Norwegian University
of Life Sciences

Master's Thesis 2016 60 ECTS
Faculty of Veterinary Medicine and Biosciences
Department of Chemistry, Biotechnology and Food Science

Establishment of Multiplexed Reduced Representation Bisulfite Sequencing Protocol on T-cells from Rheumatoid Arthritis Patients

Eirik Elias Hansen
Microbiology

Acknowledgments

This thesis marked the completion of a master's degree in microbiology at the Norwegian University of Life Sciences (NMBU), department of Chemistry, Biotechnology and Food Science over the period from August 2015 to May 2016. It was written in collaboration with the department for medical genetics at Oslo University Hospital (OUS), Ullevål.

I would like to thank my supervisor at OUS, Benedicte A. Lie for her incredible support and counseling through the work with this thesis. Despite her busy schedule, she always had time to answer questions and provide feedback. Her extensive knowledge, quick responses as well as constructive and helpful criticism has been a huge help through the whole process.

Siri T. Flåm also deserves a big thanks for all the support and guidance she has provided in the lab and on everything surrounding the method. Her insight and knowledge into all we did in and around the lab has rescued me on several occasions.

Additionally I thank Kari Guderud, Line H. Sunde and Fatima Heinicke of the EPIRA group for all of their support and help on various topics, such as theory, bioinformatics, work in the lab and more.

Simon Rayner I thank for all of his help with the bioinformatics. He always has an answer to my questions and helped me grasp the basics of how to handle bioinformatical data.

I would also like to thank everyone at EPIRA and IMMGEN for their various support, discussion, motivation and for really making me feel like a part of the group throughout the last year. A special thanks goes to Ingvild Gabrielsen and Hanna Helgeland for lending me their data on gene expressions in the CD4 blood cells.

In the end I would like to thank my supervisor at NMBU, Tor Lea for his swift replies whenever I needed his counseling.

Oslo, 12.05.2016

Eirik Elias Hansen

Sammendrag

DNA metylering er en viktig epigenetisk faktor med en regulatorisk effekt over ekspresjonen av gener. Det har også blitt påvist at metyleringsmønsteret i DNAet til et individ vil forandre seg over tid. I pattedyr forekommer DNA metylering kun på cytosiner, og nesten alltid i en CpG kontekst. Differensiell metylering har blitt vist å være viktig under utvikling og differensiering av pluripotente stamceller. I pasienter med revmatoid artritt (RA) har det blitt påvist at T-celler har redusert metylering i forhold til hos friske kontroller. RA er en alvorlig sykdom assosiert med sterke smerter og funksjonshemming. Anslagsvis 0,5-1% av befolkningen er rammet av denne sykdommen.

Arbeidet med å etablere en protokoll for multiplekset begrenset representativ bisulfittsekvensering (mRRBS), for å kunne produsere metylomprofiler for pasienter med RA er beskrevet i denne oppgaven. Det ekstraherte DNAet ble kløyvd med restriksjonsenzymet MspI. Dette restriksjonsenzymet anriket for genomiske områder som inneholdt CpG øyer. DNAet ble så behandlet med natrium bisulfitt som konverterte alle umetylerte cytosiner til uraciler, mens metylerte cytosiner sto uberørt. Gjennom PCR amplifikasjon etterfulgt av sekvensering, ble alle de umetylerte cytosinene lest som tyminer av sekvensatoren og metylerte cytosiner lest som normalt.

Fire forskjellige DNA ekstraksjonsmetoder ble testet i tillegg til to opprensingsprotokoller. Det var i denne studien et mål å få ekstrahert RNA og protein i tillegg til DNA fra det samme prøvematerialet, slik at videre studier kunne gjennomføres på et senere tidspunkt. Det beste resultatet ble oppnådd ved bruk av QIAamp DNA micro kit, men sammenlignbare resultater ble oppnådd fra DNA ekstraksjon med Norgen DNA/RNA/Protein purification plus kit etterfulgt av en opprensing med QIAamp DNA micro kit. Denne oppgaven har demonstrert at den etablerte mRRBS protokollen kunne generere metylomprofiler av T-celler fra RA pasienter med en oppnådd bisulfittkonverteringsrate på >99,9% og sekvenssammenstilling på minimum 60%. En undersøkende analyse av $CD4^+$ T celler viste at metyleringsnivået stort sett fulgte det forventede mønsteret med høy metylering i lavt utrykte gener, og motsatt, men støttet også teorien om at andre regulerende faktorer er involvert i tillegg til metylering.

Abstract

DNA methylation is an important epigenetic mark with a regulatory effect on the expression of genes. It has also been shown that individuals change their DNA methylation pattern over time. In mammals, methylation only occurs in cytosines, and almost exclusively in the CpG context. Differential methylation has been shown to be important for the development and differentiation of pluripotent stem cells. Rheumatoid arthritis (RA) patients have also been shown as having T-cells with reduced methylation compared to healthy controls. RA is a severe disease associated with great pain and disability and affects about 0.5-1% of the population.

In this thesis, the process of establishing a protocol for multiplexed reduced representation bisulfite sequencing (mRRBS) in order to create a methylome profile of RA patients is described. The extracted DNA was cleaved by a restriction enzyme, MspI, which enriched for areas of the genome containing CpG islands. This DNA was then treated with sodium bisulfite which converted all unmethylated cytosines to uracils, while leaving the methylated cytosines intact. Through PCR amplification, and subsequent sequencing, all unmethylated cytosines was read by the sequencer as thymines, while methylated cytosines was read as normal.

Four different DNA extraction methods were tested, as well as two DNA cleanup protocols. It was an aim in this study to extract RNA and protein, in addition to DNA, from the same sample, to enable further studies. The best results were achieved through the use of the QIAamp DNA micro kit, but comparable results were achieved by extracting the DNA with the Norgen DNA/RNA/Protein purification plus kit followed by a cleanup with the QIAamp DNA micro kit. In this thesis, it has been demonstrated that the established mRRBS protocol could generate methylome profiles of T cells from RA patients with a bisulfite conversion ratio >99.9%, and sequence alignment of at least 60%. An exploratory analysis of CD4+ T cells showed that the methylation level largely followed the expected pattern of high degree of methylation for genes with low expression, and vice versa, but also supported the notion that other regulatory factors were involved in addition to methylation.

Abbreviations

ACPA	Anticitrullinated peptide antibody
BAM	Binary Alignment/Map
Bp	Base pair
CD	Cluster of Differentiation
CLP	Common lymphoid precursor
CpG	Cytosine-phosphate-guanine
DNMT	DNA methyltransferase
(ds)DNA	(Double stranded) Deoxyribonucleic acid
DZ	Dizygotic
Gb	Gigabase pair
GRCh	Genome Reference Consortium Human reference
GWAS	Genome Wide Associaction Study
HCDM	Human cell differentiation molecules
HLA	Human leukocyte antigen
HLDA	Human leukocyte differentiation antigens
HS	High Sensitivity
HSC	Hematopoietic stem cell
HSP	Heat shock protein
Ig	Immunoglobulin
IUIS	International Union of Immunological Societies
Kb	Kilo base pair
(m)RRBS	(Multiplexed) Reduced Representation Bisulfite Sequencing
MHC	Major histocompability complex
MZ	Monozygotic
NEB	New England Biolabs
NGS	Next generation sequencing
NK cell	Natural killer cell
NSC	Norwegian Sequencing Center
PAD	Peptidyl arginine deiminase
PBMC	Peripheral blood mononuclear cells
PCR	Polymerase chain reaction
PEG	Polyethylen glycol

RA	Rheumatoid arthritis
RF	Rheumatoid factor
RNA	Ribonucleic acid
SAM	Sequence Alignment/Map
SE	Shared Epitope
SNP	Single nucleotide polymorphism
TCR	T cell receptor
WGBS	Whole Genome Bisulfite Sequencing
WHO	World Health Organization

Table of contents

Introduction

Rheumatoid arthritis

Rheumatoid arthritis (RA) is a systemic autoimmune disease. Autoimmune diseases are defined by the loss of tolerance towards the affected individuals own antigens. About 0.5-1% of the world's population is affected by RA (Willemze et al. 2012). Females are twice as likely to get the disease as men, and the prevalence is highest in Europe and North America (Messemaker et al. 2015). The common denominator for the disease is that it can be recognized by the chronic inflammation and infiltration of immune cells of the synovial joints, leading to joint destruction. This damage and disability of the joints may get progressively worse over the course of the disease. The disease is a chronic autoimmune disease, and is associated with a reduced life expectancy (reviewed in (Messemaker et al. 2015)). Importantly, RA patients are also known as being affected by several comorbidities with important examples including, but not limited to cardiovascular disease, depression or cancer. Depending on the comorbidity, quality of life and life expectancy can be affected (Michaud & Wolfe 2007).

In 1987, Arnett et al. (1988) made a list of seven criteria for being diagnosed with RA, 4 of these had to be met by the patient in order to be diagnosed. In their own words, the criteria were as follows:

> "1) morning stiffness in and around joints lasting at least 1 hour before maximal improvement; 2) soft tissue swelling (arthritis) of 3 or more joint areas observed by a physician; 3) swelling (arthritis) of the proximal interphalangeal, metacarpophalangeal, or wrist joints; 4) symmetric swelling (arthritis); 5) rheumatoid nodules; 6) the presence of rheumatoid factor; and 7) radiographic erosions and/or periarticular osteopenia in hand and/or wrist joints. Criteria 1 through 4 must have been present for at least 6 weeks." (Arnett et al. 1988)

In 2010, Aletaha et al. (2010) developed a new system and set of criteria in order to classify patients with RA. They have an initial condition for at least 1 joint with definitive clinical synovitis, which cannot be better explained by another disease. If that condition is met, a scoring system with 4 main categories ensues. These categories were: Joint involvement,

serology, acute phase reactants and duration of symptoms. Depending on the severity from each category, a score will be given, the maximum score is 10, and everyone achieving a score ≥6 will be said to have RA. A low score patient can however increase the score, passing the threshold at a later time. In order to detect the disease at an earlier stage, this system places a greater emphasis on the serology and acute phase reactants tests than in the list devised by Arnett et al. (1988).

Historically, rheumatoid factor (RF) was an important antibody for diagnosis (Sparks & Costenbader 2014). This autoantibody occurs in several different isoforms in different immunoglobulin (Ig) molecules, and it is targeting the Fc receptor of IgG (Mannik et al. 1988). However, RF may also occur as an aging effect or in people with other diseases (Sparks & Costenbader 2014). Anticitrullinated peptide antibody (ACPA) is another autoantibody, more specific for RA. Citrulline, which the ACPAs react against, occurs when the amino acid arginine is converted through post-translational modifications. This modification occurs during different biological processes, notably also during inflammation (reviewed by (Willemze et al. 2012)). The modification is performed by the peptidyl arginine deiminase (PAD) (Bicker & Thompson 2013). Of the patients who have detectable antibodies present in blood, 50-80% are positive for RF, ACPA or both (Scott et al. 2010). The presence of ACPA and RF is highly correlated. Due to ACPAs higher specificity for RA, detection results in less non-RA individuals wrongly diagnosed (reviewed by (Sparks & Costenbader 2014)). ACPAs are identifiable many years before patients develop the disease, however the amount and diversity of the ACPAs will increase shortly in advance of disease onset (Figure 1) (reviewed by (Koning et al. 2015)). Evidence suggests that ACPA-positive and –negative patients have genetically distinct diseases. This might indicate a different pathogenesis (reviewed by (Sparks & Costenbader 2014)). This is consistent with the notion of RA describing not a single disease, but rather a collection of several different conditions (Cope 2008). van der Woude et al. (2009) has, however, showed that there is no difference of the heritability in ACPA-positive and –negative patients.

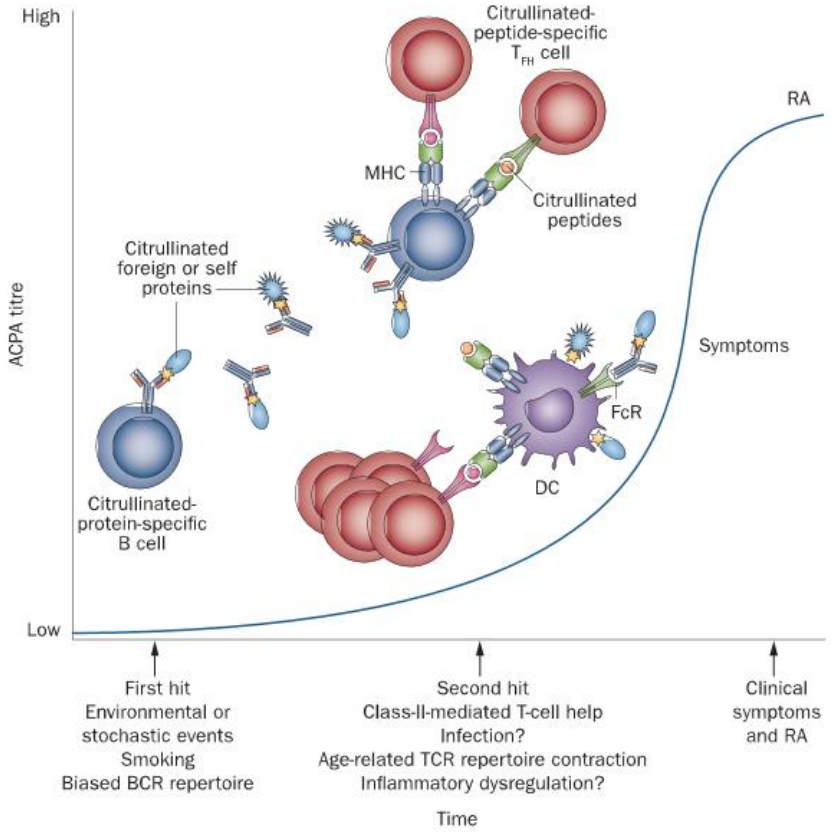


Figure 1 - The correlation between ACPA presence and development of RA (Koning et al. 2015).

$CD4^+$ and $CD8^+$ T-cells has been identified as important for RA through biopsies taken from the synovial tissue of patients, as healthy individuals will not have any lymphoid cells present in the same tissues (Duke et al. 1982). Other lymphocytes have also been identified in this target tissue of RA patients, but will not be the focus of this thesis. Both the naïve and memory T-cells found in RA patients shows telomere erosion, usually linked with aging, regardless of the age of onset. This suggests that antigen response might not be the sole reason for RA, but maybe replicative stress is also a factor (Cope 2008). The central role of T-cells, especially $CD4^+$ T-cells, is illustrated in Figure 2.

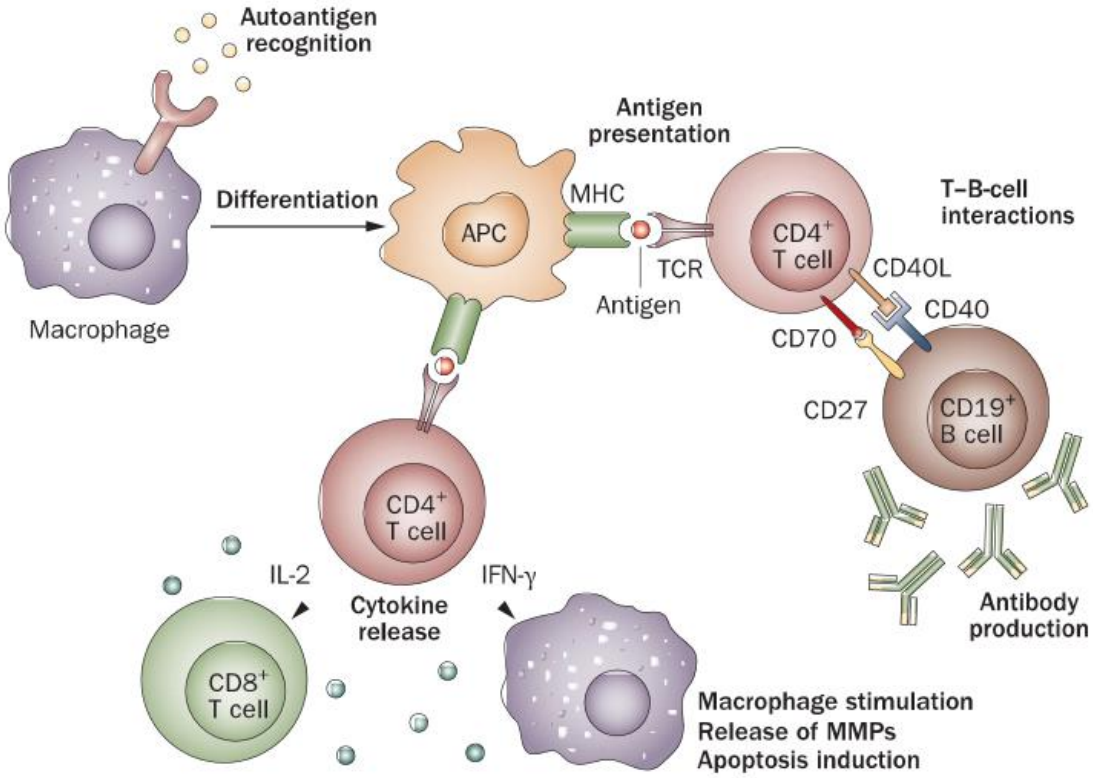


Figure 2 - Illustration of the interactions between different immune cells in RA patients (Ballestar 2011)

The general belief is that a combination of genetics and the environment causes RA (Messemaker et al. 2015). However, it has proven difficult to identify environmental factors for the disease due to several conditions such as a low disease prevalence, difficulty in predicting disease before onset and more (Sparks & Costenbader 2014). On the genetic side, the studies performed by Silman et al. (1993) and Aho et al. (1986) found a concordance rate of approximately 15% for RA between monozygotic (MZ) twins and about 3.5% for dizygotic (DZ) twins. These studies were performed on UK and Finnish populations, respectively. However, according to MacGregor et al. (2000) the common interpretation of this as meaning that there is a low genetic contribution to the disease and that the environmental factor is correspondingly larger is not correct. This, they continue, is due to the fact that the concordance rate between the twins will be dependent on the overall prevalence of the disease in a population. In other words, the concordance rate will increase and decrease together with the prevalence of the disease in the population. MacGregor et al. (2000) further propose through their own calculations on the data from the two

aforementioned studies that the genetic contribution is about 60%. This places genetics as the major contributor to the disease.

The best studied environmental factor, which is known for certain to increase the risk of the disease, is smoking. The association is particularly strong for patients with ACPA-positive RA (Figure 1), while it is weaker for the ACPA-negative patients. Other suggested, but less studied, risk-factors include reproductive factors in women, excess body mass and exposure to silica (reviewed by (Sparks & Costenbader 2014)).

Genetics of RA

RA is a multifactorial disease, and there are several genetic variants with an association to RA. The vast majority of them were discovered after 2007 due to methods such as genome-wide association studies (GWAS), which were made available due to microarray technology, large International cohorts and single nucleotide polymorphism (SNP) information from projects such as HapMap and the human genome project. Another important factor was better defined patient groups (Messemaker et al. 2015). These advances led to the discovery of more than 100 RA associated loci as described by Okada et al. (2014), in a GWAS comprising DNA from almost 100 000 individuals. Of these 101 loci, 42 were discovered for the first time in this large GWAS. However, the gene variants known from before 2007 still show, by far the strongest association with the disease (reviewed by (Messemaker et al. 2015)). The PTPN22 gene is an example which gives an odds ratio of 1.78 for the RA, but even this is overshadowed by the far most important genetic factor of the Human Leukocyte Antigen (HLA), with alleles increasing the risk of disease about 4 times (reviewed by (Sparks & Costenbader 2014)). When looking at the list provided by Okada et al. (2014), it is apparent that the odds ratios of these newly discovered SNPs are lower, with the most significant contribution at 1.47, but rarely exceeding 1.1. In addition, only about 5% of the total heritability can be explained by the genetic components located outside of the major histocompatibility complex (MHC) (Okada et al. 2014). Comparison of several GWAS studies performed on different autoimmune diseases reveals that many of the identified genetic factors are shared between the diseases. For RA, genes identified through the use of GWAS include, but are not limited to: TRAF1, TNFAIP3, CD40 and CCR6. CD40 is expressed in monocytes, B- and other immune-cells, while TRAF1 and TNFAIP3 has been

shown to be involved in the expression of CD40. CCR6 is expressed in Th17 cells which are abundant in the synovial tissue of RA patients (Suzuki et al. 2011).

Although the discoveries from the more recent genetic research might not have revealed any genetic variants with large contributions to the condition, the research has been helpful in identifying important RA pathways (Messemaker et al. 2015). Furthermore, drugs are being developed to interfere with one of the most important pathways, the JAK-STAT signaling pathway. When a gene or pathway is discovered, it is also possible to determine whether it is up- or downregulated in order to fully comprehend how the different mechanisms functions (Messemaker et al. 2015). The GWAS in RA highlights immune genes as being important and also points out pathogenic cell types based on overlap between risk loci and epigenetics marks (Okada et al. 2014). Based on these data, T cells appear to be the most important cell type.

The HLA genes are encoded in the MHC on chromosome 6 and is, as mentioned, the most important genetic factor associated with RA (Sparks & Costenbader 2014). Because of a high degree of linkage disequilibrium, it has been very hard to identify exactly which, of among about 250 genes, within the region is responsible for the disease susceptibility (Messemaker et al. 2015). GWAS and deep sequencing have been helpful tools however, and some specific causal positions have been identified for autoantibody positive disease. The HLA-DRB1 gene is especially important, and the specific positions of its amino acid sequence, position 11, 71 and 74 as discovered by Raychaudhuri et al. (2012), which is partly overlapping with what was discovered in the 1980s as the so-called the HLA shared epitope (SE), spanning amino acids 70-74 (Gregersen et al. 1987). These positions are within the antigen-binding groove of the HLA molecule (Figure 3), which reinforces the theory of the involvement of T-cells in RA (Raychaudhuri et al. 2012). Outside of the HLA SE, they also found that position 9 of the two genes HLA-DPβ1 and HLA-B had an association to the disease. The HLA SE has also been shown as contributing in the lack of tolerance towards citrullinated proteins in RA patients (Huizinga et al. 2005).

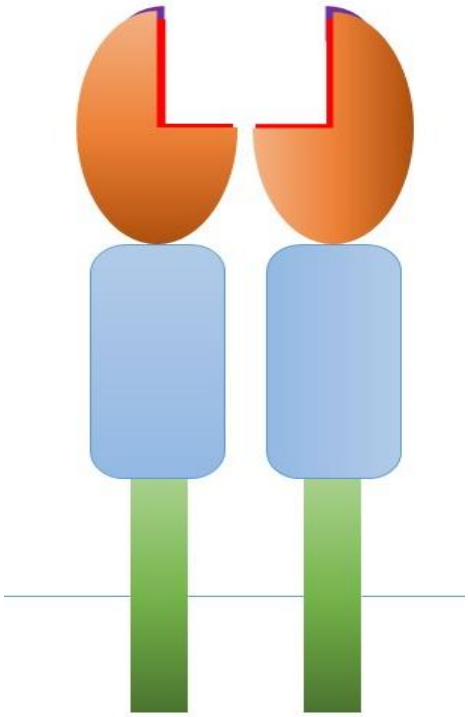


Figure 3 – A simple illustration of a membrane bound HLA class II molecule. The area marked in red is the variable region where the peptide antigen is being presented, while the area marked in purple is the less variable region where the T-cell receptor will recognize the HLA molecule.

Immunology and T-cells

The generally accepted theory for development of immune cells is that they all start out as pluripotent hematopoietic stem cells (HSC). The HSCs then mature into common lymphoid precursors (CLP), which in turn matures to T, B and natural killer (NK) cells (Blom & Spits 2006). There are two primary lymphoid organs. The bone marrow where all the lymphoid stem cells are originating, and the thymus, where the maturation of T-cells take place (Lea 2006). There, T-cells with receptors to all possible peptides, including the ones corresponding to those created by the body itself, are created through rearrangements of the α and β part of the TCR (Starr et al. 2003). 95% of these cells are then euthanized in order to avoid autoimmunity. The result is T-cells with a significant amount of receptors for foreign antigens (Lea 2006).

The B- and T-lymphoid cells are part of the adaptive immune system, and are the important cells related to RA. This is because autoreactive cells escaping the negative selection process in thymus could potentially lead to autoimmunity. It is however also worth mentioning that the adaptive immune system is complemented by the innate immune system which is non-

specific, but it will not be described in any further detail as it is outside the scope of this thesis.

The lymphocytes are continually regenerated, even in a healthy body (Lea 2006). The cells will normally circulate in a naïve state, searching for matching antigens. The fact that the TCR can only recognize antigens as shorter peptides presented in the context of an HLA molecule is known as HLA restriction. This means that the T-cells are unable to recognize antigens in free form which has yet to be processed by the APCs (Van Laethem et al. 2012). The cytotoxic $CD8^+$ T-cells recognize intracellular derived peptides in the context of HLA class I, while $CD4^+$ recognize extracellular peptides presented by HLA class II molecules (Van Laethem et al. 2012). This stands in contrast to the B-cells, which can bind its antibodies to free foreign antigens in the environment (Lea 2006).

The studies described by Brown et al. (1993) found that the class II HLA DR1 consisted of a heterodimer with ends that creates antigen binding grooves, and edges connecting with the TCR (Figure 3). The TCR structure is similar to that of the HLA (Janeway et al. 2001). The recombination of the gene sequence coding for the antigen receptors has some degree of randomness linked to it. This increases the diversity of the specificity of these receptors and makes the T-cells able to recognize ever-evolving threats to the body (Lea 2006).

In addition to the reaction between the HLA, antigenic peptide and TCR, T-cells need signals from co-receptors in order to activate (Smith-Garvin et al. 2009). It is interesting to note that naïve T-cells apparently are stricter when it comes to the signals from co-receptors than the memory cells (Berard & Tough 2002). Several surface molecules have been shown as being able to function as co-receptors, Smith-Garvin et al. (2009) lists a few, including several CD molecules. However, they also mention that CD28 seems to be the most important of all the costimulants.

When the T-cells are reacting with a complementing antigen, together with costimulatory signals from the APC, the activation will begin. During activation, the cell will start multiplying and evolve its effector functions, and in the end it will become a fully developed effector cell (Berard & Tough 2002). After dealing with the threat, apoptosis will be induced for most cells. The rest will become memory cells with a specificity for the particular infectious agent (Berard & Tough 2002).

Different immune cells have different molecules on their membrane. These surface molecules can be used to isolate particular immune cells. The membrane molecules are called "Cluster of Differentiation" (CD) followed by a specific number. The CD nomenclature is approved by the international union of immunological societies (IUIS) and the world health organization (WHO), and was created by the human leukocyte differentiation antigens (HLDA) workshop, now human cell differentiation molecules (HCDM) workshop. It is HCDM that decides and confirms CD assignments for molecules and antibodies (Zola et al. 2007). These molecules exist in several different versions which are unique for different kind of cells and stages of development. By looking at the presence of specific CD molecules, or the combination of CD molecules, the stage and identity of a cell can be determined. For example, all CLP cells in the bone marrow has been shown as being CD34 positive (reviewed by (Payne & Crooks 2002)), and all T-cells are CD3 positive (Dong & Marinez 2010).

The T-cells can be divided into several sub-categories depending on their function, and presence of CD molecules. So called T-cytotoxic cells can be discovered through the presence of CD8 and T-helper cells are identified through the presence of CD4 (Dong & Marinez 2010). These two markers are usually not present at the same time and serves as a good differentiator (Lea 2006). The regulatory T-cells are a third group, which can be identified through the presence of both CD4 and CD25 markers in addition to transcription factor Foxp3 and membrane bound molecule CTLA-4 (Dong & Marinez 2010).

Tolerance and autoimmunity

Immunological tolerance is the immune system's ability to not react towards self-antigens, as earlier mentioned a breakage of this system leads to autoimmune disease. It seems that immature lymphoid cells possess a greater ability to develop tolerance than the immune competent mature cells (Lea 2006). This results in two main types of tolerance, the central tolerance in the primary lymphoid organs, and the peripheral tolerance in the rest of the body. The central tolerance is developed in the thymus and is induced by both positive and negative selection. Positive selection involves precursor T-cells with an MHC class I or II restriction producing a secondary signal preventing apoptosis and inducing maturation into T-cells (Starr et al. 2003). Negative selection happens through clonal deletion where cell death is triggered should the antigenic peptide presented to the cell be present in the thymus at the

time of development (Starr et al. 2003). After this selection is complete, less than 5% of the original cells are left, and will both be in position of functional TCRs and able to distinguish self-antigens from other antigens, and thus avoiding reactivity towards self-antigens (Starr et al. 2003).

The peripheral tolerance is of importance because it is necessary for the body to let some autoreactive T-cells through the central tolerance in order to not risk the deletion of T-cells with receptors towards important pathogenic signals (Walker & Abbas 2002). The peripheral tolerance can either work directly on the T-cell or through dendritic cells or regulatory T-cells. Some examples of mechanisms are lack of an adequate amount of antigen in order to trigger a sufficient response, or the antigen is inaccessible for the TCR, this is known as ignorance. Anergy is another mechanism where either lack of secondary signal, or signaling through alternative receptors leads to a functional inactivation of the cell. Activation induced cell death can also happen where apoptosis is triggered instead of inactivation (Walker & Abbas 2002).

There are a few instances where autoreactivity might occur. One is if some sort of tissue damage occurs, which in turn leads to antigens from cells usually not available for the T-lymphocytes, seeping into the circulatory system where the immune cells are present (Lea 2006). Another potential source of autoreactivity is through molecular mimicry. This occurs as a result of infecting agents producing molecular structures or sequences similar to those of the body's own antigens (Cusick et al. 2012). A specific example is the heat shock protein (HSP) Hsp65. The HSPs are produced by mammals as well as bacteria when exposed to stress factors such as inflammation. The specific protein mentioned above can be found in the synovial tissue of RA patients, and it has a high sequence similarity to HSPs produced by bacteria (de Graeff-Meeder et al. 1990). A last instance for autoreactivity is through post-transcriptional modifications of proteins such as the earlier mentioned citrulination of arginine. Because of the change made after the transcription, they will differ from the ones produced in the thymus, and as such can lead to the immune cells reacting with them (Lea 2006).

DNA-methylation

In Figure 4, the chemical structure of the methylated cytosine, known as 5-methylcytosine, is given. It differs from the regular cytosine in that it has an additional methyl group in the fifth position.

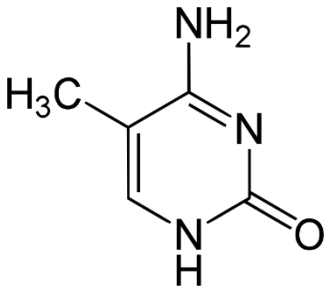


Figure 4 - The chemical structure of 5-methylcytosine

The first ever reported natural occurrence of the 5-methylcytosine was in the tubercle bacillus as discovered by Johnson and Coghill (1925). Later Hotchkiss (1948) also discovered what he hypothesized to be the same base in calf thymus samples. This has later been confirmed and he was thus the first to discover this modified base in a higher eukaryote (Moore et al. 2013). The cytosine is also the only base that has been found as methylated in multicellular animals as of yet (Jeltsch 2002). With a few exceptions, such as in pluripotent stem cells, DNA methylation only exists in the context of CpGs in vertebrates (Ziller et al. 2011).

CpGs occur at a ratio below what would be expected in the human genome (Lander et al. 2001). The reason for this is that the cytosines in this context are mostly methylated, and through spontaneous deamination, they will be converted to thymines (Gardiner-Garden & Frommer 1987). This is in contrast to the unmethylated cytosines, which as will be described later, deaminates to uracils. The uracils can in turn be repaired by the base excision repair machinery (Weber et al. 2007). Nothing of this applies to CpG islands however, where methylation is less common. Lander et al. (2001) identified 28890 CpG islands where the majority had a length of approximately 1kb and GC content at about 60-70%. The CpG islands also overlap with about 60-70% of the promoter regions of genes in the human genome (Illingworth & Bird 2009).

Methylation of the genome is a very important process for regulating the expression of genes. The general rule is that a promoter region that is methylated will have reduced expression. There are two ways in which this reduction could be happening. Either the methyl group could be physically blocking the transcriptional regulators, or the group has been proposed to

interact with methylation-specific binding proteins, creating protein complexes which blocks for transcription (Cribbs et al. 2015).

When it comes to the maturation of HSCs, methylation patterns also turn out to be a very important factor for deciding which type of cell it will differentiate into. Depending on this pattern, the HSC might turn into a myeloid or lymphoid cell (Suarez-Alvarez et al. 2012). As a consequence of DNA methylation patterns guiding the differentiation process, it follows that different cell types have differential methylation patterns (Lister et al. 2009). Lister et al. (2009) demonstrated this by comparison of two different human cell types. Other forms of epigenetic mechanisms, such as histone modifications are also important factors for gene regulation, but that will not be the focus of this thesis. The focus here will be on the DNA methylation due to the fact that deviations in these patterns has been associated with autoimmune disorders and immune deficiencies (Suarez-Alvarez et al. 2012).

The T-cells, to varying extents, are also prone to differentiation due to modifications of the methylation pattern. Especially the naïve $CD4^+$ T-cells are able to turn into other specialized T-cells through one or more steps. This happens as a result of differential gene expression (reviewed in (Suarez-Alvarez et al. 2012)). The methylation pattern might for example help dictate whether a T cell turns into a T-helper 1 or 2 cell. Many of the specialized T-cells are also able through differential gene expression, to move in one or more directions and mature into other kinds of T-cells. An example is a T-helper 17 cell which can become an inducible regulatory T-cell and vice versa (reviewed in (Suarez-Alvarez et al. 2012)).

The methylation pattern is modified by DNA methyltransferases (DNMTs) and DNA demethylation for adding and removing methyl groups respectively. The DNA demethylation is divided into two distinct categories, the active and passive demethylation. The latter is dependent of cell division and cannot continue after the mitosis is complete. This is to say that the methylation in the DNA is not conserved after the division of the cell has completed. This kind of demethylation can remove a maximum of 50% of the methyl groups for each round of cell division. This means that any excess amount of reduction in addition to this has to be explained, at least partly, by active demethylation. Proof of passive methylation is lacking for vertebrates, but there are some cases of documented occurrence, although there is not yet total agreement on the mechanism (reiewed by (Suarez-Alvarez et al. 2012)). An example is the study of the methylation levels of mice embryos during preimplantation

development, where methylation levels were dropping during the first four days of development. This methylation loss correlated with loss of DNMTs in the cells, and by inhibiting DNA replication it was shown that methylation levels were largely unchanged, pointing towards passive demethylation from replication (Howlett & Reik 1991). Active demethylation on the other hand takes place when the cell is not dividing. However, the mechanisms for this kind of demethylation is not sufficiently understood for mammalian cells as of yet, but it is suspected to be related to the DNA repair machinery (Suarez-Alvarez et al. 2012).

The findings of Fraga et al. (2005) showing that the DNA methylation changes over time is very interesting. They discovered through analysis of both DNA methylation and histone acetylation that during the first years, monozygotic twins are epigenetically indistinguishable, but as they get older they accumulate individual epigenetic patterns. This differential pattern also increased with age, lifestyle differences, and time spent apart. These results strongly indicate the important role of epigenetics, not only for gene expression, but also disease susceptibility, even in individuals with identical genomes.

Already in 1990 T-cells from RA patients were shown as having reduced methylation as compared to healthy controls (Richardson et al. 1990). For specific genes in RA, one study found a single CpG position upstream of the gene IL6 (-1990C) that had differential methylation between patients and controls (58% versus 98%, $P = 1 \cdot 10^{-6}$) (Nile et al. 2008). This study did, however, use peripheral blood mononuclear cells (PBMC), and not specifically isolated cell types. As methylation patterns are specific for each cell type, this could lead to problems when interpreting the data. Especially because RA patients have been shown to have an altered proportion of T-cells compared to healthy individuals (Cribbs et al. 2015). Another study using whole blood tried to address the problem of differential methylation between cell types by identifying the methylation patterns of the different cell types in the sample through a statistical algorithm. They claimed the identified differentiation of cells from this algorithm was comparable to results from flow cytometry. They were able to identify two differentially methylated clusters, containing CpGs, affecting the risk of developing RA, in the MHC region (Liu et al. 2013). These results indicate that DNA methylation is indeed an important factor, also in already identified risk disease associated genes (Cribbs et al. 2015). There has also been a study showing that the CD40L gene on the X-chromosome in $CD4^+$ T-cells had reduced methylation levels as compared to healthy

controls. Interestingly however, this only applied to female, and not male patients, and they found similar results for gene expression levels. This could partly explain why there are more females than males contracting the disease (Liao et al. 2012). These are just a few examples of studies performed on RA and methylation, however there is still a lack in knowledge regarding DNA methylations role in T-cells for RA patients (Cribbs et al. 2015).

Method theory

Reduced Representation Bisulfite-Sequencing (RRBS)

Unmethylated cytosines will upon reaction with bisulfite under specific conditions be converted to uracils (Hayatsu et al. 1970). Sodium bisulfite is the first described chemical that can convert a specific, common nucleic acid into another. The conversion happens through deamination of the cytosine (Figure 5), in contrast to adenine and guanine, which will not react at all with the chemical (Shapiro et al. 1970). The two studies cited above, both came to this conclusion independently, but at the same time. Under the same conditions, methylated cytosines will also be converted, but instead of uracils they will become thymines (Hayatsu et al. 1970). This however will happen at a much slower pace. As showed by Wang et al. (1980) the conditions which will give a conversion rate of >96% for the unmethylated cytosines will for the methylated versions only lead to a conversion rate at about 2-3%. When amplifying the converted fragments using PCR, the product will contain thymines in all positions which originally had an unmethylated cytosine, thus all cytosines that are left will be the ones which are methylated (Frommer et al. 1992).

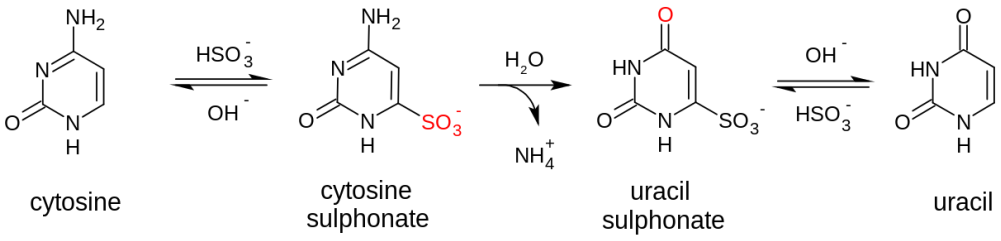


Figure 5 - The bisulfite mediated conversion of cytosine to uracil.

Reduced representation bisulfite-sequencing (RRBS) creates as the name implies a “reduced representation” of the genome by digesting it with the use of restriction enzymes before size selecting the fragments (Meissner et al. 2005). Several different restriction enzymes can be used to achieve this goal for example BgIII as used by Meissner et al. (2005) or more recently MspI used by Gu et al. (2011) and Boyle et al. (2012). MspI is a good choice as it cuts

irrespectively of methylation status in the restriction site of C^CGG (Waalwijk & Flavell 1978). It then follows, as stated by Gu et al. (2011), that each end of each sequenced fragment will contain at least one nucleotide with information on the methylation status. They also inform that an *in silico* digestion of vertebrate genomes shows that fragments with sizes from 40-220 bp should be have a representative coverage, and be enriched with most CpG-island regions and promoter sequences. Selecting for the same fragment sizes, *in* silico digestion of mouse genome has shown that by sequencing 36bp ends 90% of CpG islands will be covered, and 4.8% of the total CpG positions (Meissner et al. 2008). A coverage of 10-20% CpGs has also been claimed from the use of mRRBS (reviewed by (Plongthongkum et al. 2014)).

Before the actual bisulfite conversion, the sticky ends created by the MspI restriction enzyme have to be repaired, and at the same time, A-tailing and adapter ligation is performed (Gu et al. 2011). In order to achieve a conversion rate of the unmethylated cytosines to uracils of >99% during the bisulfite conversion, Gu et al. (2010) found that two successive rounds of bisulfite treatments at 5 hours each was optimal. During subsequent sequencing, an important error source which should be considered is the fact that longer reads could sequence into the adapters (Gu et al. 2011).

Multiplexed RRBS (mRRBS)

The multiplexed RRBS (mRRBS) differs from the original protocol in that the original RRBS protocol was designed for isolating fragments by separation on a gel. The mRRBS on the other hand skips this entire step, and in addition manages to both simplify and remove a lot of steps from the original protocol. It follows from these modifications that the whole protocol becomes much faster to perform, and also are able to produce more samples at a time than the old one. Boyle et al. (2012), the authors of the protocol, claims that they have managed to reduce the time needed for preparation from 9 to 6 days. They also claim that the mRRBS protocol should be cheaper than the original RRBS, while still producing the same amount of coverage of CpG positions.

Illumina sequencing

The Illumina sequencer is a next-generation sequencing (NGS) platform. This involves it being able to sequence at a higher speed, resolution and increased throughput in comparison

to the regular capillary electrophoresis based Sanger sequencing (Metzker 2010). Illumina sequencing can be performed in parallel for millions of fragments (Illumina inc 2016a).

As described in "An introduction to next-generation sequencing technology" an Illumina flow cell consists of lots of oligo sequences spotted to the surface. Sample preparation begins with fragmenting the DNA and ligating adapters on each end of the fragment. After amplification and purification, the library is loaded to a flow cell spotted with oligos containing regions complementary to the adapters. The oligos and adapters hybridize, and the sequence ligated to the adapter can function as a template for expansion of the spotted oligo (Illumina inc 2016a). There is two oligo sequences spotted to the flow cell, one for each end of the fragment (Metzker 2010). When the oligo extension is finished, the template is removed. Through a process called bridge amplification, the free end of the oligo will now hybridize to the neighbouring oligo sequence and duplicate once again (Figure 6). When denatured, this leaves a forward and reverse duplicate of the original fragment clustered together. This is repeated throughout the flow cell, and results in amplification of the sequence (Illumina inc 2016a). This amplification is necessary in order to get a sufficiently strong signal for reading (Metzker 2010).

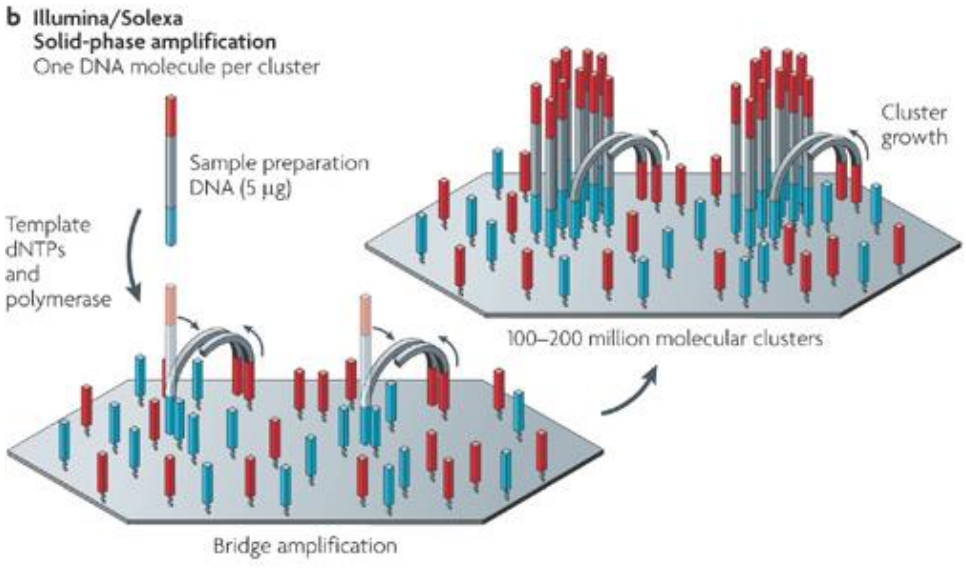


Figure 6 - Clustering through bridge amplification (Metzker 2010).

In the end, a sequencing primer hybridizes with the DNA strand, and initializes synthesis in a process known as cyclic reversible termination, (or sequencing by synthesis as Illumina themselves calls it) (Figure 7) (Metzker 2010). Each nucleotide is labeled with a unique

fluorescent signal that is excited when incorporated to the strand, this signal can be detected, and the sequence decided. All four nucleotides are added together and will compete naturally for incorporation with the template DNA strand (Illumina inc 2016a). An important aspect for this type of sequencing is that each nucleotide also contains a terminator stopping the DNA polymerase after the addition of only one nucleotide. Before detecting the signal leftover nucleotides are washed away. The fluorescent dye is cleaved of and washed away together with the termination component before the process is repeated (Metzker 2010).

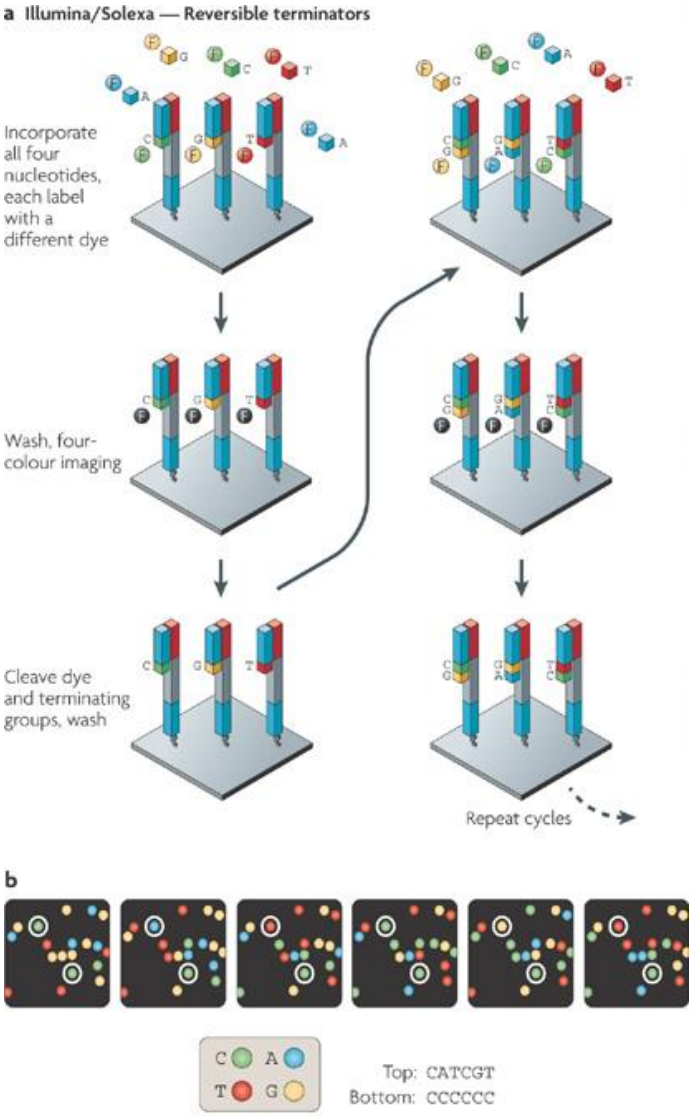


Figure 7 - Illustration of cyclic reversible termination sequencing (sequencing by synthesis). a. the incorporation of fluorescently labeled nucleotides, imaging and removal of the dye before repeat of the process. b. the four color imaging of each amplified template (Metzker 2010).

Through the addition of index sequences during library preparation, several samples can be sequenced and separated on a single flow cell during a single run. This is known as multiplexing (Illumina inc 2016a).

Bioinformatics

According to (Illumina Inc 2015), the MiSeq system produces up to 8.5Gb from up to 50M reads, depending on the input material and configuration. As such, bioinformatic software tools are necessary to make sense of the raw sequencing data, e.g. through mapping of reads to a reference genome, extracting CpG sites, calculating methylation values and more. Below follows a description of the software used during this thesis.

Mapping

The first step in the interpretation of raw sequencing data is to map the reads to a reference genome. RRBSMAP is a specialized version of BSMAP, a tool for mapping methylation sequences, for use with samples sequenced through the use of RRBS (Xi et al. 2012). In order to optimize performance of the mapping process, RRBSMAP does not align with the whole genome. Instead, it rather guides the alignment based on the restriction pattern of the selected restriction enzyme, usually MspI. This approach also improves the runtime greatly. According to the authors, RRBSMAP is much more straightforward to use for mapping of RRBS data than custom made pipelines. RRBSMAP is more user-friendly and delivers the same or better quality alignments, at a faster rate than an internal custom pipeline the authors had previously used for several publications (Xi et al. 2012). RRBSMAP works for aligning both single- and paired-end reads with varying read lengths (Xi et al. 2012). The final alignment output is delivered as a Sequence Alignment/Map (SAM) file, and contains data on read alignments against reference sequences. This file format is also the one used by the 1000 genomes project (Li et al. 2009).

The human reference genome, which was used as mapping reference in this thesis, is continually updated and revised. In this thesis, the Genome Reference Consortium Human Reference 37 (GRCh37) assembled by the genome reference consortium in 2009 was used for the mapping, even though it was not the most recent revision (Myers et al. 2015). The reason for this was that a parallel RNA study was also to be performed, and an already established internal pipeline was using this edition. The GRCh37 reference genome consists

of 27478 contigs with a total length at 3.2Gb. The chromosome length totals at 3.1Gb. Ensembl is a service providing tools and datasets for reference genomes, such as genome browsers and genome annotations and much more (Flicek et al. 2014). They report that there is about 20000 coding genes, and 14000 pseudogenes in the assembly (Flicek et al. 2016). The reference genome was created by hierarchical based assembly (Myers et al. 2015).

Calculation of methylation ratios

SAMtools is a software package containing tools for post-processing of SAM files. The package has a big range of uses such as converting from other alignment formats to the SAM format, merging of alignments, call SNPs and more (Li et al. 2009). It is also able to convert the SAM file into a binary alignment/map (BAM) file, as well as sorting the data in the file.

A python script delivered as part of the BSMAP package, methratio.py, uses the sorted BAM files as input, and is able to determine the methylation ratio of cytosine positions in the aligned sequences. These ratios are given together with positions in the reference genome, strand information, and the context of the cytosine (e.g. CpG, CHG etc.). Based on the information in this files, it is possible to extract for example only the CpG positions, or look into specific regions of interest. The methylation ratio is calculated by dividing the number of cytosines with the sum of cytosines and thymines in the same position.

Aims of this thesis

The main aim of the thesis was to establish a protocol for mRRBS of T cells from RA patients. This was done through the following objectives:

1. Identification of a DNA extraction method providing clean, uncontaminated DNA with a sufficient yield for performing mRRBS. Preferably a method also capable of extracting RNA from the same samples, for use in later studies.
2. Establishment of parameters such as incubation times and pcr settings in the mRRBS protocol.
3. Sequencing of prepared samples and quality control of these results.
4. Initial analyses of sequencing results, laying the foundation for further studies.

Materials and methods

A complete list of materials and reagents used in the work of this thesis is given in appendix 1, and a list of equipment can be found in appendix 2.

Patient samples

Table 1 - Information about the samples used in the experiments of this thesis.

Sample	Time point	Age	Sex	Smoking status	ACPA status	Cohort
RA5111	Baseline, isolated 2015	50	Female	Quit	Unknown	NOR-VEAC
RA5509	Baseline, isolated 2014	46	Female	Never	Positive	Arctic Rewind
RA5511	Baseline, isolated 2014	50	Female	Never	Positive	Arctic Rewind
RA5512	Baseline, isolated 2014	65	Female	Never	Positive	Arctic Rewind
RA5516	Baseline, isolated 2014	35	Female	Yes	Unknown	Arctic Rewind
Control Sample I*	Isolated 2013	40	Female	No	-	-
Control sample II	Isolated 2015	40	Female	No	-	-

*The table contains information about age, sex, timepoint for sampling, Smoking- and ACPA-status. *Sampled directly to EDTA tub,e as opposed to sampling in a blood bag manually prepared with EDTA.*

Patient samples were collected from two different cohorts: NOR-VEAC and Arctic Rewind, information about age, sex, smoking- and ACPA-status as well as time point of sampling was recorded for each sample (Table 1).

NOR-VEAC is a prospective cohort study consisting of patients who are newly diagnosed with RA. The first sample is taken immediately after the diagnosis is set (baseline), before methotrexate treatment begins. The next sample is taken after 3 months of methotrexate

treatment where the patients are also evaluated clinically in regards to their response to the treatment.

Arctic Rewind on the other hand is a randomized controlled trial, where the goal is to find out whether or not RA patients in remission can reduce the methotrexate dosage given. They are sampled after 12 months of remission. After this they are randomized with half of the patients further receiving the same dose as before, while the other half reduce the dose. The patients are sampled again if a flare occurs, or after 8 months if none occur. If a new flare occurred, an increase in medication is given and a sample is once again taken at renewed remission. Control sample I and II are from the same individual sampled at two different time points (Table 1).

The project is approved by the regional ethics committee (2015/1546-4), and written informed consent has been given by the participants.

Experimental overview

The DNA methylation experiment consisted of four main steps: 1. the cells were isolated and sorted by CD status. 2. the DNA was extracted from the cells. 3. the mRRBS procedure was performed 4. the samples were sequenced. In Figure 16, a flowchart showing how the pilot studies of each step relate to each other is given. The different methods are described in further detail below.

Cell isolation

For collection of patient samples, a blood bag (500mL) (Fresenius Kabi, Bad Homburg, Germany) was prepared by adding 2mL of 0.5M pH 8.0 EDTA (Thermo Fisher Scientific, Waltham, USA). Approximately 200mL of blood was drawn. The bag was then filled with STEM buffer (0.2% EDTA and 2% FBS (Biowest SAS, Nuaillé, France) in DPBS (No calcium, no magnesium, Thermo Fisher Scientific) to a mixture of 1:1.5 of blood and STEM buffer respectively. The cell separation procedure using SepMate™-50 tubes (STEMCELL technologies) is illustrated in Figure 8. As shown in the figure, the tubes were prepared by adding 14mL, to the point of the insert, of lymphoprep (Axis-Shield, Oslo, Norway). The remaining volume was then filled with blood/STEM buffer mixture. After centrifugation, to create a density gradient, plasma and PBMC was removed and mixed together with the

PBMC wash (0.40% EDTA in DPBS) with a 50:50 ratio. A centrifugation at 340g for 10 minutes in order to pellet the cells and a subsequent pooling of samples by solving the pellets in STEMbuffer was performed. The cells were counted on a Countess automated cell counter (Thermo Fisher Scientific).

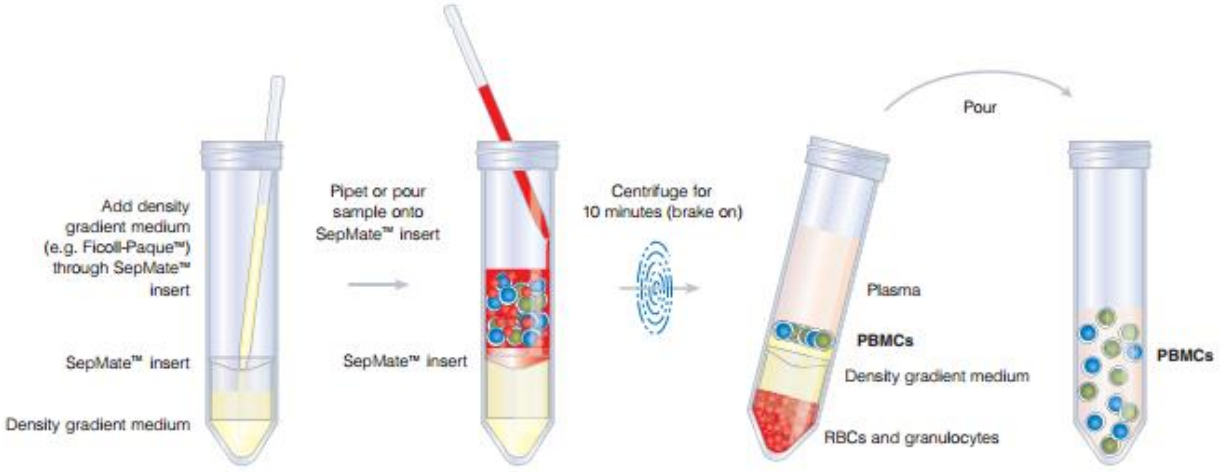


Figure 8 – PBMC isolation using SepMate tubes (STEMCELL technologies 2013).

The "EasySep positive selection for human $CD4^{+}CD25^{high}$ T cell isolation kit" (STEMCELL technologies, Cambridge, United Kingdom) was used to isolate the $CD4^{+}$ cells through negative selection, followed by isolation of $CD4^{+}CD25^{high}$ cells through positive selection of the $CD25^{high}$ cells. The "EasySep Human CD8 positive isolation Kit" (STEMCELL technologies) was used to isolate the $CD8^{+}$ cells through positive selection. The selection was performed through specific antibody complexes on the surface of the magnetic particles with an affinity towards matching cell surface antigens. In this case the cell surface antigens are CD4, CD25 and CD8. The general procedure for the isolation of specific types of T-cells consists of an incubation with an enrichment cocktail for the specific cell type being isolated (e.g. Human CD4+ T cell enrichment cocktail) before the addition of magnetic particles. The mixture including the magnetic particles was incubated in a tube magnet before pouring the supernatant off (Figure 9).

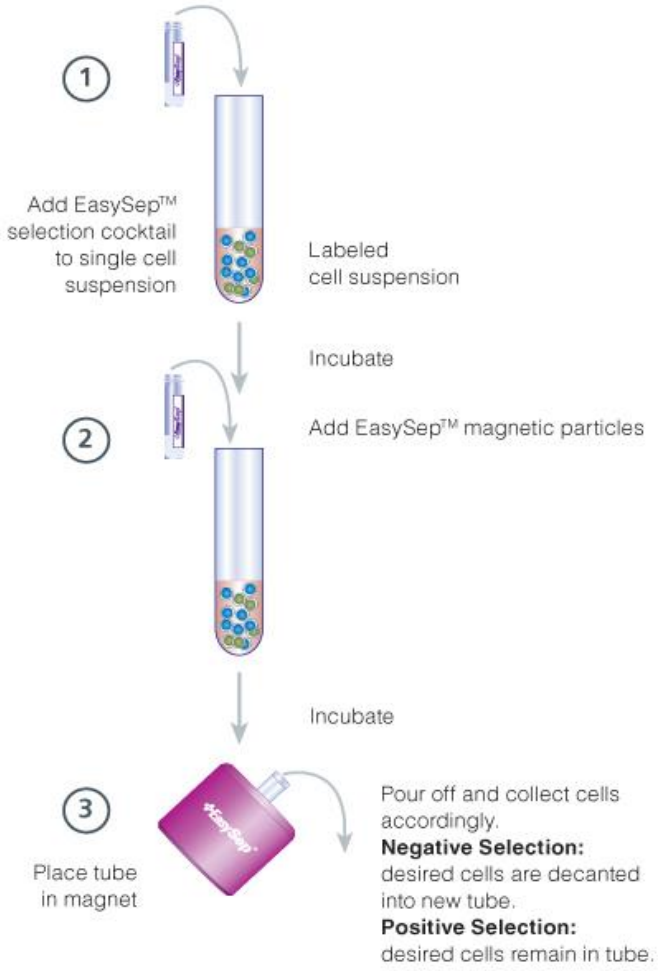


Figure 9 – The general procedure for isolating subtypes of cells using the EasySep kit (STEMCELL Technologies 2015).

In order to subtype the isolated $CD4^+$ and $CD8^+$ cells into naïve and memory cells the "EasySep Human PE Positive Selection kit" (STEMCELL technologies) coupled with CD45RO PE antibody (BioLegend, San Diego, USA) was used. The memory cells were positively selected based on presence of the antibody, while the naïve cells were negatively selected. When positively selecting the cell types, the cells remain in the original tube. When negatively selecting, they were transferred with the supernatant to a new tube. The procedure for the subtyping was the same as the one described for the cell sorting (Figure 9).

After the isolation process, the cell suspensions were centrifuged at 500g (Heraeus Biofuge Fresco, Thermo Fisher Scientific), and the supernatant removed. The dry pellets were stored at -80°C for sampe RA5509B, RA6616B and RA5111B. Sample RA5511B and RA5512B were stored at the same temperature, but on RNAprotect rather than as dry pellets. The control sample I and II samples were stored as both dry pellets and on RNAprotect. How the samples used were stored is specified for the description of each protocol.

DNA extraction

Four DNA extraction methods and two clean up protocols were tested in order to optimize the protocol so that sufficient concentration with the best possible quality of DNA could be attained for the downstream bisulfite conversion.

Manual extraction protocol

The cells used in this protocol was stored as dry pellets. DNA was extracted from the cells by the steps provided in an internal protocol for manual isolation of genomic DNA. In short, this protocol performs chemical lysis of the cells with a lysis buffer (1.6M Sucrose (Merck, Darmstadt, Germany), 5% v/v TRITON X (Merck) 25mM $MgCl_2$ (Merck) and 60mM Tris-hydrochloride (Merck)) before pelleting the material by centrifugation at 1000g. The pellet was then solved in a mixture containing 1x proteinase K buffer (0.375M NaCl (Merck) and 0.12M EDTA (Thermo Fisher Scientific)), 267µg/mL proteinase K (Merck) and 0.7% SDS (Bio-Rad Laboratories, Hercules, USA) and incubated overnight at 37°C. Next, 6M NaCl (Merck) was added before centrifugation at 1000g. The supernatant was transferred to a tube containing absolute ethanol (Ethanol AnalaR NORMAPUR® ACS, VWR, Radnor, USA) providing a final concentration of 70% ethanol. DNA then precipitated and was fished out with a glass rod. The DNA was rinsed by dipping in 70% ethanol before 1 minute air drying. In the end the DNA was solved in 1x low TE buffer (Thermo Fisher Scientific).

QIAamp DNA mini kit

The QIAamp DNA blood mini kit (QIAGEN, Manchester, United Kingdom) was tested in two rounds, differing in initial storage condition for the cells, either dry pelleted or on RNAprotect cell reagent (QIAGEN). They also differed in final elution volumes. In order to have the cells in liquid form, the dry pelleted samples were solved in 100µL buffer ATL before starting the procedure. The general procedure is illustrated in Figure 10.

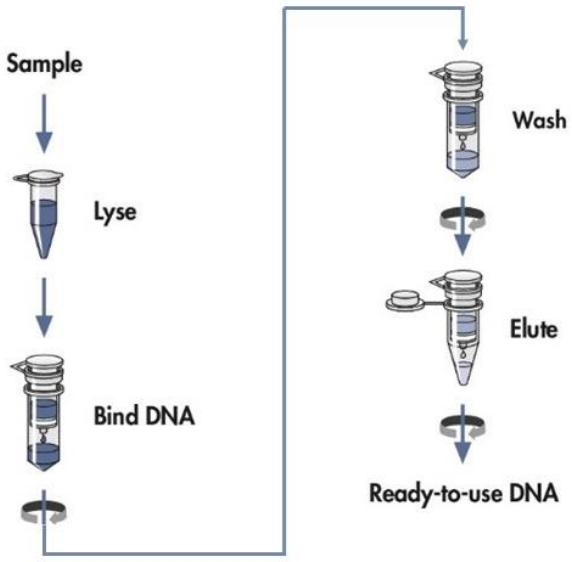


Figure 10 - The general procedure for column extraction of DNA. The circular arrows between the steps symbolize a centrifugation step. The first step consists of lysing the cells, the second is transfer of the solution to a spin column and binding of the DNA to the filter. The third step is to wash the filter containing the DNA before the last step which is the elution of the DNA bound to the membrane. Modified from QIAGEN (2015).

In short, the cell solution was mixed with proteinase K (Merck) and buffer AL before a 10 minute incubation at 56°C. Absolute ethanol (VWR) was then added to the mix before applying the whole volume to a QIAamp spin column. The spin column was then centrifuged at full speed (13000rpm, Heraeus Biofuge Fresco, Thermo Fisher Scientific), and the flow-through discarded. The spin and discard procedure was repeated with buffer AW1 and AW2 at full speed (13000rpm). An optional dry spin was not performed. In the end, buffer AE was applied to the spin column for elution of the DNA. In an attempt to generate samples with higher concentrations and purity, the final elution was done with either 2 x 100µL of buffer AE for the cells stored on RNAprotect cell reagent (QIAGEN), or once in 180µL buffer AE for the samples stored as dry pellets. When centrifuging at max speed, the samples eluted in 180µL buffer AE, were spun at 21000g instead of 20000g as the protocol indicated.

Column clean-up of manually extracted DNA

In an attempt to clean up the samples from the manual extraction, the “cleanup of genomic DNA” protocol from the QIAamp DNA micro kit (QIAGEN) was followed. However the mini, not the micro columns were used. This protocol was similar to the one described for the QIAamp DNA mini kit, but as the DNA was already isolated it lacks the first few steps.

The samples were mixed with buffer AW1 and AW2 before transferring to spin columns. The samples were then centrifuged at full speed (13000rpm, Heraeus Biofuge Fresco, Thermo Fisher Scientific) and the flow-through discarded. The process was repeated with the addition of more buffer AW2, followed by a dry spin at 20000g. Buffer AE was applied to the columns and incubated for 5 minutes before centrifugation at 20000g for eluting the DNA.

QIAamp DNA micro kit

Extraction of DNA was performed with the QIAamp DNA micro kit (QIAGEN). This kit was similar to the QIAamp DNA mini kit, but it supported a lower sample input and elution volume. The general procedure is the same as for the QIAamp DNA mini kit (Figure 10). The procedure was done according to the manufacturer's instructions for "isolation of genomic DNA from small volumes of blood" with a few modifications described in detail below.

Just as with the dry pelleted cells in the mini kit procedure, these cells were already dry-pelleted and as thus 100µL of Buffer ATL was added directly. However, for sample RA5111B and RA5516B (Table 1), the pellet was thawed by adding 100µL RNAprotect cell reagent (QIAGEN) before splitting two aliquots of 50µL each. One was stored at -80°C while the other was solved in additional 50µL of buffer ATL. Proteinase K (QIAGEN) and buffer AL was added, and an extra mechanical lysis step, performed by pipetting up and down a few times with a Sterican g21 syringe (VWR), was added before vortexing. 10 minutes of incubation at 56°C was performed before the addition of absolute ethanol (VWR). The whole volume was then applied to a QIAamp MinElute column. The column was then centrifuged at full speed (19980g, Hettich MIKRO 200, Hettich Instruments, LP, Tuttlingen, Germany) and the flow-through discarded. The spin and discard procedure was repeated with buffer AW1 and AW2 before a dry spin for three minutes at full speed. 80µL of Buffer AE was applied for elution of the DNA, and in order to increase the final DNA yield, the incubation time was increased from one to five minutes.

The DNA was eluted directly into a sterile, nuclease-free tube (Eppendorf® Biopur® Safe-Lock microtubes, Merck) and was stored in the freezer at -20°C.

A clean-up of a manually extracted sample was also performed at the same time as extraction with this kit. The clean-up was performed by following the "cleanup of genomic DNA" from

the same handbook as above. Just as for the rest of the samples, 80µL of buffer AE was used for elution.

Norgen RNA/DNA/Protein Purification Plus Kit

The Norgen RNA/DNA/Protein Purification Plus Kit (Norgen, Thorold, Canada) is also based on the use of spin-columns, but were able to isolate RNA and protein in addition to DNA. The manufacturer's instructions was followed, with the modifications of adding the mechanical lysis of the cells, and extra centrifugations before and after washing with PBS. The cell input was stored in 100µL of RNAprotect cell reagent (QIAGEN).

In short, the cells were first pelleted by centrifugation (5000g, Hettich MIKRO 200) for 5 minutes. After removal of the supernatant, the pellet was solved in RNAprotect and the centrifugation step repeated. The supernatant was removed once again, and the pellet was washed with PBS and the centrifugation repeated once again. The supernatant was removed before the addition of 300µL Lysis buffer Q. In addition to vortexing, a g21 syringe was used to help in lysing the cells. The lysate was added to a gDNA purification column and centrifuged at 5800g. The flow through was then transferred for RNA purification, while the column was put back in the collection tube. The gDNA and RNA purification protocols were then performed in parallel.

500µL of wash solution A was added to the gDNA column before centrifugation (3500g), the flow through was discarded and the process repeated once. A dry spin at 14000g was performed to ensure that the column was dry before transferring to a clean tube. 100µL Elution Buffer F was then added to the column and incubated in room temperature for 2 minutes. A centrifugation starting with 1 minute at 200g, then 2 minutes at 5800g and last 30 seconds at 14000g was performed. The eluate was then transferred back on top of the column, and the process repeated in order to increase the yield. The samples were then stored at 4°C.

A similar protocol was followed for the isolation of the RNA, but it will not be described in detail here. The isolated RNA was treated with DNase I in order to avoid gDNA contamination and stored at -80°C. The flow-through from the beginning of this protocol

contained the protein and was stored together with the RNA, available for further processing if needed at a later time.

Based on the results from the DNA extraction, a QIAamp micro column clean-up using the same protocol as earlier was performed, as well as a vacuum centrifugation on a CentriVap DNA Vacuum Concentrator (Labconco, Kansas City, USA) in an attempt to concentrate samples with less than 20ng DNA/µL.

Quality control of extracted DNA

The quality and quantity of the isolated DNA after each of the extractions was controlled by measurement on nanodrop ND-1000 (Thermo Fisher Scientific), and/or Qubit 2.0 fluorometer (Thermo Fisher Scientific). The nanodrop measures concentration of nucleic acids based on the absorbance in the 260nm, the UV-C area. Quality control was done by measuring the 260/280- and 260/230-ratios which describes the purity of the samples. The 260/280-value should be between 1.8, and 2.0. The 260/230-value is often higher than the 260/280-value, and should be somewhere between 1.8 and 2.2. Deviation from these values could point towards a difference in pH between the blanking buffer and the sample or the presence of contaminants (Thermo Fisher Scientific 2010). However, in general, the interpretation of too high 260/230 values is not covered in the literature, and as such we have assumed that too be less of a problem than if they are too low. However, a too high level could be due to problems with the blanking buffer (Thermo Fisher Scientific). Qubit measurement was performed by marking the dsDNA with fluorescent dyes before measuring the fluorescent signal to determine the concentration in the sample. In contrast to the nanodrop which measures unselectively by UV-light, the qubit measures the concentration of the DNA only, providing a more accurate measurement. However, the nanodrop is better suited for detection of contaminants in the sample (Life Technologies 2014).

Multiplexed Reduced Representation Bisulfite Sequencing

The mRRBS was performed based upon the description by Boyle et al. (2012), however some details were not described in the article, and several pilot studies was necessary in order to optimize the protocol. Problems encountered included, but were not limited to, reagents that are no longer produced and lack of details regarding settings and incubation times. The mRRBS procedure consists of five main stages: 1. MspI digestion, 2. filling of the gap

created by the restriction pattern, A-tailing and adapter ligation, 3. bisulfite conversion, 4. bisulfite cleanup and amplification and 5. final cleanup and stock library creation. This general procedure is illustrated in Figure 11 and described in detail further below.

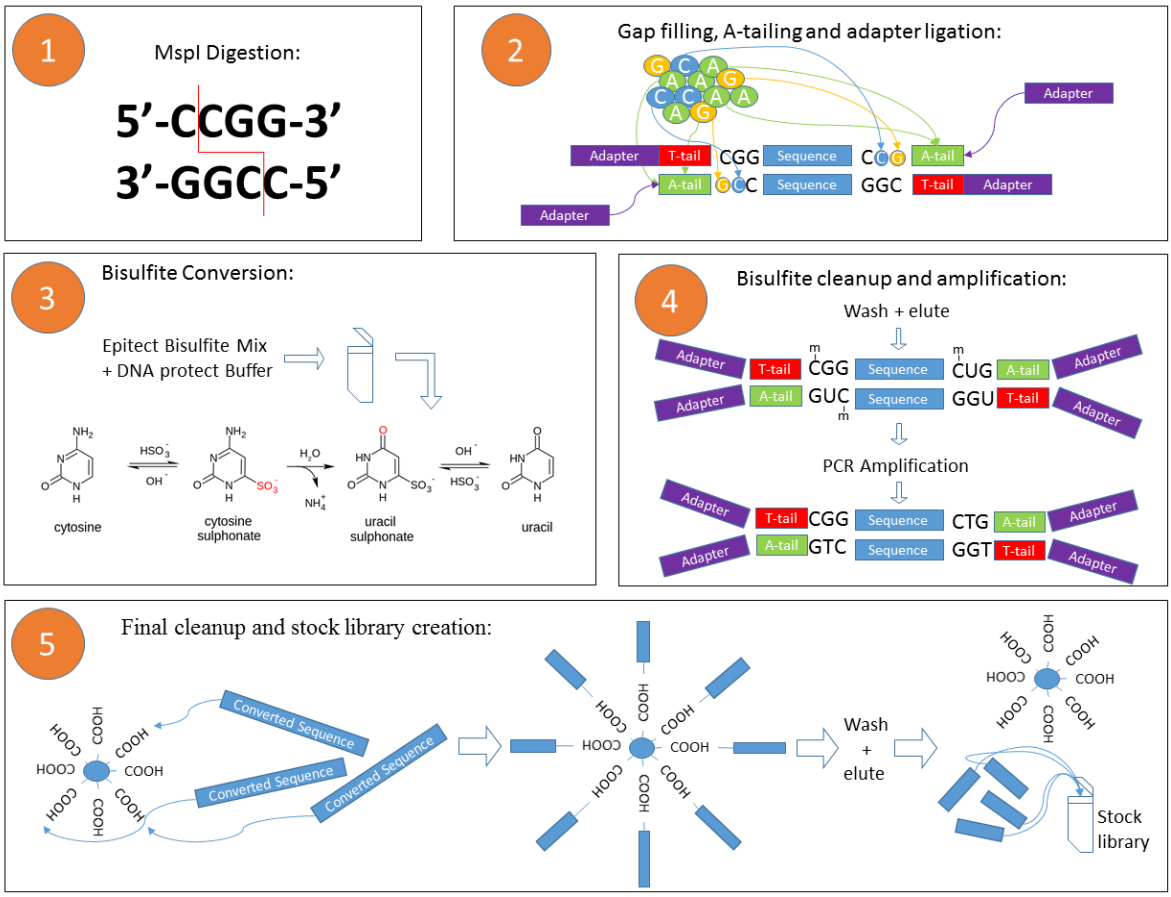


Figure 11 - Overview of the mRRBS procedure. !. MspI digestion, 2. Gap filling, A-tailing and adapter ligation, 3. The bisulfite conversion of unmethylated cytosines to uracils, 4. Cleanup from the bisulfite reaction and amplification where all uracils are swapped with thymines and 5. Final cleanup using AMPure XP beads and library creation.

MspI digestion

It was important that each sample had a concentration of 20ng/µL, i.e. 5µL with 100ng of DNA input. The samples were diluted in the appropriate amount of Low TE buffer (Thermo Fisher Scientific/QIAGEN) in order to achieve this. The concentration of each sample were measured on Qubit (Thermo Fisher Scientific) before dilution to ensure that the concentration was correct.

The 20ng/µL of DNA was mixed together with nuclease free water (QIAGEN/Thermo Fisher Scientific), 10x NEB buffer 2 (New England Biolabs (NEB), Ipswich, USA) and 20 U/µL MspI (NEB), to a total concentration of 3.3 ng/µL, 1x and 667 U/mL respectively. The

reaction was then incubated at 37°C for 19 hours. The restriction site which the MspI was targeting in this step is illustrated in Figure 11.1.

Gap filling, A-tailing and adapter ligation

To control the MspI digestion, 1µL was removed and diluted 1:4 in nuclease free water before testing on a 2200 tapestation (Agilent Technologies) using a D1000 ScreenTape (Agilent Technologies, Santa Clara, USA).

An overview at what happened with the gap filling, A-tailing and adapter ligation in the following sections is illustrated in Figure 11.2.

5000 U/mL Klenow fragment (NEB) was added to the sample reaction together with a dNTP mix consisting of 1mM dCTP (NEB) and dGTP (NEB), 10mM dATP (NEB) and no dTTP. The total concentration was then 156 U/mL for the Klenow fragment, 31.25µM dCTP and dGTP and 312.5µM dATP. The samples were then placed 20 minutes at 30°C and 20 minutes at 37°C before a hold of 4°C, all incubations were performed without a heated lid. Agencourt AMPure XP beads (Beckman Coulter, Indianapolis, USA) were added to the mix in twice the amount of the total reaction volume, before incubating at room temperature for 30 minutes.

The samples were then placed on a DynaMag2 magnet (Thermo Fisher Scientific), and the supernatant removed after five minutes. A double wash of fresh 70% ethanol was performed, with 5 minutes incubation during the second wash. The beads were then air dried for 10 minutes before the addition of 20µL EB buffer (QIAGEN).

Boyle et al. (2012) used Illumina TruSeq adapters from catalogue number PE-940-2001 (Illumina, San Diego, USA) which they claimed to be at a stock concentration of 9µM. However, these adapters are no longer for sale, so instead it was necessary to use TruSeq nano DNA adapters from FC-121-4001 (Illumina). The concentrations of these adapters were not possible to obtain, but Illumina support advised us to use the same amounts as Boyle et al. (2012), so we assume the same concentration. The stock adapter solutions were diluted 1:20 in nuclease free water. While working on ice, nuclease free water, 10x T4 DNA ligase reaction buffer (QIAGEN), 400000 U/mL T4 DNA ligase (QIAGEN) and 1:20 TruSeq nano DNA adapters (Illumina) were mixed together with the sample reaction in order to ligate the

adapters to the DNA fragments. This resulted in a final concentration of 1x, 13333 U/mL and 0.17% for the respective components. The reaction was then incubated at 16°C without a heated lid for 20 hours.

Bisulfite conversion

The T4 DNA ligase was inactivated by increasing the temperature of the reaction mix to 65°C for 20 minutes. 20% polyethylene glycol/2.5M NaCl (PEG) (KAPA PEG/NaCl SPRI® Solution, Kapa Biosystems, Wilmington, USA) was then added in double the volume of the reaction mix, and incubated for 30 minutes at room temperature.

Next, the samples were placed on a magnet, and the supernatant removed. A wash with fresh 70% ethanol was performed before the beads were left to air dry until they cracked. When the beads had dried sufficiently, 25 µL EB buffer was added for elution of the DNA. 23µL of each sample eluate was removed, of this, 20µL was set up for bisulfite conversion, while 3µL was used to perform a test-PCR on a thermal cycler (Applied Biosystems 2720 Thermal Cycler/Applied Biosystems Veriti Thermal Cycler, Thermo Fisher Scientific), to determine the optimal amount of PCR-cycles for amplification of the final converted product.

The bisulfite conversion of the DNA fragments was performed as described by the “Sodium Bisulfite Conversion of Unmethylated Cytosines in DNA Isolated from FFPE Tissue Samples” protocol in the EpiTect Bisulfite kit (QIAGEN). The bisulfite conversion reaction is illustrated in Figure 11.3.

For the initial DNA input, the whole eluted volume of 20µL was used, without any dilution, and mixed with 85µL dissolved bisulfite mix. When adding 35µL DNA protect buffer, the color changed from green to blue, indicating correct pH for the reaction. The sample was repeatedly denatured and incubated using a Thermal Cycler (Figure 12). The reaction volume was set to the maximum, at 100µL on the machine.

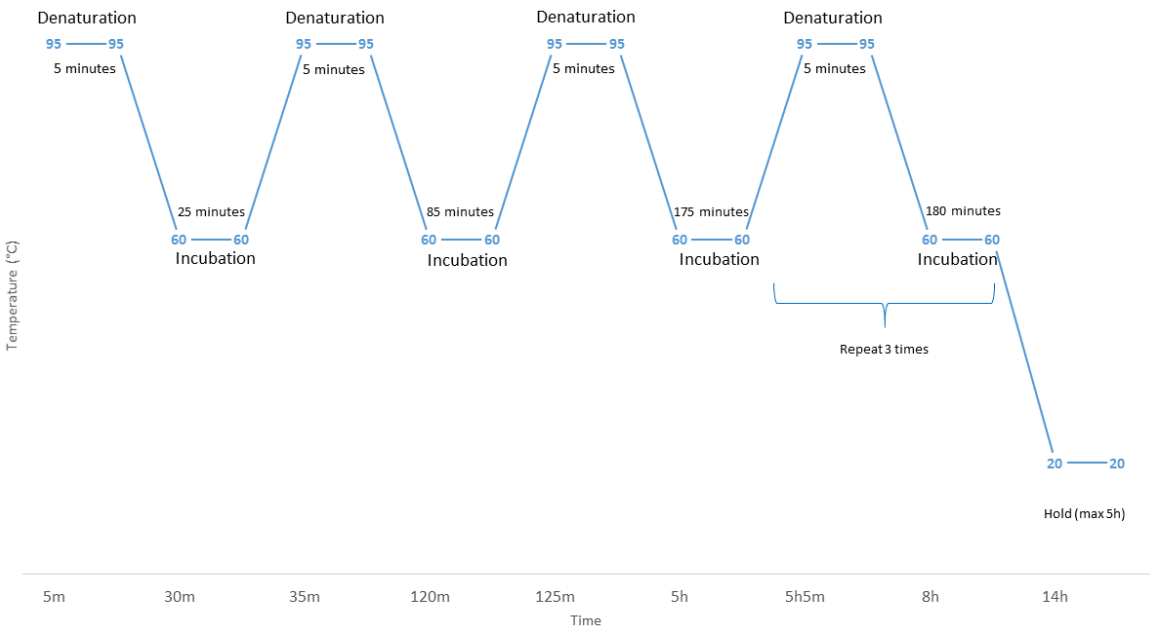


Figure 12 - Settings for the thermal cycler during Bisulfite Conversion of unmethylated DNA. The number on the graph represents the temperature in °C.

Test-PCR

A test-PCR was performed in order to determine the optimal amount of cycles for the amplification of the final converted library (Figure 11.4). This had to be done for each round of conversion. As mentioned earlier, the test-pcr was performed on unconverted DNA, before the bisulfite treatment. While working on ice, 34.75µL of PCR grade water (Roche, Basel, Switzerland), 5µL 10x PfuTurbo C_x reaction buffer (Agilent Technologies), 0.5µL 100mM dNTP (Agilent Technologies) with 25mM of each dNTP, 5µL PCR primer cocktail (Illumina) and 1µL 2.5U/µL PfuTurbo C_x Hotstart DNA polymerase (Agilent Technologies) was added to the 3µL of eluted DNA. The concentration of the PCR primer cocktail was not known, but based off Boyle et al. (2012) statements it was assumed to be 3µM. The total reaction volume of 50µL was aliquoted to five 10µL reactions and each tested for a different amount of cycles of the PCR-program shown in Figure 13.

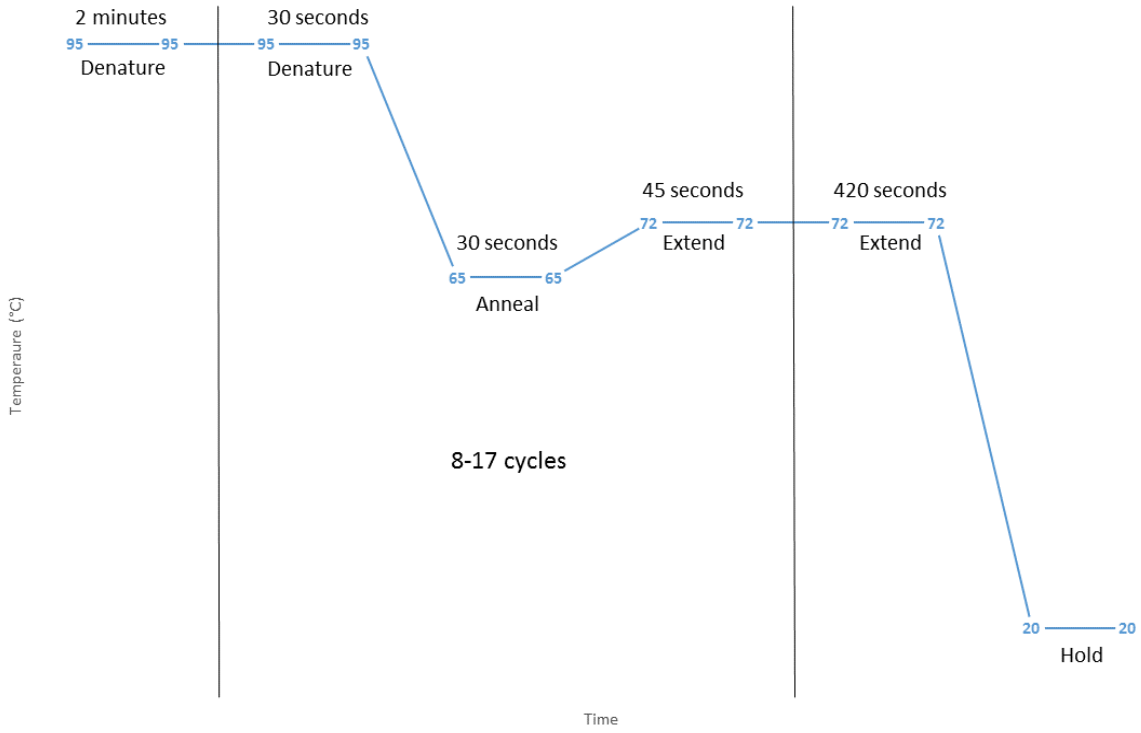


Figure 13 - PCR settings for the test-PCR and amplification of bisulfite converted DNA. Each of the test-PCR samples should be covering a different cycle number in the variable region. The numbers on the graph represents the temperature in °C.

After the test-PCR had completed, the products were controlled by high-sensitivity qubit measurement and/or tapestation D1000 ScreenTape. These results were used to determine the optimal amount of cycles for each experiment.

Bisulfite cleanup and amplification

The cleanup from the bisulfite conversion was performed as according to the same protocol as was used for the startup of the conversion. In short, the protocol used EpiTect spin columns for the cleanup. Buffer BL with 10µg/mL carrier RNA and absolute ethanol was added to each sample before transferring to the spin columns. After centrifugation (16000g), buffer BW was added and another centrifugation (16000g) was performed. Buffer BD was added and incubated for 20 minutes at room temperature before another centrifugation (16000g). This was a modification from the 15 minutes stated by the protocol. Buffer BW was added and centrifuged twice (20000g, Hettich MIKRO 200, Hettich Instruments, LP). A dry centrifugation was performed before a recommended 5 minutes incubation with open lids at 56°C. When eluting, the EB buffer was heated to 65°C before addition of 22µL and 2 minute incubation. The DNA was then eluted with a 1 minute centrifugation (15000g). In

order to provide a backup solution, 20µL of additional EB buffer was added and eluted in new tubes using the same settings.

A 200µL reaction for amplifying the 20µL of bisulfite converted DNA was prepared by mixing with PCR grade water, 10x PfuTurbo C_x reaction buffer, 100mM dNTP mix with 25mM of each dNTP, 3µM PCR primer cocktail and 2.5U/µL PfuTurbo C_x Hotstart DNA polymerase. Each component was added so that they had a final concentration of 0.1ng/µL for the DNA and 1x, 1mM total, 0.25mM each, 0.3µM and 0.05U/µL for each of the respective remaining components.

The thermal cycler was set to the same settings as for the test-PCR (Figure 13), with the variable part set to the amount of cycles decided based on test-PCR results. However, about three additional cycles should always be added for amplification of the bisulfite converted DNA in order to reach the same amount of product (Gu et al. 2011). This was due to the test-PCR being performed on unconverted DNA. The unmethylated cytosines are replaced with uracils during conversion before amplification, and after the amplification they become thymines (Figure 11.4).

Final cleanup and stock library creation

In order to perform the final cleanup, 1.2 times the total reaction volume of Ampure XP beads were added before incubating for 15 minutes. The tubes were put on a magnet and the supernatant was removed. Fresh 70% ethanol was used to wash the beads twice. After removing the supernatant again, the beads air dried for 15 minutes. 40µL of EB buffer was added before incubation at room temperature for 3 minutes, immediately followed by 2 minutes on a magnet. The supernatant was transferred without any beads to new tubes and 1.5 times (60µL) the eluate volume of beads was added before the mixture was incubated for 15 minutes. The same procedure as the above section was then repeated. In the end, as much as possible of the 40µL of eluate was transferred to yet another sterile, nuclease free tube for storage of the final bisulfite converted library. It was important that there were no beads transferred together with the eluate. The final libraries were marked and stored at -20°C. A simple overview of the cleanup using these ampure XP beads are illustrated in Figure 11.5. 4µL of the eluate was used for quality control with Qubit dsDNA High Sensitivity (HS) Assay kit, tapestation D1000 and nanodrop ND-1000.

The pilots

Details surrounding each pilot experiment follow below. This includes description of the experiences and methodological deviations from the standard protocol described above. An overview of which samples that were used in each of the pilot experiments, as well as DNA extraction method used is given in Table 2.

Table 2 - overview of which samples were used in the different mRRBS pilot experiments, and which DNA extraction protocol was used.

Sample ID	Cell type	DNA extraction method	mRRBS pilot
Control sample II	CD4 total	QIAamp DNA mini kit	Pilot I
Control sample II	CD8 total		
Control sample I	CD8 total		
Control sample I manual + column cleaned	CD8 total		
RA5511B	CD4 naïve	QIAamp DNA micro kit	Pilot II
RA5509B	CD4 naïve		
RA5512B	CD8 memory		
Control sample II	CD8 total		
RA5111B	CD4 naïve	QIAamp DNA micro kit	Pilot III
RA5111B	CD8 memory		
RA5516B	CD4 naïve		
RA5516B	CD8 memory		

RA5111B	CD4 naïve	QIAamp DNA micro kit (QIA-micro)	Pilot IV
RA5111B	CD8 memory		
RA5516B	CD4 naïve		
RA5516B	CD8 memory		
RA5516B	CD8 memory	Norgen RNA/DNA/Protein Purification Plus Kit (Norgen)	
RA5516B	CD4 naïve		
RA5111B	CD4 naïve		
RA5516B	CD8 memory	up concentrated Norgen RNA/DNA/Protein Purification Plus Kit (Norgen)	
RA5111B	CD4 naïve	up concentrated Norgen RNA/DNA/Protein Purification Plus Kit + QIAamp micro kit (NorgenClean)	
RA5111B	CD8 memory		

Pilot I

Being the nature of a first pilot study (Table 2), many elements differ between this and the final protocol described earlier. For example, as earlier mentioned, because of vague details in the descriptions in the original protocol by Boyle et al. (2012), the bisulfite conversion was done twice with a cleanup in between. The bisulfite conversion also followed another protocol “Sodium Bisulfite Conversion of unmethylated cytosines in DNA” from the same kit as described for the final procedure. The cleanup in between the conversions was done according to the end cleanup from this protocol. Other parts of the procedure differed as well. 20µL of EB buffer was used at the end of the cleanup for eluting the sample, and the flow-through volume was measured afterwards, as 20µL was needed for the PCR. The test-PCR was also performed with 10-20 cycles instead of the 8-17 cycles stated in the final protocol

earlier in this text. The final converted libraries were quality controlled with high sensitivity qubit and D1000 ScreenTape.

Pilot II

In this pilot study (Table 2), the optimal incubation time for digestion of genomic DNA with MspI was decided. This was done by testing the CD8 Control sample II digested for different amounts of time on the tapestation D1000. 17.5, 19 and 21 total hours of digestion was tested.

During the last cleanup step before the library creation, there was most likely made an error in the preparation of the fresh 70% ethanol for washing so that it instead got a concentration of 30%.

Pilot III

During this experiment (Table 2), when preparing for the bisulfite conversion, the DNA protect buffer had not been equilibrated to room temperature beforehand. This in turn led to there being approximately 12 minutes between the addition of the bisulfite mix to the sample and the DNA protect buffer. It was still observed that the buffer changed color as usual when added to the mixture, so the pH should still be correct.

After bisulfite conversion, precipitates was observed in the bottom of each well and for sample RA5111B CD8 a dry pellet was observed along the wall of the well, above the liquid. This pellet is shown in Figure 14. The precipitates and pellet was attempted resuspended as well as possible before continuing. According to the conversion protocol, presence of precipitates is okay.

Figure 14 - Picture of the observed dry pellet along the wall of the sample well of RA5111B CD8. The red ring marks the dry pellet.

The test-PCR did not show any signal, and a test was performed to see if the 10X PfuTurbo C_x reaction buffer was to blame. This was done through splitting the converted samples in 2 x 100µL samples instead of 1 x 200µL reaction for amplification. All other reagents were then also halved, and two different vials of the suspected buffer was used, one for each sample. The rest of the reagents were the same for both samples. There was however a problem with the DNA inputs as they all had closer to a volume of 15µL than the 20µL they were supposed to. For this reason, each sample was distributed as according to Table 3. There is also a small potential source of error in that after adding water to the RA5111B CD4 sample with buffer vial 1, the pipette tip touched the top of the water. The water was still used for the remaining samples using the same buffer. Also, for some unknown reason the RA5111B CD8 sample using buffer vial 2 had less volume than the rest of the samples after preparing the amplification mix. All samples were amplified with 15 cycles.

Table 3 - distribution of DNA sample for each of the amplification mixtures in order to test the different vials of 10X PfuTurbo C_x reaction buffers.

Sample ID	Cell type	$V_{\text{DNA sample with buffer 1}}$ (µL)	$V_{\text{DNA sample with buffer 2}}$ (µL)
RA5111B	CD4 Naïve	6.9	10
RA5111B	CD8 memory	10	5.1
RA5516B	CD4 naïve	10	4.1
RA5516B	CD8 memory	4.9	10

The final libraries were only quality controlled on tapestation D1000 for this experiment.

Pilot IV

Samples in pilot IV extracted with the Norgen RNA/DNA/Protein Purification Plus Kit were called Norgen, those extracted with the QIAamp DNA micro kit were called QIA-micro, and those extracted by both kits were called NorgenClean.

All samples were diluted in low TE so that each had a total of 20ng/µL except the ones with an original concentration bellow that threshold. This applied to Norgen RA5516B CD8, RA5111B CD4 and RA5111B CD8 as well as QIA-micro RA5516B CD8 (Table 7).

The samples QIA-micro RA5111B CD4 and RA5516B CD8 were situated on ice for a few minutes after addition of klenow fragment while gathering a new vial of reagent for the rest of the samples when the first one was emptied. This new vial had the same lot number as the first one.

The film covering each well when mixing the bisulfite mix, sample and DNA protect buffer did not cover the wells adequately, and as thus there was a chance of cross-contamination between Norgen RA5111B CD4 and RA5516B CD4 as well as the QIA-micro RA5111B CD4 and RA5516B CD4. This was discovered by observation of liquid between the film and the top of the plate covering the area of the four samples. After spinning the plate, the liquid was dried off and the film replaced. The RA5111B samples had index 5 while the RA5516B samples had index 19 (Figure 15).

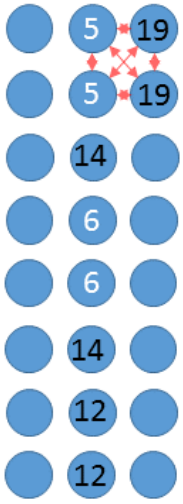


Figure 15 - Illustration of the 24-well PCR plate and arrangement of the samples together with markings where there was a potential cross contamination. The wells are marked with the adapter used for each sample. The samples were, from the top down as follows: RA5111B CD4 Norgen, RA5111B CD4 QIA-micro, RA5111B CD4 NorgenClean, RA5111B CD8 Norgen, RA5111B CD8 QIA-micro, RA5516B CD8 NorgenClean, RA5516B CD8 Norgen, RA5516B CD8 QIA-micro, RA5516B CD4 Norgen and RA5516B CD4 QIA-micro.

After the bisulfite conversion Norgen RA5111B CD8 was discovered with dry material along the wall of the well, above the liquid mixture. This material was attempted resolved into the solution by vortexing. For the amplification mix, the total volume of sample was used for every single sample as none of them had a sufficient 20µL. The PCR primer cocktail mix ran out and as such Norgen RA5111B CD4 and RA5111B CD8, QIA-micro RA5111B CD4, and NorgenClean RA5111B CD4 received a mix from a different lot. There was also a delay between these and the rest of the samples for the addition of the mix as the new mix had to be thawed. Only three samples: Norgen RA5111B CD4, QIA-micro RA5111B CD4 and NorgenClean RA5111B CD4 received the proper DNA polymerase. The rest got an expired polymerase after a substantial delay.

During sample pooling after the conversion, one of four wells from QIA-micro RA5516B CD8 was accidentally pooled with Norgen RA5516B CD8. Afterwards everything went on as normal and 36µL of each sample was frozen down at -20°C apart from for NorgenClean RA5516B CD8 where only 35µL was stored.

Quality control of each sample was performed on qubit and tapestation.

Cleanup of samples from pilot III and IV

Sample RA5111B CD8 using buffer 1 from pilot III and sample QIA-micro RA5111B CD4, Norgen RA5111B CD4 and NorgenClean RA5111B CD4 from pilot IV were all attempted cleaned with ampure XP beads to get rid of primer dimers. The same protocol as usual when performing cleanup of the final libraries was performed once more with 1.2x volume of beads. In the end 40µL EB buffer was added to each sample. The first two samples mentioned above had a final eluate volume of 39µL removed, while the two others had 38µL.

Each sample was quality controlled with qubit, tapestation and nanodrop.

Sequencing

The samples with extra cleanup from pilot III and IV was sequenced in the end (Table 4). Due to need for complementary adapters on the sequencer, the RA5111B CD8 sample was also included, even though the concentration was low (Table 9). This sample, as well as NorgenClean RA5111B CD4, was for the same reason split between the two sequencing lanes (Table 4). Because of this, the data needed to be pooled for these samples after the sequencing had been performed.

Table 4 - sample ID, cell type, lane- and indexes numbers for each of the sequenced samples

Sample ID	Cell type	DNA extraction method	Index	Lane
RA5111B	CD4 naïve	QIAamp DNA micro kit	5	1
RA5111B	CD4 naïve	Norgen DNA/RNA/Protein purification plus kit + QIAamp DNA micro kit	14	1
RA5111B	CD8 memory	QIAamp DNA micro kit	6	1
RA5111B	CD4 naïve	Norgen DNA/RNA/Protein purification plus kit + QIAamp DNA micro kit	14	2
RA5111B	CD4 naïve	Norgen DNA/RNA/Protein purification plus kit	5	2
RA5111B	CD8 memory	QIAamp DNA micro kit	6	2

The sequencing of the final samples was performed on a MiSeq system (Illumina) using the 600 cycles MiSeq Reagent Kit v3 (Illumina). However, 150bp paired end reads were performed. A spike in of 50% PhiX control v3 (Illumina) was used as cluster generation control. This was because of the anticipated unbalance of AT- and GC-content as a result of the bisulfite conversion. Boyle et al. (2012) suggested performing a method known as "dark sequencing" in order to deal with the cluster generation problem. However, as this method affects all of the lanes for a sequencing run, economic considerations lead to the choice of the PhiX control.

The sequencing service was provided by the Norwegian Sequencing Centre (www.sequencing.uio.no), a national technology platform hosted by the University of Oslo and supported by the "Functional Genomics" and "Infrastructure" programs of the Research Council of Norway and the Southeastern Regional Health Authorities.

Data analysis

The results from the sequencing are presented in FastQ format. Due to being sequenced on two lanes in the MiSeq sequencer, NorgenClean RA5111B CD4 had two sequencing output files, These were combined before further in the analysis process.

RRBSMAP was used as the software tool for mapping bisulfite converted sequences to the reference genome (Xi & Li 2009). The mapping was performed as single end even though the sequencing was performed as paired end, this was because single end reading on the HiSeq are the standard procedure and will be performed for later experiments. For the same reason RRBSMAP was set to map the 75 first nucleotides of each read. The default setting of 2 mismatches allowed was also used.

After alignment with RRBSMAP, quality reports for the alignment were given together with the SAM file. The output file was converted to BAM and sorted with the help of SAMtools (Li et al. 2009). The sorted BAM file was used with methratio.py to obtain the methylation ratio of each cytosine in the alignment.

The CpG and non-CpG positions from the methratio.py output file was extracted and stored in separate files. The positive strand was also isolated from the non-CpG file. Because

cytosine methylation in mammals mainly occurs in the CpG context (Ziller et al. 2011), the non-CpG sites are not expected to be methylated, and all cytosines in these positions should be converted to thymines after bisulfite treatment. By calculating the mean methylation ratio value for all these non-CpG positions, an estimate of the percentage of cytosines that was unsuccessfully converted was created, and should ideally be at 0%. This value can then be subtracted from 100% to find the bisulfite conversion ratio (Leontiou et al. 2015).

Start and stop positions for the codons of RA genes were extracted from the human reference genome (GRCh37). All methylation sites with at least 5x coverage, together with relevant annotation data, such as strand information, chromosome number, location and methylation ratios contained within the first 5000bp upstream of each gene were then extracted from the methratio.py output file. The genes were selected by comparing the list of the 42 RA risk loci provided by Okada et al. (2014) with a list of RA genes found to be expressed by $CD4^+$ T-cells in the blood from healthy human adults (Helgeland et al, unpublished data). In addition, some randomly chosen genes with no expression in the CD4 cells from the same dataset were chosen, as well as some of the genes mentioned in the introduction. Lastly, all of the expression data from the list was plotted against methylation for all of the genes by chromosome. The identified genes are shown in Figure 26 together with the expression levels and identified methylation sites in each sample for all of the genes.

Results

Figure 16 contains a flowchart illustrating the relation between the cell isolation, DNA extraction and mRRBS pilot experiments, as well as which pilots that were sequenced in the end.

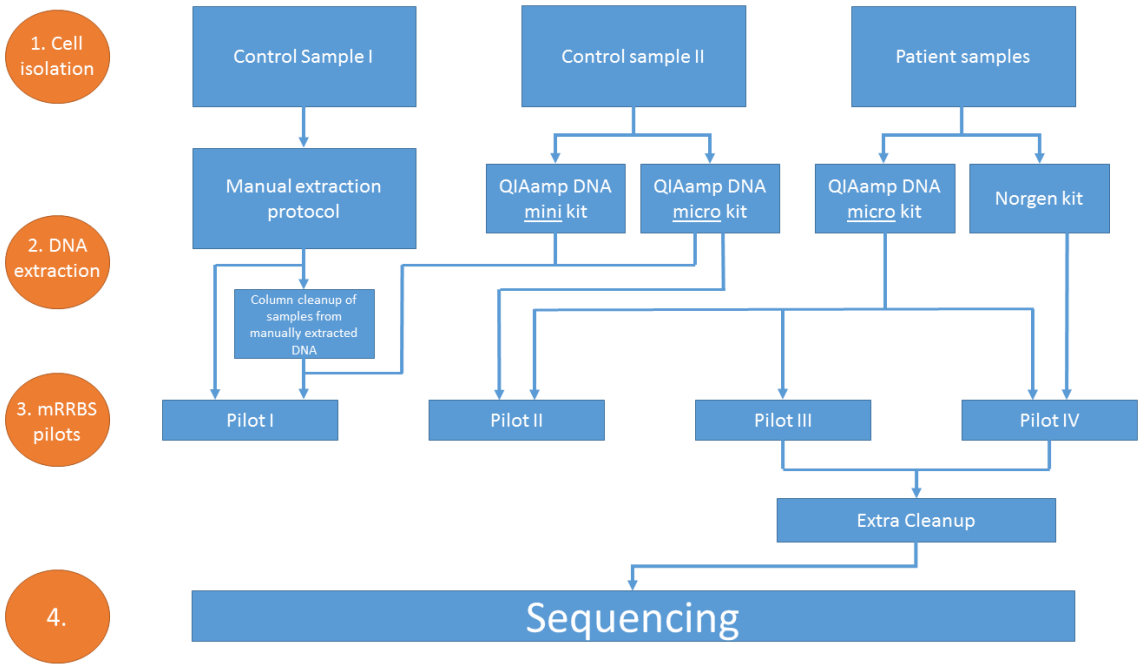


Figure 16 - A flowchart showing the connection between each of the pilot studies for every step of the experimental procedure.

Cell counts after isolation

The total cell count after separation was consistently in the area of 2 - 5· 10^6 for the isolation experiments (Table 5). These are appropriate numbers as the manual protocol was scaled for approximately 2· 10^6 cells, while the QIAamp DNA mini and Norgen RNA/DNA/Protein purification plus kits supported up to 5· 10^6 cells. The control samples were not subtyped beyond that of $CD8^+$ and $CD4^+$ cells

Table 5 - Cell counts from the countess automated cell counter for each of the isolated cell types from each sample.

Sample ID	Cell type	Cell count
Control sample I	CD4 total	$4 \cdot 10^6$
Control sample I	CD8 total	$4 \cdot 10^6$
Control sample II	CD4 total	$2 \cdot 10^6$
Control sample II	CD8 total	$5 \cdot 10^6$
RA5509B	CD4 naïve	$2 \cdot 10^6$
RA5511B	CD4 naïve	$4.5 \cdot 10^6$
RA5512B	CD8 memory	$3 \cdot 10^6$
RA5111B	CD4 naïve	$4.05 \cdot 10^6$
RA5111B	CD8 memory	$5 \cdot 10^6$
RA5516B	CD4 naïve	$3 \cdot 10^6$
RA5516B	CD8 memory	$3 \cdot 10^6$

Quality of the DNA obtained from different extraction procedures

The Nanodrop measurements for all the samples from the different DNA extraction methods are given in Table 6. As earlier mentioned, the 260/280-value should be between 1.8 and 2.0, while the 260/230 should be between 1.8 and 2.2. For the manual extraction protocol both of these values were well within the accepted range for both samples. However, contaminants were visually observed in the DNA samples from the CD8 cells as they had a brown tint which could not be observed in the DNA sample from the CD4 cells. The coloring also persisted in the CD8 control sample solution.

Table 6 - Nanodrop results for all DNA extraction experiments on cells from the control samples.

Sample ID	Cell type	DNA extraction experiment	$C_{nanodrop}$ (ng/µL)	260/280	260/230
Control sample I	CD4 total	Manual extraction protocol	270.29	1.84	2.26
Control sample I	CD8 total		128.59	1.81	1.86
Control sample I 1. eluate	CD8 total	QIAamp DNA mini kit. Stored on RNAprotect cell reagent before extraction.	17.27	1.85	2.16
Control sample I 2. eluate	CD8 total		30.84	1.55	0.78
Control sample I 1. eluate	CD4 total		30.94	1.83	2.52
Control sample I 2. eluate	CD4 total		29.15	1.61	0.76
Control sample II	CD8 total	QIAamp DNA mini kit. Stored as dry pellets before extraction.	33.72	1.81	3.12
Control sample II	CD4 total		51.72	1.8	2.61
Control sample II bottom layer	CD8 total		104.87	1.52	0.99
Control sample I	CD4 total	Column clean-up of manually extracted DNA	45.73	1.7	1.64
Control sample I	CD8 total		19.65	1.64	2
Control sample I	CD8 total	QIAamp DNA micro kit	3.08	2.8	2.85

manual + cleanup					
Control sample II	CD8 total		116.22	1.81	1.88
Control sample II	CD4 total		31.44	1.91	1.46
Control sample II + mechanical lysis	CD4 total		48.44	1.85	2.44

The table consists of sample ID, cell type, associated DNA extraction experiment, and nanodrop data in the form of concentrations, 260/280- and 260/230-values. 260/280 values should be between 1.8 and 2.0 while 260/230 values should be between 1.8 and 2.2.

After having isolated the DNA from the cells through the use of the QIAamp spin columns during the first experiment, a visual inspection of the columns was performed. From this inspection it became evident that the contaminants from the CD8 cells indeed were left in the filter (Figure 17).

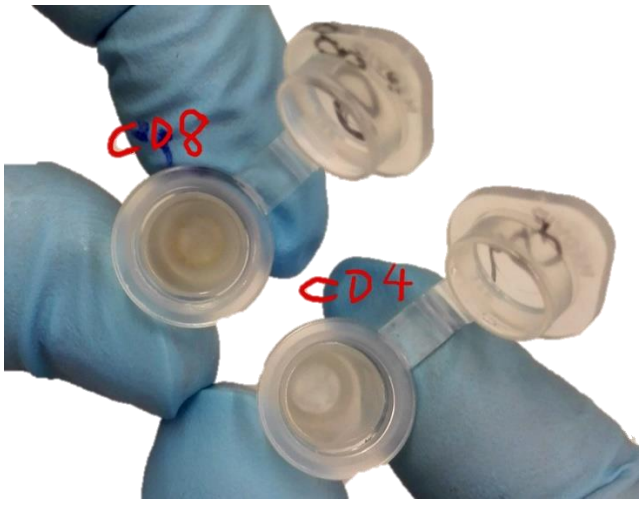


Figure 17 - Visual inspection of the two spin columns for extraction of DNA from the CD8 and CD4 cells from Control sample I. The samples are marked according to which cell type that had been spun. There is a visibly darker tint to the CD8 sample, indicating that the beads are left in the filter. The contrast has been slightly adjusted in the picture in order to emphasize the difference.

The total DNA yield was lower for the mini kit, than from the manual extraction protocol (Table 6). Furthermore, the quality dropped to unusable levels with 260/230-values below 0.8 when eluting for the second time. The first eluate however has acceptable 260/280- and 260/230-values at about 1.8 and above 2 respectively.

Both leftovers from the cell pellet and the EasySep Magnetic particles from the CD8 Control sample II passed through the filter during centrifugation in the second round. This was probably due to the centrifugation being performed at 21000g instead of 20000g as indicated by the protocol. As a result, the samples became dirty and contaminated. As all contaminants gathered in the bottom layer after centrifugation, the supernatant and bottom layer was split into two tubes (CD8 Control sample II and CD8 Control Sample II bottom layer). TheCD8 control sample II bottom layer had poor quality as measured by the 260/280- and 260/230-ratios, but it had a high concentration, probably due to a high degree of contamination. The concentrations for the rest of the samples in the second run were all better than for the samples in the first round. The 260/280-values from the second round were acceptable, but the 260/230-values on the other hand were too high. However, as mentioned in the methods this was assumed to be okay due to the literature rarely focusing on the effect of too high values.

Control sample I CD4 and CD8 which were cleaned on the spin column after following the manual extraction protocol had slightly too low 260/280 values. Likewise for the 260/230 value for the CD4 Control sample I (Table 6). This sample was however the only sample that meets the minimal required concentration for further use downstream with the mRRBS. The CD8 control sample I almost met the concentration requirements. Even though the spin column got rid of the visible contamination of the CD8 sample, the quality measurements had been degraded (Table 6).

For the samples where the DNA was extracted with the QIAamp DNA micro kit, the clean-up of the manually extracted DNA from the CD8 cells had a very low concentration, and poor quality measurements. The directly extracted CD8 sample had a very high DNA yield, and good quality. The CD4 sample, on the other hand, had lower concentration, and the 260/230 value was too low. Another round of extraction with the same protocol, with the addition of the mechanical lysis which had been used for the CD8 sample was performed, and an improvement in both the concentration and quality measurements was observed (Table 6).

In summary, the QIAamp DNA mini kit did give DNA of sufficient quantity and quality. The QIAamp DNA micro kit provided DNA of similar quality, but with higher concentration (Table 6).Therefore, in order to obtain samples with DNA yield, the QIAamp DNA micro kit was initially chosen.

Extraction of DNA from patient samples

As the QIAamp DNA micro kit did give DNA samples with sufficient quality and concentrations, extraction of patient samples were initiated. However, after having extracted a few of the samples it was decided to test the Norgen RNA/DNA/Protein Purification Plus kit in order to also extract the RNA from the same samples.

All the samples extracted with only the Norgen RNA/DNA/Protein Purification Plus kit had too low 260/230-values, and varying 260/280-values (Table 7). The concentration was acceptable for all samples, except RA5111B CD8 memory. The samples cleaned with the QIAamp micro spin columns obtained slightly better 260/280- and 260/230-values, but still not quite acceptable. In addition, they had a drop in concentration making them too low for further use.

For three samples, the DNA was attempted concentrated by vacuum centrifuge, but with little effect (Table 7). There was a small increase in the concentration for two of the samples, while sample RA5111B $CD8^+$ memory had a decrease in the concentration. All three samples still had too low concentration for the downstream bisulfite conversion.

Table 7 – Quantitative and qualitative measurements of patient DNA samples.

DNA Extraction method	Sample ID	Cell type	Qubit C_{Stock} (ng/µL)	$C_{nanodop}$ (ng/µL)	260/280	260/230
Norgen	RA5111B	CD4 naïve	34.8	106.45	1.57	0.60
Norgen	RA5111B	CD8 memory	17.0	23.78	1.69	0.56
Norgen	RA5516B	CD4 naïve	28.0	26.30	1.70	0.34
Norgen	RA5516B	CD8 memory	60.8	34.99	1.85	0.56
Norgen + QIAamp DNA micro kit	RA5516B	CD8 memory	14.6	9.94	1.66	1.99
Norgen + QIAamp DNA micro kit	RA5111B	CD4 naïve	13.8	15.54	1.58	1.32
Norgen + QIAamp DNA micro kit and vacuum centrifugation	RA5516B	CD8 memory	15.8	27.34	1.82	0.59
Norgen and vacuum centrifugation	RA5111B	CD8 memory	15.5	34.1	1.65	0.78
Norgen + QIAamp DNA micro kit and vacuum centrifugation	RA5111B	CD4 naïve	17.6	23.79	1.82	0.8
QIAamp DNA micro kit	RA5509B	CD4 naïve	55	45.83	1.87	3.37
QIAamp DNA micro kit	RA5511B	CD4 naïve	41	69.55	1.86	2.7
QIAamp DNA micro kit	RA5512B	CD8 memory	37	32.58	1.78	2.52
QIAamp DNA micro kit	RA5111B	CD4 naïve	43.2	24.41	1.95	1.81

QIAamp DNA micro kit	RA5111B	CD8 memory	59.6	79.98	1.65	1.04
QIAamp DNA micro kit	RA5516B	CD4 naïve	39.2	24.07	1.96	1.66
QIAamp DNA micro kit	RA5516B	CD8 memory	19.8	20.92	1.86	1.87

Sample ID, cell types, concentrations, 260/280- and 260/230-values as measured on nanodrop are given for all patient samples. Qubit concentrations are also given for each sample. 260/280 values should be between 1.8 and 2.0 while 260/230 values should be between 1.8 and 2.2.

Of the samples extracted using the QIAamp DNA micro column kit (Table 7), all samples, except RA5111B CD8 memory and RA5516B CD4 naïve, had acceptable 260/230-values. RA5111B CD8 memory also had a too low 260/280 value. The concentrations, as measured on qubit, were also sufficient for further bisulfite conversion for all of the samples.

In summary the QIAamp DNA micro kit gave samples containing a high enough concentration at above 20ng DNA/µL for the downstream mRRBS procedure. The 260/280 values were between 1.8 and 2.0 for all samples, and all but two samples had acceptable 260/230 values above 1.8 (Table 7). For the two other extraction methods however, this was not the case. The Norgen samples had low 260/230 values far below 1.8 indicating contamination as well as varying 260/280 values. They did however overall have sufficient DNA concentrations. The NorgenClean samples had better quality measurements, but the concentration was too low. The concentrations did not improve with vacuum centrifugation and the qualities dropped (Table 7). Regardless of these problems, the samples extracted by Norgen and NorgenClean were brought along for mRRBS as isolation of RNA for use in other projects was an aim for this study.

Parameter testing and quality control for Multiplexed Reduced Representation Bisulfite Sequencing

An overview of which extraction method that were used for the samples for each of the mRRBS pilot experiments can be found in Figure 16.

In mRRBS pilot II, 17.5, 19 and 20 hours of incubation with MspI was tested in order to determine the optimal digestion time, giving the most uniform arrangement of fragment sizes

possible (Figure 18). All three time points were tested for the same sample, CD8 Control sample II (Table 2).

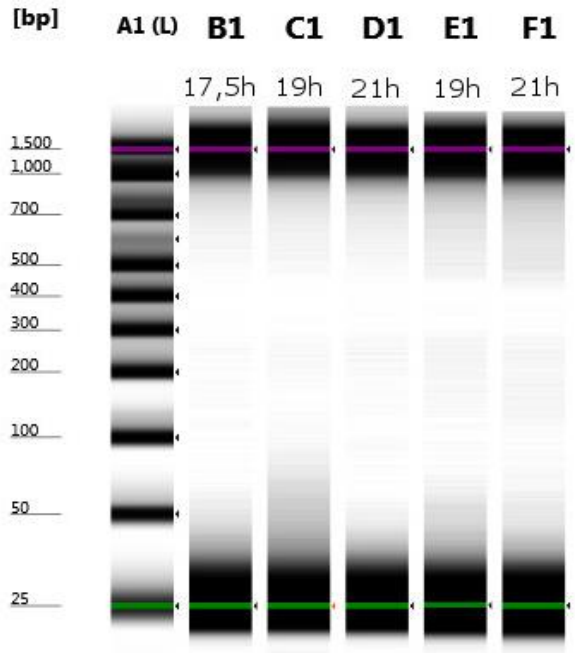


Figure 18 - test of MspI digestion of CD8 Control sample II at three different time-points: 17.5, 19 and 21 hours.

Good fragmentation and distribution of fragment sizes was achieved from 19 hours of digestion (Figure 18), and therefore this number of hours was used for the genomic digestion during later mRRBS pilot experiments. Next, a number of samples from the different DNA extraction experiments were digested. A varying degree of fragmentation and sizes can be observed in the samples from pilot IV, but to some extent all samples show signs of digestion from the restriction enzyme, with the size distribution covering the largest amount of fragment sizes observed in NorgenClean RA5516B CD8 sample (Figure 19).

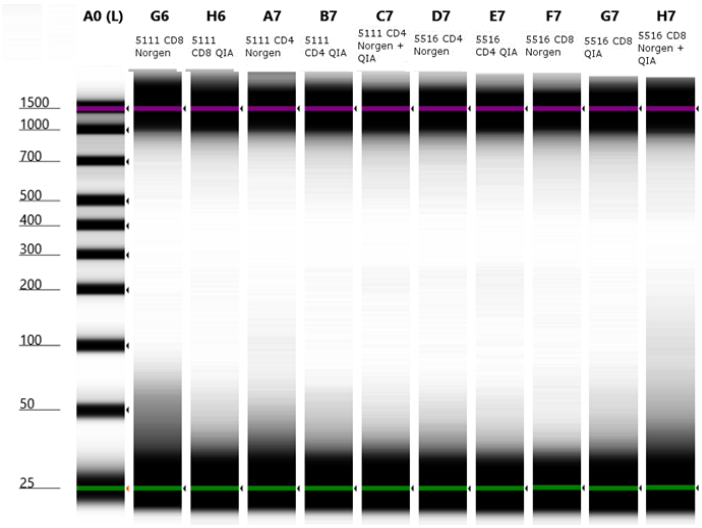


Figure 19 - Fragment sizes on a D1000 tapestation ScreenTape after 19 hours of MspI digestion of each sample from pilot IV.

A test-PCR was then performed to determine the appropriate number of cycles of amplification of the converted DNA. An example of this test after evaluation of the gel results is shown for the Norgen RA5111B CD4 sample (Figure 20). This figure shows the results from 8, 10, 12, 15 and 17 cycles of amplification. As the cycle number increases, the amount of fragments also visibly increases, especially in the size area around 200 - 300bp.

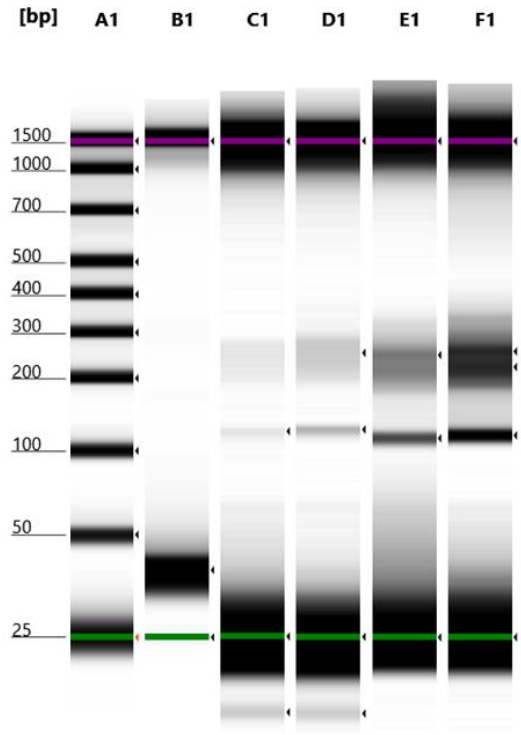


Figure 20 – Agilent D1000 gel image of Test-PCR results from pilot IV. All lanes (except ladder) contains sample RA5111B CD4⁺ extracted with the Norgen kit. A1: ladder, B1: 8 cycles, C1: 10 cycles, D1: 12 cycles, E1: 15 cycles, F1: 17 cycles.

The amount of amplification cycles needed to be individually chosen for each of the pilot experiments based on similar results as those presented in Figure 20, alternatively qubit measurements were used to evaluate the amplification. Maximum amount of DNA product at minimal number of PCR cycles are wanted, to optimize the quality. For the amplification of the converted product in pilot I, 16 cycles was chosen, for pilot II half of the samples were amplified at 12 cycles, and the other half at 15 cycles (due to uncertainty about the DNA amount needed), pilot III used 15 cycles. During test-PCR of pilot IV, only the sample shown in Figure 20 provided any fragments. Based on these test results, 19 cycles of amplification was chosen for the protocol based on the recommendations by Gu et al. (2011) to add three additional cycles to that of the optimal test-PCR.

As earlier mentioned in the methods chapter, the eluate volume during the first bisulfite clean-up of pilot I was measured (Table 8). With 20µL of elution buffer input, the eluate will have a volume of less than 20µL. As the PCR reaction was scaled for a 20µL DNA input, it was decided that the eluate volume should be increased to 22µL for the second cleanup after the second conversion in pilot I as well as after the bisulfite cleanup of the subsequent pilot experiments.

Table 8 - Eluted volumes of first eluates during clean-up of bisulfite product in pilot I.

Sample	CD8 Control sample I	CD8C Control sample I	CD4 Control sample II	CD8 Control sample II
Measured volume of eluate (µL)	19	19.5	18	16.5

CD8C refers to the column cleaned sample.

Quality and concentration of final mRRBS libraries

The concentrations of the final libraries for each mRRBS experiment, except pilot III, were measured on qubit (Table 9).

Table 9 - Qubit concentrations from the final bisulfite converted libraries

Sample ID	Cell type	DNA extraction method	$C_{stock,\ qubit}$ (ng/µL)	mRRBS pilot
Control sample II	CD4 naïve	QIAamp DNA mini kit	0.002	Pilot I
Control sample II	CD8 memory		0.001	
Control sample I	CD8 total		8.96	
Control sample I manual + column cleaned	CD8 total		2.14	
RA5511B	CD4 naïve	QIAamp DNA micro kit	<0.05 before stock calculation	Pilot II
RA5509B 15 sykler	CD4 naïve		0.186	
RA5512B	CD8 memory		<0.05 before stock calculation	
Control sample II	CD8 Total		<0.05 before stock calculation	
RA5111B	CD4 naïve		67.0	

		QIAamp DNA micro kit (QIA-micro)		Pilot IV
RA5111B	CD8 memory		<0.05 before stock calculation	
RA5516B	CD4 naïve		0.660	
RA5516B	CD8 memory		<0.05 before stock calculation	
RA5516B	CD8 memory	Norgen RNA/DNA/Protein Purification Plus Kit (Norgen)	<0.05 before stock calculation	
RA5516B	CD4 naïve		0.228	
RA5111B	CD4 naïve		86.2	
RA5516B	CD8 memory	up concentrated Norgen RNA/DNA/Protein Purification Plus Kit (Norgen)	<0.05 before stock calculation	
RA5111B	CD4 naïve	up concentrated Norgen RNA/DNA/Protein Purification Plus Kit + QIAamp micro kit (NorgenClean)	1.39 probably a misreading, see pilot IV cleanup sample	
RA5111B	CD8 memory		<0,05 before stock calculation	
RA5111B	CD4 naïve	QIAamp DNA micro kit (QIA-micro)	45.0	Pilot IV cleanup
RA5111B	CD4 naïve	Up concentrated Norgen RNA/DNA/Protein Purification Plus Kit + QIAamp micro kit (NorgenClean)	35.8	
RA5111B	CD4 naïve	Norgen RNA/DNA/Protein Purification Plus Kit (Norgen)	53.6	
RA5111B	CD8 memory	QIAamp DNA micro kit (QIA-micro)	0.676	Pilot III cleanup

In pilot I, the final mRRBS libraries of the two Control sample II samples obtained very low concentrations, and the amount of sample used for measuring had to be increased in order to be detectable (Table 9). The two samples from Control sample I had product and looked more promising at about 2 and 9 ng/µL. These results correlated well with the results shown on the tapestation gel (Figure 21), where the two Control sample II samples showed weak signals, while the two other samples showed fragmentation more akin to the successful libraries in later pilot experiments. It should be noted that the gel used had expired, and could be a potential error source, however, the size ladder appeared as expected.

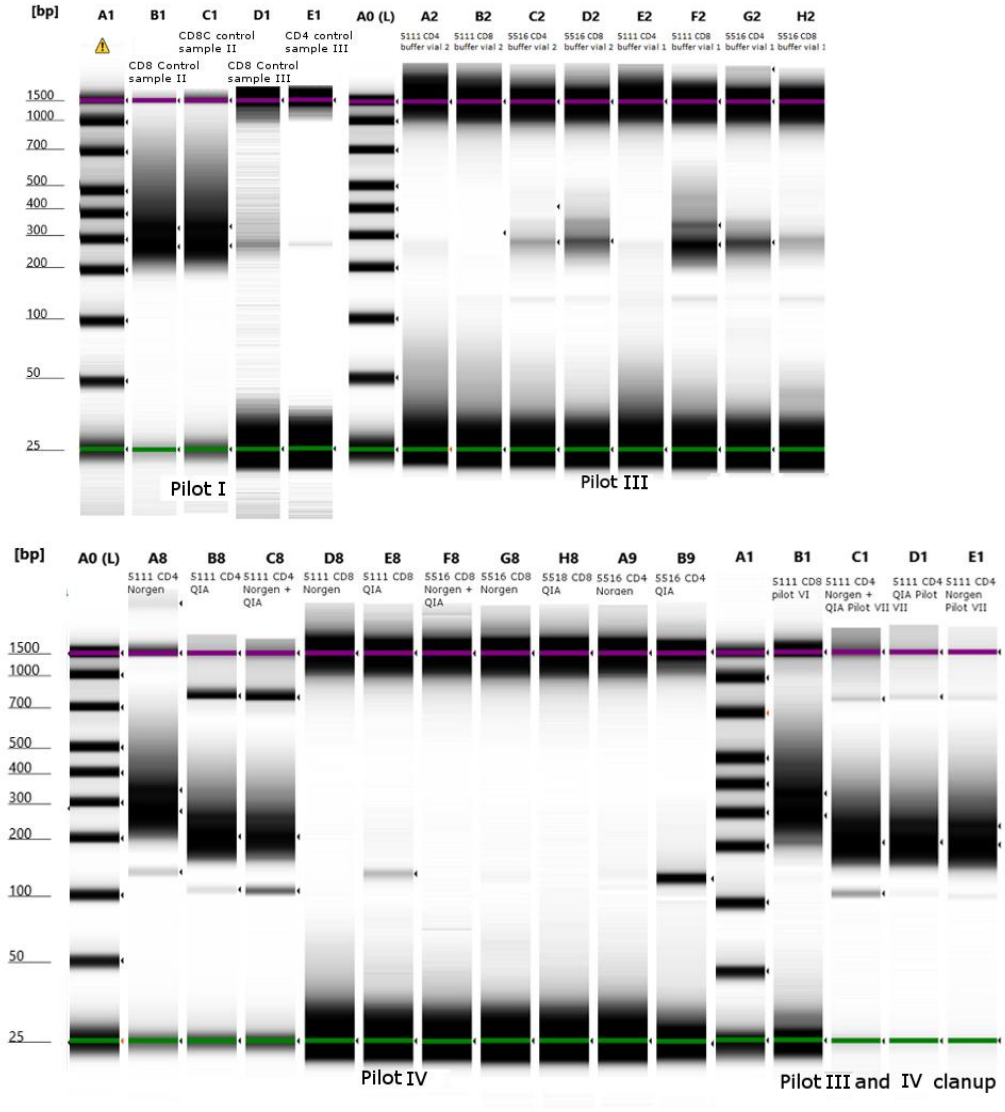


Figure 21 - The final mRRBS DNA libraries shown on Agilent D1000 ScreenTape gels for each of the pilot experiments. CD8C in pilot I refers to the column cleaned sample. For pilot III both results from samples tested with both vials of the 10X PfuTurbo C_x reaction buffer is shown. The samples for each pilot experiment follows to the right of the ladders in either well A(0) or A1 for each of the experiments.

None of the final library samples from pilot II contained any DNA (Table 9), except for sample RA5509B CD4 which had a concentration of 0.186 ng/µL, below 10ng/µL which was the minimum amount for sequencing on the MiSeq system (Illumina inc 2016b). For this reason, no gel imaging was done for these samples. The reason for the lack of sample was most likely due to the erroneous use of 30% ethanol instead of 70% during one of the wash steps, as described in the method.

Only tapestation D1000 was used in order to quality control the final bisulfite converted libraries from the pilot III experiment (Figure 21). Because of the lack of amplicon from the test-PCR, which was suspected being due to problems with the 10X PfuTurbo C_x reaction buffer, two different vials of this buffer was tested for the final library amplification. By inspecting the gel, it becomes clear that there has not been any amplification of sample RA5111B CD4 or RA5111B CD8 using buffer from vial 2 or from RA5111B CD4 or RA5516B CD8 samples using the buffer from the other vial. The other samples had to a varying degree some amplified DNA. In other words, the buffer was not to blame for lack of test-PCR signal.

There does seem to be some correlation between the input volume of DNA sample and the degree of amplification in pilot III. Sample RA5111B CD8 and RA5516B CD4 with buffer from vial 1, as well as sample RA5516B CD8 with buffer from vial 2, were the samples with the most amplicon, and all received the total of 10µL of DNA sample input. Sample RA5111B CD4 with buffer from vial 2 also received this amount of DNA sample input, but did not show any higher degree of amplification than the counterpart with 6.9µL of DNA sample input and the different buffer. As these looked quite similar on the gel, this could be attributed to something in the sample itself. The rest of the weaker samples all received less than the 10µL of DNA sample (Table 3). However, no correlation could be found beyond the difference between the full volume of 10µL added, and those with reduced volume added.. None of the samples had a high yield, and the highest concentration calculated by the tapestation software was from RA5111B CD8 with buffer from vial 1 of 0.730ng/µL.

It became apparent that regardless of DNA extraction method, the only sample from pilot IV with a significant amount of DNA left in the final library was RA5111B CD4, although it was a little on the low end for the NorgenClean sample (Table 9). This however was to be expected as it had, as earlier mentioned, a low original input to the mRRBS procedure as well

(Table 7). The RA5516B CD4 samples for both QIA-micro and Norgen also showed a very low concentration (at less than 0.7 ng/µL) the quantity was too low for sequencing. The rest of the samples had an unmeasurable concentration (Table 9). The results aligned nicely with the gel images (Figure 21). This at least applied for the three samples with noteworthy concentrations, namely RA5111B CD4- Norgen, QIA-micro and NorgenClean.

All the samples from pilot III and IV which went through the extra cleanup, except the CD8 sample still had a high enough concentration for the subsequent sequencing (Table 9). In accordance with this, all four samples also looked promising on the gel, although QIA-micro RA5111B CD8 from pilot III still a little less. The extra clean-up also seems to have successfully reduced the amount of primer dimers, as evident by the reduced band intensity at about 100bp (Figure 21).

The nanodrop results for controlling the quality of the final libraries from pilot III and IV samples with extra cleanup (Table 10) had higher concentrations than on the qubit (Table 9). They do however agree on the most abundant sample, and the data fits well with the gel results (Figure 21). All of the three RA5111B CD4 Norgen, QIA-micro and NorgenClean samples from pilot IV all had good quality measurements. This could be seen by the 260/280 values at a little above 1.8, and 260/230 values at about 2 for each sample (Table 10). All three of these samples also had good qubit concentration at 35ng/µL and above (Table 9). The RA5111B pilot III $CD8^+$ sample on the other hand had a low quality. Especially the 260/230-value, at 0.60, was too low. The nanodrop concentration was also not great at only 3.33ng/µL (Table 10). It should be noted that QIA-micro RA5111B CD4 from pilot IV had less than 1µL applied to the nanodrop, although it does seem like enough was applied for measurement.

Table 10 - Nanodrop results for the cleaned samples from pilot III and IV.

Sample ID	Pilot	Extraction method	Cell type	$C_{nanodrop}$ (ng/ul)	260/280	260/230
RA5111B	Pilot III	QIA-micro	CD8 memory	3.33	2.25	0.60
RA5111B	Pilot IV	NorgenClean	CD4 naïve	83.63	1.86	2.08
RA5111B <1µL	Pilot IV	QIA-micro	CD4 naïve	88.64	1.86	1.94
RA5111B	Pilot IV	Norgen	CD4 naïve	78.17	1.83	2.09

MiSeq sequencing output quality

The samples from the last clean up were sequenced on the MiSeq system (Table 10), however, no further analysis was performed on the QIA-micro from pilot III. Quality reports from the MiSeq sequencing output, generated with fastqc, were delivered for one of the samples of each run. However, RRBS routinely involves single-end sequencing, therefore only the data from the first read of each sample will be presented here. The data for lane one was given for QIA-micro from pilot IV, and lane two was given for the NorgenClean sample.

The quality scores for the QIA-micro and NorgenClean samples were generally quite good and there was no indication of any quality drop until about 70bp, and even then the quality stayed within reasonable levels until around 125bp for QIA-micro and 110 for NorgenClean (Figure 22). Both samples had quality scores above 30 until about 75bp. According to Illumina a quality score of 30 means that only one in 1000 bases will be erroneously called, and a higher score equals a lower probability of erroneous base calling (Illumina Inc 2014).

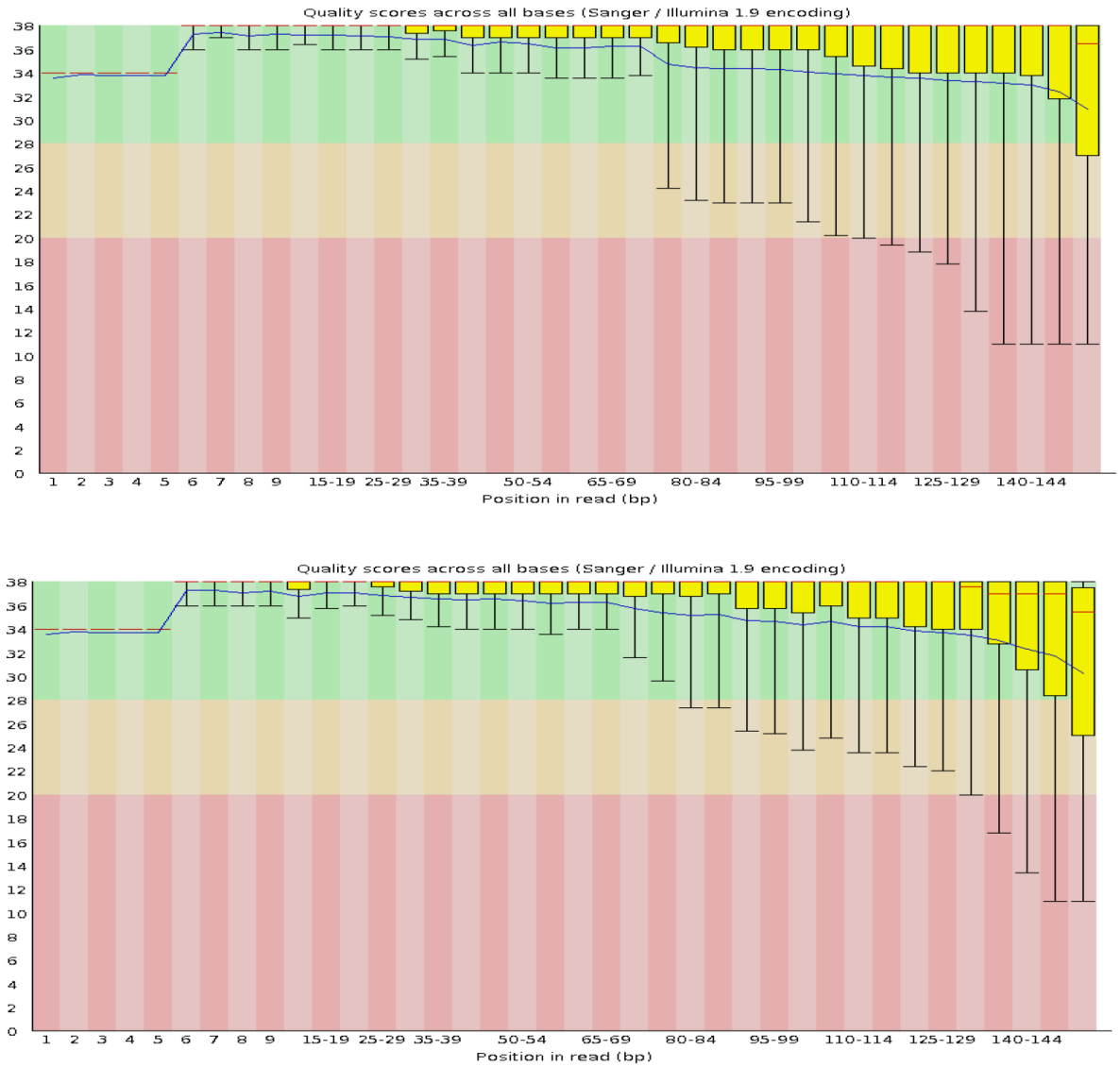


Figure 22 - Per base sequence quality scores (y-axis) for lane 1 containing sample RA5111B CD4 extracted with the QIAamp DNA micro kit on the top, and for sample RA5111B CD4 extracted with the Norgen kit followed by QIAamp cleanup on the bottom. The quality is good for at least the first 75bp in both samples.

There is as expected an altered proportion of bases read for each position of the read. There is approximately the same amount of adenines and guanines, but a higher proportion of thymines, and lower proportions of cytosines for the first 80bp, after which the proportion starts evening out (Figure 23). This is in accordance with the quality scores for each base in a read (Figure 22). The results are indicative of successful bisulfite conversion of the unmethylated cytosines. Not shown is the plot demonstrating that all bases were properly identified as one of the four standard bases, meaning the sequencer was always able to interpret a specific base to the signal. The high proportion of cytosine + thymine base calls in position one at about 80-90% in total, together with the equally high proportion of guanine in

position two and three is also reassuring in regards of proper base callings as this perfectly matches the restriction site of MspI (C^CGG).

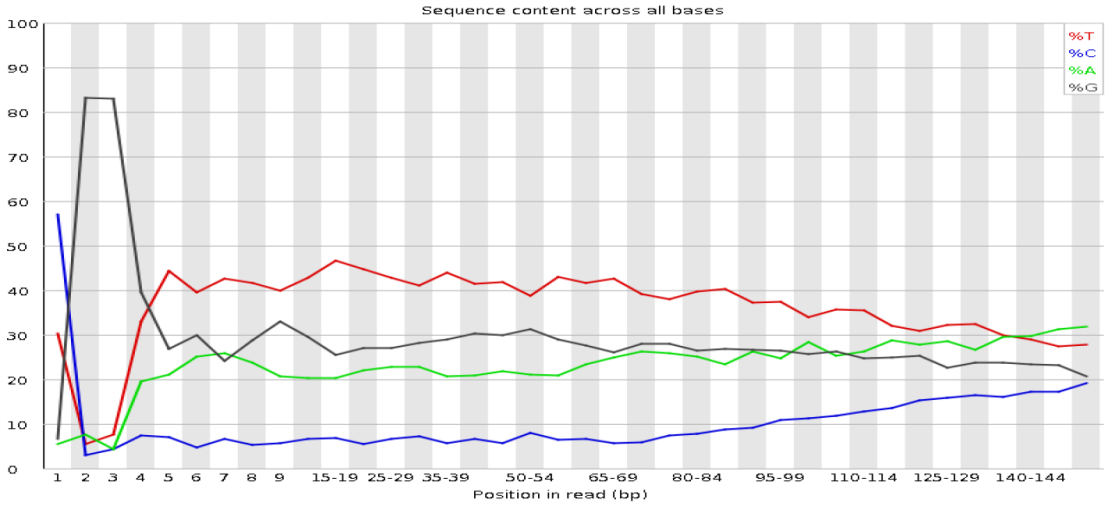


Figure 23 - proportion of each of the four bases in each read position. As the two samples from the two lanes are practically identical, only the result from QIA-micro from pilot IV is shown. There is a higher than normal amount of thymines detected, and a lower than normal amount than normal of cytosines detected. This is just as expected due to the bisulfite conversion.

Quality control of RRBS data from different DNA extraction methods

It was of interest to check whether there was evidence for the potential cross-contamination in the samples QIA-micro RA5111B and Norgen RA5111B, with suspected cross contamination as described in the methods for pilot IV. To do so a search for different indexes in the sequence files for each of the two aforementioned samples was performed. Just as expected there was a lot of hits for the adapter sequence five, which was actually used for the samples. The QIA-micro sample got somewhere between 450 – 500 thousand hits while Norgen got about 100000 less hits (Figure 24). This does however make perfect sense as there is about 80000 less total reads in the Norgen sample compared to the QIA-micro sample (Table 11). Index 19 however, which was the index in the samples suspected to have cross-contaminated, was not detected in the raw data from the samples (Figure 24). Neither does any of the randomly chosen indexes used as controls. Index 14 and 6 on the other hand, which were used for the other sequenced samples, occured in both samples. However, none of them occur in more than 0.007% of all the counted adapter sequence occurrences.

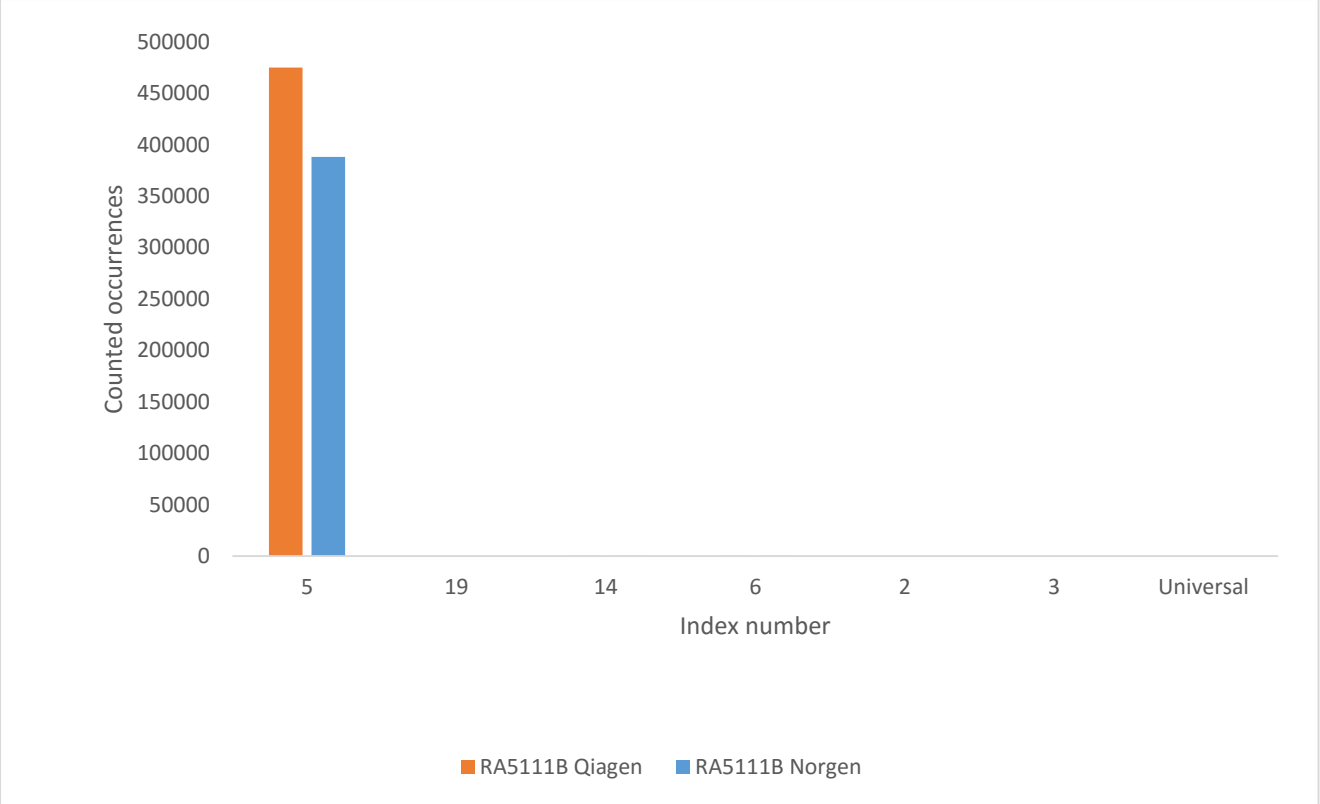


Figure 24 – visualization of the different amount of each adapter in each of the two sequenced samples with a potential for cross-contamination. As expected large amounts of index five was observed, and none of the potential cross contaminating sample. Index 14 and 6 however has a slight occurrence in both samples.

The reads for each of the samples were aligned using BSMAP (Table 11). All samples had more than 60% of their reads mapped to the genome, and more than 50% was unique reads for each of the samples. Especially, the Norgen sample had good results for the alignment. However, when comparing the actual numbers, it becomes evident that in amount of reads mapped, Norgen is actually slightly worse than the two others. Nevertheless, all three samples actually had a similar amount of reads mapped in pure numbers. The Norgen sample is however closer to the combined sample than the QIA-micro sample. The numbers fit somewhat with the number of total reads, but in that regard QIA-micro and NorgenClean is closer toward each other than the Norgen sample. In both of these occasions there is just a few kb in between the closest methods, while the difference is an order of magnitude higher between the furthest samples. Most of the reads for each sample mapped uniquely, and less than ten percent was non-uniquely mapped for each of them.

Table 11 – Total reads and alignment results for each of the samples after being mapped by BSMAP. Calculated bisulfite conversion based on the methylation ratio is also given.

	RA5111B CD4$^+$ Norgen	RA5111B CD4$^+$ QIA-micro	RA5111B CD4$^+$ NorgenClean
Total reads	4136001	4904377	4989671
Aligned reads	3057055 (73.9%)	3276122 (66.8%)	3073749 (61.6%)
Unique reads	2650782 (64.1%)	2843801 (58.0%)	2729827 (54.7%)
Non-unique reads	406273 (9.8%)	432321 (8.8%)	343922 (6.9%)
Bisulfite conversion ratio (%)	99.93841	99.99816	99.99844

The bisulfite conversion ratio was calculated for each of the sequenced samples (Table 11). All three samples showed good conversion ratios, with the NorgenClean sample having the best result, followed by QIA-micro before Norgen.

Exploratory analysis of selected genes

The total number of CpG positions identified was dropping drastically as the coverage increased (Figure 25). Already at 5 times coverage there was a serious reduction in amount of sites identified as compared to the total amount of sites. At 10 and 20 times coverage, the amount was too low to really give any meaningful results. The Norgen sample seems to have even slightly less coverage of CpG sites than the two other samples when increasing beyond a coverage of one.

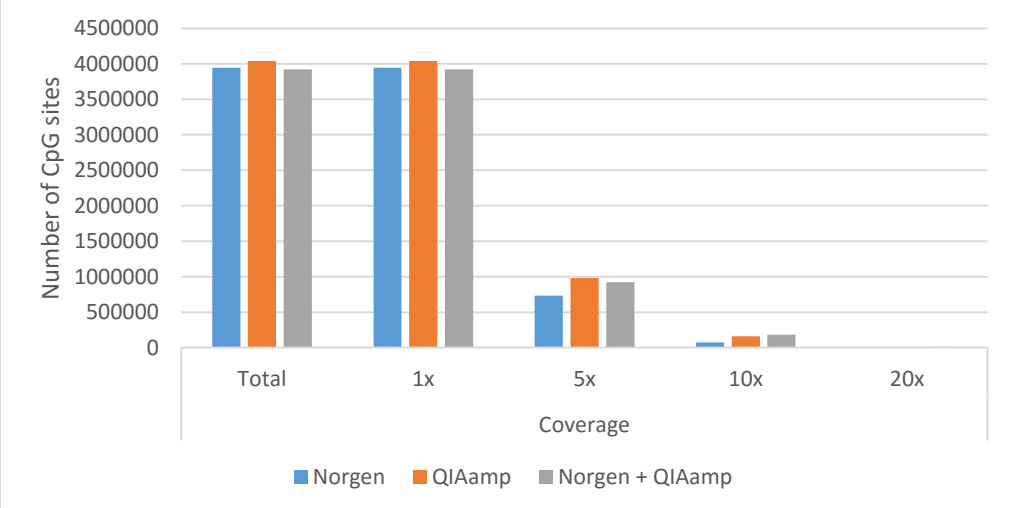


Figure 25 - Total number of CpG sites identified at different levels of coverage for each of the extraction methods. The coverage level stated in the figure is the minimum coverage, i.e. all coverage levels above the one stated is also included for each entry.

A search for CpG sites 5000bp upstream of each of the selected RA genes at minimum five times coverage, was performed for each of the three extraction methods (Figure 26). Of the total of 40 genes, 15 had no CpG sites which could be identified in any of the samples. 16 of the genes had at least one CpG site identified in each of the samples. The number of methylation sites identified was in the same order of magnitude between the samples isolated with the different DNA extraction methods.

It is evident that although there is a great variation in the amount of CpG sites between the genes, there at least seems to be some agreement between the methods number of methylation sites upstream of a certain gene (Figure 26). The QIA-micro and NorgenClean samples does in general seem to be more in agreement with each other than with the Norgen sample for amount of identified sites per gene. Norgen also tend to report fewer sites than the other methods. It is also interesting to note that on a few occasions, NorgenClean is the one not being in agreement with the two others with regard towards the amount of identified CpGs, however, this is almost never the case with QIA-micro.

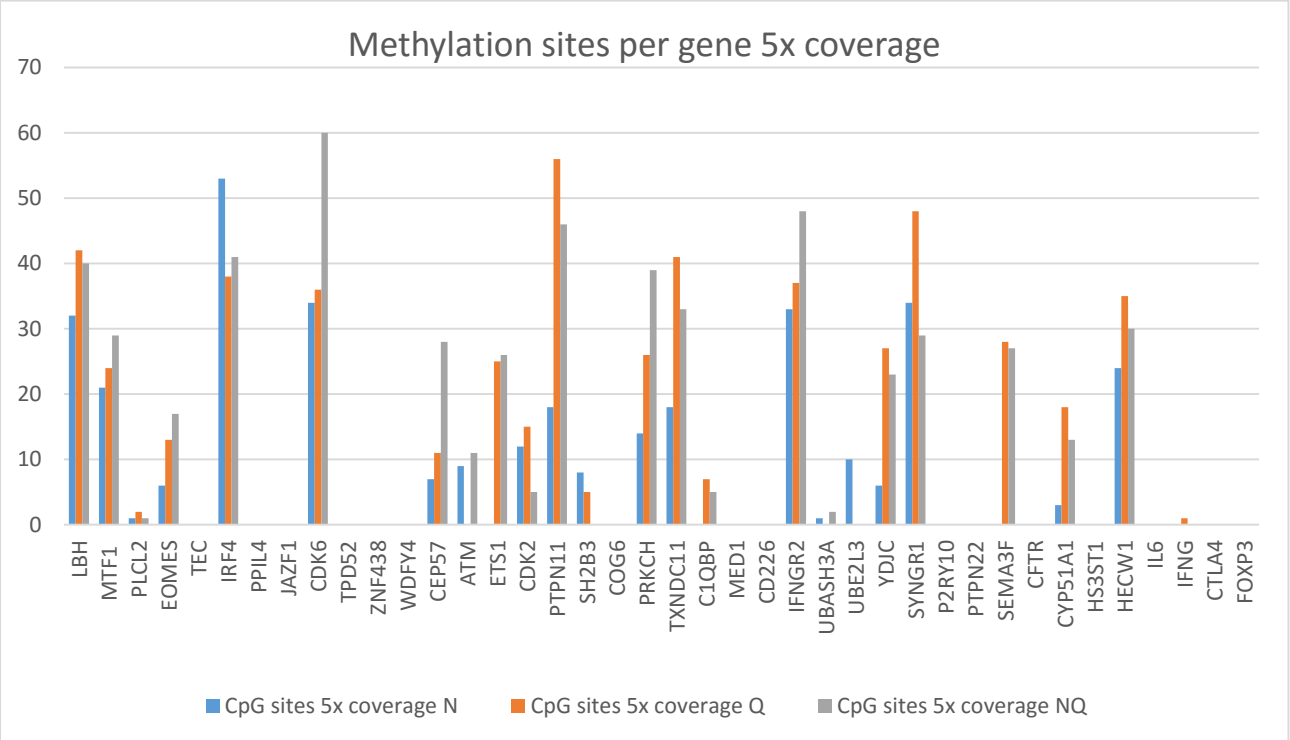


Figure 26 - Number of methylation sites 5000bp upstream of each of the selected genes at 5x coverage for each of the three samples.

The mean methylation 5000bp upstream of all the genes with at least 1 identified CpG site at 5x coverage in all three sequenced samples were plotted, for a total of 17 genes including

SEMA3F, although the Norgen sample could not identify any CpG sites upstream of this gene (Figure 27). As with the identification of CpG sites, the Norgen sample was less in agreement with the other two samples. In general, Norgen seems to be reporting a higher degree of methylation than the two other samples. However, as before, the identified magnitude of methylation is for the most part comparable between the samples (Figure 27).

A general trend seems to be that there is a correlation between an increase in expression and an increase in methylation (Figure 27). However, note that there are also several highly expressed genes with a low methylation level. The only two highly methylated genes identified were at each end of the expression scale (Figure 27).

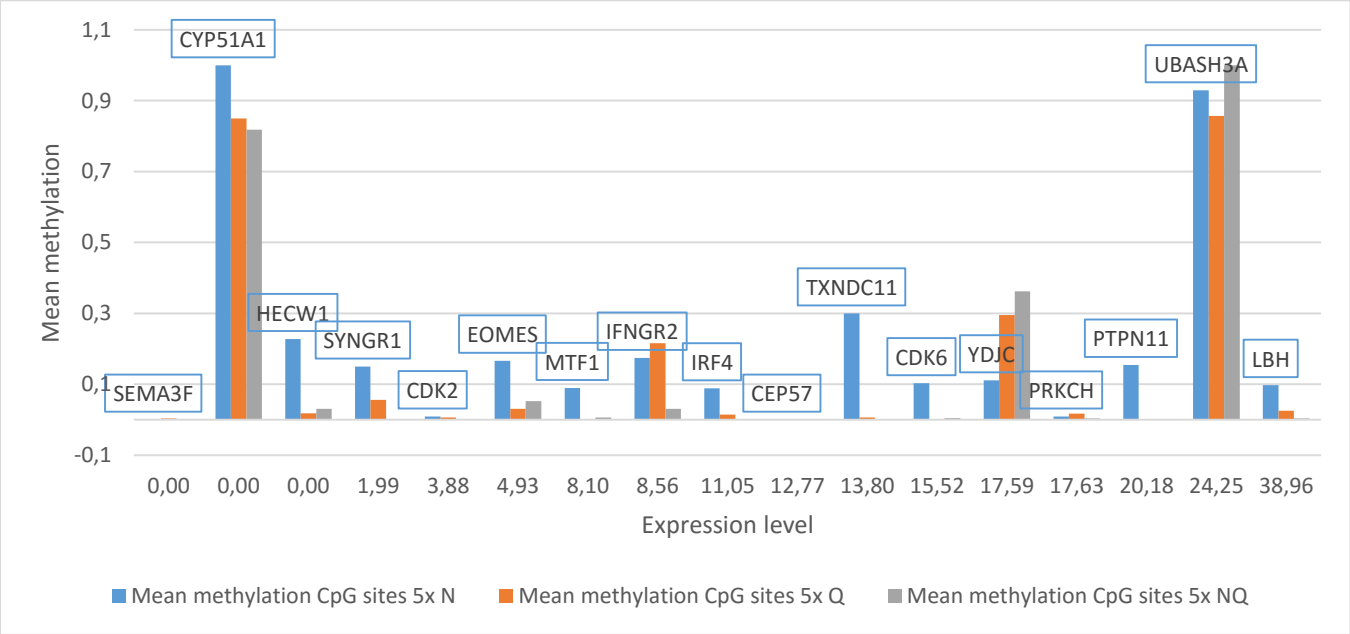


Figure 27 – Mean methylation level plotted against the expression level for all three sequenced samples. All selected genes which had identifiable CpG sites for each extraction method are included (except SEMA3F, which did not have any CpG sites identified in the Norgen sample).

A few randomly selected chromosomes with methylation data from a varying amount of genes at five times coverage was plotted against the expression data for CD4 T-cells (Figure 28). Noteworthy, these expression data are from different individuals. Only the data from NorgenClean is shown here, as these are representative, the plots for the other two samples can be seen in appendix 3. Methylation versus expression plots were created for all of the chromosomes, but only plots for chromosome 1, 8, 18 and X is shown in this thesis. The weak trend seems to be that, as expected there is a high degree of methylation in genes with

low expression, and a lower degree of expression for the genes with higher expression. However, there is also a great variability in the correlation between the individual genes. Furthermore, all of the samples show a high degree of clustering along the line of full and no methylation. There are also some outliers with very high gene expression as compared to most of the genes. It is however reassuring to see that these genes are generally not methylated.

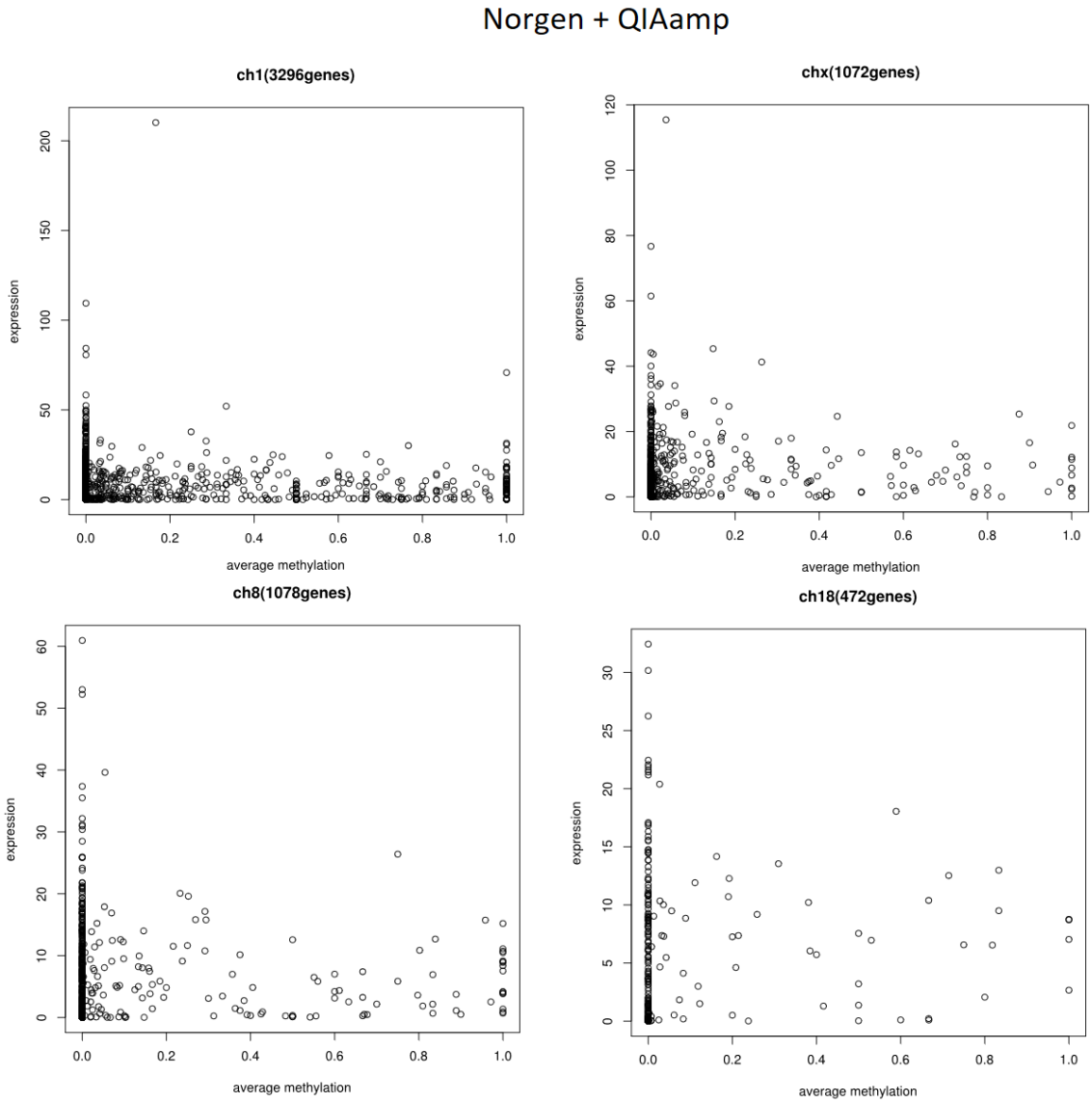


Figure 28 - average methylation ratio plotted against expression for a number of genes, only the NorgenClean sample is shown, the two other methods have comparable results and can be seen in and . Methylation ratio is plotted against the x-axis, and Expression against the y-axis. Note that the expression values vary between the graphs.

All three extraction methods seems to give more or less equal results in regards of methylation versus expression in the grander scheme (Figure 28). None of the samples differs to an obvious degree from the others in these plots.

Discussion

During the work with establishment of a mRRBS protocol in this thesis, several DNA extraction methods were tested. In order to get all details in place, several pilot experiments for the mRRBS procedure was also performed. In the end, one sample with DNA extracted by three different methods, which successfully went through the mRRBS procedure, were sequenced on the Illumina MiSeq system.

The QIAamp DNA micro kit served as a gold standard and the Norgen extraction procedures was compared against this to investigate whether it would be possible to use DNA extracted through a procedure simultaneously extracting RNA. From the sequencing results, parameters such as total number of reads, number of isolated CpG sites and methylation per gene were studied. The bioinformatic analyses pointed towards the QIA-micro and NorgenClean samples giving results which are closer to each other than the Norgen samples, which turned out to generally have a lower quality. The conclusion is that the QIAamp DNA micro kit gives the best results, but that the Norgen kit followed by an additional clean up step provided comparable results, enabling this procedure to be used for further studies due to the advantage of being able to simultaneously isolate RNA and protein from the same sample.

Choice of DNA extraction method

Five different methods for extracting DNA have been tested: a manual protocol, QIAamp DNA mini kit, QIAamp DNA micro kit, Norgen RNA/DNA/Protein plus kit in addition to a combination of the Norgen kit with a cleanup from the QIAamp micro kit. The kits were tested due to the manual protocol yielding DNA with visible contaminations, spotted through a brown tint on the sample. After use of the QIAamp DNA micro kit, it was visually confirmed that the spin columns got rid of this contamination. The QIAamp DNA micro kit was shown to give extracted DNA with high enough concentrations and good enough quality for mRRBS. The Norgen kit did for the initial extraction test not have good enough quality, although it improved with a cleanup on the QIAamp kit, but the concentration was reduced. Nevertheless, they were tested further through the complete process and actually showed a great improvement after the mRRBS procedure.

During the initial DNA extractions the only method giving samples with sufficient qualities for further use was the QIAamp DNA micro kit. However, we aimed to optimize the Norgen

kit as this also isolated RNA from the same samples, and as such both those, and the samples using both methods were brought along for the mRRBS procedure. This procedure actually cleaned up all the samples, and all three methods had samples with both high enough quality and yield of DNA for sequencing on the MiSeq. This is not surprising however, considering there are several cleanup steps during the mRRBS procedure, for example after the bisulfite conversion.

As the coloring indicating contamination was only seen in the positively selected CD8, and not the negatively selected CD4 samples of the manually extracted DNA, it was hypothesized that it was due to the magnetic particles from the EasySep kits used to isolate the cells, attaching to the DNA. This hypothesis could be reinforced by the fact that the DNA yield was about 2.2 times higher from the CD4 cells than the CD8 cells, possibly due to the particles attaching the DNA in the CD8 sample. The samples did however have the best concentrations achieved of all the DNA extractions performed. The nanodrop values were also within the acceptable levels, at between 1.8 and 2.0 for the 260/280-values and 1.8 and 2.2 for the 260/230-values. A description of how nanoparticles bound to the cell surface can be engulfed in a cell through endocytosis is given by Nimesh and Chandra (2011). As the downstream implications of the beads from the cell isolation hanging on to the extracted DNA was unknown it was decided that other DNA extraction experiments should be conducted in order to find a better suited method of DNA extraction from the cells.

At least 20ng DNA/µL was needed in order to proceed with the mRRBS procedure (Boyle et al. 2012). The QIAamp DNA mini kit using spin columns was tested for extracting DNA while also filtering out the magnetic particles. The columns worked as intended, as visually confirmed by the darker tint in the filter used for extracting the CD8 samples as compared to the CD4 samples. After testing on cells stored either on RNAprotect cell reagent or as dry pellets, as well as different DNA elution volumes, the kit was determined to give DNA with acceptable quality with 260/280-values at about 1.8. The 260/230-values were slightly high at above 2.6, but still deemed acceptable. The concentrations were also good enough. Even better quality DNA was attained by using the QIAamp DNA micro kit, and especially when combining with a manual lysis step. The 260/280-values were equal to that of the mini kit, but it also gave improved 260/230-values, that although varying, were generally lower than for the mini kit. The manually extracted samples was attempted cleaned up on both mini and

micro columns, but the quality dropped too much for the mini, and the concentration dropped to much for the micro kit.

During the DNA extraction test with the micro kit on the dry pelleted cells, the cells were thawed by the addition of 100µL RNAprotect cell reagent. They were then split into 50µL and the aliquot used further was diluted with 50µL buffer ATL, totaling to 100µL. They should, however, have been pelleted once more before adding 100µL buffer ATL in order to get the sample as equal as possible to the other samples stored on RNAprotect cell reagent.

The reasoning for testing the Norgen RNA/DNA/Protein Purification Plus kit even though the QIAamp DNA micro kit had proven to give DNA of sufficient quality for the downstream applications was that this kit would also be able to isolate RNA (as well as protein) from the same sample as the DNA. This would be beneficial for performing parallel studies based on the same samples. Through this additional sample material, expression levels could be studied based on mRNA analyses. In addition several regulatory mechanisms, like methylation and microRNA could also be studied, enabling a comprehensive overview of the changes in gene regulation in the RA cohorts upon methotrexate treatment. The Norgen samples did not provide DNA with sufficient quality. Although only one of the samples had a concentration that was too low, the quality as indicated by the 260/230-values were unacceptable (<0.6 for all samples), and only one sample had a 260/280-value above the required 1.8. Adding a cleanup step using QIAamp micro spin columns improved the quality, but reduced the concentration to beneath 20ng DNA/µL. When trying to up concentrate the samples, the quality dropped while the concentration was still too low. Even though the quality from neither methods involving the Norgen kit was acceptable for further work, it was still decided to try to take them through the mRRBS procedure. As will be described in further detail below, this actually helped improve the quality greatly for the samples.

Parameter testing and quality control of multiplexed Reduced Representation Bisulfite Sequencing libraries

Four complete pilot experiments were performed for the mRRBS procedure, but the quality of the results were varying. Although most of the samples in pilot IV got an expired polymerase, and as thus were unsuccessfully amplified, the three samples with the correct polymerase were successful and were sequenced on the MiSeq. These samples represented one of the three DNA extraction methods each: Norgen, QIAamp DNA micro kit and Norgen

combined with a QIAamp DNA micro kit cleanup. All three methods ended up having good concentrations and correct fragment sizes, however there was an abundance of primer dimers present. Through an extra cleanup step using AMPure XP beads, the primer dimers were reduced, while the samples still had good concentrations and nanodrop quality measures within accepted values.

19 hours of MspI digestion was identified as the optimal time for getting the largest specter of fragment sizes. The samples did not acquire a completely even distribution of fragment sizes, and a higher amount of smaller, as opposed to larger fragments seemed to be present. Nevertheless, there was a clear difference between each time-point in regard to the amount of digestion, where 19 hours covered fragment sizes over a larger area than 17.5 hours, which in turn covered a larger area than 21 hours, none of the samples seemed to have fragments with sizes in the area from about 100bp to 400bp. Gu et al. (2011) provided a gel image containing human genomic DNA digested by the MspI fragment. Their image showed more of an even distribution of fragment sizes, with a slight gradient from larger amount of bigger fragments and less amount of smaller fragments. Some differences between their and our images are to be expected, however, as we used tapestation D1000 gels and they used a 4-20% Criterion precast polyacrylamide TBE gel.

The final libraries of successfully bisulfite converted DNA had acceptable fragment sizes from about 200bp to 500bp. Considering that the gel images from our samples were comparable to a gel image of a final converted library provided by Gu et al. (2011), the MspI digestion was clearly sufficient despite the problems with the gels depicting the restriction fragments described above.

In pilot III, there was a delay between the addition of bisulfite mix and DNA protect buffer. This could potentially have damaged the samples. The actual effect of this is unknown, but there is a possibility that it could have downstream effects. It was, however, reassuring that the color of the DNA protect buffer still changed as it was supposed to, even though the samples might still have been affected. Precipitates were later observed in the sample, which could potentially be because of this. The precipitates themselves should not be a problem according to the EpicTect bisulfite kit protocol. After the completion of the bisulfite conversion, a dry pellet was also observed on the wall of the well in the QIA-micro RA5111B CD8 sample. This could potentially have led to a reduced effect of the conversion.

However, no proof of this hypothesis could be found, as one of the two final amplifications of QIA-micro RA5111B CD8 gave no results, while the other was the one with the best result in the final library.

The two CD8 Control sample I samples of mRRBS pilot I had DNA present at above 2 and 8 ng/µL, and the gel images showed fragment sizes at about 200bp and higher. As described above this was the fragment sizes wanted, although they also contained large fragments up to 1500bp. Considering each sample had at least 35µL, and according to Illumina the minimum input of DNA to the MiSeq is 10ng (Illumina inc 2016b), the samples were actually abundant enough for sequencing. However, the samples were not useable as the protocol differed in many aspects from the later experiments, such as the cleanup step in the middle of the conversion process, and following a different bisulfite conversion protocol. As a result, these samples were not comparable to the later samples, and no sequencing was performed.

There was either no detectable DNA or a too low amount to be usable left in the final bisulfite converted libraries from pilot II. This was probably due to the probable error in concentration, at 30% instead of 70% ethanol, during the washing step, which in turn could have led to the DNA having been eluted and thrown away together with the ethanol.

During pilot IV there was an accidental pooling of two samples, QIA-micro RA5516B CD8 and Norgen RA5516B CD8. The two Norgen and QIA-micro RA5516B CD4 samples were also involved in a potential cross-contamination. None of these samples got DNA concentrations high enough for sequencing, so the potential problems were no longer an issue. The potential cross contamination did, however, include two more samples, QIA-micro and Norgen RA5111B CD4. These both had a high concentration of DNA present, and were using the same adapters. However, as they were equally likely to be contaminated with the two RA5516B CD4 samples as with each other, a search was performed for the adapter sequences of both samples in both the sequenced samples. No evidence for the cross-contamination could be found, and as thus the assumption that there was no cross-contamination between any of the samples was made. The slight occurrence of adapters from the other samples sequenced together were neglible at only about 0.007% and less, and is probably explained by the fact that these were the adapters used in the samples pooled together for the sequencing.

RRBS are, together with whole genome bisulfite sequencing (WGBS) and arrays such as the 450K BeadChips, among the most used methods for methylation analysis, although several other methods also exist (reviewed by (Plongthongkum et al. 2014)). Arrays are cheap and provide high throughput, and as such are extensively used for analysis. However, the general problem with cross hybridization will apply (reviewed by (Plongthongkum et al. 2014)). A weakness with both arrays and the RRBS method, compared to WGBS, is the reduced information about methylation in CpG poor regions (reviewed by (Plongthongkum et al. 2014)). However, as mentioned in the introduction, 60-70% of CpG islands overlaps with the promoter regions of genes, and as thus are most likely to be informative. Although WGBS gives an unbiased coverage of all genomic regions, RRBS is a sensible choice as it is a cheaper (about 10 times reduced costs) option while also enriching interesting areas, such as the aforementioned promoter regions (Boyle et al. 2012). mRRBS did in other words provide a good compromise between price and genomic coverage.

MiSeq Sequencing quality

The per base quality of the MiSeq reads was very good with quality scores above 30, meaning no more than one erroneous call in a 1000 bases read, until somewhere around 70-80bp. None of the reads were unable to identify any positions as one of the four standard bases. The proportion of each base is also skewed as expected with more cytosines and less thymines until somewhere around 80bp.

For the first 75bp of each read, which was the amount mapped to the reference genome, the quality for each base read was very good in the samples with provided reports. As the base content was also altered with a higher amount of thymines, and less cytosines in about the same amount of bases, in addition to there being no bases called as something different than the four standard bases, it is safe to assume that the sequencing was properly calling each base. Further confirmation can be found in the clear MspI signal in the first three bases of each read.

The sequencing was performed as paired-end with a read length of 300 bp. This was due to economical and time considerations, as the test could be performed on a smaller sequencer, the MiSeq. The data was treated as 75bp single-end reads, however, as this was the standard setting for RRBS and will be performed in further experiments on a HiSeq instrument. A

reason for this is that there is no large gain in number of CpGs identified in read lengths longer than 75bp (Smith et al. 2009).

Mapping and bisulfite conversion ratios

The Norgen sample was found to have fewer total reads than the two other samples, but the amount of reads aligned to the reference genome was approximately the same for all three samples. All three extraction methods also showed a good bisulfite conversion ratio at above 99.9% for each of the methods.

Of the three extraction methods used for the samples that were sequenced, all had bisulfite conversion ratios at >99.9% each. However, Norgen had the lowest ratio of the three, while the two others had an almost equal conversion rate to each other. Our achieved conversion ratios are on par with the 99% conversion ratios reported by Boyle et al. (2012), and better than the conversion ratio reported by Leontiou et al. (2015) at 98.4% when using the same bisulfite conversion kit as in this thesis.

The conversion ratio was calculated based on the fact that non-CpG cytosines should all be unmethylated, and thus converted. However, due to the fact that the mRRBS procedure was performed, the calculation of the bisulfite conversion ratio could also be done in another manner. Namely through the use of the restriction site and the fact that the gap created needs to be filled with a nucleotide known to be unmethylated. By identifying sequences with adapters attached to the 3' end, the position with a filled in cytosine nucleotide could be studied, and the percentage of conversion could be calculated (Babraham Bioinformatics 2013). This method does, however, not seem to be used by the compared literature, and therefore the first approach was chosen.

All three samples had an alignment of at least 60%, where at least 50% was uniquely mapped. This is comparable with the 76bp alignments of human embryonic stem cells performed by Xi et al. (2012) at between 50-55%.

CpG coverage and exploratory gene analysis

Approximately the same amount of CpG sites were identified from the QIA-micro and NorgenClean samples, both when looking at single genes, and overall coverage. The Norgen

sample, on the other hand, tended to underestimate the amount of CpG sites compared to the two other samples, although there were exceptions when looking at single genes. The overall sequencing coverage of each sample was also low, and the number of CpG sites identified in each sample was drastically reduced as the coverage thresholds were increased.

Average methylation values 5000bp upstream of selected genes associated with RA were plotted against the expression of the same genes in healthy adult CD4 cells. The QIA-micro and NorgenClean were for the most part, in agreement regarding the amount of methylation. The Norgen sample, on the other hand, tended to overestimate the methylation levels compared to the other methods. This was, however, no longer evident when comparing across a high number of CD4 expressed genes. For the plot with the selected RA related genes, there was no clear connection between methylation and expression levels. For the plot including all the CD4 expressed genes, there was a slight indication of high methylation in low expressed genes, and low methylation in highly expressed genes.

The region 5000bp upstream of each of the selected genes at minimum five times coverage were chosen for the gene analysis, and CpG positions in this region were extracted. The overall coverage was too low to get reliable results from a minimum of ten times coverage, and as such the minimum of five times coverage was chosen for further use. Boyle et al. (2012) has shown that the correlation between the same samples during different runs will be high when at least five times coverage is demanded. It also gets noticeably better, although less drastically, at 10-15 times coverage.

There is no general consensus for how long the promoter regions of genes are. Some RA studies looking into specific genes have defined areas 1200bp upstream of IL-6 (Ishida et al. 2012; Nile et al. 2008) and another study looked into the area 2000bp upstream of the transcription site of CXCL2, although they were focusing on the area at 741bp upstream (Karouzakis et al. 2011). Outside of RA, one study defined 966bp upstream of TNFSF7 (Lu et al. 2005). Most of these studies only defined tens of base pairs downstream as part of the promoter region, and at most a few hundred. However, these are just a few studies on specific genes. In a study identifying SNPs in promoter regions globally in the human genome, the area 5000bp upstream and 500bp downstream of the transcription start site was identified as the promoter region (Kim et al. 2008). As we were also looking at the total effect of the DNA methylation of promoter regions in the isolated T-cells, we chose to define the area 5000bp

upstream of each start codon for our search for CpG sites. This interval should cover the promoter of most genes, although it would probably have been preferable to also include a small region downstream of the start codon. This is however probably not a major drawback for this study considering the low coverage will probably give a larger effect.

The number of CpG sites identified 5000bp upstream for each start codon varied greatly between the genes. This is, however, to be expected as these are different genes, and as such there is no reason as to why they should have an equal amount of CpG sites. Technical aspects could also be responsible for this difference, such as varying degree of targeting due to the restriction site in our reduced representation, our low sequencing coverage or potential sequencing errors. Another technical aspect is that the sequencer would have problems reading the first three bases of each read. This was due to them always being the same because of the restriction pattern of MspI. In this pilot study, the problem was tackled by adding 50% PhiX. In later studies, dark sequencing as described by Boyle et al. (2012) will be used instead, which means that the signal from the first three rounds of sequencing is not recorded. This will probably lead to an increase in the coverage as well, as 50% of the capacity was lost when the PhiX was added together with the samples in this study.

In our selection, the amount of CpG sites varied from 0 to somewhere around 50 identified sites, depending on how the count is performed, e.g. highest or average number. In the review by Cribbs et al. (2015), a few specific genes such as IL6 or CD40L, are mentioned as having known differential methylation in RA patients. However, apart from IFNG where a single CpG site was discovered in the QIA-micro sample, no CpG sites were identified for any of these genes in this study. As a consequence, no further analysis was possible to perform for these specific genes.

The mean CpG methylation level 5000bp upstream of each of the genes with at least one CpG position identified for each sample were plotted against gene expression with the data provided by Helgeland et al. (unpublished data). It is of importance to note that the gene expression data is an average expression from CD4 T-cells in blood from healthy human adults, while the methylation data is from only one female patient newly diagnosed with RA. Nevertheless, an exploratory analysis was performed. Two plots were created, first, one for the specifically selected genes gathered by comparing the 42 RA risk loci by Okada et al. (2014) with the expression data, as well as a few non expressed genes for comparison.

Second, all of the genes expressed in the CD4 cells were plotted by each extraction method and chromosome.

In the first comparison, just as described above for both the bisulfite conversion ratio, and the number of CpG sites identified, the Norgen sample once again was in less agreement with the two other samples than they were with each other. In general, for the 17 selected genes, the Norgen sample showed overall higher average methylation for most of the genes compared to the samples from the two other methods. There also seemed to be a slight trend towards an increase in the average methylation level as the expression increased. This does, however, seem counterintuitive as the general consensus is that methylation blocks transcription, and as thus would be expected to lower the expression. Possible explanations for this could be low sample size of genes, low coverage, and most importantly that the expression levels are from different individuals than the methylation data. An important note is also that although there is seemingly a correlation between the increase of the two levels, one gene, UBASH3A has a high expression and almost complete methylation, which somewhat hides the three surrounding genes with very low methylation. Also note that UBASH3A only had a few CpG sites identified. This gene has been found to have an increased methylation level in one CpG site in obese individuals (Wang et al. 2010). However, in breast cancer low methylation together with high expression in the same gene has been fund to increase lymphocyte infiltration (Dedeurwaerder et al. 2011).

It is interesting that of the genes with no expression from RNA sequencing, only one of the three genes showed a high degree of methylation, and there are no clear differences in methylation between the genes with high and low expression. Had the coverage been better, it would perhaps have been possible to identify more CpG positions, and have more precise readings of the methylation levels upstream of each of the genes. Interestingly, about the same coverage of CpG positions as Boyle et al. (2012) at a coverage of one was acquired. However, at a minimum of five times coverage, in this study, between 750000 and one million CpG positions was identified, depending on the sample, while they identified about 1.5 million positions at five times coverage exclusively. They also retained at least 500000 CpG islands at 10 times coverage exclusively, while our samples were dwindling even lower at somewhere from 250000 sites and below when this was the minimum coverage. Even though it was not possible to establish a connection between the expression and methylation

in the plot for the specifically selected genes, it is important to remember that methylation is not the only factor influencing the expression of genes.

Interestingly, when just looking at the general trend from an increased amount of expressed genes, the expected methylation levels of high degree of methylation in lowly expressed genes, and vice versa occurs. There are also some outliers with very high expression, and these are almost exclusively lowly methylated. These findings can confirm that the reason for the lack of expected correlation between methylation and expression in the first plot of only 17 genes was in fact due to a small sample size. There are some clustering along the line of full, and no methylation in the second plots. This could be due to the low coverage, leading to many genes having identified only a single CpG site, the average methylation value would then naturally be either of these options. It should however not be excluded that some of these genes truly could be fully methylated or unmethylated. This will hopefully be clarified by further experiments.

Future mRRBS analysis in RA

In this thesis, a methylome profile of a patient sample extracted by three different methods has been successfully created. Based on the results, it was decided that the extraction based on a combination of the Norgen with QIAamp clean up yielded the results with the best compromise between quality and expanded possibilities for analysis through additional extraction of RNA and protein from the same samples. However, the coverage achieved from the sequencing was not satisfactory, and lead to problems with further analysis. It was however possible to do some preliminary analyses. Future studies will get a higher coverage as a result of both transitioning from the MiSeq to the HiSeq system, as well as performing dark sequencing instead of a 50% PhiX spike in. An increase in coverage is important in order to report reliable, reproducible methylation patterns for each CpG position and/or promoter region.

Future studies will create large-scale methylome profiles of RA patients and controls based on the results from this thesis. It will for these studies be interesting to see what the correlation between methylation and expression from the same samples looks like. It will also be interesting to look at how the methylation levels of RA patients compares to controls, as RA T-cells have earlier been shown to have reduced methylation compared to healthy

controls (Richardson et al. 1990). This would be interesting both in the large scale as well as in specific genes, especially the ones with a known correlation to RA. As mentioned in the introduction the studies building on the experiences made through the work on this thesis will provide novel insights due to the isolation of specific T-cell subsets, i.e. $CD4^+$ and $CD8^+$ cells, while still providing single base resolution. The reduced methylation in RA T-cells mentioned above lacks this resolution as it was only looking at the total methylation ratio in the T-cells by HPLC, and not through sequencing (Richardson et al. 1990). Most other studies have not isolated T-cells at all. For example, both the study by Nile et al. (2008) and Liu et al. (2013) were using PBMC, although the latter study did use an algorithm for estimating cell type proportions afterwards. Nevertheless, there is at the moment a lack of studies performed on isolated T-cells and their subtypes. The studies to follow this thesis will hopefully help illuminate the role of DNA methylation in T-cells for RA patients.

Conclusion

After a series of experiments, it was concluded that the QIAamp DNA micro kit was able to both get rid of bead contaminants from the cell isolation, while still providing a good concentration and quality of samples. The Norgen DNA/RNA/Protein purification plus kit did not manage to give as good results, neither did it initially give any better quality or concentrations when combined with a cleanup from the QIAamp DNA micro kit. However, after having taken samples from all three extraction methods through the whole mRRBS procedure, it is evident that the process itself further cleans up the samples. As a result, the quality and concentration is sufficient in the last step before sequencing, even though the samples did not seem good enough to be taken into the procedure from the start. After having sequenced a sample from each of the three DNA extraction methods, it was evident that the Norgen extraction method alone differed from the two other methods, and in general delivers worse or different results. A combination of the Norgen and QIAamp method was shown to provide results on par with our gold standard, the QIAamp procedure alone, and in general they seemed to be more or less in agreement with each other. Based on these results, it has been decided that the following studies will be using the combined DNA extraction method when preparing the samples.

Although the coverage was low, some preliminary exploratory analyses were performed. The coverage problem will be improved for further studies by the use of the HiSeq system combined with the dark sequencing method. It was not possible through our analyses to identify any obvious patterns at individual genes, but to some degree the expected methylation pattern of high methylation in genes with low expression, and low methylation in genes with high expression was observed. Regardless of the coverage, very good bisulfite conversion ratios at above 99.9% for each sample were demonstrated.

Through this thesis, the foundation for further large scale methylome profiling studies of RA patients has been prepared through demonstrating a complete methylome sequencing from the isolation of T-cell subsets, extraction of DNA and conversion through mRRBS.

References

Aho, K., Koskenvuo, M., Tuominen, J. & Kaprio, J. (1986). Occurrence of rheumatoid arthritis in a nationwide series of twins. *The Journal of rheumatology*, 13 (5): 899-902.

Aletaha, D., Neogi, T., Silman, A. J., Funovits, J., Felson, D. T., Bingham, C. O., 3rd, Birnbaum, N. S., Burmester, G. R., Bykerk, V. P., Cohen, M. D., et al. (2010). 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Ann Rheum Dis*, 69 (9): 1580-8.

Arnett, F. C., Edworthy, S. M., Bloch, D. A., McShane, D. J., Fries, J. F., Cooper, N. S., Healey, L. A., Kaplan, S. R., Liang, M. H., Luthra, H. S., et al. (1988). The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum*, 31 (3): 315-24.

Babraham Bioinformatics. (2013). *Reduced Representation Bisulfite-Seq - A Brief Guide to RRBS*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/bismark/RRBS_Guide.pdf (accessed: 17.08).

Ballestar, E. (2011). Epigenetic alterations in autoimmune rheumatic diseases. *Nat Rev Rheumatol*, 7 (5): 263-71.

Berard, M. & Tough, D. F. (2002). Qualitative differences between naïve and memory T cells. *Immunology*, 106 (2): 127-138.

Bicker, K. L. & Thompson, P. R. (2013). The protein arginine deiminases (PADs): Structure, Function, Inhibition, and Disease. *Biopolymers*, 99 (2): 155-163.

Blom, B. & Spits, H. (2006). Development of human lymphoid cells. *Annu Rev Immunol*, 24: 287-320.

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

Brown, J. H., Jardetzky, T. S., Gorga, J. C., Stern, L. J., Urban, R. G., Strominger, J. L. & Wiley, D. C. (1993). Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature*, 364 (6432): 33-9.

Cope, A. P. (2008). T cells in rheumatoid arthritis. *Arthritis Research and Therapy*, 10 (1): S1.

Cribbs, A., Feldmann, M. & Oppermann, U. (2015). Towards an understanding of the role of DNA methylation in rheumatoid arthritis: therapeutic and diagnostic implications. *Ther Adv Musculoskelet Dis*, 7 (5): 206-19.

Cusick, M. F., Libbey, J. E. & Fujinami, R. S. (2012). Molecular Mimicry as a Mechanism of Autoimmune Disease. *Clinical reviews in allergy & immunology*, 42 (1): 102-111.

de Graeff-Meeder, E. R., Voorhorst, M., van Eden, W., Schuurman, H. J., Huber, J., Barkley, D., Maini, R. N., Kuis, W., Rijkers, G. T. & Zegers, B. J. (1990). Antibodies to the mycobacterial 65-kd heat-shock protein are reactive with synovial tissue of adjuvant arthritic rats and patients with rheumatoid arthritis and osteoarthritis. *The American Journal of Pathology*, 137 (5): 1013-1017.

Dedeurwaerder, S., Desmedt, C., Calonne, E., Singhal, S. K., Haibe-Kains, B., Defrance, M., Michiels, S., Volkmar, M., Deplus, R., Luciani, J., et al. (2011). DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO Molecular Medicine*, 3 (12): 726-741.

Dong, C. & Marinez, G. J. (2010). T-cells: the usual subsets. *Nature Reviews, immunology*.

Duke, O., Panayi, G. S., Janossy, G. & Poulter, L. W. (1982). An immunohistological analysis of lymphocyte subpopulations and their microenvironment in the synovial membranes of patients with rheumatoid arthritis using monoclonal antibodies. *Clinical and Experimental Immunology*, 49 (1): 22-30.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2014). Ensembl 2014. *Nucleic Acids Research*, 42 (D1): D749-D755.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2016). *Human assembly and gene annotation*. Hinxton, UK: Ensembl. Available at: http://grch37.ensembl.org/Homo_sapiens/Info/Annotation#assembly (accessed: 06.05).

Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ballestar, M. L., Heine-Suñer, D., Cigudosa, J. C., Urioste, M., Benitez, J., et al. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (30): 10604-10609.

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

Gardiner-Garden, M. & Frommer, M. (1987). CpG islands in vertebrate genomes. *J Mol Biol*, 196 (2): 261-82.

Gregersen, P. K., Silver, J. & Winchester, R. J. (1987). The shared epitope hypothesis. an approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis & Rheumatism*, 30 (11): 1205-1213.

Gu, H., Bock, C., Mikkelsen, T. S., Jager, N., Smith, Z. D., Tomazou, E., Gnirke, A., Lander, E. S. & Meissner, A. (2010). Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nat Meth*, 7 (2): 133-136.

Gu, H., Smith, Z. D., Bock, C., Boyle, P., Gnirke, A. & Meissner, A. (2011). Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat. Protocols*, 6 (4): 468-481.

Hayatsu, H., Wataya, Y., Kai, K. & Iida, S. (1970). Reaction of sodium bisulfite with uracil, cytosine, and their derivatives. *Biochemistry*, 9 (14): 2858-2865.

Hotchkiss, R. D. (1948). The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *Journal of Biological Chemistry*, 175 (1): 315-332.

Howlett, S. K. & Reik, W. (1991). Methylation levels of maternal and paternal genomes during preimplantation development. *Development*, 113 (1): 119-27.

Huizinga, T. W., Amos, C. I., van der Helm-van Mil, A. H., Chen, W., van Gaalen, F. A., Jawaheer, D., Schreuder, G. M., Wener, M., Breedveld, F. C., Ahmad, N., et al. (2005). Refining the complex rheumatoid arthritis phenotype based on specificity of the HLA-DRB1 shared epitope for antibodies to citrullinated proteins. *Arthritis Rheum*, 52 (11): 3433-8.

Illingworth, R. S. & Bird, A. P. (2009). CpG islands--'a rough guide'. *FEBS Lett*, 583 (11): 1713-20.

Illumina Inc. (2014). *Understanding Illumina Quality Scores*. San Diego, USA: Illumina, Inc. Available at: http://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_understanding_quality_scores.pdf (accessed: 01.04).

Illumina Inc. (2015). *MiSeq System. Focused power. Speed and simplicity for targeted resequencing and small-genome sequencing.* San Diego, USA: Illumina, Inc. Available at: http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_miseq.pdf (accessed: 06.05).

Illumina inc. (2016a). *An Introduction to Next-Generation Sequencing Technology*. San Diego, USA: Illumina, Inc. Available at: http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf (accessed: 10.04.2016).

Illumina inc. (2016b). *MiSeq System*. San Diego, USA: Illumina inc. Available at: http://www.illumina.com/systems/miseq.html (accessed: 01.05).

Ishida, K., Kobayashi, T., Ito, S., Komatsu, Y., Yokoyama, T., Okada, M., Abe, A., Murasawa, A. & Yoshie, H. (2012). Interleukin-6 gene promoter methylation in rheumatoid arthritis and chronic periodontitis. *J Periodontol*, 83 (7): 917-25.

Janeway, C. A., Travers, P., Walport, M. & Shlomchik, M. J. (2001). *Immunobiology: the immune system in health and disease.*

T-cell receptor gene rearrangement. 5th ed., vol. 2. New York: Garland Science.

Jeltsch, A. (2002). Beyond Watson and Crick: DNA methylation and molecular enzymology of DNA methyltransferases. *Chembiochem*, 3 (4): 274-293.

Johnson, T. B. & Coghill, R. D. (1925). Researches on pyrimidines. C111. The discovery of 5-methyl-cytosine in tuberculinic acid, the nucleic acid of the tubercle bacillus1. *Journal of the American Chemical Society*, 47 (11): 2838-2844.

Karouzakis, E., Rengel, Y., Jungel, A., Kolling, C., Gay, R. E., Michel, B. A., Tak, P. P., Gay, S., Neidhart, M. & Ospelt, C. (2011). DNA methylation regulates the expression of CXCL12 in rheumatoid arthritis synovial fibroblasts. *Genes Immun*, 12 (8): 643-652.

Kim, B. C., Kim, W. Y., Park, D., Chung, W. H., Shin, K. & Bhak, J. (2008). SNP@Promoter: a database of human SNPs (Single Nucleotide Polymorphisms) within the putative promoter regions. *BMC Bioinformatics*, 9 (Suppl 1): S2.

Koning, F., Thomas, R., Rossjohn, J. & Toes, R. E. (2015). Coeliac disease and rheumatoid arthritis: similar mechanisms, different antigens. *Nat Rev Rheumatol*, 11 (8): 450-61.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M. & FitzHugh, W. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409 (6822): 860-921.

Lea, T. (2006). *Immunologi og immunologiske teknikker*. 3. ed. Bergen: Fagbokforlaget Vigmostad og Bjørke AS. 400 pp.

Leontiou, C. A., Hadjidaniel, M. D., Mina, P., Antoniou, P., Ioannides, M. & Patsalis, P. C. (2015). Bisulfite Conversion of DNA: Performance Comparison of Different Kits and Methylation Quantitation of Epigenetic Biomarkers that Have the Potential to Be Used in Non-Invasive Prenatal Testing. *PloS one*, 10 (8): e0135058.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25 (16): 2078-9.

Liao, J., Liang, G., Xie, S., Zhao, H., Zuo, X., Li, F., Chen, J., Zhao, M., Chan, T. M. & Lu, Q. (2012). CD40L demethylation in CD4+ T cells from women with rheumatoid arthritis. *Clinical Immunology*, 145 (1): 13-18.

Life Technologies. (2014). *Comparison of fluorescence-based quantitation with UV absorbance measurements. Qubit® fluorometric quantitation vs. spectrophotometer measurements*. Waltham, USA: Thermo Fisher Scientific. Available at: https://www.thermofisher.com/content/dam/LifeTech/global/life-sciences/Laboratory%20Instruments/Files/1014/Qubit-fluorometric-quantitation-vs-spectrophotometer-measurements.pdf (accessed: 26.04).

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462 (7271): 315-322.

Liu, Y., Aryee, M. J., Padyukov, L., Fallin, M. D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., et al. (2013). Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotech*, 31 (2): 142-147.

Lu, Q., Wu, A. & Richardson, B. C. (2005). Demethylation of the Same Promoter Sequence Increases CD70 Expression in Lupus T Cells and T Cells Treated with Lupus-Inducing Drugs. *The Journal of Immunology*, 174 (10): 6212-6219.

MacGregor, A. J., Snieder, H., Rigby, A. S., Koskenvuo, M., Kaprio, J., Aho, K. & Silman, A. J. (2000). Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum*, 43 (1): 30-7.

Mannik, M., Nardella, F. A. & Sasso, E. H. (1988). *Rheumatoid factors in immune complexes of patients with rheumatoid arthritis*. Springer seminars in immunopathology: Springer. 215-230 pp.

Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S. & Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, 33 (18): 5868-5877.

Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C. & Jaffe, D. B. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454.

Messemaker, T. C., Huizinga, T. W. & Kurreeman, F. (2015). Immunogenetics of rheumatoid arthritis: Understanding functional implications. *J Autoimmun*.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet*, 11 (1): 31-46.

Michaud, K. & Wolfe, F. (2007). Comorbidities in rheumatoid arthritis. *Best Pract Res Clin Rheumatol*, 21 (5): 885-906.

Moore, L. D., Le, T. & Fan, G. (2013). DNA methylation and its basic function. *Neuropsychopharmacology*, 38 (1): 23-38.

Myers, R., Sutton, G., Eichler, E., Kent, J., Guigo, R., Bult, C., Stemple, D., Korbel, J. & Worthey, E. (2015). *Human Genome Overview*: Genome Reference Consortium. Available at: http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/ (accessed: 06.05).

Nile, C. J., Read, R. C., Akil, M., Duff, G. W. & Wilson, A. G. (2008). Methylation status of a single CpG site in the IL6 promoter is related to IL6 messenger RNA levels and rheumatoid arthritis. *Arthritis & Rheumatism*, 58 (9): 2686-2693.

Nimesh, S. & Chandra, R. (2011). *Theory, Techniques and Applications of Nanotechnology in Gene Silencing*: River Publishers.

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., et al. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506 (7488): 376-381.

Payne, K. J. & Crooks, G. M. (2002). Human hematopoietic lineage commitment. *Immunol Rev*, 187: 48-64.

Plongthongkum, N., Diep, D. H. & Zhang, K. (2014). Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat Rev Genet*, 15 (10): 647-661.

QIAGEN. (2015). *QIAamp DNA Micro Procedure*: QIAGEN. Available at: https://www.qiagen.com/no/shop/sample-technologies/dna/dna-preparation/QIAamp-DNA-Micro-Kit?cmpid=QVen9GADNAMicroKit#productdetails (accessed: 22.02).

Raychaudhuri, S., Sandor, C., Stahl, E. A., Freudenberg, J., Lee, H. S., Jia, X., Alfredsson, L., Padyukov, L., Klareskog, L., Worthington, J., et al. (2012). Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet*, 44 (3): 291-6.

Richardson, B., Scheinbart, L., Strahler, J., Gross, L., Hanash, S. & Johnson, M. (1990). Evidence for impaired t cell dna methylation in systemic lupus erythematosus and rheumatoid arthritis. *Arthritis & Rheumatism*, 33 (11): 1665-1673.

Scott, D. L., Wolfe, F. & Huizinga, T. W. (2010). Rheumatoid arthritis. *Lancet*, 376 (9746): 1094-108.

Shapiro, R., Servis, R. E. & Welcher, M. (1970). Reactions of Uracil and Cytosine Derivatives with Sodium Bisulfite. *Journal of the American Chemical Society*, 92 (2): 422-424.

Silman, A. J., Macgregor, A. J., Thomson, W., Holligan, S., Carthy, D., Farhan, A. & Ollier, W. E. R. (1993). Twin concordance rates for rheumatoid arthritis: results from a nationwide study. *Rheumatology*, 32 (10): 903-907.

Smith-Garvin, J. E., Koretzky, G. A. & Jordan, M. S. (2009). T Cell Activation. *Annual review of immunology*, 27: 591-619.

Smith, Z. D., Gu, H., Bock, C., Gnirke, A. & Meissner, A. (2009). High-throughput bisulfite sequencing in mammalian genomes. *Methods (San Diego, Calif.)*, 48 (3): 226-232.

Sparks, J. A. & Costenbader, K. H. (2014). Genetics, Environment, and Gene-Environment Interactions in the Development of Systemic Rheumatic Diseases. *Rheumatic Disease Clinics of North America*, 40 (4): 637-657.

Starr, T. K., Jameson, S. C. & Hogquist, K. A. (2003). Positive and negative selection of T cells. *Annu Rev Immunol*, 21: 139-76.

STEMCELL technologies. (2013). *Simplify and Standardize PBMC Isolation with SepMate™*: STEMcell Technologies. Available at: http://www.stemcell.com/~/media/Technical%20Resources/A/D/E/7/0/29048BR087SepMatev111Web.pdf?la=en (accessed: 19.02).

STEMCELL Technologies. (2015). *EasySep™ Human CD8+ T Cell Enrichment Kit*: STEMcell Technologies. Available at: http://www.stemcell.com/en/Products/All-Products/EasySep-Human-CD8-T-Cell-Enrichment-Kit.aspx (accessed: 19.02.16).

Suarez-Alvarez, B., Rodriguez, R. M., Fraga, M. F. & López-Larrea, C. (2012). DNA methylation: a promising landscape for immune system-related diseases. *Trends in Genetics*, 28 (10): 506-514.

Suzuki, A., Kochi, Y., Okada, Y. & Yamamoto, K. (2011). Insight from genome-wide association studies in rheumatoid arthritis and multiple sclerosis. *FEBS Lett*, 585 (23): 3627-32.

Thermo Fisher Scientific. *Assessment of Nucleic Acid Purity, T042-TECHNICAL BULLETIN, NanoDrop Spectrophotometers*. Available at: http://www.nanodrop.com/Library/T042-NanoDrop-Spectrophotometers-Nucleic-Acid-Purity-Ratios.pdf (accessed: 26.04).

Thermo Fisher Scientific. (2010). *Nucleic Acid, Thermo Scientific NanoDrop Spectrophotometers*. Wilmington, USA: Thermo Fisher Scientific. Available at: http://www.thermoscientific.com/content/dam/tfs/ATG/CAD/CAD%20Documents/Application%20&%20Technical%20Notes/Molecular%20Spectroscopy/UV%20Visible%20Spectrophotometers/Spectrophotometer%20Systems/NanoDrop/Thermo-Scientific-NanoDrop-Products-Nucleic-Acid-Technical-Guide-EN.pdf (accessed: 11.04).

van der Woude, D., Houwing-Duistermaat, J. J., Toes, R. E., Huizinga, T. W., Thomson, W., Worthington, J., van der Helm-van Mil, A. H. & de Vries, R. R. (2009). Quantitative heritability of anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis. *Arthritis Rheum*, 60 (4): 916-23.

Van Laethem, F., Tikhonova, A. N. & Singer, A. (2012). MHC restriction is imposed on a diverse TCR repertoire by CD4 and CD8 coreceptors during thymic selection. *Trends in immunology*, 33 (9): 437-441.

Waalwijk, C. & Flavell, R. (1978). MspI, an isoschizomer of HpaII which cleaves both unmethylated and methylated HpaII sites. *Nucleic acids research*, 5 (9): 3231-3236.

Walker, L. S. & Abbas, A. K. (2002). The enemy within: keeping self-reactive T cells at bay in the periphery. *Nat Rev Immunol*, 2 (1): 11-9.

Wang, R. Y., Gehrke, C. W. & Ehrlich, M. (1980). Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues. *Nucleic Acids Research*, 8 (20): 4777-4790.

Wang, X., Zhu, H., Snieder, H., Su, S., Munn, D., Harshfield, G., Maria, B. L., Dong, Y., Treiber, F., Gutin, B., et al. (2010). Obesity related methylation changes in DNA of peripheral blood leukocytes. *BMC Medicine*, 8 (1): 1-8.

Weber, M., Hellmann, I., Stadler, M. B., Ramos, L., Paabo, S., Rebhan, M. & Schubeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet*, 39 (4): 457-66.

Willemze, A., Trouw, L. A., Toes, R. E. & Huizinga, T. W. (2012). The influence of ACPA status and characteristics on the course of RA. *Nat Rev Rheumatol*, 8 (3): 144-52.

Xi, Y. & Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, 10 (1): 1-9.

Xi, Y., Bock, C., Müller, F., Sun, D., Meissner, A. & Li, W. (2012). RRBSMAP: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing. *Bioinformatics*, 28 (3): 430-432.

Ziller, M. J., Müller, F., Liao, J., Zhang, Y., Gu, H., Bock, C., Boyle, P., Epstein, C. B., Bernstein, B. E., Lengauer, T., et al. (2011). Genomic Distribution and Inter-Sample Variation of Non-CpG Methylation across Human Cell Types. *PLoS Genetics*, 7 (12): e1002389.

Zola, H., Swart, B., Banham, A., Barry, S., Beare, A., Bensussan, A., Boumsell, L., C, D. B., Buhring, H. J., Clark, G., et al. (2007). CD molecules 2006--human cell differentiation molecules. *J Immunol Methods*, 319 (1-2): 1-5.

Appendix 1: Reagent list

Table 12 contains information about all reagents used during the work on this thesis.

Table 12 – Reagents used in the work of this thesis. Included are name of the producer and the catalog number for each entry.

Reagent name	Producer	Catalog number
Agencourt AMPure XP 60mL Kit	Beckman Coulter	A63881
CD45RO PE antibody	BioLegend	304206
D1000 reagents	Agilent Technologies	5067-5583
D1000 ScreenTape	Agilent Technologies	5067-5582
100mM Deoxynucleotide (dNTP) Solution Set (25µmol of each in four searate solutions)	NEB	N0446S
100mM dNTP, 25mM each dNTP	Agilent Technologies	200415
DPBS, no calcium, no magnesium	Thermo Fisher Scientific	14190169
EasySep Human $CD4^{+}CD25^{HIGH}$ T Cell Isolation Kit	STEMCELL	18062
EasySep Human CD8 positive isolation Kit	STEMCELL	18053
EasySep™ Human PE Positive Selection Kit	STEMCELL	18551
EB buffer	QIAGEN	19086
EDTA (0.5 M), pH 8.0	Thermo Fisher Scientific	AM9262
EpiTect Bisulfite Kit	QIAGEN	59104
Ethanol AnalaR NORMAPUR® ACS	VWR	20821.310
FBS	Biowest SAS	ALB-S181H-500
5000 U/mL Klenow fragment	NEB	M0212S
1x Low TE Buffer	Thermo Fisher Scientific	2090-015

LymphoPrep	Axis-shield	1114547
Magnesum Chloride Solution ($MgCl_2$)	Merck	M1028
MiSeq Reagent Kit v3	Illumina	MS-102-3003
20U/µL MspI	NEB	R0106S
10x NEB buffer 2	NEB	M0212S
Nuclease-free water	QIAGEN	1039498
Nuclease-free water (not DEPC-treated)	Thermo Fisher Scientific	AM9937
PCR Primer Cocktail	Illumina	FC-121-4001
2.5U/µL PfuTurbo C_x Hotstart DNA Polymerase	Agilent Technologies	600412
10x PfuTurbo C_x Reaction Buffer	Agilent Technologies	600412
PhiX Control v3	Illumina	FC-110-3001
20% Polyethylene Glycol/2.5M NaCl, KAPA PEG/NaCl SPRI® Solution	Kapa Biosystems	KB8232
Proteinase K	Merck	1245680500
QIAamp DNA blood mini kit	QIAGEN	51104
QIAamp DNA micro kit	QIAGEN	56304
Qubit® dsDNA HS Assay Kit	Thermo Fisher Scientific	Q32854
RNA/DNA/Protein Purification Plus Kit	Norgen	47700
RNAprotect Cell Reagent	QIAGEN	76526
20% SDS Solution	Bio-Rad Laboratories	161-0418
Sodium Chloride (NaCl)	Merck	1064040500
Sucrose (Saccharose)	Merck	1076871000
400000 U/mL T4 DNA Ligase	QIAGEN	M0202S
10x T4 DNA Ligase Reaction Buffer	QIAGEN	M0202S
Triton-X	Merck	T8787
Trizma® Hydrochloride Solution (Tris-HCl)	Merck	T3038
TruSeq Nano DNA LT Library Prep Kit Set A	Illumina	FC-121-4001
Water, PCR Grade	Roche	03315932001

Appendix 2: Equipment list

Table 13 contains the information about all the equipment used in the work on this thesis.

Table 13 - Equipment used during the work on this thesis. Included are name of the producer and the catalog number for each entry.

Name of equipment	Producer	Catalog or model number
Applied Biosystems 2720 Thermal Cycler	Thermo Fisher Scientific	4359659
Applied Biosystems Veriti Thermal Cycler	Thermo Fisher Scientific	4375786
Blood bag	Fresenius Kabi AG	R7043
CentriVap DNA Vacuum Concentrator	Labconco	7970030
Countess Automated Cell Counter	Thermo Fisher Scientific	C10281
Direct-Q® 3UV-R	Merck	ZRQS0P3WW
DynaMag-2	Thermo Fisher Scientifc	123-21D
Eppendorf® Biopur® Safe-Lock microtubes	Merck	Z317217
Heraeus Biofuge Fresco	Thermo Fisher Scientific	75005521
Hettich MIKRO 200	Hettich Instruments, LP	2400-01
Millipak® Express 40 Filter	Merck	MPGP04001
MiSeq System	Illumina	SY-410-1003
Nanodrop	Thermo Fisher Scientific	ND-1000
Qubit 2.0 Fluorometer	Thermo Fisher Scientific	Q32866
SepMate™-50	STEMCELL	15460
Sterican G21 syringe	VWR	720-2531
2200 Tapestation	Agilent Technologies	02965A NA

Appendix 3: Methylation versus gene expression plots

The average methylation level 5000bp upstream of the genes plotted against expression for each of the three samples representing different DNA extraction methods ae given in Figure 29 for chromosomes 1 and 8, and in Figure 30 for chromosomes 18 and X.

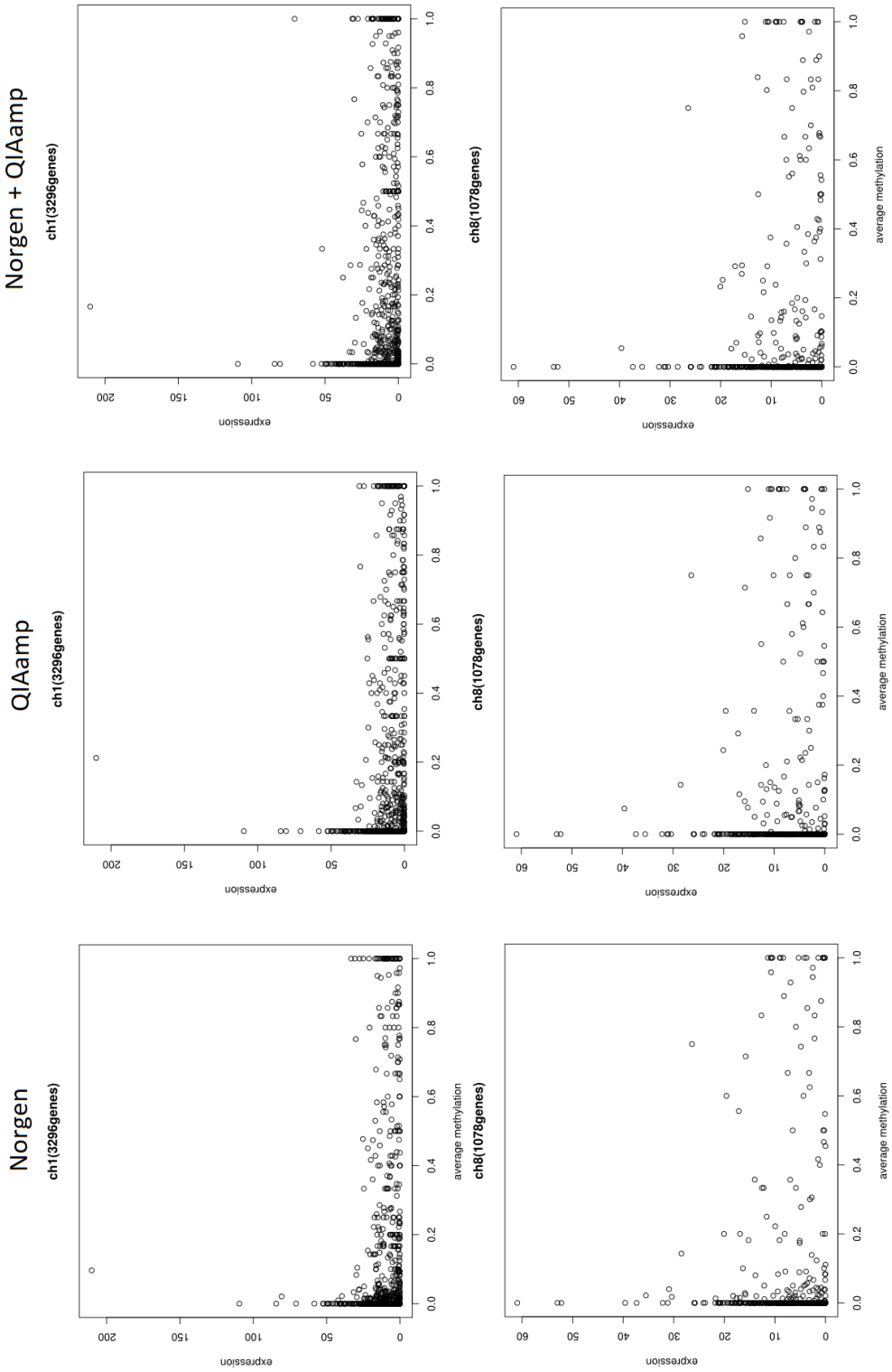


Figure 29 - Average methylation levels plotted against expression levels for Norgen, QIA-micro and NorgenClean samples. The plot for chromosome 1 and 8 is shown.

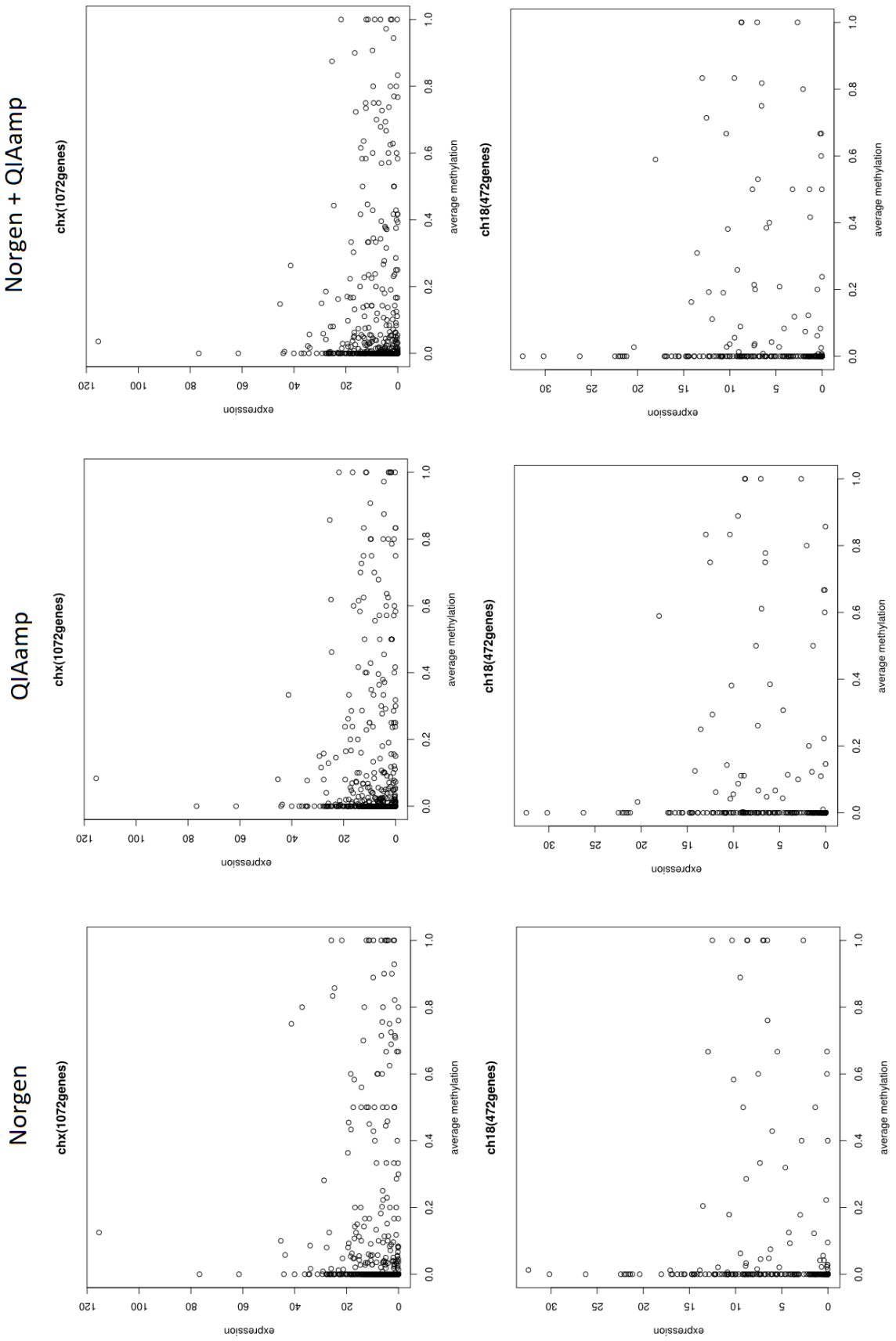
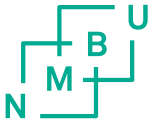


Figure 30 - Average methylation levels plotted against expression levels for Norgen, QIA-micro and NorgenClean samples. The plot for chromosome X and 18 is shown.



Norges miljø- og biovitenskapelig universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway