https://doi.org/10.15388/vu.thesis.280
https://orcid.org/0000-0001-5311-6021

VILNIUS UNIVERSITY

Povilas
GIBAS

Statistical and computational approaches for the analysis of high-throughput epigenomic data

DOCTORAL DISSERTATION

Natural sciences,
Biochemistry (N 004)

VILNIUS 2022

The dissertation was prepared between 2015 and 2020 at Vilnius University, Life Sciences Center, Institute of Biotechnology.

The research was supported by Research Council of Lithuania.

The dissertation is defended on an external basis.

Academic supervisor – Prof. Dr. Saulius Klimašauskas (Vilnius University, Natural sciences, Biochemistry, N 004).

This doctoral dissertation will be defended in a public meeting of the Dissertation Defence Panel:

Chairman – Prof. Dr. Saulius Serva (Vilnius University, Natural sciences, Biochemistry, N 004).

Members:
dr. Jonas Bačelis (Statistics Lithuania, Natural sciences, Biochemistry, N 004);
prof. Matthias Bochtler (International Institute of Molecular and Cell Biology, Poland, Natural sciences, Biochemistry, N 004);
dr. Mindaugas Margelevičius (Vilnius University, Natural sciences, Biochemistry, N 004);
dr. Mindaugas Zaremba (Vilnius University, Natural sciences, Biochemistry, N 004).

The dissertation shall be defended at a public meeting of the Dissertation Defence Panel at 2 p.m. on 4th February 2022 in auditorium R-401 of the Vilnius University Life Sciences Center. Address: Saulėtekio Ave.7, Vilnius, Lithuania. Tel. +37062156841; povilas.gibas@bti.vu.lt.

The text of this dissertation can be accessed at the library of Vilnius University, as well as on the website of Vilnius University: www.vu.lt/lt/naujienos/ivykiu-kalendorius

https://doi.org/10.15388/vu.thesis.280
https://orcid.org/0000-0001-5311-6021

VILNIAUS UNIVERSITETAS

Povilas
Gibas

Statistiniai ir kompiuteriniai metodai didelės našos epigenominių duomenų analizei

DAKTARO DISERTACIJA

Gamtos mokslai,
Biochemija (N 004)

VILNIUS 2022

Disertacija rengta 2015–2020 metais Vilniaus universitete, Gyvybės mokslų centre, Biotechnologijos institute.

Mokslinius tyrimus rėmė Lietuvos mokslo taryba.

Disertacija ginama eksternu.

Mokslinis konsultantas – prof. dr. Saulius Klimašauskas (Vilniaus universitetas, gamtos mokslai, biochemija, N 004).

Gynimo taryba:

Pirmininkas – prof. dr. Saulius Serva (Vilniaus universitetas, gamtos mokslai, biochemija, N 004).

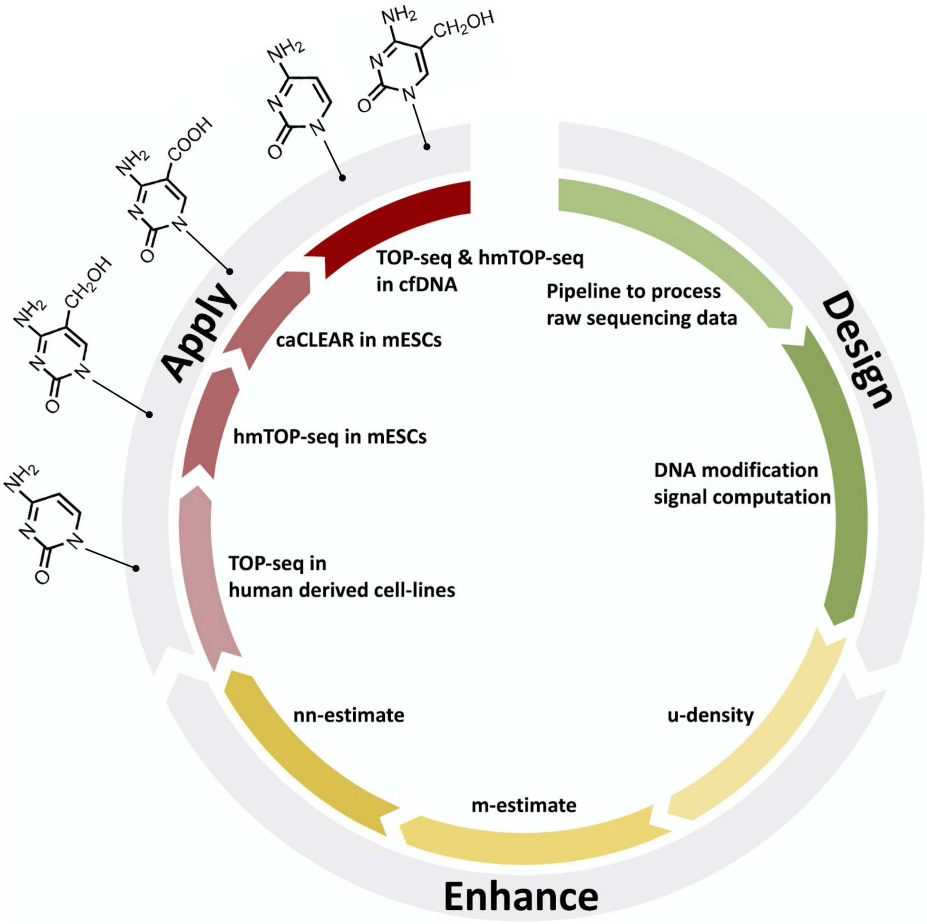
Nariai:
dr. Jonas Bačelis (Lietuvos statistika, gamtos mokslai, biochemija, N 004);
prof. Matthias Bochtler (Biochemijos ir biofizikos institutas, Lenkija, gamtos mokslai, biochemija, N 004);
dr. Mindaugas Margelevičius (Vilniaus universitetas, Gyvybės mokslų centras, gamtos mokslai, biochemija, N 004);
dr. Mindaugas Zaremba (Vilniaus universitetas, gamtos mokslai, biochemija, N 004).

Disertacija ginama viešame Gynimo tarybos posėdyje 2022 m. vasario mėn. 4 d. 14 val. Vilniaus universiteto Gyvybės mokslų centro R-401 auditorijoje. Adresas: (Saulėtekio alėja 7, LT-10257, Vilnius, Lietuva), tel. +37062156841; el. paštas povilas.gibas@bti.vu.lt.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje ir VU interneto svetainėje adresu: https://www.vu.lt/naujienos/ivykiu-kalendorius

To my parents, who gave me more than genes

Thesis at a Glance



Acknowledgements

I would like to express my sincere gratitude, first and foremost, to my supervisor Juozas Gordevičius. He served as a constant source of support and always advised me to do what is required as efficiently as possible so that time would remain to do what is meaningful. Mark Twain wrote – "Twenty years from now you will be more disappointed by the things you didn't do than by the ones you did. So throw off the bowlines, sail away from the safe harbor. Catch the trade winds in your sails. Explore. Dream.". This metaphor suits Juozas supervision very well as he is a passionate sailor and knows very well what it means to push off into open waters. I am lucky to have encountered such a dear colleague and caring supervisor.

I am deeply indebted to Saulius Klimašauskas who also supervised my PhD research. I would not have been able to achieve this thesis had he not graciously committed to sharing with me not only his considerable resources but also his invaluable experiences.

Immense gratitude goes to Edita Kriukienė who provided many inspiring ideas, impulse and support for my research. Edita treated me as a peer and this gave me the confidence to own up to my ideas. Her willingness to incorporate me into her research projects was crucial for me completing this thesis.

I was very fortunate to have worked with Artūras Petronis who proved to be a wonderful researcher. Discussions with him pushed the boundaries of my understanding of science and challenged me to think beyond the time.

The Department of Biological DNA Modification was a perfect environment to grow as a scientist. Fellow students and researchers not only shared excellent ideas, but they also provided good company along my journey.

I am particularly indebted to Petronis group at the Krembil Family Epigenetics Laboratory. I cannot stress how much joy I had working with this friendly company of researchers. I will always treasure those lunch breaks that could last for hours due to our lengthy discussions about unconventional science ideas. All the time that I spent with them is a great reminder that *work* can be entertaining.

I am very thankful to Juozas Lazutka who was my favorite professor at university. His lectures surpassed anything I could learn from a textbook. Of course, the entire department of Botany and Genetics was very important in my studies, and their classes remain the most important component of my education at university. Even though my current PhD is in biochemistry and bioinformatics, studying genetics under the supervision of this department had such a profound influence on me that, some days, I still consider myself more of a geneticist than a bioinformatician.

I would also like to thank Sonata Jarmalaitė whose lectures, guidance, and most importantly – trust, were very important for me.

During my undergraduate studies, I enjoyed attending lectures by Kastytis Beitas. His ideas on science and magic will not be forgotten.

I am eternally indebted to Jan Korbel from EMBL. While attempting to complete my bachelor's thesis, I did not have a place to perform my research, but after contacting him he graciously accepted me for an internship. As I reflect on that experience now, it feels almost like a dream. I was given the opportunity to learn so much about science, but more importantly, I was able to travel and to experience a different way of life. During my internship, I met Tobias Rausch, a brilliant programmer who supervised my work. In the end, my only regret is having met him in the early stages of my career. I am confident that with the skills and knowledge that I have now we could easily tackle those research problems that we were trying to solve. In that same department, I was able to meet and work alongside both Markus Fritz and Vladimir Benes, two brilliant scientists who welcomed me into their workspace.

Daumantas Matulis from Vilnius University was the first researcher who

supervised me. Not seeing the point of why I should wait until my third year to start an internship, I emailed him during my first month of studies. Daumantas asked me several basic interview questions which I failed to answer (this proved me why most students wait until their third year). Still, he accepted me for that internship. To this day, I do not know what he saw in me beyond, perhaps, my naive enthusiasm to get started. I would also like to thank Asta Zubrienė for supervising and for helping me with my first steps in scientific research.

During my time in Toronto, I did research under the supervision of Gabriel Oh, who is not only an inspiring researcher but one the best supervisors a person could have. Gabriel is fully committed towards improving his team through the individual development of each member. Thank you for your friendship and fair–minded criticism.

I also spent a couple of years doing an internship in Vaidutis Kučinskas lab from Vilnius University at the department of cytogenetics. I have lost count of how many happy hours I spent in that lab just looking through the microscope into the chromosomes.

I am very thankful to Igor Ulitsky from the Weizmann Institute of Science who supervised my master thesis. Clearly, I would not be at this stage of my career had he not generously offered me his time and support.

The research department at the University of Tartu was influential in my growth as a scientist. I want to especially thank Andres Metspalu and his team for allowing me to perform an internship at their department and for providing me with resources, knowledge, and guidance.

I would also like to thank Kiprianas Spiridonovas. Kiprianas was a true *techie*, one of a kind, who was a patient and understanding ally at my first job. I am very happy that I got to know and work with him. I was fortunate to have a very friendly colleague Algis Kriščiūnas with whom I spent hours traversing back and forth from work and home through the various Toronto districts. I would also like to mention Karolis Koncevičius, another brilliant and very kind colleague. Thank you both for your support and great discussions.

I will always remember and cherish the relationships I developed during my six years at university. Specifically, I fondly recall the time I spent with Kristina. *Čiurlionis* memories always bring happiness to my heart. My best friend and roommate at university was Povilas N.. I cannot express how much my personality changed while spending time with him. I arrived at university naive and somewhat oblivious. Meanwhile, Povilas was smart beyond his years. His approach to life and work influenced me a lot. Lina is another dear friend from university. Thank you for all the peace that you brought to my life. I must also mention Raimonda and Ruslanas, two good friends with whom I used to have lots of good time together.

I am lucky to have had some encounters where my supervisors have also become close allies. Tomer Gueta from the Israel Institute of Technology accepted me as a summer inter, but now he became a very good and supportive friend of mine.

Special recognition goes to Alexandra Elbakyan and Aaron Swartz. Without your sacrifice, my research and thesis would not have been possible.

I must also acknowledge all of my students. It is only through your eyes that I can see all that I do not yet know. They are my constant inspiration to improve.

I have come to admire the work of a number of researchers. Eva Jablonka from Tel Aviv University is an inspiration for me to pursue research in epigenetics. Mark Gerstein, Manolis Kellis, Lieberman Aiden are great researchers that I admire. With all my best I am just imitating the work you do.

The journey to a PhD rarely starts in university. For that reason, I am compelled to honor my favorite high school math teacher, Ms. Janė. She encouraged my scientific thinking, but more importantly, she helped and supported me throughout my angsty teenage years. With all my heart I am very thankful to you.

Most importantly, I want to thank my parents, Rita and Juozas, for loving and supporting me in everything I have ever attempted to do. I would also like to thank my sister, Ona, who means so much to me and

my dear Eugenia for bringing poetry to my life.
Finally, I would like to thank everyone that I forgot to mention here. There have been so many people who helped me in this journey and without you I would not be here today. Throughout all these years of education I have met a lot of people who have dramatically improved my outlook on the world. And now, as I enter this next stage of my journey, I hope to keep improving as a scientist, but more importantly, as a person.

. . .
Keep Ithaka always in your mind.
Arriving there is what you're destined for.
But don't hurry the journey at all.
Better if it lasts for years,
so you're old by the time you reach the island,
wealthy with all you've gained on the way,
not expecting Ithaka to make you rich.

Ithaka gave you the marvellous journey.
Without her you wouldn't have set out.
She has nothing left to give you now.

And if you find her poor, Ithaka won't have fooled you.
Wise as you will have become, so full of experience,
you'll have understood by then what these Ithakas mean.

Konstantinos P. Kavafis

Preface

This thesis consists of a selection of publications which summarise some of the research on computational genomics that I have conducted before, during and after my PhD studies (2014 — 2020). I performed most findings summarised in this work guided by my supervisors. Nonetheless, it should be noted that I did not perform any of the wet lab experiments and all the data I used was generated by my colleagues or was published a priori.

In most cases, the results are disseminated in peer–reviewed publications and some irrelevant details have been omitted. These peer–reviewed publications alone amount to 93 pages and their inclusion here would have rendered this thesis a somewhat lengthy document. Appropriate references necessary to locate them are given in the **Section 1.6** below. The chronological order in which the major parts of the works described in the publications below were performed was: I, i, II, ii. Nonetheless, most of the research was intervened and usually two or three experiments were conducted simultaneously, therefore there is no clear distinction between the publications in this thesis; usually, their results or methodology are summarised in the same sections.

Finally, as advised by D. Knuth, I have decided not to use the first person voice in this thesis, only using it in the introductory sections (Knuth et al., 1989). The word "we" is used instead to avoid a passive voice and emphasise that I am just one of many researchers that contributed to the previously mentioned publications.

Contents

List of Figures

List of Tables

Abbreviations

5caC	**5**–**ca**rboxyl**c**ytosine
5fC	**5**–**f**ormyl**c**ytosine
5hmC	**5**–**h**ydroxy**m**ethyl**c**ytosine
5mC	**5**–**m**ethyl**c**ytosine
ANN	**A**rtificial **N**eural **N**etwork
AUC	**A**rea **U**nder the **C**urve
BGT	β–**g**lucosyl**t**ransferase
bp	**B**ase **P**air
BS	**B**isulfite **S**equencing
caCLEAR	**ca**C **C**learance
cfDNA	**C**ell–**F**ree **DNA**
cffDNA	**C**ell–**F**ree **F**etal **DNA**
CG	**C**ytosine **G**uanine Dinucleotide
CGI	**C**ytosine **G**uanine **I**sland
CV	**C**horionic **V**illi
DMR	**D**ifferentially **M**odified **R**egion
DNMT	**DN**A **M**ethyl**t**ransferase
DR	**D**imensionality **R**eduction
ENET	**E**lastic **Net**
FDR	**F**alse **D**iscovery **R**ate
GLM	**G**eneralised **L**inear **M**odel
hmTOP–seq	5**hm**C **T**ethered **O**ligonucleotide **P**rimed **Seq**uencing
kb	**K**ilo **B**ase
KDE	**K**ernel **D**ensity **E**stimation
LAD	**L**amina **A**ssociated **D**omain
lincRNA	**L**ong **I**ntergenic **N**on–**C**oding **RNA**
LINE	**L**ong **I**nterspersed **N**uclear **E**lement
LTR	**L**ong **T**erminal **R**epeat

M	**M**illion
Mb	**M**ega**b**ase
MBD	**M**ethyl–**B**inding **D**omain
MeDIP	**M**ethylated **D**NA **I**mmuno**p**recipitation
NB	**N**euro**b**lastoma
ng	**N**ano**g**ram
NIPT	**N**ot **I**vasive **P**renatal **T**est
nMDS	**N**on **M**etric Multi**d**imensional **S**calling
NPC	**N**ot **P**regnant **C**ontrol
***m*–estimate**	**M**ethylation **Estimate**
mESC	**M**ouse **E**mbryonic **S**tem **C**ell
mQTL	**M**ethylation **Q**uantative **T**rait **L**oci
***nn*–estimate**	**N**eural **N**etwork Methylation **Estimate**
OR	**O**dds **R**atio
PCA	**P**rincipal **C**omponent **A**nalysis
PCR	**P**olymerase **C**hain **R**eaction
sd	**S**tandard **D**eviation
SINE	**S**hort **I**nterspersed **N**uclear **E**lement
SNP	**S**ingle **N**uncleotide **P**olymorphisms
TAB–seq	**T**ET–**A**ssisted **B**isulfite **S**equencing
T21	Chromosome **21** **T**risomy
TDG	**T**hymine **D**NA **G**lycosylase
TET	**T**en–**E**leven **T**ranslocation Methylcytosine Dioxygenase
Tet TKO	**Tet**1/2/3 **T**riple **K**nock**o**ut
TOP–seq	**T**ethered **O**ligonucleotide **P**rimed **Seq**uencing
TSS	**T**rascription **S**tart **S**ite
***u*–density**	**U**nmethylated DNA **Density**
UTR	**U**n**t**ranslated **R**egion
WGBS	**W**hole–**G**enome **B**isulfite **S**equencing
WT	**W**ild **T**ype
WRSC	**W**eighted **R**ank **S**election **C**riterion

Introduction

1.1 Study Rationale

Epigenetic control mechanisms, such as DNA modifications, play important roles in practically all living organisms in regulating various cellular, developmental, and behavioural processes. Despite the importance of DNA modification in biology, the many difficulties associated with identifying and characterising epigenetic profiles discourages researchers from pursuing this avenue of investigation.

There are two major obstacles to widespread research of epigenetic regulation effects. First, the methods to quantify the epigenetic state genome–wide, such as whole genome bisulfite sequencing (WGBS), are very expensive and generate enormous amounts of data. Second, these methods cannot optimally distinguish between different types of DNA modifications, such as 5–methylcytosine (5mC), 5–hydroxymethylcytosine (5hmC), and 5–carboxylcytosine (5caC). Furthermore, while WGBS is the most widely used and accepted as *the gold standard*, it suffers from experimental artefacts due to extensive DNA degradation and obstructed genomic mapping of sequencing reads. Due to harsh conditions required for bisulfite conversion, 90% of the template DNA can be lost (Grunau et al., 2001). Most importantly, whole genome bisulfite sequencing unavoidably generates large amounts of data as it requires blind sequencing of the whole genome, even loci lacking cytosines, and most reads (50% — 80%) provide little to no information about DNA modification (Ziller et al., 2013).

Tethered oligonucleotide–primed sequencing (TOP–seq) was the first method to use covalent tagging of individual unmodified CG sites, followed by priming of the DNA polymerase at these positions to mark sites

for sequencing and precise genomic mapping. The key novel aspect of the TOP–seq method is the combination of enrichment for unmodified DNA fraction, single base resolution and strand specificity. hmTOP–seq is a single nucleotide resolution 5hmC profiling method which is based on direct sequence readout primed at covalently labelled 5hmC sites from an in situ tethered DNA oligonucleotide. Finally, caC–Clearance (caCLEAR) is bisulfite–free, single nucleotide resolution method that enables targeted genome–wide mapping of 5caC residues. These new methods can help to profile genome–wide single nucleotide epigenetic maps with fewer resources than whole genome bisulfite sequencing.

However, to employ these techniques at their full potential, a collection of appropriate statistical and computational techniques is needed. This research presents a set of solutions to solve new challenges arising from highly–specific type of TOP–seq data. Moreover, it offers multiple applications of TOP–seq data varying from differentially modified region (DMR) identification, production of epigenomic maps, and signal normalisation by genomic context.

In essence, all the work that is presented here follows a narrative of three main parts: *design*, *enhance*, *apply*. In *design* part we introduce a data processing pipeline that transforms raw TOP–seq epigenomic data to a CG–coverage signal. In *enhance* part we propose and integrate three signal transformations that can greatly improve enrichment–based epigenomics signal. Finally, in *apply* part we present multiple case–studies where TOP–seq signal can be used to retrieve biological information.

1.2 Thesis Layout

First, the **Chapter 2** reviews the biological, technological, and statistical concepts that form the basis of this work, presenting the cellular memory hypothesis, C. H. Waddington's epigenetic landscape, DNA modification forms and molecular mechanism behind establishing or removing them.

Next, the current most widely used technologies to profile DNA modifications are discussed, closing this section with the statistical methods applied in this work, namely, linear regression, Kernel density estimation, dimensionality reduction techniques, and artificial neural networks. The **Chapter 3** introduces the general methods, genomic datasets and computational tools used in this research.

Chapter 4 presents the methodology and set of computational tools needed for efficient processing of TOP–seq sequencing data. This methodology covers read processing from the raw data format to modification level evaluation per each CG site. **Chapter 5** discusses the statistical methodologies developed to enhance TOP–seq sequencing signal, presenting statistical techniques such as Kernel density estimation, regression and artificial neural networks.

Chapter 6 details the application of TOP–seq based sequencing methods, presenting a detailed analysis of TOP–seq application to decipher epigenetic differences between various human derived cell–types. A brief introduction to the sequencing processing results for hmTOP–seq and caCLEAR methods follows. Finally and most importantly, the chapter concludes with a case study of the application of TOP–seq and hmTOP–seq methods to investigate DNA modifications in cell–free DNA (cfDNA) samples from pregnant female. An approach is presented to identify differentially modified regions specific to the fetus with a trisomy of chromosome 21. Also, the statistical techniques that can be used to estimate the fetal fraction using TOP–seq sequencing data are detailed.

1.3 Aim and Objectives

The overarching aim of the work described in this thesis was to develop a set of statistical and computational tools tailored for the analysis of TOP–seq based high–throughput epigenomic data and to apply these tools in experimental settings to gain biological knowledge. To achieve this aim, the following objectives were set:

- Develop computational methods for efficient and accurate processing of TOP–seq sequencing reads.
- Develop statistical learning techniques to enhance the quality of the TOP–seq signal in the presence of technical and biological noise.
- Apply the developed methods and techniques to compare different DNA modifications across genomic elements.
- Identify differentially modified regions across samples pertaining to distinct experimental groups using TOP–seq based high–throughput epigenomic data.

1.4 Statements to Be Defended

- The developed computational methods can be used to efficiently and accurately process TOP–seq based high–throughput epigenomic data.
- The developed statistical learning techniques can be used to enhance the quality of the TOP–seq epigenomic signal in the presence of technical and biological noise.
- TOP–seq, hmTOP–seq and caCLEAR methods provide information about DNA modification signal across different genomic elements.
- TOP–seq method could be used to identify differentially modified regions across samples pertaining to different tissues or cell–types.
- TOP–seq and hmTOP–seq methods could be used to identify fetal abnormalities in maternal cell–free DNA.

1.5 Scientific Novelty and Practical Value

The most prominent novel aspect of this work is the development and application of statistical and computational methods for the analysis and evaluation of DNA modifications in TOP–seq, hmTOP–seq and caCLEAR derived high–throughput epigenomic datasets. This work provides a detailed step–by–step description of the read processing pipeline developed specifically for TOP–seq sequencing data which can be used by other researchers to assure the accuracy and quality of the TOP–seq method. Additionally, we demonstrate library quality parameters, such as read length, CG content, read mapping rate, and distance to a CG site.

This thesis also describes a novel computational genomics approach for sequencing data transformation to enhance the coverage signal, u–density improves the accuracy of the coverage signal by leveraging modification information from the neighbouring CG sites and normalising the signal for CG–content. A novel approach for the normalisation of bandwidth selection was proposed, developed and tested. Furthermore, two supervised DNA modification transformation approaches were also designed and implemented in this thesis. This framework exploits TOP–seq data and genomic context information to estimate underlying DNA modification levels. A small fraction of the whole genome bisulfite sequencing dataset was used to train an exponential decay or an artificial neural network model which was then used to convert the TOP–seq signal into the so called CG methylation estimate signal. These enhancements transformed the relative coverage signal into an absolute scale which greatly increased correlation with a reference dataset and allowed easier signal interpretation. Our study serves as a starting point for further research to use genomic information for the coverage signal enhancement — determined genomic covariates to enhance coverage signal are provided in this thesis and could be used by other researchers.

The research presented in this thesis began with the optimised processing of the TOP–seq based high–throughput epigenomic data. Eventually,

this advancement allowed us to apply our developed methods and open new venues for a larger scale epigenomic studies. This thesis presents the first detailed study of fetal unmodified and hmC modified CG sites in maternal cell–free DNA for not invasive prenatal testing (NIPT). For the first time, we investigated the unmodified DNA fraction in chorionic villi tissue samples and compared it to a cell–free fetal DNA (cffDNA) or DNA from non–pregnant control samples. Furthermore, the methodology to identify DMRs was introduced as a promising strategy to determine fetal karyotypes. This genome–wide TOP–seq based DNA modification profiling performed in healthy and chromosome 21 trisomy positive samples led to the identification of a set of novel putative biomarkers with diagnostic value. The differentially modified regions obtained in the present study may assist in the selection of suitable diagnostic regions in a particular clinical context. It is anticipated that this method might even surpass currently available NIPT tests. Moreover, the investigation of the fetal–fraction prediction was introduced as a promising strategy to determine the amount of fetal DNA in maternal blood samples.

In addition to the practical value presented in this thesis, it is also scientifically novel. This study involved the biological systems of human tissues (e.g., chorionic villi), cancerous cell–types, mouse embryonic stem cells, and cell–free fetal DNA. DNA modification maps, both for unmodified, 5hmC modified, and 5caC modified DNA were generated and enriched. Profiles of DNA modifications were generated for genes, enhancers, and large epigenomic structures like lamina associated domains.

1.6 Approbation of the Research Results

This thesis is mainly based on the following publications:

(I) Staševskij Z.[*], **Gibas P.**[*], Gordevičius J., Kriukienė E., Klimašauskas S.; *Tethered Oligonucleotide–Primed Sequencing, TOP–Seq: A High–Resolution Economical Approach for DNA Epigenome*

[*]Shared first co–author

Profiling; **Molecular Cell**; 2017 Feb 2; 65(3):554-564.e6.

(II) Gordevičius J., Narmontė M., **Gibas P.**, Kvederavičiūtė K., Tomkutė V., Paluoja P., Krjutškov K., Salumets A, Kriukienė E.; *Identification of fetal unmodified and 5–hydroxymethylated CG sites in maternal cell–free DNA for non–invasive prenatal testing*; **Clinical Epigenetics**; 2020 Oct 20; 12(1):153.

Additional publications to which I contributed during my PhD studies:

(i) **Gibas P.***, Narmontė M.*, Staševskij Z., Gordevičius J., Klimašauskas S., Kriukienė E.; *Precise genomic mapping of 5–hydroxymethylcytosine via covalent tether–directed sequencing*; **PLoS Biology**; 2020 Apr 10; 18(4):e3000684.

(ii) Ličytė J.*, **Gibas P.***, Skardžiūtė K., Stankevičius V., Rukšėnaitė A., Kriukienė E.; *A bisulfite–free approach for base–resolution analysis of genomic 5–carboxylcytosine*; **Cell Reports**; 2020 Sep 15; 32(11):108155.

(iii) Carlucci M., Kriščiūnas A., Li H., **Gibas P.**, Koncevičius K., Petronis A., Oh G.; *DiscoRhythm: an easy–to–use web application and R package for discovering rhythmicity*; **Bioinformatics**; 2019 Nov 8; 36(6):1952-1954.

(iv) Daniūnaitė K., Dubikaitytė M., **Gibas P.**, Bakavičius A., Lazutka R. J., Ulys A., Jankevičius F., Jarmalaitė S.; *Clinical significance of miRNA host gene promoter methylation in prostate cancer*; **Human Molecular Genetics**; 2017 Jul 1; 26(13):2451-2461.

Author's contribution to the listed publications:

(I) I created a pipeline for TOP–seq high–throughput epigenomic data processing; completed the TOP–seq signal quality control on a model bacterial genome; performed sequencing data analysis for all sequenced human derived samples; calculated CG–coverage and DNA modification signal; suggested, developed and fully implemented signal enhancement techniques, such as u–density and m–estimate; calculated genomic–element covariates that could be used

to adjust high–throughput epigenomic signal; identified differentially modified regions between various sample groups and performed downstream ontological gene analysis; calculated the DNA modification signal in various genomic elements; participated in the interpretation and discussion of results, contributed to reviewing the manuscript; created the main data visualisations (figures 2 — 6); deployed data to the Gene Expression Omnibus database.

(II) Analysed sequencing data for all TOP–seq and hmTOP–seq samples; created epigenomic DNA modification maps (CG–coverage and identified CG–fraction) for all samples; calculated the DNA modification signal in various genomic elements; suggested and applied the technique to identify differentially modified regions across samples pertaining to distinct experimental groups; validated and identified differentially modified regions using the cross–validation technique; performed fetal–fraction estimation analysis; participated in the interpretation and discussion of the results, contributed to drafting the manuscript; created all main data visualisations; deployed data to the Gene Expression Omnibus database.

(i) Analysed hmTOP–seq sequencing data for all sequenced samples; created the hmTOP–seq data analysis pipeline adjusted for strand specific analysis; created the hmTOP–seq data analysis pipeline for DNA modification identification in a non–CG context; calculated the DNA modification signal in various genomic elements; participated in the interpretation and discussion of results, contributed to drafting the manuscript; created the main data visualisations (figures 2 — 4); deployed data to the Gene Expression Omnibus database.

(ii) Analysed caCLEAR sequencing data for all sequenced samples; calculated the DNA modification signal in various genomic elements; participated in the interpretation and discussion of results, contributed to drafting the manuscript; created the main data visualisations (figures 3 — 5); deployed data to the Gene Expression Omnibus database.

(iii) Contributed to developing `R shiny` data analysis platform code base; participated in reviewing the manuscript and creating data visualisations.

(iv) Contributed to bioinformatical data analysis; contributed to the first draft of the manuscript.

List of conferences related to the thesis:

(a) Oral presentation; Staševskij Z., **Gibas P.**, Gordevičius J., Kriukienė E., Klimašauskas S.; *High–Throughput Data Analysis Workflow for Large Scale Epigenome Profiling*; **NGS'17**; Barcelona, Spain; April 3 — 5 2017.

(b) Poster presentation; Šarakauskas M., **Gibas P.**, Gordevičius J.; *Estimation of DNA modification using artificial neural networks, TOP–seq data and genomic context information*; **International work–conference on bioinformatics and biomedical engineering**; Granada, Spain; April 25 — 28 2018.

(c) Oral presentation; **Gibas P.**, Šarakauskas M., Gordevičius J., Kriukienė E., Klimašauskas S.; *Estimation of DNA modification using artificial neural networks, TOP–seq data and genomic context information*; **Bioinformatics and Computational Biology Conference**; Naples, Italy; November 19 — 21 2018.

Theoretical Foundations for the Scientific Problem

2.1 Biological Aspects

2.1.1 A Brief Guide to Epigenetics

Every known living organism contains genetic information that tends to be transmitted to the progeny of cells or organisms. Usually, the definition of this genetic information is crystallised down to a DNA sequence that is stored in the cell nucleus [1]. This DNA sequence is composed of four canonical nitrogenous bases which are studied in sufficient detail and out of the 3.8×10^{13} cells estimated to be in the human body, the majority of them will contain the same DNA sequence (Sender et al., 2016). Interestingly, despite arising from a single fertilised oocyte and being almost genetically identical, these cells vary in their phenotypes – function, gene expression intensity and manage to make up various tissues and organs (Moris et al., 2016). The main reason why this primary DNA sequence can be interpreted differently is due to various epigenetic factors – mechanisms that allow adapting to the changing environment without changing the base composition. There are many epigenetic factors known to date, some of which are ubiquitous to most life forms, while others are restricted to only some species, and new epigenetic players are continually being discovered (Wu et al., 2016).

These principal epigenetic modalities are DNA modifications, histone

[1] However, at least in mammals, the term "genetic information" tends to be much broader due to mitochondrial DNA, double minute chromosomes, circulating cell–free DNA, and circulating cell–free RNA.

modifications, non–coding RNAs (such as long non–coding RNAs, microRNAs, enhancer RNAs), and RNA modifications (i.e., epitranscriptome) (Morris and Mattick, 2014; Suganuma and Workman, 2011; Wiener and Schwartz, 2020). This union of DNA sequence and epigenetic factors may be defined as chromatin (Allis and Jenuwein, 2016). It is important to note that the genome–wide pattern of epigenetic factors is highly environment–specific (e.g., tissue or time) because it reflects the function or current state of the cell while the DNA sequence is relatively stable and changes at a very different pace (Kundaje et al., 2015; Narasimhan et al., 2017; Oh et al., 2018).

DNA modifications refer to a covalent modification of the bases directly in the DNA sequence by the addition of various chemical groups at strictly defined positions without changing the sequence itself (Razin and Riggs, 1980). The most widespread form of DNA modification is the methylation of cytosine at the fifth carbon atom of the pyrimidine ring, which is commonly known as 5–methylcytosine. Other DNA modifications may include 5–hydroxymethylcytosine, 5–formylcytosine (5fC), 5–carboxycytosine, N4–methylcytosine and N6–methyladenine (Kumar et al., 2018).

2.1.2 Cellular Memory Hypothesis

The coining of the term "epigenetics" originates from "epigenesis" and should be attributed to Conrad H. Waddington (Waddington, 1942). Waddington used "epigenetics" to describe the process of "epigenesis" which was the prevailing theory of how a fertilised oocyte progressed into a complex organism. Waddington introduced an incredibly compelling way to describe the concept of the epigenetic landscape in which a cell traverses a canalised landscape and once a route is chosen, then the cell must continue down the chosen path until its ultimate fate (**Figure 2.1**) (Waddington, 1957).

The last decade was outlined by the discovery of a great variety of *epigenetic players* that affect gene expression, such as histone modifications

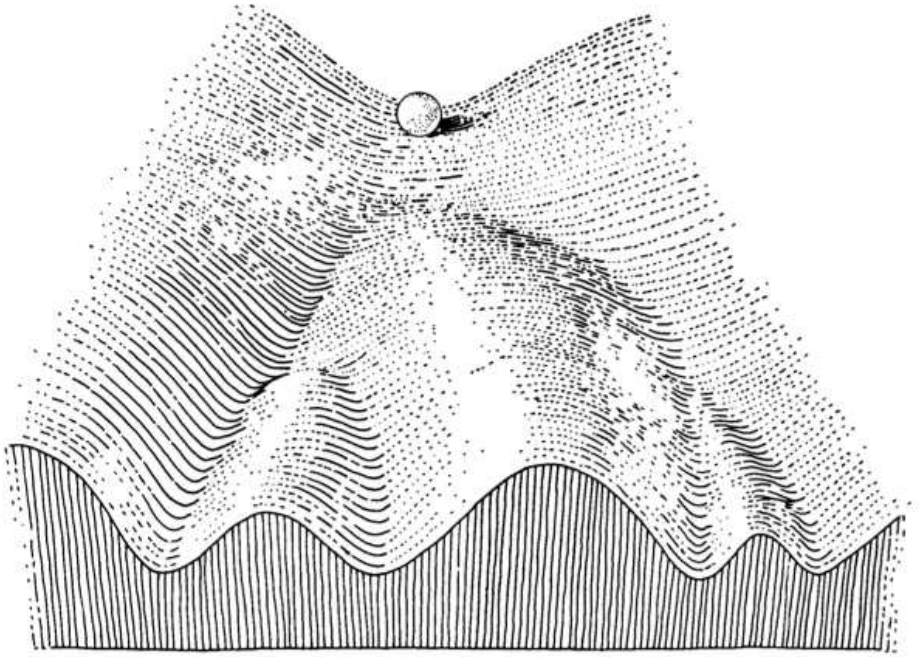


Figure 2.1 | The Epigenetic Landscape

This classical drawing by Conrad H. Waddington illustrates the process of regulation during organisms development as a landscape (Waddington, 1957). This landscape consists of valleys separated by ridges on an inclined surface through which the cell traverses on its way from undifferentiated to a fully differentiated form. A branching pathway is used as a visual metaphor that depicts a decision point of cell fate determination.

or non–coding RNAs, which culminated with multiple consortia. ENCODE is a multi–phase research project that aims to identify functional elements in the human genome, with the most recent phase expanding to other organisms and covering many different cell types (Dunham et al., 2012; Snyder et al., 2020). The National Institutes of Health Roadmap consortium aims to produce a collection of epigenomic maps for stem cells and primary tissues selected to represent the normal counterparts of tissues and organ systems frequently involved in human disease (Bernstein et al., 2010; Kundaje et al., 2015). The European Union funded BLUEPRINT consortium focused on distinct types of haematopoietic cells from healthy individuals and diseased counterparts to advance and exploit knowledge of the underlying biological processes and mechanisms

(Stunnenberg et al., 2016). Though epigenetics has been continuously redefined to accommodate ever–increasing knowledge, there is still a debate in the scientific community about the appropriate definition of epigenetics (Bird, 2007; Greally, 2018; Haig, 2004; Russo et al., 1996). For the purpose of this thesis, epigenetics is defined as: "The study of mitotically and / or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence".

To summarise, one can rephrase that conceptually, epigenetics provides an explanation of how cells interpret their genetic blueprint under the light of changing time, space, and other stochastic factors (Fraga et al., 2005). For example, thousands of cell subtypes of a multicellular organism can have variations in the readout of their genetic template in response to a large number of internal and external factors. Epigenetics thus connects DNA sequence and environmental influences to the phenotype.

2.1.3 5–methylcytosine

5–methylcytosine has been termed as the *fifth base* of the genome, as reflected by its high abundance across all domains of life — 5mC is present in most eukaryotes, including vertebrates, invertebrates, plants, fungi, and many prokaryotes (Bewick et al., 2019; Goll and Bestor, 2005; Keller et al., 2016; Suzuki and Bird, 2008). In the human genome, 5mC comprises 1% to 4% of the cytosine residues (Breiling and Lyko, 2015; Ehrlich et al., 1982).

In mammalian genomes, 5–methylcytosine is predominantly found to precede a guanine base (commonly known as a CG dinucleotide) (Bernstein et al., 2007; Doskocil and Sorm, 1962). The fact that CG dinucleotide is palindromic [2] is directly linked to the propagation of the modification patterns through the cell divisions. Interestingly, these CG

[2]Mirrors itself in an anti–parallel fashion on both forward and reverse strands of the same DNA molecule.

sites tend to be underrepresented across the genome since the CG dinucleotide is prone to mutate (i.e., mutational hotspot) — methylated cytosines have high rate of spontaneous deamination which results in C to T transitions [3] (Duncan and Miller, 1980). As a consequence, CG dinucleotides are drastically underrepresented in the vertebrate genomes and occur at only 20% — 25% of the expected frequency (Lander et al., 2001; Saxonov et al., 2006; Swarts et al., 1962). Therefore, variation in the methylation level within the genome causes variation in the local CG sites density.

In the haploid human genome there are around 28 million CG sites, of which, 60% — 80% are modified (Smith and Meissner, 2013; Zhao et al., 2014). The density of these CG dinucleotides throughout the genome is not even since CG sites tend to form clusters known as CG islands (CGI). CGIs are usually defined as DNA segments that are longer than 200 base pairs (bp), have a G + C content of 50% or higher, and a CG frequency of at least 0.6 (Gardiner-Garden and Frommer, 1987; Illingworth and Bird, 2009). There are ∼27.7 thousands CGIs in the human genome [4] (average size ∼750 bp with 1st and 3rd quartiles at ∼320 bp and ∼950 bp respectively), while in the mouse genome there are ∼16 thousand CGIs (average size ∼650 bp with 1st and 3rd quartiles at ∼330 bp and ∼820 bp respectively). CGIs are frequently associated with gene upstream regions but a significant fraction of them are found within gene bodies (i.e., intragenic CGIs) (Medvedeva et al., 2010; Saxonov et al., 2006). Around 70% of protein–coding genes have at least one CGI in their upstream region (usually two kilobases (kb) in size) and CG sites in these CGIs tend to be unmethylated, which contrasts with the high CG modification signal in the rest of the genome (Jones, 2012; Saxonov et al., 2006). These genes correspond to nearly all housekeeping genes, 93% of the genes expressed during embryogenesis, and 40% of tissue–specific genes (Larsen et al., 1992; Ponger et al., 2001). However, about half of CGIs in mammalian genomes are not associated with a known gene

[3] The transition rate of CG dinucleotides to TG, or CA on the reverse strand, is approximately twelve times the normal transition rate. Interestingly, it is much higher in human germline cells (Sved and Bird, 1990).

[4] However, it is reported that the actual number may be close to 50 thousand (Lander et al., 2001).

and have been termed *orphan CGIs* due to the uncertainty surrounding their origin (Sarda and Hannenhalli, 2018). Nevertheless, it seems that despite not being associated with specific genes, many *orphan CGIs* are actual sites of transcription initiation for unannotated protein–coding genes, non–coding RNAs or acting as enhancers (Bell and Vertino, 2017; Illingworth et al., 2010; Koerner et al., 2019; Sarda and Hannenhalli, 2018). Following the marine metaphor of the CGI, one can find the CGI shore from 0 to 2 kb on either side of a CGI, and the CGI shelf from 2 to 4 kb on either side of a CGI [5] (Bibikova et al., 2011; Price et al., 2013; Sandoval et al., 2011).

It is worth mentioning that 5mC in mammalian genomes is also found within a non–CG context (i.e., CHG or CHH sites, where H can be any nucleotide except for guanine), where it can comprise 1% to 25% of all 5–methylcytosines [6] (Guo et al., 2014a; Laurent et al., 2010; Lister et al., 2009). It was discovered that a relatively high fraction of non–CG methylation occurs in human and mouse brain or embryonic stem cells (Guo et al., 2014b; He and Ecker, 2015; Xie et al., 2012). The high abundance of non–CG modifications in the mentioned cell types and organs can be attributed to the increased expression of enzymes that can introduce this type of modification (Lister et al., 2013).

2.1.4 Mechanisms to Introduce DNA Modifications

Multiple proteins catalyze the establishment, maintenance, and removal of DNA modifications in organisms. In mammals, DNA methylation is established during embryonic development by the enzymes of the DNA methyltransferase (DNMT) family that catalyse DNA methylation by transferring the methyl–group of *S*–adenosyl–L–methionine to a cytosine base (Bestor, 2000; Bird, 2002). There are three main known DNMT enzymes that methylate CG sites in mammalian genomes – DNMT1,

[5] While *open sea* is used to refer to CG sites that do not fall into these categories and interestingly, these CG sites are mostly modified in human somatic cells (Bird, 2002).

[6] In plants, non–CG methylation can make up 25% — 50% of all 5mC (Bouyer et al., 2017).

DNMT3A, and DNMT3B (Lyko, 2018). Although these proteins are similar, they are activated during the different time points of organism's lifetime and perform different functions (Dahlet et al., 2020; Lyko, 2018). Eukaryotic DNA methyltransferase (DNMT1) was discovered first and then cloned from the human and mouse (Bestor et al., 1988; Gruenbaum et al., 1982; Yen et al., 1992). Later, another family of DNMT enzymes was discovered, which included two *de novo* DNMTs – DNMT3A and DNMT3B (Okano et al., 1999).

Expression of the maintenance methyltransferase DNMT1 peaks in the S phase, where it accordingly methylates DNA during cell division (Kishikawa et al., 2003; Lei et al., 1996). DNMT1 recognises hemimethylated DNA (i.e., when only one of the DNA strands of the CG site has 5mC) during DNA replication, and modifies cytosines on the newly synthesised DNA (Probst et al., 2009). This function provides a mechanism for maintaining DNA modification patterns during cell division, therefore making it a true epigenetic mark capable of generating cellular memory as originally hypothesised in 1975 (Holliday and Pugh, 1975; Li and Zhang, 2014; Riggs, 1975; Vilkaitis et al., 2005). DNMT1 is crucial for normal development as its deletion in mice results in a drastic loss of DNA methylation at a global level and is lethal to the organism (Brown and Robertson, 2007; Kurihara et al., 2008). DNMT1 also plays a role in imprinting mechanisms by supressing either the paternal or maternal copy of DNA, and is involved in DNA repair (Branco et al., 2008; Ha et al., 2011).

In mammals, DNA methylation patterns are rapidly erased immediately after fertilisation (Reik et al., 2001). *De novo* DNA methylation does not occur until the blastocyst stage when these patterns are re–established and eventually cell type specific methylation patterns are generated (Atlasi and Stunnenberg, 2017; von Meyenn et al., 2016). This *de novo* DNA methylation is largely established by DNMT3A and DNMT3B enzymes, which catalyse the addition of methylation marks in those CG sites that originally lack the modification in any of the two DNA strands (Okano et al., 1999; Reik et al., 2001; Saitou et al., 2012). Both enzymes

have similar domain arrangements but exhibit divergence in their N–terminal regions which carry a number conserved motifs governing their interactions with chromatin and other cellular proteins (Chédin, 2011; Gao et al., 2020). While DNMT3A is ubiquitously expressed in most cell types, DNMT3B is specifically expressed only in differentiating cells (Watanabe et al., 2002). Both enzymes are crucial for viable organisms while individuals with mutated variants exhibit pathological phenotypes (e.g., Tatton–Brown–Rahman syndrome, microcephalic dwarfism, ICF syndrome) (Jiang et al., 2005; Norvil et al., 2019; Nowialis et al., 2019). There is also evidence that DNMT3A and DNMT3B methylate non–CG sites (Laurent et al., 2010; Ramsahoye et al., 2000).

It is worth mentioning that another *de novo* DNA methyltransferase that is catalytically inactive, DNMT3L, has also been identified in mammals (Aapola et al., 2000). DNMT3L is a variant of the previously mentioned DNMT3 enzymes that lacks the N–terminal part of the regulatory domain and the C–terminal region of the catalytic domain (Lyko, 2018). DNTM3L does not bind DNA as strongly as other DNMT3 enzymes but directly interacts with both of them and enhances their activity (Chen et al., 2005; Suetake et al., 2004). DNMT3L cooperates mostly with DNMT3A to establish maternal imprints and its deletion can lead to embryo development abnormalities (Arima et al., 2006; Veland et al., 2019). Additionally, DNMT3C was recently discovered as a *de novo* DNA methyltransferase in several rodent species [7], where it plays a role in modifying young repeat elements during spermatogenesis (Barau et al., 2016). Lastly, DNMT2 is the most conserved member of the DNA methyltransferase family that acts as a transfer RNA methyltransferase and influences intergenerational epigenetic inheritance through sperm non–coding RNAs (Goll et al., 2006; Schaefer et al., 2010; Zhang et al., 2018).

[7] *Dnmt3C* arose through a duplication of *Dnmt3B* that occurred in the last common ancestor of muroid rodents (Molaro et al., 2020).

2.1.5 Mechanisms to Remove DNA Modifications

In the fertilised mammalian oocyte, the paternal pronucleus undergoes extensive genome–wide DNA demethylation hours after fertilisation before the first round of DNA replication commences, suggesting the existence of an active 5mC removal mechanism (Oswald et al., 2000). Albeit maternal pronucleus undergoes passive global methylation dilution, taking place over the first replication cycles (Howell et al., 2001; Rougier et al., 1998), both pronuclei lose almost most of the 5mC marks and preimplantation blastocysts contain only 25% of global methylation level (Lee et al., 2014). Such removal of 5mC marks can occur through either an active or passive mechanism (Saitou et al., 2012).

Passive demethylation is realised by the transcriptional or functional inhibition of DNMT1 (Kagiwada et al., 2013). In the absence of functional DNMT1, the newly synthesised DNA strand remains unmethylated, thus, methylated cytosine marks will be diluted in daughter cells resulting in gradually decreasing methylation levels (Sharif et al., 2007). Since this process is sequence independent, it results in genome–wide demethylation.

Active demethylation can occur locally in both differentiating and non–differentiating cells through a series of targeted chemical reactions via ten–eleven translocation methylcytosine dioxygenase (TET) enzymes (Branco et al., 2011; Gu et al., 2011b; Ito et al., 2010). TET enzymes catalyse the oxidation of 5mC to 5hmC, which is further oxidised to 5fC and subsequently to 5caC, in reactions that are probably also catalysed by TET family members (**Figure 2.2**) (Ito et al., 2010). These oxidised 5hmC variants can be diluted during replication or recognised and removed by thymine DNA glycosylase (TDG) to produce an abasic site [8] and later repaired by the base excision repair apparatus (He et al., 2011; Ito et al., 2010). TDG specifically recognises both 5fC and 5caC bases (but neither 5mC or 5hmC), suggesting a mechanistic link for both TET and TDG

[8] A site in DNA that consists of a deoxyribose unit lacking a purine or pyrimidine base.

to active DNA demethylation (He et al., 2011; Maiti and Drohat, 2011; Zhang et al., 2012). TDG–deficient embryos are non–viable and characterised by severe epigenetic abnormalities, highlighting the importance of TDG in normal development (Cortellino et al., 2011).

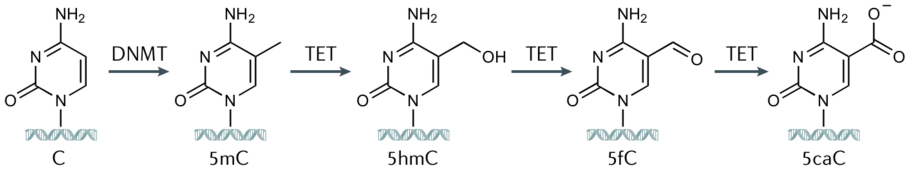


Figure 2.2 | DNA Modifications

DNA methyltransferases catalyse the methylation of cytosine to produce 5–methylcytosine. Downstream oxidation by ten–eleven translocation enzymes produces 5–hydroxymethylcytosine, 5–formylcytosine and 5–carboxylcytosine. Illustration was taken and modified from Parry et al. (Parry et al., 2021).

2.1.6 Other DNA Modifications

Through the chain of active chemical reactions, 5mC modification is oxidised to its consecutive forms. In 2009, two groups conclusively confirmed that 5hmC exist in mouse Purkinje cells, granule cells, and ESCs with levels ranging from 0.03% (in ESCs) to 0.6% (in Purkinje cells) of the total nucleotides (Kriaucionis and Heintz, 2009; Tahiliani et al., 2009) [9]. Furthermore, it was also demonstrated that TET1 oxidises 5mC to 5hmC, the first and most stable intermediate in the demethylation process (Ito et al., 2010; Tahiliani et al., 2009). Since then, 5hmC has been identified in a wide variety of mammalian tissues and cell lines, suggesting that 5hmC may be a biologically relevant epigenetic modification (Cui et al., 2020; Kriukienė et al., 2012; Li and Liu, 2011).

More important is the discovery that the hmC values deviate strongly in contrast to the stable amounts of 5mC (Globisch et al., 2010). The highest levels (0.3%–0.7%) of 5hmC are detected in the central nervous

[9] Although, 5hmC presence in rat has been detected almost for 40 before (Penn et al., 1972)

system and spinal cord (∼0.5%). These data strengthen the observation that nervous system tissues contain the highest levels of 5hmC. Medium levels of 5hmC (0.15%–0.17%) are detected in the epigenomes of kidneys, bladder, heart, skeletal muscles, and lungs. Finally, DNA from the liver, spleen, and the endocrine glands (testes and pituitary gland) possesses the lowest amounts of 5hmC with levels ranging from 0.03%–0.06%. Interestingly, the pituitary gland, which is located in the brain, has a low amount of 5hmC, supporting the hypothesis that a high 5hmC content is related to neuronal function rather than mere localisation in the brain. In contrast to 5mC, the amount of which is relatively consistent across different tissues, the amount of 5hmC is tissue specific (Globisch et al., 2010). It has been established that the 5hmC is deficient at transcription start sites and enriched at gene upstream regions and gene bodies in mammalian genomes (Pastor et al., 2011; Song et al., 2011b). Also, 5hmC shows high enrichment in regulatory chromatin states mainly marked by H3K4me1 (mark for active enhancers), active transcription start sites, and bivalent enhancer regions (Cui et al., 2020). This 5hmC enrichment has also been reported at the promoters of long interspersed nuclear elements, CTCF and pluripotency transcription factor binding sites (Ficz et al., 2011). Although most 5hmC cytosines are in the CG context, a relatively significant proportion of them also exist in a strand–specific CH context (Ficz et al., 2011; Pastor et al., 2011). Importantly, 5hmC shows a relatively high difference between the two opposite strands of a chromosome with differences up to tenfold (Mooijman et al., 2016). It is thought that 5hmC strand bias, therefore strand age, may serve as a source of chromosome–wide epigenetic memory to determine downstream protein activity and instruct biological processes such as chromosome segregation. Interestingly, 5hmC strand bias can flip within a chromosome and this sharp transition is consistent with a putative sister chromatid exchange events [10]. Most importantly, such

[10]Exchange of genetic material between two identical sister chromatids with mechanism involved in this phenomenon still largely unknown (Lazutka, 1995).

asymmetric 5hmC distribution might be an explanation for the underlying mechanisms of immortal DNA strand theory [11] that were not known before (Huh et al., 2013).

Further products of 5mC oxidation, 5fC and 5caC, can also be generated by TET enzymes, although their abundance is incredibly low (5fC levels around 0.002% and 5caC levels only at 0.0003%) in the genome (He et al., 2011; Ito et al., 2011; Pfaffeneder et al., 2011). Since TDG recognises both 5fC [12] and 5caC bases, perhaps partly explaining why neither 5fC or 5caC appear to accumulate to significant levels within DNA (He et al., 2011; Zhang et al., 2012). For 5fC modification, it was found that it predominantly clusters in enhancer regions, and several transcription factors have been shown to preferentially bind regions containing 5fC (Iurlaro et al., 2013; Song et al., 2013).

Finally, it is worth mentioning that another type of DNA modification, N6–methyladenine, has recently been identified in mammals. However, measurements of abundance are very dissimilar between studies, even when performed on DNA from identical cell types and more evidence is needed to support the presence of this modification (Douvlataniotis et al., 2020; Kweon et al., 2019; Xiao et al., 2018).

2.1.7 Biological Functions of DNA Modifications

2.1.7.1 Repression of Genetic–Information

One of the most important functions of DNA modification is dosage compensation. The methylation level of the inactivated X chromosome [13] is substantially higher compared to the other copy of the X chromosome

[11] Old DNA strands are retained by stem cells during asymmetric cell divisions to reduce the mutational load arising from genome replication (Cairns, 1975; Sherley, 2008).

[12] It was shown that TDG removes only half of 5fC residues at specific genomic cites (Su et al., 2016).

[13] By the process of X chromosome inactivation organism balances X–encoded gene products between male and female mammalian cells (Ohno et al., 1959).

(Duncan et al., 2018; Wolf et al., 1984). Interestingly, while most CGIs (68%) on the inactivated X chromosome show increased methylation, a small fraction of CGIs (7%) have significantly lower levels of methylation (Sharp et al., 2011). Another interesting example of the importance of methylation in dosage compensation is modification of duplicated genes in the honeybee *Apis mellifera* (Dyson and Goodisman, 2020). It was found that in *Apis mellifera*, levels of gene body methylation were significantly lower in duplicate genes than in single–copy genes, implicating a possible role of DNA methylation in postduplication gene maintenance. Similar methylation patterns were discovered in plants where genes that returned to single copies after the whole genome duplication event show a higher level of gene body methylation compared to the long–retained duplicates (Shi et al., 2020).

Besides regulating genetic information dosage on a large scale, methylation plays a fundamental role in local gene silencing, such as genomic imprinting [14] (Barlow and Bartolomei, 2014). The life cycle of these methylation imprints is i) erasure in the primordial germ cells, ii) establishment in mature gametes, and iii) maintenance during embryonic development (Reik and Walter, 2001). To date, 228 imprinted genes have been reported in human and 260 in mouse genomes (Tucci et al., 2019). As a consequence of this monoallelic methylation, genes will exhibit maternal, paternal or isoform–specific expression and have effects on multiple developmental stages (e.g., fetal growth, circadian machinery, and behaviour). It is important to mention that 5hmC was also discovered at imprinted loci, however, often overlapping regions are associated with parent–of–origin allelic 5mC sites (Hernandez Mora et al., 2018).

Since a large portion of the mammalian genome consists of repetitive elements, such as short interspersed nuclear elements (SINE), long interspersed nuclear elements (LINE), long terminal repeats (LTR), and satellites which endanger genomic stability, it was suggested that DNA

[14] Monoallelic DNA methylation according to parental origin.

methylation might be a protection mechanism of suppressing their parasitic functions (Yoder et al., 1997). DNA methylation facilitates transcriptional silencing of these parasitic elements, and it has been suggested that DNA methylation may have originally evolved as a defence mechanism to prevent activation and subsequent genome instability before acquiring its role in gene regulation (Slotkin and Martienssen, 2007). Important to mention that not only the 5mC mark is found in repetitive elements but significant amounts of 5hmC modification were also observed in LINE1 elements (Ficz et al., 2011)

2.1.7.2 Epigenetic Regulation at Genic and Regulatory Elements

More than half of the protein–coding genes contain a CGI in their upstream regions or transcription start sites (TSS) (Jones, 2012). Most upstream CGIs are unmethylated and nucleosome depleted, which has been associated with increased DNA accessibility and gene expression levels (Bestor et al., 2015). Hypermethylation of these CGIs can i) directly prevent the recruitment of the transcription complex scaffolding that can activate RNA polymerase II, or ii) serve as a specific binding motif for inhibitory regulation, thus inhibit gene expression. Given the observations of methylation at repressed TSSs, this raises the question of whether silencing or methylation comes first. An epigenetic "lock" model has been proposed to explain this phenomenon, where DNA methylation acts as a stabiliser of the inactive state established through other epigenetic mechanisms (Jones, 2012).

Gene bodies, as regions usually having low CG–density, are extensively methylated, which may be necessary to suppress genomic repeats that reside within introns (Jones, 2012). However, it was observed that modification of the gene bodies is positively correlated with the transcriptional level of respective genes (Wolf et al., 1984). Interestingly, while most upstream CGIs remain unmethylated as many as 34% of all intragenic CGIs are methylated in a tissue–specific manner (Maunakea et al.,

2010). Moreover, even though intragenic CGIs can be modified, this does not suppress RNA synthesis (Jones, 2012). It was suggested that intragenic methylation may be also used to avoid spurious transcription initiation (Neri et al., 2017). Thus, it seems that in mammals, it is the initiation of transcription but not transcription elongation that is sensitive to DNA methylation silencing. On the contrary, 5hmC levels within the gene body show a positive correlation with gene expression, suggesting a possible role for 5hmC in promoting transcriptional activity (Song et al., 2011b). Also, it has been observed that exons and introns have a different methylation level, and transitions in the degree of methylation occur exactly at exon–intron boundaries, possibly suggesting a role for methylation in regulating splicing (Laurent et al., 2010). Furthermore, 5hmC modification was also observed at the mammalian exon–intron cross boundary with its levels varying between tissues (Khare et al., 2012).

DNA modification was found to contain specific patterns at various genomic elements. Enhancers, which are key to the cell type–specific control of gene activity, were associated with low methylated regions [15] (Stadler et al., 2011). Similarly to enhancers, super–enhancers [16] tend to be hypomethylated or differentially methylated depending on the cell type, with hypermethylated super–enhancers being linked to the silencing of target genes (Bell et al., 2020). It has been shown that 5hmC modification also plays a role in epigenetic regulation of enhancers, with it being a mark of tissue–specific enhancers (Cui et al., 2020).

2.1.8 DNA Modifications in Higher–Order Genome Structures

Recently, it has been proven that DNA methylation plays a role in defining higher–order genomic structures (Jeong et al., 2014; Lister et al., 2009; Timp and Feinberg, 2013; Xie et al., 2013). Already in 2009, it

[15] Regions with low levels of methylation in the range of 10% – 50% that comprise around 4% of all CG sites in the mouse genome.
[16] Clusters of enhancers that control the expression of cell identity genes.

was shown that a large proportion of the lung fibroblast cell line genome displayed lower levels of CG methylation (Lister et al., 2009). Large regions with an average methylation level less than 70% were identified (average size 153 kb), which were termed partially methylated domains. These domains comprised on average ∼38% of every autosome, and 80% of chromosome X and encoded genes with relatively lower expression levels. Even larger epigenomic structures were identified a couple of years later. Genome–wide methylome analyses have revealed the existence of large hypomethylated regions, called "canyons" or "valleys" (Jeong et al., 2014; Xie et al., 2013). These regions can span up to one megabase (Mb) in size, are hypomethylated throughout development, and are very conserved across vertebrates. More than half of the genes identified in human hypomethylated regions were also present in mice regions. These genes were strongly enriched for functional groups in transcription factors, homeobox family, embryonic morphogenesis, and cancer pathways (Xie et al., 2013). Importantly, the borders of these regions are demarked by 5hmC and become eroded in the absence of DNMT3A, suggesting that there is competition between DNMT3A and TET proteins to maintain a status quo at the same loci (Jeong et al., 2014). Large hypomethylated regions have also been identified as a common epigenetic alteration in several tumour types (Timp and Feinberg, 2013). These regions were on average 28 kb in size and in normal samples exhibited methylation levels of ∼80%, while in the cancer samples methylation ranged from 40% to 60%. These regions tend to be co–localised with lamina associated domains (LADs), which confirms that DNA methylation plays a role in higher–order chromosome organisation within the nucleus.

2.2 Technological Aspects

2.2.1 Profiling Techniques for 5–methylcytosine

Multiple techniques have been developed to profile DNA methylation, which can be broadly grouped into restriction enzyme, affinity enrichment, and bisulfite conversion–based techniques.

2.2.1.1 Restriction Enzyme–Based Techniques

Methods that use enzymatic digestion employ methylation–sensitive restriction enzymes, which show variability in digestion properties at methylated and unmethylated CG sites. The first quantification of DNA methylation using restriction enzymes was performed almost half a century ago to show that DNA modification in mammals occurs in CG sites (Gautier et al., 1977). Nowadays, the two most commonly used methylation–sensitive restriction enzymes are HpaII and Msp1 (Takamiya et al., 2006). These isoschizomers recognise the same sequence (CCGG) but have different methylation sensitivity: HpaII cleaves unmethylated CG sites, and Msp1 cleaves methylated CG sites. Restriction enzyme–based methods are cost–effective and enable genome–wide methylation profiling. However, these enzymes identify only a limited fraction of genome CG sites (adjacent to restriction enzymes digestion sites), and they cannot quantify the methylation level of single CG (Yong et al., 2016).

2.2.1.2 Affinity Enrichment–Based Techniques

Affinity enrichment–based methods enrich for methylated DNA fragments by either methyl–binding domain (MBD) proteins or antibodies that target 5mC. Methylated DNA immunoprecipitation (MeDIP) uses an antibody to 5–methylcytosine targeting single–stranded DNA,

while the MBD approach uses the methylated CG binding domain of the MBD2 protein to capture double–stranded DNA (Li et al., 2010; Robinson et al., 2010). After enriching for methylated regions, DNA methylation can be quantified with either high–resolution array hybridisation or high–throughput sequencing.

MBD is more sensitive than MeDIP for high CG–density regions [17] and is not related to the DNA methylation level, while MeDIP is generally more efficient than MBD in enriching for highly methylated medium CG–density regions (Li et al., 2010). MBD is a cost–effective method that can be used with very low amounts of DNA (Aberg et al., 2012, 2017). The MeDIP method is also economical and can differentiate between CG or CH methylation contexts. On the downside, they provide low resolutions and are biased towards hypermethylated regions. Also, if the technology is biased towards CG rich regions, then regions with poor CG–density will be underrepresented or interpreted as unmethylated, therefore, computational corrections are necessary to normalise the CG content (Rauluseviciute et al., 2019). It is important to mention that peak identification algorithms that are usually used for enrichment–based methods are developed for chromatin immunoprecipitation data analysis (e.g., to locate transcription factor binding sites from the immunoprecipitation data). DNA methylation sites differ from transcription factor binding sites in that methylated CG dinucleotides are highly abundant in most differentiated cells, thus the signal peaks in affinity–based method data are densely distributed. The characteristic of this type of data raises the demand for a computational analysis programme with higher resolution since the aforementioned algorithms fail to finely detect methylation level of CG dinucleotides (Lan et al., 2011). As an alternative method methyltransferase–directed transfer of activated groups, the method offers isolation of the unmodified CG–fraction via immunoprecipitation of biotin–labelled unmodified cytosines (Kriukienė et al., 2013). This approach is much more sensitive since it isolates only the unmodified fraction of the genome and can detect more subtle changes. Furthermore,

[17] This makes MBD a highly effective method to measure the methylation status of CGIs.

MeDIP based technologies can also be adapted to hydroxymethylation by choosing an antibody specific to 5hmC (Nestor and Meehan, 2014).

2.2.1.3 Bisulfite Conversion–Based Techniques

Sodium bisulfite (BS) sequencing is considered the gold standard for DNA methylation profiling. BS converts unmethylated cytosine to uracil, which eventually turns into thymine, while methylated cytosines are protected [18] (Clark et al., 1994; Frommer et al., 1992). The resulting C to T conversion is detected either by next–generation sequencing or array hybridisation, thereby CG sites are classified as methylated or unmethylated. BS conversion–based techniques can produce single nucleotide resolution, strand specific DNA methylome that is not influenced by uneven coverage across the genome. However, while BS conversion can differentiate methylated from unmethylated cytosines, it cannot discriminate between 5mC and 5hmC modifications, which do not undergo C to T transitions after bisulfite treatment and both are read as C after sequencing (Huang et al., 2010).

Whole genome bisulfite sequencing detects C to T conversions by sequencing bisulfite–treated fragments, and aligning reads back to a reference sequence [19] (Urich et al., 2015). However, whole–genome BS is more expensive owing to genome–wide deep sequencing of bisulfite–treated fragments as most WGBS reads do not originate from CG regions, while techniques such as MeDIP and MBD produce DNA libraries covering only highly methylated genomic regions (on the other hand, WGBS is not biased towards any underlying region type) (Ziller et al., 2013). Second, the reduction in sequence complexity resulting from the conversion of cytosine into uracil can be problematic in the polymerase chain reaction (PCR) amplification step (Berney and McGouran, 2018). Also, because of the DNA degradation by purification and sodium bisulfite

[18] While the nucleotide sulfonation effect was then known for more than twenty years (Hayatsu et al., 1970).

[19] *Arabidopsis thaliana* genome was the first to be fully sequenced using WGBS (Cokus et al., 2008).

treatment, BS based techniques depends on a relatively large amount of DNA starting material [20] (Berney and McGouran, 2018). Additionally, two types of error can occur during bisulfite treatment: over–conversion (when methylated cytosines are deaminated to uracil which leads to false–negative cytosines) and under–conversion (when unmethylated cytosines are not converted to uracil which leads to false positive cytosines) (Chappell et al., 2018). These errors require adequate evaluation of the conversion rate before performing data analysis.

Reduced representation bisulfite sequencing is a cost–effective alternative to WGBS, which quantifies DNA modification only for a fraction of genomic regions (Lister et al., 2009; Meissner et al., 2005, 2008). This method utilises Msp1 and size selection of digested fragments (40 — 200 bp) [21] prior to BS conversion. Even though this method combines the sensitivity of BS methods with the cost–effectiveness of enzyme–based methods, it provides lower coverage for many other genomic regions. Also, reduced representation bisulfite sequencing coverage varies to some extent by the size selection or choice of enzymes and prior in sicilo analysis should be performed to determine the optimal experimental parameters (Gu et al., 2011a).

Illumina's Infinium HumanMethylation 450K BeadChip is the most widely used microarray for profiling DNA modification in human [22]. It uses bisulfite treatment, polymerase chain reaction amplification, hybridisation and allows the interrogation of more than 450 thousands CG sites from the human genome that cover almost all protein–coding genes and CGIs [23] (Bibikova et al., 2011). In this approach, two probes are employed to distinguish between unmodified and modified CG sites, which are marked with different fluorescence dyes and hybridised to arrays. BeadChip consists of twelve arrays making this technology suitable for

[20] It has been reported that 95% of the DNA is destroyed by sodium bisulfite treatment.

[21] Covers most of the CGIs in the human genome by isolating only 1% — 3% of the genome (Gu et al., 2011a).

[22] Before 450K two other platforms existed: GoldenGate genotyping technique with 1536 CG sites and Illumina's Infinium 27K technique (Bibikova et al., 2006, 2009).

[23] As control regions, it also includes more than three thousand non–CG sites and random SNPs.

analysing larger cohorts. Moreover, the most recent version of this technology (Infinium MethylationEPIC) covers 850 thousand CG sites in the human genome. This version contains more than 90% of the 450K sites as well as more than 300 thousand new CG sites located in enhancer regions (Moran et al., 2016). While this approach is relatively cheaper and requires less complex data processing [24], it still depends on sodium bisulfite conversion, therefore cannot be used to discriminate 5mC from 5hmC modifications.

Another method based on bisulfite conversion is bisulfite padlock probes (or molecular inversion probes) (Deng et al., 2009; Hardenbol et al., 2003; Nilsson et al., 1994). This technology was first applied for exon capturing and only then transferred to quantify DNA methylation using targeted bisulfite sequencing (Deng et al., 2009; Diep et al., 2012; Porreca et al., 2007). This technology offers the sensitivity of bisulfite sequencing and customisable selection of wanted CG sites. Similar to the polymerase chain reaction design, padlock probes are designed to target a pair of sequences flanking the target of interest, however, the hybridising segments correspond to the 5' and 3' ends of a single molecule that loops around itself in the padlock design. Custom–designed molecular inversion probes can be multiplexed (ten thousands of genomic loci can be enriched simultaneously) and such selection enrichment of genomic targets prior to sequencing substantially reduces sequencing costs. When molecular inversion probes were first used to enrich for exonic regions, one of the concerns was the high dropout rate, that is, in each replicate around 80% of the intended targets were not observed by deep sequencing (Porreca et al., 2007). Improvements to the bisulfite padlock probe design were made later with around 330 thousand probes that covered around 140 thousand non–overlapping regions with a total size of 34 mega bases (Diep et al., 2012). However, there are a couple of challenges that need to be considered when working with bisulfite padlock probes. First, off–target annealing (bisulfite converted sequences have very few cytosines so the annealing arms cannot have cytosines) and secondly,

[24] On January 17 2021 NCBI GEO data submission platform contained: 346 submitted research projects for Infinium 27K technique, 1437 for 450K, and 374 Infinium MethylationEPIC (Edgar et al., 2002).

to avoid polymorphisms and CG sites that may be methylated within annealing arms.

2.2.2 Profiling Techniques for Oxidised 5–methylcytosine Forms

The first genome–wide 5hmC profiling was performed by selective chemical labelling of 5hmC, followed by deep sequencing in the mouse cerebellum (Song et al., 2011b). In this study, the 5hmC modification was labelled with a customised glucose moiety which was later modified with biotin for enrichment and sequencing. The drawback of this technique is that custom–made reagents are needed. Later, another study on mouse embryonic stem cells suggested two new approaches for genome–wide 5hmC profiling (Pastor et al., 2011). One highly specific method used glusosylation, oxidation, and biotinylation but did not require custom–made reagents, while the other method used 5hmC conversion to 5–methylenesulphonate which was precipitated with specific antibodies, followed by high–throughput sequencing. However, there were issues with this method including the i) efficiency of antibody precipitation was dependent on 5hmC density, and the ii) antibodies cross reacted to unmodified or methylated DNA. Another method based on MeDIP protocols for 5mC profiling was developed for 5hmC identification – hMeDIP (Ficz et al., 2011; Williams et al., 2011). However, hMeDIP also faces the previously mentioned antibody–based immunoprecipitation method problems.

Additional chemical derivatisation steps before sodium bisulfite treatment made it possible to profile 5hmC modification at single base resolution (Booth et al., 2012). The oxidative bisulfite method is based on a DNA oxidation step prior to bisulfite conversion. During oxidation, 5hmC modifications are converted to 5fC and are again converted into uracil in subsequent sodium bisulfite treatment, therefore, only 5mC modification is detected as cytosine. By comparing the readouts of oxidative bisulfite and standard BS, one can estimate the 5hmC signal

at each CG site. An alternative method for 5hmC profiling, termed TET–assisted BS sequencing (TAB–seq), uses glucosylation of 5hmC by β–glucosyltransferase to protect it from oxidation by TET1, whereas all other modified cytosine residues are oxidised to 5caC (Yu et al., 2012a,b). In addition, single molecule real–time sequencing can differentiate between unmodified or methylated, hydroxymethylated cytosines (Flusberg et al., 2010; Song et al., 2011a). This method monitors the incorporation of fluorescently labelled nucleotides into newly synthesised DNA molecules. The duration of the resulting fluorescence pulse emissions yield information about polymerase kinetics and allow differentiation between modified nucleotides in the DNA template. Recently a new method, termed Jump–seq, was shown to achieve bisulfite–free, nearly base resolution detection of 5hmC at the whole genome scale (Hu et al., 2019). This method uses 5hmC labelling with an azide–modified glucose and genomic DNA tagging with biotin–P7 adapter. A hairpin DNA (with P5 adapter) carrying an alkyne is added covalently to the modified glucose. After primer extension from the hairpin and cleavage from the tethered hairpin, the newly synthesised strand is subjected to library construction and sequencing. The 5hmC single site location is inferred from the polymerases *landing* site pattern that connects the hairpin sequence and any genomic DNA sequence. However, this method showed a lower accuracy of the priming reaction and limited the resolution of the method to twenty nucleotide sized bins.

The 5fC modification can also be detected using 5fC chemical modification–assisted BS sequencing (fCABseq), in which 5fC is first protected from deamination resulting in 5fC, 5mC and 5hmC being read as cytosine. Then, the locations of 5mC or 5hmC modifications can be determined using conventional BS sequencing, and the location of 5fC modifications can then be inferred by the comparison of the two readouts (Song et al., 2013). Another variation of the CABseq method (caCABseq) can detect 5caC chemical labelling and biotin tagging. As a result, the 5caC–containing DNA is pulled down, enriched, and subjected to high–throughput sequencing (Lu et al., 2013). Methylase assisted BS sequencing (MABseq) uses methylase to modify all unmodified cytosine

bases to 5mC. Sodium bisulfite treatment then converts 5fC and 5caC modifications into uracil, enabling their identification using standard BS sequencing, although these two modifications cannot be distinguished from each other (Neri et al., 2015; Wu et al., 2014). Sodium borohydride can be used after the enzymatic methylation to convert 5fC to 5hmC enabling selective detection of 5caC because it is the only base that is read as uracil. Employing the 5hmC–selective chemical labelling can be also used to identify 5fC modification. After the protection of 5hmC modification by glucosylation, 5fC can be selectively reduced to 5hmC by sodium borohydride treatment, and the newly created 5hmC residues are then enzymatically labeled with an azide–modified glucose for the attachment of biotin tags (Song et al., 2013). Finally, the antibody–based DNA immunoprecipitation (5caC–DIP) approach was used to generate genome–wide distribution maps of 5hmC, 5fC, 5caC modifications in mouse embryonic stem cells (Shen et al., 2013).

Another newly published approach to decipher DNA modification patterns that could be used not only for 5hmC modification but also for 5mC modification is enzymatic methyl–seq. This enzymatic deamination approach, named long–read enzymatic modification sequencing, allows long–range DNA modification profiling of 5–methylcytosine and 5–hydroxymethylcytosine over multikilobase lengths of genomic DNA (Sun et al., 2021). The principle of this methodology is as follows, genomic DNA can either be treated with TET2 and β–glucosyltransferase (BGT) to protect both 5mC and 5hmC modifications or only with BGT to protect 5hmC modification, subsequent deamination by APOBEC3A, followed by an amplification step allows the distinction between the unprotected substrate from the protected cytosine derivatives. The TET2 and BGT treatment results in the distinction of 5mC and 5hmC from cytosines, whereas treatment with BGT alone results in the distinction of 5hmC from cytosines and 5mC.

2.2.3 Tethered Oligonucleotide–Primed Sequencing–Based Techniques

Since most CG sites in the human genome are methylated (approximately 60%—80%), analysis of the remaining fraction of CG sites may be more sensitive and simpler for detecting subtle changes in DNA modification profiles. The TOP–seq method has the advantage of previously developed mTAG–seq technique, which uses an engineered version of the SssI methyltransferase for biotin labelling of unmodified CG dinucleotides (Kriukienė et al., 2013; Staševskij et al., 2017). In both methods, SssI methyltransferase uses *S*–adenosylmethionine cofactor and tags the unmodified and hemimodified CG sites with a reactive azide group (**Figure 2.3**). For the TOP–seq method in the next step, a covalently tethered alkyne–bearing DNA oligonucleotide promotes non–homologous priming of a DNA polymerase at the tagged sites, thus facilitates the template–dependent polymerase action from the 3' end of the tethered DNA duplex [25]. In addition to unmodified CG sites, a 5hmC modification can be identified using the hmTOP–seq method which is based on the same sequence readout mechanism primed at covalently labelled 5hmC sites from an in situ tethered DNA oligonucleotide (Gibas et al., 2020). Finally, the caCLEAR method enables targeted mapping of 5caC sites by combining a methyltransferase–promoted C–C bond cleavage reaction, leading to the decarboxylation of 5caC that yields unmodified cytosine, followed by targeted sequencing of the introduced unmodified CG sites (Liutkevičiūtė et al., 2014; Ličytė et al., 2020).

[25] While for the mTAG–seq tecnique biotin–labelling is used.

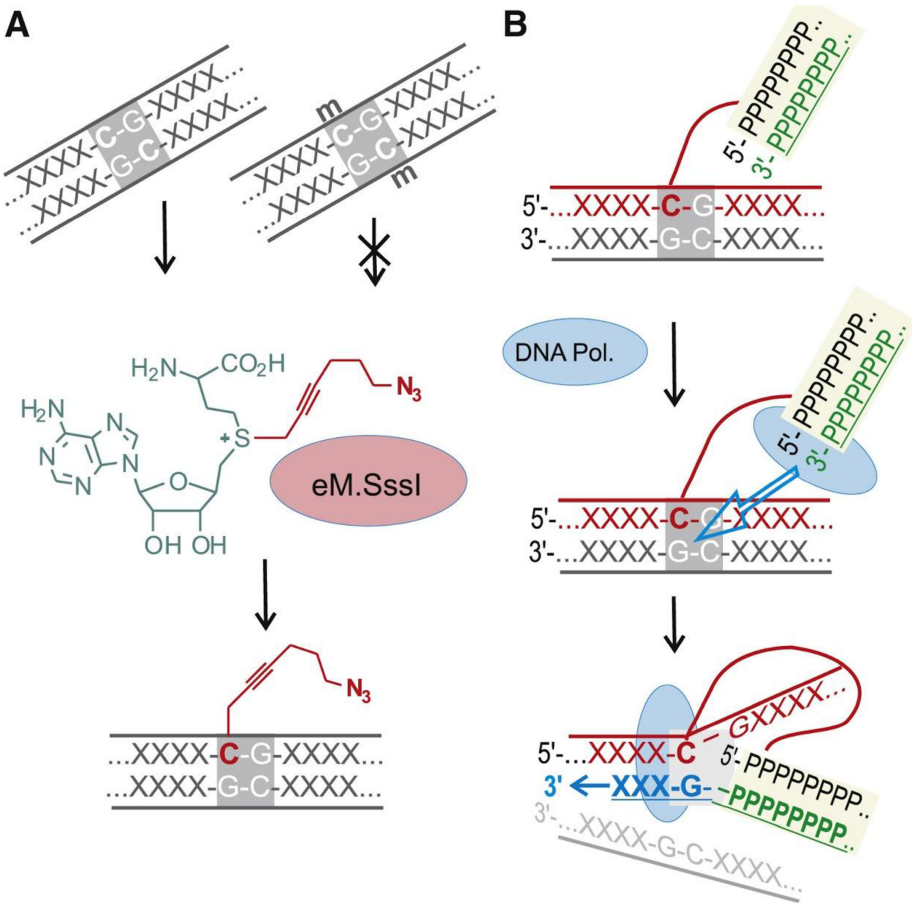


Figure 2.3 | Tethered Oligonucleotide–Primed Sequencing Method

(A) Selective tagging of unmodified CG sites with an azide group using an engineered variant of the SssI methyl–transferase and a synthetic analog of the SAM cofactor. **(B)** Tethered oligonucleotide–primed DNA polymerase activity at an internal covalently tagged CG sites. Illustration was taken and modified from Staševskij et al. (Staševskij et al., 2017).

2.3 Statistical Aspects

2.3.1 Linear Regression

This section together with ***Section 2.3.2*** *forms the foundation of* ***Section 6.2.5*** *and* ***Section 6.5.5****, describing the linear regression model training and prediction.*

Regression analysis is a statistical technique to estimate a response (dependent) variable with one or more predictors (independent variables) (Kutner et al., 2004). A regression model is based on several concepts: i) the dependent variable changes with the change in the independent variable in a systematic manner; ii) for each independent variable, there is a probability distribution for the dependent variable. Often in (bio)statistics, it is investigated how a variable Y depends on a variable X[26]. Linear regression is an approach that proves useful for providing insights about such biological relationships, as one can model the dependent variable Y given an independent variable X by assuming that X has a linear effect on Y. This linear function has the following parameters: i) $x_1, \ldots, x_k$ are the values of the k covariates; ii) interception of β_0 (value of x_i when value of $y_i = 0$); iii) slope of β_i; iv) residuals or random error ε_i – an unobserved random variable that adds *noise* to the linear relationship:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon \tag{2.1}$$

Suppose that we have observations $y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$ and want to model a linear function using $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. Then, one can try to define a value for $\hat{\beta}$ which minimises the squared error ε^2:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta x_i - \beta_0)^2 \tag{2.2}$$

Often in (bio)statistics, one response can be explained using multiple predictors, multiple linear regression should be used in such a case [27]. The *design matrix* X is used to present such multivariate data (one can

[26] For example, cytosine modification levels in prostate cancer patients and exposure to abiraterone acetate or the number of Nobel prize laureates and chocolate consumption per capita (Gordevičius et al., 2018; Maurage et al., 2013).

[27] For example, gene expression level and modification levels of all the cytosines from an associated locus.

represent n samples in rows and k observations in columns):

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \in \mathbb{R}^{n \times k}$$

Following similar notation, the response variables, random errors, and regression coefficients can be written in vector forms:

$$\begin{aligned} Y &= (Y_1, \dots, Y_n)^T \in \mathbb{R}^n \\ \varepsilon &= (\varepsilon_1, \dots, \varepsilon_n)^T \in \mathbb{R}^n \\ \beta &= (\beta_0, \beta_1, \dots, \beta_{k+1})^T \in \mathbb{R}^k \end{aligned}$$

Linear models have assumptions about the predictors, the response variables, and their relationship to simplify the modelling problems:

- *Linear relationship* – relationship between the independent and dependent variables is linear, therefore Y_i can be estimated using independent variables $x_{i1}, \dots, x_{ik}$ and the random error.
- *Normality of residuals* – error variables are normally distributed (i.e., the mean is zero and the variance is σ^2) with the expected value zero $E[\varepsilon] = 0$.
- *Independence* – error variables are independent of each other and there is no correlation between them.
- *No multicollinearity* – the independent variables do not correlate. If there is a high correlation, then it is difficult to explain the relationship between the independent and dependent variables.
- *Homoscedasticity* – the variance of the random error ε is constant and finite regardless of the values of the predictor variables: $Var(Y_i) = Var(\epsilon_i) = \sigma^2 < \infty$, for all $i = 1, \dots, n$.

2.3.2 Generalised Linear Models

Sometimes in (bio)statistics, the given relationship does not follow linear model assumptions, so the response variable comes from a non–continuous distribution and is categorical (e.g., survival status) or discrete (e.g., stages of cancer). Error variables are correlated and non–normal distributed (e.g., DNA modification level that is always non–negative). In such a case, the classical linear model cannot be applied, however, it can be replaced with a generalised linear model (GLM) which elevates the classical linear models, allowing non–linear relationships between independent and dependent variables (Nelder and Wedderburn, 1972). GLM represents a regression model family that uses transformation for a response variable that is not in the form of normal distribution.

As mentioned above, Ys are independent from each other and they come from the probability distribution with a mean $E[Y] = \mu$, Xs are predictor variables and $X\beta = \eta$ is a linear predictor. In the classical linear model, the relationship between the linear predictor and the mean is $\mu = \eta$. GLM *improves* this linear model by allowing a more complex relationship between μ and η via the so called link function $g(\cdot)$ which enhances the relationship to $\eta = g(\mu)$. Which link function is optimal depends on the underlying distributions that are explained below, however, these functions can be: identity ($\eta = \mu$), log ($\eta = \log(\mu)$), logit ($\eta = \log(\frac{\mu}{1-\mu})$), and others (Lindsey, 1997).

Probability distributions used in GLM are members of the exponential family, which represents a set of flexible probability distributions ranging through continuous or discrete variables. All these distributions follow the general formula:

$$f(x;\theta) = h(x)\ \exp[\,\eta(\theta)\cdot T(x) + A(\theta)\,] \tag{2.3}$$

In a given formula, the following representations are used:

- x – vector of measurements
- θ – canonical link

- $h(x)$ – base measurement
- η – natural parameter
- $T(x)$ – sufficient statistic of the distribution
- $A(\theta)$ – log partition function (log normaliser)

Some probability distributions that are members of this family are:

- *Bernoulli* – binary $\{0, 1\}$; logit link
- *Binomial* – counts of success or failure; logit link
- *Gaussian* – $\mathbb{R}$; identity link
- *Exponential* – $\mathbb{R}$; negative inverse link
- *Poisson* – $\mathbb{N}$; log link

A Gaussian distribution can be rewritten in terms of the general exponential format using the following parameters:

- Canonical link θ – μ
- Base measure $h(x)$ – $\frac{1}{\sqrt{2\pi}}$
- Natural parameter – $\langle \frac{\mu}{\sigma^2} \frac{-1}{2\sigma^2} \rangle$
- Sufficient statistic $T(x)$ – $\langle x x^2 \rangle$
- Log partition $A(\theta)$ – $\frac{\mu^2}{2\sigma^2} + \log|\sigma|$

Finally, the equation for the Gaussian exponential family can be stated as follows:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.4}$$

2.3.3 Kernel Density Estimation

This section forms the foundation of ***Section 5.2*** *and will describe kernel density estimators.*

Kernel density estimation (KDE) is a non–parametric method to estimate the probability density function of a variable X. Let $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ be independently and identically distributed copies of a random variable drawn from a continuous distribution with an unknown probability density f. The underlying function $f(\cdot)$ used to generate this sample can be approximated by the KDE given:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{2.5}$$

where n is the number of observations, $h > 0$ is the bandwidth parameter which determines the smoothness of the density estimate, and $K(\cdot)$ is the kernel, which is unimodal, symmetrical, usually non–negative function that integrates to one (Silverman, 1986). Intuitively, one can also explain that KDE creates a *bump* around each data point, then normalises over all the *bumps*. The shape of a *bump* depends on the used kernel function and the width of the *bump* depends on the bandwidth parameter (Ross, 2013).

Choosing the bandwidth parameter h is similar to specifying the number of bins in a histogram. Usually one wants to choose h as small as the data will allow, however, there is a trade–off between the bias of the estimator and its variance. Choosing a h that is too large will produce an over–smoothed KDE, which will reduce the variance across different samples but fail to capture local structure, whereas, a h that is too small will result in an estimate that is over fit to the actual samples available. There are several methods to select the bandwidth parameter, some of the most commonly used include:

- A rule–of–thumb method by Silverman estimates h using $h = Cn^{-\frac{1}{5}}$, with $C = 0.9 \min(sd, \frac{iqr}{1.34})$, where sd is the standard deviation of the sample of size n, and iqr is the interquartile range (Silverman, 1986). In `R 3.5`, this estimator is implemented under `stats::bw.nrd0`.

- An alternative solution by Scott uses 1.06 as a factor (implemented in `R` with `stats::bw.nrd`) (Scott, 2015).

Table 2.1 | Kernel Estimators
Most commonly used kernel functions.

Kernel	$K(x)$
Cosine	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}x\right)$
Epanechnikov	$\frac{3}{4}\left(1-x^2\right)$
Gaussian	$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$
Triangular	$(1-\lvert x\rvert)$
Tricube	$\frac{70}{81}\left(1-\lvert x\rvert^3\right)^3$
Uniform	$\frac{1}{2}$

- The method of Sheather and Jones estimates the bandwidth using a pilot estimation of derivatives (implemented in `R` with `stats::bw.SJ`) (Sheather and Jones, 1991).
- Two other methods that use biased and unbiased cross–validation are implemented in `R` with `stats::bcv` and `stats::ucv` (Scott and Terrell, 1987).

Functions that are used as kernel estimators are provided in **Table 2.1**. The Epanechnikov kernel is claimed to be optimal in the sense that it minimises the mean integrated squared error, however quite often, the Gaussian kernel is also used (Epanechnikov, 1969).

2.3.4 Dimensionality Reduction

*This section forms the foundation of **Section 6.5.2.1** data processing and will describe high dimensionality data.*

Firstly, problems that arise from high dimensionality data are introduced, followed by features for dimensionality reduction (DR) methods and finally, the section closes with the most popular techniques for DR.

2.3.4.1 The Curse of Dimensionality

In nature, combining multiple simple units allows to perform complex tasks. These units are redundant and after failing they can be replaced with others that achieve the same task [28] (in essence, the world is multidimensional). Such multidimensional data causes many, both practical and theoretical, problems.

The term *curse of dimensionality* was first used by Richard E. Bellman to describe an empty space phenomenon (Bellman, 1961). This term characterised a theoretical problem — with an increasing dimensionality, the volume between randomly distributed variables also increases [29]. To illustrate this problem in a different way, consider a circle that is embedded within a square, when they are presented within a two–dimensional space, the ratio between the two is around 0.8, in a three–dimensional space this ratio is only around 0.5, and the ratio of hyper–cube and hyper–sphere reduces exponentially for further high–dimensional spaces. Another theoretical problem named *egg and a shell problem* was used by Christopher M. Bishop to describe the hypervolume of a spherical shell (Bishop, 2006). Let us introduce two concentric spheres – first with radius r, second with slightly smaller radius $r - \varepsilon$ (i.e., ε is the thickness of the first sphere) in a space of D dimensions. Next, we can calculate the fraction of the volume occupied by inner sphere using:

$$\frac{V_D(r) - V_D\left(r - \varepsilon\right)}{V_D(r)} = \frac{r^D - \left(r - \varepsilon\right)^D}{r^D} \tag{2.6}$$

When D increases, the ratio tends to move towards one as the shell contains almost all the volume.

[28] For example, multiple CG sites in specific locus. It is expected that these CG sites perform the same task and usually failure of one CG will be balanced by the rest of CG sites.

[29] Consider a set of variables distributed within one dimensional space (i.e., a vector). When this vector is transferred to a two–dimensional space (i.e., an array), the spaces between variables increases.

2.3.4.2 Features of Dimensionality Reduction Techniques

The goal of dimensionality reduction is to identify and eliminate the redundancies among the variables. This goal requires to i) estimate the number of latent variables, ii) embed data to reduce their dimensionality, and iii) embed data to recover the latent variable [30].

A certain process in nature may be generated from a small set of independent degrees of freedom but it will usually appear in a more complex way due to a number of reasons (e.g., measurement procedure, stochastic variation). Consider a sample of D–dimensional vectors that has been generated by an unknown distribution, it is assumed that this distribution in data space is actually due to a small number $L < D$ of variables acting in combination, called *latent variables*. DR is achieved by defining a reverse mapping from data space onto latent space, so that every data point is assigned a representative in latent space. The number of latent variables can be computed from a topological point of view by estimating the intrinsic dimension(ality) of the data. When this intrinsic dimension L is equal to D, there is no structure, whereas when $L < D$, data points are often constrained to lie in a well–delimited subspace [31]. Different approaches can be used to cope with intrinsic dimension estimation: i) minimising a reconstruction error; ii) preserving pairwise distances; iii) fractal methods (Grassberger and Procaccia, 1983). Additional constraints can be imposed on the desired L–dimension representation.

DR techniques vary in many characteristics, such as i) the model that data is assumed to follow (e.g., linear or non–linear, continuous or discrete), ii) the type of algorithm (e.g., batch or online), and iii) the criterion to be optimised (e.g., mean square error, variance, pairwise distances

[30] There is no *ideal* method that can perform all three tasks optimally.

[31] This peculiar feature was well exemplified using *the swiss roll problem* (Roweis and Saul, 2000).

[32]); iv) hard or soft DR; v) global or local recovery [33].

2.3.4.3 Dimensionality Reduction Techniques

The most basic and maybe even the oldest technique for DR is principal component analysis (PCA) (Pearson, 1901). As a linear DR technique, PCA assumes that the embedded subspace is a linear subspace and looks for a linear projection. PCA converts a data matrix X with zero column–wise mean into a lower dimensional representation Y. The goal of the PCA is to find an orthonormal transformation (projection) matrix P with which $Y = PX$. Intuitively, what PCA does is that it converts high–dimensional data into a low–dimensional representation that captures as much variance of the original data as possible through a linear transform. Therefore, in the reduced representation, each column in Y can be considered as a linear combination of the columns in the original data X. The PCA algorithm works as follows:

- Subtract the mean for each data dimension to obtain the mean–adjusted data matrix B:

$$B = X - \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right)^T \tag{2.7}$$

- Calculate the covariance matrix C:

$$C = \frac{1}{n-1} BB^T \tag{2.8}$$

- Solve for the eigenvector and eigenvalues of C using spectral decomposition UOU^T, where O is the diagonal matrix and U is an orthogonal matrix. The diagonals in O represent the eigenvalues of C and the columns of U represent the associated eigenvectors

[32] Measured between the observations in the data set (from a topological point of view, the projection of the object should preserve its structure).

[33] Global techniques try to recover the global information while local techniques concentrate on recovering the local structure of the data, the global structure then emerges from the continuity of the local fits.

of the covariance matrix, which are also known as the principal components of the matrix.

- Order the eigenvector into descending order to maximise the amount of variance.
- Construct a projection matrix and derive the new dataset in a lower dimension space.

PCA with complexity $\mathcal{O}(p^2n+p^3)$ is implemented in `R` with `stats::prcomp`.

If groups are linearly inseparable in the input space R^2, then it is possible to make them linearly separable by mapping them to a higher dimension space R^3 (Schölkopf et al., 1998). In such a case, kernel PCA would use function Φ which can $\Phi : R^2 \Rightarrow R^3$ $((x_1, x_2) \Rightarrow (x_1, x_2, x_1^2 + x_2^2))$. Such mapping can be computationally expensive but one can apply *kernel trick* – the principal components can be computed from the inner product matrix $K = X^T X$. Here it is not needed to explicitly map all points into the high–dimensional space and do the calculations there, it is enough to obtain the kernel matrix. The kernel PCA algorithm works as follows:

- Select kernel function
- Calculate the kernel matrix
- Centre kernel matrix
- Solve the *eigeinproblem* for a given kernel matrix
- Project the data to each new dimension

Classical multidimensional scaling is another technique that uses eigenvalue decomposition, however not on the original data, rather on a transformed distance matrix. The algorithm for this method is:

- Set up squared matrix D^2 for a given proximity matrix D
- Calculate the double–centred matrix $B = -\frac{1}{2}JD^2J$, where J is the centring matrix

- Use eigen–decomposition and determine n largest eigenvalues and corresponding eigenvectors of B

Non–metric multidimensional scaling (nMDS) is an improvement of classical scaling in a way that relies on the ranking of distances (minimising stress function solved by iterative algorithms). A solution is found such that the rank order of distances between points in the ordination match the order of dissimilarity between the points. Another technique, locally linear embedding, reconstructs points from the high–dimensional space using their neighbourhoods. First, this technique computes the neighbours of each data point, then it computes the weights that best reconstruct each data point, and finally applies eigenvalue decomposition. One of the most popular non–linear technique is t–SNE (t–distributed stochastic neighbourhood embedding) that uses local relationships between points to create a low–dimensional mapping. In the high–dimensional space, t–SNE creates a probability distribution that dictates the relationships between various neighbouring points, then recreates a low–dimensional space that best follows that probability distribution as possible (Kobak and Berens, 2019).

2.3.5 Machine Learning

*This section forms the foundation of **Section 5.4** and will describe the theory behind artificial neural networks (ANN) for estimating DNA modification signal.*

2.3.5.1 Learning Process

Supervised learning models, in the context of machine learning [34], aim to learn a *generalised* function $f(\cdot)$ (e.g., classifier or regression model) from

[34] Machine learning can be explained as: *a computer programme is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.* (Mitchell, 1997)

a set of training pairs $(x_1, y_1), \ldots, (x_n, y_n)$. Here a goal of function $f(\cdot)$ is to perform a wanted task on unseen data x^u as well as on the original set x. By contrast, unsupervised machine learning techniques aim to discover patterns from the set x itself, without the need for output y (e.g., PCA, clustering). A machine learning model can be characterised by a set of parameters to be optimised on the training dataset (Zador, 2019).

A simple example of a supervised learning model is logistic regression. The goal of binary logistic regression is to train a classifier (in this case, a sigmoid classifier) to make a binary decision about the class of a new input observation. The classifier will use a single input x, which represents a vector of features $x_1, x_2, \ldots, x_n$. Logistic regression trains a model by learning from a training set, a vector of *weights* and a *bias* (intercept). Each weight w_i is associated with one of the input features x_i and represents how important that input feature is to the classification decision. The resulting value z expresses the weighted sum of the evidence for the class:

$$z = \left(\sum_{i=1}^{n} w_i x_i \right) + b \tag{2.9}$$

Then, one can pass calculated z to a sigmoid function to obtain values between 0 and 1:

$$y = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{2.10}$$

This simple example shows how a supervised machine learning function (classifier) works, more advanced machine learning techniques will be discussed in the next section.

2.3.5.2 Artificial Neural Networks

The previously discussed classification task can be considered as a neural network composed of only one neuron. However, not all statistical problems are linearly separable and extra layers of complexity might be needed, such a multi–layer network is called a multi–layer perceptron (Krogh, 2008). An artificial neural network is a model training technique

composed of one or more layers of neurons, with each layer containing one or more neurons [35]. Such architecture was inspired by neural networks in the brain (Farley and Clark, 1954). Neurons perform operations (calculations) on the data passed by upstream neurons given an activation function, then provide the output to downstream neurons. The connections between the neurons in different layers are represented by weights. The sensory neurons (neurons in the first layer) receive the data from the input dataset and each neuron computes a weighted sum of its inputs, applying an activation function to calculate its output. Training an ANN means optimising parameters and bias values of the activation function of each neuron so that the output of the ANN is more similar to the known value in the training set. ANN with too many layers or too many neurons can cause model overfitting, while underfitting might be caused by too few layers:

There are two categories of ANNs:

- *Feedforward ANN* – information moves one–way from the first until the last layers
- *Recurrent ANN* – information can move both ways and feedback can be given backwards from the later layers

ANNs are based on linear combinations of non–linear basis functions $\phi_j(x)$ and take the form:

$$y(x, w) = f\left(\sum_{j=1}^{M} w_j \phi_j(x)\right) \tag{2.11}$$

Here $f(\cdot)$ is a non–linear function (e.g., sigmoid). The goal of ANN is to make $\phi_j(x)$ depend on parameters, then allow these parameters to be adjusted along with the coefficients w_j (Bishop, 2006).

Basic ANN is constructed as follows:

[35] The number of layers and the number of neurons is a hyperparameter that must be optimised.

- M linear combinations of input variables $(x_1, \ldots, x_D)$ are constructed using:

$$a_j = \sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \tag{2.12}$$

  where the superscript (1) indicates that the corresponding parameters are in the first layer, w_{ji} encodes weights, and w_{j0} – bias parameter.

- Calculated activations a_j are transformed using a differentiable, non–linear activation function $h(\cdot)$ (e.g., sigmoid):

$$z_j = h(a_j) = \frac{1}{1 + exp(-a_j)} \tag{2.13}$$

- Calculated values z_j are linearly combined to give output activations:

$$a_k = \sum_{i=1}^{M} w_{ki}^{(2)} x_i + w_{k0}^{(2)} \tag{2.14}$$

  This transformation corresponds to the second ANN layer where w_{k0}^2 are bias parameters and $k = 1, \ldots, K$ is the total number of outputs.

- The output unit activations are transformed using an appropriate activation function to give a set of outputs y_k:

$$y_k = \sigma(a_k) \tag{2.15}$$

- In the last ANN layer, output is normalised (the most commonly used transformation function for the output layer is Softmax):

$$y_k(x, w) = \frac{\exp\left(a_k(x, w)\right)}{\sum_j \exp\left(a_j(x, w)\right)} \tag{2.16}$$

The ANN predicted value is compared to the true value and based on the given loss function, the ANN model adjusts the weights. Minimisation of a loss function can be achieved in two ways (Murphy, 2012):

- *Forward propagation* – at the beginning of the training *epoch* random weights for the network layers are initialised. When the input data is forward propagated through the ANN and predicted values are provided by the last layer, a comparison is performed between predicted and true values.

- *Backpropagation* – initially uses random values for model parameters and optimises them by sequentially updating the parameters from the last layer to the first ANN layer. This is achieved by measuring the squared difference between the predicted and desired values. The backpropagation is repeated with the new weights that minimise the total error until the optimum weights are found (Krogh, 2008).

2.3.5.3 Cross–Validation

This section forms the foundation of ***Section 6.5.5*** *and will describe the technique used for model appropriateness.*

Bias and variance dilemma in statistical learning is a common problem for the trade–off between good generalisation and to avoid over–training. When working with statistical learning methods, one should separate the data into training set, test set and validation set to not evaluate the model on the same data as was used in the learning process. In such a case, cross–validation is a primary approach to validate the appropriateness of the classification algorithm to the given problem (Kurtz, 1948). The main goal of cross–validation is to verify the replicability of results. Subsets of the data are held out for use as testing sets and the model is fitted to the remaining data (a training set) and used to predict for the testing set. Averaging the quality of the predictions across the validation sets yields an overall measure of prediction accuracy.

Cross–validation is applied as follows: Let us introduce a dataset $D = \{(x_i, y_i), i = 1, \ldots, n\}$ where one wants to apply a regression model M to estimate values $\hat{y}_i, i = 1, \ldots, n$. Then, it is possible to partition D

dataset into two sets: $D = D_1 \cup D_2$, with k data in D_1 and $n - k$ data in D_2. The desired regression model M can be fitted using D_2, this step is called model training and D_2 is correspondingly a training set. Then, trained model M is used to obtain estimates $\hat{y}D_1$ using data D_1. This step is known as testing and D_1 is called a testing set. There are $\binom{n}{k}$ possible partitions of the original set. The cross–validation error is the averaged prediction error over all test iterations.

When $k = 1$, the validation process is called leave–one–out cross validation (test set has cardinality 1, and each of the $i = 1, \ldots, n$ partitions are used to train and then test the model), which is calculated using:

$$\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}^{-i} \right)^2 \tag{2.17}$$

where $\hat{y}^{-i}$ is the estimated value of a given leave–one–out iteration.

General Materials and Methods

This chapter briefly summarises all the samples, genomic datasets, and computational tools used in this study. More detailed, case specific data analysis applications are provided for each application section in **Chapter 6**.

3.1 Samples Analysed

All used samples are summarized in the **Table 3.1** bellow, while more detailed information for each study is presented in the mentioned application chapter. In summary, we processed samples from two mammalian organisms – *Homo Sapiens* and *Mus Musculus*. DNA was extracted from multiple organs or cell–types (e.g., Chorionic villi, mESCs) while samples pertained to distinct genotypical or experimental groups (e.g., wild–type, knockout). In total 3.5 billion reads were processed that attributed to almost one terabyte of raw next–generation sequencing data.

Table 3.1 | Samples Analysed

Number of samples processed in specific TOP–seq application study. This table contains rough amount of samples per specific experimental group.

Study name	DNA modification	Organism	Source of DNA	Sample group	Number of samples
TOP–seq	uCG	*Homo Sapiens*	Prefrontal brain cortex	Brain	4
TOP–seq	uCG	*Homo Sapiens*	IMR90	IMR90	5
TOP–seq	uCG	*Homo Sapiens*	LA1–55n	N–type	3
TOP–seq	uCG	*Homo Sapiens*	LA1–55s	S–type	3
hmTOP–seq	5hmCG	*Mus Musculus*	mESC	+ BGT	6
hmTOP–seq	5hmCG	*Mus Musculus*	mESC	− BGT	6
caCLEAR	5caC	*Mus Musculus*	mESC	wild–type	4
caCLEAR	5caC	*Mus Musculus*	mESC	*Tdg−/−*	4
caCLEAR	5caC	*Mus Musculus*	mESC	Tet TKO	2
NIPT	uCG	*Homo Sapiens*	cfDNA	Pregnant female	13
NIPT	5hmCG	*Homo Sapiens*	cfDNA	Pregnant female	11
NIPT	uCG	*Homo Sapiens*	cfDNA	Non–pregnant female	7
NIPT	5hmCG	*Homo Sapiens*	cfDNA	Non–pregnant female	7
NIPT	uCG	*Homo Sapiens*	Chorionic villi	Pregnant female	7
NIPT	5hmCG	*Homo Sapiens*	Chorionic villi	Pregnant female	3

3.2 Genomic Datasets

Reference genome sequences were downloaded from the UCSC database (Karolchik, 2004):

- *Homo Sapiens* (build GRCh37/hg19; 2009) – 28 million CG sites
- *Mus Musculus* (build GRCm38/mm10; 2011) – 21.3 million CG sites

CG sites in a given genome sequence were identified using in–house script and only those originating from the autosomes and chromosome X were used in the downstream analysis.

Homo Sapiens genome elements were downloaded from the following sources:

- Genes – GENCODE genes (Frankish et al., 2018). All used gene biotypes (e.g., protein–coding genes, long–intergenic RNAs) were taken from this source. Accordingly, exon and intron sets were generated. Upstream (i.e., promoter), donwstream regions were defined by two kilobases upstream or downstream from the gene starting and ending coordinates.
- CGIs – the UCSC database (Gardiner-Garden and Frommer, 1987; Karolchik, 2004). CGIs were assigned promoter, intragenic, intergenic status according to their position to a nearest protein–coding gene using hierarchical assignment. First, CGI regions were intersected with protein–coding gene promoters and only those regions that did not intersect were used for intragenic CGI assignment. CGIs that were not assigned promoter or intragenic status were defined as intergenic. CGI shores were defined as region $-/+$ two kb around the CGI regions, and CGI shelves – $-/+$ two kb regions around the CGI shores.
- Major repeat families (SINE, LINE, LTR, DNA, simple repeats) (Jurka, 2000; Karolchik, 2004).

- Mappability score – the level of sequence uniqueness of the reference genome assembly (Derrien et al., 2012; Karolchik, 2004).
- GERP score – estimates of sequence evolutionary constraint (Cooper et al., 2005; Davydov et al., 2010; Karolchik, 2004).
- Single nucleotide polymorphisms (SNP) (Karolchik, 2004; Sherry et al., 2001).
- Lamina associated domains (Karolchik, 2004).
- Gaps in the assembly (Karolchik, 2004).
- Placental enhancers – the enhancer atlas (Gao et al., 2016).
- Epigenome Roadmap chromatin states – expanded chromatin model was downloaded from the Epigenome Roadmap project database for the following datasets: cell–line IMR90 (E017) and brain dorsolateral prefrontal cortex (E073) (Bernstein et al., 2010; Kundaje et al., 2015). Chromatin states were defined as follows: EnhA – active enhancers, EnhBiv – bivalent enhancer, EnhG – genic enhancers, EnhWk – weak enhancers, Het – heterochromatin, Queis – quiscent/low, ReprPC – repressed polycomb, ReprPCWk – weak repressed polycomb, TssA – active transcription start site, TssBiv – bivalent TSS, TssFlnk – flanking TSS, TssFlnkD – flanking TSS downstream, TssFlnkU – flanking TSS upstream, Tx – strong transcription, TxWk – weak transcription, ZNF/Rpts – Zinc finger genes and repeats.
- ARIES mQTL probes (Gaunt et al., 2016).
- High confidence Illumina probes (Naeem et al., 2014).

Sequence characteristics, such as GC content, CG content, cytosine ratio were computed manually using in–house scripts implement in `R`.

Additional genomic and epigenomic *Homo Sapiens* datasets were downloaded from the following sources:

- WGBS from the adult frontal lobe (GEO accession GSE46710) (Wen et al., 2014)

- WGBS IMR90 dataset 1 (GEO accession GSM432687) (Lister et al., 2009)
- WGBS IMR90 dataset 2 (GEO accession GSM1204464) (Ziller et al., 2013)
- MBD–seq IMR90 (GEO accession GSM947460) (Bert et al., 2013)
- MRE–seq IMR90 (GEO accession GSM830153) (Xie et al., 2012)
- SeqFF (Kim et al., 2015)

Mus Musculus genome elements were downloaded from the following sources:

- Genes – GENCODE genes (Frankish et al., 2018). Gene related annotations were constructed the same way as for the *Homo Sapiens*.
- CGIs – the UCSC database (Gardiner-Garden and Frommer, 1987; Karolchik, 2004). CGI related annotations were constructed the same way as for the *Homo Sapiens*.
- Major repeat families (SINE, LINE, LTR, DNA, simple repeats) (Jurka, 2000; Karolchik, 2004).
- Histone marks – were download from the ENCODE project consortium (Davis et al., 2018; Dunham et al., 2012).
- Transcription factor chromatin immunoprecipitation regions – were download the GEO database (GEO accession GSE11431) (Chen et al., 2008).
- Open chromatin ATAC–seq regions – were download the UCSC database (Karolchik, 2004).
- 5hmC TAB–seq signal – was downloaded from the GEO database (GEO accession GSE36173) (Yu et al., 2012b).

3.3 Experimental Procedures

3.3.1 Processing of Additional Datasets

In downloaded *Homo Sapiens* WGBS datasets we only considered CG methylation and averaged beta values across the strands. Additionally, IMR90 dataset 2 was filtered for CG sites with coverage greater than four. Continuous signal values downloaded from the UCSC database were in the `bigWig` file format. It was converted to the `bedGraph` format using `bigWigToBedGraph` conversion tool. For IMR90 MRE–seq dataset signals from both strands were summed and genomic coordinates were lifted to the genome build hg19 using `liftOver` tool. IMR90 MBD–seq dataset was lifted to the genome build hg19 and the same value was assigned to all the CG sites that intersect a particular MBD–seq region. mESC TAB–seq dataset was also converted from the original to the mm10 genome build using the `liftOver` tool.

3.3.2 Training Neural Network

Neural network model used to compute *nn*–estimate values discussed in **Section 5.4** was designed using the `sklearn` tool in `python` environment (Pedregosa et al., 2011). `MLPRegressor` function was used to train the model with the following parameters: i) `hidden_layer_sizes = 43` (the number of neurons in the hidden layer); ii) `activation = 'relu'` (use the rectified linear unit function as an activation function); iii) `solver = 'adam'` (use stochastic gradient–based optimizer as a solver); iv) `alpha = 0.00005`.

3.4 Computational Tools

Following bioinformatical tools were used to process and analyse next–generation sequencing data: `BEDTools`, `bwa`, `cutadapt`, `FastQC`, `FASTX--Toolkit`, `litfOver`, `samtools` (Babraham Bioinformatics, 2019; Hannon Lab, 2010; Hinrichs et al., 2006; Li and Durbin, 2009; Li et al., 2009; Martin, 2011; Quinlan and Hall, 2010). Finally, a significant part of data visualisation and inspection was performed using the UCSC Genome Browser (Kent et al., 2002).

To test similarity, statistical significance or intersection between given datasets following functions in `R` environment were used: Fisher's exact test (`fisher.test`), Student's t–Test (`t.test`), analysis of variance (`aov`), correlations (`cor`), Jaccard (in–house script implemented in `R`) (R Core Team, 2019). If not specified otherwise all other data analysis was also performed in `R` environment (versions 3.3 – 3.5) using official expansions. Some of these expansions are `data.table`, `ggplot2` (Dowle and Srinivasan, 2020; Wickham, 2016).

3.5 Hardware Infrastructure

All the data was analysed and results were generated using two computational stations: Intel(R) Core(TM) i5–4460 with 4 physical cores and 16 GB of memory running GNU/Linux (Ubuntu 16.04.4 LTS); Intel(R) Xeon(R) Gold 6126 with 48 physical cores and 284 GB of memory running GNU/Linux (CentOS 7.6).

3.6 Data Availability

Raw and processed data from which main conclusions in this thesis were drawn were deposited to the GEO database under the following accession numbers: TOP–seq signal in human derived cell–lines – GSE91023; hmTOP–seq signal in mESCs – GSE140206; caCLEAR signal in mESCs – GSE142319; TOP–seq and hmTOP–seq signals in prenatal testing analysis – GSE148964.

Processing Tethered Oligonucleotide–Primed Sequencing Data

Galbūt ne garsas skamba tyloj,
o tyla garse,
Ką apie tylą žinome mes?

Foje

4.1 Introduction

This section presents the methodology developed for the processing of the TOP–seq sequencing data. This workflow is a modified version of widely used bioinformatics pipelines and is applicable not only for the unmodified DNA sequencing method (i.e., TOP–seq) data analysis but also for other variations (i.e., hmTOP–seq – 5hmC modification, caCLEAR – 5caC modification). In this chapter, the term "TOP–seq" is used however it is important to mention that processing procedures are applicable to other variations. This workflow provides the basis for all the results described in this thesis and for the publications listed in **Section 1.6**.

The TOP–seq processing workflow consists of four major steps:

- Sequencing read processing
- Mapping reads to a reference genome
- PCR duplicate removal
- Assigning reads to CG sites

The different source of genomic DNA (e.g., bacteriophage lambda, eukaryotic cell) or different targeted DNA modifications (e.g., unmodified CG sites, 5hmCG sites) might require different TOP–seq data processing methodology. The most commonly used, which is adopted for unmodified CG sites in eukaryotic cells, is described below in detail with recommended parameters for other specific cases. The processing parameters used for each tool might affect the downstream parameters so it is necessary to have thorough knowledge of each tool. Furthermore, the TOP–seq method can generate tens of millions of sequencing reads corresponding to tens of gigabytes of data, the analysis of which requires intensive computational processing steps and semi–powerful or even powerful hardware infrastructure.

This chapter is divided into five major parts, with four sections presenting a separate workflow step (mentioned above) and one section discussing the advantages and disadvantages, recommendations and ideas that could be implemented in the future to improve processing of high–throughput tethered oligonucleotide–primed sequencing data.

4.2 Sequencing Read Processing

Raw sequencing reads come from the next–generation sequencing machine and are typically in the `FASTQ` file format which includes (for each read) a unique identifier, the nucleotide sequence, and a Phred quality score for each nucleotide.

Processing of the raw reads takes place in four steps that are visualised in a **Figure 4.1** flowchart. The initial step uses `fastq_quality_trimmer` command (implemented in `FASTX-Toolkit`) to remove sequencing reads that are too short (Hannon Lab, 2010). This step is a speed optimisation for the consequent steps as reads that are too short usually do not contain 5' and/or 3' sequencing adapters, thus cannot be classified as suitable for the analysis. We used 80 nucleotides as a filtering threshold but

this length parameter might vary between different experiments [1]. On average, this procedure discards 21.3% ($sd = 3.6$) of reads (**Figure 4.2**, **Supp. Table 1**). In **Figure 4.3 A** panel, these reads can be identified in the top row. Before short read removal, read length distribution is bimodal — the first peak is around 50 nucleotides and the second peak is around 190 nucleotides. However, after the mentioned brute force read discard approach, only the right side of this distribution is kept and further processed in the consequent workflow steps.

Then, `cutadapt` suite was used to remove or trim adapter sequences from the 5' and 3' ends of the reads (Martin, 2011). The parameters to cut 5' adapter were as follows: i) `-error-rate = 0.1` (maximum allowed error rate); ii) `-overlap = 10` (minimal overlap length between read and adapter for an adapter to be found); iii) `-trimmed-only` (retain reads only with found adapter sequence) to increase quality of maintained reads. iv) depending on experiment parameter `-front` was changed (sequence of an adapter ligated to the 5' end) (Gibas et al., 2020; Gordevičius et al., 2020; Ličytė et al., 2020; Staševskij et al., 2017). It is worth noting that `-front` parameter was prepended with `^` symbol to identify the adapter only if it was a prefix of the read. **Figure 4.3 B** visualises the general read structure before and after removal of the 5' adapter sequence. A non–random pattern of bases can be seen at the beginning of reads, which is caused by the predefined 5' adapter sequence. After removing 5' adapter sequence, only a random distribution (i.e., around 25% of each base) is left. Usually, a very high amount of reads (99%, $sd = 1.6$) contain a 5' adapter as shown in **Figure 4.2** (**Supp. Table 1**). The presence of the 3' adapter is represented by a guanine rich pattern that reaches a fractional maximum around 200 nucleotides in **Figure 4.3 B**. To trim the 3' adapter, the `cutadapt` suite was used with the same `-error-rate` and `-overlap` parameters as for 5' adapter and `-adapter` parameter (sequence of an adapter ligated to the 3' end) was adjusted according to experiment (Gibas et al., 2020; Gordevičius et al., 2020; Ličytė et al., 2020; Staševskij et al., 2017). However, it is worth noting that the `$` symbol which is usually appended to the adapter

[1] Dependency on the next–generation sequencing machine or different length of sequencing adapters.

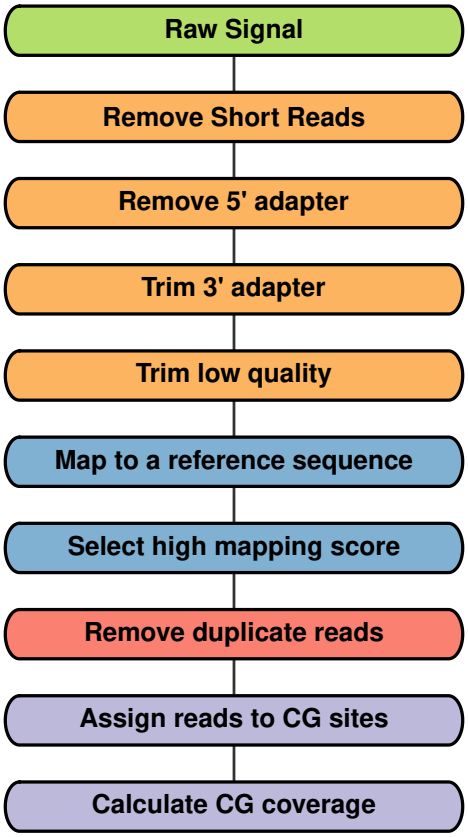


Figure 4.1 | TOP–seq Read Processing Workflow

Workflow of the current TOP–seq read processing approach to achieve genome–wide DNA (un)modification signal. Each processing phase is indicated with a distinct colour code. Raw next–generation sequencing reads are first processed to achieve higher quality set of reads that are mapped to a reference sequence. Duplicate reads are then removed using custom algorithm and finally a single CG resolution DNA (un)modification signal is computed.

sequence (equivalent of ^ symbol) was not used as it might have a negative impact on the results. Since a fraction of reads do not contain the full 3' adapter sequence as DNA polymerase may stop its synthesis after the genomic DNA part or sometimes mid 3' adapter sequence, the parameter `-trimmed-only` was not used as for the 5' adapter and a change in read number was not observed.

Finally, the `fastq_quality_trimmer` command is used to increase the quality of the reads (Hannon Lab, 2010). Typically, the 3' end of the

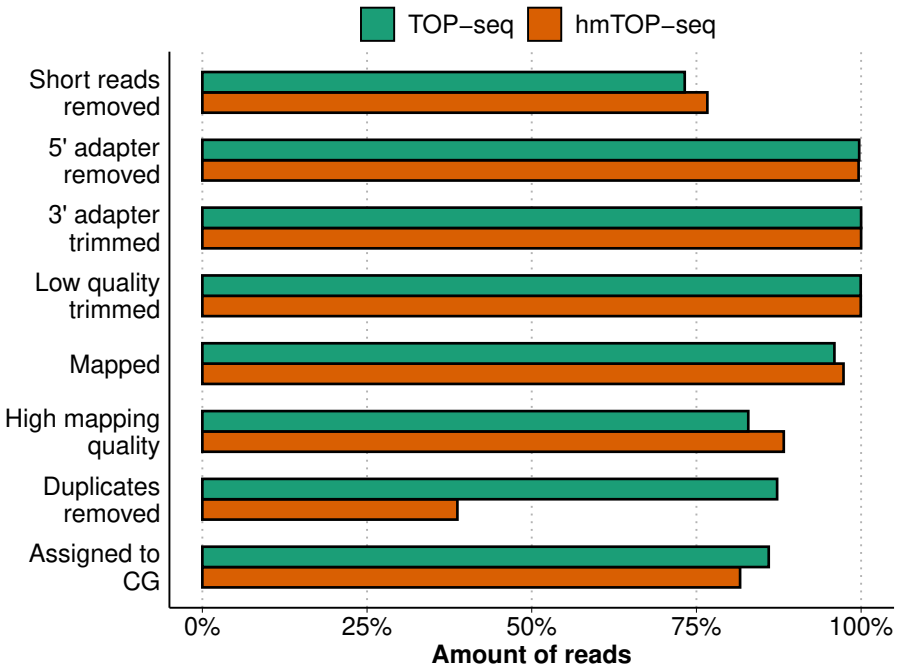


Figure 4.2 | Change in Amount of Reads

The absolute change in the amount of reads after each workflow step in cfDNA sample (sample identifier 137) from the NIPT study using TOP–seq and hmTOP–seq methods. The change was calculated as a percentage from the total amount of reads in the preceding workflow step (starting number of reads is 37 and 18 millions in TOP–seq and hmTOP–seq methods, respectively). It is apparent that 25% of reads are discarded as too short in both the TOP–seq and hmTOP–seq methods. The quantity of reads does not change much during adapter removal and quality trimming steps. In the hmTOP–seq method, around 60% of reads are classified as duplicates, while in the TOP–seq method only 15% of reads were removed in this step.

reads has a lower Phred quality score, which may cause false mapping to the reference genome due to nucleotide mismatch between a specific read and a reference sequence. The `FASTX-Toolkit` was used to trim read ends having Phred quality score below 20 (`-t 20`). Additionally, the `FASTX-Toolkit` removed reads that were too short for alignment after removing the adapter sequences and low quality nucleotides (used 15 nucleotides as a minimum length `-l 15`). This step ensures faster alignment and higher mappability score. The quality score before and after trimming according to a given threshold is visualised in **Supp. Figure 1**.

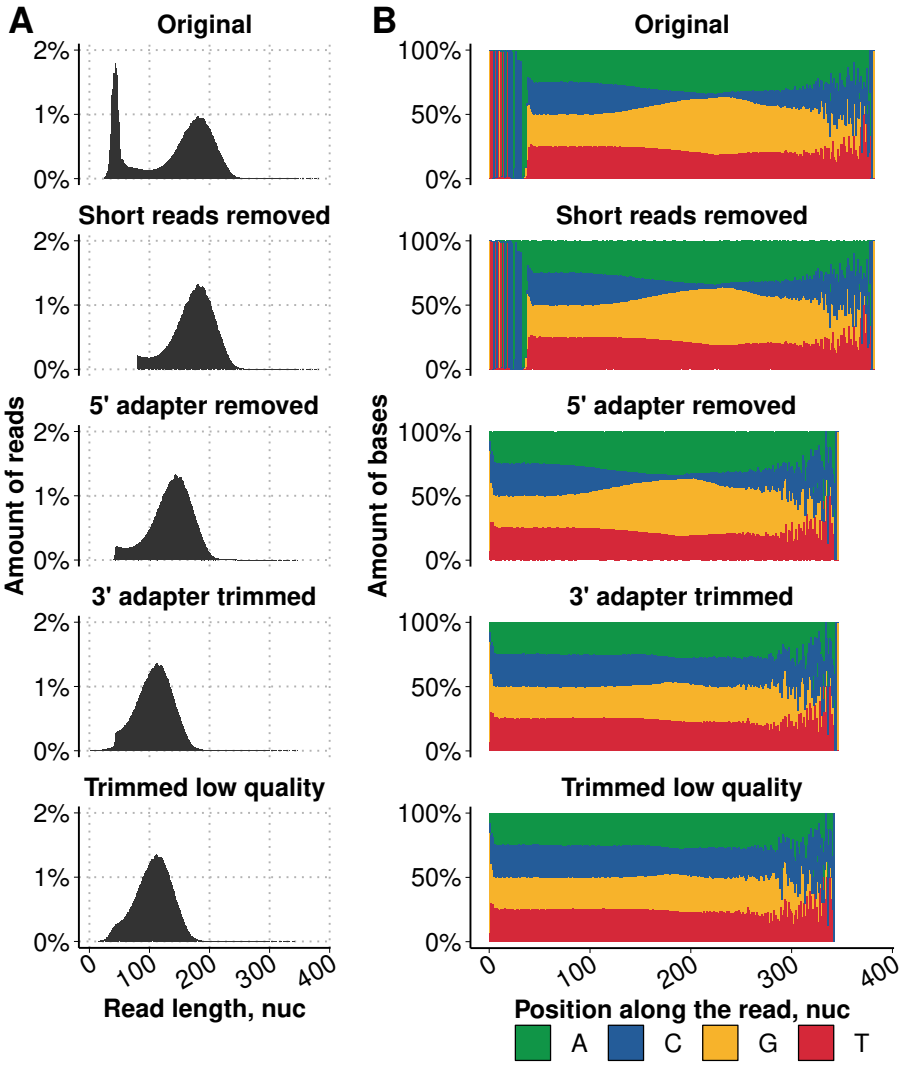


Figure 4.3 | TOP–seq Read Structure

(A) Distribution of read lengths after each processing step in cfDNA sample (sample identifier 137) analysed the using TOP–seq method. Original sequencing reads show bimodal distribution and only the longer reads are kept for the consequent data analysis. **(B)** Base composition measured in percentage along sequencing reads. In the first two processing steps, non–random base composition at the beginning of reads represent the 5' adapter sequence. At the final step, each base composition is close to the 25% — expected proportion in the human genome.

After each processing step, the `FastQC` suite was used (with default parameters) to generate a quality report (Babraham Bioinformatics, 2019). This seemingly redundant procedure ensures that adapter sequences were removed correctly, the Phred score is high enough and, in the case of TOP–seq library type, reads begin with a CG dinucleotide. At the end of this raw read processing pipeline, we are left with a normal distribution of read lengths with a peak around 100 nucleotides. Nucleotide distribution along the reads is close to random (except the beginning of the read) which is expected as reads are originating genome–wide. The read start is visualised in **Supp. Figure 2**, where it is apparent that most reads originate from a CG site.

4.3 Mapping Reads to a Reference Genome

Processed TOP–seq reads can be mapped to a reference genome using standard algorithms and tools (e.g, `bwa mem` or `bwa aln`) (Li and Durbin, 2009). However, alternative methods, such as `Salmon` quasi–aligner could be also used (Patro et al., 2017). For the standard TOP–seq analysis, the `bwa mem` command was used with default parameters (except for the `idxbase` parameter – reference genome, which depended on the experimental design) (Gibas et al., 2020; Gordevičius et al., 2020; Ličytė et al., 2020; Staševskij et al., 2017).

`samtools` tool was used to convert the `bwa` alignment `SAM` file format into a `BAM` file format (Li et al., 2009). `BAM` file is accordingly sorted and subsetted for reads with a mapping quality equal or greater than 30 (`samtools sort` and `samtools view` commands). Mapping quality represents how unique each alignment is in the genome and is equal to an integer that is closest to $-10 \log_{10} P(\text{mapping position is incorrect})$. **Figure 4.4** represents mapping quality distribution for a single cfDNA sample. In most experiments the mapping quality distribution is bimodal (first mode group at the lowest mapping quality and second mode group at the highest mapping quality), therefore it was decided to divide this

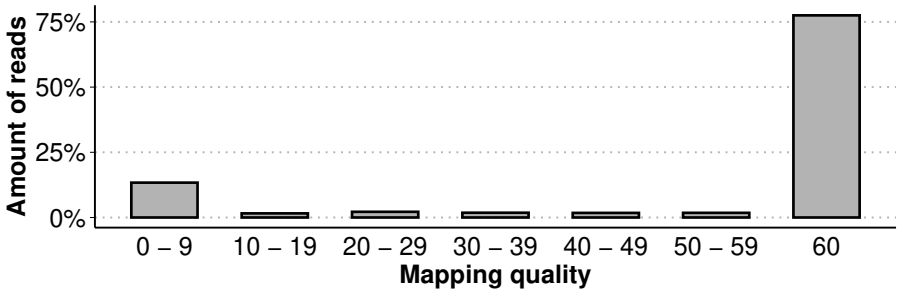


Figure 4.4 | TOP–seq Mapping Quality

Distribution of mapping quality values in cfDNA sample (sample identifier 137) computed using `bwa` tool. For better representation, values were binned into groups with a maximum mapping quality value 60 having its own group. Usually, the mapping quality forms a bimodal distribution with peaks centring at the lower and higher ends.

distribution into two parts (at 30), and use all reads that fall on the right side of the distribution.

4.4 PCR Duplicate Removal

PCR duplicates occur when shattered DNA fragments (e.g., sonicated) are amplified by the PCR method. In this case, the same DNA fragment will be amplified and sequenced multiple times. These identical reads will use space on a sequencer flow cell. Moreover, when the depth of sequencing coverage is an important factor (e.g., TOP–seq method), PCR duplicates can obstruct the true amount of DNA molecules and erroneously inflate the coverage (Marx, 2017). Most sequencing pipelines recommend marking and removing them using either unique molecular identifiers or computational tools such as `Picard` or `samtools` (Ebbert et al., 2016; Li et al., 2009).

Standard tools find PCR duplicates by identifying groups of reads that align to the same exact start and end positions in the genome (i.e., external mapping coordinates are identical) by assuming that the probability for reads to align to the same position is very low (actually, at least for the human genome, it is close to zero). However, such identification

strategy is not applicable to TOP–seq based sequencing methods since TOP–seq methods target and enrich specific genomic positions (i.e., CG dinucleotides), resulting in reads aligning to exactly the same genomic coordinates and conventional computational tools would not work in this case. The optimal solution for this PCR duplicate problem would be unique molecular identifiers (Kivioja et al., 2011), however, at the time of writing this thesis, the usage of unique molecular identifiers for the TOP–seq method was still in the early stages, so a different approach was needed. We developed a PCR duplicate identification and removal algorithm similar to canonical ones, however is not as stringent. In our PCR identification algorithm, all reads that start at exactly the same genomic coordinate on the same strand and have the same original length were classified as duplicates and only one read per each group was retained. This strategy is similar to the canonical approach as it evaluates the starting position (5' end) of the read, however, by evaluating the original read length, this algorithm takes into account the length of 5' and 3' adapters. DNA polymerase tends to prematurely stop the synthesis of the 3' adapter and sometimes skip nucleotides in the 5' adapter and such variation will be unique to a specific PCR duplicate group. **Figure 4.5** shows the variation in the 5' and 3' adapter lengths which creates Δ space to classify new groups of PCR duplicates. In summary, instead of removing all reads that have identical mapping coordinates (except the one that is left as a group representative), this algorithm retains $m \times n$ reads where m and n is number of different 5' and 3' adapter lengths accordingly.

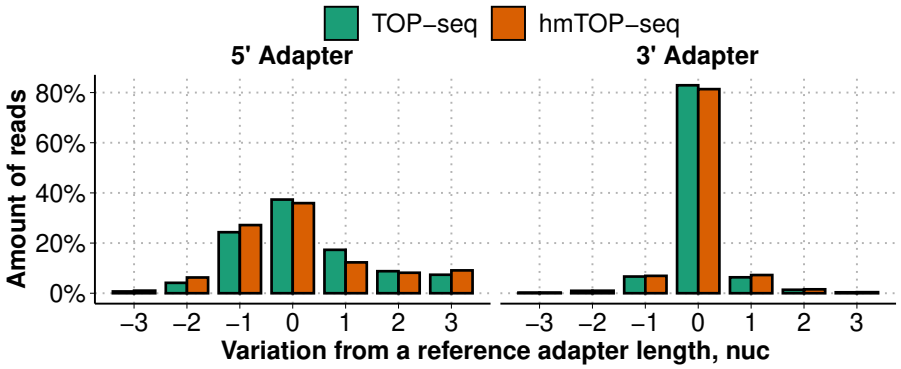


Figure 4.5 | Adapter Length Variation

Variation in adapter length as observed in cfDNA sample (sample identifier 137) analysed using TOP–seq and hmTOP–seq methods. Most encountered adapters had the expected length (i.e., 37 nucleotides for 5' adapter and 32 nucleotides for 3' adapter, thus their Δ is equal to 0). However, a fraction of encountered adapters were shorter or longer with most Δ values distributed between -3 and 3 nucleotides.

4.5 Assigning Reads to CG Sites

For each mapped read, we computed the distance from its starting position to the nearest CG dinucleotide (i.e., distance measured in nucleotides from the reads 5' end). Depending on the modification type, different distance thresholds were selected to assign reads to CG sites. For the TOP–seq library strategy, an absolute distance of three nucleotides was used and within this threshold on average 90% ($sd = 5.2$) of reads were assigned to CG sites (**Supp. Table 1**). Meanwhile for hmTOP–seq and caCLEAR methods, an absolute distance of four nucleotides was used (on average retaining 85% of reads with $sd = 4.2$). Summarised read distances can be found in **Table 4.1**.

Table 4.1 | Distance to CG sites

Distance threshold in nucleotides used to assign reads to the nearest CG site in a particular TOP–seq library strategy.

Library strategy	Distance (nuc.)
TOP–seq	3
hmTOP–seq	4
caCLEAR	4

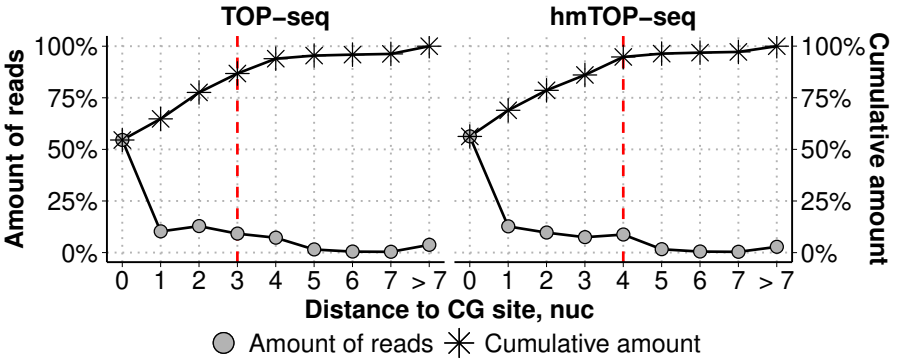


Figure 4.6 | Read Distance to CG sites

The distribution of the absolute distance from the read start to the nearest CG site. Y–axis on the left side represents the exact amount of reads that start from a nearest CG site within a given distance. Y–axis on the right side represents the cumulative sum of read quantity that starts within a given threshold from the nearest CG site. Red dashed line represents used distance threshold for read assignment to a CG site.

The decision to use different distances for different library strategies was made after inspecting scree–like plots (usually used in PCA to visualise eigenvalues for each component). The scree plot is used to determine the number of factors to retain using "elbow" rule (first sharp change in the slope indicates number of factors to use) (Cattell, 1966). The absolute distance from the CG site and amount of reads that started at exactly that distance are visualised in **Figure 4.6** and it is evident that most reads can be retained using an absolute distance of three or four. Interestingly, we also observed strand specific distance distributions (**Supp. Figure 3**). For each separate strand reads started exactly at the CG site with a very small fraction starting in the upstream direction.

After assigning reads to CG sites CG–coverage was calculated (defining coverage as the total number of reads on any strand starting within the given distance threshold). Such procedure usually divides genomic CG sites into two opposite groups (CG sites with coverage greater than 0 – identified CG sites and CG sites with coverage equal to 0 – non–identified CG sites). Depending on the study, all or only a fraction of identified CG sites were used. Strategies for selecting identified CG sites, general coverage statistics and coverage comparison with other DNA modification analysis methods are mentioned in **Chapter 6**. Finally, sequence

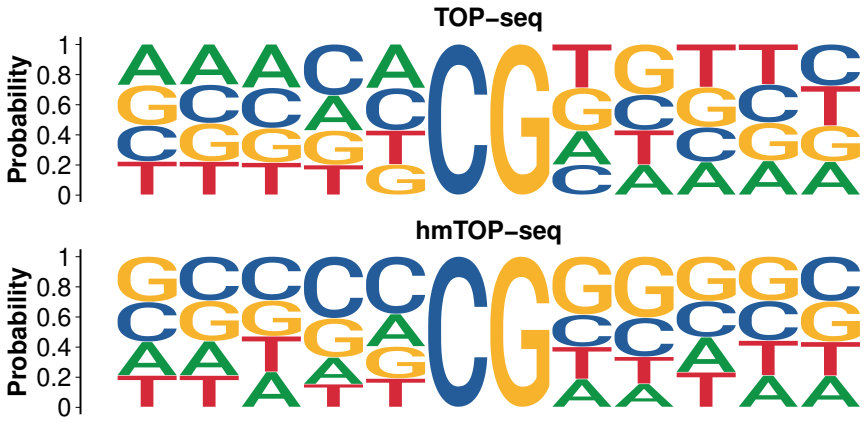


Figure 4.7 | Nucleotide Composition Around CG sites

Probability of nucleotide composition around 50 thousand randomly selected CG sites identified in cfDNA sample (sample identifier 137). At the centre of the generated sequence there is an identified CG site with most encountered nucleotides flanking it. Each nucleotide is indicated with a distinct colour code.

composition around identified CG sites was extracted (**Figure 4.7**). It is evident that for TOP–seq library reads are starting at CG sites and that there is no other observable sequence bias in flanking sequences (nucleotides around the CG site have equal probability to be observed). However, in hmTOP–seq library strategy there was a small bias for identified CG sites towards C and G rich loci.

4.6 Discussion

4.6.1 The Implications and Applications of This Methodology

This section summarised a newly developed sequencing data processing pipeline specifically tailored for TOP–seq high–throughput epigenomic data. This methodology empowered the usability and applicability of TOP–seq based techniques to process TOP–seq data from raw sequencing reads produced by next–generation sequencing technologies and to

compute the DNA modification signal by calculating the coverage of CG sites.

The workflow was presented in detail summarising the processing of the sequencing reads before alignment to a reference sequence, alignment to a reference sequence, duplicate read removal, and read assignment to CG sites. Notably, all used computational tools and parameters are specified for the full reproducibility of this methodology. It is important to mention that the duplicate read removal algorithm was specially developed for the TOP–seq method as standard duplicate removal techniques would not be applicable for this method and probably would distort the true coverage signal. Finally, this processing pipeline contains a couple of speed optimisation steps, such as removing relatively short reads in the beginning of the pipeline, to reduce the processing time of low quality reads.

4.6.2 The Difficulties in Processing TOP–seq Data

Most difficulties of analysing TOP–seq data arise from a large range of computational tools used in this data processing pipeline. Since this pipeline consists of multiple sequential steps, in theory, it could be parallelised on a multi–core computing machine. However, not all of the used tools are customised for parallelisation. For example, tools used for adapter and quality trimming cannot be parallelised and can only process sequencing reads sequentially, while tools used for read alignment are fully parallelisable.

Additionally, one peculiar problem arises from assigning reads to CG sites. It is possible that a specific read would start from its original CG site with an absolute distance greater than zero. Usually, assignment of such reads to their original CG site is straightforward, however, it is possible that within the same distance there is another CG site and in such a case, singular assignment is impossible. Consider a read that starts -2 nucleotides from its original CG site a, from the CG site a there is another CG site b in -4 nucleotides direction, then the mentioned read

would have same absolute distance 2 to the CG site a and CG site b and in such case, unique assignment would be impossible.

4.6.3 Unanswered Questions and Future Research Directions

Since we observed a specific distance distribution between the start of the read and CG sites, it could be further applied in our further data analysis. For a positive strand, most reads started exactly at the CG site with a small fraction starting in the upstream direction, the same tendency was observed for the negative strand too. It is possible to further implement this feature while assigning reads to CG sites. Instead of using a straightforward approach and assigning all reads to CG site that start within -4 to +4 bp window, they can be retained in a strand specific manner – selecting reads that start exactly at the CG site or in a small upstream direction shift. Such an approach should decrease background noise originating from the falsely assigned reads.

As the TOP–seq method is based on enrichment of target genomic regions (i.e, CG sites), additional improvement of its data analysis would be partitioning the genome into peak and non–peak regions. In such a case, the current high resolution would be lost, however, the different perspective on DNA modifications might shed light on different epigenomic processes. The TOP–seq method could be fully used at a single base resolution, however, when inspecting higher–order genomic elements, such as lamina associated domains or even genes, it might be useful to compute DNA modification peaks. A naive implementation of peak calculation was proposed during the TOP–seq data analysis process but it was never fully implemented and further tested. For such implementation, a high confidence reference dataset would be needed but more importantly, an appropriate algorithm must be fully developed. Since some of the currently most popular peak calling algorithms assume symmetrical read coverage on both strands and specific distributions around targeted sites, they are not possible to use with TOP–seq based methods (Wilbanks and

Facciotti, 2010; Zhang et al., 2008). A further investigation, algorithm development and testing is needed to confirm peak calling applicability from TOP–seq high–throughput data.

To further improve TOP–seq method, unique molecular identifiers could be introduced into the library preparation step. Such technological addition to the method could be used to reduce errors and quantitative bias introduced by the DNA amplification. The unique molecular identifier technique incorporates a unique barcode onto each molecule. By incorporating individual barcodes on each original DNA sequence, the sensitivity of the DNA modification detection should increase. Computational tools could be used to filter out duplicate reads and report unique reads. This approach would be more precise than the current solution to filter out duplicate reads by evaluating the starting position of the read and the total read length.

Instead of using conventional and more popular read aligners, it is possible to implement newer read aligning procedures. The `Salmon` tool performs quasi–alignment to rapidly determine the set of reads compatible with a given reference sequence (Patro et al., 2017). Incorporation of such a tool would lead to a faster alignment of sequencing reads with a cost of resolution. However, in cases where the TOP–seq method is used to determine DNA modification level in a set of genomic regions, such as CGIs or genes, this approach could be more preferable.

Another speed optimisation could be achieved by parallelisation of the used tools that are not parallelised already. For those tools that parallel computing is not implemented, it is possible to submit sequencing reads divided into batches, which would require changing the read processing pipeline focus from a single sample towards a single batch. In theory, each set of original reads produced by a next–generation sequencing machine could be split into k non–overlapping read batches that would be submitted to a processing pipeline as individual units. This processing pipeline could be run using k–cores on a multi–core computing machine. At the end of the pipeline, parsed batches would be merged into a singular unit that would be used to calculate the DNA modification level.

In theory, such implementation should be faster, however, one should consider the time for dividing reads into batches and merging them back into a singular unit.

Finally, to further improve TOP–seq method reproducibility and applicability for other research, a fully developed code base is needed. Considering that the TOP–seq processing pipeline was developed using `R` programming environment, it would be useful to release an `R` package with version and quality controls, unit tests and continuous integration.

4.6.4 Concluding Remarks

Herein, we provide a detailed description of a next–generation sequencing data processing pipeline that was engineered to process TOP–seq based high–throughput epigenomic data. This efficient and scalable approach combines different conceptual ideas from published methods into a comprehensive pipeline. It covers steps and tools used to process TOP–seq signal from raw sequencing data to CG–coverage profiles.

Statement I — Developed computational methods to efficiently and accurately process TOP–seq based high–throughput epigenomic data. Created strategies that enable investigation of DNA modification signal at a single cytosine resolution in a strand specific manner.

Statistical Tools to Enhance the Quality of the TOP–seq Signal

Good is good, but better is better

Old Yiddish motto

5.1 Introduction

High–throughput methods, as an approach to measure and produce high quantities of data, might suffer from the inseparable effect of measurement bias, which can be expressed as intra–sample or inter–sample variation. In such cases, the calculated TOP–seq signal might also contain unwanted variation in coverage across samples within the same sequencing chip or between different sequencing chips. Such variation might be caused by the different sequencing depth, average modification level differences or other unknown biological or technological factors. Hence, three TOP–seq signal transformations were developed to reduce such variation effect and enhance the quality of the signal. First and most basic transformation is u–density which is based on weighted coverage level normalised by CG level, m–estimate and nn–estimate are TOP–seq coverage signal projections calculated using either exponential decay model or neural networks.

This methodology was developed and adapted for an unmodified DNA sequencing method and used to estimate h–density (i.e., 5hmCG modified DNA density), however, this estimate was not used in any major analysis and will not be covered here. This chapter comprises three main sections (one for each transformation): u–density, m–estimate, nn–estimate, with each section composed of three parts: reasoning behind the calculation

of the specific estimate, algorithm to calculate estimate and comparison of estimate with a reference method. Finally, the chapter closes with a discussion of the improvements that these transformations brought to TOP–seq signal applicability, difficulties in computing and applying transformations and unanswered questions for future research.

5.2 u–density

5.2.1 Motivation for Calculating the u–density Signal

Since the TOP–seq method might be affected by several limitations faced by other enrichment–based methods (e.g., signal quality dependence on the sequencing depth, bias towards a specific sequence context), statistical adjustments were applied to the TOP–seq signal. First, the sequencing depth influence was minimised by converting the TOP–seq coverage signal into weighted–density estimates (Parzen, 1962). Such conversion equalised the signal strength between different sequencing depth experiments. Of course, the simplest solution to this problem would have been calculating $\frac{n}{M}$, where n is the number of reads assigned to a specific CG site and M is the total number of reads in sample. However, this weighted–density approach allows us to exploit information from neighbouring CG sites and in this way, it maximised usability of low–coverage regions. To remove possible sequence context (i.e., CG) bias, we additionally normalised calculated weighted–density estimates by unweighted CG–density. The obtained signal was named u–density as it reflected unmethylated DNA density (**Figure 5.1**). An identical approach was used on the hmTOP–seq signal to calculate h–density.

5.2.2 Summary of the u–density Algorithm

Weighted–density estimates of TOP–seq coverage were computed using the Epanechnikov kernel over 2^{21} points uniformly distributed across

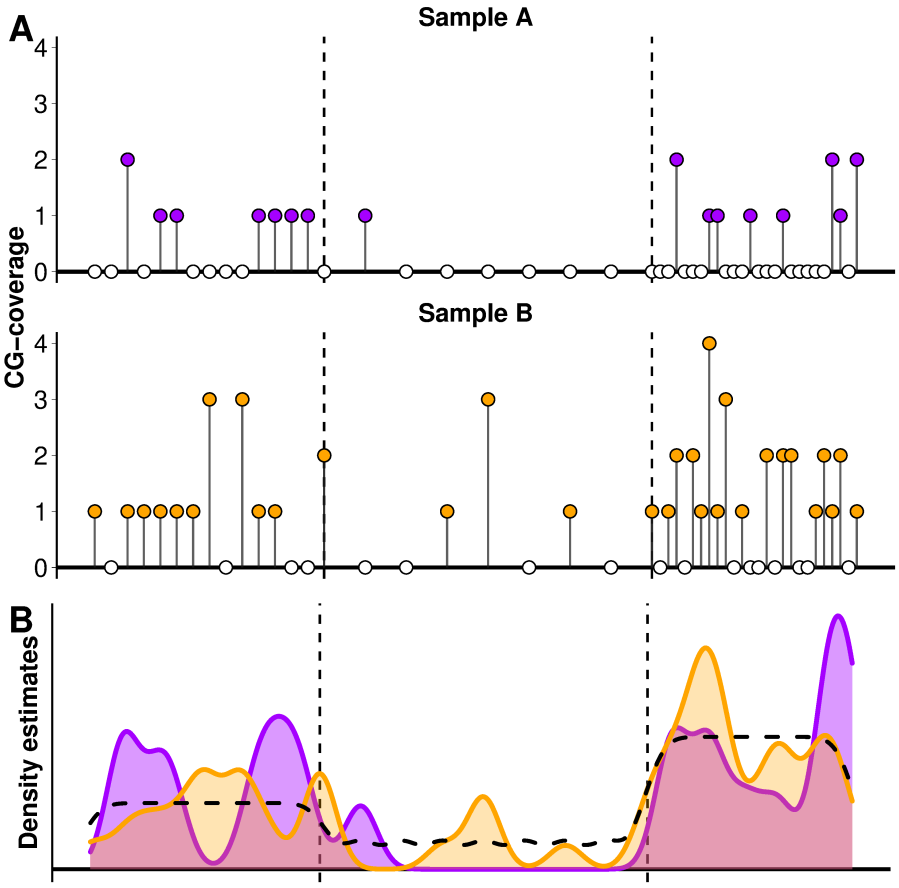


Figure 5.1 | TOP–seq Signal Along the Genomic Locus

Representation of DNA (un)modification signal at CG sites along the simulated genomic locus. **(A)** CG–coverage in two samples with different library sizes (Sample B contains 2.5 times more reads than Sample A). Circles indicate CG sites that are distributed non–randomly along the locus forming three distinct regions. Leftmost region contains average CG–density, middle region contains lowest, and rightmost region highest CG–density (regions are separated with vertical dashed lines). **(B)** Colored areas represent weighted coverage–density along the simulated locus. Even though original number of reads differed in samples their average weighted coverage–density is equal. It is apparent that regions with higher CG–density contain higher coverage–density too. Black dashed lined along the loci indicate simulated CG–density. Colour code represent signal from a specific sample.

each chromosome (computation workflow is represented in **Figure 5.2**). The number of equally spaced points at which the density is to be estimated was high enough to suppress the number of possible nucleotides [1]. The Epanechnikov kernel was chosen as it is optimal in a mean square error optimisation (Loftsgaarden and Quesenberry, 1965). Read counts were normalised to sum to 1 within each chromosome and used as weights for the density function **Equation 5.1**. The same approach with the omission of weights was used to estimate unweighted CG–density in the given chromosome **Equation 5.2**. Finally, TOP–seq unmethylation density were obtained by dividing weighted TOP–seq density by the unweighted CG–density at each CG dinucleotide **Equation 5.3**. After normalising weighted–density by CG–density, Gaussian kernel smoothing with the same bandwidth was used to interpolate respective density values at the exact positions of CG nucleotides [2]. Kernel bandwidth parameters were determined by scanning the TOP–seq *u*–density correlations in a wide range of kernel windows with the corresponding public IMR90 WGBS signal in human chromosome 1 (**Figure 5.3**) (Lister et al., 2009). After evaluating correlations at a single CG resolution, selected kernel bandwidth parameters were: 180 bp for weighted–density and 80 bp for CG–density (the same parameters were subsequently used for all samples and all chromosomes).

$$\text{Weighted}_h(x) = \frac{1}{nh}\sum_{i=1}^{n}\frac{C_i}{\sum_{j=1}^{n} C_j}K\left(\frac{x-x_i}{h}\right) \tag{5.1}$$

$$\text{CG}_h(x) = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right) \tag{5.2}$$

$$u\text{-density}_{h_1,h_2}(x) = \frac{\text{Weighted}_{h_1}(x)}{CG_{h_2}(x)} \tag{5.3}$$

After calculating the weighted–density, more representative signal distribution was noted across the samples. Observed coverage statistics (e.g.,

[1] There are only ∼249 M nucleotides in human chromosome 1.

[2] Projection from 2^{21} uniformly distributed points back to specific CG sites along the chromosome.

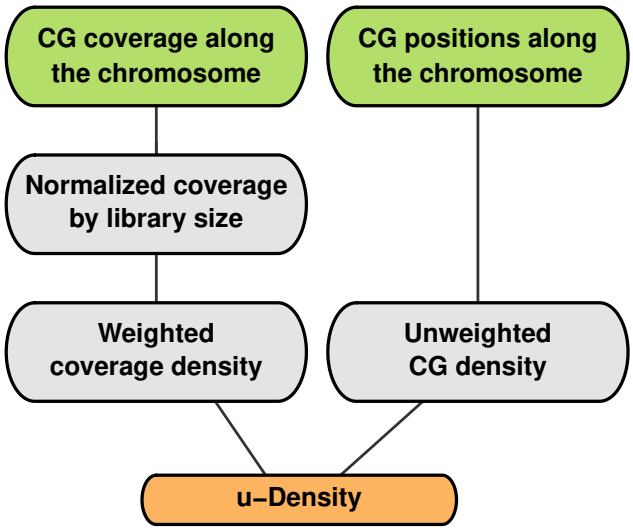


Figure 5.2 | *u*–density Computation Workflow

Workflow of the current *u*–density computation algorithm. Genome–wide CG–coverage is normalized by the total amount of reads within each sequencing library to remove global bias in sequencing depth. This normalized signal is then used to calculated weighted coverage–density. Weighted coverage–density is then normalized by unweighted CG–density to minimize for local coverage biases caused by uneven CG site distribution along the given chromosome.

mean and median coverage) per sample correlated well with the number of sequencing reads assigned to a specific sample. Such effect could easily influence result interpretation as samples with larger library sizes would have higher unmodification (or in case of hmTOP–seq — modification) signal. However, if less or more sequencing reads were used in another experiment, these two results would not be comparable. Observed coverage statistics correlated well (Pearson's $r = 0.85$, Spearman's $\rho = 0.8$, p–value 2×10^{-4}) with number of reads per sample, however this correlation decreased when weighted–density values were used instead (Pearson's $r = 0.46$, Spearman's $\rho = 0.55$, p–value 1×10^{-1}) (**Figure 5.4**).

After calculating the *u*–density, an increase in correlation between technical replicates was observed (Pearson's $r = 0.5$ and $r = 0.8$ before and after transformation for human derived TOP–seq low library depth samples). Similar effect was observed for higher depth IMR90 libraries where Pearson's correlation increased from $r = 0.62$ to $r = 0.87$. Finally,

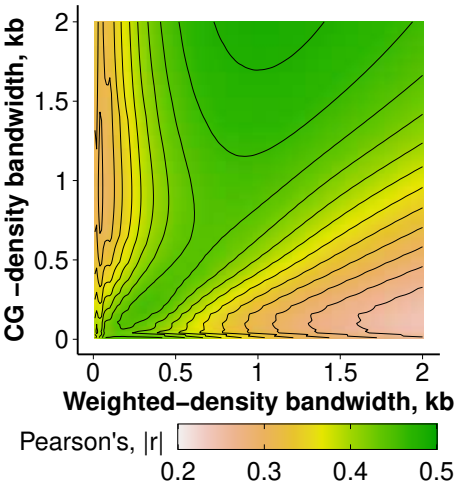


Figure 5.3 | u–density Bandwidth Optimisation

Heatmap represents similarity statistics (i.e., Pearson's correlation coefficient) between computed TOP–seq u–density signal and reference WGBS signal. To calculate u–density signal combinations of different kernel bandwidth sizes were used to estimate weighted and unweigted density. Combination of 180 bp weighted–density and 80 bp unweighted–density bandwidths was selected as an optimal one since it resulted in a relatively high similarity between the TOP–seq method and WGBS, and represented relatively high resolution.

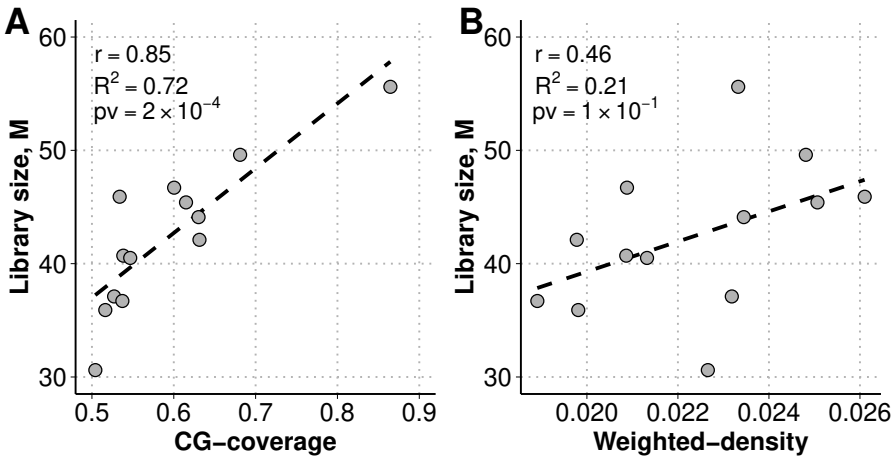


Figure 5.4 | TOP–seq Signal Dependence on the Library Size

(A) Distribution between the total library sizes (millions of reads) in human derived TOP–seq low library depth samples and average CG–coverage within given samples. Both measurements showed relatively high Pearson's correlation r and significant linear relationship R^2. **(B)** Distribution between library sizes and average weighted–density signals. After normalisation dependence between the library size and DNA modification signal decreased.

there was a great improvement in signal distribution across different genomic elements. Since the TOP–seq method is biased towards CG sites, a stronger signal would be expected in CG–rich regions only solely because the number of targets is higher there. To prove this assumption, we evaluated relationship between the two measurements in different genomic elements (**Figure 5.5**). Higher weighted–density signal was observed in elements with higher CG–density, however this bias was corrected after normalising TOP–seq signal by CG amount per each region. To further confirm CG–density normalisation effect a specific genomic loci with a high CG–density variability was interrogated (**Figure 5.6**). TOP–seq signal along the *KAZN* gene locus shows high signal bias towards CG–rich regions (i.e., CGI elements), however after CG–density normalisation signal peaks are not longer centered at CGI elements and u–density signal gradually decreases towards the end of the gene.

5.2.3 Concordance Between u–density and Other Methods

Cross–platform single nucleotide resolution correlations between TOP–seq and WGBS signal confirmed signal transformation usability. Pearson's correlation between CG–coverage and WGBS was $\mid r \mid = 0.23$, $\mid r \mid = 0.36$, $\mid r \mid = 0.44$ for the Brain, IMR90 low library depth and IMR90 high library depth samples, respectively. When using u–density this correlation increased to $\mid r \mid = 0.28$, $\mid r \mid = 0.59$, $\mid r \mid = 0.64$, accordingly. For comparison, single CG resolution IMR90 WGBS signal was compared to MRE–seq and MBD–seq DNA modification signals and computed Pearson's correlations were $\mid r \mid = 0.18$ and $\mid r \mid = 0.3$, respectively. Such low correlations between WGBS and the enrichment–based methods may in part be derived from the non–linear relationship between the data produced with different methods (Stevens et al., 2013).

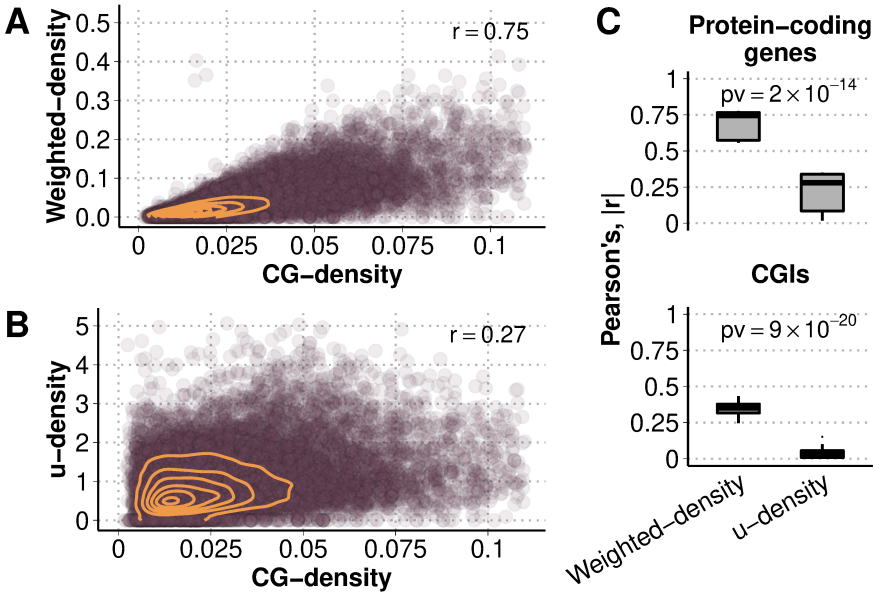


Figure 5.5 | *u*–density Dependency on the CG–density

(A) Scatter plot visualising relationship between the average weighted–density signal and CG–density signal within 19 thousand protein–coding genes in IMR90 L161 sample. Computed Pearson's correlation indicate relatively high association between the CG–density within a gene and observed TOP–seq signal. **(B)** Relationship between the average *u*–density signal and CG–density signal in IMR90 L161 sample within protein–coding genes. Normalisation by CG–density decreased TOP–seq signal dependency on genomic sequence context. **(C)** Pearson's correlations between TOP–seq signal (weighted–density or *u*–density) and CG–density within protein–coding genes or CGIs in human derived TOP–seq samples. *u*–density signal shows significantly lower correlation between TOP–seq signal and CG–density signal within given elements indicating that *u*–density signal strength is not biased towards specific sequence context anymore (*p*–values were calculated using Student's t–Test).

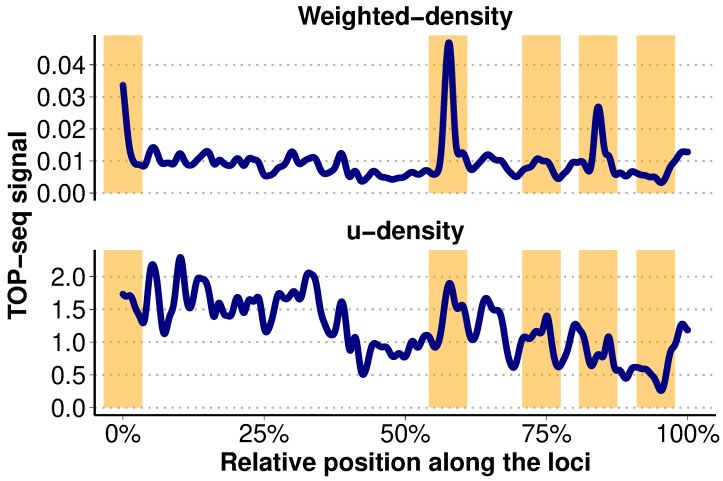


Figure 5.6 | CG–density Normalisation Along the *KAZN* Gene Locus

Weighted–density and *u*–density signals in IMR90 L161 sample along the *KAZN* gene locus shows decreased signal strength in CG–rich regions after CG–density normalisation. CGI elements are indicated with orange segments and contain strongest TOP–seq signal in weighted–density sample, especially in promoter (leftmost) and first intragenic CGI. However, after the CG–density normalisation TOP–seq signal is balanced out with stronger signal towards the beginning of the gene and gradual decrease towards the 3' end.

5.3 m–estimate

5.3.1 Motivation for Calculating m–estimate Signal

After calculating u–density values, we decided to make further signal improvements. Since TOP–seq and u–density represent enrichment–based signal, their values are distributed along the heavy–tailed Poisson distribution (i.e., coverage or u–density values are from 0 to plus infinity) with most sites receiving no or just very low positive values, while the WGBS method is able to present the same modification signal in an absolute scale (from 0% to 100%). Therefore, one of the adjustments was signal conversion from a relative to an absolute scale.

Another improvement was related to analysing epigenomes with different average modification levels. **Figure 5.7** represents two simulated epigenomes with different average modifications levels that are dependent on the position along the simulated chromosome. After performing simulation of the random coverage distribution, we observed that the epigenome with higher average hypomethylation level receives lower coverage than the epigenome with lower hypomethylation modification level. These results show that epigenomes with different modification levels are not comparable when analysed using enrichment–based methods since the observed coverage distributions are inaccurate.

5.3.2 Summary of the m–estimate Algorithm

Methylation estimates, m–estimate, were obtained by training an exponential decay model that assumes a linear decrease of WGBS methylation with exponential increase of u–density signal and other genomic feature–specific covariates **Equation 5.4**. Chromosome 20 (2.5% of all CG sites in the human genome) was used to train an exponential decay model and additional genomic feature–specific covariates were used. Covariate values were calculated for each CG site using 50 bp regions around each CG.

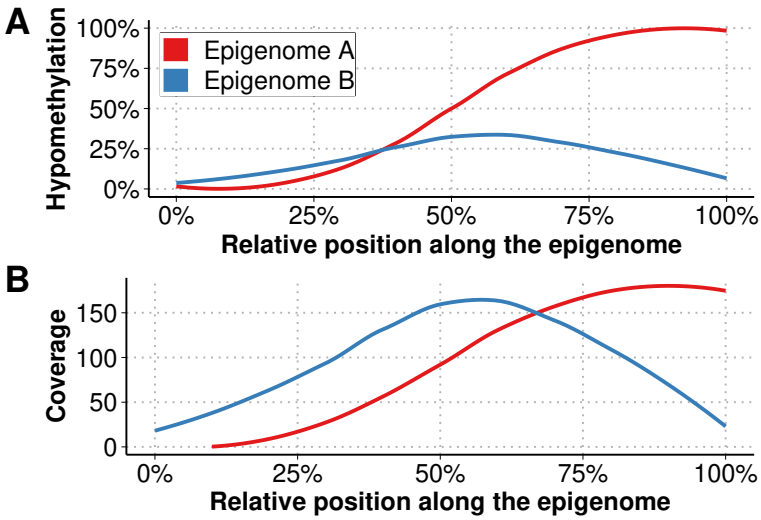


Figure 5.7 | Coverage Dependence on the DNA Modification Level

(A) Two simulated epigenomes with different average DNA modification levels. Epigenome A has 2.5 times higher hypomethylation level than Epigenome B with a tendency for higher methylation regions to cluster towards the start of the simulated genome. Meanwhile hypomethylation level in Epigenome B is normally distributed with global maximum at the middle of the genome. **(B)** Simulated TOP–seq coverage (with identical library sizes — one thousand reads) along two epigenomes. Given same library size Epigenome B shows higher coverage across most of the CG sites while containing lower hypomethylation level. Meanwhile Epigenome A shows higher coverage only in a smaller fraction of CG sites with highest hypomethylation levels.

Implemented covariates were as follows: i) GC frequency – percentage of guanine and cytosine bases per region; ii) fraction of CG dinucleotides among CN pairs within given region; iii) average sequence mappability value per region; iv) fraction of given region that intersects with SINE or LTR repeats; v) fraction of given region that intersects upstream regions (2 kb) of protein–coding genes; vi) fraction of given region that intersects 5'UTR of protein–coding genes; vii) fraction of given region that intersects intergenic regions.

$$b \backsim \exp^{b_0 + b_1 * u\text{-density} + \sum_{i=2}^{k} b_i * \text{covariate}_i} \tag{5.4}$$

5.3.3 Concordance Between m–estimate and Other Methods

Calculation of m–estimate values had a minor effect on correlation among the TOP–seq technical replicates. Pearson's correlation increased only to $r = 0.89$ for the high–depth IMR90 samples (from u–density $r = 0.87$). However, m–estimate calculation greatly improved single CG correlation with WGBS signal. Computed Pearson's r values increased to 0.69.

5.4 nn–estimate

Results presented in this section contain yet unpublished data. Nonetheless they were already presented in two conferences listed in ***Section 1.6***

5.4.1 Motivation for Calculating nn–estimate Signal

After calculating m–estimate, we solved major issues that enrichment–based methods face, however, we still saw an opportunity to increase concordance, such as correlations, between the TOP–seq and WGBS signals. Since WGBS is considered *a gold standard method*, we attempted to approximate the TOP–seq produced signal as close as possible to it. We attempted to compute nn–estimate – methylation estimates calculated using neural networks. We expected neural network based method to perform better than an exponential decay model by learning the representations of the data that we were not aware of. A disadvantage of such approach is that neural network is *a black box* phenomenon and relationships, associations of the data it creates are usually left unknown.

5.4.2 Summary of the *nn*–estimate Algorithm

A multi–layer perceptron regressor with 2 hidden layers (44 and 22 nodes respectively) was used to predict IMR90 WGBS values employing TOP–seq signal and various genomic features from chromosome 20 (**Supp. Figure 4**). Most important features in the perceptron were *u*–density and TOP–seq coverage (relative importance 6.4% and 4.4% respectively). All other used features were split into three groups — genomic elements, sequence characteristics, and base composition around the specific CG site (**Figure 5.8**). Genomic elements with the highest relative importance were CGI and, surprisingly, SINE repeats. Unexpectedly, the importance of SINE repeats was relatively high compared to other major repeat families (i.e., LTR and LINE) as their importance was one of the lowest of all the used features. Most important features from the sequence characteristics group were the amount of GC dinucleotides in a given region, amount of CG dinucleotides, and sequence mappability score. Most of the features from the base composition group showed only mediocre relative importance values.

5.4.3 Concordance Between *nn*–estimate and Other Methods

Calculation of *nn*–estimate values had a minor effect on correlation among the TOP–seq technical replicates. Pearson's correlation for the high–depth IMR90 replicates increased to $r = 0.89$. However, *nn*–estimate calculation improved single CG correlation with WGBS signal – $r = 0.71$ for the combined high–depth IMR90 dataset. Similarity between the WGBS and *nn*–estimate was even greater in higher scale genomic regions. **Figure 5.9** represents correlation between the reference IMR90 WGBS signal and TOP–seq DNA modification signal in protein–coding gene promoters. Correlation between the TOP–seq signal gradually increases with each signal transformation. Similarity remains high even using other IMR90 WGBS dataset. Finally, we divided CGI elements into DNA methylation groups according to a reference IMR90

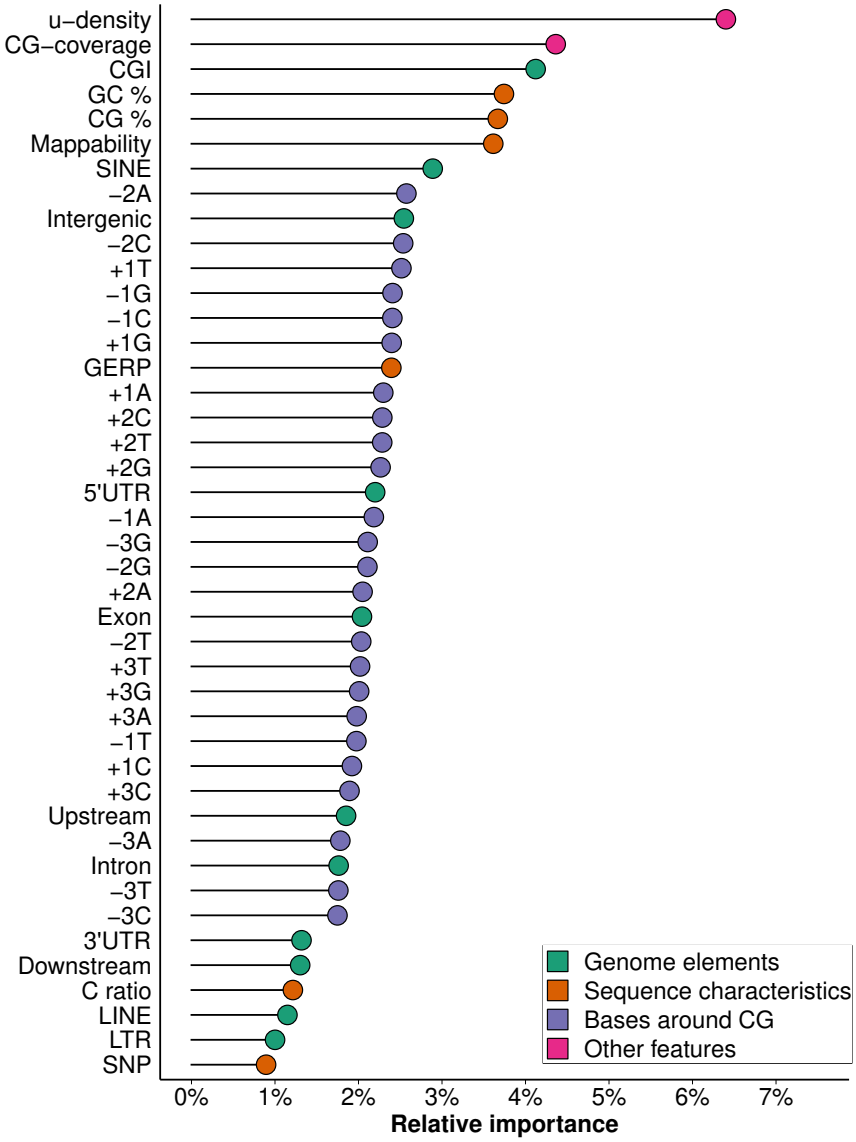


Figure 5.8 | Importance of Genomic Features in *nn*–estimate Model

Relative importance of selected genomic features in *nn*–estimate model measured in percent scale. Genomic features were calculated for each CG site in a 50 bp region. CGI, Exon, Intron, Upstream, Downstream, genomic repeats, Intergenic, 5'UTR, 3'UTR represent fraction of region covered with a specific element. Mappability, GERP, C ratio, GC%, CG% represent average sequence score for a given measurement. Absence or presence of a specific nucleotide at a given distance from a CG site was measured in a binary scale.

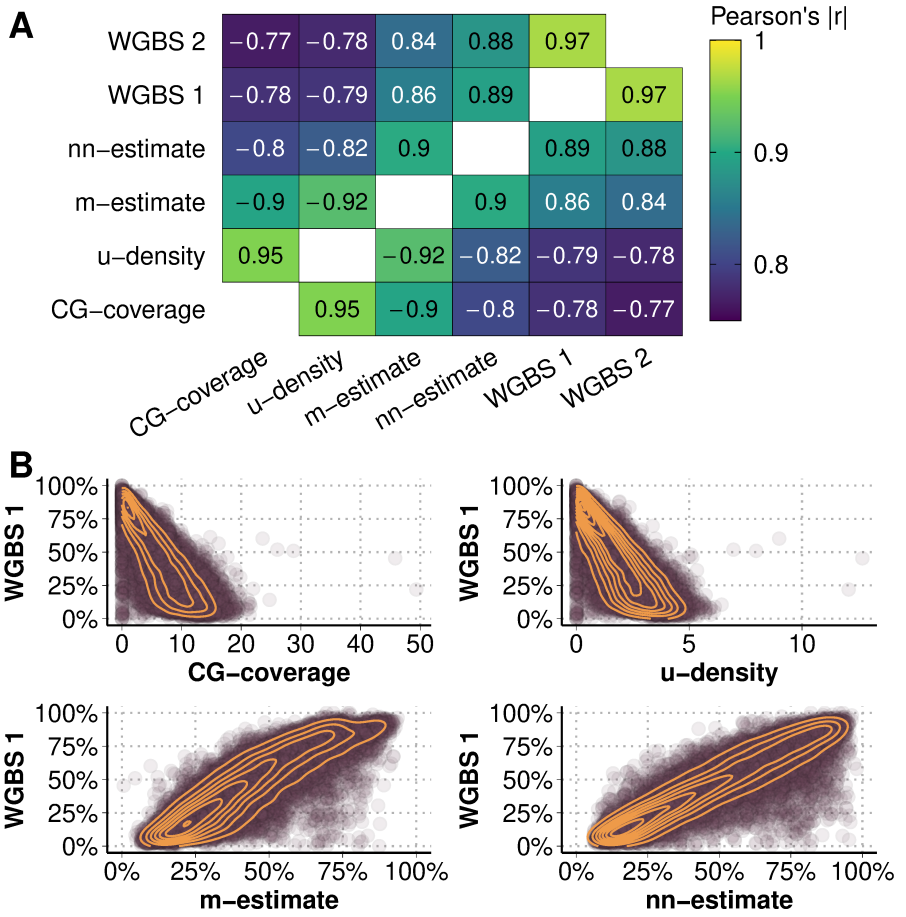


Figure 5.9 | *nn*–**estimate and WGBS Concordance in Promoters**

(A) Heatmap of Pearson's correlation between the IMR90 WGBS signal or TOP–seq DNA modification signals in 18 thousands protein–coding gene promoters. WGBS 1 was used to train *m*–estimate and *nn*–estimate models while WGBS 2 was another IMR90 dataset. Numbers on the heatmap present original Pearson's *r* values. **(B)** Scatter plots between the reference WGBS 1 and four stages of TOP–seq DNA modification signal evaluations in protein–coding gene promoters.

WGBS signal. After evaluating *nn*–estimate signal in given CGI groups we observed high agreement between the WGBS and TOP–seq methods. Next we computed *nn*–estimate values from simulated libraries with ten times lower or higher library sizes. Interestingly, we observed that trained *nn*–estimate model performs well on a lower size libraries, however in higher size libraries computed *nn*–estimate values cannot differentiate between the given CGI groups anymore (**Figure 5.10**).

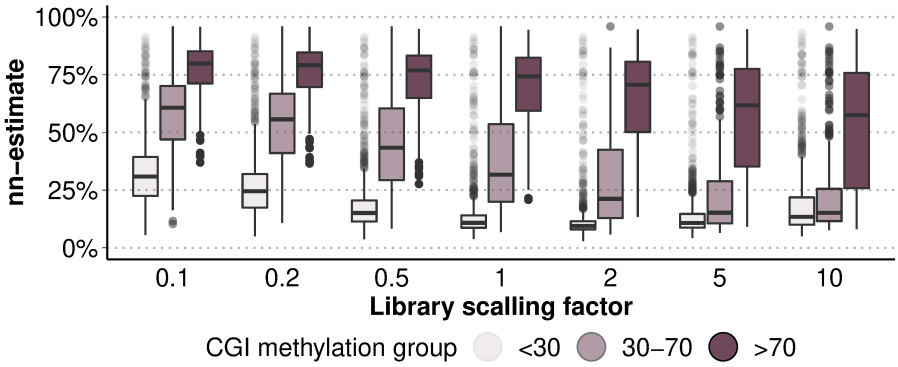


Figure 5.10 | *nn*–**estimate in CGIs**

Predicted *nn*–estimate values in CGI regions divided by their methylation intensity. *nn*–estimate signal is able to distinct different groups of CGIs when original amount of reads is used (i.e., library scaling factor is equal to one). When library size is reduced up to ten times *nn*–estimate predictions are still able to separate CGI groups. When higher library size is used *nn*–estimate model is not able to distinct CGI groups anymore.

5.5 Discussion

5.5.1 The Implications and Applications of This Methodology

In this chapter, we presented three efficient signal transformation methods used to increase the quality of TOP–seq coverage. The low sequencing coverage of TOP–seq method might make DNA modification profiling strategy challenging. Due to non–covered CG sites, sequencing experiment data might be sparse and CG sites will generate false zero counts, thereby creating a false signal. All three developed transformations — *u*–density, *m*–estimate, and *nn*–estimate attempt to solve this issue by incorporating genomic and epigenomic information from the neighbouring loci. These techniques expand the usability of the TOP–seq signal by: i) adjusting for variability in the sequencing library size; ii) compensating for lower sequencing depth; iii) normalising signal intensity by genomic background. Each section on specific transformation contains reasoning behind its computation, algorithm to transform signal,

and comparison with a reference signal. All presented transformations achieve better similarity evaluation with reference data than the original TOP–signal. This is extremely promising as with larger and more variate set of reference datasets performance and applicability of our models can increase even more.

Additionally to the designed transformations, a set of genomic features that are informative for epigenomic signal normalisation were presented. These genomic features consist of genomic elements, such as SINE repeats, sequence characteristics, and base compositions. We believe that this set could be easily applicable for other research.

5.5.2 The Difficulties in Developing Statistical Tools to Enhance the Quality of the TOP–seq Signal

A major difficulty in computing the u–density signal is maintaining the balance between single CG site resolution and using coverage information from the neighbouring CG sites. In practice, u–density computation can be divided into two *jumps* between resolutions: i) projection from a single CG site into a vector of CG sites to include their coverage into estimate computation, and ii) projection from an estimate vector back to a single CG site resolution. The most critical decision therefore is the selection of bandwidth sizes. On one side, these values might be too large and the loss of precision in measurements will occur, while using values that are too small will not deliver optimal results and the mentioned *jumps* between resolutions will only cause loss of quality in the original signal. While it was demonstrated that close to optimal selection of bandwidths values is possible, available improvements for future research are listed in the next section.

A major drawback of m–estimate and nn–estimate is the dependence on a dataset used for supervised learning. When such a dataset is not available, or its quality is not satisfactory enough, these signal transformations will not be possible. However, an appreciable amount of epigenomic

data is available in this age of genomic research and many human tissues or even cell–types are investigated using WGBS technologies. If no exact tissue type is available, then it might be possible to assemble and utilise an additional reference dataset from similar tissues, nonetheless such extrapolation should be carefully tested before its application.

Finally, the biggest issue that a researcher might face is a decision which transformation to use. With a sequencing depth high enough simple CG site coverage might be satisfactory. In such a case no *jumping* between resolutions or signal estimations will be necessary. Given a low sequencing depth one might desire to use transformations but the decision whether to use u–density or supervised techniques needs to be made on a case–by–case basis given the available resources, budget, computational infrastructure, sample availability etc.

5.5.3 Unanswered Questions and Future Research Directions

Since u–density transformation is sensitive to used parameters even wider testing of their combinations is needed. First, it is possible to test optimal bandwidth window sizes for each chromosome separately. For the current study, the kernel bandwidth window size was optimised on human chromosome 1, however each chromosome might contain different optimal weighted–density and unweighted–density bandwidths. The bandwidth window size might also be DNA modification dependent and the optimal window size for u–density might be different than for h–density. Finally, different window sizes might be optimal for different organisms and even for different epigenomes, thus an extensive study is needed to define optimal combinations of all possible u–density parameters.

Another possible optimisation for u–density transformation is computation time reduction by limiting the number of points on which density is calculated. In the present study, a brute force approach was implemented

to compute density using the Epanechnikov kernel over 2^{21} points uniformly distributed across each chromosome. This number was selected to be large enough to cover chromosome 1, however it might be significantly reduced. For example, the size difference between chromosome 1 and chromosome 21 [3] is five times and the currently used value might be unnecessarily large.

Further improvements for the *m*–estimate and *nn*–estimate transformations are also possible. For example, some of the used predictors might be unnecessary. Both *u*–density and CG site coverage were used as covariates in our model, however usage of only one of them might be sufficient. Next, we can try adding other features to further improve our models, such as implementing library size in computation and in such a way to directly normalise for the number of reads. Moreover, there are many genomic elements and features that were not tested in this study, including CGI type or their evolutionary status, various gene types, replication timing information, and many others. Current trends in genomic research tend to focus on artificial intelligence, new algorithms, and improvements, testing types are being developed constantly in this area. Thus, current concepts and algorithms of this work should be viewed as the basis of an ongoing project that needs to be further developed and tested.

5.5.4 Concluding Remarks

This chapter proposed to leverage TOP–seq data and genomic context information to estimate underlying DNA modification levels. Three transformations — *u*–density, *m*–estimate, and *nn*–estimate were designed. *u*–density is based on weighted kernel estimate of CG site coverage normalised by unweighted CG site density, *m*–estimate and *nn*–estimate are supervised learning based techniques created using either exponential decay model or a multi–layered neural network predictor

[3] Smallest autosome in the human genome.

using features as TOP–seq signal, genomic sequence and genomic context information. Our results indicate that it is necessary to combine advanced computation methods with novel sequencing technologies for cost–effective population–wide studies of DNA modification.

Statement II — Developed statistical learning techniques to enhance the quality of the TOP–seq epigenomic signal. For a model IMR90 genome the applied statistical learning approaches increased Pearson's correlation estimate between technical replicates up to $r = 0.89$, while absolute Pearson's correlation estimate at a single CG site with a reference WGBS signal increased up to $r = 0.71$.

Application of TOP–seq Based Methods

Chaos is merely order waiting to be deciphered

José Saramago

6.1 Introduction

This chapter consists of four major sections — applications of TOP–seq based high–throughput epigenome profiling methods. The first section presents the results of the unmodified DNA profiling in human tissues and cell–types, with the second and third sections briefly summarising quality control of hmTOP–seq and caCLEAR methods in mouse embryonic stem cells (mESCs). Finally, the last section present a detailed application of TOP–seq and hmTOP–seq methods in deciphering epigenomic profiles in cell–free DNA from a pregnant female. In these application sections, how previously designed computational and statistical methods can be implemented in specific cases of epigenomic research is discussed.

6.2 Application of the TOP–seq Method in Human Derived Cell–Lines

6.2.1 Introduction

To understand DNA modification dynamics, sensitive high–resolution methods are required for genome–wide mapping epigenetic residuals. This presents an application of the TOP–seq method for the sensitive detection of genome–wide unmodified cytosines. The TOP–seq method identifies genomic uCG positions.

We present genome–wide maps of unmodified cytosines in various human tissues and cell–types, first proving that the TOP–seq signal is reproducible and agrees well with other DNA modification profiling techniques. Next, DNA modification enrichment maps in genomic elements and DNA modification signal across genes or epigenetic elements, such as lamina associated domains or chromatin segments are provided. Finally, we demonstrate that TOP–seq signal is sensitive enough to identify epigenetic differences between cell–types, hence, can be applied for a genome–wide DNA modification profiling. CGI elements are used as a platform to detect differential modifications between neuroblastoma (NB) cell–types and the brain tissue.

6.2.2 Materials and Methods

6.2.2.1 Samples Analysed

TOP–seq libraries were prepared using different sources of human DNA: prefrontal brain cortex, fetal lung fibroblasts IMR90, and two clonal neuroblastoma cell–types: N–type LA1–55n, and S–type LA1–5s (both derived from the LA–N–1 cell–line). Most of the analysed samples contained at least two technical replicates to ensure signal variability within and between samples. A fraction of the IMR90 samples were sequenced

with higher than usual sequencing depth to investigate possible bias that could be caused by variable sequencing depth. A summary of all samples used is provided in **Table 6.1**. Sequencing data was processed as described in **Chapter 4** and the signal was further enhanced using statistical techniques provided in **Section 5.2** and **Section 5.3**.

Table 6.1 | Human Samples Analysed Using the TOP–seq Method

"Sample identifier" defines biological replicate, while "Replicate identifier" defines technical replicate. "DNA source" describes human tissue or cell–line from which DNA was purified. "Library depth" specifies depth of sequencing library and "GEO code" encodes sample identifier deposited under GEO accession GSE91023.

Sample identifier	Replicate identifier	DNA source	Library depth	GEO accession code
Brain 1	R1	Prefrontal brain cortex	Low	GSM2419856
Brain 1	R2	Prefrontal brain cortex	Low	GSM2419857
Brain 2	R1	Prefrontal brain cortex	Low	GSM2419858
Brain 2	R2	Prefrontal brain cortex	Low	GSM2419859
IMR90 1	R1	Fetal lung fibroblasts (IMR90)	Low	GSM2419860
IMR90 1	R2	Fetal lung fibroblasts (IMR90)	Low	GSM2419861
IMR90 2	R1	Fetal lung fibroblasts (IMR90)	Low	GSM2419862
IMR90 3	R1	Fetal lung fibroblasts (IMR90)	High	GSM2419863/4/5
IMR90 4	R1	Fetal lung fibroblasts (IMR90)	High	GSM2419866/7/8
LA1–55n	R1	Neuroblastoma (LA1–55n)	Low	GSM2419869
LA1–55n	R2	Neuroblastoma (LA1–55n)	Low	GSM2419870
LA1–55n	R3	Neuroblastoma (LA1–55n)	Low	GSM2419871
LA1–5s	R1	Neuroblastoma (LA1–5s)	Low	GSM2419872
LA1–5s	R2	Neuroblastoma (LA1–5s)	Low	GSM2419873
LA1–5s	R3	Neuroblastoma (LA1–5s)	Low	GSM2419874

6.2.2.2 DMR Calculation in Neuroblastoma Samples

To evaluate modification differences in the clonal neuroblastoma cell types promoter, intragenic, and intergenic CGIs were used. For each CGI element, the average TOP–seq u–density value per each sample was computed (if CGI had an average u–density value less than 1×10^{-4}, it was omitted from the analysis). Next, all samples were passed to `limma` tool (linear models for microarray data) for multigroup analysis and DNA modification contrasts were interrogated for N–type versus S–type, N–type versus Brain, and S–type versus Brain samples for each CGI group separately (Ritchie et al., 2015). Regions having a false discovery rate (FDR) adjusted q–value less than 1×10^{-2} and absolute fold

change greater than 20% were termed statistically significant. Each significant promoter or intragenic CGI was associated with a protein–coding gene and gene enrichment analysis was performed using the DAVID annotation tool for each CGI group separately (Huang et al., 2008). Summarised flowchart of DMR identification algorithm can be found in **Figure 6.1**.

6.2.2.3 DNA Modification Analysis Across Genomic Elements

DNA modification profiles along lamina associated domains and inter–LAD regions were computed as follows: first, from the set of all LAD elements, those that intersected gaps in the genome assembly were removed. Remaining LADs were further filtered according to their size, retaining only those that were between 0.1 and 0.9 quantiles of all the LADs. Inter–LAD domains were filtered using the same procedure. Each resulting region was divided into 10 equally sized non–overlapping bins before removal of CG sites that intersected CGIs or their 5 kb flanking regions. Due to possible $MYCN$ amplification, we also excluded chromosome 2 from analysis. For each CG site, a corresponding bin of LAD or inter–LAD region was assigned. The final profile was obtained by averaging the signal in each bin using a Gaussian kernel smoother with bandwidth 2 (in `R 3.5` this function is implemented under `stats::ksmooth`).

DNA modification profiles along gene bodies were constructed as follows: for each protein–coding gene, we selected its longest processed transcript and used it as a reference gene. This step was needed as different isoforms might have different gene–start or gene–end sites that might distort observed signal in upstream or downstream regions. Additionally, genes were removed that were shorter than 1 kb. Upstream regions were defined as 2 kb flanks from the gene start site. When computing the generic gene modification profile each specific upstream, 5'UTR, exon, intron, 3'UTR region was divided into 20 equally sized non–overlapping bins and CG modification signals were averaged within corresponding bins.

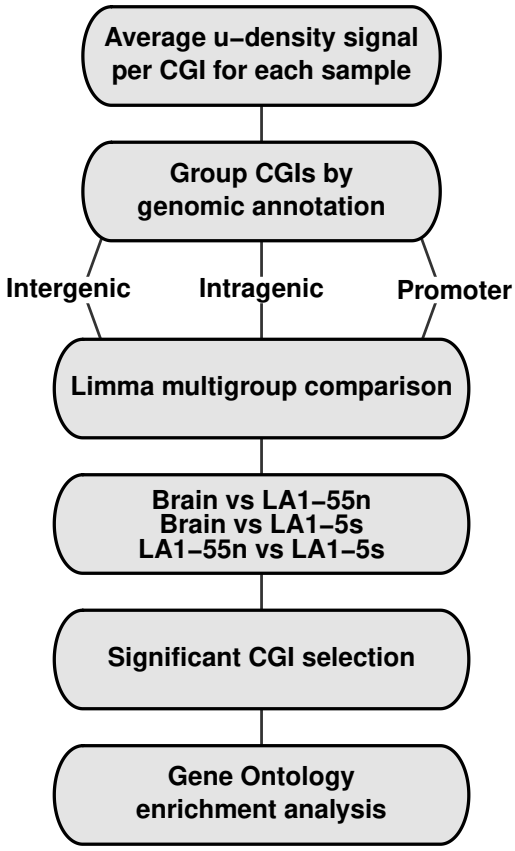


Figure 6.1 | DMR Identification in Neuroblastoma Samples Workflow

The average *u*–density signal per each CGI element was computed and all CGIs were grouped according to their position to protein–coding genes. Computed values were used in `limma` tool for multigroup analysis. DNA modification contrasts were compared for Brain versus N–type (LA1–55n), Brain versus S–type (LA1-5s), and N–type versus S–type for the intergenic, intragenic or promoter CGIs separately. Finally, statistically significant CGIs (FDR q–value less than 0.01 and absolute fold–change difference above 20%) were associated with protein–coding genes that were used in consequent gene ontology analysis.

6.2.3 Quality Control of Processed TOP–seq Sequencing Data

On average, each sample contained 42 millions raw sequencing reads ($sd = 6$), except for high–depth IMR90 samples that on average contained 238 millions raw reads ($sd = 4$). After processing and mapping reads to a reference genome, the number of reads decreased significantly (**Figure 6.2**). A large fraction of this decrease in high–depth IMR90 samples was caused by short read removal or PCR duplicate read removal. After removing duplicate reads and assigning remaining reads to CG sites, on average 16 million reads remained in low–depth ($sd = 3$) and 91 million reads in high–depth libraries ($sd = 3$).

In all analysed samples, all unmodified CG sites (coverage greater than 0) were used, resulting in on average 21% identified genomic CG sites in low–depth and 35% in high–depth samples (**Table 6.2**). The average coverage of identified CG sites per low–depth sample was 2.8 and 9.6 per high–depth sample, however, after more detailed inspection, it was discovered that the coverage was uneven between the chromosomes with neuroblastoma samples showing much higher average coverage in chromosome 2 (**Figure 6.3**). After closer examination, it was found that this higher coverage was caused by reads originating from a specific 1.6 Mb region on chromosome 2 (chr2:15026730 — 16640120) (**Figure 6.4**). This region contains proto–ongonenic *MYCN* gene which is expected to amplify exactly in the neuroblastoma samples (Spengler et al., 1997).

After selecting all identified CG sites, we measured the correlation between technical replicates (average Pearson's $r = 0.69$ and Spearman's $\rho = 0.49$) (**Figure 6.5**). Such medium correlation could be caused by the shallow sequencing depth that was used and indeed, after measuring the Jaccard coefficient between the identified CG sites, low overlap between technical replicates was observed (average Jaccard's coefficient 0.4) (**Figure 6.5 A**). The Jaccard coefficient shows that given CG sets between technical replicates are different, however this result could be attributed to the previously mentioned shallow sequencing depth. To prove

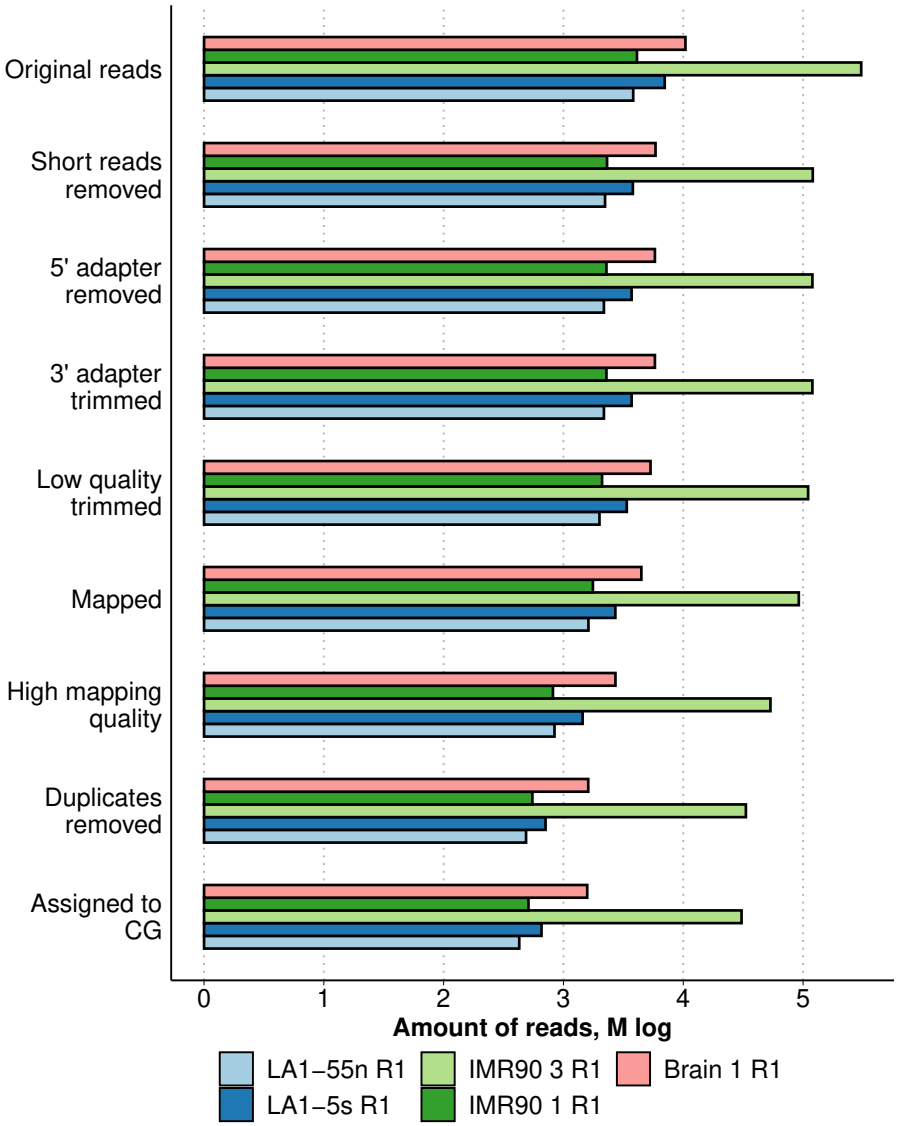


Figure 6.2 | Amount of TOP–seq Reads After Each Processing Step

The absolute change in read numbers after each processing step in a set of human derived samples. High library depth IMR90 samples have higher amount of reads compared to the other samples, however amount of reads in these samples decreases at a similar rate throughout the pipeline.

Table 6.2 | Coverage Statistics of uCG Sites

The amount of identified CG sites (coverage greater than 0) is represented in absolute and relative numbers. Coverage of identified CG sites is represented with arithmetic mean (i.e., average).

Sample identifier	Replicate identifier	Amount of uCG,%	Average coverage
Brain 1	R1	26	3.3
Brain 1	R2	20	2.5
Brain 2	R1	20	3.4
Brain 2	R2	20	3.2
IMR90 1	R1	19	2.8
IMR90 1	R2	20	2.7
IMR90 2	R1	21	3
IMR90 3	R1	34	9.1
IMR90 4	R1	32	10
LA1-55n	R1	20	2.5
LA1-55n	R2	21	2.5
LA1-55n	R3	23	2.7
LA1-5s	R1	17	3.4
LA1-5s	R2	17	3.1
LA1-5s	R3	17	3.1

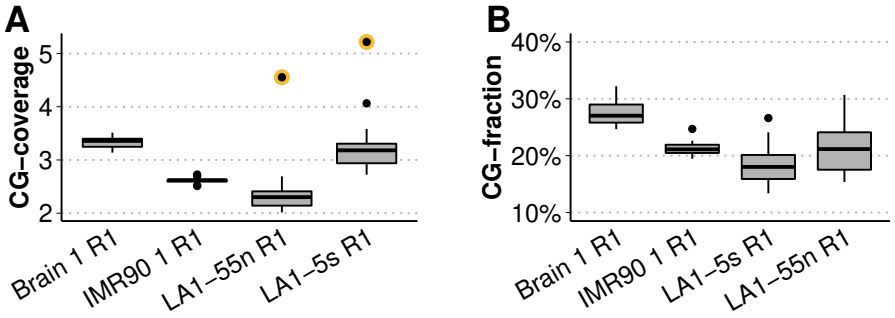


Figure 6.3 | Unmodified DNA Signal in Human Samples

(A) The average coverage of identified CG sites per each chromosome in selected low–depth samples. Neuroblastoma derived samples show unusually high coverage in chromosome 2 (marked with a yellow circle), possibly due to the amplification in the *MYCN* gene locus. **(B)** The amount of identified CG sites per each chromosome. All samples show normal distribution of identified CG sites without any outlier chromosomes.

this assumption, we simulated very low sequencing depth datasets and monotonically increased their library size, demonstrating a clear linear dependency between the library size and Jaccard's coefficient (**Figure 6.6**). To further confirm the reproducibility of the TOP–seq method we calculated correlation between the technical replicates using various sizes of

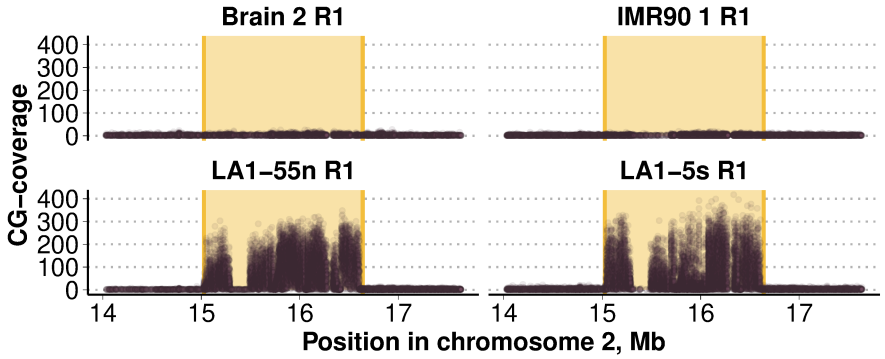


Figure 6.4 | Coverage Along the *MYCN* Gene Locus

Genomic region of *MYCN* gene locus from chromosome 2. Four panels represent coverage in samples from Brain, IMR90, and neuroblastoma sample groups and *MYCN* gene is represented with a colored area. Neuroblastoma derived samples show much higher CG–coverage at exactly the proto–ongonetic gene region.

genomic regions (**Figure 6.5 B**). Technical replicates of neuroblastoma samples reached Pearson's $r = 0.9$ with 1 kb region sizes while other sample groups needed much larger region sizes to increase their reproducibility. Additionally, we calculated Fisher's exact test between the identified CG sites between all samples to show that CG identification is not a random process. All tested overlaps between the sets were significant (all Fisher's exact test p–values $< 2.2 \times 10^{-16}$) and on average, the Fisher's estimate was 9.7 for samples pertaining same sample groups (**Figure 6.5 C**). It is expected that for enrichment–based methods, such as TOP–seq, a greater library size might not result in higher overlap between samples, due to the saturation effect — number of identified CG sites grows only until a specific threshold. This was confirmed using high–depth IMR90 samples. The number of identified CG sites monotonically increased with the total number of reads (closely resembling a logarithmic function), suggesting that at increasing sequencing depths, the unmodified CG calling progressively expands from low– to moderate, maybe even to high–modification CG sites (**Figure 6.7**). However, after reaching 60 million reads, the curve for identified CG sites plateaus and increases only minimally.

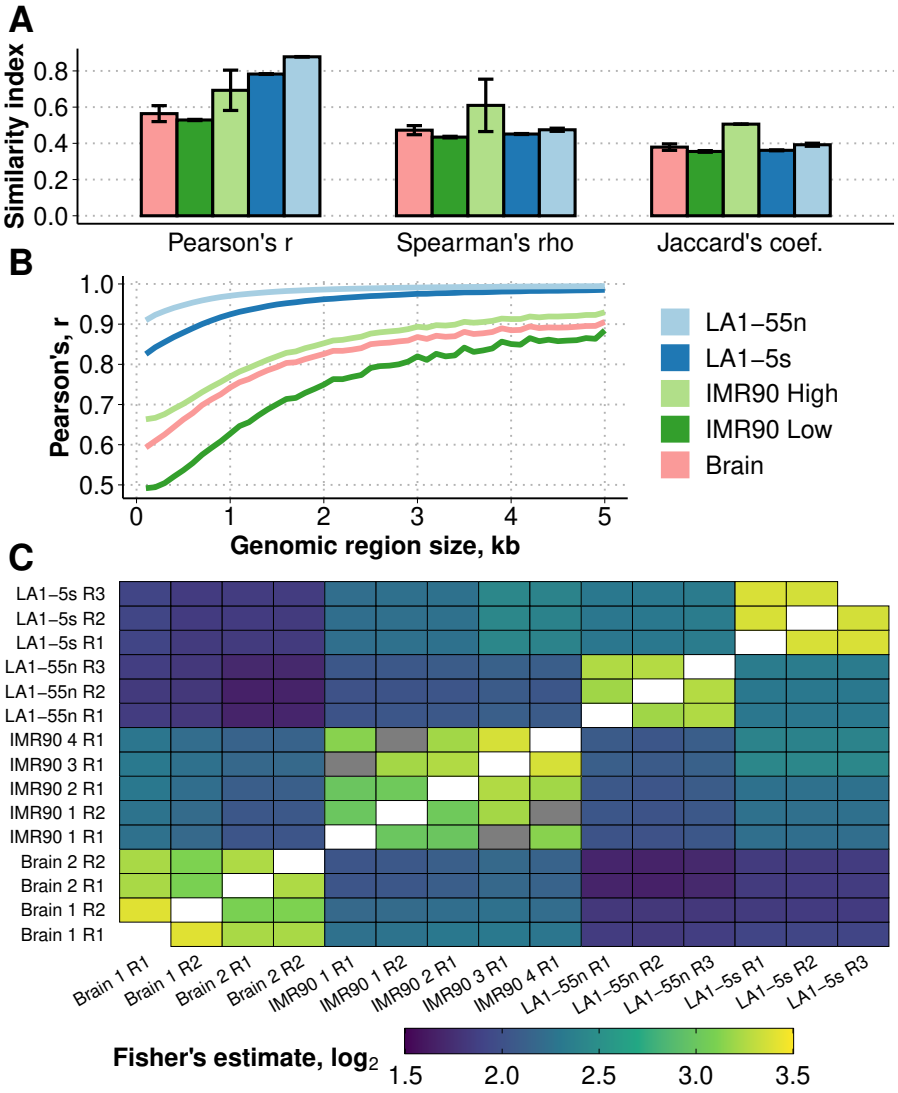


Figure 6.5 | Similarity Statistics of Samples Analysed Using the TOP–seq Method

(A) Similarity index measured in Pearson's correlation, Spearman's correlation or Jaccard's coefficient between replicates in given sample groups at a single CG resolution. On average low–depth libraries showed Pearson's r around 0.53, except for neuroblastoma derived samples that had higher Pearson's r probably to the *MYCN* gene locus amplification. This was confirmed with Spearman's correlation as all low–depth libraries showed similar Spearman's ρ. Jaccard's coefficient between shared CG sites was low–medium to medium for all used sample groups. **(B)** Correlation coefficient between replicates at varying sizes of genomic bins. **(C)** Fisher's estimates between all identified CG sites.

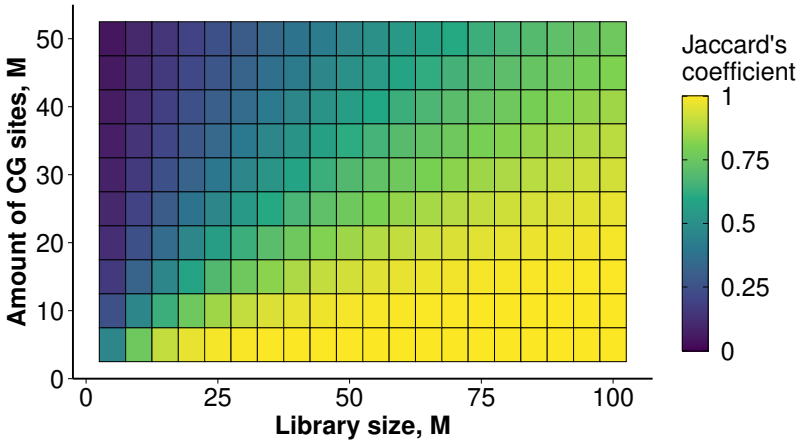


Figure 6.6 | Jaccard's Coefficient Between Identified CG Sites in Simulated Datasets

Jaccard's similarity coefficient between identified CG sites in simulated replicates computed using varying genome and library sizes. Higher library size (i.e., amount of reads) results in higher Jaccard's coefficient which is proportional to available CG sites in the genome.

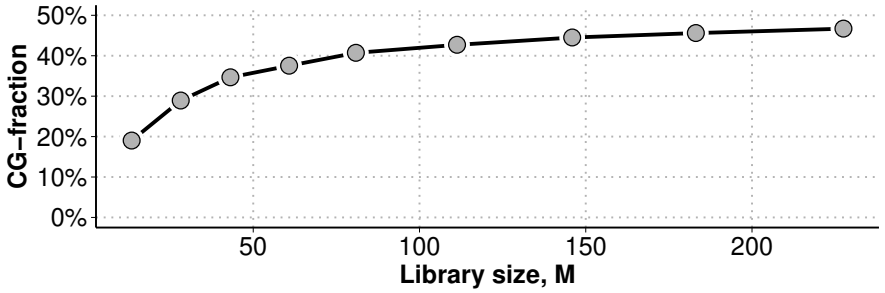


Figure 6.7 | Identified CG Amount Dependence On the Library Size

The number of identified CG sites monotonically increases with the total number of reads in the IMR90 library. After reaching ∼50 million reads this relationship plateaus and the amount of newly identified CG sites increases only marginally.

Next, we performed a computational adjustment of the TOP–seq coverage data to generate a high–resolution genome–wide prediction of DNA modification levels. Using kernel density estimation, weighted density estimates from the TOP–seq coverage signal were computed and normalised by the unweighted CG–density estimates to obtain the TOP–seq unmethylome density (u–density) signal. This adjustment enhanced the Pearson correlation of the low–depth replicates to $r = 0.8$ (Spearman's

$\rho = 0.77$). Correlation of the high–depth IMR90 replicates increased to $r = 0.9$ (Spearman's $\rho = 0.85$) (**Figure 6.8**).

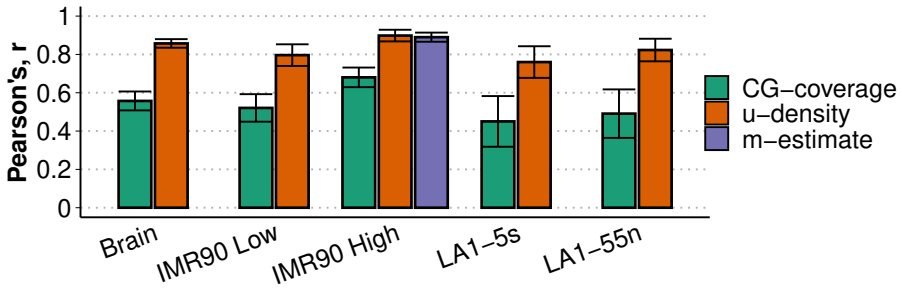


Figure 6.8 | Concordance Between TOP–seq Replicates Using Various Signal Modifications

Single CG site correlation coefficients between replicates using TOP–seq coverage, u–density, and m–estimate signals. u–density transformation improves TOP–seq signal reproducibility on average by 55%.

Cross–platform single CG absolute correlation of the low–depth and high–depth TOP–seq u–density datasets with IMR90 WGBS data was $\mid r \mid = 0.59$ and $\mid r \mid = 0.64$, respectively (**Figure 6.9**). In a further adjustment step, we sought to account for possible sequence–specific variations that may influence the TOP–seq signal. We used a small fraction of the WGBS dataset (chromosome 20) to train an exponential decay model containing additional genomic feature–specific covariates which was then used to convert the TOP–seq u–density into so–called CG methylation estimates (m–estimate, methylation values presented in the absolute scale from 0 — 100). Although the second enhancement step had a minor effect on the correlation among the TOP–seq technical replicates, it improved single CG site correlation with the IMR90 WGBS to $r = 0.69$ (Spearman's $\rho = 0.65$) high–depth set (**Figure 6.9**).

After observing the increased correlation between TOP–seq and WGBS methods within genomic bins, we compared the agreement between these methods within various genomic elements. Dissection of the whole–genome profiles across major genomic features showed good agreement of the TOP–seq signal and WGBS in CGIs, enhancers, 3'UTRs, exons, introns, upstream, and downstream regions of protein–coding genes (**Figure 6.10**). Interestingly, TOP–seq signal correlation was low with

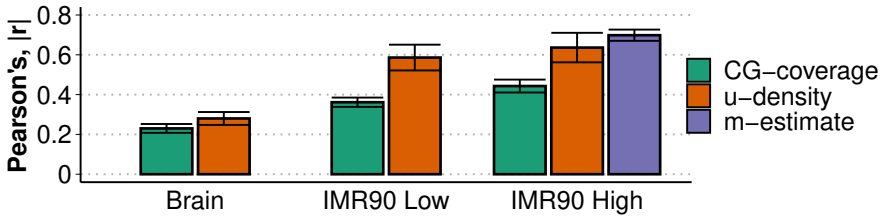


Figure 6.9 | Concordance Between the TOP–seq and WGBS Methods at Single CG resolution

Single CG site absolute correlation coefficients between given TOP–seq signals and whole–genome bisulfite signal. u–density transformation increased TOP–seq signal similarity with WGBS on average by 42% and m–estimate transformation by 10%.

WGBS signal in TSS chromatin segments, while MBD showed medium correlation in these elements. However, such low correlation should not be attributed to higher CG content in these elements as TOP–seq reported relatively high correlation in CGIs and therefore there should be another unknown effect that results in such low agreement between the methods.

Conversion of the u–density signal to the m–estimate signal was also successful (led to improved correlation with WGBS, $r = 0.69$, $\rho = 0.63$) using another independently produced IMR90 WGBS dataset (Ziller et al., 2013). However, a similar conversion of the Brain u–density data based on the published brain WGBS map did not lead to satisfactory m–estimate maps, which was understandable given the poor correlation of the u–density and the WGBS dataset (Pearson's $\mid r \mid = 0.28$, $\mid \rho \mid = 0.24$). Altogether, the presented examples suggest that this optional adjustment step is only feasible when a high–quality reference WGBS map derived from a related tissue is available. Accordingly, the TOP–seq u–density profiles were used in all further comparative tissue analyses due to lack of a suitable Brain and neuroblastoma WGBS maps.

As the ultimate validation of the predictive power of the method, we evaluated how the bottom and top 10% of unmethylated CGIs, genes and 10 kb sized regions in IMR90 cells, as well as in Brain, identified by

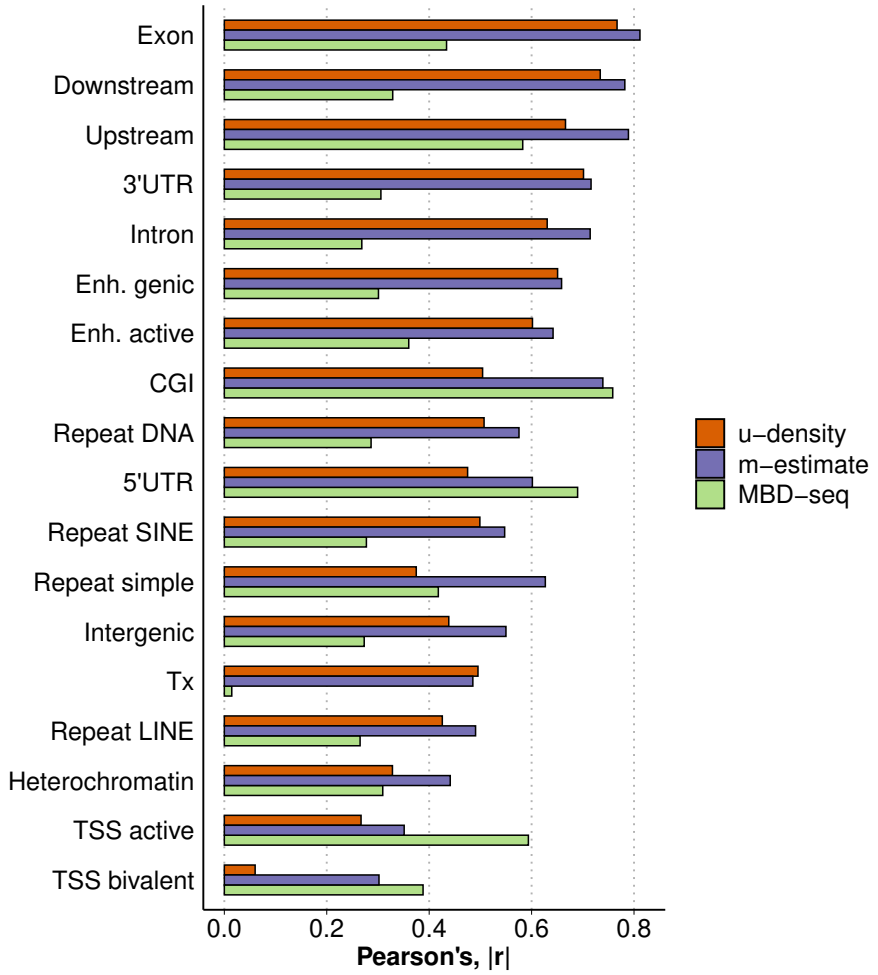


Figure 6.10 | Concordance Between the TOP–seq and WGBS Methods Within Various Genomic Elements

Absolute correlations between average IMR90 WGBS and TOP–seq or MBD–seq signal in gene–coding, Epigenome Roadmap chromatin states or repeat elements. Both TOP–seq signal transformations correlate relatively well with a WGBS signal in gene related elements and CGIs, but show decreased similarity in TSS associated elements. The Epigenome Roadmap chromatin states are defined in **Chapter 3**.

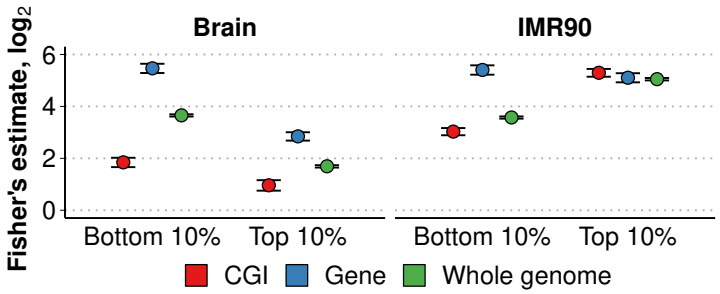


Figure 6.11 | Intersection Between the Top Modified and Unmodified Regions

Fisher's estimate for overlaps between the 10% of lowest and 10% of highest modified regions as computed with TOP–seq and WGBS methods. Lowest and highest modification retaining regions were selected from CGI, protein–coding gene set or genomic bins divided into 10 kb regions (represented as a whole genome). All reported Fisher's exact test estimates have p–values less than 2×10^{-19}.

the TOP–seq method, overlap with the bottom and top 10% of unmethylated regions derived by WGBS. In IMR90, we observed a very strong association between the TOP–seq u–density and WGBS in all top 10% used elements, as well as bottom 10% genes (Fisher's exact test odds ratio (OR) $\sim$ 37) (**Figure 6.11**). For the Brain samples only the bottom 10% genes showed such high and, interestingly, very similar enrichment to IMR90 samples. Also, Brain samples showed relatively low enrichment for the top 10% of CGIs while CGIs had the highest enrichment in top 10% of elements in IMR90 samples.

6.2.4 Epigenomic Maps

After selecting identified CG sites their enrichment and distribution throughout various genomic elements was tested (**Figure 6.12**). In most of the tested genomic elements enrichment or depletion of selected CG sites was similar between different sample groups. Biggest enrichment was observed in elements associated with the beginning of genes (5'UTRs, CGIs, promoters of various gene biotypes). Interestingly only

CG sites from the Brain sample showed enrichment in other protein–coding gene related elements (i.e., exons, 3'UTRs and introns). Highest depletion of identified CG sites was observed in pseudogenes and SINE repeats with all sample groups showing similar tendencies. The TOP–seq signal was detectable in 96% of 26,641 autosomal CGIs. As expected, promoter CGIs were the most enriched in uCG sites (50% — 100% CG sites identified in $\sim$ 85% of CGIs), indicating their highly unmodified state (**Figure 6.13**). The variation in identified uCG sites was higher among intragenic and intergenic CGIs, attesting their diversity and, on average, higher methylation levels. Interestingly, a relatively high proportion of intergenic CGIs showed either absolute modification or only very light modification forming bimodal distributions in Brain and IMR90 sample groups, but not in neuroblastoma derived samples. These findings showed that the TOP–seq data are generally consistent with established genome methylation patterns.

We also compared the TOP–seq *u*–density profiles with WGBS across different gene–associated elements (**Figure 6.14**). As expected, the TOP–seq and WGBS profiles of the corresponding tissues showed inverse patterns throughout the analysed regions. We further determined the TOP–seq *u*–density in and around segments representing a range of chromatin states (Kundaje et al., 2015). Among the active promoter states, active TSS, bivalent/poised TSS promoters, and flanking TSS upstream segments showed higher TOP–seq *u*–density signals, indicating their lower methylation levels (**Figure 6.15**).

To assess the power of TOP–seq to discern large–scale DNA modification patterns, we investigated LAD elements. It has been noted previously that LAD elements correspond to partly modified DNA regions and are directly involved in gene repression and usually range from 80 kb to 30 Mb in size (Berman et al., 2011; Guelen et al., 2008).

Since dynamic association with the nuclear lamina has been implicated as a key mechanism in the developmental regulation of long–range gene silencing that can be perturbed in cancer cells, we sought to investigate the DNA modification status in LAD elements (**Figure 6.16**). The

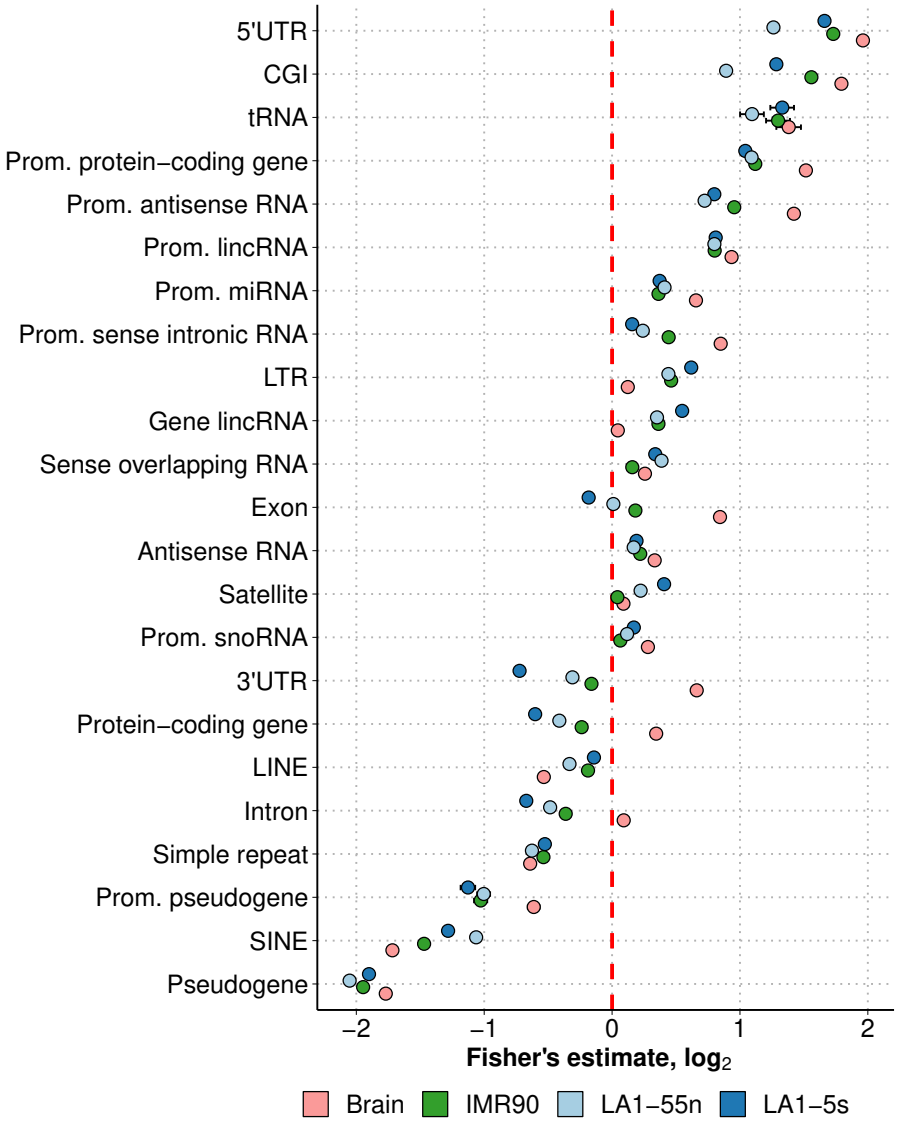


Figure 6.12 | TOP–seq Signal Across Genomic Elements

Fisher's estimate of enrichment or depletion of identified CG sites in various genomic elements (p–values for all reported Fisher's estimates are less than 0.05). Highest TOP–seq CG enrichment was observed in gene start associated elements — 5'UTR, CGIs, promoters of various gene biotypes.

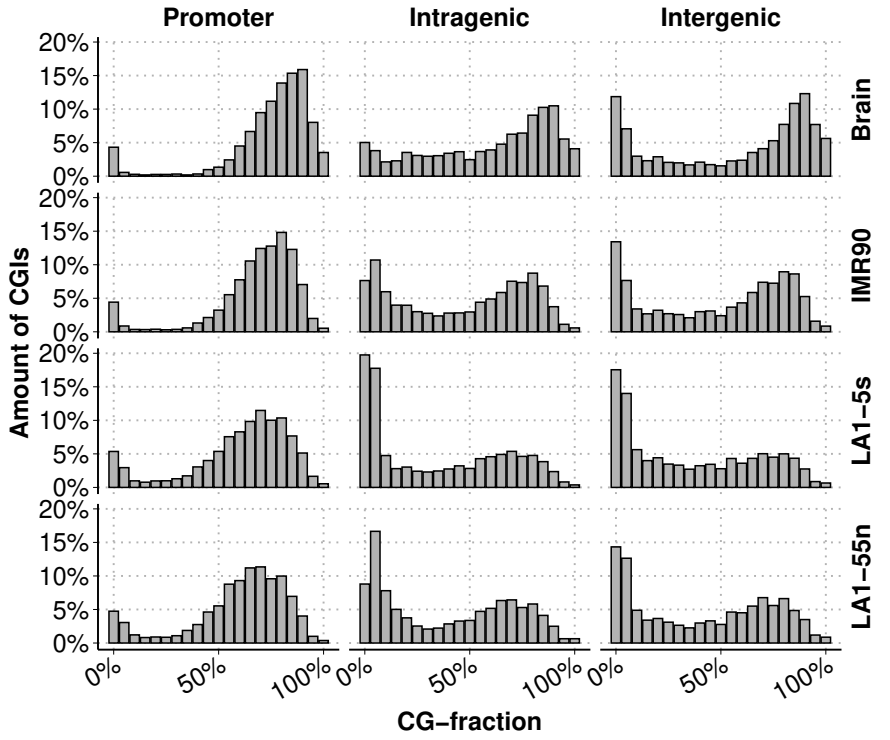


Figure 6.13 | Identified CG Sites in CGIs

DNA modification level represented as a percentage of identified CG sites in CGIs using the TOP–seq method. CGIs are divided into three groups according to their position relatively to protein–coding genes. X–axis represents amount of CG sites identified in a particular CGI and Y–axis represents amount of CGIs with a specific amount of identified CG sites. Highest CG–fraction was observed in promoter CGIs, while intragenic and intergenic CGIs tend to be more modified in neuroblastoma derived cell–lines.

analysis showed strong hypomethylation of the LAD regions compared to inter–LAD regions in the IMR90 cells, while no comparable changes in the TOP–seq u–density were detected in the brain sample. The observed methylation differences were mirrored by the WGBS data further confirming that LADs are absent in the cells of the adult brain cortex. Similar analysis of NB cells revealed the presence of LADs in both these tissues.

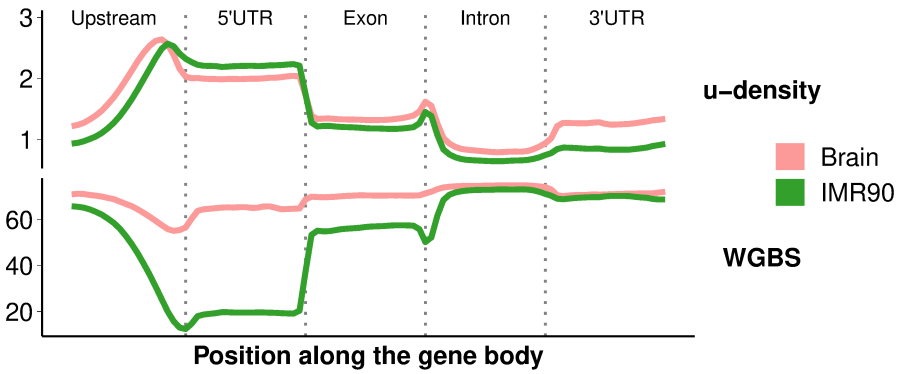


Figure 6.14 | TOP–seq Signal Along the Protein–Coding Gene Body

u–density and WGBS signals along the generalised protein–coding gene body. TOP–seq and WGBS signals mirror each other in high compatibility across all gene parts, with both methods showing relatively lower DNA modification in upstream and 5'UTR regions. X–axis represents relative position along the gene body or particular element, while Y–axis represents specified signal intensity.

6.2.5 Differentially Modified Regions in Neuroblastoma Samples

Given that neuroblastoma is a neuroendocrine tumour arising from neural crest cells, we focused our analysis on promoter and intragenic CGI–DMRs identified between N, S, and the Brain samples and assigned them to their host genes (**Table 6.3**). We performed functional annotation analysis of the genes with the identified CGI–DMRs first focusing on promoter CGI characterisation. Gene ontology functional enrichment analysis for the sets of S/B–hypoM and N/B–hypoM indicated a significant enrichment for components of the intracellular organelle lumen and cytoskeleton. HyperM promoter CGIs for both N and S cells were significantly enriched in groups of homeobox domain containing proteins, glycoproteins, signal peptides, and biological processes covering neuron differentiation, development, and axonogenesis. This is in line with the nature of this developmental tumour, which is associated with the impaired maturation of the neuronal phenotype. Intriguingly, analysis of the N/B–hyperM CGIs identified hypermethylation of genes involved in

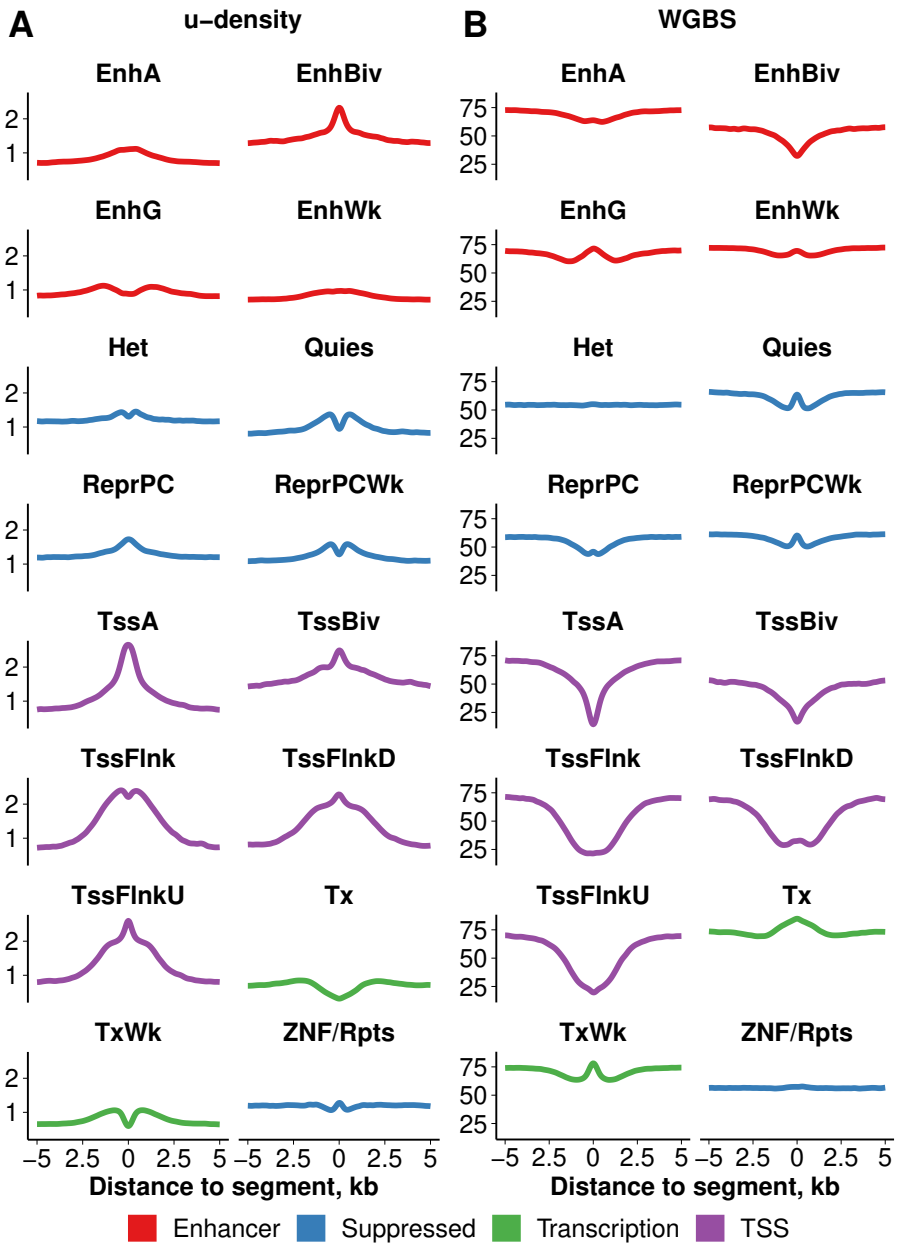


Figure 6.15 | TOP–seq Signal Along the Epigenome Roadmap Chromatin Segments

u–density **(A)** and WGBS **(B)** signals in IMR90 samples across the loci that contain various chromatin segments as identified in the Epigenome Roadmap project. Specified chromatin segment is centralized at the X–axis and 5kb of flanking regions are shown around it. In most of the segments TOP–seq and WGBS methods mirror each other and show similar DNA modification distributions. The Epigenome Roadmap chromatin states are defined in **Chapter 3**.

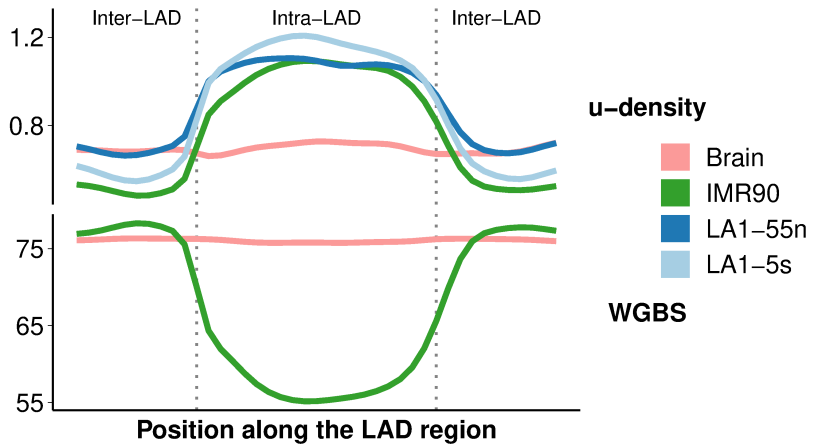


Figure 6.16 | TOP–seq Signal in LAD Elements

TOP–seq and WGBS methods show decreased DNA modification levels across LAD elements in IMR90 samples, while Brain samples show no change in DNA modifications compared to inter–LAD regions. Both neuroblastoma derived cell–lines show similar tendencies to the IMR90 signal.

neural crest development and migration, which are absent in the S/B–hyperM CGI–DMRs.

Table 6.3 | Amount of Neuroblastoma CGI DMRs

In pairwise comparisons between cancerous and normal tissues, four types of DMRs were analyzed: N/B– or S/B–specific hypomethylated regions (hypoM, regions that show higher TOP–seq *u*–density in N or S cells than in Brain) and N/B– or S/B–specific hypermethylated regions (hyperM, higher *u*–density in Brain relative to N or S). The table shows numbers of hypoM and hyperM CGI–DMRs for N–type, S-type and IMR90 cells (relative to Brain reference) and N/S–hypoM and S/N–hypoM CGI regions (N and S cells compared to each other). Number of genes with assigned tissue–specific CGI–DMRs is shown in parentheses.

CGI type	DMR direction	N/B	S/B	N/S	IMR90/B
Promoter	**hyperM**	203 (202)	1414 (1407)	487 (483)	3017 (2988)
Promoter	**hypoM**	4285 (4155)	2541 (2460)	4156 (4073)	538 (523)
Intragenic	**hyperM**	827 (567)	1153 (940)	2752 (1760)	2168 (1796)
Intragenic	**hypoM**	5321 (3509)	4729 (3055)	3104 (2456)	2790 (1751)
Intergenic	**hyperM**	543	593	1042	990
Intergenic	**hypoM**	2208	2037	1469	946

It has been reported that tissue– and cell–type specific methylation is present in a small proportion of CGI promoters, whereas a far greater proportion occurs across gene bodies, which include potential alternative

CGI promoters (Maunakea et al., 2010). Importantly, functional annotation analysis of the intragenic CGI–DMRs of the N and S cells (with respect to Brain and each other) revealed substantial differences between the NB cell types. In contrast to the S/B–hypoM (and S/N–hypoM) comparisons, for N/B–hypoM (and N/S–hypoM) CGI–DMRs, we identified significantly enriched terms related to glycoproteins, extracellular matrix structure, collagens, EGF–like domain proteins which included many growth factors, developmental and receptor proteins. Comparison of the intragenic N/B–hyperM and S/B–hyperM CGI DMRs found a strong overlap in GO terms associated with sequence–specific DNA binding proteins, neuron differentiation/development, and cell adhesion. However, N–type specific hypermethylated CGIs with respect to S (N/S–hyperM) were in large gene clusters involved in neuron differentiation and development, cell–cell signalling, synaptic transmissions, and neurological system process, pointing at potential downregulation of these genes compared to the non–tumorigenic S–type cells.

6.2.6 Discussion

6.2.6.1 New Insights Provided by This Study

In the first part of this section, the quality control results were presented for samples analysed using the TOP–seq method. Even though coverage correlation and Jaccard's coefficient of the identified CG sites were in the range of medium values, Fisher's estimates proved that CG site identification is not a random process. Additionally, Jaccard's coefficient values obtained using a simulated dataset proved that samples with larger library sizes would produce higher similarity. Comparison with a reference WGBS datasets revealed that the TOP–seq signal correlates much better than other analysed methods. Importantly, similarity to the WGBS signal depends on genomic element type suggesting possible DNA sequence bias towards enrichment–based methods.

Further signal transformations proved to be useful as correlations with

a WGBS signal were significantly higher. Higher correlations were observed even for another WGBS dataset that was not used to optimise transformation parameters. Obtained signal transformations showed relatively high agreement with WGBS signal while investigating gene or other genomic elements DNA modification profiles. Finally, we employed the TOP–seq signal and identified a wide range of CGI DMRs pertaining to specific types of neuroblastoma derived cell–types.

6.2.6.2 The Difficulties of Applying the TOP–seq Method in This Study

Since the TOP–seq method, in contrast to bisulfite conversion–based methods, cannot directly determine the absolute methylation levels, a signal transformation might be required. The u–density technique seemed to be a perfect solution for this problem as it takes in account variation in different library sizes and DNA sequence bias. However, the m–estimate transformation worked only for the IMR90 samples since they showed relatively high agreement with a corresponding WGBS signal, while this transformation did not show as good results for Brain samples. This suggests that a good reference dataset is needed for such a supervised learning technique.

The TOP–seq method, as many other read–count–based epigenome profiling approaches, is sensitive to copy number variations. When a very high coverage pattern is observed (e.g., *MYCN* locus), it is possible that such signal enrichment is solely caused by the high abundance of targeted DNA sites. Therefore, *de novo* discovered DMRs should be verified to fall outside genetic aberrations or validated by an independent method.

6.2.6.3 Unanswered Questions and Future Research Directions

One simple adjustment to the TOP–seq analysis that would greatly improve the overall quality of results would be removal of possibly amplified

sites. In this study, the *MYCN* locus was removed due to a possible amplification, however a larger set of regions is needed. It would be of a great use to compile a collection of regions that are usually amplified in cancers or regions that vary in copy numbers across different populations to help adjust for possible signal enrichment caused by more targeted sites.

One important piece of future work could be TOP–seq method applicability to identify another type of genetic variation — single nucleotide polymorphisms. Within a sample with a very deep coverage, it is possible to align sequencing reads and identify polymorphisms compared to a reference sequence. Such analysis would not only provide the modification status of CG sites but also provide information regarding DNA sequence variants that are in linkage with a given CG site. Such application would result in genomic and epigenomic profiles and, taking in account that the TOP–seq method is more economically accessible, makes TOP–seq an exceptional method both for genome–wide and epigenome–wide studies. TOP–seq method projection from the epigenomic to genomic applicability for large–scale population–level analyses would be of great scientific interest as the relationship between DNA sequence variability, DNA modification level and phenotype is not well elucidated.

6.2.6.4 Concluding Remarks

Herein, we described the application of the TOP–seq method in human derived tissues and cell–types. The obtained results suggest that TOP–seq provides a combination of single CG resolution and genome–wide coverage. Epigenomic maps generated using this TOP–seq method provide insights about DNA modification variability across samples pertaining to distinct experimental groups or different types of genomic elements.

Statement III — The TOP–seq method provides information about DNA modification signal across different genomic elements.
Statement IV — The TOP–seq method could be used to identify differentially modified regions across samples pertaining pertaining to distinct

sample groups.

6.3 Application of the hmTOP–seq Method in mESCs

6.3.1 Introduction

5hmC is the most abundant form of oxidative DNA modification. It is involved in multiple biological processes, including embryogenesis, neurological processes, and cancerogenesis. Profiling of this relatively scarce genomic modification requires sensitive, high–resolution techniques. This chapter describes analysis of the new sequencing technique called hmTOP–seq that can be used to identify 5hmC at single base resolution genome–wide. To validate our approach, we used mouse ESC genomic DNA and compared generated hmTOP–seq signal with data obtained from other DNA modification profiling method and found good correlation between 5hmC mapped regions. We also compared 5hmCG distributions in genic and epigenomics features such as histone modifications. Based on this analysis, we conclude that hmTOP–seq could be used as a genome–wide 5hmCG modification profiling technique.

6.3.2 Materials and Methods

6.3.2.1 Samples Analysed

hmTOP–seq libraries were prepared using various amounts of DNA from mESCs: 5, 50, and 500 nanograms (ng). For each specified DNA amount, two technical replicates were generated. Additionally, negative control libraries were prepared using the same hmTOP–seq library preparation pipeline but without the BGT labelling step. A summary of all used samples can be found in **Table 6.4**. All sequencing data were processed as described in **Chapter 4**.

Table 6.4 | mESCs Samples Analysed Using the hmTOP–seq Method

"Sample identifier" defines biological replicate, while "Replicate identifier" defines technical replicate. "DNA input" specifies the amount of DNA used for a given sample. "Library type" describes library preparation type and "GEO code" encodes sample identifier deposited under GEO accession GSE140206.

Sample identifier	Replicate identifier	DNA input, ng	Library type	GEO accession code
hmC ctrl 5	K1	5	hmTOP–seq (−BGT)	GSM4156657
hmC ctrl 5	K2	5	hmTOP–seq (−BGT)	GSM4156658
hmC 5	R1	5	hmTOP–seq	GSM4156659
hmC 5	R2	5	hmTOP–seq	GSM4156660
hmC ctrl 50	K1	50	hmTOP–seq (−BGT)	GSM4156661
hmC ctrl 50	K2	50	hmTOP–seq (−BGT)	GSM4156662
hmC 50	R1	50	hmTOP–seq	GSM4156663
hmC 50	R2	50	hmTOP–seq	GSM4156664
hmC ctrl 500	K1	500	hmTOP–seq (−BGT)	GSM4156665
hmC ctrl 500	K2	500	hmTOP–seq (−BGT)	GSM4156666
hmC 500	R1	500	hmTOP–seq	GSM4156667
hmC 500	R2	500	hmTOP–seq	GSM4156668

6.3.2.2 Identifying 5hmC Modified Cytosines

To analyse the 5hmC modification in the CG context, we used all CG sites that were identified in both technical replicates. To identify the 5hmC modification in the CH context, we used a more sophisticated filtering approach. First, we selected all mapped sequencing reads that started exactly at cytosines in the CH context that contained no CG site closer than seven nucleotides in the downstream direction. Next, we removed all CH sites that were identified in at least one control sample and finally selected only those CH sites that were identified in both technical replicates.

6.3.3 Quality Control of Processed hmTOP–seq Sequencing Data

On average, each non–control sample contained 59 million raw sequencing reads ($sd = 2.2$) that decreased significantly after removing short sequencing reads (**Figure 6.17**). The mapping rate for all samples was

relatively high and after removing reads due to low mapping quality, the read number decreased only marginally.

Average coverage of identified CG sites differed greatly between the different DNA input libraries, as well as, between the samples and their corresponding controls (**Figure 6.18 A**, **Supp. Table 2**). Average CG–coverage in samples varied from 4.5 to 12.8, while in the corresponding control samples CG–coverage was on average two times lower. After further investigation a great difference between samples and controls was found in identified CG–fraction. While identified CG–fraction varied from 5.6% to 34% in controls this fraction on average was only 0.01%.

Technical replicates of the higher–input hmTOP–seq libraries correlated well at a single CG resolution (Pearson's $r = 0.46$ and $r = 0.8$ for 50 ng and 500 ng input libraries, respectively **Figure 6.19 A**). While the 5 ng DNA input libraries showed considerably lower correlation between technical replicates (Pearson's $r = 0.11$). A further improvement in correlations between technical replicates was noticed when larger genomic regions were used. Pearson's correlation in higher input libraries on averaged increased up to $r = 0.92$ using 5 kb resolution, while 5 ng sample this measurement increased only up to $r = 0.55$ (**Figure 6.19 B**). We also found that overlap between the identified CG sites increases with the amount of input DNA. Jaccard's coefficient between in 500 ng technical replicates was 0.08 while in 50 ng samples it was 0.056, and in 5 ng samples only 0.02 (**Figure 6.19 A**). Even though overlap between the identified CG sites was very low it was significant and non–random. We additionally tested overlap between the identified CG sites using Fisher's exact test and discovered that 500 ng samples overlapped well not only between the technical replicates, but also with other samples (**Figure 6.19 C**). Additionally, subsampling of the original higher input datasets was performed (**Figure 6.20**). This subsampling showed that using higher amount of input DNA, but lower amount of reads higher correlations between the technical replicates could still be achieved. After decreasing library sizes 50% we observed only 10% decrease in correlation between the technical replicates. Further subsampling only marginally

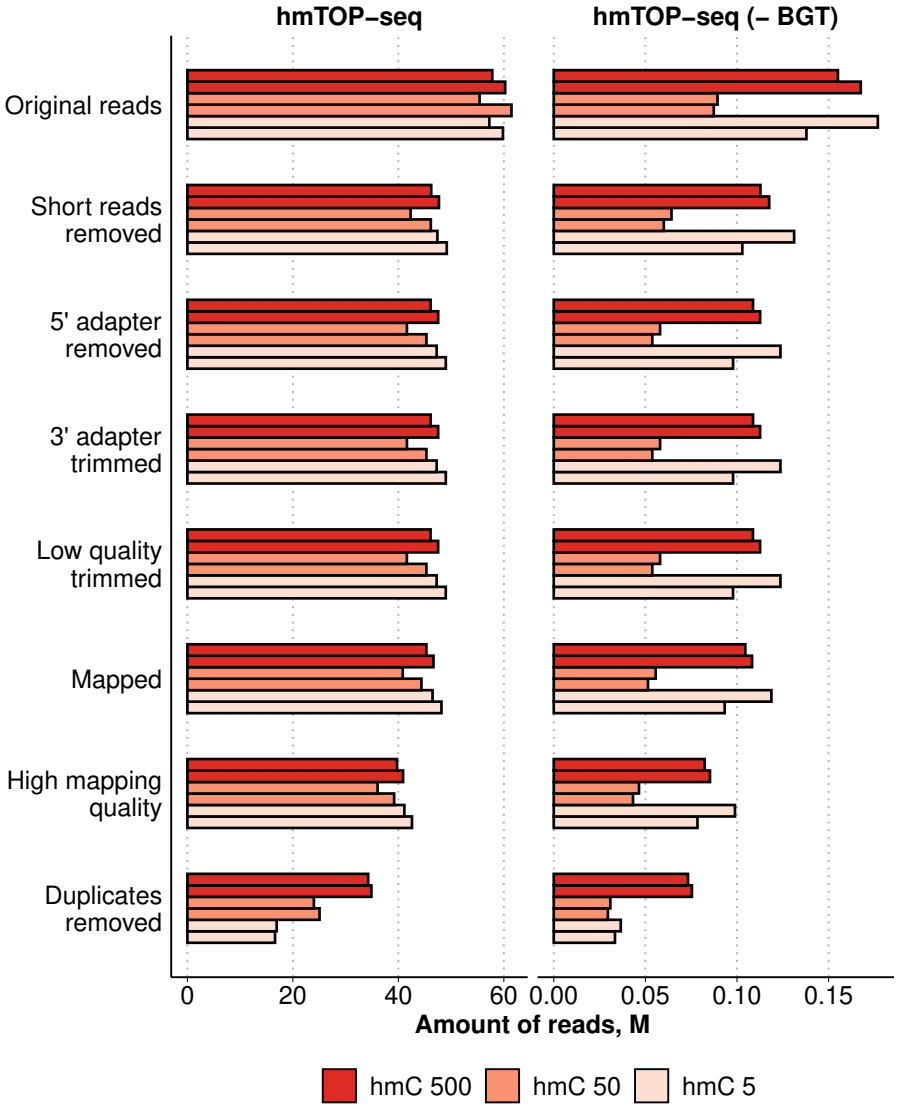


Figure 6.17 | Amount of hmTOP–seq Reads After Each Processing Step

The absolute change in read numbers after each processing step in mESC samples analysed using different amount of input DNA (color scale encodes amount of used DNA in nanograms). Samples and controls have different library sizes, but proportional change in read numbers is similar between the two sample groups.

decreased correlation and significant drop in correlation was observed only reducing libraries down to 1%.

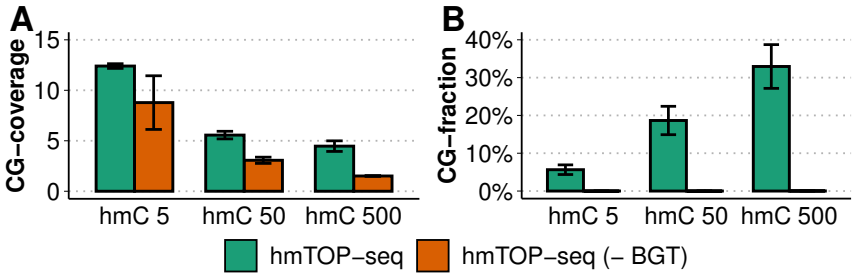


Figure 6.18 | 5hmCG Modified DNA Signal in mESC Samples

(A) The average coverage of identified CG sites per each chromosome. Lower input DNA libraries show higher average coverage than greater input DNA libraries. X–axis represents amount of input used in particular sample. **(B)** The amount of identified CG sites per chromosome. Control samples show close to zero amount of identified CG sites while samples with BGT show higher amount and variability of identified CG sites with greater amount of used DNA.

We then compared our 5hmCG datasets with the bisulfite treatment–based TAB–seq data, showing that hmTOP–seq recovered 50% and 25% of TAB–seq identified 5hmCG sites in the 500 ng and 50 ng input DNA datasets, respectively (Fisher's exact test estimates 4 and 3.8, respectively, p–values $< 2.2 \times 10^{-16}$). Direct comparison between the hmTOP–seq and TAB–seq signal indicated good agreement between the two methods (Pearson's $r = 0.94$, Spearman's $\rho = 0.96$, **Figure 6.21**).

6.3.4 Epigenomic Maps

Analysis of 5hmCG distribution across various genomic elements demonstrated good agreement with the published data. The highly hydroxymethylated CG sites (top 20% of hmTOP–seq data) were enriched in poised enhancers marked by histone H3 lysine 4 monomethylation (H3K4me1) histone marks, active enhancers (marked by histone H3 lysine 27 acetylation (H3K27ac) and histone H3 lysine 4 trimethylation (H3K4me3), exons, 3'UTRs , downstream regions of protein–coding genes, shores of CGIs, and non–active promoters depleted in histone H3 lysine 9 acetylation (H3K9ac) histone mark (**Figure 6.22**). CGIs,

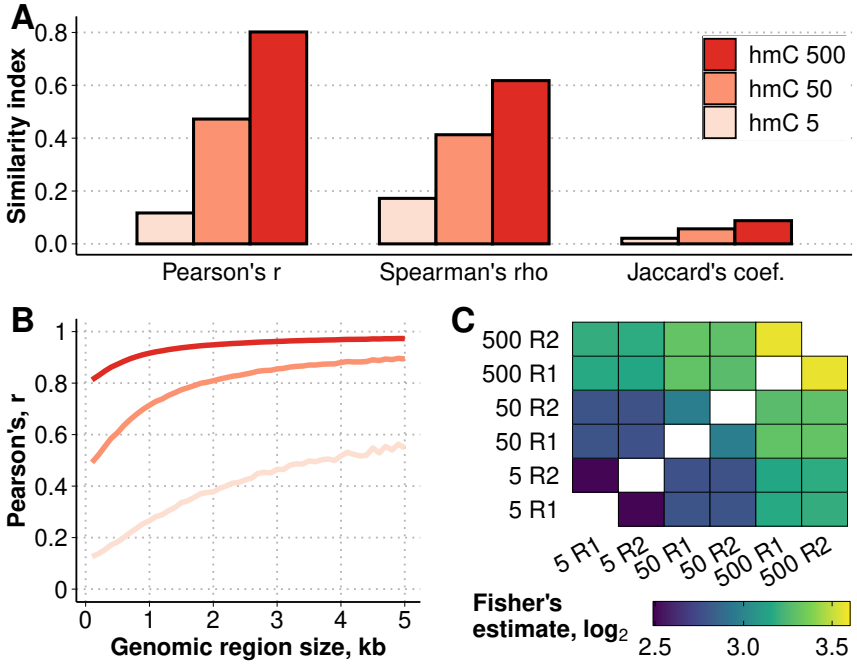


Figure 6.19 | Similarity Statistics of Samples Analysed Using the hmTOP–seq Method

(A) Similarity index measured in Pearson's correlation, Spearman's correlation or Jaccard's coefficient between replicates at a single CG site resolution. **(B)** Correlation coefficient between replicates at varying sizes of genomic bins. **(C)** Fisher's estimates between all identified CG sites.

active promoters marked by H3K9ac, intergenic regions, and all major type of repeats were depleted in 5hmCG sites (data shown only for long terminal repeats), except for SINE repeats that demonstrated moderate enrichment for less hydroxymethylated 5hmCG sites.

Using six hmTOP–seq control libraries, we observed 55,025 non–CG sites of which only 190 overlapped in at least two control libraries, and only 284 of them overlapped with the hydroxymethylated CH sites identified in the 500 ng hmTOP–seq libraries, suggesting that these sites resulted from random priming events rather than from BGT directed covalent labelling and could be defined as false positives. Of all CH sites detected in the 500 ng hmTOP–seq libraries, we selected only those sites which overlapped between technical replicates for further analysis, resulting in a final set of 76,665 5hmCHs (average coverage 2.7, Pearson's r =

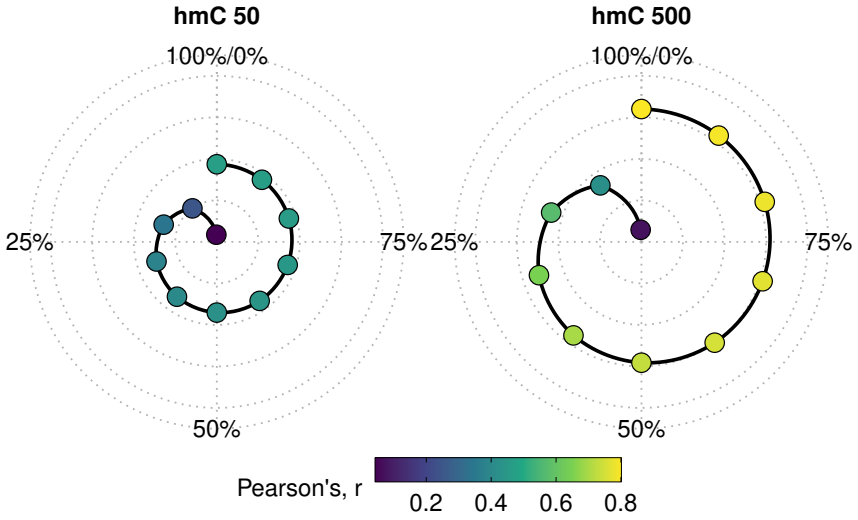


Figure 6.20 | Concordance Between hmTOP–seq Technical Replicates

Correlation between technical replicates using subsamples of the original library sizes. Original library sizes in 500 ng and 50 ng samples were gradually reduced every 10% up to 1%. Pearson's correlation decreases only marginally with up to 50% of the library sizes and reaches 0 with only 1% of the original library size. Percentage displayed in a clockwise manner represents amount of used library size which decreases from 100% to 1%. Color scale and distance towards the centre of the graph represents Pearson's correlation between technical replicates.

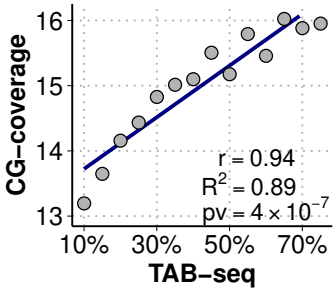


Figure 6.21 | Concordance Between hmTOP–seq and TAB–seq

Comparison of hmTOP–seq average CG–coverage and 5hmC percentages estimated by bisulfite–based TAB–seq. Relationship between both signals showed high Pearson's correlation r and significant linear relationship R^2.

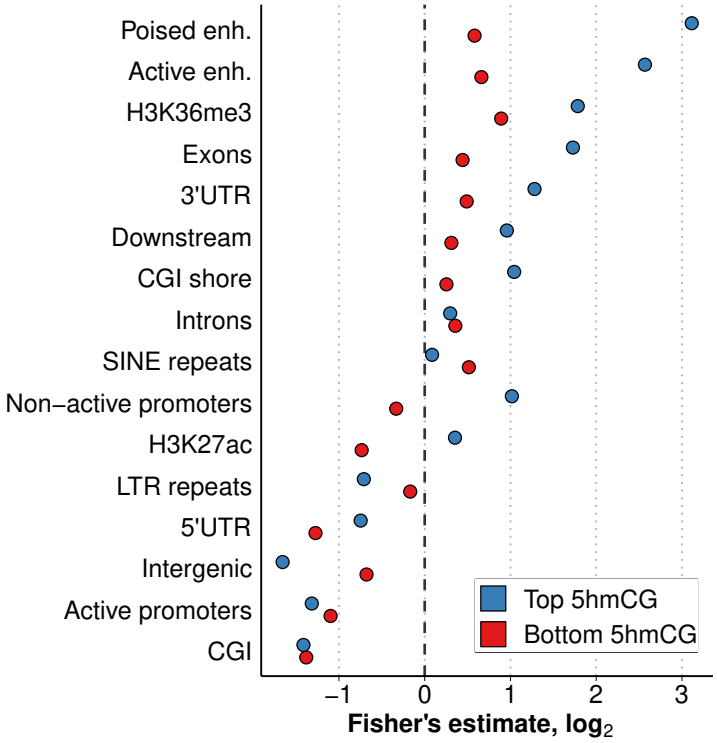


Figure 6.22 | hmTOP–seq Signal Across Genomic Elements

Fisher's exact test estimates for the enrichment or depletion of the high–(top 20%), and low–coverage (bottom 20%) 5hmCG sites across various genomic elements. Poised enhancers ("enh."): regions with H3K4me1 mark only; active enhancers: regions with H3K4me1 and H3K27ac histone marks; active promoters: 2 kb regions upstream of the gene start that overlap H3K9ac histone mark; inactive promoters: 2 kb upstream regions depleted in H3K9ac. All estimates (presented for a 500 ng input hmTOP–seq library) have Fisher's exact test p–values less than 0.05.

0.76). Half of detected 5hmCHs were found in CA sites (CA:CT:CC = 0.50:0.33:0.17) and distributed in a ratio 1:2 in CHG and CHH context.

6.3.5 Discussion

6.3.5.1 New Insights This Study Provided

In this chapter, we provided an application of the hmTOP–seq method to profile 5hmCG modification at a single CG resolution genome–wide.

We show that the computed 5hmCG signal is reproducible and correlates well in higher input DNA libraries. Moreover, agreement in DNA modification signal between the hmTOP–seq and other methods is relatively high. We also proved that the hmTOP–seq method can identify 5hmC modification in the CH context in a reproducible manner. Finally, we presented genome–wide 5hmCG modification enrichment maps across various genomic elements. Computed enrichments agree well with previously reported results, proving that the hmTOP–seq method could be used as an alternative method for currently applied epigenome–wide profiling techniques.

6.3.5.2 Concluding Remarks

Herein, we presented an application of the hmTOP–seq method to profile 5hmCG modification in mouse embryonic stem cells. The obtained results suggest that the hmTOP–seq method could be used as an epigenome–wide single nucleotide profiling technique for 5hmC DNA modification.

Statement V — The hmTOP–seq method provides information about the DNA modification signal across different genomic elements.

6.4 Application of the caCLEAR Method in mESCs

6.4.1 Introduction

This chapter presents caCLEAR, a new method for sequence–specific detection of 5caC, a rare DNA modification. To date, most mapping technologies of 5caC have relied on the use of bisulfite which converts 5caC differentially from other modified cytosines. Herein, we present quality control results for the caCLEAR method in mouse embryonic stem cells of different states of pluripotency as well as cells lacking TDG, an enzyme that can remove 5caC. Additionally, 5caC enrichment maps

are presented across various genomic elements and show that caCLEAR could be used as genome–wide DNA modification profiling technique.

6.4.2 Materials and Methods

6.4.2.1 Samples Analysed

The caCLEAR method was applied for the analysis of 5caC in mouse ESCs. Tdg depleted (*Tdg −/−*; Tdg) and Tet1/2/3 triple knockout (Tet TKO) mESCs were used as positive and negative controls in addition to wild–type (WT) mESCs. Depending on the culture conditions, mESCs can adopt two inter–convertible states resembling two different developmental stages (Habibi et al., 2013). For instance, mESCs grown in serum supplemented with leukaemia inhibitory factor are similar to cells from the early epiblast, while mESCs cultivated in serum–free medium with two small molecule kinase inhibitors (2i) closely resemble cells from the inner cell mass (Martello and Smith, 2014). The mESCs were cultivated in serum/LIF and serum/LIF/2i conditions (termed later as serum and serum–2i mESCs) and to increase the quality of analysis and reproducibility of results, all samples contained two technical replicates. A summary of all the samples used can be found in **Table 6.5**. All sequencing data was processed as described in **Chapter 4**.

6.4.2.2 Identifying 5caC Modified CG sites

The 5caCG sites were selected by evaluating various combinations between CG sites identified in different samples (i.e., Tet TKO, Tdg and WT). First, CG sites identified in Tet TKO libraries were removed from all the samples. Since Tet TKO samples were negative controls, these CG sites should be random noise, thus are false–positive sites. For Tdg samples, we only used those CG sites that in both technical replicates had CG–coverage equal or greater than the average coverage in that particular sample. A similar strategy was applied for wild–type samples,

Table 6.5 | mESCs Samples Analysed Using the caCLEAR Method

"Sample identifier" defines biological replicate, while "Replicate identifier" defines technical replicate. "Genotype" specifies genetic buildup, and "Growth conditions" environmental conditions of a given sample. "GEO code" encodes sample identifier deposited under GEO accession GSE142319.

Sample identifier	Replicate identifier	Genotype	Growth conditions	GEO accession code
Serum WT	R1	Wild type	Serum	GSM4225201
Serum WT	R2	Wild type	Serum	GSM4225202
Serum 2i WT	R1	Wild type	Serum 2i	GSM4225203
Serum 2i WT	R2	Wild type	Serum 2i	GSM4225204
Serum Tdg	R1	*Tdg−/−*	Serum	GSM4225205
Serum Tdg	R2	*Tdg−/−*	Serum	GSM4225206
Serum 2i Tdg	R1	*Tdg−/−*	Serum 2i	GSM4225207
Serum 2i Tdg	R2	*Tdg−/−*	Serum 2i	GSM4225208
Tet TKO	R1	Tet 1/2/3 triple knockout	Serum	GSM4225209
Tet TKO	R2	Tet 1/2/3 triple knockout	Serum	GSM4225210

however additional filtering steps were included. First, CG sites that in both technical replicates had coverage equal or greater than the average coverage were selected, then all CG sites that were not identified in the corresponding Tdg sample were removed. Finally, 10% of identified CG sites that had the largest coverage difference between the technical replicates were removed.

6.4.3 Quality Control of the Processed caCLEAR Sequencing Data

On average, each sample contained 34 million raw sequencing reads that varied greatly between different sample groups (**Figure 6.23**). Both Tdg sample groups and 2i wild–type sample group on average compromised 53 million reads ($sd = 8$), while remaining sample groups on average contained only 15 million reads ($sd = 1.5$).

Highest average coverage of identified CG sites was observed in Tdg samples, where it was very similar between the 2i and not 2i groups. In wild–type samples coverage was lower and varied greatly between two previously mentioned groups (**Figure 6.24 B**, **Supp. Table 3**).

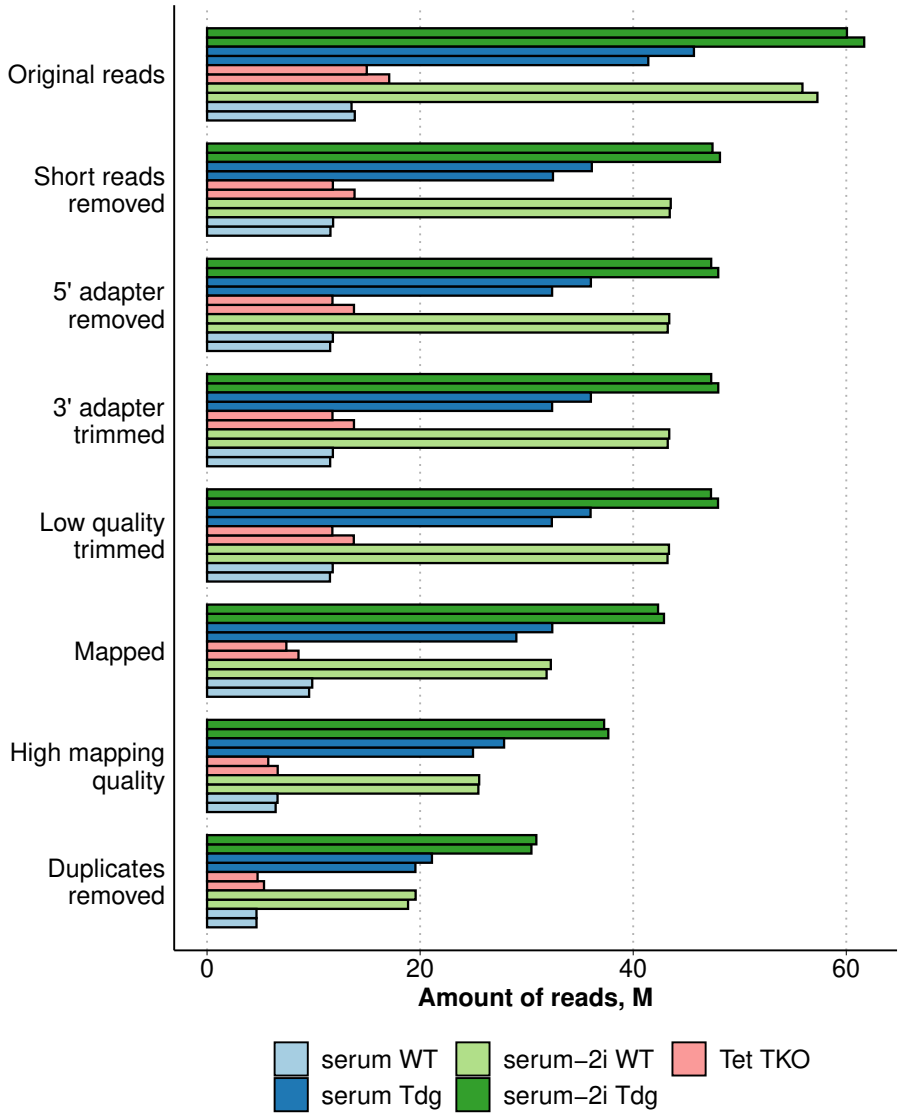


Figure 6.23 | Amount of caCLEAR Reads After Each Processing Step

The absolute change in read numbers after each processing step in mESC samples pertaining to different experimental groups.

Interestingly we observed great difference in the amount of identified CG sites when comparing 2i and not 2i groups (**Figure 6.24 B**). Both Tdg and wild–type samples with 2i contained higher amount of identified CG sites than samples without 2i (almost two times for Tdg and three times for wild–types samples).

Technical replicates from both Tdg groups showed similar medium to high correlation (Pearson's $r = 0.87$ and Spearman's $\rho = 0.49$, **Figure 6.25 A**). Meanwhile both wild–type groups showed different correlations between the technical replicates. Pearson's $r = 0.42$ (Spearman's $\rho = 0.12$) for 2i group and Pearson's $r = 0.29$ (Spearman's $\rho = 0.23$) for not 2i group. Correlation between the technical replicates expectedly increased when computed and higher resolutions (**Figure 6.25 B**). Pearson's r reached nearly maximum correlation when evaluated in 5 kb regions for both Tdg groups. Meanwhile for wild–type and Tet samples correlation increased with the size of used genomic bins and expectedly could be even higher when computed in even larger regions. Finally, overlap between the identified CG sites was calculated (**Figure 6.25 A, C**). Both Tdg sample groups showed higher Jaccard coefficient between the technical replicates and with another Tdg group when tested with Fisher's exact test. Overlap for the wild–types groups was lower than in the corresponding Tdg libraries, but higher than in Tet control proving that identified CG sites were not selected randomly.

6.4.4 Epigenomic Maps

Although the distribution of the called 5caCG sites varied in gene regulatory elements and genomic features for different conditions and cell types, most 5caCG sites were enriched in poised and active enhancers (marked by H3K4me1 and H3K27ac/H3K4me1, respectively) and binding regions of various pluripotency–related transcription–factors, such as Sox2, Nanog, Oct4, and SINE repeats (**Figure 6.26**).

Finally we compared identified CG sites with open chromatin regions from various mouse tissues and organs generated with the ATAC–seq

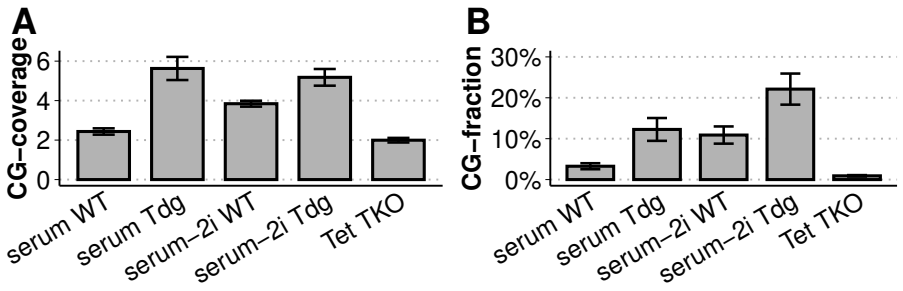


Figure 6.24 | 5caCG Modified DNA Signal in mESC Samples

(A) The average coverage of identified CG sites per each chromosome. Tdg samples show higher average coverage compared to the wild–type samples, while Tet control sample shows the lowest CG–coverage. On average each Tdg sample had 1.8 times higher CG–coverage than a corresponding wild–type sample. **(B)** The amount of identified CG sites per chromosome. On average each Tdg sample identified 2.5 times more CG sites than a corresponding wild–type sample. Coverage and identified CG site statistics is computed using all identified CG sites in the genome (i.e., before applying stringent filtering steps documented in section **Section 6.4.2.2**).

method (**Figure 6.27**). We observed great variability between sample groups and used mouse organs. Both wild–types samples showed lower Fisher's exact test estimates in all tested organs with serum–2i sample showing significant depletion. Serum–2i Tdg sample was between wild–type and serum samples — it showed significant enrichment or depletion depending on the organ. Identified CG sites were enriched in gastrointestinal organs (e.g., stomach and intestine), lungs, but lightly depleted in neural system organs.

6.4.5 Discussion

6.4.5.1 New Insights Provided by This Study

This chapter described the application of the caCLEAR method to profile 5caCG modification at a single CG resolution genome–wide. We show that the computed 5caCG signal is reproducible and correlates well between technical replicates. Additionally, we presented genome–wide 5caCG modification enrichment maps across various genomic elements

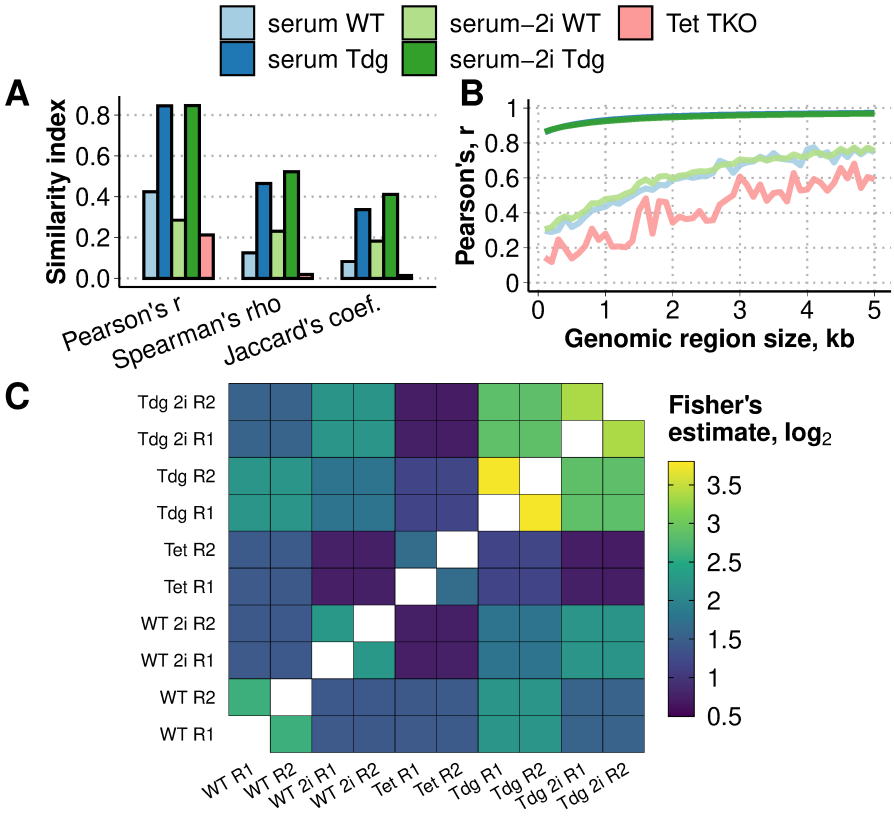


Figure 6.25 | Similarity Statistics of Samples Analysed Using the caCLEAR Method

(A) Similarity index measured in Pearson's correlation, Spearman's correlation or Jaccard's coefficient between replicates at a single CG site resolution. **(B)** Correlation coefficient between replicates at varying sizes of genomic bins. **(C)** Fisher's estimates between all identified CG sites. Tdg samples show highest similarity between replicates and with other Tdg samples, while Tet control sample show lowest similarity with all the other sample groups.

together with 5caCG modification enrichment within open chromatin regions specific to distinct mouse tissues.

6.4.5.2 Concluding Remarks

Herein, we provide an application of the caCLEAR method in mouse embryonic stem cells, demonstrating that the caCLEAR method could be

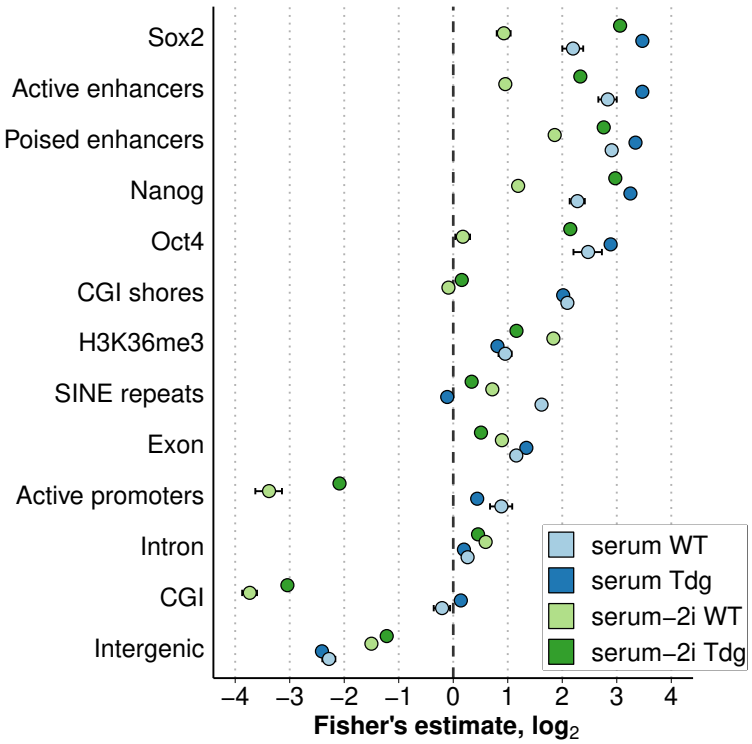


Figure 6.26 | caCLEAR Signal Across Genomic Elements

Fisher's exact test estimates for the enrichment or depletion of 5caCG sites across various genomic elements in wild–type and Tdg samples. Poised enhancers, regions with H3K4me1 histone marks; active enhancers, regions with H3K4me1 and H3K27ac marks; active promoters, 2 kb regions upstream of genes that overlap the H3K9ac histone mark. All shown enrichment or depletion values have Fisher's exact test p–values less than 0.05.

used as genome–wide profiling technique to improve our understanding of epigenomic patterns.

Statement VI — caCLEAR method provides information about DNA modification signal across different genomic elements.

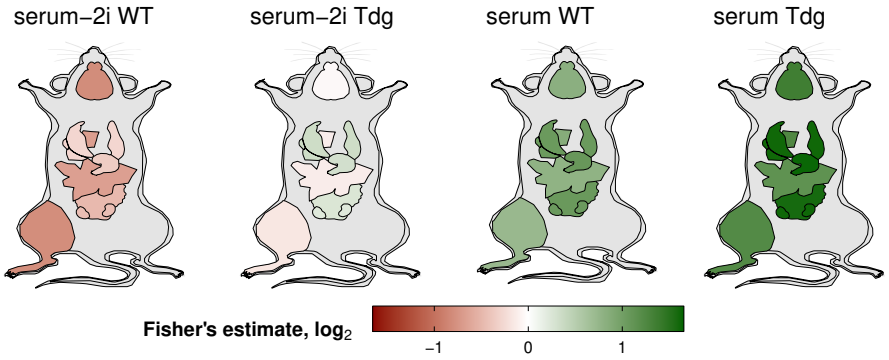


Figure 6.27 | caCLEAR Signal Enrichment Within Open Chromatin Loci

Average Fisher's exact test estimates of the caCLEAR identified 5caCG site enrichment or depletion within open chromatin loci identified in different mouse tissues using the ATAC–seq method. Both serum–2i samples show lower overlap with open chromatin loci (wild–type sample shows significant depletion, while 2i Tdg sample results in odds ratios around zero). Both serum samples show significant enrichment in all reported mouse tissues.

6.5 Application of hmTOP–seq and TOP–seq Methods for Prenatal Testing

6.5.1 Introduction

Trisomy of chromosome 21 (T21) is the most common human autosomal aneuploidy that results in a collection of phenotypical features (physical and intellectual disabilities) known as Down syndrome (Antonarakis et al., 2020). Invasive diagnostic procedures such as amniocentesis and chorionic villus sampling followed by genetic analysis (e.g., karyotyping) are currently used to confirm the diagnosis of T21 (Alfirevic et al., 2003). Although the safety of invasive procedures has improved, the risk of miscarriage (0.3% to 0.9% for amniocentesis and chorionic villus sampling) still remains (Salomon et al., 2019). Hence, to reduce the number of invasive diagnostic procedures, non–invasive and highly reliable prenatal screening tests are required.

Since the discovery of fetal genomic material in the form of circulating

cell–free fetal DNA in the blood plasma of pregnant female, many efforts have been made to employ cffDNA for non–invasive prenatal testing of fetal genomic mutations (Lo et al., 1997; Norton and Wapner, 2015). Such screening has a detection rate for T21 of more than 99%, with a false positive rate as low as 0.1% (Gil et al., 2015). Thus, NIPT based diagnostic technologies represent a substantial improvement over traditional screening. However, the detection of cffDNA in maternal blood circulation is a considerable challenge as only 10% of the DNA in the plasma of pregnant female is fetally derived (Lo et al., 2010).

Here, we applied TOP–seq and hmTOP–seq technologies to analyse DNA modifications in maternal cfDNA for the identification of fetal–derived genomic regions. Genome–wide 5hmCG and uCG modification maps of cfDNA or chorionic villus DNA samples were created and we also calculated differential modification signal enrichment in various genomic elements for the employed sample groups. Most importantly, fetal trisomy of chromosome 21 was detected with excellent specificity/sensitivity using regional modification differences. In addition, the fetal–fraction from cfDNA was calculated using uCG and 5hmCG signal.

6.5.2 Materials and Methods

6.5.2.1 Samples Analysed

In this study, TOP–seq and/or hmTOP–seq technologies were applied to construct epigenomic maps of the various sample groups **Table 6.6**. Four non–pregnant controls (NPC) were analysed using TOP–seq, another four NPCs were analysed using hmTOP–seq technologies and another three NPCs were analysed using both TOP–seq and hmTOP–seq technologies. Seven chorionic villi (CV) samples were analysed using TOP–seq (three of them were also analysed using the hmTOP–seq method). Finally, thirteen cfDNA samples from pregnant female were analysed using TOP–seq and eleven cfDNA samples hmTOP–seq technologies (six samples were analysed using both technologies). A fraction of pregnant

female was carrying T21 positive fetuses (five TOPseq and four hmTOPseq). Four additional cfDNA samples from pregnant female were also obtained, each of them containing on average 2.5 million raw reads (two T21 fetuses and two non–T21 fetuses). These samples with fewer reads were necessary to provide sensitive analysis of the fetal fraction and detection of fetal abnormalities. All sequencing data was processed as described in **Chapter 4**.

Sample outlier identification was performed separately for uCG and 5hmCG samples. CG–coverage matrices were transformed using Hellinger transformation, then represented in two–dimensional space using nMDS with Bray–Curtis similarity index (Bray and Curtis, 1957; Legendre and Gallagher, 2001). Samples that were further than two standard deviations away from the mean of their own sample group (cfDNA of NPCs, cfDNA of pregnant female, CVs) in either the first or second nMDS dimensions were deemed outliers and removed from further analysis. There were three outlying samples in uCG and one in the 5hmCG dataset (two uCG cfDNA samples, one uCG CV samples and one 5hmCG samples) that were removed from further analysis.

6.5.2.2 DMR Calculation in Cell–Free DNA

Chromosome 21 was partitioned into 100 bp non–overlapping regions and the log transformed CG–coverage and CG–fraction was calculated for each region. The CG–coverage was normalised by the total read count in a reference chromosome and the CG–fraction was normalised by the overall identified fraction in a reference chromosome. Chromosomes 20 and 16 were used as references for uCG and 5hmCG data, respectively. Next, for each region, two logistic regression models were fitted. The full model included CG–coverage, CG–fraction, and, for T21–specific DMRs, fetal sex and fetal fraction as independent variables. CG–coverage and CG–fraction were excluded from the null model. ANOVA test was used to compare full and null models to obtain a p–value. In cases where the models did not converge, fetal sex was removed. FDR was used to

Table 6.6 | Human Samples Analysed in the NIPT Study

"Sample identifier" defines biological replicate and "Sample status" defines biological state of the individual from which cffDNA sample was obtained (e.g., pregnant female, non–pregnant control female). "Fetal karyotype" and "Fetal sex" describe fetal genotype. "Source of DNA" specifies tissue (e.g., circulating cell–free DNA or chorionic villus). "Analysed modifications" column defines if unmodified DNA, 5hmCG modified DNA or both were analysed in a given sample.

Sample identifier	**Sample status**	**Fetal karyotype**	**Fetal sex**	**Source of DNA**	**Analysed modifications**
002	Pregnant	46, XY	Male	cfDNA	uCG
004T21	Pregnant	47, XY, +21	Male	cfDNA	uCG & 5hmCG
006	Pregnant	46, XY	Male	cfDNA	uCG
007	Pregnant	46, XX	Female	cfDNA	5hmCG
009	Pregnant	46, XX	Female	cfDNA	uCG
011T21	Pregnant	47, XX, +21	Female	cfDNA	5hmCG
016T21	Pregnant	47, XX, +21	Female	cfDNA	uCG & 5hmCG
022	Pregnant	46, XY	Male	cfDNA	uCG
023	Pregnant	46, XY	Male	cfDNA	5hmCG
025T21	Pregnant	47, XY, +21	Male	cfDNA	uCG
031	Pregnant	46, XX	Female	cfDNA	uCG
041	Pregnant	46, XX	Female	cfDNA	uCG
049T21	Pregnant	47, XX, +21	Female	cfDNA	uCG
050	Pregnant	46, XX	Female	cfDNA	5hmCG
050T21	Pregnant	47, XY, +21	Male	cfDNA	5hmCG
068T21	Pregnant	47, XX, +21	Female	cfDNA	uCG
083	Pregnant	46, XX	Female	cfDNA	uCG
130	Pregnant	46, XY	Male	cfDNA	5hmCG
136	Pregnant	46, XY	Male	cfDNA	5hmCG
137	Pregnant	46, XY	Male	cfDNA	uCG & 5hmCG
144	Pregnant	46, XX	Female	cfDNA	5hmCG
001	NPC	NA	NA	cfDNA	5hmCG
004	NPC	NA	NA	cfDNA	5hmCG
005	NPC	NA	NA	cfDNA	5hmCG
007ctrl	NPC	NA	NA	cfDNA	5hmCG
011	NPC	NA	NA	cfDNA	uCG
017	NPC	NA	NA	cfDNA	uCG
035	NPC	NA	NA	cfDNA	uCG
E	NPC	NA	NA	cfDNA	uCG & 5hmCG
E_R2	NPC	NA	NA	cfDNA	uCG
J	NPC	NA	NA	cfDNA	uCG & 5hmCG
M	NPC	NA	NA	cfDNA	uCG & 5hmCG
10	Pregnant	46, XX	Female	CV	uCG & 5hmCG
19	Pregnant	46, XX	Female	CV	uCG
20	Pregnant	46, XY	Male	CV	uCG
22	Pregnant	46, XX	Female	CV	uCG & 5hmCG
23	Pregnant	46, XY	Male	CV	uCG
24	Pregnant	46, XX	Female	CV	uCG
6	Pregnant	46, XY	Male	CV	uCG & 5hmCG

adjust p–values for multiple testing and q–value < 0.05 was used as a significance threshold if not specified otherwise.

For each placenta–specific DMR, leave–one–out cross–validation procedure was performed as described above to determine its ability to diagnose T21. For each cross–validation cycle, a Bayesian generalised linear model with normalised CG–coverage and CG–fraction as independent variables was constructed (Gelman et al., 2008). DMRs with area under the curve (AUC) equal to one were selected as discriminatory of fetal karyotype.

6.5.2.3 Fetal Fraction Calculation in Cell–Free DNA

The fetal fraction was predicted using the SeqFF method, an optimal method for the used technologies as it is applicable to both fetal sexes and does not require parental genotype (Kim et al., 2015). SeqFF fetal fraction prediction is based on two estimates – elastic net (ENET) and weighted rank selection criterion (WRSC).

To calculate ENET and WRSC estimates, the genome was divided into 50 kb non–overlapping bins (although different bin sizes could be evaluated, but 50 kb was chosen to mirror the data partitioning of the original SeqFF publication) (Kim et al., 2015). Within each bin, the GC–content and total coverage (uCG and/or 5hmCG) for each sample was calculated. Next, weighted coverage values were normalised for GC–content using polynomial regression (`R stats::loess` function) (Chambers et al., 1990). Normalised values were used in ENET and WRSC estimators and the final estimated fetal fraction was an average between the two estimators.

ENET is a regularised regression that is a combination of least absolute shrinkage and selection operator and ridge regression (Friedman et al., 2010), which can be simply written as $\sum_{i=0}^{n} \beta_i X_i + \delta$ where β is determined by ENET for ith 50 kb genomic region among n autosomal bins, X is the normalised coverage for bin i and δ is a correction parameter

defined by ENET. In WRSC, bin values for chromosome Y were predicted using reduced–rank regression (Izenman, 1975). Then, for both sexes, chromosome representations were evaluated as the ratios of total read counts between sex and autosomal chromosomes. SeqFF model parameters were calculated using the fetal–fraction dataset from 25,312 pregnant female and further validated on two independent sets of pregnant female (233 and 272 number of samples) (Kim et al., 2015).

6.5.2.4 Enrichment Analyses

Enrichment of genomic elements with the strongest signal was performed as follows. First, the genome was divided into 1 kb wide non–overlapping regions and the total coverage was computed per sample within each region. The total coverage values were then averaged per group of samples (cfDNA of NPCs, cfDNA of pregnant female, CVs) and regions falling among the top 10% most covered regions were designated as those having the highest signal. Then, a contingency table was computed for each genomic CG falling into one of the highest signal regions and overlapping specific genomic elements. Fisher's exact test was performed to estimate the OR and p–value. Enrichment of DMRs with genomic regions was computed by forming a contingency table which contained information regarding whether each DMR is significant and intersects a specific genomic element. As above, Fisher's exact test was used to estimate the ORs and p–values.

From the set of ARIES mQTL birth and pregnancy probes only high–quality probes were selected (Gaunt et al., 2016; Naeem et al., 2014). In total, there were 4,243 Illumina Infinium HM450 array probes in chromosome 21 and 2,642 after selecting only high–quality probes (238 birth mQTL and 291 pregnancy mQTLs). Enrichment of mQTL probes with DMRs was calculated by creating a contingency table which evaluated whether each Illumina Infinium HM450 array probe is an mQTL and intersects a DMR.

6.5.3 Quality Control of Processed Sequencing Data

After processing sequencing reads and calculating CG–coverage, we firstly compared similarity statistics between the biological replicates (**Figure 6.28**). On average uCG biological replicates were more similar than 5hmCG replicates in all measured statistics. Interestingly, correlation between the biological replicates was highest for the cfDNA from the pregnant female uCG samples, while in 5hmCG samples this group had lowest correlation compared to two other experimental groups.

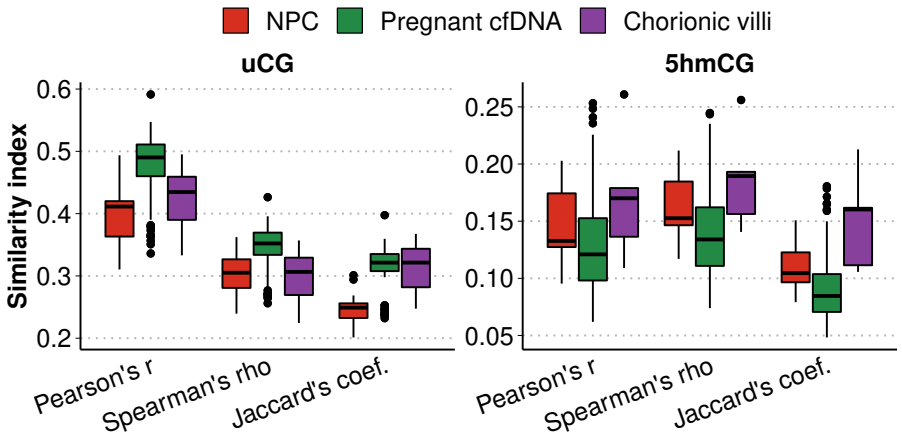


Figure 6.28 | Similarity Statistics of NIPT Samples

Similarity index measured in Pearson's correlation, Spearman's correlation or Jaccard's coefficient between biological replicates in given sample groups at a single CG site resolution.

To test whether uCG and 5hmCG modification differences could distinguish between the sample groups, we first looked at the total sequencing coverage of the uCG and 5hmCG sites. The mean total uCG–coverage was different across the three groups of samples (ANOVA p–value 7×10^{-7}); it was the lowest among NPCs and the highest among CVs (**Figure 6.29**). Importantly, the mean total coverage of the pregnant female cfDNA was in between the NPCs and CVs. Furthermore, the fraction of identified uCG sites covered by at least one read showed a very similar difference among all groups (ANOVA p–value 7.4×10^{-7}). For the 5hmCG samples, the total coverage was not significant between groups (ANOVA p–value 0.05) but the fraction of identified CG sites

increased from NPCs towards CVs (ANOVA p–value 8.9×10^{-3}). Considering both a higher total coverage and higher fraction of covered CG sites, CV tissue is relatively hypomethylated and also shows increased hydroxy–methylation compared to cfDNA of NPCs or pregnant female.

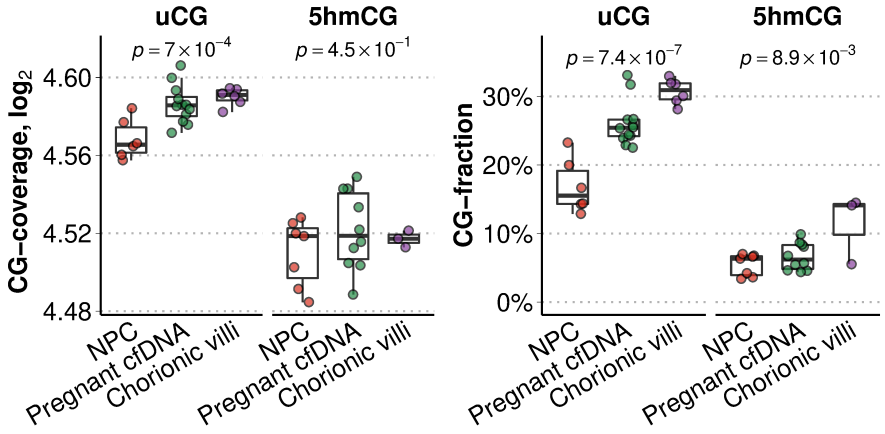


Figure 6.29 | Coverage Statistics of NIPT Samples

Total CG–coverage in cfDNA and CV tissue sample groups for the uCG and 5hmCG signals. Total log_2 transformed sequencing coverage of autosomes was computed for each sample and ANOVA was used to test for differences in distributions across sample groups. CG–fraction in cfDNA and CV tissue sample groups for the uCG and 5hmCG data. Fraction of CG sites covered by at least one read to the total number of CG sites was computed for each sample and ANOVA was used to test for differences in distributions across sample groups.

6.5.4 Epigenomic Maps

Next, we explored the distribution of the identified uCG sites and 5hmCG sites across genomic elements (**Figure 6.30**). As expected, the hypomethylated regions concentrated in CGIs, 5'UTRs and promoters of protein–coding genes and long intergenic non–coding RNAs (lincRNAs). Importantly, the enrichment difference between uCG and 5hmCG was also largest in these four element groups. The 5hmCG sites were mostly observed in 3'UTRs, exons and introns, however enrichment difference between uCG and 5hmCG was not as high in these elements (with exons showing the same uCG enrichment level as 5hmCG). Therefore, both

uTOP–seq and hmTOP–seq approaches may provide distinct but complementary data for the detection and analysis of cffDNA fragments in maternal circulation.

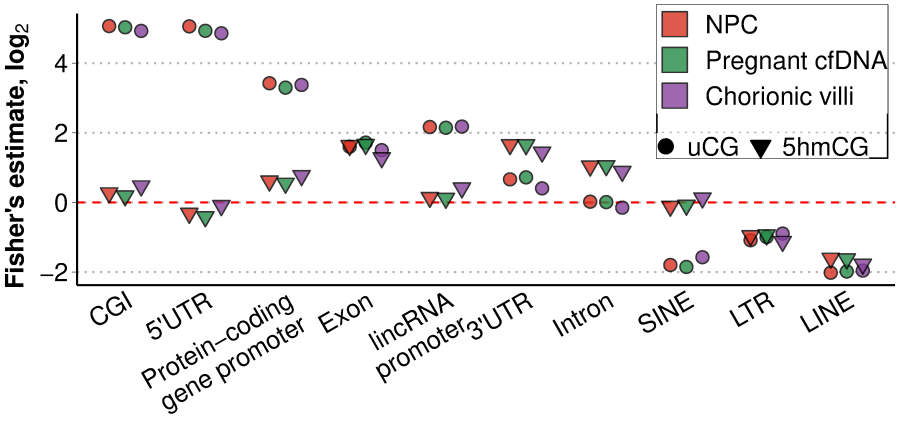


Figure 6.30 | Identified CG Sites Across Genomic Elements in NIPT Samples

Fisher's exact test estimates for the enrichment or depletion of uCG and 5hmCG sites across genomic elements (all Fisher's p–values less than 1.6×10^{-10}). The genome was divided into one kilobase regions and the total coverage per sample was averaged across sample groups for each region. Fisher's exact test was used to test whether the regions with the 10% of the strongest signal are enriched within a particular genomic element.

6.5.5 Differentially Modified Regions in Cell–Free DNA

We sought to identify fetal–specific genomic loci that could be used as uCG– and/or 5hmCG biomarkers for fetal phenotype. As our study contained DNA samples from different phenotypic groups (i.e., women pregnant with a non–T21 fetus, women pregnant with a T21 fetus, non–pregnant female, chorionic villi) this enabled identification of the desired biomarkers by applying inter–group comparisons, which are visualised in **Figure 6.31**.

Firstly, by comparing NPC samples with healthy pregnancy samples, we identified DMRs that were pregnancy–specific. This result was based on the assumption that such a comparison would remove the signal related to the control female epigenome and expose the DNA modification signal

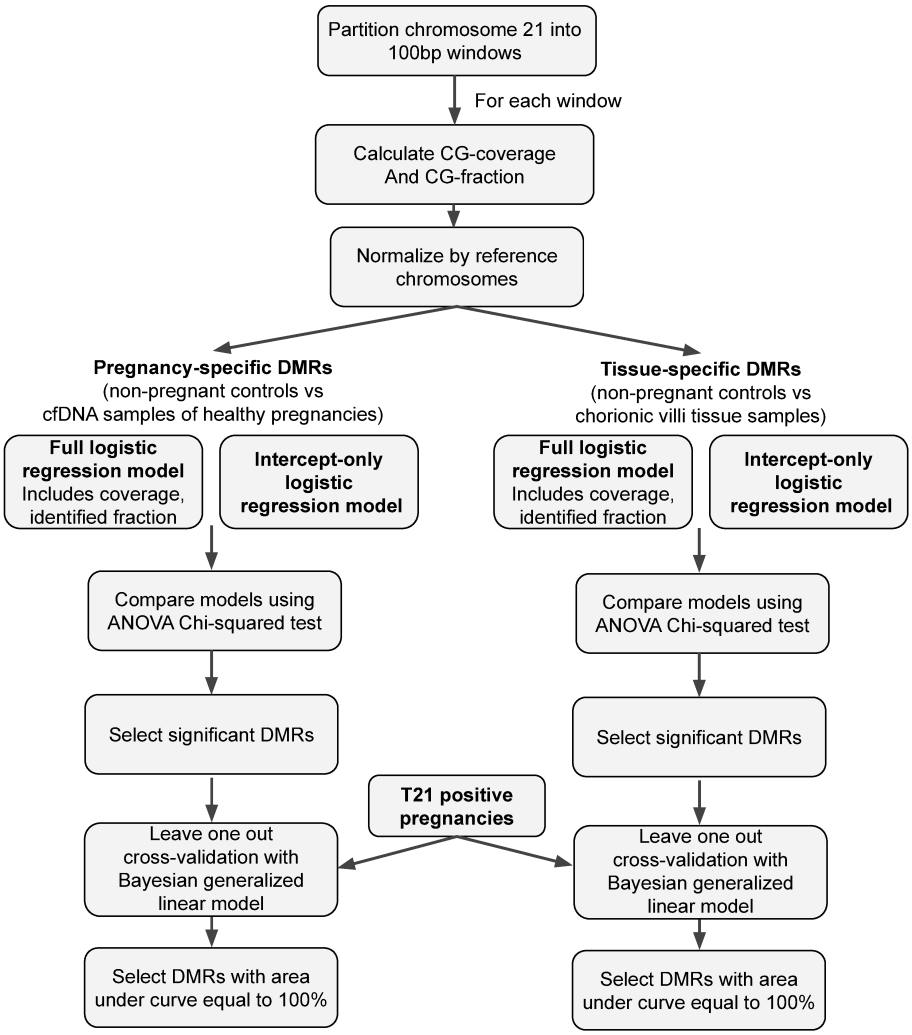


Figure 6.31 | DMR Identification in NIPT Samples Workflow

Workflow of the implement DMR identification algorithm in NIPT analysis. Genome–wide CG–coverage and CG–fraction is compared between the controls and healthy pregnancy samples to identify pregnancy–specific DMRs, while comparison between the controls and chorionic villi tissue samples leads to identification of tissue–specific DMRs. Finally, leave one out cross–validation is applied to select DMRs that are informative of the fetus karyotype.

that is specific to pregnant female or the fetus. Using logistic regression with the normalised CG–coverage and CG–fraction as independent variables, we identified 2,761 pregnancy–specific uCG DMRs (FDR q–value $< 5 \times 10^{-2}$) (**Table 6.7**). The same analytic approach did not yield

FDR–significant DMRs from the hmTOP–seq data, therefore, we used nominal p–value $< 5 \times 10^{-2}$ threshold and identified 4,930 pregnancy–specific 5hmCG DMRs.

Next, by comparing the same NPC samples with CV samples, it was revealed that the DNA modification signal is CV–specific (i.e., fetus–specific). This comparison revealed 16,555 CV–specific uCG DMRs (FDR q–value $< 5 \times 10^{-2}$) and 15,986 CV–specific 5hmCG DMRs (FDR p–value $< 5 \times 10^{-2}$).

Table 6.7 | Amount of NIPT DMRs

Amount of DMRs identified in specified group for both analysed DNA modification groups.

DNA modification	Pregnancy	Chorionic villi	Placenta	T21
uCG	2,761	16,555	2,164	3,490
5hmCG	4,930	15,986	1,589	2,002

Finally, by intersecting pregnancy–specific regions with CV–specific regions, we extracted fetus–specific DNA modification signal that could be found in cfDNA of pregnant female, we termed those regions placenta–specific DMRs. This intersection between the pregnancy–specific and CV–specific DMRs for both uCG and 5hmCG DMR sets was larger than could be expected by chance alone (n = 2,164, OR = 43; n = 1,589, OR = 5.5, for uCG and 5hmCG, respectively; p–values less than 1×10^{-15}). For the placenta–specific uCG DMRs, the difference between the NPCs and cfDNA samples of pregnant female was concordant with the difference between NPCs and CV samples (Pearson's $r = 0.82$ and Pearson's $r = 0.89$, for CG–coverage and CG–fraction, respectively, **Figure 6.32**). Similar results were observed for 5hmCG DMRs (Pearson's $r = 0.8$ and $r = 0.8$, for CG–coverage and CG–fraction, respectively).

Identified pregnancy–specific and CV–specific DMR sets intersected more than expected by chance, however, this result might have been influenced by genetic variation. It is possible that identified DMRs were derived from mQTL regions and observed DNA modification signal is related to DNA sequence variability. In this case, the used NPC samples could be characterised by one genetic background and pregnant female samples

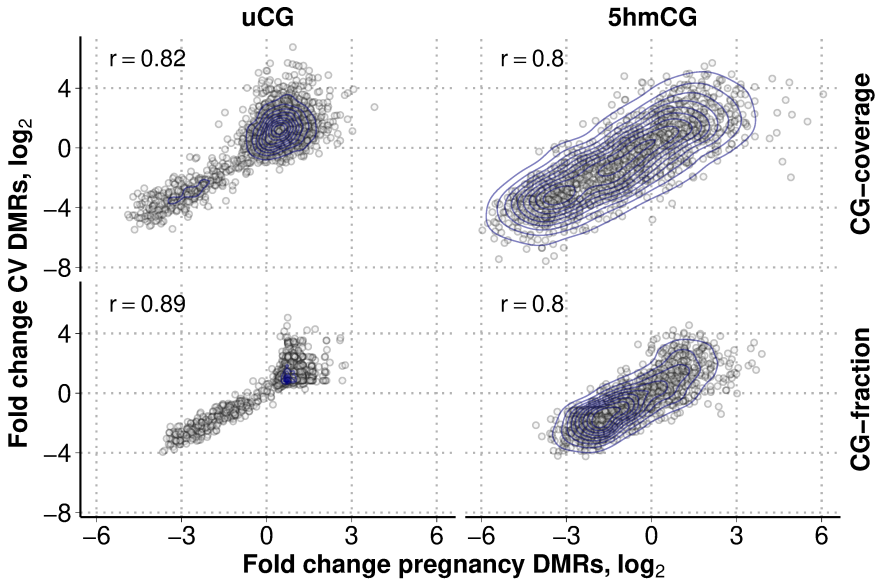


Figure 6.32 | Concordance Between CG–coverage and CG–fraction

Pearson correlation of the modification differences observed in the placenta–specific DMRs. Pregnancy–specific changes on the X–axis indicate modification differences between NPCs and cfDNA samples of pregnant female. CV–specific changes on the Y–axis indicate modification differences between NPCs and CVs.

could be characterised by another genetic background, with the intersection between the pregnancy–specific and CV–specific sets driven solely by signal influenced by DNA sequence variability (i.e., mQTLs). To test this assumption, we calculated the identified DMR intersection with known mQTL sets (mQTLs identified in cord blood and mQTL identified blood of pregnant female). Observed intersections between mQTLs and DMR groups were not significant (all calculated p–values from Fisher's exact test $> 5 \times 10^{-2}$). Moreover, birth and pregnancy mQTLs intersected less than 1% of DMRs, irrespective of the DMR group.

After identifying placenta–specific uCG and 5hmCG DMRs, we tested the overlap with different genomic elements **Figure 6.33**. The uCG DMRs were enriched in elements related to the 5' end of genes (5'UTRs, promoters of protein–coding and lincRNA genes and strongest enrichment in promoter CGIs), as well as placental enhancers. In contrast, 5hmCG DMRs were enriched in other protein–coding gene parts, exons,

introns and 3'UTRs.

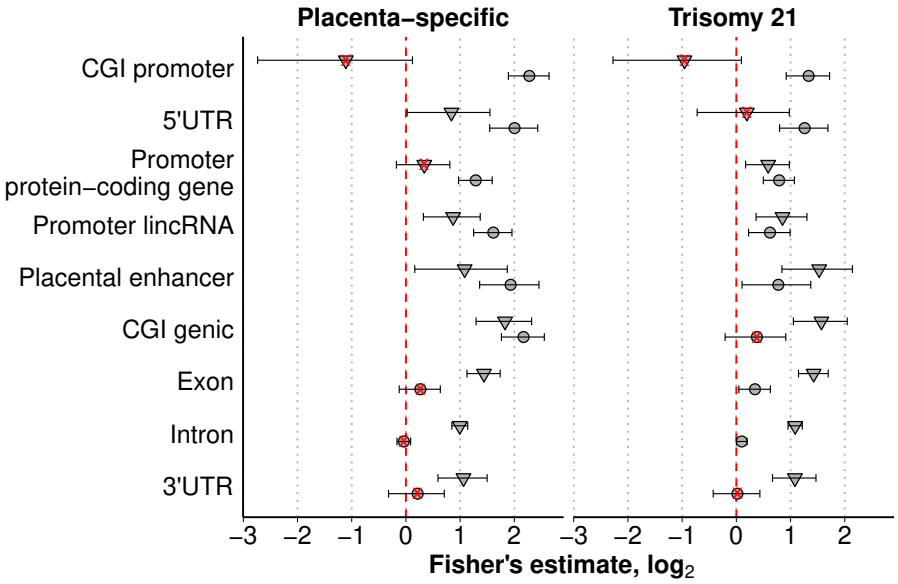


Figure 6.33 | Enrichment of NIPT DMRs in Genomic Elements

Enrichment of genomic elements for the placenta–specific and T21–specific DMRs using Fisher's exact test. All reported p–values are less than 0.05 except for ones marked with a red cross.

Finally, we asked whether the placenta–specific DMRs are informative of fetal karyotype (i.e., T21). Using leave–one–out cross–validation, we constructed and evaluated a logistic regression model for each placenta–specific DMR with the CG–coverage and CG–fractions as independent variables and fetal karyotype as the response variable. In total, 376 uCG and 496 5hmCG DMRs were discovered in chromosome 21 that classified samples according to fetal karyotype with 100% accuracy (AUC = 1) **Figure 6.34**.

After identifying placenta–specific DMRs and the subset that can classify samples according to the fetal karyotype with 100% accuracy, we then took a different approach, directly evaluating modification differences between the cfDNA samples of healthy and T21–positive pregnancies and computed the T21–specific DMRs. A logistic regression model was used with the CG–coverage and CG–fraction as independent variables and karyotype as the response variable. In addition, we adjusted for

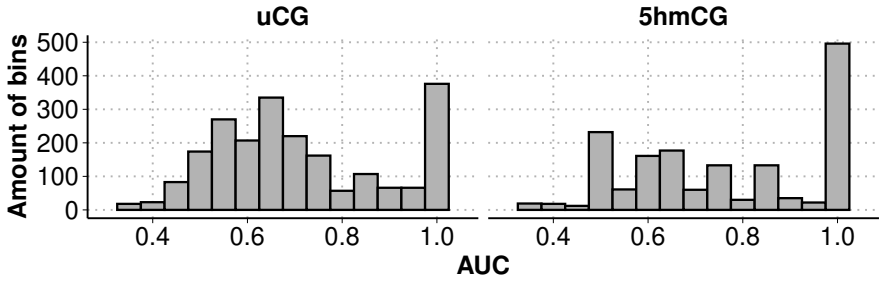


Figure 6.34 | Distribution of AUC Values

Computed AUC values for the fetal karyotype prediction using placenta–specific DMRs.

possible confounding effects of the fetal fraction and fetal sex which could not be accounted for in the previous analyses (i.e., in comparison with NPC samples). We identified 3,490 uCG and 2,002 5hmCG DMRs (FDR q–value $< 5 \times 10^{-2}$), of which only 82 intersected between the two sets (OR = 2.3, p–value $= 1.1 \times 10^{-10}$).

Only 216 and 124 T21–specific DMRs intersected with placenta–specific DMRs for uCG and 5hmCG respectively (OR = 6.1 and OR = 8.2; p–value $< 2.2 \times 10^{-16}$), demonstrating that different DMR identification strategies lead to different DMR sets in chromosome 21 which can be complementary for detecting fetal karyotype. It was noted that T21–specific DMRs exhibit higher CG–coverage and CG–fraction differences than regions that did not exhibit differential modification (**Figure 6.35**).

Interestingly, both uCG and 5hmCG DMR sets better intersected the pregnancy–specific DMR sets (OR = 6.6 and OR = 9, for uCG and 5hmCG, respectively) than the CV–specific DMR sets (OR = 2.4 and OR = 2.9 for uCG and 5hmCG, respectively; for all comparisons p–value $< 2.2 \times 10^{-16}$). This result suggests two possibilities, that fetal tissues other than the placenta–derived trophoblasts might contribute to the cfDNA mixture of maternal blood or it could be an artifact of used tissue type – the pregnancy–specific DMRs are also measured in cfDNA, like the T21–specific DMRs, where the CV–specific DMRs are identified by comparing two different tissues. Additionally, we tested the overlap of T21–specific DMRs with known mQTL loci and no significant

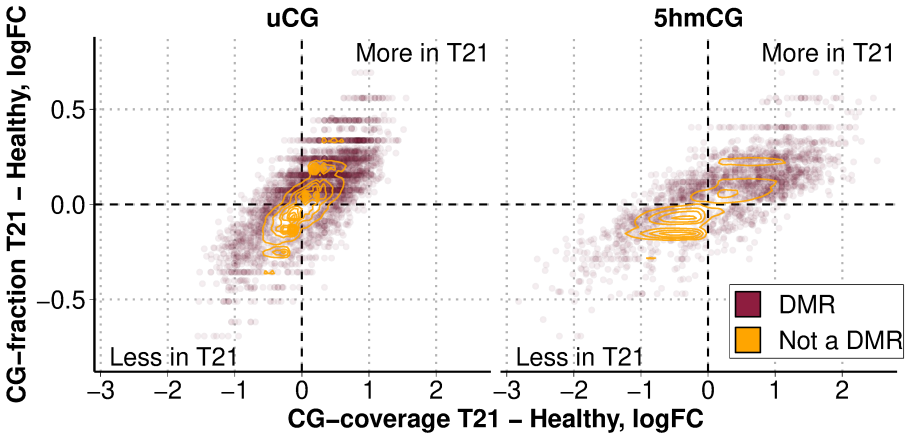


Figure 6.35 | Distribution of CG–coverage and CG–fraction in T21–Specific DMRs

T21–specific DMRs exhibit higher CG–coverage and CG–fraction differences than non–differentially modified regions. logFC represents a log fold–change difference between T21–diagnosed pregnancies and healthy pregnancies.

enrichment was found (Fisher's exact test p–value $> 5 \times 10^{-2}$) and only 0.5% of identified T21–specific DMRs were covered by mQTL loci.

Enrichment of genomic elements across the T21–specific DMRs was similar to that of the pregnancy–specific DMRs (**Figure 6.33**). T21–specific uCG DMRs were significantly enriched in 5'UTRs, promoter CGIs and promoter of protein–coding genes (OR = 2.4, OR = 2.4, and OR = 2.5, respectively) and T21–specific 5hmCG DMRs were significantly enriched in other protein–coding gene parts: exons, introns and 3'UTRs. However, contrary to placenta–specific DMRs, T21–specific uCG DMRs were less enriched in promoters on lincRNAs, placental enhancers and genic CGIs.

The distribution of the T21–specific DMRs along chromosome 21 was different for the uCG and 5hmCG datasets (p–value $= 1.6 \times 10^{-12}$; Kolmogorov–Smirnov test). Most of the identified 5hmCG DMRs tended to cluster at the end of the long arm, whereas the uCG DMRs were more evenly distributed along the long arm of chromosome 21, with

46% and 64% of the uCG and 5hmCG T21–specific DMRs, respectively, overlapping protein–coding genes. Interestingly, DMRs that intersected the T21–specific uCG and 5hmCG DMR sets (82 DMRs in total) showed very high enrichment for protein–coding exons (OR = 4.4, p–value = 8×10^{-4}). These exons corresponded to seven genes, three of which have been previously associated with Down syndrome: *GART*, *DNMT3L* and *AIRE*. *GART* genes encodes trifunctional enzyme (glycinamide ribonucleotide synthetase, aminoimidazole ribonucleotide synthetase and glycinamide ribonucleotide formyltransferase) that is expressed during normal prenatal cerebellum development and is undetectable in cerebellum shortly after birth (Brodsky et al., 1997). However, this complex is overexpressed in the cerebellum during the postnatal development of individuals with Down syndrome. *DNMT3L* genes encode DNA–methyltransferase that is overexpressed in neural progenitors of individuals with Down syndrome (Lu et al., 2016). Finally, the *AIRE* gene encodes a transcription factor expressed in the medulla of the thymus, however its expression is significantly reduced in the thymus of individuals with Down syndrome (Lima et al., 2011).

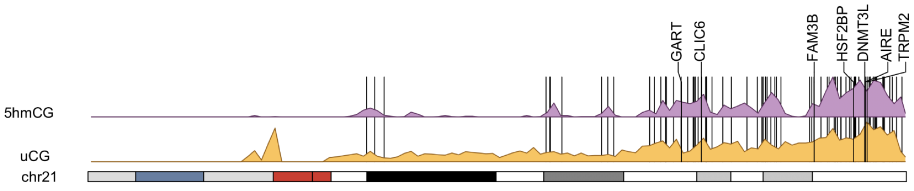


Figure 6.36 | Distribution of T21–Specific DMRs Across Chromosome 21

Ideogram of chromosome 21 (centromeric region marked in red) showing distributions of the T21–specific uCG and 5hmCG DMRs. DMRs shared between the sets are indicated with dark vertical bars. 7 genes containing the shared DMRs in their exons are specified above the graph.

6.5.6 Fetal Fraction

The reliability of NIPT depends on biological as well as experimental variations and it has been speculated that the fetal fraction should be at least 4% to allow for the reliable detection of common trisomies

(Palomaki et al., 2011). Laboratory tests to determine the fetal fraction directly on isolated DNA have been developed but have the risk of adding an error prone step in the diagnostic workflow. When splitting the isolated DNA in two different lab flows (one for determination of fetal fraction and the other for the library preparation), an error or sample swap in either flow might result in a mismatch between the fetal fraction and the NIPT result. Furthermore, this procedure reduces the amount of analysed DNA which is already precious and complicated to obtain. Therefore, the fetal fraction should preferably be determined from the same next–generation sequencing data used for the determination of the chromosomal aberrations (van Beek et al., 2017).

Having established that the uTOP–seq and hmTOP–seq signals are higher among pregnant female, we further sought to determine the correlation between the signal strength and fetal fraction. SeqFF was applied to the uTOP–seq and hmTOP–seq data, observing a high correlation between the predicted and reference fetal fractions (Pearson's $r = 0.86$, p–value $= 3.2 \times 10^{-4}$ and $r = 0.9$; p–value $= 3.9 \times 10^{-4}$), for uCG and 5hmCG, respectively (**Figure 6.37**). Importantly, a simple linear regression revealed that an increase in the reference fetal fraction by 0.01 corresponded to an increase in the fetal fraction predicted from uCG profiles by 0.079. For 5hmCG data, the predicted foetal fraction decreased by 0.226 for every 0.01 increase of the reference fraction. Interestingly, an increasing fetal fraction would acquire increasing read counts in uTOP–seq but decreasing read counts in hmTOP–seq. Such inverse relationship in hmTOP–seq most likely indicates that the regions used by SeqFF are highly enriched in uCG sites but depleted in 5hmCG sites in cffDNA.

These results indicated that both uTOP–seq and hmTOP–seq enable enrichment of fetal circulating DNA from maternal cfDNA. Importantly, hmTOP–seq may be more sensitive for the evaluation of the fetal fraction, most likely due to the well–known role of tissue specificity of 5hmC. Consequently, fewer reads would be necessary to provide sensitive analysis of the fetal fraction and detection of fetal abnormalities. To further test this hypothesis, four additional cfDNA samples were analysed by hmTOP–seq obtaining on average 2.5 million raw reads for each sample.

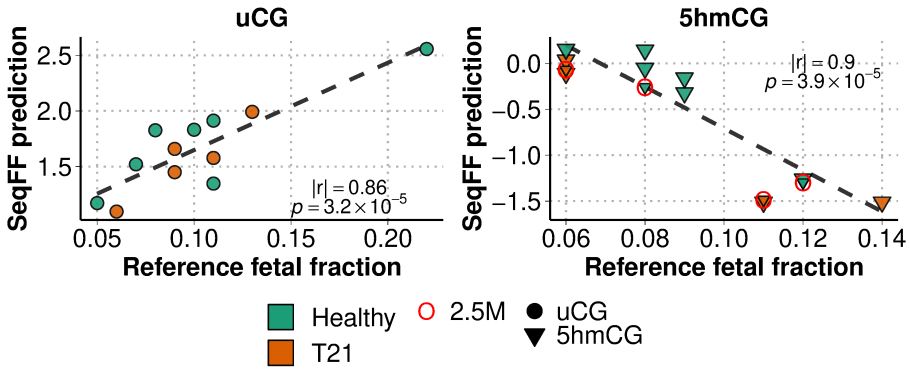


Figure 6.37 | Similarity Between the Reference Fetal Fraction and SeqFF Prediction

Correlation of the reference fetal fraction and SeqFF prediction from a uTOP–seq and hmTOP–seq data indicates the enrichment of cffDNA in maternal cfDNA mixture. The dashed line indicates linear regression. hmTOP–seq samples of shallow sequencing (on average 2.5 million raw reads) are indicated with red circles and were not used in the estimation of Pearson correlation.

As expected, there was a very high correlation between the reference and predicted fetal fraction in these samples prepared with a shallow sequencing depth (Pearson's $r = 0.95$, p–value $= 0.05$).

6.5.7 Discussion

6.5.7.1 New Insights Provided by This Study

To the best of our knowledge, this study is the first to demonstrate the covalent derivatisation and targeted sequencing of uCG sites in maternal cfDNA with the aim to detect fetal karyotype. We also showed that 5hmC profiling in maternal cfDNA can also accurately inform the fetal karyotype. The hmTOP–seq method, which covalently targets 5hmC residues, enabled the construction of genome–wide 5hmCG maps of relatively low 5hmCG levels in chorionic villus samples and cfDNA. Most importantly, hmTOP–seq was most discriminatory in the detection of T21 fetuses independently of the fetal fraction. Therefore, prenatal tests

based on 5hmCG analysis could potentially maximise the diagnostic sensitivity in relation to cost and be an optimal choice for sequencing–based epigenetic approaches of NIPT. Furthermore, fetal fraction can be measured directly from the read count of hmTOP–seq and uTOP–seq using a computational method that estimates the fetal fraction independent of fetal sex.

A large panel of placenta–specific uCG– and 5hmCG–biomarkers were identified and utilised for detection of fetal karyotype. To ascertain global methylation changes in T21 fetuses, the computation of DMRs specific for the T21–affected fetuses was also included. Interestingly, these DMRs better overlapped the healthy pregnancy–specific DMR sets than those of CV–specific DMRs, suggesting that DNA fragments of other tissue– origin than placenta might contribute to cffDNA. This points to a need for future comprehensive investigation of the tissue composition of maternal cfDNA. Analysis of the pregnancy–specific and T21–specific DMRs indicated the highly perturbed epigenome of T21–affected fetuses. Thus, disease–specific epigenetic characteristics should certainly be taken into account for the development of reliable NIPT of fetal aneuploidies, including T21.

6.5.7.2 Difficulties in Applying TOP–seq Based Methods

Differentially modified regions were identified in 100 bp non–overlapping bins. Such a naive approach sets a size limit of the expected modification region, which might miss larger scale epigenetic differences. In future applications, other effective supervised methods, such as the parsimonious temporal aggregation, could be used to capture more sensitive changes of DNA modifications (Gordevičius et al., 2009).

The observation that T21 DMRs overlapped more with the pregnancy–specific than the CV–specific DMRs is very compelling but there is another possible reason than non–trophoblast tissues contributing to fetal cfDNA. It could be an artefact of tissue type — the pregnancy specific

DMRs are also measured in cfDNA, like the T21 DMRs, while the CV–specific DMRs originate from a different tissue. A wider set of analysed tissues is needed to fully accept or reject these observations.

6.5.7.3 Unanswered Questions and Future Research Directions

Further validation of our findings in a large clinical cohort is necessary as this study is limited by sample size. Additionally, the study can be expanded to other common fetal aneuploidies such as Patau and Edwards syndromes. Another interesting application of TOP–seq and hmTOP–seq methods would be to interrogate imprinted regions in cfDNA analysis.

6.5.7.4 Concluding Remarks

This study using our previously developed TOP–seq and hmTOP–seq approaches obtained whole–genome uCG and 5hmCG maps of 10 CV tissue and 38 cfDNA samples in total. Our results indicated that such epigenomic analysis enriches fetal DNA fragments from maternal cfDNA, with both methods yielding 100% accuracy in detecting Down syndrome in fetuses. We identified 2,164 and 1,589 placenta–specific differentially modified and 5hmC modified regions, respectively, in chromosome 21. As well as 3,490 and 2,002 Down syndrome–specific differentially modified and 5hmC modified regions that can be used as biomarkers for the identification of Down syndrome or other fetal epigenetic diseases.

Statement VII — TOP–seq and hmTOP–seq methods can be used to generate single nucleotide epigenomic maps to decipher epigenetic differences across genetic elements between different sample groups.
Statement VIII — TOP–seq and hmTOP–seq methods could be used to identify fetal abnormalities in maternal cfDNA.

General Conclusions

The contributions of this research to science are summarised as follows:

- Developed computational methods to efficiently and accurately process TOP–seq based high–throughput epigenomic data. Created strategies that enable investigation of DNA modification signal at a single cytosine resolution in a strand specific manner.
- Developed statistical learning techniques to enhance the quality of the TOP–seq epigenomic signal. For a model IMR90 genome the applied statistical learning approaches increased Pearson's correlation estimate between technical replicates up to $r = 0.89$, while absolute Pearson's correlation estimate at a single CG site with a reference WGBS signal increased up to $r = 0.71$.
- TOP–seq based methods can provide information about different DNA modifications across various genomic elements and features.
- TOP–seq and hmTOP–seq methods could be used to identify differentially modified regions across samples pertaining to distinct sample groups. Both methods are able to employ CG site coverage and identification information to classify samples originating from different tissues or karyotypic groups. Identified 100 bp sized regions could be later used for prenatal diagnostics or to evaluate tissue composition within given sample.

Santrauka

8.1 Įvadas

8.1.1 Tyrimo Pagrindimas

Epigenetiniai kontrolės mechanizmai, tokie kaip DNR modifikacijos, atlieka svarbų vaidmenį praktiškai visuose gyvuose organizmuose reguliuojant įvairius ląstelinius, vystymosi ir elgsenos procesus. Nepaisant DNR modifikacijų svarbos biologijoje, daugybė sunkumų, susijusių su epigenetinių procesų identifikavimu ir apibūdinimu, atbaido tyrėjus nuo šių tyrimų. Egzistuoja dvi pagrindinės kliūtys plačiam epigenetinių procesų tyrimui. Pirma, epigenomo kiekybinio įvertinimo metodai, tokie kaip viso genomo bisulfitinė sekoskaita (angl., *WGBS*), yra labai brangūs ir sukuria didelį kiekį duomenų. Antra, šie metodai negali optimaliai atskirti skirtingų tipų DNR modifikacijų. Be to, nors WGBS ir yra plačiausiai naudojamas metodas bei priimamas kaip aukso standartas, jis kenčia nuo eksperimentinių artefaktų kaip didelė DNR degradacija.

TOP–seq (angl., *Tethered Oligonucleotide–Primed Sequencing*) yra pirmasis metodas galintis praturtinti nemodifikuotus citozinus vienos bazės skiriamaja geba išlaikant grandinės specifiškumą. hmTOP–seq yra vieno nukleotido skiriamosios gebos 5hmC profiliavimo metodas. Galiausiai, caC–Clearance (angl., *caCLEAR*) yra vieno nukleotido skiriamosios gebos metodas, leidžiantis tiksliai per visą genomą atvaizduoti 5caC modifikaciją. Šie nauji metodai gali padėti nustatyti genomo masto vieno nukleotido epigenetinius žymenis su mažiau išteklių nei viso genomo bisulfitinė sekoskaita. Tačiau norint išnaudoti visas šių metodų galimybes, reikia sukurti tinkamus statistinius ir kompiuterinius skaičiavimo metodus. Šis tyrimas pateikia siūlymus, kaip galima išspręsti uždavinius, kylančius dėl gana specifinių TOP–seq metodo duomenų. Be to, šiame

darbe taip pat pristatoma keletas TOP–seq metodo pritaikymų — diferenciškai modifikuotų regionų (angl., *DMR*) identifikavimas, epigenominių profilių sudarymas, signalo normalizavimas pagal genominį kontekstą.

Iš esmės čia pateiktas darbas susideda iš trijų pagrindinių dalių: *suprojektuoti*, *patobulinti*, *pritaikyti*. *Projektavimo* dalyje pristatome duomenų apdorojimo metodiką, kuri neapdorotus TOP—seq epigenominius duomenis paverčia CG padengimo signalu. *Patobulinimo* dalyje mes pasiūlome ir integruojame tris CG padengimo signalo transformacijas, kurios gali žymiai pagerinti praturtinimu pagrįstą DNR modifikacijų signalą. Galiausiai, *pritaikymo* dalyje pateikiame kelias metodų aplikacijas, kai sugeneruotas TOP–seq signalas gali būti naudojamas biologinei informacijai gauti ir interpretuoti.

8.1.2 Tikslas ir Uždaviniai

Pagrindinis šiame darbe aprašyto tyrimo tikslas buvo išplėtoti statistinius ir kompiuterinius įrankius, pritaikytus analizuoti TOP–seq metodu pagrįstus didelio našumo epigenominius duomenis, ir pritaikyti šias priemones eksperimentinėse aplinkose biologinėms žinioms įgyti. Šiam tikslui pasiekti buvo iškelti šie uždaviniai:

- Sukurti kompiuterinius metodus, kaip efektyviai ir tiksliai apdoroti TOP–seq sekoskaitos fragmentus.
- Sukurti statistinio mokymosi metodus, kad būtų pagerinta TOP–seq signalo kokybė esant techniniam ir biologiniam triukšmui.
- Taikyti sukurtus metodus ir technikas, kad būtų galima palyginti skirtingas genominių elementų DNR modifikacijas.
- Identifikuoti diferentiškai modifikuotus regionus, susijusius su skirtingomis eksperimentinėmis grupėmis, naudojant TOP–seq metodo didelio našumo epigenominius duomenis.

8.1.3 Ginamieji Teiginiai

- Sukurti kompiuteriniai metodai gali būti naudojami efektyviai ir tiksliai apdoroti TOP–seq metodo didelio našumo epigenominius duomenis.
- Sukurti statistinio mokymosi metodai gali būti naudojami siekiant pagerinti TOP–seq epigenominio signalo kokybę esant techniniam ir biologiniam triukšmui.
- TOP–seq, hmTOP–seq ir caCLEAR metodai suteikia informaciją apie DNR modifikacijas skirtinguose genomo elementuose.
- TOP–seq metodas gali būti naudojamas identifikuoti diferentiškai modifikuotus regionus tarp mėginių, kurie kilę iš skirtingų audinių ar ląstelių tipų.
- Norint nustatyti vaisiaus anomalijas motinos mėginių DNR, kylančioje ne iš ląstelių, gali būti naudojami TOP–seq ir hmTOP–seq metodai.

8.1.4 Mokslinis Naujumas ir Praktinė Vertė

Pagrindinis naujas aspektas šiame darbe yra statistinių ir kompiuterinių skaičiavimo metodų kūrimas ir taikymas DNR modifikacijų analizei TOP–seq, hmTOP–seq ir caCLEAR gautose didelio našumo epigenominiuose duomenų rinkiniuose. Šiame darbe pateikiama išsami sekoskaitos fragmentų apdorojimo metodika, sukurta specialiai TOP–seq sekoskaitos duomenims. Be to, mes pateikiame sekoskaitos bibliotekos kokybės parametrus, tokius kaip fragmentų ilgis, CG tankis, atstumas iki CG pozicijos genome.

Šiame moksliniame darbe taip pat aprašomas naujas sekoskaitos duomenų transformacijos metodas — u–density. u–density pagerina genomo padengimo signalo tikslumą, pasitelkdamas DNR modifikacijos informaciją iš kaimyninių CG vietų ir normalizuodamas signalą pagal CG tankį. Be to, šiame darbe taip pat buvo sukurti ir pritaikyti du statistiniu apsimokymu paremti

DNR modifikacijų signalo transformacijos metodai. Šie metodai naudoja TOP–seq duomenis ir genomo konteksto informaciją, kad įvertintų DNR modifikacijos lygius. Nedidelė WGBS rinkinio dalis buvo panaudota eksponentiniam ar dirbtinio neuroninio tinklo modeliui apmokyti, kuris tada buvo naudojamas transformuoti TOP–seq signalą į vadinamąjį CG metilinimo lygio signalą. Šie patobulinimai transformavo realiatyvų padengimo signalą į absoliučių verčių skalę, o tai labai padidino koreliaciją su referentiniu duomenų rinkiniu ir leido lengviau interpretuoti signalą.

Šiame darbe pateikiamas pirmasis išsamus vaisiaus nemodifikuotų ir hmC modifikuotų CG vietų tyrimas motinos kraujo mėginiuose, skirtas neinvaziniam prenataliniam tyrimui (angl., *NIPT*). Pirmą kartą mes ištyrėme nemodifikuotą chorioninių gaurelių mėginių DNR frakciją ir palyginome ją su ląstelių neturinčia vaisiaus DNR (angl., *cffDNA*). Be to, metodika nustatyti DMR buvo pristatyta kaip perspektyvi strategija vaisiaus kariotipams aptikti. Šis viso genomo mastu atliktas TOP–seq DNR modifikacijų profiliavimas, vykdytas sveikuose ir 21 chromosomos trisomijos teigiamuose vaisiuose, leido nustatyti naujus biologinius žymenis, turinčius diagnostinę vertę. Šiame tyrime gauti diferentiškai modifikuoti regionai gali padėti parinkti tinkamus diagnostinius žymenis tam tikrame klinikiniame kontekste. Tikimasi, kad šis metodas netgi gali pranokti šiuo metu naudojamus NIPT testus.

8.2 TOP–seq Duomenų Apdorojimas

8.2.1 Įvadas

Šiame skyriuje pateikiama metodika, sukurta apdoroti TOP–seq sekoskaitos duomenis. Ši metodika yra modifikuota plačiai naudojamų bioinformatikos metodikų versija ir gali būti taikoma ne tik nemodifikuotų CG dinukleotidų duomenų analizei, bet ir kitiems variantams (t.y., hmTOP—seq — 5hmC modifikacijai, caCLEAR — 5caC modifikacijai). Šiame

skyriuje dažniausiai vartojamas terminas TOP–seq, tačiau svarbu paminėti, kad duomenų analizės procedūros yra taikomos ir kitiems metodo variantams.

TOP–seq sekoskaitos duomenų apdorojimo eigą sudaro keturi pagrindiniai žingsniai:

- Sekoskaitos fragmentų apdorojimas
- Fragmentų prilyginimas prie referentinio genomo
- PGR duplikatų pašalinimas
- Fragmentų priskyrimas CG dinukleotidams

Skirtingam DNR šaltiniui (pvz., lambda bakteriofagui, eukariotinei ląstelei) arba skirtingoms DNR modifikacijoms (pvz., nemodifikuotoms CG vietoms, 5hmCG vietoms) gali reikėti kitokios TOP–seq duomenų apdorojimo metodikos. Svarbu paminėti, kad TOP–seq sekoskaitos metodas gali sugeneruoti dešimtis milijonų fragmentų, kurių analizei reikalingi intensyvūs skaičiavimo procesai ir infrastruktūra. Šis skyrius suskirstytas į penkias pagrindines dalis, keturiuose skyriuose pateikiamas atskiras darbo eigos etapas, o paskutiniame skyriuje aptariami pranašumai ir trūkumai, rekomendacijos ir idėjos, kurias būtų galima įgyvendinti ateityje siekiant pagerinti TOP–seq metodo duomenų analizę.

8.2.2 Sekoskaitos Fragmentų Apdorojimas

Neapdoroti sekoskaitos fragmentai gaunami iš naujos kartos sekoskaitos aparato yra `FASTQ` failo formato, kuriame yra (kiekvienam fragmentui) unikalus identifikatorius, nukleotidų seka ir kiekvieno nukleotido Phred kokybės balas. Šių neapdorotų fragmentų procesavimas vyksta keturiais etapais, kurie pavaizduoti **Figure 4.1** schemoje. Pirmasis žingsnis naudoja komandą `fastq_quality_trimmer` (įdiegtą `FASTX-Toolkit`), kad pašalintų per trumpus sekoskaitos fragmentus. Šis žingsnis yra greičio

optimizavimas tolesniems etapams, nes per trumpuose fragmentuose paprastai nėra 5' ir (arba) 3' sekos adapterių, todėl jų negalima klasifikuoti kaip tinkamų analizei. Kaip filtravimo ribą naudojome 80 nukleotidų, tačiau šis ilgio parametras įvairiuose eksperimentuose gali skirtis.

Tada `cutadapt` programa buvo naudojama pašalinti ar apkarpyti adapterių sekas iš 5' ir 3' sekoskaitos fragmentų galų. Galiausiai komanda `fastq_quality_trimmer` naudojama siekiant pagerinti fragmentų kokybę. Paprastai 3' fragmento galas turi mažesnį Phred kokybės balą, kuris gali sukelti klaidingą prilyginimą prie referentinio genomo dėl nukleotidų neatitikimo tarp konkretaus fragmento ir referentinės sekos. `FASTX-Toolkit` buvo naudojamas apkarpyti fragmento galus, kurių Phred kokybės balas buvo mažesnis nei 20. Be to, pašalinus adapterio sekas ir žemos kokybės nukleotidus, `FASTX-Toolkit` pašalino fragmentus, kurie buvo per trumpi prilyginimui (naudota 15 nukleotidų kaip slenksčio parametras). Šis žingsnis užtikrina greitesnį prilyginimą ir aukštesnį prilyginimo įvertį. Prilyginimo kokybės įvertis prieš ir po kirpimo pagal tam tikrą ribą yra pavaizduotas **Supp. Figure 1**. Po kiekvieno apdorojimo etapo kokybės ataskaitai sukurti buvo naudojama `FastQC` programa (su numatytaisiais parametrais). Ši, atrodytų, nereikalinga procedūra užtikrina, kad adapterių sekos būtų pašalintos teisingai, Phred balas būtų pakankamai aukštas, o TOP–seq bibliotekos fragmentai prasidėtų CG dinukleotidu.

8.2.3 Fragmentų Prilyginimas prie Referentinio Genomo

Suprocesuotus TOP–seq sekoskaitos fragmentus galima prilyginti prie referentinio genomo sekos, naudojant standartinius algoritmus ir įrankius (pvz., `bwa mem` ar `bwa aln`). Standartinei TOP–seq analizei buvo naudojama `bwa mem` komanda su numatytaisiais parametrais (išskyrus `idxbase` parametrą — referentinio genomo seka, kuri priklausė nuo eksperimento). `samtools` įrankis buvo naudojamas konvertuojant `bwa` programos sugeneruota palyginio `SAM` failo formatą į `BAM` failo formatą. `BAM` failas yra atitinkamai surūšiuotas ir atrinktas fragmentams, kurių prilyginimo kokybė yra lygi arba didesnė nei 30 (`samtools sort` ir `samtools`

`view` komandos). **Figure 4.4** paveiksle pavaizduotas vieno mėginio prilyginimo kokybės pasiskirstymas. Daugumoje eksperimentų prilyginimo kokybės pasiskirstymas yra bimodalinis (pirmojo moda ties žemiausia prilyginimo kokybe, o antroji — aukščiausia prilyginimo kokybe), todėl buvo nuspręsta šį pasiskirstymą padalyti į dvi dalis (atskaitos taškas — 30) ir naudoti visus fragmentus, kurie patenka į dešinę pasiskirstymo pusę.

8.2.4 PGR Duplikatų Pašalinimas

PGR duplikatai atsiranda, kai suskaidyti DNR fragmentai padauginami PGR metodu. Tokiu atveju tas pats DNR fragmentas bus amplifikuotas ir sekvenuojamas kelis kartus. Šie identiški fragmentai užims vietą sekvenavimo bibliotekoje. Be to, kai sekoskaitos padengimo gylis yra svarbus veiksnys (pvz., TOP–seq metodas), PGR duplikatai gali trukdyti nustatyti tikrąjį DNR molekulių kiekį ir klaidingai amplifikuoti padengimo lygį. Pagal daugumą sekoskaitos metodikų rekomenduojama pažymėti ir pašalinti PGR kopijas naudojant unikalius molekulinius identifikatorius arba skaičiavimo įrankius, tokius kaip `Picard` ar `samtools`.

Standartiniai įrankiai randa PGR duplikatų kopijas, nustatydami fragmentų grupes, kurios priskiriamos toms pačioms genomo pradžios ir pabaigos koordinatėms, darant prielaidą, kad tikimybė, jog fragmentai susilygins su ta pačia padėtimi, yra labai maža (iš tikrųjų, bent jau žmogaus genomo atveju, ji yra arti nulio). Tačiau tokia identifikavimo strategija negali būti taikoma metodams, pagrįstiems TOP–seq, nes TOP–seq metodai yra orientuoti į specifines genomo pozicijas (t.y., CG dinukleotidus). Todėl sekvenavimo fragmentai daugumoje atvejų prasideda tose pačiose genomo koordinatėse, taigi įprastiniai skaičiavimo įrankiai šiuo atveju neveiktų. Mes sukūrėme PGR duplikatų identifikavimo ir pašalinimo algoritmą, panašų į kanoninius, tačiau jis nėra toks reiklus. Mūsų PGR duplikatų identifikavimo algoritme visi fragmentai, prasidedantys tiksliai ta pačia genomo koordinate, toje pačioje grandinėje ir turintys tą patį pradinį ilgį, buvo klasifikuojami kaip PGR kopijos ir kiekvienoje

tokioje grupėje buvo paliktas tik vienintelis sekoskaitos fragmentas. Ši strategija yra panaši į kanoninį metodą, nes vertinama fragmento pradinė padėtis (5' galas), tačiau, įvertinant pradinį fragmento ilgį, šiame algoritme atsižvelgiama į 5' ir 3' adapterių ilgius. DNR polimerazė yra linkusi per anksti sustabdyti 3' adapterio sintezę ir kartais praleisti nukleotidus 5' adapteryje, ir toks pokytis bus būdingas konkrečiai PGR kopijų grupei. **Figure 4.5** paveiksle parodyta 5' ir 3' adapterių ilgių pasiskirtymai, kurie sukuria erdvę klasifikuoti naujas PGR duplikatų kopijų grupes. Apibendrinant galima pasakyti, kad užuot pašalinęs visus fragmentus, turinčius identiškas prilyginimo koordinates (išskyrus tą, kuris paliekamas kaip grupę reprezentuojantis fragmentas), šis algoritmas palieka $m \times n$ fragmentus, kur m ir n yra atitinkamai kiekiai skirtingų 5' ir 3' adapterių ilgių.

8.2.5 Fragmentų Priskyrimas CG Dinukleotidams

Kiekvienam suprocesuotam ir atrinktam sekoskaitos fragmentui apskaičiavome atstumą nuo jo pradinės padėties iki artimiausio CG dinukleotido (t.y., atstumą, išmatuotą nukleotidais nuo fragmento 5' galo). Priklausomai nuo modifikacijos tipo, buvo pasirinktos skirtingos atstumo ribos, kad priskirtume fragmentus CG vietoms. TOP–seq metodui buvo naudojamas absoliutus trijų nukleotidų atstumas. Su šia ribą vidutiniškai 90 proc. fragmentų buvo priskirti CG vietoms (**Supp. Table 1**). Tuo tarpu hmTOP–seq ir caCLEAR metodams buvo naudojamas absoliutus keturių nukleotidų atstumas (vidutiniškai atsirenkant 85 proc. fragmentų).

Priskyrus sekoskaitos fragmentus CG vietoms, buvo apskaičiuota CG padengimas (apibrėžiant padengimą kaip bendrą fragmentų skaičių bet kurioje grandinėje, pradedant nuo nurodyto atstumo ribos). Tokia procedūra paprastai suskirsto CG vietas į dvi priešingas grupes (CG vietos, kurių padengimas yra didesnis nei 0, – identifikuotos CG vietos, ir CG vietos, kurių padengimas yra lygus 0 – neidentifikuotos CG vietos). Priklausomai nuo eksperimento, buvo naudojamos visos arba tik dalis identifikuotų

CG vietų.

8.2.6 Diskusija

8.2.6.1 Šios Metodikos Taikymai

Šiame skyriuje apibendrinta TOP–seq sekoskaitos duomenų apdorojimo metodika. Ši metodika suteikia galimybę naudoti TOP–seq ar juo pagrįstus kitus sekoskaitos metodus ir neapdorotus sekoskaitos fragmentus paversti į DNR modifikacijų signalą vieno CG skiriamojoje geboje.

Sekoskaitos fragmentų procesavimo eiga pateikta išsamiai – apibendrinta sekoskaitos fragmentų apdorojimas prieš sulyginimą su referentine seka, sulyginimas su referentine seka, PGR duplikatų kopijų pašalinimas ir prieš priskirimas CG taikiniams. Svarbu paminėti, kad PGR kopijų šalinimo algoritmas buvo specialiai sukurtas TOP–seq metodui, nes standartiniai duplikatų šalinimo būdai šiam metodui negali būti taikomi ir iškraipytų tikrąjį padengimo signalą. Galiausiai, šioje metodikoje yra pateikti keli greičio optimizavimo veiksmai, pavyzdžiui, pašalinant gana trumpus sekoskaitos fragmentus procesavimo pradžioje, siekiant sumažinti žemos kokybės fragmentų apdorojimo laiką.

8.2.6.2 TOP–seq Duomenų Apdorojimo Sunkumai

Daugiausia sunkumų procesuojant TOP–seq duomenis kyla dėl daugybės skaičiavimo įrankių, naudojamų šioje metodikoje. Kadangi ši metodika susideda iš kelių žingsnių, teoriškai ji galėtų būti paralelizuota. Tačiau ne visi naudojami įrankiai pritaikyti paraleliam skaičiavimui. Pavyzdžiui, įrankiai, naudojami adapterių kirpimui, negali būti paralelizuoti ir gali tik nuosekliai apdoroti sekoskaitos fragmentus.

Be to, viena savita problema kyla dėl fragmentų priskyrimo CG taikiniams. Gali būti, kad konkretus sekoskaitos fragmentas prasideda nuo

genomominės pozicijos, kurios absoliutus atstumas iki CG taikinio yra didesnis nei nulis. Paprastai tokių fragmentų priskyrimas jų pradinei CG vietai yra nesudėtingas, tačiau gali būti, kad su tokiu pat atstumu yra dar vienas CG taikinys ir tokiu atveju unikalus priskyrimas yra neįmanomas.

8.2.6.3 Baigiamosios Pastabos

Čia mes pateikiame išsamią naujos kartos sekoskaitos duomenų apdorojimo metodiką, kuri buvo sukurta apdoroti TOP–seq metodo epigenominius duomenis. Šis efektyvus ir lankstus metodas sujungia skirtingas idėjas iš jau žinomų metodikų į vieną išsamią procesavimo schemą. Tai apima veiksmus ir įrankius, naudojamus apdoroti TOP–seq signalą nuo neapdorotų sekoskaitos duomenų iki CG padengimo profilių.

8.3 Statistiniai Įrankiai, Skirti Pagerinti TOP–seq Signalo Kokybę

8.3.1 Įvadas

Didelio našumo genominiai metodai, gali būti įtakoti įvairiais šališko matavimo poveikiais, kurie gali būti išreikšti kaip intra–mėginio arba inter–mėginių variaciją. Tokiais atvejais TOP–seq signalas taip pat gali turėti nepageidaujamų CG padengimų variacijų tarp tos paties ar kelių skirtingų sekoskaitos gardelių. Tokius pokyčius gali lemti skirtingas sekoskaitos gylis, vidutinio modifikacijos lygio skirtumai ar kiti nežinomi biologiniai ar technologiniai veiksniai. Taigi, siekiant sumažinti tokį variacijos efektą ir pagerinti signalo kokybę, buvo implementuotos trys TOP–seq signalo transformacijos. Pirmoji ir pagrindinė transformacija yra u–density, pagrįsta svertiniu padengimo lygiu, normalizuotu pagal CG lygį; m–estimate ir nn–estimate yra TOP–seq padengimo signalo

projekcijos, apskaičiuotos naudojant eksponentinio modelį arba dirbtinius neuroninius tinklus. Šį skyrių sudaro trys pagrindiniai poskyriai (po vieną kiekvienai transformacijai): *u*–density, *m*–estimate, *nn*–estimate, o kiekvienas poskyris susideda iš trijų dalių: įverčio skaičiavimo pagrindimas, įverčio apskaičiavimo algoritmas ir įverčio palyginimas su referentiniu metodu. Galiausiai skyrius baigiamas diskusija apie šių transformacijų atliktus patobulinimus, kurie prisidėjo prie TOP–seq signalo pritaikomumo, apie sunkumus skaičiuojant ir pritaikant transformacijas bei apie neatsakytus klausimus būsimiems tyrimams.

8.3.2 *u*–density

8.3.2.1 Motyvacija Apskaičiuojant *u*–density Signalą

Kadangi TOP–seq metodui gali turėti įtakos apribojimai, su kuriais susiduria ir kiti praturtinimu pagrįsti metodai (pvz., signalo kokybės priklausomybė nuo sekoskaitos gylio, šališkumas konkretaus sekos konteksto atžvilgiu), TOP–seq signalui buvo pritaikytos statistinės korekcijos. Pirma, sekoskaitos gylio įtaka buvo sumažinta konvertuojant TOP–seq padengimo signalą į svertinio tankio įverčius. Toks konvertavimas išlygino signalo stiprumą tarp skirtingų sekoskaitos gylio eksperimentų. Šis svertinio tankio metodas leidžia mums panaudoti informaciją iš šalia esančių CG vietų ir tokiu būdu maksimaliai padidinti mažo padengimo regionų panaudojamumą. Norėdami pašalinti galimą sekos konteksto (t.y., CG) šališkumą, mes papildomai normalizavome apskaičiuotus svertinio tankio įverčius pagal nesvertinį CG tankį. Gautas signalas buvo pavadintas *u*–density, nes atspindėjo nemetilintos DNR tankį.

8.3.2.2 *u*–density Apskaičiavimo Algoritmo Santrauka

TOP–seq padengimo svertinio tankio įverčiai buvo apskaičiuoti naudojant Epanechnikov kernelį, per 2^{21} taškų, tolygiai paskirstytų kiekvienoje

chromosomoje (skaičiavimo darbo eiga pavaizduota **Figure 5.2** paveiksle). Fragmentų skaičius buvo normalizuotas taip, kad kiekvienoje chromosomoje jis būtų lygus 1. Panašus metodas buvo naudojamas įvertinant ir nesvertinį CG tankį konkrečioje chromosomoje. Galiausiai, TOP–seq nemetilinimo tankis buvo gautas dalijant svertinį TOP–seq tankį iš nesvertinio CG–tankio. Normalizavus svertinį tankį pagal CG tankį, Gauss kernelis buvo panaudotas interpeliuoti galutinį signalą ties CG pozicijomis. Kernelio parametrai buvo nustatyti įvertinant TOP–seq *u*–density koreliacijas plačiame kernelio langų diapazone su atitinkamu IMR90 WGBS signalu žmogaus 1 chromosomoje (**Figure 5.3**). Įvertinus koreliacijas, esant vienai CG skiriamajai gebai, pasirinkti kernelio parametrai buvo: 180 bp svertiniam tankiui ir 80 bp CG–tankiui (vėliau tie patys parametrai buvo naudojami visiems mėginiams ir visoms chromosomoms).

Apskaičiavus svertinį tankį, buvo pastebėtas tolygesnis signalo pasiskirstymas mėginiuose. Padengimo statistika (pvz., vidutinis padengimas) kiekvienam mėginiui gerai koreliuoja su sekoskaitos gyliu. Toks efektas gali lengvai paveikti rezultatų interpretavimą, nes didesnių bibliotekų dydžių mėginiai turėtų didesnį nemodifikacijos signalą. Tačiau jeigu kitame eksperimente būtų naudojama mažiau ar daugiau sekoskaitos fragmentų, šie du rezultatai nebūtų palyginami. Stebima padengimo statistika gerai koreliuoja (Pearson $r = 0.85$) su fragmentų mėginių skaičiumi, tačiau ši koreliacija sumažėjo, kai vietoj to buvo naudojamos svertinio tankio vertės (Pearson $r = 0.46$) (**Figure 5.4**).

Apskaičiavus *u*–density, pastebėtas koreliacijos tarp techninių pasikartojimų padidėjimas (Pearson $r = 0.5$ ir $r = 0.8$ prieš ir po transformacijos). Galiausiai pastebėtas labai pagerėjęs signalo pasiskirstymas tarp skirtingų genominių elementų. Kadangi TOP–seq metodas yra nukreiptas į CG pozicijas, CG turinčiuose regionuose gali būti tikimasi stipresnio signalo vien tik todėl, kad ten yra daugiau taikinių, todėl toliau mes įvertinome ryšį tarp dviejų matavimų skirtinguose genomo elementuose (**Figure 5.5**). Didesnio svertinio tankio signalas buvo pastebėtas elementuose, turinčiuose didesnį CG tankį, tačiau šis nukrypimas buvo ištaisytas normalizavus TOP–seq signalą pagal CG kiekį kiekviename regione. Norint

toliau patvirtinti CG tankio normalizavimo efektą, buvo tirtas specifinis genomo lokusas, turintis netolygų CG tankio pasiskirtymą (**Figure 5.6**). TOP–seq signalas palei *KAZN* geno lokusą rodo didelį signalo poslinkį link CG turinčių regionų (t.y., CGI elementų), tačiau po CG tankio normalizavimo signalo smailės nebėra centruotos į CGI elementus, o *u*–density signalas palaipsniui mažėja link geno pabaigos.

8.3.2.3 *u*–density ir kitų metodų atitikimas

Vieno nukleotido skiriamosios gebos koreliacijos tarp TOP–seq ir WGBS signalo patvirtino signalo transformacijos naudingumą. Pearson koreliacija tarp CG padengimo ir WGBS buvo $\mid r \mid = 0.23$, $\mid r \mid = 0.36$, $\mid r \mid = 0.44$ Smegenų, IMR90 mažo bibliotekos gylio ir IMR90 didelio bibliotekos gylio mėginiams, atitinkamai. Naudojant *u*–density, ši koreliacija padidėjo iki $\mid r \mid = 0.28$, $\mid r \mid = 0.59$, $\mid r \mid = 0.64$, atitinkamai. Palyginimui, vienas CG skiriamosios gebos IMR90 WGBS signalas buvo lyginamas su MRE–seq ir MBD–seq DNR modifikacijos signalais, o apskaičiuotos Pearson koreliacijos buvo $\mid r \mid = 0.18$, atitinkamai.

8.3.3 *m*–estimate

8.3.3.1 Motyvacija Apskaičiuojant *m*-estimate Signalą

Apskaičiavę *u*–density vertes, nusprendėme toliau patobulinti TOP–seq signalą. Kadangi TOP–seq yra praturtinimu pagrįstas signalas, jo vertės pasiskirsto išilgai sunkiauodegiame Poisson skirstinyje (t.y., padengimo arba *u*–density vertės yra nuo 0 iki plius begalybės), daugumai vietų negaunant jokios arba labai mažas teigiamas vertes, tuo tarpu WGBS metodas gali išmatuoti tą patį modifikacijos signalą absoliučioje skalėje (nuo 0 proc. iki 100 proc.). Todėl viena iš korekcijų buvo signalo konversija iš realiatyvios į absoliučią skalę. Kitas patobulinimas buvo susijęs su skirtingais vidutiniais modifikacijos lygiais skirtinguose epigenomuose. **Figure 5.7** paveiksle pavaizduoti du susimuliuoti epigenomai su

skirtingais vidutiniais modifikacijų lygiais, kurie priklauso nuo padėties išilgai susimuliuotos chromosomos. Atlikę atsitiktinio padengimo pasiskirstymą, mes pastebėjome, kad epigenomas su didesniu vidutiniu hipometilinimo lygiu gauna mažesnį padengimą nei epigenomas su mažesniu hipometilinimo lygiu. Šie rezultatai rodo, kad epigenomai su skirtingais modifikacijos lygiais nėra sulyginami, nes pastebėti padengimo pasiskirstymai yra netikslūs.

8.3.3.2 *m*–estimate Apskaičiavimo Algoritmo Santrauka

Metilinimo įverčiai, *m*–estimates, buvo gauti vystant eksponentinio irimo (angl., *exponential decay*) modelį. 20 chromosoma (2.5 proc. visų žmogaus genomo CG vietų) buvo naudojama eksponentiniam modeliui apmokyti. Taip pat buvo naudojamos papildomos genominėms kovariantės (t.y., genomo elementų informaciją) padidinti modelio tikslumą. Kovariančių vertės buvo apskaičiuotos kiekvienam CG taikiniui, naudojant 50 bp regionus aplink kiekvieną CG.

Pasirinktos kovariantės buvo šios: i) GC tankis – guanino ir citozino bazių procentas regione; ii) CG dinukleotidų dalis tarp CN porų tam tikrame regione; iii) vidutinė sekos prilyginimo vertė regione; iv) tam tikro regiono dalis, persidengianti su SINE arba LTR pasikartojimais; v) tam tikro regiono dalis, persidengianti baltymus koduojančių genų promotorius; vi) tam tikro regiono dalis, persidengianti baltymus koduojančių genų 5'UTR; vii) tam tikro regiono dalis, persidengianti intergeninius regionus.

8.3.3.3 *m*–estimate ir Kitų Metodų Atitikimas

Apskaičiavus *m*–estimate reikšmes, koreliacija tarp TOP–seq techninių pasikartojimų pakito nežymiai. Pearson koreliacija padidėjo tik iki $r = 0.89$ didelio gylio IMR90 mėginiams (nuo *u*–density $r = 0.87$). Tačiau

apskaičiavus m–estimate, labai pagerėjo TOP–seq signal koreliacija su WGBS signalu. Apskaičiuotos Pearson r vertės padidėjo iki 0.69.

8.3.4 nn–estimate

8.3.4.1 Motyvacija Apskaičiuojant nn–estimate Signalą

Apskaičiavę m–estimate signalą, mes išsprendėme pagrindinius klausimus, su kuriais susiduria praturtinimo metodai, tačiau vis tiek matėme galimybę patobulinti TOP–seq metodą, pvz., koreliacijas tarp TOP–seq ir WGBS signalų. Kadangi WGBS yra laikomas aukso standarto metodu, mes bandėme priartinti TOP–seq signalą kuo arčiau jo. Mes siekėme apskaičiuoti nn–estimate — metilinimo įverčius, apskaičiuotus naudojant dirbtinius neuroninius tinklus. Neuroniniu tinklu pagrįstas metodas teoriškai galėtu būti geresnis nei eksponentinio irimo modelis, nes jis sugebėtų panaudoti nežinomas asociacijas duomenyse. Tokio metodo trūkumas yra tas, kad neuroninis tinklas ir jo gautas rezultatas yra *juodosios dėžės* fenomenas ir jo sukurtos duomenų asociacijos paprastai lieka nežinomos.

8.3.4.2 nn–estimate Apskaičiavimo Algoritmo Santrauka

Nustatyti IMR90 WGBS vertes buvo panaudotas daugiasluoksnis neuroninis tinklas su 2 paslėptais sluoksniais (atitinkamai 44 ir 22 mazgai), naudojant TOP–seq signalą ir įvairias genomo characteristikas iš 20 chromosomos (**Supp. Figure 4**). Svarbiausios perceptrono ypatybės buvo u–density ir TOP–seq padengimas (santykinė svarba atitinkamai 6.4 proc. ir 4.4 proc.). Visos kitos naudojamos ypatybės buvo suskirstytos į tris grupes – genomo elementus, sekos charakteristikas ir bazių sudėtį aplink konkrečią CG vietą (**Figure 5.8**). Genominiai elementai, turintys didžiausią santykinę vertę, buvo CGI ir SINE pasikartojimai. Svarbiausios sekos charakteristikų grupės ypatybės buvo GC dinukleotidų

kiekis tam tikrame regione, CG dinukleotidų kiekis ir genomo prilyginomo įvertis balas. Dauguma bazių sudėties grupės ypatybių parodė tik vidutines santykinės svarbos vertes.

8.3.4.3 *nn*–estimate ir Kitų Metodų Atitikimas

nn–estimate verčių apskaičiavimas turėjo nedidelę įtaką koreliacijai tarp TOP–seq techninių replikų. Pearson koreliacija didelio gylio IMR90 replikose padidėjo iki $r = 0.89$. Tačiau *nn*–estimate skaičiavimas pagerino koreliaciją su WGBS signalu — $r = 0.71$. Didesnio masto genomo regionuose WGBS ir *nn*–estimate panašumas buvo dar didesnis. **Figure 5.9** paveiksle pavaizduota koreliacija tarp referentinio IMR90 WGBS signalo ir TOP–seq DNR modifikacijos signalo baltymus koduojančiuose genų promotoriuose. Koreliacija tarp TOP–seq signalo palaipsniui naudojant signalo transformaciją. Panašumas išlieka didelis net naudojant kitus IMR90 WGBS duomenų rinkinius. Galiausiai mes suskirstėme CGI elementus į DNR metilinimo grupes pagal referentinį IMR90 WGBS signalą. Įvertinę *nn*–estimate signalą tam tikrose CGI grupėse, mes pastebėjome aukštą WGBS ir TOP–seq metodų sutaptį. Toliau apskaičiavome *nn*–estimate vertes iš susimuliuotų bibliotekų, turinčių dešimt kartų mažesnius ar didesnius bibliotekų dydžius. Įdomu tai, jog sudarytas *nn*–estimate modelis gerai veikia mažesnio dydžio bibliotekų atžvilgiu, tačiau didesnio dydžio bibliotekose apskaičiuotos *nn*–estimate vertės nebegali gerai atskirti nurodytų CGI grupių (**Figure 5.10**).

8.3.5 Diskusija

8.3.5.1 Šios Metodikos Taikymai

Šiame skyriuje mes pristatėme tris efektyvius signalo transformavimo metodus, naudojamus TOP–seq signalo kokybei gerinti. Žemas TOP–seq sekoskaitos padengimas gali sukelti DNR modifikavimo profiliavimo

iššūkių. Dėl nepadengtų CG vietų sekoskaitos duomenys gali būti negausūs, o CG vietose bus rodomas perdėtas nulinių verčių skaičius, o tai sukurs klaidingą signalą. Visos trys sukurtos transformacijos — *u*–density, *m*–estimate ir *nn*–estimate padėtų išspręsti šią problemą įtraukiant genominę ir epigenominę informaciją iš kaimyninių lokusų. Šie metodai praplečia TOP–seq signalo naudojimą: i) sumažinimas sekoskaitos bibliotekos dydžio variacijos įtaka; ii) kompensuojamas mažesnis sekoskaitos gylis; iii) normalizuojamas signalo intensyvumas pagal genomo foną. Kiekviename konkrečios transformacijos poskyryje pateikiami jos skaičiavimo argumentai, signalo transformavimo algoritmas ir palyginimas su referentiniu signalu. Visų pateiktų transformacijų palyginimas yra geresnis referentinių duomenų atžvilgiu, palyginti su pirminiu TOP–seq signalu. Tai labai perspektyvu, nes naudojant didesnius ir įvairesnius referentinių duomenų rinkinius mūsų modelių našumas ir pritaikomumas gali dar labiau padidėti.

8.3.5.2 Statistinių priemonių kūrimo sunkumai, siekiant pagerinti TOP–seq signalo kokybę

Didžiausias sunkumas apskaičiuojant *u*–density signalą yra išlaikyti pusiausvyrą tarp vieno CG skiriamosios gebos ir naudoti padengimo informaciją iš kaimyninių CG vietų. Praktiškai *u*–density skaičiavimą galima suskirstyti į du perėjimus tarp skiriamųjų gebų: i) projekcija iš vieno CG į kaimyninių CG pozicijų vektorių, siekiant įtraukti jų padengimą į vertės skaičiavimą, ir ii) projekcija iš įverčio vektoriaus atgal į vieno CG skiriamąją gebą. Todėl vienas iš svarbiausių momentų yra atitinkamų langų dydžių pasirinkimas. Viena vertus, šios vertės gali būti per didelės ir sumažės matavimų tikslumas, tuo tarpu naudojant per mažas vertes, rezultatai nebus optimalūs, o minėti perėjimai tarp skiriamųjų gebų tik praras pirminio signalo kokybę.

Pagrindinis *m*–estimate ir *nn*–estimate trūkumas yra priklausomybė nuo duomenų rinkinio, naudojamo vykdant modelio apmokymą. Kai tokio duomenų rinkinio nėra arba jo kokybė nėra pakankamai patenkinama,

šios signalo transformacijos nebus įmanomos. Tačiau šiame genominių tyrimų amžiuje yra pakankamai daug epigenominių duomenų, o daugelis žmogaus audinių ar net ląstelių tipų tiriami naudojant WGBS technologijas. Jeigu nėra tikslaus audinių tipo, gali būti įmanoma surinkti ir panaudoti papildomą referentinį audinių rinkinį iš panašių audinių, tačiau tokį ekstrapoliavimą reikėtų kruopščiai išbandyti prieš jį taikant.

Galiausiai, didžiausia problema, su kuria gali susidurti tyrėjai, yra sprendimas, kurią transformaciją naudoti. Jeigu sekoskaitos gylis yra pakankamai didelis, gali būti patenkinamas paprastas CG pozicijos padengimas. Atsižvelgiant į mažą sekoskaitos gylį, galima norėti naudoti transformacijas, tačiau sprendimą, ar naudoti *u*–density, ar apmokymu pagrįstas transformacijas, reikia priimti kiekvienu atveju atskirai, atsižvelgiant į turimus išteklius, biudžetą, skaičiavimo infrastruktūrą, mėginių prieinamumą ir kt.

8.3.5.3 Baigiamosios Pastabos

Šiame skyriuje buvo pasiūlyta panaudoti TOP–seq duomenis ir genomo konteksto informaciją, siekiant įvertinti DNR modifikacijos lygius. Buvo suintegruotos trys signalo transformacijos — *u*–density, *m*–estimate ir *nn*–estimate. *u*–density grindžiamas svertiniu kernelio CG padengimo įvertinimu, normalizuotu pagal nesvertinį CG tankį; *m*–estimate ir *nn*–estimate yra apmokymu pagrįsti metodai, sukurti naudojant eksponentinio irimo modelį arba daugiasluoksnį neuroninį tinklą, naudojant ypatybes kaip TOP–seq signalą, genomo seką ir genomo konteksto informaciją. Mūsų rezultatai rodo, kad norint atlikti ekonomiškai efektyvius populiacinius DNR modifikavimo tyrimus, būtina derinti pažangius skaičiavimo metodus su naujomis sekoskaitos technologijomis.

8.4 TOP–seq Pagrįstų Metodų Taikymas

8.4.1 Įvadas

Šį skyrių sudaro keturi pagrindiniai poskyriai — didelio našumo epigenomo profiliavimo metodų taikymas naudojant TOP–seq. Pirmame poskyryje pateikiami nemodifikuotų CG profiliavimo žmogaus audiniuose ir ląstelių linijose rezultatai, o antrame ir trečiame poskyriuose trumpai apibendrinama hmTOP–seq ir caCLEAR metodų kokybės kontrolė pelių embrioninėse kamieninėse ląstelėse. Galiausiai paskutiniame poskyriuje pateikiamas išsami TOP–seq ir hmTOP–seq metodų analizė nėščių moterų kraujyje cirkuliuojančioje DNR. Šiuose taikomuosiuose poskyriuose aptariama, kaip anksčiau sukurti skaičiavimo ir statistiniai metodai gali būti pritaikomi konkrečiais epigenominių tyrimų atvejais.

8.4.2 TOP–seq Metodo Taikymas Žmogaus Audiniuose

8.4.2.1 Įvadas

Norint suprasti DNR modifikacijos dinamiką, reikalingi jautrūs didelės skiriamosios gebos metodai, leidžiantys išanalizuoti epigenetinius virsmus visame genome. Vienas iš tokių metodų yra TOP–seq — šis metodas identifikuoja genomines uCG (t.y., nemodifikuoti CG) pozicijas.

Čia pateikiame nemodifikuotų citozinų palyginimus įvairiuose žmogaus audiniuose ir ląstelių linijose, pirmiausia įrodydami, kad TOP–seq signalas yra atkartojamas ir gerai sutampa su kitomis DNR modifikavimo profiliavimo metodikomis. Toliau pateikiami DNR modifikavimo palyginimai įvairiuose genomo elementuose — DNR modifikacijos signalas per genus ar epigenetinius elementus, tokius kaip su lamina susijusios sritys ar chromatino segmentai. Galiausiai, mes parodome, kad TOP–seq signalas yra pakankamai jautrus, kad būtų galima nustatyti epigenetinius skirtumus tarp ląstelių tipų, todėl jį galima pritaikyti viso genomo

masto DNR modifikacijų profiliavimui. CGI elementai naudojami kaip platforma aptikti diferencines modifikacijas tarp neuroblastomos ląstelių tipų ir smegenų audinio.

8.4.2.2 TOP–seq Sekoskaitos Duomenų Kokybės Kontrolė

Vidutiniškai kiekviename mėginyje buvo 42 milijonai neapdorotų sekoskaitos fragmentų, išskyrus didelio gylio IMR90 mėginius, kuriuose vidutiniškai buvo 238 milijonai neapdorotų fragmentų. Apdorojus ir prilyginus fragmentus prie referentinio genomo, fragmentų skaičius žymiai sumažėjo (**Figure 6.2**). Pagrinde šio didelio gylio IMR90 mėginių fragmentų kiekio sumažėjimo dalį nulėmė trumpų fragmentų pašalinimas arba PGR duplikatų pašalinimas. Pašalinus pasikartojančius fragmentus ir priskyrus likusius fragmentus CG vietoms, vidutiniškai 16 milijonų fragmentų liko mažo gylio bibliotekose ir 91 milijonas fragmentų didelio gylio bibliotekose.

Visose analizuojamose mėginiuose buvo naudojamos visi nemodifikuoti CG (t.y., padengimas didesnis nei 0), vidutiniškai 21 proc. genomo CG mažo gylio, o 35 proc. — didelio gylio mėginiuose (**Table 6.2**). Vidutinis nustatytų CG vietų padengimas kiekvieno mažo gylio mėginyje buvo 2.8 ir kiekvieno didelio gylio mėginyje 9.6, tačiau po išsamesnio patikrinimo buvo nustatyta, kad chromosomų iš neuroblastomos mėginių padengimas buvo nevienodas — pastebėtas daug didesnis vidutinis padengimą 2 chromosomoje (**Figure 6.3**). Atidžiau išnagrinėjus, buvo nustatyta, kad šį didesnį padengimą sukėlė fragmentai, kilę iš konkretaus 1.6 Mb regiono 2 chromosomoje (chr2:15026730 — 16640120) (**Figure 6.4**). Šiame regione yra proto-ongoneninis *MYCN* genas, kuris amplifikuojasi būtent neuroblastomos ląstelių linijose.

Pasirinkę visas identifikuotas CG pozicijas, išmatavome koreliaciją tarp techninių pasikartojimų (Pearson vidutinis $r = 0.69$) (**Figure 6.5**). Tokią vidutinę koreliaciją galėjo sukelti nedidelis sekoskaitos gylis, kuris buvo naudojamas ir iš tiesų, išmatavus Jaccard koeficientą tarp nustatytų CG vietų, buvo pastebėtas nedidelis techninių pasikartojimų sutapimas

(vidutinis Jaccard koeficientas 0.4) (**Figure 6.5 A**). Jaccard koeficientas rodo, kad tam tikri CG rinkiniai tarp techninių pasikartojimų yra skirtingi, tačiau šį rezultatą galima būtų susieti su anksčiau minėtu mažu sekoskaitos gyliu. Norėdami įrodyti šią prielaidą, mes susimuliavome labai mažo sekoskaitos gylio duomenų rinkinius ir monotoniškai padidinome jų bibliotekos dydį, pademonstruodami aiškią linijinę priklausomybę tarp bibliotekos dydžio ir Jaccardo koeficiento. Norėdami dar labiau patvirtinti TOP–seq metodo atkartojamumą, apskaičiavome koreliaciją tarp techninių replikų, naudojant įvairaus dydžio genomo regionus (**Figure 6.5 B**). Koreliacijos tarp neuroblastomos mėginių buvo Pearson $r = 0.9$ 1 kb regionuose, o kitoms mėginių grupėms reikėjo daug didesnių regionų dydžių, kad koreliacija būtų pakankamai aukšta.

Toliau mes atlikome TOP–seq signalo transformacijas, kad gautume didelės skiriamosios gebos viso genomo DNR modifikacijos lygį. Taikant kernelio tankio metodą, apskaičiuoti svertinio tankio įverčiai iš TOP–seq padengimo signalo ir normalizuoti pagal nesvertinio CG tankio įvertį, siekiant gauti TOP–seq nemetilomo tankio signalą. Šis koregavimas sustiprino mažo gylio pasikartojimų Pearson koreliaciją iki $r = 0.8$, o didelio gylio IMR90 pasikartojimų koreliacija padidėjo iki $r = 0.9$ (**Figure 6.8**).

Vieno CG absoliuti žemo gylio ir didelio gylio TOP–seq *u*–density koreliacija su IMR90 WGBS duomenimis buvo $\mid r \mid = 0.59$ ir $\mid r \mid = 0.64$, atitinkamai (**Figure 6.9**). Tolesniame koregavimo etape mes siekėme atsižvelgti į galimus sekai būdingus variantus, kurie gali turėti įtakos TOP–seq signalui. Mes panaudojome nedidelę WGBS duomenų rinkinio dalį (20 chromosomą), kad pritaikytumėme eksponentinio irimo modelį, kuriame būtų panaudotos papildomos genominės kovariantės. Šis modelis vėliau buvo naudojamas konvertuojant TOP–seq *u*–density į vadinamuosius CG metilinimo įverčius (*m*–estimate metilinimo vertes absoliučioje skalėje nuo 0 iki 100). Nors antrasis patobulinimo žingsnis turėjo nedidelę įtaką koreliacijai tarp TOP–seq techninių replikų, jis pagerino vieno CG gebos koreliaciją su IMR90 WGBS $r = 0.69$ didelio gylio rinkiniu (**Figure 6.9**).

Toliau mes palyginome šių metodų sutaptį įvairiuose genomo elementuose. Viso genomo profiliai tarp pagrindinių genominių ypatybių parodė gerą TOP–seq signalo ir WGBS sutaptį CGI, enhanceriuose, 3'UTRs, egzonuose, intronuose, baltymus koduojančių genų prieš srovę ir pasroviui regionuose (**Figure 6.10**). Įdomu tai, kad TSS chromatino segmentuose TOP–seq signalo koreliacija buvo maža su WGBS signalu, o MBD parodė vidutinę koreliaciją šiuose elementuose.

Galiausiai mes įvertinome, kaip apatiniai ir viršutiniai 10 proc. nemetilintų CGI, genų ir 10 kb dydžio regionų identifikuojami naudojant TOP–seq metodą, sutampa su apatiniu ir viršutiniu 10 proc. nemetilintų regionų, gautų iš WGBS. IMR90 mėginiuose mes pastebėjome labai stiprų ryšį tarp TOP–seq u–density ir WGBS visuose 10 proc. naudojamų elementų, taip pat apatinių 10 proc. genų (tiksliojo Fisher testo koeficientas $\sim$ 37) (**Figure 6.11**). Tačiau smegenų mėginiuose tik apatiniai 10 proc. genų parodė tokį didelį ir, įdomu, labai panašų praturtinimą kaip IMR90 mėginiai. Smegenų mėginiai taip pat parodė santykinai mažą 10 proc. geriausių CGI praturtinimą, tuo tarpu CGI buvo labiausiai praturtinta 10 proc. elementų IMR90 mėginiuose.

8.4.2.3 Epigenominiai Žemėlapiai

Pasirinkus identifikuotas CG pozicijas, buvo ištirtas jų praturtinimas ir pasiskirstymas įvairiuose genomo elementuose (**Figure 6.12**). Daugumoje tirtų genominių elementų pasirinktų CG pozicijų praturtinimas ar trūkumas tarp skirtingų mėginių grupių buvo panašus. Didžiausias praturtinimas pastebėtas elementuose, susijusiuose su genų pradžia (5'UTR, CGI, įvairių genų biotipų promotoriai). Įdomu tai, kad tik smegenų mėginiuose esančios CG vietos parodė praturtinimą kituose su baltymais—koduojančiame genais susijusiuose elementuose (t. y., egzonus, 3'UTR ir intronus). Didžiausias nustatytų CG vietų trūkumas pastebėtas pseudogenuose ir SINE pasikartojimuose, kai visos mėginių grupės turėjo panašias tendencijas. TOP–seq signalą buvo galima aptikti 96 proc. iš visų autosominių CGI. Kaip ir tikėtasi, promotoriaus

CGI buvo labiausiai praturtinti uCG, patvirtinant jų labai nemodifikuotą būseną (**Figure 6.13**). Nustatytų uCG vietų skirtumai buvo didesni tarp intrageninių ir tarpgeninių CGI, tai patvirtina jų modifikacijų įvairovę ir vidutiniškai aukštesnį metilinimo lygį. Įdomu tai, kad santykinai didelė tarpgeninių CGI dalis parodė arba absoliučią modifikaciją, arba tik labai lengvą modifikaciją, formuojančią bimodalinius pasiskirstymus smegenų ir IMR90 mėginių grupėse, bet ne iš neuroblastomos gautuose mėginiuose.

Mes taip pat palyginome TOP–seq *u*–density profilius su WGBS įvairiuose su genais susijusiuose elementuose (**Figure 6.14**). Kaip ir tikėtasi, atitinkamų audinių TOP–seq ir WGBS profiliai parodė atvirkštines tendencijas visuose analizuojamuose regionuose. Mes taip pat sudarėme TOP–seq *u*–density profilius įvairiuose chromatino segmentuose ir aplink juos. Tarp aktyvių promotorių būsenų aktyvūs TSS, dvivalenčiai TSS promotoriai parodė aukštesnius TOP–seq *u*–density signalus, rodančius jų žemesnį metilinimo lygį (**Figure 6.15**). Norėdami įvertinti tolimesnį TOP–seq pritaikojamumą mes ištyrėme LAD elementus. Jau anksčiau buvo pastebėta, kad LAD elementai atitinka iš dalies modifikuotus DNR regionus ir yra tiesiogiai susiję su genų represijomis ir paprastai būna nuo 80 kb iki 30 Mb dydžio. Analizė parodė stiprų LAD regionų regionų hipometilinimą, palyginti su tarp LAD esančiais regionais IMR90 ląstelėse, o smegenų mėginiuose nebuvo nustatyta jokių TOP–seq *u*–density pokyčių (**Figure 6.16**). Pastebėtas DNR modifikacijų tendencijas atspindėjo ir WGBS duomenys, dar labiau patvirtinantys, kad suaugusiųjų smegenų žievės ląstelėse nėra LAD.

8.4.2.4 Skirtingai Modifikuoti Regionai Neuroblastomos Mėginiuose

Atsižvelgiant į tai, kad neuroblastoma yra neuroendokrininis navikas, atsirandantis dėl nervinių keterinių ląstelių, analizę sutelkėme į promotorinių ir intragenininių CGI—DMR, nustatytus tarp N, S ir smegenų mėginių, ir priskyrėme juos jų genams šeimininkams (**Table 6.3**). Mes atlikome

genų funkcinę anotacijų analizę su identifikuotais CGI-–DMR, pirmiausia sutelkdami dėmesį į promotoriaus CGI. Genų praturtinimo analizė S/B–hypoM ir N/B–hypoM rinkiniams parodė reikšmingą tarpląstelinių organelių ir citoskeleto komponentų praturtinimą. HyperM promotorinės CGI tiek N, tiek S ląstelėms buvo žymiai praturtintos grupėse, kuriose buvo baltymų, glikoproteinų, signalinių peptidų ir biologinių procesų, apimančių neuronų diferenciaciją, vystymąsi ir aksonogenezę. Tai atitinka šio vystymosi naviko pobūdį, kuris yra susijęs su sutrikusiu neuronų fenotipo brendimu. Įdomu tai, kad analizuojant N/B–hyperM CGI, nustatyta genų, dalyvaujančių nervų keteros vystymesi ir migracijoje, hipermetilinimas, kurių nėra S/B–hyperM CGI-–DMR.

8.4.2.5 Diskusija

Pirmoje šio skyriaus dalyje buvo pateikti pavyzdžių, analizuotų naudojant TOP–seq metodą, kokybės kontrolės rezultatai. Nors nustatytų CG pozicijų padengimo koreliacija ir Jaccard koeficientas buvo vidutinių verčių diapazone, Fisher testo įverčiai įrodė, kad CG vietos nustatymas nėra atsitiktinis procesas. Be to, Jaccard koeficiento vertės, gautos naudojant susimuliuotą duomenų rinkinį, įrodė, kad didesnių bibliotekų dydžių pavyzdžiai sukurs didesnį panašumą. Palyginus TOP–seq signalą su referentiniais WGBS duomenų rinkiniais, paaiškėjo, kad jis koreliuoja daug geriau nei kiti palyginti metodai. Svarbu tai, kad panašumas į WGBS signalą priklauso nuo genomo elemento tipo, o tai rodo galimą DNR sekos konteksto įtaką praturtinimo metodams.

Kitos signalo transformacijos buvo taip pat naudingos, nes koreliacijos su WGBS signalu buvo žymiai didesnės. Didesnės koreliacijos pastebėtos net kitame WGBS duomenų rinkinyje, kuris nebuvo naudojamas modelių parametrams optimizuoti. Gautos signalo transformacijos parodė gana didelę sutaptį su WGBS signalu tiriant genų ar kitų genomo elementų DNR modifikacijos profilius. Galiausiai mes panaudojome TOP–seq signalą ir nustatėme platų CGI DMR spektrą, susijusį su specifiniais iš neuroblastomos kilusių ląstelių tipais. Tačiau svarbu paminėti, kad TOP–seq

metodas, kaip ir daugelis kitų fragmentų skaičiavimu pagrįstų epigenomų profiliavimo metodų, yra jautrus kopijų skaičiaus kitimams. Kai pastebimas labai didelis padengimo lygis (pvz., *MYCN* lokusas), gali būti, kad tokį signalo praturtinimą lemia tik didelis tikslinių DNR kopijų kiekis. Todėl *de novo* atrasti DMR turėtų būti patikrinti, ar jie nepatenka į genetinių aberacijų karštuosius taškus.

Apidendrinant šiame skyriuje mes apibūdinome TOP–seq metodo taikymą žmogaus audiniuose ir ląstelių tipuose. Gauti rezultatai rodo, kad TOP–seq metodas suteikia informaciją apie DNR modifikacijas vieno CG skiriamosioje geboje ir viso genomo masto. Epigenominiai palyginimai, sudaryti naudojant šį TOP–seq metodą, suteikia įžvalgų apie DNR modifikacijos kintamumą mėginiuose, susijusiuose su skirtingomis eksperimentinėmis grupėmis arba skirtingų tipų genominiais elementais.

8.4.3 hmTOP–seq Metodo Taikymas mESC

8.4.3.1 Įvadas

5hmC yra gausiausia oksidacinės DNR modifikacijos forma. Ji dalyvauja daugelyje biologinių procesų, įskaitant embriogenezę, neurologinius procesus ir vėžį. Profiliuojant šią palyginti negausią genomo modifikaciją reikia naudoti jautrius, didelės skiriamosios gebos metodus. Šiame skyriuje aprašoma naujos kartos sekoskaitos technologijos, vadinamos hmTOP–seq, analizė, kuri gali būti naudojama norint nustatyti 5hmC vienos bazės skiriamosios gebos tikslumu. Norėdami patvirtinti savo metodą, mes panaudojome pelės ESC genominę DNR ir palyginome sugeneruotą hmTOP–seq signalą su duomenimis, gautais naudojant kitą DNR modifikavimo profiliavimo metodą, ir nustatėme gerą koreliaciją tarp 5hmC susietų regionų. Mes taip pat palyginome 5hmCG pasiskirstymą geniniuose ir epigenominiuose elementuose, tokiose kaip histonų modifikacijos. Remdamiesi šia analize darome išvadą, kad hmTOP–seq gali būti naudojama kaip viso genomo 5hmCG profiliavimo technika.

8.4.3.2 hmTOP–seq Sekoskaitos Duomenų Kokybės Kontrolė

Vidutiniškai kiekviename tiksliniame mėginyje buvo 59 milijonai neapdorotų sekoskaitos fragmentų, kurių žymiai sumažėjo pašalinus trumpus sekoskaitos fragmentus (**Figure 6.17**). Visų mėginių prilyginimo įvertis buvo palyginti didelis, o pašalinus fragmentus dėl prastos prilyginimo kokybės, fragmentų skaičius sumažėjo tik nežymiai.

Vidutinis nustatytų CG vietų padengimas labai skyrėsi skirtingose DNR kiekio bibliotekose, taip pat tarp tikslinių mėginių ir jų atitinkamų kontrolių (**Figure 6.18 A**, **Supp. Table 2**). Vidutinis CG padengimas mėginiuose svyravo nuo 4.5 iki 12.8, tuo tarpu atitinkamuose kontroliniuose mėginiuose CG padengimas buvo vidutiniškai du kartus mažesnis. Atlikus tolimesnį tyrimą, nustatytas tikslinių mėginių ir kontrolinių mėginių skirtumas identifikuotoje CG–frakcijoje. Nors identifikuota CG–frakcija svyravo nuo 5.6 proc. iki 34 proc., šis įvertis kontrolėse vidutiniškai buvo tik 0.01 proc.

Didesnio DNR kiekio hmTOP–seq bibliotekų techniniai pasikartojimai gerai koreliavo esant vienai CG skiriamajai gebai (Pearson $r = 0.46$ ir $r = 0.8$ 50 ng ir 500 ng bibliotekose, atitinkamai **Figure 6.19 A**). Nors 5 ng DNR kiekio bibliotekos parodė žymiai mažesnę koreliaciją tarp techninių pasikartojimų (Pearson $r = 0.11$), tačiau buvo pastebėtas pagerėjimas kai buvo naudojami didesni genomo regionai. Pearson koreliacija didesnio kiekio bibliotekose vidutiniškai padidėjo iki $r = 0.92$, naudojant 5 kb skiriamąją gebą, o 5 ng mėginyje šis matavimas padidėjo tik iki $r = 0.55$ (**Figure 6.19 B**). Mes taip pat nustatėme, kad nustatytų CG vietų sutaptis didėja, atsižvelgiant į panaudotą DNR kiekį. Jaccardo koeficientas tarp 500 ng techninių pasikartojimų buvo 0.08, tuo tarpu 50 ng mėginiuose jis buvo 0.056, o 5 ng mėginiuose – tik 0.02 (**Figure 6.19 A**).

Nors nustatytų CG vietų sutaptis buvo labai maža, ji buvo reikšminga ir neatsitiktinė. Mes papildomai patikrinome nustatytų CG vietų sutaptį naudodami tikslų Fisher testą ir nustatėme, kad 500 ng mėginiai gerai

sutapo ne tik tarp techninių pasikartojimų, bet ir su kitais mėginiais (**Figure 6.19 C**). Be to, buvo atliktas didesnių DNR kiekio duomenų rinkinių sumažinimas (**Figure 6.20**). Ši analizė parodė, kad su mažesnio dydžio bibliotekomis vis tiek galima pasiekti pakankamai aukštas koreliacijas. Sumažinę bibliotekos dydį 50 proc., pastebėjome, kad koreliacija tarp techninių pasikartojimų sumažėjo tik 10 proc. Tolesnis bibliotekų mažinimas tik nežymiai sumažino koreliaciją.

Toliau mes palyginome mūsų 5hmCG duomenų rinkinius su TAB–seq duomenimis, parodydami, kad hmTOP–seq identifikavo 50 proc. ir 25 proc. TAB–seq nustatytų 5hmCG vietų atitinkamai 500 ng ir 50 ng DNR rinkiniuose (tiksliojo Fisher testo įverčiai atitinkamai 4 ir 3.8). Tiesioginis hmTOP–seq ir TAB–seq signalo palyginimas parodė gerą abiejų metodų sutaptį (Pearson $r = 0.94$, **Figure 6.21**).

8.4.3.3 Epigenominiai Žemėlapiai

5hmCG pasiskirstymo įvairiuose genominiuose elementuose analizė parodė gerą sutaptį su viešais duomenimis. Labai hidroksimetilintos CG vietos (20 proc. stipriausio hmTOP–seq signalo) buvo praturtintos paruoštais (angl., *poised*) enhanceriais (pažymėtais H3K4me1), aktyviais enhanceriais (pažymėtais H3K27ac), egzonais, 3'UTR (**Figure 6.22**).

8.4.3.4 Diskusija

Šiame skyriuje mes pateikėme hmTOP–seq metodo taikymą 5hmCG modifikacijai profilizuoti viena CG skiriamąja geba genomo mastu. Mes parodome, kad apskaičiuotas 5hmCG signalas yra atkuriamas ir gerai koreliuoja didesnio DNR kiekio bibliotekose. Be to, DNR modifikacijos signalo sutaptis tarp hmTOP–seq ir kitų metodų yra gana aukšta. Galiausiai, mes pristatėme genomo masto 5hmCG modifikacijos praturtinimo prilyginimus įvairiuose genomo elementuose. Apskaičiuoti praturtinimai gerai sutampa su anksčiau praneštais rezultatais, įrodydami, kad

hmTOP–seq metodas gali būti naudojamas kaip alternatyvus metodas, taikant šiuo metu epigenomo mastu naudojamus profiliavimo metodus.

8.4.4 caCLEAR Metodo Taikymas mESC

8.4.4.1 Įvadas

Šiame skyriuje pateikiama caCLEAR metodo duomenų analizė. caCLEAR — naujas 5caC modifikacijai specifinis sekoskaitos metodas. Čia pateikiame caCLEAR metodo kokybės kontrolės rezultatus pelių embrioninėse kamieninėse ląstelėse iš skirtingų pliurpotencijos būsenų. Be to, pateikiami 5caC modifikacijos praturtinimai įvairiuose genomo elementuose ir įrodoma, kad caCLEAR gali būti naudojamas kaip viso genomo masto DNR modifikavimo profiliavimo technika.

8.4.4.2 caCLEAR Sekoskaitos Duomenų Kokybės Kontrolė

Vidutiniškai kiekviename mėginyje buvo 34 milijonai neapdirbtų sekoskaitos fragmentų, kurių kiekis labai varijavo skirtingose mėginių grupėse (**Figure 6.23**). Tdg mėginių grupėse ir 2i laukinio tipo mėginių grupėje vidutiniškai buvo 53 milijonai fragmentų, o likusiose mėginių grupėse vidutiniškai buvo tik 15 milijonų fragmentų.

Didžiausias vidutinis identifikuotų CG vietų padengimas buvo pastebėtas Tdg mėginiuose, kur jis buvo labai panašus tarp 2i ir ne 2i grupių. Laukinių mėginių padengimas buvo mažesnis ir labai skirtingas tarp dviejų ankstesnių grupių (**Figure 6.24 B**, **Supp. Table 3**). Įdomu tai, kad lyginant 2i ir ne 2i grupes pastebėjome didelį nustatytų CG vietų kiekio skirtumą (**Figure 6.24 B**). Tiek Tdg, tiek laukinio tipo mėginiuose su 2i nustatyta daugiau identifikuotų CG vietų nei mėginiuose be 2i (beveik du kartus daugiau Tdg ir tris kartus daugiau laukinių tipų mėginiuose).

Abiejų Tdg grupių techniniai pakartojimai parodė koreliaciją (Pearson $r = 0.87$, **Figure 6.25 A**). Tuo tarpu abi laukinio tipo grupės parodė skirtingą koreliaciją tarp techninių pakartojimų. Pearson $r = 0.42$ 2i grupei ir Pearson $r = 0.29$ ne 2i grupei. Koreliacija tarp techninių pakartojimų padidėjo apskaičiuojant ir didesnėse skiriamosiose gebose (**Figure 6.25 B**). Pearson r pasiekė beveik maksimalią koreliaciją, kai buvo vertinama 5 kb regionuose abiejose Tdg grupėse. Galiausiai buvo apskaičiuota identifikuotų CG vietų sutaptis (**Figure 6.25 A, C**). Abi Tdg mėginių grupės parodė aukštesnį Jaccard koeficientą tarp techninių pasikartojimų ir su kita Tdg grupe, kai buvo panaudotas tikslusis Fisher testas. Laukinių tipų grupių sutaptis buvo mažesnė nei atitinkamose Tdg bibliotekose, bet didesnė nei Tet kontrolėje, įrodant, kad identifikuotos CG pozicijos buvo ne atsitiktinės.

8.4.4.3 Epigenominiai Žemėlapiai

Nors dauguma 5caCG pozicijų varijavo tarp genų, reguliatorinių ir kitų elementų, bet daugumą identifikuotų pozicijų buvo praturtinta — paruoštuose ir aktyviuose enhanceriuose. Taip pat patikima CG dalis buvo rasta su pliuripotencija susijųsių veiksnių jungimosi vietose ir SINE pasikartojimuose (**Figure 6.26**). Galiausiai mes palyginome identifikuotas CG pozicijas su atviro chromatino regionais iš įvarių pelės audinių ir organų (**Figure 6.27**). To pasekoje buvo pastebėta didžiulė variacija tarp mėginių grupių ir analizuotų organų. Identifikuotos CG pozicijos buvo praturtintos virškinimo trakto organuose (pvz., skrandyje, žarnyne), plaučiuose, bet rodė trūkumą nervinės sistemos organuose.

8.4.4.4 Diskusija

Šiame skyriuje aprašytas caCLEAR metodo taikymas 5caCG modifikacijos profiliavimui, esant vienai CG skiriamajai gebai viso genomo mastu. Mes parodomėme, kad apskaičiuotas 5caCG signalas yra atkuriamas ir gerai koreliuoja tarp techninių pasikartojimų. Be to, pristatėme genomo

masto 5caCG modifikacijos praturtinimą įvairiuose genomo elementuose kartu su 5caCG modifikavimo praturtinimu atviruose chromatino regionuose, kurie būdingi specifiniams pelės audiniams.

Teiginys VI — caCLEAR metodas suteikia informacijos apie DNR modifikacijos signalą įvairiuose genomo elementuose.

8.4.5 TOP–seq ir hmTOP–seq Metodų Taikymas Prenataliniame Testavime

8.4.5.1 Įvadas

21 chromosomos trisomija yra dažniausia žmogaus autosominė aneuploidija, dėl kurios susidaro fenotipinių ypatybių (fizinių ir intelektinių sutrikimų) rinkinys, žinomas kaip Dauno sindromas. Šiuo metu T21 diagnozei patvirtinti naudojamos invazinės diagnostikos procedūros, tokios kaip amniocentezė ir choriono gaurelių mėginių ėmimas, o po to atliekama genetinė analizė (pvz., kariotipo nustatymas). Nors invazinių procedūrų saugumas pagerėjo, vis dar išlieka persileidimo rizika (0.3 proc., 0.9 proc. atliekant amniocentezę ir imant choriono gaurelius). Taigi, norint sumažinti invazinių diagnostikos procedūrų skaičių, reikalingi neinvaziniai ir labai patikimi prenataliniai atrankos tyrimai.

Kai nėščiųjų moterųjų kraujo plazmoje buvo rasta vaisiaus genominė medžiaga — cirkuliuojanti ne ląstelinė vaisiaus DNR (cffDNA), buvo dedama daug pastangų panaudojant cffDNA neinvaziniam prenataliniam vaisiaus genomų mutacijų tyrimui. Tokios atrankos metu T21 aptikimo dažnis yra didesnis nei 99 proc., o klaidingai teigiamas rodiklis siekia vos 0.1 proc. Taigi NIPT pagrįstos diagnostikos technologijos yra reikšmingas patobulinimas, palyginti su tradicine patikra. Tačiau cffDNA nustatymas motinos kraujyje yra nemenkas iššūkis, nes tik iki 10 proc. nėščios moters plazmoje esančios DNR yra gaunama iš vaisiaus.

Šiame tyrime mes pritaikėme TOP–seq ir hmTOP–seq technologijas,

norėdami analizuoti motinos cfDNA DNR modifikacijas, kad nustatytume iš vaisiaus gautus genomo regionus. Buvo sukurti viso genomo masto 5hmCG ir uCG modifikacijų profiliai, taip pat apskaičiavome diferencinių modifikacijų signalo praturtinimą įvairiuose genomo elementuose tarp įvairių eksperimentinių grupių. Svarbiausia, kad 21 chromosomos vaisiaus trisomija buvo nustatyta labai specifiškai / jautriai, naudojant regioninius modifikacijų skirtumus. Be to, vaisiaus frakcija iš cfDNA buvo apskaičiuota naudojant uCG ir 5hmCG signalą.

8.4.5.2 Sekoskaitos Duomenų Kokybės Kontrolė

Apdorojęs sekoskaitos fragmentus ir apskaičiavęs CG padengimą, pirmiausia palyginome biologinių pasikartojimų panašumo statistikas (**Figure 6.28**). Vidutiniškai uCG biologiniai pasikartojimai buvo daugiau panašesni nei 5hmCG pasikartojimai visose išmatuotose statistikose. Įdomu tai, kad koreliacija tarp biologinių pasikartojimų buvo didžiausia cfDNA iš nėščių moterų uCG mėginių, tuo tarpu 5hmCG mėginiuose ši grupė turėjo mažiausią koreliaciją, palyginti su dviem kitomis eksperimentinėmis grupėmis.

Norėdami patikrinti, ar uCG ir 5hmCG modifikacijos skirtumus galima atskirti tarp mėginių grupių, pirmiausia apžvelgėme bendrą uCG ir 5hmCG vietų sekoskaitos padengimą. Vidutinis bendras uCG padengimas tarp tirtų trijų grupių buvo skirtingas (ANOVA p–vertė 7×10^{-7}); jis buvo mažiausias tarp NPC ir didžiausias tarp CV mėginių (**Figure 6.29**). Toliau mes palyginimo DNR modifikacijų signalo pasiskirstymą tarp genomo elementų. Hipometilinti regionai geriausiai persidengė su CGI, 5'UTR ir genų promotoriais. 5hmCG modifikuotos pozicijos labiausiai persidengė su 3'UTR, egzonų ir intronų regionais.

8.4.5.3 Diferenciškai Modifikuotų Regionų Analizė

Šiame tyrime mes siekėme nustatyti vaisiui būdingus genomo lokusus, kurie galėtų būti naudojami kaip vaisiaus fenotipo uCG ir (arba) 5hmCG

biomarkeriai. Kadangi mūsų tyrime buvo skirtingų fenotipinių grupių DNR mėginiai (t. y. moterys, nėščios su ne T21 vaisiu, moterys, nėščios su T21 vaisiu, nenėščios moterys, choriono gaurelių mėginys), o tai leido identifikuoti norimus biologinius žymenis taikant grupių palyginimus, kurie pavaizduoti **Figure 6.31** paveiksle.

Pirmiausia palyginę NPC mėginius su ne T21 nėštumo mėginiais mes identifikavome DMR, kurie buvo nėštumui specifiniai. Ši analizė atskleidė 2,761 uCG DMR (FDR q–vertė $< 5 \times 10^{-2}$) (**Table 6.7**), tačiau toks statistinis slenkstis neleido identifikuoti 5hmCG DMR ir reikėjo naudoti žemesnį patikimumą (p–vertė $< 5 \times 10^{-2}$), kas mums leido identifikuoti 4,930 nėštumui specifinius 5hmCG DMRs. Toliau palyginę NPC mėginius su CV mėginiais mes sugebėjome identifikuoti CV–specifinį (t.y., vaisiui–specifinį) signalą. Ši analizė atskleidė 16,555 CV–specifinius uCG DMR (FDR q–vertė $< 5 \times 10^{-2}$) ir 15,986 CV–spefinius 5hmCG DMR (FDR p–vertė $< 5 \times 10^{-2}$).

Galiausiai, sukryžmindami nėštumui būdingus regionus su CV specifiniais regionais, išskyrėme vaisiui būdingą DNR modifikacijos signalą, kurį galima rasti nėščios moters cfDNA, tuos regionus pavadinome specifiniais placentos DMR. Šis susikirtimas tarp nėštumo ir CV specifinių DMR tiek uCG, tiek 5hmCG DMR rinkiniuose buvo didesnis, nei buvo galima tikėtis vien atsitiktinai (n = 2,164, OR = 43; n = 1,589, OR = 5.5, uCG ir 5hmCG, atitinkamai; p–vertės mažiau nei 1×10^{-15}). Placentai būdingų uCG DMR skirtumas tarp NPC ir cfDNA mėginių sutapo su skirtumu tarp NPC ir CV mėginių (Pearson r = 0.82 ir Pearson r = 0.89, CG–padengimui ir CG–frakcijai, atitinkamai, **Figure 6.32**). Panašūs rezultatai buvo pastebėti ir 5hmCG DMR (Pearson r = 0.8 ir r = 0.8, atitinkamai CG–padengimui ir CG–frakcijai).

Nustatyti nėštumo ir CV specifiniai DMR rinkiniai persidengia labiau, nei tikėtasi atsitiktinai, tačiau šiam rezultatui įtakos galėjo turėti genetinė variacija. Gali būti, kad nustatyti DMR buvo gauti iš mQTL regionų, o

pastebėtas DNR modifikacijos signalas yra susijęs su DNR sekos kintamumu. Šiuo atveju naudotiems NPC mėginiams būdingas vienas genetinis fonas, o nėščių moterų mėginiams -– kitas genetinis fonas, o susikirtimą tarp nėštumui ir CV būdingų rinkinių lemia signalas, kuriam daro įtaką DNR sekos kintamumas (t.y., mQTL). Norėdami patikrinti šią prielaidą, apskaičiavome nustatytą DMR persidenigmą su žinomais mQTL rinkiniais (virkštelės kraujyje nustatyti mQTL ir mQTL identifikuotas nėščių moterų kraujyje). Stebimi mQTL ir DMR grupių susikirtimai nebuvo reikšmingi (visos apskaičiuotos p-–vertės iš tiksliojo Fisher testo buvo daugiau nei 0.05).

Nustatę placentai būdingus uCG ir 5hmCG DMR, mes išbandėme persidengimą su skirtingais genominiais elementais **Figure 6.33**. uCG DMR buvo praturtinti elementais, susijusiais su genų 5' galu (5'UTR, baltymus koduojančių ir lincRNA genų promotoriai ir stipriausias praturtinimas promotoriaus CGI), taip pat placentos enhanceriais. Priešingai, 5hmCG DMR buvo praturtinti kitomis baltymus koduojančiomis genų dalimis, egzonais, intronais ir 3'UTRs.

Galiausiai mes paklausėme, ar placentai būdingi DMR yra informatyvūs apie vaisiaus kariotipą (t.y., T21). Naudodami kryžminę validaciją, mes sukonstravome ir įvertinome logistinės regresijos modelį kiekvienam placentai būdingam DMR, naudodami CG–padengimą ir CG–frakciją kaip nepriklausomus kintamuosius, o vaisiaus kariotipą -– kaip priklausomąjį kintamąjį. Iš viso 21 chromosomoje buvo atrasti 376 uCG ir 496 5hmCG DMR, kurie 100 proc. tikslumu klasifikavo mėginius pagal vaisiaus kariotipą (AUC = 1) **Figure 6.34**.

Po to, kai nustatėme placentai būdingus DMR ir pogrupį, kuris 100 proc. tikslumu gali klasifikuoti mėginius pagal vaisiaus kariotipą, mes pasirinkome kitokį metodą. Tiesiogiai įvertinome sveikų ir T21 teigiamų nėštumų cfDNA mėginių modifikacijos skirtumus ir apskaičiavome T21 specifinius DMR. Buvo naudojamas logistinis regresijos modelis, kai CG–padengimas ir CG–frakcija buvo nepriklausomi kintamieji, o kariotipas -– priklausomas kintamasis. Taip mes nustatėme 3,490 uCG ir 2,002 5hmCG DMR, iš kurių tik 82 persidengė tarp dviejų rinkinių (OR = 2.3,

p–vertė $= 1.1 \times 10^{-10}$). Tik 216 ir 124 T21 specifiniai DMR persidengia su placentai būdingais uCG ir 5hmCG DMR (OR = 6.1 ir OR = 8.2; p–vertė $< 2.2 \times 10^{-16}$), o tai rodo, kad skirtingos DMR identifikavimo strategijos lemia skirtingus DMR rinkinius 21 chromosomoje. Pabrėžtina, kad T21 specifiniai DMR rodo didesnius CG–padengimo ir CG–frakcijos skirtumus nei regionai, kuriuose nebuvo diferencijuotų modifikacijų (**Figure 6.35**).

Įdomu tai, kad tiek uCG, tiek 5hmCG DMR rinkiniai geriau persidengia su nėštumui būdingais DMR rinkiniais (OR = 6.6 ir OR = 9, uCG ir 5hmCG, atitinkamai) nei su CV specifiniai DMR rinkiniais (OR = 2.4 ir OR = 2.9, uCG ir 5hmCG, atitinkamai; p–vertės $< 2.2 \times 10^{-16}$). Šis rezultatas pateikia du galimu scenarius: kad vaisiaus audiniai be placentos trofoblastų gali prisidėti prie motinos kraujo cfDNA mišinio arba kad tai gali būti naudojamo audinio tipo artefaktas — nėštumui būdingi DMR taip pat yra matuojami cfDNA, kaip ir T21 specifiniai DMR, kai CV specifiniai DMR nustatomi lyginant du skirtingus audinius.

Genominių elementų praturtinimas su T21 specifiniais DMR buvo panašus į nėštumui būdingų DMR (**Figure 6.33**). T21 specifiniai uCG DMR buvo žymiai praturtinti 5'UTR, promotorinėmis CGI ir baltymus koduojančių genų promotoriais (atitinkamai OR = 2.4, OR = 2.4 ir OR = 2.5). O T21 specifinių 5hmCG DMR promotoriai buvo žymiai praturtinti kitomis baltymus koduojančiomis genų dalimis: egzonais, intronais ir 3'UTR. Tačiau, priešingai nei placentai būdingiems DMR, T21 specifiniai uCG DMR buvo mažiau praturtinti lincRNR promotoriais, placentos enhanceriais ir geniniais CGI.

T21 specifinių DMR pasiskirstymas palei 21 chromosomą uCG ir 5hmCG duomenų rinkiniams buvo skirtingas (p–vertė $= 1.6 \times 10^{-12}$; Kolmogorov–Smirnov testas). Dauguma nustatytų 5hmCG DMR buvo linkę kauptis chromosomos ilgojo peties gale, o uCG DMR buvo tolygiau pasiskirstę per 21 chromosomos ilgąjį petį, atitinkamai 46 proc. ir 64 proc. specifinių uCG ir 5hmCG T21 DMR, persidengiantys baltymus koduojančius genus. Įdomu tai, kad DMR, kertantys T21 specifinius uCG ir 5hmCG DMR rinkinius (iš viso 82 DMR), parodė labai didelį baltymus koduojančių

egzonų praturtinimą (OR = 4.4, p–vertė = 8×10^{-4}). Šie egzonai atitiko septynis genus, iš kurių trys anksčiau buvo susiję su Dauno sindromu: *GART*, *DNMT3L* ir *AIRE*.

8.4.5.4 Vaisiaus Frakcijos Analizė

Nustatę, kad nėščių moterų tarpe uTOP–seq ir hmTOP–seq signalai yra didesni, mes toliau siekėme nustatyti koreliaciją tarp signalo stiprumo ir vaisiaus frakcijos. SeqFF metodas buvo pritaikytas uTOP–seq ir hmTOP–seq duomenims, stebint didelę sąsają tarp numatomų ir referentinių vaisiaus frakcijų (Pearson $r = 0.86$, p–vertė = 3.2×10^{-4} ir $r = 0.9$; p–vertė = 3.9×10^{-4}, atitinkamai uCG ir 5hmCG, **Figure 6.37**). Svarbu tai, kad paprasta linijinė regresija atskleidė, kad referentinės vaisiaus frakcijos padidėjimas 0.01 atitinka vaisiaus frakcijos padidėjimą 0.079. 5hmCG duomenimis, prognozuojama vaisiaus frakcija sumažėjo 0.226 kiekvienam 0.01 referentinės frakcijos padidėjimui. Įdomu tai, kad didėjanti vaisiaus frakcija įgytų didėjantį fragmentų skaičių uTOP–seq, bet mažėjantį fragmentų skaičių hmTOP–seq. Toks atvirkštinis ryšys hmTOP–seq greičiausiai rodo, kad SeqFF naudojami regionai yra labai praturtinti uCG vietose, bet jų trūksta 5hmCG vietose.

8.4.5.5 Diskusija

Kiek mums yra žinoma, šis tyrimas yra pirmasis analizuojantis uCG sekoskaitą motinos cfDNA, siekiant nustatyti vaisiaus kariotipą. Mes taip pat parodėme, kad 5hmC profiliavimas motinos cfDNA gali tiksliai informuoti apie vaisiaus kariotipą. hmTOP–seq metodas leido choriono gaurelių mėginiuose ir cfDNA sukurti viso genomo masto 5hmCG profilius. Svarbiausia, kad nustatant T21 vaisius hmTOP–seq buvo labiausiai diskriminuojantis, nepriklausomai nuo cirkuliuojančios vaisiaus DNR frakcijos. Todėl prenataliniai tyrimai, pagrįsti 5hmCG analize, gali maksimaliai padidinti diagnostinį jautrumą, palyginti su sąnaudomis, ir būti

optimalus pasirinkimas sekoskaitos pagrindu atliekamiems NIPT epigenetiniams metodams. Be to, vaisiaus frakciją galima išmatuoti tiesiogiai per hmTOP–seq ir uTOP–seq fragmentų skaičių. Pabrėžtina, kad nustatyta didelė placentai būdingų uCG ir 5hmCG biologinių žymenų grupė, kuri buvo naudojama vaisiaus kariotipui identifikuoti. Įdomu tai, kad šie DMR turėjo geresnę sutaptį su sveikais nėštumui būdingais DMR rinkiniais nei su CV susijusiais DMR, taigi daroma prielaida.

Svarbu paminėti, kad diferentiškai modifikuoti regionai buvo identifikuoti 100 bp genomo regionuose. Toks paprastas metodas savaime nusistato regiono ribą, kuri galimai neleis nustatyti didesnių epigenetinių skirtumų. Ateityje būtų galima naudoti kitus veiksmingus metodus, tokius kaip lygiagretus laikinas agregavimas, kad būtų galima užfiksuoti jautresnius DNR modifikacijų pokyčius.

Pabrėžtina, kad čia aprašytus rezultatus būtina patvirtinti didelėje klinikinėje studijoje, nes šį tyrimą riboja mėginių dydis. Be to, studiją galima išplėsti į kitas įprastas vaisiaus aneuploidijas, tokias kaip Patau ir Edwards sindromai. Kitas įdomus TOP–seq ir hmTOP–seq metodų pritaikymas būtų gauti imprintuotų regionų analizė iš nėščiųjų kraujo išskirtoje DNR.

8.5 Bendrosios Išvados

Šio tyrimo įnašas į mokslą yra:

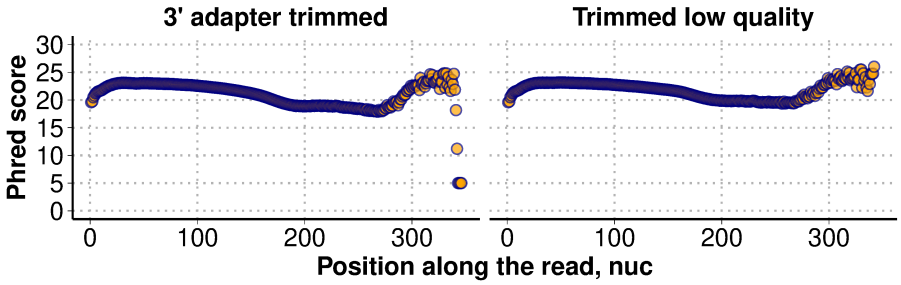
- Sukurti kompiuteriniai metodai efektyviai ir tiksliai apdoroti TOP–seq metodu sugeneruotus didelio našumo epigenominius duomenis. Išvystytos strategijos įgalina DNR modifikacijų analizę vieno citozino skiriamojoje geboje išlaikant grandininės specifiškumą.
- Išvystytos statistinio mokymosi metodikos pagerino TOP–seq metodo epigenominio signalo kokybę. Modeliniam IMR90 genomui pritaikytos statistinio mokymosi technikos padidino Pearson koreliacijų įvertį tarp techninių pakartojimų iki $r = 0.89$, o absoliuti Pearson

koreliacija vieno CG skiriamojoje geboje su referentiniu WGBS signalu padidėjo iki $r = 0.71$.

- TOP–seq metodu paremtos technologijos gali suteikti informacijos apie DNR modifikacijų signalą įvairiuose genominiuose elementuose ir struktūrose.

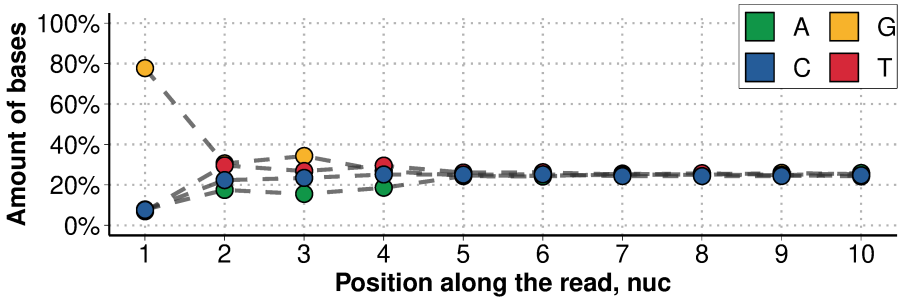
- TOP–seq ir hmTOP–seq metodai gali būti naudojami identifikuoti diferentiškai modifikuotus regionus tarp įvairių mėginių grupių. Abu metodai gali panaudoti CG padengimo ir identifikavimo informaciją suklasifikuoti mėginius iš skirtingų audinių ar kariotipinių grupių. Identifikuoti 100 bp regionai gali būti panaudoti prenatalinei diagnostikai arba įvertinti audinių kompoziciją tam tikrame mėginyje.

Supplemental Figures



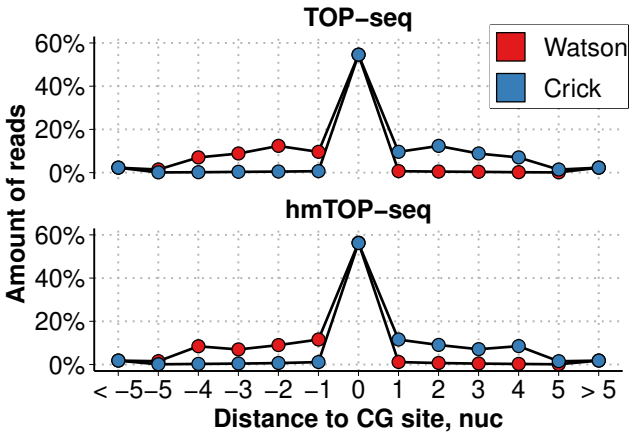
Supp. Figure 1 | TOP–seq Read Quality

Average Phred quality score along the reads in cfDNA sample (sample identifier 137). After trimming lower quality nucleotides from the end of the read Phred score increases and higher mapping quality is expected.



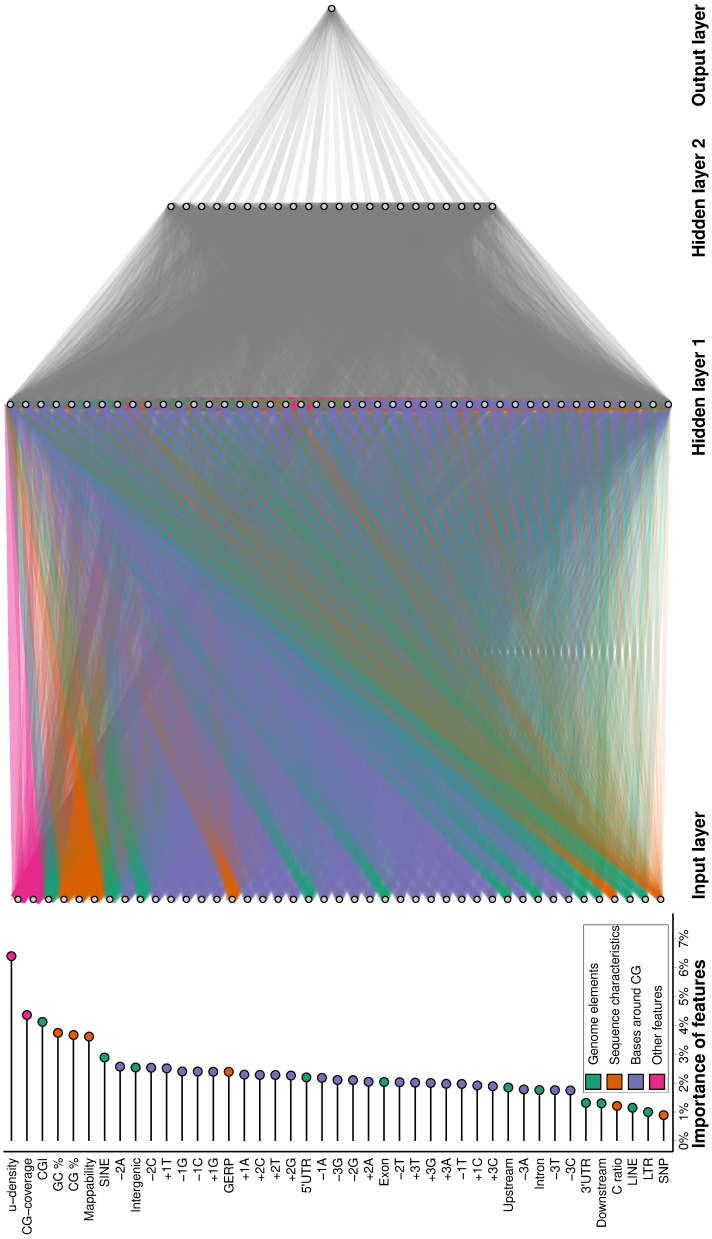
Supp. Figure 2 | Beginning of the TOP–seq Read Structure

Composition of first 10 bases along the generalised read in cfDNA sample (sample identifier 137). Majority of reads contain guanine as their first base since they originate from a cytosine in a CG context.



Supp. Figure 3 | Strand Specific Read Distance to CG sites

Relative distribution between the read start and nearest CG site in cfDNA sample (sample identifier 137). Reads have non–symmetrical distance to CG sites with both TOP–seq and hmTOP–seq reads starting at exactly CG sites or with a strand specific shift.



Supp. Figure 4 | Neural Network Used to Train *nn*–estimate Model

Graphical representation of a designed neural network that was used to estimate DNA modification signal. Leftmost section shows genomic feature information that was extract for each CG site. Length of the segment represents feature importance. "Input layer" contains given features and its line thickness reflects feature importance, while color reflects status of a feature. Next there are two "Hidden layers" that pass final prediction to an "Output layer".

Supplemental Tables

Supp. Table 1 | Change in Amount of Reads

Amount of reads processed in each TOP–seq method or its variations. Given number of reads (in millions) is presented as the total amount of reads from all the samples in a given study at a particular processing step.

Study name	DNA modification	Original reads	Short reads removed	5'adapter removed
TOP–seq	uCG	1,029	777	774
hmTOP–seq	5hCG	353	280	278
caCLEAR	5caCG	382	301	300
NIPT	uCG	1,060	818	815
NIPT	5hmCG	671	540	538

Study name	DNA modification	3'adapter trimmed	Low quality trimmed	Mapped
TOP–seq	uCG	774	748	695
hmTOP–seq	5hCG	278	278	273
caCLEAR	5caCG	300	300	247
NIPT	uCG	815	815	778
NIPT	5hmCG	538	538	526

Study name	DNA modification	High mapping quality	Duplicates removed	Assigned to CG
TOP–seq	uCG	542	412	394
hmTOP–seq	5hCG	241	153	140
caCLEAR	5caCG	205	160	103
NIPT	uCG	645	545	477
NIPT	5hmCG	472	264	223

Supp. Table 2 | Coverage Statistics of 5hmCG Sites

The number of identified CG sites (coverage greater than 0) is represented in absolute and relative numbers. The coverage of identified CG sites is represented with arithmetic mean (i.e., average).

Sample identifier	Replicate identifier	Amount of 5hmCG,%	Average coverage
hmC ctrl 5	K1	0.002	8
hmC ctrl 5	K2	0.002	9.8
hmC 5	R1	5.7	12.1
hmC 5	R2	5.6	12.8
hmC ctrl 50	K1	0.005	3
hmC ctrl 50	K2	0.005	3
hmC 50	R1	18	6
hmC 50	R2	19	5.3
hmC ctrl 500	K1	0.03	1.5
hmC ctrl 500	K2	0.03	1.5
hmC 500	R1	32	4.7
hmC 500	R2	34	4.4

Supp. Table 3 | Coverage Statistics of 5caCG Sites

The number of identified CG sites (coverage greater than 0) is represented in absolute and relative numbers. The coverage of the identified CG sites is represented by the arithmetic mean (i.e., average).

Sample identifier	Replicate identifier	Amount of 5caCG,%	Average coverage
Serum WT	R1	3.3	2.4
Serum WT	R2	3.2	2.5
Serum 2i WT	R1	11	3.8
Serum 2i WT	R2	10.8	4
Serum Tdg	R1	12.2	5.6
Serum Tdg	R2	12.4	5.9
Serum 2i Tdg	R1	22.2	5.2
Serum 2i Tdg	R2	22.1	5.3
Tet TKO	R1	1	2
Tet TKO	R2	0.9	2

Bibliography

1. Aapola U., Kawasaki K., Scott H. S., et al. Isolation and initial characterization of a novel zinc finger gene, DNMT3L, on 21q22.3, related to the cytosine-5-methyltransferase 3 gene family. *Genomics*, 65(3):293–298, May 2000.

2. Aberg K. A., McClay J. L., Nerella S., et al. MBD-seq as a cost-effective approach for methylome-wide association studies: demonstration in 1500 case–control samples. *Epigenomics*, 4(6):605–621, Dec 2012.

3. Aberg K. A., Chan R. F., Shabalin A. A., et al. A MBD-seq protocol for large-scale methylome-wide studies with (very) low amounts of DNA. *Epigenetics*, 12(9):743–750, 09 2017.

4. Alfirevic Z., Sundberg K., and Brigham S. Amniocentesis and chorionic villus sampling for prenatal diagnosis. *Cochrane Database Syst Rev*, (3):CD003252, 2003.

5. Allis C. D. and Jenuwein T. The molecular hallmarks of epigenetic control. *Nat Rev Genet*, 17(8):487–500, 08 2016.

6. Antonarakis S. E., Skotko B. G., Rafii M. S., et al. Down syndrome. *Nat Rev Dis Primers*, 6(1):9, 02 2020.

7. Arima T., Hata K., Tanaka S., et al. Loss of the maternal imprint in Dnmt3Lmat-/- mice leads to a differentiation defect in the extraembryonic tissue. *Dev Biol*, 297(2):361–373, Sep 2006.

8. Atlasi Y. and Stunnenberg H. G. The interplay of epigenetic marks during stem cell differentiation and development. *Nat Rev Genet*, 18(11):643–658, 11 2017.

9. Babraham Bioinformatics. Fastqc - a quality control tool for high throughput sequence data. `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`, 2019. Accessed: 2020-08-26.

10. Barau J., Teissandier A., Zamudio N., et al. The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. *Science*, 354(6314):909–912, 11 2016.

11. Barlow D. P. and Bartolomei M. S. Genomic imprinting in mammals. *Cold Spring Harb Perspect Biol*, 6(2), Feb 2014.

12. Bell E., Curry E. W., Megchelenbrink W., et al. Dynamic CpG methylation delineates subregions within super-enhancers selectively decommissioned at the exit from naive pluripotency. *Nat Commun*, 11(1):1112, 02 2020.

13. Bell J. S. K. and Vertino P. M. Orphan CpG islands define a novel class of highly active enhancers. *Epigenetics*, 12(6):449–464, 06 2017.

14. Bellman R. *Adaptive Control Processes*. Princeton University Press, 1961.

15. Berman B. P., Weisenberger D. J., Aman J. F., et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat Genet*, 44(1):40–46, Nov 2011.

16. Berney M. and McGouran J. F. Methods for detection of cytosine and thymine modifications in dna. *Nature Reviews Chemistry*, 2 (11):332–348, 2018.

17. Bernstein B. E., Meissner A., and Lander E. S. The mammalian epigenome. *Cell*, 128(4):669–681, Feb 2007.

18. Bernstein B. E., Stamatoyannopoulos J. A., Costello J. F., et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol*, 28(10):1045–1048, Oct 2010.

19. Bert S. A., Robinson M. D., Strbenac D., et al. Regional activation of the cancer genome by long-range epigenetic remodeling. *Cancer Cell*, 23(1):9–22, Jan 2013.

20. Bestor T., Laudano A., Mattaliano R., and Ingram V. Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *J Mol Biol*, 203 (4):971–983, Oct 1988.

21. Bestor T. H. The DNA methyltransferases of mammals. *Hum Mol Genet*, 9(16):2395–2402, Oct 2000.

22. Bestor T. H., Edwards J. R., and Boulard M. Notes on the role of dynamic DNA methylation in mammalian development. *Proc Natl Acad Sci U S A*, 112(22):6796–6799, Jun 2015.

23. Bewick A. J., Hofmeister B. T., Powers R. A., et al. Diversity of cytosine methylation across the fungal tree of life. *Nat Ecol Evol*, 3(3):479–490, 03 2019.

24. Bibikova M., Lin Z., Zhou L., et al. High-throughput DNA methylation profiling using universal bead arrays. *Genome Res*, 16(3): 383–393, Mar 2006.

25. Bibikova M., Le J., Barnes B., et al. Genome-wide DNA methylation profiling using Infinium assay. *Epigenomics*, 1(1):177–200, Oct 2009.

26. Bibikova M., Barnes B., Tsan C., et al. High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4): 288–295, Oct 2011.

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

28. Bird A. Perceptions of epigenetics. *Nature*, 447(7143):396–398, May 2007.

29. Bishop C. M. *Pattern Recognition and Machine Learning*. Springer Verlag, New York, 2006.

30. Booth M. J., Branco M. R., Ficz G., et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science*, 336(6083):934–937, May 2012.

31. Bouyer D., Kramdi A., Kassam M., et al. DNA methylation dynamics during early plant life. *Genome Biol*, 18(1):179, 09 2017.

32. Branco M. R., Oda M., and Reik W. Safeguarding parental identity: Dnmt1 maintains imprints during epigenetic reprogramming in early embryogenesis. *Genes Dev*, 22(12):1567–1571, Jun 2008.

33. Branco M. R., Ficz G., and Reik W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nat Rev Genet*, 13(1): 7–13, Nov 2011.

34. Bray J. R. and Curtis J. T. An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, 27(4): 325–349, Feb. 1957.

35. Breiling A. and Lyko F. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics Chromatin*, 8:24, 2015.

36. Brodsky G., Barnes T., Bleskan J., et al. The human GARS-AIRS-GART gene encodes two proteins which are differentially expressed during human brain development and temporally overexpressed in cerebellum of individuals with Down syndrome. *Hum. Mol. Genet.*, 6(12):2043–2050, Nov 1997.

37. Brown K. D. and Robertson K. D. DNMT1 knockout delivers a strong blow to genome stability and cell viability. *Nat Genet*, 39 (3):289–290, Mar 2007.

38. Cairns J. Mutation selection and the natural history of cancer. *Nature*, 255(5505):197–200, May 1975.

39. Cattell R. B. The Scree Test For The Number Of Factors. *Multivariate Behav Res*, 1(2):245–276, Apr 1966.

40. Chambers J., Hastie T., and Pregibon D. Statistical models in s. In *Compstat*, pages 317–321. Physica-Verlag HD, 1990.

41. Chappell L., Russell A. J. C., and Voet T. Single-Cell (Multi)omics Technologies. *Annu Rev Genomics Hum Genet*, 19:15–41, 08 2018.

42. Chen X., Xu H., Yuan P., et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117, Jun 2008.

43. Chen Z. X., Mann J. R., Hsieh C. L., et al. Physical and functional interactions between the human DNMT3L protein and members of the de novo methyltransferase family. *J Cell Biochem*, 95(5):902–917, Aug 2005.

44. Chédin F. The DNMT3 family of mammalian de novo DNA methyltransferases. *Prog Mol Biol Transl Sci*, 101:255–285, 2011.

45. Clark S. J., Harrison J., Paul C. L., and Frommer M. High sensitivity mapping of methylated cytosines. *Nucleic Acids Res*, 22(15): 2990–2997, Aug 1994.

46. Cokus S. J., Feng S., Zhang X., et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, 452(7184):215–219, Mar 2008.

47. Cooper G. M., Stone E. A., Asimenos G., et al. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*, 15(7):901–913, Jul 2005.

48. Cortellino S., Xu J., Sannai M., et al. Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell*, 146(1):67–79, Jul 2011.

49. Cui X. L., Nie J., Ku J., et al. A human tissue map of 5-hydroxymethylcytosines exhibits tissue specificity through gene and enhancer modulation. *Nat Commun*, 11(1):6161, 12 2020.

50. Dahlet T., Argüeso Lleida A., Al Adhami H., et al. Genome-wide analysis in the mouse embryo reveals the importance of DNA methylation for transcription integrity. *Nat Commun*, 11(1):3153, 06 2020.

51. Davis C. A., Hitz B. C., Sloan C. A., et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res*, 46(D1):D794–D801, 01 2018.

52. Davydov E. V., Goode D. L., Sirota M., et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol*, 6(12):e1001025, Dec 2010.

53. Deng J., Shoemaker R., Xie B., et al. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat Biotechnol*, 27(4):353–360, Apr 2009.

54. Derrien T., Estellé J., Marco Sola S., et al. Fast computation and applications of genome mappability. *PLoS One*, 7(1):e30377, 2012.

55. Diep D., Plongthongkum N., Gore A., et al. Library-free methylation sequencing with bisulfite padlock probes. *Nat Methods*, 9(3): 270–272, Feb 2012.

56. Doskocil J. and Sorm F. Distribution of 5-methylcytosine in pyrimidine sequences of deoxyribonucleic acids. *Biochim Biophys Acta*, 55:953–959, Jun 1962.

57. Douvlataniotis K., Bensberg M., Lentini A., et al. No evidence for DNA N6-methyladenine in mammals. *Sci Adv*, 6(12):eaay3335, 03 2020.

58. Dowle M. and Srinivasan A. *data.table: Extension of 'data.frame'*, 2020. R package version 1.13.6.

59. Duncan B. K. and Miller J. H. Mutagenic deamination of cytosine residues in DNA. *Nature*, 287(5782):560–561, Oct 1980.

60. Duncan C. G., Grimm S. A., Morgan D. L., et al. Dosage compensation and DNA methylation landscape of the X chromosome in mouse liver. *Sci Rep*, 8(1):10138, 07 2018.

61. Dunham I., Kundaje A., Aldred S. F., et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414): 57–74, Sep 2012.

62. Dyson C. J. and Goodisman M. A. D. Gene Duplication in the Honeybee: Patterns of DNA Methylation, Gene Expression, and Genomic Environment. *Mol Biol Evol*, 37(8):2322–2331, 08 2020.

63. Ebbert M. T., Wadsworth M. E., Staley L. A., et al. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*, 17 Suppl 7:239, Jul 2016.

64. Edgar R., Domrachev M., and Lash A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–210, Jan 2002.

65. Ehrlich M., Gama-Sosa M. A., Huang L. H., et al. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res*, 10(8):2709–2721, Apr 1982.

66. Epanechnikov V. A. Non-parametric estimation of a multivariate probability density. *Theory of Probability and Its Applications*, 14 (1):153–158, 1969.

67. Farley B. and Clark W. Simulation of self-organizing systems by digital computer. *Transactions of the IRE Professional Group on Information Theory*, 4(4):76–84, 1954.

68. Ficz G., Branco M. R., Seisenberger S., et al. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature*, 473(7347):398–402, May 2011.

69. Flusberg B. A., Webster D. R., Lee J. H., et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods*, 7(6):461–465, Jun 2010.

70. Fraga M. F., Ballestar E., Paz M. F., et al. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A*, 102(30):10604–10609, Jul 2005.

71. Frankish A., Diekhans M., Ferreira A.-M., et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1):D766–D773, Oct. 2018.

72. Friedman J., Hastie T., and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 2010.

73. Frommer M., McDonald L. E., Millar D. S., et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A*, 89 (5):1827–1831, Mar 1992.

74. Gao L., Emperle M., Guo Y., et al. Comprehensive structure-function characterization of DNMT3B and DNMT3A reveals distinctive de novo DNA methylation mechanisms. *Nat Commun*, 11 (1):3355, 07 2020.

75. Gao T., He B., Liu S., et al. EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. *Bioinformatics*, 32(23):3543–3551, 12 2016.

76. Gardiner-Garden M. and Frommer M. CpG islands in vertebrate genomes. *J Mol Biol*, 196(2):261–282, Jul 1987.

77. Gaunt T. R., Shihab H. A., Hemani G., et al. Systematic identification of genetic influences on methylation across the human life course. *Genome Biology*, 17(1), Mar. 2016.

78. Gautier F., Bünemann H., and Grotjahn L. Analysis of calf-thymus satellite DNA: evidence for specific methylation of cytosine in C-G sequences. *Eur J Biochem*, 80(1):175–183, Oct 1977.

79. Gelman A., Jakulin A., Pittau M. G., and Su Y.-S. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.

80. Gibas P., Narmontė M., Staševskij Z., et al. Precise genomic mapping of 5-hydroxymethylcytosine via covalent tether-directed sequencing. *PLoS Biol*, 18(4):e3000684, 04 2020.

81. Gil M. M., Quezada M. S., Revello R., et al. Analysis of cell-free DNA in maternal blood in screening for fetal aneuploidies: updated meta-analysis. *Ultrasound Obstet Gynecol*, 45(3):249–266, 03 2015.

82. Globisch D., Münzel M., Müller M., et al. Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS One*, 5(12):e15367, Dec 2010.

83. Goll M. G. and Bestor T. H. Eukaryotic cytosine methyltransferases. *Annu Rev Biochem*, 74:481–514, 2005.

84. Goll M. G., Kirpekar F., Maggert K. A., et al. Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science*, 311(5759):395–398, Jan 2006.

85. Gordevičius J., Gamper J., and Böhlen M. Parsimonious temporal aggregation. page 1006–1017, 2009.

86. Gordevičius J., Kriščiūnas A., Groot D. E., et al. Cell-Free DNA Modification Dynamics in Abiraterone Acetate-Treated Prostate Cancer Patients. *Clin Cancer Res*, 24(14):3317–3324, 07 2018.

87. Gordevičius J., Narmontė M., Gibas P., et al. Identification of fetal unmodified and 5-hydroxymethylated CG sites in maternal cell-free DNA for non-invasive prenatal testing. *Clin Epigenetics*, 12(1):153, Oct 2020.

88. Grassberger P. and Procaccia I. Characterization of strange attractors. *Phys. Rev. Lett.*, 50:346–349, Jan 1983.

89. Greally J. M. A user's guide to the ambiguous word 'epigenetics'. *Nat Rev Mol Cell Biol*, 19(4):207–208, 04 2018.

90. Gruenbaum Y., Cedar H., and Razin A. Substrate and sequence specificity of a eukaryotic DNA methylase. *Nature*, 295(5850):620–622, Feb 1982.

91. Grunau C., Clark S. J., and Rosenthal A. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Res*, 29(13):E65–65, Jul 2001.

92. Gu H., Smith Z. D., Bock C., et al. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat Protoc*, 6(4):468–481, Apr 2011a.

93. Gu T. P., Guo F., Yang H., et al. The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature*, 477(7366):606–610, Sep 2011b.

94. Guelen L., Pagie L., Brasset E., et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197):948–951, Jun 2008.

95. Guo H., Zhu P., Yan L., et al. The DNA methylation landscape of human early embryos. *Nature*, 511(7511):606–610, Jul 2014a.

96. Guo J. U., Su Y., Shin J. H., et al. Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat Neurosci*, 17(2):215–222, Feb 2014b.

97. Ha K., Lee G. E., Palii S. S., et al. Rapid and transient recruitment of DNMT1 to DNA double-strand breaks is mediated by its interaction with multiple components of the DNA damage response machinery. *Hum Mol Genet*, 20(1):126–140, Jan 2011.

98. Habibi E., Brinkman A. B., Arand J., et al. Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell*, 13(3):360–369, 2013.

99. Haig D. The (dual) origin of epigenetics. *Cold Spring Harb Symp Quant Biol*, 69:67–70, 2004.

100. Hannon Lab. Fastx-toolkit - fastq/a short-reads preprocessing tools. `http://hannonlab.cshl.edu/fastx_toolkit/index.html`, 2010. Accessed: 2020-08-26.

101. Hardenbol P., Banér J., Jain M., et al. Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat Biotechnol*, 21(6):673–678, Jun 2003.

102. Hayatsu H., Wataya Y., Kai K., and Iida S. Reaction of sodium bisulfite with uracil, cytosine, and their derivatives. *Biochemistry*, 9(14):2858–2865, Jul 1970.

103. He Y. and Ecker J. R. Non-CG Methylation in the Human Genome. *Annu Rev Genomics Hum Genet*, 16:55–77, 2015.

104. He Y. F., Li B. Z., Li Z., et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science*, 333(6047):1303–1307, Sep 2011.

105. Hernandez Mora J. R., Sanchez-Delgado M., Petazzi P., et al. Profiling of oxBS-450K 5-hydroxymethylcytosine in human placenta and brain reveals enrichment at imprinted loci. *Epigenetics*, 13(2): 182–191, 2018.

106. Hinrichs A. S., Karolchik D., Baertsch R., et al. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*, 34(Database issue):D590–598, Jan 2006.

107. Holliday R. and Pugh J. E. DNA modification mechanisms and gene activity during development. *Science*, 187(4173):226–232, 1975.

108. Howell C. Y., Bestor T. H., Ding F., et al. Genomic imprinting disrupted by a maternal effect mutation in the Dnmt1 gene. *Cell*, 104(6):829–838, Mar 2001.

109. Hu L., Liu Y., Han S., et al. Jump-seq: Genome-Wide Capture and Amplification of 5-Hydroxymethylcytosine Sites. *J Am Chem Soc*, 141(22):8694–8697, 06 2019.

110. Huang D. W., Sherman B. T., and Lempicki R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2008.

111. Huang Y., Pastor W. A., Shen Y., et al. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS One*, 5(1): e8888, Jan 2010.

112. Huh Y. H., Cohen J., and Sherley J. L. Higher 5-hydroxymethylcytosine identifies immortal DNA strand chromosomes in asymmetrically self-renewing distributed stem cells. *Proc Natl Acad Sci U S A*, 110(42):16862–16867, Oct 2013.

113. Illingworth R. S. and Bird A. P. CpG islands–'a rough guide'. *FEBS Lett*, 583(11):1713–1720, Jun 2009.

114. Illingworth R. S., Gruenewald-Schneider U., Webb S., et al. Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet*, 6(9):e1001134, Sep 2010.

115. Ito S., D'Alessio A. C., Taranova O. V., et al. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*, 466(7310):1129–1133, Aug 2010.

116. Ito S., Shen L., Dai Q., et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, 333(6047):1300–1303, Sep 2011.

117. Iurlaro M., Ficz G., Oxley D., et al. A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol*, 14(10): R119, 2013.

118. Izenman A. J. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, June 1975.

119. Jeong M., Sun D., Luo M., et al. Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nat Genet*, 46(1):17–23, Jan 2014.

120. Jiang Y. L., Rigolet M., Bourc'his D., et al. DNMT3B mutations and DNA methylation defect define two types of ICF syndrome. *Hum Mutat*, 25(1):56–63, Jan 2005.

121. Jones P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*, 13(7):484–492, May 2012.

122. Jurka J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet*, 16(9):418–420, Sep 2000.

123. Kagiwada S., Kurimoto K., Hirota T., et al. Replication-coupled passive DNA demethylation for the erasure of genome imprints in mice. *EMBO J*, 32(3):340–353, Feb 2013.

124. Karolchik D. The UCSC table browser data retrieval tool. *Nucleic Acids Research*, 32(90001):493D–496, 2004.

125. Keller T. E., Han P., and Yi S. V. Evolutionary Transition of Promoter and Gene Body DNA Methylation across Invertebrate-Vertebrate Boundary. *Mol Biol Evol*, 33(4):1019–1028, Apr 2016.

126. Kent W. J., Sugnet C. W., Furey T. S., et al. The human genome browser at UCSC. *Genome Res*, 12(6):996–1006, Jun 2002.

127. Khare T., Pai S., Koncevicius K., et al. 5-hmC in the brain is abundant in synaptic genes and shows differences at the exon-intron boundary. *Nat Struct Mol Biol*, 19(10):1037–1043, Oct 2012.

128. Kim S. K., Hannum G., Geis J., et al. Determination of fetal DNA fraction from the plasma of pregnant women using sequence read counts. *Prenat Diagn*, 35(8):810–815, Aug 2015.

129. Kishikawa S., Murata T., Ugai H., et al. Control elements of dnmt1 gene are regulated in cell-cycle dependent manne. *Nucleic acids research. Supplement (2001)*, 3:307–308, 2003.

130. Kivioja T., Vähärautio A., Karlsson K., et al. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, 9(1):72–74, Nov 2011.

131. Knuth D. E., T. L., and Roberts P. M. Mathematical writing. *Mathematical Association of America*, 1989.

132. Kobak D. and Berens P. The art of using t-SNE for single-cell transcriptomics. *Nat Commun*, 10(1):5416, 11 2019.

133. Koerner M. V., Chhatbar K., Webb S., et al. An Orphan CpG Island Drives Expression of a let-7 miRNA Precursor with an Important Role in Mouse Development. *Epigenomes*, 3(1):7, Mar 2019.

134. Kriaucionis S. and Heintz N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*, 324(5929):929–930, May 2009.

135. Kriukienė E., Liutkevičiūtė Z., and Klimašauskas S. 5-Hydroxymethylcytosine–the elusive epigenetic mark in mammalian DNA. *Chem Soc Rev*, 41(21):6916–6930, Nov 2012.

136. Kriukienė E., Labrie V., Khare T., et al. DNA unmethylome profiling by covalent capture of CpG sites. *Nat Commun*, 4:2190, 2013.

137. Krogh A. What are artificial neural networks? *Nat Biotechnol*, 26 (2):195–197, Feb 2008.

138. Kumar S., Chinnusamy V., and Mohapatra T. Epigenetics of Modified DNA Bases: 5-Methylcytosine and Beyond. *Front Genet*, 9: 640, 2018.

139. Kundaje A., Meuleman W., Ernst J., et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, Feb 2015.

140. Kurihara Y., Kawamura Y., Uchijima Y., et al. Maintenance of genomic methylation patterns during preimplantation development requires the somatic form of DNA methyltransferase 1. *Dev Biol*, 313(1):335–346, Jan 2008.

141. Kurtz A. K. A research test of Rorschach test. Personnel Psychology. *Personnel Psychology*, 1948.

142. Kutner M. H., Nachtsheim C. J., Neter J., and Li W. *Applied Linear Statistical Models (5th ed.)*. McGraw-Hill Irwin, Boston, 2004.

143. Kweon S. M., Chen Y., Moon E., et al. An Adversarial DNA N6-Methyladenine-Sensor Network Preserves Polycomb Silencing. *Mol Cell*, 74(6):1138–1147, 06 2019.

144. Lan X., Adams C., Landers M., et al. High resolution detection and analysis of CpG dinucleotides methylation using MBD-Seq technology. *PLoS One*, 6(7):e22226, 2011.

145. Lander E. S., Linton L. M., Birren B., et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

146. Larsen F., Gundersen G., Lopez R., and Prydz H. CpG islands as gene markers in the human genome. *Genomics*, 13(4):1095–1107, Aug 1992.

147. Laurent L., Wong E., Li G., et al. Dynamic changes in the human methylome during differentiation. *Genome Res*, 20(3):320–331, Mar 2010.

148. Lazutka J. R. Sister chromatid exchanges (SCEs) and high frequency cells (HFCs) in human population studies: principles of their analysis. *Mutat Res*, 331(2):229–231, Oct 1995.

149. Lee H. J., Hore T. A., and Reik W. Reprogramming the methylome: erasing memory and creating diversity. *Cell Stem Cell*, 14 (6):710–719, Jun 2014.

150. Legendre P. and Gallagher E. D. Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129(2):271–280, Oct. 2001.

151. Lei H., Oh S. P., Okano M., et al. De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. *Development*, 122(10):3195–3205, Oct 1996.

152. Li E. and Zhang Y. DNA methylation in mammals. *Cold Spring Harb Perspect Biol*, 6(5):a019133, May 2014.

153. Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009.

154. Li H., Handsaker B., Wysoker A., et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

155. Li N., Ye M., Li Y., et al. Whole genome DNA methylation analysis based on high throughput sequencing technology. *Methods*, 52(3): 203–212, Nov 2010.

156. Li W. and Liu M. Distribution of 5-hydroxymethylcytosine in different human tissues. *J Nucleic Acids*, 2011:870726, 2011.

157. Lima F. A., Moreira-Filho C. A., Ramos P. L., et al. Decreased AIRE expression and global thymic hypofunction in Down syndrome. *J. Immunol.*, 187(6):3422–3430, Sep 2011.

158. Lindsey J. K. *Applying Generalized Linear Models*. Springer, 1997.

159. Lister R., Pelizzola M., Dowen R. H., et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, Nov 2009.

160. Lister R., Mukamel E. A., Nery J. R., et al. Global epigenomic reconfiguration during mammalian brain development. *Science*, 341(6146):1237905, Aug 2013.

161. Liutkevičiūtė Z., Kriukienė E., Ličytė J., et al. Direct decarboxylation of 5-carboxylcytosine by DNA C5-methyltransferases. *J Am Chem Soc*, 136(16):5884–5887, Apr 2014.

162. Ličytė J., Gibas P., Skardžiūtė K., et al. A Bisulfite-free Approach for Base-Resolution Analysis of Genomic 5-Carboxylcytosine. *Cell Rep*, 32(11):108155, Sep 2020.

163. Lo Y. M., Corbetta N., Chamberlain P. F., et al. Presence of fetal DNA in maternal plasma and serum. *Lancet*, 350(9076):485–487, Aug 1997.

164. Lo Y. M., Chan K. C., Sun H., et al. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci Transl Med*, 2(61):61ra91, Dec 2010.

165. Loftsgaarden D. O. and Quesenberry C. P. A nonparametric estimate of a multivariate density function. *Ann. Math. Statist.*, 36 (3):1049–1051, 06 1965.

166. Lu J., Mccarter M., Lian G., et al. Global hypermethylation in fetal cortex of Down syndrome due to DNMT3L overexpression. *Hum. Mol. Genet.*, 25(9):1714–1727, 05 2016.

167. Lu X., Song C. X., Szulwach K., et al. Chemical modification-assisted bisulfite sequencing (CAB-Seq) for 5-carboxylcytosine detection in DNA. *J Am Chem Soc*, 135(25):9315–9317, Jun 2013.

168. Lyko F. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nat Rev Genet*, 19(2):81–92, 02 2018.

169. Maiti A. and Drohat A. C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J Biol Chem*, 286(41): 35334–35338, Oct 2011.

170. Martello G. and Smith A. The nature of embryonic stem cells. *Annual Review of Cell and Developmental Biology*, 30(1):647–675, 2014.

171. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, 2011. ISSN 2226-6089.

172. Marx V. How to deduplicate PCR. *Nat. Methods*, 14(5):473–476, 04 2017.

173. Maunakea A. K., Nagarajan R. P., Bilenky M., et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, 466(7303):253–257, Jul 2010.

174. Maurage P., Heeren A., and Pesenti M. Does chocolate consumption really boost Nobel Award chances? The peril of over-interpreting correlations in health studies. *J Nutr*, 143(6):931–933, Jun 2013.

175. Medvedeva Y. A., Fridman M. V., Oparina N. J., et al. Intergenic, gene terminal, and intragenic CpG islands in the human genome. *BMC Genomics*, 11:48, Jan 2010.

176. Meissner A., Gnirke A., Bell G. W., et al. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res*, 33(18):5868–5877, 2005.

177. Meissner A., Mikkelsen T. S., Gu H., et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770, Aug 2008.

178. Mitchell T. M. *Machine Learning*. The McGraw-Hill Companies, Inc., New York, international edition 1997 edition, 1997.

179. Molaro A., Malik H. S., and Bourc'his D. Dynamic Evolution of De Novo DNA Methyltransferases in Rodent and Primate Genomes. *Mol Biol Evol*, 37(7):1882–1892, 07 2020.

180. Mooijman D., Dey S. S., Boisset J. C., et al. Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. *Nat Biotechnol*, 34(8):852–856, 08 2016.

181. Moran S., Arribas C., and Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*, 8(3):389–399, Mar 2016.

182. Moris N., Pina C., and Arias A. M. Transition states and cell fate decisions in epigenetic landscapes. *Nat Rev Genet*, 17(11):693–703, 11 2016.

183. Morris K. V. and Mattick J. S. The rise of regulatory RNA. *Nat Rev Genet*, 15(6):423–437, Jun 2014.

184. Murphy K. P. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning. The MIT Press, 1 edition, 2012. ISBN 0262018020,9780262018029.

185. Naeem H., Wong N. C., Chatterton Z., et al. Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. *BMC Genomics*, 15(1):51, 2014.

186. Narasimhan V. M., Rahbari R., Scally A., et al. Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nat Commun*, 8(1):303, 08 2017.

187. Nelder J. A. and Wedderburn R. W. M. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3): 370–384, 1972. ISSN 00359238.

188. Neri F., Incarnato D., Krepelova A., et al. Single-Base Resolution Analysis of 5-Formyl and 5-Carboxyl Cytosine Reveals Promoter DNA Methylation Dynamics. *Cell Rep*, 10(5):674–683, Feb 2015.

189. Neri F., Rapelli S., Krepelova A., et al. Intragenic DNA methylation prevents spurious transcription initiation. *Nature*, 543(7643): 72–77, 03 2017.

190. Nestor C. E. and Meehan R. R. Hydroxymethylated DNA immunoprecipitation (hmeDIP). *Methods Mol Biol*, 1094:259–267, 2014.

191. Nilsson M., Malmgren H., Samiotaki M., et al. Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science*, 265(5181):2085–2088, Sep 1994.

192. Norton M. E. and Wapner R. J. Cell-free DNA Analysis for Noninvasive Examination of Trisomy. *N Engl J Med*, 373(26):2582, 12 2015.

193. Norvil A. B., Saha D., Dar M. S., and Gowher H. Effect of Disease-Associated Germline Mutations on Structure Function Relationship of DNA Methyltransferases. *Genes (Basel)*, 10(5), 05 2019.

194. Nowialis P., Lopusna K., Opavska J., et al. Catalytically inactive Dnmt3b rescues mouse embryonic development by accessory and repressive functions. *Nat Commun*, 10(1):4374, 09 2019.

195. Oh G., Ebrahimi S., Carlucci M., et al. Cytosine modifications exhibit circadian oscillations that are involved in epigenetic diversity and aging. *Nat Commun*, 9(1):644, 02 2018.

196. Ohno S., Kaplan W. D., and Kinosita R. Formation of the sex chromatin by a single X-chromosome in liver cells of Rattus norvegicus. *Exp Cell Res*, 18:415–418, Oct 1959.

197. Okano M., Bell D. W., Haber D. A., and Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257, Oct 1999.

198. Oswald J., Engemann S., Lane N., et al. Active demethylation of the paternal genome in the mouse zygote. *Curr Biol*, 10(8): 475–478, Apr 2000.

199. Palomaki G. E., Kloza E. M., Lambert-Messerlian G. M., et al. DNA sequencing of maternal plasma to detect Down syndrome: an international clinical validation study. *Genet Med*, 13(11):913–920, Nov 2011.

200. Parry A., Rulands S., and Reik W. Active turnover of DNA methylation during cell fate decisions. *Nat Rev Genet*, 22(1):59–66, 01 2021.

201. Parzen E. On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33(3):1065–1076, 09 1962.

202. Pastor W. A., Pape U. J., Huang Y., et al. Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature*, 473 (7347):394–397, May 2011.

203. Patro R., Duggal G., Love M. I., et al. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, 14(4):417–419, Apr 2017.

204. Pearson K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

205. Pedregosa F., Varoquaux G., Gramfort A., et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

206. Penn N. W., Suwalski R., O'Riley C., et al. The presence of 5-hydroxymethylcytosine in animal deoxyribonucleic acid. *Biochem J*, 126(4):781–790, Feb 1972.

207. Pfaffeneder T., Hackner B., Truss M., et al. The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew Chem Int Ed Engl*, 50(31):7008–7012, Jul 2011.

208. Ponger L., Duret L., and Mouchiroud D. Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res*, 11(11):1854–1860, Nov 2001.

209. Porreca G. J., Zhang K., Li J. B., et al. Multiplex amplification of large sets of human exons. *Nat Methods*, 4(11):931–936, Nov 2007.

210. Price M. E., Cotton A. M., Lam L. L., et al. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin*, 6(1):4, Mar 2013.

211. Probst A. V., Dunleavy E., and Almouzni G. Epigenetic inheritance during the cell cycle. *Nat Rev Mol Cell Biol*, 10(3):192–206, Mar 2009.

212. Quinlan A. R. and Hall I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, Mar 2010.

213. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

214. Ramsahoye B. H., Biniszkiewicz D., Lyko F., et al. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc Natl Acad Sci U S A*, 97(10):5237–5242, May 2000.

215. Rauluseviciute I., Drabløs F., and Rye M. B. DNA methylation data by sequencing: experimental approaches and recommendations for tools and pipelines for data analysis. *Clin Epigenetics*, 11 (1):193, 12 2019.

216. Razin A. and Riggs A. D. DNA methylation and gene function. *Science*, 210(4470):604–610, Nov 1980.

217. Reik W. and Walter J. Genomic imprinting: parental influence on the genome. *Nat Rev Genet*, 2(1):21–32, Jan 2001.

218. Reik W., Dean W., and Walter J. Epigenetic reprogramming in mammalian development. *Science*, 293(5532):1089–1093, Aug 2001.

219. Riggs A. D. X inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet*, 14(1):9–25, 1975.

220. Ritchie M. E., Phipson B., Wu D., et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, Jan. 2015.

221. Robinson M. D., Stirzaker C., Statham A. L., et al. Evaluation of affinity-based genome-wide DNA methylation data: effects of CpG density, amplification bias, and copy number variation. *Genome Res*, 20(12):1719–1729, Dec 2010.

222. Rougier N., Bourc'his D., Gomes D. M., et al. Chromosome methylation patterns during mammalian preimplantation development. *Genes Dev*, 12(14):2108–2113, Jul 1998.

223. Roweis S. T. and Saul L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, Dec 2000.

224. Russo V. E. A., Martienssen R. A., and Riggs A. D. Epigenetic mechanisms of gene regulation. 1996.

225. Saitou M., Kagiwada S., and Kurimoto K. Epigenetic reprogramming in mouse pre-implantation development and primordial germ cells. *Development*, 139(1):15–31, Jan 2012.

226. Salomon L. J., Sotiriadis A., Wulff C. B., et al. Risk of miscarriage following amniocentesis or chorionic villus sampling: systematic review of literature and updated meta-analysis. *Ultrasound Obstet Gynecol*, 54(4):442–451, Oct 2019.

227. Sandoval J., Heyn H., Moran S., et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*, 6(6):692–702, Jun 2011.

228. Sarda S. and Hannenhalli S. Orphan CpG islands as alternative promoters. *Transcription*, 9(3):171–176, 2018.

229. Saxonov S., Berg P., and Brutlag D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A*, 103(5):1412–1417, Jan 2006.

230. Schaefer M., Pollex T., Hanna K., et al. RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes Dev*, 24(15):1590–1595, Aug 2010.

231. Schölkopf B., Smola A., and Müller K.-R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10 (5):1299–1319, 1998.

232. Scott D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics. Wiley, 2 edition, 2015. ISBN 0471697559,9780471697558.

233. Scott D. W. and Terrell G. R. Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, 82(400):1131–1146, 1987. ISSN 01621459.

234. Sender R., Fuchs S., and Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol*, 14(8): e1002533, 08 2016.

235. Sharif J., Muto M., Takebayashi S., et al. The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature*, 450(7171):908–912, Dec 2007.

236. Sharp A. J., Stathaki E., Migliavacca E., et al. DNA methylation profiles of human active and inactive X chromosomes. *Genome Res*, 21(10):1592–1600, Oct 2011.

237. Sheather S. J. and Jones M. C. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):683–690, 1991. ISSN 00359246.

238. Shen L., Wu H., Diep D., et al. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell*, 153(3):692–706, Apr 2013.

239. Sherley J. L. A new mechanism for aging: chemical "age spots" in immortal DNA strands in distributed stem cells. *Breast Dis*, 29: 37–46, 2008.

240. Sherry S. T., Ward M. H., Kholodov M., et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, 29(1):308–311, Jan 2001.

241. Shi T., Rahmani R. S., Gugger P. F., et al. Distinct Expression and Methylation Patterns for Genes with Different Fates following a Single Whole-Genome Duplication in Flowering Plants. *Mol Biol Evol*, 37(8):2394–2413, 08 2020.

242. Silverman B. W. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986. ISBN 0-41224620-1.

243. Slotkin R. K. and Martienssen R. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet*, 8(4):272–285, Apr 2007.

244. Smith Z. D. and Meissner A. DNA methylation: roles in mammalian development. *Nat Rev Genet*, 14(3):204–220, Mar 2013.

245. Snyder M. P., Gingeras T. R., Moore J. E., et al. Perspectives on ENCODE. *Nature*, 583(7818):693–698, 07 2020.

246. Song C. X., Clark T. A., Lu X. Y., et al. Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. *Nat Methods*, 9(1):75–77, Nov 2011a.

247. Song C. X., Szulwach K. E., Fu Y., et al. Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat Biotechnol*, 29(1):68–72, Jan 2011b.

248. Song C. X., Szulwach K. E., Dai Q., et al. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell*, 153 (3):678–691, Apr 2013.

249. Spengler B. A., Lazarova D. L., Ross R. A., and Biedler J. L. Cell lineage and differentiation state are primary determinants of MYCN gene expression and malignant potential in human neuroblastoma cells. *Oncol. Res.*, 9(9):467–476, 1997.

250. Stadler M. B., Murr R., Burger L., et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480(7378):490–495, Dec 2011.

251. Staševskij Z., Gibas P., Gordevičius J., et al. Tethered Oligonucleotide-Primed Sequencing, TOP-Seq: A High-Resolution Economical Approach for DNA Epigenome Profiling. *Mol. Cell*, 65 (3):554–564, Feb 2017.

252. Stevens M., Cheng J. B., Li D., et al. Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome Res*, 23(9): 1541–1553, Sep 2013.

253. Stunnenberg H. G., Hirst M., Abrignani S., et al. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell*, 167(5):1145–1149, Nov 2016.

254. Su M., Kirchner A., Stazzoni S., et al. 5-Formylcytosine Could Be a Semipermanent Base in Specific Genome Sites. *Angew Chem Int Ed Engl*, 55(39):11797–11800, 09 2016.

255. Suetake I., Shinozaki F., Miyagawa J., et al. DNMT3L stimulates the DNA methylation activity of Dnmt3a and Dnmt3b through a direct interaction. *J Biol Chem*, 279(26):27816–27823, Jun 2004.

256. Suganuma T. and Workman J. L. Signals and combinatorial functions of histone modifications. *Annu Rev Biochem*, 80:473–499, 2011.

257. Sun Z., Vaisvila R., Hussong L. M., et al. Nondestructive enzymatic deamination enables single-molecule long-read amplicon sequencing for the determination of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Genome Res*, Jan 2021.

258. Suzuki M. M. and Bird A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet*, 9(6):465–476, Jun 2008.

259. Sved J. and Bird A. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc Natl Acad Sci U S A*, 87(12):4692–4696, Jun 1990.

260. Swarts M. N., Trautner T. A., and Kornberg A. Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids. *J Biol Chem*, 237: 1961–1967, Jun 1962.

261. Tahiliani M., Koh K. P., Shen Y., et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, 324(5929):930–935, May 2009.

262. Takamiya T., Hosobuchi S., Asai K., et al. Restriction landmark genome scanning method using isoschizomers (MspI/HpaII) for DNA methylation analysis. *Electrophoresis*, 27(14):2846–2856, Jul 2006.

263. Timp W. and Feinberg A. P. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nat Rev Cancer*, 13(7):497–510, 07 2013.

264. Tucci V., Isles A. R., Kelsey G., et al. Genomic Imprinting and Physiological Processes in Mammals. *Cell*, 176(5):952–965, 02 2019.

265. Urich M. A., Nery J. R., Lister R., et al. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nat Protoc*, 10(3):475–483, Mar 2015.

266. van Beek D. M., Straver R., Weiss M. M., et al. Comparing methods for fetal fraction determination and quality control of NIPT samples. *Prenat Diagn*, 37(8):769–773, Aug 2017.

267. Veland N., Lu Y., Hardikar S., et al. DNMT3L facilitates DNA methylation partly by maintaining DNMT3A stability in mouse embryonic stem cells. *Nucleic Acids Res*, 47(1):152–167, 01 2019.

268. Vilkaitis G., Suetake I., Klimašauskas S., and Tajima S. Processive methylation of hemimethylated CpG sites by mouse Dnmt1 DNA methyltransferase. *J Biol Chem*, 280(1):64–72, Jan 2005.

269. von Meyenn F., Berrens R. V., Andrews S., et al. Comparative Principles of DNA Methylation Reprogramming during Human and Mouse In Vitro Primordial Germ Cell Specification. *Dev Cell*, 39(1):104–115, 10 2016.

270. Waddington C. H. Canalization of development and the inheritance of acquired characters. *Nature*, 150(3811):563–565, 1942.

271. Waddington C. H. The cybernetics of development. 1957.

272. Watanabe D., Suetake I., Tada T., and Tajima S. Stage- and cell-specific expression of Dnmt3a and Dnmt3b during embryogenesis. *Mech Dev*, 118(1-2):187–190, Oct 2002.

273. Wen L., Li X., Yan L., et al. Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. *Genome Biol*, 15(3):R49, Mar 2014.

274. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4.

275. Wiener D. and Schwartz S. The epitranscriptome beyond m6A. *Nat Rev Genet*, Nov 2020.

276. Wilbanks E. G. and Facciotti M. T. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One*, 5(7):e11471, Jul 2010.

277. Williams K., Christensen J., Pedersen M. T., et al. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature*, 473(7347):343–348, May 2011.

278. Wolf S. F., Jolly D. J., Lunnen K. D., et al. Methylation of the hypoxanthine phosphoribosyltransferase locus on the human X chromosome: implications for X-chromosome inactivation. *Proc Natl Acad Sci U S A*, 81(9):2806–2810, May 1984.

279. Wu H., Wu X., Shen L., and Zhang Y. Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nat Biotechnol*, 32(12):1231–1240, Dec 2014.

280. Wu T. P., Wang T., Seetin M. G., et al. DNA methylation on N(6)-adenine in mammalian embryonic stem cells. *Nature*, 532(7599): 329–333, Apr 2016.

281. Xiao C.-L., Zhu S., He M., et al. N(6)-methyladenine dna modification in the human genome. *Molecular cell*, 71(2):306–318.e7, Jul 2018. ISSN 1097-4164. 30017583[pmid].

282. Xie W., Barr C. L., Kim A., et al. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell*, 148(4):816–831, Feb 2012.

283. Xie W., Schultz M. D., Lister R., et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*, 153(5):1134–1148, May 2013.

284. Yen R. W., Vertino P. M., Nelkin B. D., et al. Isolation and characterization of the cDNA encoding human DNA methyltransferase. *Nucleic Acids Res*, 20(9):2287–2291, May 1992.

285. Yoder J. A., Walsh C. P., and Bestor T. H. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet*, 13(8): 335–340, Aug 1997.

286. Yong W. S., Hsu F. M., and Chen P. Y. Profiling genome-wide DNA methylation. *Epigenetics Chromatin*, 9:26, 2016.

287. Yu M., Hon G. C., Szulwach K. E., et al. Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. *Nat Protoc*, 7(12):2159–2170, Dec 2012a.

288. Yu M., Hon G. C., Szulwach K. E., et al. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell*, 149 (6):1368–1380, Jun 2012b.

289. Zador A. M. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat Commun*, 10(1):3770, 08 2019.

290. Zhang L., Lu X., Lu J., et al. Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. *Nat Chem Biol*, 8(4): 328–330, Feb 2012.

291. Zhang Y., Liu T., Meyer C. A., et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, 9(9):R137, 2008.

292. Zhang Y., Zhang X., Shi J., et al. Dnmt2 mediates intergenerational transmission of paternally acquired metabolic disorders through sperm small non-coding RNAs. *Nat Cell Biol*, 20(5):535–540, 05 2018.

293. Zhao L., Sun M. A., Li Z., et al. The dynamics of DNA methylation fidelity during mouse embryonic stem cell self-renewal and differentiation. *Genome Res*, 24(8):1296–1307, Aug 2014.

294. Ziller M. J., Gu H., Müller F., et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500(7463): 477–481, Aug 2013.

Vilnius University Press
9 Saulėtekio Ave., Building III, LT-10222 Vilnius
Email: info@leidykla.vu.lt, www.leidykla.vu.lt
Print run 15 copies