DISSERTATIONES
INFORMATICAE
UNIVERSITATIS
TARTUENSIS
9

VIJAYACHITRA MODHUKUR

Profiling of DNA methylation patterns as biomarkers of human disease

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

9

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS
9

VIJAYACHITRA MODHUKUR

Profiling of DNA methylation patterns as biomarkers of human disease

UNIVERSITY OF TARTU
Press

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia.

Dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in informatics on April 17, 2019 by the Council of the Institute of Computer Science, University of Tartu.

Supervisor

Prof. PhD.	Jaak Vilo University of Tartu Estonia
PhD.	Balaji Rajashekar University of Tartu Estonia

Opponents

Prof. PhD	Stephan Beck University College London United Kingdom
Assoc. Prof. PhD	Anagha Joshi University of Bergen Norway

The public defense will take place on June 14, 2019 at 2.15 p.m in J. Liivi 2-405.

The Publication of this dissertation was financed by the Institute of Computer Science, University of Tartu.

ISSN 2613-5906
ISBN 978-9949-03-028-6 (print)
ISBN 978-9949-03-029-3 (PDF)

University of Tartu Press
http://www.tyk.ee/

To my family

CONTENTS

LIST OF PUBLICATIONS

Publications included in this thesis

I K. Lokk, **V. Modhukur**, B. Rajashekar, K. Märtens, R. Mägi, R. Kolde, M. Koltšina, T. K. Nilsson, J. Vilo, A. Salumets, and N. Tõnisson. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. ***Genome Biology***. 2014 Apr 1;15(4):1.

II M. Saare[1], **V. Modhukur**[1], M. Suhorutshenko, B. Rajashekar, K. Rekker, D. Sõritsa, H. Karro, P. Soplepmann, A. Sõritsa, C. M. Lindgren, N. Rahmioglu, A. Drong, C. M. Becker, K. T. Zondervan, A. Salumets, and M. Peters. The influence of menstrual cycle and endometriosis on endometrial methylome. ***Clinical Epigenetics***. 2016 Jan 12;8(1):1.

III V. Kukushkina[1], **V. Modhukur**[1], M. Suhorutshenko, M. Peters, R. Mägi, N. Rahmioglu, A. Velthut-Meikas, S. Altmäe, F. J. Esteban, J. Vilo, K. T. Zondervan, A. Salumets, and T. Laisk-Podar. DNA methylation changes in endometrium and correlation with gene expression during the transition from pre-receptive to receptive phase. ***Scientific Reports***. 2017 Jun 20;7(1):3916.

IV **V. Modhukur**, T. Iljasenko, T. Metsalu, K. Lokk, T. Laisk-Podar, and J. Vilo. MethSurv: A web-tool to perform multivariable survival analysis using DNA methylation data. ***Epigenomics***. 2018 Mar;10(3):277-288.

Publications not included in this thesis

I P. Pappu, D. Madduru, M. Chandrasekharan, **V. Modhukur**, S. Nallapeta, and P. Suravajhala. Next generation sequencing analysis of lung cancer datasets: A functional genomics perspective, ***Indian Journal of Cancer***. 2016 Jan 1;53(1):1.

II M. Lundborg, **V. Modhukur**, and G. Widmalm. Glycosyltransferase functions of E. Coli O-antigens, ***Glycobiology***. 2010 Mar 1;20(3):366-8.

[1]Equal contribution

ABSTRACT

Deoxyribonucleic acid (DNA) forms the blueprint of life, wherein the genetic information required for the growth and development of organism is stored. *Epigenetics* deals with the heritable phenotypic changes in the organism without altering the DNA sequence. It serves as a mediator between environment and genomes and controls which genes are turned on or off. The most studied epigenetic modification is the DNA methylation - the addition of a methyl group to the cytosine base of the DNA. Methylation patterns are crucial for the normal functioning of an organism and are susceptible to change owing to environmental changes, disease progression, and aging. Therefore, methylation patterns can be used as a biomarker indicating the normal biological process, disease progress, and prognosis. The current thesis aims to explore the role of DNA methylation in various human biological conditions and to predict biologically driven biomarker candidates by employing various computational and statistical methods.

In the first part of the thesis, three biological studies were considered to predict the DNA methylation-based biomarkers. The first study included several rare healthy tissues, whereas the following one is related to menstrual cycle-specific methylation patterns in endometriosis (a gynecological disease associated with endometrial dysfunction) patients and healthy controls. Subsequently, the final study pertains to the identification of methylation changes during the transition from pre-receptive to the receptive endometrium. A few of the methods used to achieve the aforementioned aim include customized data processing (normalization and batch effect correction), clustering analysis, differential methylation analysis, gene enrichment analysis, and integration with gene expression data combined with powerful data visualization approaches.

After predicting the DNA methylation biomarkers in selected biological conditions, we aimed to understand the applicability of DNA methylation patterns as a biomarker to predict cancer patients' survival in different cancer types. DNA methylation and clinical data relevant to this study were identified from "The Cancer Genome Atlas" (TCGA) database. In order to assist the scientific community to explore methylation-based prognostic biomarkers, *MethSurv* a web tool was developed. The technical features in the *MethSurv* include Cox-proportional hazard model fitting for understanding patients' survival, clustering and principal component analysis (PCA). The developed tool remains a prominent platform to integrate the statistical and computational tools on the DNA methylation data from TCGA. Further, it helps in facilitating the initial assessment of methylation-based cancer biomarkers.

In conclusion, the studies included in this thesis develop and combine numerous bioinformatics and statistical methods enabling a rapid and cost-effective way to identify DNA methylation patterns as a potential biomarker candidate.

INTRODUCTION

Epigenetics can be formally defined as "the study of heritable phenotype changes that do not involve alterations in the DNA sequence" (Dupont et al., 2009). DNA methylation patterns are known to control the regulation of gene expression through the epigenetic mechanism. Moreover, methylation patterns are associated with various biological processes such as embryonic development, genomic imprinting, and X-chromosome inactivation. Certain genomic methylation patterns are known to be associated with diseases such as diabetes, neurological disorders, and cancer development. Therefore, DNA methylation patterns are proposed as a biomarker candidate that indicates the normal biological process as well as disease progression. High-throughput technologies, such as microarrays or next-generation sequencing (NGS), provide methylation profiles on a genomic scale. Methylation microarrays, primarily the Illumina HumanMethylation450K (HM450K) array (Illumina Inc., San Diego, CA, USA), have been a popular choice owing to their cost-effectiveness and quantification in well-characterized regions of the human genome.

The current thesis aims to identify DNA methylation patterns as a potential biomarker candidate in healthy as well as disease states with the use of HM450K methylation data, by utilizing various computational and statistical approaches.

The first part of the thesis focuses on identifying DNA methylation biomarker candidates in healthy and diseased tissues using bioinformatic methods. For this, we utilized the methylation data generated from different project partners, which resulted in Publications I, II and III. To elaborate, in Publication I, using bioinformatic methods such as clustering analysis, linear regression, correlation analysis, and integration with gene expression data, we showed that methylation patterns can be used for tissue classification. On the other hand, in Publication II, the main analytical problem was the discovery and management of the batch effect, since the methylation data were generated from different labs and cohorts. Applying the *Empirical Bayes* method, the said effect was successfully corrected. Further, the methylation markers in endometriosis pathogenesis were evaluated using customized workflow from Publication I. In Publication III, bioinformatic methods obtained from the previous publications were further improved and tailored to identify methylation biomarkers associated with endometrial receptivity. Additionally, multiple differential methylation analysis methods were utilized and the overlap of the same were applied to identify biomarkers with higher confidence.

The second aim of this thesis was to identify survival biomarkers in different cancer types. For this purpose, methylation and clinical data were utilized from the TCGA consortium, which are described in Publication IV. Here, we developed an intuitive and exploratory web tool to assist scientists lacking bioinformatics, computational, or statistical skills. Furthermore, we processed multiple methylation and clinical datasets generated from different cancer types. The created web tool has multiple functionalities including analysis of methylation patterns in relation to patient survival in cancer types, visualization of methylation patterns, perform

cluster analysis, and the most important biomarkers for each type of cancer.

The current thesis is structured as follows:

In its first part, we provide the biological background on DNA methylation, summarize the methods used to measure methylation levels, and emphasize the role of methylation in human health and disease. Subsequently, we summarize different bioinformatic methods applied to further explore methylation biomarkers used in different studies included in this thesis (Lokk et al., 2014; Saare et al., 2016; Kukushkina et al., 2017; Modhukur et al., 2017) with a brief description of context-based methods. Chapters II–V summarize the publications included in this thesis. Finally, the thesis ends with the discussion highlighting strengthens and weakness of the study design, future perspectives and concluding remarks. Reprints of the Publications I–IV are included at the end of this thesis.

1. PRELIMINARIES

This chapter introduces some of the biological background as well as the bioinformatic methods that are necessary to understand the research related contribution of this thesis.

1.1. Biological background

1.1.1. The human genome and the epigenome

The hereditary information of living beings is stored and preserved by the DNA, which comprises four chemical bases, namely adenine (A), guanine (G), cytosine (C), and thymine (T), and is packaged inside the nucleus of the cell. The human genome consists of approximately 3.2 billion base pairs (bp). Furthermore, the DNA is compacted into discrete sections of different lengths called chromosomes. Humans have 46 chromosomes, among which 22 pairs are termed as autosomes, while the 23rd pair carries the sex chromosomes (X and Y). Moreover, the cell nucleus comprises two copies of chromosomes, each of which is inherited from the parents. The process through which it makes two identical copies of itself is called DNA replication. The flow of genetic information (central dogma) takes place from DNA to ribonucleic acid (RNA) (transcription) and subsequently from RNA to protein (termed as translation). This process occurs in all living organisms and forms the basis for biological inheritance. The genetic information derived from the DNA is read and processed in individual cell and tissue types differently. Epigenetic modification acts as an information control that governs the way in which DNA can be processed. Epigenetic modifications include DNA methylation, histone modifications, and RNA modifications that control the cellular phenotype by regulating the gene expression (Zaidi et al., 2010) (Figure 1). DNA methylation involves the addition of methyl group to the DNA. In the case of histone modification, histone proteins (which packages the DNA in nucleosome) are prone to various types of chemical modifications such as methylation and acetylation (Figure 1).

1.1.2. DNA methylation

Methylation that involves the addition of the methyl group at the C-5 of cytosine, resulting in 5mc (See Figure 2), termed as *CpG methylation* (where p denotes the phosphodiester bond). Methylation can also take place in the non-CpG context such as such as CpA, CpT, and CpC (Jang et al., 2017). In addition, *hydroxymethylation* replaces the C5-position in cytosine by a hydroxy methyl group (hm5C). The highest level of hydroxy methylation is known to occur in the brain (Lister et al., 2013; Guibert and Weber, 2013).

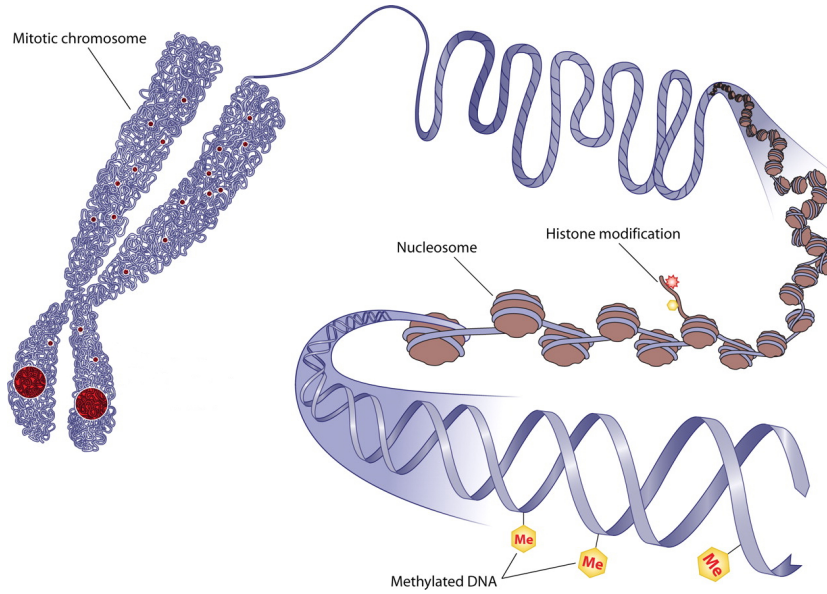

Figure 1. Epigenetic modifications. This figure depicts epigenetic modifications such as DNA methylation and histone modifications taking place in the chromosome of the cell. This figure has been adapted from Zaidi et al. (2010) and reprinted with permission from American Society for Microbiology.

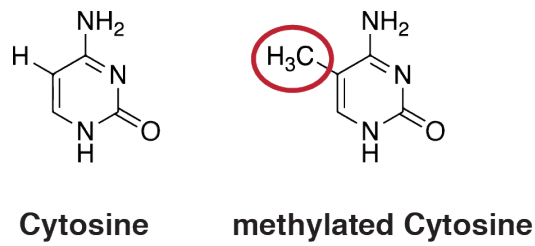

Figure 2. DNA methylation. The figure shows a methyl group (CH3) added to the 5th position of cytosine through DNA methyltransferase enzymes (DNMT), creating 5mc. This figure has been adapted from Wikimedia (2016), distributed under a CC BY-SA 4.0 license.

1.1.3. DNA methylation in health and disease

In this section, we provide a short overview of DNA methylation's role in health and disease, with a focus on the biological studies regarded in the current thesis.

Tissue-specific DNA methylation patterns. DNA methylation patterns are essential to direct the cells towards their respective lineages and ultimately leading to the development of a mammalian organism (Okano et al., 1999; Messerschmidt et al., 2014). Moreover, DNA methylation changes play a crucial role in establishing cell type or tissue-specific epigenomes (Reik, 2007). Furthermore, methylation patterns are suggested to be highly variable among different tissues of the same individual, in comparison to different organisms concerning the same tissue.

Age-related methylation. Aging is a natural process in human life representing the accumulation of changes over time (Bowen and Atwood, 2004). However, aging is also considered as a disease according to Bulterijs et al. (2015). Several studies suggested a strong correlation between age and DNA methylation levels (Jung and Pfeifer, 2015). Age-related methylation changes are partly caused by environmental changes (Gabbianelli and Damiani, 2018). Moreover, age-related methylation changes are also suggested to contribute for gene expression changes in diseases, particularly in type 2 diabetes (Nilsson et al., 2014) linked with epigenetic aging in alcohol dependence (Rosen et al., 2018).

Tissue-specific methylation changes are suggested to be a useful tool for biological age prediction, also known as *Hovarth's epigenetic clock* (Horvath, 2013). Briefly, an age estimator was developed using 8,000 samples from 82 Illumina DNA methylation array datasets, containing 51 healthy tissues and cell types. In Hovarth's clock, set of 353 CpGs were identified as the age stimulator within the organism. However, Hovarth's clock is not accurate for tissues which are subject to hormonal changes such as endometrium or breast tissue (Olesen et al., 2018). Therefore, more studies and pipelines are needed to understand the methylome changes in such tissues.

DNA methylation patterns in diseases. DNA methylation plays a vital role in genomic imprinting, which causes genes to be expressed in a parent-of-origin- specific manner (Reik et al., 1987; Sapienza et al., 1987; Hadchouel et al., 1987).Moreover, normal methylation patterns are essential to control the gene expression of the paternal and maternal alleles of imprinted genes (Li et al., 1993). Gain or loss of methylation at imprinting-control regions may result in several imprinting disorders such as Prader-Willi syndrome, Angelman syndrome and Beckwith-Wiedemann syndrome (Henry et al., 1991).

Autoimmune disease is a condition which arises when the body's immune system attacks healthy cells. Aberrant DNA methylation changes are linked to autoimmune diseases and are suggested to be caused by the strong interplay between environmental factors, genetic variants, drugs, and miRNAs, resulting in aberrant DNA methylation (Sun et al., 2016). A genome-wide case-control study by Ellis et al. (2012) suggested differential T cell DNA methylation is an essential feature in juvenile idiopathic arthritis. Likewise, Cai et al. (2017) reported hypomethylation in the promoter regions of *IL-6* gene observed in the peripheral blood. Furthermore, Meng et al. (2017) showed that methylation patterns acts as a link between genetic and environmental factors by mediating the interaction between the genotype and smoking behavior in rheumatoid arthritis.

DNA methylation patterns are important for normal brain functions (Weng et al., 2013). Research in the recent decades suggested a link between DNA methylation and neurodegenerative diseases, which features the progressive loss of neurons (Lu et al., 2013). It has been suggested that environmental factors such as exposure to pesticide during mothers' pregnancy may cause DNA methylation changes in placental tissue, leading to higher risk for autism spectrum disorder for the child

(Schmidt et al., 2016). Further, *COMT* gene was shown to be hypomethylated in schizophrenia patients when compared to healthy controls, observed in peripheral blood. It has also been shown that *SNCA* and *PARK2* genes exhibits decreased levels of methylation in the early onset of Parkinson's disease (Eryilmaz et al., 2017).

DNA methylation in the endometrium. The endometrium (inner lining of the uterus) includes epithelial, stromal, and endothelial cells. It is controlled by a series of hormones and undergoes cyclic as well as structural changes during every menstrual cycle (which lasts for 28 days on an average during reproductive years) (Caplakova et al., 2016). A menstrual cycle is divided into three main phases in the following chronological order: menstrual, proliferative, and secretory (Figure 3).

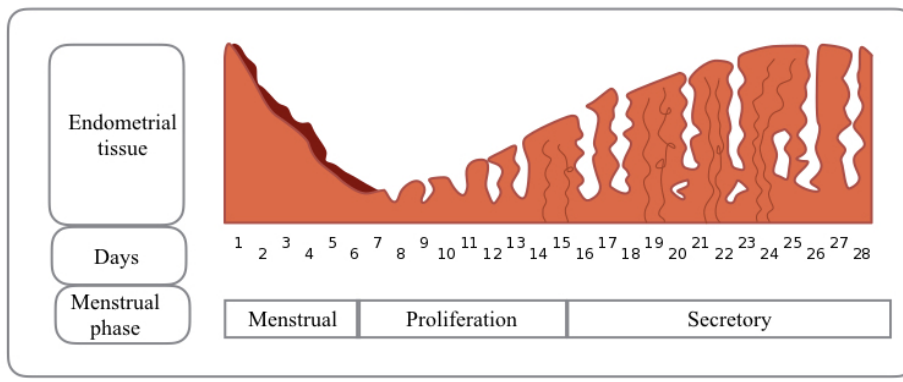

Figure 3. Structural changes in endometrium during a normal menstrual cycle (about 28 days). Menstrual cycle days (ranging about 28 days) are represented by numbers and the respective menstrual phase is shown below. Figure modified from Wikimedia (2004), distributed under a CC-BY 2.0 license.

During the proliferative phase of normal endometrium, estrogen controls the growth of epithelial and stromal cells (Baca-García et al., 1998; Caplakova et al., 2016), while the secretory phase is controlled by the hormone progesterone (Caplakova et al., 2016). Furthermore, methylation changes in healthy human cycling endometrial tissue has been studied byHoushdaran et al. (2014), and the substantial changes in DNA methylation are reported to occur between proliferative and mid-secretory phase wherein the endometrial tissue reaches its maximal thickness and is ready for embryo implantation (Houshdaran et al., 2014).

Moreover, methylation changes are also hypothesized in endometriosis disease progression. Endometriosis is a gynecological disease, wherein the endometrial-like tissues grows outside the uterine cavity (Guo, 2009). Several studies conducted in the past decades suggested endometriosis as the hormonal and even genetic disease (Guo, 2009). However, epigenetic changes are also assosiated with the endometriosis. Specifically, DNA methylation patterns are shown to regulate the gene expression, thereby playing a possible role in endometriosis pathogenesis.

DNA methylation in cancer. Cancer is a complex disease that remains a major public health concern. Cancer development is known to be influenced by both epigenetic and genetic alterations (Feinberg et al., 2006; Jones and Baylin, 2007). In particular, abnormal methylation patterns form the hallmark of cancer development. Cancer cells are known to exhibit genome-wide hypomethylation and site-specific CpG island promoter hypermethylation (Rodríguez-Paredes and Esteller, 2011; Kulis and Esteller, 2010). In cancer cells, DNA hypomethylation is known to occur in multiple genomic loci including repetitive genomic elements, retrotransposons, satellite sequences, and so on, leading to overall genomic instability (Rodríguez-Paredes and Esteller, 2011; Kulis and Esteller, 2010).

1.1.4. DNA methylation biomarkers

A biomarker may refer to the characteristic feature of biological processe(s), which can be quantifiable (Strimbu and Tavel, 2010). According to the National Institutes of Health Biomarkers Definitions Working Group, biomarkers can be defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.", (Group et al., 2001). In the previous section, we briefly explained the way DNA methylation patterns are associated with several normal biological processes as well as disease progression, contributing towards a possible biomarker candidate. In this thesis, we identified the biomarker potential of DNA methylation using different biological conditions.

1.1.5. Profiling of DNA methylation

The advent of NGS and microarray technologies allows us to assess DNA methylation changes in different biological conditions on a scale at the genome-wide level (Yong et al., 2016). The experimental methods to profile genome-wide methylation include the following:

- Enrichment based methods;
- Bisulphite sequencing and
- Bisulphite microarrays

In an enrichment based method, methylated or unmethylated DNA fragments are enriched using several molecules such as methyl-CpG-binding domain (MBD) proteins, restriction antibodies or methylation specific antibodies in a DNA library (Bock, 2012). The enrichment step is followed by quantification using the NGS approach.

In bisulphite sequencing based-approach, the DNA is treated with sodium bisulphite which introduces mutations in unmethylated C's and are further quantified using NGS based methods (Bock, 2012).

On the other hand, in bisulphite microarrays, bisulphite treatment is combined with microarrays, which enables measuring DNA methylation levels at a preselected fraction of Cs throughout the genome. Moreover, the said microarrays are

cheaper in terms of their cost per sample compared to the whole genome bisulphite sequencing approach (Bock, 2012). One of the major drawback in using microarrays is that it is not possible to discriminate between 5mC and 5hmC (Bock, 2012). Furthermore, the HM450K array is based on bisulphite conversion where bisulphite treatment converts unmethylated cytosines into thymines, while methylated cytosines remain unaffected (Frommer et al., 1992). In bisulphite microarrays, a set of probes are predesigned for bisulphite conversion. The HM450K array profiles 485,577 probes of which 482,421 CpG sites, 3091 non-CpG sites and 65 randomly designed single nucleotide polymorphism (SNP) sites (Bibikova et al., 2011), serve as a part of experimental control. Additionally, the said array includes 21,231 genes (99 % RefSeq genes), 26,658 CpG islands (CGI, defined as >200 bp long, >50% GC composition and an observed-to-expected CpG ratio greater than 0.6 (Gardiner-Garden and Frommer, 1987) (96% CGIs), 26,249 CGI shores (0-2 kb from CGI) and 24,018 CGI shelves(2-4 kb from CGI) (Bibikova et al., 2011). Further, probes covering gene-centric regions were further targeted across multiple sub-regions. To summarize, promoter regions were divided into TSS200 (region from the transcription start site [TSS] to -200 nucleotides upstream of the TSS) and TSS1500 (covering -200 to -1500 nucleotides upstream of TSS) respectively) (Bibikova et al., 2011). In addition, 5' and 3' (Untranslated region) UTR, first exon and gene body were also targeted (Bibikova et al., 2011). Figure 4 depicts the gene-centric and CGI regions covered in the HM450K array. A brief summary of methylation based arrays available from Illumina is presented in Table 1; the current thesis utilizes DNA methylation data profiled using the HM450K array.

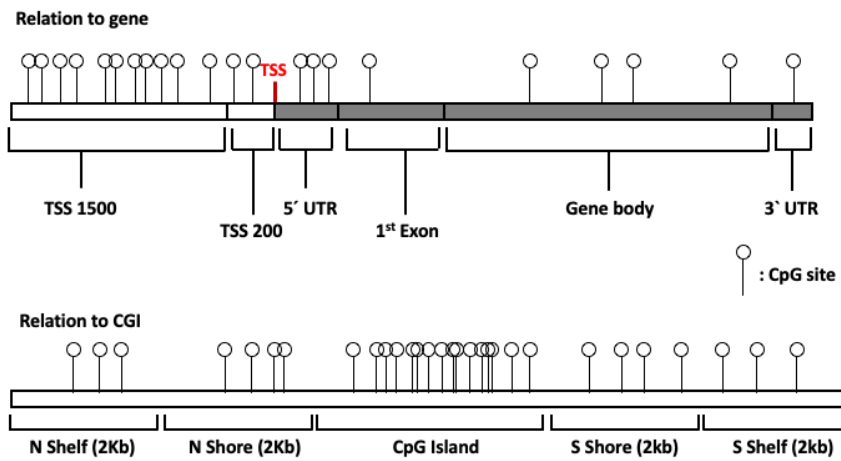

Figure 4. An illustrative representation of the gene-centric and CGI regions covered in the HM450K array. TSS: Transcription start site; UTR: Untranslated region. This figure has been redraw from (Huang et al., 2014).

Technique	Release date	Nr. Probes	Nr. Studies	Nr. Samples
HM27K	April 27, 2009	27,578	341	18,858
HM450K	May 13, 2011	485,577	1,223	86,637
EPIC	Nov 16, 2015	850,000	103	3899

Table 1. Details of the methylation based arrays published by Illumina. Release date, number of samples, and number of studies are retrieved from GEO database, as on 11 April 2019 (Edgar et al., 2002). HM27K: HumanMethylation27 BeadChip; HM450K: HumanMethylation450K BeadChip; EPIC: Infinium MethylationEPIC Kit.

1.2. Bioinformatics methods

In this section, a general overview of the analytical pipeline (See Figure 5) utilized in this thesis will be provided. These methods are most suited for the analysis of the HM450K array and can be tailored for the analysis of other genome-wide methylation quantification methods described in Section 1.1.5. Extensive data processing steps such as quality control, normalization, and batch effect correction are required to address biological questions. Following the quality control and normalization procedures, statistical methods including exploratory analysis, such as data clustering, correlation analysis, and confirmatory analysis including differential methylation analysis, have been performed for an overall interpretation of the data obtained. The details pertaining to these methods will be discussed in the following section.

1.2.1. Quality control and normalization

Quality control of methylation data is necessary to ensure reliable and high quality data necessary for subsequent analysis (Cazaly et al., 2016). The main steps of the quality control procedure have been explained below.

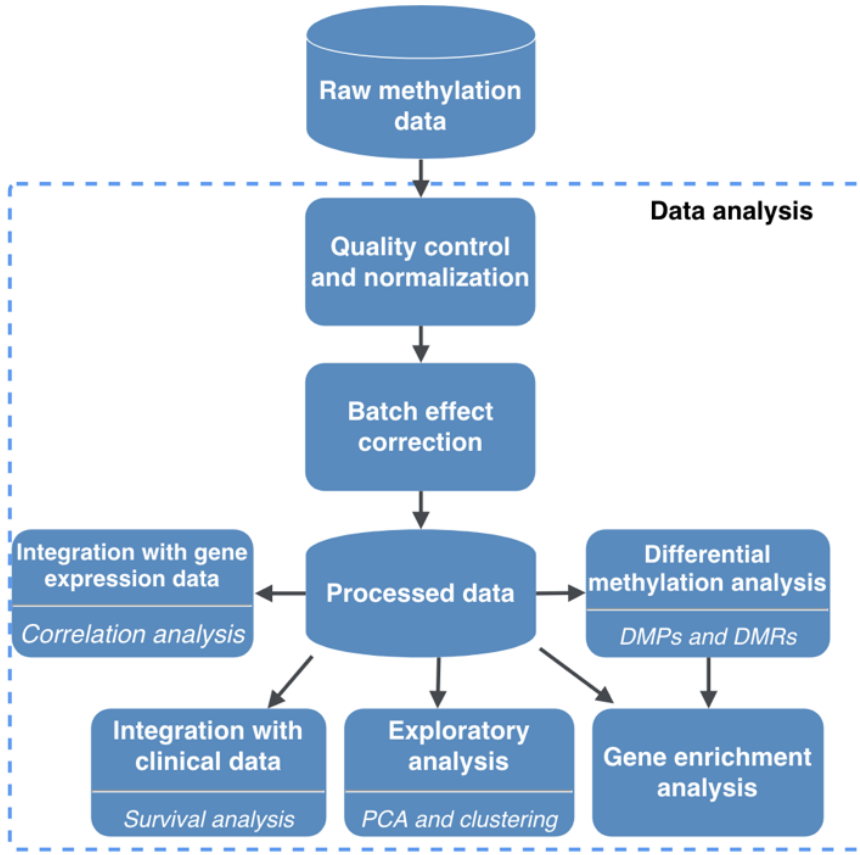

Figure 5. A flow diagram illustrating anaytical pipeline utilized in the current thesis. DMP: Differentially methylated position; DMR: Differentially methylated region.

Raw data processing. Methylation measurement data are usually presented in a binary format called IDAT files, which contain intensity signals from the red (methylated signal) and green channels (unmethylated signal). These IDAT files contains intensity signals from the red (methylated signal) and green channel (unmethylated signal), respectively (Dedeurwaerder et al., 2013). Such IDAT files can be conveniently loaded in the R programming environment (Team, 2017) using any of the R bioconductor packages which includes *minfi* (Aryee et al., 2014), *methylumi* (Davis et al., 2015) and *ChAMP* (Morris and Beck, 2015). The methylation levels for every CpG are represented by the β value, which can be calculated as follows:

$$\beta = \frac{\max(m,0)}{\max(m,0)+\max(u,0)+\alpha} \tag{1.1}$$

In this equation, m and u denotes methylated and unmethylated signal, respectively. A β value of 0 denote an unmethylated CpG, while the β value 1 denote

a fully methylated CpG site. Additionally, the offset α (usually set equal to 100) enables the stabilization of beta values when both m and u are smaller (Du et al., 2010). Alternatively, methylation values can also be represented in M values, as described by the following equation:

$$M = log_2 \left(\frac{\max(m,0) + \alpha}{\max(u,0) + \alpha} \right) \tag{1.2}$$

Based on the aforementioned equations, the relationship between β and M (Du et al., 2010) can be represented as follows:

$$\beta = \frac{2^M}{2^M + 1} \tag{1.3}$$

and

$$M = log_2 \left(\frac{\beta}{1 - \beta} \right) \tag{1.4}$$

Assessing experimental quality. This step includes visual inspection of control probes that enable one to assess the quality control of different sample preparation steps such bisulphite conversion, hybridization, staining, among others. This step can be conveniently visualized using the minfi package (Aryee et al., 2014).

Probe filtering. The filtering of probes is necessary to eliminate any possible noisy signal. Probes with low quality signals should be removed by using the detection P-value threshold (P > 0.05) (Dedeurwaerder et al., 2013). Further, probes residing in the single nucleotide polymorphism (SNP) loci, which may also result in noisy signals and cross-reactive probes (Chen et al., 2013), showing spurious cross hybridization (Wilhelm-Benartzi et al., 2013) also need to be removed. In addition, probes localized at the X and Y chromosomes are removed if the researcher wants to eliminate sex-associated methylation effect.

Normalization. This is one of the crucial steps in data processing, which involves the removal of variation that is not related to biological properties, but rather the technical variation (Dedeurwaerder et al., 2013). Normalization of methylation data includes within-array normalization and between-array normalization. (See Table 1.2.1)

Within-array normalization involves background correction as well as type I and II probe scaling (Wilhelm-Benartzi et al., 2013; Dedeurwaerder et al., 2013). The former involves the removal of any non-specific signal from the overall signal, correcting any possible between array artifacts (Wilhelm-Benartzi et al., 2013) which can be facilitated by the *minfi* package(Aryee et al., 2014). Further, Illumina First Sample Normalization (IFSN) is a classical within-array normalization procedure used for the HM450K array, which adjusts for this variability of color signals (Yousefi et al., 2013) and can also be used in this regard.

The HM450K array utilizes two types of assays, Infinium I and Infinium II, for methylation profiling. About 72% HM450K probes use the latter in cases where

unmethylated and methylated signals are measured by a single bead (Bibikova et al., 2011). On the other hand, the remaining probes are profiled using the Infinium I assay, wherein unmethylated and methylated signals are measured by different beads in the same color channel. However, the two types of probes used the in the HM450K array may not have the same methylation distribution, causing biased estimation of methylation measure (Dedeurwaerder et al., 2013). The most commonly used methods to eliminate type I and II bias, which are discussed below.

Unlike gene expression microarrays, direct application of quantile normalization is not suitable in this regard, since type I and II probes do not measure the same CpG, thereby varying the distribution of β values between these probe types (Dedeurwaerder et al., 2013). However, Subset-quantile Within Array Normalization (*SWAN*) implemented in the *minfi* package (Aryee et al., 2014) matches the Type I and Type II β distributions , by applying a within-array quantile normalization separately for different subsets of probes (Maksimovic et al., 2012). However, one of the pitfalls of the *SWAN* based approach is that it is not suitable when a global methylation difference is expected, such as cancer/normal studies (Dedeurwaerder et al., 2013) or between-tissue studies. Functional normalization, can overcome this problem, which by default applies to the preprocessNoob, functions as a first step for background subtraction and uses the first two principal components of the control probes in order to determine the unwanted variation. Another alternative normalization approach is beta-mixture quantile normalization (*BMIQ*) (Teschendorff et al., 2012), which can be used if a variation is present in the shape of methylation distribution. This method decomposes the density profiles of Infinium I and Infinium II probes into two mixtures of three distributions based on the three methylation states: unmethylated (close to 0), partially methylated (close to 0.5), and fully methylated (close to 1). Furthermore, (*BMIQ*) employs a quantile normalization to fit each distribution of the Infinium II profile to the corresponding distribution of the Infinium I profile. Unlike (*SWAN*), (*BMIQ*) is assumption free and does not depend on arbitrary choices concerning biological characteristics in order to perform normalization (Dedeurwaerder et al., 2013).

Between array normalization is used to normalize the data between the arrays and can be performed using *shift and scaling* using *lumi* R package (Du et al., 2008). For more information on this procedure, refer to Wilhelm-Benartzi et al. (2013); Dedeurwaerder et al. (2013).

In the publications referred to in this thesis, we utilized Illumina First Sample Normalization (*IFSN*) which provided good quality data to reveal tissue-specific methylation patterns in Publication I. However, for Publication II, we observed that methylation differences in the type I and type II probes resulted in large bias and variation. Therefore, *BMIQ* normalization resulted in the generation of the most suitable data for further analysis. In Publication III, between array normalization, in conjunction with a *shift and scaling* approach, was utilized since this method fitted our data according to the clustering approach (Figure 13).

Method	Normalization type	Description	R Packages
Illumina	W and BA	Background correction and reference normalization	*minfi*
PBC	W and BA	Type I and II probe bias correction	*ChAMP*
QN	W and BA	The probe intensities for all the samples are made identical	*minfi*
SQN	W and BA	Type I and II probe bias correction	*minfi*
SWAN	W	Type I and II probe bias correction	*minfi*
FN	BA	Removes unwanted technical variation using control probes	*minfi*
NOOB	W	Background correction and adjustment of dye bias	*Methylumi*
BMIQ	W	Type I and II probe bias correction	*waterRmelon*, *ChAMP*
DASEN	BA	Background correction and quantile normalization separately performed for Type I and II probes	*waterRmelon*

Table 2. Summary of different normalization procedures available for the HM450K array. PBC: Peak-based correction; QN: Quantile normalization; SQN: Subset-quantile normalization; SWAN: Subset-quantile witharray-normalization; FN: Functional normalization; NOOB: Normal-exponential using out-of-band probes; BMIQ: Beta-mixture quantile normalization; DASEN: Data-driven separate normalization; W: Within array; BA: Between array normalization.

1.2.2. Batch effect correction

The batch effect corresponds to non-biological variation that may arise among the batches of samples that are not processed on the same day, or when different scanners are used in this process, when it is performed by different persons (Dedeurwaerder et al., 2013), or when it is performed at different labs. Moreover, the position of the array on the slide from the same batch of samples may lead to the generation of non-biological variations (Dedeurwaerder et al., 2013). A comprehensive definition of the batch effect can be found in a review article Lazar et al. (2012). Owing to the aforementioned reasons, it is important to correct these confounding factors before proceeding to any further analysis.

Methods to correct the batch effect have been briefly described below:

Empirical Bayes method. The *Empirical Bayes* method implemented in *ComBat* (Johnson et al., 2007), originally designed for gene expression microarrays can be used for batch effect correction in HM450K data implemented using the *ChAMP* package (Morris and Beck, 2015). *ComBat* uses parametric and/or non-parametric *Empirical Bayes* frameworks for adjusting the batch effects within the data. Furthermore,the input data for *ComBat* needs to be cleaned and normalized prior to the batch effect removal. *ComBat's* performance has been noted to be robust towards outliers (Dedeurwaerder et al., 2013).

BEclear. BEclear is a batch effect correction method used for HM450K data, which was developed over the recent past. To summarize, *BEclear* adjusts the portions of the data identified to differ significantly from the other batches (Akulenko et al., 2016). More details pertaining to *BEclear* is available in Akulenko et al. (2016).

To sum up, batch effect correction enables minimization of any possible unwanted non-biological variation. Further, it provides an opportunity to combine samples from two or more studies that lie with in the same phenotype of interest, thereby increasing statistical power and yielding robust results. In this thesis, we utilized the *Empirical Bayes method* (Johnson et al., 2007) to adjust the batch effect between Oxford and Tartu samples in order to analyze menstrual cycle specific methylation patterns in Publication II.

1.2.3. Correlation analysis

Correlation analysis is a statistical approach adopted to explore the relation between two variables. The most widely used methods for performing the said analysis have been described below:

Pearson's correlation coefficient (PCC). Let us assume that we want to determine the linear relationship between two variables x and y. The Pearson-product moment correlation coefficient (r) can be expressed as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}} \tag{1.5}$$

In this equation, n denotes the sample size $\bar{x}$ and $\bar{y}$ represent the sample means of x and y respectively.

Spearman's rank correlation. Since *Pearson's correlation* assesses the linear relationship between the samples, one can alternatively use *Spearman's correlation* that assesses monotonic relationships (whether linear or not). *Spearman's correlation* can also be defined as the *PCC* between the ranked variables, and it can be computed using the following formula:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{1.6}$$

In this equation, ρ denotes the *PCC* applied to the ranked variables. d_i signifies the distances between the ranked variable and n represents the number of observations.

The correlation measure is always expressed between -1 and +1. A value close to 0 can be interpreted to signify no relation between the tested variables, while a value close to +1 and -1 denotes high positive correlation and negative correlation, respectively.

In this thesis, we applied *PCC* to compute the correlation across methylation levels in tissues and also to integrate gene expression with methylation profiles in Publication I. In Publication III, we computed *Spearman's rank correlation* in order to calculate the correlation between gene expression and methylation levels in pre-receptive and receptive endometrium to obtain insights on the effects of methylation at transcriptional levels. Finally, both *PCC* and *Spearman's rank correlation* were implemented to generate heat map visualization for the web tool *MethSurv* described in Publication IV.

1.2.4. Principal component analysis

Principal component analysis (PCA), a method to reduce the dimensionality of data (Pearson, 1901; Yeung and Ruzzo, 2001) can be applied to genome-wide methylation data for meaningful interpretation and exploration. PCA reduces high dimensionality by identifying directions, called principal components, along which the variation in the data is maximal. These principal components are linear combinations of the original variables present in the high dimensional data (Ringnér, 2008). An illustrative example of a three dimensional PCA plot from the subset of tissue methylation data obtained from Publication I, highlighting the separation of brain and adipose tissues has been provided in Figure 6. In this figure, the first principal component describes the direction along which the samples show the greatest variation. Moreover, every component is uncorrelated to the previous ones, which maximizes the variance of the samples when projected onto the component. In PCA, the direction depicting the variances are eigenvector shown by PC1 and PC2 in Figure 6 and the corresponding the eigen value explains the

variance of the data in the respective direction. Additionally, in this figure, PC1 has the highest eigenvalue and therefore, the first principal component. On the contrary, PC2 contain the second highest eigenvalue is the second principal component and so on (Peterson, 2015). A detailed description of PCA in genome-wide data is provided in Ringnér (2008).

In this thesis, we employed PCA in Publication II to visualize any possible distinction between endometriosis patients and healthy controls. In Publication IV, PCA was implemented in the web tool *MethSurv* to identify patterns in the gene methylation with respect to patient's characters such as age, gender, among others.

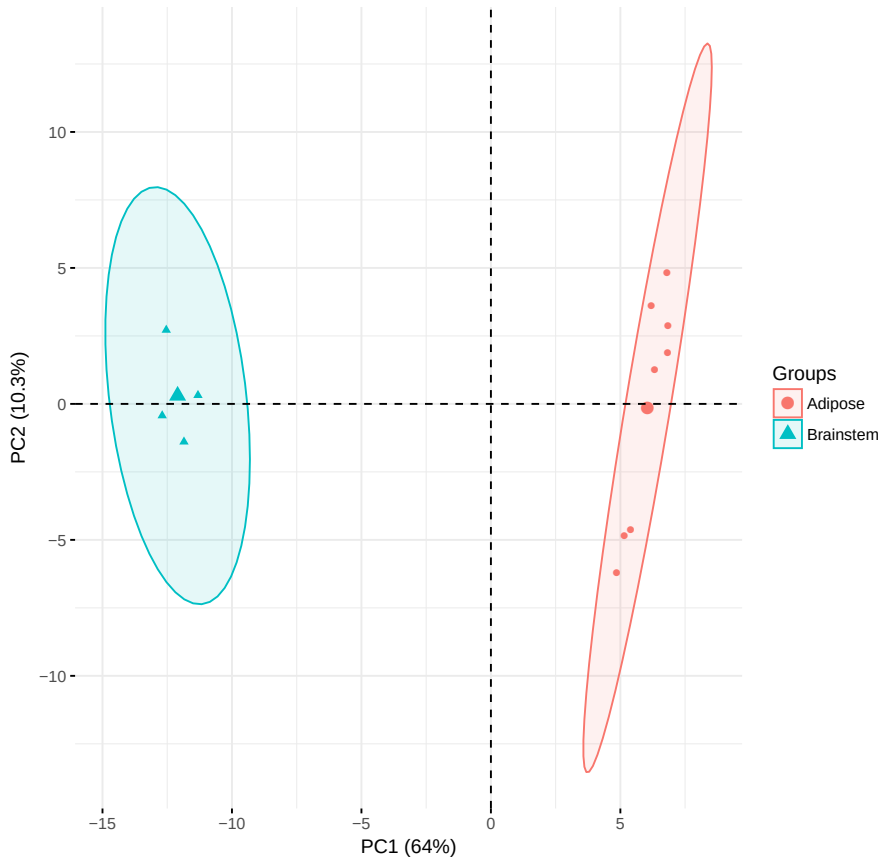

Figure 6. PCA plot of the tissue methylation data obtained from Publication I, showing the separation between adipose and brain tissues. PC1 denotes 64% variability, PC2 denotes 10.3% variability.

1.2.5. Cluster analysis

Cluster analysis or data clustering involves grouping a set of data points that are similar to each other based on certain criteria. Several algorithms are available to perform data clustering. Hierarchical clustering, a prominent approach in this

regard, which is also known as connectivity-based clustering wherein the clusters are built on the basis of hierarchy, will be briefly discussed in this section. In the said clustering, a binary tree is constructed by successive merging of similar samples or probes based on similarity measure (Eisen et al., 1998). Hierarchical clustering includes agglomerative and divisive hierarchical clustering (Wang and Petronis, 2008) which have been briefly discussed below.

Agglomerative hierarchical clustering. It is based on the bottom-up approach wherein the underlying method tends to find and merge two clusters based on the shortest distance. Moreover, between-cluster distance is calculated based on the centroid of the merged clusters (Wang and Petronis, 2008). For instance, one member from a particular cluster is paired with another member of the other cluster after which the pairwise-distance is thus calculated. This approach can also be termed as complete linkage clustering (Wang and Petronis, 2008).

Divisive hierarchical clustering. The underlying principle of this clustering is exactly the opposite to that of agglomerative hierarchical clustering. To elaborate, this approach begins with a single cluster and the point comprising the largest pair-wise distance to the other point is chosen. Subsequently, it is split from the initial cluster resulting in the formation of a new cluster.

The results of hierarchical clustering can be conveniently visualized using a dendrogram, which provides a visual overview of relationship between the samples. A detailed description of hierarchical clustering can found in (Wang and Petronis, 2008).

In this thesis, hierarchical clustering analysis was a significant method for data visualization. In Publication I, the hierarchical clustering approach revealed tissue-specific methylation patterns, while in Publication II, it provided a visual overview of the strong batch effect between Oxford and Tartu methylome profiles. Furthermore, in the case of the latter, it also demonstrated menstrual cycle methylation patterns following batch effect correction. Lastly, in Publication IV, we employed the hierarchical clustering approach to visualize the methylation patterns of CpGs within the vicinity of a gene, in order to capture methylation differences across different genomic regions.

1.2.6. Differential methylation analysis

The said analysis is performed to identify the genomic regions associated with differential methylation status across different biological samples. Differential methylation can be conducted either at single CpG level, called differentially methylated positions (DMP) or at the region level called differentially methylated regions (DMR). DMP analysis can be performed using a simple *t-test*, or *Bayes moderated t-test* (Smyth, 2005) which can be implemented by the *ChAMP* package (Morris and Beck, 2015). It must be noted that the *t-test* is based on the assumption that the samples are normally distributed. However, the condition of normality may not always hold true for certain kind of methylation datasets (Wang and

Petronis, 2008). Therefore if one wants to perform statistical tests based on no prior assumptions regarding the methylation data distribution, non-parametric tests such as *Wilcoxon signed-rank test* or *Wilcoxon rank-sum test* can be used (Wang and Petronis, 2008). The former can be performed if the methylation data comprises matched data (for instance, methylation levels are measured before and after fasting for the same set of individuals) whereas the latter can be used for unmatched methylation data.

Since DMP analysis focuses on individual CpG sites, ignoring any possible correlation with nearby CpG sites, it may produce biased and redundant results. In particular, the HM450K array provides spatially distributed DNA methylation structure that is often expected to have similar methylation profiles (Kolde et al., 2016). On the other hand, the DMR methods are based on the assumption that nearby probes on a given window size tend to demonstrate the same behavior (uniformly hyper or hypomethylated) (Dedeurwaerder et al., 2013), thus yielding statistically more sensitive and powerful results (Kolde et al., 2016). The region-based analysis pipeline presented in the IMA package (Wang et al., 2012) treats each predefined genomic regions such as CGI, as a separate unit and performs differential methylation analysis accounting for the statistical significance ($P < 0.05$). Such analysis may often result in larger regions covering the same set of differentially methylated sites (Kolde et al., 2016). Alternatively, one can use the bump hunting algorithm (Jaffe et al., 2012), which performs differentially methylation analysis on spatially smoothed data and subsequently aggregates the individual sites into a region. Further, the results of the said algorithm depend on the parameters supplied by the user (Jaffe et al., 2012; Kolde et al., 2016). These parameters may include parameters such as effect size cut-off and smoothing window size (Kolde et al., 2016).

Another powerful DMR based approach is *seqlm* (Kolde et al., 2016), which provides flexible parameters to determine DMRs. The main steps of DMR identification in seqlm (Kolde et al., 2016) can be briefly described as follows:

1. Genomic segments are established according to the user specified threshold (e.g. 500 bp)
2. The segments obtained from the aforementioned step are divided into regions based on fitting a linear model into the sliding window. The minimum description length (MDL) principle (For more details, See Kolde et al. (2016)) is applied to the fitted model accounting for longer regions as well as the goodness of the model fit (Kolde et al., 2016). This is because methylation is known to be regulated in longer regions (Lienert et al., 2011) and may have better biological implications.
3. Once an optimal genomic regions are derived using the aforementioned steps, a linear mixed effect model as depicted below is fitted to derive statistically significant DMRs.

$$y_{ij} = \mu_j + \beta x_i + b_i + \varepsilon_{ij} \tag{1.7}$$

In this equation y signifies the response variable denoting the methylation value of the sample i and site j. While μ denotes the baseline methylation value, β represents the average effect size within the region and x signifies the phenotype of interest (e.g. age). Moreover, b_i denotes the random effect which accounts for within-sample correlation and ε represents the standard error.

In this thesis, *seqlm* was initially developed in Publication I to establish tissue specific DMRs but was subsequently established into a separate method that was published in Kolde et al. (2016). Further, *seqlm* was used to evaluate the difference in methylation between both endometriosis patients and healthy controls in addition to the difference among the menstrual cycle groups in Publication II. On the other hand, in Publication III, *seqlm*, along with other methods was used to determine the true set of differently methylated CpG sites between the pre-receptive and receptive states of the endometrial tissue.

1.2.7. Gene enrichment analysis

Once differential methylation analysis has been performed, it is important to statistically relate the results (usually a genomic loci or CpG sites, that are further related to genes) to gene functional categorie;, this process is termed as *gene enrichment analysis*. There are several ways to relate the gene list in order to make meaningful biological interpretation which have been discussed below.

Enrichment using Gene Ontology (GO) terms. The gene list can be related to Gene Ontology (GO) terms which can be facilitated by the GO (Ashburner et al., 2000),i.e., a large collection of genes described by a controlled vocabulary (Peterson, 2015). GO comprises of the vocabularies (also termed as ontologies) that are hierarchically structured with the molecular function (MF), biological process (BP) and cellular component (CC). GO terms constitute the main elementary unit of GO wherein the relationship between each of these terms is represented by a directed acyclic graph (DAG). In other words, GO terms constitute a hierarchy in which every gene could be annotated with one or more terms in each ontology (Eden et al., 2009) and may further associated with the parent term. For instance, cell death (GO:0008219) is a child term of cellular process (GO:0009987) while the term cellular process is linked to other cellular processes such as regulation of euchromatin binding (GO:1904793), regulation of core promoter binding (GO:1904796) and so on.

Enrichment with biological Pathway. Apart from GO, one may wanted to relate the gene list derived from differential methylation analysis to biological pathways (e.g, steroid hormone biosynthesis) using pathway knowledge bases such as *KEGG* (Kanehisa and Goto, 2000), *BioCarta* (Team, 2001) and *Reactome* (Joshi-Tope et al., 2005).

Linking with other resources. In addition to GO and pathway analysis, it is possible to evaluate and associate the enrichment of the gene list procured from differential methylation analysis with protein complexes facilitated by the knowledge-base, *CORUM* (Ruepp et al., 2007), transcription factor enrichment using *TRANSFAC* (Matys et al., 2003) and human diseases using *The Human Phenotype Ontology* (Robinson et al., 2008).

A web tool *g:Profiler*, (Reimand et al., 2007, 2016) can be conveniently used for gene set enrichment analysis. The tool *g:Profiler* utilizes hypergeometric tests to determine whether the inputted gene list overlaps with functional categories (such as GO term, pathways, transcription factor enrichment and so on) followed by multiple testing correction when several categories are tested (Reimand et al., 2007). The options available for multiple testing correction in *g:Profiler* includes *Bonferroni*, *Benjamin and Hochberg* (BH) and *g:SCS* that is a custom threshold (Reimand et al., 2007). Apart from *g:Profiler*, there are many other tools available for the research community to facilitate gene enrichment analysis - such as, *GOrilla* (Eden et al., 2009), *DAVID* (Huang et al., 2008), *topgo* (Alexa and Rahnenfuhrer, 2010) to name a few.

In some cases, gene enrichment analysis may result in the generation of several hundred terms, which may be difficult to interpret. One possible approach to tackle this issue is to visualize gene list enrichment analysis in the form of word clouds, facilitated by *GOsummaries* (Kolde and Vilo, 2015) implemented in a bioconductor package. *GOsummaries* implements custom methods to filter GO enrichment results which depend on the size of the terms and the relationship between each term (See Kolde and Vilo (2015) for more details). Further, the font size in the resulting word cloud is directly proportional to the strength of the enrichment. (Kolde and Vilo, 2015).

We utilized *DAVID* to perform gene enrichment analysis pertaining to tDMRs in Publication I and *g:profiler* to understand the functional relevance of menstrual cycle specific genes in Publication II. On the other hand in Publication III, we utilized *g:Profiler*, *GOsummaries* to understand the functional relevance of differential methylation between pre-receptive and receptive endometrium.

1.2.8. Survival prediction using methylation data

As mentioned in section 1.1.3, methylation patterns are a promising biomarker candidate for patient survival prediction. To associate methylation levels with patient survival, one possible approach is to perform Cox-proportional model fitting for the patient i as shown below:

$$h_i(t) = h_0(t)exp(\beta_i X_i) \tag{1.8}$$

In this equation $X = (X_1, X_2)$ and X_1 is the CpG methylation level while X_2 represents clinical covariates such as age, BMI, stage, grade and so on. Moreover, $h_0(t)$ is the baseline hazard, $h_0(.)$ is the arbitrary baseline hazard function, while

β is a vector of regression coefficient (not be to confused with methylation β values). Survival analysis can be performed accounting for both the univariable and multivariable models. In univariable analysis, probe's methylation status is an explanatory variable and while survival time function is the response variable. In multivariable analysis, in addition to the methylation status, clinical covariates such as age, sex, stage among others can be included in the model. Additionally differential survival can be assessed by dichotomizing the patient's methylation level into higher and lower methylation. The cut-off point for such dichotomization can be computed based on the mean or data quantiles. However, to evaluate of all possible cut-off points in the continuous methylation β values, the method concerning Maximally Selected Rank Statistics (maxstat) (Hothorn and Lausen, 2003) can be used. For more details on maxstat, See Hothorn and Lausen (2003). In principle, maxstat assesses all of the data points obtained from the continuous methylation data and establishes a cut off point where the standardized statistics take their maximum significance regarding the separation of patient groups. The resulting Cox model fit provides, hazard ratio (HR) with 95% CI. On the other hand, the likelihood-ratio (LR) test and Wald test allows to assess the goodness of Cox model fit. The assumption of Cox proportionality is tested using proportional hazards assumption test.

2. TISSUE SPECIFIC METHYLATION PATTERNS (PUBLICATION I)

This chapter focuses on the understanding methylation patterns that are specific to different somatic tissues. The data was provided by our project collaborators from Estonian Genome Centre, University of Tartu. We developed and provided the bioinformatics workflow to support the data analytics and interpretation of this project. The methylation data used in this research were unique at that time of the study, concerning the tissue samples and the choice of the array. For instance, it included methylation data obtained from the same set of individuals covering somatic tissues from numerous tissue systems of the body such as nervous system (medulla oblongata and ischiatic nerve), circulatory system (aortas and arteries), digestive system (gastric mucosa), skeletal system (adipose tissue, bone and joint cartilage), excretory system (bladder and gallbladder), and so on (See Figure 7). The primary goal of this study was to understand whether there were tissue specific-methylation patterns (tissue-specific biomarkers). A few of the fundamental questions while initiating the project were concerned with data processing and analysis such as how the tissues are related to each other regard to the methylation patterns are how tissue-specific methylation patterns are related to gene expression. During this study, data processing which includes normalization, data cleaning, and filtering before downstream analysis played a very crucial role in identifying tissue-specific biomarkers.

Some of the key methodologies, results and the contribution to Publication I are described in the following sections.

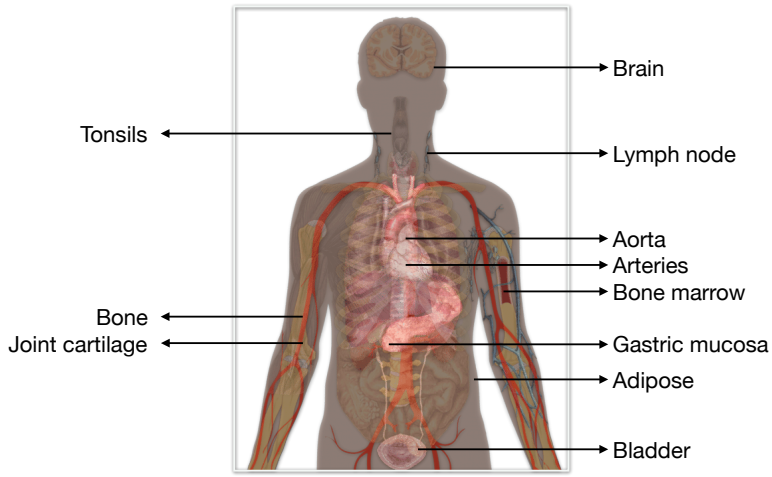

Figure 7. This figure depicts the list of tissues studied for methylation profiling in the current chapter. Figure modified from Wikimedia (2012).

2.1. Data processing

The methylation data were pre-processed using the standard pipeline suggested by the *minfi* package (Aryee et al., 2014) and the probes residing on the X and Y chromosomes as well as SNP regions were filtered out. (See section 1.2.1). The probability density plot of the processed methylation data showed that most of the CpGs were either fully unmethylated or fully methylated in the somatic tissues shown in Figure 8.

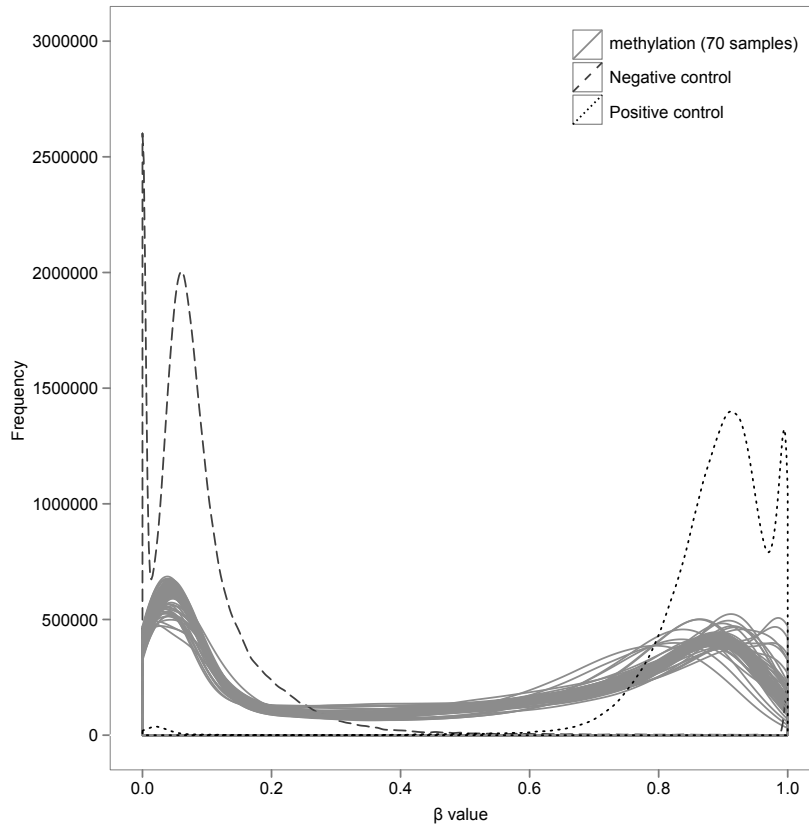

Figure 8. Methylation distribution of the tissues. Global methylation distributions of tissues (70 samples) along with the positive (fully methylated) and negative (fully unmethylated) controls are represented as the density plot. The X axis corresponds to methylation beta values ranging from 0 to 1, while the Y axis represents the frequency. From the global distribution plot, It can be seen that most of the CpGs are either unmethylated or fully methylated in the somatic tissues.

2.2. Clustering analysis

Hierarchical clustering analysis (See Section 1.2.5) along with complete linkage was performed on the processed methylation data in order to obtain a general relatedness profile between the tissues. Moreover, the hierarchical clustering of tissue methylation data showed that similar tissues were clustered together, instead of individuals being clustered together (eg. bone marrow red and yellow. See Figure 9), suggesting that tissues with similar functions exhibit similar methylation profiles. Therefore, methylation patterns are sufficient for the classification of tissues.

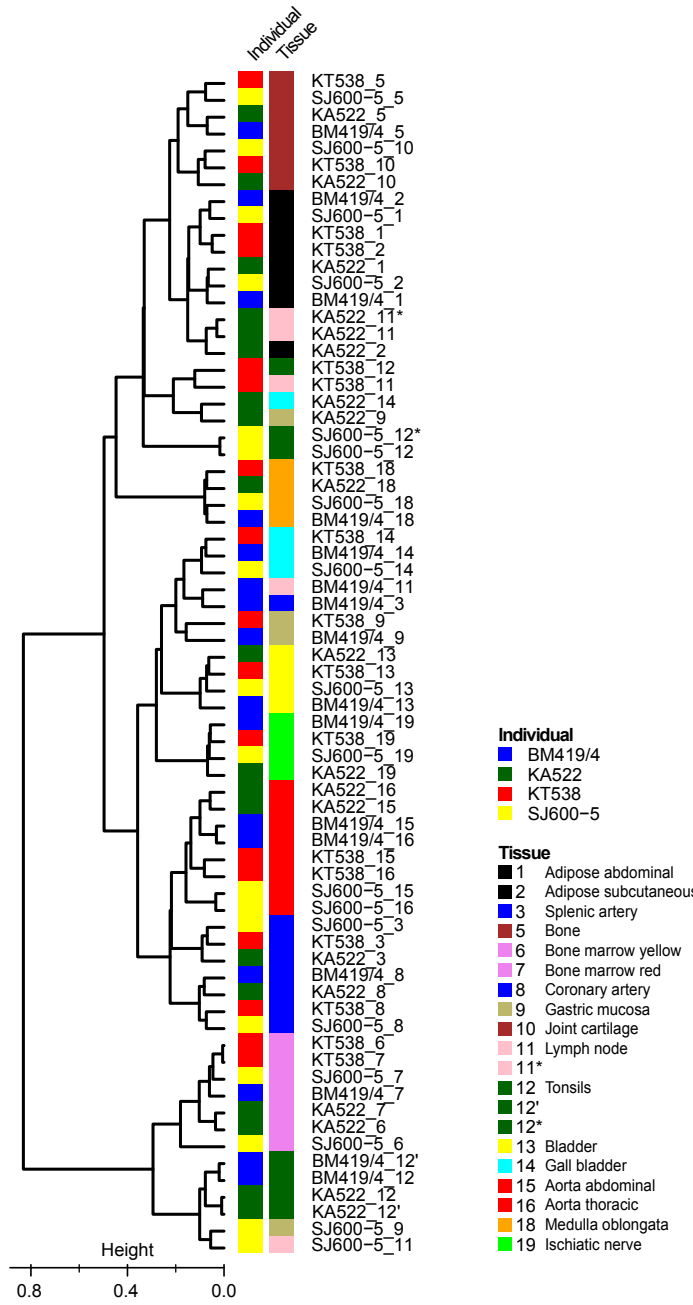

Figure 9. Hierarchical clustering of 17 tissues based on methylation levels. Hierarchical clustering dendrogram was generated using complete linkage. Similar tissues (e.g., arteries, aortas and brain tissues) are clustered together based on their methylation patterns. It is also evident that tissues exhibit higher level of similarity compared to individuals. The tissue number followed by the special character (*, ') denotes a technical or biological replica.

2.3. Comparison of methylation distribution in different genomic regions

We compared the methylation distribution of tissues in different gene sub-regions and CGI using a novel visualization approach represented by the binned box and whisker plot (binbwplot). The plot provides a clear understanding of the low and high methylation patterns, which can be compared within and between different gene sub-regions. The bin-bw plot showed that tissues exhibited high methylation levels in places where transcription is not usually initiated (gene body and 3'UTR) and low methylation levels in the vicinity of promoter regions (TSS200, TSS1550 and 3'UTR) and CGI regions (See Figure 4 from the Publication I). These findings are in line with the classical gene model in which lower methylation in the promoter and CGI regions may favor the binding of transcriptional proteins to the gene.

2.4. Tissue specific differentially methylated regions (tDMRs)

Differential methylation analysis was performed to identify genomic regions with distinct methylation profiles across the studied tissues. First, we evaluated some of the standard differential analysis methods including *Bayes moderated t-test* and *Wilcoxon rank sum test* (data not shown). Due to the disadvantages of the aforementioned methods, our co-authors developed a novel method for differential methylation analysis (See Section 1.2.6). The approach of this method aimed to identify longer genomic regions that are differentially methylated in one tissue compared to other tissues. Subsequently this method was published as *Seqlm* (Kolde et al., 2016) for exploring differentially methylated regions. DMR analysis revealed that every tissue has its distinctive methylation patterns (See Figure 10 for an illustrative visualization of a tDMR region differentiating bladder vs. all other tissues). However, the number of tDMRs varies significantly among different tissues (See Table 1 of Publication I). The highest number of tDMRs was found in tonsils followed by medulla oblongata, whereas the lowest number of tDMRs was found in the lymph nodes. In addition to tDMRs present in promoter regions we reported higher preference of tDMRs in gene body (*Fisher's exact test*, $P < 2.2 \times 10^{-16}$), which are further supported by recent studies by Yang et al. (2014); Teissandier and Bourc'his (2017). However, the biological functions of gene body's tissue methylation need to be elucidated experimentally.

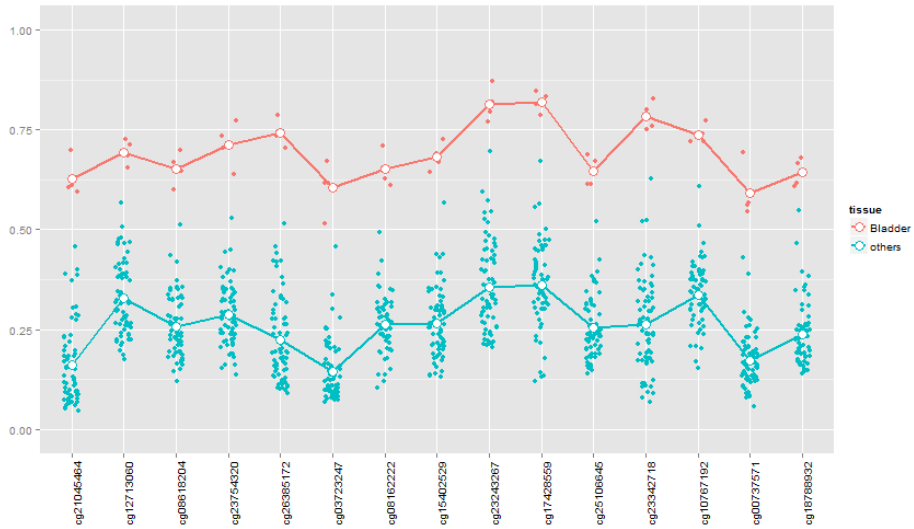

Figure 10. Visualization of the bladder specific DMR region visualized using *seqlm* (Kolde et al., 2016). This illustrative figure show the bladder specific DMR region (red) in comparison with the other tissues (blue). The horizontal axis corresponds to the CpG site, while the vertical axis corresponds to the methylation beta values, ranging from 0 to 1.

2.5. Integrating tissue methylation profiles with gene expression data

We performed integrated analysis of gene expression levels and tissue methylation profiles in order to understand the functional role of methylation in gene expression regulation. Since we do not have gene expression measures from the original tissues used in this study, we searched the expression of matching tissues from the public gene expression repository, gene expression omnibus (GEO) (Edgar et al., 2002) and array express (Brazma et al., 2003).

The association between methylation and gene expression levels was performed using *PCC* (See Section 1.2.3 using two different approaches

- By correlating global tissue methylation levels with gene expression profiles and
- By correlating tDMRs with gene expression profiles

Thus, the aforementioned approaches clearly distinguishes the functional role of tDMRs compared to global methylation.

The methylation and gene expression correlation analysis showed more number of negative correlations than positive correlations (See Table 3 and Table 4 from Publication I for more details) (Fisher's exact test, $P < 2.2 \times 10^{-16}$).

2.6. Summary and impact

In this article, we explored the significance of tissue-specific methylation patterns that are crucial for the classification of tissues. The methylation levels of selected tDMRs (See Publication I for more details) in this study were validated by our collaborators using traditional Sanger sequencing. This validation gave us further confidence on our bioinformatics workflow. Hence our workflow can be adapted for similar studies. Additionally, the data used in this study have also been used for determining imprinting genes by Pervjakova et al. (2016). Further, the method of DMR analysis applied here was developed into an independent tool and published by Kolde et al. (2016). Moreover, our article highlighting tissue-specific methylation patterns (Lokk et al., 2014) is an unique study and was one of the top ten highly accessed of the journal *Genome Biology* journal for that particular year. Further, Publication I has been cited by over 170 publications till date (retrieved from google scholar, 11/04/2019).

2.7. Contribution

In this project, I worked on processing the raw methylation data, evaluating differential methylation methods, and identifying gene expression data from publicly available databases. I also downloaded, normalized, performed further integrated analysis of methylation with gene expression data, and worked on data visualization. In addition, I drafted the manuscript related to my contribution.

3. MINING DISEASE SPECIFIC METHYLATION PATTERNS IN ENDOMETRIUM (PUBLICATION II)

This chapter aims to find biomarkers differentiating endometriosis patients from healthy individuals, based on methylation patterns. The methylation data used in this project were generated by the project collaborators from Tartu and Oxford. In the recent years, several research works conducted transcriptomics studies to understand transcriptional changes in endometriosis disease progression. However, altered methylation patterns have been proposed to be a mechanism that is possibly responsible for the initiation of endometriosis, apart from classical transcriptomics. Moreover, considering the fact that endometrium undergoes cyclic changes (Houshdaran et al., 2014), menstrual phase-specific changes must be considered while analyzing endometrial methylome in relation to the disease status. Therefore, we aimed to examine menstrual cycle-specific DNA methylation patterns in the endometrium of endometriosis patients and controls. This study included endometrial methylation profiles obtained from 31 endometriosis patients and 24 controls. It also contained methylation data throughout the menstrual cycle phase (28 days) described as follows: menstrual (M, n = 5), proliferative (P, n = 5), early secretory (ES, n = 8), mid secretory (MS, n = 26), and late secretory (LS, n = 11). In this project, we provided complete bioinformatics analytical support to achieve data processing, visualization, and differential methylation analysis between endometriosis patients and controls as well as between different menstrual cycle phases. The key challenges included data processing such as normalization, in addition to batch effect correction and data visualization.

3.1. Data processing

This process begins with the normalization of the methylation data. We normalized the said data using Beta-Mixture Quantile (*BMIQ*) normalization procedure (Teschendorff et al., 2012) (See Section 1.2.1). Furthermore, the samples used in this project were prepared in two different laboratories, i.e., in Tartu and Oxford. We observed a strong batch effect in the samples, which was evident from the sample clustering analysis, wherein samples were clustered according to the lab instead of the disease status or menstrual cycle phase, as shown in Figure 12. The batch effect was corrected using ComBat (Johnson et al., 2007) (See Section 1.2.2).

3.2. PCA and clustering

PCA and clustering was performed to visualize the methylation data patterns with respect to menstrual cycle phases and/or disease status. In this study, the PCA (See Section 1.2.4) showed no distinction between the endometriosis patients and

controls (Figure 11). Further, the hierarchical clustering analysis (See Section 1.2.5) clearly demonstrated that the methylation profiles were clustered together by the menstrual cycle phases rather than the disease status (See Figure 2 from Publication II).

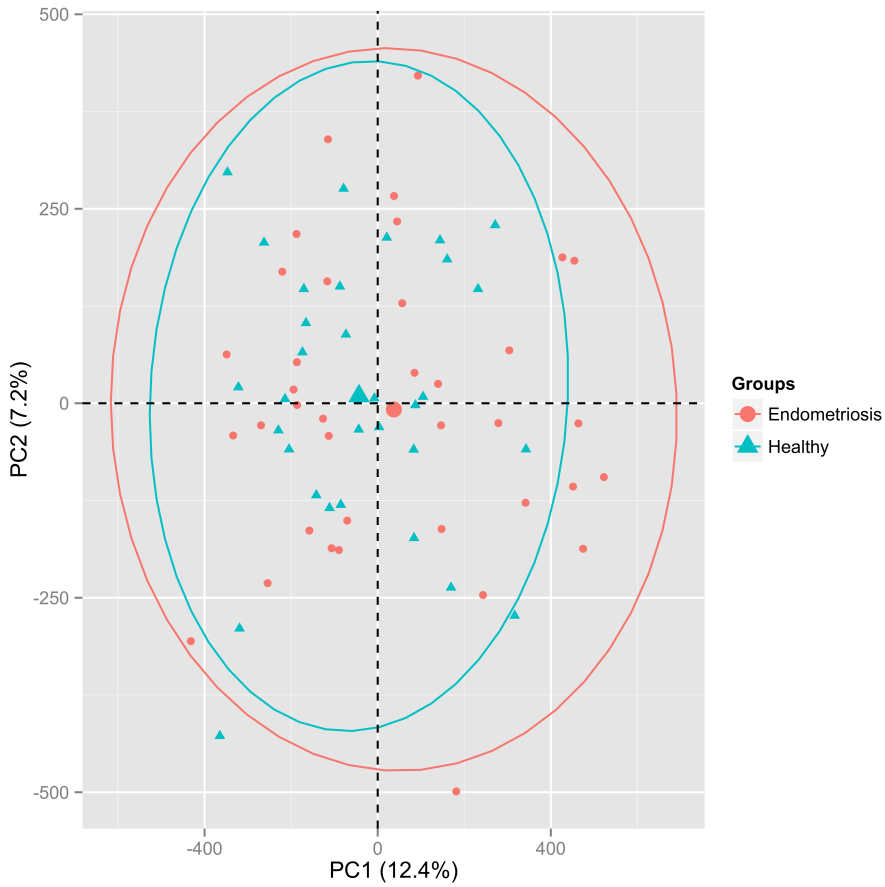

Figure 11. PCA plot of all endometrial and control samples after normalization. Endometriosis samples are denoted by red circles, while healthy controls are denoted by blue triangles. There is no distinction between the endometriosis patients and controls according to the PCA.

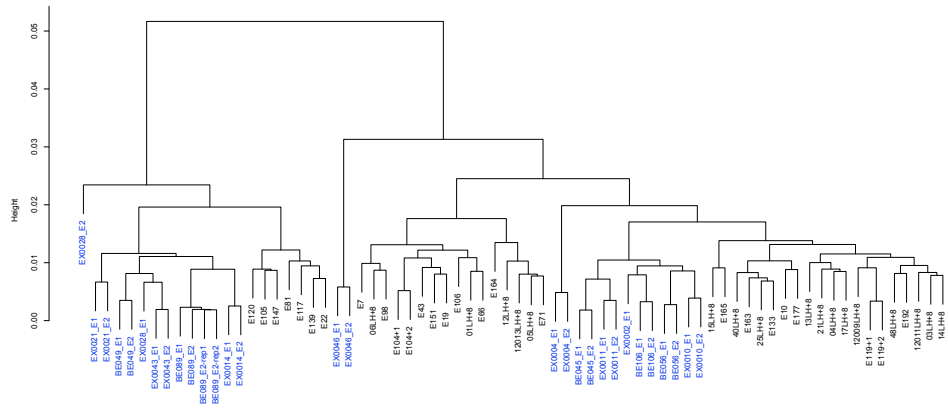

Figure 12. Hierarchical clustering analysis of all endometrial and control samples before batch effect correction. Samples in blue and black corresponds to Oxford and Tartu samples, respectively. The cluster dendrogram demonstrates samples were clustered according to collection center and not by disease status

3.3. Differential methylation analysis

Differential methylation analysis was performed between endometriosis patients and controls as well as between the menstrual cycle phases using *seqlm* method (Kolde et al., 2016) (See Section 1.2.6). Moreover, DMR analysis between endometriosis patients and controls revealed only 28 differentially DMRs. However, none of the DMRs remained significant after adjusting for co-variates such as age, nationality and body mass index (BMI). On the other hand, DMR analysis between the menstrual cycle phases, particularly adjacent menstrual cycle phases (M vs. P, MS vs. LS and LS vs. P) revealed large scale substantial methylation changes between the menstrual cycle groups, when either or both M and LS are included in the analysis (Saare et al., 2016) (See Table 2 from Publication II). This could be owing to the fact that the endometrial tissue undergoes major morphological changes including thickening during the LS phase and desquamation during the M phase, which are reflected in their methylation patterns as well. The entire DMR analysis presented here was corrected for false discovery rate (FDR<0.05).

3.4. Summary and impact

In this study, we analyzed the methylation changes taking place in the endometrium of endometriosis patients and healthy controls by considering menstrual cycle specific methylation changes. The results of this study suggests that overall endometrial DNA methylation patterns are highly similar between patients with endometriosis and healthy women. However, methylome changes may be primarily influenced by the menstrual cycle phases. Although DNA based diagnostic biomarkers are largely suggested for clinical set ups, especially for cancer and

mental disorders (Kim et al., 2018), our results suggest that methylation-based biomarkers may not be beneficial for the classification of endometriosis patients.

3.5. Contribution

In this project, I was responsible for the entire data analytics process, including data pre-processing, batch effect identification and correction, clustering analysis, PCA, evaluation of DMR methods and data visualization. I also drafted the manuscript related to my contribution.

4. METHYLOME ANALYSIS DURING THE TRANSITION FROM PRE-RECEPTIVE TO RECEPTIVE ENDOMETRIUM (PUBLICATION III)

In the previous chapter, we focused on identifying disease-specific endometrial biomarkers. In the current chapter, we focus on the methylation patterns of healthy and fertile endometrial tissue, emphasizing methylation changes that take place during the transition from the pre-receptive to the receptive endometrium. It is important to study the methylation changes during this transition since successful implantation of an embryo requires synchronization between a healthy embryo and a functionally competent endometrium (Mahajan, 2015). This phenomenon is termed as *window of implantation* (WOI) or endometrial receptivity. The WOI corresponds to the mid secretory phase (19–24 days) of a regular menstrual cycle. Therefore, our study included endometrial methylation profiles from the early secretory and to the mid secretory phase.

Numerous studies suggested a multitude of transcriptional changes that take place during the transition from the pre-receptive to the receptive endometrium (Carson et al., 2002; Díaz-Gimeno et al., 2014; Hu et al., 2014). Epigenetic modulators, mainly DNA methylation, may play an essential role in the transcriptional changes that take place during the WOI. For this reason in Publication III, we investigated the methylation changes taking place during the said transition. We also evaluated the methylation effect on gene expression changes by correlating DNA methylation profiles with the gene expression data profiled using RNA sequencing. This study comprises endometrial methylome profiles of 17 healthy and fertile women corresponding to the early secretary and mid secretory phase of the menstrual cycle generated by our project partners from the Competence Centre on Health Technologies, Tartu, Estonia. Our study utilizes an endometrial sample from the same set of individuals, thus eliminating any inter-individual variation. For gene expression-methylation correlation analysis, 14 biopsies of seven women profiled using RNA sequencing were utilized.

4.1. Data processing

Preprocessing and normalization of the methylation data were conducted using the bioconductor *RnBeads* package (Assenov et al., 2014).

4.2. Analysis of global methylation profiles of the pre-receptive and receptive endometrium

First, we visualized global methylation patterns in the pre-receptive and receptive endometrium using a probability density plot. The said plot (See Figure 1 of Publication II) clearly showed that the methylation profiles between the two time

points were nearly similar. Further, the hierarchical clustering analysis (See Section 1.2.5) demonstrated that the methylation patterns were mostly clustered according to the individuals (See Figure 13) .

4.3. Differential methylation analysis

As global methylation profiling showed that there may not be large-scale differences between the pre-receptive and receptive endometrium, we expected to observe only small-scale methylation differences between them. To explore these small-scale methylation changes, we utilized site-level analysis methods such as *RnBeads* and *Wilcoxon signed rank-test* to compute differential methylation. The former utilizes a *moderated t-test* to compute differential methylation (DM). Moreover, the latter was also used in the site-level analysis since the global methylation distribution deviated slightly from the normal distribution (data not shown) (See Section 1.2.6). In addition, we used the region-level analysis method *seqlm* (Kolde et al., 2016) and extracted individual CpG sites from the DMRs for computing DM. Thus we employed a combination of *RnBeads* and *Wilcoxon signed rank-test* and *seqlm* for differential methylation and selected the intersection of the same to derive the DM CpG sites to reduce any potential false positives and select the differential methylation analysis results with greater confidence (See Figure 14).

The clustering analysis of DM sites clearly showed two main branches that divided the methylation samples according to the pre-receptive and receptive phase except for one sample that clustered together with receptive phase samples. Additionally, three samples from the receptive phase were also clustered in the first branch (See Figure 13).

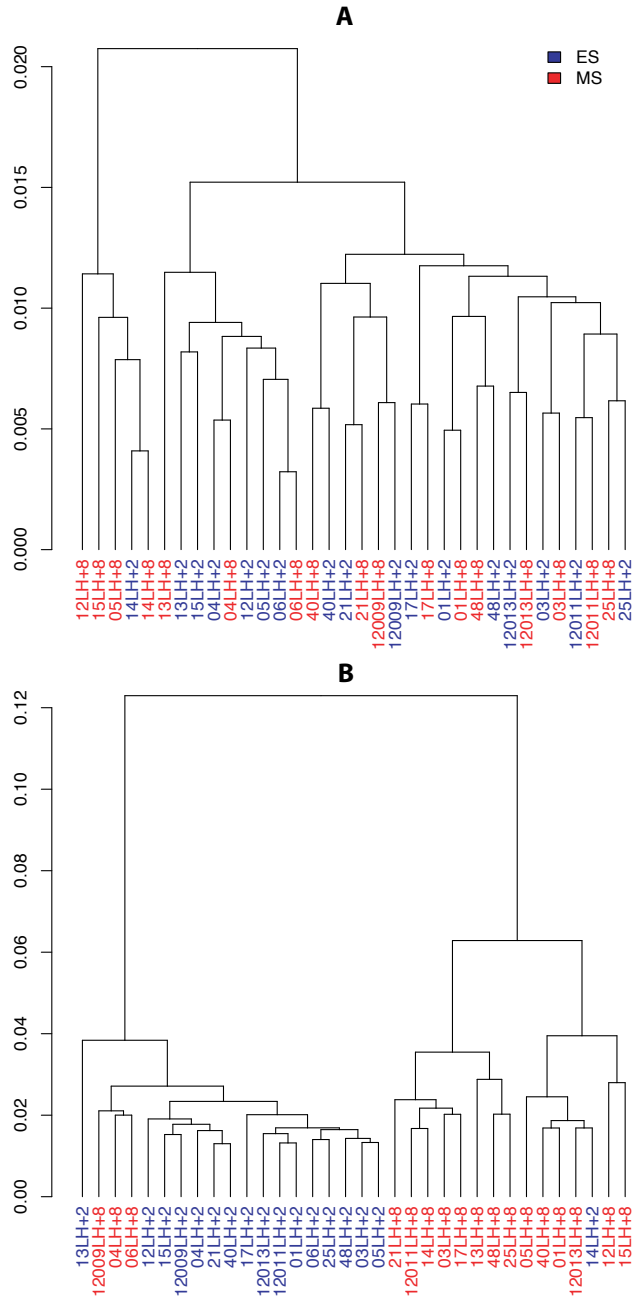

Figure 13. Hierarchical clustering analysis of the processed (A) methylation data and differentially methylated data (B) from the pre-receptive and receptive endometrium. ES – Early-secretory (pre-receptive LH+2) samples; MS – Mid-secretory (receptive LH+8) samples.

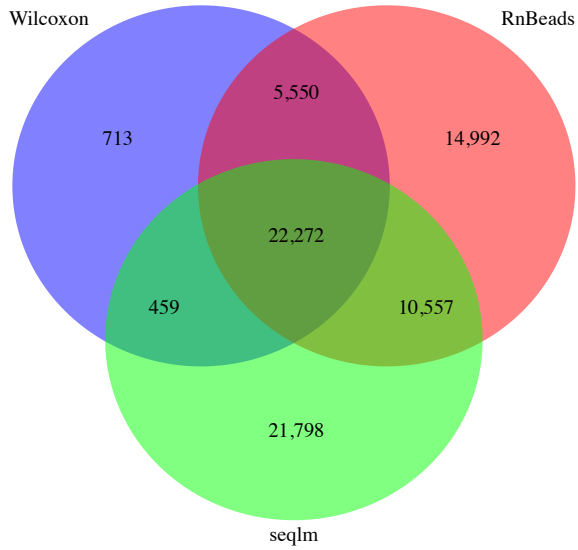

Figure 14. Venn diagram highlighting the overlap between the three methods (*Wilcoxon*, *Rnbeads* and *seqlm*) used for identifying differentially methylated CpG sites between the pre-receptive (LH+2) and receptive phases (LH+8) of endometrium.

4.4. Gene enrichment analysis

GO enrichment analysis was performed by *g:Profiler* (Reimand et al., 2007, 2016) for the DM CpGs. Additionally, *GOsummaries* (Kolde and Vilo, 2015) was used for summarized word cloud visualization (See Figure 5 from Publication III) based on the results of *g:Profiler*. In site-level analyses, we found that CpGs mapped with decreased methylation were primarily associated with immune response regulation and cell activation and adhesion. On the other hand, in site-level analyses with increased methylation, extracellular matrix organization, cellular signaling, regulation, and development were prominent. The same trend was reflected in the DMR analysis results as well (See Figure 5 from Publication III). In the case of GO analysis with genes correlating with the gene expression, positively correlated genes are related to extracellular matrix organization and immune response. Conversely, we did not see any enrichment for negative correlations (Kukushkina et al., 2017).

4.5. Summary and impact

In Publication III, we analyzed the methylation changes taking place in pre-receptive (LH+2) and receptive (LH+8) endometrium of fertile women within the same menstrual cycle. Overall, our study's results suggest methylation profiles have quite similar pre-receptive and receptive endometrium on a global scale. However, small scale changes were detected using DM analysis.

4.6. Contribution

In this publication, I contributed towards data processing, visualization and clustering analysis. Moreover, I performed differential methylation using *seqlm* and performed gene ontology analysis. I also drafted this manuscript related to my contribution.

5. MULTIVARIABLE SURVIVAL ANALYSIS ON LARGE COLLECTIONS OF DNA METHYLATION DATA (PUBLICATION IV)

In the previous chapters (Publications I–III), we provided the analytical pipeline for evaluating DNA methylation biomarkers in healthy as well as diseased samples, primarily using the methylation data derived from the project collaborators. It is also possible to determine methylation biomarkers by using publicly available methylation data. Moreover, cancer-based biomarkers are assessed by various research groups, since aberrant methylation patterns serve as a hallmark for several cancer types and remain an attractive biomarker candidate for cancer-risk prediction, therapy management, and early diagnosis (Laird, 2003). In this project, we used methylation and clinical data obtained from "The Cancer Genome Atlas" (TCGA) database to evaluate methylation-based prognostic biomarkers to determine the survival time of cancer patients. The current chapter is based on Publication IV, which addresses the second aim of this thesis, i.e., to develop a web tool for correlating methylation levels with cancer survival time. Using the developed web tool *MethSurv*, users can evaluate the prognostic potential of cancer biomarker candidates across 25 different cancer types.

5.1. Survival analysis methods and visualization

After methylation data as well as clinical data was acquired from TCGA database, we then matched the downloaded methylation data with the available clinical data including survival status, patient characteristics (age, sex, height, weight, race, etc.) and clinicopathologic features such as the stage and grade of the cancer, and so on. (See Supplementary Table 1 from Publication IV for more details).

To perform survival analysis, we fitted Cox proportional hazards models (Cox, 1972) using the R survival package (Therneau, 2014). The patient's methylation levels were stratified according to mean, median, and lower or upper percentiles. Moreover, the outcome oriented method was implemented using Maximally Selected Rank Statistics (maxstat) (Hothorn and Lausen, 2002). The best cut-off is derived according to the highest hazard ratio. *MethSurv* facilitates both univariable (using only methylation levels in the prediction model) and multivariable survival analyses (accounting for patient characteristics and clinicopathologic features in addition to methylation levels in the prediction models) (See section 1.2.8).

Apart from survival analysis, it is also possible to explore methylation pattern across a given gene in the form of a heat map using seamless integration with *ClustVis* (Metsalu and Vilo, 2015). The entire pipeline adapted for the construction of *MethSurv* has been shown in Figure 15.

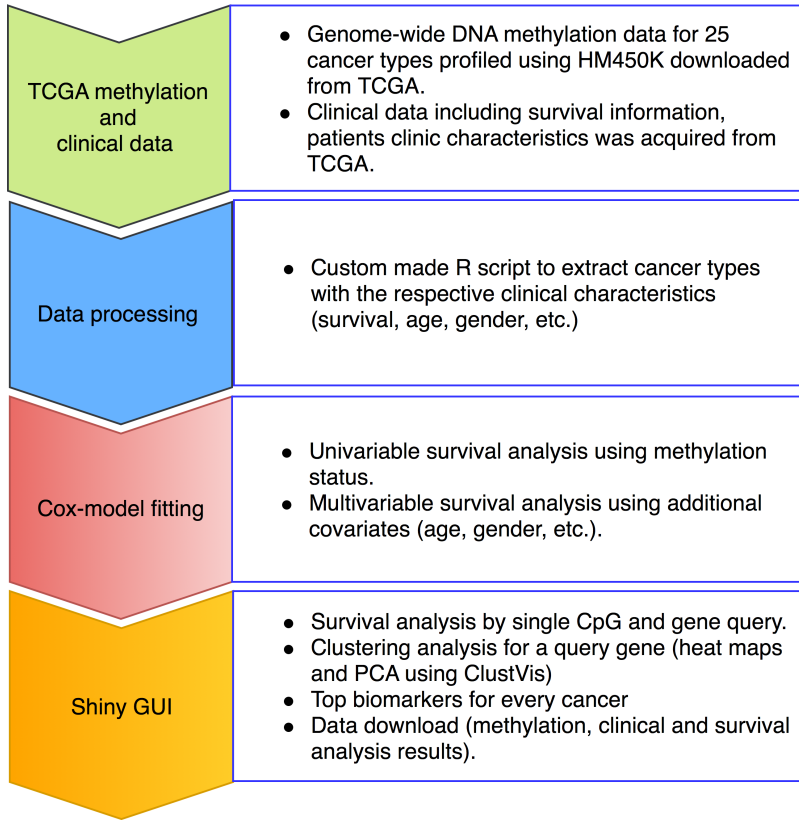

Figure 15. The pipeline adapted for the construction of *Methsurv*. Methylation, survival and clinical data were acquired from TCGA. Customized R script was used to match methylation data with relevant survival and clinical data. Cox-proportional hazard model was utilized to perform univariable and multivariable survival analysis implemented by shiny web-interface. *MethSurv* contains five main components: (A) survival analysis for a single CpG. (B) summarized tabular view for a query gene and genomic region. (C) clustering analysis (PCA and heat map) for a query gene (D) top biomarkers for every cancer and (E) data download options to download processed methylation data and pre-computed survival analysis summary for every type of cancer.

5.2. Description of the user interface

Survival analysis for a single CpG query. Using the single CpG analysis tab of *Methsurv*, users can conduct survival analysis for a single CpG for any of the available genes. Using this feature, users can choose cancer type, gene, relation to island, gene sub-region and a CpG site. In terms of statistical parameters, they can choose which cut-off point to dichotomize patient's methylation profiles (mean, median, upper and lower quantiles, *maxstat* and best split). Upon selecting all the required parameters, *Methsurv* generates the Kaplan-Meier plot (Figure 16) with the statistical summary (HR with CI, proportional hazard P-value, Wald test P-value, log-likelihood ratio test's P-value for the model fit, proportional hazard

test's P-value and mean, median and range of methylation beta values). In addition, the following have been displayed: a distribution plot highlighting the cut-off points used for dichotomization and a violin plot depicting the distribution, median and interquartile range of methylation profiles of the query CpG site in relation to patient characteristics such as age, sex and stage (See Figure 2 from Publication IV).

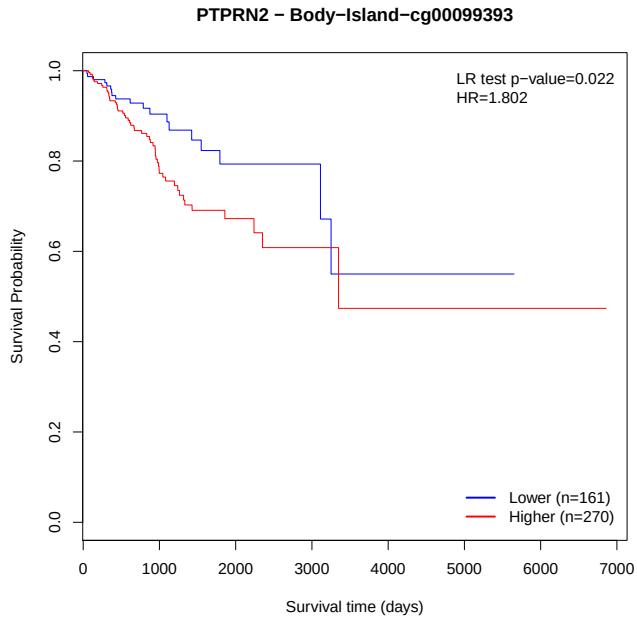

Figure 16. Survival plot of *cg00099393-PTPRN2* for endometrial carcinoma generated using *Methsurv*.

Users can view a tabular summary of the survival analysis' results of all 25 cancers for a query gene using *Methsurv's* "All cancers" tab. In addition, users can also retrieve the survival analysis' summary within a queried genomic region (e.g. Chromosome 17 location: 41278000 to 41278000).

Identifying top survival biomarkers. Users can quickly browse through the most significant biomarkers for each type of cancer using the "Top biomarkers" tab of *Methsurv*.

Clustering analysis visualization for a query gene. Clustering analysis of a given cancer and a query gene is facilitated in *Methsurv* using the "Gene visualization" tab which allows users to visualize the clustering of individual CpGs with in a query gene. This enables users to visually associate methylation levels with patient characteristics (age, gender, and so on.) as well gene sub-regions. In addition we also integrated *Clustvis* (Metsalu and Vilo, 2015) for advanced clustering visualization in the form of a heat map and PCA (See Figure 17).

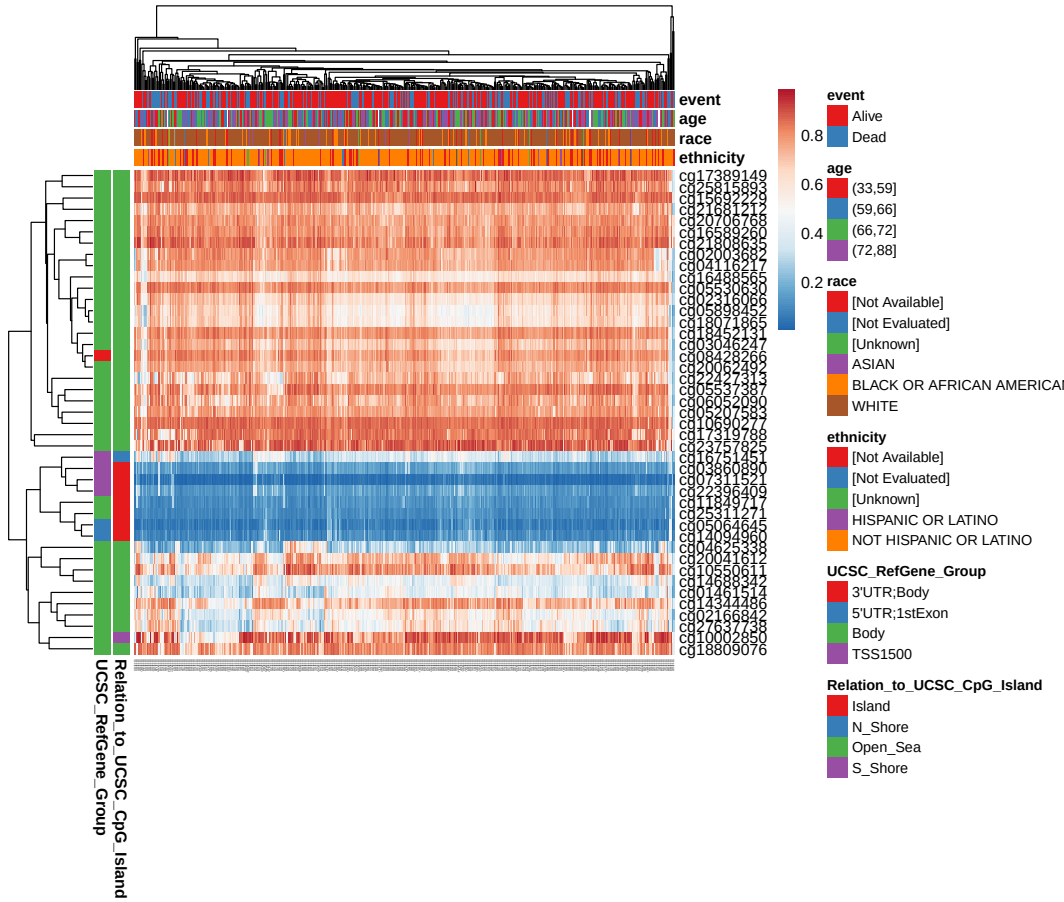

Figure 17. A heat map depicting clustering analysis of *EGFR* using TCGA data on lung adenocarcinoma. This heat map was generated using average linkage method concenering correlation distance. Methylation levels (1 = fully methylated; 0 = fully unmethylated) are shown as a continuous variable from ranging from a blue to a red color. The rows correspond to the CpGs and the columns correspond to the patients.

Data download. Using the download option in *Methsurv*, users can download processed methylation data matrix and clinical data for any of the available cancers. It is also possible to download pre-computed survival analysis results in the form of a table for any particular cancer type.

5.3. Example of *MethSurv* using the known biomarker

As an example of the evaluation of *MethSurv's* usability, we compared *MethSurv* results with the study results provided by Li et al. (2014). included 98 primary breast tumor samples obtained from Shenzhen Maternal and Child Health Hospital (independent of the TCGA samples). Li et al. (2014) showed that *PTPRO* promoter hypermethylation is associated with poorer overall survival (HR = 2.7; 95% CI:

1.1-6.2; P = 0.023). This effect can be observed in *Methsurv* (See Figure 4A from Publication IV) using TCGA data for *cg22374861-TSS200* annotated to *PTPRO* (number of patients = 782; HR = 1.8; 95% CI: 1.2-2.7; P = 0.0054). More such examples evaluating *Methsurv* results in conjunction with previously published biomarkers are presented in Publication IV.

5.4. Summary and impact

In Publication IV, we analyzed large-scale methylation datasets for 25 different cancer types from the TCGA. We observed that this large-scale data shows great promise with respect to performing robust statistical analysis for survival prediction. We further presented a web tool *MethSurv* to correlate overall cancer survival with DNA methylation levels. Using *MethSurv*, users can perform univariable and multivariable survival analyses with only methylation levels and methylation levels as well as patient characteristics, respectively. In this way, we provide an assessment of the way methylation, in combination with relevant clinical covariates, may influence a cancer patient's survival rate. We believe that *MethSurv* can serve as a valuable resource for cancer biologists in generating hypotheses pertaining to cancer biomarkers. To enable the usage of this web tool across a wider scientific community, we made the tool interactive and easy to use, so that even non-bioinformaticians can use it without any difficulty.

5.5. Contribution

I downloaded the data from the TCGA, processed and prepared the database for the web interface, designed the query system, and drafted the entire manuscript. Further, I researched the literature to prepare the datasets for the aforementioned case study.

6. DISCUSSION

Bioinformatics methods are required for analysis of large-scale biological data and therefore remain central to accelerating biomarker discovery. The current thesis identified methylation based biomarker candidates by using high quality tissue samples from several experimental conditions. The thesis contributions were twofold. Firstly, a comprehensive pipeline was introduced to identify methylation biomarkers in different experiments. Secondly, a novel web-tool *Methsurv* was developed.

The discussion part of the thesis starts with exploring the strengths and weaknesses of the experimental design followed by suggestions on future research and final remarks.

6.1. Strengths and weaknesses of the studies

As mentioned in the preliminaries section (See Section 1.1.5), the entire thesis utilizes methylation data profiled using the HM450K array. The HM450K array is a powerful and cost-effective method of quantifying the DNA methylation status from different cell or tissue types from the human genome. The HM450K array covers several functionally known regulatory regions of genome (Bibikova et al., 2011). Some of the major drawbacks of using this array includes unequal distribution of the probes and coverage of fewer enhancer regions. Whole genome bisulfite sequencing may successfully overcome the aforementioned probe selection bias and comprehensive genomic coverage. However, higher costs and the need for sophisticated computational resources compared to the HM450K array (Kurdyukov and Bullock, 2016) must be borne in mind. Alternatively, one could choose a recent and cost-effective technology Illumina MethylationEPIC BeadChip, which contains an additional 350,000 CpGs, considered as potential enhancers, compared to the HM450K array (McCartney et al., 2016).

The work presented in Publication I provided an opportunity to explore methylation patterns of rare tissues or tissue subtypes. Moreover, the existence of tissue specific methylation patterns highlighted in our study can be used for tissue identification in a forensic context (Sijen, 2015). However, there are drawbacks which can be improved for future studies. Firstly, the study was performed when the HM450K array was newly introduced. Therefore, much of the technical advances made in the recent years in terms of data analysis such as the discovery of cross-reactive probes (Chen et al., 2013) and incorrectly reported methylation status due to the presence of germ line or somatic mutations (Zhou et al., 2018) may therefore confound region specific methylation patterns. Our hierarchical clustering and DMR analysis results strongly support tissue-specificity. One of the interesting future work would be to combine methylation data from our current study with NIH Roadmap Epigenomics data (Kundaje et al., 2015) to describe extended tissue specific methylation profiles. Some of the relevant tissues to combine include,

oesophagus, thymus, liver and kidney.

The next part of the thesis focused on identifying methylation based biomarkers in endometriosis disease prediction. Since endometrium tissue undergoes hormonal changes throughout the menstrual cycle, our study included samples from the entire menstrual cycle phase. Moreover, our study was the first to evaluate both the menstrual cycle influences and endometriosis disease status at the same time (Saare et al., 2016). However, some of the limiting factors of the experimental study design include inter-individual variability, because of the diverse nature of the samples from both Tartu and Oxford cohorts. Additionally a major drawback of the study includes complexity of endometrium tissue, since endometrium consists of different cell types such as epithelial, stromal and immune cells (Mortlock et al., 2019). Though we demonstrated the elimination of cohort specific batch effect (See Figure 12 and Figure 2 from Publication II), the results from Publication II suggest methylation may not be a powerful indicator of biological changes taking place in endometriosis according to, PCA (Figure 11), hierarchical clustering (Figure 12) and differential methylation analysis (See Section 3.3). We next identified methylation changes during the transition from pre-receptive to receptive endometrium. Compared to Publication II, samples from Publication III included endometrial tissues from the same set of individuals which eliminates any possible inter-individual variability (Kukushkina et al., 2017). However, the problem of endometrial tissue heterogeneity still exists. One of the suggestions to eliminate cell type heterogeneity is to use single cell DNA methylation sequencing accounting for epithelial and stromal cell types and subsequently use the same for reference based cell-type composition adjustment (Teschendorff and Zheng, 2017; Titus et al., 2017).

Next, in Publication IV, we utilized big data from TCGA, to identify methylation based survival biomarkers for different cancer types. The *Methsurv* web tool complements several existing web tools (Huang et al., 2014; Díez-Villanueva et al., 2015; Koch et al., 2015; Xiong et al., 2016) for analyzing cancer methylation patterns by including comprehensive visualization panels for survival outcome and heat map visualization of CpG methylation patterns. Some of the drawbacks of our tool are that *Methsurv* is applicable only for overall survival (Modhukur et al., 2017) which limits survival analysis for specific cases such as treatment type, and drug specific survival because of the limited number of samples. Although *Methsurv* is a valuable intuitive tool for methylation based biomarkers, prospective studies are needed to properly evaluate biomarkers for clinical practice. As a future work, it would be interesting to compare survival predictions using the Cox-proportional hazard model with the machine learning method, for example, random forest. As an update, we will consider processing user's own data for survival prediction. Further, it will be intriguing for the bioinformaticians to develop novel methods to classify tumor types based on TCGA methylation profiles (Angermueller et al., 2016, 2017; Celli et al., 2018).

6.2. Final remarks

The thesis has demonstrated the power of bioinformatics to identify methylation based biomarkers. We also hope that the thesis provided a comprehensive set of data analysis methods and web tool to understand how methylation patterns can be translated into a diagnostic tool for clinical research. Moreover, the pipeline presented in this thesis can be adapted for other high throughput data such as NGS or other types of microarrays. We hope that we have convinced the readers about the need for bioinformatics tools and methodologies to understand an exciting field of science. Although we acknowledge the power of methylation based biomarkers (Laird, 2003), we suggest the investigation of other epigenetic modifications including histone modifications, chromatic modifications and microRNA expression to understand the overall biological mechanism. Given enough computational resources and funding, the next generation sequencing will be a promising technology to understanding epigenetic changes in the lab which potentially could be used for diagnosis in the clinic.

CONCLUSIONS

Methylation patterns are essential for a myriad of biological processes such as growth and development; they are also associated with disease progression. This thesis aimed to identify DNA methylation patterns as biomarker candidates in normal and diseased human tissues using bioinformatics and statistical approaches. To fulfill this aim, we worked on different biological experiments that are of high value, since the tissue types and the experimental set-up was unique.

The first part of this thesis involves the application of bioinformatics methods to identify methylation-based biomarker candidates. The primary bioinformatics analysis challenge is the appropriate selection of methods to decipher the methylation patterns and eliminate any possible non-biological variation in order to identify condition-specific biomarkers. To overcome these challenges and offer solutions for the biological hypothesis, the following steps were followed:

- Performed hierarchical clustering analysis, which showed tissues can be classified according to the methylation patterns provided in Publication I.
- Integrated analysis between gene expression patterns and tissue methylation profiles using Pearson correlation coefficient in Publication I, which showed an overall inverse correlation between gene expression and methylation.
- Performed batch effect correction of the said effect existing between the experimental samples in endometriosis patients and controls and further detected menstrual cycle methylation patterns of the endometrium in Publication II.
- Applied a careful selection of differential analysis methods which detected small-scale methylation changes during the transition from pre-receptive to the receptive endometrium in Publication III.

Overall, the methodological steps of the aforementioned pipeline can be customized for other methylation datasets for the detection of methylation-based biomarker candidates.

Subsequently the second part of this thesis focuses on the identification of methylation-based marker candidates in order to detect survival time in cancer patients. The work accomplished on these lines has contributed to the development of a novel user-friendly web tool *MethSurv*, which is the first web tool for evaluating methylation based survival biomarkers in Publication IV. *MethSurv*, tool enables the scientific community easy to assess methylation based biomarkers by using *Cox-proportional hazard models*. *MethSurv* uses methylation and clinical data derived from TCGA, wherein the data processing, statistical calculations, and visualizations can be a cumbersome task for non-bioinformaticians. In particular, *MethSurv*, provides multiple options for the users including survival analysis based on individual CpG, gene-specific survival, visualization of methylation patterns for a queried gene in the form of a heat map and PCA. In this tool, top biomarkers for each for cancer type are readily available. Moreover, all the recomputed results are

available for the users to downloaded by users enabling further analysis queries.

Overall, the work presented in this thesis presented a comprehensive bioinformatics workflow used to identify methylation patterns as potential biomarker candidates in various biological conditions. Further biological experiments conducted in this regard will provide more evidence concerning the accessibility of these biomarkers.

BIBLIOGRAPHY

R. Akulenko, M. Merl, and V. Helms. Beclear: batch effect detection and adjustment in dna methylation data. *PloS one*, 11(8):e0159921, 2016.

A. Alexa and J. Rahnenfuhrer. topgo: enrichment analysis for gene ontology. *R package version*, 2(0), 2010.

C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle. Deep learning for computational biology. *Molecular systems biology*, 12(7):878, 2016.

C. Angermueller, H. J. Lee, W. Reik, and O. Stegle. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome biology*, 18 (1):67, 2017.

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

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.

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

E. Baca-García, A. S. González, P. G. Diaz-Corralero, I. G. García, and J. d. Leon. Menstrual cycle and profiles of suicidal behaviour. *Acta Psychiatrica Scandinavica*, 97(1):32–35, 1998.

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

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

R. L. Bowen and C. S. Atwood. Living and dying for sex. *Gerontology*, 50(5): 265–290, 2004.

A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, et al. ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic acids research*, 31(1):68–71, 2003.

S. Bulterijs, R. S. Hull, V. C. Björk, and A. G. Roy. It is time to classify biological aging as a disease. *Frontiers in genetics*, 6:205, 2015.

X. Cai, Y. Lu, C. Tang, X. Lin, J. Ye, W. Li, Z. He, and F. Li. Effect of interleukin-6 promoter DNA methylation on the pathogenesis of systemic lupus erythematosus. *Zhonghua yi xue za zhi*, 97(19):1491–1495, 2017.

V. Caplakova, E. Babusikova, E. Blahovcova, T. Balharek, M. Zelieskova, and J. Hatok. DNA methylation machinery in the endometrium and endometrial cancer. *Anticancer research*, 36(9):4407–4420, 2016.

D. D. Carson, E. Lagow, A. Thathiah, R. Al-Shami, M. C. Farach-Carson, M. Vernon, L. Yuan, M. A. Fritz, and B. Lessey. Changes in gene expression during the early to mid-luteal (receptive phase) transition in human endometrium detected by high-density microarray screening. *Molecular human reproduction*, 8(9): 871–879, 2002.

E. Cazaly, R. Thomson, J. R. Marthick, A. F. Holloway, J. Charlesworth, and J. L. Dickinson. Comparison of pre-processing methodologies for illumina 450k methylation array data in familial analyses. *Clinical epigenetics*, 8(1):75, 2016.

F. Celli, F. Cumbo, and E. Weitschek. Classification of large DNA methylation datasets for identifying cancer drivers. *Big data research*, 13:21–28, 2018.

Y.-a. Chen, M. Lemire, S. Choufani, D. T. Butcher, D. Grafodatskaya, B. W. Zanke, S. Gallinger, T. J. Hudson, and R. Weksberg. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*, 8(2):203–209, 2013.

D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34:187–220, 1972. doi: 10.2307/2985181. URL `http://www.jstor.org/stable/2985181`.

S. Davis, S. Bilke, and T. Triche Jr. Bootwalla, M. methylumi: Handle Illumina methylation data. R package version 2.14. 0, 2015.

S. Dedeurwaerder, M. Defrance, M. Bizet, E. Calonne, G. Bontempi, and F. Fuks. A comprehensive overview of Infinium HumanMethylation450 data processing. *Briefings in bioinformatics*, 15(6):929–941, 2013.

P. Díaz-Gimeno, M. Ruíz-Alonso, D. Blesa, and C. Simón. Transcriptomics of the human endometrium. *International Journal of Developmental Biology*, 58 (2-3-4):127–137, 2014.

A. Díez-Villanueva, I. Mallona, and M. A. Peinado. Wanderer, an interactive viewer to explore DNA methylation and gene expression data in human cancer. *Epigenetics & chromatin*, 8(1):22, 2015.

P. Du, W. A. Kibbe, and S. M. Lin. lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, 24(13):1547–1548, 2008.

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

C. Dupont, D. R. Armant, and C. A. Brenner. Epigenetics: definition, mechanisms and clinical perspective, 2009.

E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC bioinformatics*, 10(1):48, 2009.

R. Edgar, M. Domrachev, and A. E. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30 (1):207–210, 2002.

M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.

J. A. Ellis, J. E. Munro, R. A. Chavez, L. Gordon, J. E. Joo, J. D. Akikusa, R. C. Allen, A.-L. Ponsonby, J. M. Craig, and R. Saffery. Genome-scale case-control analysis of CD4+ T-cell DNA methylation in juvenile idiopathic arthritis reveals potential targets involved in disease. *Clinical epigenetics*, 4(1):20, 2012.

I. E. Eryilmaz, G. Cecener, S. Erer, U. Egeli, B. Tunca, M. Zarifoglu, B. Elibol, A. Bora Tokcaer, E. Saka, M. Demirkiran, et al. Epigenetic approach to early-onset parkinson's disease: low methylation status of snca and park2 promoter regions. *Neurological research*, 39(11):965–972, 2017.

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

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

R. Gabbianelli and E. Damiani. Epigenetics and neurodegeneration: Role of early-life nutrition. *The Journal of nutritional biochemistry*, 2018.

M. Gardiner-Garden and M. Frommer. Cpg islands in vertebrate genomes. *Journal of molecular biology*, 196(2):261–282, 1987.

B. D. W. Group, A. J. Atkinson Jr, W. A. Colburn, V. G. DeGruttola, D. L. DeMets, G. J. Downing, D. F. Hoth, J. A. Oates, C. C. Peck, R. T. Schooley, et al. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics*, 69(3):89–95, 2001.

S. Guibert and M. Weber. Functions of DNA methylation and hydroxymethylation in mammalian development. *Curr Top Dev Biol*, 104:47–83, 2013.

S.-W. Guo. Epigenetics of endometriosis. *Molecular human reproduction*, 15(10): 587–607, 2009.

M. Hadchouel, H. Farza, D. Simon, P. Tiollais, and C. Pourcel. Maternal inhibition of hepatitis b surface antigen gene expression in transgenic mice correlates with de novo methylation. *Nature*, 329(6138):454, 1987.

I. Henry, C. Bonaiti-Pellie, V. Chehensse, C. Beldjord, C. Schwartz, G. Utermann, and C. Junien. Uniparental paternal disomy in a genetic cancer-predisposing syndrome. *Nature*, 351(6328):665, 1991.

S. Horvath. DNA methylation age of human tissues and cell types. *Genome biology*, 14(10):3156, 2013.

T. Hothorn and B. Lausen. Maximally selected rank statistics in R. *R News*, 2(1): 3–5, 2002.

T. Hothorn and B. Lausen. On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*, 43:121–137, 2003. ISSN 01679473. doi: 10.1016/S0167-9473(02)00225-6. URL `www.elsevier.com/locate/csda`.

S. Houshdaran, Z. Zelenko, J. C. Irwin, and L. C. Giudice. Human endometrial DNA methylome is cycle-dependent and is associated with gene expression regulation. *Molecular Endocrinology*, 28(7):1118–1135, 2014.

S. Hu, G. Yao, Y. Wang, H. Xu, X. Ji, Y. He, Q. Zhu, Z. Chen, and Y. Sun. Transcriptomic changes during the pre-receptive to receptive transition in human endometrium detected by rna-seq. *The Journal of Clinical Endocrinology & Metabolism*, 99(12):E2744–E2753, 2014.

D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4(1):44, 2008.

W.-Y. Huang, S.-D. Hsu, H.-Y. Huang, Y.-M. Sun, C.-H. Chou, S.-L. Weng, and H.-D. Huang. MethHC: a database of DNA methylation and gene expression in human cancer. *Nucleic acids research*, 43(D1):D856–D861, 2014.

A. E. Jaffe, P. Murakami, H. Lee, J. T. Leek, M. D. Fallin, A. P. Feinberg, and R. A. Irizarry. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International journal of epidemiology*, 41(1): 200–209, 2012.

H. S. Jang, W. J. Shin, J. E. Lee, and J. T. Do. CpG and Non-CpG Methylation in Epigenetic Gene Regulation and Brain Function. *Genes*, 8(6):148, 2017.

W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

P. A. Jones and S. B. Baylin. The epigenomics of cancer. *Cell*, 128(4):683–692, 2007.

G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. Gopinath, G. Wu, L. Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl_1):D428–D432, 2005.

M. Jung and G. P. Pfeifer. Aging and DNA methylation. *BMC biology*, 13(1):7, 2015.

M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

H. R. Kim, X. Wang, and P. Jin. Developing DNA methylation-based diagnostic biomarkers. *Journal of Genetics and Genomics*, 2018.

A. Koch, T. De Meyer, J. Jeschke, and W. Van Criekinge. MEXPRESS: visualizing

expression, DNA methylation and clinical TCGA data. *BMC genomics*, 16(1): 636, 2015.

R. Kolde and J. Vilo. Gosummaries: an R package for visual functional annotation of experimental data. *F1000Research*, 4, 2015.

R. Kolde, K. Märtens, K. Lokk, S. Laur, and J. Vilo. seqlm: an MDL based method for identifying differentially methylated regions in high density methylation array data. *Bioinformatics*, 32(17):2604–2610, 2016.

V. Kukushkina, V. Modhukur, M. Suhorutšenko, M. Peters, R. Mägi, N. Rahmioglu, A. Velthut-Meikas, S. Altmäe, F. J. Esteban, J. Vilo, et al. DNA methylation changes in endometrium and correlation with gene expression during the transition from pre-receptive to receptive phase. *Scientific reports*, 7(1):3916, 2017.

M. Kulis and M. Esteller. DNA methylation and cancer. *Adv Genet*, 70(10):27–56, 2010.

A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317, 2015.

S. Kurdyukov and M. Bullock. DNA methylation analysis: choosing the right method. *Biology*, 5(1):3, 2016.

P. W. Laird. The power and the promise of DNA methylation markers. *Nature Reviews Cancer*, 3(4):253–266, 2003.

C. Lazar, S. Meganck, J. Taminau, D. Steenhoff, A. Coletta, C. Molter, D. Y. Weiss-Solís, R. Duque, H. Bersini, and A. Nowé. Batch effect removal methods for microarray gene expression data integration: a survey. *Briefings in bioinformatics*, 14(4):469–490, 2012.

E. Li, C. Beard, and R. Jaenisch. Role for dna methylation in genomic imprinting. *Nature*, 366(6453):362, 1993.

S.-y. Li, R. Li, Y.-l. Chen, L.-k. Xiong, H.-l. Wang, L. Rong, and R.-c. Luo. Aberrant ptpro methylation in tumor tissues as a potential biomarker that predicts clinical outcomes in breast cancer patients. *BMC genetics*, 15(1):67, 2014.

F. Lienert, C. Wirbelauer, I. Som, A. Dean, F. Mohn, and D. Schübeler. Identification of genetic elements that autonomously determine DNA methylation states. *Nature genetics*, 43(11):1091, 2011.

R. Lister, E. A. Mukamel, J. R. Nery, M. Urich, C. A. Puddifoot, N. D. Johnson, J. Lucero, Y. Huang, A. J. Dwork, M. D. Schultz, et al. Global epigenomic reconfiguration during mammalian brain development. *Science*, 341(6146): 1237905, 2013.

K. Lokk, V. Modhukur, B. Rajashekar, K. Märtens, R. Mägi, R. Kolde, M. Koltšina, T. K. Nilsson, J. Vilo, A. Salumets, et al. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome biology*, 15(4):3248, 2014.

H. Lu, X. Liu, Y. Deng, and H. Qing. DNA methylation, a hand behind neurodegenerative diseases. *Frontiers in aging neuroscience*, 5:85, 2013.

N. Mahajan. Endometrial receptivity array: Clinical application. *Journal of human reproductive sciences*, 8(3):121, 2015.

J. Maksimovic, L. Gordon, and A. Oshlack. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome biology*, 13(6):R44, 2012.

V. Matys, E. Fricke, R. Geffers, E. Gößling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, et al. Transfac®: transcriptional regulation, from patterns to profiles. *Nucleic acids research*, 31(1):374–378, 2003.

D. L. McCartney, R. M. Walker, S. W. Morris, A. M. McIntosh, D. J. Porteous, and K. L. Evans. Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. *Genomics data*, 9:22–24, 2016.

W. Meng, Z. Zhu, X. Jiang, C. L. Too, S. Uebe, M. Jagodic, I. Kockum, S. Murad, L. Ferrucci, L. Alfredsson, et al. DNA methylation mediates genotype and smoking interaction in the development of anti-citrullinated peptide antibody-positive rheumatoid arthritis. *Arthritis research & therapy*, 19(1):71, 2017.

D. M. Messerschmidt, B. B. Knowles, and D. Solter. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes & development*, 28(8):812–828, 2014.

T. Metsalu and J. Vilo. ClustVis: A web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic Acids Research*, 43(W1):W566–W570, 2015.

V. Modhukur, T. Iljasenko, T. Metsalu, K. Lokk, T. Laisk-Podar, and J. Vilo. MethSurv: a web tool to perform multivariable survival analysis using DNA methylation data. *Epigenomics*, 2017.

T. J. Morris and S. Beck. Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data. *Methods*, 72:3–8, 2015.

S. Mortlock, R. Restuadi, R. Levien, J. E. Girling, S. J. Holdsworth-Carson, M. Healey, Z. Zhu, T. Qi, Y. Wu, S. W. Lukowski, et al. Genetic regulation of methylation in human endometrium and blood and gene targets for reproductive diseases. *Clinical epigenetics*, 11(1):49, 2019.

E. Nilsson, P. A. Jansson, A. Perfilyev, P. Volkov, M. Pedersen, M. K. Svensson, P. Poulsen, R. Ribel-Madsen, N. L. Pedersen, P. Almgren, et al. Altered DNA methylation and differential expression of genes influencing metabolism and inflammation in adipose tissue from subjects with type 2 diabetes. *Diabetes*, 63 (9):2962–2976, 2014.

M. Okano, D. W. Bell, D. A. Haber, and E. Li. DNA methyltransferases Dnmt3a

and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257, 1999.

M. S. Olesen, A. Starnawska, J. Bybjerg-Grauholm, A. P. Bielfeld, I. Agerholm, A. Forman, M. T. Overgaard, and M. Nyegaard. Biological age of the endometrium using dna methylation. *Reproduction*, 155(2):167–172, 2018.

K. Pearson. Principal components analysis. *The London, Edinburgh and Dublin Philosophical Magazine and Journal*, 6(2):566, 1901.

N. Pervjakova, S. Kasela, A. P. Morris, M. Kals, A. Metspalu, C. M. Lindgren, A. Salumets, and R. Mägi. Imprinted genes and imprinting control regions show predominant intermediate methylation in adult somatic tissues. *Epigenomics*, 8 (6):789–799, 2016.

H. Peterson. Exploiting high-throughput data for establishing relationships between genes, 2015.

W. Reik. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447(7143):425, 2007.

W. Reik, A. Collick, M. L. Norris, S. C. Barton, and M. A. Surani. Genomic imprinting determines methylation of parental alleles in transgenic mice. *Nature*, 328(6127):248, 1987.

J. Reimand, M. Kull, H. Peterson, J. Hansen, and J. Vilo. g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research*, 35(suppl 2):W193–W200, 2007.

J. Reimand, T. Arak, P. Adler, L. Kolberg, S. Reisberg, H. Peterson, and J. Vilo. g:Profiler–a web server for functional interpretation of gene lists (2016 update). *Nucleic acids research*, 44(W1):W83–W89, 2016.

M. Ringnér. What is principal component analysis? *Nature biotechnology*, 26(3): 303–304, 2008.

P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615, 2008.

M. Rodríguez-Paredes and M. Esteller. Cancer epigenetics reaches mainstream oncology. *Nature medicine*, pages 330–339, 2011.

A. D. Rosen, K. D. Robertson, R. A. Hlady, C. Muench, J. Lee, R. Philibert, S. Horvath, Z. A. Kaminsky, and F. W. Lohoff. DNA methylation age is accelerated in alcohol dependence. *Translational psychiatry*, 8(1):182, 2018.

A. Ruepp, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, M. Stransky, B. Waegele, T. Schmidt, O. N. Doudieu, V. Stümpflen, et al. Corum: the comprehensive resource of mammalian protein complexes. *Nucleic acids research*, 36(suppl_1):D646–D650, 2007.

M. Saare, V. Modhukur, M. Suhorutshenko, B. Rajashekar, K. Rekker, D. Sõritsa, H. Karro, P. Soplepmann, A. Sõritsa, C. M. Lindgren, et al. The influence

of menstrual cycle and endometriosis on endometrial methylome. *Clinical epigenetics*, 8(1):2, 2016.

C. Sapienza, A. C. Peterson, J. Rossant, and R. Balling. Degree of methylation of transgenes is dependent on gamete of origin. *Nature*, 328(6127):251, 1987.

R. J. Schmidt, D. I. Schroeder, F. K. Crary-Dooley, J. M. Barkoski, D. J. Tancredi, C. K. Walker, S. Ozonoff, I. Hertz-Picciotto, and J. M. LaSalle. Self-reported pregnancy exposures and placental DNA methylation in the MARBLES prospective autism sibling study. *Environmental epigenetics*, 2(4):dvw024, 2016.

T. Sijen. Molecular approaches for forensic cell type identification: on mRNA, miRNA, DNA methylation and microbial markers. *Forensic Science International: Genetics*, 18:21–32, 2015.

G. K. Smyth. Limma: linear models for microarray data, 2005.

K. Strimbu and J. A. Tavel. What are biomarkers? *Current Opinion in HIV and AIDS*, 5(6):463, 2010.

B. Sun, L. Hu, Z.-Y. Luo, X.-P. Chen, H.-H. Zhou, and W. Zhang. DNA methylation perspectives in the pathogenesis of autoimmune diseases. *Clinical Immunology*, 164:21–27, 2016.

B. Team. Biocarta: Charting pathways of life, 2001.

R. C. Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2017). R Foundation for Statistical Computing, 2017.

A. Teissandier and D. Bourc'his. Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription. *The EMBO journal*, 36(11): 1471–1473, 2017.

A. E. Teschendorff and S. C. Zheng. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. *Epigenomics*, 9(5):757–768, 2017.

A. E. Teschendorff, F. Marabita, M. Lechner, T. Bartlett, J. Tegner, D. Gomez-Cabrero, and S. Beck. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, 29(2):189–196, 2012.

T. Therneau. A package for survival analysis in S. R package version 2.37-4, 2014.

A. J. Titus, R. M. Gallimore, L. A. Salas, and B. C. Christensen. Cell-type deconvolution from DNA methylation: a review of recent applications. *Human molecular genetics*, 26(R2):R216–R224, 2017.

D. Wang, L. Yan, Q. Hu, L. E. Sucheston, M. J. Higgins, C. B. Ambrosone, C. S. Johnson, D. J. Smiraglia, and S. Liu. IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics*, 28(5): 729–730, 2012.

S.-C. Wang and A. Petronis. DNA methylation microarrays: experimental design and statistical analysis, 2008.

Y.-L. Weng, R. An, J. Shin, H. Song, and G.-l. Ming. DNA modifications and neurological disorders. *Neurotherapeutics*, 10(4):556–567, 2013.

Wikimedia. Menstrualcycle., 2004. URL `https://commons.wikimedia.org/wiki/File:MenstrualCycle2.png#/media/File:MenstrualCycle.png`.

Wikimedia. Internal organs., 2012. URL `https://commons.wikimedia.org/wiki/File:Internal_organs.svg#/media/File:Internal_organs.svg`.

Wikimedia. DNA methylation., 2016. URL `https://en.wikipedia.org/wiki/DNA_methylation#/media/File:DNA_methylation.svg`.

C. S. Wilhelm-Benartzi, D. C. Koestler, M. R. Karagas, J. M. Flanagan, B. C. Christensen, K. T. Kelsey, C. J. Marsit, E. A. Houseman, and R. Brown. Review of processing and analysis methods for DNA methylation array data, 2013. ISSN 00070920.

Y. Xiong, Y. Wei, Y. Gu, S. Zhang, J. Lyu, B. Zhang, C. Chen, J. Zhu, Y. Wang, H. Liu, et al. DiseaseMeth version 2.0: a major expansion and update of the human disease methylation database. *Nucleic acids research*, 45(D1):D888–D895, 2016.

X. Yang, H. Han, D. D. De Carvalho, F. D. Lay, P. A. Jones, and G. Liang. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer cell*, 26(4):577–590, 2014.

K. Y. Yeung and W. L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, 2001.

W.-S. Yong, F.-M. Hsu, and P.-Y. Chen. Profiling genome-wide DNA methylation. *Epigenetics & chromatin*, 9(1):26, 2016.

P. Yousefi, K. Huen, R. A. Schall, A. Decker, E. Elboudwarej, H. Quach, L. Barcellos, and N. Holland. Considerations for normalization of DNA methylation data by Illumina 450K BeadChip assay in population studies. *Epigenetics*, 8 (11):1141–1152, 2013.

S. K. Zaidi, D. W. Young, M. Montecino, J. B. Lian, J. L. Stein, A. J. Van Wijnen, and G. S. Stein. Architectural epigenetics: mitotic retention of mammalian transcriptional regulatory information. *Molecular and cellular biology*, 30(20): 4758–4766, 2010.

W. Zhou, T. J. Triche Jr, P. W. Laird, and H. Shen. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic acids research*, 46(20):e123–e123, 2018.

ACKNOWLEDGEMENTS

The work in this thesis was supported by the European Social Fund's Doctoral Studies and Internationalization Programme DoRa, carried out by Archimedes Foundation, Estonian Doctoral School of Information and Communication Technology (IKTDK), Centre of Excellence in Computer Science (EXCS), Complexity-Net programme (Complexity of Independent Epigenetic Signals in Cancer Initiation (CIESCI) project), ERDF through CoE EXCS, Estonian Centre of Excellence in ICT Research (EXCITE) and BioMedIT projects, IUT34-4 (Data Science Methods and Applications (DSMA)). Large-scale analysis results presented in this thesis are performed using the resources from the University of Tartu High-Performance Computing Center.

I thank my supervisor Jaak Vilo for providing me with an exciting bioinformatics research topic on epigenetics to pursue my Ph.D. thesis. I am also grateful for his support and ideas to carry out my research. Next, I am very appreciative to Triin Laisk-Podar, for providing guidance, which helped to improve my Ph.D. thesis. I also thank Balaji Rajashekar for his time to improve the quality of my Ph.D. thesis and his innovative ideas for bringing out the best possible outcome in the publications. I thank Meelis Kull, for being a kind mentor during the initial phase of my Ph.D. studies. My sincere gratitude to Lili Milani, for driving my motivation in epigenetics, when I begin to work on research projects in Tartu. I thank Raivo kolde for being a helpful colleague, especially during 17 tissues project and developing *seqlm*, along with Kaspar Märtens. It was a pleasure working with exciting collaborative projects, especially, 17 tissues project with Kaie Lokk and endometriosis project with Merli Saare. I appreciate Viktorija Kukushkina for being lovely collaborator while working for endometrial receptivity project, and it was excellent teamwork!. I am grateful to the PI's, Andres Salumets, Neeme Tönisson, and Maire Peters for allowing me to work with most interesting methylation studies. I am thankful to both present and past BIIT members for being fantastic colleagues. I want to thank Tatjana Iljasenko for the statistical contribution in *MethSurv*, Tauno Metsalu for the guidance towards *RShiny*, and Ivan Kuzmin for suggestions to solve technical problems. I thank my colleague Ahto Salumets for his fruitful discussions on methylation analysis, Estonian translations and his recommendations to improve my thesis. I am very thankful to my pre-reviewers Leopold Parts, Stephen Beck and Anagha Joshi, for their insightful suggestions and motivating feedback which helped to improve the quality of my thesis

Last but not least, I cannot thank enough my father, Suryanarayanan Modhukur, mother, Rukmini Modhukur and sister, Vijayalakshmi Modhukur for moral support and endless love!. I also thank my husband, Prakash Lingasamy and my child Vikas Prakash for giving me the emotional strength, patience, and constant support to pursue my doctoral studies.

SUMMARY IN ESTONIAN

Haiguste ja koespetsiifiliste DNA metülatsioonil põhinevate biomarkerite uurimine

DNA metülatsioon on epigeneetiline modifikatsioon, mis osaleb geenikspressiooni regulatsioonis. Seda modifikatsiooni seostatakse väga paljude oluliste bioloogiliste protsessidega nagu näiteks embrüonaalne areng, genoome vermimine ning X kromosoomi inaktivatsioon. Teatud metülatsioonimustrid on seotud haigustega, nagu diabeet, neuroloogilised häired või vähk. Seetõttu peetakse DNA metülatsioonimustreid ka headeks biomarkeri kandidaatideks, sobides iseloomustama näiteks teatud haiguse kujunemist (või varajast staadiumi). Suuremahulised tehnoloogiad, nagu mikrokiibid ja teise põlvkonna sekveneerimine (NGS), võimaldavad luua ülegenoomse DNA metülatsiooniprofiili, mis on heaks vahendiks mõistmaks geeniregulatsiooni. Üheks populaarseimaks DNA metülatsiooni mikrokiibiks on Illumina HumanMethylation450K (HM450K) kiip (Illumina Inc., San Diego, CA, USA), mis võimaldab kulutõhusalt hinnata metülatsiooni taset hästiiseloomustatud genoomi regioonides.

Käesolev doktoritöö käsitlebki DNA metülatsiooninimustreid kui potentsiaalseid biomarkeri kandidaate. Selleks on kasutatud HM450K metülatsiooniandmeid ning erinevaid arvutuslikke ja statistilisi meetodeid.

Väitekirja esimene osa keskendub DNA metülatsiooni analüüsile tervetes ja haigetes kudedes. Selleks kasutati DNA metülatsiooniandmeid erinevatelt koostööpartneritelt ning analüüsitulemused on avaldatud publikatsioonides I, II ja III. Esimeses artiklis (publikatsioon I) näitasime, et metülatsioonimustrite alusel saab kudesid klassifitseerida. Selles töös kasutasime näiteks andmete klasteranalüüsi, lineaarset regressiooni, korrelatsioonianalüüsi ning integreerisime need andmed ka geenikspressiooniandmetega. Teises artiklis (publikatsioon II) oli peamiseks analüütiliseks probleemiks tehnilise varieeruvuse kõrvaldamine, nimelt andmed olid pärit erinevatest laboritest. Selle eemaldasime kasutades meetodit nimega Empirical Bayes. Lisaks uurisime endometrioosiga seonduvaid metülatsioonipõhiseid biomarkereid kasutades esimeses publikatsioonist pärit lähenemist. Kolmanda artikli (publikatsioon III) jaoks kohandasime varasemalt kasutatud metoodikaid selleks, et identifitseerida endomeetriumi vastuvõtlikkusega seotud metülatsioonimustreid. Lisaks rakendasime mitmeid erinevaid differentsiaalanalüüsimeetodeid ning võtsime nende tulemuste ühisosa, et leida suurema usaldusväärsusega biomarkereid.

Tõõ teiseks eesmärgiks oli identifitseerida vähipatsientide elumusega seotud biomarkereid. Selleks otstarbeks kasutasime andmeid TCGA (The Cancer Genome Atlas) konsortsiumist. Antud tõõ tulemusena loodi intuitiivne veebirakendus (esitatud publikatsioonis IV), mis abistab teadlasi, kellel puudub vastav bioinformaatika- ja statistikaalane kompetents. See rakendus võimaldab uurida, kuidas on DNA metülatsioonimustrid seotud vähihaigete elumusega, visualiseerida metülatsiooni-

mustreid, teha klasteranalüüsi ning kuvada olulisemaid biomarkereid iga vähitüübi kohta.

Käesolev väitekiri on struktureeritud järgmiselt:
Esimeses osas tutvustatakse töö bioloogilist tausta (DNA metülatsiooni), võetakse kokku metülatsiooni uurimise metoodikad ning rõhutatakse DNA metülatsiooni mõju inimese tervisele. Lisaks antakse ülevaade kasutatud bioinformaatilistest meetoditest koos põgusa ülevaatega, milliseid probleeme nende meetoditega lahendati (Lokk et al., 2014; Saare et al., 2016; Kukushkina et al., 2017; Modhukur et al., 2017). Järgnevad neli peatükki võtavad kokku doktoritöös kasutatud publikatsioonid, nendele järgnevad arutelu ning kokkuvõte. Samuti on lisatud eespool mainitud neli publikatsiooni.

PUBLICATIONS

CURRICULUM VITAE

Personal data

Name:	Vijayachitra Modhukur
Date of Birth:	October 24th, 1984
Citizenship:	Indian
Languages:	English, Tamil,Telugu and Hindi
Contact:	vijayachitra.m@gmail.com

Education

2010–	University of Tartu, Estonia PhD candidate in Computer Science
2007–2010	Stockholm University, Sweden MSc in Bioinformatics
2003–2007	Sathyabama University, India B.Tech in Bioinformatics

Employment

2014–	University of Tartu, Estonia Junior Research Fellow in Bioinformatics
2012–2013	University of Tartu, Estonia Programmer
2009–2010	Genetwister Technologies B.V, Netherlands Bioinformatics researcher

Scientific work

Main fields of interest:

- Bioinformatics
- Analysis of High-Throughput Biological Data
- Data mining
- Machine learning

ELULOOKIRJELDUS

Isikuandmed

Nimi:	Vijayachitra Modhukur
Sünniaeg:	24. oktoober 1984
Kodakondsus:	India
Keeled:	Inglise, Tamili, Telugu ja Hindi keel
Kontakt:	vijayachitra.m@gmail.com

Haridus

2010–	Tartu Ülikool informaatika doktorant
2007–2010	Stockholmi Ülikool, Rootsi MSc Bioinformaatikas
2003–2007	Sathyabama ülikool, India B.Tech Bioinformaatikas

Teenistuskäik

2014–	Tartu Ülikool Bioinformatika nooremteadur
2012–2013	Tartu Ülikool Programmeerija
2009–2010	Genetwister Technologies B.V, Holland, Bioinformaatika teadur

Teadustegevus

Peamised uurimisvaldkonnad:

- Bioinformaatika
- Suure läbilaskevõimega tehnoloogiate poolt produtseeritud andmete analüüs
- Andmekave
- Masinõpe

DISSERTATIONES INFORMATICAE PREVIOUSLY PUBLISHED IN DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 lk.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.

74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.

113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.
121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

1. **Abdullah Makkeh**. Applications of Optimization in Some Complex Systems. Tartu 2018, 179 p.
2. **Riivo Kikas**. Analysis of Issue and Dependency Management in Open-Source Software Projects. Tartu 2018, 115 p.
3. **Ehsan Ebrahimi**. Post-Quantum Security in the Presence of Superposition Queries. Tartu 2018, 200 p.
4. **Ilya Verenich**. Explainable Predictive Monitoring of Temporal Measures of Business Processes. Tartu 2019, 151 p.
5. **Yauhen Yakimenka**. Failure Structures of Message-Passing Algorithms in Erasure Decoding and Compressed Sensing. Tartu 2019, 134 p.
6. **Irene Teinemaa**. Predictive and Prescriptive Monitoring of Business Process Outcomes. Tartu 2019, 196 p.
7. **Mohan Liyanage.** A Framework for Mobile Web of Things. Tartu 2019, 131 p.
8. **Toomas Krips.** Improving performance of secure real-number operations. Tartu 2019, 146 p.