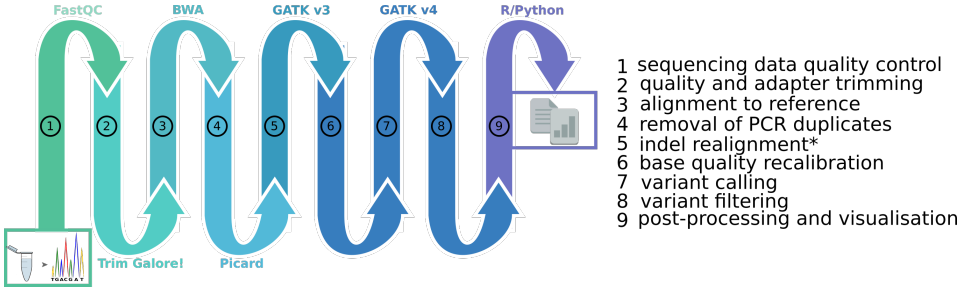


Figure 2.1: A schematic of the generalised WGS variant calling pipeline used. The quality of sequencing data was assessed (1), followed by quality and adapter trimming of reads (2) and their alignment (3). Then PCR and optical duplicates were removed (4) and the base qualities (BQ, Phred scores) were recalibrated to account for sequencing biases (6), while the need for indel realignment (5) was tested and found to be unnecessary due to tool updates. The data pre-processing allowed for variant calling (7), and quality filtering of genomic variants (8). Finally, statistical analyses were run on the data and the results were visualised (9). Drawn with diagrams.net (open-source) [6].

2.7.2 Whole-genome bisulfite sequencing data

Bismark

The WGBS data were processed using Bismark (version 0.19) [185] using the –no_overlap option. Bismark aligned to two different sets of the genome given the conversion and the read orientation against the reference. After alignment,

reads with possible non-conversion at CpH positions were removed using the 'three-C' filter, which removes sequencing reads where three or more cytosines in non-CpG context appear methylated. The original experiments used only the alignment option of Bismark that wraps around the short-read aligner Bowtie2 [197], while at a later stage a more sensitive aligner was used and tested on bisulfite data (HISAT [174], [175]).

BSMAP

BSMAP (version 2.6) is the most commonly used 'wild-card' aligner for the analysis and processing of bisulfite sequencing data [375]. It used the IUPAC symbol for either cytosine or thymine Y and was applied on both TAPS and WGBS data.

asTair

asTair is a tool for alignment and processing of TAPS and WGBS data developed for the needs of TAPS analysis by us. When applied on WGBs data, the reads were aligned to the reference genome using the default pair-end settings of astair align –method CtoT which wraps around bwa-meth.py (version 0.2.2) and ensures alignment of WGBS data that is comparable to TAPS and WGS.

2.7.3 TET-assisted pyridine borane sequencing data

In order to align TAPS sequencing reads asTair align –method mCtoT was used. AsTair align wraps around bwa (version 0.7.12).

2.8 Modification calling

DNA modification levels or β values in TAPS (also conversion) were calculated using the formula 2.1 and in WGBS—using the ratio described as equation 2.2.

$$\beta = \frac{\text{modified cytosines}}{(\text{modified cytosines} + \text{unmodified cytosines})} \tag{2.1}$$

$$\beta = \frac{\text{unmodified cytosines}}{(\text{modified cytosines} + \text{unmodified cytosines})} \tag{2.2}$$

The cytosine modification calling using Bismark was conducted only on CpG positions using bismark_methylation_extractor in pair-end mode after the reads in the BAM file were sorted by read name using samtools sort -n option. For the TAPS initial method development analyses and the Nature Biotechnology publication (asTair versions 1 and 2), the reads were split into coming from the original top strand (OT) and the original bottom strand (OB), followed by the removal of PCR duplicates with Picard MarkDuplicates as explained above. The final pre-processing step was the removal of any overlap between the reads in the read pair was clipped based on higher base quality using BamUtil clipOverlap [165]. Then pileups per positions of deduplicated non-overlapping reads were taken using samtools mpileup -v -t DP,AD,ADF,ADR -x options. The VCF file was then read, and modification levels were calculated per positions. In the asTair version 3, which was used for the variant calling in TAPS part of the thesis, all filtering and calling steps were done internally after the removal of PCR duplicates. In the TAPS case, the CpG, CHG, and CHH contexts were obtained together unless stated otherwise.

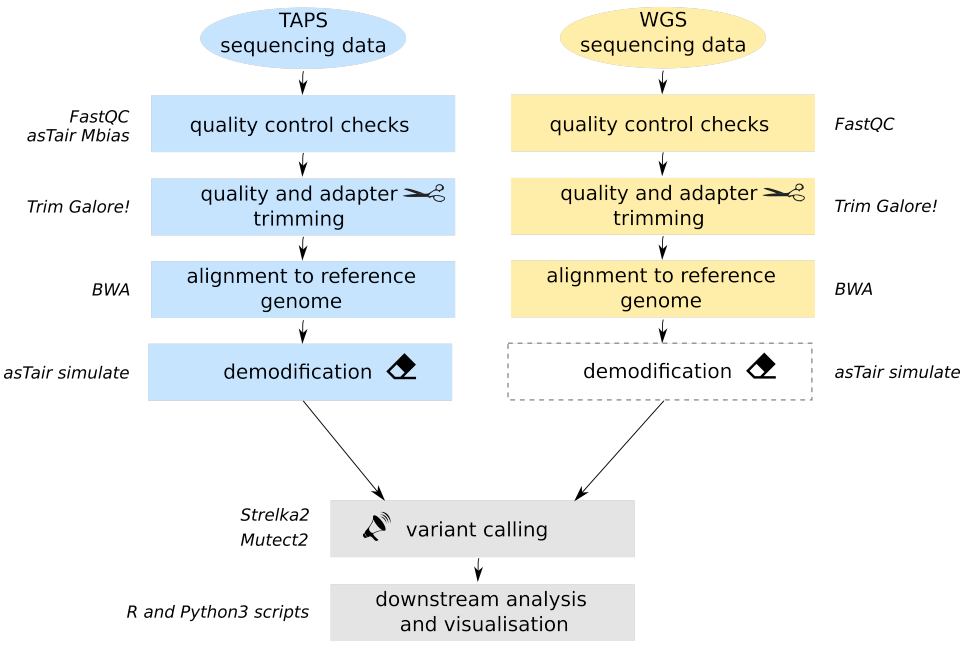


Figure 2.2: Summary of the generalised variant calling pipeline used for TAPS (applicable to WGBS with a modification of the aligner to bwa-meth) and WGS. TAPS, WGS, and WGBS, were processed as described in the methods (quality control, trimming of reads, alignment, and removal of PCR duplicates). Conversion markings in TAPS and WGBS were removed in a process called demodification before the application of variant callers, while in WGS the demodification algorithm on variant calling can be used for testing purposes to evaluate the degree of bias the removal of C>T changes would bring to the overall performance of the caller. After the variant calling step results were visualised in both groups.

2.9 Variant calling

2.9.1 Germline variants

Germline variants were identified using 1) deduplicated and for WGS base recalibrated reads and GATK HaplotypeCaller or 2) Strelka2 [177] (an upgrade from Strelka [301] that could only be applied for somatic calling), or 3) raw reads and Octopus (version 0.6.3b) [80]. HaplotypeCaller is called on the base recalibrated reads using the same dbSNP information using the following additional param-

eters: –annotation AlleleFraction –annotation FisherStrand –annotation OrientationBiasReadCounts –annotation PossibleDeNovo –annotation QualByDepth –output-mode EMIT_VARIANTS_ONLY. The raw haplotype calls were then filtered using the default settings of the VariantRecalibrator and using priors for their model based on hapmap_3.3.hg38, 1000G_omni2.5.hg38, 1000G_phase1.SNPs.high_confidence.hg38 and dbsnp_146 (more info), with the additional parameters –use-annotation QD –use-annotation MQ –use-annotation MQRankSum –use-annotation ReadPosRankSum –use-annotation FS –use-annotation SOR –mode SNP. The variant quality filtering was applied using ApplyVQSR from a recalibration table generated by VariantRecalibrator. The variant recalibration procedure uses a probabilistic model that assigns a probability score to observed variants compared to a control truth set and removes variants bellow a certain sensitivity threshold. GATK HaplotypeCaller calls indels and SNPs simultaneously through local realignment and a model that calculates probabilities of possible haplotypes. Octopus and Strelka2 are bioinformatic tools that can be used for both somatic and germline variant identification. GATK HaplotypeCaller is more conservative to the other callers used, due to the simultaneous use of several different genetic variation datasets. Both GATK HaplotypeCaller and Octopus rely on the local de-novo assembly of haplotypes. The three variant callers used for the discovery of germline variants allow for the calling of both SNPs and short indels.

2.9.2 Somatic variants

For matched tumour-normal sample pairs somatic variant identification Octopus, Mutect2 (GATK version 4.1.4.1), and Strelka2 (version 2.9.10) were applied to the datasets. Mutect2 is a somatic variant caller that was developed from the initial version Mutect [76] with additional local assembly computations similarly

to the GATK HaplotypeCaller. The three variant callers used for the discovery of somatic variants allow for the calling of both SNVs and short indels, as well as the discovery of *de novo* mutations.

2.9.3 Mutational signatures

The contributions of mutational processes to the available Barrett's oesophagus patients were deconvolved using the R package deconstrutSigs [292], which was applied on WGS and specifically pre-processed TAPS and WGBS data. The signatures were based on somatic variant calls from Strelka2 and Mutect2. The deconstructSigs tool was used with the following parameters: signatures.ref = signatures.nature2013, contexts.needed = TRUE, tri.counts.method = 'genome'. The first parameter gives the original list of signatures as defined in [23], and the second ensures the mutations are normalised to the trinucleotide context of the genome.

2.9.4 Copy number and structural variation

Manta [69] was used to get additional information about the structural variation in the WGS samples. The collected information was used for testing purposes to identify what is the contribution of larger genomic variants to the differences of WGS to TAPS or WGS to WGBS. For this purpose, Manta's variant calls were used for secondary Strelka2 somatic calling, and the difference between the two runs was assessed.

Sequenza [95] is a tool that allows for the discovery of copy number variants in matched tumour-normal sample pairs, and it was applied only on WGS samples to longer indels due to the sensitivity of CNV calls to SNVs (DNA modifications in both TAPS and WGBS can be mistaken as SNVs by the tool).

2.10 Computation and statistics

All experiments requiring random sampling of positions and reads from BAM or CRAM and BED files were conducted with the Linux command *shuf*. Numerical handling and statistical analyses of data were performed using Python's packages NumPy [9] and Scipy [349], or R default statistical functions.

2.11 Machine Learning

Multiple Python packages were used for feature extraction, pre-processing and train-test splitting of datasets, and consequent training and evaluation of the performance of ML models. For feature extraction from BAM or CRAM files, a custom Python script was written. Data were cleaned and restructured using Pandas [228], while all of the data pre-processing (normalisation and spitting), and the feature selection were made using scikit-learn [262]. The synthetic multiplication of minority samples to decrease the imbalance in the training set was done using imblearn [13]. The training and validation of logistic regression and random forest models were again conducted with scikit-learn. The training of neural network models was performed using Keras [74] wrapping around TensorFlow [221]. The hypertuning tests for the NN model were run using keras-tuner [254].

2.12 Downsampling

In order to study the relationship between coverage and the overlap of call sets by different somatic variant callers, as well as to measure the expected error rate per sequencing depth given the individual sample (with its cellularity

and mutational load), the WGS samples matching TAPS were downsampled to depths of 5x, 10x, 15x, 20x, 25x, 30x, 40x, and where it was applicable—50x. The downsampling with repeats was done with samtools view -s 0.x where x is the *proportion* of reads from the sample to be included for the desired depth. The overlap of these downsampled calls was expected to be higher, given the random chance for the inclusion of the same read pairs in different downsampled depths. Thus, the lower bound of variant calls intersection was achieved by having a downsampling experiment without repeats. This was done with a custom Python script, which split the original input into two halves by read names keeping read pairs and separating singleton reads. The WGS samples were downsampled without repeats to 1X ($\approx$ 25x), 2X ($\approx$ 12.5x), 3X ($\approx$ 6.25x) and 4X ($\approx$ 3.125x). The data obtained from the downsampling without repeats samples was further downsampled with samtools view -s 0.x to achieve mean sequencing depths of 20x, 15x, 10x, and 5x.

2.13 Quality metrics

A variety of quality metrics were used to assess sequencing quality, alignment rate, and for TAPS and WGBS data—chemical conversion of cytosines (modified or unmodified). The majority of those were based on extracting features from the aligned reads in BAM or CRAM format with custom scripts followed by data pre-processing (like binning and normalisation) and visualisation.

2.13.1 FastQC

The quality of raw and post-trimming reads was examined with FastQC [28], where biases per read length, base, read orientation in pairing (first versus second in pair), sequencing cells, fragment length, and overrepresented sequencing

distributions are visualised in individual reports per sample.

2.13.2 Read quality metrics

The read quality metrics category included data processing, and visualisation of sequencing read characteristics extracted directly from the BAM or CRAM format.

Alignment rates The ratio of aligned reads to total reads described the alignment rate. Furthermore, the requirement to use unique reads gave an estimate of alignment errors due to library preparation and the mappability of reads.

Fragment and coverage distribution The fragment length distribution was obtained through direct handling of the BAM or CRAM files (extracting the absolute values from the ninth column). The genomic coverage was obtained by either bedtools genomecov or samtools depth (the second had higher processing speeds).

Mapping and base quality The mapping qualities (MAPQ) depend on the aligner. Both the MAPQ and base qualities (BQ) were visualised as mean values per read after sampling from the total number of reads or were directly visualised as distributions for comparisons.

Conversion The conversion or non-conversion was calculated on specific spike-ins: conversion on modified spike-ins for TAPS and unmodified for WGBS; non-conversion was measured on unmodified spike-in in TAPS and WGBS it was measured through the β values in CpH contexts.

ID- and M-bias The methylation bias or M-bias is an observed property of bisulfite sequencing reads that describes the variable methylation towards

the ends of the reads [47]. The per read length modification levels for CpG, CHG, and CHH contexts were visualised using Bismark and asTair on WGBS and TAPS data. Similarly, the indel distribution was studied per read length and visualised as an indel bias (ID-bias) plot. Further analyses of deletions and insertions in sequenced reads were done through a custom Python script extracting indel information from the reads' CIGAR string together with other descriptive characteristics of the reads for visualisation.

2.14 Visualisation

The figures included in this thesis were either data analysis graphics and per region genomic visualisation of reads and features, or schematics drawn using Inkscape [7] and diagrams.net [6].

2.14.1 Sequencing reads and genomic tracks

The Integrative Genome Browser (IGV) [286] was used to visualise sequencing reads in BAM or CRAM format, as well as different genomic feature tracks in bed or bedGraph format and to take screenshots of the genomic tracks.

2.14.2 Figures generation

All of the data analysis results presented were visualised using ggplot2 [116] in R, or matplotlib [147] and seaborn [357] in Python. The visualisation of the NN model was done using Keras function plot_model. Venn diagrams were drawn with VennDiagram [66], and heatmaps using NMF [110] packages in R.

Chapter 3

Development of the TAPS modification calling method

The TAPS method for studying cytosine modifications was developed with stepwise chemical optimisation, which was driven by the results obtained by data analysis (see Figure 3.1). Quality metrics and estimates of conversion were established using spike-in sequences (used only for experiment quality control and not for normalisation at present). TAPS data differs from WGBS in the expected conversion (modified cytosines versus unmodified cytosines); thus, it required the development of a custom-made analysis pipeline. An analysis tool for TAPS data, asTair, was made in parallel with chemical developments of the method. This chapter presents the data analysis of five complete whole-genome runs (including library preparation, PCR, sequencing) and two test runs (li-

Part of the results of this chapter was published as *Liu, Y., Siejka-Zielińska, P., Velikova, G. et al. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. Nat Biotechnol 37, 424–429 (2019)*, https://doi.org/10.1038/s41587-019-0041-2 and presented as a talk at the *3rd Epigenetics and Bioengineering Conference, 2019, Spain*, and as a poster at the NDM organised *DPhil day, 2018*. The bioinformatics tool asTair for methylation calling can be found at https://bitbucket.org/bsblabludwig/astair.

brary preparation and PCR after pull-down and sequencing) using DNA from E14 mouse embryonic (mESC) cell line unless stated otherwise. These runs varied in the library preparation conditions, such as the TET enzymes used, temperature, incubation time, pH, and the reducing agents (see Table 3.1 for a summary of the run conditions).

3.1 Data-driven chemical development of TAPS

TAPS performance on genomic DNA was optimised by analysis of sequencing data. The novelty of the sequencing method required building a data processing pipeline which took into account specific features of TAPS data. The need for reliable performance evaluation of the data was further complicated by the short time frame for development and optimisation (see Figure 2).

It is essential to provide additional information about specific parameters of the TAPS development pipeline before explaining the results of the sequencing runs. The choice of the following parameters was crucial for method development.

TET enzymes TET enzymes oxidise 5-mC in consecutive steps to cytosine in a process known as the active demethylation pathway. Tet1 is the main methylation level maintenance enzyme in mammals. TAPS requires highly active TET enzymes for full conversion (similarly to TAB-seq [380]), and two different Tet1 forms were used and assessed for applicability in TAPS: NgTet1 and mTet1. NgTet1 (Naegleria TET-like protein) originally found in a non-pathogenic amoeba *Naegleria gruberi* and can be recombinantly expressed in *E. coli*. NgTet1 is a homologous protein to the mammalian TET enzymes and converts 5-mC and 5-hmC to 5-caC, which is reduced by a TAPS-specific compound to dihydrouracil (DHU). The mTet1 (catalytic domain) was also generated in-

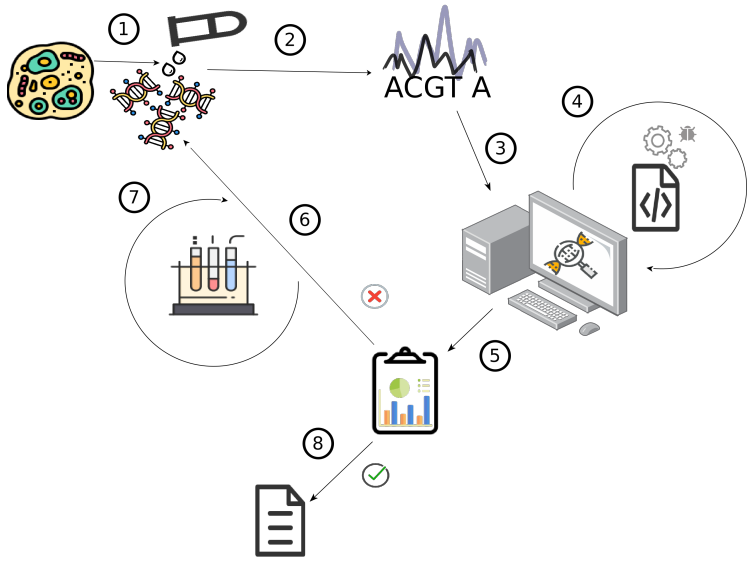


Figure 3.1: TAPS method development as a schema: (1) DNA was isolated from cells, and the TAPS sequencing library was prepared; (2) DNA was sequenced and (3) the data were analysed; (4) the code is under active development and (5) the results from the data analysis could result in (6) going back to the library preparation and (7) method optimisation; (8) the manuscript was written after the TAPS method was significantly improved. The diagram was created with diagrams.net (open-source) [6].

house in Expi293F cells for the needs of this study. The catalytic domain of the enzyme is needed to induce demethylation by itself [161] and was applied for Tet1 expression in [156].

Chemistry The overall chemical environment during the experiment including the incubation time, pH, temperature, ion strength of the solution, and the solvent itself can cause variation in both the conversion rate and the methylation. More aggressive treatment with higher temperatures and longer incubation times could lead to DNA damage. Also, pH values outside the physiological range can cause depurination, strand breakage, and destabilisation of the double-stranded duplex [26], [285]. Also, it is possible that the organic solvent, enzymes or other components in the solution can cause demethylation (such as

DMSO [336] and as a side reaction of some enzymes) which can serve as another validation of why the chemical conditions should be strictly monitored and maintained during the reactions in TAPS library preparation and conversion.

DNA polymerase DNA polymerases amplify the DNA sequences of interest during the PCR step, and they show differences in their error rates and biases [226]. The high fidelity polymerases used in PCR amplification have low error frequencies (approximately a single incorrect base per million or less [77]). Still, the overall PCR performance is highly dependent on the DNA sequence. High (>65%) or low GC contents (<12%) [22] can lead to amplification imbalance where the reads with extreme GC contents are not representatively amplified in the final library. Adding betaine or changing the PCR reaction conditions such as extending the time for the denaturation step and the use of particular polymerases could help with the amplification of GC-skewed libraries. A known problem in bisulfite sequencing is that there is a bias for higher amplification of methylated sequences (because of their more balanced GC content compared to the overall highly AT-rich context) which potentially skews the expected modification values, particularly at lower coverage [160]. Certain uracil-insensitive DNA polymerases that can read uracils can lead to a more balanced PCR library. Given the chemical conversion of modified cytosines to DHU in TAPS, the application of bisulfite specific uracil-insensitivity polymerase was used. The choice of the polymerase for TAPS and WGBS amplification in this study was Kapa HiFi Uracil+ (Kapa Biosystems) because of its high sensitivity and low error rate and lesser bias of higher amplification of methylated fragments and the PCR cycle [160].

DNA input The multiple rounds of optimisation during method development required a large and reliable source of DNA, which nonetheless had to be similar to the DNA that would be used in clinical applications. The requirement for

Experiment	DNA input	TET enzyme	Reducing agents	Organism	DNA	Incubation time
run1	1 μg	NgTet1	pic-BH3	mouse	genomic	1h
run2	1 μg	NgTet1	pic-BH3	mouse	genomic	8h
run3	1 μg	NgTet1	pic-BH3	mouse	genomic	5h
run4	1 μg	NgTet1	pic-BH3	mouse	genomic	3h
run5	200 ng	mTet1CD	pyr-BH3	mouse	genomic	16h
test run 1	200 ng	NgTet1	pyr-BH3	mouse	genomic	16h
test run 1	200 ng	NgTet1	pyr-BH3	mouse	genomic	12h
test run 1	200 ng	NgTet1	pyr-BH3	mouse	genomic	8h
test run 1	200 ng	NgTet1	pyr-BH3	mouse	genomic	16h
test run 1	200 ng	mTet1CD	pyr-BH3	mouse	genomic	16h
test run 1	200 ng	mTet1CD	pyr-BH3	mouse	genomic	16h
test run 1	10 ng	NgTet1	pyr-BH3	human	cfDNA	16h
test run 1	10 ng	NgTet1	pyr-BH3	human	cfDNA	16h
test run 2	200 ng	NgTet1	pic-BH3	mouse	genomic	2h
test run 2	200 ng	mTet1CD	pic-BH3	mouse	genomic	2h
test run 2	200 ng	NgTet1	pyr-BH3	mouse	genomic	16h
test run 2	200 ng	mTet1CD	pyr-BH3	mouse	genomic	16h

Table 3.1: A complete description of samples produced during TAPS development with specific reaction conditions and input DNA information. During the course of TAPS method development, the effect of multiple reaction parameters were tested, such as the incubation conditions, TET enzyme, reducing agent, and input DNA quantity.

well-studied, reliable, and a mammalian DNA source suggested DNA extraction from cultured cells as an option. The E14 cell line (E14Tg2A) was used as it is considered to be amongst the most widely used cell lines in molecular biology [154] for over three decades [352] (and it even was used for cloning of live mice). Given its widespread use, the DNA from the E14 cell line enables direct comparisons of TAPS with TAB-seq data, as well as with WGBS data [126]. However, embryonic stem cells are expected to have higher 5-hmC levels than the majority of the genomic DNA, as well as can dynamically change their modification status under certain conditions [332].

Sequencing library preparation On the one hand, the library preparation protocol can have a direct effect on the alignment and processing of the sequencing data, on the other hand, some library preparation techniques enable higher sensitivity towards low input quantities but at the expense of library directionality (such as the post-bisulfite adapter tagging method, PBAT) which

may complicate data analysis. In the case of TAPS, the time and the step of TET oxidation can also affect the final conversion rate. Finally, the choice between a pair-end and a single-end library can result in higher coverage, or it can have minor effects on the information content (especially in cases where pair-end reads have considerable overlap).

Reducing agents Two reducing agents are discussed in this chapter—pyridine borane and picoline borane (others were tested but not included because of their lower efficiency), which differ in their conversion rates, toxicity, and mildness of conditions.

Spike-in sequences During the development of the TAPS method, different short DNA sequences with known sequence and properties were spiked-in with the genomic DNA to assess the success of the method (more information on the spike-ins can be found in the following methods subsection). Very short spike-ins were named after their length in base pairs as DNA sequences as 120-mers and 160-mers. Here is a list of the spike-ins used given that they have changed between runs (* was used in all runs, $^{t1/2}$ indicates test run 1 or test run 2, and ^{2} shows run 2):

- M.SssI methylated at CpG positions bacteriophage λ was used to assess TAPS method reading of methylation in CpG contexts*;
- completely hydroxymethylated 2 kb spike-in was used to assess possible bias against 5-hmC of the TAPS method[1];
- completely methylated 2 kb spike-in was used to assess possible bias against methylation in non-CpG contexts of the TAPS method[1];
- completely unmodified 2 kb spike-in was used to assess the false-positive rate of the TAPS method*;

- M.SssI methylated at CpG positions 2 kb spike-in was used to assess CpG methylation conversion of the TAPS method[2−t1];
- 1:1 mixture of M.SssI methylated at CpG positions and unmethylated 2 kb spike-in was used to assess PCR bias against CpG methylation of the TAPS method[t2−5];
- M.CviPI methylated at GpC positions 2 kb spike-in was used to assess the non-CpG bias of the TAPS method[2−t1];
- two 120-mers to assess TAPS method reading of methylation and hydroxymethylation in CpG and non-CpG contexts*;
- NCNN to assess TAPS method conversion rates at multiple cytosine contexts at higher depths[t2−5].

It is important to note that TAPS conversion rate is a result of three consecutive dependent processes (which is a significant difference from WGBS where the conversion rate measures only the bisulfite conversion of unmodified cytosines): 1) cytosine modification, 2) TET oxidation, 3) reduction with picoline or pyridine borane. Thus, incomplete oxidation or reduction could skew the estimate of the real modification levels.

The calculation of the modification (β) values is presented in the following methods section.

3.1.1 First whole-genome run

The data from the first sequencing run of the new base-resolution bisulfite-free method were analysed to evaluate the conversion rate as well as the suitability of the spike-ins added. Additionally, the sequencing data could provide more sensitive results about the actual conversion rate compared to the laboratory-

based HPLC-MS/MS (high performance liquid chromatography combined with mass spectrometry) readings used as pre-sequencing estimates. By default, the data were quality trimmed by Illumina BaseSpace with read lengths in the range 35-79 bp, with mode 79 bp. The reads were then aligned with both a wild-card bisulfite sequencing aligner BSMAP and a whole-genome aligner BWA. BSMAP permissively aligns tentative cytosines or thymines using the IUPAC symbol Y thus allowing the alignment of TAPS data as well as WGBS data, while BWA was used due to its reliability and the relatively small expected difference between TAPS and WGS in terms of sequence complexity. Both aligners showed mapping rates of 97-98%, suggesting either an incomplete conversion or that very few reads are unmappable by BWA as a result of their high conversion and difference from the reference sequence. The coverage and conversion rates of the spike-ins and the genomic coverage were assessed in a step-wise manner. The initial estimates of the conversion using spike-ins allow for a fast evaluation of the experiment, which helped the laboratory method development. The highest conversion was observed on the CpG positions in 120_1-mer (91-93%). Still, the first run showed poor overall performance. In the other short spike-in, 120-2, the non-CpG conversion rates were lower than the CpG conversion, which could partially be explained by NgTet1 lower activity on non-CpG methylation (see Figure 3.2).

In comparison to the high conversion rates in 120-1 short spike-in, the 2 kb spike-ins that have to be fully methylated or hydroxymethylated seemed unmodified (or partially modified), indicating that there was an undetected issue with the conversion. A broad coverage gap between the unmodified and 5-hmC and 5-mC modified 2 kb spike-in showed an over 50 times higher read count for the non-modified cytosine, suggesting an issue with the PCR amplification depending on the sequence context (see Figure 3.16). The λ phage CpG conversion

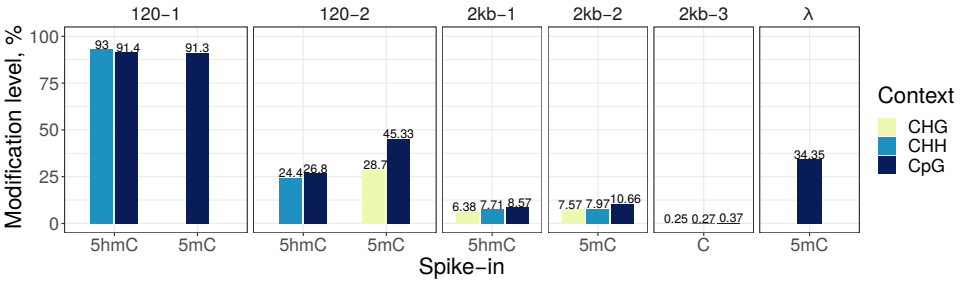


Figure 3.2: The conversion rates, as seen in the first whole-genome run showed by spike-in, context, and modification type. A complete conversion (β value of ≈ 1.0) was expected for the short 120-mer spike-ins, the λ phage, and the 2 kb spike-ins where all cytosine positions were hydroxymethylated or methylated. The lower observed modification values were influenced by the short incubation time (1h) and the GC bias.

was lower than expected in the 20-30% range while it was fully enzymatically methylated by M.SssI. Careful observation revealed that the λ phage spike-in appeared to be divided roughly in half of well and poorly converted cytosine. The higher observed conversion in the short spike-ins (120-mers) was assumed to be a result of their short length or sparse CpG positions compared to 5-hmC, and 5-mC modified 2 kb spike-ins and fully modified at CpG positions the λ phage. Grouping the spike-in fragments into short ($<$ 150 bp), medium (150$<$x$<$200 bp), and long ($>$ 200 bp), and visualising the dependency of fragment size distribution to conversion (even number of modifications per fragment was calculated) did not show an association between fragmentation patterns and conversion (see Figure 3.15). In order to test the hypothesis that conversion was negatively associated with the GC content, the λ phage was binned into 200 bp windows with 100 bp sliding windows, and the average GC content and CpG count per bin were calculated. Although real separation of CpG and GC is impossible, as they are related, there was a more considerable negative correlation between the averaged modification value per bin and GC (ρ=-0.878, p-value $\leq 2.2e^{-16}$, Pearson's product-moment correlation, two-sided hypothe-

sis, CI=0.95), while fragment length did not appear to be the leading factor for the conversion differences. A possible explanation is that high GC content, in combination with artificial modification at every CpG, might cause difficulties with denaturation and introduce PCR bias. Furthermore, large, highly modified GC-rich regions are not common in mammalian genomes (apart from specific CGIs, shores, and shelves), while this description fits the sequence of the longer spike-ins. As a proxy for the estimated coverage and conversion rate in the mm9 genome, again binning into 200 bp windows with 100 bp sliding windows was used. This analysis showed that 14% of mm9 and 44% of lambda have a GC content of 50% and above. As an additional proof to this, the conversion in λ phage and 120_2-mer was higher at the lower GC half (first 61.7% and 48.3% second half of 120_2-mer, first 55.1% and second 44.8% half of λ phage), and the overall GC content was lower in 120_1-mer where the highest conversion at CpG was observed (first 38.3% and 36.7% second half).

Following the discovery of the negative association between the expected modification levels and the GC content of the DNA sequence, the laboratory protocol was adjusted to ensure a more complete TAPS conversion. Longer incubation times showed higher conversion as they increased between the first three runs.

3.1.2 Second whole-genome run

The first TAPS run showed that GC bias might interfere with the successful conversion of the method. After optimising the library preparation conditions for high GC-content regions, a new batch of TAPS data was sequenced. In the second run, two of the 2 kb spike-ins have changed: the 2 kb spike-in which was fully 5-hmC modified has been enzymatically methylated at GpC positions by M.CviPI methyltransferase instead, and the fully 5-mC modified spike-in has

been enzymatically methylated at CpG positions by M.SssI. This change was required because the high density of modified cytosines was interfering with TAPS conversion under the incubation conditions in the first run.

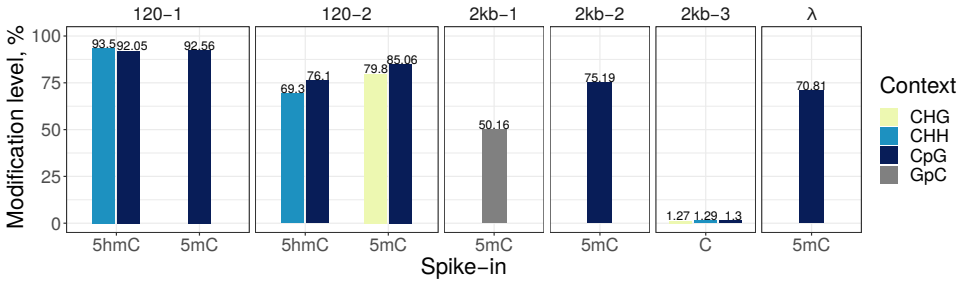


Figure 3.3: The conversion rates, as seen in the second whole-genome run showed by spike-in, context, and modification type. A complete conversion (β value of $\approx$ 1.0) was expected for the short 120-mer spike-ins, the λ phage, and the 2 kb spike-ins where GpC or CpG positions were enzymatically methylated. The library preparation conditions influenced the lower observed modification values, together with the uneven GC content, but to a lesser extent than the first run.

The initial exploration of the spike-ins in IGV seemed promising, as the GC bias in the lambda genome seemed to have diminished, but was not entirely removed (see Figure 3.16). The sequencing data were BaseSpace trimmed with read lengths 35-80 bp, which aligned to the concatenated mm9 and spike-ins genomes with mapping rates of 98-99% with BWA (not much improvement was seen when an additional mapping with BSMAP of the BWA-unmapped reads was done; BSMAP salvaged $\approx$ 5000 reads). No further analysis of the location of these BSMAP-specific reads, and how these reads differ from the BWA-successfully mapped ones were conducted due to their low numbers.

The spike-ins showed higher conversion rates in both CpG and CpH contexts compared to the first sequencing run (see Figure 3.3). Again, the highest conversion was observed in the short 120-1 (92-93.5%), while the CpG methylation in bacteriophage λ and the 2 kb CpG methylated spike-in was in the range 70-85%. However, the false-positive rate (measured on the unmodified 2 kb spike-in) has

increased from the first sequencing run (≈ 0.3% compared to ≈ 1.3%) which was a direct result of the change in the library preparation conditions such as longer incubation times.

The next step was to compare public E14 WGBS data [126] to TAPS, and the two methods seemed to agree on CGIs as seen in IGV, with TAPS having generally more even coverage throughout the genome. Base-resolution genomic comparison of modification values of TAPS and WGBS required initial filtering of true modified positions using binomial testing (for WGBS pbinom(T, C+T, 1-non-conversion rate, lower.tail=T), and TAPS pbinom(C, C+T, 1-non-conversion rate, lower.tail=T)) with no multiple testing correction with p-value cutoffs of 0.1 and 0.05. Only two of the publicly available datasets were used, because of their low non-conversion rates in the λ phage. Binomial testing is a common method for distinguishing modified from unmodified cytosines given the background noise from the non-conversion rate, sequencing, and alignment errors [305]. The results showed some similarity of the modified cytosines between the WGBS and the TAPS datasets but the significantly lower conversion rate in TAPS was an obstacle to make any conclusions for the mouse genome.

3.1.3 Third whole-genome run

After further optimisation of TAPS preparation conditions in order to increase the conversion rate and to enable direct comparisons with WGBS data, the third round of sequencing data was prepared. The conversion rates observed on all spike-ins were the highest of the three runs with rates over CpG methylation of 82-95% (lowest in the methylated 2 kb spike-in of the three runs), and CpH rates of 86-95% (see Figure 3.3). The GpC spike-in again showed lower modification levels compared to CpGs, which may be a result of both incomplete methylations with the methyltransferase and incomplete conversion by TET or picoline

borane. Unfortunately, the false-positive rates were higher than in the previous two runs 2.1-2.5% indicating that the same conditions with picoline borane that increase the conversion have increased cytosine deamination as well. The higher false-positive rate would mean that a higher proportion of cytosines can be filtered out as unmodified when applying binomial testing. After binomial filtering using an average false-positive rate of 2.5%, there were around $35\text{x}10^6$ CpGs left with a p-value cutoff of 0.1 showing that the higher non-conversion rate did not filter out a large proportion of the CpGs. The GC bias was less pronounced than in the previous runs but still higher GC regions showed lower coverage as observed in bacteriophage λ (see Figure 3.16). Furthermore, the coverage of the third run looked more uneven than the second run, which might be a result of both the harsher conditions and alignment with higher conversion rates. Harsher chemical treatment may introduce DNA damage which could affect PCR amplification and alignment [335].

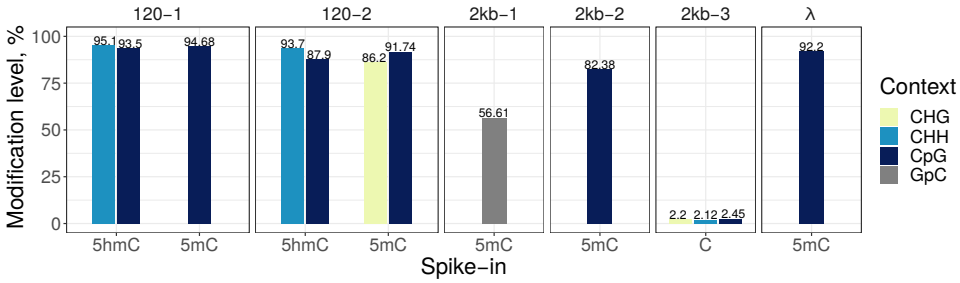


Figure 3.4: The conversion rates, as seen in the third whole-genome run showed by spike-in, context, and modification type. A complete conversion (β value of $\approx$ 1.0) was expected for the short 120-mer spike-ins, the λ phage, and the 2 kb spike-ins where GpC or CpG positions were enzymatically methylated. The change in the library preparation conditions significantly increased the conversion of modified CpGs, with higher performance on single non-CpG positions. The lower GpC modification level may be a result of both incomplete methylation and conversion. Also, a higher false-positive rate was observed compared to the previous runs.

The per base comparisons of TAPS and the public WGBS data showed that

the ground state modification levels of WGBS and TAPS data differed, which decreased the reliability of any discoveries in terms of modification in TAPS before they could be validated in a more similar dataset (see Figure 3.5). The 2i-adapted cells from [126] data presumably should have been the most similar in terms of sequencing depth and growth media, so they were downloaded and processed, but there seemed to be an issue with the conversion rate and the WGBS data showed different than expected modification values distribution. A possible explanation for the phenomenon was that the 15% serum compared to 10% in the WGBS paper (they did not use the serum in the 2i medium) were enough to trigger the post-implantation modification state of the cells, as fetal bovine serum (FBS) may trigger differentiation. Tamm et al. 2013 [332] claims that the growth factors and other components of FBS can promote mESC growth and other authors have suggested that it contains potential differentiation factors, which can modulate growth, differentiation, and implantation of embryonic cells [71]. These results indicated that the 2i cultured cells were growing in demethylating and hypomethylating media, and a comparison of methylated (TAPS) to unmethylated (2i WGBS) would be uninformative. These computational findings were confirmed by the HPLC detection of around 4% of 5-mC compared to the 1% 5-mC given for the cells grown in the 2i media. As a result of this analysis, the public data of serum-grown cells was used as a secondary validation in the next runs, and a separate in-house WGBS data were prepared from the same input DNA to ensure minimal differences between the TAPS and WGBS which are not a result of their chemistry. Given that the testing set data were an mESC cell line with an unstable methylome, the methylation profile of this cell line differed from the usually observed bimodal distribution of CpG modification values with modes at 0 and 1.

After the inconsistency in the modification distributions between the initial

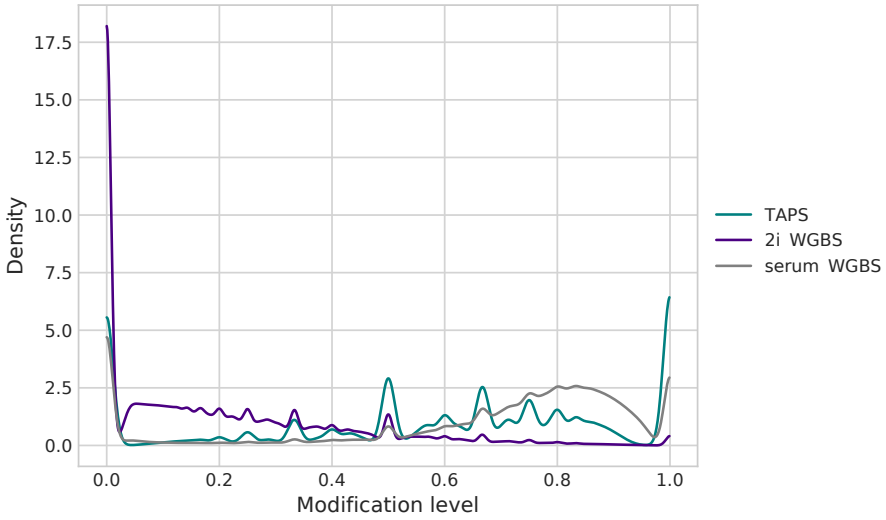


Figure 3.5: This figure shows a comparison between TAPS and WGBS in terms of the distribution of modification values, where the E14 TAPS cells were grown in media containing serum, similarly to one of the WGBS conditions. TAPS showed higher similarity to the WGBS serum case, compared to the 2i WGBS. These results showed that the initial dataset (2i WGBS) was not suitable for comparison to TAPS due to its lower methylation levels.

control WGBS sample and TAPS was discovered, in-house WGBS samples were prepared for the next two whole-genome sequencing runs.

3.1.4 Fourth whole-genome run

After several runs of TAPS were sequenced to remove the GC bias, to improve the conversion rate, and the false-positive rate, the fourth whole-genome TAPS run was sequenced together with WGBS from the same cells. The undetermined and indexed reads were merged by Illumina for the WGBS and were separate (and processed separately until the modification calling) for TAPS. The sequencing library was prepared with fragment lengths 200-400 bp to test TAPS conversion on sequences with longer lengths, and also it would decrease the proportion of overlapping pair-end reads. The conversion rates measured on TAPS spike-ins were lower than these in the third sequencing run, how-

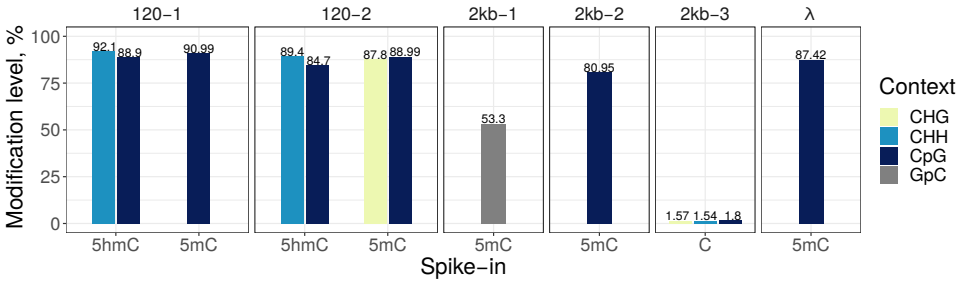


Figure 3.6: The conversion rates, as seen in the fourth whole-genome run showed by spike-in, context, and modification type. A complete conversion (β value of $\approx$ 1.0) was expected for the short 120-mer spike-ins, the λ phage, and the 2 kb spike-ins where GpC or CpG positions were enzymatically methylated. The change in the library preparation conditions significantly increased the conversion of modified CpGs, with higher performance on single non-CpG positions. The lower GpC modification level may be a result of both incomplete methylation and conversion. The fourth whole-genome sequencing run performed on par to the third run but with a decreased false-positive rate.

ever, on the short 120-mer spike-ins no obvious bias in conversion against CpH positions was seen (88-92% in CpH versus 84-91% CpG), but the conversion rates for 5-hmC were lower than those of 5-mC which may be a result of the lower processivity of TET on 5-hmC [369] (see Figure 3.6). The false-positive rate was decreased from 2.5% in the third run to $\approx$ 1.5% in the fourth run, with the highest error rate in CpG positions. The WBGS data were mapped with Bismark and then deduplicated, 3C-filtered, and Bismark was used again to do the methylation calling. The unmodified bacteriophage λ showed over 99% conversion rate in WGBS. The modification levels profiles of the in-house WGBS showed higher agreement with the WGBS converted serum-treated cells, as previously suggested (see Figure 3.8). The Pearson correlation between the in-house and the serum WGBS was $\approx$ 0.5, while the correlation with the 2i media cultured cells was $\approx$ 0.3 ($\leq$2.2x10^{-16} p-value, Pearson's product-moment correlation, two-sided hypothesis, CI=0.95). An additional analysis of randomly selected mean base quality (Phred scores) showed significantly lower base qual-

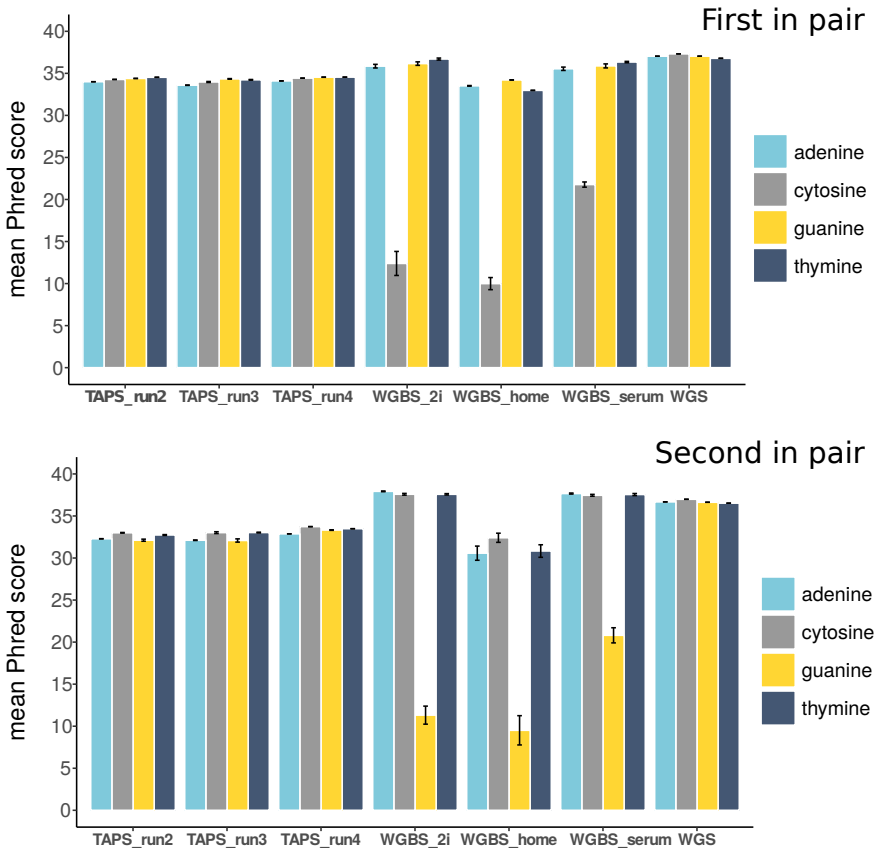


Figure 3.7: A comparative analysis of the base quality (Phred scores) of TAPS runs 2-4, public and in-house WGBS samples, and a public E14 cell line WGS sample. Individual base quality scores were averaged per read after the selection of random 10^7 reads. There was no observable bias in the base qualities of TAPS and WGBS for the different bases, while in WGBS, the C>T conversion (observed on the OT strand) and the G>A change (visible on the OB strand) lead to a lower mean Phred quality of cytosines and guanines in the first in pair and second in pair reads.

ities of thymines in the first in pair reads and guanines in the second in pair read in WGBS (see Figure 3.7). TAPS similarly to WGS did not show a significant variation in Phred score values between reads. The difference between WGBS and TAPS could be explained by the too unbalanced sequencing library after conversion in WGBS where 10-20% non-modified sequences are added in order to sequence (15% for the in-house data) [16]. The slightly lower maxi-

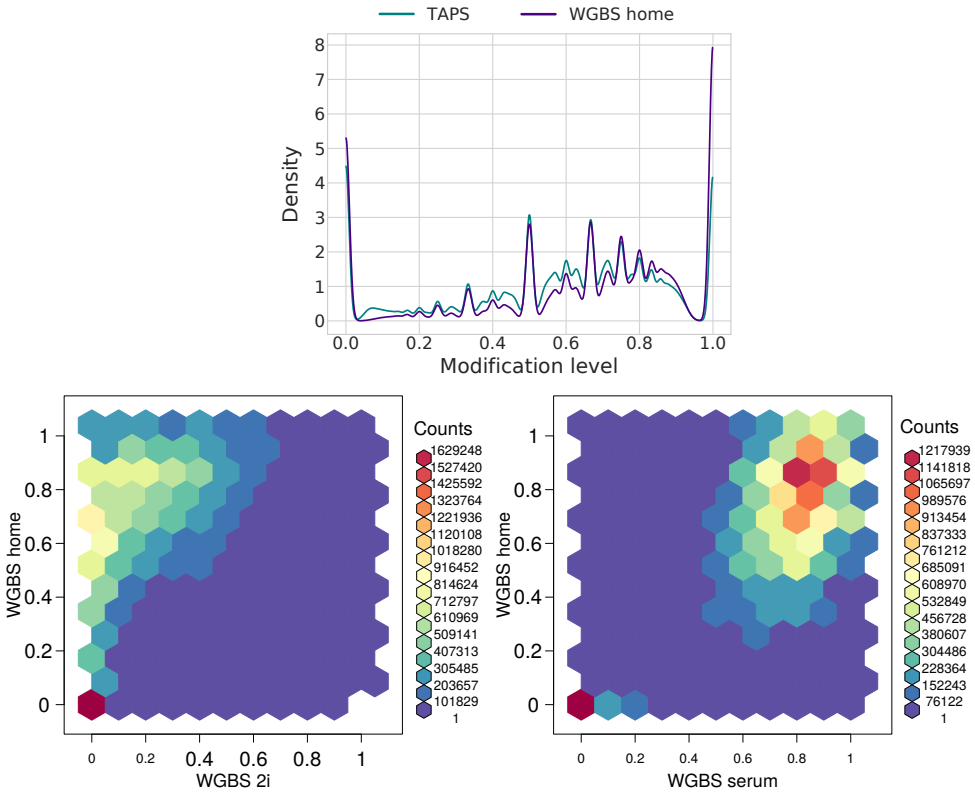


Figure 3.8: This figure shows a comparison of the methylation level distribution between TAPS and in-house WGBS (top), and the matched modification values distribution between the in-house WGBS and the publicly available 2i and serum WGBS (bottom). The in-house prepared WGBS showed more similar methylation patterns with TAPS and with the WGBS prepared from cultivated with serum E14 mESC cells (Habibi et al. paper [126]).

mum base qualities between the in-house samples and the public data are most probably a result of differences in the sequencing machines (NextSeq500 versus HiSeq2500-4000).

The mapping rate of all TAPS reads to mm9 + spike-ins was 0.978, and the addition of ϕX (Illumina) elevated it to 0.986, compared to WGBS mm9 + spike-ins 0.643, and with ϕX 0.685. Also, BSMAP was used to test how many of the remaining unmapped reads would be salvaged, and the answer was 60895, or 0.0119%, again proving that BWA could be used by itself for

TAPS alignment. The fragment distributions of TAPS and WGBS without repeat masking (converting repetitive sequences into strings of lower case letters or Ns for downstream analysis and identification as repetitive sequences) and blacklisting (removal of genomic regions with poor alignment) showed some differences as TAPS fragments are shorter on average than the WGBS ones, which might be explained by any differences in the sonication protocol. In order to make the analysis more comparable TAPS was subsampled to the number of reads in WGBS (above $22.9 \text{x} 10^6$). The evenness of coverage after alignment was of interest; therefore, the downsampled TAPS was used for comparison to WGBS as they were of the same median sequencing depth. It was found that the distribution of TAPS was narrower, suggesting less variance than WGBS, but also being slightly lower than the full coverage TAPS. Starting with the almost identical starting material TAPS had nearly twice more reads and better coverage than WGBS. A similar distribution was seen for the CpG coverage. As CGIs' modification levels are of research interest, their metagene profiles were analysed in terms of coverage and CpG modification. TAPS coverage was more even on the CGI, in contrast to WGBS, which is known to have a bias in coverage towards high modification regions (at least if these are preferentially PCR amplified [160]). The most important property of WGBS is capturing epigenetic modifications on cytosine, and it was essential to measure how well TAPS captures modification. In order to juxtapose the two methods, a coverage cutoff filtering (modified and unmodified bases covering positions needed to be at least three). These analyses again included CGI modification metagene profiles, where TAPS and WGBS seem comparable. Also, their correlation was high, given that WGBS may overestimate the modification level, even having a higher conversion rate than TAPS. Finally, 3x CpG distributions per genomic features in TAPS, WGBS, their non-overlapping region in the Venn diagram were explored. There

was a subsection of CpGs which was highly variable including modified CpGs with 0 in WGBS and 1 in TAPS. These positions suggested that TAPS reveals cytosine modifications in genomic regions where WGBS is unable to capture them due to coverage.

3.1.5 First test run

Further optimisation of the method's conditions was undertaken in order to decrease the false-positive rate and to test different TAPS conditions and applications. These test runs were directly sequenced after the pull-down step of TAPS. During the first sequencing test run, DHU_PCR and DHU_PCR-free libraries were prepared from the single-strand DNA sequence and used to validate the sensitivity of the sequencing machine towards DHU. Two other application-based samples were tested, which were single-strand prepared cfDNA and double-strand cfDNA. The overall conversion was lower in the first test run compared to whole-genome runs three and four (see Figure 3.9, 3.10). Furthermore, the difference in the conversion of CpGs vs non-CpGs appeared more pronounced than in the fourth run in all conditions and appeared even lower in the tested PCR-free conditions. The error rates achieved were the lowest since the first sequencing run suggesting that the optimisation of conditions could lead to a change in the false-positive and false-negative rates. It is important to notice that longer incubation times appear to be correlated with the false-positive rate and that mTet1 has a higher oxidation activity resulting in higher conversion.

3.1.6 Second test run

The second test run was used to evaluate the combination of a TET enzyme (NgTet1 and mTet1) and a reducing agent (picoline or pyridine borane) so that the optimal sequencing condition would be selected for the fifth whole-genome

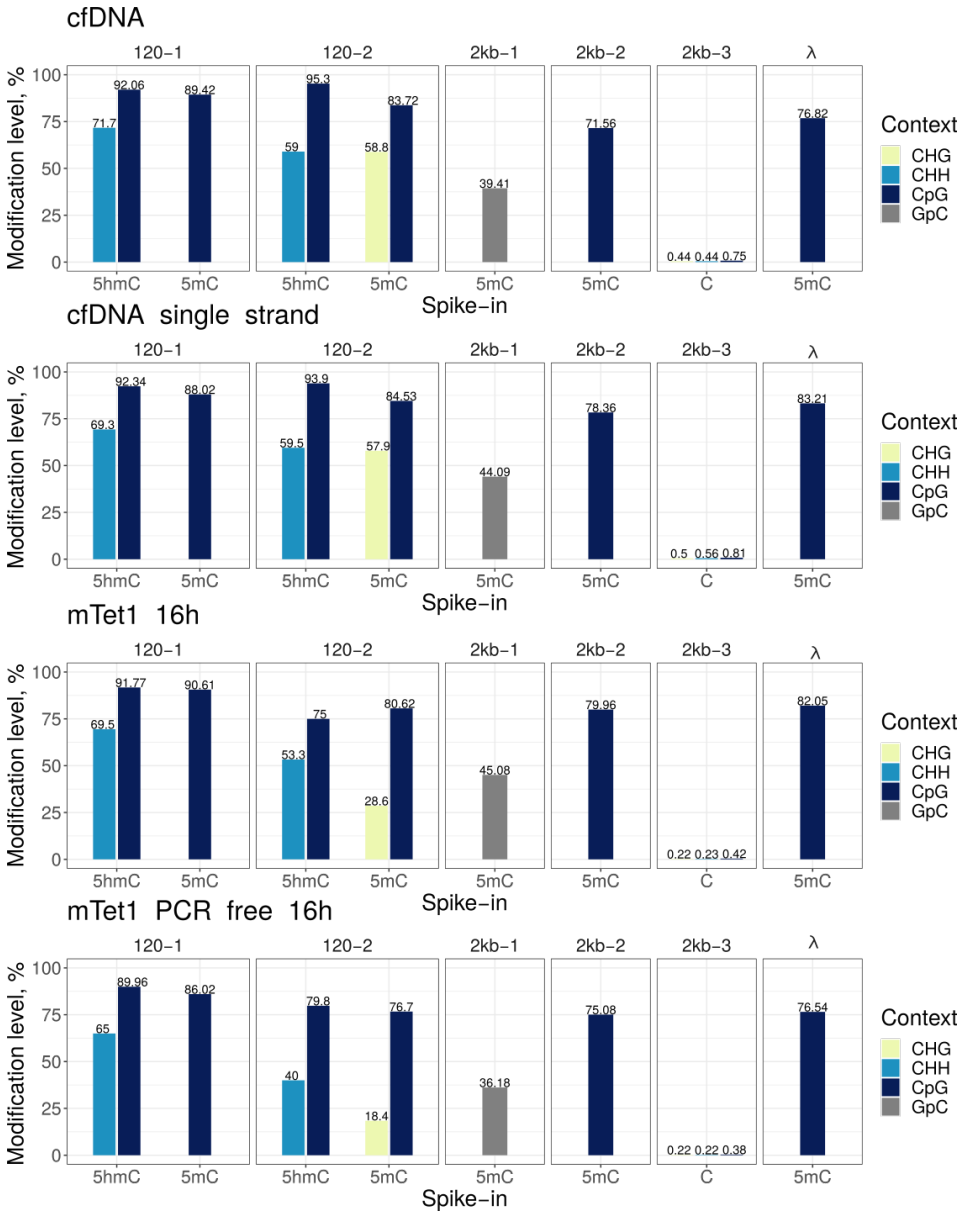


Figure 3.9: The conversion rates, as seen in the first test run showed by spike-in, context, and modification type. A complete conversion (β value of ≈ 1.0) was expected for the short 120-mer spike-ins, the λ phage, and the 2 kb spike-ins where GpC or CpG positions were enzymatically methylated. The first test run showed a more significant difference between CpG and non-CpG conversion. The first test run validated the applicability of TAPS to cfDNA.

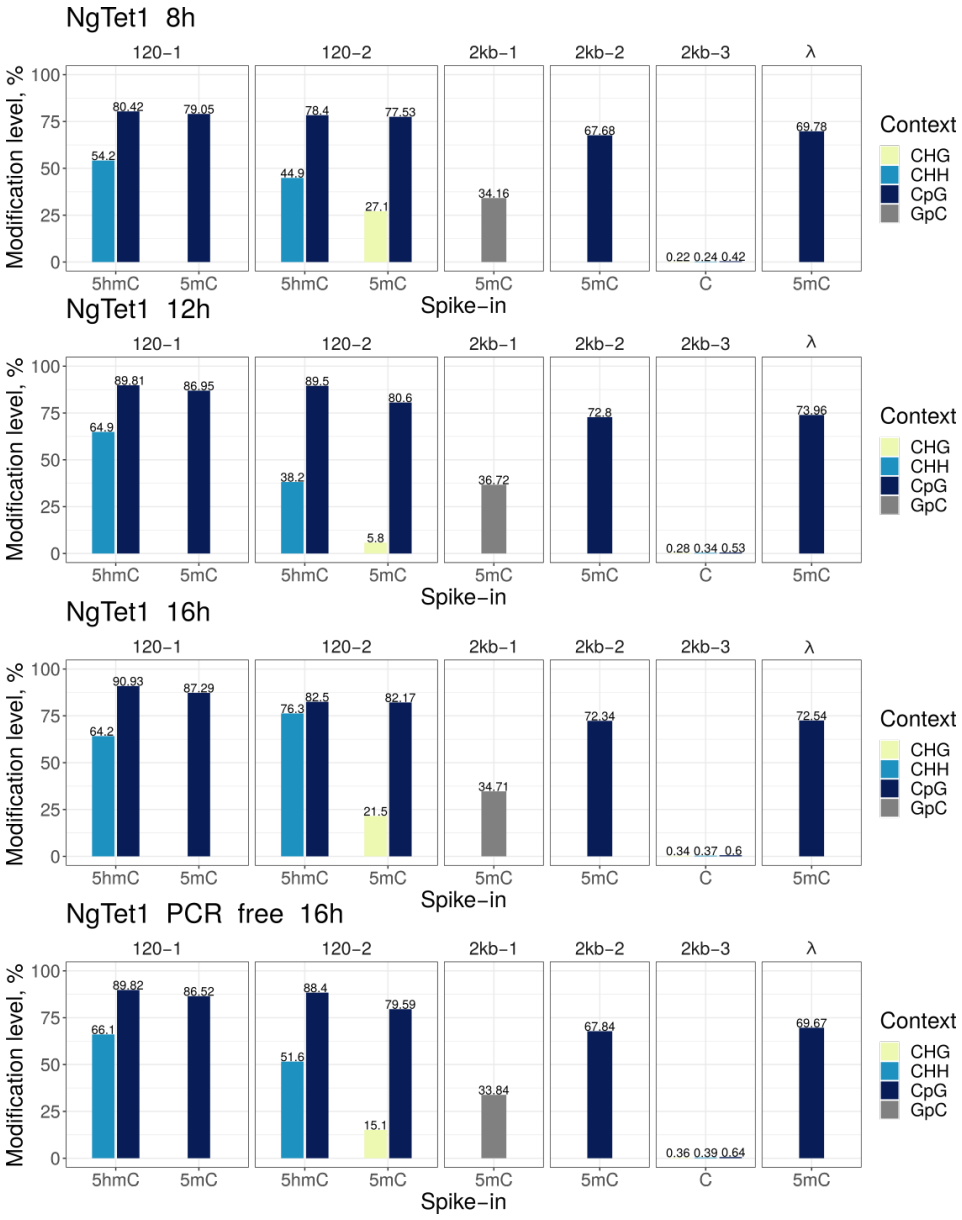


Figure 3.10: The conversion rates, as seen in the first test run showed by spike-in, context, and modification type. A complete conversion (β value of ≈ 1.0) was expected for the short 120-mer spike-ins, the λ phage, and the 2 kb spike-ins where GpC or CpG positions were enzymatically methylated. The lower GpC modification level may be a result of both incomplete methylation and conversion. This run showed the dependency of incubation times to conversion.

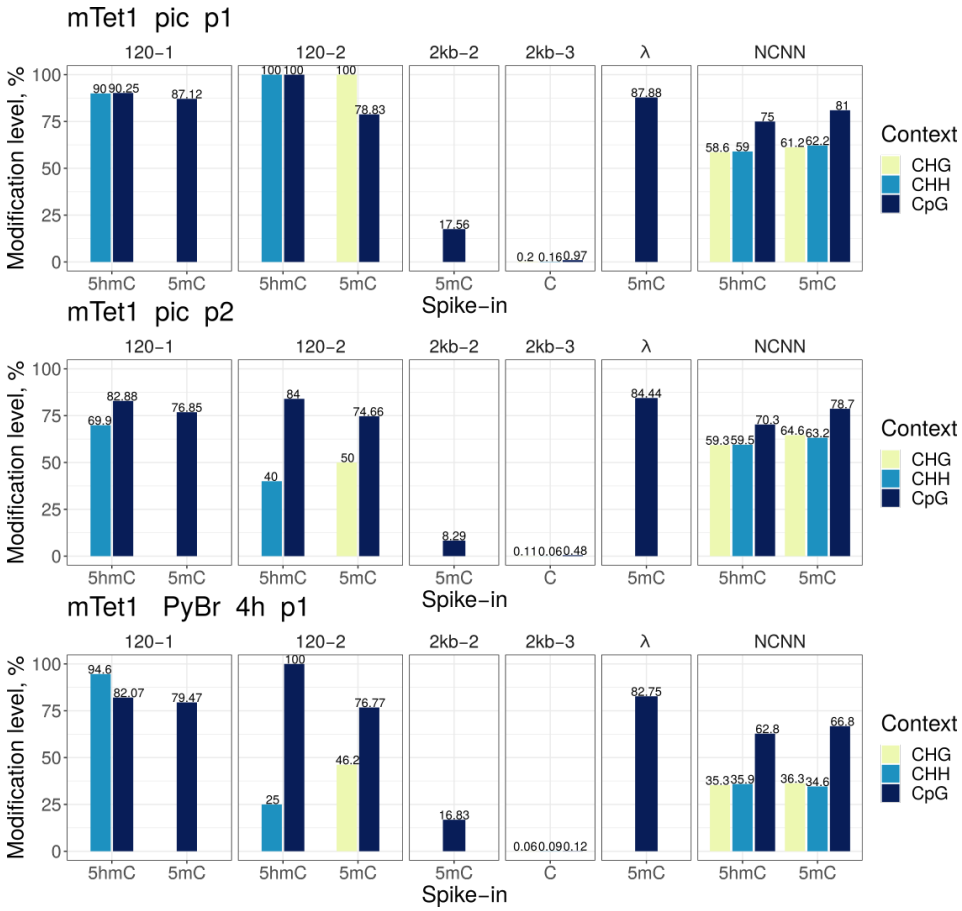


Figure 3.11: The conversion rates, as seen in the second test run showed by spike-in, context, and modification type. A complete conversion (β value of $\approx$ 1.0) was expected for the short 120-mer spike-ins and the λ phage. An additional half methylated at CpG 2 kb spike-in and a highly variable in context NCNN spike-in were added to study PCR bias and non-CpG conversion. The second test run was primarily used to compare TET enzymes and reducing agents. The mTet1 and pyridine borane condition was selected as the best performing one due to its higher conversion rates and lower false-positive rate. This condition was sequenced in the fifth whole-genome run.

sequencing run. Two changes were made to the spike-ins: the 2 kb CpG methylated spike-in was a mixture of methylated and unmethylated DNA fragments (expected distribution 1:1) which was created to evaluate PCR bias in TAPS, and a synthetically generated highly variable NCNN-containing 160-mer that

was added to assess biases in CpH contexts for 5-mC and 5-hmC modifications (see Figures 3.9, 3.10). In some cases, there was insufficient coverage over certain modified positions which hindered their proper analysis. Again the observations from the previous test run that incubation time increases the false positives and that mTet1 gives a higher conversion rate are confirmed in this test run. Furthermore, the new reducing agent pyridine borane improved conversion compared to picoline borane and has a lower error rate on average. Therefore the fifth sequencing run proceeded with the best condition that was mTet1 + pyridine borane (incubated for 16h) after this combination of condition showed the highest conversion coupled with lowest false-positive rate.

3.1.7 Fifth whole-genome run

As a result of the continuous optimisation, TAPS had achieved higher conversion rates (above 95%) and lower false-positive rates (1% $\leq$) in the second test run, where different chemicals (pyridine borane and picoline borane), incubation times, library preparation protocols (PTAT and directional) were tested. The condition selected for deeper sequencing was pyridine borane, a directional library with 16 hours of incubation. Four aligners were used to map the sequencing data to the combined spike-ins and mm9 genome: Bowtie2, BWA, BSMAP, and Bismark (only to look at mapping rates compared to WGBS). Several alignments runs in parallel of WGBS were performed: BSMAP and Bismark on mm9 + spike-ins + ϕX (only for the fifth run), Bismark on mm9 + spike-ins + ϕX on the fourth run. As an approach to decrease the loss of reads mapping to short reference sequences such as spike-ins, five Ns were added to the ends of the spike-ins for another run of testing as padding. Although these pads helped with keeping more uniquely mapped reads, they had the following issues: 1) the modified positions shifted right with five, 2) marked as modified positions

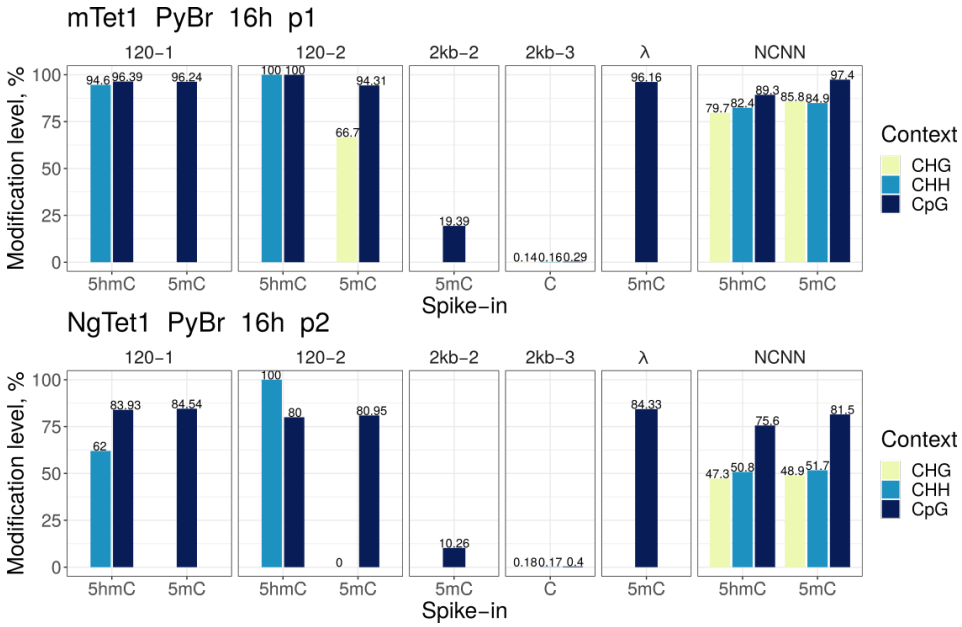


Figure 3.12: The conversion rates, as seen in the second test run showed by spike-in, context, and modification type. A complete conversion (β value of $\approx$ 1.0) was expected for the short 120-mer spike-ins and the λ phage. An additional half methylated at CpG 2 kb spike-in and a highly variable in context NCNN spike-in were added to study PCR bias and non-CpG conversion. The second test run primarily was used to compare TET enzymes and reducing agents. The mTet1 and pyridine borane condition was selected as the best performing one due to its higher conversion rates and lower false-positive rate. This condition was sequenced in the fifth whole-genome run.

that were not modified by design (on the 120-mers). Interestingly, such 'ghost' modified positions were seen after regular alignment to all spike-ins, only the 120-mers and using Bismark versions 0.19 and 0.20. These anomalies seemed to occur at shallow coverage (such as one read at a position), which could be an artefact of removing too many reads after alignment rather than incorrect read alignment.

The fragment size distribution for each of the aligners for 2.5x10^7 random reads for BSMAP, Bowtie2, BWA was plotted and showed that the general

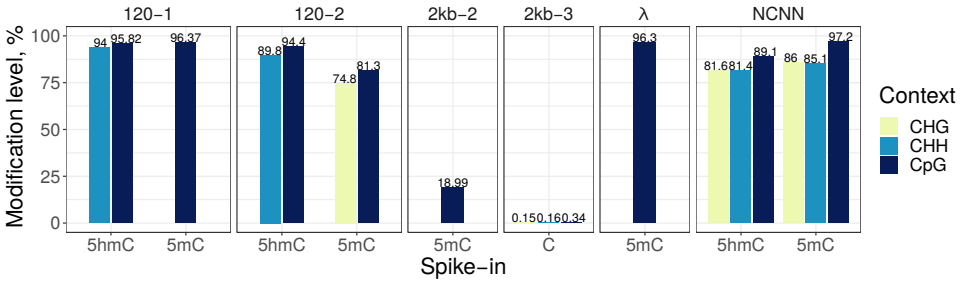


Figure 3.13: The conversion rates, as seen in the fifth whole-genome run showed by spike-in, context, and modification type. A complete conversion (β value of ≈ 1.0) was expected for the short 120-mer spike-ins and the λ phage. A 2 kb spike-in with 50% expected methylation showed lower values most probably as a result of an error during library preparation. The CpG methylation appeared to get a higher conversion in TAPS compared to non-CpG modifications and hydroxymethylation. The false-positive rate was significantly decreased in the fifth run due to the change in the conditions and reagents.

profile shape remained the same. Next, the total mapped to trimmed reads and the properly paired to trimmed reads were evaluated, and BWA and Bowtie2 had higher rates than Bismark as previously observed. The coverage of the short 120-mers appeared divergent between the mappers with Bowtie2 showing an entirely different coverage profile from BSMAP and BWA.

In the fifth sequencing run, WGBS contained the same spike-ins as TAPS. However, there seemed to be some quality issues in read 2 that were persistent after trimming and were observed in the reads that aligned to mm9. The FastQC reports suggested that there was a flow-cell issue (maybe because of low sequence complexity in the second read in the WGBS), where bases 'A' and 'G' formed loop shapes at read positions 60-72 bp. Another reason for the phenomenon could be a high cluster density. The higher false-positive rate (2% non-conversion rate) observed in the fifth run was probably due to the sequencing issue as the non-conversion rate which was low in the second test run (the samples were prepared from the same library). Given that the sequencing reads needed to

be trimmed 20 bp in order to remove the sequencing anomaly and this would result in reads with lengths of 50-60 bp which would make them harder to align correctly, it was decided against using the WGBS data from the fifth run and instead of comparing to the one from the fourth run.

The spike-ins showed high conversion and low false-positive rate as seen in the second test run with rates calculated as a total number of modified cytosines at a context divided by the sum of the total number of modified and unmodified cytosines (before these values were calculated as the beta values per position averaged per spike-in and context); see Figure 3.13). The change of calculation might help with lowly covered regions, which is less typical of spike-ins but can be observed in the genome and was directly comparable to Bismark's results.

As a proxy to evaluate genomic modification analysis in the new TAPS data prior to detailed analysis, the modification calls of chromosome 1 were used as a proxy. The mean modification value for WGBS was $\approx 64\%$, whereas, for TAPS, it was 61%.

3.1.8 Sequencing runs comparison

There has been a significant improvement in terms of conversion rates during the TAPS method development and optimisation process with an observable decrease in false-positive and false-negative error rates (see Figure 3.14). The enzymatically methylated at CpG positions bacteriophage λ spike-in was expected to be completely modified, and the observed differences from a modification value of β=1 were due to the incomplete chemistry (TET enzyme oxidation and picoline or pyridine borane reduction). During development, the false-negative rate of not detecting truly modified positions was decreased from 65.652% in run1 to 3.7% in run5. During that time the false-positive rate (measured as observed modification level at unmodified cytosines in the unmodified 2 kb spike-in) was

non-linearly decreased from 0.371%, 0.251%, 0.267% (CpG, CHG, CHH contexts) in run1 to 0.336%, 0.154%, 0.164% in run5, passing through 1-2.5% error rates in runs 2-4. The higher conversion observed in run5 was a result of the change in both the Tet1 enzyme (mTet1 against NgTet1) and reductive compound (picoline borane versus pyridine borane) which lead to complete reaction. Possibly, the observed false-positive rate of run1 could have been higher if the general conversion of the method was comparable with runs 2-4, which would indicate that the low false-positive rate of run5 compared to the others was a result of a lower rate of deamination by hydrolysis of unmodified cytosines to uracils due to the lower incubation temperature ($60 - 70°$C in runs 1-4 with incubation times of 1-8 hours versus 37°C and 16h incubation) [209]. The cytosine deamination under a physiological temperature is expected to be lower compared to higher temperatures irrespective of the pH of the solution.

The nucleotide content of the underlying DNA sequence appeared to affect both TAPS conversion and sequencing coverage. The genome of bacteriophage λ (which was used in all runs as a spike-in to assess conversion success) is roughly divided into two parts in terms of GC content—the fist 20 kb have a mean GC content of $\approx$ 57% and the remaining sequence $\approx$ 45%. The high GC content was negatively correlated with the observed CpG modification values (see Figure 3.16) with Pearson correlation ranging between -0.878 and -0.799 for run1 and run2, and $\approx$ 0.41-0.45 for the remaining three whole-genome runs (all p-values were $\leq 2.2e^{-16}$, Pearson's product-moment correlation, two-sided hypothesis, CI=0.95). The successful modification in the chemistry of the method has lead to a substantial increase in the modification rates at CpGs. They were measured in the bacteriophage λ spike-in, which was fully enzymatically methylated (methylation rates were tested by mass spectrometric methods before sequencing). The lowest conversion rate in run1 showed a Pearson correlation of

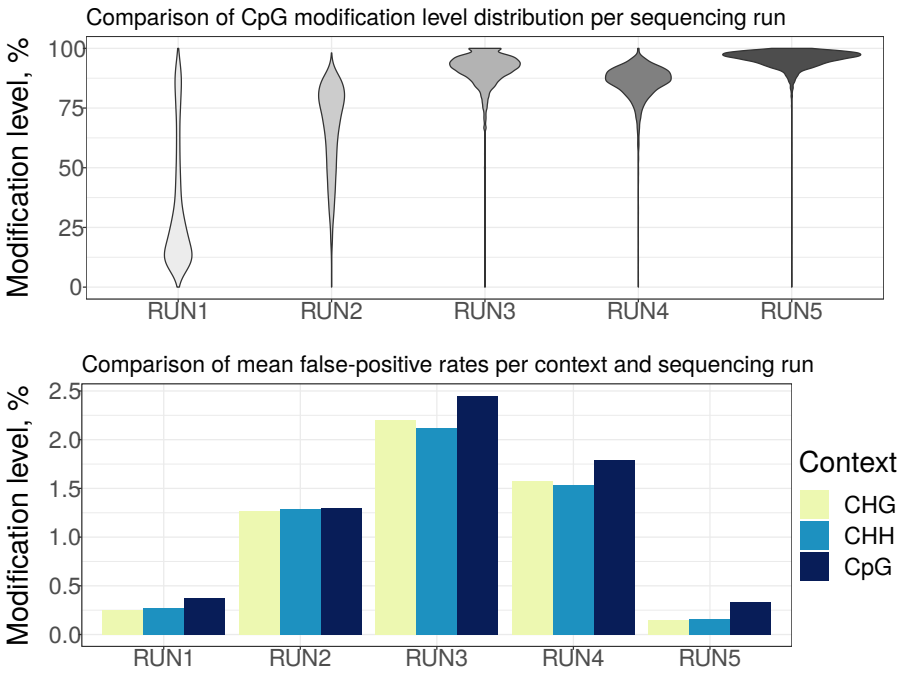


Figure 3.14: A comparative analysis between TAPS whole-genome runs in terms of their CpG modification levels distribution in the bacteriophage λ spike-in (top), and the false-positive rate at different cytosine contexts as measured in the unmodified 2 kb spike-in (bottom). A significant increase in TAPS conversion was observed after the first run, but the conditions of the fifth run ensured a high conversion and a low false-positive rate at milder conditions.

0.76 to the conversion rate of the second run (see Figure 3.18). In contrast, its correlation with the remaining runs conversion values was low in the range 0.27-0.35, and was lowest with run5, while the correlation between runs 3-5 was very high $\approx$ 0.86-0.89 (all p-values were $\leq 1.063e^{-9}$, Pearson's product-moment correlation, two-sided hypothesis, CI=0.95). The higher conversion rates in the later whole-genome sequencing runs, however, were negatively correlated with coverage at the high GC region in the bacteriophage λ spike-in (see Figure 3.17).

The Pearson correlation of coverage and GC content was performed in 100 bp bins, where the highest negative correlation was observed in run1—0.361. At the same time, for the remaining runs, it was lower in the range -0.7384-0.842 (all

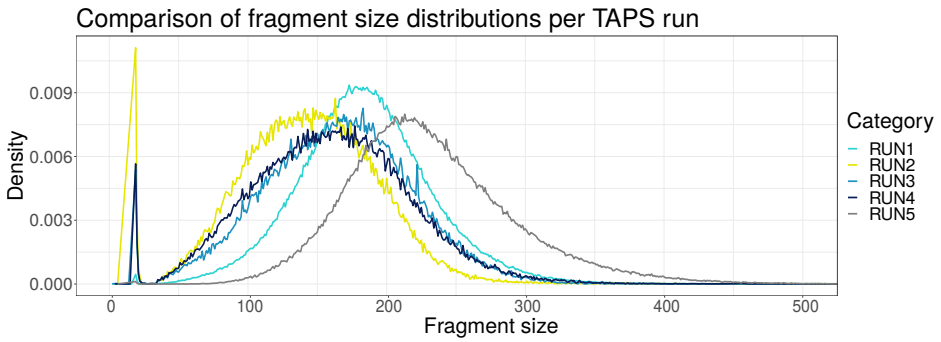


Figure 3.15: A comparative analysis between TAPS whole-genome runs in terms of their fragmentation profiles. The third and the fourth runs showed very similar fragment distributions, while runs 1-2 and 5 had specific distributions. The fifth run was size-selected for the longer fragments to decrease the overlap between the pair-end reads.

p-values were $\leq 2.2\text{x}10^{-16}$, Pearson's product-moment correlation, two-sided hypothesis, CI=0.95). At first, it was suggested that observed coverage unevenness was primarily due to alignment, because of the different fragment size distributions (see Figure 3.15) that might have contributed to the lower alignment rates. Whether the difference of sequencing coverage was primarily an alignment issue was tested with BSMAP which allows for the wild-card alignment of both modified and unmodified cytosines. However, the BSMAP alignment rates did not change significantly the observed coverage suggesting that the uneven coverage was a product of library preparation and not bioinformatic data analysis. A possible explanation was that in the fully converted high GC region of bacteriophage λ the sequence after TAPS conversion and PCR would be primarily AT ($\geq$ 70% with long stretches of TG and AC depending on the read orientation), but also the very high density of CpGs and the DHU may influence the activity of the polymerase.

The uracil-insensitive polymerases used in bisulfite sequencing usually read through converted non-CpG regions (at least in vertebrate animals) and the remaining opposite to the origin strand cytosines and guanines may require

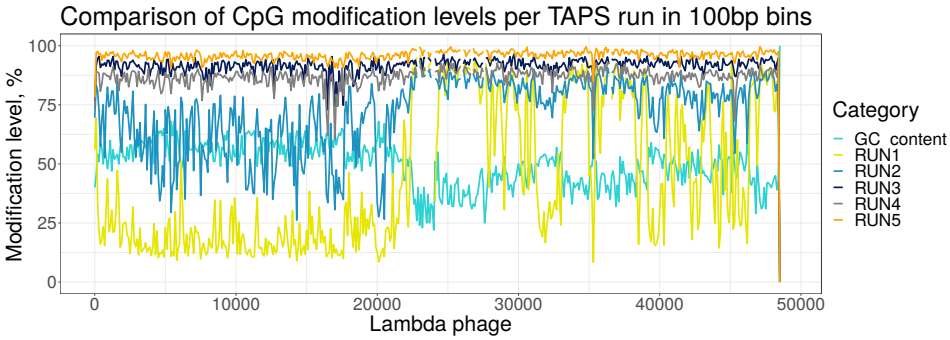


Figure 3.16: A comparative analysis between TAPS whole-genome runs in terms of their CpG modification level (a proxy for conversion) and GC content in 100 bp bins in the bacteriophage λ spike-in. There was a pronounced negative association between GC content and TAPS conversion (measured by the CpG modification level) in the first two runs.

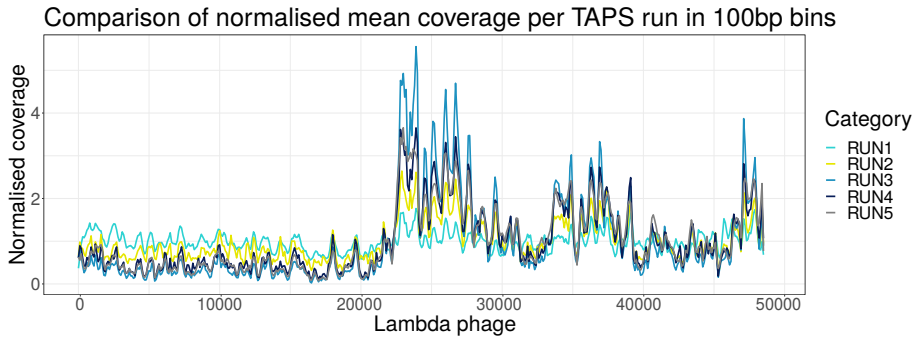


Figure 3.17: A comparative analysis between TAPS whole-genome runs in terms of their average sequencing coverage normalised per 100 bp bins in the bacteriophage λ spike-in. For visualisation purposes and to create similar profiles between runs, the individual coverages were divided by the mean coverage observed in each sample. The lower GC content in the second half of the λ phage genome was associated with a higher sequencing coverage.

slightly different reaction conditions such as higher temperature. Furthermore, DHU which is the molecule resulting from the reduction of 5-caC with a picoline or pyridine borane has a slightly different conformation than uracil which could affect the polymerase processivity particularly in higher densities of modified positions. Thus, it is expected that the TAPS method's original chemistry would be biasing for unmodified sequences and not overly high modified position

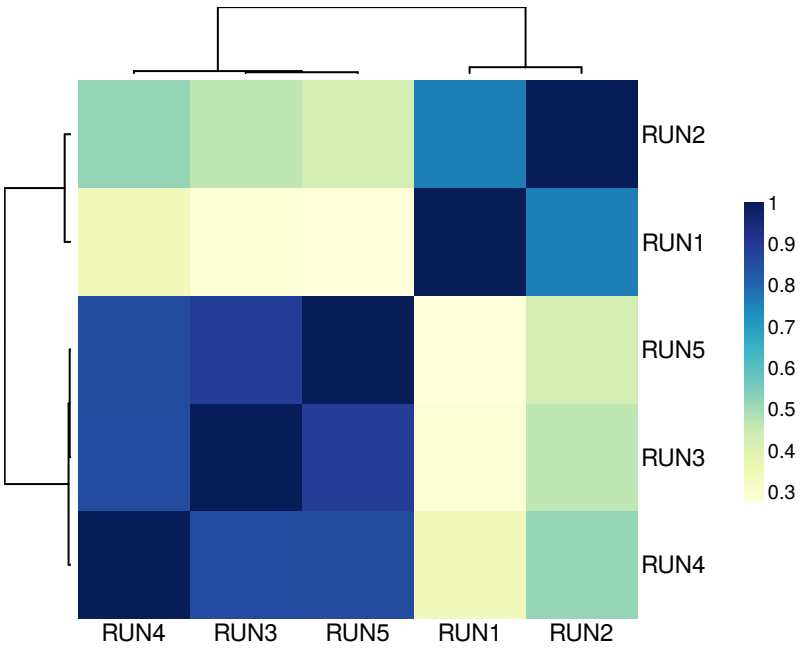


Figure 3.18: A visualisation of the Pearson's product-moment correlation of CpG modification levels in all five whole-genome runs as observed in the bacteriophage λ spike-in. Runs 1-2 and 3-5 showed the highest correlation of observed conversion rates. As expected, the first two runs were dissimilar to the subsequent runs with improved reaction conditions.

densities, which would be the opposite of the biases observed in WGBS. Ideally, a laboratory test will be needed to evaluate the maximum density of modified positions where more substantial bias in coverage occurs and whether using a high fidelity polymerase, particularly for AT-rich sequences could handle the bias.

3.2 Development of the modification calling computational pipeline

The following section discusses the computational tests and simulations which took part during the development of the tool for aligning and calling modifi-

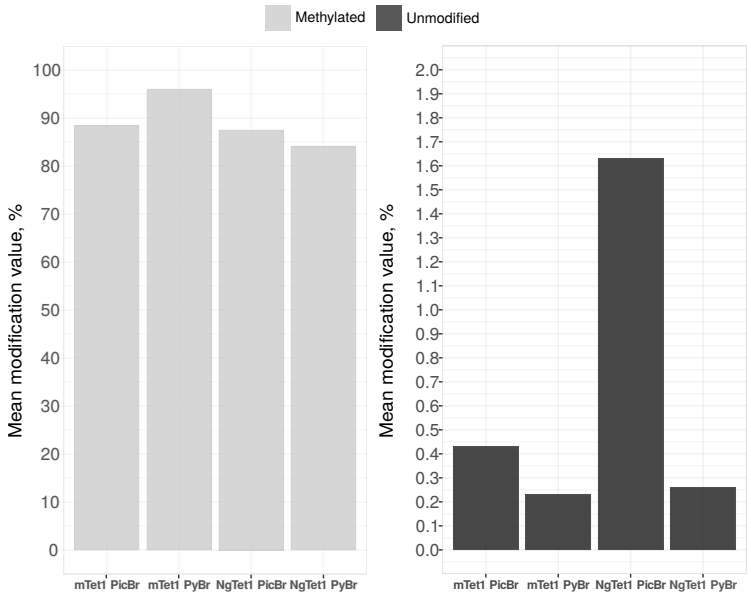


Figure 3.19: The barplot shows a comparison of TAPS performance as a combination of a Tet1 enzyme (mTet1CD, NgTet1) and the reducing agent (picoline or pyridine borane). The combination of mTet1 and pyridine borane used in the fifth whole-genome sequencing run showed the highest conversion and the lowest false-positive rate of all conditions. Conversely, the combination used for the first four runs (NgTet1 and picoline borane) had a relatively lower conversion and the highest observed false-positive rate.

cation on TAPS and WGBS data asTair. TAPS sequencing data differs from both WGBS and WGS data to a different degree which required the creation of a TAPS-specific data analysis pipeline. The main steps of the data processing pipeline are common between TAPS, WGS, and WGBS, but there are certain modifications in terms of alignment and filtering and modification calling.

3.2.1 Alignment of TAPS data

During the analysis of the whole-genome TAPS data from the fourth run, it was found that some genomic regions are covered in TAPS but not covered in WGBS (see Figure 3.20). The difference can be explained as a result of both alignment and the difference in chemistry. However, without further experiments, the role

of the controllable factor—alignment could not be determined. Furthermore, during the development of the TAPS method BSMAP, BWA, and a combination of both aligners were tried empirically on the data (first round with BWA and unmapped reads are aligned with BSMAP on the second step of alignment), suggesting that applying only BWA with the default WGS settings does not result in losing a significant proportion of the reads. Still, aligners were not evaluated on a testing dataset prior to processing of the TAPS sequencing runs. In order to study alignment performance on TAPS data, a genomic test region of 2 Mb was selected as a testing dataset (more information on the simulation experiment can be found in the following methods subsection).

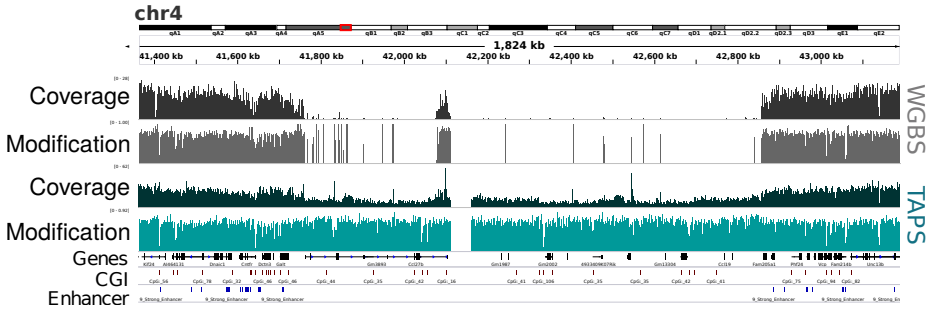


Figure 3.20: An IGV screenshot of a 2 Mb region in chromosome 4 showing that TAPS has a sequencing coverage in contrast to WGBS. The region was selected due to the large difference in coverage between TAPS and WGBS when visualised in IGV. The difference was found to be a result of read filtering during the alignment step of Bismark.

With a 100 bp pair-end mappability (normalised to 2 Mb) of ≈ 0.6 (0.581 UMAP and 0.579 BISMAP averaged scores) and a proportion of repetitive regions spanning this region, the overall mappability of the selected region was below the mean for both the mm9 and hg39 genome in bins of the same size (see Figure 3.21 top left). The decision to use this region (chr4 41 000 000 to 43 000 000 bp) was due to the high variability in terms of both coverage and modification levels between the TAPS and WGBS methods observed in a larger

part of the region (see Figure 3.20). Although TAPS had shown higher and more even coverage than WGBS, the possibility that some of the differences observed between the two methods were a result of the choice of read filtering and the aligner remained. Bismark is a tool commonly used for bisulfite sequencing data processing which was the reason for its selection for WGBS data pre-processing. However, the suboptimality of some of the algorithm parameters was observed while working on previous projects. The following parameters appear to lower the expected mapping rates independently of the chemistry of the method used:

1. extraction of aligned reads—in the mapping report output by Bismark the alignment rate is higher than what can be calculated with other methods (such as samtools flagstat) because some reads are removed from the BAM file after they were included in the report. These can commonly be reads mapped to short references as they have higher chances of being removed as the context at their first two positions and last two positions cannot be extracted. Furthermore, adding padding to the ends of the short sequences results in position shifting and miscalling seen on short spike-ins;

2. filtering of uniquely aligned reads based on their AS and XS tags [15]—reads tagged as aligning to multiple places in the genome are removed as the default setting which may be too stringent from a biological perspective where variation is expected and cannot be overridden.

In the light of studying optimal alignment and filtering in TAPS and the extent to which TAPS and WGBS differ due to alignment, the performance of several aligners was evaluated: Bowtie2 (BWT2), BWA, BWA meth, and BSMAP (TAPS), and BWA meth, BSMAP, Bismark (BSMK) + Bowtie2, and Bismark + HISAT2 (WGBS). Furthermore, three different filtering options were tested on the TAPS data in order to remove sequencing read that aligns non-

uniquely to the genome. The removal of multialigners was as follows:

1. BWA alignment excluding reads with tags 'XA:Z:' and 'SA:Z:'—the XA tag labels reads with alternative hits and show their location in the genome, and their CIGAR string and Edit distance [8], and the SA tag labels chimeric (split reads) [17]. Chimeric reads occur when a sequencing read aligns to distant regions in the genome and can be a sign of structural variation.

2. BWA alignment excluding reads with MAPQ=1; empirical evaluation of the MAPQ distribution of the incorrectly aligned reads suggested that on average the incorrectly mapped reads have lower MAPQs (BSMAP does not have a MAPQ scale), but in Bismark, this rule does not apply.

3. Bowtie2 alignment keeping uniquely mapped reads with tags 'XS:i:0' and 'AS:i:0'. AS and XS are alignment scores (primary and secondary), where the XS value can be >0 when the read has more than one possible alignment [4]. The scoring procedure of Bowtie2 starts with a search for an optimal alignment for a read, which continues after such an alignment is found to calculate the read mapping quality and to assign the alignment score tags. The same filtering where only reads with tags 'XS:i:0' and 'AS:i:0' are kept is used internally in Bismark during alignment [15].

The most important metrics for successful alignment in both TAPS and WGBS concern both the correctness of read coordinates after mapping as well as the preservation of the actual modification level at a cytosine position. In some cases, the misalignment may not change the modification level but would decrease the information content of the region in terms of natural genetic variation. However, in the following experiments, no preservation of positional SNP information was considered as a performance metric. In order to compare aligners in

their alignment sensitivity, alignment rate, and influence over the modification levels, WGS data were simulated with ART over the lowly mappable region of interest. The simulated data were created to be as close as possible to the real data at hand: with a mean coverage of 20x, specifying the NextSeq500 sequencing platform errors, which was then modified as 0, 20, 50, 80, 100% of all CpGs in the region applying TAPS- and WGBS-like conversion at the same positions. After the modifications were inserted, the read information concerning the exact read names and positions was extracted. The reads were sorted by query names and written as pair-end FASTQ files. Afterwards, the reads are aligned with several aligners, and the modification calls for BWA meth, BSMAP, and Bismark are done using the default scripts for each tool. In contrast, for TAPS, we used our custom pipeline of mpileup and calling, which will be presented in the next subsection.

In terms of alignment, the majority of the aligners had total alignment rates (the ratio of aligned reads regardless of their position in the genome to all available reads for alignment) of $\approx$ 40-50% suggesting that the high density of repetitive regions and the lower mappability interfered with alignment regardless of the method and the proportion of modified CpGs (see Figure 3.22 left). The highest observed total alignment rates were for BSMAP and Bowtie2 ($\approx$ 60%), followed by BWA with 55%. The lowest values were detected in all methods using the AS+XS unique read filtering (Bismark and Bowtie2 + additional filtering) where the alignment rate was 20%. If an exact matching of the read position after alignment and the truth set is required, of the total aligned reads the majority of the aligners, and filtering methods showed that only around 50-60% of the reads were correctly mapped (see Figure 3.22 right). The best performer was BWA with a MAPQ filtering which was tested before including in the experiment where it was discovered that the majority of the reads with

MAPQ=0 are incorrectly aligned (see Figure 3.21 top right). None of the remaining filtering techniques decreased the proportion of incorrect reads when compared to the default settings of the aligner. BSMAP and BWA meth showed similar alignment rates for both TAPS and WGBS, and the default settings of Bowtie2 showed similar results with BWA when reads with XA and SA tags are removed. The AS+XS unique read filtering used by Bismark (regardless of whether it wraps around Bowtie2 or HISAT2) resulted in having only 10% correctly aligned reads of the total available reads. At the same time, BWA with the MAPQ filter preserved 40% of the reads where almost 100% of them were correctly aligned (the remaining did not align at all). These discrepancies suggested that the coverage of bisulfite sequencing experiments (particularly in hard to map regions) can be underestimated. The higher proportions of modified cytosines had minimal influence over the alignment rate suggesting that the leading cause of low mappability was the underlying DNA sequence of the region, due to the alignment of short reads.

Another metric that can be used to compare the performance of aligners on WGBS and TAPS data concerns CpG coverage. The highest error rate of CpGs that were covered outside of the genomic region of interest to all covered CpGs was seen in BSMAP on WGBS with 27%, followed by Bowtie2 and Bowtie2 with AS+XS tags filtering both with $\approx$ 18% and for the remaining aligners the error rate was in the range 1-5%. It appeared that BSMAP makes more mistakes on WGBS than TAPS, and Bowtie2 filtered or with the default settings makes more errors than BWA filtered by tags or unfiltered. These results further confirm the observation that the choice of aligner can influence the observed modification values, and aligners that appear to align more can make more mistakes, particularly in lowly mappable regions. Also, the BSMAP error rates on CpGs can be treated as another confirmation that WGBS data may prove

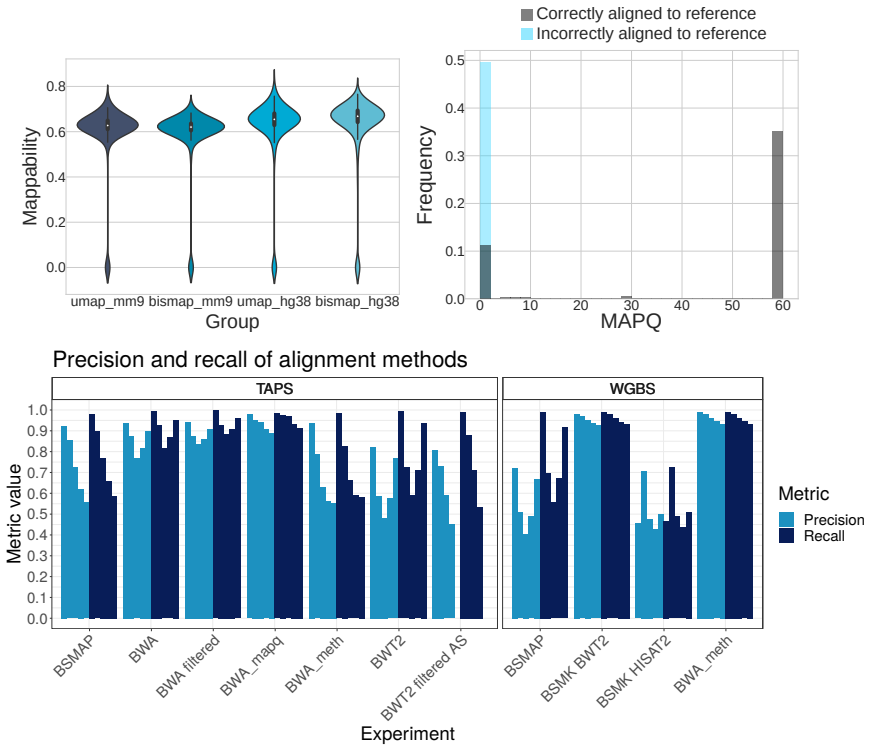


Figure 3.21: A visualisation of the averaged mappability for the regular and bisulfite converted genomes at 2 Mb windows (top left), together with the observation that for BWA aligned reads removing reads with MAPQ=0 removes false positives at 50% CpG modification in the lowly mappable region of interest (top right). The frequencies of the false and the correct reads sum to 1 for visualisation purposes. The performance of different aligners in terms of similarity to the true CpG modification (absolute allowed error of 1% difference) is given in the bottom panel. The different CpG modification proportions (0, 0.2, 0.5, 0.8, 1.0) are grouped together as five consecutive bars with metrics starting with the unmodified dataset. The correctness of alignment was influenced as expected by the modification level of the reads.

to be more difficult because of its higher proportion of converted positions than TAPS.

The purpose of applying epigenomic methods is to uncover the epigenetic patterns in the underlying data. In this case, the primary objective of this computational experiment was to get an estimate of how many of the modified CpGs in the tested regions and to what extend have preserved their modifica-

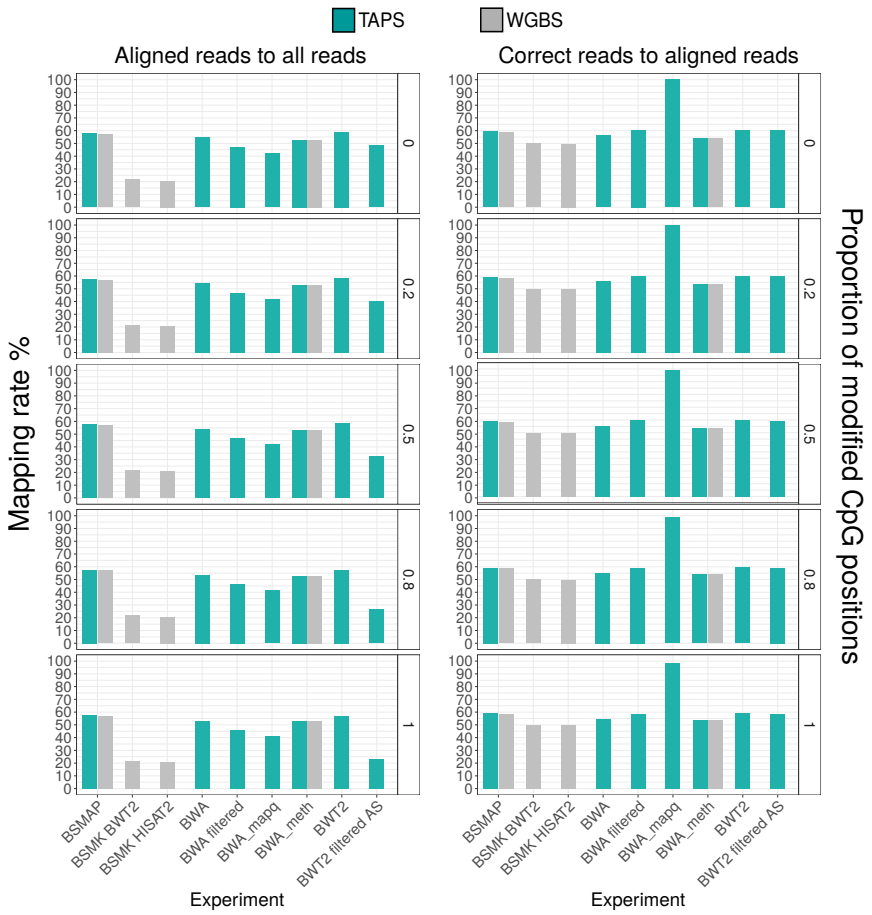


Figure 3.22: A comparison between the performance of different aligners and filtering procedures on simulated TAPS and WGBS data. The total mapping rate (left) which is defined as the ratio of the aligned reads to all reads, and the correct alignment rate (right) that is the number of exactly matching reads by read name and staring position to all alignment reads is presented at five different modified CpGs proportions: 0, 0.2, 0.5, 0.8, 1.0. The choice of the aligner and read filtering method influenced the alignment rate more than the chemistry of the method.

tion values after alignment. To this end, the absolute difference between the truth set modification levels, and the observed values were measured. The CpG modification values were considered to have been preserved if the error was $\leq$ 1%.

The highest precision and recall rates in terms of preserved modification lev-

els were seen in BWA (filtered or default), Bismark (wrapping around Bowtie2), and BWA meth (only on WGBS) (see Figure 3.21). In BWA-meth (WGBS), Bismark (wrapping around Bowtie2), and BWA (with MAPQ filtering) precision and recall decreased slightly with the increase in the proportion of modified CpGs; the same pattern was observed in BSMAP, Bowtie2 with the AS+XS filter, and BWA meth on TAPS but with a steeper difference. The remaining aligners showed the worst preservation performance when 50% of the CpGs in the region were modified with an increase in the lower and higher modification values. Most probably, this was a result of read substitution that has a less pronounced effect when the reads are more homogeneous in terms of modification values. An outlier was Bismark wrapping around HISAT2 where the precision and recall were around 0.5, excluding the case when 20% of all CpGs were modified. The Bowtie2 AS+XS tag filtering produced the worst results in TAPS and these results can be used as a final argument against using this unique read filtering in TAPS. Also, a comparison of the modification value distributions showed that BWA has closer to the expected distribution than Bowtie2. In order to obtain a single metric to measure the performance of aligners preserving CpG modification levels on TAPS and WGBS, the Mathew correlation coefficient (MCC) was calculated. The Mathew correlation coefficient was used to represent the confusion table of a classifier. It is a preferred metric because of its lack of bias towards classes. A high Mathew correlation coefficient can be obtained only when both the positives and the negatives were correctly classified [72]. The highest MCC scores were calculated for BWA meth on WGBS (99-96%, lower values observed at higher modification proportions), Bismark wrapping Bowtie2 (99-96%, lower values observed at higher modification proportions), and BWA with MAPQ filtering on TAPS (98-93%, lower values observed at higher modification proportions). These results correspond to the high precision and recall

values measured for these methods.

The experiments with simulated TAPS and WGBS data confirmed the importance of the choice of the aligner. The observed coverage differences between TAPS and WGBS can be partially explained by the application of Bismark filtering on WGBS data which removes a large proportion of reads that can be aligned. BWA and BWA meth showed higher mapping rates compared to Bismark and performed on par in terms of the preservation of cytosine modifications. Previous studies on highly repetitive regions [337] showed that BWA could align with high correctness to such regions. Still, it was influenced by the read length and the type of repetitive regions. Finally, in all cases where a WGBS-specific aligner was applied on TAPS data, the results showed worse performance on TAPS compared to the same methods applied to WGBS data, or when WGS aligners were used on TAPS. This confirms the hypothesis that alignment and modification calling in TAPS requires custom tools and WGS-modified processing pipelines.

3.2.2 Modification calling in TAPS data

As a result of the computational methods presented in the previous subsection and the need for the development of a TAPS-specific processing pipeline, a tool called asTair was created. asTair was purpose-built to handle TAPS data, but its functionalities were expanded, and it currently contains functions that handle WGBS data. In terms of speed, asTair runs 3x faster than Bismark (as measured on v2), and allows whole-genome or per chromosome analyses, and its detailed modification level outputs provide information on both CpG and CpH positions and can be directly visualised and interpreted. The main functions in the tool are the aligner and the modification caller. Given that BWA meth (without any filtering) performed on par with BWA without filtering and it

was implemented as part of our tool asTair for the alignment of WGBS data, while BWA was selected for TAPS (both methods use the MAPQ≥1 filter). Initially, the modification calling pipeline used external splitting of reads into the original top and original bottom reads, and external removal of overlapping ends of pair-end reads which were removed with the clipOverlap tool. Then the files were split by chromosomes and changes in positions were counted with samtools mpileup, the resulting VCF files were then processed with a Python script to get the modification values. The modification calling on TAPS (and directional WGBS libraries) uses the information about the origin of the read (coming from the original top or bottom strand compared to the reference) to deduce whether a conversion has taken place at a position of interest. A read pileup where reads pass quality filters was divided by their base and read flag (the initial version appropriately used paired reads only as they make up to ≈ 99% of all aligned pair-end reads). The individual bases and their flags are counted, and the modification ratios are calculated according to the method conversion chemistry. In asTair v3 all steps from the aligned reads to the output of modification calls run internally using Python.

3.3 TAPS data analysis of mESC E14 line

After the successful optimisation of the TAPS method protocol, the data from the fifth TAPS whole-genome run was compared against WGBS sequencing data where the DNA input was precisely the same. Part of the results in this section was published in [216].

The TAPS data showed a high alignment rate and higher base qualities compared to WGBS (see Figure 3.23). Also, the sequencing-only error provided by the sequencer on ϕX (TAPS read1 0.25%, read2 0.38%; WGBS read1 1.25%,

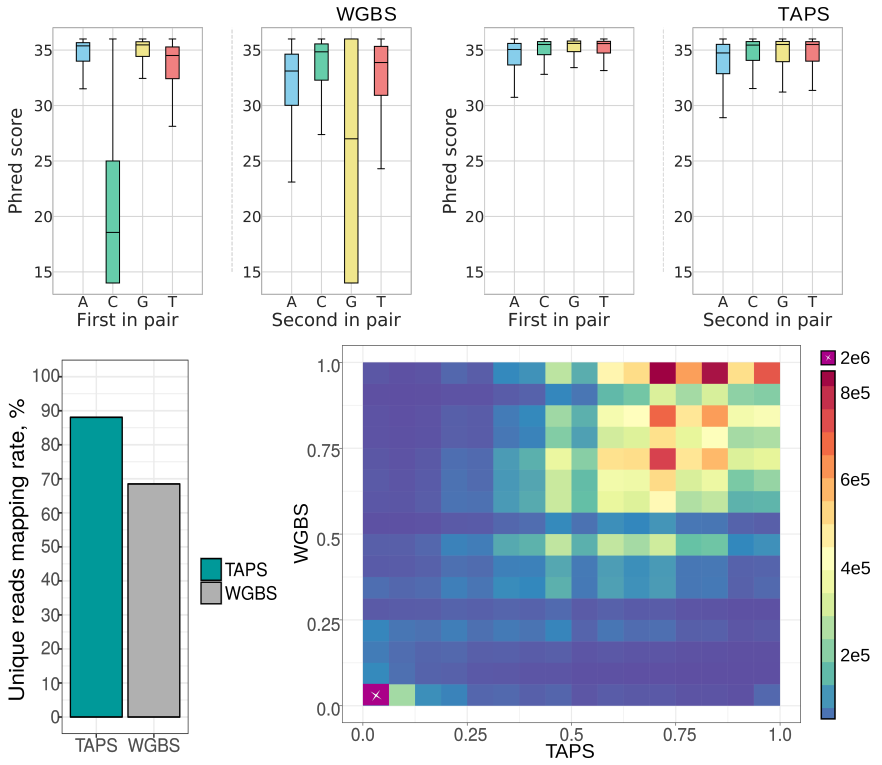


Figure 3.23: Summary of the main results from the comparison of TAPS and WGBS data: TAPS base qualities of the bases which are converted during library preparation do not differ between the first and second in pair strands in contrast with WGBS (top) the figure shows mean base quality per read for 10^7 randomly selected sequencing reads; TAPS showed higher unique mapping rate as a result of both alignment and chemistry (bottom left); there was a high correlation between TAPS and WGBS in terms of modified CpGs (bottom right).

read2 1.1%) showed a higher error rate as a result of the chemistry of the method that would affect base qualities and alignment. The unique alignment rate (the ratio of the uniquely filtered reads as explained in the previous section to trimmed reads) was 88.08% for TAPS and 68.49% for WGBS (the mapping rates without filtering were 68.49% and 99.57%, and after the removal of PCR duplicates 63.15% and 81.31% for WGBS and TAPS correspondingly). The lower WGBS uniqueness rate (and the lower T or A base quality scores in the first in pair read and the second in pair read) was a result of both alignment and the

requirement to sequence 5-20% non-bisulfite treated sequences which are usually sequences from ϕX. After repeating the alignment and the removal of multi-mapping reads on WGBS using asTair '–method CtoT' the unique alignment rate was 78% (mapping rate 98.31% and unique mapping rate after removal of PCR duplicates 71.88%). The difference between the two alignment rates was due to Bismark conservative filtering of reads which potentially could be aligned incorrectly. The results from asTair align with filtering suggested that TAPS has a $\approx 10\%$ higher alignment rate due to its chemistry. The higher alignment rates (both with Bismark and asTair) showed that TAPS has a higher mean genomic coverage than WGBS (≈ 13 and ≈ 21 for WGBS and TAPS respectively using Bismark as showed in Figure 3.24 bottom, and 16.85 for WGBS using asTair). In order to evaluate whether TAPS has a more even coverage compared to WGBS given the chemistry and the alignment rate of the bisulfite sequencing method, the TAPS sequencing data were downsampled to the same median coverage of 13. TAPS showed more even coverage with a mean of 12.411 and variance of 482.24 (IQR=9) while WGBS mean of 13.078, and the variance was 1988.24 (IQR=11). The original analysis showed that the mean sequencing depth of CpGs in both WGBS and TAPS followed a similar distribution to the per base distribution (see Figure 3.24 top) and was lower for WGBS. An additional analysis on CpH positions showed similar pattern (see Figure 3.25).

The next important comparison between TAPS and WGBS was in terms of conversion and similarity between the datasets. TAPS has covered 88.3% of all 42,741,339 CpG sites in the genome of the E14 cell line, while for WGBS, the figure was 77.6% (see Figure 3.26). Given the difference in coverages between TAPS and WGBS, only CpG positions with coverage of at least three reads were used in order to increase the reliability of the modification values used in the comparison. Using this coverage cutoff, the overlap between the two

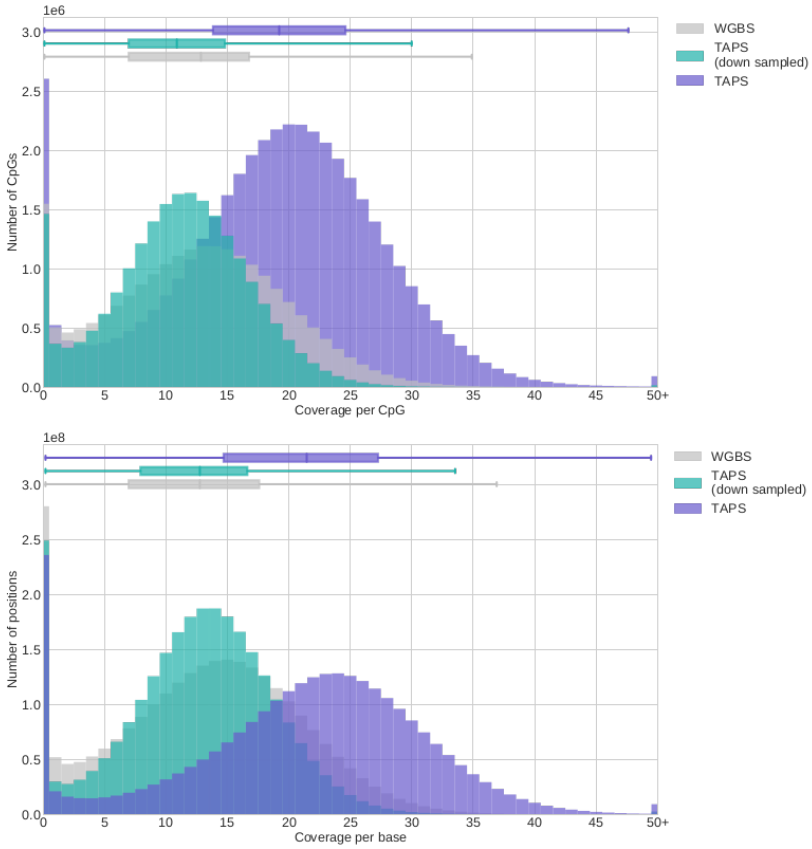


Figure 3.24: A comparison of TAPS and WGBS in terms of per base (bottom) and per CpG sequencing coverage (top). Due to its higher coverage, the TAPS sample was downsampled to the same median coverage as the WGBS sample. Positions with coverage higher than 50 were showed in the last bin.

methods was high, with a total of 32,428,076 CpG positions. Next, in order to answer the question of whether the modification states between TAPS and WGBS are the same given that they used the same input DNA, the CpG positions were divided into being unmodified and modified based on their β value. The cutoff to consider a CpG position as modified was ≥ 0.1 as described in [363] which should exclude the majority of sequencing and alignment errors, but would not penalise TAPS which has a slightly lower conversion rate (particularly for 5-hmC and non-CpG contexts). This threshold showed that 97.0%

of the CpG with coverage of at least three reads had a matching modified or unmodified state between TAPS and WGBS, were 96.2% of the CpGs fulfilling the coverage and modification level criteria in WGBS were also seen as modified in TAPS, suggesting that the two methods are comparable and highly similar. Indeed, a high per base correlation was observed between TAPS and WGBS on reads with coverage of at least 3x (Pearson ρ=0.68, P<2x10^{-16}, two-sided Pearson correlation test; see Figure 3.23 bottom right). Even though a higher correlation was to be expected given the same DNA source and the high TAPS conversion, the observed correlation confirmed that the two methods output similar results in terms of CpG modification levels. For example, WGBS technical replicates from [126] showed a Pearson correlation of 0.82 (p-value<2x10^{-16}, two-sided Pearson correlation test). Furthermore, other bisulfite-free methods have shown correlation values per bins and not per CpGs. When the method described in [306] where modification values normalised per 10 kb genomic windows were compared, WGBS and TAPS showed a Pearson correlation of 0.88 (Pearson ρ=0.88, p-value<2x10^{-16}, two-sided Pearson correlation test; see Figure 3). However, coverage alone cannot explain completely the observed per base pattern where positions fully modified in WGBS are modified at 0.7-0.8 β level in TAPS and *vice versa*. Even after downsampling of TAPS to the same median coverage as in WGBS, differences in terms of modification levels were still visible although a greater agreement exists in terms of fully methylated positions (see Figure 4). A possible explanation for the observation could lie in the combined effect of PCR bias towards modified fragments in WGBS and the incomplete conversion of TAPS under certain circumstances and possibly in particular sequences in the genome.

A particular group of interest was a subset of CpGs which showed modification values of $\geq$ 0.8 while their matching positions in WGBS had modification

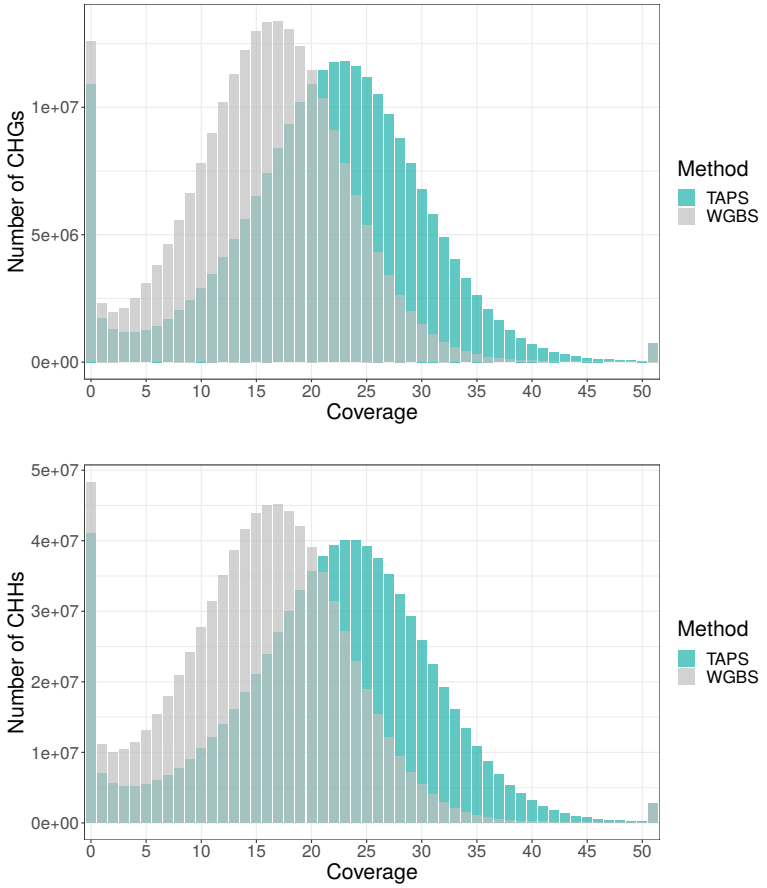


Figure 3.25: A comparison of the CpH coverage distribution in TAPS and WGBS: the distribution of CHG positions is showed in the top figure and the CHH on the bottom. Positions with coverage higher than 50 where showed in the last bin.

values of ≈ 0. A restriction enzyme HpaII assay was prepared on selected highly modified only in TAPS CpGs where the expectation was that HpaII would cut the second cytosine in the CCGG context if it is methylated. The list of CCGG positions prepared for testing included only positions where TAPS had a modification value of 1 and WGBS of 0 after any possible alignment artefacts were excluded as an explanation for them (removal of blacklisted positions in the mm9 genome and read filtering based on MAPQ). An initial laboratory test was conducted using the selected CCGG positions, which suggested that these

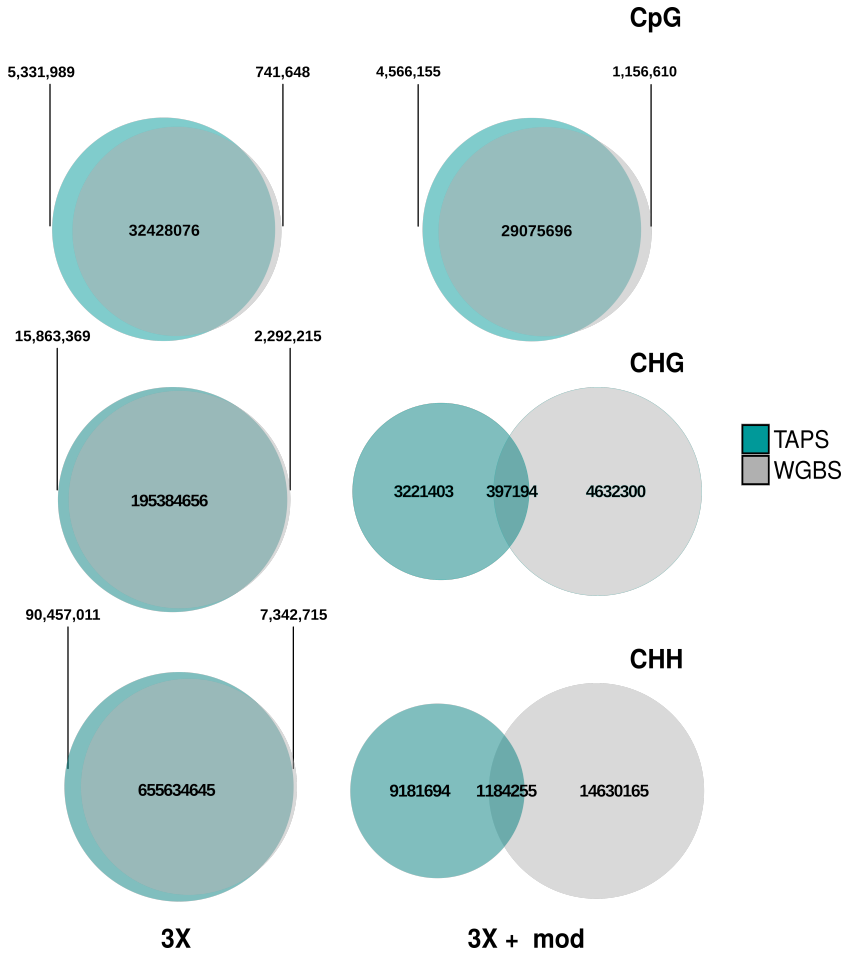


Figure 3.26: A comparison of the coverage of TAPS and WGBS CpG, CHG, and CHH positions at ≥3x coverage (left) and having ≥3x coverage and a modification value of ≥ 0.1 for cytosines to be considered modified (right).

positions were truly highly modified in TAPS only. However, after running a computational experiment to test whether the TAPS-specific highly modified CpGs were enriched in any genomic feature (the comparison was against the mean count of overlap of 100 random draws with repeats of CpGs with coverage ≥ 3x spanning the features of interest) no clear pattern of higher density was seen in any of the regions (see Figure 3.27). The TAPS-specific highly modified CpGs appear to have a random dispersal in most features or were depleted in

promoters, CGI, and enhancers. These patterns suggested that the observed positions were SNPs as CGI-rich sequences are expected to be mutated less frequently due to lower methylation levels and the CpG density [302]. In order to test this hypothesis, specific SNPs for the cell line E14 were obtained (from publicly available WGS data and E14 genome reference [153]) and were intersected in a subset of CpGs which were highly modified only in TAPS. These results confirmed that the large proportion of these CpGs with $\beta > 0.8$ were SNPs and they were removed from the final analysis.

After removing the known E14 SNPs, the next step of the analysis was to look into non-CpG modification. Despite the fact that the current TAPS protocol is not optimised for non-CpG modifications, its higher coverage together with the fact that the input DNA was from mESC cells could potentially show new non-CpG positions. For this analysis, the '3-C' filter (removing reads with over thee modifications at non-CpG positions) was not applied to the WGBS data. TAPS covered more CHG and CHH positions at $\geq$ 1x and $\geq$ 3x than WGBS, where 98.84% of the CHG positions covered at least by three reads in WGBS were covered by TAPS, and 98.89% of the CHH positions under the same conditions (see Figure 3.26). Generally, TAPS has significantly lower modification values (p-value $< 2.2e^{-16}$ Kruskal-Wallis test; see Figure 3.28). However, when the modification threshold of 0.1 was used for the non-CpG modifications, the overlap between WGBS and TAPS decreases, which was most probably due to TAPS conversion rate on non-CpGs. Interestingly, the cytosine positions with coverage of at least 3x and β value of at least 0.1 that was shared between TAPS and WGBS show Pearson correlations of 0.85 for CHG and 0.81 for CHH (p-value$<2x10^{-16}$, two-sided Pearson correlation test).

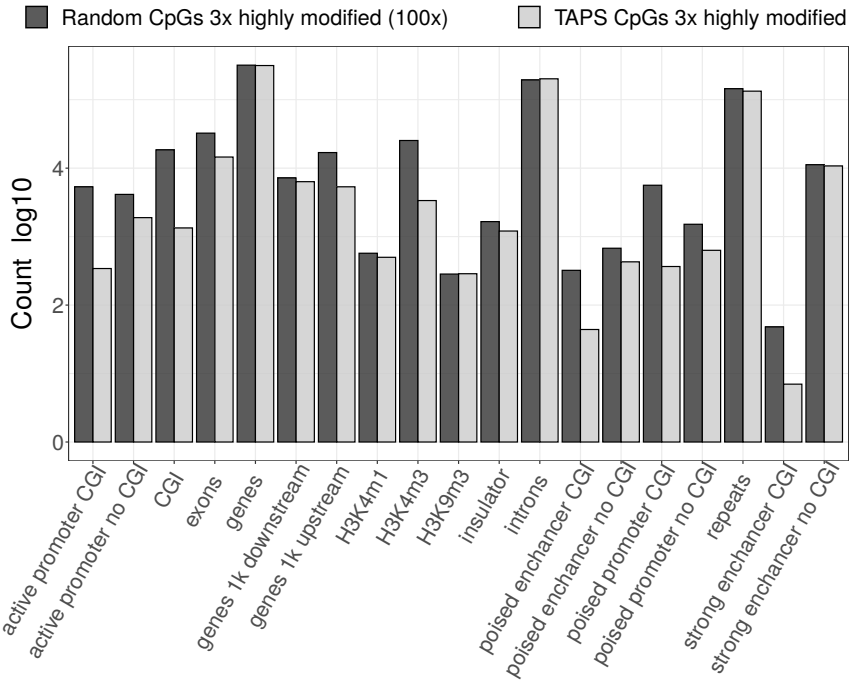


Figure 3.27: A comparison between the distribution of TAPS-specific highly modified CpGs to 100 randomly drawn CpG sets and their intersection with genomic features. The CpG positions that appeared highly modified only in TAPS were present in CGIs and certain regulatory regions at a lower rate than what is expected by random. These results suggested that the positions of interest could be SNPs.

3.4 Discussion and future directions

The TAPS sequencing method can be viewed as both a general-purpose method and a technique for specialised applications. TAPS can be used as a direct substitute for bisulfite sequencing in studies where the main focus is CpG methylation which covers a vast range of applications in biomedical sciences. The vital points of TAPS are the lower DNA damage together with more efficient conversion only of modified positions that would allow for both studying genetic variants and modifications. For applications such as cfDNA and single-cell sequencing, TAPS could prove useful due to its ability to preserve low initial DNA inputs. Long-

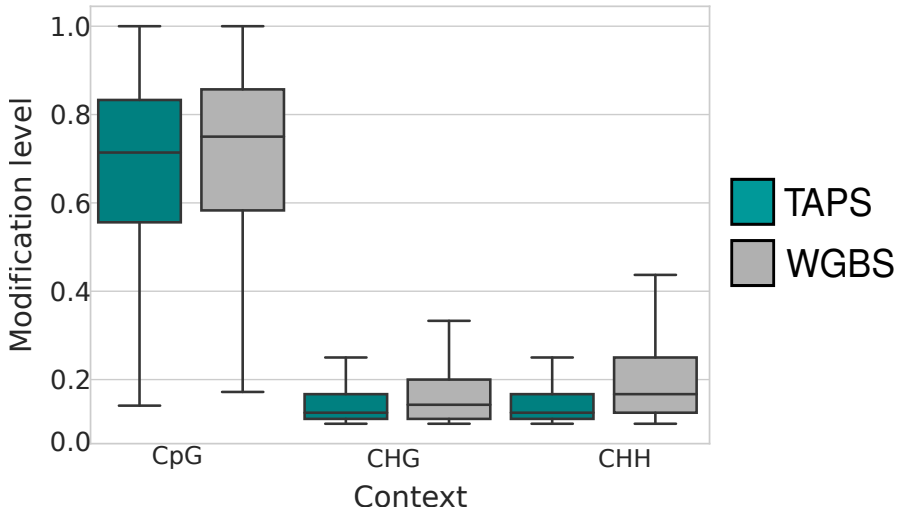


Figure 3.28: These boxplots show the comparison of the distribution of modified cytosine positions (having a modification value of ≥ 0.1) in TAPS and WGBS at different contexts. TAPS has lower modification values at cytosine positions which are considered as modified, with a more pronounced bias towards the non-CpG positions.

reads TAPS [214] can be applied to study more reliably modifications at lowly mappable sections of the genome such as the repetitive regions, while cfDNA TAPS and reduced representation TAPS are under active development. The presented TAPS method, however, was based on analyses over a few samples and its low sample numbers cannot ensure statistical reproducibility and the completeness of discovered biases. Furthermore, the cell line used for testing of the TAPS method due to its popularity in research appears to be a poor choice for the method development of sequencing approaches to study cytosine methylation and its derivatives due to its dynamic changes in epigenetic patterns in embryonic stem cells. The E14 mESC modification levels were sensitive to changes in the environment as the cells could be switched between a pluripotent and a differentiated state that differs in the methylation levels. Given the possible perturbation of epigenetic patterns between cell passages, the TAPS against WGBS modification analysis could have been influenced negatively. Higher per

base Spearman correlation coefficients were observed when comparing the CpG methylation levels between different Barrett's oesophagus patients ($\rho \approx 0.8$).

Also, methods are needed to decrease or eliminate TAPS GC-related coverage bias and the bias against the CpH modifications as the sequence context bias against non-CpGs which restricts its applicability for the study of plant genomes. Furthermore, the lower conversion on 5-hmC decreases TAPS sensitivity in cells rich with this modification, such as neural and embryonic stem cells. Site-directed mutagenesis can be applied to the TET enzyme in order to increase its sensitivity for CpH and 5-hmC modifications as well as for higher densities of cytosine modifications [353]. In addition, optimisation of chemical conditions could improve TET kinetics towards 5-hmC such as longer incubation times. In terms of the reaction chemistry of TAPS, milder reaction conditions (pH, temperature, and time of incubation) should be sought in order to decrease the cytosine deamination rate and the introduction of less potential DNA damage. Next, the potential bias of TAPS towards lower GC content can skew coverage at certain genomic regions. Given that the coverage bias was largely due to the polymerase used, which is specific for uracil-rich sequences such as converted WGBS DNA, two different directions can be explored in a laboratory setting. The first one concerns a chemical reaction that would convert DHU molecules to uracil applying dihydropyrimidine dehydrogenase which should increase the processing ability of the polymerase (Kappa HiFi is a uracil-tolerant polymerase that is used in TAPS and bisulfite sequencing). The second approach uses directed mutagenesis that can be applied on the polymerase enzyme in order for it to accept with better conformational fit the DHU as a substrate in its catalytic domain. Regardless of the method optimisation direction taken, different types of spike-ins need to be included to study CpG and CpH modification levels biases as dependent on the density of the modified positions.

From a computational perspective, the conversion bias against a certain context can be mitigated through the application of generative probabilistic models. Such model-based approaches exist for different applications to study DNA modifications in bisulfite-treated data such as estimating the modification level of 5-hmC and 5-mC using WGBS and oxBS [21], and variant calling [123], discovery of differentially methylated regions [99], [326], [256], to calibrate PCR bias [250], bias-corrected methylation estimation [315], or to distinguish heterogeneous populations in SMRT sequencing [100]. A Bayesian model which takes as run-specific parameters the conversion rates as observed at specific spike-ins and contexts and the genomic coverage could provide an estimate of the true modification value at each position with a credible interval. In a similar fashion, any negative bias of the higher non-modified coverage could be computationally fixed as well by probabilistic and machine learning-based methods. A known limitation for applying generative models is the requirement for sufficient data (such as sequencing read depth) in order for the data to influence the posterior (otherwise, the results gained by running the models would depend solely on the model priors) [187]. Bias-correcting and modification sensitivity increasing algorithms can improve the performance of the current computational approach. The tool developed for the analysis of TAPS data asTair has basic alignment and modification calling functionalities for TAPS and WGBS data. Furthermore, at present it is only applicable to directional libraries; thus the development of tagmentation-based TAPS methods (PTAT, similar to PBAT which is a PCR-free method that has high sensitivity towards low inputs [241]) would require additional work. Also, the ability of the modification caller to distinguish between true modifications and C>T mutations in TAPS and the reference and C>T mutations in WGBS would decrease the false-positive rate of modification calls. Other avenues for the development of the current computational tool for

analysis of TAPS data would be for specific applications such as cfDNA TAPS and scTAPS. CfDNA poses a particular difficulty because of its overall lower sequencing coverage, and high level of heterogeneity due to the non-random representation and mixing of DNA fragments in the bloodstream [317]. The direct analysis of cfDNA modification levels could 'silence' the true modification levels in minority DNA fragment populations which would be averaged out. The distribution and abundance of DNA fragments in cfDNA depend not only on the size of the cellular population, but also on the way DNA leaves the cell, of the cell turnover, nucleosome binding, and even organ size and its shedding rate (cellular turnover). Thus, the separation of DNA fragments by their respective origin could give a richer picture of epigenetic and genetic changes. The separation of DNA fragments by the tissue of origin using the 5-mC patterns of DNA is possible through deconvolution using available tissue maps [243]. Still, it could be complicated by cellular heterogeneity and malignant transformation of epigenetic patterns. The problem of heterogeneous DNA molecule mixtures also exists in malignancies such as cancer and for natural occurrent populations such as plastids and mitochondria. Through elucidating the modification and mutation status at population levels more could be understood about the individual cell populations and their interplay in health and disease. A possible computational approach to such mixture problems would be to use a probabilistic model incorporating both modifications and mutations, as well as locus information similarly to mixture models in populational genetics [324], [323].

At present, there are more sensitive, and low-input bisulfite sequencing methods [329], [178], as well as there are novel bisulfite-free methods that have been developed in the recent years such as the APOBEC3A-based methods EM-seq [290] and ACE-seq [306], and the further development of the capabilities of Nanopore and PacBio sequencing [112], [343] that are able to directly read

DNA modifications (including novel DNA modifications [212]) without chemical conversion. In comparison to these methods TAPS strong points remain the conversion of modified cytosines only and the cost-effectiveness and sensitivity to low inputs that would suggest that TAPS should be further optimised for specific applications such as cfDNA TAPS and scTAPS. However, the most significant limitations to the applicability of TAPS are the requirement for synthesis of potent mTet1 enzyme which is not commercially available at present, and the toxicity of the main reductive compound pyridine borane [10]. Currently, efforts are made to switch to a less toxic borane compound. The primary purpose of both pyridine and 2-picoline borane is the carbonyl reduction of 5-fC and 5-caC. The application of sodium borohydride to 5-fC and 5-caC, which converts them to reduction products has been described previously and was used to study oxidative metabolites of 5-mC and 5-hmC [157]. This compound has been tested for applicability in TAPS, but it showed only 60% conversion which can be explained by its inability to react to carboxylic acids (5-caC) but is reactive towards aldehydes (5-fC) [111]. Of the tested boranes during the TAPS method development, a possible candidate for a less toxic reducing agent is ammonia borane which has a lower toxicity than 2-picoline and pyridine borane [2]. A combined approach of a TET homologue from the fungi *Coprinopsis cinerea* that preferentially converts 5-mC and 5-hmC to 5-fC under slightly acidic conditions (pH 5.8) [385] that may simplify the search for carboxylic acid-reducing agents, and presumably would improve the reductive ability of commonly used compounds such as sodium borohydride. Alternatively, new TET-assisted methods relying only on enzymatic or immunochemical labelling of TET-oxidised DNA modifications to 5-caC could selectively label true modifications without the need for a reducing borane agent, and such antibody labelled sequences can be directly sequenced by Nanopore or with Ion Torrent sequencing [81]. Such

a method will remove the issue of distinguishing C>T mutations which is an issue in both TAPS and WGBS, would enable the detection of the four types of cytosine modifications, and by being PCR-free could eliminate some of the observed GC biases. Furthermore, novel nanozyme-based [41] or electrochemical [387] oxidation approaches to study cytosine modifications could replace TET enzymes applied to NGS methods decreasing costs and substrate bias.

Chapter 4

Model development for variant calling on WGBS and TAPS data

TAPS preserves sequence complexity to a greater extent compared to WGBS data owing to the chemical alteration of DNA only at modified cytosine positions. The lower divergence of TAPS from WGS is expected to allow for reliable combined variant and modification calling from samples. The ability to extract both modification and variant calls from a single sequencing data source is validated by the existence of several tools that allow for variant and modification calling on WGBS data. Furthermore, approaches, which can recognise modifications and variants in TAPS and WGBS could boost biomedical research as more information about the interplay of modifications and genetic variants during development, healthy ageing and diseases such as cancer would be collected and

Part of the results of this chapter was presented as a poster presentation at the *20th Genome Informatics Conference, 2020, UK (virtual conference).*

interpreted. Also, using a single data source could help minimise sample noise due to technical and biological variation introduced primarily through random sampling and sequencing library preparation. However, WGBS and TAPS both rely on the conversion of cytosines to thymines when they are unmodified, or they are modified respectively. The C>T substitution is the most common mutation observed, which could lead to mislabeling of modifications and variants. Thus, it is essential to identify modified cytosines in TAPS and WGBS data correctly. The modifications and genetic variants in TAPS and WGBS data have different expected distributions and other sequencing read-related properties that could be used to separate them.

The following chapter shows the development of methods to distinguish modifications from genetic variants in both TAPS and WGBS data by incorporating selected features from the sequencing data.

4.1 Description of the used WGS, TAPS and WGBS datasets

This section summarises descriptive characteristics of the datasets used in this and the following chapters comprising 11 TAPS, four WGBS samples, and their matched WGS data (labelled as tWGS and bWGS respectively, with 16 tWGS and nine bWGS samples). The samples in these datasets were either clinical cases (Barrett's oesophagus and oesophageal adenocarcinoma) or patient-matched controls (healthy solid tissues and blood). Oesophageal adenocarcinoma (OAC) is the eighth most common cancer worldwide with increasing numbers of people affected [171]. In both Barrett's oesophagus and OAC, abnormal methylation patterns have been observed [171]. Barrett's oesophagus is a premalignant condition where the oesophageal epithelium changes its morphol-

ogy and function as a result of external factors, such as stomach acid reflux. Patients with Barrett's oesophagus have 11 times higher chance for the development of the oesophageal adenocarcinoma compared to the general population (with a risk of progression of 0.11% per year) [148]. Endoscopy is the standard of care for diagnostics and monitoring of the disease. Still, efforts are taken for the discovery of biomarkers and drug targets that would enable non-invasive diagnostics and stratification of patients who will progress to malignancy. Both genetic and epigenetic variability between patients could be used in the clinic if it achieves separability of the groups of interest. Discovering genetic and epigenetic patterns could happen only if the methods used can reliably distinguish signal (true variants) from noise (sequencing and alignment errors, modifications).

The validity of variant and modification calling depends on both whether the distribution of the input data matches the assumptions made by the variant calling model, and on signal strength (and availability) to distinguish between the classes. Thus, the presence of sequencing artefacts and other kinds of errors should be taken into account during data processing. In order to decrease the effect of possible confounders, the data need to pass through multiple quality filters that remove known errors prior to variant calling. A common type of confounder is duplicated DNA sequences, which can inflate and propagate sequencing errors that could mask true variants or increase the false-positive rate. The analysis step of duplicate reads removal is also known as *deduplication*. PCR duplicates are commonly introduced as a result of a low input or complexity library preparation and can introduce coverage bias at low GC regions [257]. After the removal of PCR duplicates (generated in the PCR step, which are distinguished by their exactly matching start and end coordinates after alignment) and optical duplicates (introduced during sequencing when an optical sensor mislabels amplification clusters) from the sequencing reads, the

samples alignment rate is calculated as the ratio of deduplicated to trimmed reads (reads with removed adapters and passing quality filters) as shown in equation 4.1. The opposite measure the *PCR duplication rate* can be expressed as a 1 - alignment rate after demodification.

$$\text{alignment rate} = \frac{\text{deduplicated reads}}{\text{all reads}} \tag{4.1}$$

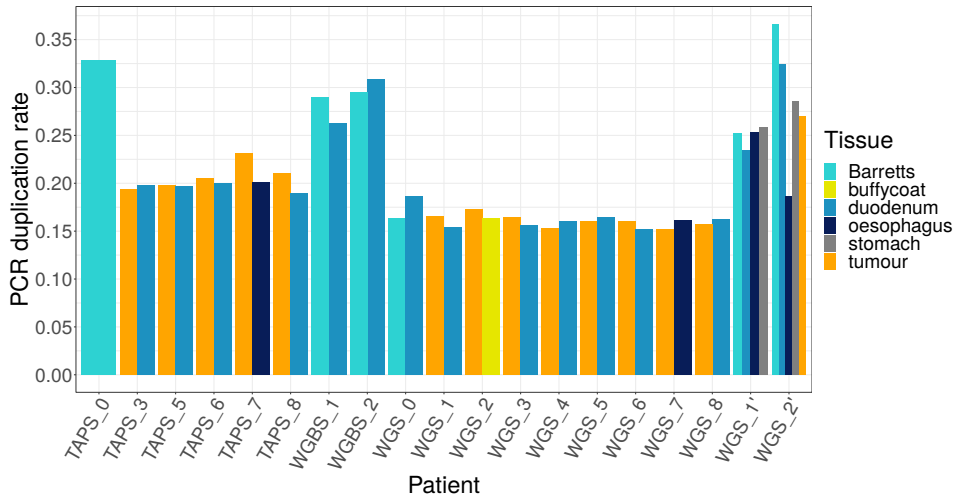
$$\text{mean sequencing coverage} = \frac{\sum \text{coverage per position}}{\text{covered positions}} \tag{4.2}$$

All samples had ≥ 0.6 alignment rate after deduplication with the lowest values for bWGS patient 2 Barrett's oesophagus and duodenum, and patient 0 TAPS Barrett's oesophagus sample (see Figure 4.1). The higher PCR duplication rates suggested low library complexity resulting from amplification of limited starting genetic material [86], [304]. The higher PCR duplication rate can influence downstream data analysis primarily through lowering the sequencing coverage of certain positions. Insufficient sequencing depth may lead to the false representation of true allele frequencies and thus increase variant and modification calling false-positive and false-negative rate depending on whether the reads with the alternative or the reference allele are affected by duplication [280], [54]. Given the link between correct variant calling and representative allele frequencies, another critical metric for correct variant and modification calling is the mean coverage defined as the ratio of the total coverage and the positions covered (see equation 4.2). The TAPS and tWGS samples had sequencing depths of $\approx$ 30x for the control samples and $\approx$ 45-60x for the OAC samples and 20-25x for Barrett's oesophagus tissue (patient 0). Lower than 30x coverage of control and case tissues from patient 0 could cause issues with somatic variant calling (identification of tumour specific variants) where a genomic coverage of $\geq$ 30x

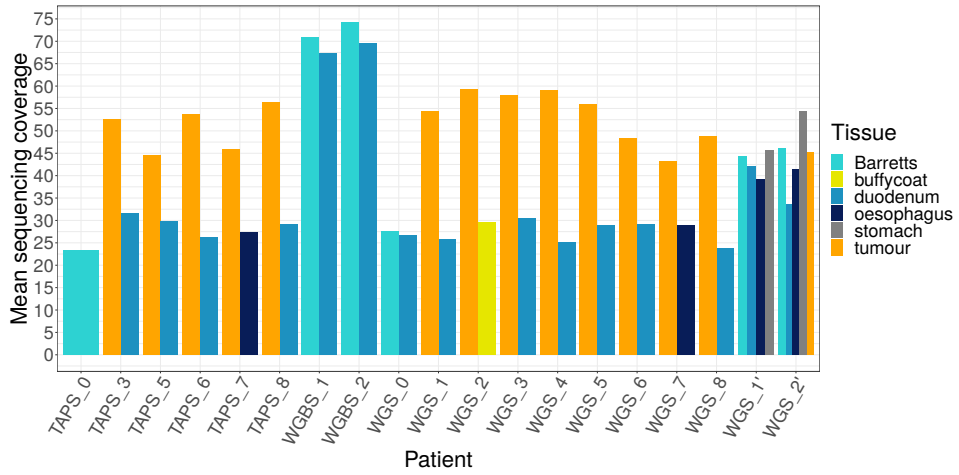
of the matched healthy tissue sample is required for variant discovery [150], as samples with low complexity and low coverage are expected to have less reliable variant calls [229]. The bWGS tissue samples fell within the 30-55x range, and the WGBS samples within 65-75x, both groups of samples without a clear difference in depth between the controls and cases (see Figure 4.1).

Another difference between the samples was the fragment size distribution, where a different library size selection approaches were used on TAPS, WGBS, and WGS. At least 90% of the read fragments found in TAPS and WGBS had template lengths $\leq$ 500 bp, while fragment lengths between 500 bp and 1000 bp were more common in the WGS samples (see Figure 4.2). Results from Tan et al. indicate that fragments $\geq$ 500 bp in pair-end Illumina sequencing can lower the base quality scores of the second read, and could potentially affect variant calling at low sequencing depths [334]. In order to account for issues resulting from the base qualities assigned by the sequencing machines, GATK base quality score recalibration (BQSR) [3] was applied to the WGS data as part of the pre-processing pipeline before variant calling.

Tumours and pre-malignant tissues such as Barrett's oesophagus consist of heterogeneous cell populations. In the case of solid tumours, the cancer cells are surrounded by a non-malignant environment, also known as stroma created by immune cells, circulatory vessels and connective tissues. Cellularity (tumour purity) is the proportion of cancer cells within the tumour environment. The cellularity of cancer affects somatic and copy number variation calling, as well as modification calling (the discovery of differentially modified regions, DMRs, between the healthy and the cancer sample). Low tumour purity would prove insufficient for sensitive mutation discovery as the differences between the cancer cells, and the stromal cells would be averaged out given the larger non-malignant population. Traditionally, tumour cellularity can be evaluated through staining



a) Alignment rate after removal of PCR and optical duplicates



b) Mean sequencing coverage

Figure 4.1: Descriptions of the datasets used in terms of a) alignment and b) mean sequencing coverage. The numbers indicate patient numbers for each of the samples, where 1' and 2' label WGS samples that match WGBS instead of TAPS. The highest duplication rates were observed in WGBS and their matched WGS samples, and the TAPS sample of patient 0. The highest coverages were seen in the WGBS samples and the tumour samples for the OAC trial.

and a pathological examination under a microscope. However, in the case of NGS paired healthy (normal) and cancer samples, the tumour purity can be

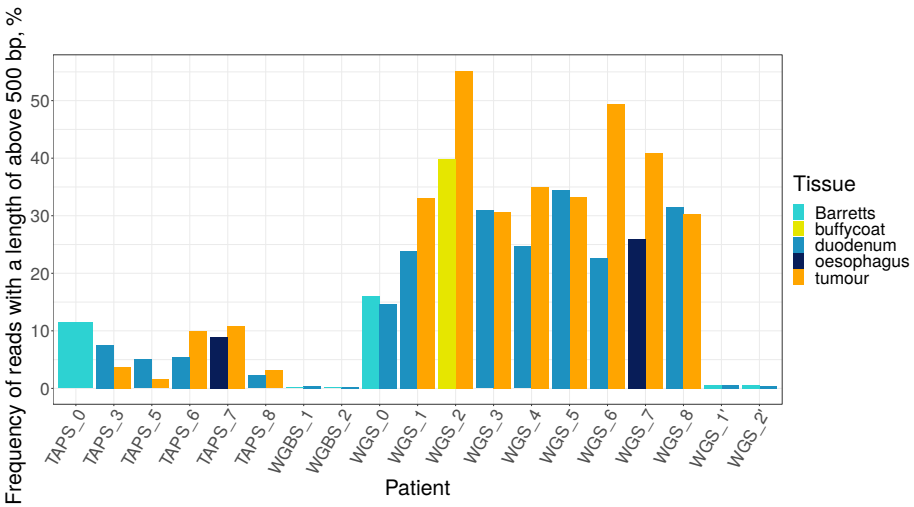


Figure 4.2: Descriptions of the datasets used in terms of their fragment distribution above and below 500 bp. The numbers indicate patient numbers for each of the samples, where 1' and 2' label WGS samples that match WGBS instead of TAPS. The TAPS matched WGS samples showed a different fragmentation profile as a result of alternative size selection.

estimated during a copy number variant calling step. Another neoplastic characteristic that can be obtained from this analysis is ploidy that is the level of change in the number of each set of chromosomes (the difference from diploidy). From the CNV analysis performed with Sequenza, four (patients 0, 6, 7, and 8) of the tWGS samples have lower ploidy (1.9), together with relatively low but comparable to the remaining samples tumour purity apart from sample 8 (see Figures 4.3 and 4.4). The WGS and TAPS samples for patient 0 were sequenced to 2-3X lesser sequencing depth than the remaining WGS samples studied. In patient 8, the high cellularity and low ploidy together with the relatively flat CNV profile suggested that there was possible contamination of the control tissue sample, or that there are very few cancer cells in the sample. The remaining two (6 and 7) also have ploidy close to diploid and potentially have lower cellularity than what was detected by the algorithm. In the case of the WGBS matched WGS samples, three different healthy control tissues were available

together with a secondary malignant tissue from patient 2 (see Figures 4.3 and 4.5). The CNV cellularity and ploidy estimates in WGBS were very close regardless of the healthy tissue used as a control, and the three Barrett's oesophagus and tumour samples showed relatively low cellularity and ploidy close to the diploid genome.

Additionally, the CNV information could be used to identify Barrett's oesophagus patients who would progress to cancer, as well as it shows association with advanced malignancy stages as observed histologically [261].

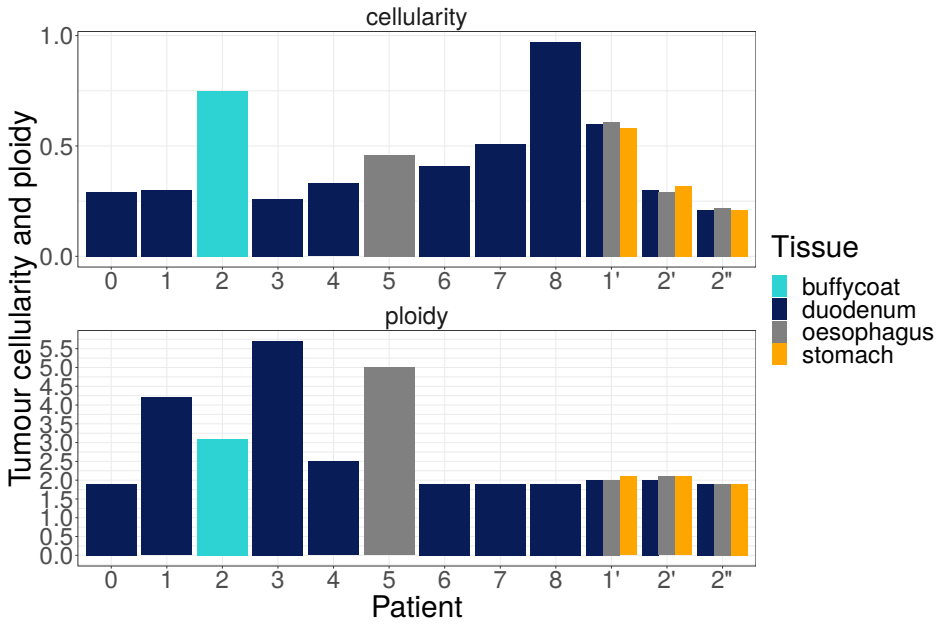


Figure 4.3: Cellularity and ploidy estimates in WGS samples obtained with Sequenza. Here, numbers indicate patients, where 1' and 2' indicate WGS samples matching the WGBS Barrett's oesophagus, while 2" labels WGS matching WGBS cancer sample from patient 2. Patients 0, 6, 7, and 8 had ploidy close to the normal diploid genome, while the cellularity estimates varied. A reason for the observed cellularity estimates was the low purity of the sample that throws off the Sequenza CNV calling algorithm.

Given the differences in the chemistry of TAPS and WGBS compared to

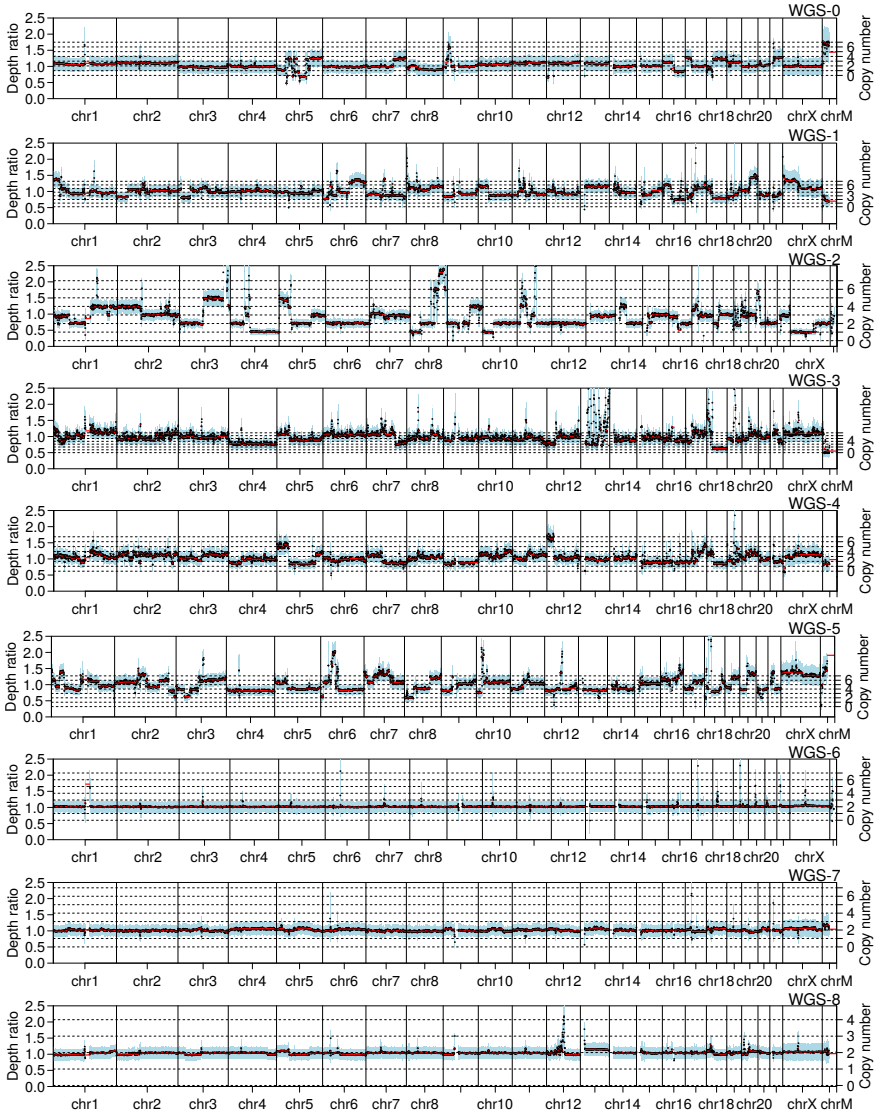


Figure 4.4: Genomic copy number profiles of WGS samples matching TAPS as a result of Sequenza CNV variant calling for each patient. Patients 6, 7, and 8 showed CNV profiles closer to a copy number of two in comparison to other OAC patients that might be indicative of lower tumour cellularity.

WGS, it is expected that insertions and deletions[1] could be introduced chemically or through the alignment and pre-processing steps. In order to study the

[1]Insertions are a type of mutations where new nucleotides are added to the DNA sequence, while deletions occur with the removal of reference bases. These two types of genetic errors are known as indels.

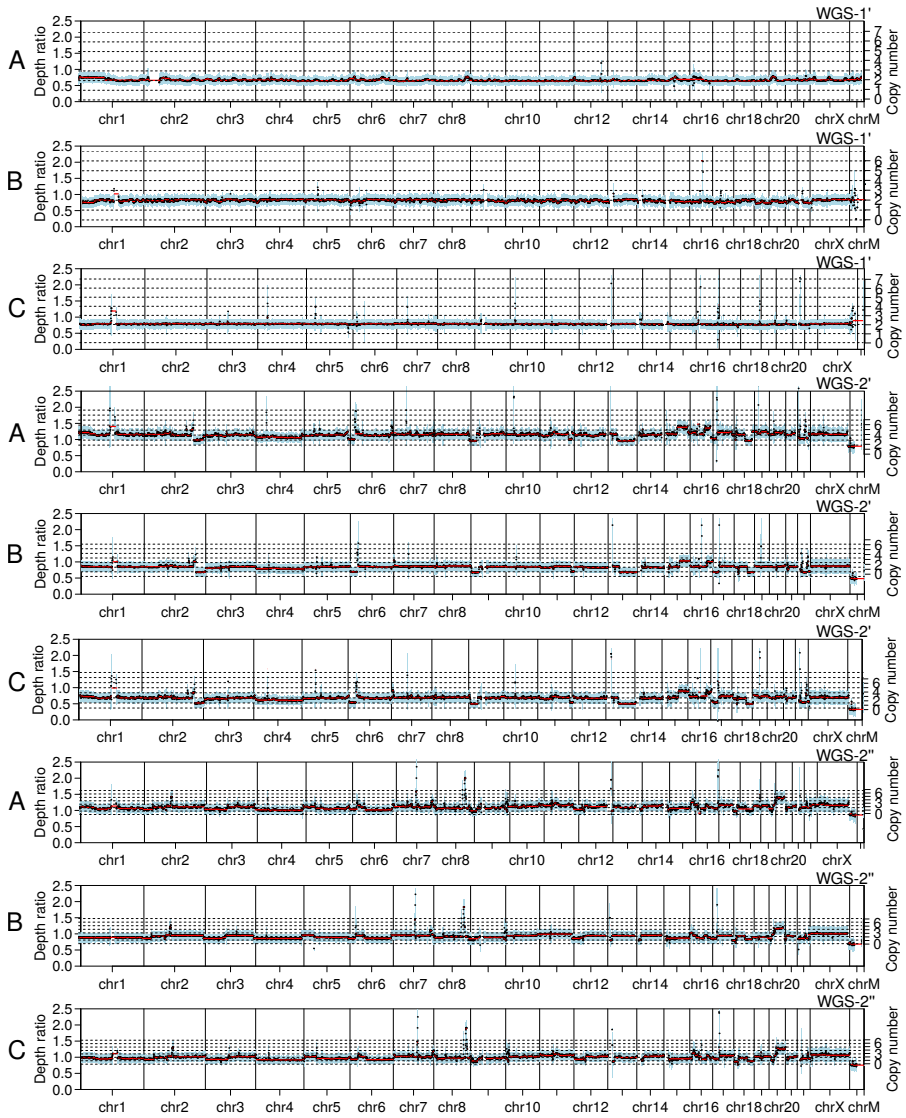


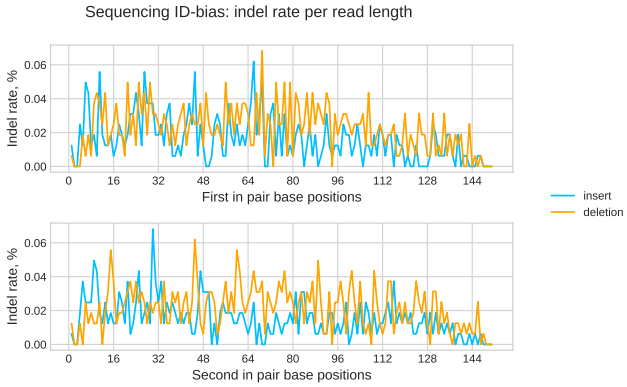
Figure 4.5: Genomic copy number profiles of WGS samples matching WGBS as a result of Sequenza CNV variant calling for each patient. Here, numbers indicate patients, where 1' and 2' indicate WGS samples matching the WGBS Barrett's oesophagus, while 2" labels WGS matching WGBS cancer sample from patient 2. A, B, C mark that the healthy control used for CNV calling was duodenum, oesophagus, or stomach, respectively. These results showed that general tumour CNV patterns were captured using any of the available matched controls.

indel frequency and distribution between different types of data, information about the indels was taken from the CIGAR string (a short string for each

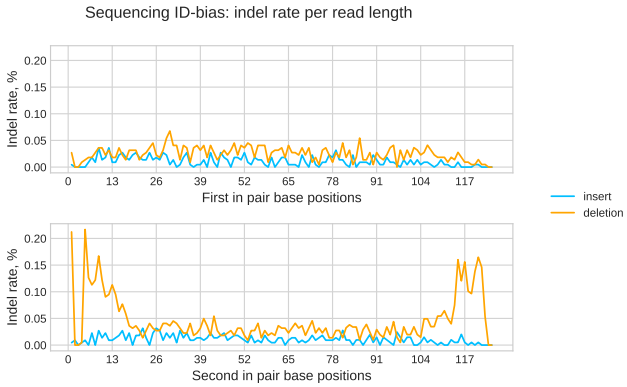
aligned read giving mapping information about soft and hard clipping, indels, matching to reference, and other characteristics) of the aligned reads (start and end position, length, base and mapping qualities). These data were summarised as counts and descriptive statistics. Separately, another script visualised the dependency of the indel frequency and their position in the read called *indel-bias*. The indel-bias plots show much higher indel rates, particularly in the second read in WGBS (see Figure 4.6). These high rates of deletions in the ends of the second read in WGBS closely match the lower base qualities at these positions as observed in FastQC reports. TAPS and WGS show lower levels of insertions compared to deletions with the highest incidence of indels occurring in the middle of the read.

The frequency of reads containing indels to all reads showed the lowest relative abundance for insertions and deletions in the bWGS, and the highest rate of deletions in the WGBS samples (see Figure 4.7). In the case of TAPS, the indel rates are comparable as frequency, and the observed differences with the bWGS could be a result of differences in the sequencing technology (NovaSeq and HiSeq). However, the highly elevated number of deletions in WGBS compared to its matched WGS suggested that these indels are a result of the specific chemistry and data processing of bisulfite-treated samples such as depyrimidination events and general DNA degradation [335]. Furthermore, no apparent difference in the number of indels spanning the individual read was seen between case and control samples, but there was an association between the frequency of deletions and coverage of the sample ($\rho = 0.588$, Pearson's product-moment correlation, p-value = 0.00026).

Finally, some characteristics of the reads carrying indels were distinctive between the groups in comparison (see Figure 4.8). The WGBS samples had shorter insertions on average, as well as longer deletions in comparison with



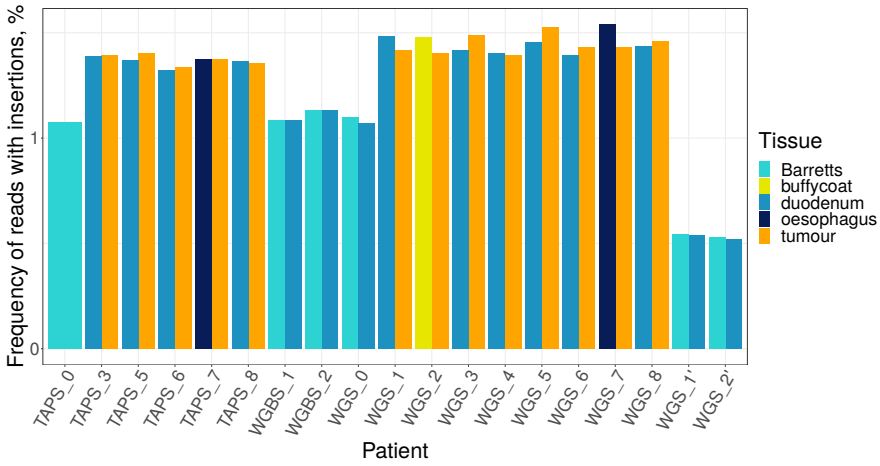
a) Indel bias per read length in TAPS.



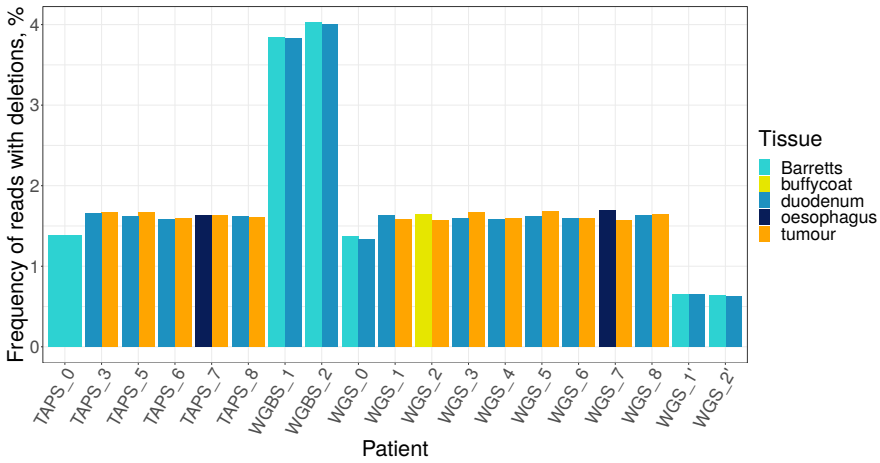
b) Indel bias per read length in WGBS.

Figure 4.6: The dependency of the insertions and deletions to the position in the sequencing read in a) TAPS and b) WGBS. The rate of deletions observed is higher than the rate of insertions. The second in pair read of the tested WGBS sample showed a higher rate of deletions towards the beginning and the end of the read.

TAPS and WGS (four sample groups indel lengths p-value $< 2.2e^{-16}$ Kruskal-Wallis test; p-value $< 2.2e^{-16}$, CI 95%, two-sided two-sample t-test comparing WGBS and bWGS reads with indels). In order to compare reads with indels and reads without indels in terms of their features, 8-10 million reads without indels were randomly selected. The reads with indels have a higher mean base quality which may be a result of alignment penalisation in samples with more differences



a) Insertions frequency



b) Deletions frequency

Figure 4.7: Frequency of sequencing reads a) containing insertions or b) deletions to all reads in the sample showed as a percentage. Here, numbers indicate patients, where 1' and 2' indicate WGS samples matching the WGBS Barrett's oesophagus, while 2" labels WGS matching WGBS cancer sample from patient 2. The highest rate of observed deletions was detected in the WGBS data, while the lowest indel rates were seen in the WGBS-matched WGS data. This difference between the TAPS- and WGBS-matched WGS samples might be explained by the different sequencing machine used.

from the reference—reads with indels are aligned if their overall quality was above a certain threshold where the remaining non-indel bases in the read should

have higher base qualities to compensate thus biasing the group means. In TAPS and WGBS, the reads without indels have a lower mean BQ compared to their matched WGS samples suggesting alignment differences. Furthermore, in all cases (TAPS and TAPS_n, tWGS and tWGS_n, bWGS and bWGS_n, WGBS and WGBS_n) reads with indels had higher mean BQ values compared to reads without indels (p-value $<2.2\text{x}10^{-16}$, CI 95%, two-sided two-sample t-test on reads with versus reads with indels in terms of BQ).

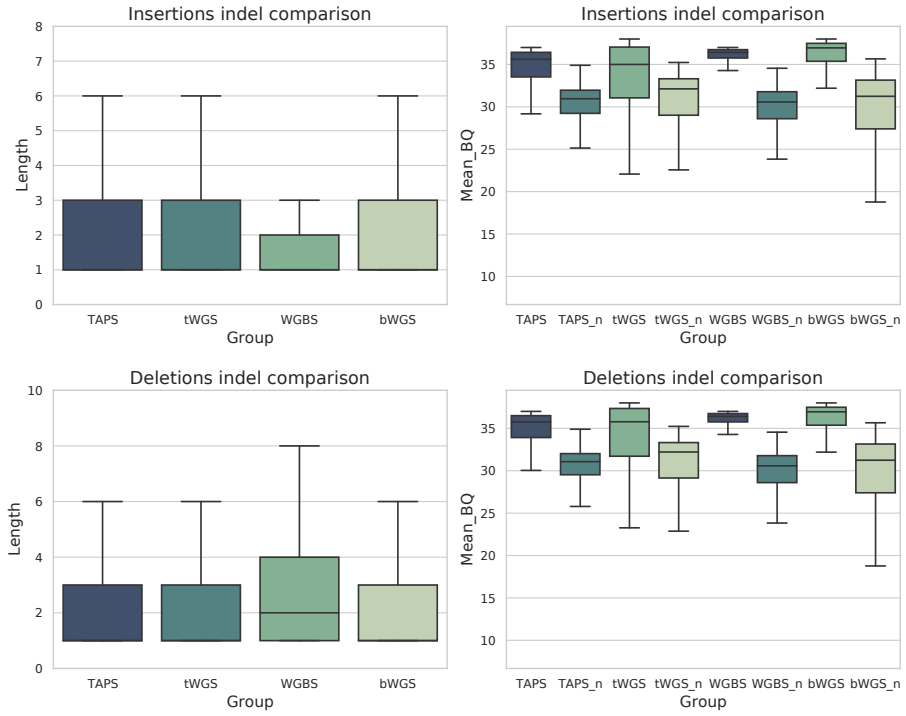


Figure 4.8: Indel distributions per characteristics. Legend: TAPS, WGBS, tWGS (TAPS-matched WGS samples), bWGS (WGBS-matched WGS samples). Reads that do not contain indels are further marked by *_n*. The lengths of the WGBS indels showed a significant difference from the other groups. Also, the reads with indels had, on average, slightly higher mean base quality than reads without indels. This difference could be explained by penalisation of reads during alignment where reads with certain scores would be aligned to the reference.

4.2 Demodification

The most straightforward approach to variant calling in TAPS was to revert observable cytosine conversion before using WGS-specific variant calling tools. The process of changing back converted cytosines (read as thymines in both WGBS and TAPS data) to cytosines was called *demodification*. The WGS variant calling following the demodificaion step assumes that TAPS or WGBS are alike WGS in terms of library and alignment bias and differ only in their modification status. This is a strong assumption, but it allows for the discovery of specific biases in and issues with TAPS and WGBS data and the further development of custom tools specifically build to resolve these issues. To amend the problem of modifications falsely being called as high-quality variants, a tool that was able to revert tentative chemical conversions in the sequencing data was developed. At first, the tool was using a simple deterministic approach where demodification, D, was a function of the nucleotide in the sample, S $\in$ {A,C,G,T} versus the nucleotide at the reference, R $\in$ {C,G} (equation 4.3).

$$D \coloneqq f(S, R) \tag{4.3}$$

The demodification occurs at every T at a position that shows C in the reference (on the OT strand, for the OB we are considering A at a G position) (equation 4.4).

$$D = \begin{cases} C & \text{if T at a C position} \\ G & \text{if A at a G position} \\ base & \text{otherwise} \end{cases} \tag{4.4}$$

The expected function of the demodifier was validated on TAPS and WGBS

sequencing reads coming from the first 300 bp of bacteriophage λ that was used as a spike-in control during sequencing of the samples prepared during TAPS method development. The results suggested complete demodification of all cytosine contexts in both TAPS and WGBS (see Figure 4.9).

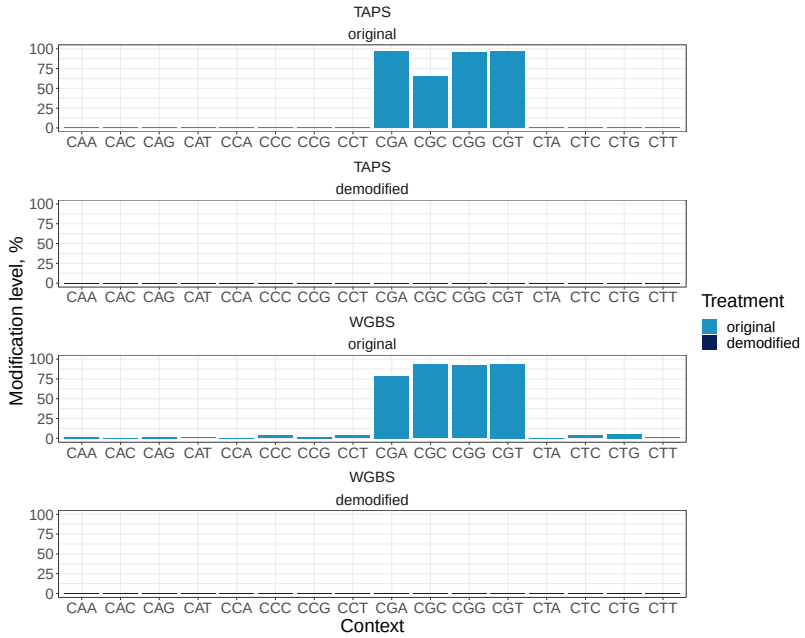


Figure 4.9: Demodification efficiency on WGBS and TAPS data from real sequencing data from bacteriophage λ. The deterministic demodification successfully reverted all thymines to cytosines at modified positions regardless of the cytosine context.

In order to distinguish between genetic variants and modification (which would decrease the false-negative rate of detected variants due to real variants being 'reverted' back in T>C, A>G manner), another variable class K was included, which is functionally dependent on features a_i from the sequencing data (equation 4.5).

$$D \coloneqq f(S, R, K) \tag{4.5}$$

In this expanded model, the demodification occurs at positions indicating chemical conversion only when the classifier shows that the highest probability of the position of being a modification (equations 4.6, 4.7). Here the classes modification, heterozygous and homozygous SNP, and the reference are labelled as a mod, het, hom, and ref respectively.

$$K(S, R, a_i) \coloneqq \arg\max_k P(k|S, R, a_i), k \in \{mod, ref, het, hom\} \tag{4.6}$$

$$D = \begin{cases} C & \text{if T at a C position iff K(S, R, } a_i) = \text{mod} \\ G & \text{if A at a G position iff K(S, R, } a_i) = \text{mod} \\ base & \text{otherwise} \end{cases} \tag{4.7}$$

In order to build a model that would be able to distinguish SNPs from modifications, different sets of features, model types, and architectures were explored, and were finally benchmarked against other datasets. Different discriminative machine learning (ML) models were applied to the data for the classification of variants and chemical conversions. The full set of features that was assembled was based on previous knowledge about TAPS and WGBS sequencing data and on assumptions about the distribution of features within the classes of interest.

4.2.1 Feature exploration

A fundamental step for developing an algorithm to distinguish modifications from genetic variants is finding and curating a list of features that can reliably separate genetic variants and chemical conversions. The feature engineering for the problem included exploration and step-wise addition of features, where the feature sets were generated through the following steps:

1. initial exploration based on domain knowledge (features extracted from chromosome 1 from Barrett's patient sample 0) with labels based on WGS GATK HaplotypeCaller;
2. expansion of the feature set (features extracted from chromosome 1 from Barrett's patient 3 (sample 3), or 10 Mb random regions from autosomes, sex chromosomes, and the mitochondria) where again the labels were based on WGS GATK HaplotypeCaller;
3. expansion of the feature set (random positions from all autosomes, sex chromosomes, and the mitochondrion from patient 3 duodenum and Barrett's oesophagus, where the labels were the overlap of the haplocalls made by WGS GATK HaplotypeCaller, Octopus, and Strelka2;
4. feature selection using Random Forest feature ranking and scikit-learn automated feature selection.

The expansion of the training set with the sex chromosomes and the mitochondrion was based on their properties differing from the autosomes (mitochondrion – haploid, very high coverage, exists as a heterogeneous population within cells, a property known as *heteroplasmy*; Y and X – hemizygosity, different methylation patterns), thus introducing richer information into the model. Using domain knowledge about NGS sequencing and alignment, TAPS, and WGBS conversion, an initial feature set was created and modified by empirical evidence for the separation ability of the features during model training. The features used in the classifiers fall into several main groups based on whether they describe the reference position or the individual read (the superscript number shows at which feature set development step the feature was added with $*$ labelled features included from the beginning):

1. strand orientation (OT or OB)*;
2. trinucleotide contexts*;

3. base quality (statistics including mean, mode, median, standard deviation)*;
4. BWA mapping quality (statistics including mean, mode, median, standard deviation)*;
5. modification ratio or β value (on the strand of interest and the opposite strand), including[2–3] or excluding* reads which are not marked as proper pairs;
6. read counts and read ratios grouped by flags*;
7. BWA primary and secondary alignment scores (AS and XS)*;
8. Edit distance of all reads covering the position of interest (statistics including mean and median)*;
9. normalised sequence errors (Edit distance to read length)[1];
10. count of all XA tag (an alternative match of the read with genomic coordinates) labeled reads at a position*;
11. count of possibly modified CpGs on the same read*;
12. GC content of the reference region spanning the aligned read*;
13. proportion of reads covering a position of interest which are insertions*, deletions*, reference skipping[2–3], secondary[2–3] or supplementary[2–3] (all of these read characteristics are Boolean).

These features can be grouped further into 1) primary separating features (containing strong signal to separate the classes) and 2) quality filtering features that ensure that only high-quality reads are used to minimise sequencing and alignment errors. In the first category fall the strand and trinucleotide context information, the read counts and ratios, and the count of chemically modified cytosines at CpG positions. The strand orientation and read counts and ratios give information on whether the observed cytosine can be modified (only when C>T or G>A changes are observed in the OT and OB strand respectively), or

whether it matches the reference. Both genetic variants and modifications do not match the expected reference; however, their distribution differs depending on the underlying DNA context. There are 4^3 possible trinucleotide contexts, of which the four CpG contexts are expected to be methylated in the genomes of mammals (up to 60-80%), the remaining 12 cytosine contexts have a lower probability of methylation, but can be modified in tissue and state-dependent manner (like in neural tissue and embryonic cells). The remaining 48 non-cytosine contexts are not expected to be modified under sodium bisulfite or pyridine borane treatment. Thus, they can either differ from the reference as a result of genetic variation or can be a product of errors. The modification values of neighbouring CpGs are correlated, which is the reason why the count of C>T changes in CpGs in the same read was included as a feature. The filtering features such as the Edit distance, base and mapping quality, indel information can further improve the separability between chemical conversions and modifications as observed empirically and were included as features. Also, at a later stage of model development, the publicly available SNP information was used to improve the classification sensitivity.

Using read information from one chromosome from a TAPS patient 0 Barrett's oesophagus sample, the differences between the feature correlations for heterozygous and homozygous SNPs, and non-variants (modifications and reference positions) were studied in terms of read ratios, abundance, and feature distributions. This sample was selected due to its availability being the first solid human tissue sample. Its matched WGS sample germline calls were used as a truth set for labelling heterozygous and homozygous SNPs. In the TAPS sample used, around 80% of all CpGs were modified compared to $\leq 1\%$ of the positions spanning non-CpG contexts (see Figure 4.10). The rate of true heterozygous SNPs was around twice the rate of homozygous SNPs, and the probability of

SNP was 5-6 times higher in CpG contexts. Given that higher modification levels were expected at CpG contexts, the positions were separated into two groups (CpG and non-CpG positions) and observed correlation patterns between the groups (see Figures 10, 11, 12, 13, 14, 15). In all of the SNP feature correlation plots, a high degree of association was observed between the modification ratio on the opposite strand, and the number of G, T, and A bases to the reference ratio. In the correlation of features describing modifications at CpG contexts, the modification level was correlated with the ratio of T to reference and C to reference. In all of the correlation plots, the BWA primary and secondary alignment scores (AS, XS) were negatively correlated with Edit distance and the standard deviation of the base and mapping quality of reads spanning a position, but they were positively associated with the mean and median base and mapping qualities as expected. It is important to see whether any features are correlated, which would mean that keeping them together does not increase usable information. The count of chemical conversions at CpGs in the individual reads in all described cases showed a slight negative correlation with alignment metrics. Overall, the correlation plots for modified positions showed a higher correlation between features compared to SNPs, particularly in the case of homozygous SNPs. Given the higher correlation between the different base and mapping qualities, only the mean, median, and standard deviation were used further. Nevertheless, such simple feature correlation plots may not provide direct information on how important each feature is for classification, which was later done during the feature selection step.

If we assume the simplest case, where we have a diploid organism with the expected genotypes AA for the homozygous reference genotype, with AB the heterozygous mutant and with BB the homozygous mutant, we can further develop our strand-specific genotypes that can be observed in TAPS data and class

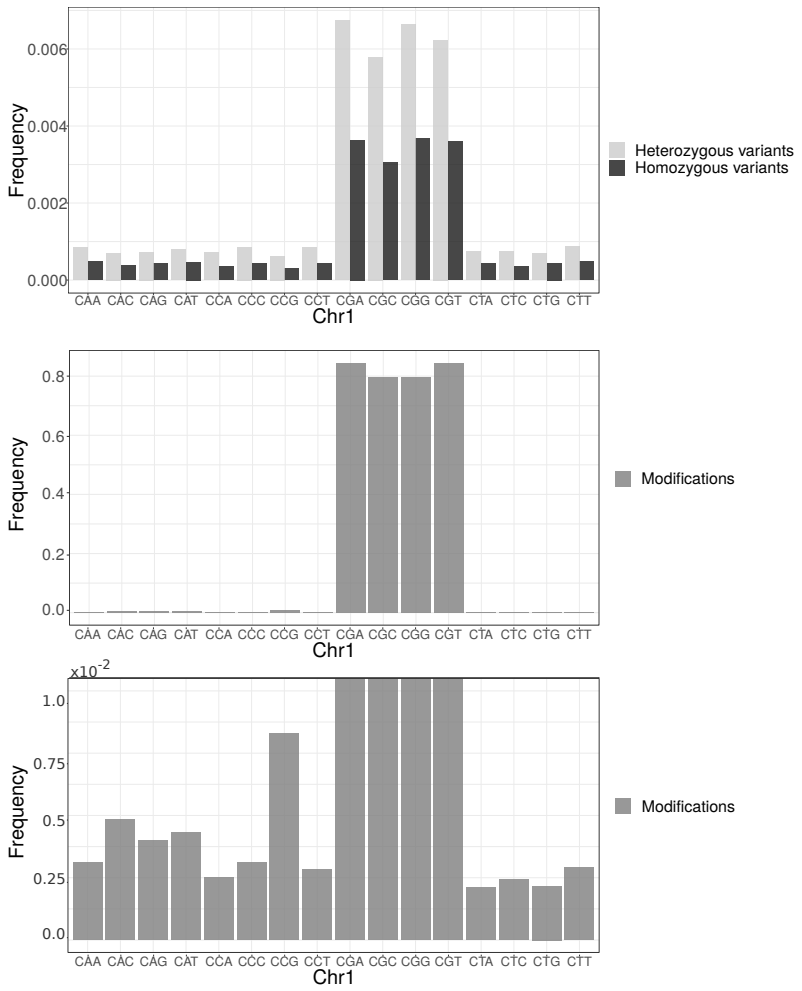


Figure 4.10: Distribution of SNPs and modifications per cytosine trinucleotide context in TAPS using positions from chromosome 1 of Barrett's oesophagus sample. From top to bottom: germline variant frequency, modification levels in cytosine positions with $\beta \geq 0.1$, and again modification values showing the non-CpG contexts. The heterozygous SNPs appear to occur about two times more frequently than homozygous SNPs, with a higher rate on CpG contexts (top). In comparison to them, the modification values vary considerably with the highest rates again seen at CpGs (middle and bottom panel).

expectations in a table (see Tables 4.1 and 4.2). The influence of larger copy number variations on expected read counts is excluded at present. Also, as the reference in the simplest case, the known and publicly available genetic sequence

of an organism was used (supplied as genomic build in FASTA format), and we did not distinguish between germline and somatic polymorphisms. Apart from the expected diploid variants, reference, and modification genotypes, there are also genotypes where modification happens on one DNA strand while there is a variant on the other side. In carcinogenesis theory, it is expected that two different mutations have a negligible probability for occurring in the same position. Usually, variant callers discard variants when there are multiple at a position (by probabilistic ranking or breaking ties) [284]. However, mutational and modification processes are separate. Although they cannot be treated as fully independent, they allow for the co-occurrence of modification and mutation in the same place in the genome. There were $\approx$ 1000 positions per TAPS sample (using a conservative overlap approach), wherein TAPS C>T or G>A change was observed, but at the same position in the WGS sample the substitution was C>A, C>G, or G>C, G>T. These positions may show co-occurrence of a modification and a variant, but before that, any alignment issues and low coverage should be excluded as a possible cause. These events also primarily happen at CpG positions, and further lab-based experiments are needed to study their frequency.

Feature selection A feature extraction script was created that takes sequencing reads in BAM or CRAM format and outputs a feature table that can then be labelled and used for training and testing of the model. Feature selection was performed in order to decrease the noise of features and remove dependent features, thus resulting in higher classification performance and lower computational complexity [61]. The most important features (which maximise the separation of classes) were used based on feature ranking during the training of the random forest to simplify the feature space. BWA alignment scores (AS, XS), mapping, and base quality statistics seem to be of higher relative importance

Class	Observed change	A	C	G	T	a	c	g	t	Mistaken with
unmodified cytosine	None	*	C	*	*	*	*	g	*	None
modified cytosine	$C \to T$	*	*	*	T	*	*	g	*	heterozygous $C \to T$
	$g \to a$	*	C	*	*	a	*	*	*	heterozygous $g \to a$
cytosine heterozygous variant	$C \to T$	*	*	*	T	*	*	g	*	modification $C \to T$
	$C \to A$	A	*	*	*	*	*	g	*	None
	$C \to G$	*	*	G	*	*	*	g	*	None
	$g \to t$	*	C	*	*	*	*	*	t	None
	$g \to c$	*	C	*	*	*	c	*	*	None
	$g \to a$	*	C	*	*	a	*	*	*	modification $g \to a$
cytosine homozygous variant	$C \to T$	*	*	*	T	a	*	*	*	mixed modification and heterozygous variant $C \to T$, $g \to a$
	$C \to A$	A	*	*	*	*	*	*	t	None
	$C \to G$	*	*	G	*	*	c	*	*	None
mixed modification and heterozygous variant	$C \to T$, $g \to t$	*	*	*	T	*	*	*	t	None
	$C \to T$, $g \to a$	*	*	*	T	a	*	*	*	homozygous variant $C \to T$
	$C \to T$, $g \to c$	*	*	*	T	*	c	*	*	None
	$g \to a$, $C \to T$	*	*	*	T	a	*	*	*	None
	$g \to a$, $C \to A$	A	*	*	*	a	*	*	*	None
	$g \to a$, $C \to G$	*	*	G	*	a	*	*	*	None
mutation followed by modification	$A \to C$, $C \to T$	*	*	*	T	*	*	g	*	modification $C \to T$
	$G \to C$, $C \to T$	*	*	*	T	*	*	g	*	modification $C \to T$
	$T \to C$, $C \to T$	*	*	*	T	*	*	g	*	modification $C \to T$
	$a \to g$, $g \to a$	*	C	*	*	a	*	*	*	modification $g \to a$
	$c \to g$, $g \to a$	*	C	*	*	a	*	*	*	modification $g \to a$
	$t \to g$, $g \to a$	*	C	*	*	a	*	*	*	modification $g \to a$

Table 4.1: Each of the four major cytosine classes (unmodified, modified, heterozygous variant, and homozygous variant) contains one to multiple possible genotypes based on expected allele counts per strand. Furthermore, in several cases, the observed genotype may be shared between classes, but occurs with a different probability. Lowercase letters represent OB strand bases, while uppercase letters are OT strand bases.

in WGBS compared to TAPS, which might be a direct result from the lower sequence complexity and higher alignment error following bisulfite treatment

Class	Observed change	A	C	G	T	a	c	g	t	Mistaken with
non-cytosine heterozygous variant	$A \to C$	*	C	*	*	*	*	*	t	None
	$A \to G$	*	*	G	*	*	*	*	t	None
	$A \to T$	*	*	*	T	*	*	*	t	None
	$G \to A$	A	*	*	*	*	c	*	*	None
	$G \to C$	*	C	*	*	*	c	*	*	None
	$G \to T$	*	*	*	T	*	c	*	*	None
	$T \to A$	A	*	*	*	a	*	*	*	None
	$T \to C$	*	C	*	*	a	*	*	*	None
	$T \to G$	*	*	G	*	a	*	*	*	None
	$a \to c$	*	*	*	T	*	c	*	*	None
	$a \to g$	*	*	*	T	*	*	g	*	None
	$a \to t$	*	*	*	T	*	*	*	t	None
	$c \to a$	*	*	G	*	a	*	*	*	None
	$c \to g$	*	*	G	*	*	*	g	*	None
	$c \to t$	*	*	G	*	*	*	*	t	None
	$t \to a$	A	*	*	*	a	*	*	*	None
	$t \to c$	A	*	*	*	*	c	*	*	None
	$t \to g$	A	*	*	*	*	*	g	*	None
non-cytosine homozygous variant	$A \to C$	*	C	*	*	*	*	g	*	None
	$A \to G$	*	*	G	*	*	c	*	*	None
	$A \to T$	*	*	*	T	a	*	*	*	None
	$G \to A$	A	*	*	*	*	*	*	t	None
	$G \to C$	*	C	*	*	*	*	g	*	None
	$G \to T$	*	*	*	T	a	*	*	*	None
	$T \to A$	A	*	*	*	*	*	*	t	None
	$T \to C$	*	C	*	*	*	*	g	*	None
	$T \to G$	*	*	G	*	*	c	*	*	None

Table 4.2: Possible genotypes on non-cytosine bases based on expected allele counts per strand. Errors resulting from sequencing and alignment were excluded.

(see Figure 4.12). The most important features in both datasets concerned read ratios, strand, and context information, where CpG contexts have higher importance compared to other trinucleotides as expected. The modification ratios (on the strand of interest and the opposite strand) were the two highest-ranking features in both TAPS and WGBS, which agreed with the recall and precision values achieved by the simple heuristic that was implemented for calling modification and distinguishing homozygous SNPs from modifications. If highly modified CpGs ($\beta \geq 0.1$) were compared with heterozygous and homozygous SNPs (where the change from the reference is either C>T or G>A they appear as modified) in terms of their ratios of observed modification calls, there was a visible difference in distributions (see Figure 4.11 top). Homozygous SNPs show a bimodal distribution of appearing either fully modified or unmodified. At the same time, in the heterozygous SNP case, the majority of the tentative modifications are around 50%, while in true modifications a higher proportion shows a complete modification. Sequencing coverage of ≥ 3 reads decreased the number of false-positive homozygous SNPs (see Figure 4.11, bottom for an overall and zoomed-in coverage distributions).

Feature normalisation The data need to be normalised before the training of any machine learning model, given that the ranges of the individual features are not on the same scale (mainly if the differences are within several orders of magnitude), as this leads to slower convergence of the algorithm and poor performance. During the development of the models, three types of data normalisation have been applied: min-max, L2-norm, and max-norm scaling.

The vector norm measures the size of a vector (the following definitions are presented from MathWorld [12]). Max (infinity) norm is defined as the maximum absolute value of the vector $\mathbf{x}$ with elements x_i for i $\in$ {1, 2, ... ,n}:

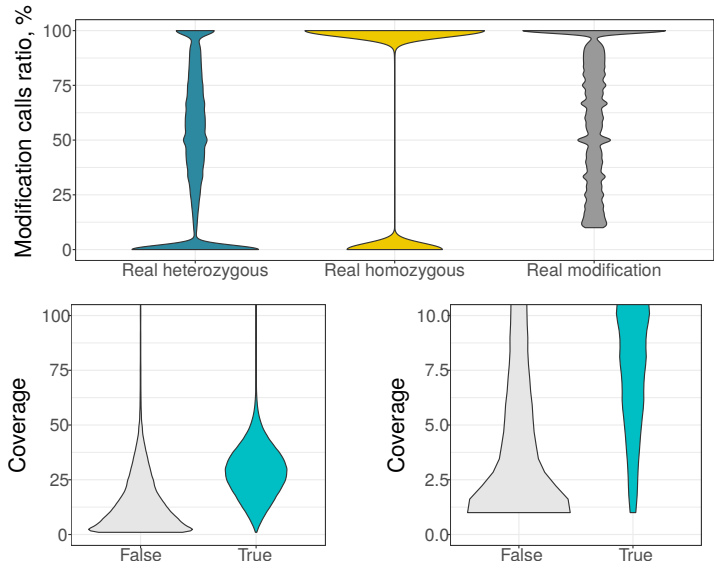


Figure 4.11: Modification calls levels on the strand of interest observed at cytosine positions labeled as modified ($\beta \geq 0.1$), or a heterozygous or a homozygous SNP in the matched WGS samples for patient 0 (top), and the distribution of true and false homozygous SNP is dependent on coverage (bottom). The left bottom figure shows the total coverage distribution of correctly and incorrectly labeled positions as homozygous SNPs, while the right is zoomed-in to show that lower coverages may increase the false-positive rate of the heuristic.

$$\|\mathbf{x}\|_\infty := \max_i \|x_i\| \tag{4.8}$$

L2-norm or Euclidean norm in $\mathbb{R}^n$ is defined as the square root of the sum of squares of the elements of a vector $\mathbf{x}$ with elements x_i for i $\in$ {1, 2, ..., n} or as the dot product between two vectors:

$$\|\mathbf{x}\|_2 := \sqrt{x_1^2 + x_2^2 + ... + x_n^2} \tag{4.9}$$

Min-max scaled data $\mathbf{X}'$ can be presented as an equation featuring the standard deviation σ of the input data $\mathbf{X}$, the range of values of the dataset, and the minimal value:

$$\mathbf{X}' \coloneqq \sigma_x * (max_x - min_x) + min_x \tag{4.10}$$

The min-max normalisation scales the values of the features within a given range, such as [0,1] and [-1,1] with the minimum and the maximum values equal to the boundary values of the interval. In order to apply min-max normalisation, the full dataset needs to be used for transformation which inhibits small batch and online warning unless some modifications are made to the algorithm. One of the main differences between min-max scaling and L2-norm or max-norm is that with unit norms, the data are scaled on each row separately, which enables online learning. Depending on the implementation of the model, each of the normalisation techniques here would be applicable to transform the data before classification.

Another critical issue is the encoding of categorical variables such as strand, context, and class which can be represented either as dummy vectorised variables or one-hot encoding (conversion of categorical data to a sparse vector through an indicator function). One-hot encoding was primarily used for context and strand information, which has the benefit of having the same numeric value for each of the classes but increases the feature space with sparse features. A final important point is the need for regularisation, which decreases the chance of overfitting on the training data. A combination of lasso and ridge regression with different weights (known as elastic net) was used for the logistic regression and the NN model.

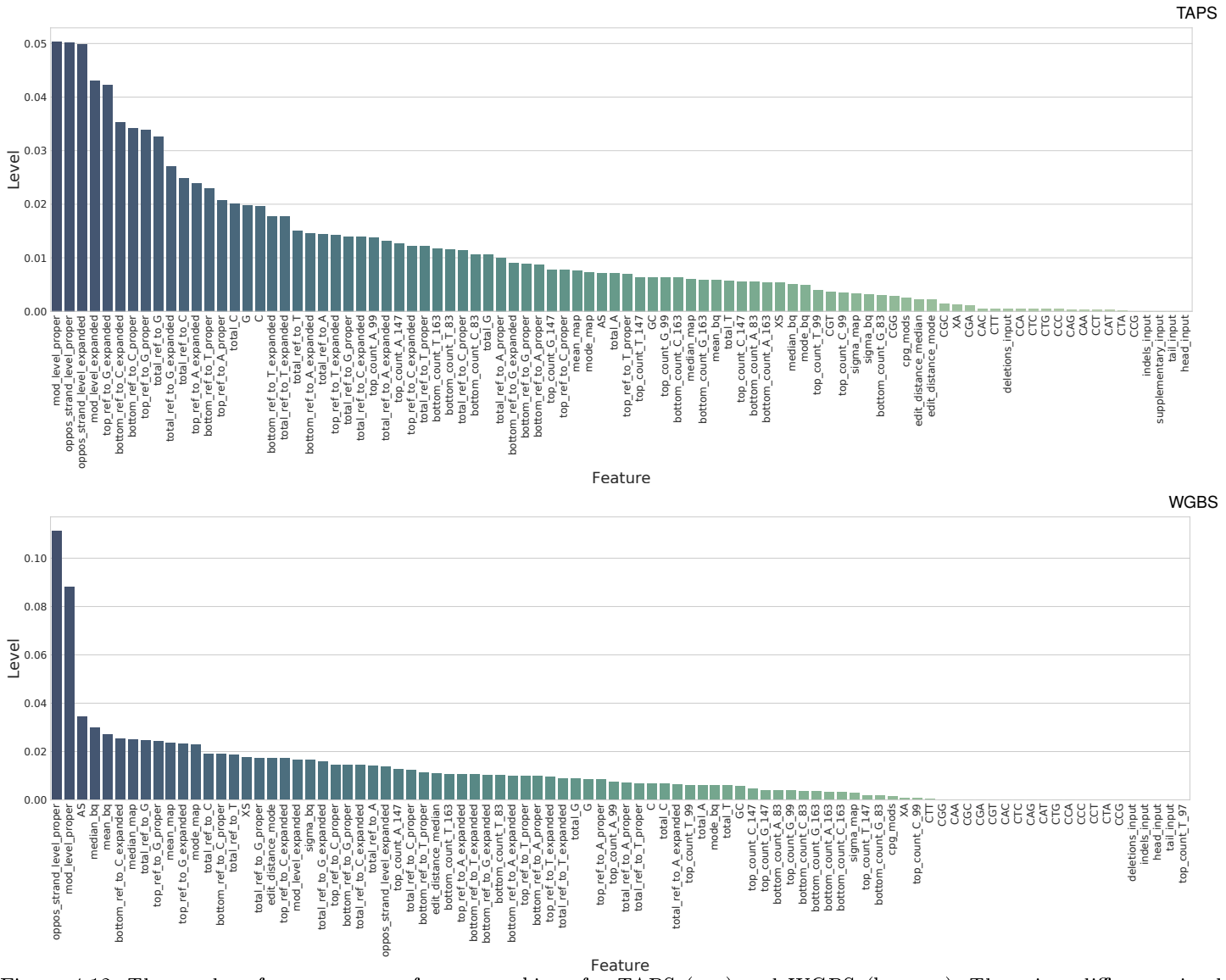


Figure 4.12: The random forest non-zero feature rankings for TAPS (top) and WGBS (bottom). There is a difference in the relative importance of features for classification within the two datasets, with random forest placing more importance on alignment scores and read quality metrics in WGBS data.

4.2.2 Machine learning model

Supervised machine learning was used for model development, as machine learning methods have been used successfully in recent years for biological applications, because of their ability to discover patterns in large datasets. The ML packages and approaches used are described in the following methods section. Different ML techniques were applied to the modification-variant classification problem: neural networks (NN), logistic regression (LR), and random forest (RF). The ML models were trained on a list of cytosine positions, which were labelled as being one of *homozygous SNP*, *heterozygous SNP*, *modified position* and *reference position*. The list of positions labelled as variants were taken from the germline calling of the WGS samples. The remaining cytosines that were not intersected with the (true) SNPs were split into modified and reference, based on their β value with the cutoff being $\beta \geq 0.1$. The training data were highly unbalanced having $\approx 0.1\%$ SNPs and $\approx 4\%$ modifications with the remaining cytosines being representatives of the reference class. Most ML algorithms require balanced datasets for training, although class imbalance may not be the only factor for having poor classification performance [101]. Other contributing factors are dataset sizes, overlapping characteristics between classes, outlier samples, and the existence of subclusters and insufficient data cleaning and preprocessing [37]. In order to solve the class imbalance problem, the Synthetic Minority Oversampling Technique (SMOTE) algorithm [64] was used to generate more synthetic SNPs for training (to the same number as the modifications), and different weights for each class were set. SMOTE is widely used to solve the class imbalance problem by the generation of minority samples through interpolation. Three other methods based on SMOTE were tested: borderline SMOTE [127], SMOTE + Tomek links [339] and SMOTE + ENN (Edited Nearest Neighbor Rule) [366]. These three methods (SMOTE + Tomek and SMOTE + ENN

are described in [37]) use distance metrics to ensure that neighbouring samples have at least n neighbours of the same class or are at close vicinity by similarity. Both SMOTE + ENN and SMOTE + Tomek are combinations of undersampling of the majority class and oversampling of the minority. Borderline SMOTE is an oversampling method that generates new synthetic samples at two steps–1) SMOTE step generation of a new synthetic sample by 5 kNN, and 2) further generation of 'border' samples. These samples have more majority type kNN than minority ones and may prove difficult to classify; thus, they are selectively amplified by the algorithm. The four imbalanced-dataset methods were tested (SMOTE, Borderline SMOTE, SMOTE+Tomek, SMOTE+ENN) and showed similar performance per class on the testing set. The main difference was in run time with Borderline SMOTE, SMOTE+Tomek, and SMOTE+ENN being computationally expensive. Furthermore, Borderline SMOTE was very sensitive to changes in the input weights—the lowering of the weights of the minority classes leads to an increase in false positives. Thus, the regular SMOTE method was used for selective amplification of the minority classes.

General description The data consisting of either sequencing read features on cytosines only (for modification calling and demodification) or the four nucleotides were initially selected from an autosome (chromosomes 1,2 or 3). Still, at a later stage, a random selection of positions from the genome was used. At each training, the data were split in proportion 0.8:0.2 or 0.9:0.1 training and testing and 0.2 of the training samples were used for cross-validation during training. The models were exported and loaded in Python for any demodification and SNP calling experiment.

Logistic regression for multiple classes Logistic regression (LR) is a simple supervised ML algorithm used for classification. Given its clarity, the LR using the full set of features could provide information on performance with

less non-linearity. The model used was multinomial logistic regression allowing for the classification of four classes, and the class assigned-scores by LR can be treated as probabilities. The logistic regression was run with the following parameters: *penalty='elasticnet', solver='saga',* l_1*_ratio=0.1* on the full set of features.

Random forest Random forest (RF) is another type of supervised ML algorithm that is widely used for classification and regression, and it can also provide dimensionality reduction and ranking of features. Algorithms such as RF can be applied to small or imbalanced datasets which explain the popularity of RF for solving problems in the biomedical sciences. The default scikit-learn implementation of a RF with parameters (*bootstrap=True, criterion='entropy', class_weight='balanced_subsample', n_estimators=100*) was trained on the full set of features. The feature importance of RF was used for feature selection for the NN model.

Performance evaluation In order to compare the performance of models to distinguish modifications from variants in 1) a cytosine-only case (suitable for modification calling and demodification) and in 2) a four-nucleotide case (suitable for germline variant calling) different metrics were used including accuracy, F_1 score, precision, recall and AUC. If we consider the class of interest to be a positive class, all positive class members that are labelled correctly are *true positives* (TP), while the misclassified as positives are *false positives* (FP). At the same time, the correctly classified as being of the negative class are *true negatives* (TN), and when mislabelled—*false negatives* (FN). A commonly used method to visualise the TP, TN, FP, FN is to use a contingency table known as a *confusion table*. The accuracy represents the number of correct predictions to all predictions made by the model (equation 4.11) and could be problematic for highly imbalanced datasets where misclassifying all of the minority classes

as the majority would still result in high accuracy.

$$accuracy = \frac{\sum \text{true positives} + \sum \text{true negaties}}{\sum \text{all cases}} \tag{4.11}$$

Other types of metrics used to evaluate classification performance for imbalanced class data are precision, recall, and F_1 scores. We can describe precision (the positive predictive value) as the ratio of the true positive predictions to all positive predictions made by the classifier (equation 4.12); recall (also known as sensitivity and true positive rate, TPR) as a measure of how many true positive predictions were made to all real positives (equation 4.13). The F_1 score is a specific balanced ratio of the product of precision and recall and their sum (equation 4.14).

$$precision = \frac{\sum \text{true positives}}{\sum \text{true positives} + \sum \text{false positives}} \tag{4.12}$$

$$recall = \frac{\sum \text{true positives}}{\sum \text{true positives} + \sum \text{false negatives}} \tag{4.13}$$

$$F1 = 2\frac{\text{precision x recall}}{\text{precision} + \text{recall}} \tag{4.14}$$

ROC curves are used to visualise classifier performance (or the general discriminative ability of a model) at different thresholds, with the area under the curve (AUC) used as an evaluation metric [129], [51]. When comparing models, the higher AUC scores mean better predictive performance, as AUC can be interpreted as the probability that a random member of the positive class will receive a higher ranking compared to a random member of the negative class. Given the multi-class classification problem, AUC is measured for classifying the class of interest versus all other classes. The main components of the

ROC curves are the TPR and the false-positive rate (FPR). The FPR (which is 1-specificity) of a classification problem can be described as follows:

$$FPR = \frac{\sum \text{false positives}}{\sum \text{all negatives}} \tag{4.15}$$

Finally, the receiver operating characteristic (ROC) curves for different models were used to visualise the tradeoff of sensitivity and specificity of the models, as sensitivity is inversely proportional to specificity.

4.2.3 Neural networks model

Feed-forward networks with different architectures and hyperparameters were used over the course of model development. In order to increase the training speed and decrease the chance of overfitting batch normalisation [155] was used. The architecture of the currently evaluated model is given in Figure 7.

General description Given the multi-class classification problem, the softmax function was used as an activation function in the outer layer and the network was trained using the categorical cross-entropy metric. The softmax function takes a vector of real numbers **z** with elements z ∈ {1, 2, ..., k} and returns a probability distribution where the elements of the vector were normalised proportionally to their exponent. The softmax function can be written as:

$$softmax(\mathbf{z})_i = \frac{e^{z_i}}{\sum_k e^{z_k}} \tag{4.16}$$

The categorical cross-entropy in classification problems measures the information difference between the probability distribution of the predicted and the true labels (for multinomial classification) and is commonly used as a loss func-

tion. The use of softmax as an activation function in the output layer together with the cross-entropy loss enables the treatment of the class assigned scores as probabilities.

Model development steps The design of the model used for the classification of variants and modifications required considering the following model characteristics.

1. The classification type: the problem of discovering SNPs and modifications can be broken down into a series of binary classification problems, or a single multiclass (categorical) classification. There are two main issues that relate to the choice of how to solve the problem—1) the majority of the single binary comparisons would be the classification of a reference versus non-reference, which could speed up the algorithm, 2) subpopulations of samples that resemble different class than their own could bias the results by a different degree compared to multiclass classification. At present, only multiclass classification models are discussed.

2. The choice of architecture: larger layers with many neurons show greater flexibility in learning finely adjusting weights. However, such networks are at risk of overfitting and require regularisation to prevent poor generalisation due to model sensitisation to training data's noise. Deeper networks also require more time and resources to train, and transfer learning (the use of pre-trained models when data of certain type is abundant compared to others but the data or the problem needs to be similar) for TAPS data was not available. Empirically testing model architecture with one to ten layers, with 2-4 to several hundred neurons per layer showed that NN needs at least two layers to learn the problem effectively on the full dataset. The selection of features is of great importance to classification, together with the choice of the sample used for training, because of its intrinsic biases.

3. The issue of class imbalance: mutations are rare events and modifications which occur more frequently follow specific patterns (being largely dependent on the DNA sequence), and we treat all positions as being independent which decreases the model complexity (if no independence was assumed, the problem would have required a temporal model like LSTM, an acronym for long short-term memory), removing location-specific characteristics which may aid classification. Furthermore, sampling and handling of the data can introduce bias that can make members of two classes more similar (such as low coverage, alignment bias, low cellularity for tumours, general mixed populations such as cell-free DNA). In order to decrease the number of false negatives, the model needs to be sensitive to the minority classes, which was ensured though weights and algorithms for handling imbalanced data. Taking the frequency of each class as weight could increase the number of reference calls; thus, the selection of weights for the model is needed prior to classification on the full dataset.

4. The type of regularisation to apply: regularisation is needed to prevent overfitting, but the selection of regularisation type depends on the problem and the parameters. L1-regularisation can act as a feature ranking technique putting zero weights to features of low importance, while L2-regularisation only changes the value of the weights. Using a combination of both types of regularisation is expected to decrease the chance for overfitting by simultaneous shrinkage and feature selection [390].

5. The hyperparameter tuning: hyperparameter tuning is a process that can be automated but requires empirical evaluation on the problem at hand. Deciding on the activation function, batch size, learning rate, learning algorithm, and others could significantly change the results of the classification. Two stages of hyperparameter tuning were conducted: empirical

on a subset of the dataset where certain parameters were varied and automated using a specific Python package.

Given the importance of class weights and NN architecture, further details about the experiments with which they were studied are presented in supplementary section I.

Hyperparameter tuning Following the empirical testing of hyperparameters, automated hyperparameter tuning was done for validation. It was used on a subsample of the primary training sample ($1x10^6$ positions) through RandomSearch from keras-tuner on a subset of the data. The hyperparameter search space was over the number of neurons, the activation function, L1 and L2 regularisation values, and the learning rate, and the best performing configurations were tested. The model created empirically was performing on par with the predicted best models.

4.3 Demodification variant discovery

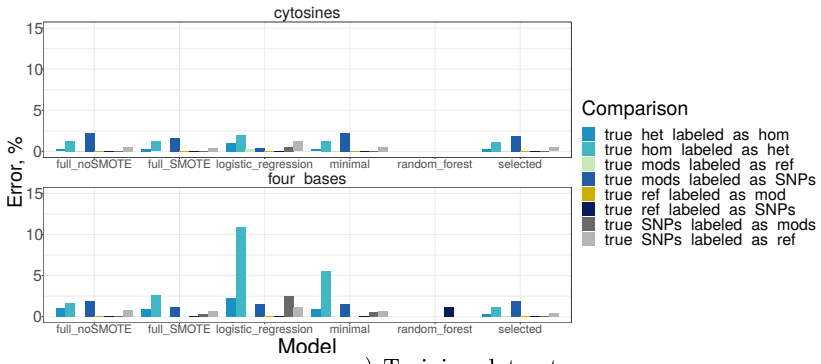
The performance of the described NN, LR, and RF models was assessed on WGBS and TAPS data containing only labelled cytosine positions and their extracted features. In both TAPS and WGBS, the main source of error were heterozygous SNPs, and the worst-performing model was logistic regression. Modifications and reference both had high precision and recall values ($\geq 99\%$) using the full, selected, and minimal features NN model as well as with RF, suggesting that a reason for the higher misclassification of SNPs was their limited number in the training dataset. Another clue that the scarcity of heterozygous and homozygous SNPs could partially explain the worse performance for both classes was further confirmed by the full NN model without SMOTE amplification of the minority classes. This model performed better on the training set compared to the testing set due to insufficient generalisation (see Figure 4.14).

The performance of the model with selected features through the RF feature ranking was comparable to the full model with high precision and recall in both WGBS and TAPS. This observation confirmed that the selected features have a high information load for class separation. Overall WGBS classification on cytosines-only showed lower precision and recall compared to TAPS. The error rates for true modifications labeled as reference were $\approx$ 0.0005 to 0.3% in TAPS and between 0.002 and 0.7% in WGBS; the reverse case of true reference positions labeled as modifications was between 0 and 0.02% in TAPS and between 0 and 0.57% in WGBS showing again that the higher error rates in WGBS complicate classification even for the reference compared to TAPS (top row on Figures 4.13, 4.15). Given the problem of true SNPs being incorrectly labelled in chemically converted sequencing data, the most important errors are the true modifications labelled as SNPs, the true SNPs labelled as modifications, and true SNPs labelled as reference. Given the position labelling based on the SNPs found in the matching WGS samples, there was a possibility for a true variant or modification considered as an error because of natural variation and biological sampling. However, a high agreement between the datasets in terms of genetic variation was expected. The SNP categories heterozygous and homozygous are combined for simplicity. Not demodifing a SNP regardless of its class would result in a correct variant call, but separate misclassification rates were calculated for heterozygous and homozygous SNPs. The error rates between the two SNP groups were relatively low in the cytosines-only data (0.3-1% TAPS, 0.18-2% WGBS for true heterozygous labelled as a homozygous SNP, and 1-2% in TAPS and 0.65-1.2% for WGBS for true homozygous SNPs misclassified as heterozygous). It appeared that in the WGBS data, there was a higher variation in between SNP class performance than in TAPS. In the demodification case, no known SNPs information was provided using dbSNP. Still, the true

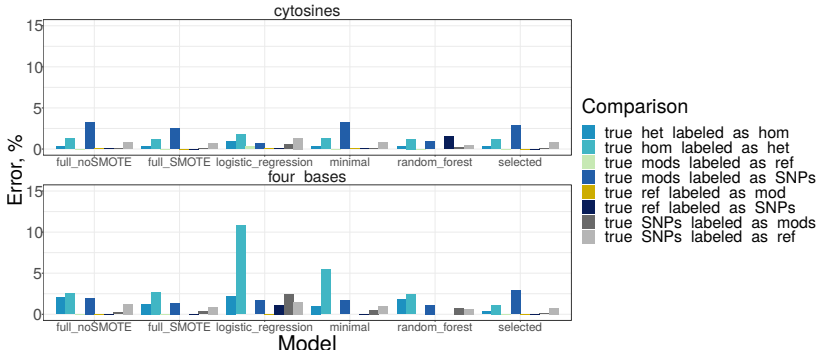
SNP labelled as modifications were very few (error rates in TAPS 0-0.09% and 0-0.02% in WGBS) in all models apart from the logistic regression ($\approx$ 0.58% TAPS and $\approx$ 0.44% WGBS). The true modification labelled as an SNP error rate was higher (error rates 0-3.3% in TAPS and 0-2.2% WGBS), and the true SNP misclassified as reference was 0-1.3% for TAPS and 0-3% in WGBS. These two types of error were partially resulting from the limitations of the dataset where fewer diverse SNPs exist (and these were computationally amplified), but there was a decrease in the 'majority' types of errors because of SMOTE. Only in the random forest model, a sizable population of reference positions was misclassified as SNPs. The performance was slightly lower overall for WGBS, with a higher percentage of SNPs seen as a reference compared to TAPS. A possible explanation why the overall SNP to modification rates were lower in WGBS compared to TAPS (where any errors with the reference were low) was that the majority of cytosines in the bisulfite-converted genome do not match the reference (the reference positions would be converted and appear as C>T variants), while only modifications would stay as cytosines. In TAPS the highest expected error rate would be with C>T variants, which are common, while the reference would remain unchanged apart from the sequencing and alignment errors. Another contributing factor was the higher sequencing depth of WGBS data compared to TAPS which can provide more information about the true state of the cytosine position.

A higher error rate was observed in OAC and Barrett's oesophagus tissue in comparison to the duodenum. The primary inflation in incorrect calls, as expected, was due to heterozygous SNPs, where additional SNP specific information can be used to improve the reliability of classification. Given the importance of the known SNPs information, a second model was created by adding dbSNP information as an additional feature to the existing models. This modification

expanded the possible applications of the model not only for modification calling and demodification on cytosines but also for calling SNPs on TAPS and WGBS. More details about its performance are provided in the next section.



a) Training dataset



b) Testing dataset

Figure 4.13: Error rates comparisons for TAPS classification using either cytosines or all four bases. The logistic regression had the highest error rates, but the error rates of all tested models were low for TAPS.

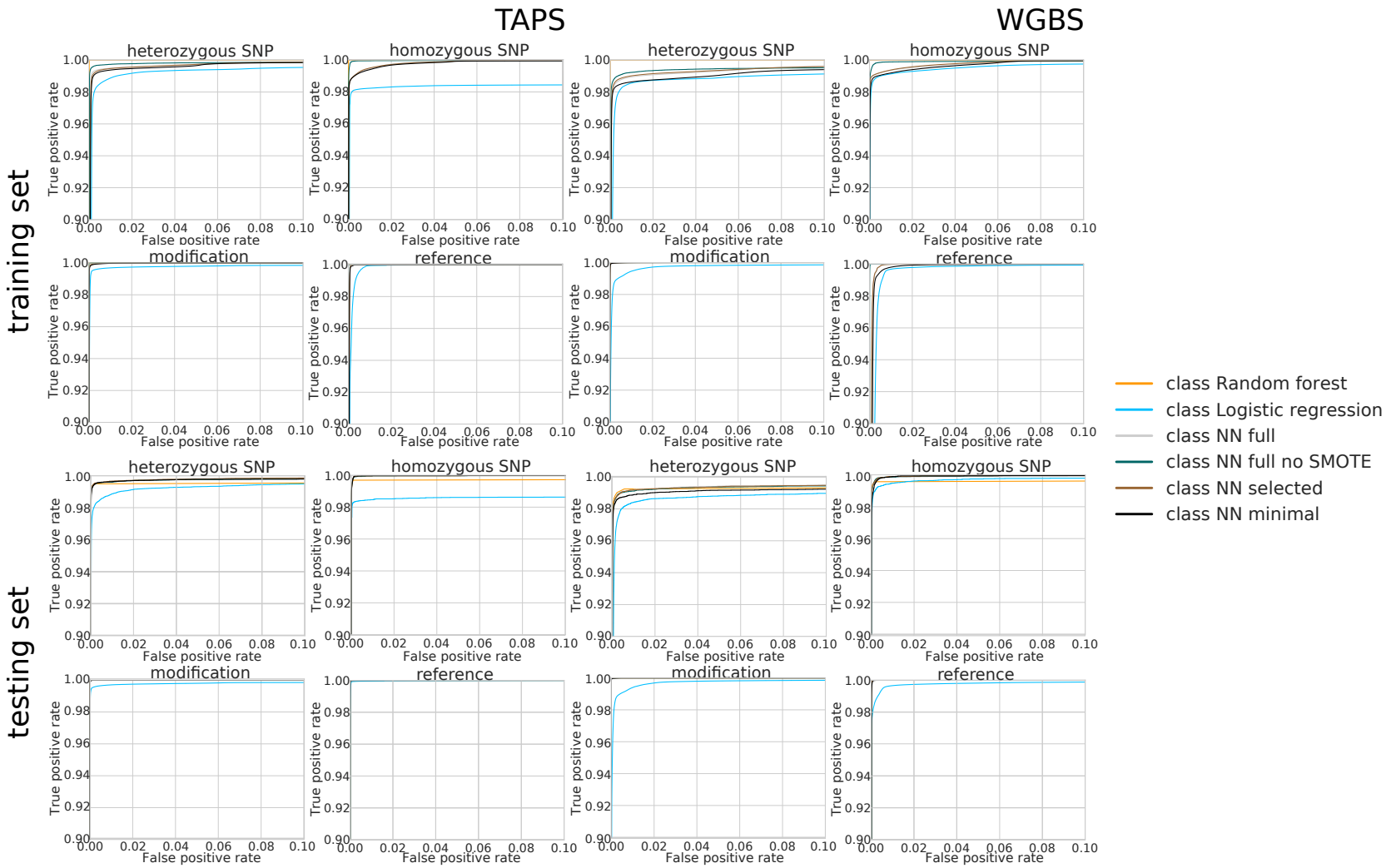
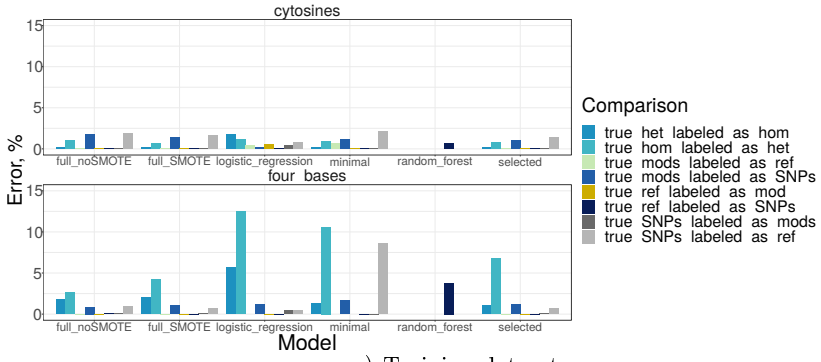
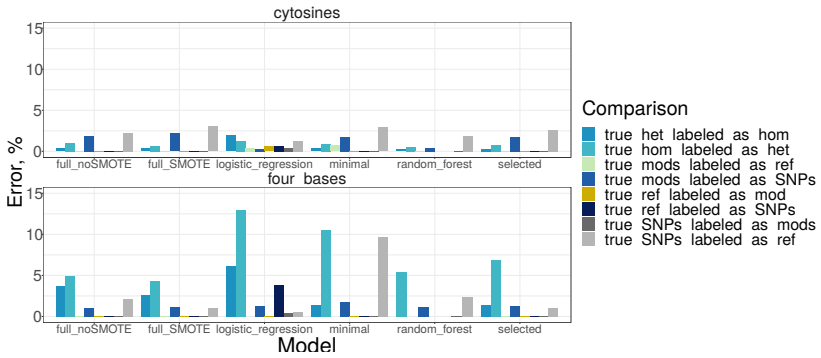


Figure 4.14: FPR and TPR (ROC) curves for TAPS and WGBS classification only on cytosines. The best performing models were the random forest and the full and selected features NN models. The classification of cytosine positions had fewer errors in TAPS compared to WGBS.



a) Training dataset



b) Testing dataset

Figure 4.15: Error rates comparisons for WGBS classification using either cytosines or all four bases. The logistic regression had the highest error rates, but the error rates of all tested models were higher for WGBS compared to TAPS.

4.4 Germline variant discovery

In the case of the germline variant classifier, random genomic positions from the healthy matched tissues were used for training. The labels used were the agreement of the germline classifiers on the WGS data. As an additional feature, the count of known SNPs from dbDNP v146 intersecting the position was added to increase the sensitivity of the model. The performance differences between germline variant calling in TAPS and WGBS were greater compared to the cytosines-only case with lower values observed in WGBS relative to TAPS precision and recall in both the test and the train sets (see Figure 4.16). Both

modification and reference classes have $\geq 99\%$ precision and recall for all models except for the logistic regression in TAPS. In contrast, in WGBS again, the reference class has lower recall compared to modifications. The real reference positions labeled as SNPs in TAPS have an error rate of 0-0.009% in TAPS (logistic regression $\approx 1.1\%$) and 0-0.05% (logistic regression $\approx 3.8\%$) in WGBS (see Figures 4.13, 4.15). Reciprocally, the true variants labelled as reference were 0-1.5% in TAPS and 0-9.6% in WGBS. In order for the model to be used for haplocalling, the two classes of SNPs need to be separated with sufficient accuracy. However, the general inflation of the misclassification error between heterozygous and homozygous SNPs was observed. The most probable reason was the absolute count of variants and their relative distribution given the trinucleotide context, which was used as a feature. The differences in the SNP rarity depending on the context could potentially bias the model for certain more commonly mutated contexts. The error rate of true heterozygous SNPs marked as homozygous was 0-2.2% in TAPS and 0-6% in WGBS. In the opposite case, when homozygous SNPs were misclassified as heterozygous variants are 0-10% in TAPS and 0-12.9% for WGBS data. The logistic regression errors of misclassifying one type of a germline variant for another was the highest reaching over 10-12%. Although a possible explanation of why there was slight inflation in the error rates of incorrectly classified homozygous SNPs could be sampling and sequencing coverage, another explanation was that homozygous SNPs are rarer compared to the heterozygous variants. The rate of real SNPs labelled as modifications was 0-0.55% (LR 2.5%) in TAPS and 0-0.1% (LR 0.44%) in WGBS. Finally, the rate of modifications labelled as SNPs was again higher in WGBS with 0-1.5% in TAPS versus 0-9.6% in WGBS confirming the suggestion that bisulfite chemistry may complicate the classification of C>T SNPs because of their similarity to the reference.

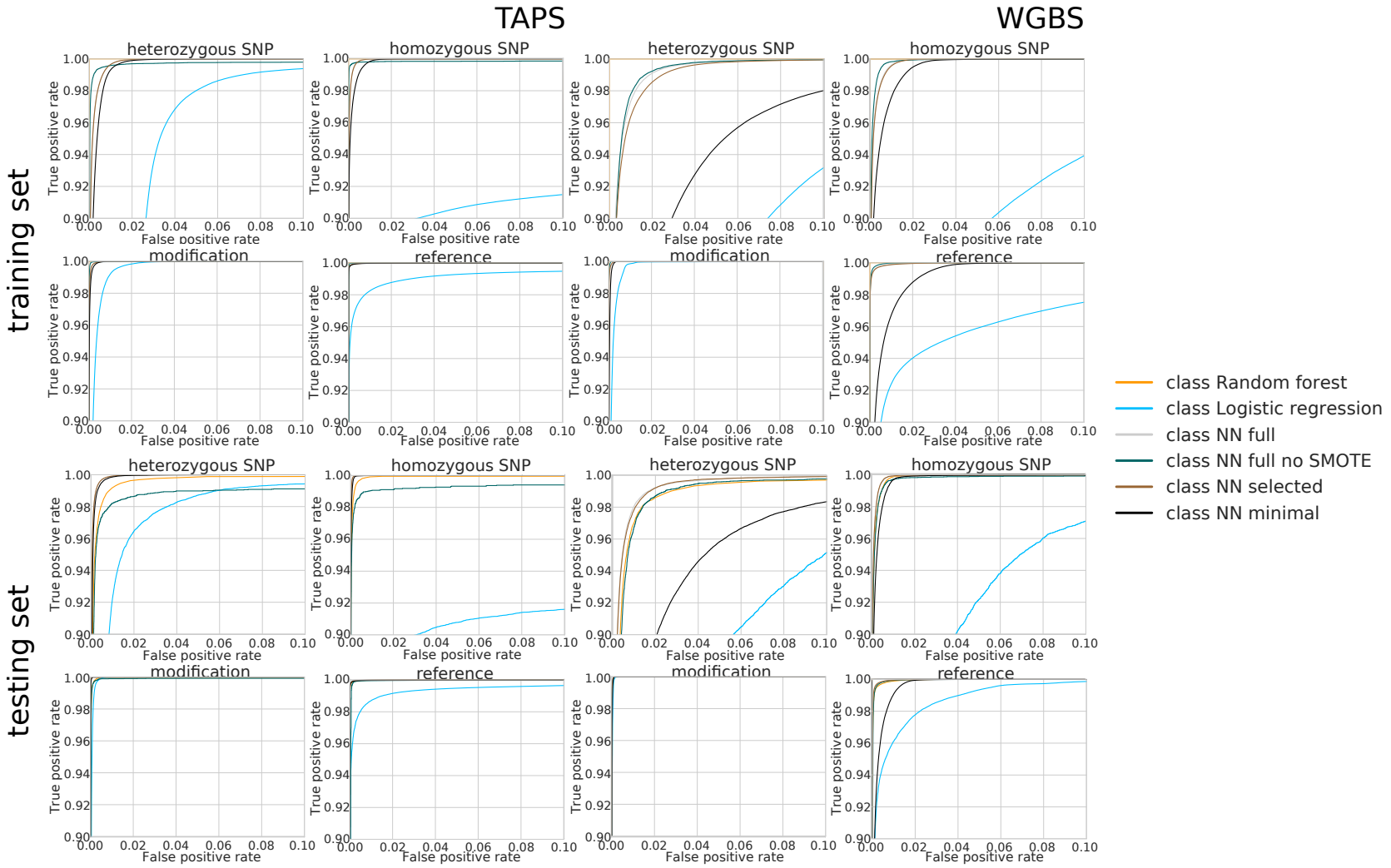


Figure 4.16: FPR and TPR (ROC) curves for TAPS and WGBS classification on all four nucleotides. The best performing models were the random forest and the full and selected features NN models. The classification of cytosine positions had fewer errors in TAPS compared to WGBS.

4.5 Error analysis

Identifying exactly where and why the errors arise could help to improve the performance of the models and the generalised demodification pipeline through additional filtering steps. For the error analysis, only the full NN model with SMOTE was applied on chromosome 21 of the same patient where the random training data were taken. Although there was a risk of duplication of training positions, this dataset can be treated as a validation set to study better the models before applying them on sequencing data from other patients.

After comparing feature distributions between the correctly and incorrectly classified SNPs in the cytosines-only and the germline calling TAPS model (full NN with SMOTE), specific differences were visible: variants having twice or more the mean sample coverage have a higher chance of being incorrectly labelled as SNPs particularly for homozygous SNPs (see Figure 4.17). Furthermore, misclassified SNPs have higher BWA secondary alignment scores (XS) as well as more prevalent secondary alignments (XA), and more often have no real known SNPs from dbSNP. In the WGBS case, again, the incorrectly labelled as homozygous SNPs positions appear to have a lower base and MAPQ qualities, and higher XA and XS ratios (see Figure 4.18). These results indicated that alignment errors and possible copy number variation can inflate the number of false-positive calls.

Taking all classified positions and comparing their label with the expected label was used for this analysis. The number of correctly and incorrectly labelled positions per context was normalised to the total number of each group separately, which enabled the comparison of frequency distribution between them. In the TAPS samples, there was a different error pattern given the origin of the sample (OAC or duodenum) (see Figure 4.19) that was not observed in WGBS

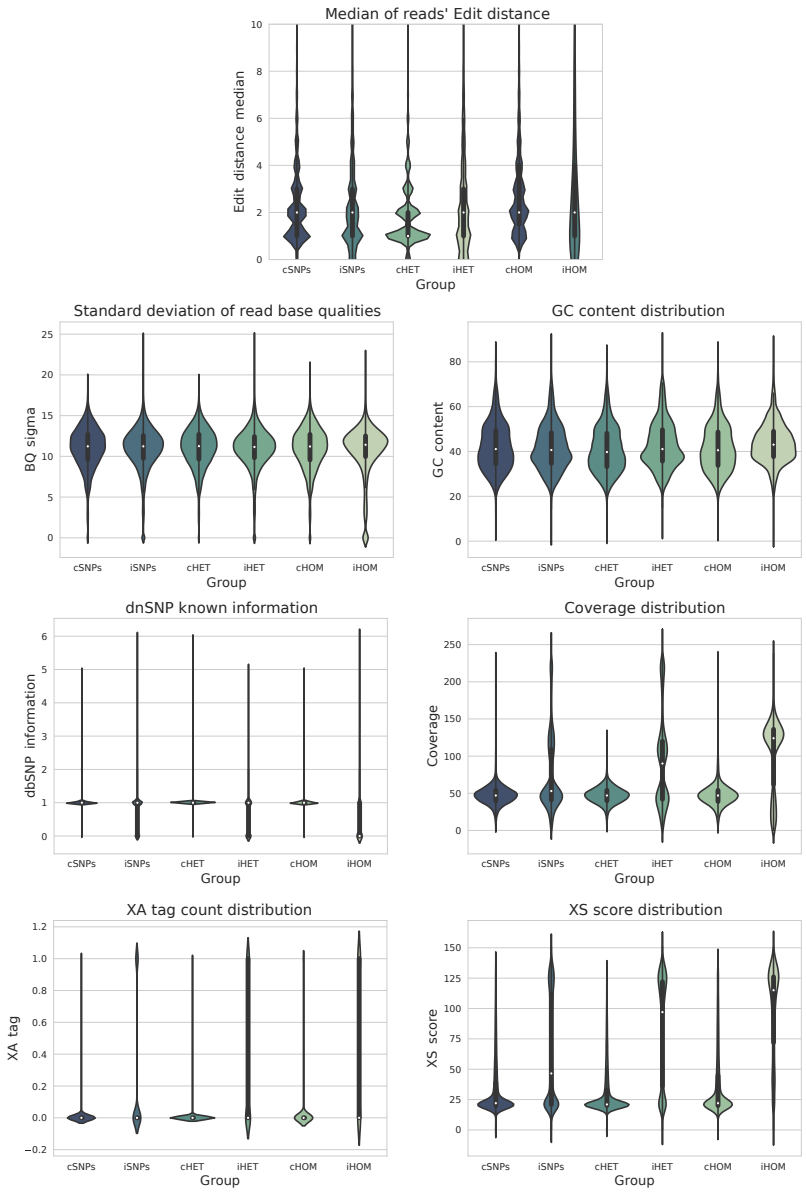


Figure 4.17: Distributions of different features of correct and incorrect SNPs in TAPS. Features such as the Edit distance, the inclusion of the position in the dbSNP database, the coverage, and certain alignment tags and scores such as the XA and the XS showed promising differences between true and false variants.

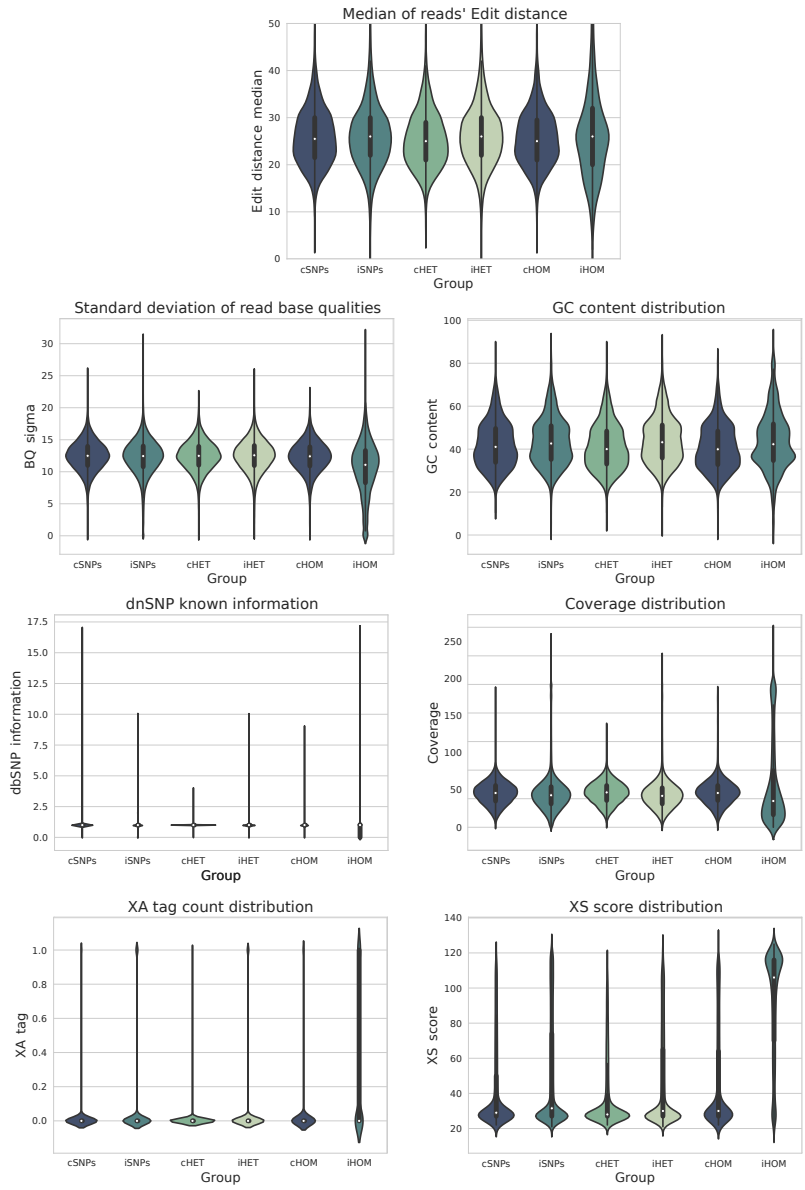


Figure 4.18: Distributions of different features of correct and incorrect SNPs in WGBS. Features such as the Edit distance, the inclusion of the position in the dbSNP database, the coverage, and certain alignment tags and scores such as the XA and the XS showed promising differences between true and false variants. The differences were less pronouced in WGBS compared to TAPS.

where the differences were mostly quantitative (see Figure 4.20). In both WGBS and TAPS, regardless of the tissue of origin, the most common positions of errors were cytosine contexts and TTT, and the highest relative error frequencies were in the homozygous case. In WGBS, two additional contexts had elevated error rates in comparison to others: TAT and TCT only in the homozygous SNP case. A higher level of errors in CpG contexts was observed in the heterozygous SNP case in TAPS, wherein in the OAC tissue both CpG and non-CpG errors were elevated while in duodenum only the CpG errors were higher. For homozygous SNPs in TAPS, the pattern showed a higher non-CpG error rate, particularly in OAC sample. In WGBS the non-CpG context was elevated in all cases, as the more common classification error is expected to be with the reference in comparison to TAPS where the errors are highest when classifying SNPs versus modifications.

Next, the exact locations of the misclassified positions were used to evaluate whether a particular class of errors was more prevalent in any type of genomic feature. For this purpose, the incorrect SNP calls were intersected with hg38 genomic features, and the count of the intersection was normalised to the total number of positions spanning the feature class in chromosome 21. In TAPS, approximately 4-5% of the incorrect SNP calls fell within CGIs compared to only 0.37-0.39% in the true variants. There were more significant proportions of correct SNP calls within transcripts (7-9% in true variants and 0.5% in the incorrect calls), DNase accessible regions (12-13% versus 0% in the incorrect calls), and longer repeats such as LINEs (9-11% versus 3-5%) and LTR (8% versus 2-3%). Conversely, the simple and satellite repeats have a greater proportion of incorrectly labelled SNPs (1.18-2.80 times more in the low complexity repeats, 3.87-4.51 times more in the satellite repeats, and 1.9-2.28 time more frequent in the simple repeats) particularly in homozygous SNPs except for the

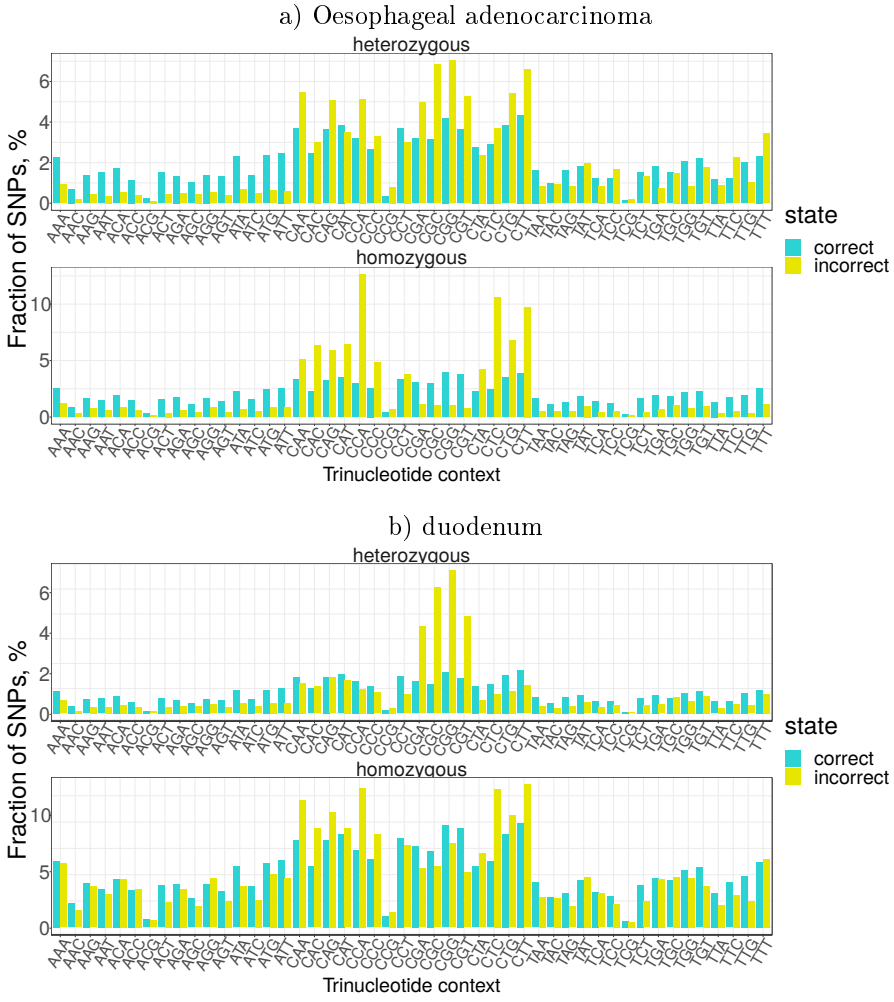


Figure 4.19: The dependency of the mislabeled positions and their trinucleotide context in TAPS. The correct and incorrect counts were normalised separately with the sum of all positions per group. Higher overall fractions of incorrect SNPs are observed in CpG contexts, and specific CC and CA contexts.

simple repeats. Also, incorrectly labelled homozygous SNPs appear to be three times less common in exons compared to the true homozygous SNPs. These results confirmed that regions with higher CpG methylation would have more false-positive SNP calls in TAPS due to the change in the expected frequency of

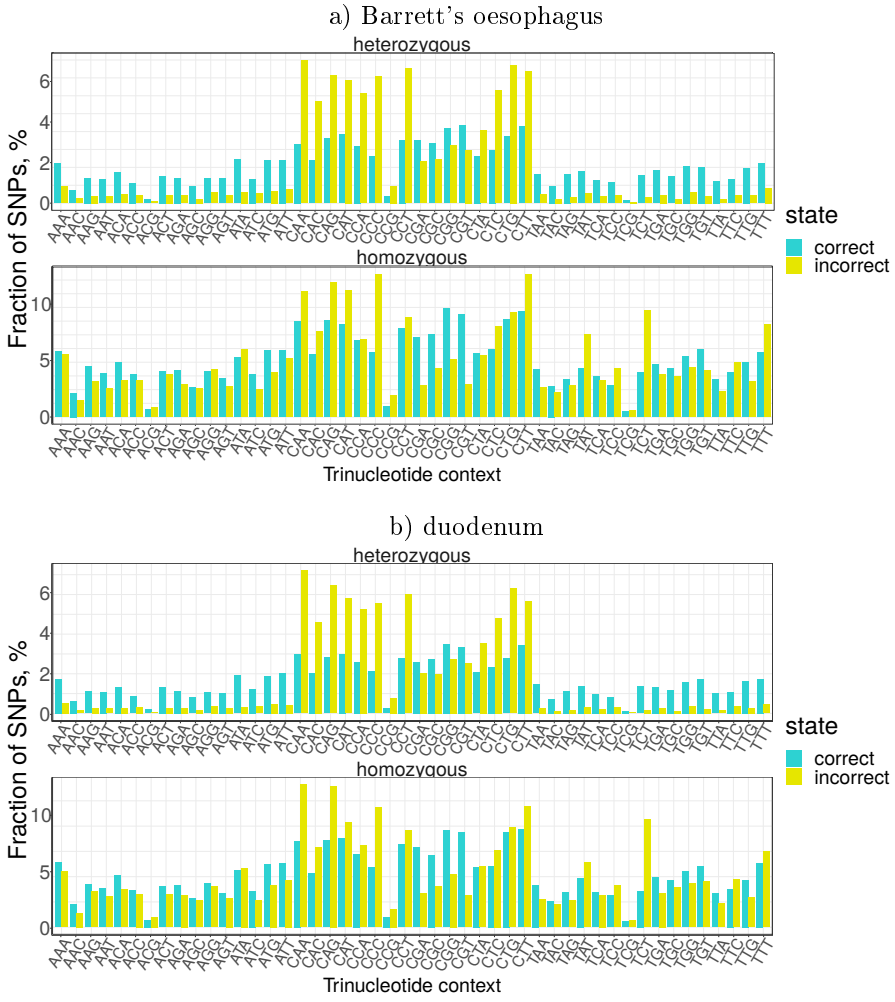


Figure 4.20: The dependency of the mislabeled positions and their trinucleotide context in WGBS. The correct and incorrect counts were normalised separately with the sum of all positions per group. Higher overall fractions of incorrect SNPs are observed in CpG contexts, and specific CC, CA, TT and TC contexts.

modification versus germline variation events and that alignment errors in lowly mappable regions may contribute to the false-positive calls, particularly for false homozygous SNPs. In WGBS fewer incorrect SNP calls (both heterozygous and homozygous) were made on the introns (6% in the true calls versus 3-5% in

the false calls), DNase accessible regions (11-13% versus 6-9%), LINEs (9% in the true calls versus 4-6% in the false calls), LTR (7% in the true calls versus 2-5% in the false calls), but the count of incorrect calls was elevated in the simple repeats. The errors in the low complexity repeats were 1.14-3.69 times more common than the true calls, 1.68-2.75 times more frequent in the simple repeats, and enriched 1.55-2.32 times in the satellite repeats with the higher values for the homozygous SNPs. These results suggested that false-positive calls in WGBS were influenced by misalignment, where the reference is mistaken with variants.

4.6 Discussion and future directions

The models presented in this chapter achieved classification metrics suggesting good performance in the situation where SNPs and modifications need to be identified in TAPS and WGBS data. Applying model-based or hard filtering as a second step would further decrease the number of misclassification events. Given the performance on the testing sets and the possibility of secondary inclusion of a filtering step, the models are applicable to real-world epigenomic sequencing data. However, the current models were trained on limited datasets where the germline variants were naturally ≥ 100 times fewer than the modifications. This imbalance was partially addressed by synthetically amplifying the number of variants in the dataset. Nevertheless, such an approach cannot increase the natural variability of variants as it generates novel positions by similarity to n neighbours of the same class.The synthetic amplification of variants was required due to the limited TAPS data at present. Although databases containing publicly available variant calls are available, their direct utility would be limited for TAPS or WGBS variant calling due to the specific characteristics of these epigenomic methods. An approach to learn variant distributions from data

would be to use transfer learning where models are trained on WGS data and then modified for TAPS and WGBS data. The error rate for the classification of SNPs is expected to decrease when there is a representative sample for germline variants in each trinucleotide context. Also, a training set including not only more SNPs but positions with different coverages and sequencing and mapping qualities may improve model performance by providing additional information with varying levels of noise. A model trained on more diverse data would be less likely to overfit and would capture to a greater extent general patterns in the data. A possible approach would be to use data augmentation where the training dataset includes similar data created by manipulation of the existing dataset [200]. Furthermore, the current models focused only on human samples, where the methylation occurs almost exclusively on CpGs, and databases containing rich information on genetic variation exist. However, for some organisms, SNP information may not be readily available, and that would require the addition of further conservative filtering of the discovered variants. Also, organisms such as plants have non-CpG modifications predominantly, thus creating a plant-specific model would allow for studying simultaneously SNPs and DNA modifications in other taxa. The current TAPS dataset included a single training sample (due to limited availability at the time of the analysis). Still, in the case of WGBS data, it is possible to obtain additional data from publicly available datasets to be included in the training dataset. Furthermore, the conversion rates per context were not provided as features in the classifiers, given the data limitations (no differences of the conversion rates exist within a sample). Such information may improve prediction if some estimate of the false-positive and false-negative rates can be calculated for the observed modification calls. Finally, given that a fraction of the false SNP calls occur in regions of simple repeats, the mappability of the region can be added as a feature to the model.

At present, the proposed variant classification is a one-step process, but having a filtering step with another model that distinguishes with high accuracy between correct and incorrect variants could further improve overall performance. Given that there are differences between the true and false SNP calls, in terms of their feature distributions that could potentially be explored in the future. Theoretically, classifying modifications correctly from SNPs should be eased by the more significant differences between the reference and SNP positions in WGBS compared to modification versus heterozygous SNP problem in TAPS. However, due to the greater degree of DNA degradation and pervasive conversion of unmodified cytosines in WGBS, variant calling is complicated by alignment and systematic errors resulting from the chemistry of the method [335]. A practical optimisation, particularly for biomedical applications, would include the simultaneous running of TAPS sequencing with a SNP array [193] ideally including cytosine-only positions (such customised array can be used for both methods converting modified cytosines to thymines and unmodified cytosines to thymines, *e.g.* TAPS, WGBS, RRBS). The application of SNP arrays in TAPS they will boost SNP and somatic calls discovery by providing true SNP information that can be applied for direct demodification, or used as a model parameter of variant callers. An alternative would be to use panel sequencing, where specific genomic regions can be studied at lower costs [327]. As a final point, none of the classifiers presented in this chapter is directly applicable for somatic variant calling with matched tumour-normal samples. In order for the methods to be relevant for such problems, the combined feature sets of the control and the tumour samples have to be used for training.

From a computational perspective, different approaches for the sensitive discovery of SNPs in TAPS and WGBS exist. One of these directions is to apply a machine learning model that uses positional information which is expected to

improve classification for modified positions versus SNPs in the CpG context (*e.g.* RNN, LSTM models) in TAPS, and including information about the genomic region for WGBS. Machine learning models using spatial and temporal dependencies between elements are commonly used for the analysis of sequential data such as text in natural language processing where the probability of certain words depends on the surrounding words [381], [312]. The greatest benefit of such models is that no independence is assumed between the elements in space [354], and adding sequence context information in the models can potentially improve the classification of modifications and variants. Several examples for successful application of RNN and LSTM models in the biomedical sciences are for disease detection [149], [356], protein structure prediction [124], [207], RNA binding [107], and correction of errors in sequencing data [19].

Another future direction for creating a sensitive classifier for epigenomic data is to use a generative model. The most vital points of using a generative probabilistic model are interpretability of results and handling missing data. Thus, data availability and complete representation (having representative labels of all true classes) of data are not limiting factors [45], [208]. Another limitation of the current method is its dependence on SNP information (here dbSNP). However, such information may not be available or extensive enough for some organisms of scientific interest. A possible solution would be the incorporation of a probabilistic model as one feature of the ML model. Such a probabilistic model may output probabilities of whether a position is a modification or a variant, that could further improve the ML classifier. Bayesian networks represented by both simple mixture models and complex multi-parameter networks can capture the uncertainty of the underlying processes and have been applied widely for biological questions such as gene-protein interactions, medical diagnosis, and epidemiological research [183]. Given the stochastic nature of mutational pro-

cesses, a possible approach to express our assumptions is through a Bayesian mixture model, where homozygous variants, heterozygous variants, and modifications are subpopulations within the cytosine population. The probability of a cytosine site falling within a particular class depends on the calculation of the joint distribution. Given the hierarchical structure of the model, the joint likelihood can be calculated through marginalisation or maximum likelihood estimation techniques. Here the unmodified cytosine, mixed mutation, and modification cases are also included. The priors for each class are dependent on the observed context with higher expected mutation and particularly modification rates in CpG (in vertebrate genomes and particularly mammalian samples). The *error* category denotes cytosine positions that do not have enough sequencing depth to be confidently classified in any group where the Bayesian model would provide a classification label based on the priors.

The current TAPS protocol is optimised for CpG methylation with different levels of non-conversion for non-CpG contexts and 5-hmC, 5-fC, and 5-caC (the last two were not empirically validated with sequencing and bioinformatic analysis). The observed modification level Z_{obs} is equal to the sum of true modification level Z_{tr} and an error term ϵ. The error term can be further explained by context, modification type, PCR bias, sequencing bias, and errors from the bioinformatic pipeline (base quality scores, alignment scores), but in order to simplify our model for the modification versus mutation calling, we can assume that the error term is constant in all cases. The status C of a cytosine given that there is a difference from the reference at this position can be predicted by several parameters: context $c \in \{CHH, CHG, CG, CN\}$ where $H \in \{A, C, T\}$, allele frequency f ($f_i \in \{A_{OT|OB|total}, C_{OT|OB|total}, G_{OT|OB|total}, T_{OT|OB|total}\}$), base b and mapping m quality scores (their σ at present), primary p and secondary s alignment scores, beta scores β and opposite strand β scores. In this model, we

can consider the strand ρ as a hyperparameter that will modify the allele frequency expectation for each of the classes. Furthermore, the context influences the prior expectation of modification versus mutations with the CpG contexts having a considerably higher rate for both single-nucleotide changes and modifications compared to non-CpG contexts. Another hyperparameter that would make sense to include is sequence uniqueness u, which will influence the primary and secondary alignment scores and our general confidence in the classification. One description of uniqueness would be the n-mer Shannon entropy for the DNA sequence spanning the read length, with uniqueness being continuous and non-negative. Such a Bayesian model can assume that there is no spatial relationship between neighbouring cytosines and that all data points are *iid*. This simplification may represent the modification status of neighbouring CpGs incorrectly, for which is known that the modification levels are highly correlated, but would help with the initial iterations of the model development. Also, a hybrid model combining a Bayesian model's classification as a feature for an ML model could improve probability and increase sensitivity regardless of the limited training data.

Chapter 5

Somatic variant calling on TAPS and WGBS data

Somatic variant calling is commonly used to evaluate the mutational load and diversity in malignant tissues. The results from somatic variant calling are then interpreted in the light of cancer development and progression. The majority of current computational approaches use whole-exome sequencing (WES, that targets the coding regions of the genome) or WGS data from matched cancer and healthy control tissue with some which enable cancer-only variant discovery mode. The agreement between variant callers depends on the properties of the samples such as tumour heterogeneity, purity, mutational load (the total number of somatic variants), the stage and the type of cancer and certain characteristics of the sequencing run such as sequencing coverage and possible alignment artefacts and biases. The combination of lower coverage and fewer mutations (as a result of contamination of the tumour sample or the cancer type and stage) leads to a more considerable discrepancy between the variants discovered by dif-

ferent callers [70]. As an addition to the intrinsic difficulties in finding somatic variants in non-chemically treated sequencing (WGS, WES) tumour data, an unresolved problem with TAPS is that pyridine borane conversions of modified cytosines can mask true variants (C>T transition), especially in a CpG context in mammals (conversely in WGBS the unmodified cytosines are converted to thymines and can be mistaken for C>T substitutions). Also, other single nucleotide and larger structural variants need to be discovered and reliably filtered. At present, the working solution is to convert back thymines to cytosines (a process of *demodification*) and apply WGS somatic variant callers. As a performance control and objective truth, a set of quality passing variants from matched WGS data were used for comparison to TAPS- and WGBS-generated variant calls. As with other experiments, the known unknowns (such as sampling randomness, and TAPS and bisulfite chemistry that may result in differences between the two datasets) were ignored for simplicity.

This chapter aims to apply and compare somatic calling pipelines on TAPS and WGBS data and presents the results from somatic variant calling on WGS, TAPS, and WGBS healthy, Barrett' oesophagus and OAC patient samples.

Details about the variant calling pipeline used can be found in the following methods section.

5.1 Somatic calling comparisons on WGS data

Somatic variant calling is complicated by 1) the number of somatic versus germline calls, 2) their heterogeneity and clonality within the tumour and 3) the purity of the sample. Often, the allelic and error distributions of the matched control are used for the prediction and filtering of the somatic variants. Thus, having a high purity tumour sample and matched healthy control having simi-

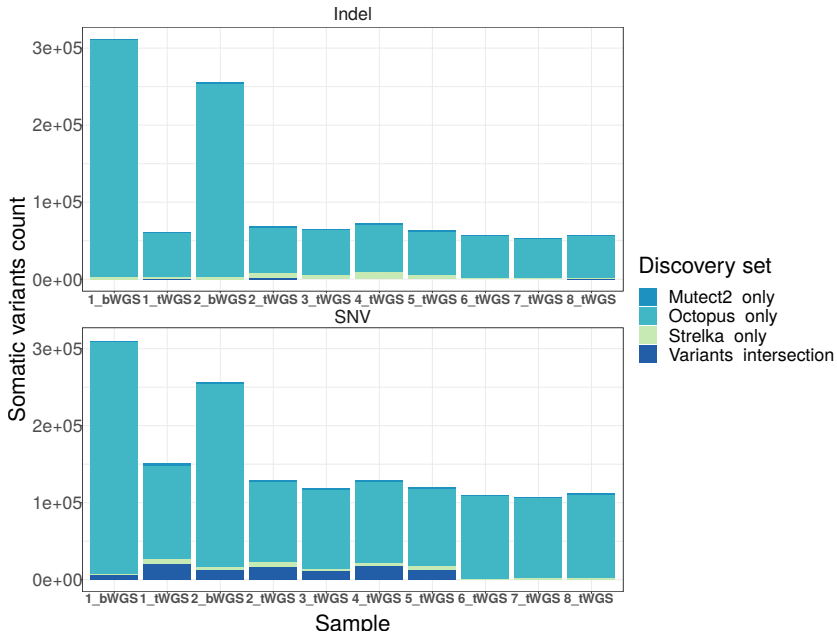
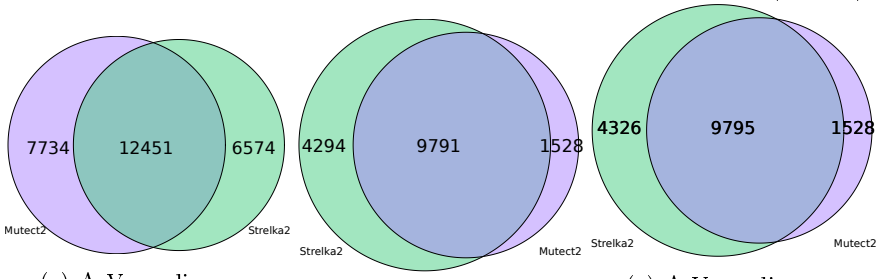


Figure 5.1: Agreement between the SNVs and indels found by the three somatic callers (Octopus, Strelka2, Mutect2) on tWGS and bWGS. In all cases, the highest number of somatic calls was made by Octopus. The highest number of somatic variants were found in the WGBS-matched WGS samples (bWGS).



(a) A Venn diagram showing the agreement between the high quality variants discovered by Strelka2 and Mutect2 if only GATK base recalibration is used.

(b) A Venn diagram showing the agreement between the high quality variants discovered by Strelka2 and Mutect2 if only GATK base recalibration is used.

(c) A Venn diagram showing the agreement between the high quality variants discovered by Strelka2 and Mutect2 if GATK base recalibration and indel realignment are used

Figure 5.2: GATK base recalibration decreased the number of somatic variants found by Mutect2 (purple) and Strelka2 (green). There is a small difference in the number of quality variants found in the base recalibrated WGS data and the base recalibrated and indel dealigned data.

lar sequencing and alignment profile as the tumour, can improve the calling of variants.

The following analysis required matched healthy-tumour sample pair somatic variant calling using Mutect2, Strelka2, and Octopus. The intersection between variants discovered by the three callers was expected to have higher confidence than each of the sets separately. It can be used for filtering and preparing data for training of TAPS- and WGBS-specific models. Furthermore, the proportion of the overlap compared to the whole variant call set can be used to evaluate the expected lower bound of disagreement between WGS callers for TAPS- and WGBS-specific tools.

The results from the three somatic variant callers showed that above certain cellularity and coverage, the single nucleotide variant (SNV) calls of different algorithms have a high overlap (88-91%) except for samples 6, 7, 8 (see Figures 5.1, 5.5) The outlier results of these samples appeared to be a result of lower cellularity or contamination of the healthy sample (ploidy 1.9 for all of the three samples). Indel matching even for the high-quality samples had an intersection between callers in the range 30-50%. Furthermore, Octopus discovered thousands of high-quality variants that were not matched by Strelka2 and Mutect2, which could be clonal or germline variants.

A comparison between the results obtained with data pre-processing with 1) no base recalibration after read deduplication, 2) with base recalibration and 3) with base recalibration and indel realignment followed by somatic calling was done on Barrett's oesophagus sample (patient 0) used previously for method development, to assess the extent of GATK pre-processing on variant discovery and the quality of variants. After the data were prepared for somatic variant calling, variants were called on Barrett's tissue sample using the duodenum as a control, and filtered the variants with the default quality passing filters. There

was a high agreement between the somatic SNVs found by Mutect2 and Strelka2 that varied with the base recalibration strategy that was used (see Figure 5.2). The WGS base recalibration was the main reason for the decrease of identified high-quality variants in Strelka2 and Mutect2 compared to the non-GATK pre-processing, where base recalibration diminished the difference between Strelka2 and Mutect with 20%, which was largely a result of decreasing the number of Mutect2 variants. Base recalibration decreased the number of variants found by both callers by 24% for Strelka2 calls and 44% for Mutect2 calls. However, as stated by GATK indel realignment was not required in version 4, and no significant difference between the base recalibration only and base recalibration and indel realignment was seen (see Figure 20), thus the base recalibrated results were used, and base recalibration only was decided for as a general approach for data preparation before variant calling.

5.2 Base error estimation comparisons on WGS data

The development of new TAPS- or WGBS-specific pipelines for variant discovery required a set of comparison metrics for performance evaluation. As a proxy for random sampling, sample preparation, and alignment biases the overlap between the somatic calls discovered in the same sample after downsampling was applied; this metric was named *chance of discovery*. The observation that a common variant within the tumour population could be detected at lower sequencing coverage than a rare variant was used to define a second performance metric. The metric *variant agreement* gave the overlap of variants discovered by different somatic callers. Again, highly represented variants spanned by reads with certain sequencing quality would be identified by more tools, compared

to rare variants and variants not passing certain quality filters based on depth, alignment metrics, and others. Ideally, any TAPS-specific (or WGBS-specific) approach would fall within the variant discovery, and agreement ranges observed for WGS variant calling tools.

Information about the downsampling experiments can be found in this methods section.

The following comparisons using WGS samples were expected to give more information on the disagreement rate of variant discovery between somatic callers. Different combinations of analyses over high and low depth samples resulting from downsampling with or without repetition of reads of the same high depth sample measured:

1. chance of discovery lower bound (without repetition)
2. chance of discovery upper bound (with repetition)
3. variant agreement lower bound (without repetition)
4. variant agreement upper bound (with repetition)

Additionally, a measure of randomness and biological differences was obtained through comparison between somatic calls where more than one control tissue was available. After downsampling the original WGS samples allowing repetition of reads to depths of 5x, 10x, 15x, 20x, 25x, 30x, 40x, and in some cases 50x, matched pair healthy-tumour somatic calling with Strelka2 and Mutect2 was performed. The number of quality passing somatic calls in Mutect2 was positively associated to the sequencing depth (see Figure 5.3), whereas Strelka2 calls were elevated in numbers at lower depths ($\leq$10x). A possible explanation of the phenomenon was that the elevated calls were an artefact of the model based on the expected depth which by default is 30x where certain SNPs in the healthy controls were identified as somatic variants (see Figure 5.4). Furthermore, the number of false positives in Strelka2 decreased with sequencing coverage (disagreement rates at 5x 51-98%, while at 40x or 50x, the range is 2-

32%). The variants found by Mutect2 appeared to be more stringently filtered, as the disagreement rates for 5x coverage in samples 1, 2, 3, 4, 5 were 3-13% (the disagreement for the remaining samples was several times higher as a result of the very low number of variants discovered at any depth). In both Mutect2 and Strelka2 call sets, lower general count of samples 6, 7, 8 implied sample impurity.

The next step of the analysis included downsampling without repeating sequencing reads, where the downsampling factors 0-4X were used (0X=total coverage, 1X $\approx$ 25x, 2X $\approx$ 12.5x, 3X $\approx$ 6.25x, 5X $\approx$ 3.125x). The original file was split in a binary manner keeping separate reads by read names but allowing read pairs to remain in the same file. In this case both Mutect2 and Strelka2 made errors, discovering more variants at lower sequencing coverage (see Figures 5.7 and 5.8). Strelka2 showed the lowest number of variant calls at 2X downsampling for all samples except for patients 6, 7, 8. Furthermore, Strelka2 had made more error calls compared to the case where repeating reads were allowed (see Figure 5.9). The chance of discovery of a variant for samples of higher cellularity was above 50% at depths $\approx$ 25x of the tumour sample, but at around 25% with depths $\approx$ 12.5x, with an upper bound of $\approx$ 95% Mutect2 (Strelka2 $\approx$ 91%). Samples 6, 7, 8 had an upper bound for variant discovery at $\approx$ 50% Mutect2 (30-40% Strelka2) and a lower bound of $\approx$ 5% due to their cellularity.

A more precise further sampling from the downsampled files without repeating reads generated files with depths of coverage 20x, 15x, 10x, and 5x (the values presented will be excluding TAPS matched WGS samples 6,7 and 8). By making direct comparisons between the calls discovered at each depth and their overlap with the theoretically present variants (overlap of calls between downsampling with and without repeats), an assessment of performance at low depths was made. At coverage 20x, the upper and lower bound of discovery were

86-95% for Mutect2 and 67-91% for Strelka2; for 5-15x Mutect2 had a range of 83-95%, while Strelka2 performance rapidly worsened with lower sequencing depths (15x 55-87%, 10x 36-76%, and 5x 8-48%).

The variant agreement lower bound defined as the overlap between samples resulting from downsampling without repeats suggested that at least 25x coverage was needed to get an overlap of at least 50% (see Figure 5.9). All of the downsampling experiments were conducted without downsampling of the reference, which remained at $\approx$ 30x. In some of the shallow sequencing depths, it was possible that the inflated false-positive variants are actual SNPs found in the healthy sample. However, these results suggested that in order to achieve 50-95% discovery of variants, a depth of $\geq$ 25x was required for the tumour together with sufficient cellularity of the sample.

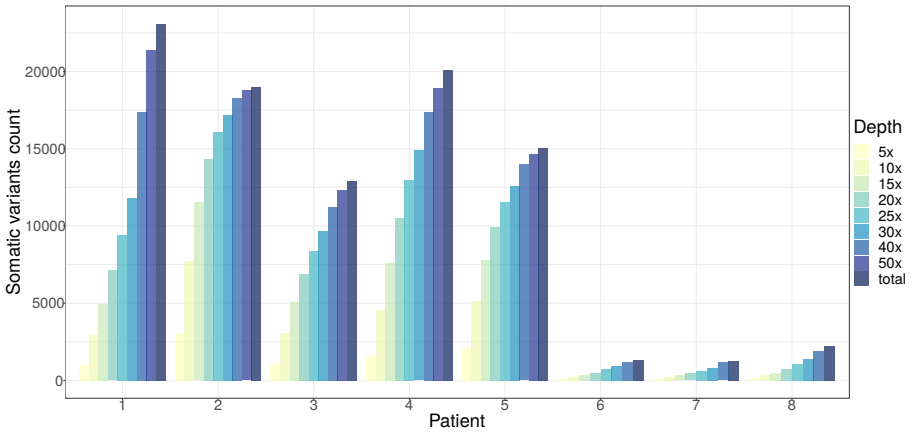


Figure 5.3: The somatic calls found by Mutect2 at certain mean genomic coverage after downsampling with repeats. There was an association between the number of discovered variants and the depth of the sample. In the WGS samples of patients 6, 7, and 8, the lower observed number of variants might be a result of lower tumour cellularity.

The final part of the base error evaluation concerned natural variation, which includes biological differences and sampling. There were three different control tissue samples in the bWGS dataset that could be used for both somatic and

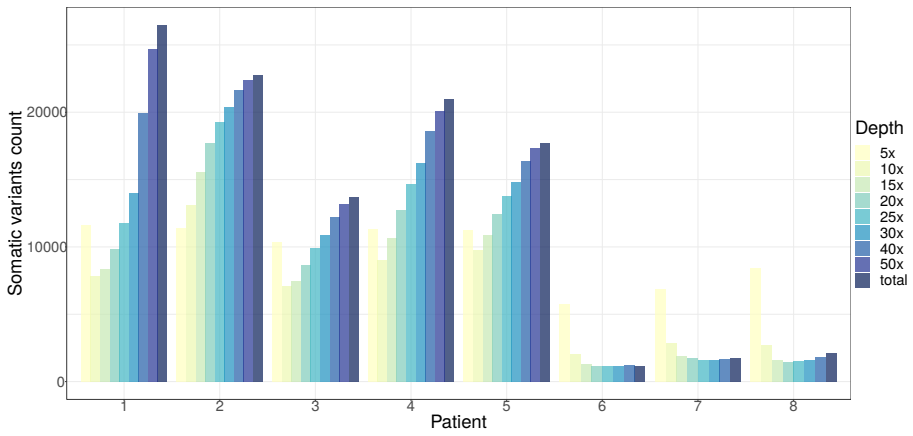


Figure 5.4: The somatic calls found by Strelka2 at certain mean genomic coverage after downsampling with repeats. There was an association between the number of discovered variants and the depth of the sample for coverages $\geq$ 10x. The increased number of variants at lower depths could be a result of detecting true SNPs on the matched healthy sample, or not fulfilling the model conditions. In the WGS samples of patients 6, 7, and 8, the lower observed number of variants might be a result of lower tumour cellularity.

germline calling and comparisons between them. By looking at the somatic variants found by Strelka2, there was a high overlap between the different experiments (see Figure 5.12). The sampling and biological differences accounted for 5-17% difference in patient1 and 4-5% for patient2, confirming the general rule that higher sequencing coverage increased the overlap between replicates.

5.3 Direct somatic calling on TAPS and WGBS data

As the first step of variant calling analysis of TAPS and WGBS data direct calling using WGS-specific tools was performed. Given the novelty of TAPS and the limited availability of germline and somatic WGBS-specific variant callers, such step is justified for comparison reasons. However, the results from WGS-specific calling on TAPS and WGBS data were expected to suffer from the

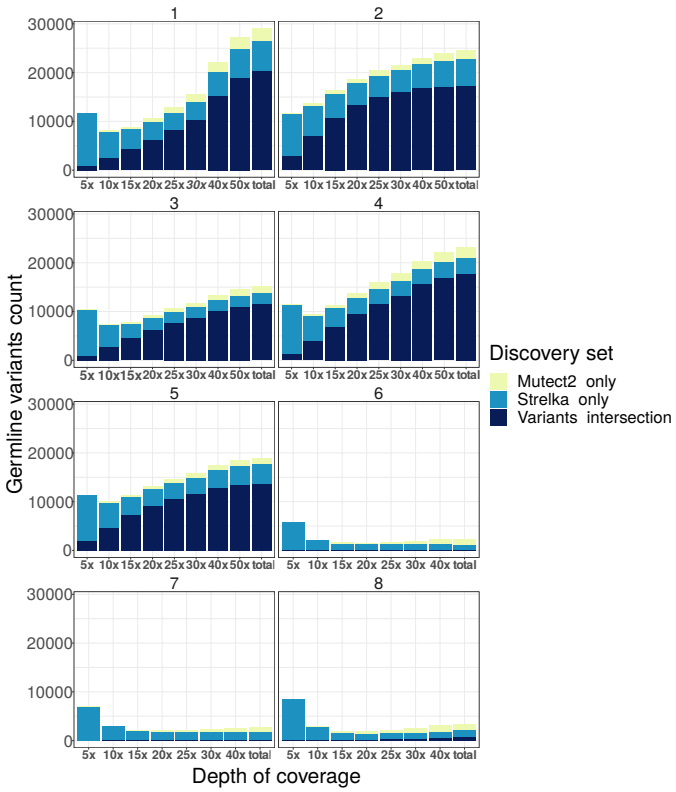


Figure 5.5: The somatic calls found by Strelka2 and Mutect2 at certain mean genomic coverage after downsampling with repeats. The increase in sequencing depth was associated with both the total number of variants discovered and their overlap. These results suggested that calling high-quality variants required sequencing depths of at least 25x for samples with enough tumour purity. In the WGS samples of patients 6, 7, and 8, the lower observed number of variants might be a result of lower tumour cellularity.

following issues:

- the elevated C>T conversion a result of the underlying chemistry of the method would be falsely labelled as variants (C>T is the most common type of mutation);
- strand bias resulting from non-complementarity of DNA strands after chemical treatment (modifications on the top strand do not match modifications on the bottom strand and after conversion, they are not comple-

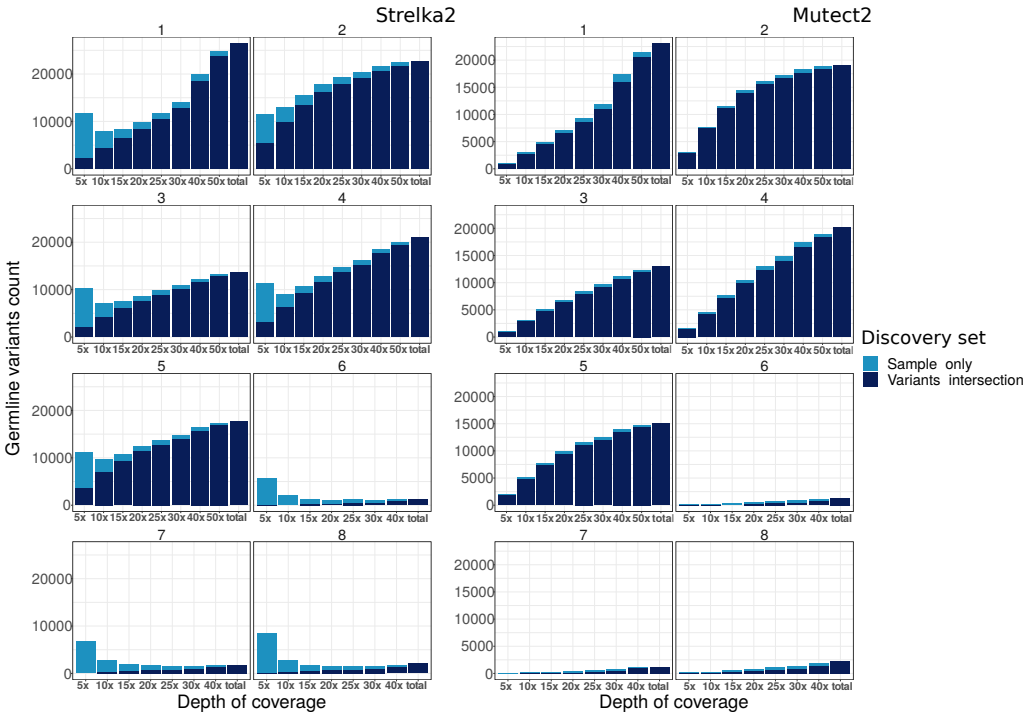


Figure 5.6: The somatic calls found by Strelka2 (left) and Mutect2 (right) at certain mean genomic coverage after downsampling with repeats. The increase in sequencing depth was associated with discovering a higher proportion of the variants in the initial (not downsampled) sample. These results suggested that calling high-quality variants required sequencing depths of at least 25x for samples with enough tumour purity. In the WGS samples of patients 6, 7, and 8, the lower observed number of variants might be a result of lower tumour cellularity.

mentary)

- any library preparation-induced errors if they were at a different rate than what was observed in WGS;
- sequencing and alignment biases such as mapping or base quality scores and Edit distances could cause high quality and biologically important reads to be ranked lower by the WGS calling algorithm;
- a higher proportion of indels observed on reads as a result of the mapping.

In order to evaluate the somatic variant calling tool error rates on TAPS and

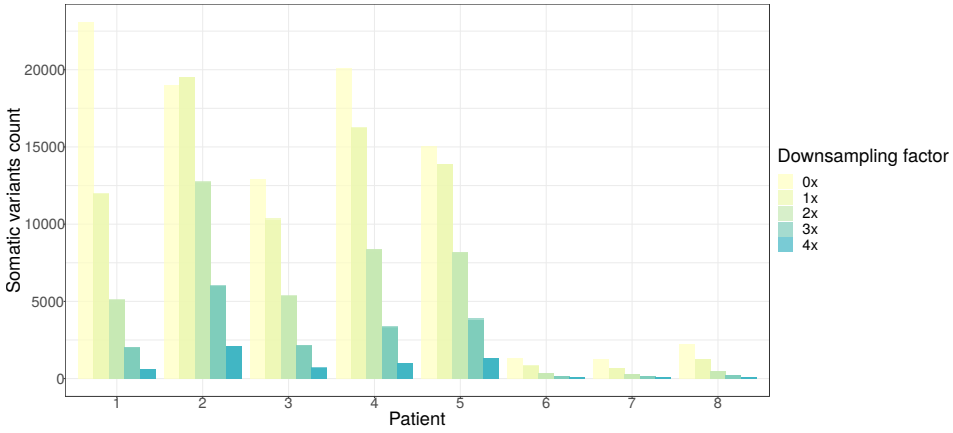


Figure 5.7: The somatic calls found by Mutect2 at certain mean genomic coverage after downsampling without repeats. There was an association between the number of discovered variants and the depth of the sample. In the WGS samples of patients 6, 7, and 8, the lower observed number of variants might be a result of lower tumour cellularity.

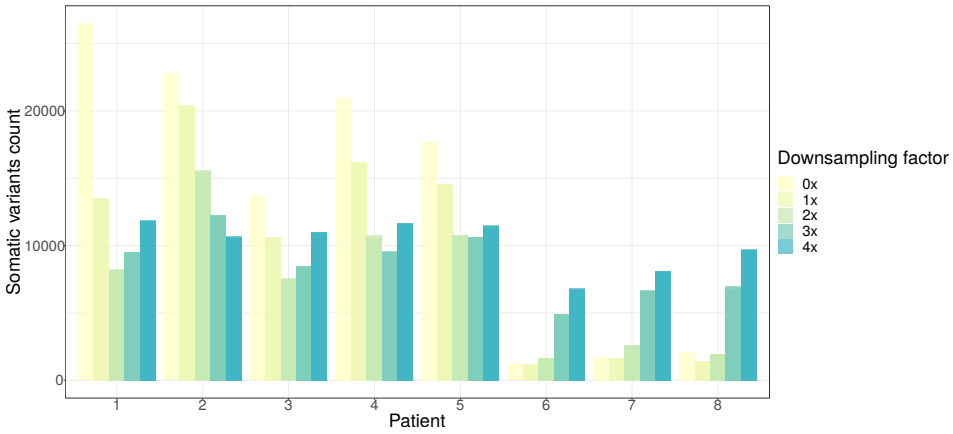


Figure 5.8: The somatic calls found by Strelka2 at certain mean genomic coverage after downsampling without repeats. There was an association between the number of discovered variants and the depth of the sample for coverages $\geq$ 10x. The increased number of variants at lower depths could be a result of detecting true SNPs on the matched healthy sample, or not fulfilling the model conditions. In the WGS samples of patients 6, 7, and 8, the lower observed number of variants might be a result of lower tumour cellularity.

WGBS, either a TAPS- or WGBS-modified healthy tissue or its WGS equivalent was used as a control sample. In the following section, only Strelka2 results which have passed the default filtering settings are discussed. The overlap with WGS

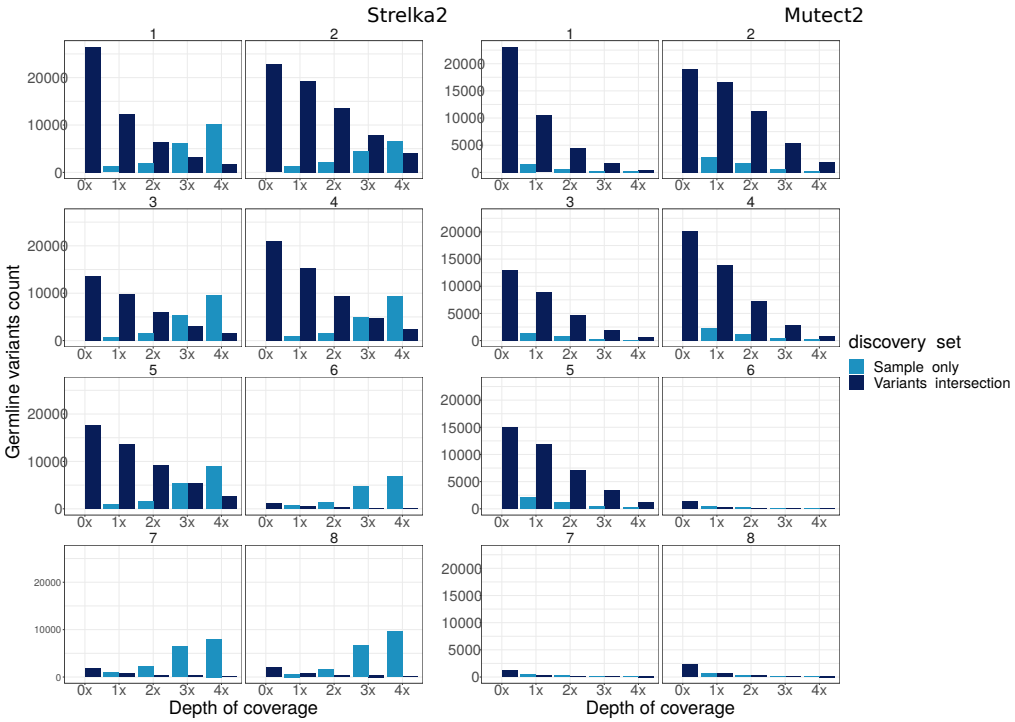


Figure 5.9: The somatic calls found by Strelka2 (left) and Mutect2 (right) at certain mean genomic coverage after downsampling without repeats and their intersection with the original depth sample. The increase in sequencing depth was associated with discovering a higher proportion of the variants in the initial (not downsampled) sample. These results suggested that calling high-quality variants required sequencing depths of at least 25x for samples with enough tumour purity. In the WGS samples of patients 6, 7, and 8, the lower observed number of variants might be a result of lower tumour cellularity.

to WGS matched healthy-tumour somatic calls was higher in the direct calling of TAPS matched to WGS ($\approx$ 84% in samples 3 and 5) compared to 72-73% in the TAPS-TAPS matched sample (see Figure 23). However, the false positives in the TAPS-WGS case were much higher where the vast majority of false positives are C>T conversions. Still, there are also alignment and chemistry-specific differences. The false-positive rate in the TAPS-TAPS case was around 60-70 times lower, mostly because of the more extensive shared methylation profiles between the control and healthy sample.

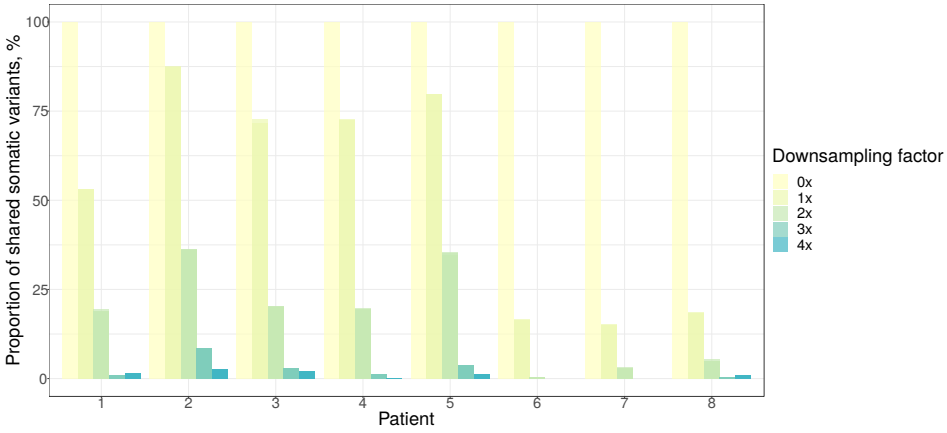


Figure 5.10: The matching somatic calls found by Mutect2 at certain mean genomic coverage after downsampling without repeats. There was an association between the number of discovered variants and the depth of the sample. In the WGS samples of patients 6, 7, and 8, the lower observed number of variants might be a result of lower tumour cellularity.

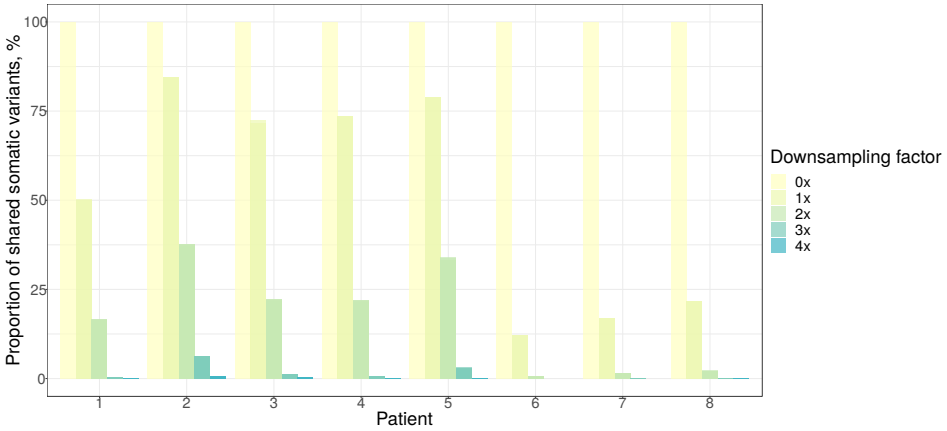
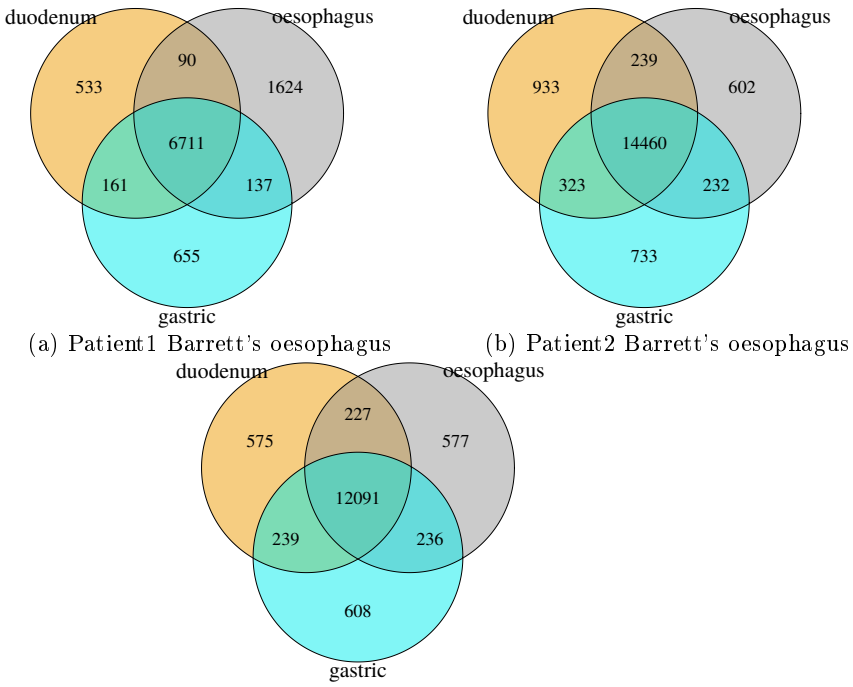


Figure 5.11: The matching somatic calls found by Strelka2 at certain mean genomic coverage after downsampling without repeats. There was an association between the number of discovered variants and the depth of the sample. In the WGS samples of patients 6, 7, and 8, the lower observed number of variants might be a result of lower tumour cellularity.

In WGBS the use of the WGBS-matched sample gave only 54000-59000 variants, however only around 3-4% of them were matching the WGS reference (in the WGS matched samples the overlap with the truth set was 7.7-9.5%, but the total number of variants was around 4.9-5x10^7). The difference between



(a) Patient1 Barrett's oesophagus (b) Patient2 Barrett's oesophagus

(c) Patient2 cancer

Figure 5.12: Using different healthy controls for variant calling leads to small differences in highly covered samples (Strelka2 calls). For the natural variation experiment, the bWGS samples were used, where three different healthy matched tissues were available—duodenum, oesophagus, and stomach.

calling in WGBS-WGBS versus the WGBS-WGS was 800-900 times smaller, while the difference between the real sets and the direct variant call sets were in the range 4-6600 times. Strelka2 uses the healthy sample distribution to build its mixture model of the somatic sample variant expectation, and bias profiles resulting from misalignment and library preparations would be more similar between TAPS or conversely WGBS samples resulting in lower false-positive rate.

The contribution of structural variation to the differences between WGS and the epigenomic datasets was evaluated by comparing Strelka2-only calls and Strelka2 calls using structural information from Manta. There were several

tens of variants differing between the two experimental conditions, thus showing that structural variation did not have a significant contribution to the difference between the WGS and TAPS, WGS and WGBS. For this reason, Strelka2-only calls without the additional information of Manta's output were used in the following analyses.

The direct somatic variant calling approach was also tested on the TAPS Barrett's sample 0 matched with the WGS duodenum. As expected, millions of modified cytosines were called as high quality variants. In contrast, true variants (the ones discovered in the WGS case) did not pass the quality filters set by the tools, especially in case of Strelka2 calling where the low overall depth of the sample was one of the reasons.

5.4 Deterministic demodification of TAPS and WGBS data

The known issue with the complete reversal of possible modifications was that all real C>T variants would be 'reverted' back to the human genome reference. However, even a demodifier with such limitations could provide valuable information on variant callers' error rates resulting from the specific alignment and library biases of WGBS and TAPS data. This information might be used to modify the existing default configurations of the tools when they are run with non-WGS data. The demodifier had been developed over time using real available data, and the following issues had to be resolved:

1. Sequencing depth: the TAPS sample used for initial method development (Barrett's oesophagus patient 0) had a mean coverage of approximately 20x with a higher duplication rate than what was observed in the WGS

samples and previous genomic TAPS samples, which increased the uncertainty of somatic variant calls particularly when Strelka2 was used.

2. Read flags: the specific read flags assigned to each read during alignment contain important information both on the DNA strand of origin and the pairing, and on the DNA fragmentation pattern. General-purpose aligners such as BWA have certain fragment size distribution expectations, which if unmet result in labelling the reads as improperly paired. However, these reads are not necessarily incorrectly paired and with the wrong orientation but can suggest shorter or longer fragments than expected. Such cases include but are not limited to short spike-ins, cfDNA, and different kinds of structural variation. The original version of demodifier only looked at properly paired-reads that have F1R2 (flags 99 and 147) or F2R1 orientation (flags 83 and 163) in order to evaluate the modification levels, which was the method used in other tools looking at bisulfite sequencing data such as Bismark. The reads with the four properly paired flags were over 99% of all reads (see Figures 21 and 22). However, the use of only four flags might be problematic for the reasons mentioned above, thus improperly paired reads for which the strand origin was known also had to be included. The list of recognised flags by the tools was expanded and was used for both modification and demodification in order to reduce the noise of variant calls after demodification and improve the sensitivity of modification calling.

3. The need for realignment: initially the whole demodified file was realigned regardless of whether a read was modified or not. However, this step might not only increase the computational cost at the next steps but also might result in a higher divergence from the original alignment. At a later stage, a modification of some read tags directly in the BAM file was used to

decrease the false positives without re-mapping to the reference sequence.

4. Base to position mapping: the information in the CIGAR string was used to ensure better mapping of bases to their reference position, especially in the cases where the reads had indels or soft or hard clipping.

5. The MD tags (strings describing mismatches from the reference for each sequencing read giving positional information and the nucleotide that is not matching) need to be changed after demodification: if the sequencing data are directly used for variant calling without realignment then all bases converted back would contain their original MD tag, so it was crucial to implement an MD tag converter which used the correct base to position mappings.

Another critical issue was whether changing the MD tags of the reads alone accounted for any demodification taken place in the sequence of the read, or that re-alignment was necessary to ensure variant calling correctness. Finally, deciding whether base recalibration of demodified data would improve the overlap between the call and the WGS 'truth set' was also needed before any more complicated demodification strategies were applied.

After the complete demodification of C>T substitutions, in the not-remapped and not base recalibrated case using Strelka2 on TAPS suggested that roughly, around 60% of all variants that were not C>T were found in both the TAPS-TAPS and TAPS-WGS case, where again the error rate for the TAPS-TAPS case was lower presumably because of the shared modification and error profiles (see Figure 23). This trend was preserved in all tested samples and experimental settings (with and without realignment after demodification and with or without GATK base recalibration after demodification or realignment). SNVs discovered against a TAPS reference after the demodification of the case and

control in samples 3 and 5 were 35-43% (Strelka2) and 47-57% (Mutect2) of the variants discovered in tWGS, while the values were lower when the calling was against the tWGS duodenum—21-25% (Strelka2) and 14-16% (Mutect2) (see Figure 24). However, the overlap between the Mutect2 and Strelka2 calls on the demodified TAPS data were overlapping at $\approx$ 20-50% suggesting that existing differences in TAPS from alignment, library preparation, and biology, had introduced 'true' variation from WGS. Realignment by itself or in combination with base recalibration following demodification had lead to $\approx$ 4-15% increase in the discovered variants in both Mutect2 and Strelka2, which was not followed by an increase in the overlap with the WGS calls; thus these steps have introduced more noise into the system. Base recalibration directly after demodification has decreased the false positives with 4-5% in Strelka2 and increased them by the same degree in Mutect2.

Taking into account the results of base recalibration following demodification in TAPS, the error rate and the overlap with the WGS variant set were evaluated without this step in WGBS (once after direct demodification and then after demodification and realignment). Similarly to TAPS, there were $\approx$ 2.6 times fewer variants discovered when the WGBS duodenum sample was used in comparison to using the WGS healthy sample for the tumour matched calling. The largest contributor of false-positive calls were A>G and T>C substitutions ($\approx$ 70-75%) which were a direct result of the demodification. However, in the case of WGBS the demodification events had a more pervasive effect on the DNA sequence with over $3x10^6$ A>G and T>C changes per sample in the WGBS-WGS case and around $1.3x10^6$ events in the WGBS-WGBS case. In the direct calling after demodification, $\approx$ 51-52% of WGS variants were observed in Strelka2 call set when the WGBS healthy sample was used as a control and around 55-63% when the WGS sample was used (a slightly higher overlap was seen for patient 2). The

realignment did not lead in a significant difference in the number of overlapping variants between Strelka2 WGBS and WGS calls (the increases were in the range of few tens to hundreds more matching variants). However, the realignment step decreased the number of false-positive calls two times. Also, there were around twice as many variants discovered in the WGBS-WGS case compared to the WGBS-WGBS ($7x10^5$ to $3.5x10^5$ with the most considerable contribution of A>G and T>C variants), and around six times fewer variants in after realignment. This result, however, could be attributed mainly to the effect of alignment where a proportion of the fully demodified WGBS sequencing reads were not realigned in the same place. In terms of overlap, the realignment increased the intersection between the WGS and the demodified WGBS set with $\approx 15\%$ ($\approx 30\%$ to $\approx 45\%$), with the WGBS-WGBS case having only 2% fewer overlapping variants. Overall, Mutect2 had fewer demodification driven somatic calls compared to Strelka2 in WGBS, while in TAPS, the two methods had a comparable performance with fewer errors seen in Strelka2 sets. Furthermore, while realignment increased the false-positive rate in TAPS, it had the opposite effect in WGBS, which can be explained primarily by the similarity of reads before and after demodification. A more significant difference in realignment positions was expected in WGBS (in the next chapter, this suggestion was tested).

5.5 Probabilistic demodification of TAPS and WGBS data

Demodification of TAPS and WGBS sequencing data where every difference from the reference at cytosine positions was considered a result of conversion did not take into account the natural genetic variation existing in the samples. A more sophisticated approach that was able to distinguish modifications from

single-base genetic variants was expected to decrease both the false-positive rate during modification calling and the false-negative rate during demodification, which would lead to a higher overlap with the WGS somatic calls.

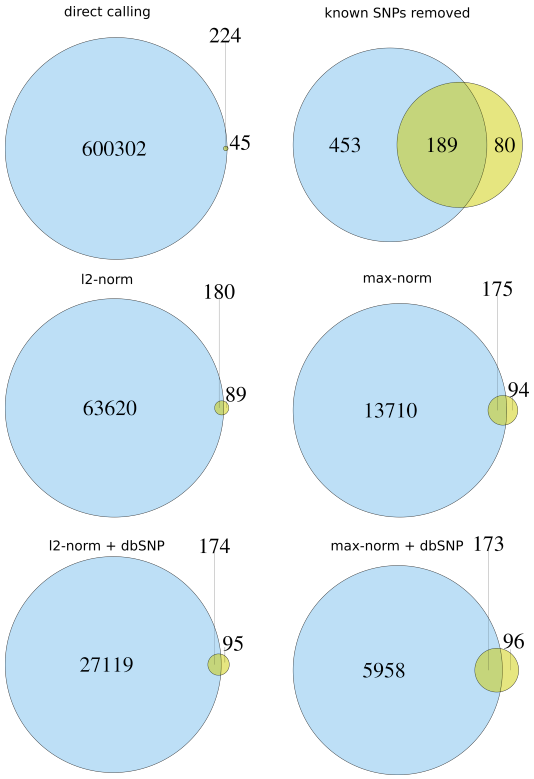


Figure 5.13: Comparison of Strelka2 somatic calls on TAPS (blue) OAC using as a control WGS duodenum sample against WGS somatic calls (yellow). The default parameters of the tools were used for calling. The figure also shows the dependence of the data normalisation method and performance (fewer errors were made when the data were max-norm normalised).

In order to test the demodification algorithm presented in chapter 4, chromosome 21 from TAPS patient 3 and WGBS patient 1 were demodified using the NN model with the selected through RF importance ranking features. This model was used because of its performance and decreased parameter space compared to the full model. As part of the model evaluation in TAPS, different normalisation methods (L2-norm versus max-norm) were tested. The first ex-

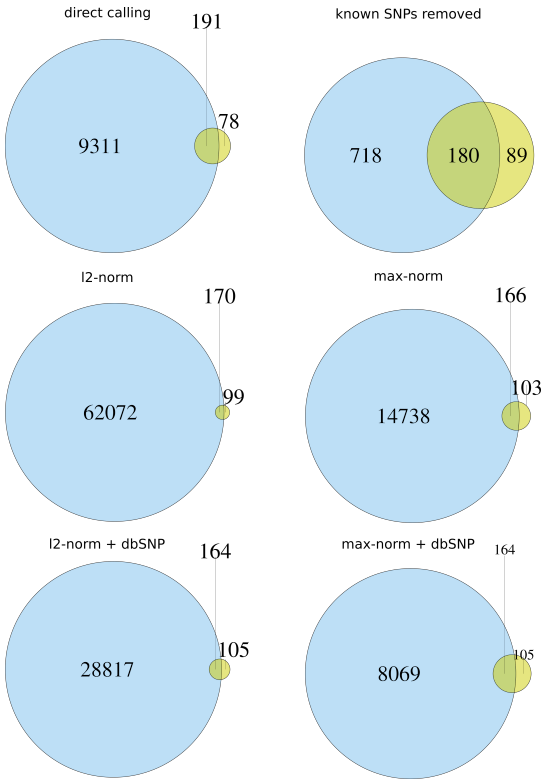


Figure 5.14: Comparison of Strelka2 somatic calls on TAPS (blue) OAC using as a control TAPS duodenum sample against WGS somatic calls (yellow). The figure also shows the dependence of the data normalisation method and performance (fewer errors were made when the data were max-norm normalised).

periment used TAPS data that was not base recalibrated after demodification, and a WGS duodenum control for somatic calling. In order to get a more realistic reference frame for the performance of the model, given the training constraints (a single sample for feature extraction and selection, a high class imbalance with the potential overlap between samples from different classes, particularly modifications and heterozygous SNPs, and sample purity which was not taken into account) the lower and the upper bounds of the demodification were established. The direct TAPS calling was used to assess the worst performance for somatic calling with WGS variant callers. In contrast, the best performance was ex-

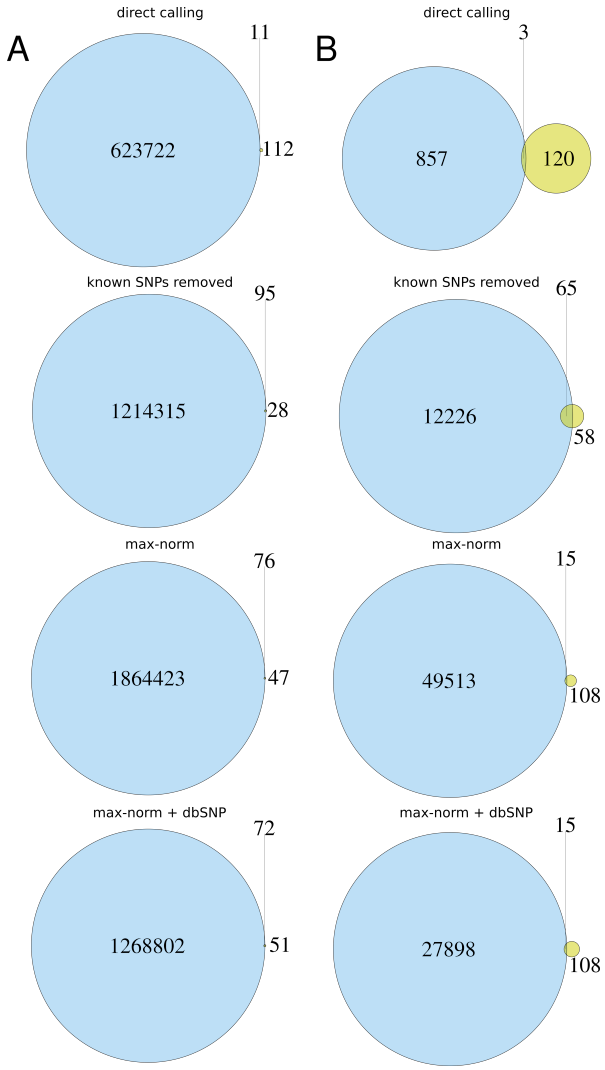
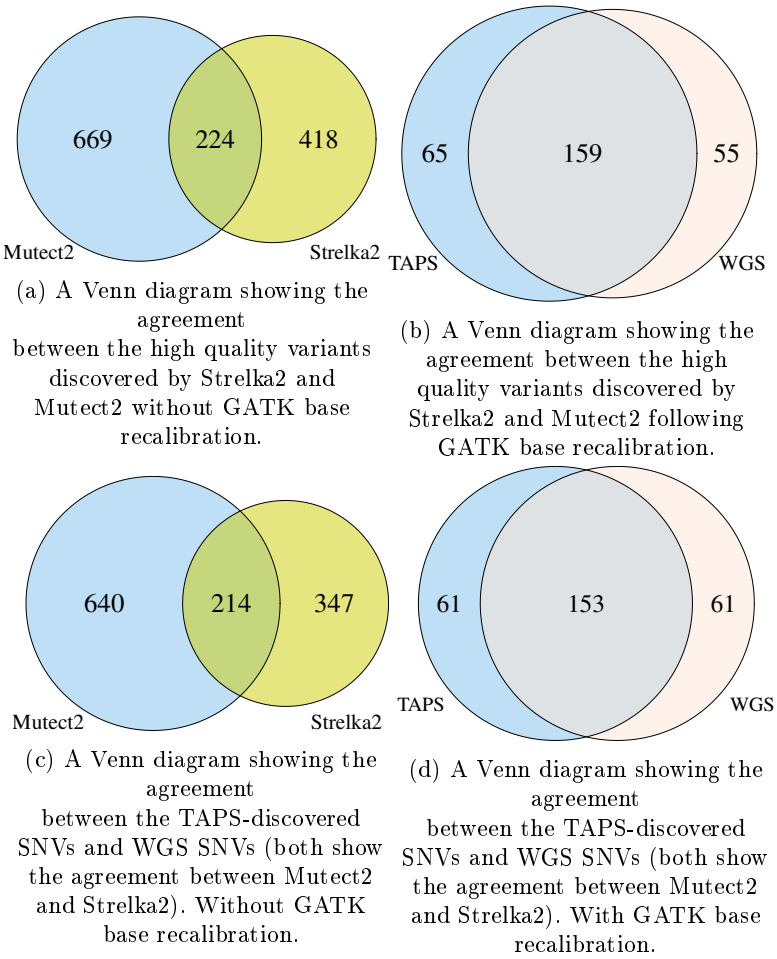


Figure 5.15: Comparison of Strelka2 somatic calls on WGBS (blue) Barrett's oesophagus using as a control A) WGS duodenum sample and B) WGBS duodenum sample against WGS somatic calls (yellow). The default parameters of the tools were used for calling.

pected after a demodification where known SNPs in the WGS were not reverted to the reference. Also, the overlap of WGS Mutect2 and WGS Strelka2 calls for patient 3 showed that only 5% of the somatic variants were non-overlapping. This performance framework measured the expected false positives with the di-



(a) A Venn diagram showing the agreement between the high quality variants discovered by Strelka2 and Mutect2 without GATK base recalibration.

(b) A Venn diagram showing the agreement between the high quality variants discovered by Strelka2 and Mutect2 following GATK base recalibration.

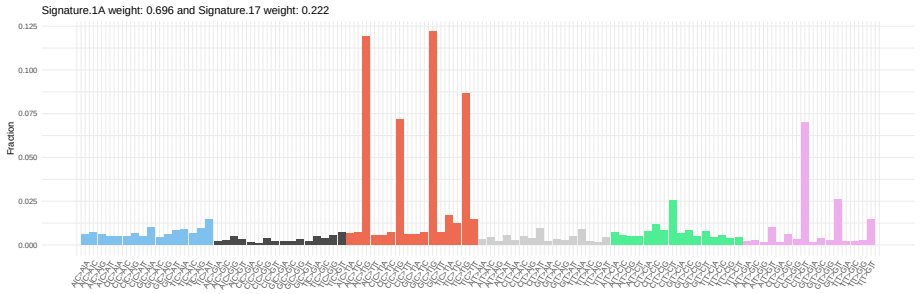
(c) A Venn diagram showing the agreement between the TAPS-discovered SNVs and WGS SNVs (both show the agreement between Mutect2 and Strelka2). Without GATK base recalibration.

(d) A Venn diagram showing the agreement between the TAPS-discovered SNVs and WGS SNVs (both show the agreement between Mutect2 and Strelka2). With GATK base recalibration.

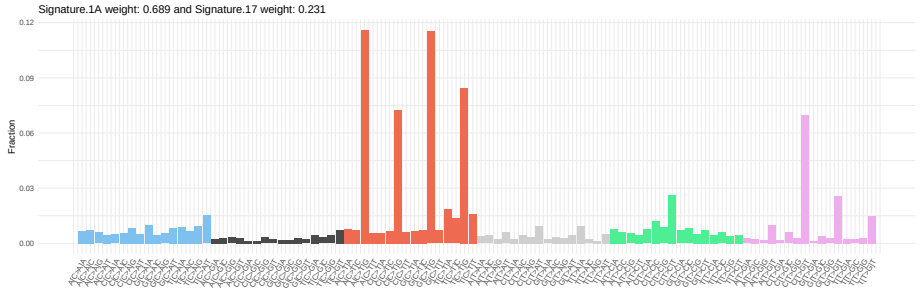
Figure 5.16: Comparison of the SNV calls between demodified TAPS at known positions and WGS. The overlap between Mutect2 and Strelka2 (left) intersects to a greater extent the known WGS somatic variants (right), regardless of base recalibration state.

.

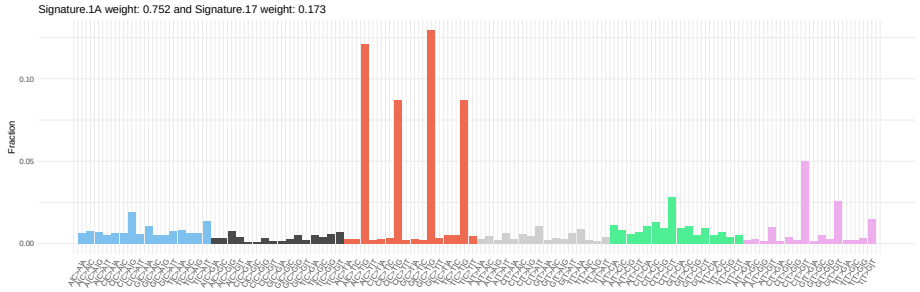
rect calling giving $>$ 2000 times more false positives than what was expected from the WGS overlap of somatic callers, while the TAPS skipping demodification on known WGS positions has $\approx$ 2 times more false positives (see Figure 5.13). The best performance of the current models has $>$ 10 times the number of false positives compared to the TAPS sample with removed known SNPs. At the



(a) Patient 3 mutational signatures from Mutect2 somatic calls.



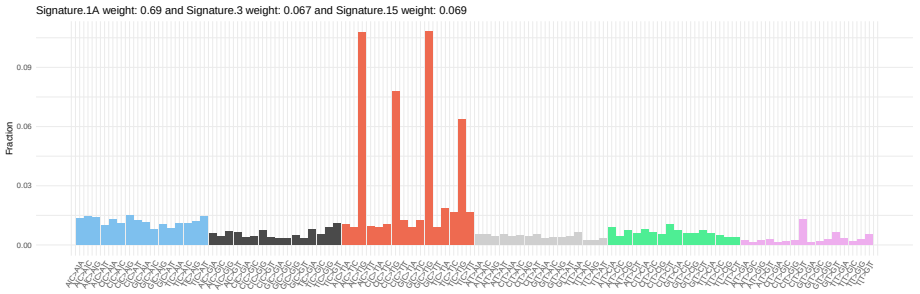
(b) Patient 3 mutational signatures from Strelka2 somatic calls



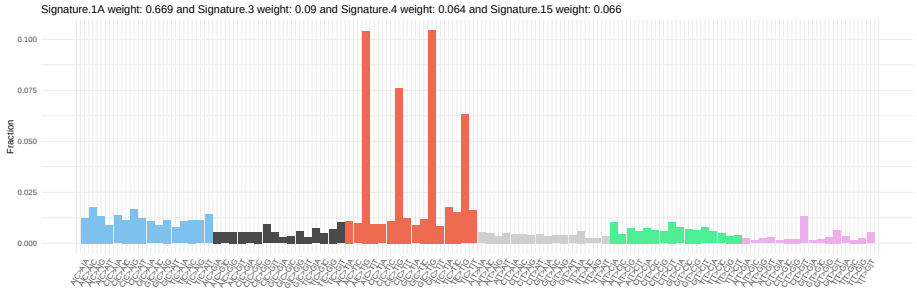
(c) Patient 3 mutational signatures on demodified TAPS using the overlap of Strelka2 and Mutect2 somatic calls

Figure 5.17: Mutational signatures created with deconstructSigs package illustrates the similarity between the patterns found in demodified TAPS compared to WGS for patient 3.

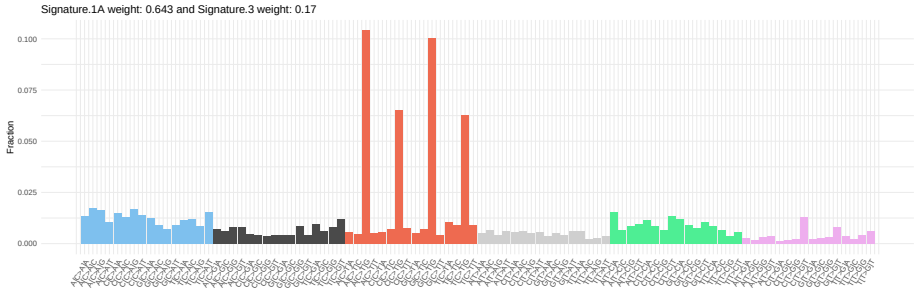
same time, the overlap with Strelka2 somatic calls did not decrease significantly. Filtering out positions labelled as heterozygous or homozygous SNPs by their presence in a dbSNP dataset reduced the false-positive calls by a half. These results suggested that known SNP information from public databases should be



(a) Patient 5 mutational signatures from Mutect2 somatic calls.



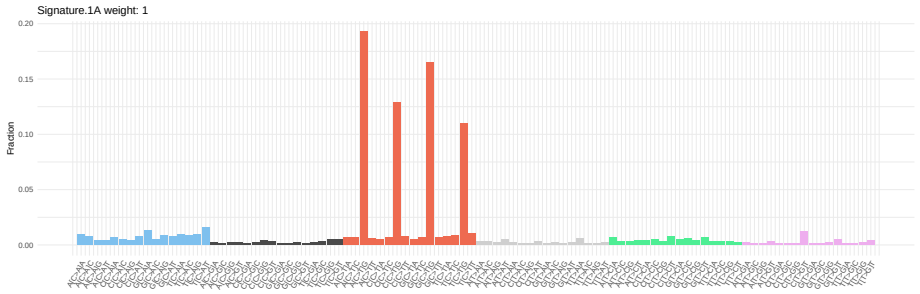
(b) Patient 5 mutational signatures from Strelka2 somatic calls



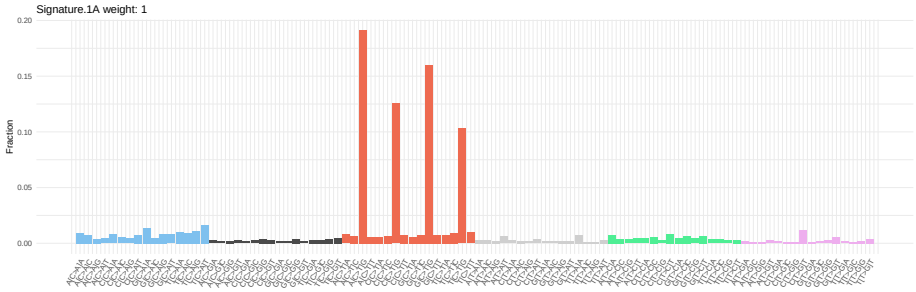
(c) Patient 5 mutational signatures on demodified TAPS using the overlap of Strelka2 and Mutect2 somatic calls

Figure 5.18: Mutational signatures created with deconstructSigs package illustrates the similarity between the patterns found in demodified TAPS compared to WGS for patient 5.

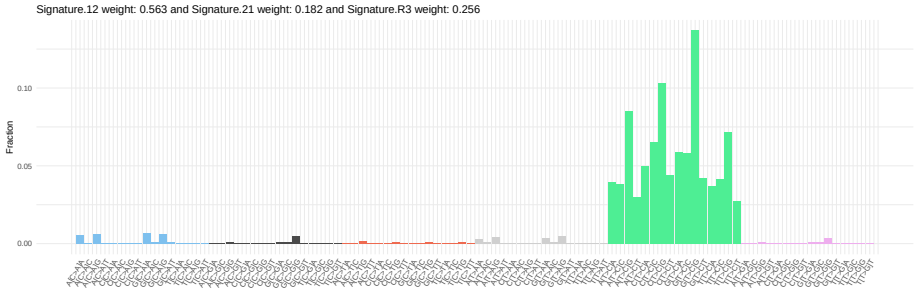
incorporated as a feature in the model for more precise results (the dbSNP information was used for the development of a germline variant caller). The results using TAPS OAC against TAPS duodenum showed similar trends with slightly higher false-positive counts and less overlap with WGS (see Figure 5.14). The



(a) Patient 1 mutational signatures from Mutect2 somatic calls.



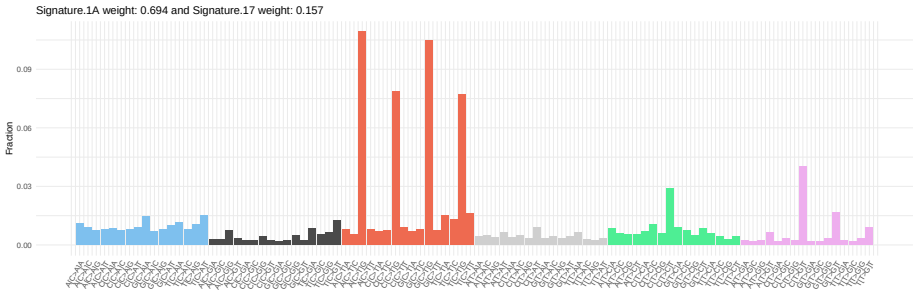
(b) Patient 1 mutational signatures from Strelka2 somatic calls



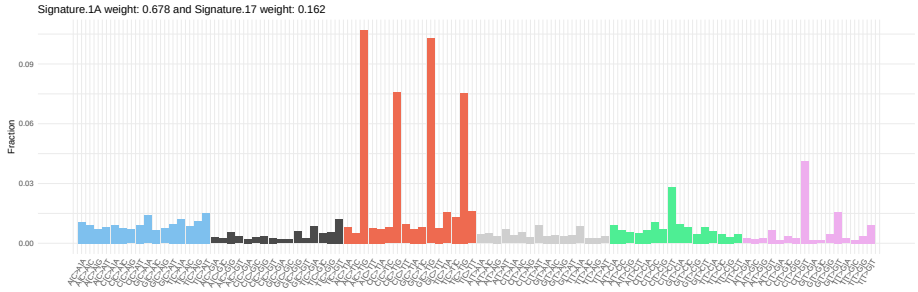
(c) Patient 1 mutational signatures on demodified WGBS using the overlap of Strelka2 and Mutect2 somatic calls

Figure 5.19: Mutational signatures created with deconstructSigs package illustrates the similarity between the patterns found in demodified WGBS compared to WGS for patient 1.

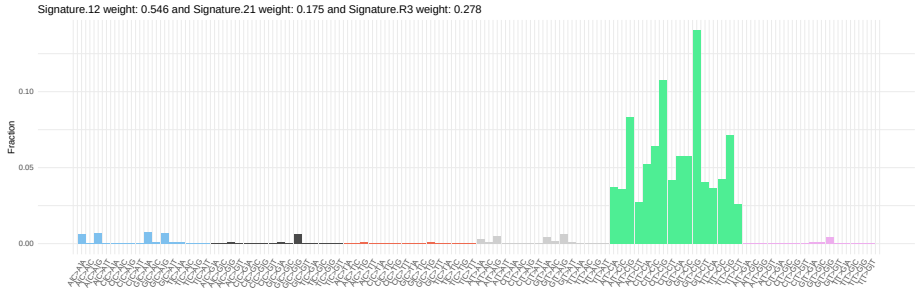
similarity between the results for TAPS/TAPS and TAPS/WGS demodification and calling may indicate that a more stringent classification would enable purely TAPS versus TAPS analyses. Mutect2 calls using as a control WGS duodenum file had between 7% and 20% fewer calls without a considerable difference in



(a) Patient 2 mutational signatures from Mutect2 somatic calls.



(b) Patient 2 mutational signatures from Strelka2 somatic calls



(c) Patient 2 mutational signatures on demodified WGBS using the overlap of Strelka2 and Mutect2 somatic calls

Figure 5.20: Mutational signatures created with deconstructSigs package illustrates the similarity between the patterns found in demodified WGBS compared to WGS for patient 2.

.

the overlap with the exception of the TAPS file which was demodified using the known WGS positions. There was a 39% increase in false positives. The false positives in both Mutect2 and Strelka calls on the demodified positions

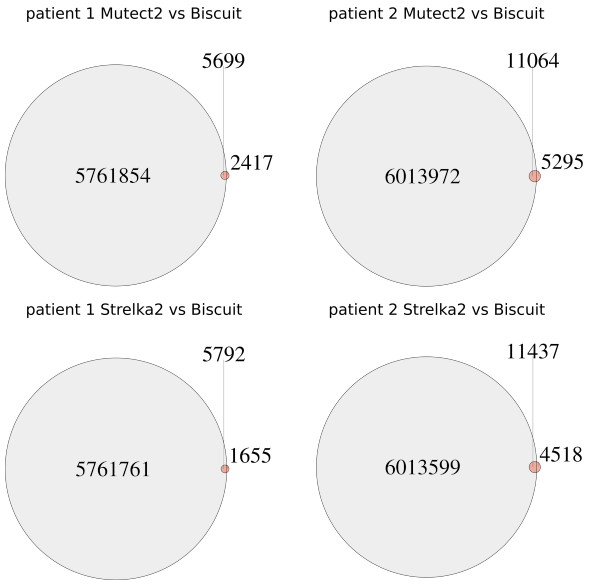


Figure 5.21: Comparison of Biscuit somatic calls on WGBS (grey) Barrett's oesophagus using as a control WGBS duodenum sample against WGS somatic calls (red). The default parameters of the tools were used for calling.

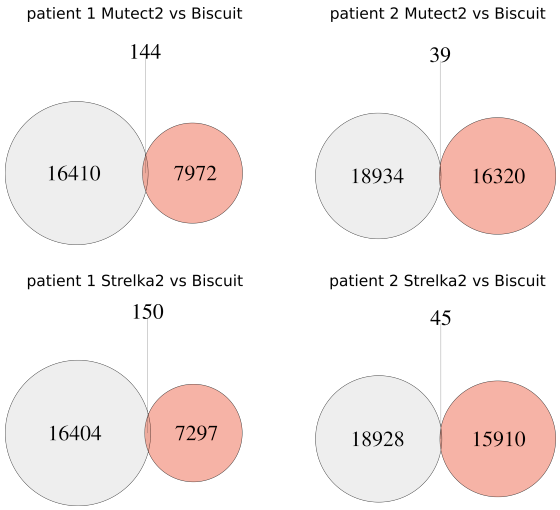


Figure 5.22: Comparison of Biscuit somatic calls on WGBS (grey) Barrett's oesophagus using as a control WGBS duodenum sample against WGS somatic calls (red). The parameters of the tools were used for calling assumed lower ploidy with an admixture from the normal sample of 0.3.

with known SNPs TAPS samples were: Mutect2 {G>A, T>C, T>G, A>G}, Strelka2 {A>C, A>G, T>C, T>G, C>T}. Base recalibration of the demodified files lead to a small decrease in the number of false positives (≈0.5-3%) for all cases except for the known SNPs demodification where the reduction was 13%, combined with a 32% decrease in the overlap positions. In TAPS, overlapping Mutect2 and Strelka2 calls reduced the number of false-positive variants by 21% for the model-only approach, and by 17% for the classification followed by dbSNP filtering.

Regardless of the lower observed intersection between the variants discovered in TAPS where known SNPs were removed, and WGS calls, there was a significant overlap between demodified TAPS Mutect2 and Strelka2 calls (see Figure 5.16). Approximately 70% of the variants in the overlap between Mutect2 and Strelka2 on demodified at known positions TAPS data intersected true WGS somatic variants. The conservative approach of using only the overlap between Mutect2 and Strelka2 variant calls was applied. As an approach to compare the biological similarity of OAC in TAPS after demodification and in WGS (the samples were produced from the same starting DNA), a visualisation of the discovered SNVs as mutational signatures was drawn. For TAPS patients 3 and 5 there was a high agreement in the mutational patterns discovered in the demodified sample and WGS (see Figures 5.17, 5.18). Further improvement of signature calling would be to use TAPS-specific sequence weights. These results suggested that somatic variant callers can be applied to demodified TAPS data. Still, at present, a more conservative approach using only the overlap of several callers significantly decreases the noise seen as elevated false-positive calls. The intersection of somatic variants contains relevant biological information such as preserved signatures of mutational processes.

Given the better performance of the model using max-norm normalised data

in TAPS, only four cases were considered: 1) direct somatic calling, 2) known SNPs removed, 3) max-norm NN model with selected features, and 4) max-norm NN model with selected features and removal of variants not included in dbSNP. Overall, the demodification approach appeared to perform worse in WGBS due to the greater degree of difference from the reference. There was a high level of false positives in the demodified WGBS data regardless of the algorithm (see Figure 5.15). However, intersecting Mutect2 and Strelka2 variant calls after the model-only and model followed by dbSNP filtering, removed respectively 99% and 96% of the false positives. Furthermore, by using a similar approach to TAPS and intersecting the calls from of Mutect2 and Strelka2 over the WGBS demodified data where known SNPs were kept, an inflated number of T>C and G>A substitution was observed (see Figures 5.19 and 5.20). Also, the ratio of A>G was twice as high in number as T>C. Around $1.2\text{x}10^6$ positions perceived as somatic variants indicated that the error profile of WGBS differed from what from the patterns observed in WGS. The lack of A>G and T>C variant elevation in the overlap of Mutect2 and Strelka2 when the control was WGBS provided further evidence for this hypothesis. A possible explanation for the high number of A>G observed was the deamination of adenine to hypoxanthine. Hypoxanthine would be read as guanine during the PCR step; the same type of error has been observed in formalin-fixed, paraffin-embedded (FFPE) tissues [137].

Also, calling with other somatic variant callers and tuning their parameters could improve the false-positive rate and understanding WGS biases towards TAPS data. Other options to improve the performance included additional data pre-processing, a larger training dataset, preparing synthetic data with certain qualities for testing, and training the filtering of the WGS somatic calling. For testing purposes, a bisulfite-specific somatic calling with Biscuit using its de-

fault parameters or changing the expected tumour purity to account for 0.3 normal sample admixture was applied on the WGBS data. The somatic calls found with the default parameters had high overlap with the WGS calls with 77.8% (Strelka2) and 70.2% (Mutect2) for patient 1, and 71.7% (Strelka2) and 67.6% (Mutect2) for patient 2 (see Figure 5.21). However, there was a high level of false-positive calls in A>G, T>C, T>G (>1x10^6 each) where the first two substitutions could be indicative differences in the modification level in Barrett's oesophagus versus the duodenum. The most common differences between WGS and WGBS somatic calls were C>T and G>A (possibly misclassified as conversions on unmodified cytosines in the tumour sample), or T>C and A>G (possibly misclassified as modifications in the tumour) that were not labelled as variants by Biscuit. Given the option to use a lower tumour purity in the model, Biscuit significantly decreased the number of false positives, but also reduced the intersection between the WGS and WGBs somatic calls (see Figure 5.22). The most common false positives in the WGBS-only set remained C>T, G>A, T>C, and A>G, but in the WGS-only additional to these classes, there were high numbers of A>C, C>A, T>G, G>T variants.

5.6 Discussion and future directions

The demodification approach, combined with filtering based on known SNPs from dbSNP, showed that WGS-specific variant callers could be applied on TAPS data after reversal of modified bases only after certain modifications are made. There was an increased background noise consisting of misclassified DNA modifications and errors generated by TAPS chemistry and alignment. The contribution of the characteristic errors of TAPS to the inflation of erroneous somatic calls was measured after the removal of known SNPs from the matched WGS reference. In this case, around two-thirds of the variant calls were not

observed in the WGS truth set. Also, combining the variant calls of two somatic callers (Mutect2 and Strelka2) resulted in decreasing the false positives while keeping the overlap with WGS. Likewise, there was a high degree of information preservation of OAC mutational patterns (as seen in the mutational signatures test) in the agreement of Mutect2 and Strelka2 sets.

The result that the conservative agreement of Mutect2 and Strelka2 callers decreased the number of false positives suggested that systematic error patterns are observed in TAPS. Thus, the model-based demodification at present would require the use of several variant callers, and an additional filtering step to test for the correctness of presumed C>T SNPs. In WGBS, similarly to TAPS, systematic errors were observed, together with misclassification of positions. The false-positive rate was higher than in TAPS which would be a result of the proportion of converted cytosines by each method. Realigning WGBS reads after demodification decreased the number of false positives and increased the overlap with WGS. Generally, the approach where a combination of somatic callers is applied to demodified data together with a stringent filtering step would decrease the number of false positives in both demodified WGBS and TAPS.

Yet, the demodification approach assumes that the conversions in DNA are the only difference between WGBS and WGS, and TAPS and WGS. So, the complete reversal of the modifications in the epigenomic data would result in data that is indistinguishable from WGS. The result of this demodification would be that WGS-specific somatic variant calling tools can be applied without limitations. However, the base qualities (mainly in WGBS), and the systematic errors resulting from the chemical treatment and alignment in both TAPS and WGBS create a degree of dissimilarity to WGS. Calling somatic mutations in TAPS (or bisulfite sequencing data) was complicated by off-target effects of chemicals and

enzymes, biased conversion, and PCR amplification. These epigenomic methods can suffer disproportionally from coverage and GC biases. Therefore, alignment in TAPS or WGBS may be affected additionally from the number of base changes in reads. Sequencing reads containing real biological information (modifications or variants) can receive a low mapping score and be misaligned or discarded by the aligner. Thus, even in the case of perfect demodification, error profile filtering would be required to remove the false positives generated by WGS somatic calling. Another limitation of the demodification process described was that the somatic variant callers Strelka2 and Mutect2 are WGS-specific and their variant calling models were built with assumptions for WGS data, some of which could not be held by TAPS and WGBS data. The direct application of these tools without pre-training on the specific input data (which may have different weights of features throwing off the somatic callers) could have led to filtering out of true variants. A greater overlap between WGS and the demodifed TAPS was observed if all Strelka2 variants were included. Many real TAPS somatic mutations were not passing the quality filters set, which may be a result of the RF expectations that are used for filtering in Strelka2.

Even after running a perfect demodifier for TAPS and WGBS data that distinguishes between variants and modifications, there would be a degree of mismatch due to unique germline and somatic variants in the sample.

In terms of the demodification model, several different future directions can be outlined. The first approach would be to improve the classification of SNPs and modifications through additional filtering, as described in the previous chapter. Another method to decrease the false-positive and false-negative rates would be to perform step-wise binary classification starting from classifying reference to non-reference (this classification would be the most common). In case the classification result suggests that the position is a non-reference, non-cytosine

positions could be either heterozygous SNP, homozygous SNP, or a sequencing error. The most challenging case would be on cytosines, where two consecutive classification steps would determine the final output. These would be homozygous SNP versus a non-homozygous position, followed if needed by separating modifications and heterozygous SNPs. Given the differences between reference and non-reference in terms of features, and the reference positions frequency, such a model could be simpler and computationally efficient. On a performance note, parallelisation and splitting of chromosomes into smaller independently processed bins could speed up classification and demodification.

The limitations of the demodification as a method for pre-processing of epigenomic data for variant calling with WGS tools, together with the small number of WGBS somatic callers justify the creation of new somatic callers. These variant callers should incorporate as part of their probabilistic model or filtering step TAPS- and WGBS-specific error profiles for the sensitive discovery of variants. The ability to call simultaneously somatic variants and modifications from DNA would improve the understanding of their dynamics of interaction in disease progression.

As a result of the demodification experiments, an error model for TAPS and WGBS can explain their dissimilarity to WGS. Other more detailed efforts have been made for describing errors in NGS data and their effects on data analysis and interpretation [239], [231], [192], [68]. The differences between the demodified data and WGS are a product of multiple levels of randomness where the error or the discrepancy (E) is a function that takes as arguments variables describing the sample preparation and analysis. The variables of the equations are mutational and modification processes (P), data sampling (D), library preparation (L), sequencing (S), alignment (A), data processing (DP), intrinsic assumptions of the calling algorithm (C), and unknown errors (ϵ). The

error model is as follows:

$$E \coloneqq f(P, D, L, S, A, DP, C, \epsilon)$$

The known unknowns P, D, L, S, A, DP, and C fall into the main categories of biological differences (P), systematic experimental bias (D, L, and S) and computational systematic bias (S, A, DP and C). The mutational and modification processes are partially explained patterns by the DNA sequence itself. Experimenter bias through data sampling may lead to different outputs given the person who has taken the probe and has pre-processed it. In the library preparation step, chemicals and enzymes may introduce novel errors to the DNA sequence [335], [160]. The composition (GC and AT content) and natural conformation of DNA sequences (*e.g.* cfDNA fragment size and abundance depending on the nucleosomal attachment, and the level of methylation and compactisation for structures like Barr bodies) may change the representation of the fragments during sequencing (*e.g.* PCR bias affecting sequences with extreme GC or AT content) [160], [317]. Next-generation sequencing influences both in the experimental and the computational bias through uneven amplification of specific sequences, and reading and base calling [388], [160], [280]. Sequencing alignment of short reads uses heuristics and can give inexact results because of both the length of the fragments and their sequence compositions. The choice of alignment and read filtering affects the overall alignment rate and the correctness of alignment as was showed in chapter 3. Alignment errors may lead to non-uniqueness, which can mask real variation. These examples of bias that can be introduced at different steps of the sample preparation and data analysis should be studied futher and the results from them taken into account for the development of TAPS and WGBS specific and sensitive somatic callers.

Chapter 6

Germline variant calling on TAPS and WGBS data

Variant calling methods are used for the discovery of somatic or germline SNVs, CNV, or large structural variations. These types of variations are abundant within malignant tissues, and their identification, annotation, and relation to clinical and treatment outcomes explain the need for variant calling bioinformatic analysis. This section presents the results of germline variant calling on tWGS, TAPS, bWGS, and WGBS patient samples which were described in the previous chapters.

Details about the variant calling pipeline used can be found in the following methods section.

6.1 Germline variant calling comparison on WGS data

Germline variant discovery (also known as haplocalling) is an important step for distinguishing tumour-specific variants from the natural individual variation observed in healthy tissues. To get a clearer understanding of the number of germline SNVs and indels, Octopus, Strelka2, and GATK HaplotypeCaller were applied to the sequencing data. Combining the results from the three SNP calls datasets enabled:

- comparisons between the calling algorithms that might result in modification of the run arguments if required;
- the creation of a list with high confidence variants as a result of an intersection between the three datasets;
- the application of the high confidence SNP list to curate a set of 'true' variants that could be used for training of TAPS or WGBS specific models for identification of modifications and variants;
- annotation of all identified variants, followed by the discovery of their relevant clinical and biological functions.

The germline variants and short indels discovered by Octopus and by Strelka2 (ranging between 4.1 and 5 million germline variants) were more than the variants discovered by GATK HaplotypeCaller (ranging between 3.1 and 3.3 million variants) (see Figures 6.1, 6.2 and 6.3). The results from GATK HaplotypeCaller were 72-82% of Strelka2 calls and 64-67% of Octopus calls. The most significant contribution to the discrepancy in the number of variants was the proportion of indels reported by each caller. The percentage of indels in Strelka2 germline

calls was 15-20%, while for Octopus the values were 20-22%, and GATK HaplotypeCaller showed only 0.13-0.16% indels (see Figure 26). The WGS samples matching WGBS (bWGS) showed a comparable number of germline variants to the tWGS samples. The observed similarity was regardless of the bWGS lower average genomic coverage (tWGS $\approx$ 48-60x and bWGS $\approx$ 33-46x) suggesting that the SNP callers used were less error-prone or sensitive to coverage differences above a certain threshold. Independent haplocalling on both the normal and the tumour samples was done to validate the high expected agreement in shared germline variants. There was a significant overlap between them (80%-99%) (see Figure 6.4). However, the tumour sample germline calling output contains a proportion of the somatic calls which were not separated from the SNPs without providing a control healthy tissue sample. Any somatic calls need to be removed from the list before variants in the tumour samples were labelled and were used for model training.

Germline variant calling using GATK HaplotypeCaller on a single Barrett's oesophagus sample (patient 0) was done using two data pre-processing pipelines. The differences were in whether indel realignment was applied after base recalibration. Approximately 3.2 million variants were found in Barrett's oesophagus and the duodenum samples of the tested patient with around 3.1 million of them agreeing between the two sampling sites (see Figure 27). These results were used as a confirmation that additional indel realignment is not required for GATK v4 before SNP calling with GATK HaplotypeCaller.

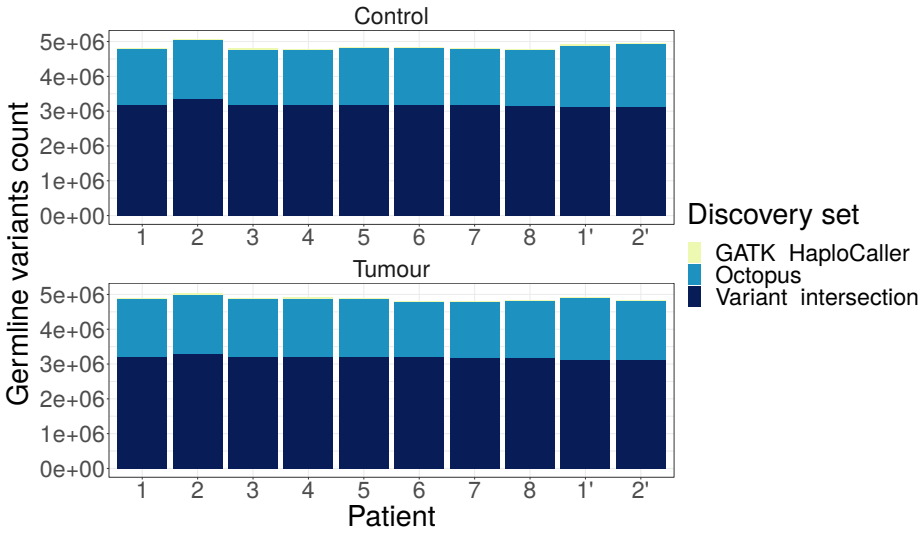


Figure 6.1: The germline variants found by GATK and Octopus in each of the patients. Calls on the healthy control tissues (duodenum, buffy coat, oesophagus) and Barrett's oesophagus samples are presented. Patients 1' and 2' are WGS samples matching WGBS while the remaining samples are WGS matching TAPS.

6.2 Deterministic demodification of TAPS and WGBS data

After the deterministic demodification step, the sequencing reads were realigned and base recalibrated. GATK HaplotypeCaller and Strelka2 germline callers were used to discover germline variants, and they showed high agreement. Around 65% of the SNPs discovered in TAPS after demodification were matching the WGS reference (the agreement between the three germline callers). In WGBS, approximately 61% of the discovered variants after demodification were confirmed in WGS. In TAPS, as expected, the majority of the variants unique to WGS consisted primarily of C>T variants ($\approx$ 2% were other types). Still, there was a different distribution of germline SNVs distinct to TAPS (see Figure 6.5

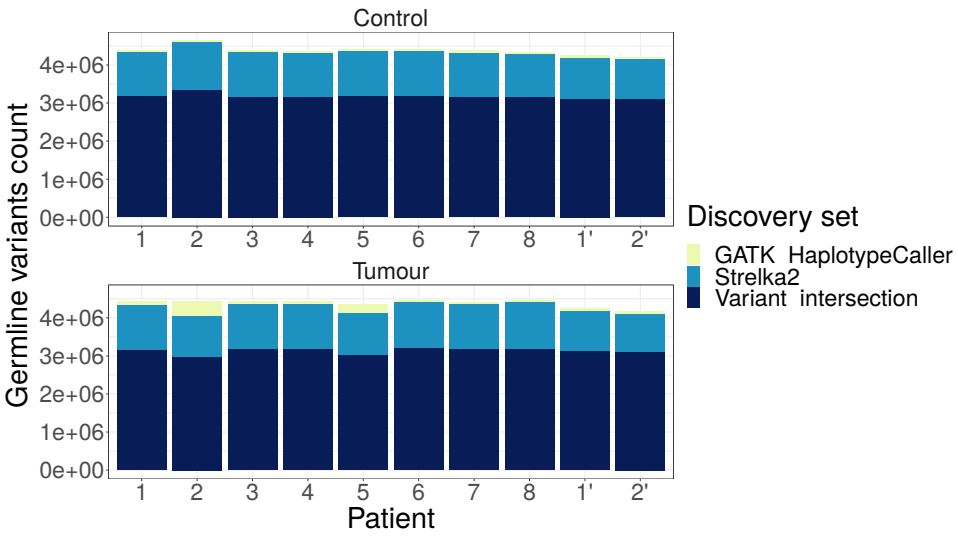


Figure 6.2: The germline variants found by GATK and Strelka2 as observed in the healthy and tumour samples in each of the patients. Patients 1' and 2' are WGS samples matching WGBS while the remaining samples are WGS matching TAPS.

left) with the highest proportion of T>C and G>A substitutions which were expected given the complete nature of the demodification). The new C>T and G>A SNPs should be a result of the alignment, but generally, it appeared that there was a pattern within the 'noise' of unique to TAPS germline calls. The pattern seemed to be group-specific rather than individual and could be a result of TAPS chemistry and alignment, as suggested in the previous chapter. In the WGBS data compared to WGS, there was a different group-specific pattern observed. Again, in the WGS unique set, the most considerable difference was visible in the number of C>T and G>A changes due to the demodification of WGBS. However, the T>C and A>G differences were much more pronounced than in the WGS matching TAPS data. These two patterns of change might indicate the incorrect location of the reads after demodification and realignment. The mismatch due to realignment would be expected to a much greater extent in

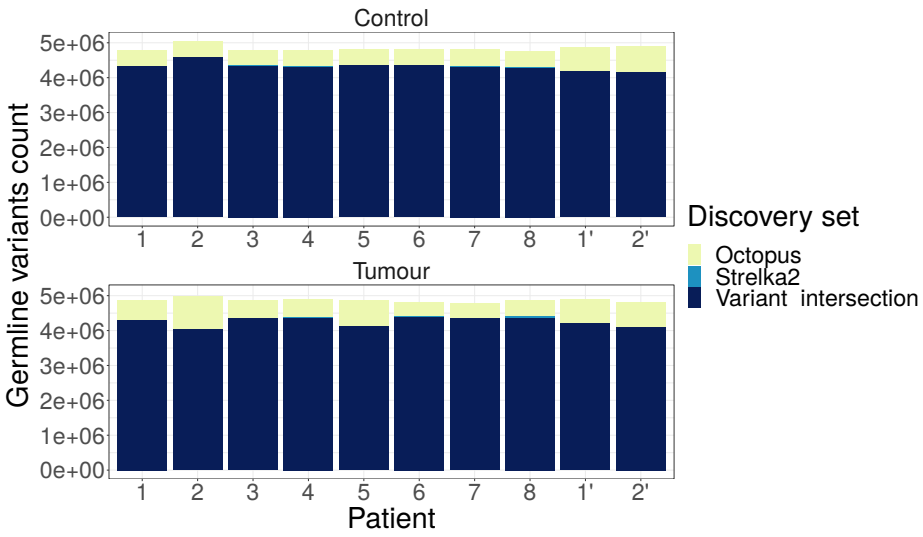


Figure 6.3: The germline variants found by Strelka2 and Octopus in each of the patients. Calls on the healthy control tissues (duodenum, buffy coat, oesophagus) and the Barrett's samples are presented. Patients 1' and 2' are WGS samples matching WGBS while the remaining samples are WGS matching TAPS.

WGBS as a result of the more significant change in the initial bisulfite converted sequence. Both the C>A, G>T, and T>A, A>T changes were more prevalent as a difference from the WGS reference in WGBS. They could be indicative of the oxidative damage of bisulfite treatment and depyrimidination [335].

Ten million randomly selected reads from patient 3 for TAPS and patient 1 for WGBS before and after demodification were used to study the misalignment rate after demodification. The proportion of reads matching exactly their primary location after realignment was calculated for each case. A single WGS sample was also selected for comparison to measure the degree of discrepancy between genomic alignment experiments due to randomness only. By requiring exact matching by read name, starting position, and read flag, TAPS, after demodification and realignment had 98.07975% of the reads aligning to the same genomic coordinates. At the same time, for WGBS, the match was only

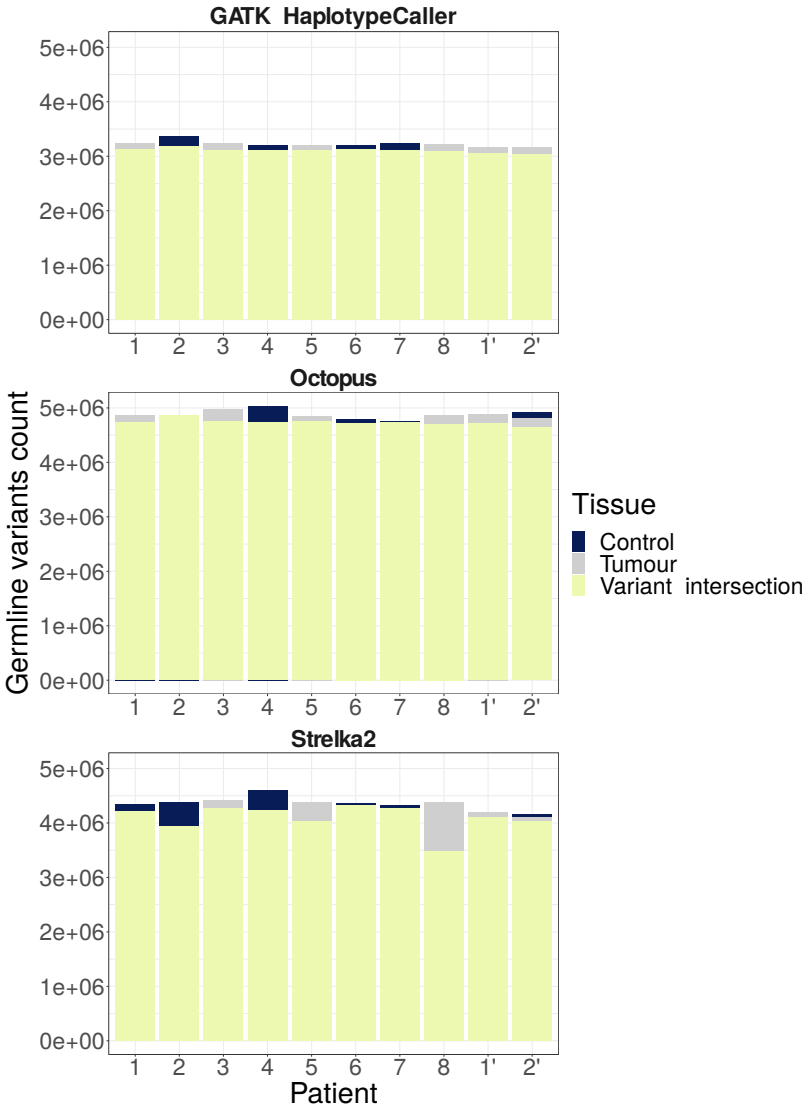


Figure 6.4: The intersection between germline variants found by Strelka2, GATK HaplotypeCaller and Octopus as observed in the healthy and tumour samples in each of the patients. The calls are a result from independent haplocalling of individual samples. Patients 1' and 2' are WGS samples matching WGBS while the remaining samples are WGS matching TAPS.

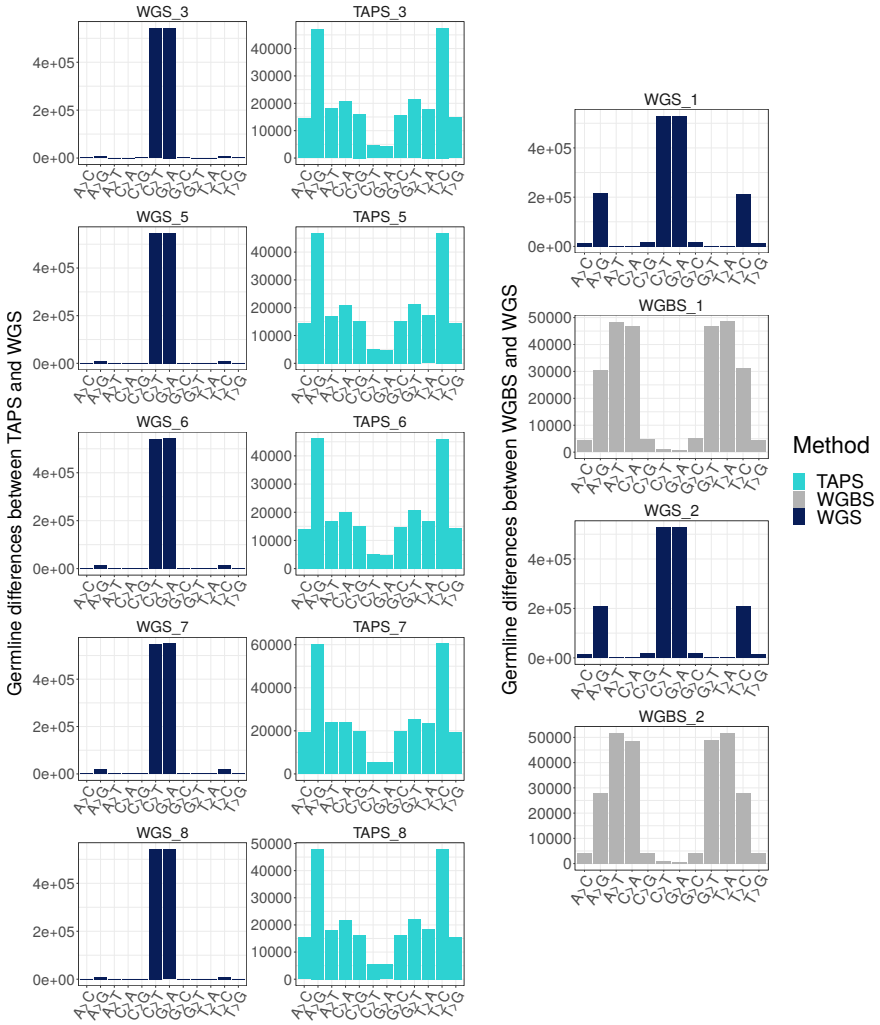


Figure 6.5: Differences between demodified and realigned TAPS and WGS (left), and demodified and realigned WGBS and WGS (right) in terms of germline SNVs (only healthy control tissues showed). The TAPS, WGBS, and WGS label the SNPs uniquely found in the named dataset after the intersection with the WGS germline reference. The names displayed show the patient number and the sequencing method used.

56.32423%. In the TAPS file where alignment was repeated without demodification 99.99917% of the reads aligned to the positions where they were observed initially (only 83 out of 10^6 were aligned to a different position most probably due to random tie-breaking). Sill, in the WGBS this alignment rate match, was

slightly lower at 99.75314%. In the WGS sample used for testing for alignment randomness, the mapping rate was 99.96495%. These results confirm that the demodification process could increase the misalignment rate of bisulfite reads to a greater degree than in TAPS.

6.3 Probabilistic demodification of TAPS and WGBS data

The deterministic T>C and A>G reversion in the WGBS and TAPS datasets removed ≈17% of all germline variants that were occurring on cytosines or guanines.

On the TAPS and WGBS datasets, it was found that around 51.4% of all single-nucleotide germline variants occur at cytosine contexts (cytosines and guanines combined). The SNPs occurring on CpG were disproportionate to their number (≈15% in CpGs, ≈8% in CHG and ≈28% in CHH) as cytosines in CpG dinucleotides were only 4.8% of all covered cytosines in the samples (see Figure 6.6). Furthermore, the majority of germline variants are C>T and G>A. They can be masked by conversion by both sodium bisulfite and pyridine (or picoline) borane. A high percentage (>80%) of SNPs in CpGs were C>T or G>A, while the values for both CHG and CHH were around 60% (see Figure 6.7). In the dbSNP v146 that was used for filtering of SNPs found by the GATK HaplotypeCaller and other tools, 37.98% of all single-nucleotide changes were C>T or G>A, with 56.23% of all SNPs occurring on a cytosine or a guanine position. These results were following previous studies showing that the most common substitution is C>T (respectively G>A), which occurs more frequently in CpG contexts [386].

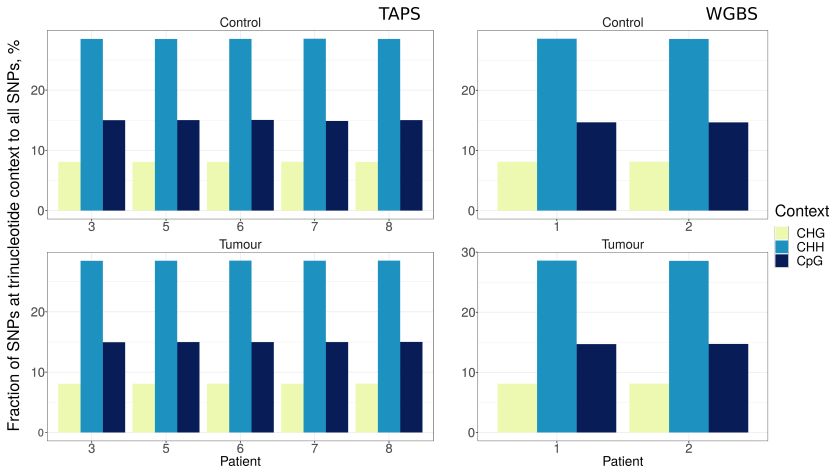


Figure 6.6: The fraction of the germline variants in the intersection between the SNPs discovered by Strelka2, GATK HaplotypeCaller, and Octopus which are at different cytosine contexts in tissue samples (TAPS on the left and WGBS on the right).

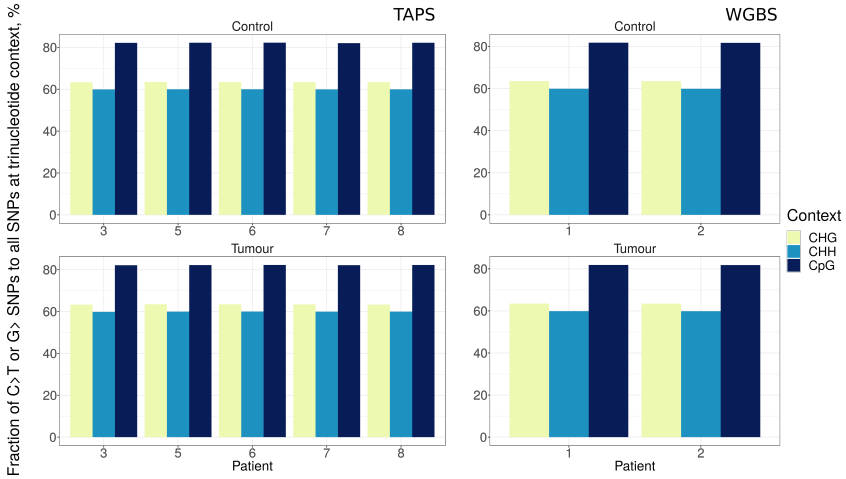


Figure 6.7: The fraction of C>T and G>A germline variants within the intersection between the SNPs discovered by Strelka2, GATK HaplotypeCaller, and Octopus which are at different cytosine contexts in tissue samples (TAPS on the left and WGBS on the right).

Expected read distributions per strand The primary TAPS method is directional where F1 and R2 reads align to the original top (OT) strand and F2 and R2 reads align to the original bottom (OB) strand. The count distributions of reads from chromosome 1 of TAPS Barrett's patient 0 were used to study the

expected nucleotide frequencies between SNPs, modifications, and reference positions. As highly modified positions were considered cytosines at CpG positions with a β value ≥ 0.8, and the SNP information was taken from the haplocalling on the matched WGS sample. The read distributions of homozygous SNPs were symmetric on the top and bottom DNA strands, while this condition was not true for modifications and heterozygous SNPs (see Figure 6.10). The difference between modifications and heterozygous SNPs became more apparent with the inclusion of a read flag information). The three non-reference classes also showed different β value distribution densities where the majority of heterozygous SNPs had a modification value around 0.5, while homozygous SNPs seemed fully modified or unmodified (depending on whether the base substitution was C>T or G>A or else). Modifications similarly to homozygous SNPs also showed a bimodal distribution with modes around 0 and 1 (see Figure 28).

A heuristic approach based on the expected read distribution in case of homozygous SNPs was used to label homozygous SNPs at cytosines during asTair call. The precision and recall of the heuristic for labelling homozygous SNPs in both TAPS and WGBS samples was measured. The C>T and G>A SNPs were $\approx$ 67% of all C or G germline variants found in the intersection between the three WGS germline callers. The overlap between what was labelled as a homozygous SNPs (only C>T and G>A) using the heuristic and the total number of true germline variants with reference C or G (heterozygous and homozygous) suggested good performance of the algorithm. The intersection between them was between 38%, and 46% with $\approx$ 94-99% of the positions labelled as homozygous SNP and overlapping the WGS set being true homozygous variants (see Figures 6.8). The precision defined here as the number of overlapping variants with WGS germline that were labelled by the heuristic was $\approx$ 80% and the recall that was the true homozygous SNPs labelled correctly by the heuristic di-

vided by the number of cytosine or guanine homozygous SNPs was between 60% and 70% (see Figure 6.9). These results confirm the importance of the observed modification ratio on both the strand of interest and the opposite strand.

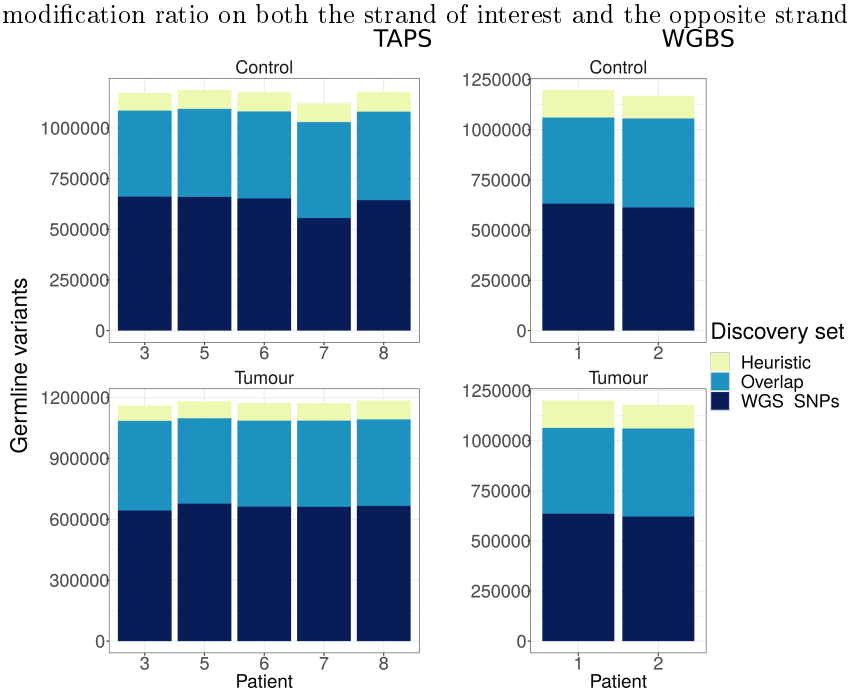


Figure 6.8: The homozygous SNP labelling heuristic in TAPS (left) and WGBS (right) overlaps with the subset of germline variants with reference C or G in WGS. The heuristic was able to capture a large proportion of the true homozygous SNPs found in WGS.

Then the dependency of read distributions on the trinucleotide context and the type of change using cytosine positions information from patient 0 was evaluated. There was a significant difference between the reference and any of the non-reference groups by looking only at read counts (see Figures 29, 31, 32, 30).

Next, the demodification approach was tested on both TAPS and WGBS where after model-based reversal of converted cytosines GATK HaploCaller and Strelka2 germline caller were applied. The performance of the demodification for SNP calling was evaluated using the overlap of both variant callers

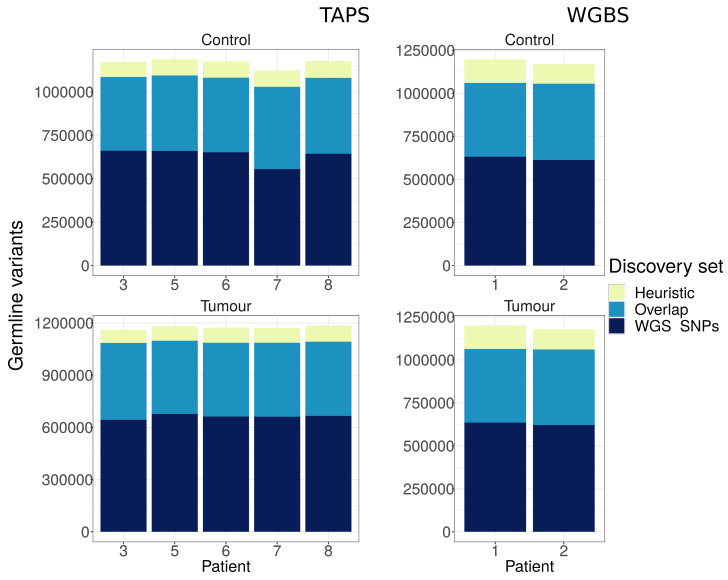
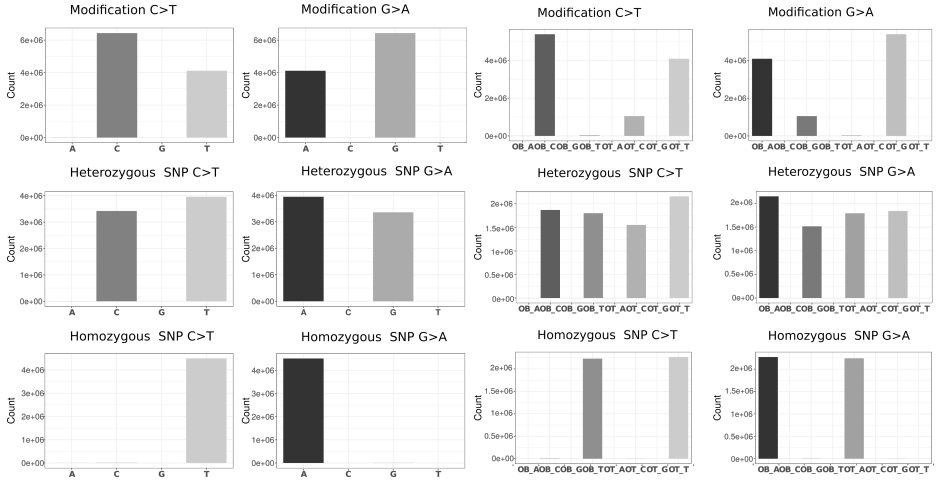


Figure 6.9: The homozygous SNP labelling heuristic in TAPS (left) and WGBS (right) precision and recall metrics by comparison with germline variants in WGS. The precision was ≈ 0.8 while the recall was ≈ 0.6 for the heuristic performance in both TAPS and WGBS.



(a) Read distributions per strand and base.

(b) Read distributions per strand, base and read orientation.

Figure 6.10: Read counts distributions per base and read orientation for SNPs and modifications. There are observable differences between highly modified cytosines, heterozygous and homozygous SNPs on cytosines when strand and read orientation are used.

similarly to the somatic variant calling in order to remove spurious calls. The model-only approach identified 73% of all WGS SNPs in TAPS and 67% in WGBS. The model-based method followed by dbSNP found 73% in TAPS and 69% in WGBS. The false-positive rate was $\approx$ 17-30% for TAPS and 75-97% for WGBS. The TAPS results compare fairly to the demodification with the exclusion of known WGS SNPs where 96% of TAPS and 90% WGBS were discovered, with a false-positive rate of 3% for TAPS and 24% for WGBS respectively. The demodification approach appeared to give fewer false positives due to the algorithms implemented. The somatic variant callers expect low-frequency true variants that are cancer clones or a result of lower cellularity. This sensitivity to lower variant frequencies can be inflated by sequencing and alignment errors if the error profiles of TAPS and WGBS differ from WGS. This observation appeared particularly true for bisulfite-treated data.

After studying the performance of the heuristic approach and the demodification combined with WGS SNP callers, the next step of the analysis was concerned with testing the NN model for SNP discovery. The model was tested on synthetic data, and for the WGBS data—comparing to existing WGBS variant calling tools. The NN model with a selection of features was used for its high performance in both WGBS and TAPS, combined with a smaller number of features compared to the random forest and the full NN model, and its flexibility.

6.4 Benchmarking against other tools

6.4.1 Benchmarking on synthetic data

Given that differences between the matched WGS-TAPS and WGS-WGBS samples would exist as a result of random sampling from highly heterogeneous tissue

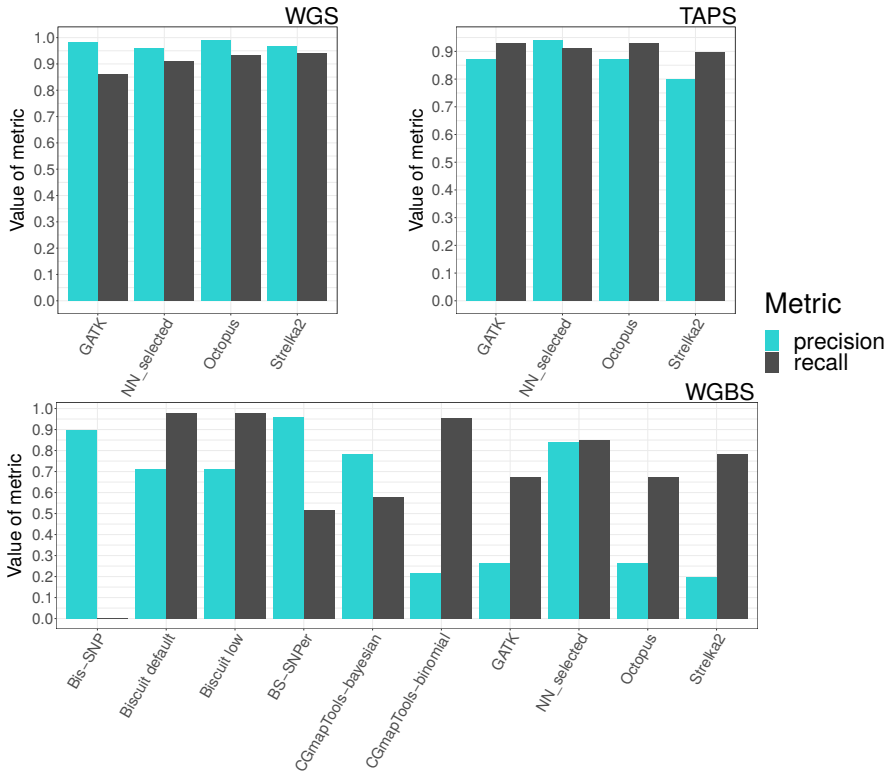


Figure 6.11: Comparison of germline calls on TAPS, WGBS and WGS data. The TAPS NN selected model was used on both the TAPS and the WGS datasets. The NN model achieved comparable precision and recall scores on both TAPS and WGS data (top). The NN model performed better than direct variant calling with WGS tools on WGBS data (bottom). Also, the NN model performed better or on par with existing germline calling tools for WGBS data.

such as Barrett's oesophagus and OAC, the benchmarking of the new SNP calling algorithms was done on synthetic data (more information on the simulation experiment can be found in the following methods subsection). The performance of WGS specific germline callers, the NN model, and WGBS-specific callers, was measured in terms of their precision and recall of SNP discovery. WGBS is used as a gold standard to study cytosine modifications (primarily methylation), but it also can provide information about genetic variants. The low sequence complexity, DNA degradation, and the lack of strand complementarity result-

ing from bisulfite conversion complicate variant calling in WGBS. Nevertheless, there are several germline and somatic variant caller tools created specifically for WGBS data, such as BS-SNPer [108], Biscuit [14], CGmapTools [123] and Bis-SNP [215]. Bis-SNP is based on a modified version of the GATK HaplotypeCaller, whereas the remaining tools employ custom probabilistic models (Biscuit uses samtools pileup and model-based filtering). The tools were run with their default pair-end settings, with an additional run of Biscuit with a higher healthy tissue component of 0.3 resulting in lower expected purity of the sample. The tested NN selected TAPS model showed comparable performance to the WGS germline callers applied on the synthetic WGS data. This result suggested that the model had high generalisation ability on new data (see Figure 6.11). The NN model on the WGBS data showed slightly lower performance compared to TAPS, which can be explained by the lower similarity of WGBS to WGS data. Nevertheless, the NN selected models on TAPS and WGBS outperformed WGS- and WGBS-specific SNP callers.

6.4.2 Benchmarking on real data

As another measure of performance of variant calling tools were tested on the available WGBS data. The results labeled as passing quality filters from Bis-SNP, BS-SNPer, CGmapTools-binomial and Biscuit showed $>$ 90% (Bis-SNP $\approx$ 94%, BS-SNPer $\approx$ 98%, Biscuit default $\approx$ 99%, Biscuit lower purity $\approx$ 96%, CGmapTools-binomial $\approx$ 96%, CGmapTools-bayesian $\approx$ 76%) overlap with the WGS haplocalls (agreement between Octopus, Strelka2, GATK HaplotypeCaller). Still, these calls are a relatively small proportion of all calls being $<$ 45% (fraction of overlapping positions to all positions labelled as SNPs: Bis-SNP $\approx$ 34%, BS-SNPer $\approx$ 12%, Biscuit default $\approx$ 10%, Biscuit lower purity $\approx$ 44%, CGmapTools-binomial 0.26%, CGmapTools-bayesian 0.45%). The overlap be-

tween all the WGBS callers with the WGS SNPs was ≈ 65%, and excluding the worst-performing caller (CGmapTools) ≈ 90%. The relatively higher proportion of false-positive reads and their reduction in numbers with the assignment of tumour purity may indicate that the WGBS samples used for testing are of lower purity than what is expected by the models. Another possibility was the existence of other differences that may need finer tuning of model parameters before repeating the germline calling (see Figure 6.12). Even though the individual WGBS germline callers had labelled spurious calls as passing quality filters, they had a high agreement with the WGS SNPs. Furthermore, their agreement with the overlap of GATK Haplocaller and Strelka2 when applied directly to WGBS data was only on 0.5% of positions (see Figure 6.13).

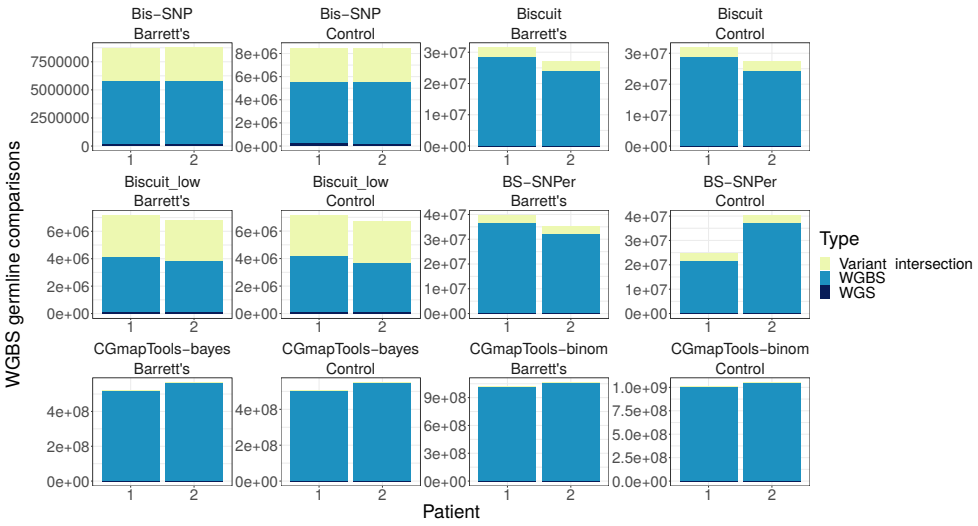


Figure 6.12: Comparison of WGBS variant callers to WGS (only SNPs overlapping between Strelka2, GATK HaplotypeCaller and Octopus) germline calls. A high proportion of false-positive calls was discovered in the individual call sets.

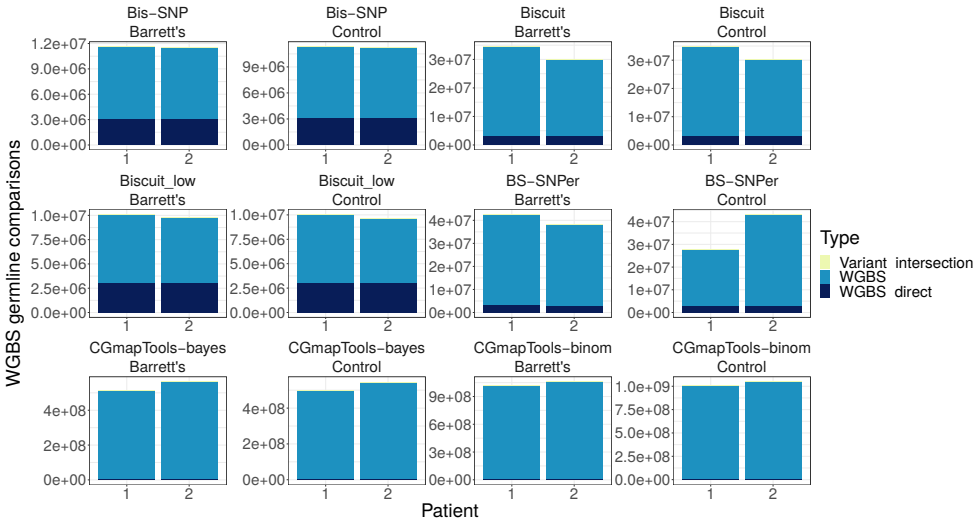


Figure 6.13: Comparison of WGBS variant callers to WGBS when WGS callers were applied (only SNPs overlapping between Strelka2 and GATK HaplotypeCaller). Less similarity in terms of variant calls discovered in WGBS through variant calling with WGS tools without demodification and WGBS-specific variant tools was observed.

6.5 Discussion and future directions

The level at which 'randomness' is introduced during alignment of TAPS, and WGBS files as opposed to being a side effect from the demodification was measured. It was shown that demodification followed by realignment leads to misalignment, particularly for WGBS data. The higher misalignment rate, together with the higher level of systematic errors in the bisulfite-treated data suggest lower applicability of the demodification approach unless a stringent filtering algorithm is developed. In comparison, relying on WGBS-specific SNP calling may give better results, where the use of several WGBS-variant callers removes the number of false-positive calls through a conservative intersection between call sets.

Several different techniques for studying SNPs in TAPS and WGBS data

were compared. Using synthetic data and real patient data to check the performance of WGBS-callers (precision and recall for the models described in this work were presented in chapter 4), showed that the suggested NN models could be reliably applied for SNP calling. A limitation in the generation of synthetic data was that it was based on WGS data later modified in TAPS- and WGBS-specific way. The similarity between TAPS and WGS is higher in terms of base quality and library complexity, as shown in chapter 3. Thus, variant calling on the synthetic WGBS data was expected to show worse results than on real bisulfite data. This hypothesis was later confirmed on the real dataset, where the majority of WGS SNPs were found, while the false-positive rate of individual callers was decreased by using only overlapping SNP calls. A better approach that would amend the issues listed would be to create fully synthetic TAPS and WGBS data with the same base quality, read length, and indel rate as observed in TAPS and WGBS, which was then subjected to the controlled introduction of modifications and SNPs, or to use WGBS and TAPS specific 'platinum' genomes used for method testing. Such reference information is not available at present for epigenomic sequencing data but is commonly used to benchmark WGS variant callers [91].

As discussed previously, additional training data would benefit the germline calling models through additional information about SNPs compared to errors and modifications.

Chapter 7

Discussion

The presented work focused on exploring specific characteristics of a novel epigenomic method TET-assisted pyridine borane sequencing (TAPS) in comparison to the 'gold standard' method for base resolution study of DNA modifications—bisulfite sequencing. The current study found that TAPS sequencing data has a higher alignment rate and more even coverage to WGBS. However, the results obtained with both methods and particularly WGBS depend strongly on the alignment and filtering methods applied to sequenced reads. It seems possible that previous observations about the relatively low alignment rate in bisulfite sequencing were due to the choice of an aligner to a greater extent than the chemistry of the method. TAPS showed fewer chemistry and alignment-based errors and a higher overall similarity to WGS in terms of base quality when compared to WGBS. Despite these promising results, further chemical optimisation of the TAPS method would be required to ensure unbiased conversion of non-CpG DNA contexts.

Next, as part of the project, a tool for analysis of TAPS and WGBS data, asTair, was created. The tool enables users to align sequencing data and call

modifications in both methods. As an objective, sensitive modification and variant calling were sought using three different techniques—a heuristic based on sequencing reads ratios, demodification followed by WGS variant calling tools, and machine learning SNP calling models. The heuristic alone correctly recognised the majority of homozygous SNPs matching the truth set. However, a limitation was that this algorithm could be implemented only for homozygous SNPs, while in TAPS, a greater misclassification rate was observed between heterozygous SNPs and modifications. Two different demodification approaches were tested—a complete reversal of any possible conversion C>T, and restoration of positions that had a high probability of being true modifications. The first method was used mostly for comparison and as a development step, given that the largest proportion of single base substitutions are known to be C>T [386]. For the second method, three machine learning approaches were used: logistic regression, random forest, and neural networks. Random forests and neural networks outperformed logistic regression, while neural networks showed higher generalisation ability to novel datasets and are expected to improve performance through training on richer data. The false-positive rate of the algorithms inflated the number of discovered somatic variants compared to the approach where known SNPs from the WGS were excluded from demodification. Post-classification filtering using dbSNP information reduced considerably the error rate and suggested that an additional filtering step might improve the performance of the algorithms. This hypothesis was corroborated by the observation made as part of the error analysis of models that differences in certain features between correctly classified and misclassified SNPs (base qualities, alignment scores, coverage and certain alignment tags) exist.

The demodification approach gave more promising results in TAPS compared to WGBS due to the higher similarity of the error profiles of TAPS and

WGS compared to WGBS and WGS. Main culprits for the difference were DNA degradation under high temperatures, and acidic conditions in bisulfite treatment and alignment of short sequencing reads. Less than 60% (56%) of the demodified and realigned WGBS sequencing reads ended at their exact previous locations compared to 98% in TAPS. The DNA degradation due to bisulfite treatment was evaluated by Tanaka et al. [335] and was measured around 90% due to the harsh preparation conditions. Given the alignment issues and the effects of chemical treatment on DNA, a relatively high false-positive rate was found in the case of WGS somatic variant callers applied on demodified data. A possible approach to increase the signal to noise ratio of the found variants was to trust only variant calls found by two or more WGS variant callers. The suggested method presented better results in TAPS compared to WGBS due to the lesser degree of systematic errors mislabeled by the somatic variant callers as variants.

When demodification followed by WGS germline variant calling were applied to TAPS and WGBS data, again a lower false-positive rate was observed in TAPS compared to WGBS. In the bisulfite data, even in the case where known SNPs from the WGS matched samples were used to preserve true C>T SNPs, the false positives were eight times more than in TAPS. Given that WGBS-specific variant callers were able to discover $\approx$ 90% of WGS SNPs, the discrepancy in the number of SNPs found in the bisulfite-treated data suggested that more information could potentially be lost using demodification and WGS variant calling in WGBS compared to applying custom variant calling methods.

The algorithms developed for demodification showed high precision and recall ($\geq$ 99%) for discovering true modifications in both WGBS and TAPS data; thus they can be applied directly as part of a modification calling pipeline to decrease the chance of mislabeling cytosine positions. It is expected that the er-

ror rate for misclassifying modifications and reference positions would fall when more training data are available, that would enable more sensitive demodification, and variant calling was relying only on the presented algorithms.

The final step of studying the differences between C>T base changes resulting from chemistry compared to mutational processes included the building and testing of a SNP calling algorithm. These results were promising in terms of correctly detecting germline variants. On a simulated testing dataset, the TAPS and WGBS trained neural models (with a selected subset of features) outperformed WGS- and WGBS-specific SNP callers. Similarly to the demodification algorithm, diverse training data and post-classification filtering can improve the performance of the algorithm.

Also, a comparison of existing somatic and germline variant callers on the available WGBS data were tested. The results from this analysis suggested that the conservative approach of using only the intersection of two or more variant callers decreased the false-positive rate while preserving $\approx 90\%$ similarity to the WGS truth set.

Additionally, this work studied the error rates of somatic variant calling as dependent on coverage and showed that for tumour samples above the detection limit for ploidy and tumour purity require at least 25x mean sequencing coverage to detect 50-95% of variants. The sampling and natural variability of using different matched healthy tissues as control could alter the somatic calling by 5-17%. By comparing the somatic variant callers, Mutect2 appeared more conservative when calling variants and showed a lower error rate in low coverages (particularly $\leq$ 10x) and after demodification of WGBS and TAPS data compared to Strelka2.

Finally, the observation that Illumina reads may suffer from a higher rate of deletions compared to insertions [239] was confirmed on the data used for

analysis. The highest rate of deletions, however, was seen in WGBS and was assumed to be at least partially influenced by bisulfite chemistry.

The thesis focused on single nucleotide changes as part of both germline and somatic variation. The main reason for this choice was the natural similarity between converted cytosines (modified or unmodified) in TAPS and WGBS to single nucleotide variants. One possible future direction for this project is the development of callers of large scale structural and copy number variation for both WGBS and TAPS. Through solving the problem of distinguishing DNA modifications from SNVs and SNPs, the development of TAPS-specific tools for copy number discovery would be enabled. Standard copy number calling algorithms rely on both genomic coverage (which is usually GC corrected to ensure PCR and context biases are removed), as well as somatic variant calls when available [95], [20]. The study of TAPS-specific biases could also aid the development of callers able to discover large scale genomic rearrangements in epigenomic data. Algorithms identifying structural variation rely on information about the alignment of paired reads, mainly if the reads are located far apart (on the same or different chromosomes), or have an incorrect orientation towards each other (such reads are known as discordant reads) [20]. The number of discordant reads can be affected by the alignment, where the conversion events would decrease the alignment rates of both WGBS and TAPS compared to WGS and may change the alignment score of the reads.

The model development together with its performance evaluation for somatic and germline variant calling presented in this work should be seen as a first step towards the development of more comprehensive models and tools for variant discovery in epigenomic data. The limited available data could have biased the models and decreased the scope of the observable errors. Moreover, the reliance on data availability could be diminished by taking a more structured approach

incorporating a Bayesian model for variant calling either as stand-alone model or as a feature for the ML models. Generally, the ability to discover genetic variants and modifications from the same data could boost clinical applicability due to the higher utility of limited data sources. Nevertheless, extensive data science and model development work would be required prior to applying such methods to precious samples.

Furthermore, the use of variants that are in the intersection of the variant lists generated by different variant callers as reference or truth set may be problematic. On the one hand, such approach is conservative as it removes a proportion of possibly high confidence variants, and on the other hand, there is a risk of keeping variants because of a shared dependence between callers (*e.g.* dependence on the same SNP data or models). The second issue would mean that the variant calls were not independently confirmed, but kept because of majority voting.

The results of this thesis could be summarised in the spirit of the 'no free lunch' theorem [367], where two optimisation algorithms have the same averaged performance over all possible problems. The analysis showed that TAPS as a method retains a higher similarity to the initial untreated DNA sequence that simplifies alignment. The ability of TAPS to preserve the information complexity of low input samples could be used for particular applications such as cfDNA and single-cell sequencing. However, apart from the conversion bias against non-CpG positions, variant calling in TAPS is complicated by the higher similarity between C>T variants and modifications. In comparison, in WGBS the difference between C>T genetic variants and the reference is greater (unmodified cytosines are converted to thymines), but so is the proportion of chemically and alignment induced errors present in the data. Thus, the choice of the epigenomic method depends on their specific performance over a particular set of problems.

Chapter 8

Conclusion

The main goal of this thesis was to create computational methods for the analysis of novel epigenomic data (TAPS). The second objective was to study the differences between TAPS, WGS, and WGBS. An important point to make is that the TAPS method development was driven by observations made through data analysis. As a result of the presented work, a tool (asTair) for modification calling and alignment of both TAPS and WGBS data was created. Also, models for sensitive modification and SNP calling on TAPS and WGBS were built and tested. The thesis showed that distinguishing modifications from single-base variants could be applied to TAPS in a more computationally efficient way than in WGBS. Possible future directions were outlined in the discussion of each of the results chapters. The development and testing of a model for filtering of candidate variants for SNP calling, together with the building of a reliable somatic variant caller for TAPS and WGBS data are seen as a direct continuation of this work.

The structure and individual contributions per chapter of the thesis are as follows. Chapter 1 introduced epigenetics and its links with cancer, sequencing

technologies for epigenomics, WGBS and TAPS chemistry, and an overview of machine learning with a focus on applications in the biomedical sciences. Chapter 2 presented a complete description of the methods and tools used for analysis and visualisation of sequencing data. Chapter 3 showed the development of the TAPS method and the modification calling pipeline and the tool for modification calling in both TAPS and WGBS—asTair. Analyses of possible DNA context, coverage and alignment biases were studied in TAPS. TAPS showed higher alignment rates and more even coverage compared to WGBS while having a similar conversion rate on methylated CpGs, which are of primary interest in biomedical research. Chapter 4 gave a description of the datasets used for the analysis and presented the machine learning model development and testing on a subset of the data. Two different types of models were created—1) for modification calling and demodification, and 2) for SNP calling. Chapters 5 and 6 were focused on applying the machine learning models build in chapter 4 and evaluating of their performance against currently available tools. The main focus of chapter 5 was calling somatic variants in TAPS and WGBS data and comparing these results against a ground truth set consisting of WGS variant calls. Chapter 6 presented germline calling and potential systematic biases in TAPS and WGBS, the effect of which can be minimised using filtering methods. Chapters 5 and 6 showed the applicability of the demodification approach for TAPS, and in a lesser degree for WGBS; for both types of data, additional filtering is expected to decrease the error rate.

The results from this thesis could ease the future development of algorithms for analysis of TAPS data through providing information for expected biases and giving a list of potential features for classification of genomic positions. Building upon these findings could enable the extraction of combined information of modifications and genetic variants from a single data source. The ability

to simultaneously read and interpret DNA modifications and genetic variation might shed more light in the development of complex diseases such as cancer, along with elucidating natural processes like ageing.

Appendices

"At this rate, by Tuesday it will be Thursday. By Wednesday it will be August, and by Thursday it will be the end of existence as we know it!"

Professor Farnsworth, Futurama

I Additional results

I Development of the NN model

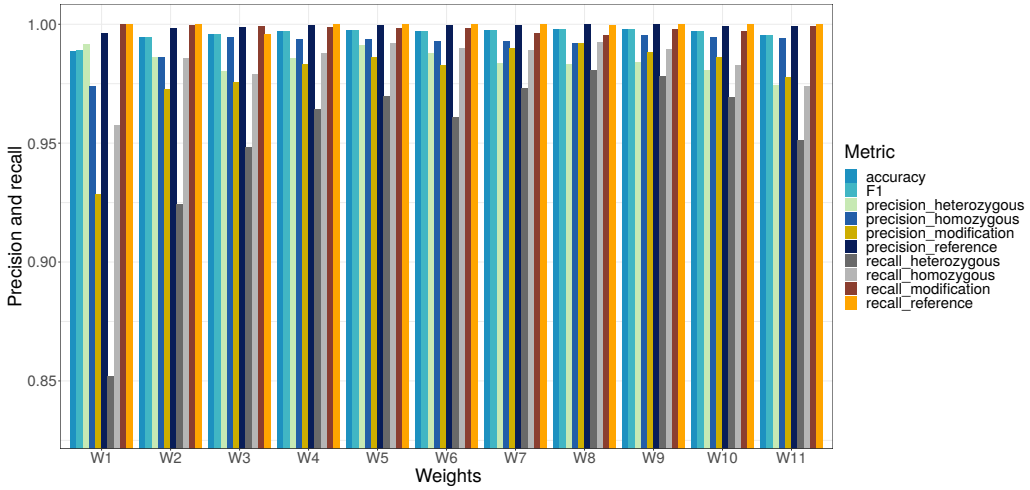
In order to study the dependency between the class weights and the performance of the classifier, 11 different weight proportions were chosen (see Table 1) and tested on a subset of the positions (one million) used for training. The best performing cases were observed when higher ratios were given to the minority types SNPs and modifications which were artificially brought to the same numbers using SMOTE (see Figure 1). The most sensitive to the value of the weight metric was the recall for the heterozygous class, and very low values (<0.1) were decreasing the overall performance of the model as they also decreased substantially the modification calling precision and homozygous recall. The most probable cause is that SNPs are confused as being modifications if they are considered to be too rare by the model using the small values for the weights. The metrics for the reference were high in all cases suggesting that a significant issue to be expected with the model would occur because of data insufficiency for the minority classes.

Next, on the subject of model architecture, more testing was performed additional to the empirical observation of the classification performance and the number of layers used. A random subsample of a few million positions from TAPS and WGBS labelled data were used to show the dependency of classification performance on the number of layers and the neurons in a layer in a NN. Feed-forward networks with one to ten layers and 5, 10, 20, 50, 100, and 200 neurons per layers were tested. Given the small datasets, comparable performance in terms of accuracy, F1, precision, and recall was observed. In TAPS having five neurons per layer lead to unstable performance for the minority classes for

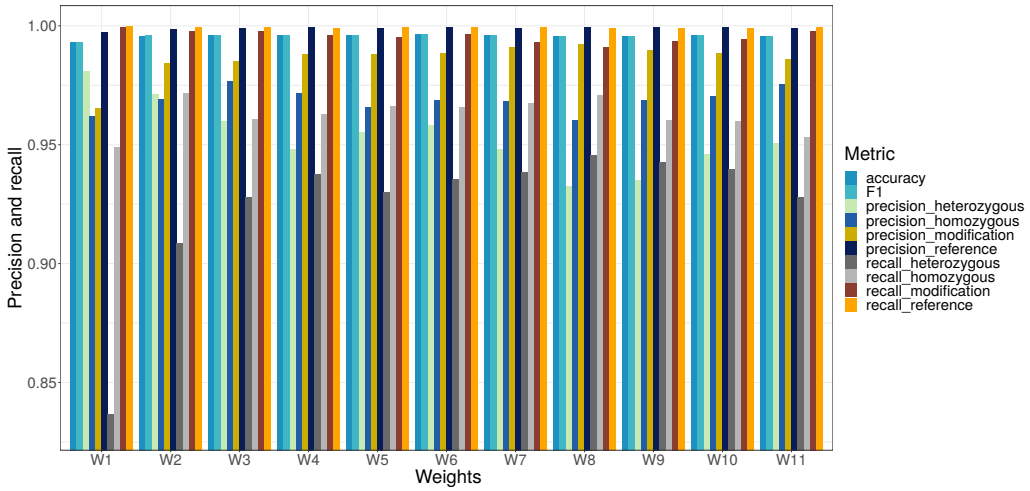
experiment	weights
W1	het:0.01,hom:0.01,mod:0.4,ref:0.58
W2	het:0.02,hom:0.02,mod:0.4,ref:0.56
W3	het:0.05,hom:0.05,mod:0.4,ref:0.5
W4	het:0.1,hom:0.1,mod:0.4,ref:0.4
W5	het:0.2,hom:0.2,mod:0.4,ref:0.2
W6	het:0.1,hom:0.1,mod:0.1,ref:0.7
W7	het:0.2,hom:0.2,mod:0.2,ref:0.4
W8	het:0.3,hom:0.3,mod:0.3,ref:0.1
W9	het:0.3,hom:0.2,mod:0.4,ref:0.1
W10	het:0.1,hom:0.05,mod:0.35,ref:0.5
W11	het:0.05,hom:0.02,mod:0.23,ref:0.7

Table 1: Weights per class used for different training experiments using a sample from TAPS training data. Testing different weight combinations ensured that optimal values were used for the training of the model.

all layers, and the same was observed when having 200 neurons per layer and $>$6 layers due to overfitting (see Figure 17). Judging by the accuracy and F1 scores, in most cases 4-7 layers had the best performance (see Figures 16 and 8). The slightly lower accuracy scores on the training set in the cases 5,10, 20, 50 neurons per layer may be a result of the regularisation of the training set. The expected difference in speed was over 3X between training a model with ten layers and five neurons each and having 200 neurons at each layer. Similar dependencies were seen in the WGBS data, with overall lower accuracy and F1 scores, because of the relatively higher misclassification rate of the minority (SNP) classes compared to TAPS (see Figures 18, 9, 19). Another difference between WGBS and TAPS was that the 7-10 layers having $\geq$ 100 neurons had a lesser degree of overfitting in the test set, which suggested that the genetic variant versus modification classification problem was of higher complexity in WGBS. One reason for that was the higher number of sequencing and alignment errors and the lower similarity to the reference genome in terms of the sequence.



a) Training dataset



b) Testing dataset

Figure 1: Different performance metrics show classification success depending on the class weights in the NN model using TAPS data. The worst performance overall was observed for the heterozygous and homozygous precision (light green and medium blue), and recall (dark and light grey). The trend in these four categories is visible between the training and the testing set, where the training sets have better performance as expected.

II Additional figures and tables

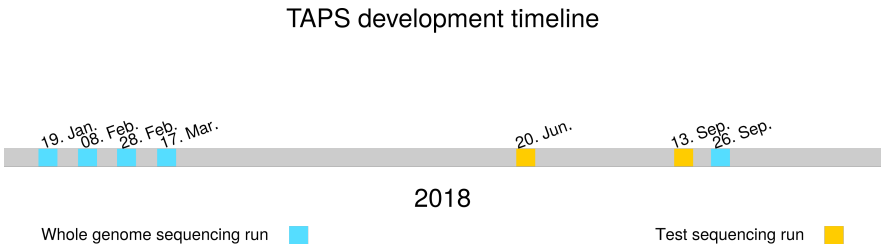


Figure 2: The TAPS method development timeline: the analysis of the five full-genome sequencing runs and two library preparation test runs has helped the development.

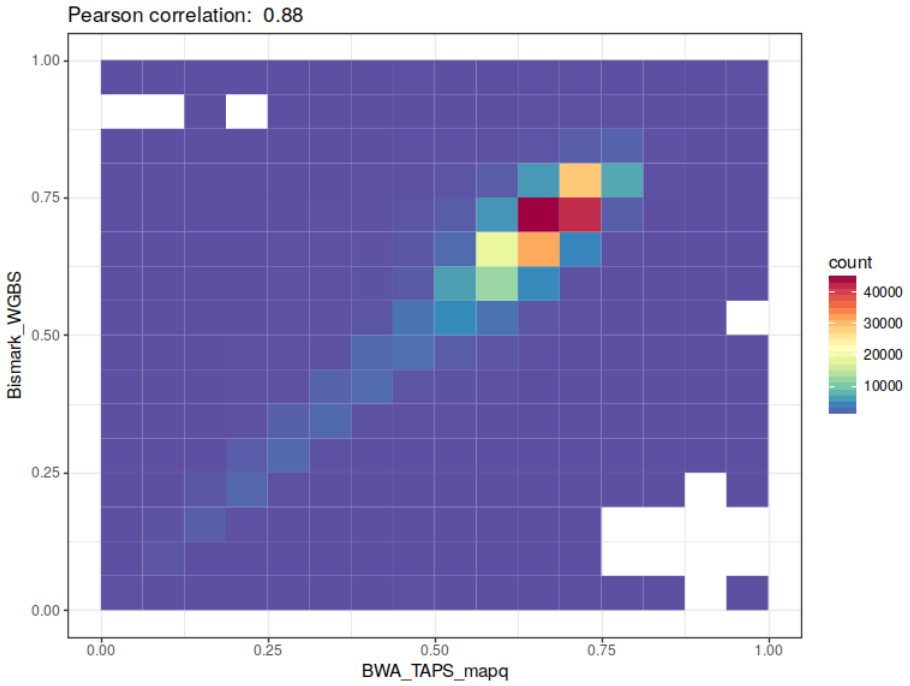


Figure 3: A visualisation of the Pearson correlation of TAPS and WGBS CpG modification values normalised at 10k bins. The mm9 genome was binned into 10k bins, and the CpG modification level was calculated as the total number of modified to modified + unmodified reads, divided by the number of CpGs per bin. The normalised β values of TAPS and WGBS showed a higher correlation at 10k (0.88, p-value $\leq 2.2e^{-16}$, Pearson's product-moment correlation, two-sided hypothesis, CI=0.95) compared to per base (0.68, p-value $\leq 2.2e^{-16}$, Pearson's product-moment correlation, two-sided hypothesis, CI=0.95). Possible explanations are both the lower mean coverage of WGBS and the higher false-positive rate of TAPS, as both factors contribute to the difference of observed modification values per individual CpGs.

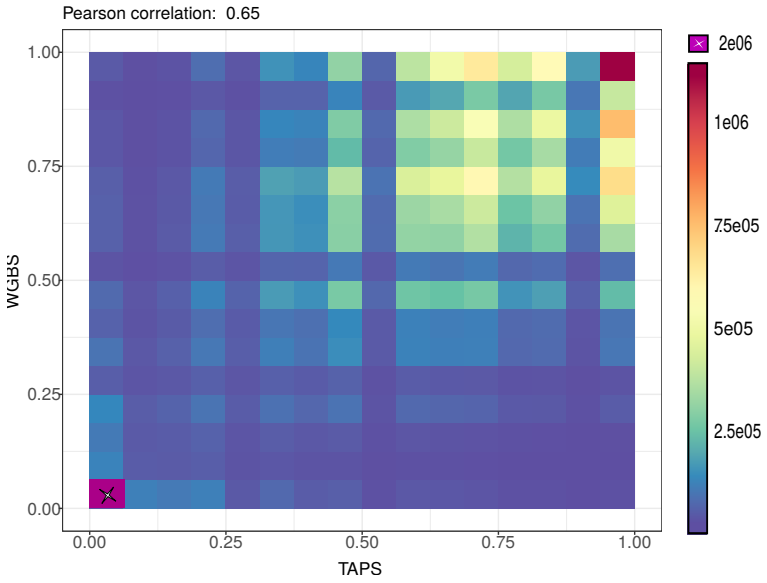
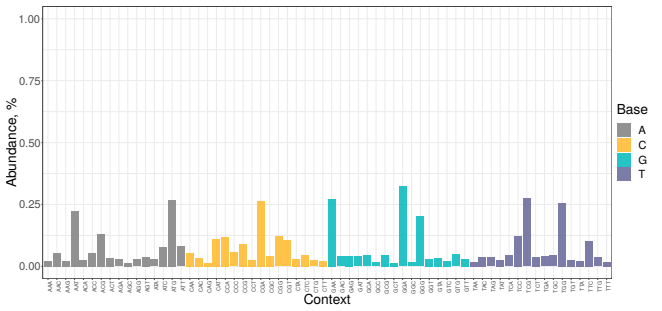
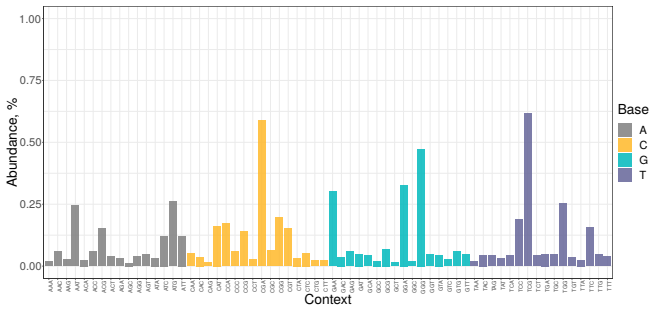


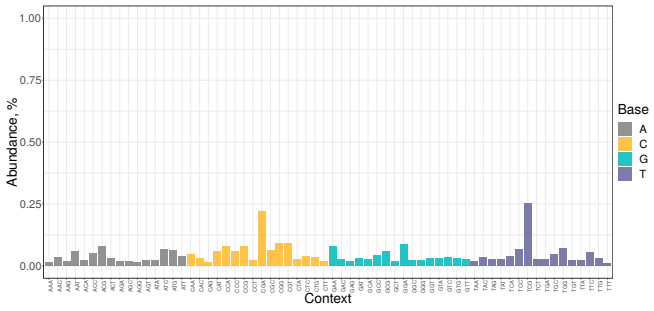
Figure 4: The high correlation between TAPS and WGBS in terms of modified CpGs is preserved after TAPS is downsampled to the median coverage of WGBS.



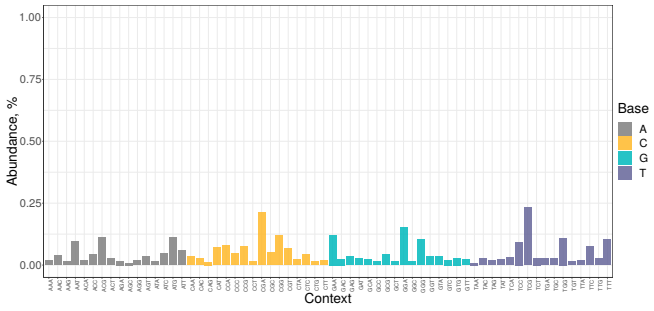
(a) TAPS insertions



(b) tWGS insertions

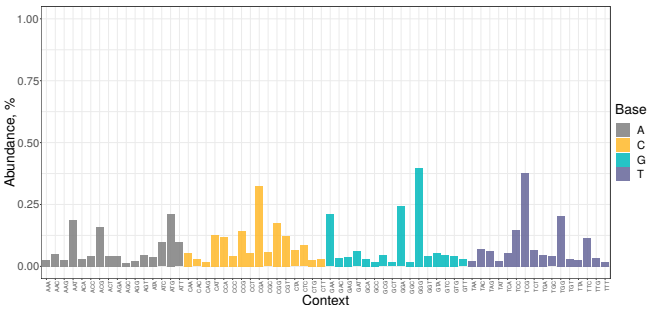


(c) WGBS insertions

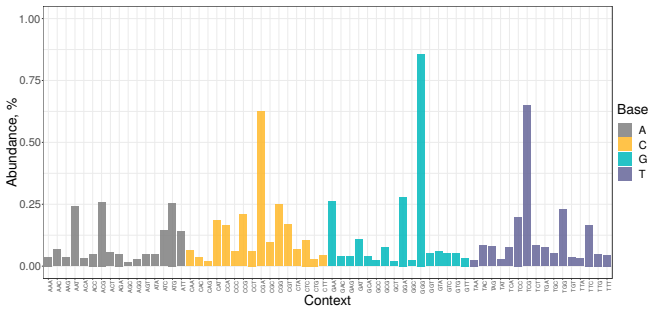


(d) bWGS insertions

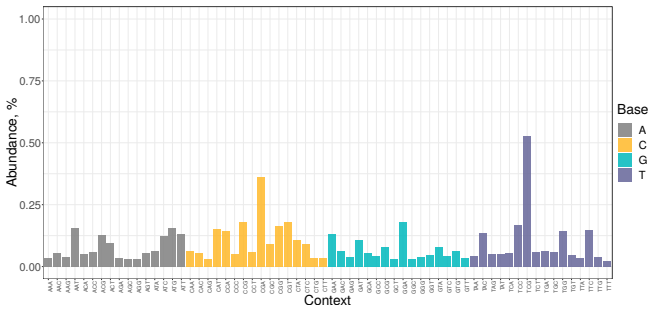
Figure 5: A visualisation of the abundance of insertions per trinucleotide context. The number of substitutions per context was normalised to the total number of substitutions for each group prior to plotting.



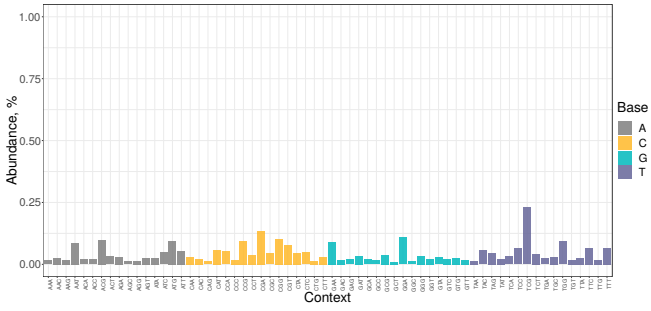
(a) TAPS deletions



(b) tWGS deletions



(c) WGBS deletions



(d) bWGS deletions

Figure 6: A visualisation of the abundance of deletions per trinucleotide context. The number of substitutions per context was normalised to the total number of substitutions for each group prior to plotting.

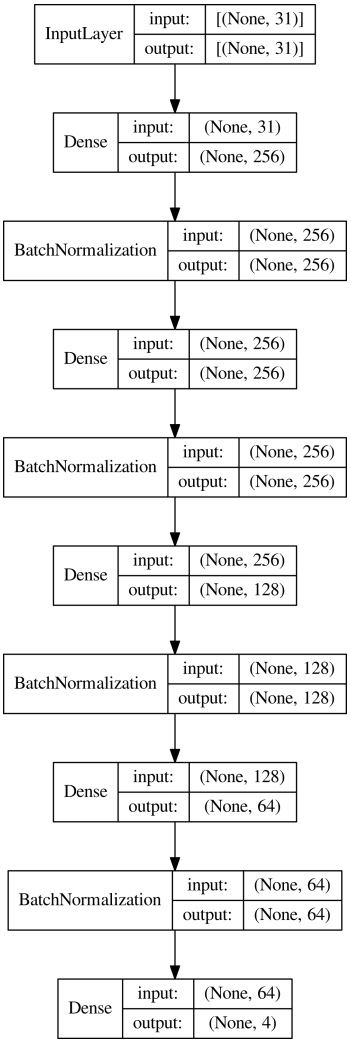
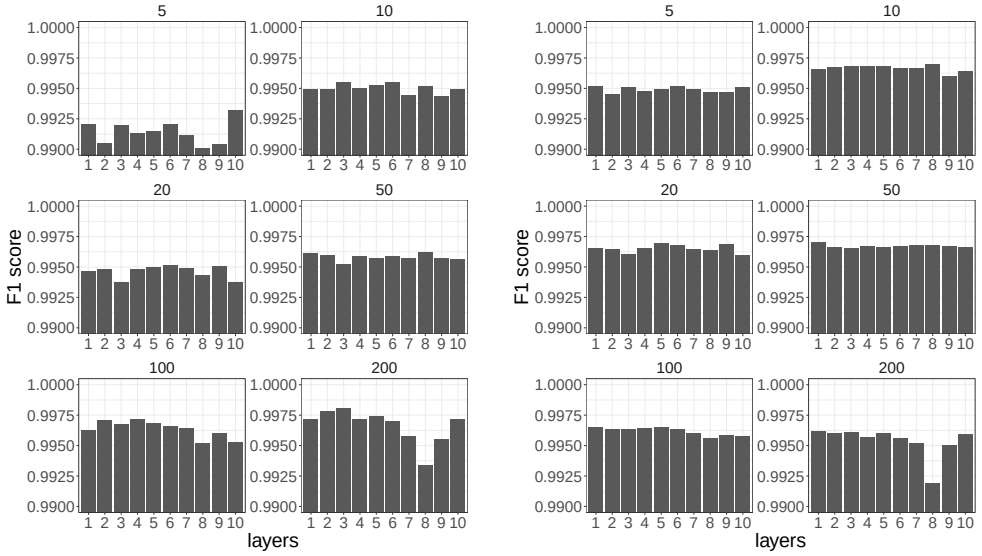


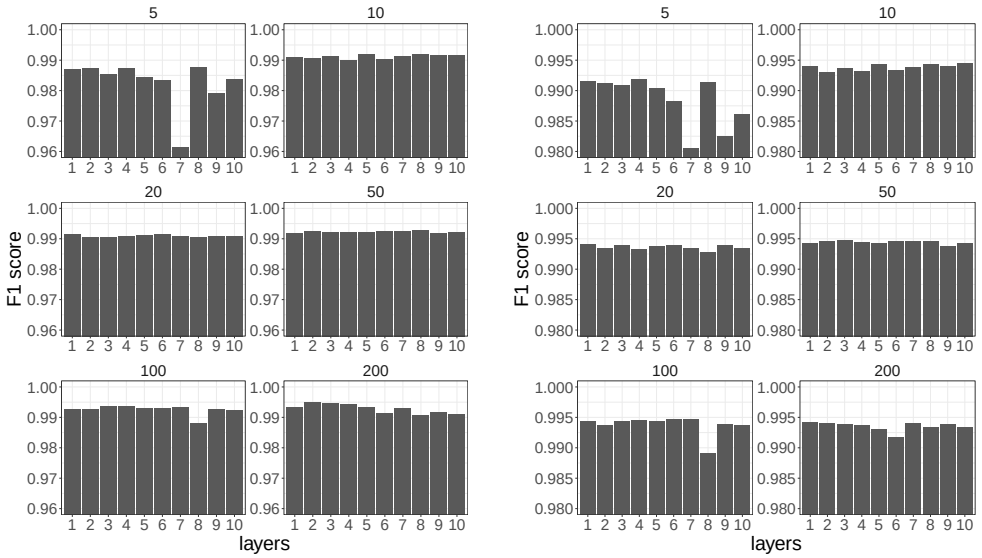
Figure 7: A general scheme of the feedforward network used for classification. The NN contained input and output layers, and four hidden dense layers with a batch normalisation step between them. The figure shows the input and output dimensions (for the input layers these are features, for the remaining layers these units correspond to the neurons in the layer, where for the output these four neurons match the four classes—reference, modification, heterozygous, and homozygous SNP). The scheme was drawn using keras model plotting utility *plot_model*.



a) Training dataset

b) Testing dataset

Figure 8: F1 score of the classifier depending on the number of layers and neurons in the NN model using TAPS data



a) Training dataset

b) Testing dataset

Figure 9: F1 score of the classifier depending on the number of layers and neurons in the NN model using WGBS data

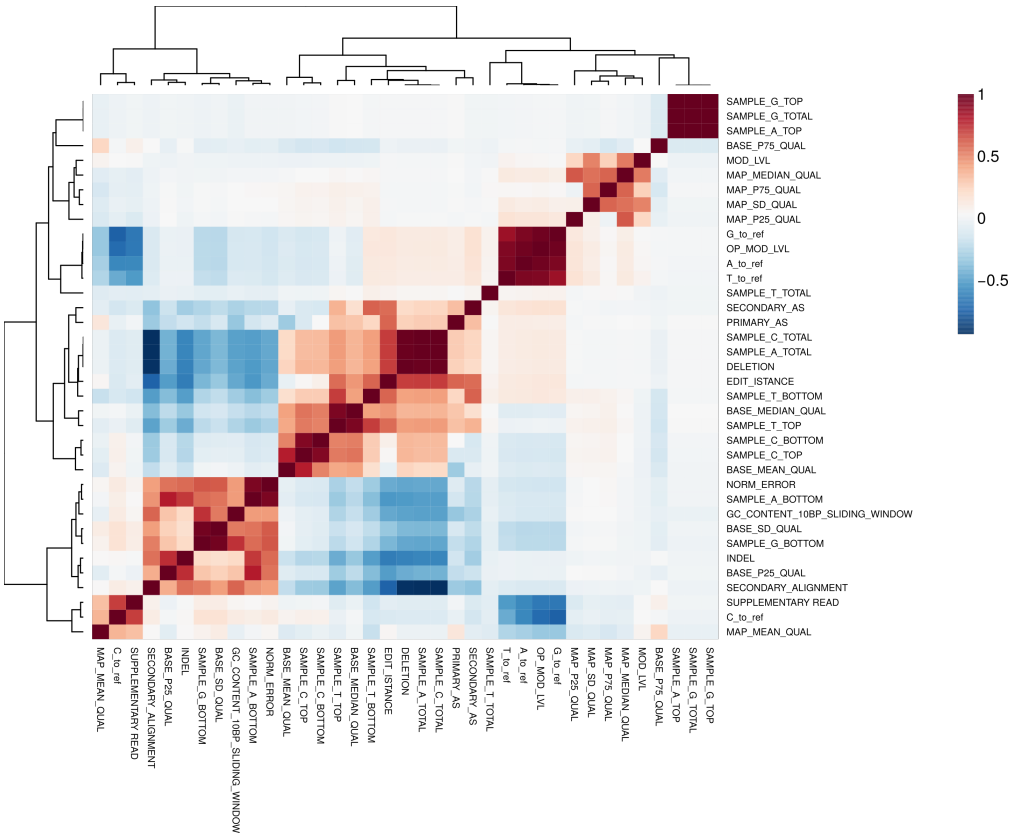


Figure 10: The Pearson correlation between features from the initial feature set was visualised. The CpG positions were labelled as heterozygous SNPs by the GATK HaplotypeCaller. Only cytosine positions from chromosome 1 of patient 0 were used for the visualisation.

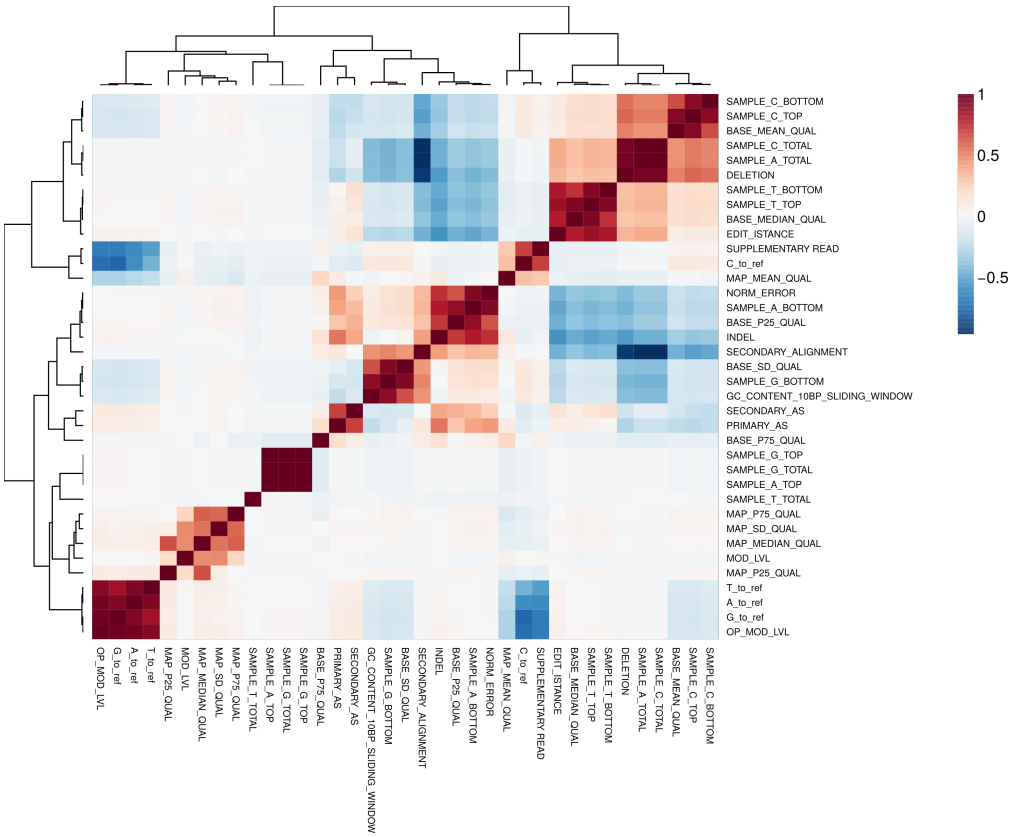


Figure 11: The Pearson correlation between features from the initial feature set was visualised. The non-CpG positions were labelled as heterozygous SNPs by the GATK HaplotypeCaller. Only cytosine positions from chromosome 1 of patient 0 were used for the visualisation.

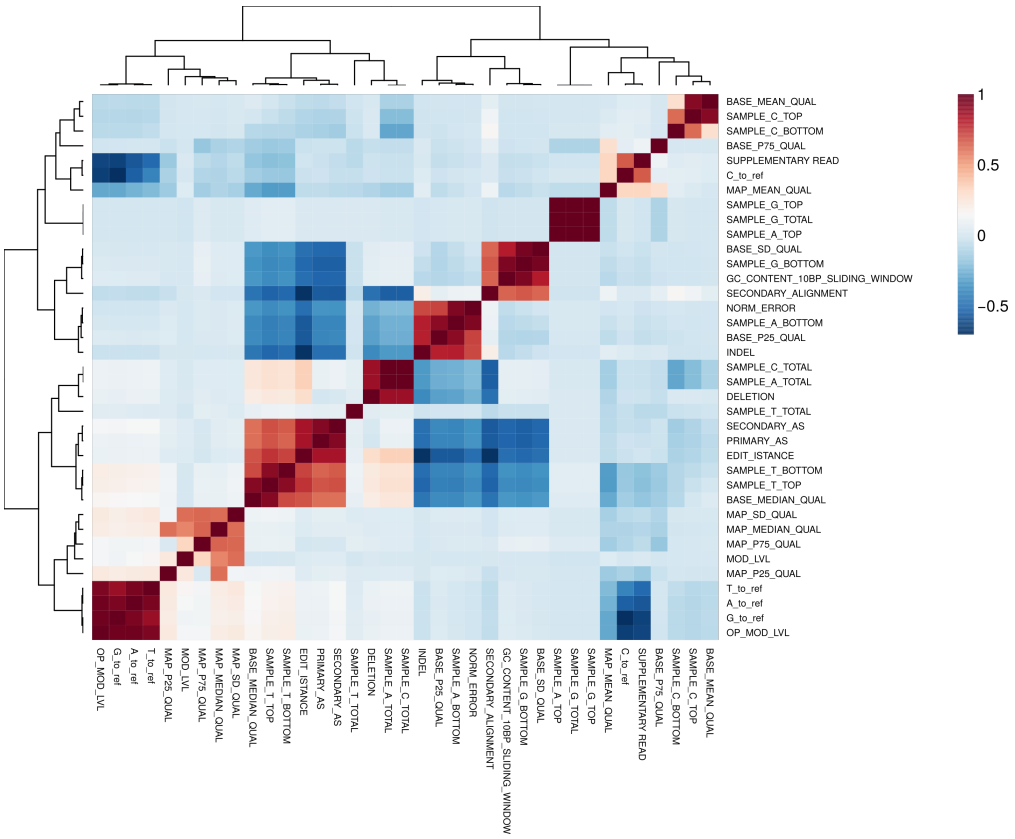


Figure 12: The Pearson correlation between features from the initial feature set was visualised. The CpG positions were labelled as homozygous SNPs by the GATK HaplotypeCaller. Only cytosine positions from chromosome 1 of patient 0 were used for the visualisation.

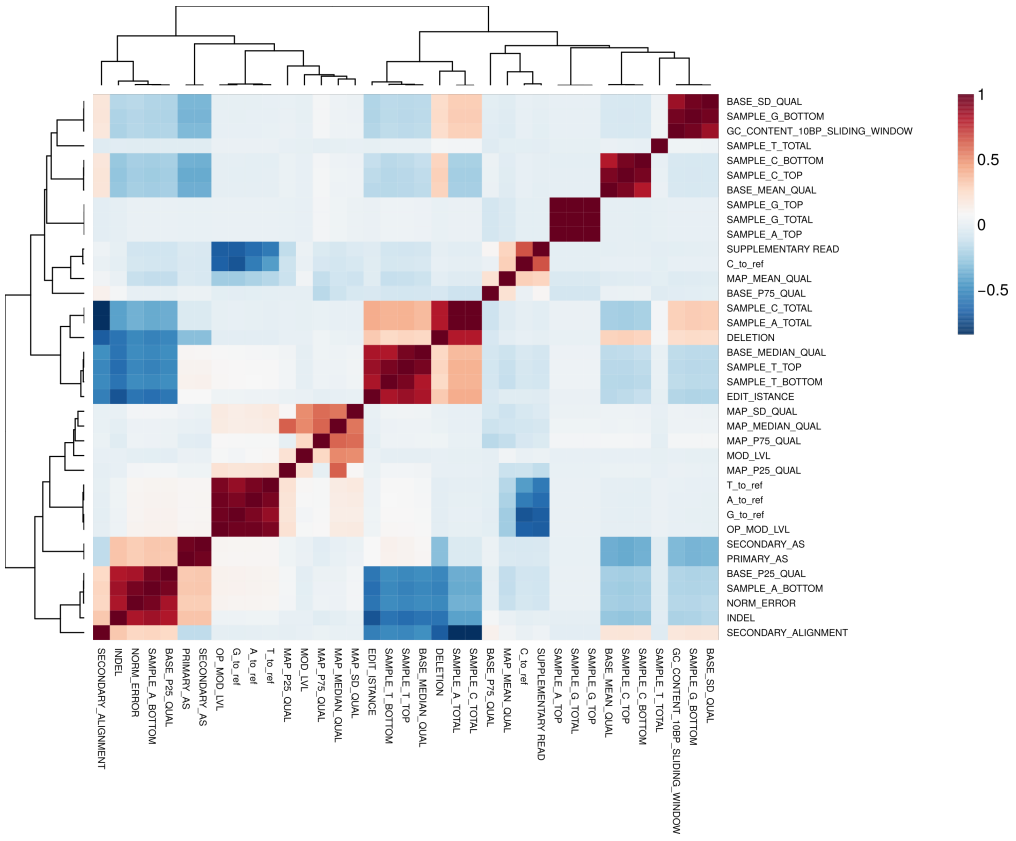


Figure 13: The Pearson correlation between features from the initial feature set was visualised. The non-CpG positions were labelled as homozygous SNPs by the GATK HaplotypeCaller. Only cytosine positions from chromosome 1 of patient 0 were used for the visualisation.

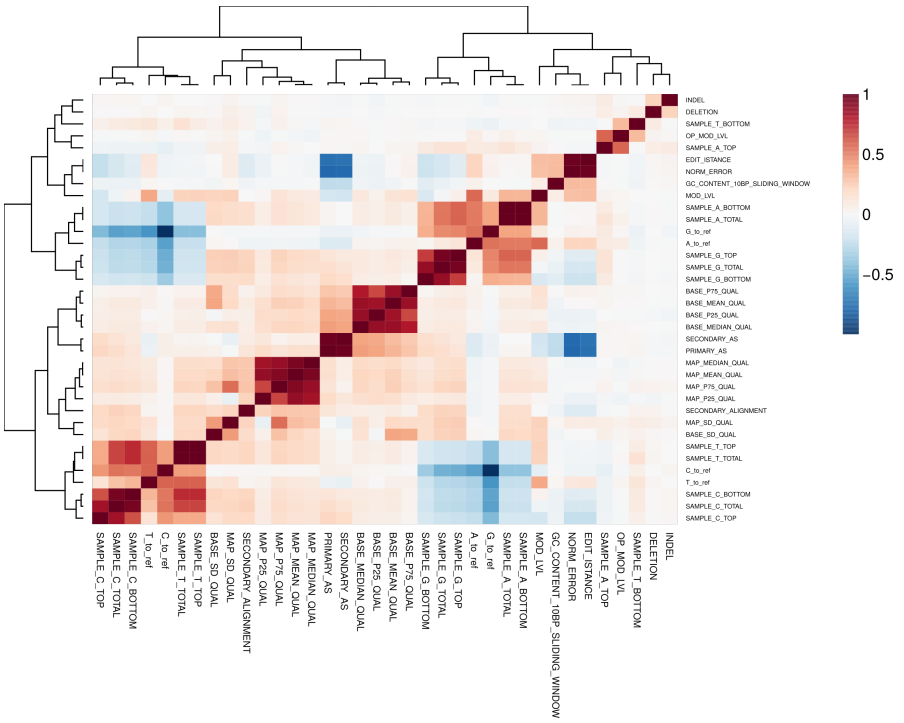


Figure 14: The Pearson correlation between features from the initial feature set was visualised. The CpG positions were labelled as modifications after SNPs discovered by the GATK HaplotypeCaller were removed from the list of positions. Only cytosine positions from chromosome 1 of patient 0 were used for the visualisation.

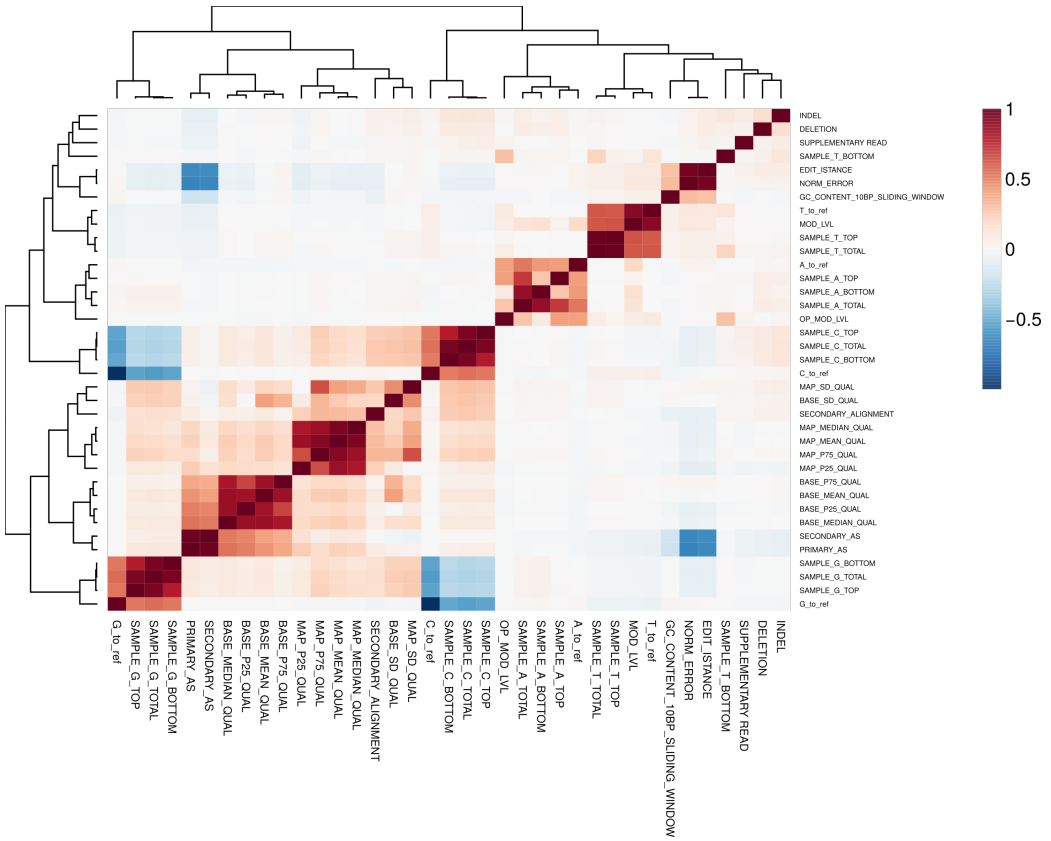
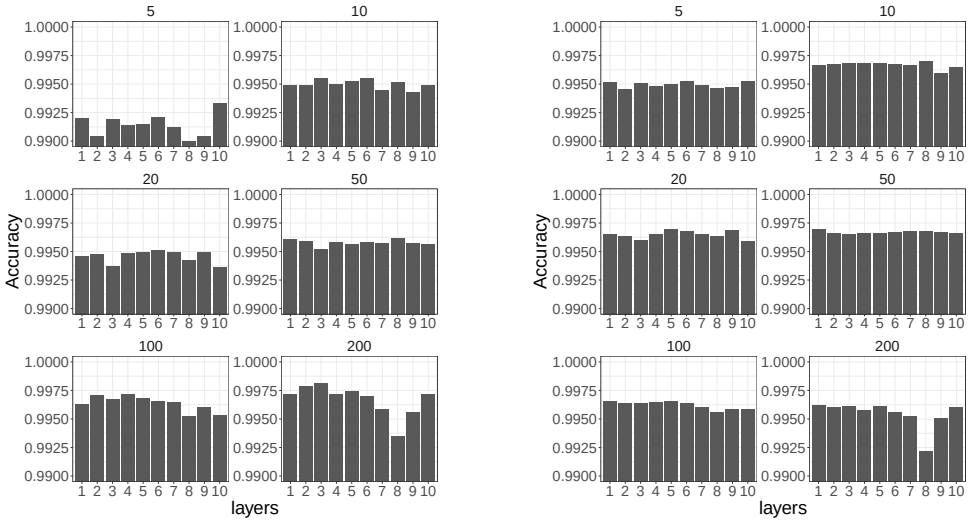


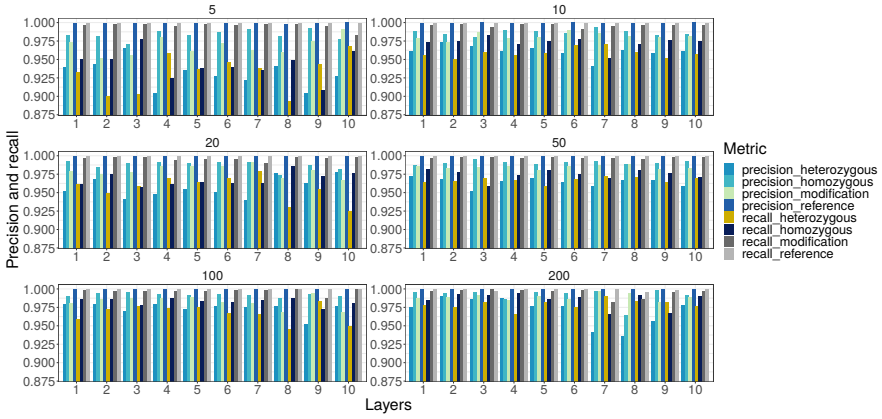
Figure 15: The Pearson correlation between features from the initial feature set was visualised. The non-CpG positions were labelled as modifications after SNPs discovered by the GATK HaplotypeCaller were removed from the list of positions. Only cytosine positions from chromosome 1 of patient 0 were used for the visualisation.



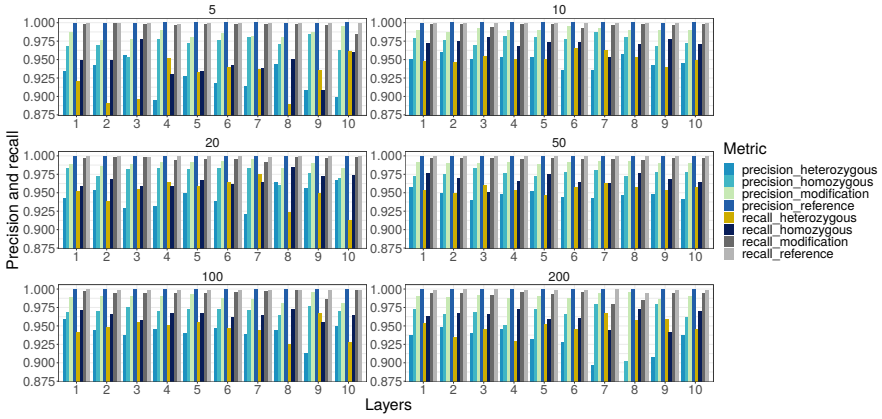
a) Training dataset

b) Testing dataset

Figure 16: Accuracy of the classifier depending on the number of layers and neurons in the NN model using TAPS data. The small number of neurons per layer 5-10 resulted in highly variable accuracy scores on the training set, while their performance was comparable with the cases where the neurons were 50 per layer on the testing set. Above 100 neurons per layers with $\geq$ seven layers, the results show variability which may be due to overfitting.

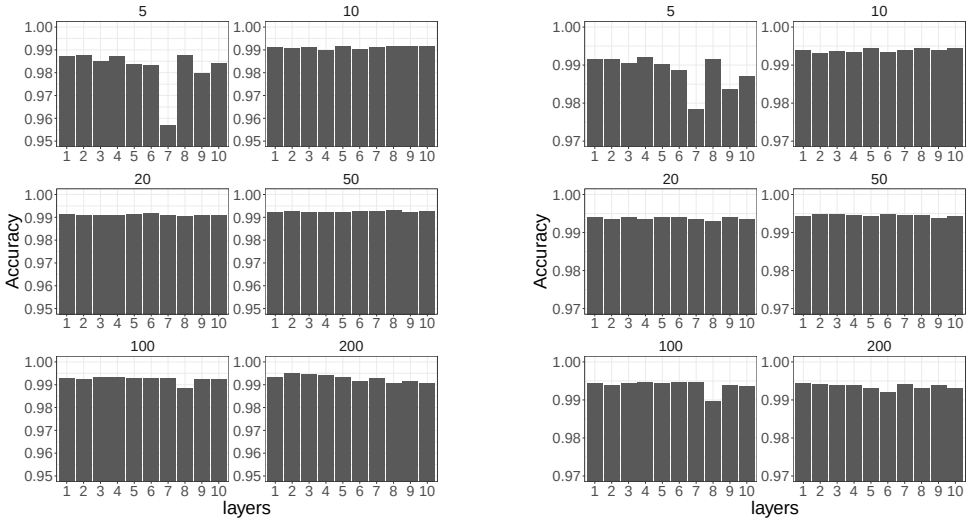


a) Training dataset



b) Testing dataset

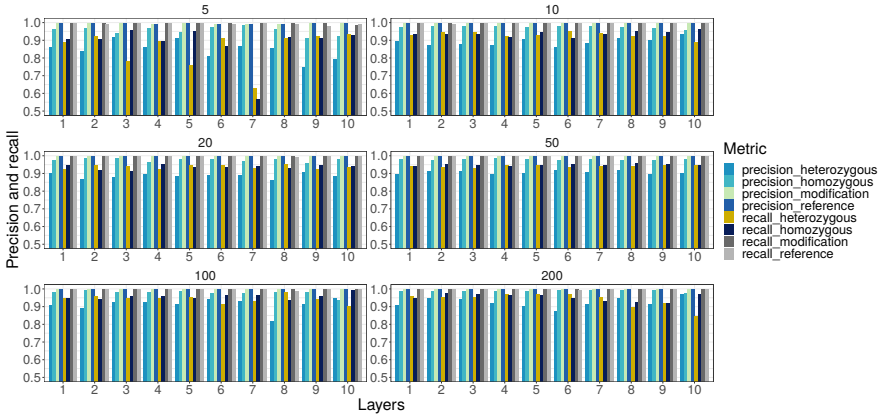
Figure 17: Precision and recall scores per class to measure the performance of the classifier depending on the number of layers and neurons in the NN model using TAPS data. Both the cases with five and over 100 neurons per layer appeared to give worse performance due to either underfitting or overfitting.



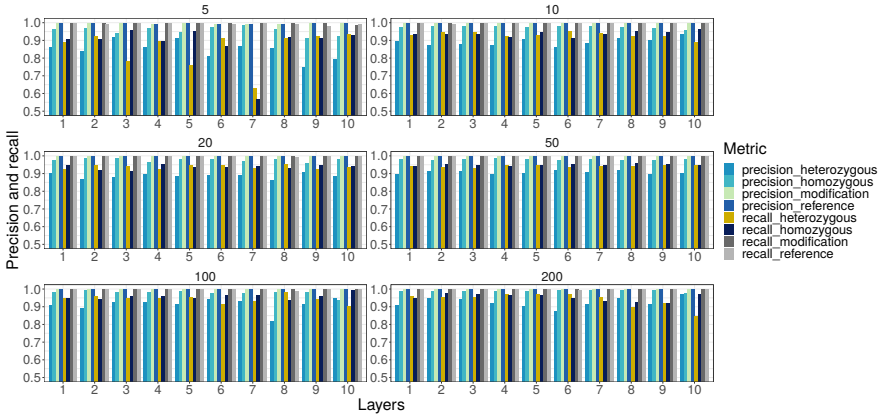
a) Training dataset

b) Testing dataset

Figure 18: Accuracy of the classifier depending on the number of layers and neurons in the NN model using WGBS data. The small number of neurons per layer 5-10 resulted in highly variable accuracy scores on the training set, while their performance was comparable with the cases where the neurons were 50 per layer on the testing set. Above 100 neurons per layers with $\geq$ seven layers, the results show variability which may be due to overfitting.

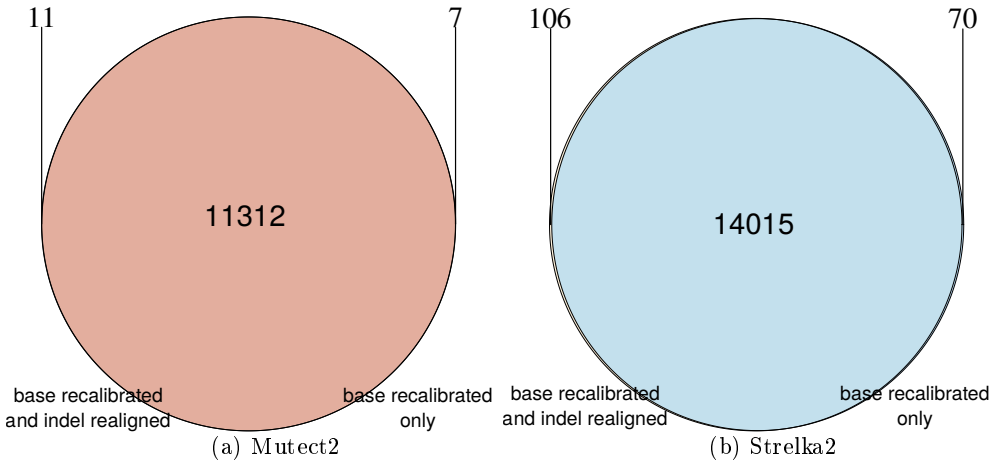


a) Training dataset



b) Testing dataset

Figure 19: Precision and recall scores per class to measure the performance of the classifier depending on the number of layers and neurons in the NN model using WGBS data. Both cases with five and over 100 neurons per layer appeared to give worse performance due to either underfitting or overfitting.



(a) Mutect2 (b) Strelka2

Figure 20: The dependence between WGS GATK pre-processing and the somatic variant calling results obtained using Strelka2 and Mutect2 were studied on the WGS matched duodenum and Barrett's oesophagus data of patient 0. There was a small difference in the number of variants passing quality filters found in the base recalibrated WGS data and the base recalibrated and indel dealigned data.

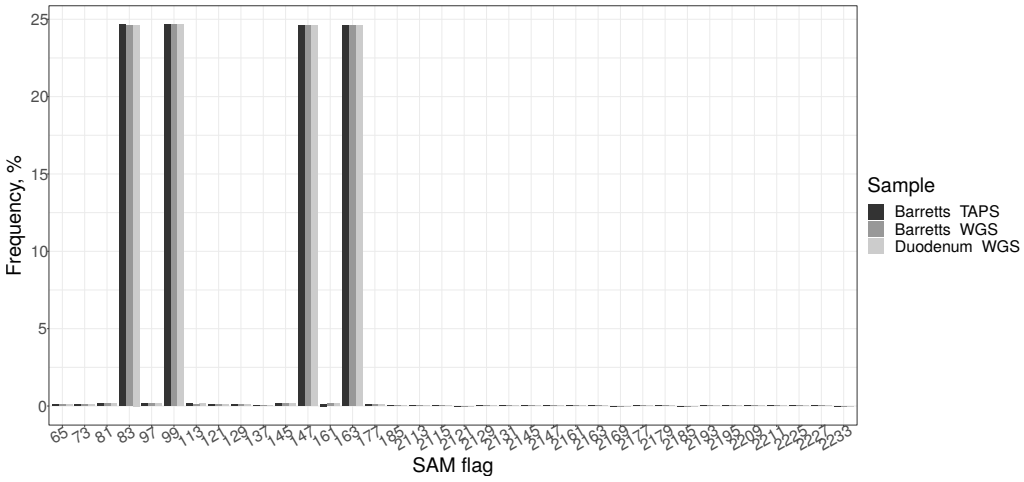


Figure 21: The flag distributions of all pair-end sequencing reads. The read flag distribution was extracted directly from the BAM file of one TAPS file. The majority of the reads were aligned as properly paired.

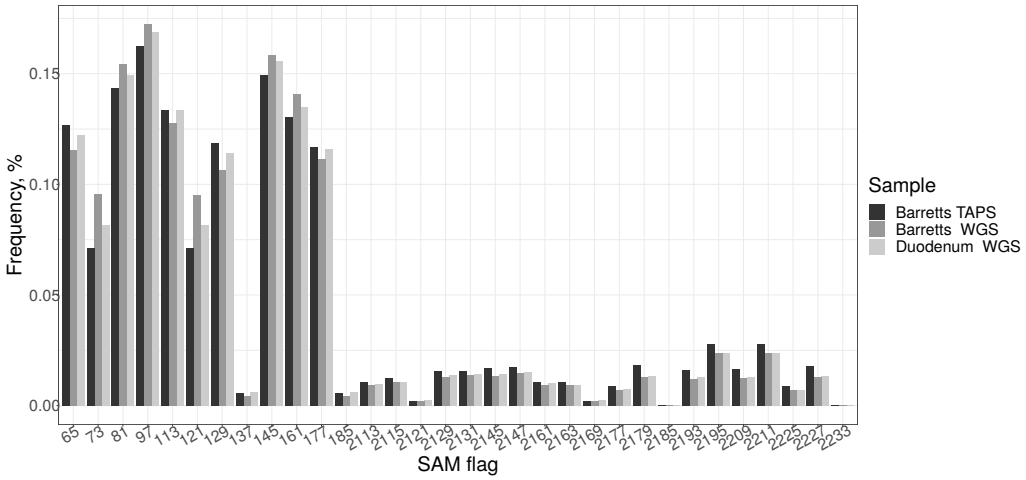


Figure 22: The flag distributions of all pair-end sequencing reads with a focus on read flags observed with lower frequencies. The read flag distribution was extracted directly from the BAM file of one TAPS file. The majority of the reads were aligned as properly paired. The figure shows flag distributions of not properly paired pair-end and supplementary sequencing reads.

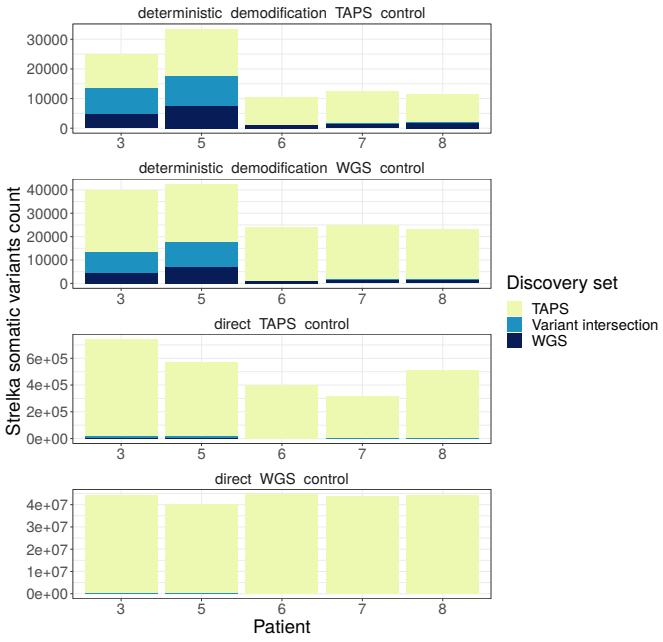


Figure 23: The somatic calls found by Strelka2 on TAPS-TAPS and TAPS-WGS matched samples after deterministic demodification (top two panels) and with direct variant calling (bottom two panels). As expected, the direct variant calling resulted in an inflation of false-positive calls which were fewer in the TAPS-TAPS case due to the more significant similarity of the tumour and control samples.

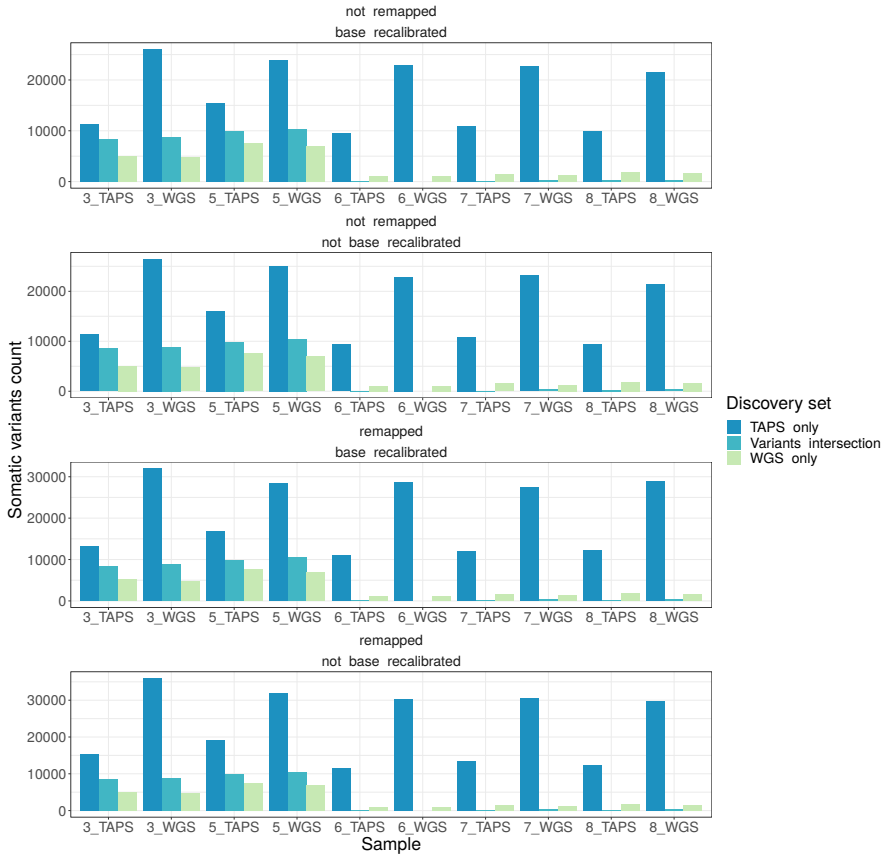


Figure 24: The somatic calls found by Strelka2 on TAPS-TAPS and TAPS-WGS matched samples following different types of pre-processing. Base recalibration and re-alignment of demodified reads introduced more errors in TAPS.

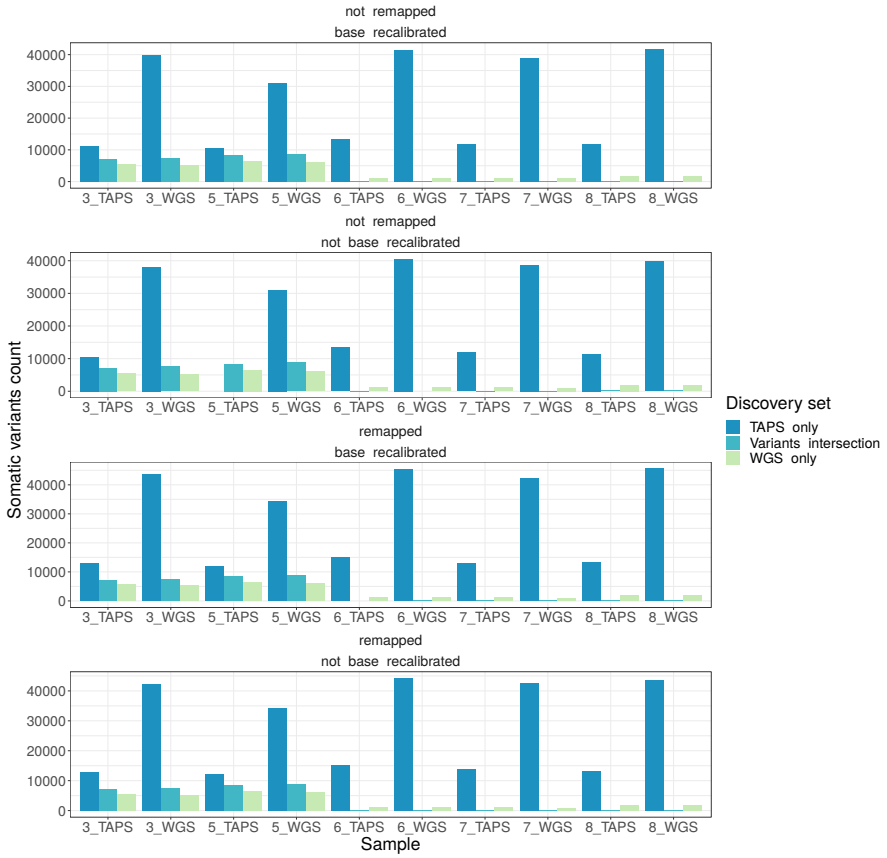


Figure 25: The somatic calls found by Mutect2 on TAPS-TAPS and TAPS-WGS matched samples following different types of pre-processing. Base recalibration and re-alignment of demodified reads introduced more errors in TAPS.

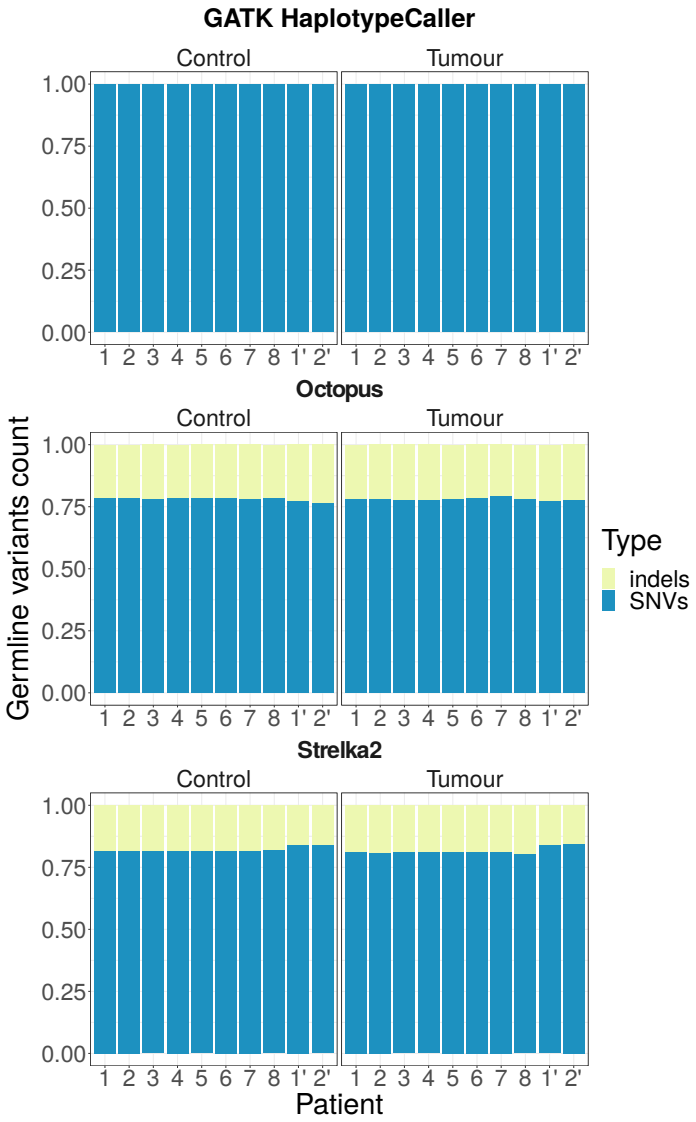
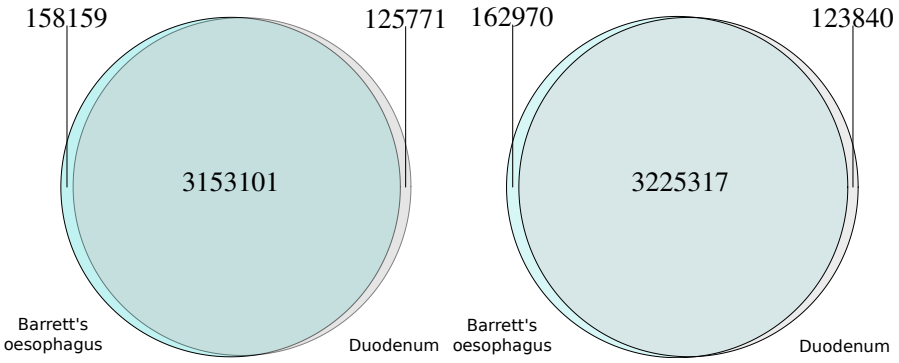


Figure 26: The proportion of indels to all germline variants differs between callers. Patients 1' and 2' are WGS samples matching WGBS while the remaining samples are WGS matching TAPS.



(a) A Venn diagram showing the agreement between the high quality variants discovered by GATK's HaplotypeCaller if only base recalibration is used.

(b) A Venn diagram showing the agreement between the high quality variants discovered by GATK's HaplotypeCaller if base recalibration and indel realignment is used.

Figure 27: There was a small difference in the number of quality variants found in the base recalibrated WGS data and the base recalibrated and indel dealigned data using GATK HaplotypeCaller.

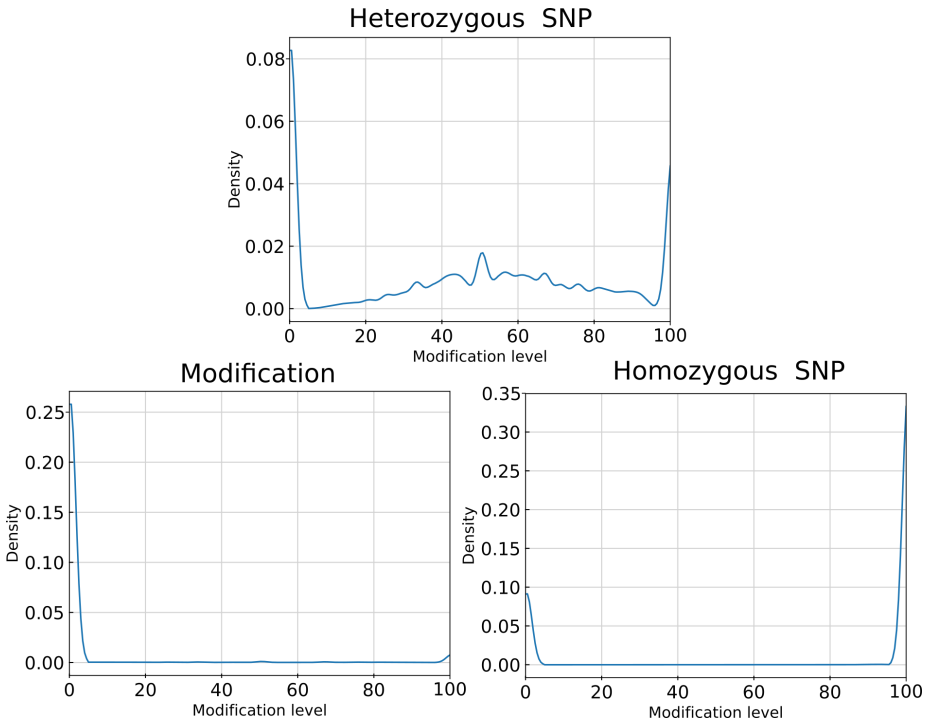


Figure 28: The observed β value ratios in modified and SNP cytosine positions were visualised. The distribution of the ratios of T to T+C on the OT strand and A to A+G on the OB strand per individual positions, were plotted for modifications, heterozygous and homozygous SNPs. Both modifications and homozygous SNPs showed a bimodal distribution, while the observed modification level for heterozygous SNPs had the highest densities at $\approx 0, 0.5$ and 1 β values.

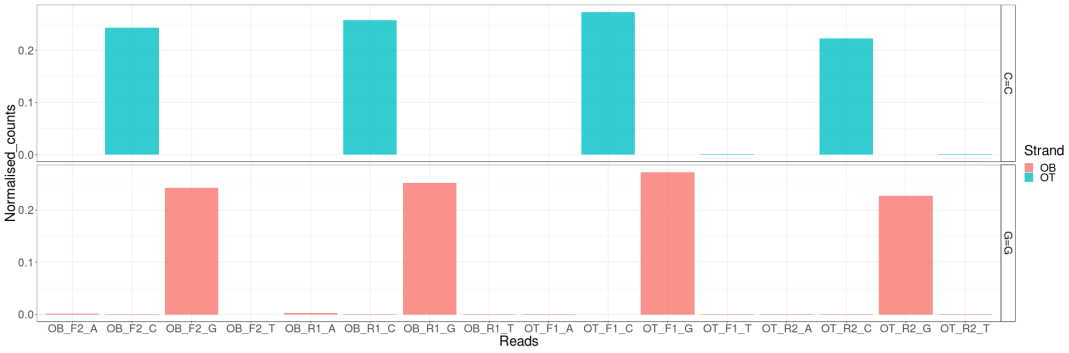


Figure 29: A visualisation of reference read distributions by a read flag and read-pair orientation. The read counts per base and read orientation were normalised by the total number of sequencing reads per position, and the mean values per each group of interest were visualised. Few non-reference reads were observed which were presumably a result of sequencing and alignment errors.

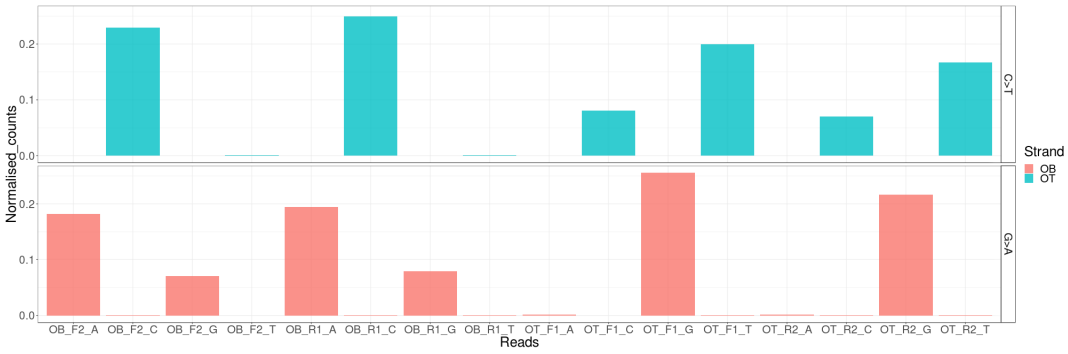


Figure 30: A visualisation of modification read distributions by a read flag and read-pair orientation at highly modified ($\beta \geq 0.8$) CpG positions. The read counts per base and read orientation were normalised by the total number of sequencing reads per position, and the mean values per each group of interest were visualised. Only C>T changes on the OT strand and G>A on the OB strand were observed in TAPS.

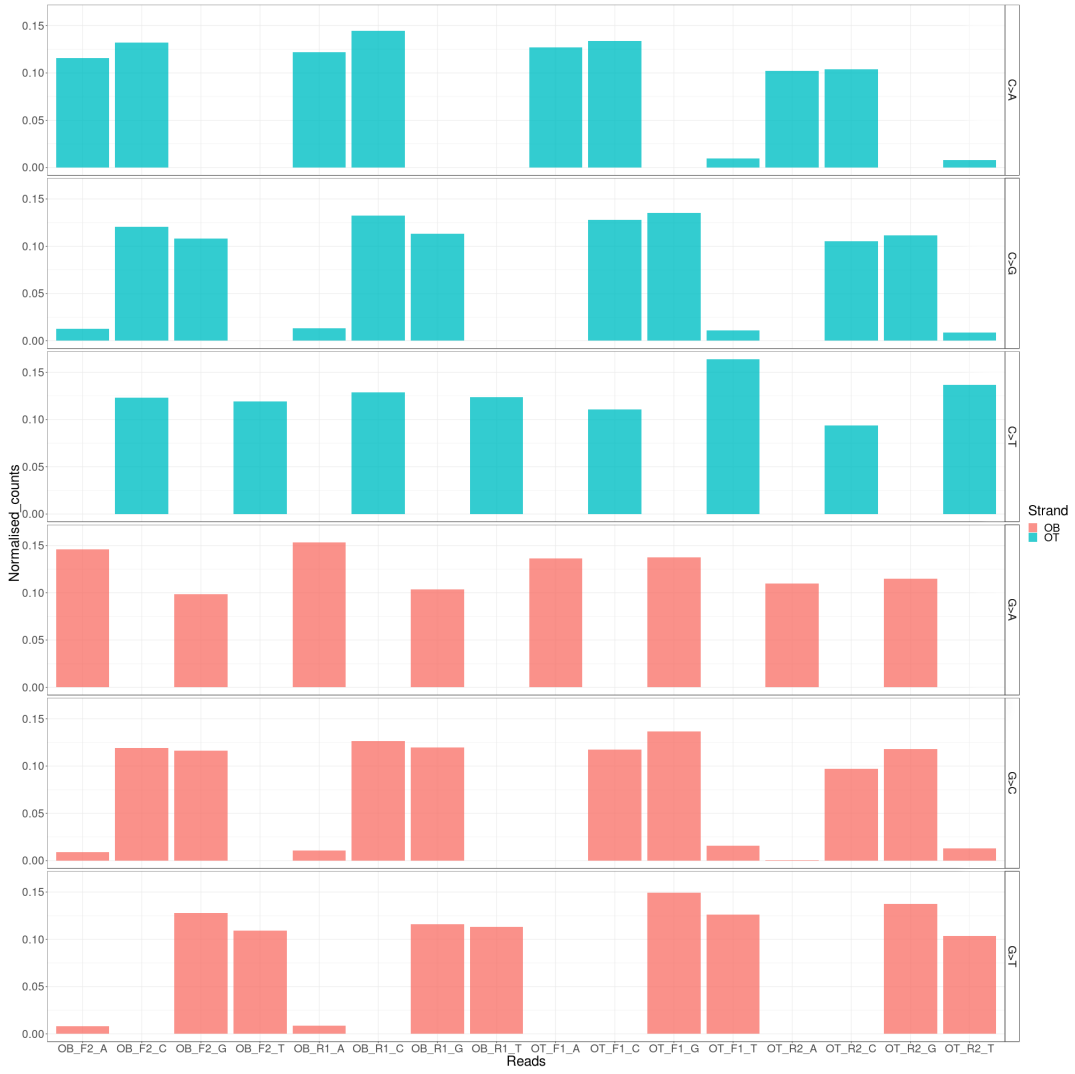


Figure 31: A visualisation of heterozygous SNPs read distributions by a read flag and read-pair orientation. The read counts per base and read orientation were normalised by the total number of sequencing reads per position, and the mean values per each group of interest were visualised. Specific read count patterns were observed depending on the type of change (C>A, C>T, C>G), and the strand (OT or OB).

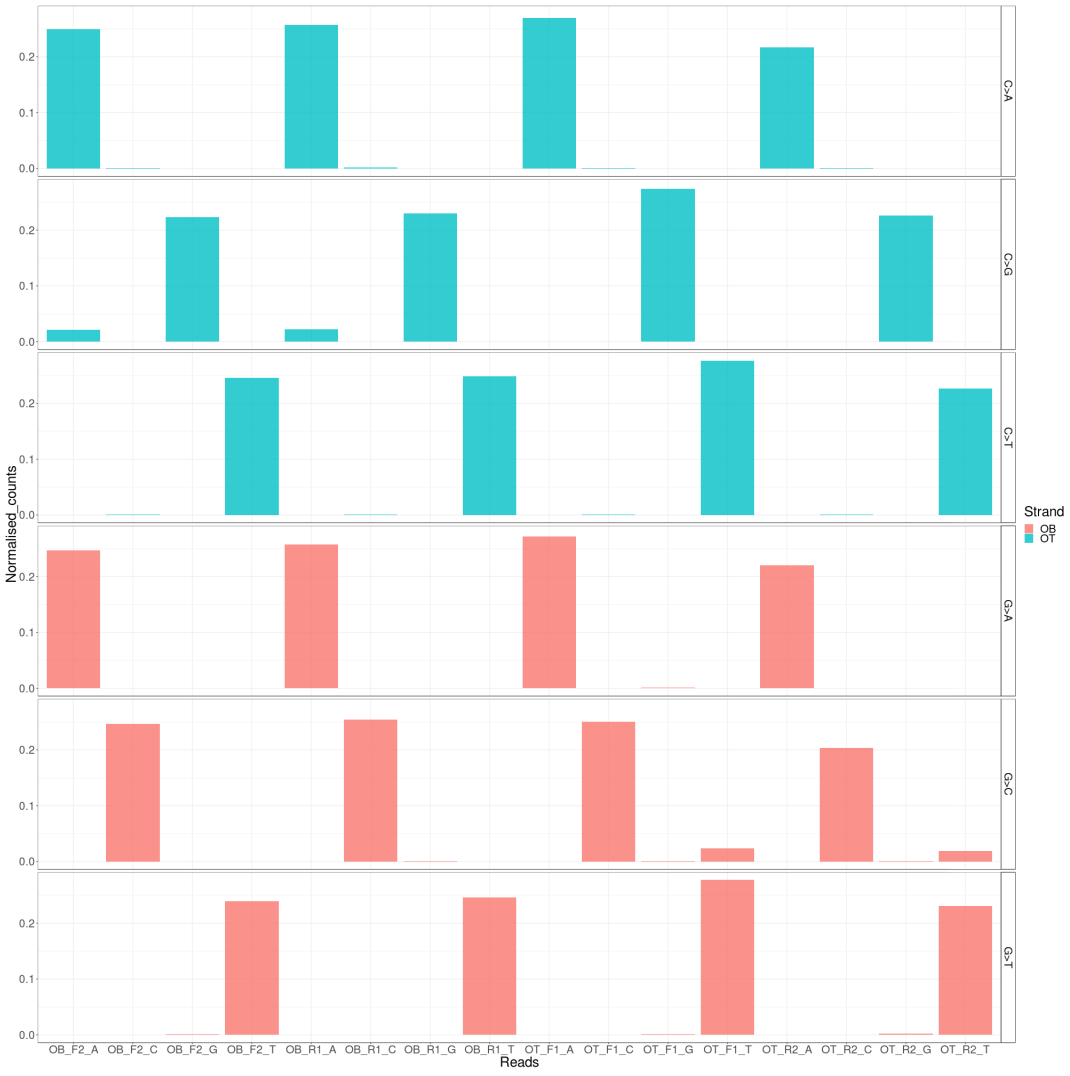


Figure 32: A visualisation of homozygous SNPs read distributions by a read flag and read-pair orientation. The read counts per base and read orientation were normalised by the total number of sequencing reads per position, and the mean values per each group of interest were visualised. Specific read count patterns were observed depending on the type of change (C>A, C>T, C>G), and the strand (OT or OB). The homozygous SNP read distributions were distinct from the reference, modified, and heterozygous SNP positions.

III cfDNA projects

The initial work for this thesis was focused on analysing data from cell-free DNA (cfDNA) epigenomic methods. The technologies used to generate the data were developed by Dr Chun-Xiao Song and Paulina Siejka-Zielińska. The 5-hmC-Seal (hMe-Seal) and 5-mC-Seal (Me-Seal) methods enable sensitive enrichment of DNA fragments that contain DNA modifications of interest [318]. These methods were applied to cfDNA samples due to their ability to preserve epigenomic information from a low input sample.

I Feature exploration studies

The application of cfDNA epigenomic methods to the clinic for early diagnostics of cancers requires the discovery of features (markers) that distinguish between groups of medical interest, such as patients versus healthy people. Thus, the first part of the project was to explore differential regions, coverage, fragment lengths, CNVs, and the microbiome and their usability as features for classification. The data used was published in [320] (the 5-hmC-Seal was also applied to cancer cfDNA samples in [204]). The general data analysis steps used for the cfDNA enrichment data analyses followed the methods outlined in [320].

The read counts were normalised for each sample depth using fragments per kilobase of gene per million mapped reads (FPKM) normalisation. Following the normalisation step, they were counted in genomic bins of 2 kb, 5 kb, and 10 kb, as well as over genomic regions such as gene bodies and CGIs. The enrichment methods did not provide information about the modification state of the individual CpGs. However, the relative abundance of reads spanning regions was indicative of 5-hmC or 5-mC at a position. By applying clustering

Part of the results of this section was included in the Transfer of Status report submitted in Michaelmas 2017.

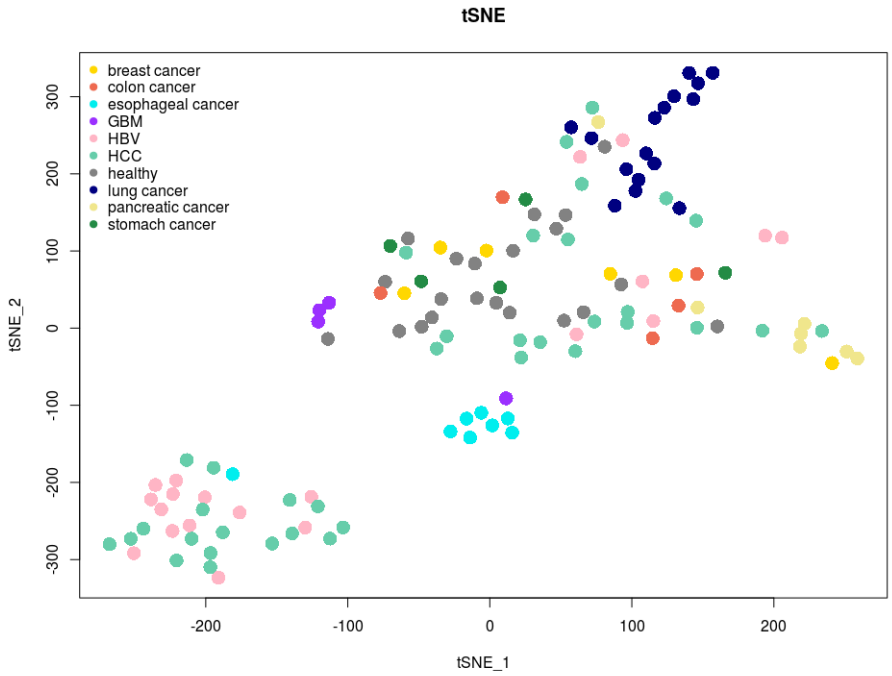


Figure 33: The figure presents a visualisation of 5-hmC enrichment at gene bodies using tSNE. The read counts per gene body for each of the samples were normalised to the sample sequencing coverage, and the gene length using FPKM normalisation. Certain cancer types (lung, oesophagus, pancreatic, and HCC together with the pre-malignant HBV) appear to separate from the healthy samples and some of the other malignancies. These results suggested that 5-hmC read counts might be used to distinguish between groups of clinical interest. Legend: GBM stands for glioblastoma, HBV for hepatitis B, and HCC for hepatocellular carcinoma.

and dimensionality reduction approaches (principal component analysis, PCA, and t-distributed stochastic neighbour embedding, tSNE), certain types of cancers and pre-malignant conditions showed higher separation from the healthy samples (see Figure 33). However, a significant constraint of this explorative study is the limited number of samples for some of the groups. Also, the healthy category includes patients with other diseases such as diabetes but not cancer. Thus, a greater separability can be observed when cancer samples are compared to clinically healthy samples as other illnesses may trigger the immune response

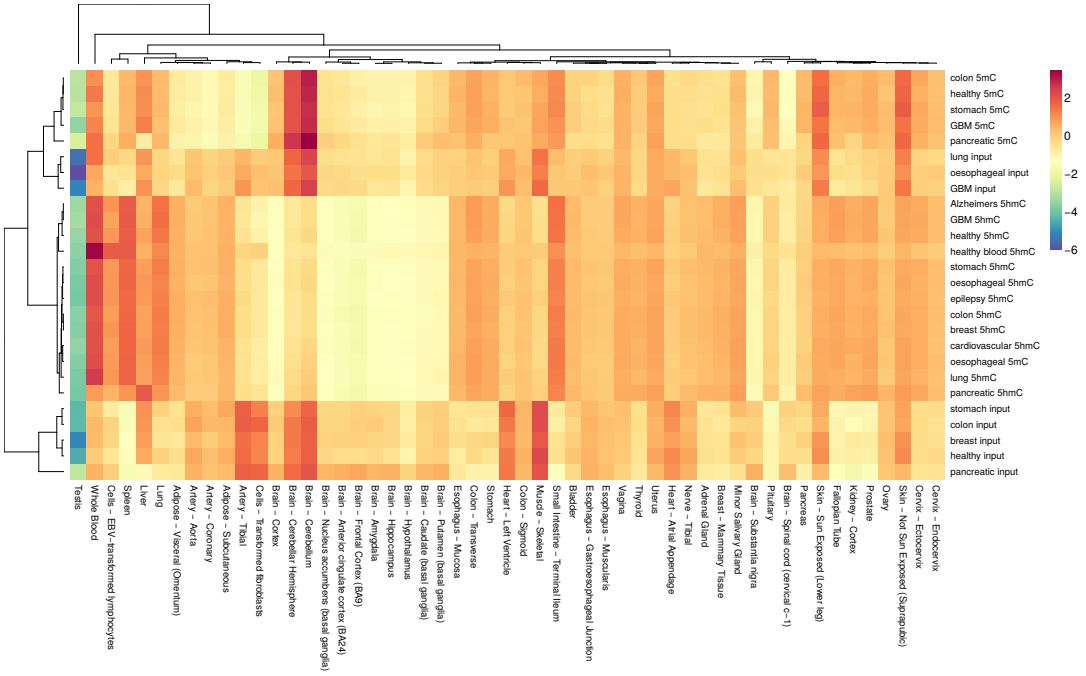


Figure 34: Spearman correlation of cfDNA normalised counts on coding genes, pseudogenes and genes coding for lncRNA with median RPKM values for the GTEx v6p study. CfDNA values are centred and standardised to row Z-scores.

and change 5-mC and 5-hmC patterns.

The next step of the analysis included a comparison of the available cfDNA samples with tissue-specific expression and open chromatin data in order to explore possible tissue contributions. To this end, the median reads per kilobase per million mapped reads (RPKM) values for the GTEx v6p study [57] were correlated to the corresponding median cfDNA FPKM values (edgeR package [287], autosomes only) (see Figure 34). A high correlation was observed between whole blood and 5-hmC enriched cfDNA after centring and Z-score standardisation, suggesting that the majority of cfDNA fragments were released by blood cells. Also, correlations of normalised read counts from the available samples within CGIs (gene expression) and DNAse I accessibility regions (open chromatin from Roadmap Epigenomics [40]) were studied. The use of CGI instead

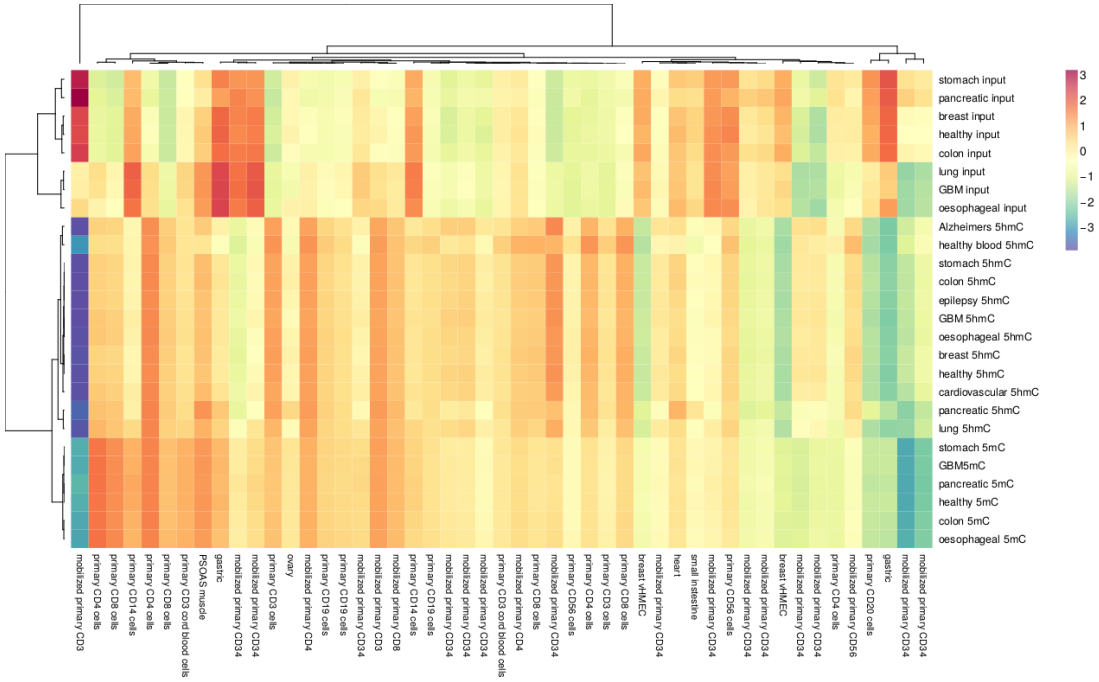


Figure 35: Spearman correlation of cfDNA and DNAse I accessibility from Roadmap Epigenomics for adult tissue types and cells. CfDNA values were centred and standardised to row Z-scores.

of genes did not improve separation between sample types. In contrast, a correlation between oesophageal, stomach, pancreatic, and colorectal cancer input samples and DNAse-seq gastric tissues was observed using 10 kb normalised for sequencing depth read counts (autosomes only, normalised using the DESeq2 package [217]). There was an observable association between the majority of the immune cells and 5-hmC enrichment samples, suggesting the contribution of immune cells with an observable difference in the level of association between the input and the enrichment samples for cytotoxic T-cell CD3 (see Figure 35).

Furthermore, the exploration studies over 5-hmC sequencing data included alignment of sequencing reads that did not align to the human genome to a complete bacterial and viral genomes database to study any possible contributions of the microbiome. The tool used to map DNA fragments to the viral and

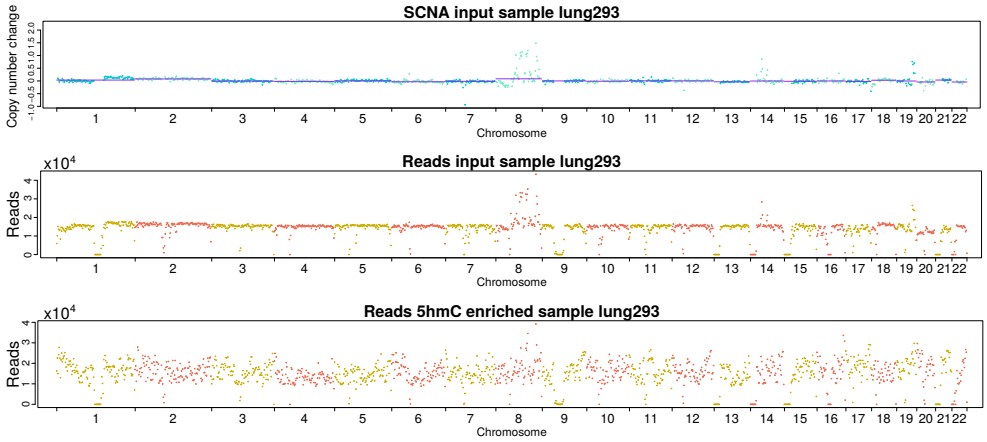


Figure 36: The figure shows a visualisation of low depth WGS ($\approx$ 1x sequencing coverage labelled as input) and 5-hmC enrichment data from cfDNA from a lung cancer patient. A 2 Mb CNV is observed in chromosome 8 and is detected by both HMMcopy and by raw read counts in the input cfDNA (top to panels). Some signal is observed in the 5-hmC enrichment data (bottom panel), but without further normalisation, the signal to noise appears higher than in the input WGS sample.

bacterial database was Kraken [368]. The results showed few reads aligning to the database which might be a result of size selection during library preparation, where bacterial and viral genomes (and mitochondrial DNA) are expected to be shorter than genomic nucleosome bound DNA sequences, and naturally observed lower abundance of naked DNA (DNA lacking histones) fragments, as they are more readily degraded by nucleases in the blood [53]. Another direction of the exploratory study was to look at larger copy number variation (at a 1-2 Mb scale). This bin size was selected due to the sparse nature of the cfDNA enrichment data (mean coverage $\leq$ 1x) where reads are observed as genomic *peaks* (regions with higher sequencing coverage than the mean) which by themselves can be mistaken as a smaller scale CNVs. By using a CNV caller suitable for shallow sequencing WGS data (HMMcopy [194]), large CNVs were detected in two lung cancer samples—lung293 and lung417 (see Figure 36).

The 5-hmC and 5-mC Seal methods were applied to a limited number of

oesophageal cancer patients in order to compare how well read counts over genomic regions can separate patients from healthy volunteers. There was a separation between healthy and oesophageal cancer patients, but also a patient-specific separation of samples was visible if dimensionality reduction techniques were applied.

The exploratory studies suggested that the signal available in enrichment cfDNA could be applied for the classification of malignant conditions. At the same time, more work is needed in order to compare epigenomic data to DNA expression and methylation maps in order to study tissue contribution.

II Classification of HCC

The aim of the feature exploration study was to discover genomic regions and characteristics of epigenomic cfDNA data that would enable the correct classification between groups of clinical interest. The 5-hmC Seal method was applied on hepatocellular carcinoma (HCC) samples together with samples with patients with hepatitis B (HBV) in order to study disease progression and to test whether the method could be applied for diagnostics. Hepatocellular carcinoma is the fifth most common type of cancer worldwide, with 1.6 million deaths per year [176]. The main factors for the development of the disease included infection with HBV, hepatitis C (HCV) and liver damage resulting from alcohol consumption and obesity. At present, the test used for the detection of HCC is the serum alfa-fetoprotein (AFP) which, when used as a diagnostic marker for HCC suffers from limited sensitivity and specificity. One reason for that is that there are two distinct subtypes of HCC that differ in size and the production of AFP.

In order to study whether 5-hmC counts within genomic regions could distinguish between HCC and hepatitis patients, 57 (pre-treatment and 77 in total)

HCC and 40 HBV patient samples were used for the analysis. The data analysis was focused on finding different features between patients with hepatocellular carcinoma (HCC) and hepatitis type B (HBV) carriers, in order to build a classifier and compare its performance against commonly used protein marker AFP. A strong batch effect was observed between sample batches (see Figure 37), which might be a direct result from DNA degradation of cfDNA (see Figures 38, 39). In order to resolve the batch effect, different techniques were applied such as using a coverage cutoff ($\geq$ 4.5x10^6 unique, non-blacklisted fragments), downsampling, removal of genomic regions with mappability in hg19 of less than 0.8 over a sliding window with a bin size of 100 bp. The batch effects were still observable, which lead to the sequencing of new HBV samples due to the initial unbalanced design where the majority of the HBV samples were in the same batch. A further comparison looked at different fragment size distributions and whether 5-hmC counts of fragments falling in each category could improve the separation. Also, a comparison to healthy samples was drawn to ensure there was enough detectable signal to classify HCC. Given the differences in the fragment size distributions between batches, we decided to take only fragments from the categories: 1) under 100 bp, 2) 140-155 bp, 3) 155-185 bp, 4) 260-360 bp, 5) 1000-2000 bp in order to capture more intra-type similarities reducing the technical noise. These fragment sizes were selected to include mitochondrial, microbial and degraded DNA fragments (under 100 bp), nucleosomal wrapped DNA (140-360 bp), and DNA fragments resulting from necrosis that is expected to occur more often in tumour compared to healthy cells [166], [158]. Nevertheless, segment separation by itself was insufficient to separate the HBV, HCC, and healthy samples. These results suggested that further analysis would be required before 5-hmC enrichment could be ruled out as a non-invasive diagnostic for HCC detection and prognostication.

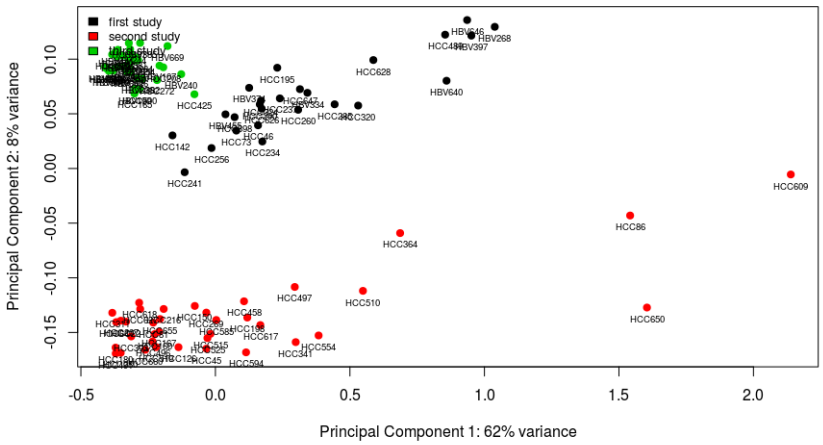


Figure 37: The figure presents a visualisation of 5-hmC enrichment at 2 kb bins using PCA. The data were adapter and quality trimmed with Trim Galore!, keeping fragments with length $\geq$ 35 bp. Then blacklisted regions for hg19 were removed and the read counts per bin for each of the samples were normalised to the sample sequencing coverage, and the bin length using FPKM normalisation.

IV Enrichment-bisulfite method development project

The enrichment-bisulfite method based on 5-hmC and 5-mC Seal was a new method for studying the most frequent cytosine modifications in the mammalian genome—5-hmC and 5-mC. It combined chemical labelling of the modified bases and the enrichment of the DNA fragments incorporating them, followed by TET oxidation (not on glucose protected cytosines) and bisulfite treatment, which is then followed by PCR amplification and Illumina sequencing. Thus, this novel method was expected to provide a cost-effective option to study DNA modifications at base resolution. The method was developed and tested on sonicated genomic DNA extracted from mouse E14 mES cells before it was abandoned with the development of TAPS.

The enrichment-bisulfite data (from two runs) was analysed following a stan-

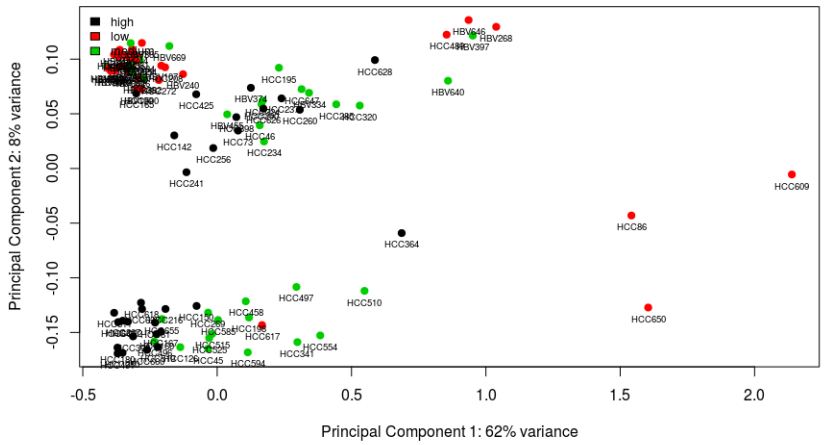


Figure 38: The figure presents a visualisation of 5-hmC enrichment at 2 kb bins using PCA. The data were adapter and quality trimmed with Trim Galore!, keeping fragments with length $\geq$ 35 bp. Then blacklisted regions for hg19 were removed and the read counts per bin for each of the samples were normalised to the sample sequencing coverage, and the bin length using FPKM normalisation. The colouring shown represents the batch of the individual sample.

dard bisulfite sequencing data analysis pipeline, where reads were aligned to the reference genome using Bismark, and the same tool was used for modification calling on CpG positions after the removal of PCR duplicates. The public data used for comparison to the enrichment-bisulfite method included MethylC-seq (accession number GSM1027572), TAB-seq (accession number GSE36173), hMeDIP and MeDIP (accession number PRJEB2462) datasets. Also, previous input and 5-hmC Seal samples (accession number GSE41545) were downloaded and used in the analysis. TAB-seq and MethylC-seq samples were processed in the same way as the enrichment-bisulfite samples, while the hMeDIP, MeDIP, input, and 5-hmC Seal samples were processed using the cfDNA enrichment pipeline (aligned with Bowtie2, blacklisted and deduplicated with per bin extraction of read counts). The false-positive rate of the bisulfite conversion was calculated (the number or higher of modified cytosines occurring by chance) as

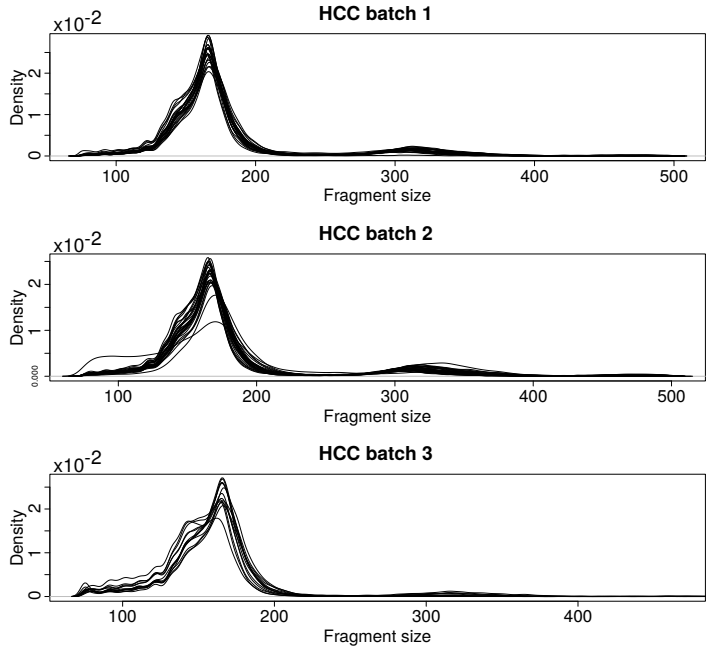


Figure 39: The figure presents the insert size distribution of the HCC 5-hmC enrichment samples used for the study. Some DNA degradation was observed as smoothed density profiles and the lack of a pronounced peak at 320-360 bp in the samples from the third batch.

pbinom(T, C+T, 1-non-conversion rate, lower.tail=T), while the false-negative rate was calculated (the number or higher of unmodified cytosines occurring by chance) as pbinom(C, C+T, protection rate, lower.tail=T).

In order to evaluate the success of the enrichment, oxidation and the bisulfite steps, different groups of spike-ins were used:

- short synthetic sequences with three 5-hmC modified cytosines on one strand (97-mer and 100-mer named after their length in bases);
- eight bacmid-based (a transfer vector used to transfect insect cells) spike-ins to examine the affinity of the method (their names indicate the number of consecutive CpG positions—1CG, 2CG, 5CG, 10CG, and 19CG);
- pUC19 and pUC19 300-mer (a common transfer vector used in *E.coli*)

were utilised to study the protection rate in the enrichment-TAB samples, as all cytosines were 5-hmC in these spike-ins;

- a CpG methylated lambda genome (from the M.Sssl methyltransferase) was employed to assess the non-conversion rate (success of bisulfite conversion), and enrichment-BS protection rate (protection of 5-hmC and 5-mC from the TET enzyme).

There were issues observed with the spike-ins after the first enrichment-bisulfite run: 1) no reads were mapped to bacmid 10CG and 5-hmC; 2) there were very few if any reads mapped to the 97-mer and the 100-mer. After finding out that no reads were aligned to 5-hmC spike-ins in the enrichment-TAB samples with Bismark, but 23 000 of them were detected in the enrichment-only samples, we tried using another aligner. The wildcard mapper BSMAP was able to map all of the 5-hmC spike-in reads, which were labelled as secondary by Bismark. To evaluate the possible primary location of these reads, we used MUMer [191] and assessed the sequence similarity between the spike-ins. We revealed that pUC-19 was identical in reverse to 10CG and 5-hmC (Figure 40). Hence, pUC-19 and pUC19 300-mer were removed from the reference genome in the second trial. Another major issue appeared to be with the lambda control, which showed on the enrichment and regular bisulfite CpG methylation of less than 20% when it was expected to be fully methylated. The cause of this error was not found, but an off-target enzymatic reaction during library preparation was suspected. The 97-mer and 100-mer were discriminated against during the size selection step in the first trial but were extended to 120 bp for the second trial. After testing for the sequence similarity of the spike-ins, and considering the size-selection step, we learned more about the spike-ins controls of the experiment, and how their detection can be improved. Also, the second trial reached a lower non-conversion rate of 1% indicating a complete bisul-

fite conversion, compared to the relatively high 4.5-4.7% from the first trial. The protection rate (glucose protection from the TET enzyme) of around 80% from the lambda converted genome of the enrichment-BS samples was in line with the expected values. These measurements are vital because a low bisulfite non-conversion rate and a high TET protection rate enable us to distinguish significant cytosine modifications while diminishing technical noise.

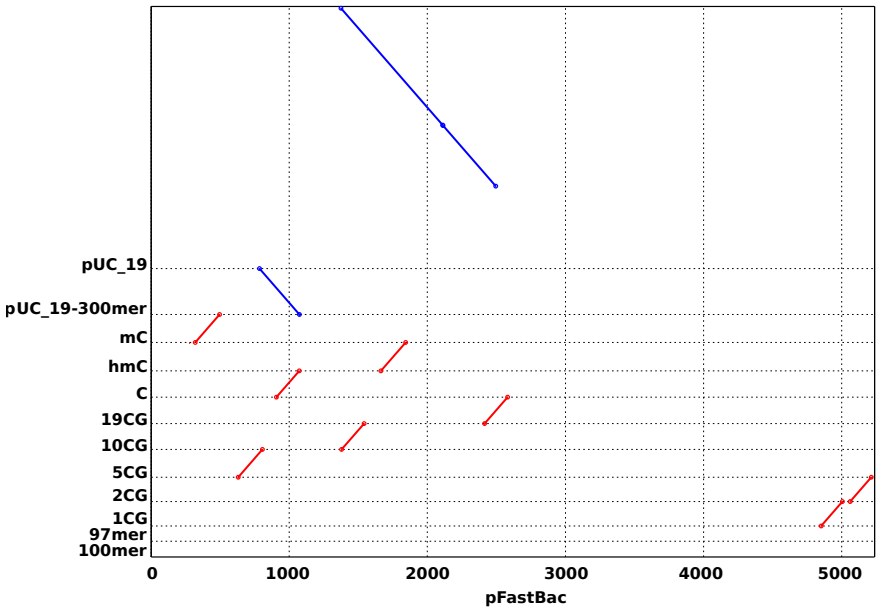


Figure 40: The sequence similarity between the short spike-ins explained why some spike-ins were not aligned using Bismark. The similarity plot was output by MUMer. The pUC-19 spike-ins were identical in reverse to 5-hmC and 10CG, which were not aligned to in the first trial. Only the short spike-ins were visualised for simplification.

For the comparison with existing methods, TAB-seq (WGTABS) and methylC-seq (WGBS) data were used. We expected to see a similarity between the input, WGTABS and WGBS coverage profiles, and consistency between the coverage and the CpG methylation profiles for the enrichment-only and the enrichment bisulfite samples.

The modification counts distributions vary considerably between the whole-

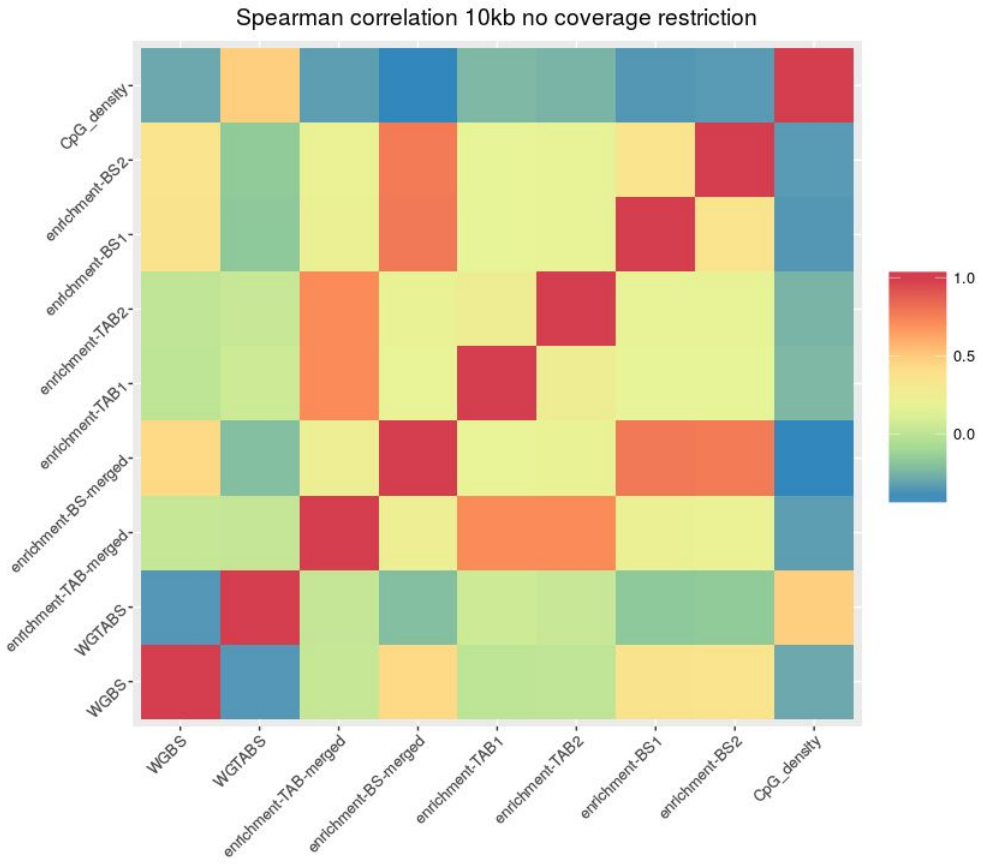


Figure 41: The Spearman correlation of CpG modification values averaged over 10 kb windows. Only genomic regions which were covered by all samples were kept for the analysis. Additionally, the averaged CpG density per 10 kb bin is given.

genome and the enrichment based methods, suggesting that the new sequencing methods are cost-effective, capturing a higher proportion of modified cytosines at a lower sequencing depth. Next, we were interested in what is the association of the modified to unmodified cytosines (β values) of the whole-genome and the enrichment-bisulfite methods. We used Spearman correlation between the modification values mapped to 10 kb windows per sample. As a result, we observed that the WGTABS method was more similar to the enrichment-TAB, analogical

observation was made for the WGBS and the enrichment-BS (see Figure 41). Moreover, the CpG density (CpG number per 10 kb) seemed to be associated positively only with the modification values of the WGTABS, implying that WGTABS may lack the sensitivity to capture signals from low CpG content, which may be biologically relevant. After gaining some understanding of the data from the enrichment-bisulfite study, our next step was to look for possible ways to transform the two methods, rendering them comparable at base resolution. In order to juxtapose the modification score distributions between whole-genome and the enrichment-bisulfite method, we used binomial filtering without multiple testing correction, taking into account the highly context-dependent coverage of the enrichment-bisulfite data. We decided on a significance cutoff of 0.1 p-values for the binomial probability of modified cytosines to all cytosines at a CpG, as we took into account the high non-conversion rate of the enrichment-bisulfite methods in the first trial. This filtering step enabled us to observe a large overlap between the WGBS and the enrichment-BS, which was supposed to increase with sequencing depth. However, in the case of WGTABS and the enrichment-TAB, the convergence was smaller, which may be due to the higher sensitivity of the new enrichment-TAB method to capture 5-hmC. The number of the modified CpGs after the binomial filtering may corroborate this hypothesis, showing $8.7 \text{x} 10^6$ modified CpGs in the merged enrichment-TAB sample compared to $2.5 \text{x} 10^6$ in WGTABS.

V Publications and contributions

Here is a summary of publications, conference presentations and other participation that included my work during the course of the DPhil. Publications and presentations related to the thesis:

Using neural networks for discovery of DNA modifications and genetic variants in DNA sequencing data, 20th Genome Informatics (virtual conference), 2020 (poster presentation and a lightning talk)

Payne, K., Velikova, G., Song, C., Nankivell, P., Mehanna, H., High-throughput methylation profiling of cell-free plasma DNA in head and neck cancer: a pilot study, British Journal of Oral and Maxillofacial Surgery 57 (10), e101, 2019, https://doi.org/10.1016/j.bjoms.2019.10.292

Liu, Y., Siejka-Zielińska, P., Velikova, G. et al. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. Nat Biotechnol 37, 424–429 (2019). https://doi.org/10.1038/s41587-019-0041-2

Development of a Computational Tool for the Analysis of Bisulfite-Free DNA Sequencing Data, 3rd Epigenetics and Bioengineering Conference, Spain, 2019 (oral presentation, selected speaker)

Development of a computational tool for the analysis of bisulfite-free DNA sequencing data, Kellogg College Graduate Research Symposium, 2019 (oral presentation)

A new bisulfite-free method to detect cytosine modifications at base resolution, DPhil day, 2018 (poster presentation)

Selective chemical labeling approach for genome-wide profiling of 5-methylcytosine in circulating cell-free DNA, Ludwig Institute 10 years anniversary, 2017 (poster presentation)

Publications not related to the thesis:

Tihomir Ivanov, Neli Dimitrova, Gergana Velikova, Mathematical models with non-linear growth rates in population dynamics, Biomath Communications, 2018

Tihomir Ivanov, Gergana Velikova, Data fitting in Monod-type models with nonlinear growth rates, Biomath Communications, 2018 http://dx.doi.org/10.11145/bmc.2018.04.187

Lloyd Hamilton, Katy R. Astell, Gergana Velikova, and Dirk Sieger, Zebrafish, 2016. http://doi.org/10.1089/zeb.2016.1339

Other:

asTair - a bioinformatic tool for modification calling and analysis in TAPS and WGBS data https://bitbucket.org/bsblabludwig/astair

participation in Merck's Innovation cup 2020 (virtual event), Team Digitalisation, with a project for laboratory protocol recording and optimisation with ML

participation in L.E.V8 WOMEN program, Oxford Foundry, UK, 2019

Bibliography

[1] A Brief History of Machine Learning.
[2] Ammonia borane | BH3N - PubChem.
[3] Base Quality Score Recalibration (BQSR) – GATK.
[4] Bowtie 2: Manual.
[5] Conda — Conda documentation.
[6] Diagram Software and Flowchart Maker.
[7] Draw Freely | Inkscape.
[8] Manual reference pages - bwa.
[9] NumPy.
[10] Pyridine borane | C5H5BN - PubChem.
[11] Tesla P100 Data Center Accelerator | NVIDIA.
[12] Vector Norm – from Wolfram MathWorld.
[13] Welcome to imbalanced-learn documentation! — imbalanced-learn 0.5.0 documentation.
[14] zhou-lab/biscuit: BISulfite-seq CUI Toolkit.
[15] Bismark Bisulfite Mapper-User Guide-v0.15.0. Tech. rep., 2016.
[16] Illumina Whole-genome Bisulfite Sequencing on the HiSeq 3000/HiSeq 4000 Systems. Tech. rep., 2016.
[17] Sequence Alignment/Map Format Specification. Tech. rep., 2020.
[18] D. P. A., D. R., et al. Relationship between 5-fluorouracil disposition, toxicity and dihydropyrimidine dehydrogenase activity in cancer patients. *Annals of Oncology*, 2001. ISSN 0923-7534.
[19] M. Abdallah, A. Mahgoub, et al. Athena: Automated Tuning of k-mer based Genomic Error Correction Algorithms using Language Models. *Scientific Reports*, 2019. ISSN 20452322.
[20] H. J. Abel and E. J. Duncavage. Detection of structural DNA variation from next generation sequencing data: A review of informatic approaches, 2013.
[21] T. Äijö, Y. Huang, et al. A probabilistic generative model for quantification of DNA modifications enables analysis of demethylation pathways. *Genome Biology*, 2016. ISSN 1474760X.
[22] D. Aird, M. G. Ross, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, 2011. ISSN 14747596.

[23] L. B. Alexandrov, S. Nik-Zainal, et al. Signatures of mutational processes in human cancer. *Nature*, 2013. ISSN 14764687.
[24] C. D. Allis and T. Jenuwein. The molecular hallmarks of epigenetic control. *Nature Reviews Genetics*, 17(8):487–500, 2016. ISSN 1471-0056.
[25] H. M. Amemiya, A. Kundaje, et al. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports*, 2019. ISSN 20452322.
[26] R. An, Y. Jia, et al. Non-enzymatic depurination of nucleic acids: Factors and mechanisms. *PLoS ONE*, 2014. ISSN 19326203.
[27] G. Ananda, Y. Takemon, et al. Whole-genome sequence of the C57L/J mouse inbred strain. *G3: Genes, Genomes, Genetics*, 2014. ISSN 21601836.
[28] S. Andrews. FastQC - A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, 2010.
[29] F. Antequera, M. Tamame, et al. DNA methylation in the fungi. *Journal of Biological Chemistry*, 1984. ISSN 00219258.
[30] Ari and M. Arikan. Next-generation sequencing: Advantages, disadvantages, and future. In *Plant Omics: Trends and Applications*. 2016. ISBN 9783319317038.
[31] P. Armitage and R. Doll. The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer*, 1954. ISSN 15321827.
[32] B. G. Barrell. Fractionation and sequence analysis of radioactive nucleotides. In *Procedures in nucleic acid research*, pp. 751–779. 1971.
[33] M. Bachman, S. Uribe-Lewis, et al. 5-Formylcytosine can be a stable DNA modification in mammals. *Nature Chemical Biology*, 2015. ISSN 15524469.
[34] A. Bacolla, D. N. Cooper, et al. Mechanisms of base substitution mutagenesis in cancer genomes, 2014.
[35] M. H. Bailey, C. Tokheim, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, 2018. ISSN 10974172.
[36] L. Balakrishnan and B. Milavetz. Epigenetic regulation of viral biological processes, 2017.
[37] G. E. A. P. A. Batista, R. C. Prati, et al. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 2004. ISSN 1931-0145.
[38] P. E. Bayer, A. A. Golicz, et al. Plant pan-genomes are the new reference, 2020.
[39] B. E. Bernstein, A. Meissner, et al. The Mammalian Epigenome, 2007.
[40] B. E. Bernstein, J. A. Stamatoyannopoulos, et al. The NIH roadmap epigenomics mapping consortium, 2010.
[41] R. Bhattacharjee, S. Tanaka, et al. Porous nanozymes: The peroxidase-mimetic activity of mesoporous iron oxide for the colorimetric and electrochemical detection of global DNA methylation. *Journal of Materials Chemistry B*, 2018. ISSN 2050750X.
[42] M. Bibikova, Z. Lin, et al. High-throughput DNA methylation profiling using universal bead arrays. *Genome Research*, 2006. ISSN 10889051.

[43] A. Bird. DNA methylation patterns and epigenetic memory, 2002.
[44] A. Bird, P. Tate, et al. Studies of DNA methylation in animals. *Journal of Cell Science*, 1995. ISSN 00219533.
[45] C. Bishop and J. Lasserre. Generative or discriminative? getting the best of both worlds. *Bayesian Statistics*, 2007.
[46] K. Blaschke, K. T. Ebata, et al. Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells. *Nature*, 2013. ISSN 00280836.
[47] C. Bock. Analysing and interpreting DNA methylation data., 2012.
[48] G. K. Bogu, P. Vizán, et al. Chromatin and RNA Maps Reveal Regulatory Long Noncoding RNAs in Mouse. *Molecular and Cellular Biology*, 2016. ISSN 0270-7306.
[49] M. J. Booth, M. R. Branco, et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science*, 2012. ISSN 10959203.
[50] I. Bozic, J. M. Gerold, et al. Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. *PLoS Computational Biology*, 2016. ISSN 15537358.
[51] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 1997. ISSN 00313203.
[52] A. Brunet and S. L. Berger. Epigenetics of aging and aging-related disease, 2014.
[53] P. Burnham, M. S. Kim, et al. Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. *Scientific Reports*, 2016. ISSN 20452322.
[54] L. Cai, W. Yuan, et al. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Scientific Reports*, 2016. ISSN 20452322.
[55] E. B. Cambareri, B. C. Jensen, et al. Repeat-induced G-C to A-T mutations in Neurospora. *Science*, 1989. ISSN 00368075.
[56] C. Cao, F. Liu, et al. Deep Learning and Its Applications in Biomedicine, 2018.
[57] L. J. Carithers, K. Ardlie, et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation and Biobanking*, 2015. ISSN 1947-5535.
[58] J. Casadesús and D. Low. Epigenetic Gene Regulation in the Bacterial World. *Microbiology and Molecular Biology Reviews*, 2006. ISSN 1092-2172.
[59] G. Cavalli and E. Heard. Advances in epigenetics link genetics to the environment and disease, 2019.
[60] A. O. O. Chan, R. R. Broaddus, et al. CpG island methylation in aberrant crypt foci of the colorectum. *American Journal of Pathology*, 2002. ISSN 00029440.
[61] G. Chandrashekar and F. Sahin. A survey on feature selection methods. *Computers and Electrical Engineering*, 2014. ISSN 00457906.

[62] E. Chargaff, S. Zamenhof, et al. Human desoxypentose nucleic acid: Composition of human desoxypentose nucleic acid. *Nature*, 1950. ISSN 00280836.

[63] N. Chatterjee and G. C. Walker. Mechanisms of DNA damage, repair, and mutagenesis, 2017.

[64] N. V. Chawla, K. W. Bowyer, et al. SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 2002. ISSN 10769757.

[65] C. C. Chen, K. Y. Wang, et al. The mammalian de novo DNA methyltransferases DNMT3A and DNMT3B are also DNA 5-hydroxymethylcytosine dehydroxymethylases. *Journal of Biological Chemistry*, 2012. ISSN 00219258.

[66] H. Chen and P. C. Boutros. VennDiagram: A package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics*, 2011. ISSN 14712105.

[67] H. Chen and X. He. The convergent cancer evolution toward a single cellular destination. *Molecular Biology and Evolution*, 2016. ISSN 15371719.

[68] L. Chen, P. Liu, et al. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*, 2017. ISSN 10959203.

[69] X. Chen, O. Schulz-Trieglaff, et al. Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 2016. ISSN 14602059.

[70] Z. Chen, Y. Yuan, et al. Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Scientific Reports*, 2020. ISSN 20452322.

[71] J. Cheng, A. Dutra, et al. Improved generation of C57BL/6J mouse embryonic stem cells in a defined serum-free media. *Genesis*, 2004. ISSN 1526954X.

[72] D. Chicco and G. Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 2020. ISSN 14712164.

[73] T. Ching, D. S. Himmelstein, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface*, 2018. ISSN 17425662.

[74] F. Chollet. Keras: The Python Deep Learning library. *Keras.Io*, 2015.

[75] J. K. Christman. 5-Azacytidine and 5-aza-2-deoxycytidine as inhibitors of DNA methylation: Mechanistic studies and their implications for cancer therapy, 2002.

[76] K. Cibulskis, M. S. Lawrence, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 2013. ISSN 10870156.

[77] J. Cline, J. C. Braman, et al. PCR fidelity of Pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Research*, 1996. ISSN 03051048.

[78] S. J. Cokus, S. Feng, et al. Shotgun bisulphite sequencing of the Ara-

bidopsis genome reveals DNA methylation patterning. *Nature*, 2008. ISSN 14764687.

[79] G. A. Colditz. History of Cancer. In *The SAGE Encyclopedia of Cancer and Society*. 2015.

[80] D. P. Cooke, D. C. Wedge, et al. A unified haplotype-based method for accurate and comprehensive variant calling. *bioRxiv*, 2018.

[81] M. J. Corley, W. Zhang, et al. Semiconductor-based sequencing of genomewide DNA methylation states. *Epigenetics*, 2015. ISSN 15592308.

[82] A. Cornish-Bowden. Nomenclature for incompletely specified bases in nucleic acid sequences: Recommendations 1984. *Nucleic Acids Research*, 1985. ISSN 03051048.

[83] C. Coulondre, J. H. Miller, et al. Molecular basis of base substitution hotspots in Escherichia coli. *Nature*, 1978. ISSN 00280836.

[84] J. Craig Venter, M. D. Adams, et al. The sequence of the human genome. *Science*, 2001. ISSN 00368075.

[85] R. Dahm. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research, 2008.

[86] T. Daley and A. D. Smith. Predicting the molecular complexity of sequencing libraries. *Nature Methods*, 2013. ISSN 15487091.

[87] M. A. Dawson. The cancer epigenome: Concepts, challenges, and therapeutic opportunities, 2017.

[88] A. M. Deaton and A. Bird. CpG islands and the regulation of transcription. *Genes and Development*, 2011. ISSN 08909369.

[89] R. Dias and A. Torkamani. Artificial intelligence in clinical and genomic diagnostics, 2019.

[90] J. Dunn and S. Rao. Epigenetics and immunotherapy: The current state of play, 2017.

[91] M. A. Eberle, E. Fritzilas, et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Research*, 2017. ISSN 15495469.

[92] R. Edgar, P. P. C. Tan, et al. Meta-analysis of human methylomes reveals stably methylated sequences surrounding CpG islands associated with high gene expression. *Epigenetics and Chromatin*, 2014. ISSN 17568935.

[93] L. Edler and A. Kopp-Schneider. Origins of the mutational origin of cancer, 2005.

[94] M. Esteller. Epigenetics in evolution and disease. *The Lancet*, 2008. ISSN 01406736.

[95] F. Favero, T. Joshi, et al. Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. *Annals of Oncology*, 2015. ISSN 15698041.

[96] M. J. Fazzari and J. M. Greally. Introduction to epigenomics and epigenome-wide analysis., 2010.

[97] A. P. Feinberg and B. Tycko. The history of cancer epigenetics, 2004.

[98] A. P. Feinberg and B. Vogelstein. of some human cancers from their normal counterparts. (October 1982):3–6, 1983.

[99] H. Feng, K. N. Conneely, et al. A Bayesian hierarchical model to detect

differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Research*, 2014. ISSN 13624962.

[100] Z. Feng, J. Li, et al. Qdnamod: A statistical model-based tool to reveal intercellular heterogeneity of DNA modification from SMRT sequencing data. *Nucleic Acids Research*, 2014. ISSN 13624962.

[101] A. Fernández, S. García, et al. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary, 2018.

[102] G. Ficz, M. R. Branco, et al. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature*, 2011. ISSN 00280836.

[103] J. T. Finch and A. Klug. Solenoidal model for superstructure in chromatin. *Proceedings of the National Academy of Sciences of the United States of America*, 1976. ISSN 00278424.

[104] W. A. Flavahan, E. Gaskell, et al. Epigenetic plasticity and the hallmarks of cancer, 2017.

[105] B. A. Flusberg, D. R. Webster, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, 2010. ISSN 15487091.

[106] M. Frommer, L. E. McDonald, et al. 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*, 1992. ISSN 00278424.

[107] S. Gandhi, L. Lee, et al. cDeepbind: A context sensitive deep learning model of RNA-protein binding. *bioRxiv*, 2018.

[108] S. Gao, D. Zou, et al. BS-SNPer: SNP calling in bisulfite-seq data. *Bioinformatics*, 2015. ISSN 14602059.

[109] M. Gardiner-Garden and M. Frommer. CpG Islands in vertebrate genomes. *Journal of Molecular Biology*, 1987. ISSN 00222836.

[110] R. Gaujoux. An introduction to NMF package. *BMC Bioinformatics*, 2010.

[111] N. G. Gaylord. Reduction with complex metal hydrides. *Journal of Chemical Education*, 1957. ISSN 00219584.

[112] S. Gigante, Q. Gouil, et al. Using long-read sequencing to detect imprinted DNA methylation. *Nucleic Acids Research*, 2019. ISSN 13624962.

[113] W. Gilbert and A. Maxam. The Nucleotide Sequence of lac Operator. *Proc. Natl. Acad. Sci. U. S. A.*, 1973. ISSN 0027-8424.

[114] D. Globisch, M. Münzel, et al. Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS ONE*, 2010. ISSN 19326203.

[115] S. Gombar, T. MacCarthy, et al. Epigenetics Decouples Mutational from Environmental Robustness. Did It Also Facilitate Multicellularity? *PLoS Computational Biology*, 2014. ISSN 15537358.

[116] V. Gómez-Rubio. ggplot2 - Elegant Graphics for Data Analysis (2nd Edition) . *Journal of Statistical Software*, 2017. ISSN 1548-7660.

[117] A. M. Goodman, S. Kato, et al. Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Molecular*

Cancer Therapeutics, 2017. ISSN 15388514.

[118] Q. Gouil and A. Keniry. Latest techniques to study DNA methylation, 2019.

[119] M. V. Greenberg and D. Bourc'his. The diverse roles of DNA methylation in mammalian development and disease, 2019.

[120] S. I. Grewal. RNAi-dependent formation of heterochromatin and its diverse functions, 2010.

[121] M. M. Gubin, M. N. Artyomov, et al. Tumor neoantigens: Building a framework for personalized cancer immunotherapy, 2015.

[122] J. U. Guo, Y. Su, et al. Emerging roles of TET proteins and 5-hydroxymethylcytosines in active DNA demethylation and beyond, 2011.

[123] W. Guo, P. Zhu, et al. CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data. *Bioinformatics*, 2018. ISSN 14602059.

[124] Y. Guo, W. Li, et al. DeepACLSTM: Deep asymmetric convolutional long short-term memory neural models for protein secondary structure prediction. *BMC Bioinformatics*, 2019. ISSN 14712105.

[125] H. L. Wheeler and T. B. Johnson. Researches on Pyrimidine Derivatives: 5-Methyl-cytosine. *Journal of the American Chemical Society*, pp. 591–606, 1904.

[126] E. Habibi, A. B. Brinkman, et al. Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell*, 2013. ISSN 18759777.

[127] H. Han, W. Y. Wang, et al. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Lecture Notes in Computer Science*. 2005. ISSN 03029743.

[128] D. Hanahan and R. A. Weinberg. Hallmarks of cancer: The next generation, 2011.

[129] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 1982. ISSN 00338419.

[130] K. D. Hansen, W. Timp, et al. Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, 2011. ISSN 10614036.

[131] J. Harrow, A. Frankish, et al. GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research*, 2012. ISSN 10889051.

[132] K. Hayashi, K. Kidouchi, et al. Possible prediction of adverse reactions to pyrimidine chemotherapy from urinary pyrimidine levels and a case of asymptomatic adult dihydropyrimidinuria. *Clinical Cancer Research*, 1996. ISSN 10780432.

[133] H. Hayatsu, Y. Wataya, et al. Reaction of Sodium Bisulfite with Uracil, Cytosine, and Their Derivatives. *Biochemistry*, 1970. ISSN 15204995.

[134] S. Heerboth, K. Lapinska, et al. Use of epigenetic drugs in disease: An overview. *Genetics and Epigenetics*, 2014. ISSN 1179237X.

[135] I. R. Henderson and S. E. Jacobsen. Epigenetic inheritance in plants, 2007.

[136] A. Hermann, R. Goyal, et al. DNA Processively with High Preference for Hemimethylated. *Journal of Biological Chemistry*, 2004. ISSN 0021-9258.
[137] C. Heydt, N. Kumm, et al. Massively parallel sequencing fails to detect minor resistant subclones in tissue samples prior to tyrosine kinase inhibitor therapy. *BMC Cancer*, 2015. ISSN 14712407.
[138] M. Hirst and M. A. Marra. Next generation sequencing based approaches to epigenomics. *Briefings in Functional Genomics*, 2010. ISSN 20412649.
[139] T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998. ISSN 01628828.
[140] S. J. Hogg, P. A. Beavis, et al. Targeting the epigenetic regulation of antitumour immunity. *Nature Reviews Drug Discovery*, pp. 1–25, sep 2020. ISSN 1474-1776.
[141] R. W. Holley, J. Apgar, et al. Structure of a ribonucleic acid. *Science*, 1965. ISSN 00368075.
[142] R. Holliday and J. E. Pugh. DNA modification mechanisms and gene activity during development. *Science (New York, N.Y.)*, 1975. ISSN 0036-8075.
[143] D. W. Hosmer and S. Lemeshow. Introduction to the Logistic Regression Model. In *Applied Logistic Regression*. 2005.
[144] R. D. Hotchkiss. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *The Journal of biological chemistry*, 1948. ISSN 00219258.
[145] W. Huang, L. Li, et al. ART: A next-generation sequencing read simulator. *Bioinformatics*, 2012. ISSN 13674803.
[146] Y. Huang, W. A. Pastor, et al. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE*, 2010. ISSN 19326203.
[147] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, 2007. ISSN 15219615.
[148] F. Hvid-Jensen, L. Pedersen, et al. Incidence of adenocarcinoma among patients with Barrett's esophagus. *New England Journal of Medicine*, 2011. ISSN 15334406.
[149] O. Iizuka, F. Kanavati, et al. Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours. *Scientific Reports*, 2020. ISSN 20452322.
[150] Illumina. Evaluating Somatic Variant Calling in Tumor/Normal Studies. Tech. rep.
[151] —. An introduction to Nex-Generation Sequencing Technology. *Genome*, 2016.
[152] a. Illumina and P. Control. Using a PhiX Control for HiSeq ® Sequencing Runs. *Technical Note: Sequencing*, 770-2011-0:1–2, 2012.
[153] D. Incarnato, A. Krepelova, et al. High-throughput single nucleotide variant discovery in E14 mouse embryonic stem cells provides a new reference genome assembly. *Genomics*, 2014. ISSN 10898646.
[154] D. Incarnato and F. Neri. High-throughput whole-genome sequencing of E14 mouse embryonic stem cells. *Genomics Data*, 2015. ISSN 22135960.

[155] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *32nd International Conference on Machine Learning, ICML 2015*. 2015. ISBN 9781510810587.

[156] S. Ito, A. C. Dalessio, et al. Role of tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*, 2010. ISSN 00280836.

[157] S. Ito, L. Shen, et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, 2011. ISSN 00368075.

[158] M. Ivanov, A. Baranova, et al. Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC Genomics*, 2015. ISSN 14712164.

[159] R. Jaenisch and A. Bird. Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. 2003. ISSN 15461718.

[160] L. Ji, T. Sasaki, et al. Methylated DNA is over-represented in whole-genome bisulfite sequencing data. *Frontiers in Genetics*, 2014. ISSN 16648021.

[161] C. Jin, Y. Lu, et al. TET1 is a maintenance DNA demethylase that prevents methylation spreading in differentiated cells. *Nucleic Acids Research*, 2014. ISSN 13624962.

[162] T. B. Johnson and R. 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*, 1925. ISSN 15205126.

[163] P. A. Jones. Functions of DNA methylation: Islands, start sites, gene bodies and beyond, 2012.

[164] J. Josse, K. A. D., et al. Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *The Journal of biological chemistry*, 1961. ISSN 00219258.

[165] G. Jun, M. K. Wing, et al. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Research*, 2015. ISSN 15495469.

[166] K. Jung, M. Fleischhacker, et al. Cell-free DNA in the blood as a solid tumor biomarker-A critical appraisal of the literature. *Clinica Chimica Acta*, 411(21-22):1611–1624, 2010. ISSN 00098981.

[167] R. Kanwal, K. Gupta, et al. Cancer epigenetics: an introduction, 2015.

[168] M. Karimzadeh, C. Ernst, et al. Umap and Bismap: Quantifying genome and methylome mappability. *Nucleic Acids Research*, 2018. ISSN 13624962.

[169] M. B. Kastan, B. J. Gowans, et al. Methylation of deoxycytidine incorporated by excision-repair synthesis of DNA. *Cell*, 1982. ISSN 00928674.

[170] Y. Kato, M. Kaneda, et al. Role of the Dnmt3 family in de novo methylation of imprinted and repetitive sequences during male germ cell development in the mouse. *Human Molecular Genetics*, 2007. ISSN 09646906.

[171] A. M. Kaz, W. M. Grady, et al. Genetic and Epigenetic Alterations in Barrett's Esophagus and Esophageal Adenocarcinoma, 2015.

[172] R. Ke, M. Mignardi, et al. Fourth Generation of Next-Generation Sequencing Technologies: Promise and Consequences, 2016.
[173] D. A. Khavari, G. L. Sen, et al. DNA methylation and epigenetic control of cellular differentiation, 2010.
[174] D. Kim, B. Langmead, et al. HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, 2015. ISSN 15487105.
[175] D. Kim, J. M. Paggi, et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 2019. ISSN 15461696.
[176] J. U. Kim, M. I. Shariff, et al. Hepatocellular carcinoma: Review of disease and tumor biomarkers, 2016.
[177] S. Kim, K. Scheffler, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods*, 2018. ISSN 15487105.
[178] T. Klobučar, E. Kreibich, et al. IMPLICON: an ultra-deep sequencing method to uncover DNA methylation at imprinted regions. *Nucleic Acids Research*, 2020. ISSN 1362-4962.
[179] D. C. Koboldt. Best practices for variant calling in clinical sequencing, 2020.
[180] K. P. Koh, A. Yabuuchi, et al. Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. *Cell Stem Cell*, 2011. ISSN 19345909.
[181] N. Kono and K. Arakawa. Nanopore sequencing: Review of potential applications in functional genomics, 2019.
[182] R. D. Kornberg. Chromatin structure: A repeating unit of histones and DNA. *Science*, 1974. ISSN 00368075.
[183] T. J. Koski and J. Noble. A Review of Bayesian Networks and Structure Learning. *Mathematica Applicanda*, 2012. ISSN 1730-2668.
[184] S. Kriaucionis and N. Heintz. The nuclear DNA base 5-hydroxymethylcytosine is present in purkinje neurons and the brain. *Science*, 2009. ISSN 00368075.
[185] F. Krueger and S. R. Andrews. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 2011. ISSN 13674803.
[186] —. Quality Control , trimming and alignment of Bisulfite-Seq data. *Practical Info NICE!!*, (July):1–13, 2012.
[187] J. K. Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, second edition*. 2014. ISBN 9780124058880.
[188] Y. Kudo, K. Tateishi, et al. Loss of 5-hydroxymethylcytosine is accompanied with malignant cellular transformation. *Cancer Science*, 2012. ISSN 13479032.
[189] R. M. Kuhn, D. Haussler, et al. The UCSC genome browser and associated tools. *Briefings in Bioinformatics*, 2013. ISSN 14675463.
[190] G. Kunde-Ramamoorthy, C. Coarfa, et al. Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. *Nucleic Acids Research*, 2014. ISSN 13624962.
[191] S. Kurtz, A. Phillippy, et al. Versatile and open software for comparing

large genomes. *Genome biology*, 2004. ISSN 1465-6914.

[192] D. Laehnemann, A. Borkhardt, et al. Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Briefings in Bioinformatics*, 2016. ISSN 14774054.

[193] T. LaFramboise. Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances, 2009.

[194] D. Lai, S. Shah, et al. HMMcopy: copy number prediction with correction for GC and mappability bias for HTS data. *R package version*, 2012.

[195] G. Lancia, V. Bafna, et al. SNPs problems, complexity, and algorithms. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2001. ISBN 9783540424932. ISSN 16113349.

[196] E. S. Lander, L. M. Linton, et al. Initial sequencing and analysis of the human genome. *Nature*, 2001. ISSN 00280836.

[197] B. Langmead and S. Salzberg. Bowtie2. *Nature methods*, 2013. ISSN 1548-7091.

[198] R. Lardenoije, A. Iatrou, et al. The epigenetics of aging and neurodegeneration, 2015.

[199] M. S. Lawrence, P. Stojanov, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 2013. ISSN 00280836.

[200] J. Lemley, S. Bazrafkan, et al. Smart Augmentation Learning an Optimal Data Augmentation Strategy. *IEEE Access*, 2017. ISSN 21693536.

[201] C. Li and S. M. Williams. Human Somatic Variation: It's Not Just for Cancer Anymore. *Current Genetic Medicine Reports*, 2013. ISSN 2167-4876.

[202] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 2009. ISSN 13674803.

[203] H. Li, B. Handsaker, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 2009. ISSN 13674803.

[204] W. Li, X. Zhang, et al. 5-Hydroxymethylcytosine signatures in circulating cell-free DNA as diagnostic biomarkers for human cancers. *Cell Research*, 2017. ISSN 17487838.

[205] Y. Li, S. Kumar, et al. Active DNA demethylation: mechanism and role in plant development, 2018.

[206] Y. Li and T. O. Tollefsbol. DNA methylation detection: Bisulfite genomic sequencing analysis. *Methods in Molecular Biology*, 2011. ISSN 10643745.

[207] Z. Li and Y. Yu. Protein secondary structure prediction using cascaded convolutional and recurrent neural networks. In *IJCAI International Joint Conference on Artificial Intelligence*. 2016. ISSN 10450823.

[208] M. W. Libbrecht and W. S. Noble. Machine learning applications in genetics and genomics, 2015.

[209] T. Lindahl and B. Nyberg. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry*, 1974. ISSN 15204995.

[210] R. Lister, R. C. O'Malley, et al. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell*, 2008. ISSN 00928674.

[211] J. Liu and P. W. Doetsch. Escherichia coli RNA and DNA polymerase

bypass of dihydrouracil: Mutagenic potential via transcription and replication. *Nucleic Acids Research*, 1998. ISSN 03051048.

[212] L. Liu, Y. Zhang, et al. Recent Advances in the Genomic Profiling of Bacterial Epigenetic Modifications, 2019.

[213] X. Liu, S. Han, et al. Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS ONE*, 2013. ISSN 19326203.

[214] Y. Liu, J. Cheng, et al. Accurate targeted long-read DNA methylation and hydroxymethylation sequencing with TAPS. *Genome Biology*, 2020. ISSN 1474760X.

[215] Y. Liu, K. D. Siegmund, et al. Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biology*, 2012. ISSN 1474760X.

[216] Y. Liu, P. Siejka-Zielińska, et al. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nature Biotechnology*, 2019. ISSN 15461696.

[217] M. I. Love, S. Anders, et al. *Differential analysis of count data - the DESeq2 package*. 2014. ISBN 0000000000.

[218] M. F. Lyon. Gene action in the X-chromosome of the mouse (mus musculus L.). *Nature*, 1961. ISSN 00280836.

[219] T. Marschall, M. Marz, et al. Computational pan-genomics: Status, promises and challenges. *Briefings in Bioinformatics*, 2018. ISSN 14774054.

[220] M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 2011.

[221] P. B. E. B. Z. C. C. C. G. S. C. A. D. J. D. M. D. S. G. I. G. A. H. G. I. M. I. Y. J. R. J. L. K. M. K. J. L. D. M. R. M. S. M. D. M. C. O. M. S. J. S. B. S. I. S. K. T. P. T. V. V. V. V. F. V. O. V. P. W. M. W. M. W. Y. Y. Martín Abadi, Ashish Agarwal and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[222] A. M. Maxam and W. Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 1977. ISSN 00278424.

[223] W. Mayer, A. Niveleau, et al. Demethylation of the zygotic paternal genome. *Nature*, 2000. ISSN 00280836.

[224] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 1943. ISSN 00074985.

[225] J. McGrath and D. Solter. Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell*, 1984. ISSN 00928674.

[226] P. McInerney, P. Adams, et al. Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. *Molecular Biology International*, 2014. ISSN 2090-2182.

[227] A. McKenna, M. Hanna, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 2010. ISSN 10889051.

[228] W. McKinney and P. D. Team. Pandas - Powerful Python Data Analysis

Toolkit. *Pandas - Powerful Python Data Analysis Toolkit*, 2015.

[229] S. N. McNulty, P. R. Mann, et al. Impact of Reducing DNA Input on Next-Generation Sequencing Library Complexity and Variant Detection. *Journal of Molecular Diagnostics*, 2020. ISSN 19437811.

[230] C. E. Meacham and S. J. Morrison. Tumour heterogeneity and cancer cell plasticity, 2013.

[231] F. Meacham, D. Boffelli, et al. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, 2011. ISSN 14712105.

[232] A. Meissner, A. Gnirke, et al. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, 2005. ISSN 03051048.

[233] P. Melamed, Y. Yosefzon, et al. Tet enzymes, variants, and differential effects on function, 2018.

[234] I. Mellman, G. Coukos, et al. Cancer immunotherapy comes of age, 2011.

[235] D. Meulendijks, L. M. Henricks, et al. Pretreatment serum uracil concentration as a predictor of severe and fatal fluoropyrimidine-associated toxicity. *British Journal of Cancer*, 2017. ISSN 15321827.

[236] F. Miescher. Ueber die chemische Zusammensetzung der Eiterzellen. *Medicinisch-chemische Untersuchungen*, 1871.

[237] J. L. Miller and P. A. Grant. The role of DNA methylation and histone modifications in transcriptional regulation in humans. *Subcellular Biochemistry*, 61:289–317, may 2013. ISSN 03060225.

[238] L. Mills. Common file formats. *Current Protocols in Bioinformatics*, 2014. ISSN 1934340X.

[239] A. E. Minoche, J. C. Dohm, et al. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology*, 2011. ISSN 14747596.

[240] T. B. Miranda and P. A. Jones. DNA methylation: The nuts and bolts of repression, 2007.

[241] F. Miura, Y. Enomoto, et al. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Research*, 2012. ISSN 03051048.

[242] S. H. Moolgavkar and A. G. Knudson. Mutation and cancer: A model for human carcinogenesis. *Journal of the National Cancer Institute*, 1981. ISSN 14602105.

[243] J. Moss, J. Magenheim, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nature Communications*, 2018. ISSN 20411723.

[244] J. C. Mu, M. Mohiyuddin, et al. VarSim: A high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. *Bioinformatics*, 2015. ISSN 14602059.

[245] C. B. Mulholland, F. R. Traube, et al. Distinct and stage-specific contributions of TET1 and TET2 to stepwise cytosine oxidation in the transition from naive to primed pluripotency. *Scientific Reports*, 2020. ISSN 20452322.

[246] D. Muzzey, E. A. Evans, et al. Understanding the Basics of NGS: From Mechanism to Variant Calling. *Current Genetic Medicine Reports*, 2015. ISSN 2167-4876.

[247] F. N. Naguib, M. H. El Kouni, et al. Enzymes off Uracil Catabolism in Normal and Neoplastic Human Tissues. *Cancer Research*, 1985. ISSN 15387445.

[248] I. Ngs. Technology Spotlight: Illumina Sequencing Technology. *Manual*, 2009. ISSN 01677799.

[249] R. Nielsen, J. S. Paul, et al. Genotype and SNP calling from next-generation sequencing data, 2011.

[250] E. Ochoa, V. Zuber, et al. MethylCal: Bayesian calibration of methylation levels. *Nucleic acids research*, 2019. ISSN 13624962.

[251] R. Ohno, M. Nakayama, et al. A replication-dependent passive mechanism modulates DNA demethylation in mouse primordial germ cells. *Development (Cambridge)*, 2013. ISSN 09501991.

[252] M. Okano, D. W. Bell, et al. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 1999. ISSN 00928674.

[253] A. L. Olins and D. E. Olins. Spheroid chromatin units (v bodies). *Science*, 1974. ISSN 00368075.

[254] T. O'Malley, E. Bursztein, et al. Keras Tuner, 2019.

[255] S. Pal and J. K. Tyler. Epigenetics and aging, 2016.

[256] J. Park and S. Lin. Detection of Differentially Methylated Regions Using Bayesian Curve Credible Bands. *Statistics in Biosciences*, 2018. ISSN 18671772.

[257] N. J. Parkinson, S. Maslau, et al. Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. *Genome Research*, 2012. ISSN 10889051.

[258] K. Paschos and M. J. Allday. Epigenetic reprogramming of host genes in viral and microbial pathogenesis, 2010.

[259] W. A. Pastor, Y. Huang, et al. The GLIB technique for genome-wide mapping of 5-hydroxymethylcytosine. *Nature Protocols*, 2012. ISSN 17542189.

[260] H. Patani, M. D. Rushton, et al. Transition to naïve human pluripotency mirrors pan-cancer DNA hypermethylation. *Nature Communications*, 2020. ISSN 20411723.

[261] T. G. Paulson, C. C. Maley, et al. Chromosomal instability and copy number alterations in Barrett's esophagus and esophageal adenocarcinoma. *Clinical Cancer Research*, 2009. ISSN 10780432.

[262] F. Pedregosa, G. Varoquaux, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011. ISSN 15324435.

[263] N. W. Penn, R. Suwalski, et al. The presence of 5-hydroxymethylcytosine in animal deoxyribonucleic acid. *The Biochemical journal*, 1972. ISSN 02646021.

[264] T. Pfaffeneder, B. Hackner, et al. The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angewandte Chemie - International Edition*, 2011. ISSN 14337851.

[265] C. S. Pikaard and O. M. Scheid. Epigenetic regulation in plants. *Cold Spring Harbor Perspectives in Biology*, 2014. ISSN 19430264.
[266] N. Plongthongkum, D. H. Diep, et al. Advances in the profiling of DNA modifications: Cytosine methylation and beyond, 2014.
[267] R. Poplin, P. C. Chang, et al. A universal snp and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 2018. ISSN 15461696.
[268] A. Portela and M. Esteller. Epigenetic modifications and human disease, 2010.
[269] Y. Qi. Random forest for bioinformatics. In *Ensemble Machine Learning: Methods and ApplicatiOns*. 2012. ISBN 9781441993267.
[270] A. R. Quinlan. BEDTools: The Swiss-Army tool for genome feature analysis. *Current Protocols in Bioinformatics*, 2014. ISSN 1934340X.
[271] A. Rada-Iglesias, F. G. Grosveld, et al. Forces driving the three-dimensional folding of eukaryotic genomes. *Molecular Systems Biology*, 2018. ISSN 1744-4292.
[272] E.-A. Raiber, R. Hardisty, et al. Mapping and elucidating the function of modified bases in DNA. *Nature Reviews Chemistry*, 2017.
[273] E. A. Raiber, G. Portella, et al. 5-Formylcytosine organizes nucleosomes and forms Schiff base interactions with histones in mouse embryonic stem cells. *Nature Chemistry*, 2018. ISSN 17554349.
[274] V. R. Ramirez-Carrozzi, D. Braas, et al. A Unifying Model for the Selective Regulation of Inducible Transcription by CpG Islands and Nucleosome Remodeling. *Cell*, 2009. ISSN 00928674.
[275] K. D. Rasmussen and K. Helin. Role of TET enzymes in DNA methylation, development, and cancer, 2016.
[276] A. Ratan, Y. Zhang, et al. Calling SNPs without a reference sequence. *BMC Bioinformatics*, 2010. ISSN 14712105.
[277] D. Ray and R. Yung. Immune senescence, epigenetics and autoimmunity. *Clinical Immunology*, 2018. ISSN 15217035.
[278] A. Razin and A. D. Riggs. DNA Methylation and gene function. *Science*, 1980. ISSN 00368075.
[279] W. E. Razzell and H. G. Khorana. Studies on polynucleotides. X. Enzymic degradation. Some properties and mode of action of spleen phosphodiesterase. *The Journal of biological chemistry*, 1961. ISSN 00219258.
[280] N. Rieber, M. Zapatka, et al. Coverage Bias and Sensitivity of Variant Calling for Four Whole-genome Sequencing Technologies. *PLoS ONE*, 2013. ISSN 19326203.
[281] J. J. M. Riethoven. Regulatory regions in DNA: promoters, enhancers, silencers, and insulators., 2010.
[282] A. D. Riggs. X inactivation, differentiation, and DNA methylation. *Cytogenetic and Genome Research*, 1975. ISSN 1424859X.
[283] A. Rimmer, H. Phan, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, 2014. ISSN 15461718.
[284] G. Ritter, R. Wilson, et al. The multistage model of cancer development: Some implications. *Toxicology and Industrial Health*, 2003. ISSN

07482337.
[285] R. W. Roberts and D. M. Crothers. Stability and properties of double and triple helices: Dramatic effects of RNA or DNA backbone composition. *Science*, 1992. ISSN 00368075.
[286] J. T. Robinson, H. Thorvaldsdóttir, et al. Integrative genomics viewer, 2011.
[287] M. D. Robinson, D. J. McCarthy, et al. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 2010. ISSN 1367-4803.
[288] M. D. Robinson and M. Pelizzola. Computational epigenomics: Challenges and opportunities, 2015.
[289] L. Rokach and O. Maimon. *Data mining with decision trees : theroy and applications*. 2008. ISBN 9789812771711\r9812771719.
[290] Z. S. B. W. L. L. S. S. G. N. D. M. A. C. B. S. K. M. M. S. J. C. S. H. E. C. E. T. I. R. C. J. S. P. E. T. D. T. C. E. J. L. W. Romualdas Vaisvila, V. K. Chaithanya Ponnaluri and T. B. Davis. EM-seq: Detection of DNA Methylation at Single Base Resolution from Picograms of DNA.
[291] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 1958. ISSN 0033295X.
[292] R. Rosenthal, N. McGranahan, et al. deconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology*, 2016. ISSN 1474760X.
[293] J. J. Salk, M. W. Schmitt, et al. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations, 2018.
[294] R. M. Samstein, C. H. Lee, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types, 2019.
[295] Samtools Organisation. The Variant Call Format Specification v4.3. *Online Resource*, 2018.
[296] M. A. Sánchez-Romero and J. Casadesús. The bacterial epigenome, 2020.
[297] J. Sandoval, L. Peiró-Chova, et al. Epigenetic biomarkers in laboratory diagnostics: Emerging approaches and opportunities, 2013.
[298] F. Sanger, G. G. Brownlee, et al. A two-dimensional fractionation procedure for radioactive nucleotides. *Journal of Molecular Biology*, 1965. ISSN 00222836.
[299] F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 1975. ISSN 00222836.
[300] F. Sanger, S. Nicklen, et al. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977. ISSN 0027-8424.
[301] C. T. Saunders, W. S. Wong, et al. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 2012. ISSN 13674803.
[302] S. Saxonov, P. Berg, et al. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Pro-*

ceedings of the National Academy of Sciences of the United States of America, 2006. ISSN 00278424.

[303] E. W. Sayers, M. Cavanaugh, et al. GenBank. *Nucleic Acids Research*, 2020. ISSN 13624962.

[304] S. Sayols, D. Scherzinger, et al. dupRadar: A Bioconductor package for the assessment of PCR artifacts in RNA-Seq data. *BMC Bioinformatics*, 2016. ISSN 14712105.

[305] M. D. Schultz, R. J. Schmitz, et al. 'Leveling' the playing field for analyses of single-base resolution DNA methylomes, 2012.

[306] E. K. Schutsky, J. E. Denizio, et al. Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. *Nature Biotechnology*, 2018. ISSN 15461696.

[307] R. Shapiro, B. Braverman, et al. Nucleic acid reactivity and conformation. II. Reaction of cytosine and uracil with sodium bisulfite. *Journal of Biological Chemistry*, 1973. ISSN 00219258.

[308] R. Shapiro and R. S. Klein. The Deamination of Cytidine and Cytosine by Acidic Buffer Solutions. Mutagenic Implications. *Biochemistry*, 1966. ISSN 15204995.

[309] R. Shapiro, R. E. Servis, et al. Reactions of Uracil and Cytosine Derivatives with Sodium Bisulfite. A Specific Deamination Method. *Journal of the American Chemical Society*, 1970. ISSN 15205126.

[310] L. Shen, H. Wu, et al. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell*, 2013. ISSN 00928674.

[311] R. M. Sherman and S. L. Salzberg. Pan-genomics in the human genome era, 2020.

[312] A. Sherstinsky. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 2020. ISSN 01672789.

[313] C. Shih and R. A. Weinberg. Isolation of a transforming sequence from a human bladder carcinoma cell line. *Cell*, 1982. ISSN 00928674.

[314] M. Shiraishi and H. Hayatsu. High-speed conversion of cytosine to uracil in bisulfite genomic sequencing analysis of DNA methylation. *DNA Research*, 2004. ISSN 13402838.

[315] M. Singer, D. Boffelli, et al. MetMap enables genome-scale methyltyping for determining methylation states in populations. *PLoS Computational Biology*, 2010. ISSN 1553734X.

[316] H. O. Smith and K. Welcox. A Restriction enzyme from Hemophilus influenzae. *Journal of Molecular Biology*, 1970. ISSN 00222836.

[317] M. W. Snyder, M. Kircher, et al. Cell-free DNA Comprises an in Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell*, 2016. ISSN 10974172.

[318] C. X. Song, K. E. Szulwach, et al. Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nature Biotechnology*, 2011. ISSN 10870156.

[319] —. Genome-wide profiling of 5-formylcytosine reveals its roles in epige-

netic priming. *Cell*, 2013. ISSN 00928674.

[320] C. X. Song, S. Yin, et al. 5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages. *Cell Research*, 2017. ISSN 17487838.

[321] G. Song, G. Wang, et al. An All-to-All Approach to the Identification of Sequence-Specific Readers for Epigenetic DNA Modifications on Cytosine, 2021.

[322] A. J. Sood, C. Viner, et al. DNamod: The DNA modification database. *Journal of Cheminformatics*, 2019. ISSN 17582946.

[323] V. C. Sousa, M. Fritz, et al. Approximate Bayesian computation without summary statistics: The case of admixture. *Genetics*, 2009. ISSN 00166731.

[324] M. Stephens and P. Donnelly. A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data. *American Journal of Human Genetics*, 2003. ISSN 00029297.

[325] S. Sumi, K. Kidouchi, et al. Urinary Screening for Pyrimidine Metabolism Disorders. 1998.

[326] D. Sun, Y. Xi, et al. MOABS: Model based analysis of bisulfite sequencing data. *Genome Biology*, 2014. ISSN 1474760X.

[327] Y. Sun, C. A. Ruivenkamp, et al. Next-Generation Diagnostics: Gene Panel, Exome, or Whole Genome? *Human Mutation*, 2015. ISSN 10981004.

[328] M. A. Surani, S. C. Barton, et al. Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. *Nature*, 1984. ISSN 00280836.

[329] M. Suzuki, W. Liao, et al. Whole-genome bisulfite sequencing with improved accuracy and cost. *Genome Research*, 2018. ISSN 15495469.

[330] M. Szyf. Nongenetic inheritance and transgenerational epigenetics, 2015.

[331] M. Tahiliani, K. P. Koh, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, 2009. ISSN 00368075.

[332] C. Tamm, S. P. Galitó, et al. A comparative study of protocols for mouse embryonic stem cell culturing. *PLoS ONE*, 2013. ISSN 19326203.

[333] S. A. Tammen, S. Friso, et al. Epigenetics: The link between nature and nurture. 2013. ISSN 00982997.

[334] G. Tan, L. Opitz, et al. Long fragments achieve lower base quality in Illumina paired-end sequencing. *Scientific Reports*, 2019. ISSN 20452322.

[335] K. Tanaka and A. Okamoto. Degradation of DNA by bisulfite treatment. *Bioorganic and Medicinal Chemistry Letters*, 2007. ISSN 0960894X.

[336] R. Thaler, S. Spitzer, et al. DMSO is a strong inducer of DNA hydroxymethylation in pre-osteoblastic MC3T3-E1 cells. *Epigenetics*, 2012. ISSN 15592308.

[337] S. Thankaswamy-Kosalai, P. Sen, et al. Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. *Genomics*, 2017. ISSN 10898646.

[338] W. Timp and A. P. Feinberg. Cancer as a dysregulated epigenome allowing

cellular growth advantage at the expense of the host, 2013.
[339] I. Tomek. TWO MODIFICATIONS OF CNN. *IEEE Transactions on Systems, Man and Cybernetics*, 1976. ISSN 00189472.
[340] T. Ueki, K. M. Walter, et al. Aberrant CpG island methylation in cancer cell lines arises in the primary cancers from which they were derived. *Oncogene*, 2002. ISSN 09509232.
[341] A. Unnikrishnan, N. Hadad, et al. Revisiting the genomic hypomethylation hypothesis of aging. *Annals of the New York Academy of Sciences*, 2018. ISSN 17496632.
[342] C. Valiente-Mullor, B. Beamud, et al. One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. *PLOS Computational Biology*, 2021.
[343] E. L. van Dijk, Y. Jaszczyszyn, et al. The Third Revolution in Sequencing Technology, 2018.
[344] S. J. van Dijk, R. L. Tellam, et al. Recent developments on the role of epigenetics in obesity and metabolic disease. *Clinical Epigenetics*, 2015. ISSN 18687083.
[345] A. B. Van Kuilenburg, H. Van Lenthe, et al. Activity of pyrimidine degradation enzymes in normal tissues. *Nucleosides, Nucleotides and Nucleic Acids*, 2006. ISSN 15257770.
[346] J. Van Vliet, N. A. Oates, et al. Epigenetic mechanisms in the context of complex diseases, 2007.
[347] A. Varriale and D. N. A. Methylation. DNA Methylation , Epigenetics , and Evolution in Vertebrates :. 2014:1–7, 2014.
[348] G. Vernikos, D. Medini, et al. Ten years of pan-genome analyses, 2015.
[349] P. Virtanen, R. Gommers, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 2020. ISSN 15487105.
[350] E. Vischer and E. Chargaff. The separation and quantitative estimation of purines and pyrimidines in minute amounts. *The Journal of biological chemistry*, 1948. ISSN 00219258.
[351] B. Vogelstein, N. Papadopoulos, et al. Cancer genome landscapes, 2013.
[352] T. Wakayama, I. Rodriguez, et al. Mice cloned from embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, 1999. ISSN 00278424.
[353] K. W. Walker. Site-Directed Mutagenesis. In *Encyclopedia of Cell Biology*. 2016. ISBN 9780123944474.
[354] J. Wan, D. Wang, et al. A Critical Review of Recurrent Neural Networks for Sequence Learning arXiv : 1506 . 00019v2 [cs . LG] 29 Jun 2015. *International Journal of Computer Vision*, 2015. ISSN 18792782.
[355] L. Wang, Y. Zhou, et al. Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex. *Nature*, 2015. ISSN 14764687.
[356] T. Wang, R. G. Qiu, et al. Predictive Modeling of the Progression of Alzheimer's Disease with Recurrent Neural Networks. *Scientific Reports*, 2018. ISSN 20452322.
[357] M. Waskom and the seaborn development team. Seaborn.
[358] R. H. Waterston, K. Lindblad-Toh, et al. Initial sequencing and compar-

ative analysis of the mouse genome. *Nature*, 2002. ISSN 00280836.

[359] J. D. Watson and F. H. Crick. The structure of DNA. *Cold Spring Harbor symposia on quantitative biology*, 1953. ISSN 00917451.

[360] M. Weber, J. J. Davies, et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat.Genet.*, 2005.

[361] A. Weismann, W. N. Parker, et al. The Germ-Plasm: A Theory of Heredity. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 1894. ISSN 09595295.

[362] B. Weiss and C. C. Richardson. The 5-terminal dinucleotides of the separated strands of T7 bacteriophage deoxyribonucleic acid. *Journal of Molecular Biology*, 1967. ISSN 00222836.

[363] L. Wen, X. Li, et al. Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. *Genome Biology*, 2014. ISSN 1474760X.

[364] W. R. Wikoff, D. Grapov, et al. Metabolomic markers of altered nucleotide metabolism in early stage adenocarcinoma. *Cancer Prevention Research*, 2015. ISSN 19406215.

[365] M. H. Wilkins and J. T. Randall. Crystallinity in sperm heads: Molecular structure of nucleoprotein in vivo. *BBA - Biochimica et Biophysica Acta*, 1953. ISSN 00063002.

[366] D. L. Wilson. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man and Cybernetics*, 1972. ISSN 21682909.

[367] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1997. ISSN 1089778X.

[368] D. E. Wood and S. L. Salzberg. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3), 2014. ISSN 1474760X.

[369] H. Wu, Y. Zhang, et al. Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell*, 2014. ISSN 1097-4172.

[370] H. X. Wu, Y. X. Chen, et al. Alteration in TET1 as potential biomarker for immune checkpoint blockade in multiple cancers. *Journal for ImmunoTherapy of Cancer*, 2019. ISSN 20511426.

[371] R. Wu. Nucleotide sequence analysis of DNA. I. Partial sequence of the cohesive ends of bacteriophage λ and 186 DNA. *Journal of Molecular Biology*, 1970. ISSN 00222836.

[372] R. Wu and A. D. Kaiser. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *Journal of Molecular Biology*, 1968. ISSN 00222836.

[373] G. R. Wyatt. Occurrence of 5-methyl-cytosine in nucleic acids, 1950.

[374] G. R. Wyatt and S. S. Cohen. A new pyrimidine base from bacteriophage nucleic acids, 1952.

[375] Y. Xi and W. Li. BSMAP: Whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, 2009. ISSN 14712105.

[376] C. L. Xiao, S. Zhu, et al. N6-Methyladenine DNA Modification in the Human Genome. *Molecular Cell*, 2018. ISSN 10974164.

[377] L. Xu and M. Seki. Recent advances in the detection of base modifications using the Nanopore sequencer, 2020.

[378] I. Y., K. S., et al. Plasma dihydrouracil/uracil ratio as a predictive biomarker for the effect and pharmacokinetics of 5-fluorouracil in colorectal tumor growth. *Drug Metabolism Reviews*, 2014. ISSN 0360-2532.

[379] W. Yan, J. G. Herman, et al. Epigenome-based personalized medicine in human cancer, 2016.

[380] M. Yu, G. C. Hon, et al. Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. *Nature Protocols*, 2012. ISSN 17502799.

[381] Y. Yu, X. Si, et al. A review of recurrent neural networks: Lstm cells and network architectures, 2019.

[382] D. Zeiberg, T. Prahlad, et al. Machine learning for patient risk stratification for acute respiratory distress syndrome. *PLoS ONE*, 2019. ISSN 19326203.

[383] H. Zeng, B. He, et al. Bisulfite-Free, Nanoscale Analysis of 5-Hydroxymethylcytosine at Single Base Resolution. *Journal of the American Chemical Society*, 2018. ISSN 15205126.

[384] J. Zeng, H. K. Nagrajan, et al. Fundamental diversity of human CpG islands at multiple biological levels. *Epigenetics*, 2014. ISSN 15592308.

[385] L. Zhang, W. Chen, et al. A TET homologue protein from coprinopsis cinerea (CcTET) that biochemically converts 5-methylcytosine to 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxylcytosine. *Journal of the American Chemical Society*, 2014. ISSN 15205126.

[386] Z. Zhao and E. Boerwinkle. Neighboring-nucleotide effects on single nucleotide polymorphisms: A study of 2.6 million polymorphisms across the human genome. *Genome Research*, 2002. ISSN 10889051.

[387] X. Zheng and L. Wang. Direct Electrocatalytic Oxidation and Simultaneous Determination of 5-Methylcytosine and Cytosine at Electrochemically Reduced Graphene Modified Glassy Carbon Electrode. *Electroanalysis*, 2013. ISSN 10400397.

[388] L. Zhou, H. K. Ng, et al. Systematic evaluation of library preparation methods and sequencing platforms for high-throughput whole genome bisulfite sequencing. *Scientific Reports*, 2019. ISSN 20452322.

[389] Y. Zhou, F. He, et al. The impact of DNA methylation dynamics on the mutation rate during human germline development. *G3: Genes, Genomes, Genetics*, 2020. ISSN 21601836.

[390] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 2005. ISSN 13697412.