Genetic study of T cell receptor (TCR) in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)

Marthe Ueland

Master Thesis
Genetics and Developmental Biology
60 credits

Department of Biosciences
Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO

May 2021

Genetic study of T cell receptor (TCR) in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)

Oslo University Hospital,
Department of Medical Genetics

and

University of Oslo,
Faculty of Mathematics and Natural Sciences,
Department of Biosciences,
Master in Genetics and Developmental Biology

2021

Genetic study of T cell receptor (TCR) in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)

Marthe Ueland

http://www.duo.uio.no/

Trykk: Reprosentralen, Universitetet i Oslo

Abstract

Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) is a disabling disease affecting patients physically and cognitively by e.g fatigue, post-exertional malaise (PEM), pain, memory-loss and concentration difficulties. It is currently no treatment for ME/CFS, and manifestation differs between individuals, which makes it difficult to identify its aetiology. Multiple genetic and environmental factors are believed to contribute to its development, thus categorizing it as a complex disease, which also is the case for autoimmune diseases (AID).

A hypothesis that ME/CFS is an immune-mediated disease has been suggested and are supported by various findings. Immunological alterations such as altered T cell response have been reported in patients. Additionally, has an increased occurrence of autoimmune diseases (AIDs) been observed in families with ME/CFS. The hypothesis is further supported by an identified association between human leukocyte antigen (HLA) class I and II and ME/CFS, a hallmark for most AIDs.

As HLA molecules present antigens to the T cell receptor (TCR), this receptor is of interest for further investigation. Furthermore, studies of TCRs have shown that both HLA molecules and single nucleotide polymorphisms (SNPs) located within the germline DNA can influence the gene usage in TCRs. Associations have been found between the TCR α chain (TRA) region and immune-mediated diseases. For example has a genome-wide association study (GWAS) conducted in narcolepsy, which also has an HLA class II-association, identified associations ($p<10^{-21}$) to three single nucleotide polymorphisms (SNPs) in TRA (rs1154155, rs12587781 and rs1263646), which was the first documented involvement of this region in disease. An additional small GWAS in ME/CFS showed association between three SNPs in TRA (rs17255510, rs11157573 and rs10144138) and the disease (adjusted $p<0.05$).

The aim of this thesis was to find methods that can be used to study genetic variants in the T cell receptor α region (TRA) to identify possible associations with ME/CFS. This was done by genotyping and sequencing. Association analysis of 30 SNPs genotyped using Illumina Immunochip (Ichip) and Taqman assays in a Norwegian cohort of 408 ME/CFS cases and 721 controls failed to show any association between TRA and ME/CFS. Since these included two of the SNPs previously associated with ME/CFS (rs17255510 and rs11157573), we did not replicate the findings.

Analysis of Ichip´s coverage of the TRA gene showed that it was inadequate with only 27 SNPs covered in this region, although 737 has been identified in the 1000 genomes CEU dataset. The TCR genetic regions are generally understudied due to homology and repetitive regions, which is problematic to cover with existing methods. Hence, two sequencing protocols were established in the TRA region. PacBio´s No-amp targeted sequencing utilizing the CRISPR-Cas9 system with SMRT sequencing was tested for fragments ranging from 4.8 to 20.1 kb, however, the highest read depth was obtained for fragments <6 kb. We conclude that this protocol is not suited for screening but can be a good complement to other sequencing methods. Long-range PCR with Illumina Miseq sequencing resulted in read depth able to detect genetic variant for some fragments, however, the approach required a lot of optimization. The obtained sequences were not studied in detail during this work and would therefore be of interest to investigate further to identify genetic variants. Future studies in this region would include targeted enrichment using capture probes.

In conclusion, we did not detect any association between ME/CFS and TRA, however, we revealed that the genetic variants tested thus far does not capture the genetic variation in this region. Furthermore, the sequencing protocols tested pave the way for further optimization and characterization of TRA by sequencing.

Acknowledgements

I would like to express deep gratitude to my main-supervisor, Marte K. Viken, for all her time devoted to me throughout this project. My understanding of genetics and my laboratory skills have greatly advanced due to all her knowledge, which she has graciously shared. For this I am very grateful.

I would also like to thank Benedicte A. Lie for being one of my co-supervisors. She has offered me extensive support and valuable inputs, and I am grateful for having been given the opportunity to join her research group.

I appreciate all the technical guidance and help which Siri Tennebø Flåm has provided me. She was always available, and willing to assist, whenever I had a question related to my laboratory work.

Thanks also to Riad Hajdarevic for being one of my co-supervisors, and for always keeping his door open for me to ask about anything.

I would also like to thank my internal supervisor at UiO, Finn-Eirik Johansen.

I wish to thank the rest of the immunogenetics group for making me feel welcome, as well as for their support and motivation, although most of it was kept on screen. I have especially enjoyed the company of my fellow master student, Invild Ringen Jøråsen. Our walks and talks about fantasies have been much appreciated during these times.

Lastly, I would like to thank my family and friends for always checking in on me, and for offering me company through study dates.

Marthe Ueland

Oslo, May 2021

Abbreviations

ABS	Antigen-Binding Site
AID	Autoimmune Disease
APC	Antigen Presenting Cell
BAM	Binary Alignment Map
BSA	Bovine Serum Albumin
BWA	Burrows-Wheeler Alignment Tool
CCC	Canadian Consensus Criteria
CCS	Circular Consensus Sequence
CEU	Caucasian of European Descent
CDR	Complementarity-Determining Regions
CD3	Cluster of Differentiation 3
CFS	Chronic Fatigue Syndrome
crRNA	CRISPR RNA
DIN	DNA Integrity Number
DPBS	Dulbecco´s Phosphate-Buffered Saline
FBS	Fetal Bovine Serum
gRNA	Guide RNA
GWAS	Genome Wide Association Studies
HiFi	High Fidelity
HLA	Human Leukocyte Antigen
HMW	High Molecular Weight
HTS	High-Throughput Sequencing
ICC	International Consensus Criteria
Ichip	Immunochip
ICD-10	International Statistical Classification of Diseases and Related Health Problems
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
ME	Myalgic encephalomyelitis
MHC	Major Histocompatibility Complex
NBMDR	The Norwegian Bone Marrow Donor Register
PEM	Post Exertional Malaise
RPMI	Roswell Park Memorial Institute
RSS	Recombination Signal Sequence
SAM	Sequence Alignment Map
sgRNA	Single-Guide RNA
SMRT	Single Molecule, Real-Time
SNP	Single Nucleotide Polymorphism
TCR	T Cell Receptor
TRA	T Cell Receptor α Chain
TRAC	T Cell Receptor α Chain Constant
tracrRNA	Trans-Activating CRISPR RNA
TRAJ	T Cell Receptor α Chain Joining
TRAV	T Cell Receptor α Chain Variable

Table of contents

1 Introduction

1.1 T cells and autoimmune diseases

The immune system consists of organs, tissues and cells participating in a complex arrangement to eliminate intruding microorganisms called pathogens, such as bacteria, viruses and parasites. The contributors have different roles and will be activated at certain time points during an infection. The cells of the innate immune system are activated within hours of infection and will start the process of killing the pathogens in a non-specific manner, while at the same time activating the more specialised part of the immune system, the adaptive immune system.

The adaptive immune system takes longer to be activated but will in turn be more specific towards the type of pathogen causing the infection and will result in immunological tolerance. The immunological tolerance offers an immediate and stronger response against a second infection with the same pathogen. The contributors of the adaptive immune system are B- and T cells, which are descendants of pluripotent hematopoietic stem cells found in the bone marrow (Murphy & Weaver, 2017). Both cell types have important roles in fighting infection but only the T cells will be presented in this thesis.

The T cells are circulating the blood, lymphatic system and secondary lymphatic organs searching for antigens bound to and presented by surface molecules called the major histocompatibility complex (MHC) (Alcover, Alarcón, & Bartolo, 2018) or human leukocyte antigen (HLA) in human. There are two kinds of HLA molecules: HLA class I and HLA class II. HLA class I is located on the surface of all nucleated cells, while HLA class II molecules are found on the surface of antigen presenting cells (APCs). APCs are immune cells such as dendritic cells, macrophages and B cells (Murphy & Weaver, 2017) which have the ability to internalize extracellular proteins or pathogens (e.g bacteria) by phagocytosis and present antigens derived from these on HLA class II. The antigens bound on HLA class II will be recognized by CD4 (helper) T cells and will lead to the activation of other immune cells such as B cells and macrophages, which both will help clear the infection. HLA class I molecules will on the other hand present antigens derived from intracellular proteins, from for example

infecting viruses. CD8 (cytotoxic) T cells will thus upon activation by HLA class I on infected cells, initiate apoptosis of these cells and eventually kill them.

Antigens are not only belonging to external organisms. They are also expressed on our own cells, called an autoantigen. In the event of an autoantigen being bound by a T cell, the T cell should be tolerant, meaning that it should be able to recognize it and not initiate an immune response (Murphy & Weaver, 2017). In some cases, however, T cells do not make this distinction, thus causing an immune response towards own cells and tissues, resulting in what is called an autoimmune disease (AID). In order to prevent an autoimmune response taking place, mechanisms exist to identify and eliminate the T cells recognizing autoantigens during their development in thymus.

1.1.1 T cell maturation in thymus

When the T cell progenitors enters the thymus, thymic cells signal for the progenitors to commit to the T cell linage and start the T cell receptor (TCR) rearrangement process (Figure 1.1). Prior to this signalling event, the cells are in a double-negative state, which means that they do not have co-receptors (CD4 or CD8) or a TCR on their surface. Initiation of the rearrangement result in T cells (now called thymocytes) starting to produce TCRs by rearranging gene segments that make up the two polypeptide chains the TCR consist of. This event result in the thymocytes having a middle state where it has both co-receptors on its surface, CD4 or CD8, so called double positive thymocytes.

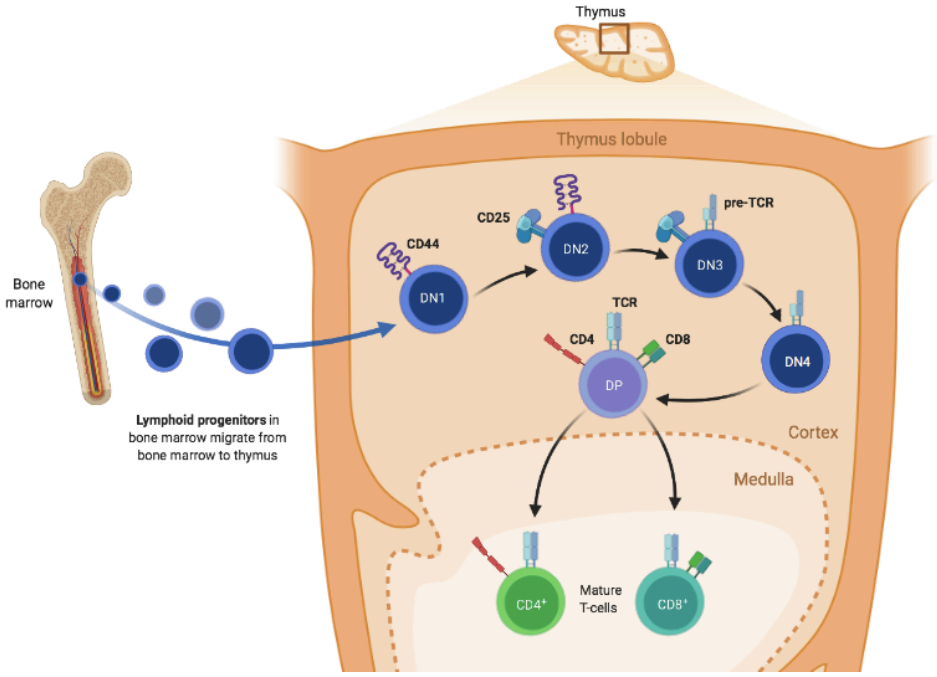


Figure 1.1 Illustration of T cell development in thymus (created in Biorender). Lymphoid progenitors migrate from bone marrow to thymus where they upon signaling from thymic cells commit to the T cell lineage and start expressing markers and rearrange gene segments to make the T cell receptor (TCR).

When the TCR is rearranged and located on the cell surface as a protein complex with an invariant cluster of differentiation 3 (CD3) chain molecule, the thymocyte goes through two selection processes. The first one is a positive selection where the aim is to make sure that the TCR receptor is functional. Thymocytes unable to bind HLA molecules located on thymus residential cells die by neglect while the ones able to bind, downregulate their unbound co-receptor. The latter ones will in turn go through a negative selection where thymocytes binding autoantigens too strongly will undergo apoptosis, which is essential to prevent an autoimmune response.

Thymocytes that pass both selection processes (~5% (Borghans, Noest, & De Boer, 2003)) are now more specialized with only one kind of co-receptor, either CD4 or CD8, and the mature, single positive T cells now migrate from the thymus and into the periphery where they circulate the blood, lymph and secondary lymphoid organs, such as the spleen and lymph nodes (Ruddle & Akirav, 2009) searching for APCs (Murphy & Weaver, 2017).

1.1.2 T cell receptor

The TCR is a protein complex on the T cell plasma membrane, as illustrated in Figure 1.2. The TCR itself is a heterodimer able to bind antigens when presented as peptides in the context of HLA. In order for the TCR to initiate signal transduction upon antigen-binding, a complex is needed (Alcover et al., 2018). This complex is the CD3 and consist of four different polypeptide chains (Alcover et al., 2018)

The two polypeptide chains making up the TCR heterodimer can either be an α and a β chain or a γ and a δ chain. γδ T cells comprises <5% of the peripheral lymphocyte population (Paul, Shilpi, & Lal, 2015). All four chains consist of a variable (V) and a constant (C) region and are produced the same way.

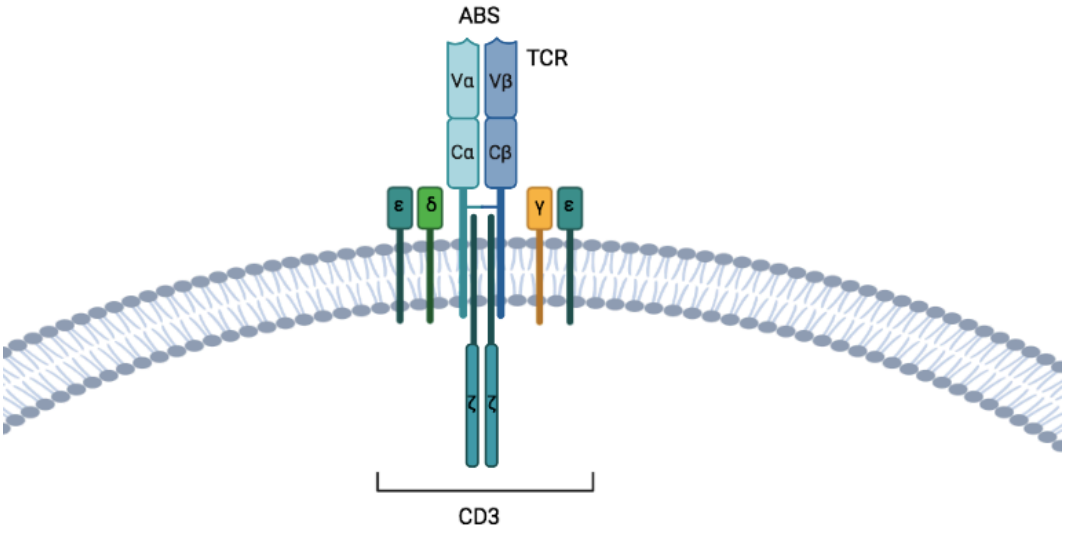


Figure 1.2 Schematic of the T cell receptor (TCR) protein complex consisting of an αβ TCR and a surrounding cluster of differentiation 3 (CD3) complex (created in Biorender). The TCR is made up of an α and a β chain with a variable (V) and constant (C) region, while the CD3 complex consists of four different chains, γ, δ, ε and ζ.

The TCR binds to an antigen presented by either HLA class I or class II through the antigen-binding site (ABS). The ABS is consisting of three hypervariable loops called complementarity determining regions (CDRs)(Murphy & Weaver, 2017), found in all TCR chains (Rosati et al., 2017), and are complementary to the antigens they bind (Murphy & Weaver, 2017). Two of them (CDR1 and CDR2) are found to provide the basic affinity of the TCR for the HLA allele and is likely to be responsible for positive selection, whereas the third (CDR3) are positioned to primarily contact the peptide and play a more important role in

negative selection (Rudolph & Wilson, 2002). CDR3α loop is found to be the most variable in the investigated TCR structures (Rudolph & Wilson, 2002).

1.1.3 TCR diversity

TCR genes are encoded at multiple locations in the human genome (Lefranc & Lefranc, 2001). The TCRβ and TCRγ chain region (TRB and TRG) are both located on chromosome 7, while the TCRα and δ chain region (TRA and TRD) are located on chromosome 14. The TRD region is embedded in TRA and they are spanning 960 kb (from chr14:21,620,000 to 22, 580,000, GRCh38) at chromosome site 14q11.2. Including the *TRD* genes, is a total of 127 genes are located here (Lefranc, 2020), however, when excluding the *TRD* and non-functional genes (e.g pseudogenes), between 94 and 96 *TRA* genes makes up the germline repertoire (Lefranc, 2020) that can be rearranged to make the TRA chain. The invariant CD3 chain molecules are encoded by the *CD3γ*, *CD3δ* and *CD3ε* genes located close to each other on chromosome 11 (Evans, Lewis, & Lawless, 1988; Weissman et al., 1988) the remaining CD3ζ chain is transcribed from chromosome 1 (Weissman et al., 1988).

In order for a TCR to be expressed on the T cell surface, germline DNA of the two polypeptides goes through rearrangements of gene segments, transcription, splicing and translation, as presented in Figure 1.3. All of the four polypeptide chains consist of one variable (V), one joining (J) and one constant (C) gene segments. The β and δ chains has an additional diversity (D) gene segment. The V, J and D gene segments make up the variable (Rosati et al., 2017) domain. Different number of gene segments is making up the germline repertoire of the four chains, however, the rearrangement occurs in the same way. As the α chain is the focus of this thesis, this will be used to describe the process.

The germline TRA repertoire contains 43-45 *TRAV*, including 5 that can become *TRAV* or *TRADV* gene segments, 50 *TRAJ* and 1 *TRAC* functional gene segments (Lefranc, 2020). The V(D)J rearrangement occurs to bring one V and J gene segments together to make a functional V region exon (Murphy & Weaver, 2017). The Vα exon will be transcribed and spliced to Cα to form an mRNA which is translated to the TCRα polypeptide chain (Murphy & Weaver, 2017).

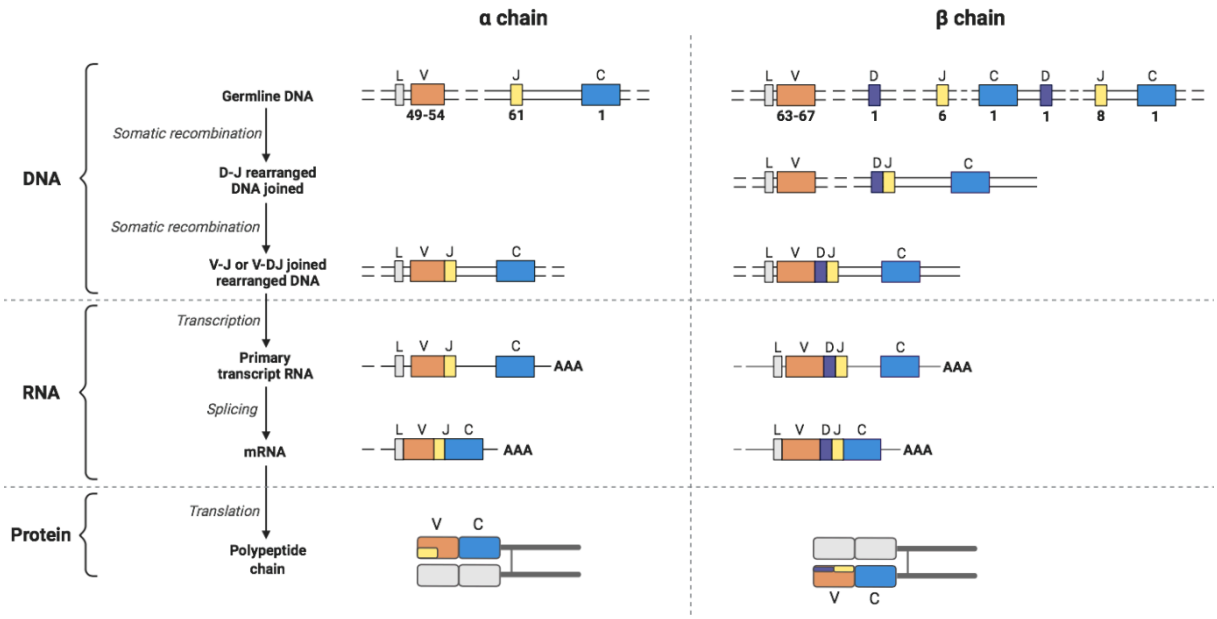


Figure 1.3 Schematic drawing of an αβTCR and the genomic regions that make up the two polypeptide chains (created in Biorender). Non-functional genes are included in their germline DNA. The αβTCR is made up by V, D, J and C segments through recombination.

V(D)J rearrangement is guided by two conserved recombination signal sequence (RSS) motifs of 12 or 23 bp located adjacent to the V, D and J gene segments (Murphy & Weaver, 2017). During rearrangement, a gene segment with a 23 bp RSS always be recombined with one of 12 bp length to fulfil the 12/23 rule (Murphy & Weaver, 2017).

The TCR repertoires in individuals are influenced by the gene usage in the interaction between the peptide, HLA and the T cell (Murphy & Weaver, 2017) created by the three hypervariable regions. CDR3 is encoded by the junctional region between the V and J segment (D and J in β chain) and is highly variable (Rosati et al., 2017). Thus, the possibility of two T cells to express the same nucleotide sequence in this loop is highly unlikely (Rosati et al., 2017). The two other loops are encoded by V genes (Rosati et al., 2017).

1.1.4 TCR in autoimmunity

Although there is negative selection to prevent T cells with self-reactive TCRs from leaving the thymus, there are still some that escape and can cause autoimmunity. Mechanisms both in the developing thymocytes and in other cells have been proposed to play a part to how T cells are surpassing negative selection process (Arnold, 2002; Klein, Kyewski, Allen, & Hogquist,

2014) or how their rearrangement is being influenced (McMurry, Hernandez-Munain, Lauzurica, & Krangel, 1997; Posnett et al., 1994)

The expression levels of some autoantigens in thymic cells may be influenced by SNPs in regulatory regions. Thus, causing them to be expressed in low levels in the thymus but not in the periphery, resulting in the TCRs not binding these autoantigens before they are in the periphery and thus able to cause disease (Klein et al., 2014). Post-translational modification of autoantigens in various cells or tissues, but not in thymic cells (Klein et al., 2014) can also result in autoreactive T cells because the antigens being presented in thymus is not the exact same as the ones found in the periphery.

The generation of TCR repertoires through rearrangement and recombination of the germline segments might also not be completely random. HLA alleles have been found to influence the preferred usage of segments in the TCR rearrangement (Sharon et al., 2016) which is interesting given the strong association between autoimmune diseases. SNPs in intron, exons, RRS and enhancer are also suggested to influence the TCR repertoires (Sharon et al., 2016).

Enhancers and promoters are known to influence gene expression levels, and they have been found to have an important role in the V(D)J recombination, where enhancers have been hypothesized to have an additional role in accessibility of the RSS, which are needed to direct recombination (McMurry et al., 1997). Physical constraints in the RSS region can prevent the enzymes involved from accessing the site (Posnett et al., 1994), thus inhibiting the use of the adjacent gene segment.

SNPs located in RSS have been found to skew gene expression towards certain segments in the TCR (Posnett et al., 1994), which has also been seen in B cells receptors (Watson, Glanville, & Marasco, 2017). Because antibodies produced in B cells are highly homologous to TCRs in terms of their production, function and structure, it provides additional strength to the finding. These findings, together with the observation that more SNPs are located in non-coding than in coding regions (Mu & Zhang, 2013), could support the hypothesis that there may be SNPs in the TRA region that could affect the TCR rearrangement by skewing the production towards certain receptors which binds autoantigens.

1.2 ME/CFS

Myalgic encephalomyelitis or chronic fatigue syndrome (ME/CFS) is a disease to which there is currently no cure. The disease has through decades been debated whether it is a pure psychologic disease or if it has biomedical roots. A major reason for this is its unknown aetiology (cause) and pathogenesis (development), and the difference in clinical manifestation, both within an individual as well as between individuals. The World Health Organization (WHO) classified ME/CSF or post viral fatigue syndrome (PVS), as it is also called, a neurological disease in the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) from 2016. This is due to the fact that encephalomyelitis is an inflammation of the central nervous system (CNS) which includes the brain and spinal cord and associates with muscle pain (myalgia). As there are no biomarkers available in aiding the diagnosis, ME/CFS diagnosis is based on exclusion of other conditions and fulfilment of certain criteria.

There are today several sets of criteria that are used in diagnosis of ME/CFS; the 1994 Fukuda Criteria (Fukunda et al., 1994), the Canadian Consensus Criteria (CCC) (Carruthers, 2007) and the International Consensus Criteria (ICC) (Carruthers et al., 2011).

In all three criteria, there are differences in the number of and which symptoms that have to be present in patients for them to get an ME/CFS diagnosis. Typical symptoms are fatigue, post-exertional fatigue (PEM) and cognitive effects like memory-loss. Fatigue results in a decrease in the physical and mental activity level of a patient as compared to before illness and is a common symptom between all three criteria. The difference between the criteria in terms of this requirement, is that the fatigue must be present for at least 6 months before making the diagnosis according to the CCC (Carruthers, 2007) and Fukunda criteria (Fukunda et al., 1994) in adults.

The use of different diagnosis criteria and limited sample size in studies of ME/CFS (Schlauch et al., 2016; A. K. Smith, Fang, Whistler, Unger, & Rajeevan, 2011) make it difficult to compare between studies and draw any conclusions regarding the aetiology or pathogenesis of the disease, but there are findings suggesting that it can be an autoimmune disease (Lande et al., 2020; J. Smith et al., 2005). Lande et al. (2020) discovered significant association ($p<0.05$) with both an HLA class I (HLA-C*07:04) and class II (HLA-

DQB1*03:03) allele and ME/CFS when conducting a study of 426 cases compared with 4511 controls. This association may potentially affect the presentation of peptides to CD8 and CD4 T cells in patients. Increased number of autoreactive T cells in patients (Morris, Berk, & Galecki, 2014) provides additional support to it being an immune-mediated disorder.

There is strong support for a genetic contribution to disease susceptibility in ME/CFS (Albright, Light, Light, Bateman, & Cannon-Albright, 2011), further supported by family members being diagnosed with ME/CFS (Walsh, Zainal, Middleton, & Paykel, 2001). ME/CFS diagnoses are, however, not following a Mendelian pattern, which suggests that there is not one genetic variant that increases ME/CFS risk (Dibble, McGrath, & Ponting, 2020). Thus, making it likely to be a complex multifactorial disorder where there are many and different genetic contributors, as is the case for many autoimmune diseases for example (Dibble et al., 2020).

1.3 Genetics of complex autoimmune diseases

The human genome consists of around three billion base pairs (NIH). 0.1% of these are what contributes to genetic diversity between any two individuals (Goris & Liston, 2012). Of these are SNPs (MAF>1%) contributing to 90% of this difference (Goris & Liston, 2012). The study of genes and SNPs have provided a lot of information about their roles and how they contribute to human traits.

Traits can be either monogenic or polygenic, meaning that they are affected by variants in one or multiple genes, respectively. Polygenic traits can often be challenging to study due to the number of genes involved. Traits influence by both risk SNPs and environmental factors that affect the susceptibility are called complex traits. Extra challenges are added to the study of these traits due to phenotypic differences in clinical presentation and outcome between individuals with the same trait, as is often seen in complex diseases such as autoimmune or in the case of ME/CFS. Such phenotypic heterogeneity could be reflected in genetic heterogeneity, pleiotropy and phenocopies. The combination of risk factors hence varies greatly between patients.

In order to identify the causal variant(s) one can take advantage of linkage disequilibrium (LD). LD means that alleles at two or more loci are inherited together more frequently than

expected by chance (Ardlie, Kruglyak, & Seielstad, 2002; Slatkin, 2008). This is often more likely to occur between alleles that are located in close proximity to each other on a chromosome (Slatkin, 2008) as they are less prone to be split by gene rearrangement events.

LD is typically measured by unidirectional D prime (D´) and bidirectional R squared (r^2) (Ardlie et al., 2002). Both measures provide information about the history between two alleles. In both cases, a value of 1 means that the alleles have never been separated by recombination (Ardlie et al., 2002). If $r^2=1$, maximum two out of four haplotypes are observed in the population, which is called perfect LD. When $D´=1$, it indicates that maximum three out of four haplotypes are observed in the population, also known as complete LD.

An advantage with LD is that it allows for SNP tagging, where genotyping of one SNP that is in strong LD with the nearby SNPs in the same LD block provides information about their alleles as well (Hirschhorn & Daly, 2005). This allows for a more efficient screening of genome regions, or even the whole genome, by reducing the number of SNPs needed to be genotyped, which also affect the time spent and the cost of genotyping. In order to get as much information as possible from a genomic region by SNP tagging, all genetic variation within the region are covered by the SNPs.

Even though LD can be exploited to screen the human genome for disease associations, it can also make it difficult to pinpoint the actual causal variant(s). A SNP can be believed to cause disease when it in reality is in LD with the actual SNP involved in the development of the disease (Maynard Smith & Haigh, 1974).

1.3.1 Study designs

In genetic studies of monogenic diseases families are investigated in order to find the causal variant, while for polygenic diseases case-control studies are mostly used to uncover the susceptibility variants. A case-control study is a retrospective study method where an observed disease is investigated with the goal of finding the exposure leading to it (Lewallen & Courtright, 1998). To be able to perform a case-control study, cases (group with the outcome) and controls (without the outcome) need to be identified. As allele frequencies can

vary between populations it is important to account for the ethnic background. Hence, cases and controls should be collected from the same population background.

Once the cases and controls are identified, one can find out how many have been exposed to the identified risk factors and compare the frequencies of the groups. This can be done by calculating the frequencies of variables in both groups to find the odds ratio (OR) "ratio of the odds of an exposure in the case group to the odds of an exposure in the control group"(Lewallen & Courtright, 1998). An OR>1 and OR<1 indicates that the investigated exposure is a risk and protective factor, respectively. In genetic studies this is typically performed by comparing allele- or genotype frequencies between cases and controls (Hirschhorn & Daly, 2005), and if the occurrence of an allele or a genotype is significantly more frequent in cases than in controls it is considered to be a risk variant for the disease being studied.

Genetic case-control studies can be performed as a genome-wide association study (GWAS) where hundreds or thousands of variants at different loci are being studied with the aim of finding an association between the variants and phenotypes being studied (i.e affected and not affected) (Donnelly, 2008). An association is detected when a variant at a locus is found more frequently in one of the phenotypes than in the other (Donnelly, 2008). GWASs are mostly performed using common SNPs with a minor allele frequency (MAF)>5% (Trynka et al., 2011) without a prior hypothesis of them being associated with the phenotypes.

Another type of case-control study that can be performed is a candidate gene study. In these studies, one or a few genes or gene regions are being investigated based on previous knowledge and/or hypothesis of them having a plausible role in aetiology or pathogenesis of the phenotype being investigated. Allele frequencies of SNPs within the gene(s) or regions can then be compared to see whether they are more frequently present in one phenotype than in others.

Independent of the study performed, the sample size is important to ensure appropriate statistical power needed for the findings to be representative and true. This proves a challenge when studying complex or rare diseases where there are few affected, or the manifestations of disease is different between patients. Which in the case of ME/CFS makes it hard to draw any conclusions regarding the disease´s aetiology or pathogenesis.

1.3.2 Findings from GWAS

GWASs have been performed for different phenotypes. Some of which are interesting for this thesis as they have identified associations with SNPs located in TRA. Hallmayer et al. (2009) used the Genome-Wide SNP Array 6.0 (Affymetrix) to genotype a Caucasian cohort consisting of 1830 narcolepsy cases and 2164 controls from Europe and the United States. As narcolepsy has been associated with HLA-DQB1*06:02 (Hallmayer et al., 2009), all their cases were HLA positive. The three most significant SNPs identified in this study were located in the TRA region, around the *TRAJ* gene segments, and were all in high LD. The significance of the three SNPs was successfully replicated in another Caucasian and Asian cohort using TaqMan assays. Rs1154155 was the most significant SNP ($p<10^{-21}$). This study was the first to document the genetic involvement of TRA in a disease (Hallmayer et al., 2009).

Schlauch et al. (2016) genotyped 42 ME/CFS cases and 38 controls, all Caucasian, using the same genotyping array as Hallmayer et al. 442 SNPs, at loci covering most of the chromosomes, were found associated to the cases with statistical significance (adjusted p-value $p<0.05$) after quality control. Three of these (rs17255510, rs11157573 and rs10144138) were found to be in the TRA or TRA/TRD region on chromosome 14. LD was measured between the three SNPs and showed almost complete and almost perfect LD ($D'=0.999$ and $r^2=0.999$) between rs17255510 and rs10144138.

Studies like these provides information about genomic regions which can be associated with disease. Both narcolepsy and ME/CFS have an already suggested HLA-association (Hallmayer et al., 2009; Lande et al., 2020), which can indicate an immunologic contribution to their aetiologies, this may provide more support for a possible involvement of variants in TRA.

Both studies described above utilized the Genome-Wide SNP Array 6.0 (Affymetrix) for genotyping. The array includes 906,600 SNPs (Schlauch et al., 2016) and is one of the genotyping arrays which can be used to screen the genome for association between a phenotype and chromosomal regions tagged by the SNPs on the array. Another genotyping array is the Illumina Immunochip (Ichip), which was made after an initiative by the Immunohip Consortium and contains close to 200,000 SNPs at 186 loci. The SNPs are all in

regions showing GWAS significant ($p<5x10^{-8}$) association with at least one of twelve immune-mediated diseases, like lupus erythematosus (SLE), rheumatoid arthritis (RA), type 1 diabetes (T1D), celiac disease (CD) and multiple sclerosis (MS)(Trynka et al., 2011). Genetic studies of autoimmune diseases have shown that most associated variants are regulatory (Frazer, Murray, Schork, & Topol, 2009), hence likely to affect the fine tuning of the immune system.

Genotyping arrays have been a great tool to identify regions the phenotype is associated with, as exemplified by the two GWASs revealing the TRA region in narcolepsy and possibly also in ME/CFS. A continuation of such studies may be to try and replicate the findings in other cohorts, as well as genotype more SNPs in the same LD block to identify actual causal variants.

1.3.3 Detecting novel genetic variation

Another way of identifying variations within the genome is, by sequencing, to visualize the genetic code of which the genome is made up. An advantage of sequencing compared to arrays is that also novel genetic variants can be detected. This is particularly useful when investigating regions, like the TCR genes, where all genetic variation and haplotypes have not yet been characterized (Omer et al., 2020).

The sequencing methods available today are high throughput sequencing (HTS) methods, meaning that they are sequencing thousands of fragments covering hundreds to thousands of genes simultaneously. Sequencing can be used for a range of different research questions depending on the choice of template (e.g genomic DNA (gDNA), RNA or exomes) while allowing to sequence everything from only an area of interest to the whole genome of an organism. Independent of the target size, DNA is fragmented into short fragments before they are amplified and sequenced massively in parallel to generate “reads”. These reads are then mapped to the human reference genome (GRCh37) to make up a continuous genomic sequence.

Advances in the sequencing technology have not only resulted in optimization of already existing methods able to sequence short DNA fragments (called short-read sequencing), such

as Illumina, but it has also led to the development of additional methods able to cover longer sequences (long-read sequencing). Both of which have their advantages and limitations.

The short-read sequencing techniques have gone through optimizations that have led to an increase in throughput and accuracy as well as a huge price reduction for sequencing a whole genome, which have been important for getting us where we are today and makes them important tools in clinics. However, their limited read length (150-600 bp) makes them unable to detect all structural variants in the genome and makes it challenging to map back to the reference genome (Logsdon, Vollger, & Eichler, 2020). Uncovered regions (e.g centromeres and telomeres) are estimated to make up more than 15% (Logsdon et al., 2020) of the genome, and could possibly contain variants involved in disease.

Long-read sequencing (third-generation sequencing) approaches can generate long continuous sequences covering multiple kilobases of DNA (Logsdon et al., 2020). These methods can be beneficial to use when covering large structural variants such as insertions, deletions, duplications and translocations, as well as repeats and pseudogenes which may be too long for short-read methods to cover. An example of such a method is single-molecule real-time (SMRT) sequencing provided by Pacific Bioscience (PacBio).

Compared to next generation sequencing (NGS) methods like Illumina, PacBio use the creation of circularised DNA fragments which the polymerase can cover multiple times to create long reads (Figure 1.4). Each subread can then be compared and a consensus sequence from which a consensus read can be constructed (Travers, Chin, Rank, Eid, & Turner, 2010), offering higher accuracy compared to other long-read sequencing methods because sequencing errors can be identified.

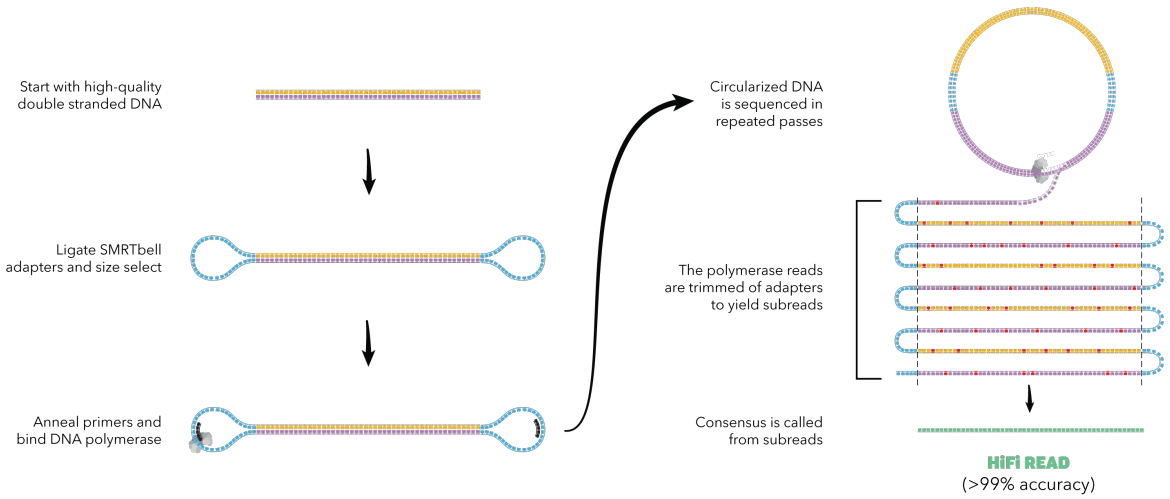


Figure 1.4 Schematics of the circularization of DNA fragments for consensus sequencing (Courtesy of Pacific Biosciences of California, Inc., Menlo Park, CA, USA). A consensus sequence is made from all subreads and allows for a reduction in sequencing errors, thus increasing the accuracy of the read.

Also, for PacBio sequencing different templates can be used, and will influence the cost. A resent protocol established by PacBio (PacBio, 2020), was selected to be tested in this thesis. This is the protocol "no-amp targeted sequencing utilizing the CRISPR-Cas9 system", which includes the need to design CRISPR RNA oligonucleotides and uses Cas9 endonuclease for targeting the region of interest. This method requires more input than for other PCR-based methods, which may be challenging if one has a limited amount of material accessible. PacBio also recommends using high molecular weight (HMW) DNA, with fragments of about 50 kb in length (PacBio, 2020), making other routine DNA extraction methods resulting in smaller fragments unsuitable for this application.

CRISPR RNA oligonucleotides (crRNA) are, similar to PCR primers, designed to be complementary and bind specifically to or near a target sequence (Figure 1.5). For crRNA and primers to work, they both require two oligonucleotides to bind on each side of the target sequence in a 5′→3′ direction. The two sequencing methods, Illumina and PacBio, differs in oligo function. While primers are being used for amplification of the target fragment, crRNA is used for fragmentation.

In order to fragment the DNA, trans-activating CRISPR RNA (tracrRNA), a universal sequence necessary for Cas9-nuclease-recruitment has to be annealed with a crRNA to create a guide RNA (gRNA).

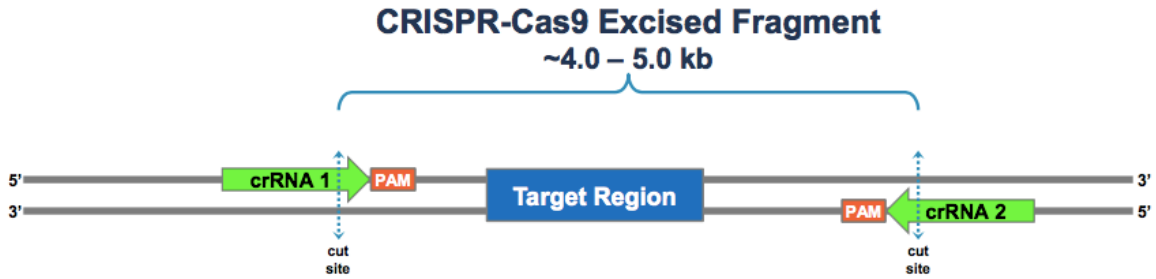


Figure 1.5 Illustration of CRISPR RNA oligonucleotides (crRNA) required for fragmentation of target sequence by Cas9. Two crRNAs are designed to bind on either side of the sequence. (Courtesy of Pacific Biosciences of California, Inc., Menlo Park, CA, USA)

Generally, the PacBio sequencing approach is known to have lower accuracy and higher price than short-read sequencing. While short read-sequencing have been popular due to their cost-effectivity and accuracy, long-read sequencing are often used to aid in de novo genome assembly (Amarasinghe et al., 2020). And these methods can complement each other when studying difficult regions in the human genome (Roberts et al., 2021) thus, the preferred method to use will depend on the aim of the research question. In this study, we have included both techniques as they have different pros and cons which we wanted to evaluate.

2 Aims

The overall hypothesis is that ME/CFS is an immune-mediated and complex disease in which genetic variants in the genes encoding TCR can contribute to susceptibility.

The aim of this thesis was to find methods that can be used to study genetic variants in the T cell receptor α region (TRA) to identify possible associations with ME/CFS.

Thus, to further understand the role of the TRA in ME/CSF, we therefore had the following objectives:

- Investigate the actual SNP coverage of the TRA region on immunochip
- Evaluate if publicly available genotyping data is representative for the Norwegian population
- Study whether genotyping of SNPs in the rearranged TRA region is reliable
- Establish PacBio´s “no-amp targeted sequencing utilizing the CRISPR-Cas9 system” protocol in the TRA region
- Compare long-read and short-read sequencing in TRA

3 Materials and methods

3.1 Material

For this study we had genomic DNA (gDNA) previously collected from 408 ME/CFS patients available. The patients were diagnosed in Norway according to the 2003 Canadian Consensus Criteria and recruited through recent or ongoing trials, the ME/CFS biobank at Oslo University Hospital (OUH) or announcements in patient networks. Clinical information about the patients were obtained though questionnaires filled by the patient or close relatives. Based on information provided by the patients, only one person per extended family (up to third relative) were included in the study. Additionally, we had genomic DNA from 721 healthy Norwegian controls recruited from The Norwegian Bone Marrow Donor Register (NBMDR) and ethnically matched to the cases. Age and gender distributions are listed in Table 1.

Table 1 Demographics of 408 ME/CFS patients and 721 healthy controls. [1]Mean age is calculated for 719 controls.

	Patients	**Controls**
Mean age, years (min, max)	39.5 (17, 79)	37.97[1] (18, 72)
Female : Male	335:73	394:327

A genotyping dataset generated using Ichip for 427 ME/CFS cases and 972 controls was also available for data analyses (Hajdarevic et al., Unpublished; International Multiple Sclerosis Genetics et al., 2013; Liu et al., 2013)

Furthermore, for the sequencing, blood was drawn from 5 healthy Norwegian anonymous women, between 30 and 60 years of age using EDTA collection tubes (Vacutest). The blood draw occurred according to regulations. Notably, blood volume was less than it should with about 3 mL per vial due to the collection tubes being expired by about two and a half months. One vial with whole blood was immediately placed at 4°C, while the other was subject to immediate CD3 depletion. The CD3 depletion and DNA extraction are described in Chapter 3.2.

3.2 DNA extraction from healthy individuals

In order to investigate if the presence of T cells with rearranged TCRs would affect the sequencing and genotyping results, gDNA was extracted from both whole blood (WB) and CD3 depleted (CD3(-)) blood where the T cells have been removed. As described in the introduction, mature T cells found in the periphery have rearranged TCRs. This removal of gene segments cause the DNA extracted from T cells in the whole blood to be shorter than the germline sequence present in all other white blood cells. Whether the presence of these rearranged TCRs affect sequencing of this region is not known, but it has the potential of causing an amplification bias towards the shorter sequences when using sequencing approaches requiring PCRs.

3.2.1 CD3 depletion of whole blood

To remove T cells from whole blood samples, CD3 depletion was performed immediately after the blood draw. CD3 depletion was performed using the Dynabeads® FlowComp™ Human CD3 kit by Invitrogen (Catalog nr. 11365D), Thermo Fisher Scientific (Waltham, MA, USA) on blood from the five healthy, anonymous individuals, now referred to as individual 1 to 5.

The depletion occurred mostly in accordance with the manufacturer´s protocol, with the exception of the initial handling of the blood, in which blood plasma was removed prior to additions of antibodies due to the protocol for product number 11151D being used.

More specifically, the depletion was performed by preparing an isolation buffer consisting of Dulbecco´s phosphate-buffered saline (DPBS) 0.1% bovine serum albumin (BSA) and 2mM EDTA, pH 7.4 to use for washing of Dynabeads and the bound cells.

The Dynabeads were vortexed to homogenize the solution before beads were transferred to an Eppendorf tube and washed with isolation buffer. The beads and buffer were mixed by pipetting before the tube was placed in a DynaMag™ -2 magnet (Thermo Fisher Scientific) in which the beads would be drawn to the tube wall allowing the supernatant to be removed while in the rack. The beads were resuspended in isolation buffer and ready for use.

In accordance with the protocol for product number 11151D, isolation buffer was added to cooled whole blood, kept on ice, in a 2:1 ratio before the samples were centrifuged at 600 x g for 10 minutes at 4°C using the Centrifuge 5810 R by Eppendorf. The plasma (top layer) was decanted out before 50 µL CD3 antibodies was added to the remaining blood cell pellet. The volume of antibodies added was adjusted from 37.5 µL to 50 µL because the kit was expired and the impact on its quality was unknown. The tubes mixed by rotation before they were placed at 4°C for 10 minutes.

Thereafter, 4 mL of isolation buffer was added to the samples before they were centrifuged at 350 x g for 15 minutes with no brakes.

Most of the supernatant was aspirated but around 1 cm was left covering the pellet. Next, pre-washed Dynabeads were added to each sample for CD3(+) cells to bind to. To the blood from individuals 1 to 5, 150 µL beads were added. The starting volume of 3 mL was used to calculate the volume, requiring 112.5 µL of beads. This volume was increased due to the expiration date of the kit.

The samples were incubated in room temperature for 15 minutes in a Hulamixer® Sample Mixer (Life Technologies, Thermo Fisher Scientific) where they were tilted and rotated. After incubation the five samples were centrifuged briefly to collect the samples before they were transferred to new polypropylene tubes. 4 mL isolation buffer was added to the old tubes and transferred to the new ones, to wash the old tubes and ensure that all sample material was collected. The samples were vortexed for 2-3 seconds before they were placed in the DynaMag™ -5 (Invitrogen, Thermo Fisher Scientific) magnetic rack for separation of beads and supernatant. The CD3(-) supernatant was transferred to new polypropylene tubes. Because the blood volume from individual 1 to 5 exceeded the height of the magnet, the new tubes containing the CD3(-) supernatant were placed in the magnet rack. The supernatant was transferred to a new tube in order to make sure that no beads were left in the sample.

Roswell Park Memorial Institute (RPMI) medium and Fetal Bovine Serum (FBS) were added to the CD3(-) blood to provide nutrition to the cells until the DNA extraction two days later, in an effort to keep the cells viable.

3.2.2 DNA extraction

The genomic DNA (gDNA) extracted in this thesis was intended for sequencing methods using long fragment templates in the library preparations. Hence, the Monarch® Genomic DNA Purification Kit by New England BioLabs (NEB, Ipswich, MA, USA) was used to extract high molecular weight (HMW) gDNA from whole and CD3 depleted (CD3(-)) blood according to manufacturer's protocol. Fresh whole and CD3(-) blood from individual 1 to 5 were kept in the fridge for two days before extraction, as recommended by NEB. An additional test of DNA quality was performed on freshly CD3 depleted blood, in which HMW gDNA was extracted both immediately following CD3 depletion and two days after from the same CD3 depleted blood sample.

A master mix containing protease K, RNase A and blood lysis buffer was made and added to decrease pipetting, as recommended by the supplier.

All vortexing was performed by pulse-vortexing 5 times using an Analog Vortex Mixer (VWR, Radnor, PA, USA). Thermomixer comfort (Eppendorf, Hamburg, Germany) was used for all sample incubations and pre-heating of elution buffer. For incubation at 56°C, agitation at 1400 rpm was used. Heraeus Pico 17 (Thermo Fisher Scientific) centrifuges was utilized for all centrifugations of CD3(-) samples, while Biofuge fresco (Thermo Fisher Scientific) was used for WB samples. Both centrifuges were used at 3200 rpm (1000 x g) when centrifuging for three minutes and 13700 rpm (12000 x g) or 13000 x g for all centrifugation steps at full speed for one minute, as indicated in the protocol. The tubes were placed in the same direction in all centrifugation steps in order for the sample to move in the same direction, which may influence the final yield.

All samples were eluted in pre-heated (at 60°C) elution buffer. The samples extracted for sequencing were eluted in two aliquots of 100 µL, while 80 µL was used for the additional test.

To ensure more uniform DNA quality, an additional gDNA clean-up step was performed on the DNA extracted for sequencing. The two HMW gDNA aliquots originating from the same

blood were combined prior to the addition of a 0.6x ratio of AMPure PB beads (Pacific Biosciences, Menlo Park, CA, USA). To compare the DNA quality of the samples before and after clean-up, 4 µL of the pre-clean up samples was set aside for quality measurements. The AMPure PB clean up procedure was performed according to the protocol from the supplier and the HMW gDNA was eluted in 80 µL elution buffer (PacBio).

3.2.3 Concentration measurements and quality control of nucleic acids

In this thesis multiple methods were used to check the sample integrity of extracted HMW gDNA, amplicons and sequencing library prior to them being used in downstream experiments or analyses. These methods use different approaches to provide information about concentration and/or quality and purity, and thus complement each other.

Nanodrop® ND-1000 (Thermo Fisher Scientific) measures quality and quantity without requiring any additional reagents. First the elution buffer used in the extraction or clean-up procedures was used as a blanking reference for the instrument, then the samples were added to the sample arm and measured.

In contrast to Nanodrop, the two other nucleic acid measurement methods used in this thesis, Qubit and TapeStation, uses fluorescence to measure the concentration and/or quality of the gDNA or amplicons. Preparation of samples, standards and ladder were performed according to protocol by supplied by the manufacturer for both these methods.

For the Qubit 2.0 Fluorometer (Thermo Fisher Scientific) measurements, all reagents in the Qubit® dsDNA HS (High Sensitivity) Assay kit (Thermo Fisher Scientific) were equilibrated to room temperature before a Qubit working solution was prepared by dilution of Qubit® dsDNA HS Reagent in Qubit® dsDNA HS buffer with a 1:200 ratio. 1 µl of each DNA sample and 10 µl of each of the two supplied Qubit® dsDNA HS standards, used for instrument calibration, were added to separate Qubit® assay tubes (Thermo Fisher Scientific) containing Qubit working solution, bringing the total volume up to 200 µl after addition of the DNA/standards. Next, all the tubes were vortexed for 2-3 seconds using the Analog Vortex Mixer (VWR) and measured after a minimum incubation of 2 minutes.

The 4200 TapeStation system (Agilent, Santa Clara, CA, USA) uses an automated electrophoresis to provide information about concentrations and fragment sizes in each sample. In this thesis the Agilent Genomic DNA ScreenTape System was used. This gives a DNA integrity number (DIN), which report the level of DNA degradation that have occurred in addition to the concentration measurement. Unwanted fragmentation of the gDNA during the depletion or extraction could influence the sequencing of long fragments and reduce the targeted sequencing output and quality.

The Tapestation Genomic DNA protocol was performed as follows: first all reagents and DNA were equilibrated to room temperature for 30 minutes. The Genomic DNA Sample Buffer and Genomic DNA Ladder were vortexed and centrifuged using the Analog Vortex Mixer (VWR) and Mini Star (VWR) centrifuge. 10 µl of genomic DNA sample buffer and 1 µl of each sample were added to a 96-well plate before it was sealed with Microseal ´F´foil seals (Bio-Rad, Hercules, CA, USA). The plate was vortexed for 1 minute at 1800 rpm and spun for about 15-20 seconds by MixMate (Eppendorf) and PCR plate spinner (VWR) prior to insertion into the 4200 TapeStation (Agilent). Genomic ScreenTape (Agilent) and Loading tips (Agilent) were added, and Genomic DNA ladder (Agilent) were mixed as instructed by the TapeStation instrument.

3.3 Sequencing of T cell receptor α (TRA) region

3.3.1 Identification of target region for sequencing

Since the 960 kb TRA region is too large to sequence within the time and economical window of this thesis, it was necessary to choose a smaller region to sequence. Based on the reasoning that functionally important regions, such as enhancers, promoters, genes or variants are conserved between organisms, a 100 kb region covering parts of the variable gene region with regulatory elements present was chosen for sequencing, as presented in Figure 3.1.

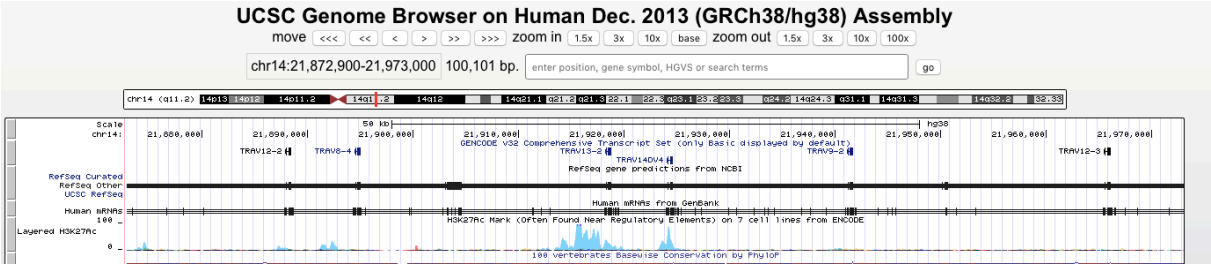


Figure 3.1 Screenshot of the 100kb TRA region to be sequenced as visualized in UCSC Genome Browser. The light blue peaks showing regulatory regions.

To evaluate if a short-read or a long-read sequencing approach is the best to use for this region, one of each was performed. For both sequencing approaches, the region of interest had to be targeted as smaller regions. Hence, primer- and gRNA design for long-range PCR and no-amp targeted sequencing, respectively, was done in such a way that it resulted in 10 or 12 smaller fragments, which all had a little overlap with the adjacent fragments to try to avoid gaps in the overall 100 kb target region. The design process for both methods is described later in this section.

3.3.2 No-amp targeted sequencing using the CRISPR-Cas9 system by PacBio

crRNA design

In order to design the crRNAs needed for targeted sequencing, a previously retrieved sequence from UCSC (with SNP markings) was used in the Genetic Perturbation Platform (GPP) single guide RNA (sgRNA) designer webtool on the Broad Institute GPP Web Portal website (https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-designas) as instructed in the reference guide by PacBio (Part Number 101-839-600 V.02).We divided the 100kb target region into 12 smaller regions (Figure 3.2) and used the CRISPR-Cas9 design tool to design the sense and antisense crRNAs for each of the 12. The suggested crRNAs with the highest combined rank was selected. In addition, the crRNAs were manually checked against the obtained reference sequence to make sure that they were not spanning SNPs and had the correct orientation with the PAM sequence located adjacent to the 3´ end of the crRNA.

The crRNAs used in this thesis are presented in Table 3.2.

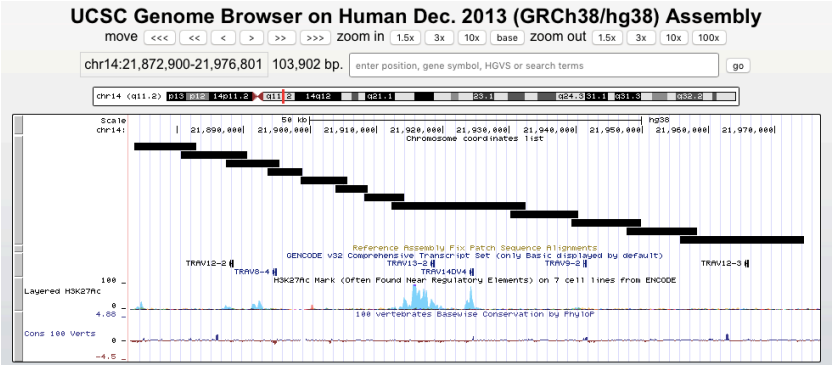


Figure 3.2 Overview of the 12 smaller target regions for no-amp targeted sequencing of the 100 kb within the TRA region using CRISPR-Cas9.

Table 3.2 CRISPR RNA oligonucleotides (crRNA) used for DNA fragmentation with the protospacer adjacent motifs (PAM) and fragment sizes per crRNA pair in base pairs (bp).

crRNA name	Sequence (5´ → 3´)	PAM sequence	Fragment size (bp)
TRA_R34F_gRNA	TGTAATTGAGTAATATCCCT	AGG	9331
TRA_R34R_gRNA	TTGAGGTTGCTATTGCAGGT	TGG	
TRA_R35F_gRNA	GCTTTTGCTACTCAGAGTCG	GGG	9906
TRA_R35R_gRNA	AATTGATATTCAAACGAGCA	TGG	
TRA_R36F_gRNA	CAGAATGGTATAGGGATGTG	TGG	8057
TRA_R36R_gRNA	AAGAACAAAAAAAGGAACAT	GGG	
TRA_R37aF_gRNA	GGCATGATGTATCAGACTGT	AGG	5193
TRA_R37aR_gRNA	CCATTCTAACTGGTGTGAGA	TGG	
TRA_R37bF_gRNA	CCCCATCAAAAAGTGGGCGA	AGG	6967
TRA_R37bR_gRNA	CTCAGGAAGCTGACTGAGGT	GGG	
TRA_R38aF_gRNA	CTCAGCAAGGAACATCCCTG	GGG	4852
TRA_R38aR_gRNA	GGCACACATAAAACCTTCTG	GGG	
TRA_R38bF_gRNA	CTCGACTTAGACATGCACCA	AGG	6026
TRA_R38bR_gRNA	TTACAGAATCATCCTCACAA	TGG	
TRA_R39F_gRNA	AACTCAGGACTGTAGCAAGT	GGG	20143
TRA_R39R_gRNA	AGCCTACATGAGTTATCCTG	AGG	
TRA_R40F_gRNA	AAGCTTTGAATGGTAATGGT	TGG	10248
TRA_R40R_gRNA	GATACATCGGCTGATAATCG	AGG	
TRA_R41F_gRNA	TGTGTGTGTTGAAACAACTT	TGG	10449
TRA_R41R_gRNA	TTTATCCATGAGACTTACTG	CGG	

TRA_R42F_gRNA	TGAAACCAAATAATGCCATG	GGG	10516
TRA_R42R_gRNA	CAGGAGAATACAACAAGCTG	AGG	
TRA_R43F_gRNA	CTTCCTGAGTCAATCTTGGG	AGG	18676
TRA_R43R_gRNA	AGATAGTCACAATAGACACT	GGG	

Preparation of crRNA, tracrRNA and barcoded adapter

The crRNAs (Integrated DNA Technologies, IDT, Coralville, IA, USA) and the universal tracrRNA (IDT) arrived freeze-dried and had to be resuspended to a final concentration of 50 µM in nuclease-free IDTE pH 7.5 buffer (1x TE solution) from IDT prior to use. Following the recommendation of the manufacturer on how to achieve more reliable final concentration, the crRNA where first resuspended to an initial concentration of around 55 µM. The concentration of each crRNA was then measured using Nanodrop ND-1000, and the volume needed to adjust the resuspensions to obtain the final concentration of 50 µM was calculated for each crRNA separately. Notably, due to calculation errors the TRA_R43_F crRNA ended up with a final concentration of around 25 µM, which had to be accounted for later in the experiment. As this was very time-consuming and error-prone, the tracrRNA was resuspended to 50 µM directly. Resuspended crRNA and tracrRNA were stored at -80°C after being aliquoted to a maximum of 10 uses, as recommended by PacBio.

Barcoded adapters ordered from IDT were resuspended in nuclease-free IDTE pH 7.5 buffer to a stock solution of 100 µM before they were diluted in 1x Annealing Buffer (PacBio) and Nuclease-Free Water, not DPEC-Treated (Ambion) to a working stock of 20 µM. Next, the adapters were annealed by incubation using a 2720 Thermal Cycler (Applied Biosystems, Thermo Fisher Scientific) at 95°C for 5 minutes, 25°C for 1 second and hold on 4°C. All handling of the barcodes occurred on ice and the annealed working stocks were stored at -20°C.

For all incubation steps at 16°C, 37°C and 65°C a Thermomixer Comfort (Eppendorf) were used. Heated lids were advised for all incubations, but as this was not possible in the laboratory used, aluminium foil was put on the inside of the lid belonging to the Thermomixer. 1.5 mL DNA LoBind tubes (Eppendorf) were utilized in all steps except for gRNA preparation. Nuclease-Free Water, not DPEC-Treated (Ambion) was used whenever the protocol required water.

Fresh 80% ethanol was prepared and AMPure PB beads (PacBio), Elution Buffer (PacBio) and gDNA were equilibrated to room temperature before use.

gDNA dephosphorylation treatment

First, dephosphorylation of the gDNA was performed to avoid fragment ends generated during DNA extraction from participating in the ligation reaction following the CRISPR-Cas9 digestion step. An input of 600 ng gDNA was used for all samples. At room temperature the following reagents were added in the named order with Nuclease-Free Water not DPEC-Treated (Ambion) first being added to the gDNA, followed by NEBuffer™ 3.1 (NEB) to a final concentration of 1x and then Shrimp Alkaline Phosphatase (rSAP) (NEB) to a final concentration of 0.05 U/µl. The tubes were inverted 22 times before they were spun down in the Mini Star (VWR), incubated at 37°C for one hour and 65°C for 10 minutes, and placed on ice.

Guide RNA (gRNA) preparation

To generate the gRNAs, the working solutions of the crRNAs and tracrRNA were thawed on ice with occasional flicking and quick spins using the Mini Star (VWR). The 24 combined gRNAs were prepared in a 0.2 mL 96-well PCR-plate (Thermo Fisher Scientific) placed on ice as follows. A master mix of tracrRNA and Nuclease-Free Duplex Buffer (IDT) was prepared. 9 µL master mix were transferred to the PCR-plate to which 1 µL of each crRNA were added. As the resuspension of TRA_R43_F resulted in a final concentration of 25 µM, 7 µL Nuclease-Free Duplex Buffer, 1 µL tracrRNA and 2 µL crRNA were added to this well separately in order to maintain a 1:1 concentration ratio of crRNA and tracrRNA. Once everything was added to the wells, it was mixed by pipetting before the plate was spun down briefly using a Heraeus™ Multifuge™ X3 Centrifuge (Thermo Fisher Scientific) and incubated for 5 minutes at 95°C in a 2720 Thermal Cycler (Applied Biosystems, Thermo Fisher Scientific). Once finished, the plate was placed on the bench to cool down and subsequently placed on ice.

From the 24 gRNAs, two different multiplexes were generated, targeting 6 regions each, as we had designed with overlap between adjacent fragments. Twelve gRNAs were combined in two separate DNA LoBind tubes by transferring 9 µL of each gRNA, resulting in a final

concentration of 5 µM for the multiplexed gRNA. The two different multiplexed gRNA mixes were named cut-mix 1 and 2 as presented in Figure 3.3.

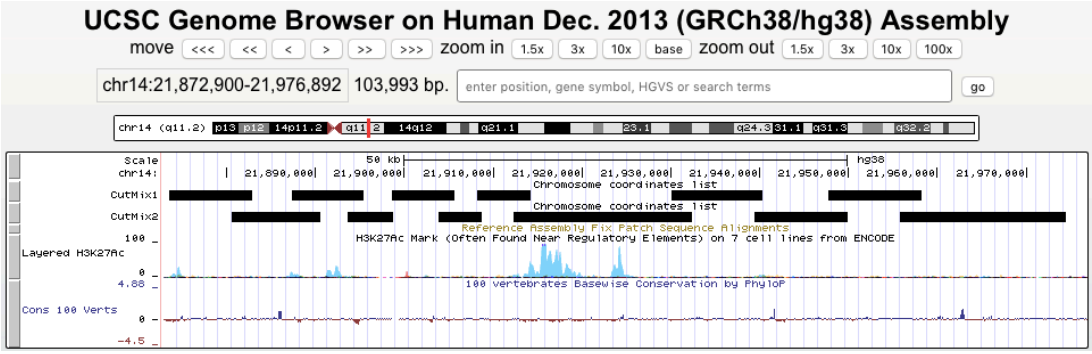


Figure 3.3 Illustration of which fragments the two different cut mixes will result in after performing the CRISPR-Cas9 digestion for the TRA region of interest.

CRISPR-Cas9 digestion and post-digestion

Separate master mixes were prepared for each cut mix, cut-mix 1 and cut-mix 2. Notably, each of the eight different samples from the four individuals had two digestion reactions each. The reagents were added in named order to generate two separate gRNA-cas9 complex mixes as follows: NEBuffer™ 3.1 (NEB) to a concentration of 1x, multiplexed gRNA (cut-mix 1 or 2) to a concentration of 400 nM, Cas9 Nuclease (NEB) to a concentration of 400 nM and Nuclease-Free Water not DPEC-Treated (Ambion). After mixing by pipetting and a brief centrifugation, they were incubated at 37ºC for 10 minutes and then place on ice. Then 20 µL from these two gRNA-cas9 complex mixes were added to sixteen DNA LoBind tubes (eight tubes per cut mix) containing 80 µl of dephosphorylated gDNA. The tubes were carefully inverted 22 times and spun down before they were incubated at 37ºC for 1 hour and then place on ice. 0.5 M EDTA, pH 8.0, Molecular Biology Grade (Millipore, Burlington, MA, USA) was added. The tubes were carefully inverted six times to avoid unspecific fragmentation and quickly spun followed by an AMPure PB bead purification as specified for the PacBio no-amp protocol with: 0.45x volume to sample of room tempered and homogenized AMPure PB beads (PacBio), freshly prepared 80% ethanol with a DynaMag™ -2 Magnet (Invitrogen, Thermo Fischer Scientific) and 1.5 mL DNA LoBind tubes (Eppendorf). For elution of the targeted DNA fragments, 31µl elution buffer was added to each sample. 1 µl of the elution was used to measure the concentration prior to subsequent steps. For the samples WB3 and WB4 in cut-mix 1 and samples CD3(-)2 and CD3(-)4 in cut-mix 2, an extra 1:1 ratio AMPure step was performed on the supernatant removed earlier in the procedure due to the gDNA not being recovered. The eluted gRNA-cas9 digested fragments were kept on ice.

Adapter ligation and post-ligation pooling

The next step in the protocol was to ligate the adapters to the gRNA-cas9 digested fragments. Importantly, as two different multiplexed gRNA-cas9 complex mixes (cut-mix 1 and 2) were used separately on each sample, barcoding was performed per sample (Table 3.3). The pre-prepared barcoded adapters (IDT, final concentration of 0.40 µM) were together with T4 DNA Ligase Reaction Buffer (NEB, final concentration of 1x) added to the DNA LoBind tubes containing the cleaved and purified gRNA-cas9 digested fragments. The tubes were mixed by inversion and spun down before Nuclease-free water and T4 DNA ligase (Thermo Fisher Scientific, final concentration of 0.90 U/µL) were added while on ice and the tubes were inverted 22 times.

Following incubation at 16°C and 65°C, the samples were placed on ice to cool before they were centrifuged for 15 minutes at 14000 x g using the Heraeus Pico 17 centrifuge (Thermo Fisher Scientific). Next the samples were pooled so a multiplexed library would be generated by transferal of around 50 µL of each sample to a new DNA LoBind tube, taking care not to disturb the pellet. The pooled samples where then subject to a 0.45x volume AMPure PB beads (PacBio) clean-up, where the tube was vortexed for 10 minutes at 2000 rpm using an IKA MS3Vortexer to ensure proper binding of DNA to the beads before elution in 200 µL elution buffer.

Once the AMPure purification was done, the SMRTbell library was stored in the fridge until the next day.

Table 3.3 Barcoded adapters used to identify sequences from the four individuals.

Barcoded adapter	Sample
bc1001	WB1
bc1002	WB2
bc1004	WB3
bc1008	WB4
bc1009	CD3(-) 1
bc1010	CD3(-) 2
bc1012	CD3(-) 3
bc1014	CD3(-) 4

Nuclease and trypsin treatment of SMRTbell library

To remove failed ligation products and unwanted gDNA fragments and enzymes, nuclease and trypsin treatments were performed. Both steps were carried out on ice. The nuclease treatment was done by adding the SMRTbell library, Nuclease-Free Water, not DPEC treated (Ambion), CutSmart® Buffer (NEB, final concentration of 1x), Exonuclease III (NEB, final concentration of 2.4 U/µl) and enzymes A, B, C and D from the SMRTbell Enzyme Cleanup Kit (PacBio) to a new DNA LoBind tube, followed by incubation at 37 °C for two hours. Immediately after the incubation the nuclease treated SMRTbell library was placed on ice and 0.5 M EDTA, pH 8.0 Molecular Biology Grade (Millipore, final concentration of 23 mM) and SOLu-Trypsin (Sigma-Aldrich, Darmstadt, Germany, final concentration of 41 µg/mL) were added. Mixing was performed by inversion 22 times, a quick spin (Mini Star (VWR)) and incubation at 37°C for 20 min followed before the tube was placed on ice. The treated SMRTbell library was then immediately transferred to a new DNA LoBind tube at room temperature, to which Elution Buffer (PacBio) was added to bring the total sample volume from 438 µL to 500 µL and the two rounds of AMPure purification were performed using 0.45x volume of AMPure PB beads for the first purification and 0.42x volume of AMPure PB beads for the second purification. The final elution volume was 6 µl.

The PacBio sequencing was performed by the core facility using SMRT sequencing. Hence, 1 kb DNA Ladder (Carrier DNA, NEB) was diluted in Elution Buffer to a final concentration of 50 ng/µl before it was brought to the sequencing core facility at the University of Oslo together with the final SMRTbell library for primer annealing, polymerase binding, sample clean-up and sequencing using the Sequel II system.

3.3.3 Long-range PCR and shot-gun sequencing using Illumina MiSeq

Long-range PCR was performed to generate targeted amplicons for subsequent shot-gun sequencing using Miseq (Illumina, San Diego, CA, USA) sequencer. The primer pairs were previously designed using the website Primer3web v.4.1.0 (https://primer3.ut.ee/) to divide the 100 kb TRA region of interest into ten smaller and overlapping fragments with sizes ranging from 8 kb to around 20 kb, as visualized in Figure 3.4. The primers that fulfilled certain criteria (i.e not including SNPs, have similar T_m as the other primer in the same pair) were tested in an in silico PCR (http://genome.ucsc.edu/cgi-bin/hgPcr) to check if they only

targeted the correct region. Successfully designed primer pairs were ordered from Eurofins Scientific (Luxembourg) and are presented in Table 3.4.

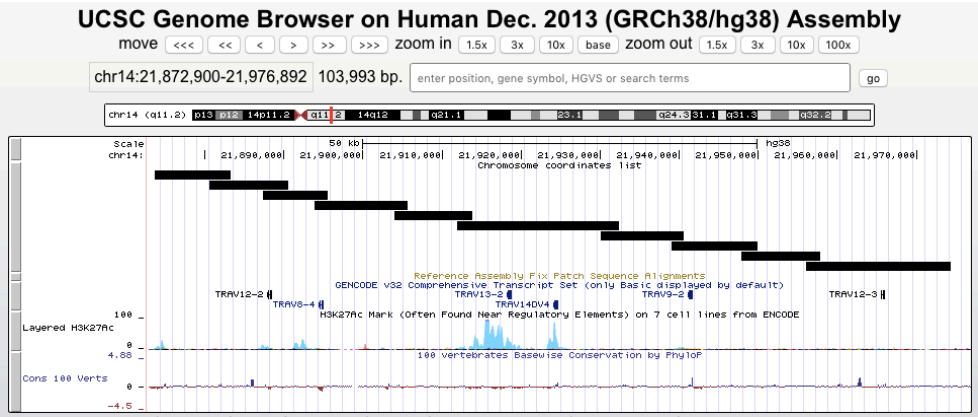


Figure 3.4 Overview of wanted amplicons for sequencing of 100 kb within the TRA region.

Table 3.4 Primer pairs for amplification of TRA fragments to be sequenced. F and R in the primer name indicates forward and revers primer, respectively and together constitute one primer pair.

Primer pair name	**Sequence (5´ → 3´)**	**Tm (°C)**	**Product size (bp)**
TRA_R34Fprimer	TATGGCTCATTTGGGGTCCC	59.4	9524
TRA_R34Rprimer	GGTGTCTGGACTGTCTTGGG	61.4	
TRA_R35Fprimer	TTTGCTACTCAGAGTCGGGG	59.4	9898
TRA_R35Rprimer	ATTCAAACGAGCATGGCAGC	57.3	
TRA_R36Fprimer	GTGGCCTGACATGAAGAGCT	59.4	8160
TRA_R36Rprimer	TGTACATTGCCATCTTCAGAGCA	58.9	
TRA_R37Fprimer	TCACCCACCGTAATGCTGAC	59.4	11718
TRA_R37Rprimer	AGCTGACTGAGGTGGGAGAA	59.4	
TRA_R38Fprimer	TCAATGACAATGGTGCCCGA	57.3	9819
TRA_R38Rprimer	AAGGAAATGGAGGCGGTCAG	59.4	
TRA_R39Fprimer	TCCCTTCATCCTCAGAGCCA	59.4	20528
TRA_R39Rprimer	TCTGCTGTCACAACACCTGA	57.3	
TRA_R40Fprimer	CCAGACCTGTTCCACCATCC	61.4	10469
TRA_R40Rprimer	GGACGCTTTCTACTGTGGTCA	59.8	
TRA_R41Fprimer	CTTTTGCCTGCTTGCTCTGG	59.4	10849
TRA_R41Rprimer	CCCAAAACTCACAGCACAGC	59.4	
TRA_R42Fprimer	GGAGAGGTGGTGAAGTGAGC	61.4	9977
TRA_R42Rprimer	GGAAGGGGCTGGTCATCTAC	61.4	
TRA_R43Fprimer	GGGGTGGGTTTACTCTTGGG	61.4	18279
TRA_R43Rprimer	CACAACCTGAGCGACAATGC	59.4	

PCR optimization

PCR optimization was performed to identify the best conditions for each primer pair to produce successful and specific amplification with only one amplicon of expected size for each target region. Both HMW gDNA from WB and CD3(-) blood were used in the optimization due to the concern that rearranged gDNA from T cells in the whole blood samples could affect the amplification of the germline HMW gDNA.

Two different long-range PCR kits were tested. Reactions with the LongAmp *Taq* PCR Kit (NEB) had to be set up on ice while the reactions with UltraRun LongRange PCR Kit (Qiagen, Hilden, Germany) could be set up at room temperature as its polymerase needs heat-activation above 93°C to become active. For all ten primer-pairs master mixes were made by scaling the reagents according to Table 3.5 for the LongAmp Taq PCR Kit master mix and Table 3.6 for the UltraRun LongRange PCR Kit without Q-solution. The concentrations used were according to the suppliers´ recommendations. The gDNA, with concentrations ranging from 10 to 49 ng/µl, was added to the reaction master mix in order to ensure uniform concentrations between the different annealing temperatures used in the PCR optimization. For all master mixes, the tube used was inverted a few times to carefully mix all reagents without causing unnecessary fragmentation to the HMW gDNA. Next, the liquid was collected by a brief spin using MiniStar (VWR) and pipetted onto a 96-well PCR plate which was sealed and quickly spun down. Additionally, two amplicons (TRA_R36 and TRA_R43) were also set up with the UltraRun LongRange PCR Kit with 1x and 1/2x Q-solution that affect the DNA melting behaviour, Table 3.7 and Table 3.8.

Table 3.5 Reaction setup for one sample with LongAmp Taq *PCR Kit (NEB).*

Reagent	12 µL reaction	Final concentration
5x LongAmp *Taq* reaction buffer	2.4 µL	1X
10 mM dNTPs	0.36 µL	300 µM
10 µM Forward primer	0.48 µL	0.4 µM
10 µM Reverse primer	0.48 µL	0.4 µM
LongAmp *Taq* DNA polymerase	0.48 µL	5 units/ 50 µL
Nuclease-free water	5.8 µL	
HMW gDNA (10 – 49 ng/µl)	2.0 µL	

Table 3.6 PCR setup for UltraRun LongRange PCR Kit (Qiagen) without Q-solution.

Reagent	10 µL reaction	Final concentration
UltraRun LongRange PCR Master Mix, 4x	2.5 µL	1X
10 µM Forward primer	0.5 µL	0.5 µM
10 µM Reverse primer	0.5 µL	0.5 µM
Nuclease-free water	4.5 µL	
HMW gDNA (10 ng/µl)	2.0 µL	

The kit from Qiagen contained three optional components – Master Mix Tracer, Sample Tracer and Q-solution. The Master Mix Tracer and Sample Tracer are dyes which offer helpful visualization while pipetting as well as working as loading dyes during electrophoresis by running at about 50 bp and 4000 bp respectively. Both tracers were used during PCR optimization. Prior to use, they were diluted to a final concentration of 1x. 4 µL Master Mix Tracer (125x) was added to 500 µL UltraRun LongRange PCR Master Mix (4x) and 0.2 µL Sample Tracer (25x) was added per 5 µL DNA solution - DNA solution being DNA diluted in Nuclease-Free Water to a final concentration of 10 ng/µL.

Table 3.7 PCR setup for UltraRun LongRange PCR Kit (Qiagen) with 1x Q-solution.

Reagent	10 µL reaction	Final concentration
UltraRun LongRange PCR Master Mix, 4x	2.5 µL	1x
10 µM Forward primer	0.5 µL	0.5 µM
10 µM Reverse primer	0.5 µL	0.5 µM
Q-solution	2.0 µL	1x
Nuclease-free water	2.5 µL	
HMW gDNA (10 ng/µl)	2.0 µL	

Table 3.8 PCR setup for UltraRun LongRange PCR Kit (Qiagen) with 1/2x Q-solution.

Reagent	10 µL reaction	Final concentration
UltraRun LongRange PCR Master Mix, 4x	2.5 µL	1x
10 µM Forward primer	0.5 µL	0.5 µM
10 µM Reverse primer	0.5 µL	0.5 µM
Q-solution	1.0 µL	1/2x
Nuclease-free water	3.5 µL	
HMW gDNA (10 ng/µl)	2.0 µL	

To find the optimal melting temperature (T_m) for each primer pair, annealing temperatures starting at 58°C covering every degree including 63°C were tested for all pairs. This was done

because the theoretic T_m may not always be the optimal one. The elongation time were adjusted to each primer pair according to their expected fragment sizes and the polymerases´ elongation speed to make sure that the polymerase had enough time to synthesize the strands. The initial PCR programs tested for the NEB and Qiagen kit are presented in Table 3.9 and Table 3.10, respectively. Notably, the reactions set up with the LongAmp *Taq* PCR Kit needed a hot-start on the PCR machine, which means the machine was heated to over 90°C prior to placing the reactions. The initial number of cycles of 30 were later changed to 35 to see if it would increase the amount of amplicon product.

Table 3.9 PCR program as presented in the LongAmp Taq PCR Kit (NEB) protocol. The annealing temperature gradient was used during PCR optimization to find the optimal temperature for each primer pair. The elongation time was adjusted according to expected amplicon size ranging from 8.1kb to 20.5 kb.

Step	Temperature	Time	Number of cycles
Initial denaturation	94°C	30 sec	
Denaturation	94°C	15 sec	X 30-35
Annealing	58°C - 63°C	15 sec	
Elongation	65°C	8,5 – 17 min	
Final extension	65°C	10 min	
	4°C	Hold	

Table 3.10 PCR program for optimization of primer pairs with UltraRun LongRange PCR Kit (Qiagen). The annealing temperature gradient was used during PCR optimization to find the optimal temperature for each primer pair. The elongation time was adjusted according to expected amplicon size ranging from 8.1kb to 20.5 kb.

Step	Temperature	Time	Number of cycles
Initial denaturation	93°C	3 min	
Denaturation	93°C	30 sec	X 30-35
Annealing	58°C - 63°C	15 sec	
Elongation	68°C	4 - 10 min	
Final extension	72°C	10 min	
	4°C	Hold	

To try and increase the primer specificity and product for unspecific amplifications, touch down PCR was attempted. This means an extra cycling step was added prior to the ordinary cycles, which included 5-10 additional cycles with higher annealing temperatures than the initial annealing temperature (Table 3.11). These additional cycles started off 3°C or 5°C higher than the annealing temperature(s) and decreased by 0.5°C or 1°C per cycle until it reached the ordinary annealing temperature and the ordinary cycles started. The number of

extra cycles added was dependent on the initial staring temperature of the samples and by how much the temperature decreased per cycle.

Table 3.11 Example of a touchdown PCR tested for UltraRun LongRange PCR Kit (Qiagen). The program has two sets of cycles where the first one is added as an attempt to increase primer specificity. The annealing temperature of the first cycles is 5°C higher (61°C-68°C) than the primer pair´s optimal annealing temperature (56°C-63°C) and decreases 0.5°C per cycle for 10 cycles until it reaches the optimal annealing temperature. The elongation time is adjusted according to expected amplicon size.

Step	Temperature	Time	Number of cycles
Initial denaturation	93°C	3 min	
Denaturation	93°C	30 sec	X 10
Annealing	Start +5°C of annealing temp. (- 0,5°C per cycle)	15 sec	
Elongation	68°C	4 – 10 min	
Denaturation	93°C	30 sec	X 35
Annealing	56°C - 63°C	15 sec	
Elongation	68°C	4 – 10 min	
Final extension	72°C	10 min	
	4°C	Hold	

Long-range PCR amplification of samples for sequencing

Following PCR optimization were the kits, temperatures and PCR programs presented in Table 3.12 used for amplification with each of the 10 primer pairs.

Table 3.12 Presentation of the combination of PCR kit, temperature and PCR program used for amplification of fragments for sequencing. These conditions provided the most PCR product for each amplicon. The reactions performed with the kit from Qiagen were without Q-solution.

Primer pair	Kit	Temperature	PCR program
TRA_R34	NEB	63°C	According to protocol (35 cycles)
TRA_R35	NEB	63°C	According to protocol (35 cycles)
TRA_R36	NEB	63°C	According to protocol (35 cycles)
TRA_R37	NEB	63°C	According to protocol (35 cycles)
TRA_R38	NEB	62°C	According to protocol (35 cycles)
TRA_R39	Qiagen	61°C	According to protocol (35 cycles)
TRA_R40	NEB	61°C	According to protocol (35 cycles)
TRA_R41	Qiagen	62°C	TD +5°C, -0.5°C/cycle in 10 cycles
TRA_R42	NEB	62°C	TD +3°C, -0.5°C/cycle in 5 cycles
TRA_R43	Qiagen	60°C	TD +5°C, -0.5°C/cycle in 10 cycles

Gel electrophoresis

For separation of the amplicons in each sample, 0.5% and 0.6% agarose gel were made using SeaKem® LE Agarose (Lonza, Basel, Switzerland), 1x TAE buffer (Invitrogen, Thermo Fisher Scientific) and stained with Gel Red Nucleic Acid Stain (Biotium, Fremont, CA, USA, final concentration of 1x). The samples were mixed with 6x DNA loading dye (Fermentas, Waltham, MA, USA) to a final concentration of 1x. Two ladders were used, GeneRuler 1kb DNA ladder (Fermentas) and 1kb plus DNA ladder (Invitrogen, Thermo Fisher Scientific). While the GeneRuler 1kb was ready-to-use, the 1kb plus ladder had to be pre-mixed with nuclease-free water (Qiagen) and 6x Gel Loading Dye (NEB) or 6x DNA loading buffer (Fermentas) to a final concentration if 1x. ImageQuant LAS 4000 (GE Healthcare, Chicago, IL, USA) and ImageQuant™ TL 1D v8.1 software (GE Healthcare) were used to visualize amplicons.

Extraction of PCR-products from agarose gel

As most of the primer pairs tested were unspecific (i.e resulted in more than the one band of interest), the bands of expected size were cut out and purified from agarose gels using Roche´s High Pure PCR-product Purification Kit (Basel, Switzerland). The extraction occurred mostly according to protocol, except that the step in which isopropanol was added was skipped because fragments larger than 30 kb have previously been successfully purified using a protocol without its addition (Stucka, 2005).

After addition of binding buffer to the excised gel, the tubes were pulse-vortexed 10 to 15 times before incubation at 56°C for 10 minutes using thermomixer comfort (Eppendorf). The tubes were vortexed every third minute during the incubation. All centrifugation was done using the Heraeus Pico 17 centrifuge (Thermo Fisher Scientific) at 17.0 x g for 40 seconds during binding, and 1 minute when the samples were washed and finally eluted in 60 µL elution buffer.

The concentration after gel purification was measured using Qubit. Some amplicons were not extracted from gel and therefore not included for further analyses. Table 3.13 shows the amplicons that were extracted and purified for each sample.

Table 3.13 Samples successfully extracted from agarose gel per amplicon. X marks missing samples.

Amplicon	WB1	WB2	WB3	WB4	CD3(-) 1	CD3(-) 2	CD3(-) 3	CD3(-) 4
R34	X				X			
R35							X	
R36								
R37								
R38			X				X	
R39								X
R40	X		X					
R41								
R42								
R43								

Clean-up and up concentration of PCR-product

Due to low sample concentrations following amplicon extraction from gel, an additional clean-up step using a 1:1 ratio of sample and AMPure XP beads (Beckman Coulter, Brea, CA, USA) was performed on all amplicons. Samples were incubated for 10-15 minutes after the beads had been added. The beads were separated from the supernatant for 5 minutes using magnetic rack before the supernatant was removed and the beads were washed with freshly prepared 80% ethanol twice. Beads were then air dried for a few minutes, prior to elusion with 12 µL nuclease-free water (Qiagen). The eluate was transferred to a 96-well PCR plate and concentration measured using Qubit was performed for three out of the eight samples. Care was taken into measuring samples that were successfully extracted for all twelve amplicons, which only were the case for WB2, WB4 and CD3(-) 2. The three concentrations should in theory be representative for all eight samples amplified with the same primer pair because the eight samples (WB1-4 and CD3(-) 1-4) had been through the same process of amplification and purification from gel. The average concentration measurements for these three samples was used to identify the pooling volumes to be used per amplicon.

Illumina Miseq library preparation

The Illumina Miseq instrument cannot sequence large fragments, hence as part of the library preparation the long-range PCR amplicons had to be fragmented into smaller fragments of 500-800 bp in size.

The NGSgo-LibrX (GenDx, Utrecht, The Netherlands) library preparation kit, was used to generate a sequencing library compatible with Illumina sequencing. First, all amplicons per sample were pooled based on the average amplicon measurements obtained after the gel purification. Optimally the input of pooled samples should be around 250 ng, but due to low concentrations and most samples missing amplicons the input ended up being between 50-60 ng, which was within the range of the kit.

A 2720 Thermal Cycler (Thermo Fisher Scientific) was used for all PCRs and incubation steps that did not occur in room temperature, and Heraeus Megafuge 8 Centrifuge (Thermo Fisher Scientific) was utilized for all spins. All steps except clean-up using solid phase reversible immobilization (SPRI) beads were carried out on ice.

A master mix for enzymatic fragmentation, end repair and dA-tailing was prepared according to Table 3.14. In a new 96-well PCR plate, 8.5 µL of master mix was added to 24 µL of the pooled samples. Notably, the WB3 sample had a volume of 21.5 µL after pooling of the amplicons due to it missing amplicons TRA_R38 and TRA_R40. Thus, an additional 3.5 µL of nuclease-free water was added to this sample.

Table 3.14 NGSgo master mix prepared for fragmentation, end repair and dA-tailing of pooled samples.

Reagent	Volume (µL) for 10 samples
NGSgo-LibrX Fragmentase Buffer	20.0
NGSgo-LibrX End Prep Buffer	32.5
NGSgo-LibrX Fragmentase Enzyme	15.0
NGSgo-LibrX End Prep Enzyme	15.0
Nuclease-free water	2.5

The plate was sealed, vortexed and centrifuged to properly mix and collect the samples before the plate was placed in the PCR machine for incubation: 25°C for 20 minutes, 70°C for 10 minutes and ramp down to 15°C. When the temperature reached 15°C the plate was placed on ice while the master mix for adapter ligation was prepared according to Table 3.15.

Table 3.15 NGSgo master mix for adapter ligation.

Reagent	Volume (µL) for 11 samples
NGSgo-LibrX Ligase Mix	82.5
NGSgo-LibrX Ligation Enhancer	5.5
NGSgo-LibrX Adapter for Illumina (AD-IL)	2.75
Nuclease-free Water	11.0

9.25 µL of the ligase master mix was added to the same eight wells used in the previous step while the plate was still on ice. The plate was sealed, vortexed and spun down briefly before the plate was incubated at 20°C for 15 minutes.

A clean-up using SPRI beads was carried out room temperature. 18.8 µL room tempered and resuspended SPRI beads were added to each sample, keeping a 0.45x beads:sample ratio. The plate was sealed and vortexed to make sure that the solutions were homogenized, before incubation for 5 minutes.

The plate was then spun down briefly and placed in Dynamag -96 Side Magnet (Thermo Fischer Scientific) for five minutes, allowing the beads to separate fully from the supernatant before it was removed. To wash the beads, 200 µL freshly prepared 80% ethanol was added to the wells while the plate was still on the magnet. After a 30 second incubation, the ethanol was carefully removed. This wash step was performed a total of three times. After removing remaining ethanol, the beads were allowed to dry for around three-four minutes in the magnet before the plate was removed from the magnet and 12.5 µL elution buffer (0.1x TE) was added. Again, the plate was sealed and vortexed until the solution was homogenized and left for incubation in room temperature for two minutes. The plate was placed in the magnet after a quick spin.

Simultaneously as the clean-up was performed, 1.25 µL i5 and i7 indexes were added to eight wells in a new 96-well PCR plate placed on ice, as shown in Table 3.16 to give the samples dual indexing. Next, 12.5 µL NGSgo-LibrX HiFi PCR Mix and 10 µL of the eluates from the clean-up were added before the plate was sealed, vortexed and spun down briefly. The plate was then placed in the 2720 Thermal Mixer with heated lid for an indexing PCR, which program is presented in Table 3.17.

Table 3.16 Indexes added to each of the eight samples for dual indexing.

Sample	Indexes used
WB1	IN-IL-501 and IN-IL-701
WB2	IN-IL-502 and IN-IL-701
WB3	IN-IL-503 and IN-IL-701
WB4	IN-IL-504 and IN-IL-701
CD3(-) 1	IN-IL-505 and IN-IL-701
CD3(-) 2	IN-IL-506 and IN-IL-701
CD3(-) 3	IN-IL-507 and IN-IL-701
CD3(-) 4	IN-IL-508 and IN-IL-701

Table 3.17 NGSgo indexing PCR program

Step	Temperature	Duration	
Initial denaturation	98°C	30 sec	
Denaturation	98°C	10 sec	X10
Annealing	65°C	30 sec	
Elongation	72°C	30 sec	
Final extension	72°C	5 min	
	15°C	Hold	

The plate with the now dual indexed samples was spun briefly before 15 µL of each sample were transferred to a new Eppendorf tube for size-selection using a 0.6x ratio of SPRI beads and washing the beads with 800 µl fresh 80% ethanol twice. The pooled library was eluted using 66 µL elution buffer (0.1x TE). 60 µL of the eluted and pooled library was transferred to a new Sarstedt tube. QC prior to the sequencing was performed by measuring 1 µL of the library on Nanodrop to identify the library concentration. Additionally, 4 µL library was mixed with 0.8 µL 6x loading dye and loaded on a 1% agarose gel, which ran for one hour on 90V to identify the library size and lack of primer dimers. The library was of adequate quality for it to be sequenced using Illumina Miseq 300 MICRO at the core facility located at the Department of Medical Genetics (https://www.sequencing.uio.no/).

3.4 Genotyping of Norwegian ME patients and controls

3.4.1 SNP selection

The main focus was to identify SNPs either previously associated with other AIDs or ME/CFS or were located in regulatory regions. SNPs within the TRA region were therefore evaluated and selected for genotyping based on 1) previously published papers with ME/CFS associations, 2) TRA associations in AIDs published in the online GWAS catalog, 3) location in TRA regulatory regions as visualized in UCSC Genome Browser, in addition to 4) minor allele frequencies (MAF) between 5 and 50%.

As a result of this research, four biallelic SNPs not previously genotyped in our Norwegian ME/CFS cohort were chosen for genotyping in this thesis: rs5742831 was found associated with Crohn´s disease (O'Donnell et al., 2019), rs11157573 and rs17255510 based on previously association with ME/CFS in an American study of 80 individuals (Schlauch et al., 2016) and rs35379740 located in the recombination signal sequence of the *TRAV14DV4* gene.

3.4.2 Allelic discrimination using Taqman assays

To genotype the ME/CFS cases and healthy controls, four different Taqman assays were ordered. These were pre-designed and functionally tested by the manufacturer, Thermo Fisher Scientific, and are listed together with their respective SNP IDs and dyes in Table 3.18.

Table 3.18 Overview of 40x TaqMan® SNP Genotyping Assays with their SNP number and corresponding alleles visualized by the VIC and FAM dyes.

Assay	SNP ID	VIC	FAM
C_25962246_10	rs35379740	G	T
C_44756205_10	rs5742831	A	C
C_34374423_10	rs17255510	C	T
C_32009806_10	rs11157573	T	C

The TaqMan reaction mixes were prepared in room temperature according to Table 3.19 with an excess volume of 10%. The reagents were mixed by pipetting before 4 µl of the reaction mix were transferred to a MicroAmp® Optical 384-Well Reaction Plate with Barcode

(Applied Biosystems, Thermo Fisher Scientific). To avoid technical bias between cases and controls, all plates had a mix of cases and controls as far as possible.

Table 3.19 TaqMan master mix set up for one sample and a full 384-well reaction plate.

Reagent	Mix for one sample (µl)	Mix for 422 samples (µl)
2X TaqPath™ ProAmp™	2.5	1055
40X TaqMan® Genotyping	0.1	42.2
Nuclease-Free Water	1.4	590.8
Total	4	1688.0

1 µl DNA (5 ng/µL) from each of the 408 Norwegian ME patients (REK 2015/1547, 2014/365, 2010/473) and 721 healthy controls (S-04279) was then pipetted into the wells with reaction mix according to a pre-defined plate setup before the plate was sealed by MicroAmp™ Optical Adhesive Film (Applied Biosystems, Thermo Fisher Scientific), centrifuged briefly and inserted into the QuantStudio™ 12K Flex (Applied Biosystems, Thermo Fisher Scientific) or GeneAmp® PCR System 9700 (Applied Biosystems, Thermo Fisher Scientific) for amplification with the program shown in Table 3.20. QuantStudio™ 12K Flex was used for post-reading of all plates.

Table 3.20 PCR program for amplification in TaqMan genotyping with TaqPath™ ProAmp™ Master Mix (Thermo Fisher Scientific). [1] Polymerase activation was mostly set to 10 minutes, although the master mix required 5 min activation. The length did not seem to make a difference.

Step	Temperature	Time	Number of cycles
Pre-read	60°C	30 sec	
Polymerase activation	95°C	10[1] min	
Denaturation	95°C	15 sec	X 40
Annealing	60°C	1 min	
Post-read	60°C	30 sec	

Notably, an additional setup was used for the Taqman assay C_44756205_10 since it performed poorly with the previous. The TaqMan® Universal Master Mix II, no UNG (Applied Biosystems, Thermo Fisher Scientific) was therefore tested and as it resulted in better allelic discrimination for this assay, all samples were run using this set up for C_44756205_10. The reaction setup was carried out the same way as for the others but the PCR program (Table 3.21) used for amplification did not have a pre-read stage.

Tabell 3.21 PCR program for amplification in TaqMan genotyping with TaqMan® Universal Master Mix II, no UNG (Applied Biosystems, Thermo Fisher Scientific).

Step	Temperature	Time	Number of cycles
Polymerase activation	95°C	10 min	
Denaturation	95°C	15 sec	X 40
Annealing	60°C	1 min	
Post-read	60°C	30 sec	

3.5 Bioinformatical online tools and software used

1000 Genomes project

The 1000 Genomes (1000G) is an international project in which the goal was to identify most genetic variations in human populations. It ran from 2008 to 2015 and characterized more than 88 million variants (including SNPs and structural variants (SVs)) from 2504 unrelated individuals from 26 populations (Genomes Project et al., 2015). Sequencing has continued after the project ended and have increased the number of individuals included to 3202 (The International Genome Sample Resource, 2020).

UCSC Genome Browser

The Genome Bioinformatics Group within the UCSC Genomics Institute developed this graphical viewing tool in order for the genome and its content to be available for everyone (http://genome.ucsc.edu/). The browser was used to inspect the TRA region of interest and SNP locations. The reference sequence of the TRA region was downloaded for USCS Genome Browser and used for gRNA- and primer design. Furthermore, their in silico PCR tool was used before ordering primers and the custom tracks to visualize expected fragments for both long-range PCR and Cas9-fragmentation.

ENSEMBL

Ensembl (https://www.ensembl.org/index.html) is a genome browser containing genome annotations and offers tools to find information about variation, regulation and conservation between vertebrates. One such tool is the VCP to PED converter (http://www.ensembl.org/Homo_sapiens/Tools/VcftoPed) which gives marker information and linkage pedigree files that can be opened in Haploview and was in this work used to extract SNP genotyping data for the 1000G human Caucasian of European descent (CEU)

population for the whole TRA region (chr14:21401841-23030791, GRCh38) and the 100 kb TRA region to be sequenced (chr14:21872900-21973910, GRCh38).

Genome-Wide Association Studies (GWAS) Catalog

The online database GWAS catalog (https://www.ebi.ac.uk/gwas/) was created by the National Human Genome Research Institute (NHGRI) and was used to search for publicly available information about human GWAS in autoimmune diseases and ME/CFS to identify potential SNPs in the TRA region for genotyping. As of September 2020, the database is comprising 4694 publications and 197708 associations between variations and phenotypes (traits and diseases).

Haploview

In this thesis Haploview (Barrett, Fry, Maller, & Daly, 2005) was used to visualize how much genetic variation in the TRA region the genotyped SNPs covered. Since it is a bioinformatics software used to analyse and estimate genetic patterns by looking at the linkage disequilibrium (LD) between variations it was also used to visualize the LD patterns.

PLINK

To handle and analyse SNP data PLINK v1.9 (Purcell et al., 2007) was used in this thesis. PLINK is a free genome association analysis toolset used to perform a variety of analyses on genotype and phenotype data. Hence, it was used to extract the region of interest (chr14:21,870,000-23,500,000, GRCh37) from the previously generated Ichip genotyping dataset and to recode the binary files into ped- and map- files prior to performing an association analyses for the ME/CFS patients and controls. PLINK was also used to perform Hardy-Weinberg equilibrium testing, calculating allele frequencies and identify missing genotypes. It was further used to split the genotyping dataset into separate files for cases and controls so that subsequent linkage disequilibrium analyses could be performed using Haploview.

LDlink

LDlink (https://ldlink.nci.nih.gov/) offers a number of web applications that can be used to investigate linkage disequilibrium in populations. Three application that was used in this thesis are LDtrait, LDmatrix and SNPchip Tool. LDtrait was used to investigate SNPs associated with a phenotype in the GWAS catalog, while LDmatrix visualized SNPs on the

chromosome and measured LD between them based on publicly available data. SNPchip Tool was used to look at different genotyping arrays´ SNP coverage in TRA.

RStudio

RStudio, a software built on the programming language R used for data management, analyses and visualization. Users can develop and share different programming packages written in R. Examples of such R packages used in this thesis are tidyverse and dplyr for data management and ggplot2 for visualization of analyses.

Unphased

Unphased v3.0.13 (Dudbridge, 2008) is an analyses software used for performing genomic association and haplotype analysis in ME/CFS cases and healthy controls.

BWA and Bowtie2 aligners

In order to see where in the genome sequencing reads belong, they need to be mapped against the human reference genome. The Burrows-Wheeler Alignment Tool (BWA)(Li & Durbin, 2010b) is a software package for mapping of reads which uses an algorithm, BWA-MEM, to perform local alignment on high-quality queries. It is suitable for the more error prone long-reads and could therefore be used on both the PacBio and the Illumina sequencing reads in this thesis. Bowtie2 (Langmead & Salzberg, 2012) is a commonly used tool for short-read sequencing reads and can perform gapped, local and pair-end alignments, and was only used for the Illumina sequencing reads. All samples sequenced with both PacBio and Illumina were aligned using BWA version 0.07.17, while the eight samples sequenced with Illumina short-reads were aligned using Bowtie2 by adjusting scripts previously generated by my main-supervisor using TSD at the University of Oslo. They both generated SAM files as output for subsequent analyses.

SAMtools

After generating SAM files with BWA or Bowtie2 aligners, SAMtools v1.8 was used to manipulate the alignments to sort the alignments (Li et al., 2009) presented in the SAM and generating BAM files – the latter being the binary equivalent to the former.

IGV

To visualize the aligned and sorted BAM files generated in this study, we used the Integrative Genomics Viewer version 2.9.4 (IGV)(Robinson et al., 2011).

4 Results

4.1 SNP coverage and LD patterns in the T cell receptor α (TRA) region

To evaluate the SNP coverage we used two different datasets, the Ichip dataset for Norwegian ME/CFS cases and controls (Hajdarevic et al., Unpublished) and 1000 genomes Caucasian of European descent (CEU) dataset obtained from Ensembl.

The Ichip and 1000G datasets cover 27 and 8958 SNPs in the TRA region (chr14:21401841-23030791, GRCh38), respectively. Haploview software had limitations when it comes to file input sizes, so all variants included in the 1000G dataset could therefore not be visualized, and, subsequently, for this dataset the SNPs in the ~100 kb region (chr14:21872900-21973910, GRCh38) selected for sequencing were explored further. While only two SNPs are covered by Ichip in this region, 1000G covers 737 SNPs. The LD between the SNPs in the latter dataset is shown in the plot found in Figure 4.1, where each small square in the plots show the measured LD between two SNPs. The darker the shade of colour is (red for D' or black for r^2), the stronger the LD is between them, meaning that they are seen more frequently together than expected by chance.

Within the plot, four distinct blocks with high LD are visible, as indicated with black lines. Each of these cover SNPs located close to each other on the chromosome, which is rather expected because there is less chance of recombination occurring between them. There is otherwise not much LD in this ~100 kb region.

Using SNPchip tool to investigate how well represented these 737 SNPs located in the ~100 kb region are on 59 Illumina arrays and 22 Affymetrix arrays, showed that 455 of them are not covered by any genotyping arrays (data not shown). As ichip only includes 27 SNPs in the whole TRA region, it is clear that region is not adequately covered.

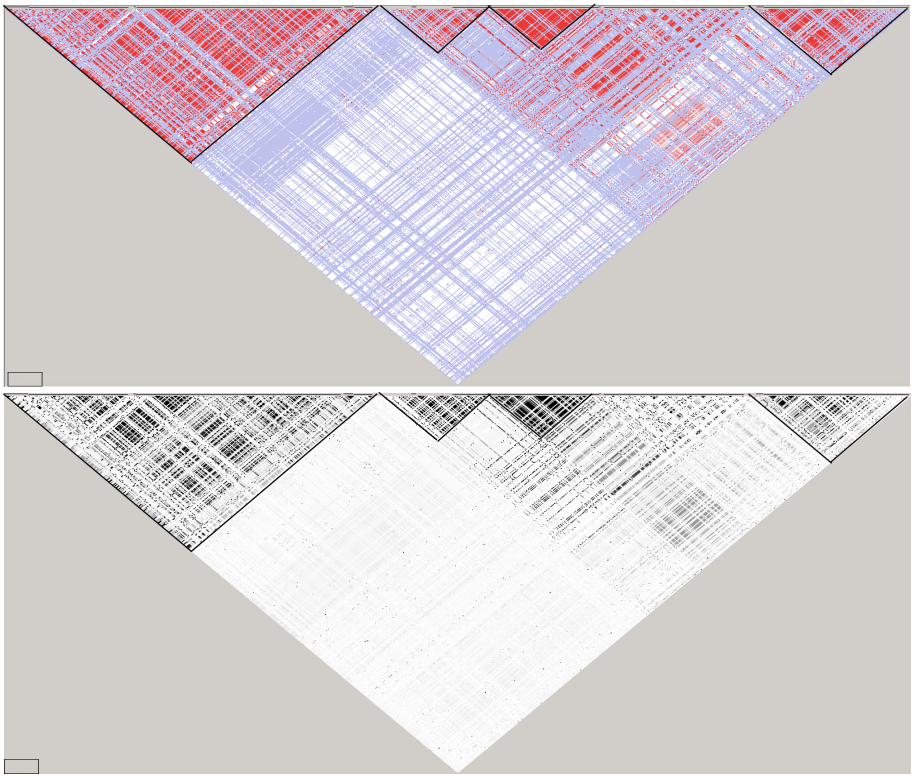


Figure 4.1 LD plot of 737 SNPs in a ~100 kb kb TRA region obtained from 1000G CEU. Measured as D´ (top) and r^2 (bottom). Four distinct LD blocks can be seen, as indicated with the black lines.

4.2 Association analysis of TRA SNPs in ME/CFS

4.2.1 Evaluation of four SNPs selected for genotyping

Four SNPs, presented in Table 4.1, were chosen for genotyping in the 408 ME patients and 721 healthy Norwegians in addition to the 27 SNPs previously genotyped using ichip. LD analyses based on publicly available data from the 1000 Genome project (CEU) on the four SNPs, showed weak LD between all SNPs when measured in both r^2 and D´. The strongest LD was measured between rs17255510 and rs11157573 with D´=0.669, as shown in Figure 4.2.

*Table 4.1 SNPs chosen for genotyping with TaqMan assay. * MAF in 1000G CEU*

SNP ID	Location (GRCh38)	MAF*	Reference
rs35379740	14:21924662	5.1%	
rs5742831	14:22570239		O'Donnell et al., 2019 (Supplementary table 4)
rs17255510	14:22194962	23.2%	Schlauch et al., 2016
rs11157573	14:22420786	19.7	Schlauch et al., 2016

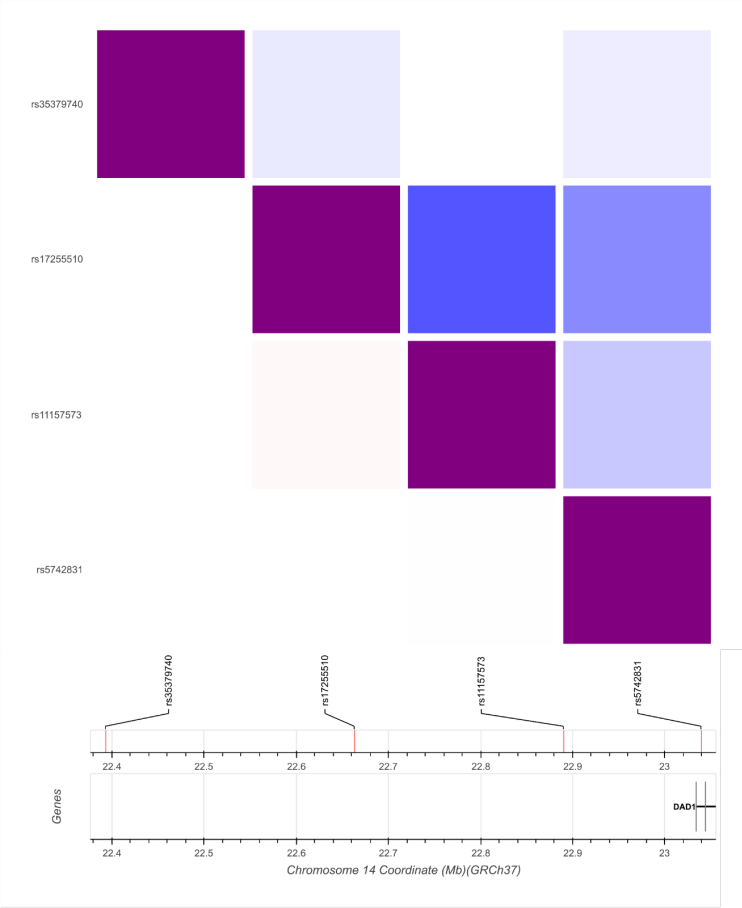


Figure 4.2 LD plot and positions on chromosome 14 (GRCh37) of the four SNPs for genotyping. Blue: LD measured as D′, red: LD measured as r^2. Brighter colour indicates stronger LD. The strongest LD is found between rs17255510 and rs11157573, while the weakest LD is measured between rs35379740 and the other SNPs.

4.2.2 Genotyping quality control

The quality of the genotyping clustering is important to obtain good genotyping results. The genotyping clustering is visualized in allelic discrimination plots (Figure 4.3), where each dot in the plot is indicating the genotype of an individual based on the dye(s) bound to the DNA during the analyses. Three out of four assays gave good clustering (Figure 4.3 B-D), however one SNP assay (rs5742831), had low fluorescence levels and bad clustering (Figure 4.3 A, Figure 4.4 A). Using an alternative master mix for this assay first indicated that it would give better results (Figure 4.4 B). However, the clustering and fluorescence still proved to be poor when applied to the other samples, thus resulting in high numbers of undetermined genotyping calls samples and in turn causing the genotypes for this SNP to be unreliable and not included in subsequent analyses.

For the other three SNPs genotyped in this project (rs35379740, rs17255510 and rs11157573) a genotyping success rate of >98% was obtained both for all three assays combined and separately. For each of the 1129 individuals, a success rate of minimum 67% was achieved, allowing maximum one of the three assays to have failed to assign a genotype. This corresponded to 19 (1.68%), 12 (1.06%) and 14 (1.24%) missed genotypes for rs11157573, rs17255510 and rs35379740, respectively.

The genotyping data obtained in this thesis was added to the genotyping data for the 27 SNPs previously genotyped using Ichip, making it a total of 30 SNPs to be included in association analyses. Across all 30 SNPs a total genotyping success rate of >99.8% was achieved for the 408 ME/CFS cases and 721 controls, and they proved to be in Hardy-Weinberg equilibrium, meaning that the observed heterozygosity is close to the expected one for both cases and controls.

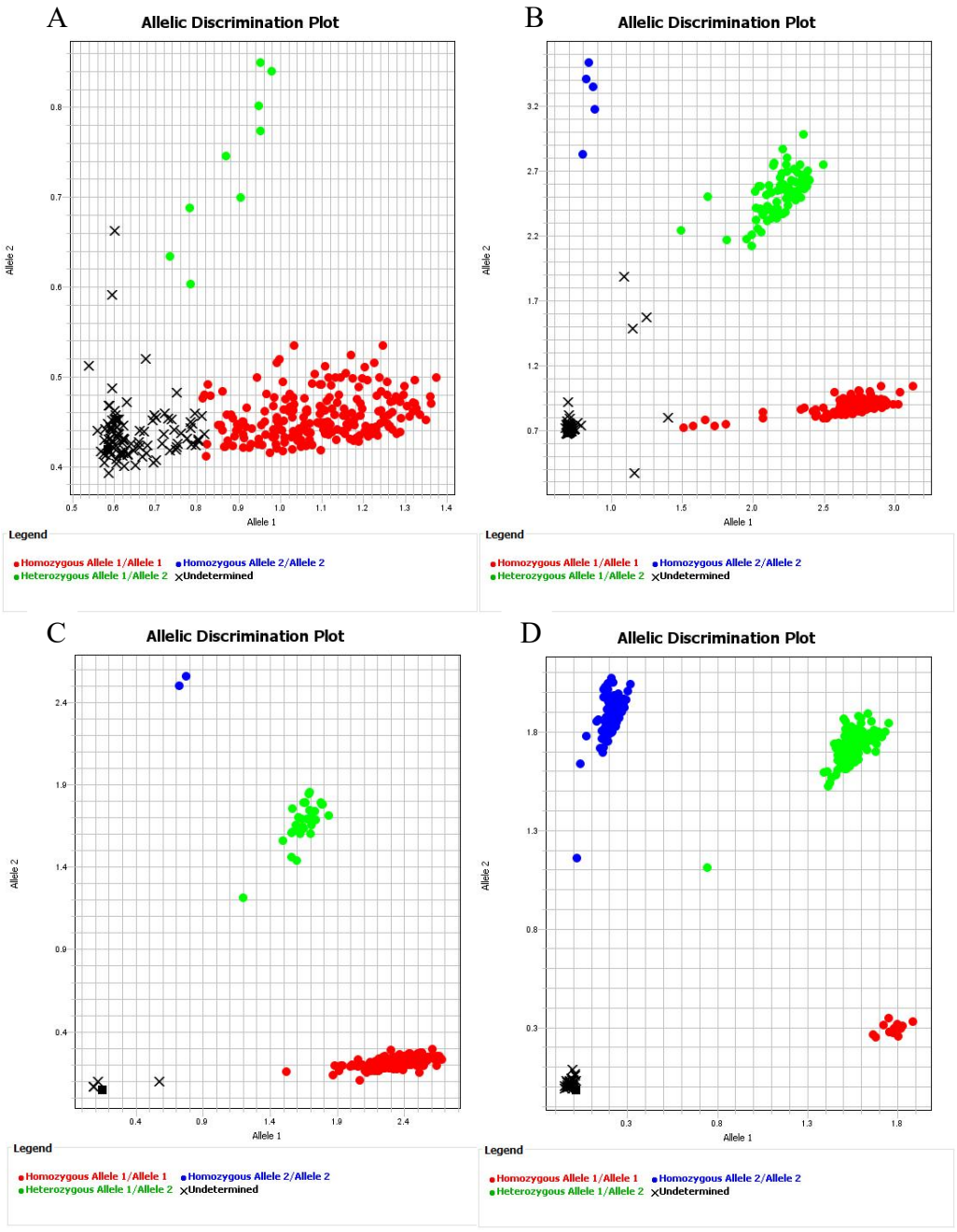


Figure 4.3 Allelic discrimination (AD) plots of the four SNPs investigated in a Norwegian cohort with ME/CFS cases and healthy controls. Red and blue dots correspond to individuals being homozygous for allele 1 and 2 respectively, while green indicates heterozygous. X indicates undetermined genotype. Samples excluded in analyses are not excluded in these plots. A-C are showing genotypes called for the same individuals. A) rs5742831 amplified in QuantStuio 12K Flex (Thermo Fisher Scientific) with TaqMan Universal Master Mix II, no UNG. B) rs11157573 amplified in GeneAmp® PCR System 9700 (Applied Biosystems) and read using QuantStudio. C) rs17255510 amplified using QuantStudio. D) rs35379740 amplified and read in QuantStudio.

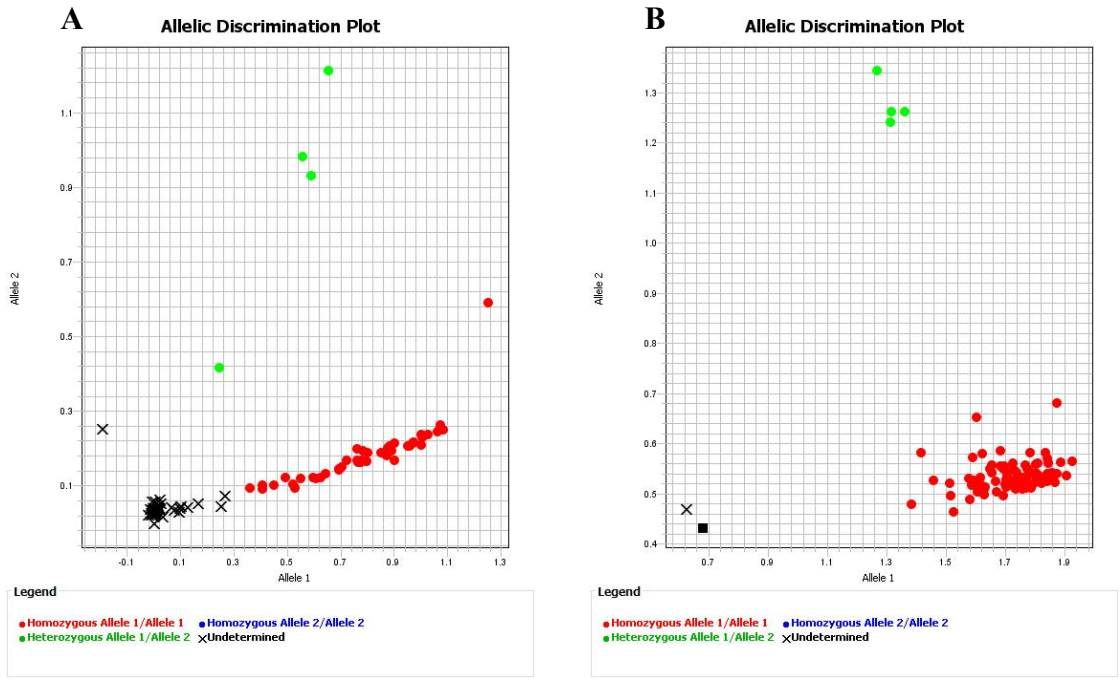


Figure 4.4 Allelic discrimination plot of 94 controls with A) TaqPath master mix (Thermo Fisher Scientific) and B) TaqMan universal master mix II, no UNG (Thermo Fisher Scientific).

4.2.3 Association analyses of TRA SNPs in ME/CFS

The association study did not show any significant findings for any of the SNPs as p>0.05 (Table 4.2), and the allele frequencies did not differ significantly between cases and controls.

Table 4.2 Association analysis between SNPs in the TRA region and ME/CFS.

SNP ID	Minor allele frequency (cases)	Minor allele frequency (controls)	p-value	OR
rs2297093	0.449	0.459	0.64	0.96
rs1263811	0.115	0.116	0.96	0.99
rs45612332	0.042	0.037	0.61	1.13
rs4429208	0.069	0.055	0.21	1.27
rs12897433	0.069	0.056	0.24	1.25
rs17255021	0.116	0.104	0.37	1.14
rs7157659	0.067	0.053	0.19	1.29
rs2293707	0.069	0.059	0.40	1.17
rs11157268	0.115	0.109	0.66	1.07
rs35379740	0.054	0.053	0.92	1.02
rs2031068	0.201	0.197	0.84	1.02
rs2031070	0.199	0.197	0.88	1.02
rs3811315	0.114	0.099	0.30	1.17
rs8022660	0.297	0.302	0.80	0.98

rs12891256	0.479	0.474	0.80	1.02
rs17183131	0.020	0.027	0.32	0.73
rs11628824	0.414	0.42	0.79	0.98
rs17255510	0.22	0.246	0.20	0.87
rs2001022	0.048	0.039	0.34	1.24
rs11626312	0.074	0.055	0.09	1.37
rs11157573	0.183	0.162	0.22	1.16
rs8021297	0.573	0.519	0.05	0.83
rs2254272	0.321	0.284	0.08	1.19
rs1154155	0.196	0.181	0.39	1.11
rs1263655	0.241	0.261	0.34	0.90
rs1263656	0.337	0.335	0.91	1.01
rs8572	0.131	0.137	0.70	0.95
rs11622421	0.357	0.369	0.57	0.95
rs12879543	0.359	0.376	0.44	0.93
rs8005677	0.423	0.42	0.91	1.01

OR= odds ratio. P>0.05 indicates no significant association with ME/CFS.

LD plots of the 30 SNPs in cases (Figure 4.5) and controls (Figure 4.6) showed the same LD pattern. The three SNPs genotyped in this work (coloured green) were not in strong LD with each other, which corresponds to the observation in the 1000G CEU dataset (Figure 4.2). However, rs11157573 is in complete and almost complete LD with the two neighbouring SNPs (rs11626312 and rs80211297) in controls (D′=1 and D′=0.97, respectively) and in strong LD in cases (D′=0.75 and D′=0.91).

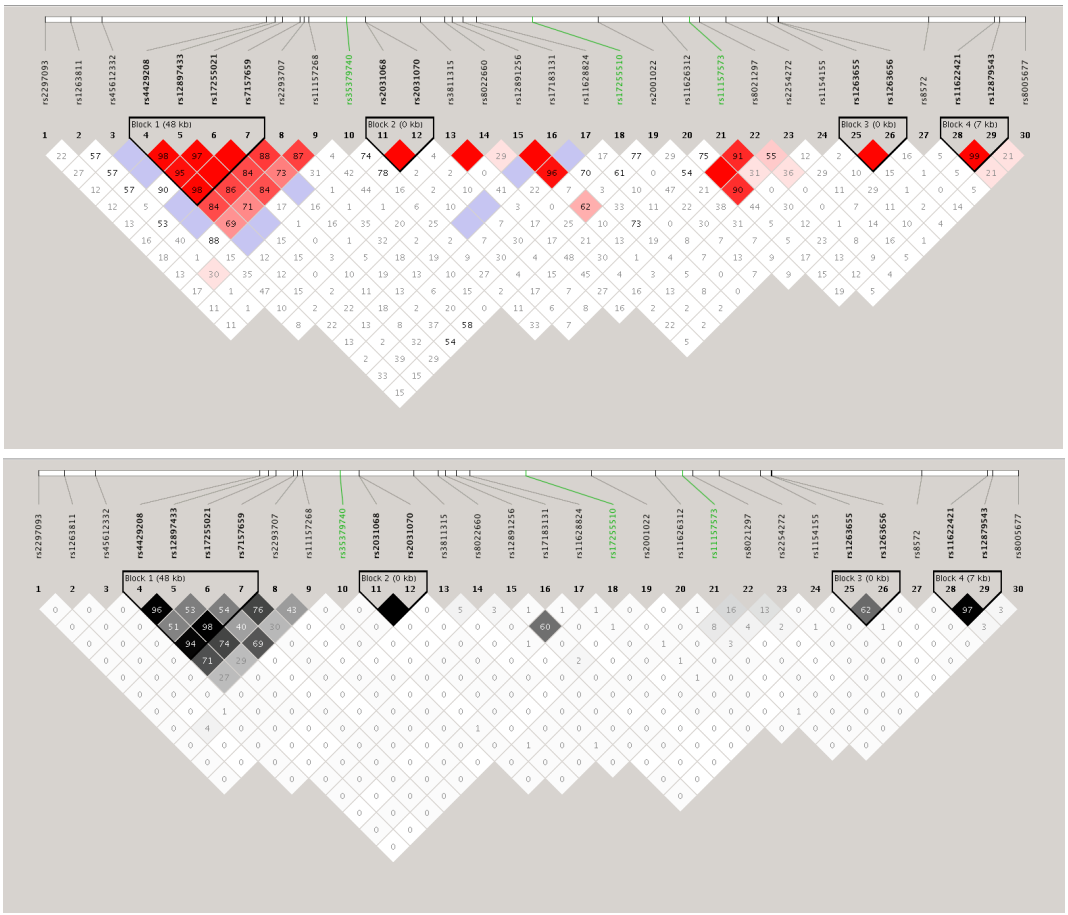


Figure 4.5 LD plots of genotyped SNPs in 408 ME/CFS-patients measured in D´ (top) and r^2 (bottom). SNPs coloured in green were genotyped with TaqMan Assay.

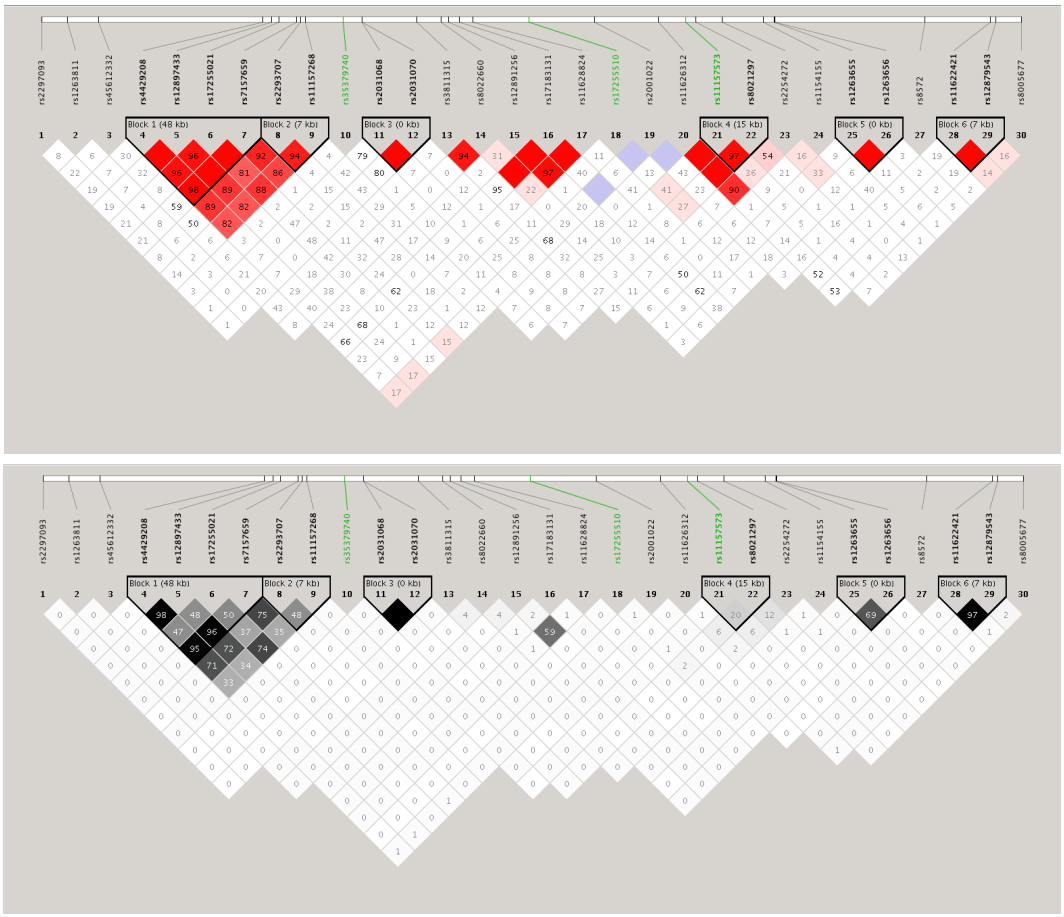


Figure 4.6 LD plots of genotyped SNPs in 721 healthy controls measured in D´ (top) and r^2 (bottom).

To compare the LD pattern of Norwegian ME/CFS cases and controls with the 1000G CEU dataset, plots with 27 out of the 30 SNPs genotyped in the Norwegian ME/CFS dataset were generated Figure 4.7. The fact that the 1000G data show the same LD pattern as our genotyping data, suggests that the 1000G dataset is representative for the Norwegian population and can be used to evaluate the LD pattern within the whole TRA region.

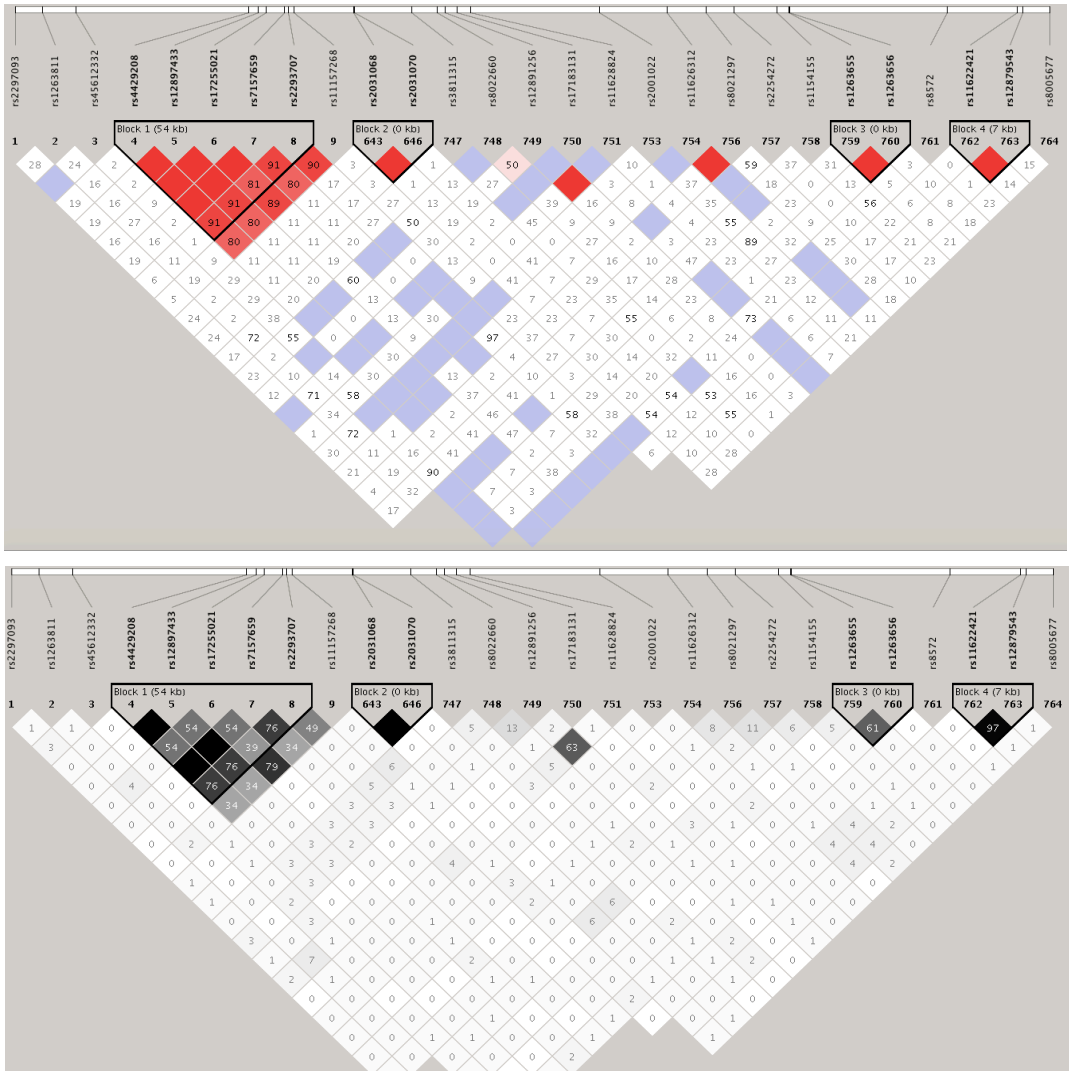


Figure 4.7 LD plots of the 27 SNPs covered by Ichip, here visualized by 1000G CEU data. LD measured in D' (top) and r^2 (bottom).

4.3 Sequencing of TRA region

Two different sequencing strategies were used to investigate genetic variation in ~100 kb of the TRA region (chr14:21,872,900-21,973,000 (GRChr38)), which included regulatory region and several *TRAV* genes as shown in Figure 3.1.

4.3.1 The quality of extracted high molecular weight genomic DNA (HMW gDNA)

First step towards sequence long fragments, was to ensure good quality HMW gDNA. The measured HMW gDNA quality and quantity after extraction with Monarch® Genomic DNA

Extraction kit (NEB) resulted in relatively clean samples but varying and somewhat low concentrations for some of the samples when measured on Nanodrop® ND-1000 (Thermo Fisher Scentific)(Table 4.3). The lowest concentrations (around 10 ng/µL) were measured for the two aliquots from CD3(-) 3, while the highest concentration was measured to be 58.43 ng/µL for WB4:1. Concentrations below 15 ng/µL could not be used as input for PacBio no-amp CRISPR/Cas9 targeted sequencing because the input volume would exceed what the protocol would allow.

All samples except two had a $A_{260/230}$ ratio above 1.8, indicating DNA of high purity. This is further supported by many of the samples having an A260/280 ratio above 2, also indicating pure DNA sample. The samples with a lower $A_{260/230}$ ratio may be contaminated by small proteins, salts, aromatic compounds or haemoglobin.

Table 4.3 Measured gDNA quality and quantity on Nanodrop® ND-1000 (Thermo Fisher Scientific) after HMW DNA extraction from fresh blood. WB=whole blood, CD3(-)=CD3 depleted blood. The number in front of the colon refers to the five anonymous individuals, the number after is the aliquot number.

Sample	Concentration (ng/µL)	$A_{260/280}$	$A_{260/230}$
WB1:1	14.55	1.8	2.52
WB2:1	24.88	1.91	2.48
WB3:1	13.5	1.85	2.19
WB4:1	58.43	1.84	2.07
WB5:1	17.2	1.99	2.53
WB1:2	15.27	1.81	0.85
WB2:2	21.27	2.19	2.04
WB3:2	16.66	1.93	1.57
WB4:2	36.12	1.93	1.49
WB5:2	36.17	1.68	1.11
CD3(-) 1:1	24.4	1.86	2.48
CD3(-) 2:1	19.95	1.93	1.44
CD3(-) 3:1	9.58	2	0.81
CD3(-) 4:1	15.66	2	2.43
CD3(-) 5:1	16.71	1.94	2.93
CD3(-) 1:2	15.17	1.9	2.14
CD3(-) 2:2	21.72	1.91	2.13
CD3(-) 3:2	11.41	2.04	1.67
CD3(-) 4:2	18.13	1.95	2.44
CD3(-) 5:2	17.24	1.87	0.6

Pooling and clean-up of aliquots extracted from the same blood sample resulted in increased concentrations for most samples (Table4.4), bringing it up to at least the minimum concentration needed for sequencing (15 ng/µl) for all samples except CD3(-) 5, which decreased during the process. The absorbance ratios, especially $A_{260/230}$, were also more uniform after clean-up and only one sample (CD3(-) 4) showed indications of contaminants.

Table 4.4 Sample quality and quantity after pooling and cleanup as measured by NanoDrop. Two aliquots of each sample except PCRopt (WB5) had been pooled and purified by AMPure PB beads (PacBio) and eluted in 80 µL Elution Buffer (PacBio) before measuring. PCRopt(WB5) had was made by combining four aliquots and purified. Sample input is 1 µL for all samples.

Sample	Concentration	$A_{260/280}$	$A_{260/230}$
WB1	18.46	1.91	2.26
WB2	31.46	1.84	2.3
WB3	17.35	2.09	2.61
WB4	32.56	1.8	2.35
WB5	18.97	2.04	2.36
CD3(-) 1	38.41	1.87	2.15
CD3(-) 2	30.52	1.93	2.19
CD3(-) 3	16.68	1.89	2.08
CD3(-) 4	24.67	1.7	1.71
CD3(-) 5	8.35	1.66	2.45
PCRopt (WB5)	49.0	1.92	2.35

Next, the TapeStation gel pictures (Figure 4.8) and electropherograms (Figure 4.9) showed that HMW gDNA (>50 kb) of good quality was obtained for the samples, which was necessary to be able to sequence fragments up to 20 kb. The DNA integrity number (DIN) and smeared bands indicated that more sample degradation had occurred for CD3(-) samples than for WB, which is also seen on the broader peaks in the electropherograms. The concentration measurements (Table 4.5) clearly show an increase in the concentration after purification with AMPure PB beads (PacBio) which correspond to the Nanodrop measurements (Table 4.4). The increase in concentration is even visually apparent when comparing the band and sample intensities before and after clean-up (indicated as NP and P, respectively). Taken together, there is a higher concentration of HMW gDNA (>50kb) after clean-up for all samples except for WB1. However, the DIN was not affected by the extra purification step (Table 4.5),

The other electropherograms can be found in appendix II.

Table 4.5 Sample quality and quantity after pooling and cleanup as measured by TapeStation.

	gDNA concentration ng/µl		DIN	
Sample	**Not purified**	**Purified**	**Not purified**	**Purified**
WB1	21.30	18.70	9.6	9.4
WB2	12.90	39.60	9.5	9.4
WB3	10.60	21.00	9.5	9.7
WB4	5.50	32.40	9.1	9.4
WB5	7.98	25.60	9.5	9.4
CD3-1	6.79	23.90	8.3	8.1
CD3-2	16.20	37.00	7.4	7.5
CD3-3	3.15	15.60	8.5	7.8
CD3-4	6.29	22.70	7.8	7.4
CD3-5	11.20	10.40	7.8	7.8

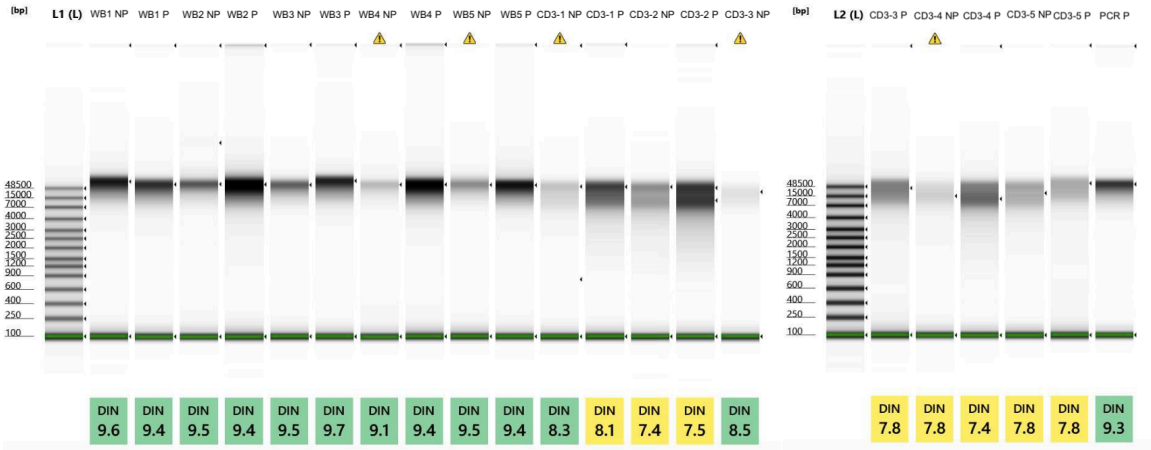


Figure 4.8 Gel picture showing fragment sizes and DNA integrity number (DIN) of samples before and after clean-up. The picture obtained by using the 4200 TapeStation system (Agilent). L1 and L2 contain ladders, WB= whole blood, CD3 = CD3 depleted blood, NP=not purified, P=purified. The samples are loaded so that the two samples from det same blood type from an individual are next to each other, NP first. The higher the DIN is, the less DNA degradation has occurred.

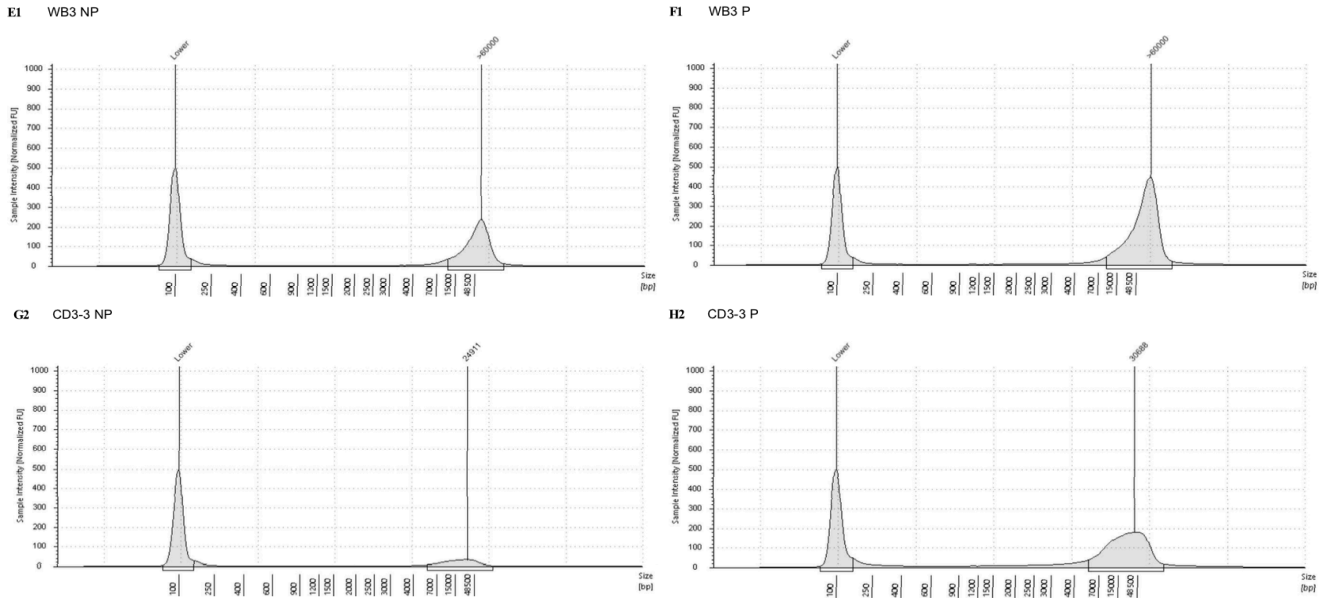


Figure 4.9 Electropherogram showing sample intensity of gDNA extracted from whole blood (WB) and CD3 depleted (CD3(-)) blood from individual 3 measured by the 4200 TapeStation System. NP = not purified, P = purified.

HMW DNA extraction of two aliquots directly following (DE) and one aliquot two days after (E48t) CD3 depletion both resulted in HMW DNA of good quality with most of the fragments being >60 kb for all three samples. The electropherogram peaks and gel bands of the CD3(-) DE samples were more intense than for E48t, as seen in figure 4.10, indicating that more large fragments were extracted directly after.

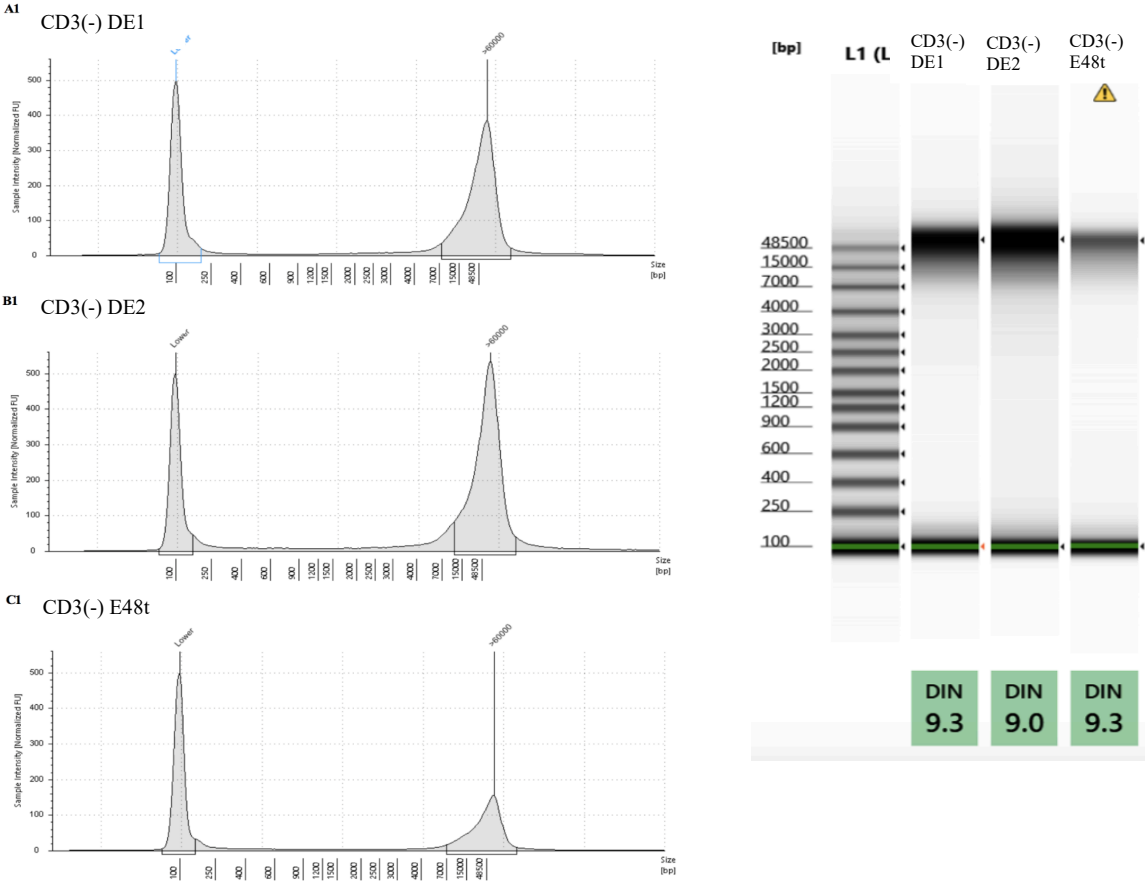


Figure 4.10 Fragment sizes and DNA integrity of two aliquots directly after (DE) CD3 depletion and one aliquot two days after (E48t) gDNA.

4.3.2 Optimization of amplicons prior to short-read sequencing

All primer pairs resulted in amplicons of expected length, although all but two amplicons (TRA_R36 and TRA_R43) had additional bands for all the conditions tested during optimization. The primer specificity improved with increased temperatures for some of the prime pairs, as exemplified for TRA_R43 (expected product size of ~18 kb) in Figure 4.11 A, which at the same time decreased the amount of product. Figure 4.11 B on the other hand, shows an increase in the amount of product for all temperatures without Q-solution, while the addition of Q-solution did not result in amplicon of expected size. Extra cycles added to the PCR program and increased temperature further increased the amount of product for samples amplified without Q solution as seen in Figure 4.11 C.

The conditions giving the highest amount of PCR products of the expected size for each primer pair, were chosen for amplification of the eight samples (WB1-4 and CD3(-) 1-4), even if they resulted in unspecific binding. Therefore, in order to sequence the whole 100 kb

region, only the bands with expected size per amplicon was extracted from gel. As the eluted amplicons gave low amplicon concentrations, subsequent clean-up step resulted in a doubling of concentrations as presented in Table 4.6.

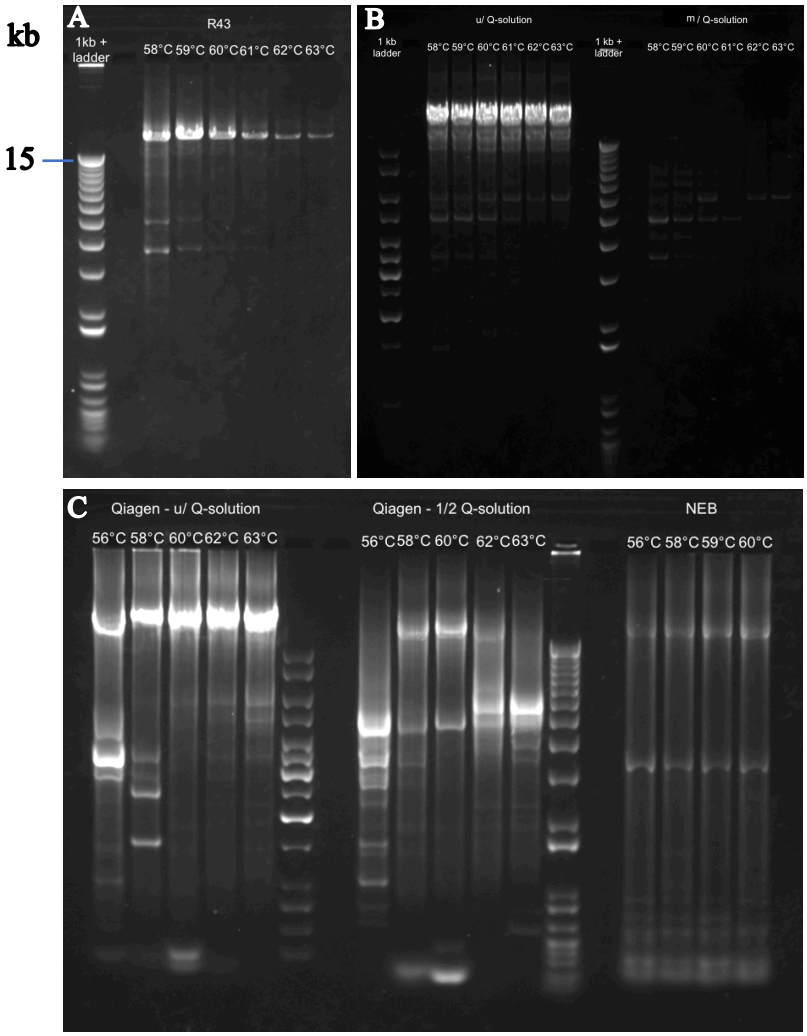


Figure 4.11 Gel electrophoresis pictures following amplification of TRA_R43 with expected product size of about 18 kb. A) gDNA from WB5 (18 ng/µL) amplified using LongAmp Taq polymerase kit (NEB) and PCR-program according to protocol. B) gDNA from CD3(-) blood (10 ng/µL) amplified using UltraRun LongRange PCR kit (Qiagen) with and without Q-solution and PCR program according to protocol with 35 cycles. C) Touch-down (TD) PCR of gDNA from CD3(-) blood (10 ng/µL) using both NEB and Qiagen (without and with ½ Q-solution). TD annealing temperature was +5°C above annealing temperature, decreasing 0,5°C/cycle for 10 cycles.

Table 4.6 Concentration of amplicons measured on Qubit prior to sample pooling. Measured before and after AMPure XP PCR purification

Sample	Concentration prior to clean-up (ng/µl)	Concentration after clean-up (ng/µl)
TRA_R34 WB2	0.908	1.41
TRA_R34 WB4	1.53	5.04
TRA_R34 CD3(-) 2	2.86	6.30
TRA_R35 WB2	0.584	1.39
TRA_R35 WB4	0.430	0.916
TRA_R35 CD3(-) 2	0.232	0.580
TRA_R36 WB2	6.02	20.2
TRA_R36 WB4	7.30	25.8
TRA_R36 CD3(-) 2	7.24	27.0
TRA_R37 WB2	0.888	1.92
TRA_R37 WB4	0.874	2.22
TRA_R37 CD3(-) 2	0.560	1.73
TRA_R38 WB2	0.360	0.740
TRA_R38 WB4	0.272	0.576
TRA_R38 CD3(-) 2	0.130	0.316
TRA_R39 WB2	3.94	9.32
TRA_R39 WB4	3.40	8.48
TRA_R39 CD3(-) 2	4.14	9.70
TRA_R40 WB2	Too low to measure	0.264
TRA_R40 WB4	Too low to measure	0.322
TRA_R40 CD3(-) 2	Too low to measure	0.162
TRA_R41 WB2	2.88	8.30
TRA_R41 WB4	1.85	5.28
TRA_R41 CD3(-) 2	1.59	5.40
TRA_R42 WB2	2.08	6.56
TRA_R42 WB4	1.71	5.46
TRA_R42 CD3(-) 2	1.17	4.58
TRA_R43 WB2	2.58	8.98
TRA_R43 WB4	4.60	14.5
TRA_R43 CD3(-) 2	3.90	11.9

4.3.3 Short-read sequencing on an Illumina Miseq

Despite the low input concentration of amplicons with expected length was the number of reads generated per sample ranging from 104,545 to 1,281,586, which was pretty good. These

reads were aligned against the human reference genome (GRChr37) and visualized in IGV (Figure 4.12) where the reads depth per amplicon could be estimated, as shown in Table 4.7.

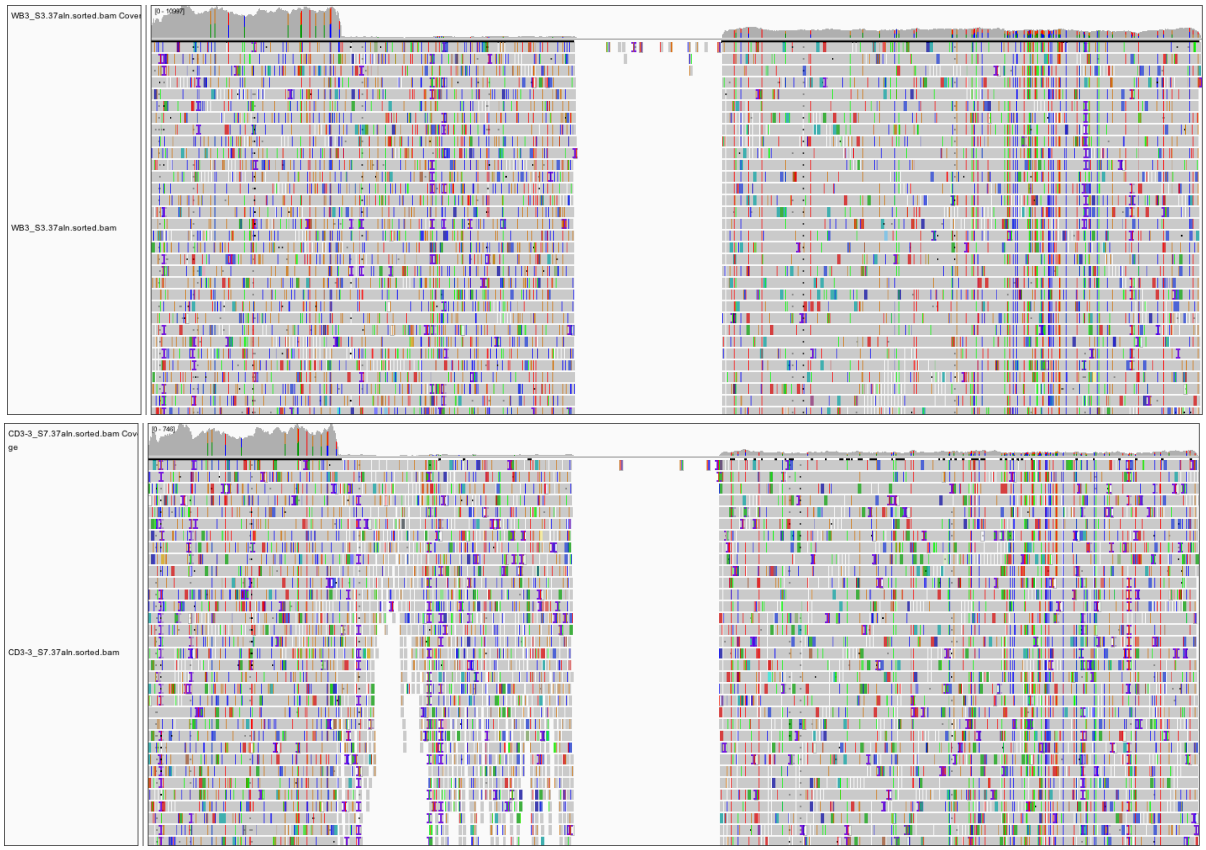


Figure 4.12 Visualization of reads from fragments TRA_R36 to TRA_R39 in WB3 (upper panel) and CD3(-) 3 (lower panel) aligned against the human reference genome (GRCh37) as shown in IGV 2.9.4.

*Table 4.7 Approximate of maximum amplicon read depth for both gDNA from whole blood (WB) and CD3 depleted blood (CD3(-)) for all four sequenced individuals as seen in IGV. * = amplicon not extracted from gel, # = no clear band but extracted from gel, • = the sequence has a small section where read depth is 0.*

		Read depth							
Amplicon	Position Chr14 (build37)	WB1	CD3(-) 1	WB2	CD3(-) 2	WB3	CD3(-) 3	WB4	CD3(-) 4
TRA_R34	22,341,867-22,351,390	*0	*0	510	2070	1110	36	1800	1700
TRA_R35	22,348,809-22,358,704	1080	240	1030	300	340	#6	800	240
TRA_R36	22,355,532-22,363,689	17800	3900	3280	3600	7600	400	3350	2400
TRA_R37	22,363,063-22,373,780	450	330	•800	•600	560	40	845	275

TRA_R38	22,372,218-22,382,030	1040	150	500	120	*0	*0	330	110
TRA_R39	22,380,031-22,400,558	4610	1680	2370	2070	3310	130	2200	*0
TRA_R40	22,398,227-22,408,706	*0	#0	#0	#0	*0	#16	#0	#8
TRA_R41	22,407,278-22,418,148	2140	450	690	300	860	35	420	270
TRA_R42	22,416,055-22,426,052	6800	930	1660	940	1740	100	1210	700
TRA_R43	22,424,277-22,442,559	860	1450	1360	1820	6540	450	1970	640

The read depth obtained for some of the amplicons were adequate for detecting genetic variation within the region, but as many amplicons failed, this resulted in read depths varying between 0 – 17800, and it would be difficult to collect reliable information about genetic variants for the whole region.

4.3.4 No-amp CRISPR-Cas9 sequencing

Initial results covering all eight samples (not separated) following SMRT sequencing showed the presence of sequences that could suggest reads corresponding to our expected fragment sizes. As illustrated in Figure 4.13, half of the bases were in reads above 160 000 bp and the top 5% bases were in reads > 245 000 bp, indicating a good quality sequencing run.

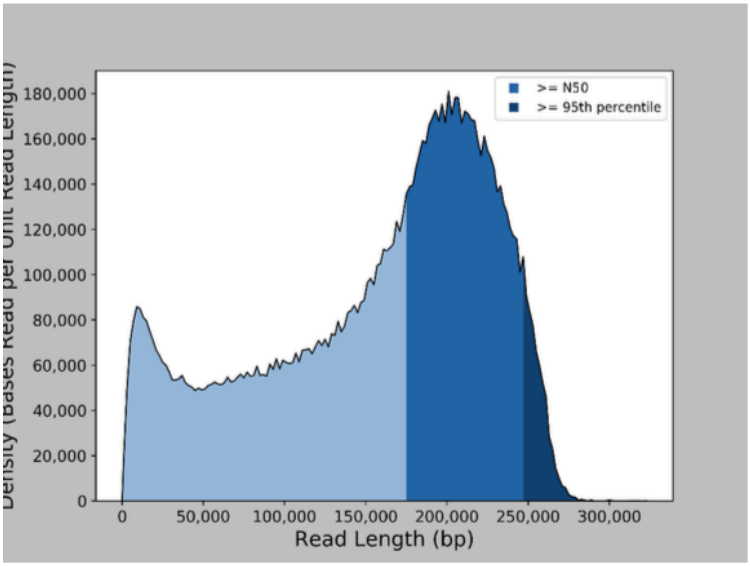


Figure 4.13 Base density plot showing the number of bases sequenced in the samples according to the length of the read in which they were observed.

The insert size distributions are visualized in Figure 4.14. The colour shades indicate the counted number of reads with the same read length, ranging from the highest (red) and lowest (purple) number of reads. From the plot one can see that most of the reads are between 3 kb and 20 kb, which corresponds well to the expected read length of 4.8 kb to 20.1 kb. The fact that most reads have a maximum read length of 20 kb may indicate that either the longest reads have only been passed once or that the shorter fragments have been passed multiple times, creating long subreads. There are few subreads of 50 kb which leads to the assumption that the longest fragments (TRA_R39 and TRA_R43) have not been passed multiple times in the same subread.

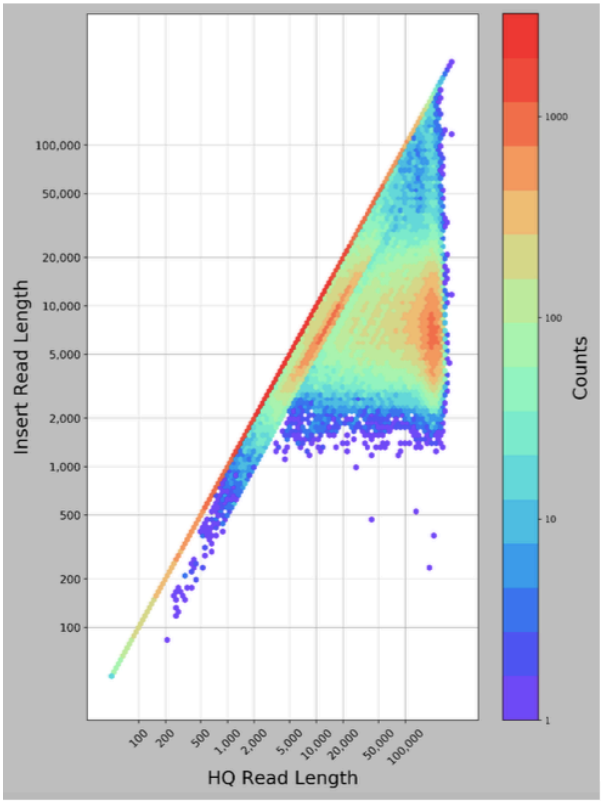


Figure 4.14 Insert Read Length Density (IRLD) Plot displaying the density plot of reads. Reads consisting of a single subread falls along the diagonal. The majority of reads are between 3 kb and 20 kb.

Reads were obtained for most of the fragments with SMRT sequencing. The reads were aligned against the human reference genome (GRChr37) and visualized in IGV (Figure 4.15) where reads for the fragments passed at least once (Table 4.8) and three times (HiFi, Table 4.9) were noted for all eight samples. The highest read depth was found for fragments of maximum 6 kb length, which was the length recommended by PacBio. Notably, the longest fragments (TRA_R39 and TRA_R43) had the least reads, with a maximum of 3 reads passing the whole fragment once and no reads covering the fragment three times.

These results show that targeted enrichment using CRISPR-Cas9 works for the TRA region and can be used for fragments up to 6 kb.

Figure 4.15 Visualization of reads from fragments TRA_R36 to TRA_R39 in WB3 (upper panel) and CD3(-) 3 (lower panel) aligned against the human reference genome (build 37) as shown in IGV 2.9.4.

Table 4.8 Approximate read depth for reads passing the fragments ≥1 time extracted from IGV.

		Read depth							
Fragment	Position Chr14 (build37)	WB1	CD3(-) 1	WB2	CD3(-) 2	WB3	CD3(-) 3	WB4	CD3(-) 4
TRA_R34	22,341,703-22,351,033	6	2	6	1	8	7	11	0
TRA_R35	22,348,806-22,357,811	6	2	6	2	4	0	3	1
TRA_R36	22,355,513-22,363,569	2-8	1-2	2	0	1-5	0	1-3	0-1
TRA_R37a	22,361,786-22,366,978	27	5	17	11	21	7	10	10
TRA_R37b	22,366821-22,372787	2	4	18	3	1	2	8	5
TRA_R38a	22371998-22376849	18-30	4-8	19	5	6	2-4	10-18	3-7

TRA_R38b	22376333-22382358	16	3	1	4	15	8	13	6
TRA_R39	22380467-22400609	0	0	0	1	0	0	0	0
TRA_R40	22398306-22408553	4	1	4	1	2	0	2	1
TRA_R41	22407592-22418062	2	0	6	0	1	0	1	0-1
TRA_R42	22415900-22426436	3	0	7	2	1	0	5	2
TRA_R43	22423964-22442643	1	0	0	0	1	0	2	0-3

Table 4.9 Approximate read depth for HiFi reads for each fragment. Extracted from IGV.

		Read depth							
Fragment	Position Chr14 (build37)	WB1	CD3(-) 1	WB2	CD3(-) 2	WB3	CD3(-) 3	WB4	CD3(-) 4
TRA_R34	22,341,703-22,351,033	5	2	6	1	7	7	9	7
TRA_R35	22,348,806-22,357,811	5	2	6	2	4	0	2	0
TRA_R36	22,355,513-22,363,569	1	1	1	0	0	0	0	0
TRA_R37a	22,361,786-22,366,978	21	5	16	9	18	7	8	7
TRA_R37b	22,366821-22,372787	1	4	13	3	15	2	7	2
TRA_R38a	22371998-22376849	16	4	17	5	0	2	9	2
TRA_R38b	22376333-22382358	13	3	19	3	12	8	10	8
TRA_R39	22380467-22400609	0	0	0	0	0	0	0	0
TRA_R40	22398306-22408553	0	1	4	1	2	0	2	0
TRA_R41	22407592-22418062	1	0	5	0	1	0	1	0
TRA_R42	22415900-22426436	2	0	6	2	1	0	2	0
TRA_R43	22423964-22442643	0	0	0	0	0	0	0	0

5 Discussion

5.1 No association was found with TRA in ME/CFS

No association was detected between TRA and ME/CFS with the 30 SNPs investigated in this study. Thus, we did not replicate the association between rs17255510 and rs11157573 and ME/CFS (Schlauch et al., 2016). A number of reasons can be given for the failure to replicate the findings of Schlauch et al. First, clinical manifestations in ME/CFS cases are known to be different between individuals. Although both cohorts meet the Canadian Consensus Criteria for diagnosis, it does not ensure comparable patient cohorts. Secondly, the variation in sample size between the studies affects their statistical power.

A striking difference between the two studies are the frequencies for the two alleles rs17255510-C and rs11157573-C. Our frequencies for the rs17255510-C allele in cases (22%) and controls (24.6%) were both similar to the allele frequency found in 1000G CEU (23.2%). The same can be seen for the rs11157573-C allele where our allele frequencies are 18.3% (cases) and 16.2% (controls) as compared to 19.7% in 1000G CEU. On the other hand, the allele frequencies from Schlauch et al. study differ both from our study and for the publicly available 1000G CEU. First of all, is their rs17255510-C allele frequency in controls (17.1%) lower than in our study and the CEU data. Secondly, was in fact rs17255510-C the major allele in cases with a frequency of 67.9%. Also, for rs11157573-C did the study by Schlauch et al. have a much higher frequency in cases (48.8%) and a slightly lower frequency in controls (15.8%) compared to what we observed in our and the CEU data. We cannot rule out small differences in allele frequency, however, both cohorts are Caucasian.

A huge drawback with the study by Schlauch et al. is their small sample size consisting of 42 ME/CFS cases and 38 controls, all Caucasians. Studies with few individuals are more affected by random effects like selection bias and genotyping errors (Colhoun, McKeigue, & Smith, 2003). Hence, increasing the chance of false positive discoveries. A larger portion of the individuals included in the study by Schlauch et al. may by coincidence have the minor allele than what is actually representative for the Caucasian population. Thus, providing some explanation to the difference in allele frequencies.

The sample size of our study is larger, but likewise, suboptimal to detect significance for low effect sizes typically seen for complex diseases (Hong & Park, 2012). There is a chance of false negatives when the statistical power is inadequate. Other studies have shown that negative studies were actually consistent with the level of increased risk seen in an initial positive report, but lacked the power to detect it (Altshuler, Daly, & Lander, 2008). For some of our SNPs, a 1% difference in MAF between cases and controls are seen. Although not significant now, the same MAF in a much larger population could reach a statistically significant level of association.

The lack of association in this study cannot be used to conclude that TRA is not involved in the disease. Replication studies are necessary to confirm the presence or absence of association between TRA and ME/CFS. Hence, larger study sizes, including more patients and controls as well as a denser set of SNPs across the TRA region, are necessary to provide more reliable findings. However, obtaining large data sets typically warrants international collaborations and large consortia to achieve. Recently such a large initiative called DecodeME (https://www.decodeme.org.uk/) have been established with the goal of recruiting 20,000 ME/CFS patients, which hopefully will increase the statistical power and aid in the identification of more associations to the disease.

5.2 Poor SNP coverage in TRA on genotyping arrays

GWASs have been facilitated by the production of relatively inexpensive SNP genotyping arrays. The most used arrays vary in their content, but they broadly contain 200,000 to more than 2,000,000 SNPs (Visscher et al., 2017). GWAS discoveries have and still are increasing our knowledge for a wide variety of phenotypes (Visscher et al., 2017). They have also been successfully implemented to better determine the relative role of genes and the environment in disease susceptibility (Visscher et al., 2017). Identification of large numbers of genetic loci have improved our understanding of the basic causes of autoimmune and inflammatory conditions greatly, and provided solid groundwork for hypothesis-driven research into disease mechanisms (Cortes & Brown, 2011).

In this study we employed data produced using the genotyping array Ichip to study the TRA region in ME/CFS. In the 1000G CEU dataset it was shown that 737 SNPs are located in the

~100 kb region sequenced in this thesis. However, we found that 455 of these SNPs were not covered by any of the existing 59 Illumina or 22 Affymetrix arrays evaluated using the SNPchip tool. From the remaining 282 SNPs covered by minimum one array only six were present on the Illumina Human Immunochip 24 v.2 (Ichip for short) array while 67 SNPs were covered by the Affymetrix Genome-Wide SNP Array 6.0. Since our Norwegian ME/CFS Ichip genotyping data consisted of both Ichip v.1 and v.2 generated genotypes (Hajdarevic et al., Unpublished; International Multiple Sclerosis Genetics et al., 2013; Liu et al., 2013), we could only include successfully genotyped SNPs from all datasets. As Ichip version 1 only covered three of the six SNPs found on v.2, we could maximum have three SNPs present. The reason only two SNPs were found in the analysis of our genotyping data is because the data was obtained after QC where the others have been excluded. Notably, the Affymetrix Genome-Wide SNP Array 6.0 was used in the GWASs by Schlauch et al. and Hallmeyer et al.

An explanation for this difference between coverage may be due to the design of these arrays. While Illumina includes SNPs that have been associated with immune-mediated diseases (such as AIDs) (Cortes & Brown, 2011), the Affymetrix array is not using such requirements. Inclusion of SNPs on the Ichip are therefore dependent on identified associations, which are challenging for complex diseases. However, the number of SNPs included has increased from v1 to v2, and Ichip has been useful for identification of susceptibility loci in different immune-mediated diseases, both in HLA region (Mayes et al., 2014) and outside (International Genetics of Ankylosing Spondylitis et al., 2013; Isobe et al., 2015), albeit few associations have been identified in TCR and TRA region. Exactly why few associations between TCR and immune-mediated diseases has been seen is uncertain, but a possible explanation may be that Ichip is being used when studying immune-mediated diseases because of their thought shared pathophysiological mechanisms (Zhernakova, van Diemen, & Wijmenga, 2009). Further supported by the observed occurrence of more than one immune-mediated disease in some patient and families (Zhernakova et al., 2009).

Considering that TCR genes are so immunologically important and that they are interacting with HLA at the molecular level, it may be especially interesting for diseases with an identified HLA-association to study the TCR genes further. This is further supported by the study in narcolepsy where the HLA-DQB1*06:02 allele and TRA have been associated with disease together (Hallmayer et al., 2009). With suggestions that SNPs located in RSS (Posnett

et al., 1994), enhancers or genes (Sharon et al., 2016) can affect the TCR production, it provides a background to study the region more and identify potential functional variants. Particularly the TRA V region seem to interact more with HLA and antigen than the TRB V region (Rudolph & Wilson, 2002). Hence, getting a better understanding of the genetic landscape of this immunologically important region should be important for our understanding of TCR production as well as immune-mediated diseases. This, together with other findings of TCRs role in immunity, can provide a reason to investigate the region more. One time-efficient and relatively inexpensive way to look into it could be with genotyping arrays. However, considering that there are three billion base pairs in the human genome (NIH), and Ichip covers only 195,806 of these SNPs (Cortes & Brown, 2011) other genotyping arrays and sequencing methods should aid in the identification of candidates regions. Notably, it is not necessary to genotype all SNPs in the genome due to the presence of LD blocks in the genome. Thus, one can still identify associated variant by genotyping of tagSNPs. Nevertheless, in this study we show that there are little LD in the ~100 kb TRA region, further indicating the need of better coverage by for example sequencing.

Our findings when looking into the coverage of genotyping arrays in TRA showed that Ichip is not the array covering this region the best, which may be something to consider in studies of complex diseases such as ME/CFS. There may be some benefits with using Affymetrix arrays to perform GWAS in immune-mediated diseases, although these have little coverage of TRA as well. It does cover more SNPs than Ichip, so even though it cannot provide the same insight into the shared mechanisms between immune-mediated diseases, it may identify new suggestive associations, which can be looked into further. Nevertheless, at present time the TRA region is not well studied when using most common genotyping arrays.

5.3 Two sequencing protocols for TRA were established

5.3.1 No-amp Pacbio sequencing can be used for the TRA region

PacBio sequencing is an established method, however, their no-amp targeted sequencing utilizing the CRISPR-Cas9 system protocol was only recently established. Thus, it has not been used in many studies, and as far as we know not to sequence a 100 kb region – a part of the human TRA region, like we have. Typical use has been for smaller regions of 1-4 kb in

size, such as to characterize triplet repeat in the *TCF4* gene (Hafford-Tear et al 2019) and to study repeat elements in huntingtin (HTT) gene (Höljer et al 2018).

Albeit, PacBio recommends producing fragments of 4-6 kb, we tested fragments ranging from 4-20 kb as the TRA region is large and we wanted to sequence ~100 kb. The method was straightforward and easy to use, although it is costly due to the crRNAs that are being used for fragmentation. Hence, it is understandable that other more inexpensive methods are often preferred. However, for studies of repeat elements this method provided novel insight into the dynamic nature of repeat elements not observed by standard methods (Hafford-Tear et al., 2019).

In our no-amp TRA sequencing, we only obtained a maximum read depth of 30, and this was observed for the shortest target fragments (< 6 kb in size). Why we did not get many reads for the larger fragments can have different causes. During the sequencing the polymerase might not have been able to read through the largest fragments one full time which resulted in the incomplete reads being excluded for further analysis. Furthermore, unsuccessful CRISPR-Cas9-targeted enrichment will result in lack of reads. In our study, we used one crRNA per cut site, although Pacbio recommends ordering three. Furthermore, we did not perform an evaluation of the CRISPR-Cas9 targeted enrichment prior to sequencing as suggested by PacBio. Despite sending our SMRTbell library for SMRT sequencing without any previous knowledge about the fragment sizes or quality, the reads obtained indicates that most of our cut reactions seem to have worked.

In our study, it is unlikely that the gDNA affected the CRISPR-Cas9 targeted enrichment because all evaluations performed on the template showed that it was high molecular weight with good quality. However, the amount of DNA available was limited. Thus, an increase in input may have resulted in more reads.

The protocol was easy to follow and resulted in reads for most fragments, although we did not get the read depth necessary for genetic variant detection by our design. We knew that our design to cover the 100kb region was pushing the limits of the protocol. Overall, our sequencing results indicate that the method can be used in TRA, however, it is not optimal for screening of large regions. It can instead be a useful complement to another sequencing approach because the protocol has been found to result in accurate genotyping information

and phasing (Hafford-Tear et al., 2019) in addition to offering a way to study repeat elements difficult to investigate with PCR-based methods (Hoijer et al., 2018).

5.3.2 Long-range PCR and short read sequencing are difficult to optimize for TRA

The most widely used technology is the Illumina sequencing platform (Torresen et al., 2019). It produces highly accurate (>99.9%) reads that are inexpensive to generate on a massive scale (Logsdon et al., 2020). The method can be used for whole-genome sequencing, exome sequencing, transcription sequencing and targeted sequencing, depending on the template being used. It is often combined with PCR, which has been the most popular sample preparation technique (Mamanova et al., 2010). Traditional PCR reactions had limitations in size of amplicons which has been improved by the development of long-range PCR (Jia, Guo, Zhao, & Wang, 2014). Thus, long-range PCR has facilitated studies in molecular genetics, and can when combined with sequencing, achieve higher sensitivity as well as a more cost-effective tool for detecting genetic variants (Jia et al., 2014).

In this work long-range PCR was used for enrichment of the target region before sequencing with Illumina. The same method is successfully been used to sequence the HLA region (Lande et al., 2020), which contains the largest degree of polymorphism in the human genome (Gough & Simmonds, 2007). Hence, it was thought to be useful for sequencing of TRA as well.

Generally good read depth was obtained with this method despite low input. The biggest drawbacks were that the long-range PCR optimization was time-consuming, and the primer pairs were unspecific. Even with careful primer design and in silico PCR control of the primers, they resulted in several additional bands for most of the amplicons. These unspecific amplifications may be due to homology and repetitive regions in the TRA, thus making it difficult to obtain specific amplicons in the region. Although we tried different temperatures, PCR programs and long-range PCR kits with different polymerases, the amplifications remained unspecific. If further optimization would have taken place, we could have ordered new primer pairs, however, as 8 out of 10 primer pairs gave unspecific amplifications, it is not given that the specificity would have improved. On the other hand, the magnesium (Mg^{2+}) concentration could have been adjusted, as Mg^{2+} affects the polymerase activity and can result

in increased product (Scientific). However, this was not performed with the kits used in this work because both reaction buffers already had Mg^{2+} in them. Furthermore, altered concentration has also shown to increase the unspecific products (Scientific). Therefore, it may not have made such a large difference for us either way as we ended up extracting our band of interest from agarose gel in order to see whether our target region had been amplified and we could detect genetic variants within it.

Short-read sequencing with Illumina resulted in varying read depth for the amplicons where genetic variation can be detected in some of them. However, as long-range PCR was laborious and difficult to optimize in TRA, an alternative would be to test enrichment using capture probes as it is used for many molecular diagnostics approaches and has been shown to have potential in HLA typing as well (Hogan, Cransberg, Jordan, Goodridge, & Sayer, 2015).

5.3.3 Alignment tools should be chosen based on sequencing methods

Alignment or mapping of reads sets the basis for further analysis of sequencing data (Langmead & Salzberg, 2012; Thankaswamy-Kosalai, Sen, & Nookaew, 2017). Thus, the alignment strategy used for both sequencings methods in this thesis will influence the variant detection. First, there are bioinformatical challenges when using long-read sequencing. Thus, requiring analysis tools accounting for the characteristics of long reads (Amarasinghe et al., 2020). For instance, indels are more frequent in long reads as compared to short reads (Li & Durbin, 2010a). Therefore, the aligner must allow for more alignment gaps when handling long reads, while an aligner for short reads can allow limited gaps (Li & Durbin, 2010a). In this study we used, BWA-MEM and Bowtie 2 for alignment of the reads to the human reference genome (GRChr37). BWA-MEM is one of many applications that can be used for such long reads (Li & Durbin, 2010b), while Bowtie 2 is efficiently aligning short-reads (Langmead & Salzberg, 2012). Notably, the choice of aligner and the settings used will influence the alignment. We therefore applied less stringent alignment criteria, allowing for sequencing errors, to ensure that as many reads as possible would align.

Visualization of the aligned reads showed that the majority of our reads had aligned to our target region. The reads were mapped against the human reference genome, but it is possible that another region-specific reference might be more appropriate for this region.

5.4 Can somatic TRA rearrangement confound genotyping and sequencing results?

One of our objectives was to find out whether genotyping of SNPs in the rearranged TRA region is reliable. Therefore, both sequencing and genotyping were performed on gDNA extracted from whole and CD3 depleted blood. The main question was whether the presence of rearranged TCRs in WB would influence the results of the two methods by disturbing the genotyping of our SNPs due to the somatic rearrangements or by resulting in smaller targeted fragments as compared to the germline DNA in CD3(-) samples.

Since it is known that smaller fragments can be amplified more than the larger ones, causing an amplification bias (Amarasinghe et al., 2020), it was hypothesized that using the PCR-based approach in WB would result in increased amplification of the somatic rearranged, shorter DNA than of the non-rearranged germline DNA. Thus, disturbing the sequencing and resulting in difficulties with covering the whole region. Unfortunately, for the long-range PCR we lost the possibility to evaluate whether CD3 depletion was necessary because we only sequenced the bands of interest due to unspecific binding for most primer pairs. As there was no difference in the specificity between WB and CD3(-) samples regarding the specificity, it is highly likely that the additional bands could have been due to the primers annealing outside the target region. However, it could also be caused by TCR recombinants, which is less likely as it would only explain some of the bands for WB.

On the other hand, the no-amp CRISPR-Cas9 targeted enrichment protocol does not have this issue of introducing PCR bias. Hence, making it a good method to evaluate the necessity of CD3 depletion. Notably, the reads obtained after SMRT sequencing does not show any immediate signs of there being differences between the read lengths in WB and CD3(-) samples, however, a conclusion cannot be made because the reads depth is not adequate for detection of genetic variants.

We also designed our SNP genotyping to contain one SNP outside the region being rearranged, thus in theory not being influenced by somatic rearrangements when genotyped. However, as this was the SNP that failed completely during genotyping, it was not possible to evaluate any potential influence the rearrangement has on the genotyping.

Taken together, no conclusions can be made, which still leaves the question if somatic TRA rearrangement can confound genotyping and sequencing results unanswered.

5.5 Future perspectives

As time became a limiting factor, we did not have time to take a closer look at the obtained reads. Furthermore, we were not able to evaluate if CD3 depletion is necessary, this should therefore be part of future studies. It would therefore be interesting to try and identify SNPs covered in the regions with high enough read depth and compare these between WB and CD3(-) from the same individual, as well between the four individuals. In addition to using the human reference genome for alignment, as in this study, it would be interesting to see if using de novo assembly or a region-specific reference would influence the results.

Future studies may benefit from a fully sequenced TRA. Although our results showed that the sequencing protocols tested works in this region, they are not optimal and should be adapted. A proposed alternative is to test enrichment by capture probes with following sequencing using Illumina or SMRT sequencing. Proper coverage of this region may reveal genetic variants influencing the rearrangement in TRA, thus provide more insight into the aetiology of diseases. Indications of the affect SNPs in regulatory regions of the TCR has on its production, provide good reasons for this receptor to be investigated.

Our finding of no association between TRA and ME/CFS should be replicated in larger cohorts. Additionally, should more SNPs be genotyped within this region as the coverage of identified SNPs in this region is inadequate on all genotyping arrays.

6 Conclusion

The aim of this thesis was to find methods that can be used to study genetic variants in the T cell receptor α region (TRA) to identify possible associations with ME/CFS. We have demonstrated that the publicly available 1000G CEU dataset is representative for the Norwegian population in the TRA region, by comparing the LD pattern from the 1000G CEU with our Norwegian cohort. Furthermore, we found no association between the 30 genotyped SNPs in TRA and our Norwegian ME/CFS population. However, we have shown that the genetic variation in the TRA region is not well-covered on genotyping arrays.

Both a long- and a short reads sequencing approach were established for TRA. The protocol for no-amp targeted sequencing utilizing the CRISPR/Cas9 system with SMRT sequencing by PacBio provided the highest read depth for fragments <6 kb and can be useful as a complement to other sequencing approaches. Target enrichment using long-range PCR required a lot of optimization but resulted in good read depth for some fragments, which could provide some information about genetic variants located in this region. Hence, more research is required to identify SNPs with regulatory function in this region.

References

Albright, F., Light, K., Light, A. R., Bateman, L., & Cannon-Albright, L. A. (2011). Evidence for a heritable predisposition to Chronic Fatigue Syndrome. *BMC Neurology, 11*(62).

Alcover, A., Alarcón, B., & Bartolo, V. D. (2018). Cell Biology of T Cell Receptor Expression and Regulation. *Annu. Rev. Immunol.* doi:10.1146/annurev-immunol-

Altshuler, D., Daly, M. J., & Lander, E. S. (2008). Genetic mapping in human disease. *Science, 322*(5903), 881-888. doi:10.1126/science.1156409

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biol, 21*(1), 30. doi:10.1186/s13059-020-1935-5

Ardlie, K. G., Kruglyak, L., & Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet, 3*(4), 299-309. doi:10.1038/nrg777

Arnold, B. (2002). Levels of peripheral T cell tolerance. *Transplant Immunology, 10*, 109-114.

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics, 21*(2), 263-265. doi:10.1093/bioinformatics/bth457

Borghans, J. A., Noest, A. J., & De Boer, R. J. (2003). Thymic selection does not limit the individual MHC diversity. *Eur J Immunol, 33*(12), 3353-3358. doi:10.1002/eji.200324365

Carruthers, B. M. (2007). Definitions and aetiology of myalgic encephalomyelitis: how the Canadian consensus clinical definition of myalgic encephalomyelitis works. *J Clin Pathol, 60*(2), 117-119. doi:10.1136/jcp.2006.042754

Carruthers, B. M., van de Sande, M. I., De Meirleir, K. L., Klimas, N. G., Broderick, G., Mitchell, T., . . . Stevens, S. (2011). Myalgic encephalomyelitis: International Consensus Criteria. *J Intern Med, 270*(4), 327-338. doi:10.1111/j.1365-2796.2011.02428.x

Colhoun, H. M., McKeigue, P. M., & Smith, G. D. (2003). Problems of reporting genetic associations with complex outcomes. *The Lancet, 361*(9360), 865-872.

Cortes, A., & Brown, M. A. (2011). Promise and pitfalls of the Immunochip. *Arthritis Research & Therapy, 13*.

Dibble, J. J., McGrath, S. J., & Ponting, C. P. (2020). Genetic Risk Factors of ME/CFS: A Critical Review. *Hum Mol Genet*. doi:10.1093/hmg/ddaa169

Donnelly, P. (2008). Progress and challenges in genome-wide association studies in humans. *Nature, 456*(7223), 728-731. doi:10.1038/nature07631

Dudbridge, F. (2008). Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum Hered, 66*(2), 87-98. doi:10.1159/000119108

Evans, G. A., Lewis, K. A., & Lawless, G. M. (1988). Molecular organization of the human CD3 gene family on chromosome 11q23. *Immunogenetics, 28*, 365-373.

Frazer, K. A., Murray, S. S., Schork, N. J., & Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nat Rev Genet, 10*(4), 241-251. doi:10.1038/nrg2554

Fukunda, K., Straus, S. E., Hickie, I., Sharpe, M. C., Dobbins, J. G., & Komaroff, A. (1994). The Chronic Fatigue Syndrome: A Comprehensive Approach to Its Definition and Study. *Ann Intern Med, 121*, 953-959.

Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., . . . Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. doi:10.1038/nature15393

Goris, A., & Liston, A. (2012). The immunogenetic architecture of autoimmune disease. *Cold Spring Harb Perspect Biol, 4*(3). doi:10.1101/cshperspect.a007260

Gough, S. C. L., & Simmonds, M. J. (2007). The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. *Current Genomics, 8*, 453-465.

Hafford-Tear, N. J., Tsai, Y. C., Sadan, A. N., Sanchez-Pintado, B., Zarouchlioti, C., Maher, G. J., . . . Davidson, A. E. (2019). CRISPR/Cas9-targeted enrichment and long-read sequencing of the Fuchs endothelial corneal dystrophy-associated TCF4 triplet repeat. *Genet Med, 21*(9), 2092-2102. doi:10.1038/s41436-019-0453-x

Hajdarevic, R., Lande, A., Rydland, A., Strand, E. B., Sosa, D. D., Mella, O., . . . Lie, B. A. (Unpublished). Studying known autoimmune risk genes in patients with CFS/ME.

Hallmayer, J., Faraco, J., Lin, L., Hesselson, S., Winkelmann, J., Kawashima, M., . . . Mignot, E. (2009). Narcolepsy is strongly associated with the T-cell receptor alpha locus. *Nat Genet, 41*(6), 708-711. doi:10.1038/ng.372

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet, 6*(2), 95-108. doi:10.1038/nrg1521

Hogan, H., Cransberg, R., Jordan, M., Goodridge, D., & Sayer, D. (2015). Target enrichment using capture probes: The future of HLA typing by next generation sequencing? *Human Immunology, 76*, 165.

Hoijer, I., Tsai, Y. C., Clark, T. A., Kotturi, P., Dahl, N., Stattin, E. L., . . . Ameur, A. (2018). Detailed analysis of HTT repeat elements in human blood using targeted amplification-free long-read sequencing. *Hum Mutat, 39*(9), 1262-1272. doi:10.1002/humu.23580

Hong, E. P., & Park, J. W. (2012). Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics & Informatics, 10*, 117-122.

ICD-10. G93.3. Retrieved from https://icd.who.int/browse10/2016/en#/G93.3

International Genetics of Ankylosing Spondylitis, C., Cortes, A., Hadler, J., Pointon, J. P., Robinson, P. C., Karaderi, T., . . . Brown, M. A. (2013). Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nat Genet, 45*(7), 730-738. doi:10.1038/ng.2667

International Multiple Sclerosis Genetics, C., Beecham, A. H., Patsopoulos, N. A., Xifara, D. K., Davis, M. F., Kemppinen, A., . . . McCauley, J. L. (2013). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat Genet, 45*(11), 1353-1360. doi:10.1038/ng.2770

Isobe, N., Madireddy, L., Khankhanian, P., Matsushita, T., Caillier, S. J., More, J. M., . . . Oksenberg, J. R. (2015). An ImmunoChip study of multiple sclerosis risk in African Americans. *Brain, 138*(Pt 6), 1518-1530. doi:10.1093/brain/awv078

Jia, H., Guo, Y., Zhao, W., & Wang, K. (2014). Long-range PCR in next-generation sequencing: comparison of six enzymes and evaluation on the MiSeq sequencer. *Sci Rep, 4*, 5737. doi:10.1038/srep05737

Klein, L., Kyewski, B., Allen, P. M., & Hogquist, K. A. (2014). Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nat Rev Immunol, 14*(6), 377-391. doi:10.1038/nri3667

Lande, A., Fluge, O., Strand, E. B., Flam, S. T., Sosa, D. D., Mella, O., . . . Viken, M. K. (2020). Human Leukocyte Antigen alleles associated with Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS). *Sci Rep, 10*(1), 5267. doi:10.1038/s41598-020-62157-x

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods, 9*(4), 357-359. doi:10.1038/nmeth.1923

Lefranc, M.-P. (2020). Nomenclature and overview of the human T cell receptor genes. Retrieved from http://www.imgt.org/IMGTeducation/QuestionsAnswers/_UK/sommaireTcr.html

Lefranc, M.-P., & Lefranc, G. (2001). *The T cell receptor factbook*. London: Academic Press.

Lewallen, S., & Courtright, P. (1998). Epidemiology in Practice: Case-Control Studies. *Community Eye Health, 11*.

Li, H., & Durbin, R. (2010a, 20.02.2010). Burrows-Wheeler Aligner. Retrieved from http://bio-bwa.sourceforge.net/index.shtml

Li, H., & Durbin, R. (2010b). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics, 26*, 589-595. doi:10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Subgroup, G. P. D. P. (2009). The Sequencing Aligment/Map format and SAMtools. *Bioinformatics, 16*. doi:10.1093/bioinformatics/btp352.

Liu, J. Z., Hov, J. R., Folseraas, T., Ellinghaus, E., Rushbrook, S. M., Doncheva, N. T., . . . International, I. B. D. G. C. (2013). Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. *Nat Genet, 45*(6), 670-675. doi:10.1038/ng.2616

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nat Rev Genet, 21*(10), 597-614. doi:10.1038/s41576-020-0236-x

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., . . . Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nat Methods, 7*(2), 111-118. doi:10.1038/nmeth.1419

Mayes, M. D., Bossini-Castillo, L., Gorlova, O., Martin, J. E., Zhou, X., Chen, W. V., . . . Martin, J. (2014). Immunochip analysis identifies multiple susceptibility loci for systemic sclerosis. *Am J Hum Genet, 94*(1), 47-61. doi:10.1016/j.ajhg.2013.12.002

Maynard Smith, J., & Haigh, J. (1974). The hitch-hiking effect of a favorable gene. *Genet. Res., 23*, 23-35.

McMurry, M. T., Hernandez-Munain, C., Lauzurica, P., & Krangel, M. S. (1997). Enhancer Control of Local Accessibility to V(D)J Recombinase. *Molecular and Cellular Biology, 17*, 4553-4561.

Morris, G., Berk, M., & Galecki, P. (2014). The Emerging Role of Autoimmunity in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/cfs). *Mol Neurobiol, 49*, 741-756. doi:10.1007/s12035-013-8553-0

Mu, W., & Zhang, W. (2013). Molecular Approaches, Models, and Techniques in Pharmacogenomic Research and Development. In *Pharmacogenomics* (pp. 273-294).

Murphy, K., & Weaver, C. T. (2017). *Janeway's immunobiology* (9th edition. ed.). New York: Garland Science/Taylor & Francis Group.

NBMDR. (08.06.2020). The Norwegian Bone Marrow Donor Registry. Retrieved from https://oslo-universitetssykehus.no/fag-og-forskning/nasjonale-og-regionale-tjenester/det-norske-benmargsgiverregisteret

NIH. (24.02.2020). Human Genome Project FAQ. Retrieved from https://www.genome.gov/human-genome-project/Completion-FAQ

O'Donnell, S., Borowski, K., Espin-Garcia, O., Milgrom, R., Kabakchiev, B., Stempak, J., . . . Silverberg, M. S. (2019). The Unsolved Link of Genetic Markers and Crohn's Disease Progression: A North American Cohort Experience. *Inflamm Bowel Dis, 25*(9), 1541-1549. doi:10.1093/ibd/izz016

Omer, A., Shemesh, O., Peres, A., Polak, P., Shepherd, A. J., Watson, C. T., . . . Yaari, G. (2020). VDJbase: an adaptive immune receptor genotype and haplotype database. *Nucleic Acids Res, 48*(D1), D1051-D1056. doi:10.1093/nar/gkz872

PacBio. (2020). *Procedure & Checklist - No-Amp Targeted Sequencing Utilizing the CRISPR-Cas9 System* [05].

Paul, S., Shilpi, & Lal, G. (2015). Role of gamma-delta (gammadelta) T cells in autoimmunity. *J Leukoc Biol, 97*(2), 259-271. doi:10.1189/jlb.3RU0914-443R

Posnett, D. N., Vissinga, C. S., Pambuccian, C., Wei, S., Robinson, M. A., Kostyu, D., & Concannon, P. (1994). Level of Human TCRBV3S1 (Vβ3) Expression Correlates with Alleic Polymorphism in the Spacer Region of the Recombination Signal Sequence. *The Journal of Experimental Medicine, 179*, 1707-1711.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet, 81*(3), 559-575. doi:10.1086/519795

Roberts, H. E., Lopopolo, M., Pagnamenta, A. T., Sharma, E., Parkes, D., Lonie, L., . . . Buck, D. (2021). Short and long-read genome sequencing methodologies for somatic variant detection; genomic analysis of a patient with diffuse large B-cell lymphoma. *Sci Rep, 11*(1), 6408. doi:10.1038/s41598-021-85354-8

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative Genomics Viewer. *Nature Biotechnology, 29*, 24-26. doi:10.1038/nbt0111-24

Rosati, E., Dowds, C. M., Liaskou, E., Henriksen, E. K. H., Karlsen, T. H., & Franke, A. (2017). Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnology, 17*.

Ruddle, N. H., & Akirav, E. M. (2009). Secondary lymphoid organs: responding to genetic and environmental cues in ontogeny and the immune response. *J Immunol, 183*(4), 2205-2212. doi:10.4049/jimmunol.0804324

Rudolph, M. G., & Wilson, I. A. (2002). The specificity of TCR/pMHC interaction. *Curr Opin Immunol, 14*, 52-65.

Schlauch, K. A., Khaiboullina, S. F., De Meirleir, K. L., Rawat, S., Petereit, J., Rizvanov, A. A., . . . Lombardi, V. C. (2016). Genome-wide association analysis identifies genetic variations in subjects with myalgic encephalomyelitis/chronic fatigue syndrome. *Transl Psychiatry, 6*, e730. doi:10.1038/tp.2015.208

Scientific, T. F. PCR Setup - Six Critical Components to Consider. Retrieved from https://www.thermofisher.com/no/en/home/life-science/cloning/cloning-learning-center/invitrogen-school-of-molecular-biology/pcr-education/pcr-reagents-enzymes/pcr-component-considerations.html

Sharon, E., Sibener, L. V., Battle, A., Fraser, H. B., Garcia, K. C., & Pritchard, J. K. (2016). Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nat Genet, 48*(9), 995-1002. doi:10.1038/ng.3625

Slatkin, M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet, 9*(6), 477-485. doi:10.1038/nrg2361

Smith, A. K., Fang, H., Whistler, T., Unger, E. R., & Rajeevan, M. S. (2011). Convergent genomic studies identify association of GRIK2 and NPAS2 with chronic fatigue syndrome. *Neuropsychobiology, 64*(4), 183-194. doi:10.1159/000326692

Smith, J., Fritz, E. L., Kerr, J. R., Cleare, A. J., Wessely, S., & Mattey, D. L. (2005). Association of chronic fatigue syndrome with human leucocyte antigen class II alleles. *J Clin Pathol, 58*(8), 860-863. doi:10.1136/jcp.2004.022681

Stucka, R. (2005). High Pure PCR-product Purification Kit: Changed Protocol for Purification of Large DNA Fragments (4.5 kb to >30 kb). *Biochemica, 3*.

Thankaswamy-Kosalai, S., Sen, P., & Nookaew, I. (2017). Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. *Genomics, 109*(3-4), 186-191.

The International Genome Sample Resource, I. (2020). Retrieved from https://www.internationalgenome.org/home

Torresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., . . . Linke, D. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Res, 47*(21), 10994-11006. doi:10.1093/nar/gkz841

Travers, K. J., Chin, C. S., Rank, D. R., Eid, J. S., & Turner, S. W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res, 38*(15), e159. doi:10.1093/nar/gkq543

Trynka, G., Hunt, K. A., Bockett, N. A., Romanos, J., Mistry, V., Szperl, A., . . . van Heel, D. A. (2011). Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet, 43*(12), 1193-1201. doi:10.1038/ng.998

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet, 101*(1), 5-22. doi:10.1016/j.ajhg.2017.06.005

Walsh, C. M., Zainal, N. Z., Middleton, S. J., & Paykel, E. S. (2001). A family history study of chronic fatigue syndrome *Psychiatric Genetics, 11*.

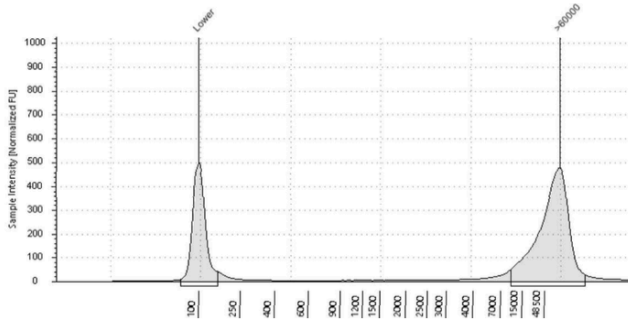
Watson, C. T., Glanville, J., & Marasco, W. A. (2017). The Individual and Population Genetics of Antibody Immunity. *Trends Immunol, 38*(7), 459-470. doi:10.1016/j.it.2017.04.003

Weissman, A. M., Hou, D., Orloff, D. G., Modi, W. S., Seuanez, H., O'Brien, S. J., & Klausner, R. D. (1988). Molecular cloning and chromosomal localization of the human T-cell receptor ζ chain: Distinction from the molecular CD3 complex. *Proc. Natl. Acad. Sci., 85*, 9709-9713.

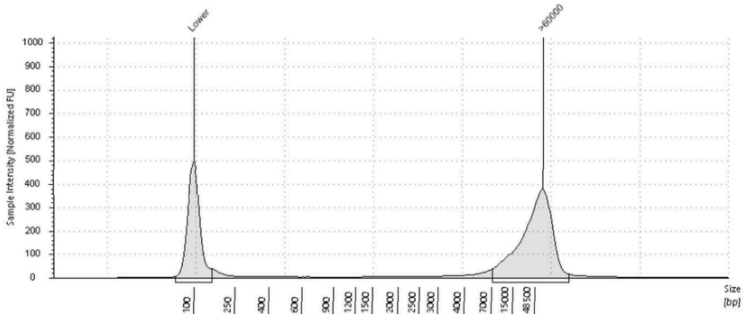
Zhernakova, A., van Diemen, C. C., & Wijmenga, C. (2009). Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat Rev Genet, 10*(1), 43-55. doi:10.1038/nrg2489

Appendix I – Electropherograms

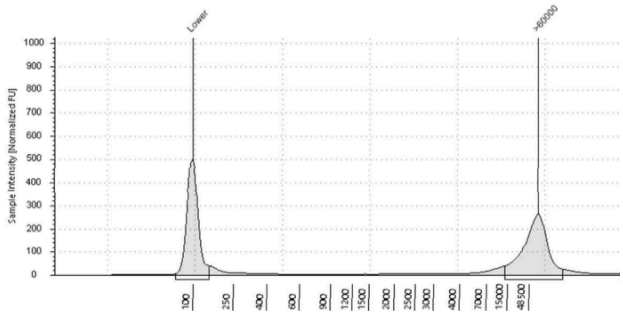
A1 WB1 NP



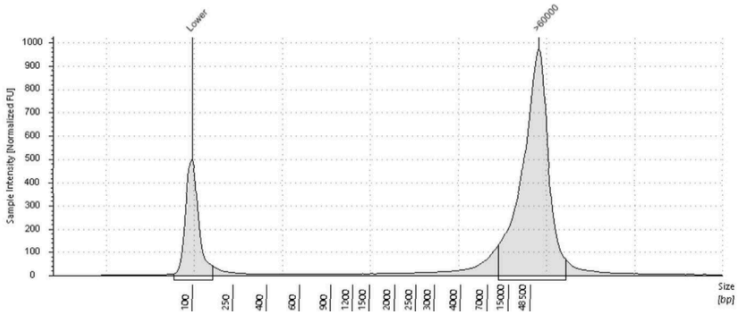
B1 WB1 P



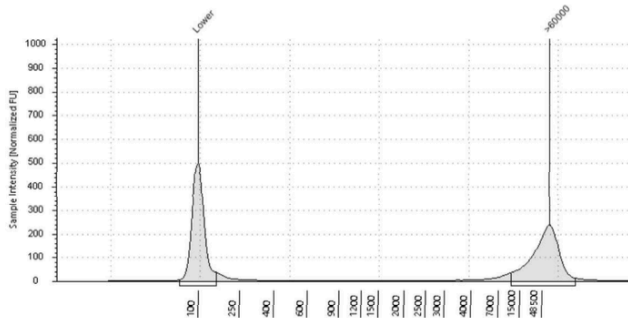
C1 WB2 NP



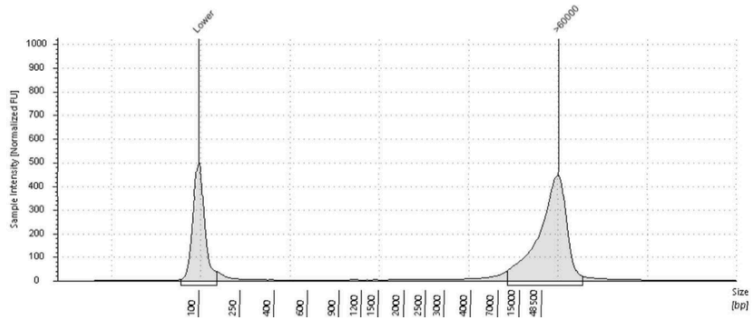
D1 WB2 P



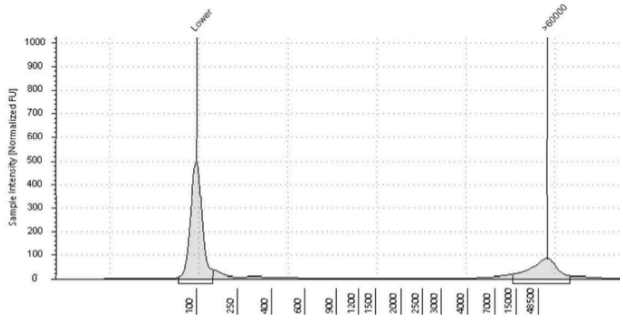
E1 WB3 NP



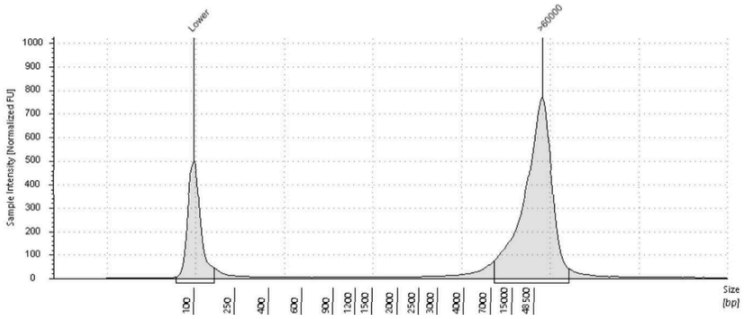
F1 WB3 P



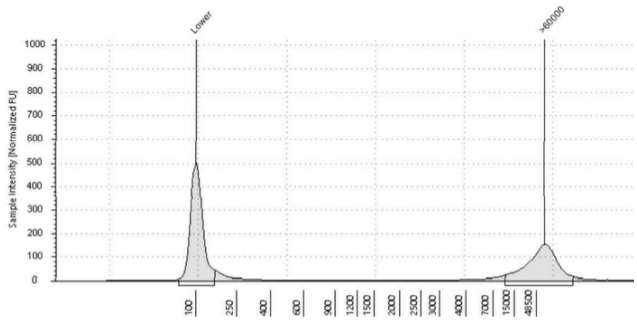
G1 WB4 NP



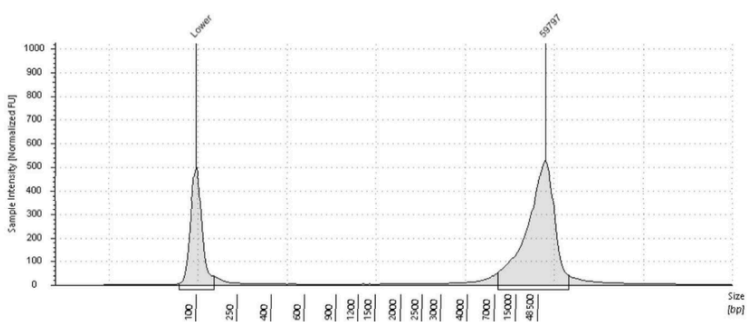
H1 WB4 P



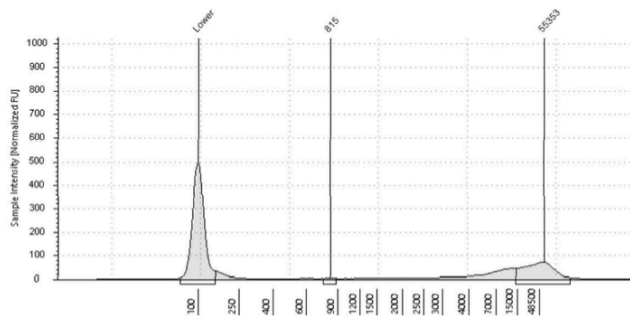
A2 WB5 NP



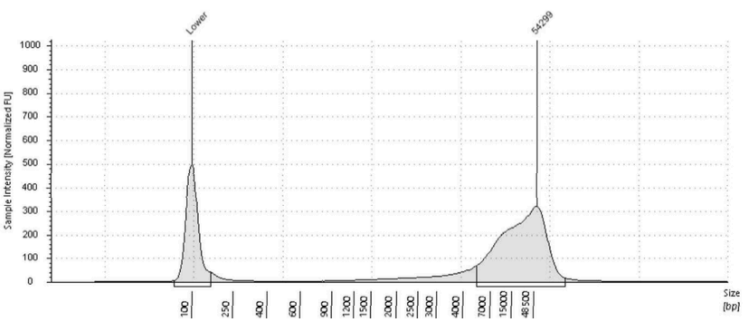
B2 WB5 P



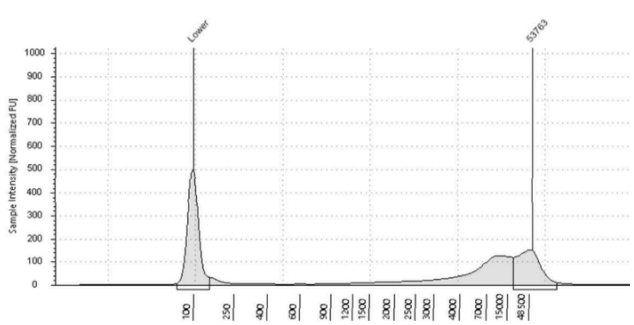
C2 CD3-1 NP



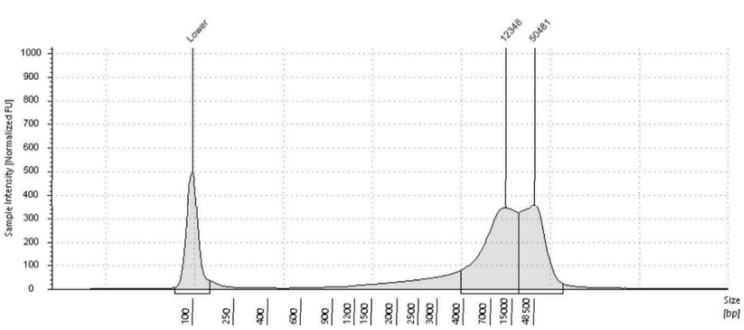
D2 CD3-1 P



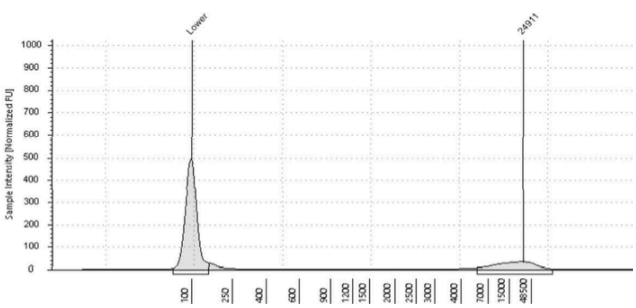
E2 CD3-2 NP



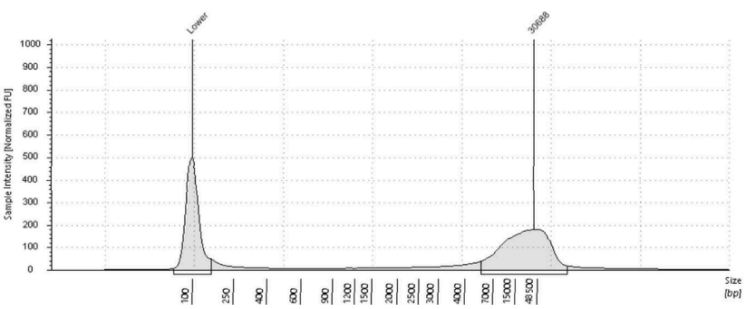
F2 CD3-2 P



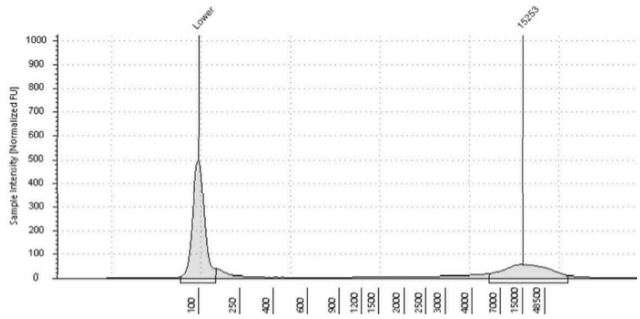
G2 CD3-3 NP



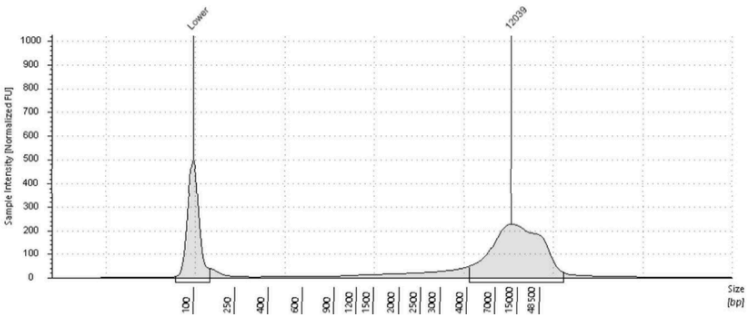
H2 CD3-3 P



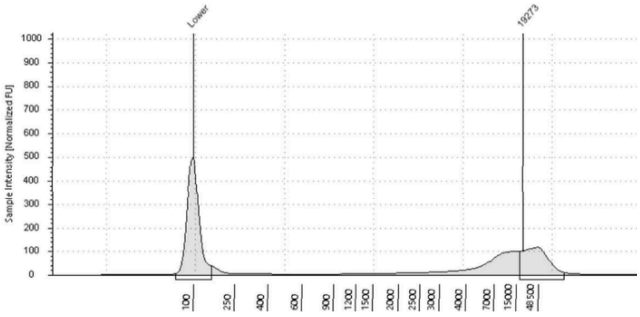
A3 CD3-4 NP



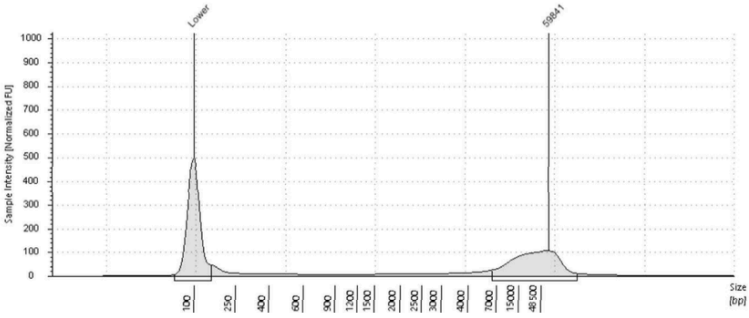
B3 CD3-4 P



C3 CD3-5 NP



D3 CD3-5 P



Appendix II - Materials

Table S1. Reagents, suppliers and catalog numbers

Product name	Supplier	Catalog nb.
Alt-R® CRISPR-Cas9 tracrRNA	Integrated DNA Technologies, Coralville, IA, USA	
AMPure PB	Pacific Biosciences, Menlo Park, CA, USA	100-265-900
Barcoded adapters	Integrated DNA Technologies, Coralville, IA, USA	
BSA		
DPBS		
Cas9 Nuclease, S. pyogenes, high conc., 500 pmol	New England BioLabs, Ipswich, MA, USA	M0386M
CutSmart® Buffer, 10X	New England BioLabs, Ipswich, MA, USA	B7204S
DNA Ladder 1 kb	New England BioLabs, Ipswich, MA, USA	N3232S
EDTA, pH 7.4		
Elution Buffer	Pacific Biosciences, Menlo Park, CA, USA	101-633-500
Exonuclease III	New England BioLabs, Ipswich, MA, USA	M0206S
Fetal Bovine Serum	Biowest	S181H-500
Genomic DNA Reagents	Agilent Technologies	5067-5366
GeneRuler 1kb DNA Ladder	Fermentas, Thermo Fisher Scientific, Waltham, MA, USA	SM0311
Gel Red™ Nucleic acid Gel Stain 10,000X in Water	Biotium Inc, Freemont, CA, USA	41001
IDTE pH 7.5 Buffer (1X TE Solution)	Integrated DNA Technologies, Coralville, IA, USA	11-01-02-02

NEBuffer™ 3.1, 10X	New England BioLabs, Ipswich, MA, USA	B7203S
Nuclease-Free Duplex Buffer	Integrated DNA Technologies, Coralville, IA, USA	11-01-03-01
Nuclease-Free Water	Ambion	AM9937
Qubit™ dsDNA HS Assay Kit	Thermo Fisher Scientific, Waltham, MA, USA	Q32854
RPMI		
SeaKem® LE Agarose	Lonzo, Basel, Switzerland	50004
Shrimp Alkaline Phosphatase (rSAP)	New England BioLabs, Ipswich, MA, USA	M0371S
SMRTbell Enzyme Clean Up Kit	Pacific Biosciences, Menlo Park, CA, USA	101-746-400
Solid Phase Reversible Immobilization (SPRI) beads	Beckman Coulter, Brea, CA, USA	
SOLu-Trypsin	Sigma-Aldrich, Darmstadt, Germany	EMS0004
TaqMan® SNP Genotyping Assay C_25962256_10	Thermo Fisher Scientific, Waltham, MA, USA	4351379
TaqMan® SNP Genotyping Assay C_32009806_10	Thermo Fisher Scientific, Waltham, MA, USA	4351379
TaqMan® SNP Genotyping Assay C_34374423_10	Thermo Fisher Scientific, Waltham, MA, USA	4351379
TaqMan® SNP Genotyping Assay C_44756205_10	Thermo Fisher Scientific, Waltham, MA, USA	4351379
T4 DNA Ligase Reaction Buffer, 10X	New England BioLabs, Ipswich, MA, USA	B0202S
T4 Ligase, HC	Thermo Fisher Scientific, Waltham, MA, USA	EL0013
UltraPure™ 10X TAE Buffer	Invitrogen, Thermo Fisher Scientific, Waltham, MA, USA	15558-042
0.5 M EDTA, pH 8.0 Molecular Biology Grade	Millipore, Burlington, MA, USA	324506

1kb Plus DNA Ladder	Invitrogen, Thermo Fisher Scientific, Waltham, MA, USA	10787-018
6X DNA Loading Dye	Fermentas, Thermo Fisher Scientific, Waltham, MA, USA	R0611
10X Blue Juice Gel Loading Buffer	Invitrogen, Thermo Fischer Scientific, Waltham, MA, USA	10787-018

Table S2. Consumables, suppliers and catalog numbers

Product name	**Supplier**	**Catalog nb.**
Genomic DNA ScreenTape	Agilent, Santa Clara, CA, USA	5067-5365
Loading Tip	Agilent, Santa Clara, CA, USA	5067-5598
MicroAmp™ Optical Adhesive Film	Applied Biosystems™, Thermo Fischer Scientific, Waltham, MA, USA	
MicroAmp® Optical 384-Well Reaction Plate with Barcode	Applied Biosystems™, Thermo Fischer Scientific, Waltham, MA, USA	
Microseal® ´F´foil Seals	Bio-Rad, Hercules, CA, USA	MSF1001
MixMate	Eppendorf, Hamburg, Germany	
Qubit™ Assay Tubes	Thermo Fisher Scientific, Waltham, MA, USA	Q32856
Safe-Lock Tubes 1.5 mL	Eppendorf, Hamburg, Germany	0030121589
0.2 Non-skirted 96-well PCR plate	Thermo Fisher Scientific, Waltham, MA, USA	AB0600
96-well Plate	Agilent, Santa Clara, CA, USA	5042-8502

Table S3. Kits, suppliers and catalog numbers

Product name	**Manufacturer**	**Catalog nb.**
Dynabeads® FlowComp™ Human CD3	Invitrogen, Thermo Fisher Scientific, Waltham, MA, USA	11365D
High Pure PCR Product Purification Kit	Roche, Basel, Switzerland	11732668001

LongAmp™ Taq PCR Kit, 100 Assays	New England BioLabs, Ipswich, MA, USA	E5200S
Monarch® Genomic DNA Purification Kit	New England BioLabs, Ipswich, MA, USA	T3010S
NGSgo-LibrX	GENDX, Utrecht, The Netherlands	
No-Amp Accessory Kit	Pacific Biosciences, Menlo Park, CA, USA	101-788-900
UltraRun LongRange PCR Kit	Qiagen, Hilden, Germany	206442

Table S4. Instruments and suppliers.

Product name	**Supplier**
Biofuge fresco	Thermo Fisher Scientific, Waltham, MA, USA
Centrifuge 5810 R	Eppendorf, Hamburg, Germany
DNA LoBind Tube, 1.5 mL	Eppendorf, Hamburg, Germany
DynaMag™ -2 Magnet	Invitrogen, Thermo Fisher Scientific, Waltham, MA, USA
DynaMag™ -5 Magnet	Invitrogen, Thermo Fisher Scientific, Waltham, MA, USA
Dynamag™ -96 Side Magnet	Thermo Fischer Scientific, Waltham, MA, USA
GeneAmp® PCR System 9700	Applied Biosystems™, Thermo Fischer Scientific, Waltham, MA, USA
Heraeus™ Multifuge™ X3 Centrifuge	Thermo Scientific™, Thermo Fischer Scientific, Waltham, MA, USA
HulaMixer® Sample Mixer	Life technologies™, Thermo Fisher Scientific, Waltham, MA, USA
ImageQuant™ LAS 4000	Bio-Rad, Hercules, CA, USA
Nanodrop® ND-1000	Thermo Fisher Scientific, Waltham, MA, USA
PCR plate spinner	VWR
Qubit® 2.0 Fluorometer	Thermo Fisher Scientific, Waltham, MA, USA
QuantStudio™ 12K Flex	Applied Biosystems™, Thermo Fischer Scientific, Waltham, MA, USA

Thermomixer comfort	Eppendorf, Hamburg, Germany
2720 Thermal cycler	Applied biosystems, Thermo Fisher Scientific, Waltham, MA, USA
4200 TapeStation	Agilent, Santa Clara, CA, USA

Table S5. Softwares, versions, manufacturers and references

Product name	**Supplier**	**Reference**
Bowtie2 v2.2.9.gnu		(Langmead & Salzberg, 2012)
BWA-MEM v0.07.17		(Li & Durbin, 2010b)
Excel	Microsoft, Redmond, WA, USA	
Haploview v4.2		(Barrett et al., 2005)
IGV v2.9.4	Broad Institute, Cambridge, MA, USA	(Robinson et al., 2011)
ImageQuant	Applied Biosystems™, Thermo Fischer Scientific, Waltham, MA, USA	
Nanodrop 3.0.0	Thermo Fisher Scientific, Waltham, MA, USA	
Plink v1.9		(Purcell et al., 2007)
Rstudio 1.3.1093	Rstudio, Inc., Boston, MA, USA	
Samtools v1.8		(Li et al., 2009)
TapeStation Controller Software v3.2	Agilent, Santa Clara, CA, USA	
Unphased v.3.0.13		(Dudbridge, 2008)
QuantStudio	Applied Biosystems™, Thermo Fischer Scientific, Waltham, MA, USA	

Table S5. Web pages and address

Page name	Web address
dbsnp	http://www.ncbi.nlm.nih.gov/snp/
ENSEMBL	http://www.ensembl.org/index.html
ENSEMBL VCF to PED converter	http://www.ensembl.org/Homo_sapiens/Tools/VcftoPed
GWAS catalog	http://www.ebi.ac.uk/gwas/
Ldlink	http://www.ldlink.nci.nih.gov/?tab=home
Primer3web	http://primer3.ut.ee
UCSC Genome browser	https://genome.ucsc.edu/
GPP sgRNA Designer	https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design