Norwegian University
of Life Sciences

Master's Thesis 2018 60ECTS
Faculty of Chemistry, Biotechnology and Food Sciences
Tor Lea

Functional studies of *ERAP2* associated with risk of autoimmune diseases

Anne Rydland
Biotechnology
Faculty of Chemistry, Biotechnology and Food Sciences

Functional Studies of *ERAP2* associated with risk of autoimmune diseases

Oslo University Hospital,
Department of Medical Genetics

and

The Norwegian University of Life Sciences,
Faculty of Chemistry, Biotechnology and Food Sciences

Acknowledgements

The work of this thesis was carried out in collaboration with the Department of Medical Genetics, at Oslo University Hospital (OUS), as part of a Master´s Degree in Biotechnology at the Norwegian University of Life Sciences (NMBU), main affiliation being the Faculty of Chemistry, Biotechnology and Food Science (KBM).

First and foremost, I would like to thank my supervisor Marte K. Viken, Phd, for doing an outstanding job devoting her time supervising and supporting me throughout this thesis. Her invaluable advice and knowledge has greatly contributed to my understanding of the field of genetics. Additionally, her lab prowess and guidance has helped me improve my laboratory skills and expanded my expertise by introducing me to several well established laboratory techniques. I am very grateful to Professor Benedicte A. Lie for being my co-supervisor, including me in the research group and giving me the opportunity to execute such an engaging research project. Her support, encouragement and feedback has been invaluable to this thesis. I would also like to thank Department Engineer Siri Tennebø Flåm for her immense patience and contribution to the study by always being present and willing to help when I was in need of assistance or advice during laboratory procedures. Furthermore, I would like to thank my supervisor at NMBU, Professor Tor Lea.

I would like to express a great deal of gratitude to the whole immunogenetics group for their warm welcome and for the way they included me in the group, involving me in both social and work related gatherings. Moreover, I have appreciated the company of my fellow master students, Cathrine McCoig and Marianne Staff Fredhjem, sharing the experience of performing a Master´s project at OUS.

Lastly, I am extremely grateful for the everlasting support and encouragement of my family, always inspiring me to strive for greatness.

Anne Rydland
Ås, May 2018

Abstract

Autoimmune diseases are prevalent in the global population affecting approximately 5-10% worldwide, with increasing incidence. The complexity of these diseases has made diagnostic assessment problematic, as they vary in their predisposing factors, environmental triggers and epigenetic influence. Consequently, many research projects have aimed to identify specific factors implemented in development and progression of autoimmune diseases. Genome wide association studies have identified a genetic association between the endoplasmic reticulum aminopeptidase gene, *ERAP2*, and several autoimmune diseases. However, the functional implications of the risk variants are not fully established. In addition, researchers have demonstrated correlation between high *ERAP2* expression and increased levels of MHC class I molecules on B-lymphocytes, and several studies have suggested an association between high MHC class I presentation and increased risk of autoimmune diseases. A genetic study detected two main haplotypes of *ERAP2* associated with differing expression levels of the gene. Already established as an *ERAP2* eQTL the researchers identified rs2248374 as an *ERAP2* splice SNP, with the G-allele causing an alternative splicing event of exon 10 resulting in an extended transcript, subsequently degraded through nonsense-mediated decay. Hence, rs2248374 is often regarded as the SNP causal to *ERAP2* expression. However, several genetic studies have identified other *ERAP2* eQTLs showing stronger correlation with *ERAP2* expression, including the novel eQTL SNP rs27302. Overall these findings have put an emphasis on acquiring knowledge concerning the regulation of *ERAP2* expression.

The aim of this thesis was to gain a better understanding of expression of *ERAP2* on both RNA and protein level, and investigate the influence of SNPs on gene expression levels. The main focus was the correlation between genotypes of the two *ERAP2* eQTLs, rs2248374 and rs27302, and *ERAP2* expression.

Microarray analysis based on thymic tissue data identified the presence of regulatory factors in *ERAP2* expression, with further eQTL analysis indicating a distinct pattern between gene expression and genotypes of rs27302. Western blot of selected thymic tissue samples supported this discovery. However, western blot of LCL samples representing additional populations to the thymic material obtained contradictory results, indicating that rs2248374 was the causal SNP. Furthermore, screening of the *ERAP2* exon 10 splice junction in all samples found the expression of alternatively spliced *ERAP2-208* transcripts to coincide with rs2248374 genotypes. Consequently, none of the SNPs were established as the sole causal

SNP regarding *ERAP2* expression. High LD was identified in the genomic region of *ERAP2* and between nine selected SNPs in the *ERAP2* region, with LD patterns varying between populations. The variation implied that a causal SNP(s) may be concealed due to strong LD between the SNP(s) and rs27302 in the thymic samples, and the SNP(s) and rs2248374 in several of the populations represented in the LCLs, meaning that the causal SNP(s) was causing an indirect signal in the rs2248374 and rs27302.

Future studies may provide a better understanding of factors regulating *ERAP2* expression, i.e. by sequencing the genomic region of the gene in the HapMap populations, possibly identifying new potential regulatory SNPs.

Sammendrag

Autoimmune sykdommer er svært utbredt i den globale populasjonen (ca. 5-10%), og forekomsten blir stadig hyppigere. Sykdommene påvirkes av genetiske risikofaktorer, miljøfaktorer og epigenetisk innflytelse, dette har gjort det vanskelig å diagnostisere pasienter. Denne problematikken har ført til at mange forskningsprosjekter har hatt som mål å identifisere spesifikke faktorer involvert i utviklingen av sykdom og sykdomsforløpet. Helgenom assosiasjonsstudier har identifisert en genetisk assosiasjon mellom endoplasmatisk retikulum aminopeptidase genet, *ERAP2*, og flere autoimmune sykdommer, men de funksjonelle implikasjonene av risiko variantene er ikke forstått fullt ut. Videre har forskere funnet en korrelasjon mellom høy ekspresjon av *ERAP2* og økte nivåer av MHC klasse I molekyler på B-lymfocytter, og også en assosiasjon mellom høy MHC klasse I presentasjon og økt risiko for autoimmune sykdommer. En genetisk studie fant at det i hovedsak er to haplotyper av *ERAP2* som uttrykkes, og at disse er assosiert med ulike ekspresjonsnivåer av genet. De identifiserte rs2238374 som en *ERAP2* eQTL og spleise SNP, der G-allelet førte til alternativ spleising i ekson 10, som resulterte i et forlenget transkript som ble degradert via nonsense-mediated decay. På grunn av disse funnene blir rs2248374 ofte referert til som den kausale *ERAP2* eQTLen. Likevel har andre genetiske studier oppdaget flere *ERAP2* eQTLer som viser sterkere korrelasjon med *ERAP2* ekspresjon, inkludert rs27302. Disse funnene har vist viktigheten av videre opparbeidelse av kunnskap om *ERAP2* regulering.

Målet med denne oppgaven var å få en bedre forståelse av *ERAP2* ekspresjon på både RNA og protein nivå, og undersøke innflytelsen av SNPer på genekspresjonsnivåene. Hovedfokuset var på korrelasjonen mellom de to *ERAP2* eQTLene, rs2248374 og rs27302, og *ERAP2* ekspresjon.

Mikromatrise analyser basert på data fra tymus vev fant at regulatoriske faktorer har innvirkning på ekspresjonen av *ERAP2*, og videre eQTL analyser observerte et distinkt mønster mellom genekspresjonen og genotypene i rs27302. Western blot analyse av utvalgte tymusprøver støttet disse funnene. Derimot fant western blot analyser av LCL prøver, fra individer fra flere populasjoner i tillegg til CEU i tymusene, bestridende funn som indikerte at rs2248374 var den kausale SNPen. I tillegg viste en undersøkelse av ekson 10 spleiseovergangen at ekspresjonen av alternativt spleisede *ERAP2-208* transkripter sammenfalt med rs2248374 genotypene. Som en konsekvens av dette ble ingen av SNPene etablert som den kausale SNPen i *ERAP2* ekspresjon. Høy LD ble observert i det genomiske

området av *ERAP2* og mellom ni utvalgte SNPer i *ERAP2* i seks HapMap populasjoner. Den observerte variasjonen i LD mønster mellom populasjonene tydet på at en eller flere kausale SNPer muligens er skjult av høy LD i området, som har ført til indirekte signaler i rs27302 i tymusprøvene og rs2248374 i LCL prøvene.

Fremtidige studier kan gi en bedre forståelse av underliggende faktorer som regulerer *ERAP2* ekspresjonen, for eksempel ved å sekvensere det genomiske området av genet i HapMap populasjonene, som kan føre til identifiseringen av nye potensielt regulatoriske SNPer.

Abbreviations

AIDs – Autoimmune diseases

AS – Ankylosing spondylitis

DNA – Deoxyribonucleic acid

cDNA – Complementary DNA

ER – Endoplasmic reticulum

ERAP – Endoplasmic reticulum aminopeptidase

eQTL – Expression quantitative trait loci

GAPDH – Glyceraldehyde-3-Phosphate Dehydrogenase

gDNA – Genomic deoxyribonucleic acid

GWAS – Genome wide association studies

HLA – Human Leucocyte Antigen

IBD – Irritable bowel disease

Ichip – Immunochip

IMDs – Immune-mediated diseases

JIA – Juvenile idiopathic arthritis

LCL – Lymphoblastoid cell line

LD – Linkage disequilibrium

MAF – Minor allele frequency

MHC – Major Histocompatibility Complex

miRNA – Micro ribonucleic acid

mRNA – Messenger ribonucleic acid

NMD – Nonsense-mediated decay

OUS – Oslo University Hospital

PCR – Polymerase chain reaction

RA – Rheumatoid arthritis

RNA – Ribonucleic acid

SLE – Systemic Lupus Erythematous

SNPs – Single nucleotide polymorphisms

T1D – Type 1 diabetes

TAP – Transporter associated with antigen processing

Table of contents

1. Introduction

1.1 The immune system

1.1.1 Innate and adaptive immunity

The immune system is divided into two main components, innate and adaptive immunity (Figure 1) (Elliott et al., 2014). Innate immunity is the primary defense against foreign substances such as microbes and other pathogens, and reacts by eliciting an immediate, rapid immune response. It consists of both humoral and cellular components such as complement proteins, acute phase reactants, cytokines, macrophages, granulocytes, mast cells and natural killer cells (Lea, 2013). Innate immunity is nonspecific as the cells express invariant receptor molecules termed pattern recognition receptors able to recognize conserved microbial molecules that are situated on or excreted by several types of microorganisms (Akira et al., 2006; Lea, 2013; Parkin & Cohen, 2001). In situations where the innate immune system is insufficient in removal of an infectious agent, adaptive immunity with its specificity, diversity and immunological memory is recruited (Khan Academy, 2017; Lea, 2013).

Adaptive immunity consists primarily of B-lymphocytes and T-lymphocytes, representing humoral and cell-mediated immunity, respectively**.** Progenitor cells of the two are derived from bone marrow. B-lymphocytes mature in the bone marrow, while T-lymphocytes migrate to thymus where the maturation process proceeds and T-cell receptors are developed. An important notion is that the development of naïve immunocompetent T-lymphocytes is age-dependent as thymus changes over time. The gland is highly active during fetal life and the first years after birth, and the organ continues to grow until puberty. During onset of puberty an involution process involving replacement of lymphocytes with fat is initiated (Gui et al., 2012). These changes result in decreased T-cell output (Aw & Palmer, 2012).

In the periphery, T-cells and B-cells are circulating between blood and lymph (passing through secondary lymphoid organs e.g. lymph nodes, spleen and Peyer´s patches*)*, surveilling the body for intruders. Antigens are transported by dendritic cells from the infected tissue to the lymph nodes where immunocompetent cells may recognize them (Lea, 2013; Santori, 2015). Specific lymphocytes are then imprinted with information, through tissue specific adhesion molecules, on the location of the infection, where they migrate to initiate an immune response by following the blood stream (Lea, 2013).

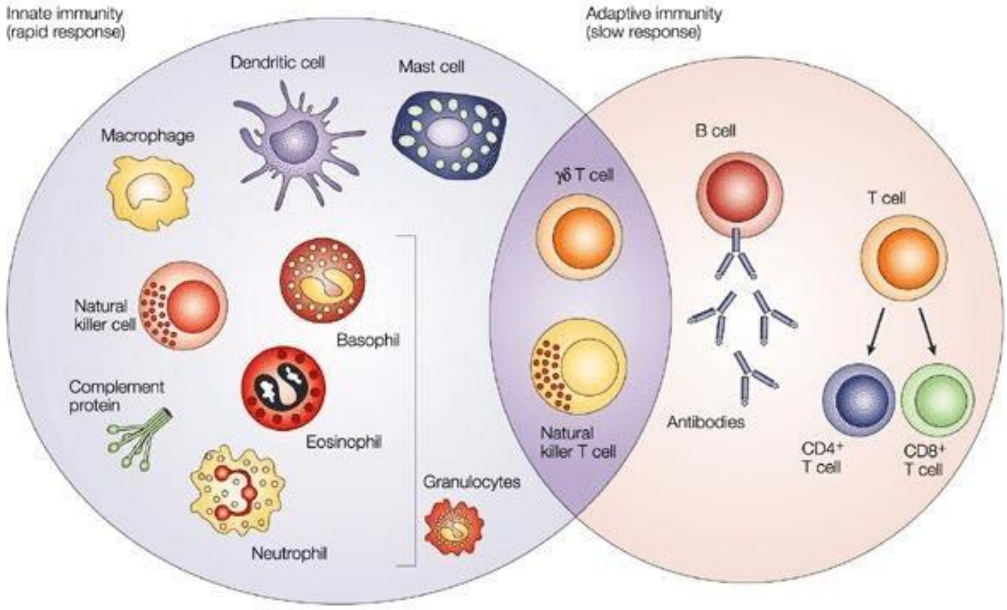

Figure 1. The immune system. The immune system is comprised of two main components, innate and adaptive immunity. Innate immunity is the nonspecific, primary defense against foreign pathogenic substances and involves granulocytes, natural killer cells, macrophages, dendritic cells, mast cells and complement proteins. The adaptive immune response which is highly specific and exhibits immunological memory consists of T-lymphocytes and B-lymphocytes, expressing both cell mediated and humoral immunity, respectively. Natural killer T-cells and gamma-delta T cells share properties with cells of both innate and adaptive immunity (Dranoff, 2004).

T-lymphocytes and B-lymphocytes are highly specialized cells, with each cell expressing a great number of identical T-cell receptors or B-cell receptors targeting one specific antigenic structure. Although the two cell types exhibit the same degree of specificity, they differ in the structure of their receptor molecules and antigen-recognition mechanisms. (Alberts B, 2002; Lea, 2013)

1.1.2 T-cells

The population of T-cells consists of several subgroups expressing different effector functions mainly as, killer cells, helper cells or regulatory cells (Lea, 2013). The helper T-cells are important contributors in B-cell and cytotoxic T-cell activation, and to the destruction of microbes by macrophages (Alberts B, 2002). Cytotoxic T-cells on the other hand kill infected cells and destroy tumors. One important property of T-cells is their ability to detect endogenous pathogenic agents as well as exogenous pathogens, enabling the immune system to surveille the endogenous environment (Lea, 2013).

Whilst activated B-cells have the ability to secrete antibodies that can bind antigen directly, the T-cells are dependent on cell-cell interaction to elicit an immune response. An important notion is that the T-cell receptor can only recognize antigen bound to a major histocompatibility complex (MHC) class I or II self-molecule on the cell surface of infected cells. (Alberts B, 2002; Janeway CA Jr, 2001; Lea, 2013)

1.1.3 The major histocompatibility complex and the HLA class I antigen processing pathway

The MHC molecules are encoded by the *MHC genes* located on chromosome 6p21 in humans. The MHC consists of more that 200 genes and encompasses the important genes encoding the MHC molecules, termed human leukocyte antigen (HLA) molecules in humans. There are a variety of HLA molecules with affinity for different antigens resulting in the presentation of a wide pool of antigens to T-cells. They bind with reduced specificity which enable them to bind molecules with similar amino acids in the HLA binding site. (Janeway CA Jr, 2001; Lea, 2013).

HLA class I molecules are expressed by all nucleated cells in the human body, whilst HLA class II molecules are expressed by macrophages, dendritic cells and B-cells (Janeway CA Jr, 2001; Lankat-Buttgereit & Tampe, 2002). The binding of the T-cell receptor to the peptide on the HLA molecule may be unstable, thereby demanding the presence of co-receptors for proper T-cell activation. Two important co-receptors are the CD4 and CD8 molecules, which are present on the surface of helper T-cells and cytotoxic T-cells, respectively. The CD4$^+$ helper T-cells only recognize antigen bound to HLA class II molecules, whilst the CD8$^+$ cytotoxic T-cells recognize antigen bound to HLA class I molecules (Lea, 2013). There is a complex process of antigen processing taking place prior to antigen presentation on the cell surface, with the two HLA classes being implemented in different pathways.

The classical HLA class II antigen processing pathway involves interaction with exogenous pathogens or pathogenic proteins engulfed by cells through endocytosis. As opposed to this, the HLA class I antigen processing pathway interacts with endogenous peptides (Figure 2). HLA class I molecules are folded and assembled in the endoplasmic reticulum (ER) lumen and consist of a heavy α-chain and β_2-microglobulin (Janeway CA Jr, 2001). In ER, the HLA class I molecules interact with specific antigens. These antigens are the product of a complex degradation cascade involving several enzymes situated in different parts of the cell. Infectious agents may reside inside cells producing or consisting of antigenic peptides. In the cytoplasm, the antigenic proteins are conjugated with ubiquitin to be recognized by a class of proteasomes termed immunoproteasomes. The immunoproteasomes have specific cleavage and recognition specificities that degrade ubiquitin-conjugated antigens into peptide fragments of approximately 15 amino acids. The small peptide fragments are transported from cytosol into the ER lumen by the transporter associated with antigen processing (TAP), a transmembrane heterodimer consisting of TAP1 and TAP2, where specialized aminopeptidases interact with fragments exhibiting certain properties, cleaving them into peptides of 8-10 amino acids that fit in the groove of the HLA class I molecules (Blum et al., 2013; Hattori & Tsujimoto, 2004; Lankat-Buttgereit & Tampe, 2002). Two important aminopeptidases, endoplasmic reticulum aminopeptidase (ERAP) 1 and 2, will be discussed in detail later. The precisely processed peptides are then loaded onto the appropriate HLA class I molecule and the complex is transported to the cell surface for presentation to CD8$^+$ cytotoxic T-cells (Blum et al., 2013; Janeway CA Jr, 2001).

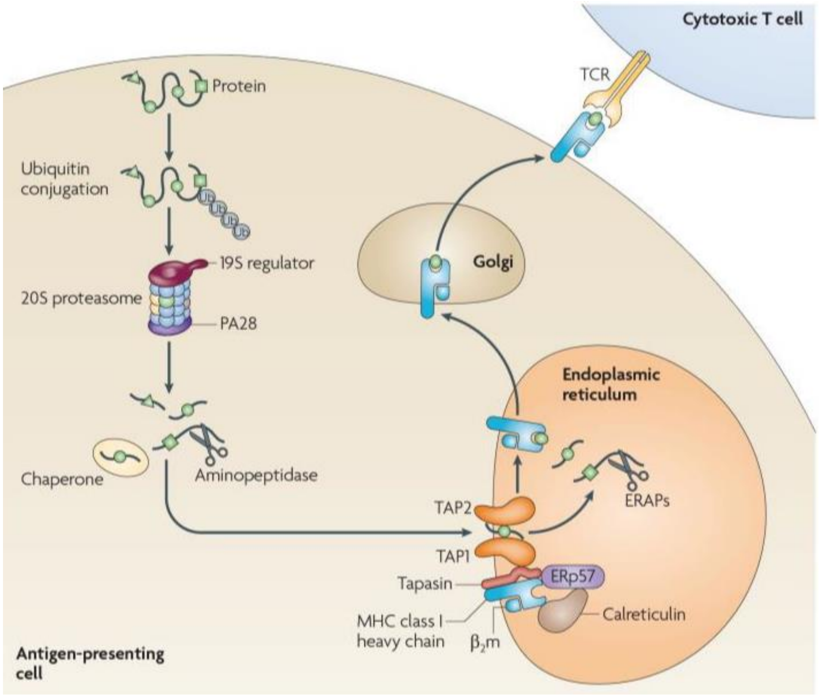

Figure 2. HLA class I antigen processing pathway. Endogenous proteins are conjugated with ubiquitin before being degraded by the proteasome and transported into the ER. ERAP1 and 2 collaborate in the ER in trimming of extended peptides by hydrolyzing specific N-termini based on properties and length, resulting in a coordinated presentation of antigenic peptides on HLA class I molecules to CD8$^+$ cytotoxic T-cells (Groettrup et al., 2010).

T-cell receptors can only recognize one population of ligands with a specific amino acid combination. Fortunately, the immune system develops a variety of T-cells with differing receptor properties, resulting in a T-cell population able to recognize a range of pathogenic antigens (Janeway CA Jr, 2001; Khan Academy, 2017; Lea, 2013).

As implied in the paragraphs above, the immune system is a complex and intricate system responsible for protecting the body against foreign invaders. It is synergistic in the collaboration between innate and adaptive immunity; adaptive immunity would not develop without innate immunity, and innate immunity would not reach its full potential without adaptive immunity (Lea, 2013).

1.2 Autoimmune diseases

Autoimmunity is the result of an aberration from self-tolerance by the immune system (Bolon, 2012). In normal, healthy individuals, the immune system exhibits tolerance to self-antigens when surveilling the body for intruders. Even though some autoreactive T-cells and B-cells are released into the peripheral lymphoid tissue under normal conditions, they are inhibited by mechanisms of peripheral tolerance (Mueller, 2010). In patients exhibiting autoimmune diseases (AIDs) one or several malfunctions in the immune systems tolerance mechanisms are present, leaving the immune cells to recognize self-molecules as foreign invaders, destroying healthy tissue (Bolon, 2012).

Tolerance is classified as either central or peripheral (Figure 3) (Lea, 2013). Central tolerance is an intricate process of adaptive immunity where lymphocytes are carefully selected in the primary lymphoid organs before being released to the circulatory system. During maturation of T-lymphocytes in thymus three processes determine their fate. Here T-cells are presented to self-antigen/HLA complexes. Depending on their affinity for these complexes they go through negative selection, positive selection or die by neglect (Boehm et al., 2013; Janeway CA Jr, 2001). Autoreactive T-cells with high affinity towards self-antigen/HLA complexes pose a risk to development of AIDs and are removed by apoptosis during negative selection. T-cells expressing no affinity toward these complexes die by neglect, while T-cells with low affinity go through positive selection and may ultimately become part of the adaptive immune system (Boehm et al., 2013; Janeway CA Jr, 2001).

Central B-cell tolerance is maintained through deletion, editing or anergy (Meffre & Wardemann, 2008). B-cells exhibiting high affinity towards self-antigens die by apoptosis in a process termed clonal deletion. Autoreactive B-cells with moderate affinity can go through secondary recombination to modify the B-cell receptor and decrease its autoreactivity or become inactivated in an anergic cell state. The anergic cells are released into the periphery where their unresponsiveness results in removal of the cells (Gay et al., 1993; Nemazee, 2017). Unfortunately, some autoreactive B-cells and T-cells may escape central tolerance and are released into circulation, where peripheral tolerance is responsible of maintaining the immunological balance. There are several mechanisms of peripheral tolerance that inhibit autoimmunity, including activation-induced cell death by apoptosis, anergy induction of lymphocytes lacking a secondary activation signal and inhibition by regulatory T-cells (Lea, 2013; Maher et al., 2002). If an individual is exposed to malfunctions in either of these mechanisms, an AID may develop.

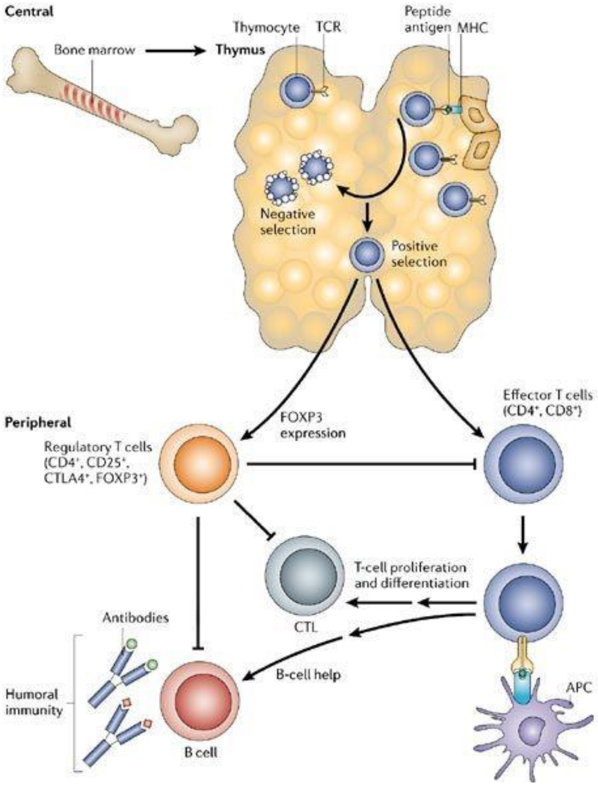

Figure 3. Immunological tolerance. Central tolerance is an important mechanism in preventing the escape of self-reactive lymphocytes into the periphery. The central tolerance mechanisms are imposed on developing B-lymphocytes and T-lymphocytes in the primary lymphoid organs; bone marrow and thymus. Peripheral tolerance is responsible for maintaining the immunological balance in the periphery, e.g. preventing activation of self-reactive circulating lymphocytes (Gregersen & Behrens, 2006).

Autoimmune diseases, such as Systemic Lupus Erythematosus (SLE), Rheumatoid Arthritis (RA) and Type 1 diabetes (T1D), are prevalent in the global population affecting approximately 5-10% worldwide and their incidence is increasing (Lerner & Matthias, 2015; Marson et al., 2015). Observations from most diseases show a higher prevalence in women than men (Bolon, 2012), although there are some exceptions e.g. Primary Sclerosing Cholangitis (Williamson & Chapman, 2015). Researchers have also discovered an association between having one AID and risk of developing a second AID (Anaya et al., 2007; Bolon, 2012; Nacu et al., 2015), and that susceptibility to autoimmune diseases runs in families, with members displaying similar or different AIDs (Bolon, 2012). Currently there is no cure for autoimmune diseases, and disease discovery is usually dependent on symptoms occurring upon progression of the disease, resulting in treatment being aimed at controlling the symptoms. This indicates a need for early discovery of AIDs as well as development of proper and effective disease treatment, and ultimately a cure.

A major issue in diagnostic assessment of AIDs is the lack of knowledge concerning the etiology and pathogenesis of these complex diseases. They vary in their predisposing factors, environmental triggers (e.g. xenobiotics, pathogens) and epigenetic influence, as well as in disease development and target tissues (Ayensu et al., 2004; Bolon, 2012). Consequently, the aim of many research projects is the identification of specific factors involved in the development and progression of AIDs, assessing individual diseases and similarities between diseases, as well as their functional role in the disease (Senolt et al., 2009; St Clair, 2009). AID patients often show disparities in treatment response, adding another layer of complexity to AID assessment and treatment (Liu et al., 2014).

Autoimmune diseases are classified as either organ-specific or systemic based on their target tissue and autoantibody production. Organ-specific AIDs affect a particular organ or tissue in the body, while systemic AIDs implicate several organ systems (Fridkis-Hareli, 2008; Janeway CA Jr, 2001). T1D is an organ-specific disease where the immune system develops autoantibodies acting against the insulin producing ß-cells of the pancreas, causing disruption of the glycemic control in the affected individual (Graham et al., 2012). RA on the other hand is a systemic disease affecting the synovial membrane, bone and cartilage of several joints, and in some instances other parts of the body. Autoantibodies are widely used as markers for AIDs (Janeway CA Jr, 2001; Sener, 2015), but the presence of autoantibodies alone is not sufficient to diagnose a patient with an AID as they can be present in healthy individuals and the absence of these do not exclude disease. The assessment needs to include other clinical findings as well in order to confirm diagnosis (Aggarwal, 2014). Like AIDs, autoantibodies are either organ-specific, targeting tissue specific autoantigens, or systemic, targeting ubiquitous autoantigens.

1.3 Genetics and Genome-wide association studies (GWAS) in autoimmune diseases

The human genome contains approximately 20,500 genes (NIH, 2016). Human diversity is based on a 0.1% difference between any two genomes (Goris & Liston, 2012; Gregersen & Olsson, 2009), where common genetic variants (minor allele frequency (MAF)>1%) account for 90%. The remaining 10% is found in rare genetic variants with MAF<1%. Even though common variants dominate the genome, the rare variants outnumber these when assessing the total number in a population (Goris & Liston, 2012).

In genetics, traits are separated into monogenic and polygenic traits. The term monogenic is assigned traits where there is a single gene or allele influencing the phenotype. Monogenic or Mendelian diseases run in families and the disease phenotype is determined by either a recessive or dominant allele, located on an autosomal or sex chromosome (Celedón JC, 2017). Polygenic traits, such as most AIDs, are determined by the influence of multiple genes, environmental factors and epigenetics. This means that the genome of an individual may be comprised of genetic variants that put them at risk of developing a complex disease, but the disease only develops in the presence of the proper environmental triggers. The consequence of the involvement of many factors is individual phenotypic differences, spanning over a continuous range of phenotypes. This heterogeneity is present in autoimmune diseases as well, where one disease can show differences in clinical and biochemical manifestations (Cho & Feldman, 2015; Gregersen & Olsson, 2009). The complexity of AIDs and the small phenotypic contribution of each gene make it difficult to detect causal genetic variants and completely understand their impact on the disease (Marson et al., 2015).

1.3.1 Genetic variation

Genetic variation is the source of individual differences between any two members of a population, with some variants being implicated in disease susceptibility. There are mainly three types of genetic variation; structural variation through copy number variation and chromosomal rearrangements, indels through insertion and deletion and single nucleotide polymorphisms (SNPs) (EMBL-EBI, 2018). SNPs are variants differing at a single nucleotide and are often biallelic. There is on average one SNP per 300 nucleotides, meaning that the human genome contains approximately 10 million SNPs (NIH, 2018a). These are often found in the noncoding regions of the genome (Rada-Iglesias, 2014).

Several methods have been and are being utilized to assess the genetic contribution of variants in disease. The development of genome-wide association studies (GWAS), which had its first successful study published in 2005 (Klein et al., 2005), has greatly contributed to the discovery of disease associated genes and genetic regions. To fully understand the magnitude of GWAS, it is important to acknowledge the important technological advancements leading to its development and the methods that were utilized prior to and during its evolution.

1.3.2 Candidate gene and linkage studies

Before the development of GWAS, mainly candidate gene approaches and family based linkage studies were utilized in the attempt to identify and map disease associated loci in the human genome. Linkage studies have been successful in mapping genomic loci associated with Mendelian diseases, by relying on the co-segregation of markers and causal variants using family and pedigree data (Dueker & Pericak-Vance, 2014; Hirschhorn & Daly, 2005; Visscher et al., 2012). Loci residing in close proximity on the same chromosome are more likely to avoid segregation by recombination during meiosis and thus more likely to be inherited together (Visscher et al., 2012). Mapping of genes and variants involved in common complex diseases, such as AIDs, has proven a more challenging task as co-segregation of chromosomal regions associated with AIDs are restricted within families (Gregersen & Olsson, 2009) and the diseases involve multiple loci with low effect size, making causal variants difficult to detect (Hirschhorn & Daly, 2005; Marson et al., 2015). Still, some quantitative trait loci have been successfully mapped by using linkage analysis but these loci usually only account for a small portion of the disease heritability (Gregersen & Olsson, 2009; Hirschhorn & Daly, 2005; Visscher et al., 2012).

Researchers have performed candidate gene studies as an alternative to linkage studies, which led to the identification of several genes involved in common disease. The method is based on the hypothesis that a certain gene, based on the knowledge of their function, is likely to be involved in disease susceptibility. Although the method has led to the identification of several important disease associated genes, it is a hypothesis-based method that only uncover a fraction of risk loci (Gregersen & Olsson, 2009; Hirschhorn & Daly, 2005).

The challenges encountered when candidate-gene approaches and family based linkage studies dominated can in large part be explained by the limited disease knowledge and restricted resources available at that time. The technological tools that are used today were in its early stages, and the methods used were time consuming and expensive. Additionally, insight into genes involved in disease and their biological function was modest, restricting the identification of candidate genes, and the absence of a reference map of common genetic variation limited the amount of polymorphisms studied (Ricano-Ponce & Wijmenga, 2013).

1.3.3 Technological advances

There are several technological advances that laid the groundwork for genome wide association studies. The completion of the Human Genome Project in April 2003, which is a map of the entire, human DNA sequence (NIH, 2015), and the development of high-throughput sequencing technology and microarrays were important contributors. Another crucial factor was the development of publicly available genotype databases, such as the HapMap and 1000 Genomes Project, describing common patterns of human genetic variation involved in health and disease (NIH, 2017; NIH, 2018b).

1.3.4 Linkage disequilibrium

GWAS rely on linkage disequilibrium (LD) (Visscher et al., 2017), which is the non-random association of alleles at separate loci in a population. If two alleles at different loci on the same chromosome are inherited together more frequently in a population than what is expected by chance they are in LD (Slatkin, 2008). An important feature of LD is that it provides the means for SNP tagging, which entails that the typing of one SNP simultaneously give information about the alleles of other SNPs within the same LD block (Hirschhorn & Daly, 2005). This enables a more effective analysis by reducing the number of SNPs that needs to be genotyped, subsequently reducing cost and time spent typing SNPs.

The degree of LD is usually measured in r^2 and D'. The first is the correlation coefficient, which is bidirectional. When $r^2=1$ only two of the four expected haplotypes are present in the population, and perfect LD is obtained. The opposite is observed when $r^2=0$ and the two loci are in equilibrium, then all four haplotypes are present in the population. The second measurement, D', is unidirectional. D'=1 indicates that one or several of the expected haplotypes are absent from the population, suggesting LD is high. By implementing the information of the correlation coefficient one can determine if one ($r^2<1$) or two ($r^2=1$) haplotypes are missing. (Gregersen & Olsson, 2009; Viken, 2008)

One of the challenges faced with LD is that it is difficult to determine the causal SNP. If a SNP in high LD with other SNPs show association with disease it is difficult to conclude which SNP(s) is causal. Still, the non-random distribution of alleles between individuals in LD blocks makes LD a very useful tool to geneticists assessing common genetic variation in complex diseases (Gregersen & Olsson, 2009).

1.3.5 GWAS

Since 2005, GWAS have been used to detect associations between genetic variants and complex diseases by performing comparative studies spanning the entire genome without any prior assumptions on candidate genes or causal variants (Liu et al., 2013; Visscher et al., 2017). A GWAS is facilitated using genotyping SNP chips containing from 200,000 to more than 2 million SNPs (Visscher et al., 2017), enabling researchers to compare healthy individuals with patients to identify genomic variants associated with the disease(s) in question. Through the last years GWAS has surveyed genetic variants that are common in populations usually with a MAF >1% by investigating samples from one population at a time (Visscher et al., 2017). It is important to separate individuals by ancestry as disease association may be confused with population stratification and can lead to false positive and false negative associations (Liu et al., 2013). An interesting discovery is that most disease associated GWAS variants are in noncoding regions of the genome, presumably having regulatory functions (Ricano-Ponce & Wijmenga, 2013; Zhang & Lupski, 2015).

The overall genomic coverage of GWAS SNPs yields a relatively low resolution, e.g. a chip with 550,000 SNPs will on average include three SNPs for each gene in the human genome. Fine-mapping of disease associated loci increase the resolution, by identifying the most significant risk loci from GWAS and resequencing the area(s) to try and identify causal variants (Ricano-Ponce & Wijmenga, 2013). In 2009, the Immunochip Consortium developed a SNP array containing approximately 200,000 SNPs for deep replication and fine-mapping of GWAS significant loci ($p<5x10^{-8}$). The immunochip (Ichip) cover 186 genomic regions containing risk variants for 12 immune-mediated diseases (IMDs) including several AIDs like T1D and RA, with approximately 3000 SNPs for each disease (Trynka et al., 2011). The disease associated SNPs were selected from available GWAS data, enabling deep replication and subsequent identification of which GWAS SNPs are in fact associated with disease. Based on the genetic relatedness of several IMDs, the Ichip has also identified genes and genetic regions that are associated with several of the diseases included on the chip (Cortes & Brown, 2011). Note that the sharing of genetic susceptibility loci between IMDs includes variating effects on the different diseases, e.g. discordancy and allelic heterogeneity (Parkes et al., 2013). Additionally, the Ichip cost is significantly lower than GWAS chips, permitting genotyping of a larger sample of individuals (Cortes & Brown, 2011; Trynka et al., 2011).

Although GWAS has expanded the knowledge concerning association between genetic variants and traits in populations, there are limitations to the method which are important to consider when analyzing GWAS data. Firstly, the SNP arrays used in GWAS target common genomic variants, resulting in a bias towards these variants. Secondly, the issues encountered by SNP tagging makes it difficult to determine the causative SNP(s) (Visscher et al., 2017). If assumptions are made without taking this into account it may lead to false positive findings. Nevertheless, GWAS has been widely used and led to the identification of hundreds of AID loci in the human genome, mostly with low effect size. An interesting finding is that several AID risk loci are shared between different AIDs (Ramos et al., 2011; Ricano-Ponce & Wijmenga, 2013; Richard-Miceli & Criswell, 2012). Studies have also shown that presence of one disease enhance the risk of developing a second AID. For instance, this has been observed in patients with Grave´s disease, RA and SLE where the presence of T1D is prevalent (Ricano-Ponce & Wijmenga, 2013).

In developing a better understanding of complex diseases, such as AIDs, researchers are performing functional studies to try and elucidate the effect of the genetic variants on disease susceptibility. GWAS is important in detecting genetic variants associated with disease, but it alone is insufficient when investigating the biological relevance of the individual variants.

Several molecular mechanisms are associated with physiological variation between individuals and populations with some phenotypes increasing disease susceptibility (Lappalainen et al., 2013; Michaelson et al., 2009). One of these mechanisms is gene expression.

1.3.6 Gene expression, regulation and alternative splicing

Nucleotides are the building blocks of DNA. Some of these are situated in gene regions and may be expressed depending on the function of the gene and its relevance in the specific cell. Gene expression is achieved through transcription by which RNA is produced. Some RNA originates from protein coding genes and is translated into proteins, while others function as various types of RNA, e.g. rRNA, tRNA, microRNA (miRNA). A balanced expression of genes is vital to cells as this dictates cell function. To maintain this balance, the cell is equipped with several checkpoints in the biological pathway from gene to RNA to protein, and additional regulation following protein synthesis (Nature, 2014). The checkpoints include chromatin accessibility, the presence of transcription factor proteins that promote or repress gene expression, the processing of RNA, miRNA activity degrading mRNA and influencing translation, and protein modification where degradation or tagging affects protein activity (Lewin, 2006; Nature, 2014).

Alleles of certain loci, termed expression quantitative trait loci (eQTL), may also affect gene expression. These SNPs are usually bi-allelic, with the two alleles exhibiting different effects on gene expression. The eQTLs are separated into two categories based on where they are located relative to the gene they are influencing. An eQTL position overlapping the gene region of the gene it is influencing is termed a *cis*-eQTL, while eQTLs located further away from the gene, often on a different chromosome, are defined as *trans*-eQTLs (Nica & Dermitzakis, 2013). eQTLs can be identified through eQTL analyses, by comparing genotypic data and expression data in a sample set attempting to detect significant correlation between polymorphic genomic regions and levels of gene expression (Figure 4) (Michaelson et al., 2009). By genotyping two groups of individuals with two specific phenotypes, e.g. patients and a healthy control group, and combining this information with data illustrating gene expression, novel eQTLs involved in the disease(s) in question may be detected (Janeway CA Jr, 2001).

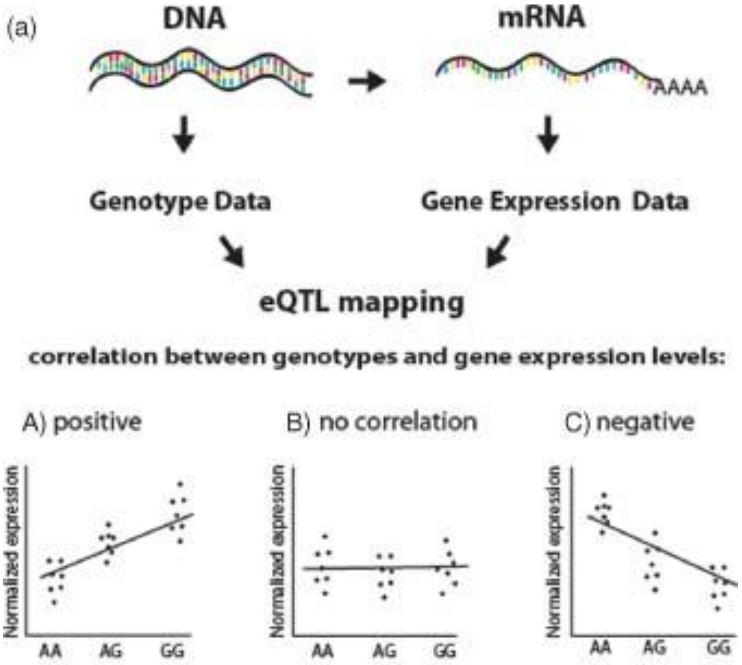

Figure 4. eQTL mapping. Comparison of genotype and gene expression data is crucial to eQTL mapping. The association between genotypes and expression levels can be either positive, negative or show no correlation. Modified from Hrdlickova et al. (Hrdlickova et al., 2011).

The effect of gene expression on disease susceptibility is not only dependent on expression level, but also on phenotypic variation. Genes contain several coding sequences termed exons and noncoding sequences termed introns. During processing of pre-mRNA exons are spliced together to form mature mRNA strands, excluding the introns. Alternative splicing contributes to phenotypic variation by being a biological event where exons are spliced in various combinations resulting in a pool of mRNA isoforms that can function as RNA or be translated into proteins with varying function and size, increasing protein diversity (Coulombe-Huntington et al., 2009; Park et al., 2018). There are several splicing mechanisms, mainly exon skipping, intron retention, alternative 3´and 5´splice sites and in rare cases mutually exclusive exons (Nilsen & Graveley, 2010; Park et al., 2018; Pohl et al., 2013). Intron retention is a process by which an intron is included in the pre-mRNA strand in contrast to its usual exclusion. This event was long believed to be a consequence of mis-splicing, but resent studies has identified it as a conserved mechanism of alternative splicing (Gregersen & Olsson, 2009; Wong et al., 2013; Wong et al., 2016). mRNA containing introns are usually removed through nonsense-mediated decay (NMD) or nuclear retention and exon degradation. NMD is a biological process elicited by the presence of a premature stop codon, that prevents the production of truncated or nonfunctioning proteins (Andres et al., 2010; Chang et al., 2007; Wang et al., 2011). The degradation mechanisms can result in reduced levels of intron retaining transcripts in the cell, complicating their detection (Wong et al., 2016).

There is evidence indicating that some phenotypic variation is based on genetic variants (splice SNPs) influencing the alternative splicing event (Andres et al., 2010; Faber et al., 2011). These findings have contributed to an increased importance of acquiring insight into the functionality of genes and genetic variants involved in disease susceptibility (Hassan et al., 2014).

1.4 ERAP2

GWAS have unveiled a genetic association between the endoplasmic reticulum aminopeptidase gene, *ERAP2,* and Ankylosing Spondylitis (AS), Psoriasis, Irritable Bowel Disease (IBD) and Birdshot chorioretinopathy (Agrawal & Brown, 2014; Kenna et al., 2015; Kuiper et al., 2014). In addition, researchers have demonstrated correlation between high *ERAP2* expression and increased levels of cell surface HLA class I molecules on B-lymphocytes (Andres et al., 2010), and several studies on AIDs have suggested that high HLA class I presentation is associated with increased risk of autoimmune diseases (Mozes et al., 2005; Napolitano et al., 2002; Skog et al., 2015).

The *ERAP2* gene is situated on chromosome 5q15, containing 19 exons (including UTR) and spanning across 43.8 kb. Located adjacent to the gene is the closely related *ERAP1* gene and *leucyl-cystinyl aminopeptidase* (*LNPEP*) (Figure 5) (Andres et al., 2010; Kuiper et al., 2014). The two major transcripts encoded by *ERAP2,* detected in lymphoblastoid cell lines (LCLs), are translated into proteins of subsequently 960 and 532 amino acids. The first is a full length protein comprised of all 19 exons, while the second undergo an alternative splicing event of exon 10. The alternatively spliced *ERAP2* mRNA is extended in exon 10 with 56 extra nucleotides. This occurs when a downstream splice site at position 56 of intron 10 is preferred above the standard splice site in position 69 of exon 10. The extended mRNA strand contains a premature stop codon, and thereby encodes a truncated protein (Andres et al., 2010).

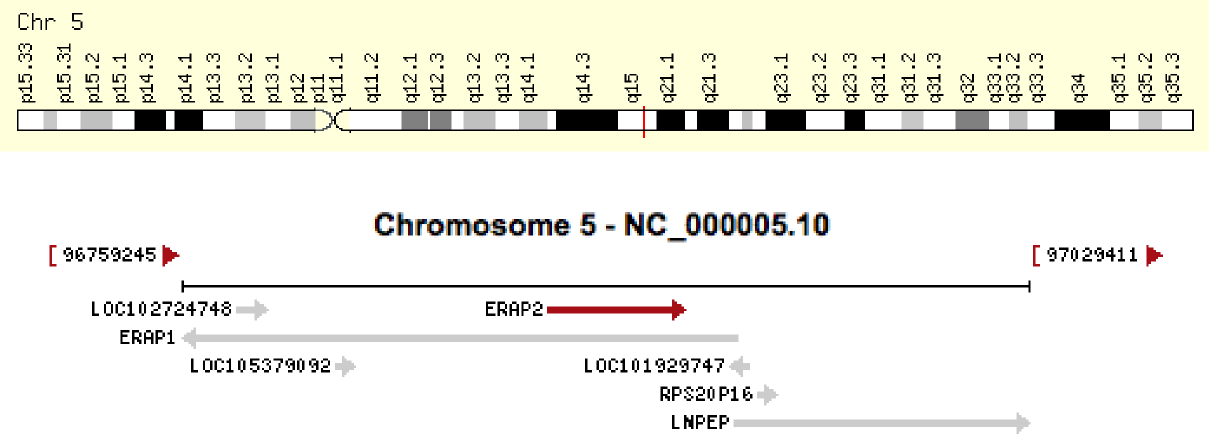

Figure 5. *ERAP2* location in the GRCh38p7 assembly. The gene is situated on chromosome 5 between *ERAP1* and *LNPEP* on chromosome 5 (5q15) (GeneCards, NA-b; NCBI, 2018).

The *ERAP2* gene encodes a zinc-metalloaminopeptidase of the M1 protease family residing in the endoplasmic reticulum. The enzyme is important for the final trimming of antigenic precursor peptides before HLA class I loading (Fierabracci et al., 2012; Haroon & Inman, 2010). As mentioned in chapter 1.1.3, the trimming of antigenic peptides is initiated by proteasomes in cytosol. This process yields peptides with the proper C-termini for HLA class I loading but the N-termini may be extended with one or several residues. ERAP2 work in a concerted fashion with ERAP1 in trimming of extended peptides by hydrolyzing specific N-termini based on properties and length, resulting in a coordinated presentation of antigenic peptides on HLA class I molecules (Fierabracci et al., 2012; Papakyriakou & Stratikos, 2017; Vitulano et al., 2017). ERAP2 show preference toward basic residues, primarily arginine and lysine, while ERAP1 trim hydrophobic residues, primarily leucine (Fierabracci et al., 2012; Tsui et al., 2010; Vitulano et al., 2017). The aminopeptidases can form heterodimers that allosterically activate ERAP1 and trim residues with a faster rate. Still, only 30% of ERAPs in live cells are heterodimers and aspects concerning their function *in vivo* are not well understood (Lopez de Castro et al., 2016).

According to Andrés et al. balancing selection has maintained two main haplotypes of *ERAP2*, Haplotype A (0.44) and Haplotype B (0.56), although with some outliers. The two haplotypes are associated with differing levels of *ERAP2* expression, with Haplotype B showing the lowest expression. Haplotype B encodes the variant of *ERAP2* that undergoes the alternative splicing event of exon 10 resulting in the extended mRNA degraded by nonsense-mediated decay (Andres et al., 2010). Genetic studies have proposed the bi-allelic rs2248374

(G/A) as the causative splice SNP, with the G-allele encoding the alternative splicing (Andres et al., 2010; Coulombe-Huntington et al., 2009). This SNP is located in the 5´canonical splice site of exon 10. It is suggested that the alternatively spliced transcript undergoes NMD, removing the mRNAs containing a premature stop codon (Andres et al., 2010). This is based on experiments on NMD inhibition using emetine as inhibitory agent. Haplotype B exhibited low *ERAP2* mRNA expression under normal conditions compared to mRNA derived from Haplotype A, while under NMD inhibiting conditions similar mRNA amounts were observed from both haplotypes (Andres et al., 2010). Because it is the alternatively spliced transcript that undergoes NMD, rs2248374 has been suggested as the eQTL causing reduced expression of *ERAP2* (Andres et al., 2010; Groettrup et al., 2010; Harvey et al., 2011).

Other studies have performed *ERAP2* eQTL analyses that resulted in the discovery of several eQTLs (rs10044354, rs2762, rs27302) showing higher correlation with *ERAP2* expression compared to rs2248374 (Cheung et al., 2005; Gabrielsen et al., 2016b; Kuiper et al., 2014). Still, the current opinion stating rs2248374 as the SNP regulating *ERAP2* expression remains unsettled.

ERAP2 eQTLs had not been explored in thymus before the research of Gabrielsen et al. in 2016. Their studies found rs27302 as the peak *ERAP2* eQTL (P=8.22x10^{-23}), showing the highest correlation with *ERAP2* expression. The *ERAP2* region contain several AID risk variants, e.g. rs2910686 (AS), rs1363907 (IBD), rs27290 (Juvenile idiopathic arthritis (JIA)) and rs27293 (JIA), and LD analysis showed that the novel eQTL, rs27302 (Figure 6), exhibited strong LD with these SNPs ((rs2910686, r^2=0.94), (rs1363907, r^2=0.94), (rs27290, r^2=0.94) and (rs27293, r^2=0.95)), indicating an overlap between the novel eQTL and the AID risk loci. Further analysis revealed a novel haplotype comprised of all AID risk loci and the rs27302 G-allele, with individuals homozygous for the rs27302 G-allele having the highest expression of *ERAP2* (Gabrielsen et al., 2016b). After discovery of rs27302 as a non-tissue specific eQTL in *ERAP2*, they investigated the significance of the previously suggested *ERAP2* eQTL, rs2248374, in their thymic data set and found it to be less correlated with *ERAP2* expression (P=2.74x10^{-9}) than several of their eQTLs, including rs27302. Their results showed that rs27302 remained statistically significant (P=1.76x10^{-4}) when conditioning on the splice SNP, but when conditioning on the novel eQTL, rs2248374 had a nonsignificant P-value (P=0.22) (Gabrielsen et al., 2016b).

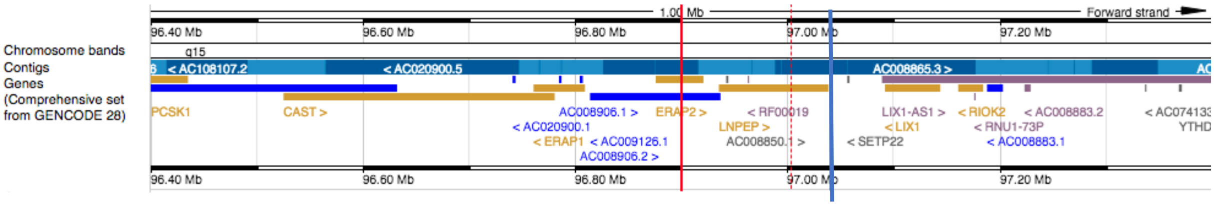

Figure 6. Genetic location of the *ERAP2* associated SNPs rs2248374 and rs27302 (GRCh38.p12). rs2248374 is indicated by the red line and is situated within the *ERAP2* gene. rs27302 is indicated by the blue line and is located in an intergenic region downstream *LNPEP*. (Ensembl, 2018b).

Taken together, *ERAP2* has shown association with several autoimmune diseases, however the functional implications of the risk variant(s) are not yet established. According to theory there are two main *ERAP2* transcripts that are expressed (Andres et al., 2010), referred to as full length *ERAP2* and alternatively spliced *ERAP2-208* throughout this thesis. Genetic studies have discovered several SNPs that are involved in *ERAP2* expression as eQTLs and/or functional SNPs, showing a greater association with *ERAP2* than rs2248374 in several studies (Cheung et al., 2005; Gabrielsen et al., 2016b; Kuiper et al., 2014). Overall, these findings have put an emphasis on acquiring knowledge concerning the regulation of *ERAP2* expression.

2. Aim of study

The hypothesis of this study was that one or more SNP(s), i.e. rs27302, are responsible for the differential expression of *ERAP2* in addition to or instead of rs2248374. The aim of this thesis was therefore to gain a better understanding of expression of the *ERAP2* gene on both RNA and protein level, and investigate the influence of SNPs on gene expression levels. The study mainly focuses on the splice SNP, rs2248374, and rs27302, investigating *ERAP2* in LCLs and human thymic tissue.

The study therefor aimed to:

- Ascertain the possible influence of regulatory factors on *ERAP2* expression by analyzing the expression pattern of the gene in thymic microarray data.
- Investigate the correlation of rs2248374 and rs27302 genotypes and *ERAP2* expression by eQTL analysis.
- Assess ERAP2 protein expression, in samples with varying genotypes in rs2248374 and rs27302, through western blotting.
- Analyze the LD pattern in the genomic region of *ERAP2* and rs27302, in addition to LD between nine SNPs in the *ERAP2* region in a thymic data set and data sets representing six HapMap populations.
- Investigate the expression of *ERAP2* transcripts based on rs2248374 genotypes, by amplifying the exon 10 splice junction of the gene.
- Initiate sequencing of the genomic region of *ERAP2* and rs27302 to identify other SNPs potentially influencing *ERAP2* expression, while simultaneously confirming the genotypes in rs2248374 and rs27302 in the samples.

3. Materials and methods

This thesis is a continuation of previous work performed by researchers at Oslo University Hospital (OUS), where results from an eQTL study on 42 human thymic tissue samples formed the basis for the current in-depth study of *ERAP2* (Gabrielsen et al., 2016b). The data from whole genome expression arrays and genotyping arrays, generated prior to this study, have been included to form an analytical starting point.

3.1 Materials

In 2005, human thymic tissue samples were collected from 42 Norwegian children (<13 years, with 26 individuals being <1 year) undergoing corrective cardiac surgery, with a gender distribution of 22 girls and 20 boys. The Regional Committee for Research Ethics approved the project, and tissue was collected under the written informed consent of the children´s parents, with the donors being made anonymous. The collected tissue was immediately submerged in RNAlater® solution, and subsequent RNA and protein isolation was achieved using TRIzol® reagent (Viken et al., 2007). The remaining tissue samples have been stored in RNAlater® in the OUS laboratory freezer since then.

Experiments executed during this thesis included material from the 42 human thymic tissue samples stored in RNAlater in the OUS laboratory freezer and 18 LCLs from HapMap individuals (Utah, Yoruba, African, Han Chinese, Japanese and Mexican) obtained from the Coriell Cell Repositories. The 18 LCL samples included eight samples used in the study of Andrés et al. (2010) and ten samples with relevant rs27302 and rs2248374 genotypes. Samples and genotypes are listed in the Appendix, Section A.

Additional samples were included in some of the experiments. gDNA from six other thymic samples, in addition to the 42 samples mentioned, isolated by researchers at OUS in year 2014, were used in several DNA experiments. Furthermore, a western blotting experiment included CD8$^+$ blood cells (2x10^6) and protein lysate isolated from thymic tissue using TRIzol in year 2005 (T27-TRIzol). Four whole blood cell samples were utilized in two separate optimization experiments, where one included two samples containing 2.5x10^6 and 1.8x10^6 cells each, while the other included two samples with cell counts of 1.8x10^6. An additional DNA sample were employed as positive control during some of the experiments.

3.1.1 Gene expression data

OUS researchers obtained thymic microarray data in collaboration with the Norwegian Genomics Consortium by use of the Illumina Human WG-6 v3 Gene Expression Beadchip array (Illumina, San Diego, CA, USA) (Amundsen et al., 2014; Gabrielsen et al., 2016a). In the present study, the data was utilized for gene expression profiling of *ERAP2* in thymus. The array contains several gene-specific probes that hybridize with labelled cDNA. Signal strength detected from each probe corresponds to the expression level of the specific transcripts, and thereby gene expression level. The study utilized data on one *ERAP2* probe (ILMN_1743245), three *Glyceraldehyde-3-Phosphate Dehydrogenase* (*GAPDH*) probes (ILMN_1343295, ILMN_1802252, ILMN_2038778) and three *beta actin* probes (ILMN_2152131, ILMN_2152132, ILMN__2152313).

3.1.2 Genotyping data *ERAP2*

Prior to this study the 42 human thymic tissue samples were genotyped for 196,524 polymorphisms, including rs27302 and rs2248374 (Gabrielsen et al., 2016a), using the Illumina Immunochip (Cortes & Brown, 2011). Some of the data obtained with the immunochip was employed in analyses concerning SNPs in the *ERAP2* region in thymus. The analyses extracted data based on the SNPs location, only including SNPs from the genomic region of *ERAP2*, chr5: 96,800,598-97,432,894 (GRCh38/hg38). This yielded a total of 882 SNPs.

3.2 Methods

Several commercial kits, reagents, consumables, instruments, software, primers and solutions were utilized during the experiments of this study. An overview including catalog number and materials was listed in the Appendix page i-ix.

3.2.1 Protein methods

Protein extraction – Optimization

In this study RIPA buffer (Sigma-Aldrich®, Darmstadt, Germany) was primarily chosen as lysis buffer since ERAP2 proteins reside in the ER and RIPA buffer is recommended for extraction of proteins that are membrane bound, nuclear or mitochondrial (Abcam, 2018).

Halt protease and phosphatase inhibitor enzyme (Thermo Fisher Scientific, Waltham, MA, USA) was added to the RIPA buffer in a 1:100 ratio to inhibit degradation of proteins.

Several isolation experiments were performed to test the amount of RIPA-inhibitor mix that yielded the highest amount of protein when isolating from cell pellets. The optimization experiments included using 200µl, 100µl and 60µl of RIPA-inhibitor mix to different tissue and cell samples (Table 1). It was equally necessary to find a volume of RIPA-inhibitor mix that efficiently isolated protein from the thymic tissue. To achieve this, two thymic samples (T27, T30) were included in the first experiment.

Table 1. Protein isolation using different amounts of RIPA-inhibitor mix. An overview of samples used during different scenarios with their approximate cell number.

200µl RIPA-inhibitor mix		**100µl RIPA-inhibitor mix**		**60µl RIPA-inhibitor mix**	
Tissue	**Cell number**	**Tissue**	**Cell number**	**Tissue**	**Cell number**
Thymus	NA	Blood	$2.5x10^6$	Blood x2	$9x10^5$
Thymus	NA	Blood	$1.8x10^6$	Blood x2	$9x10^5$
Blood (CD8$^+$ cells)	$2x10^6$	LCLs (18 samples)	$5x10^5$	LCLs (18 samples)	$5x10^5$

During the first experiment, which included western blotting, four samples were selected; two thymic tissue samples (T30, T27) that had been stored in RNAlater in freezer since 2005, one cell pellet containing approximately $2x10^6$ CD8$^+$ cells and one sample containing protein isolated with TRIzol in year 2005. The last sample was included to examine the state of the proteins isolated with TRIzol to determine if this and similar samples were applicable to western blotting. A small portion of the tissue samples were isolated from the tissue stored in RNAlater and washed in PBS (Thermo Fisher Scientific). The samples were transferred to separate Eppendorf tubes where 200µl RIPA-inhibitor mix was added. The cell pellet was washed with PBS (Thermo Fisher Scientific) and centrifuged before adding 200µl RIPA-inhibitor mix. The two tissue samples were homogenized using the TissueRuptor® (QIAGEN, Hilden, Germany) hand blender while the cell pellet was homogenized by pipetting. The samples were vortexed and kept on ice for 15 min to allow the enzyme reaction to occur. The next step involved centrifugation of the samples at 12,000rpm for 20min before transferring the protein lysate into new Eppendorf tubes.

The same protocol was followed during the remaining experiments. 100µl RIPA-inhibitor mix was added to two cell pellets derived from blood that contained approximately $2.5x10^6$ and $1.8x10^6$ cells. The same amount was used on an aliquot of the 18 LCL cell pellets. Due to the reduced number of cells (appx. $5x10^5$) in these samples in comparison to the blood samples, an investigation of protein yield when using 60µl RIPA-inhibitor mix in the LCL samples was included.

Homogenization may be achieved in several ways, e.g. by using a hand blender, syringe or by pipetting with tips of different sizes. To try and optimize the protein yield, a test using the homogenization methods mentioned was performed on two samples containing $1.8x10^6$ blood cells. These were split into two samples each and washed in PBS (Thermo Fisher Scientific) before 60µl RIPA-inhibitor mix was added. All samples were vortexed for 30 seconds as part of the homogenization step prior to applying the separate techniques; hand blender, 0.8mm syringe, 0.2-20µl pipette tip and 2-100µl pipette tip.

Protein extraction

In addition to the protein lysates from T27 and T30, acquired during the optimization experiments, protein was extracted from seven of the 42 thymic tissue samples from year 2005 (T01, T10, T18, T40, T56, T57, T58), adding to a total of nine thymic protein lysates. The samples were selected based on their genotypes in rs2248374 and rs27302. Isolation was achieved by application of 200µl RIPA-inhibitor mix and using hand blender as homogenization method. Extraction of protein from the 18 LCL cell pellets was performed by adding 60µl RIPA-inhibitor mix to the cell pellets, and the solution was subsequently homogenized by passing the sample through a sterile 0.8mm syringe. The samples were washed with PBS (Thermo Fisher Scientific) as indicated in the protocol used during the optimization experiments.

Total protein

Protein concentrations were measured through application of the Pierce BCA Protein Assay Kit (Thermo Fisher Scientific) in all protein related experiments apart from the very first, where the presence of protein was ascertained by western blotting. The method detects the protein concentration of each sample and allows for normalization of concentrations prior to western blotting. Pierce BCA Protein Assay utilizes five standards (Table 2) in a two-fold dilution series with albumin concentrations of 0.125-2mg/ml to create a standard curve used as reference when calculating the total protein concentration in the samples. A solution containing 100x Halt^{TM} Protease and Phosphatase Inhibitor cocktail (Thermo Fisher

Scientific) and RIPA lysis buffer (Sigma-Aldrich®) was used as dilution agent (1:100). 5µl of the standards and protein lysates were transferred to a non-skirted 96-well plate (Thermo Fisher Scientific), where 200µl of BCA Working Reagent, prepared by mixing Reagent A and Reagent B (50:1), was added to each well. This was followed by mixing the plate for 30sec before incubation at 37°C for 30min. Total protein concentration in the samples was assessed by measuring the absorbance at 570nm using the VersaMax microplate reader (Molecular Devices, San Jose, CA, USA), with concentrations being calculated by SoftMax Pro 6.4 software (Molecular Devices).

Table 2. Pierce BCA Protein Assay standards.

Standard	Concentration (mg/ml)		Diluent
A	2	*Albumin standard solution*	0
B	1	50µl A	50µl
C	0.5	50µl B	50µl
D	0.25	50µl C	50µl
E	0.125	50µl D	50µl
F	Blank	0	50µl

Western blotting

Presence of full length ERAP2 (110kDa) in the thymic samples and LCL cell pellets was ascertained through western blotting. Five blotting procedures were executed with samples selected based on their genotype for rs27302 and rs2248374. The first western blot experiment was performed to investigate the efficiency of protein isolation from cell pellets and thymic tissue with 200µl RIPA-inhibitor mix (CD8$^+$cells, T27, T30) and the state of the TRIzol protein lysate (T27). The second blot contained the nine thymic samples, while the remaining three blots were composed of different genotype combinations of the 18 LCL samples. Sample distribution in the three LCL blots was based on the genotype of rs27302, separating them into A/G heterozygous, G/G homozygous and A/A homozygous blots. All three blots contained one rs27302G/G-rs2248374A/A sample (GM12043) as reference.

Western blotting was initiated by loading the samples onto 10% mini-PROTEAN® TGX^{TM} Precast Gels (Bio-Rad Laboratories, Hercules, CA, USA), with protein concentrations of 3.5µg in the LCL samples and 10µg in the thymic samples. The first blot contained samples with unknown protein concentrations. Presicion Plus Protein^{TM} Dual Color Standards (Bio-Rad) was the ladder of choice during the western blotting experiments. 1xTris/Glycin/SDS buffer (Bio-Rad) was added and the gel electrophoresis were run at 200V for 30-40min. The separated proteins were blotted onto a 0.2µm nitrocellulose membrane (Bio-Rad), and successful protein transfer was determined by incubation with Ponceau S solution (Sigma-Aldrich®). Blocking was performed by incubating the blots in 5% BSA (Sigma-Aldrich®) for 60-120min. ERAP2 protein detection was achieved through antibody probing with mouse polyclonal anti-ERAP2 (ab69037, Abcam®, Cambridge, UK) as primary antibody in a 1:1000 dilution and rabbit polyclonal mouse immunoglobulins conjugated with horseradish peroxidase (HRP) (P0260, Agilent, Santa Clara, CA, USA) in a 1:1000 dilution as secondary antibody. Blots were incubated with the primary antibody overnight at 4°C, followed by a washing step using TBS-T (see Section E, Appendix). Subsequent probing with the secondary antibody was achieved by 60min incubation at room temperature. TBS-T and TBS solution (Bio-Rad) were then applied to wash the blots prior to protein visualization. The blots were developed using ECL^{TM} Prime Western Blotting Detection Reagent kit (GE Healthcare Life Sciences, Pittsburg, PA, USA) and proteins were visualized with ImageQuant LAS 4000 (GE Healthcare). Final images were created with ImageQuant TL 1D v8.1 software (GE Healthcare).

To assess protein loading on all five blots, presence of beta actin in each sample was ascertained. Secondary blocking was achieved by incubating the blots in 5% BSA (Sigma-Aldrich®) for 60 minutes, with subsequent antibody probing with mouse monoclonal beta actin (8H10D10, Cell Signaling Technology, Danvers, MA, USA) as primary antibody and rabbit polyclonal mouse immunoglobulins conjugated with HRP (P0260, Agilent) as secondary antibody. Image development was achieved using the same ImageQuant LAS 4000 instrument and software (GE Healthcare) as earlier. To ensure proper binding of beta actin antibodies to the thymic blot, it was stripped with Restore^{TM} Western Blot Stripping Buffer (Thermo Fisher Scientific), prior to secondary blocking.

ERAP2 protein expression levels were measured in samples expressing ERAP2 using ImageQuant^{TM} TL Toolbox v8.1 software (GE Healthcare) that generated values of average ERAP2 intensity and average beta actin intensity. These values were utilized to create histograms of normalized ERAP2 expression in Microsoft Excel (Microsoft, Redmond, WA, USA).

3.2.2 RNA methods

RNA isolation

An analysis of the exon 10 splice junction of *ERAP2* transcripts in 38 of the 48 thymic samples and 18 LCL cell pellets were performed to determine which transcripts were expressed in the different samples compared to their genotype. The analysis was initiated by RNA isolation. RNA was isolated using the RNeasy Plus Mini Kit (QIAGEN). The procedure was executed according to the manufacturers protocol 'Purification of Total RNA from Animal Cells'. In short, disruption of cells was achieved by adding 350µl Buffer RLT Plus containing ß-mercaptoethanol and mixing the solution by vortexing. It was homogenized by passing the lysate through a 0.8mm sterile syringe, followed by the addition of 350µl ethanol (70%). Three of the samples (GM18923, GM18870, GM19201) were eluted in 30µl RNase free water, while the remaining 15 samples were eluted in 35µl RNase free water. RNA concentrations were quantified using Nanodrop® ND-1000 (Thermo Fisher Scientific).

cDNA synthesis

The Superscript® III First-Strand Synthesis System for RT-PCR (Thermo Fisher Scientific) was used to synthesize complementary DNA from RNA extracted from the 18 LCL cell pellets. The synthesis was performed in accordance with the Invitrogen protocol using random hexamers. In short, RNA was diluted in DPEC-treated water, assuring 448ng RNA input in each reaction. Random hexamers and dNTP mix was added to the RNA solution, before subsequent incubation of the mix at 65°C. cDNA Synthesis mix was added to the samples and the solution was incubated as described in the protocol. Following incubation, the samples were cooled on ice and RNase H was added. The final solution was incubated at 37°C for 20min and subsequently stored at -20°C.

PCR across the exon 10 splice junction

Screening of the exon 10 splice junction was performed using the same primer pair as Andrés et al. (2010) (Table S10) to amplify the exon 10 splice junction of *ERAP2* in cDNA isolated from the 18 LCLs and cDNA from 11 thymic samples, prepared in year 2009. The PCR program utilized was listed in Table 3. PCR products and a 50bp GeneRuler^{TM} DNA ladder (Thermo Fisher Scientific) were loaded onto a 2% agarose gel run at 100V for 60min.

Amplicons were detected with ImageQuant LAS 4000 (GE Healthcare), and images developed with ImageQuant^{TM} TL 1D v8.1 software (GE Healthcare).

Table 3. PCR program employed to amplify the exon 10 splice junction from LCL and thymic cDNA.

Process	Temperature	Incubation time	Cycles
Initial denaturation	94°C	2min	1x
Denaturation	94°C	30sec	35x
Annealing	60°C	30sec	
Extension	72°C	15sec	
Hold	4°C	∞	1x

3.2.3 DNA methods

DNA isolation

DNA was extracted from the 18 LCL cell pellets from Coriell Cell Repositories in preparation of sequencing the genomic region of *ERAP2* (chr5: 96,875,939-96,919,716, GRCh38) and rs27302 (chr5: 97,038,046, GRCh38). Extraction was achieved using the QIAamp® DNA Micro Kit (QIAGEN) by following the user developed protocol 'Purification of genomic DNA from cultured cells using the QIAamp® DNA Micro Kit'. To lyse the cells 100µl Buffer ATL, 10µl proteinase K and 100µl Buffer AL was added to the samples, which were pulse vortexed to create a homogenous solution. The samples were incubated at 56ºC with 600rpm agitation, and 50µl ethanol (96-100%) was subsequently added before transfer of the entire lysate to a QIAamp MinElute® column. The column was washed with Buffer AW1 and Buffer AW2 in two separate steps, precluding elution of DNA in 60µl Buffer AE. DNA concentrations was quantified and purity was measured on Nanodrop® ND-1000 (Thermo Fisher Scientific), where a 260/280 ratio 1.8-2.0 was deemed acceptable.

Primer design

To amplify the genomic region of *ERAP2* and rs27302, ten primer pairs were obtained from Eurofins Genomics (Table S9, Section C, Appendix). Primer design was initiated by copying the genomic sequence of *ERAP2,* including regions flanking the gene, and a region containing rs27302 from genome.ucsc.edu (GRCh38/hg38) (UCSC Genome Browser, 2018b). Primer3 v. 0.4.0 software from bioinfo.ut.ee was utilized to select primers with Tm ~65°C producing products of 4500-6500bp (Untergasser et al., 2012). An *in silico* PCR search was performed on all ten primer pairs to verify absence of products that could interfere with the results

(UCSC Genome Browser, 2018c). Additionally, all primers were tested for self-complementarity using the Oligo Calc: Oligonucleotide Properties Calculator from biotools.nubic.northwestern.edu (Kibbe, 2007).

PCR

Primer specificity

Assessment of primer specificity of the ten primer pairs was performed using a sample of DNA isolated from blood (~25ng/µl) prior to synthesizing the *ERAP2* amplicons from gDNA extracted from the 18 LCL cell pellets and 48 thymic samples. The analysis was performed using GenDx - LongRange PCR kit (GenDx, Utrecht, Netherlands) with ~50ng gDNA as input. The PCR mix was created following the manufacturer protocol, with some modifications. PCR mix composition and volumes were listed in Table 4, while the PCR reaction was listed in Table 5.

PCR products were separated on a 1% agarose gel at 80V for 60min, with subsequent amplicon detection on ImageQuant LAS 4000 (GE Healthcare), using ImageQuant^{TM} TL 1D v8.1 software (GE Healthcare). GeneRuler^{TM} 1kb DNA Ladder (Thermo Fisher Scientific) was the ladder of choice during the gel electrophoresis. Primers yielding the expected amplicon as PCR product were used to generate *ERAP2* amplicons from the 18 LCLs and 48 thymic tissue DNA extracts following the same procedure used during assessment of primer specificity (Table 4-5).

Table 4. PCR mix based on GenDX-LongRange PCR kit (GenDx). The table contains an overview of reagents and volumes used during PCR when testing primers prior to sequencing.

Reagent	**µl/ sample**
Nuclease free H_2O	14.18
LongRange enzyme mix (5U/µl)	0.32
LongRange Buffer (10x)	2
dNTP mix (10mM)	1
gDNA (~50ng input)	2
Forward primer	0.25
Reverse primer	0.25
Total	20

Table 5. PCR reaction testing primer specificity.

Reaction	Temperature	Time	Cycles
Denaturation	95°C	15sec	1x
Denaturation Annealing Elongation	95°C 65°C 68°C	15sec 30sec 6min and 30sec	35x
Termination of elongation	68°C	10min	1x
Cooling	4°C		1x

Optimization

Optimization of PCR conditions may be required to generate the PCR products of interest by increasing specificity of primer pairs. This applied to Amp1, Amp3, Amp7 and Amp10 primers. The effect of temperature and addition of X-solution (GenDx) were the parameters explored. Eight samples were made, two for each primer pair. The first set of amplicons were generated using the PCR mix described in Table 4, while a second set of amplicons were generated from samples containing X-Solution in addition to the other reagents. In this experiment, the PCR reactions were run on a temperature gradient (62°C, 63°C, 65°C, 66°C, 67°C, 68°C) during annealing. PCR products were separated on a 1% agarose gel at 90V for 60-90min along with an aliquot of GeneRuler^{TM} 1kb DNA Ladder (Thermo Fisher Scientific). Amplicons were detected on ImageQuant LAS 4000 (GE Healthcare), using ImageQuant^{TM} TL 1D v8.1 software (GE Healthcare).

A second set of primers were obtained for Amp1, Amp3, Amp7 and Amp10 (Table S11, Section C, Appendix). Primer specificity was assessed as earlier described (Table 4-5), using samples composed of primers in combinations of original (O) and new (N) primer pairs (Table 6). PCR conditions were optimized for two combinations of Amp7 primers, Amp7_OO and Amp7_NO, by exploring the effect of temperature and addition of X-solution (GenDx) to the PCR mix. Four samples of PCR mix were created, two for each combination of Amp7 primers, with one lacking and one containing X-solution (GenDx). The PCR reaction were run on the following temperature gradient; 58°C, 59°C, 60°C, 61°C, 62°C and 63°C. PCR products from both the specificity and the optimization experiments were separated on 1% agarose gels at 85V for 70min and 60-120min, respectively. Detection of products was achieved using ImageQuant LAS 4000 (GE Healthcare) as described earlier.

Table 6. Amp1, Amp3, Amp7 and Amp10 primer combinations.

Primer pair	Direction	Order
Amp1_ON	F	Original
	R	New
Amp3_NN	F	New
	R	New
Amp3_NO	F	New
	R	Original
Amp3_ON	F	Original
	R	New
Amp7_NN	F	New
	R	New
Amp7_NO	F	New
	R	Original
Amp7_ON	F	Original
	R	New
Amp10_NN	F	New
	R	New
Amp10_NO	F	New
	R	Original
Amp10_ON	F	Original
	R	New

O = Original primer, N = New primer, F = Forward primer, R= Reverse primer.

Extended amplicons were attempted generated by combining the forward primer of Amp6 with the original reverse primer of Amp7, and the original forward primer of Amp7 with the reverse primer of Amp8. The PCR mixes were produced following Table 4, and the PCR reaction were run as described in Table 5, but with the upregulation of annealing time to 12min. The reactions were run on the following temperature gradient; 62°C, 63°C, 64°C, 65°C, 66°C and 67°C. PCR products were separated on a 0.7% agarose gel at 80V for 90min, with subsequent detection as earlier described.

Primers fulfilling the criteria for sequencing, by producing an efficient amount of the expected amplicon without excess of unspecific products, were used to generate amplicons of the *ERAP2* and rs27302 region in the 18 LCL cell pellets and 48 thymic samples following the protocol used during the investigation of primer specificity (Table 4-5).

3.3 Statistical analyses

3.3.1 *ERAP2* expression in thymus

Microarray data on the 42 thymus samples was used as input in Microsoft Excel (Microsoft) to graphically illustrate *ERAP2* expression levels based on log2 transformed values of the ILMN_1743145 probe which hybridize with cDNA generated from *ERAP2* transcripts. The plot included maximum and minimum expression values of any gene in the data set to compare *ERAP2* expression relative to the highest and lowest gene expression observed in thymus, while simultaneously giving an overview of *ERAP2* expression differences between the individual thymic samples.

Expression of the housekeeping genes *GAPDH* and *beta actin* in thymus, was used as reference. A box plot was generated using R Studio 1.1.447 (RStudio, Inc., Boston, MA, USA), utilizing data from three *GAPDH* (ILMN_1802252, ILMN_2038778, ILMN_1343295) and three *beta actin* (ILMN_2152131, ILMN_1777296, ILMN_2038777) probes hybridizing with transcripts from these genes.

To determine which *ERAP2* transcripts were represented by the microarray data, an analysis identifying the area of *ERAP2* that hybridized with the ILMN_1743145 probe was performed. The probe sequence (50bp) was inserted into BLAT software that outputs the genomic location of the probe sequence (UCSC Genome Browser, 2018a). Furthermore, redirecting the search to UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly which gave the chromosomal area of hybridization. The presence of common SNPs in the probe was also assessed, investigating the strength of hybridization, by visualizing the ‘common SNPs (150)’, ‘Common SNPs (146)’ and ‘All SNPs(150)’ on UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly.

3.3.2 eQTL analysis

The eQTL analysis utilized *ERAP2* (ILMN_1743145) expression data derived from the microarray analysis and genotype data from the immunochip analysis. The data was used to investigate possible expression differences regarding rs2248374 and rs27302 genotypes, to analyze the correlation between each SNP and *ERAP2* expression. GraphPad Prism 7 software (GraphPad Software, Inc., La Jolla, CA, USA) was used to graphically illustrate the *ERAP2* expression in eQTL dot plots of the two SNPs. The software was utilized to perform a nonparametric one-way ANOVA, Kruskal-Wallis test. The analysis was used to determine the presence or absence of statistically significant differences between the median of the three genotype groups of rs2248374 and rs27302, in two separate analyses.

3.3.3 Linkage disequilibrium analyses

To investigate the correlation between rs27302 and rs2248374, as well as the SNPs correlation to seven other SNPs in the *ERAP2* region, a linkage disequilibrium (LD) analysis was performed. The analysis involved genotype data on the 42 thymic samples from the immunochip analysis and *ERAP2* data on the Caucasian population, 1000 Genomes Project, Phase 3 (Ensembl, 2018c). The data sets contained information on 882 and 843 SNPs, respectively. Two LD plots were made for each data set, one with nine SNPs in the *ERAP2* region (Table 7) and one containing SNPs from the genomic region of the *ERAP2* gene, chromosome5: 96,211,643-96,373,800 (GRCh37). To execute the LD analysis Haploview 4.2 (Broad Institute, Cambridge, MA, USA) was employed. The following criteria were used in the analysis; HW p-value cutoff = 0.001, Min genotype % = 75, Minimum minor allele frequency = 0.001.

Five additional LD plots were generated using the same criteria as in the paragraph above. The objective of this analysis was to investigate the LD pattern in the remaining populations included in the LCL cell pellets; Yoruba, Han Chinese, Mexican, African and Japanese. To ensure that the population data (1000 Genomes Project, Phase 3) could be analyzed in Haploview 4.2 (Broad Institute), missing or triallelic data possibly causing a rejection of data was zeroed out.

Table 7. Overview of the nine SNPs of interest included in the LD analysis. The overview includes information of the SNPs function regarding ERAP2 and their genomic position on chromosome 5.

SNP	**Function**	**Position (GRCh38)**
rs27302	*ERAP2* eQTL	Downstream LNPEP
rs2248374	*ERAP2* eQTL and splice SNP	ERAP2 intron 10
rs2549782	Functional ERAP2 SNP	ERAP2 exon 6
rs27293	Juvenile idiopathic arthritis (JIA)	LNPEP intron 13
rs27290	Juvenile idiopathic arthritis (JIA)	LNPEP intron 12
rs2910686	Ankylosing spondylitis (AS)	ERAP2 intron 17
rs1363907	Irritable bowel disease (IBD)	ERAP2 intron 18

4. Results

4.1 *ERAP2* expression in thymus

ERAP2 expression was assessed in the 42 thymic samples to investigate if similar expression levels were detected in all samples or if any of the samples showed distinct expression patterns, indicating an influence of regulatory factors on gene expression. The expression was analyzed based on microarray data from the only probe that hybridized to *ERAP2* transcripts (ILMN_1743145) in the 42 human thymic tissue samples. Expression of the housekeeping genes *GAPDH* and *beta actin* was used as reference.

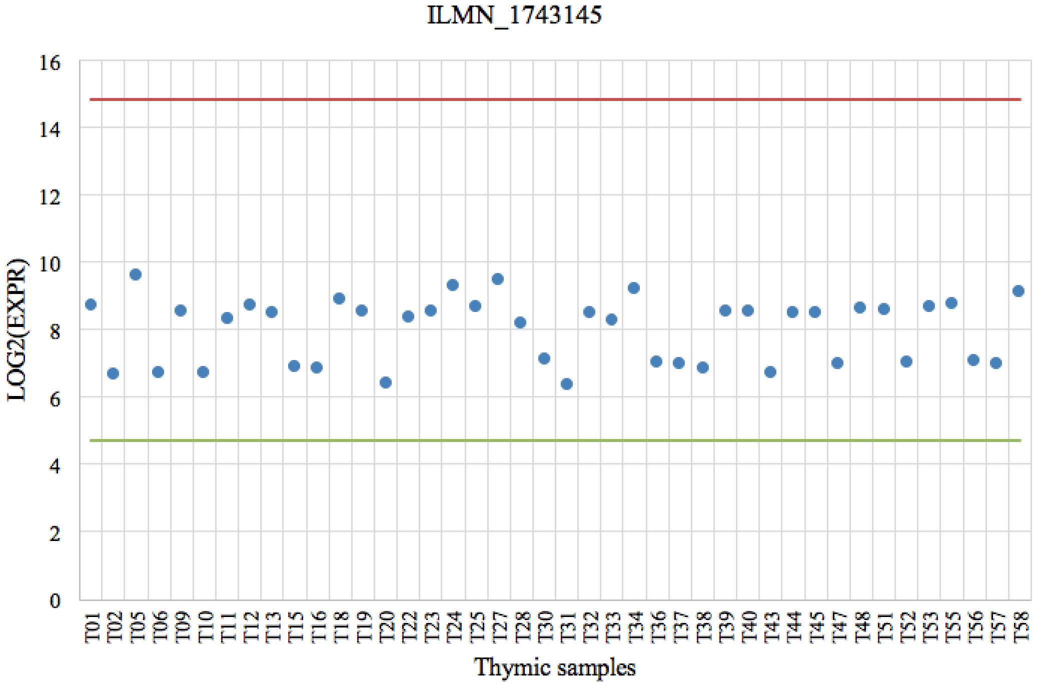

Figure 7. Expression of *ERAP2* in thymus. Log2 transformed expression levels in the microarray data from the 42 thymic samples. The blue dots represent *ERAP2* expression in each thymic sample based on signals in ILMN_1743145. The lines represent maximum (14.84) and minimum (4.69) expression of any gene (independent of samples) detected in the microarray data set, red and green respectively. Log2 transformed expression levels are indicated on the y-axis.

Assessment of *ERAP2* expression in the 42 thymic samples showed that samples segregated into two main groups, one with expression levels between 6.38 and 7.14 and one with levels ranging between 8.17 and 9.63 (Figure 7). The first group showed low or no *ERAP2* expression, while the second group showed clear expression of the gene. The overall expression of *ERAP2* in all samples was slightly skewed towards the lower part of the graph compared to the total data set maximum and minimum, ranging from 6.38-9.63.

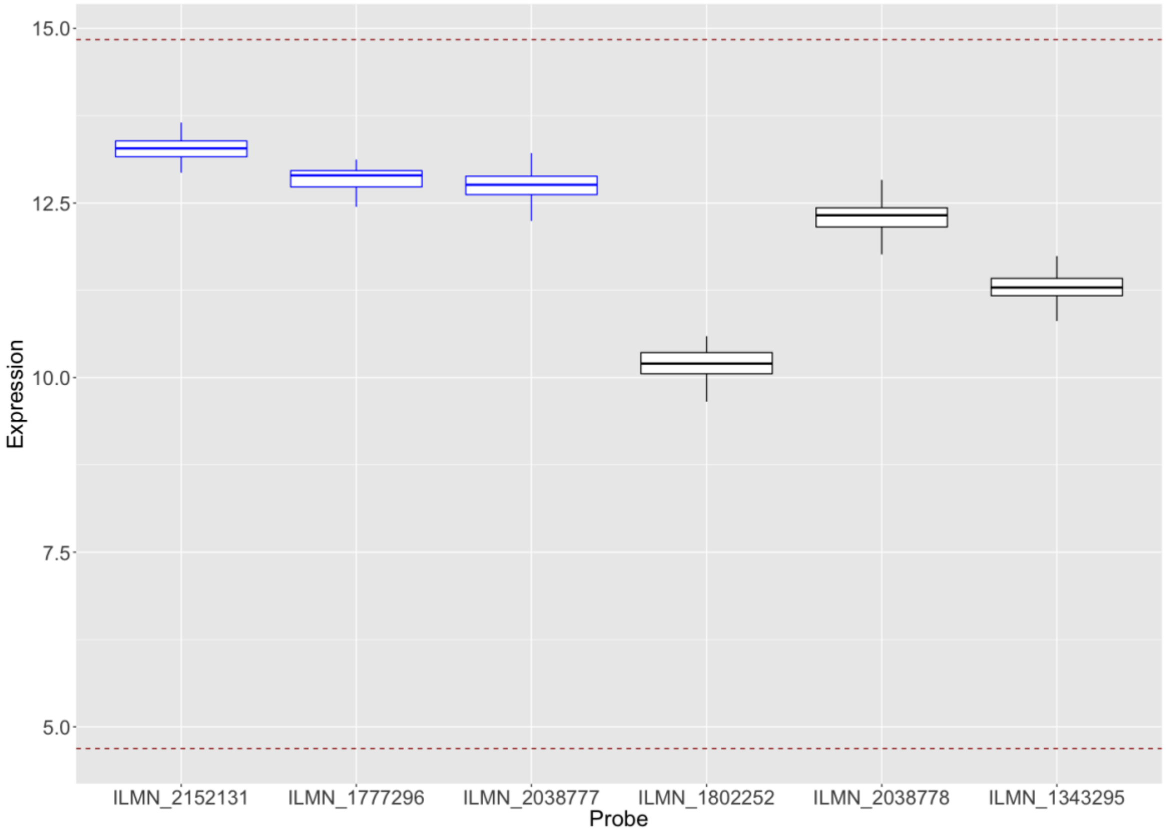

Figure 8. Expression of *GAPDH* and *beta actin* in thymus. Log2 transformed expression levels on the microarray data from the 42 thymic samples. *Beta actin* expression was based on signals in probes ILMN_2152131, ILMN_1777296 and ILMN_2038777, represented by the blue boxes. *GAPDH* expression was based on signals in probes; ILMN_1343295, ILMN_1802252 and ILMN_2038778, represented by the black boxes**.** The red dotted lines indicate the maximum and minimum expression of any gene in the microarray data set. Log2 transformed expression levels are indicated on the y-axis.

Investigation of *GAPDH* and *beta actin* expression, showed mRNA from the two housekeeping genes to be highly expressed in all 42 thymic samples, with similar expression level across samples (Figure 8). The *beta actin* probes showed the highest expression of the probes representing the two genes.

4.1.1 Probe binding analysis

Since there was only one probe (ILMN_173145) present on the microarray to measure *ERAP2* expression, it was important to obtain information on which *ERAP2* transcripts the ILMN_1743145 probe could hybridize to, thereby indicating which transcripts might have been included in the expression analysis. This was achieved through blasting of the probe sequence against the reference genome (GRCh38/hg38). Additionally, it was important to assess the binding stability of the probe by investigating the presence of SNPs in the probe.

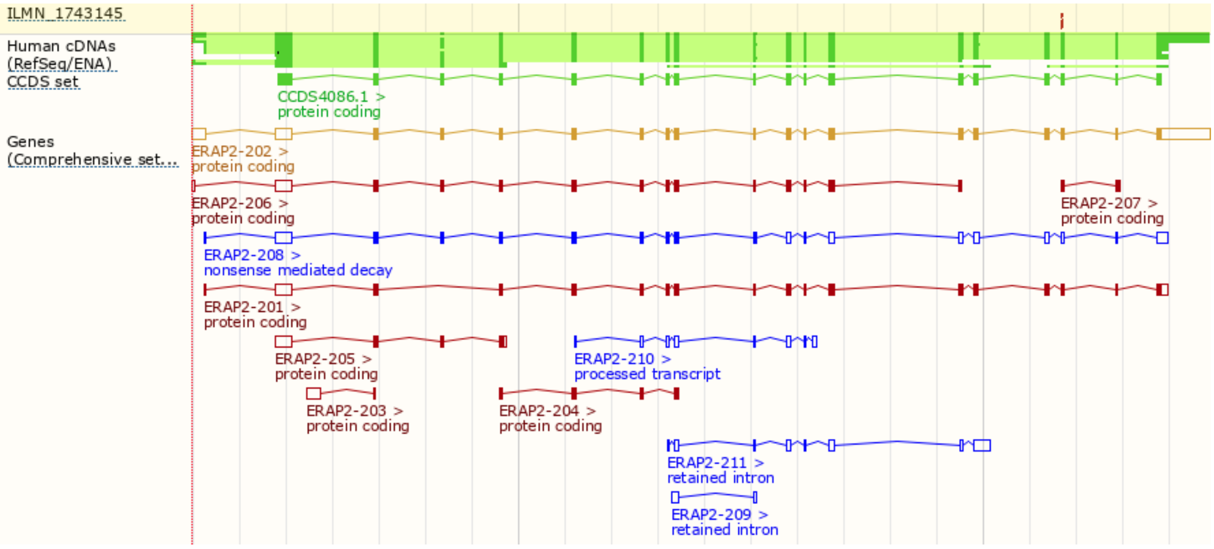

Figure 9. Probe binding analysis regarding *ERAP2* transcripts. ILMN_1743145 binding site in *ERAP2* is indicated by the red line parallel to the probe. The 11 transcript variants of *ERAP2* is illustrated below the RefSeq RNA. (Ensembl, 2018a)

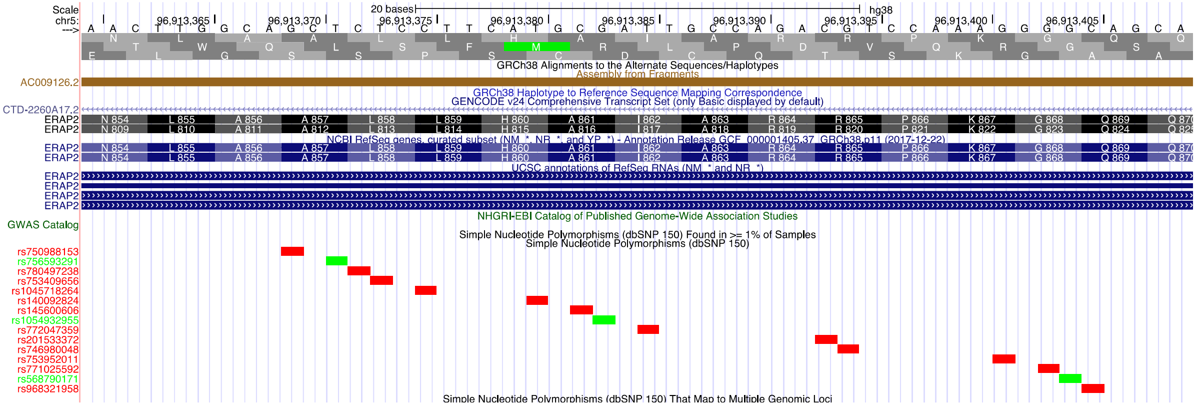
Scale
chr5:
20 bases
hg38
96,913,365
96,913,370
96,913,375
96,913,380
96,913,385
96,913,390
96,913,395
96,913,400
96,913,405
GRCh38 Alignments to the Alternate Sequences/Haplotypes
Assembly from Fragments
AC009126.2
GRCh38 Haplotype to Reference Sequence Mapping Correspondence
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
CTD-2260A17.2
ERAP2
NCBI RefSeq genes, curated subset (NM_*, NR_*, and YP_*) - Annotation Release GCF_000001405.37_GRCh38.p11 (2017-12-22)
UCSC annotations of RefSeq RNAs (NM_* and NR_*)
NHGRI-EBI Catalog of Published Genome-Wide Association Studies
GWAS Catalog
Simple Nucleotide Polymorphisms (dbSNP 150) Found in >= 1% of Samples
Simple Nucleotide Polymorphisms (dbSNP 150)
rs750988153
rs756593291
rs780497238
rs753409656
rs1045718264
rs140092824
rs145600606
rs1054932955
rs772047359
rs201533372
rs746980048
rs753952011
rs771025592
rs568790171
rs968321958
Simple Nucleotide Polymorphisms (dbSNP 150) That Map to Multiple Genomic Loci

Figure 10. SNPs in the genomic region of ILMN_1743145 (50bp). The SNPs indicated by the red and green boxes are found in >1% of samples, resulting in a total of 15 SNPs. The image was obtained from www.genome.ucsc.edu using the GRCh38 genome assembly. None of the selected SNPs investigated in this study were included in the 15 SNP located in the probe.

The results demonstrated that the ILMN_1743145 probe bind to an area of exon 17 in *ERAP2* transcript variant 1, mRNA (Figure 9). There are four out of eleven known *ERAP2* transcripts that contain this exon, and thereby may be included in the expression data as the probe have the ability to hybridize to these. The four transcripts included the alternatively spliced *ERAP2* transcript undergoing NMD (*ERAP2-208*) and the full length *ERAP2* transcript (*ERAP2-202*). Furthermore, presence of 15 SNP sites, found in ≥1% of samples in published reference populations, was detected in the probe (Figure 10).

4.2 eQTL analysis

eQTL analysis was performed to see if the differential expression levels observed in *ERAP2* (Figure 7) correlated to specific genotypes. Two *ERAP2* eQTL analyses were performed; one to assess the correlation between the splice SNP, rs2248374, and *ERAP2* expression, and another assessing the correlation between rs27302 and *ERAP2* expression. The splice SNP, rs2248374, was included based on its acceptance as the eQTL regulating *ERAP2* expression, (Andres et al., 2010), while rs27302 was included as it had recently been identified as a novel *ERAP2* eQTL showing greater correlation with *ERAP2* expression than rs2248374 (Gabrielsen et al., 2016b).

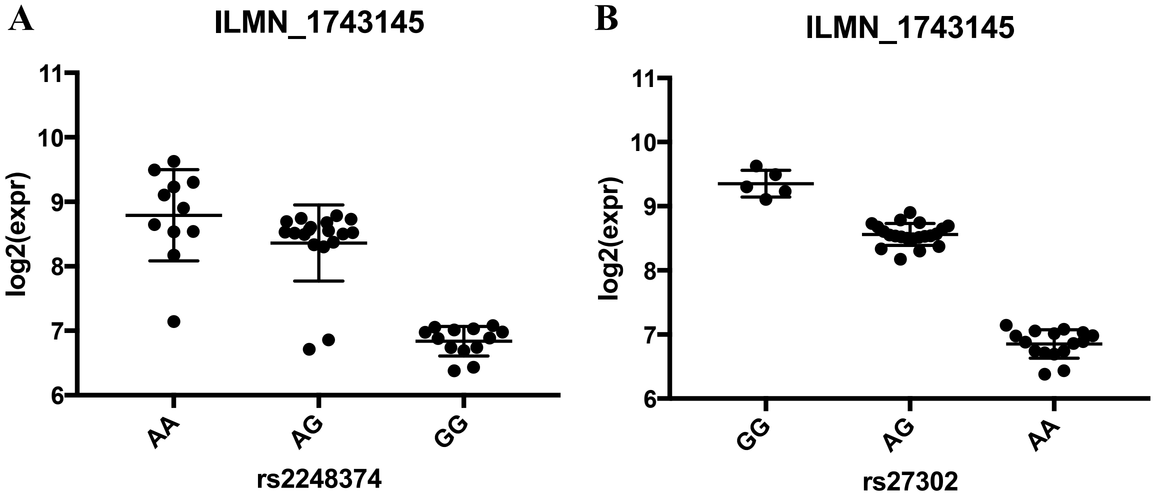

Figure 11. eQTL analysis. Log2 transformed expression data from the *ERAP2* probe, ILMN_1743145, plotted against genotype data for rs2248374 and rs27302 in the 42 thymus samples. Genotypes of rs2248374 and rs27302 are listed on the x-axis, log2 transformed expression values are on the y-axis. Both eQTLs had P-value<0.0001.

The association between rs2248374 genotypes and *ERAP2* expression, and rs27302 genotypes and *ERAP2* expression measured by ILMN_1743145 was statistically significant for both SNPs (P<0.0001). The eQTL analysis of rs2248374 showed that the samples with A/A and A/G genotypes showed higher expression of *ERAP2* compared to thymic samples with G/G genotype (Figure 11A). However, there were three samples not following this pattern i.e. one in the A/A homozygous and two in the A/G heterozygous group expressing low or no *ERAP2* (Table 8). Contributing to this pattern of divergence, several of the A/A homozygous and A/G heterozygous individuals had overlapping *ERAP2* expression levels, and the A/A homozygous individuals showed greater variation of expression levels within the group. In the eQTL analysis of rs27302 *ERAP2* expression followed a distinct and highly clustered pattern, with higher expression in G/G homozygous individuals, intermediate expression in A/G heterozygous individuals and low or no expression in A/A homozygotes (Figure 11B).

Table 8. rs2248374 and rs27302 genotypes of the three outliers in Figure 11A.

Thymic sample	rs2248374	rs27302	Log2 Expression
T30	A/A	A/A	7.15
T10	A/G	A/A	6.72
T16	A/G	A/A	6.86

4.3 Linkage disequilibrium

Since both SNPs showed significant correlation with expression levels of *ERAP2*, it was relevant to measure LD between rs2248374, and rs27302, but also other SNPs in *ERAP2*, in order to point out possible causal candidate SNPs. LD plots were generated to investigate the LD pattern in the genomic region of *ERAP2* and rs27302, chr5:96,857,939-96,919,716 (GRCh37), using two individual datasets, i.e. one for the 42 thymic samples of Norwegian origin (CEU) and one for publicly available Caucasian population data (1000 Genomes Project, Phase 3, GRCh37.p13). High LD was observed between the majority of SNPs across the *ERAP2* region in both data sets (Figure 12-13). However, some of the SNPs were in equilibrium.

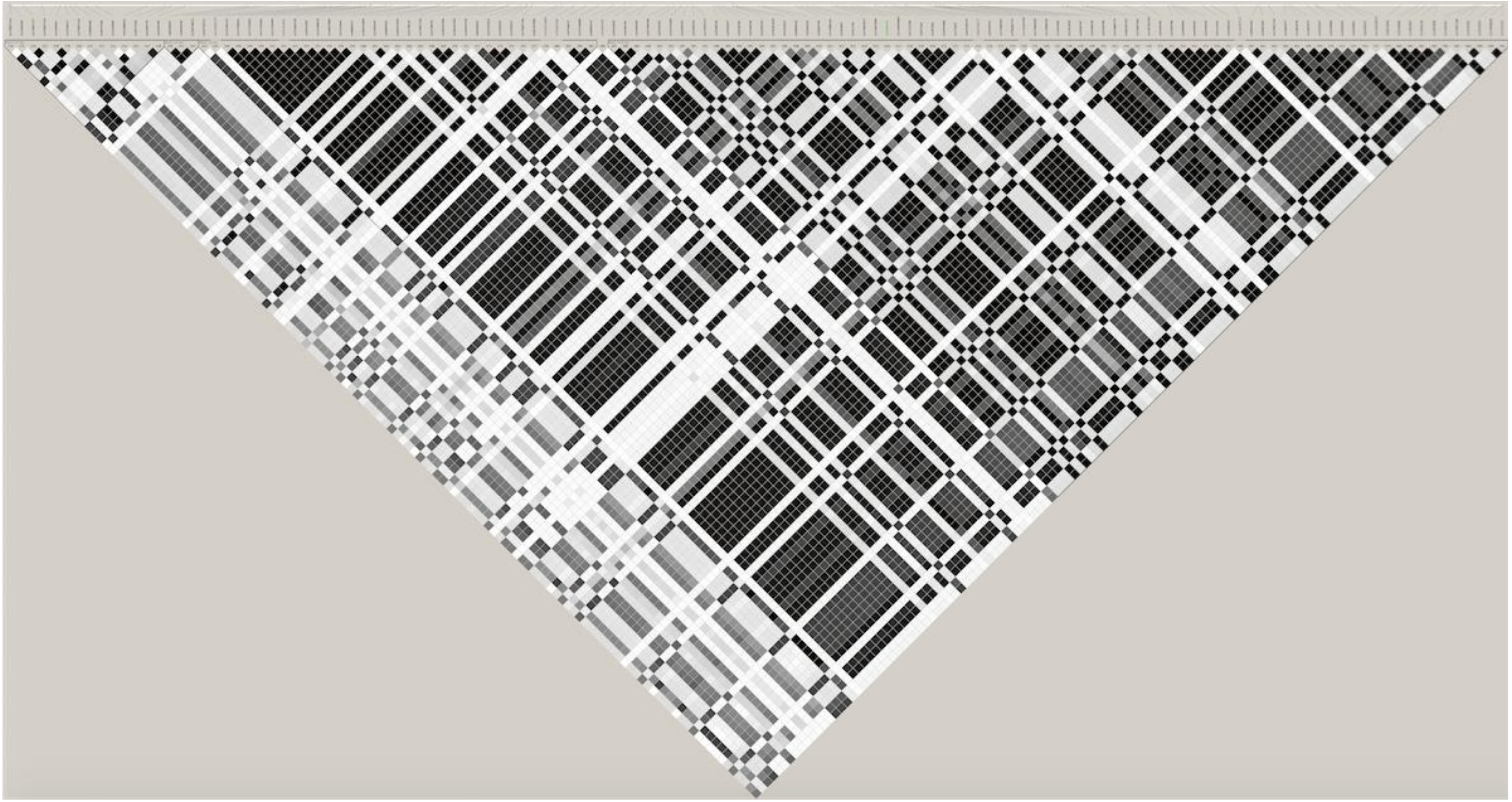

Figure 12. LD in the 42 thymic samples. The plot represents LD in the *ERAP2* gene region (chr5:96,875,939-96,919,716, GRCh37), including rs27302, based on r^2-values. Black squares indicate high LD, white squared indicate no LD (equilibrium).

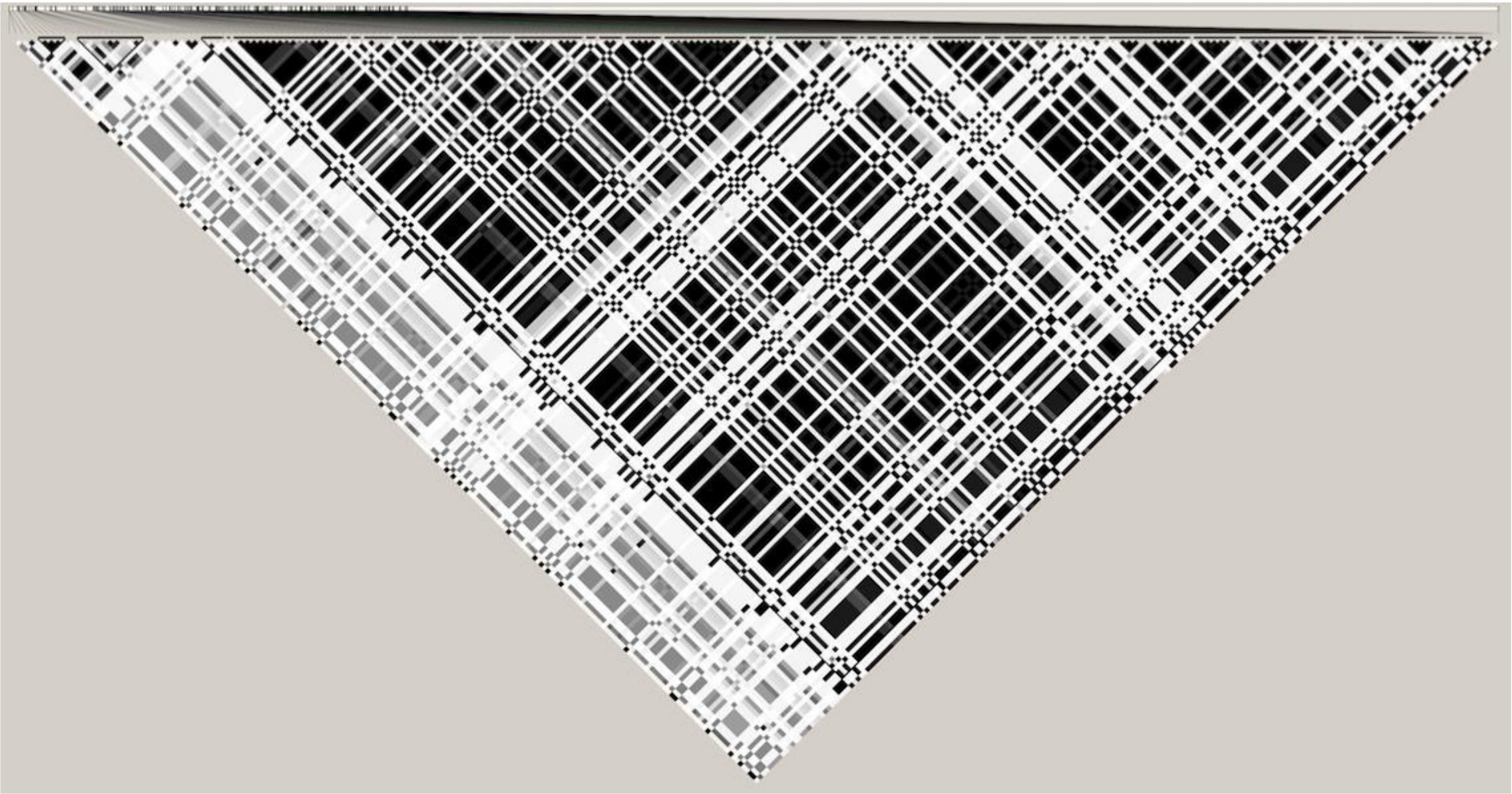

Figure 13. LD in the Caucasian population. LD in the *ERAP2* gene region (chr5:96,875,939-96,919,716, GRCh37), including rs27302, based on r^2-values. Black squares indicate high LD, white squared indicate no LD (equilibrium). The data was obtained from the 1000 Genomes Project, Phase 3, GRCh37.p13

LD analysis was also performed focusing on nine selected SNPs in the *ERAP2* region reported to be of relevance to disease risk or functional aspects, by utilizing the thymic data and CEU data described. The selected SNPs were: rs27302, rs2248374, four AID risk SNPs (rs1363907, 27290, rs27293, rs2910686) two *ERAP2* eQTLs (rs2762, rs10044354) that showed higher significance regarding *ERAP2* expression than rs2248374 in previous studies (Cheung et al., 2005; Kuiper et al., 2014) and one functional *ERAP2* SNP (rs2549782).

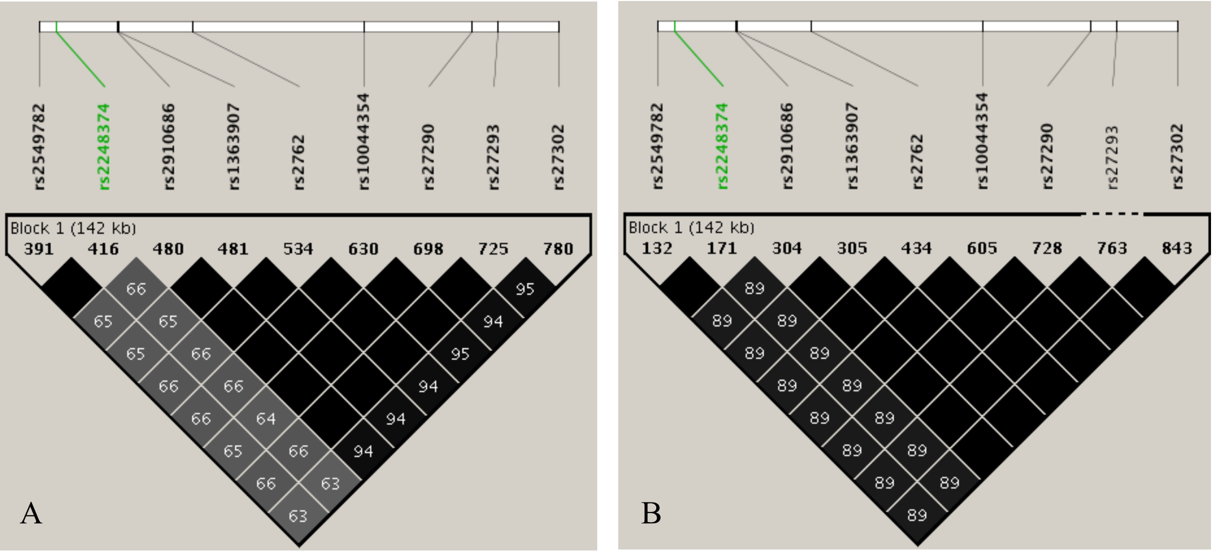

Figure 14. LD between nine SNPs in the *ERAP2* region. A) Thymic immunochip data B) Caucasian population data (1000 Genomes Project, Phase 3, GRCh37.p13). Nine SNPs in the *ERAP2* region were included, four AID risk SNPs (rs1363907, rs27290, rs27293, rs2910686), four *ERAP2* eQTLs (rs10044354, rs2248374, rs27302, rs2762) and one *ERAP2* functional SNP (rs2549782). The splice SNP, rs2248374, was highlighted in green color. LD was measured in r^2.

LD analysis of the nine SNPs using the Caucasian population data showed perfect LD (D'=1, r^2=1) between rs27302 and the AID risk SNPs, rs1363907, rs27290, rs27293 and rs2910686 (Figure 14B). High LD (D'=1 and $r^2 \geq 0.93$) was detected between the same SNPs in the thymic data set (Figure 14A). The analysis also showed that LD measured between rs2248374 and rs2549382, and the remaining seven SNPs was higher in the Caucasian population data (r^2=0.89), compared to the thymic data set (r^2=0.63-0.66). rs2248374 and rs2549782 were in perfect LD (D'=1, r^2=1) in both data sets.

The high LD observed in all plots (Figure 12-14) made it difficult to distinguish SNPs from each other. These findings influenced the evaluation of results obtained in the further experiments where the thymic tissue samples were included. The samples represent the Caucasian populations by originating from Norwegian individuals and the experiments mainly focuses on the effects of rs2248374 and rs27302, which are in high LD with several other SNPs. Hence, the high LD needs to be considered when analyzing the effects of the two SNPs.

Since the experiments of this study included 18 LCL samples representing individuals from five more populations other than the CEU population (Yoruba, Han Chinese, Mexican, African, Japanese), these were included in a separate LD analysis (Figure 15). The objective of the analysis was to evaluate the LD pattern in these populations, and determine if the patterns diverged from each other and from the LD pattern observed in the Caucasian population (Figure 12). Furthermore, LD was measured in these populations so that LD could be considered when analyzing the results from other experiments where the LCL samples were included.

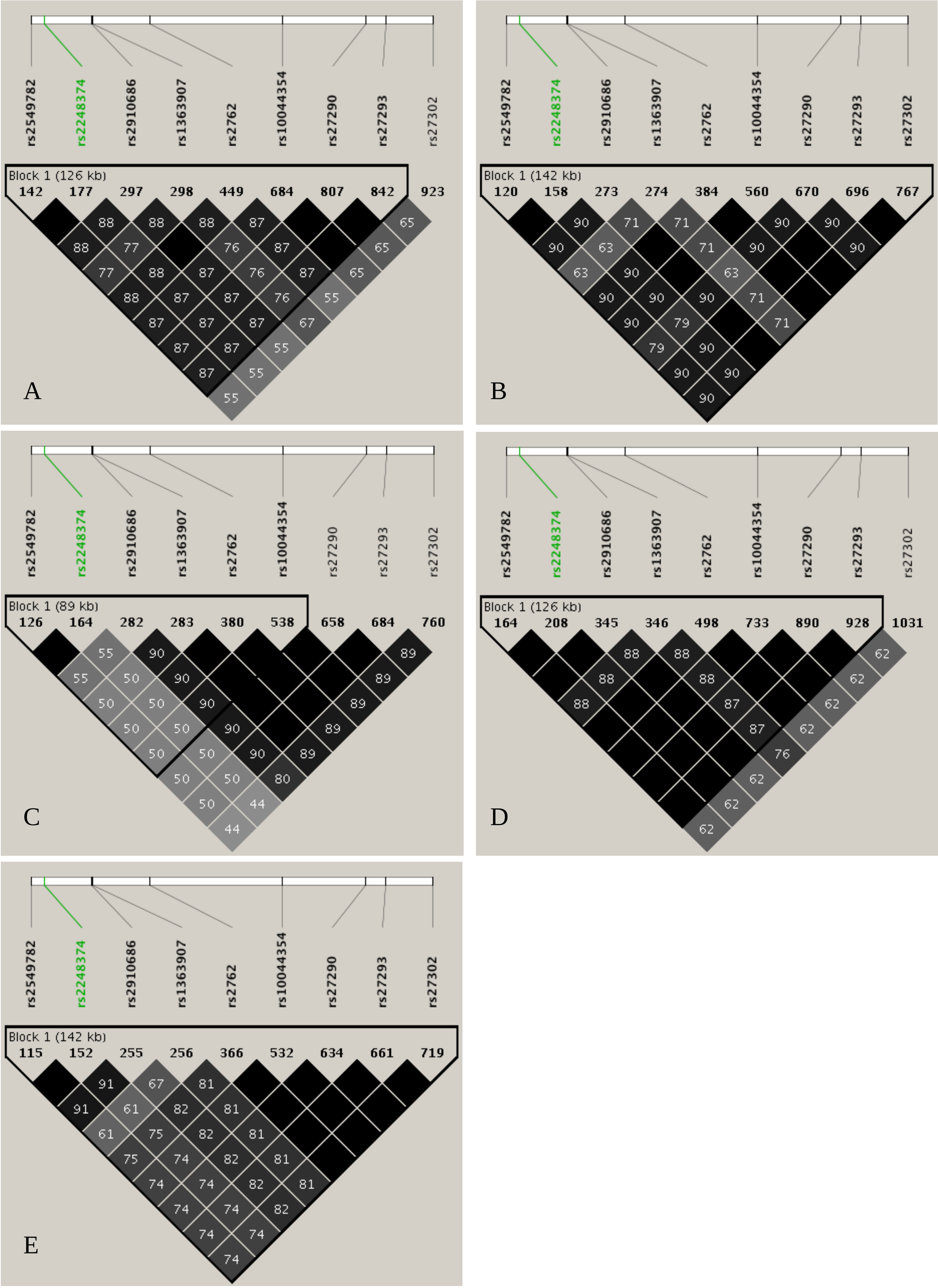

Figure 15. LD observed in populations included in the LCL cell pellets. A: Yoruba, B: Han Chinese, C: Mexican, D: African, E: Japanese. Nine SNPs in the *ERAP2* region were included, four AID risk SNPs (rs1363907, rs27290, rs27293, rs2910686), four *ERAP2* eQTLs (rs10044354, rs2248374, rs27302, rs2762) and one *ERAP2* functional SNP (rs2549782). The splice SNP, rs2248374, was highlighted in green color. LD was measured in r^2.

High LD was observed between the nine SNPs in all populations, although with variation of LD patterns between populations (Figure 15). The five populations showed varying degree of LD when focusing on rs2248374 and rs27302 isolated. Population data from Yoruba, Mexican and African populations showed r^2 values ranging between 0.44-0.62. The Japanese showed slightly higher LD between the two SNPs with r^2=0.74, while the Han Chinese showed r^2=0.90 between the two SNPs, representing high LD. An overall assessment of the plots showed several differences in LD between all SNPs, apart from rs2549782 and rs2248374 which were in perfect LD (D'=1, r^2=1) in all populations. Note that all plots were dominated by high LD, with an absence equilibrium. Comparison of the LD pattern between the nine selected SNPs in the *ERAP2* region in the various populations (1000 Genomes data), showed that overall the highest LD was within the CEU population.

4.4 ERAP2 protein expression in thymus and LCLs

ERAP2 protein expression in thymic and LCL samples stratified for rs27302 and rs2248374 genotypes were assessed through western blotting. The objective was to observe if ERAP2 protein expression followed a specific pattern with regards to the two SNPs. Presence of full length ERAP2 was ascertained through antibody probing with anti-ERAP2. To determine the amount of full length ERAP2 in each sample, the blots were probed with antibody against beta actin, with subsequent calculations of ERAP2 expression relative to beta actin expression. Prior to these investigations, protein isolation was optimized.

4.4.1 Protein isolation - optimization

Optimization experiments were performed to determine the amount of RIPA-lysis buffer with inhibitor mix necessary for efficient protein isolation from cell pellets and thymic tissue samples.

The initial optimization experiment included protein lysate from two thymic samples (T27, T30), one $CD4^+$ cell pellet and T27 protein lysate isolated with TRIzol in year 2005. Protein was isolated from the first three samples using 200µl RIPA-inhibitor mix. Beta actin was detected in two of the four samples, T27 and T30, meaning that protein had been successfully isolated in these (Figure 16). Furthermore, the results showed that application of 200µl RIPA-inhibitor mix was sufficient for protein isolation from thymic tissue. Based on the western blot results, no protein was extracted from the CD4$^+$ cells using 200µl RIPA-inhibitor mix nor detected in the T27-TRIzol protein lysate from year 2005. In addition to the observations of beta actin in the two thymic samples, ERAP2 was detected in T27, while being absent in T30.

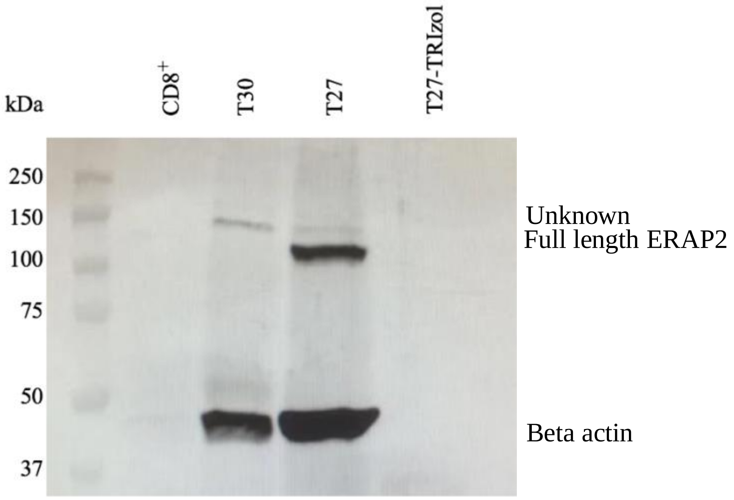

Figure 16. Western Blot of RIPA efficiency. Western Blot of selected samples testing the efficiency of RIPA lysis buffer when isolating protein from thymic tissue and cell pellets. In addition to two thymic samples (T27, T30) and one cell pellet (CD8$^+$) the test included one sample (T27-TRIzol) where protein was extracted from thymus using TRIzol in year 2005. Band at approximately 110kDa represented full length ERAP2 protein, while band at approximately 42kDa represented beta actin.

Subsequent optimization experiments were performed on various cell pellets, with varying homogenization and lysis technique. The assays found optimal protein isolation when applying 60µl RIPA-inhibitor mix to the pellets followed by use of a 0.8mm sterile syringe to homogenize the solution. Total protein concentrations were measured by Pierce BCA Protein Assay prior to western blotting (Table 12-14, Section D, Appendix).

4.4.2 Thymic samples

Presence of full length ERAP2 protein was investigated in nine selected thymic tissue samples, with different genotype combination for rs27302 and rs2248374, through western blotting. The objective was to observe if full length ERAP2 expression followed a specific pattern with regards to the two SNPs. The western blot showed presence of full length ERAP2 protein in five of the nine samples, T27, T58, T40, T18 and T01 (Figure 17). Beta actin protein, used as loading control, was observed in all samples, although in varying amounts, with lowest expression detected in T30, indicating variable loading of protein.

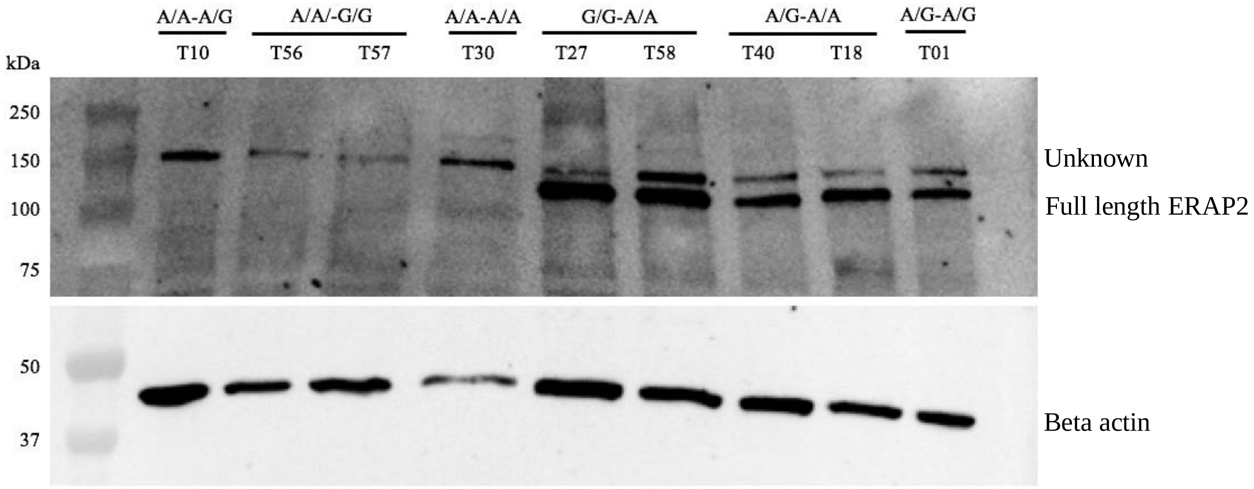

Figure 17. Western blot analysis of nine selected thymic samples. Genotypes are listed as rs27302-rs2248374. Investigation of full length ERAP2 and beta actin expression in nine selected thymic samples with relevant rs27302-rs2248374 genotypes. Protein detection was performed using primary antibodies specific to full length ERAP2 (mouse polyclonal) and beta actin (mouse monoclonal).

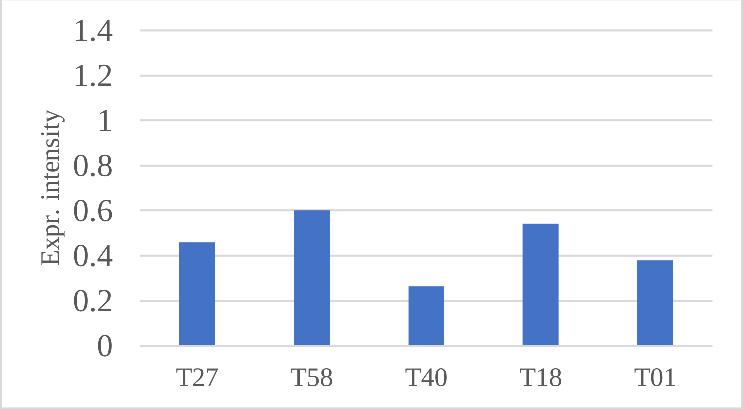

Figure 18. Normalized full length ERAP2 protein expression in thymic samples. Average full length ERAP2 expression intensity measured in the thymic samples expressing full length ERAP2; T27, T58, T40, T18 and T01. Normalized full length ERAP2 expression was based on calculations comparing average full length ERAP2 expression levels and average beta actin expression in each sample.

To normalize full length ERAP2 expression according to protein loading, calculations of full length ERAP2 expression relative to beta actin presence was performed on samples expressing full length ERAP2 (Figure 18). This showed that T58 expressed the largest amount of full length ERAP2 protein. Following T58 in declining order was T18, T27, T01 and T40 having lower full length ERAP2 expression. The samples expressing full length ERAP2 had genotypes G/G or G/A for rs27302 and A/A or A/G for rs2248374. Full length ERAP2 protein was absent in the samples A/A homozygous for rs27302 (T10, T56, T57, T30) regardless of the rs2248374 genotype (Figure 17).

The western blot experiment showed that individuals with a rs27302 G-allele, either being homozygous or heterozygous, expressed full length ERAP2, while A/A homozygotes lacked expression of the protein. No specific expression pattern was observed regarding rs2248374 genotypes. Additionally, a band of unknown origin was detected around 150kDa in all samples.

4.4.3 LCL samples

To ascertain the presence of full length ERAP2 protein in the LCL cell pellets with different rs27302-rs2248374 genotype, three western blot experiments were performed. The distribution of samples in the three blots was based on rs27302 genotypes, generating separate blots for the A/G, G/G and A/A genotypes. All blots contained one rs27302G/G-rs2248374A/A sample (GM12043) that was used as reference.

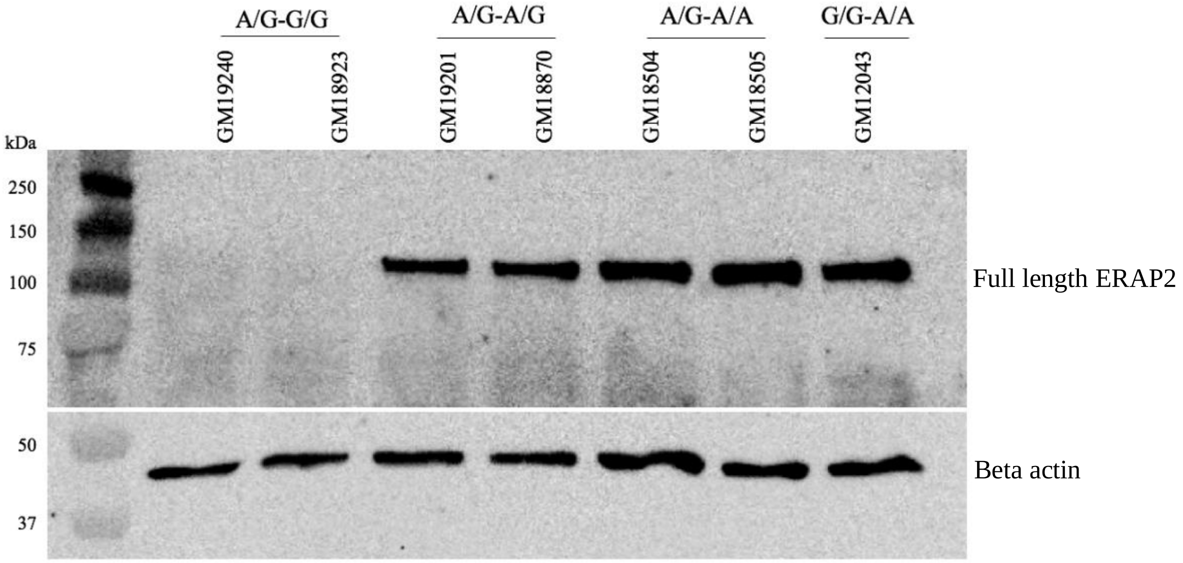

Figure 19. Western blot analysis of LCL samples A/G heterozygous for rs27302. Genotypes are listed as rs27302-rs2248374. Investigation of full length ERAP2 and beta actin expression in LCL samples with rs27302 A/G genotype; GM18504, GM18505, GM18870, GM18923, GM19201 and GM19240. Protein detection was performed using primary antibodies specific to full length ERAP2 (mouse polyclonal) and beta actin (mouse monoclonal). Sample GM12043 (G/G-A/A) was included as reference.

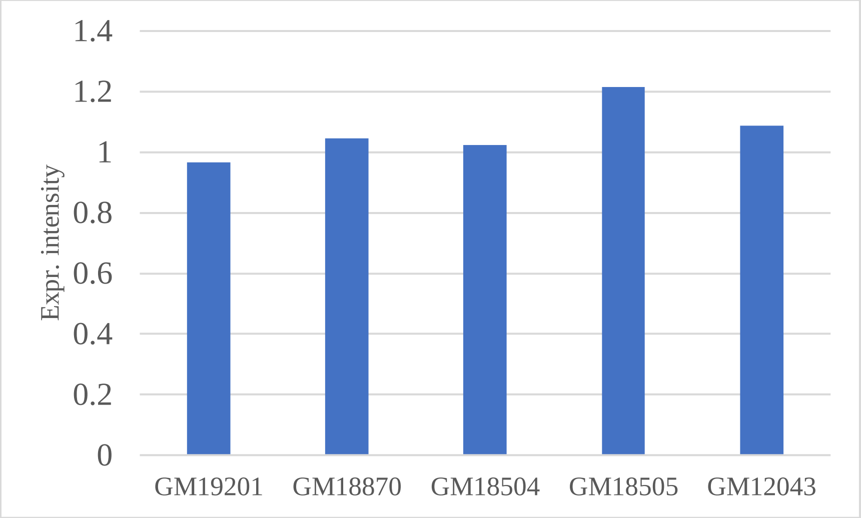

Figure 20. Normalized full length ERAP2 expression in LCL samples A/G heterozygous for rs27302. Average full length ERAP2 expression intensity measured in LCL samples with rs27302A/G genotype expressing full length ERAP2; GM19201, GM18870, GM18504 and GM18505. GM12043 (G/G-A/A) was included as reference. Normalized full length ERAP2 expression was based on calculations comparing average full length ERAP2 expression levels to average beta actin expression in each sample.

The beta actin band observed in all samples A/G heterozygous for rs27302 illustrated successful protein loading. Full length ERAP2 protein was observed in the A/G-A/G, A/G-A/A and G/G-A/A samples, but was absent in the A/G-G/G samples (Figure 19). According to intensity calculations, the five samples expressing full length ERAP2 showed similar amounts of the protein, with normalized values ranging between 1.22 (GM18505) and 0.97 (GM19201) (Figure 20).

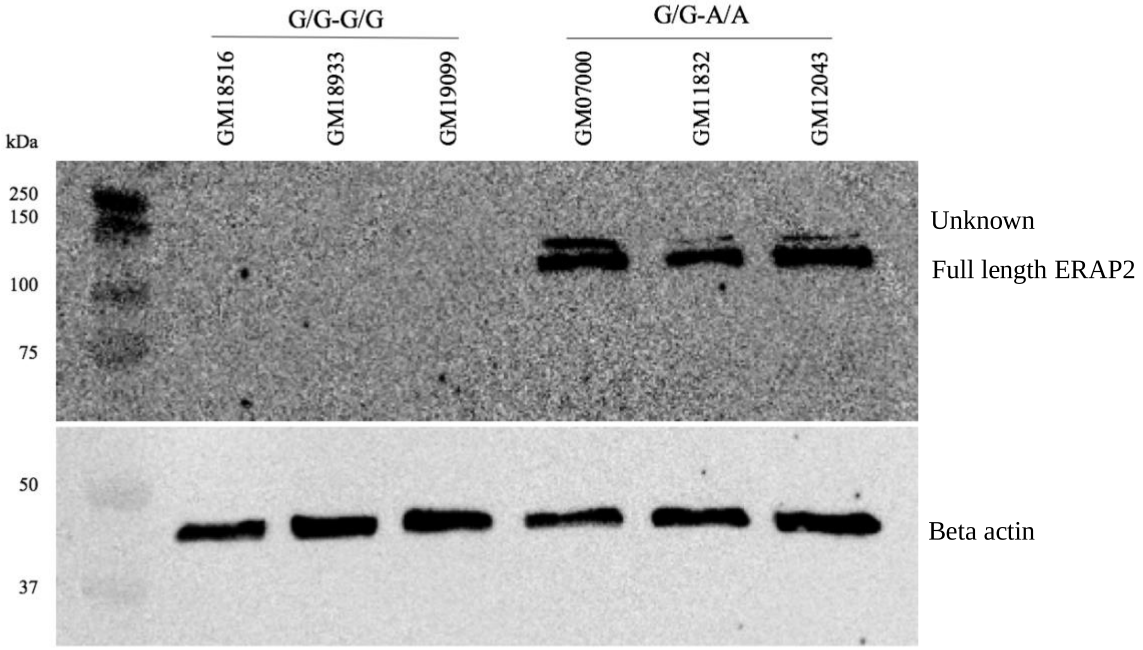

Figure 21. Western blot analysis of LCL samples G/G homozygous for rs27302. Genotypes are listed as rs27302-rs2248374. Investigation of full length ERAP2 and beta actin expression in LCL samples with G/G-G/G or G/G-A/A genotype; GM07000, GM11832, GM12043, GM18516, GM18933 and GM19099. Protein detection was performed using primary antibodies specific to full length ERAP2 (mouse polyclonal) and beta actin (mouse monoclonal).

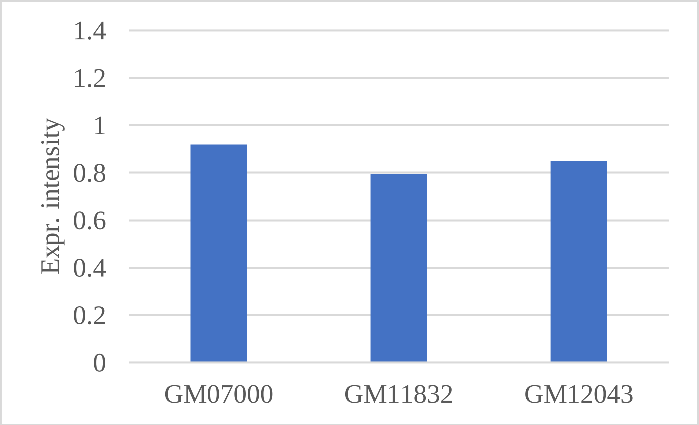

Figure 22. Normalized full length ERAP2 expression in G/G-A/A LCL samples. Average full length ERAP2 expression intensity measured in LCL samples with G/G-A/A genotype expressing ERAP2; GM07000, 11832 and GM12043. Normalized full length ERAP2 expression was based on calculations comparing average full length ERAP2 expression and average beta actin expression in each sample.

All samples G/G homozygous for rs27302 expressed beta actin. The G/G-A/A samples showed expression of full length ERAP2, while it was absent in the G/G-G/G samples (Figure 21). Little expression level variety was observed between the G/G-A/A samples, with GM07000 (0.92) and GM11832 (0.79) expressing the highest and lowest amount of full length ERAP2, respectively (Figure 22).

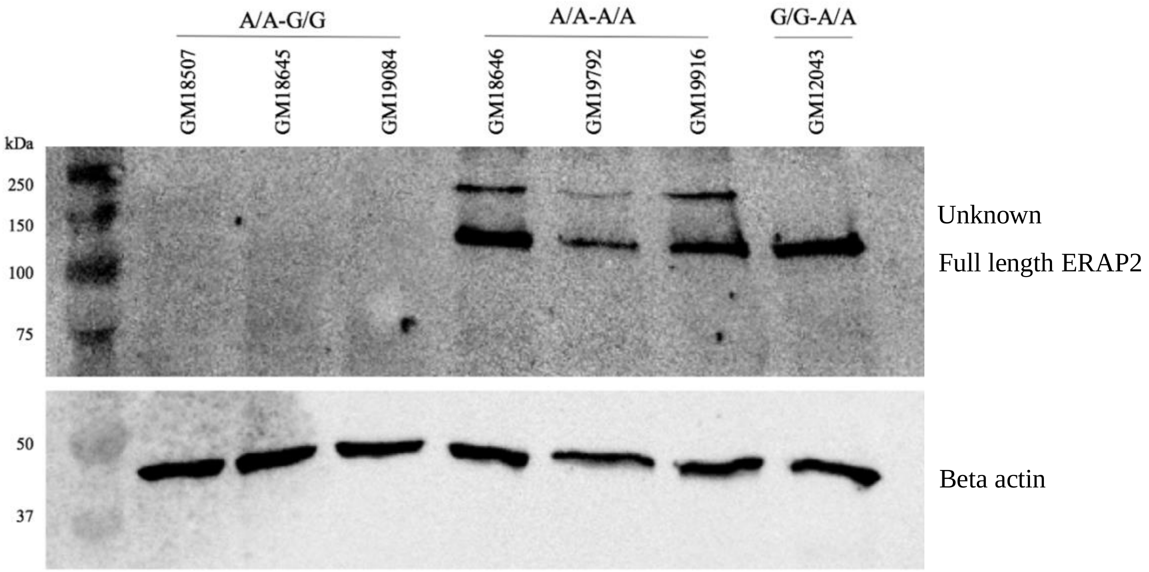

Figure 23. Western blot analysis of LCL samples A/A homozygous for rs27302. Genotypes are listed as rs27302-rs2248374. Investigation of full length ERAP2 and beta actin expression in LCL samples with A/A-G/G or A/A-A/A genotype; GM18507, GM18645, GM18646, GM19084, GM19792 and GM19916. Protein detection was performed using primary antibodies specific to full length ERAP2 (mouse polyclonal) and beta actin (mouse monoclonal). Sample GM12043 (G/G-A/A) was included as a reference.

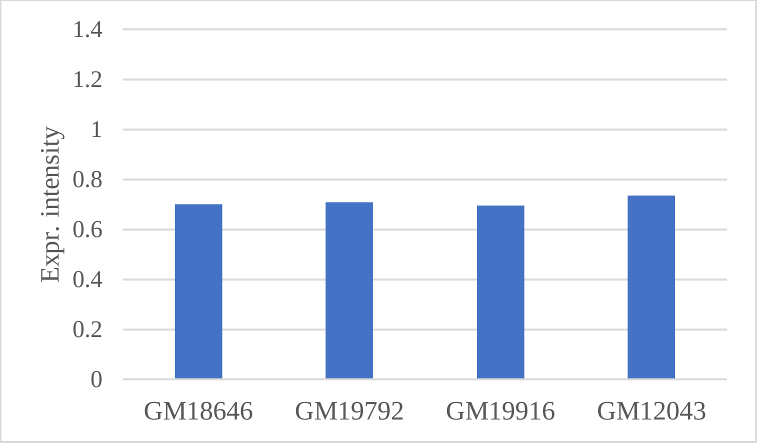

Figure 24. Normalized full length ERAP2 expression in LCL samples A/A homozygous for rs27302. Average full length ERAP2 expression intensity measured in LCL samples with rs27302A/A genotype expressing full length ERAP2; GM18646, GM19792 and GM19916. GM12043 (G/G-A/A) was included as reference. Normalized full length ERAP2 expression was based on calculations comparing average ERAP2 expression and average beta actin expression in each sample.

Successful protein loading was detected in all samples A/A homozygous for rs27302, based on the observed beta actin bands. Presence of full length ERAP2 was detected in four of the seven samples, with three being A/A homozygous for both SNPs and the forth having G/G-A/A genotype (Figure 23). The A/A-AA samples expressed similar amounts of full length ERAP2, with normalized intensity values ranging between 0.70 (GM19916) and 0.71 (GM19792) (Figure 24).

Overall, these western blot results showed that full length ERAP2 was expressed by individuals A/A homozygous or A/G heterozygous for rs2248374, but not by the G/G homozygotes. Furthermore, samples of all rs27302 genotypes showed varying expression of full length ERAP2, with all genotypes being related to both expression and absence of the protein. Unknown bands were detected around 150kDa in several samples on the G/G and A/A blots, while the band was absent from the A/G blot.

4.5 Screening of the exon 10 splice junction

The study sought to investigate the correlation between the rs2248374 and rs27302 genotypes and presence of transcripts from the alternatively spliced *ERAP2-208* variant suggested to undergo NMD and/or presence of regular exon 10 spliced *ERAP2* transcripts, potentially including full length *ERAP2*. The analysis was performed by PCR amplification of the exon 10 splice junction in total RNA isolated from the 18 LCLs and 11 thymic samples.

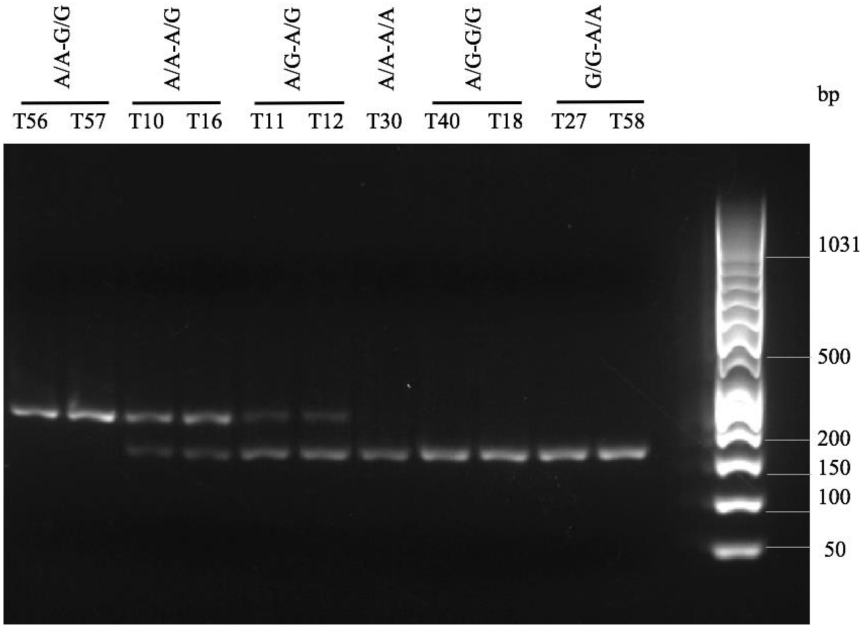

Figure 25. Exon 10 splice junction transcripts in thymus. Transcripts were isolated and amplified from thymic samples; T56, T57, T10, T16, T11, T12, T30, T40, T18, T27 and T58. Genotypes are listed as rs27302-rs2248374. Amplification was achieved using primers specific for amplifying the *ERAP2* exon 10 splice junction. Alternatively spliced *ERAP-208* transcripts were expected to produce products of 192bp, while regular exon 10 spliced *ERAP2* transcripts were expected to produce 136bp products. GeneRuler^{TM} 50bp was the ladder of choice during this experiment.

Amplification products from alternatively spliced *ERAP2-208* transcripts (expected band size 192bp) were detected in six of the eleven thymic samples (Figure 25). All six samples represented individuals carrying at least one rs2248374 G-allele, with four individuals being A/G heterozygous and two being G/G homozygous. Additionally, the A/G heterozygous group expressed regular exon 10 spliced *ERAP2* transcripts. The six individuals were either A/A homozygous or A/G heterozygous for rs27302. Regular exon 10 spliced *ERAP2* transcripts were also expressed by individuals A/A homozygous for rs2248374, with A/A, A/G, or G/G genotype for rs27302.

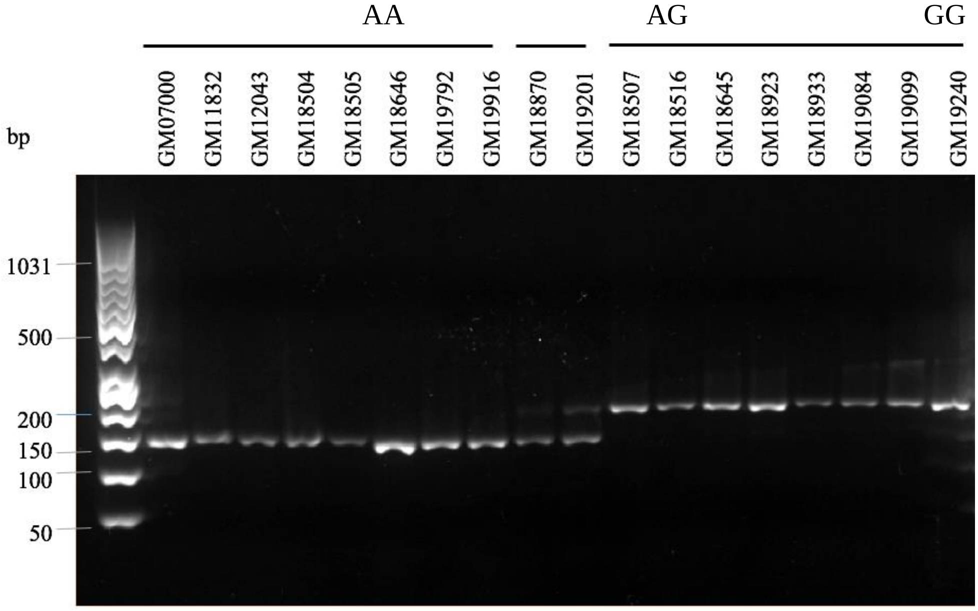

Figure 26. Exon 10 splice junction transcripts in the LCL samples. Transcripts were isolated and amplified from the 18 LCL samples. Genotypes are listed for rs2248374. Amplification was achieved using primers specific for the *ERAP2* exon 10 splice junction. Alternatively spliced *ERAP-208* transcripts were expected to produce products of 192bp, while regular exon 10 spliced *ERAP2* transcripts were expected to produce 136bp products. GeneRuler^{TM} 50bp was the ladder of choice during this experiment.

The amplification pattern detected in the thymic samples, regarding rs2248374 and rs27302 genotypes was observed in the 18 LCLs as well (Figure 26). Focusing on rs2248374 genotypes, eight A/A homozygous individuals expressed regular exon 10 spliced *ERAP2* transcripts, two A/G heterozygotes expressed both transcripts, and eight G/G homozygotes expressed alternatively spliced *ERAP-208* transcripts. The individuals in the rs2248374 A/A- and G/G homozygous groups had varying genotypes for rs27302, resulting in the groups representing all possible combinations of rs2248374 homozygosity and rs27302 genotypes. The two individuals with rs2248374 A/G genotype (GM18870, GM19201) were both A/G heterozygous for rs27302.

4.6 Optimization of amplification prior to sequencing

4.6.1 Primer specificity

Sequencing of the genomic region of *ERAP2* and rs27302 was initiated to investigate the presence of SNPs influencing *ERAP2* expression, and to confirm the genotype in rs2248374 and rs27302 in the 18 LCL samples and 48 thymic samples. Prior to sequencing of the genomic region of *ERAP2* and rs27302, it was necessary to investigate the specificity of the ten primer pairs designed, expectant to generate amplicons covering the genomic regions mentioned (Figure 27, Table 9).

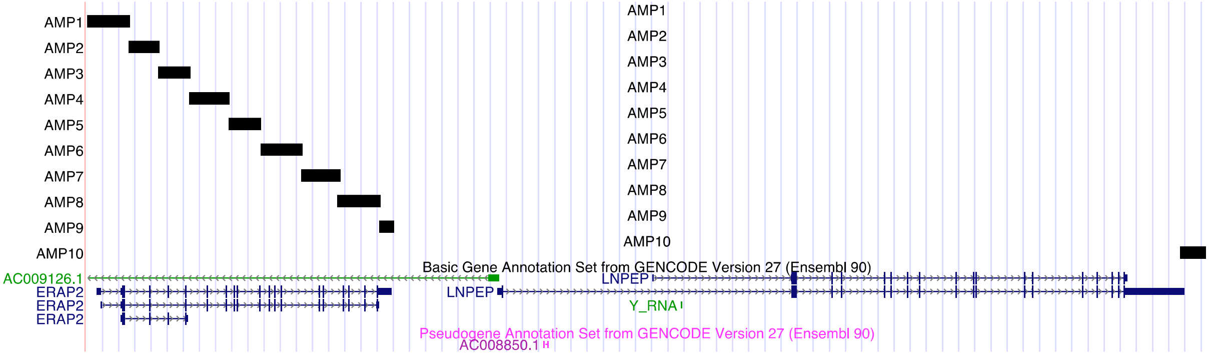

Figure 27. Genomic region covered by the ten amplicons. Primers designed were expected to generate amplicons covering the genomic region of *ERAP2* and rs27302, as illustrated by the black squares representing one amplicon each. The image was obtained from www.genome.ucsc.edu.

Table 9. Expected lengths of the ten amplicons covering the genomic region of *ERAP2* and rs27302.

Amplicon	Expected length	Products observed
Amp1	6465bp	Expected + unspecific band
Amp2	4529bp	Single expected band
Amp3	4899bp	Expected + unspecific band
Amp4	6010bp	Single expected band
Amp5	4917bp	Single expected band
Amp6	6108bp	Single expected band
Amp7	5906bp	Two unspecific bands
Amp8	6356bp	Single expected band
Amp9	2252bp	Single expected band
Amp10	3815bp	Single expected band (very weak)

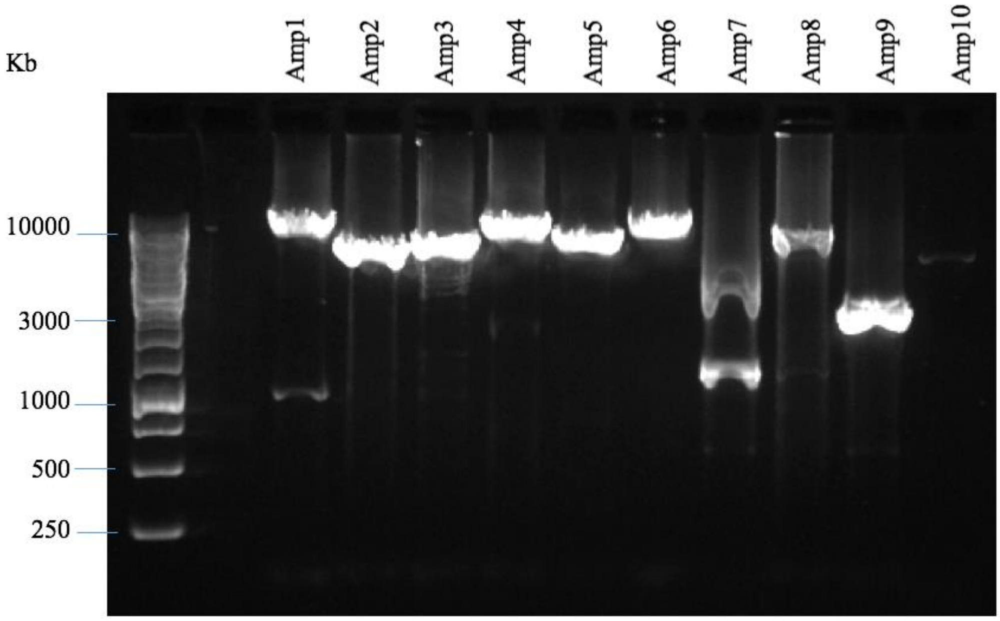

Figure 28. Specificity of the ten primer pairs. PCR products were separated on a 1% agarose gel to investigate the specificity of the individual primer pairs covering the genomic region of ERAP2 and rs27302. GeneRuler^{TM} 1kb DNA ladder was the ladder of choice during the gel electrophoresis.

The specificity varied between primer pairs (Figure 28). Successful amplification was observed for Amp2, Amp4, Amp5, Amp6, Amp8 and Amp9. Although the amplicons were perceived as longer than expected when comparing them to the 1kb ladder, the majority were of the expected size relative to one another. Four of the reactions yielded unsuccessful results; two of the PCR reactions generated the expected product but showed presence of additional amplicons as well (Amp1 and Amp3), the amplification of Amp7 resulted in several bands instead of the expected 5906bp amplicon, and lastly Amp 10 gave only a weak band albeit of the expected size (3815bp). The inconsistency observed between the DNA ladder and amplicons, led to band sizes being determined by comparing the PCR products relative to one another. This approach to determining product sizes applied to the remaining amplification experiments as well.

4.6.2 PCR optimization

The results from testing the primer specificity of the ten amplicons (Figure 28), showed that four primer pairs yielded undesirable amplicons for Amp1, Amp3, Amp7 and Amp10. Based on these results several optimization experiments were performed in an attempt to generate the four proper amplicons. Since a commercial long range PCR kit was used in the amplifications, temperature was adjusted and the addition of X-solution (GenDx) was explored in an effort to optimize the PCR conditions. There were no positive effects of neither adding X-solution (GenDx) to the PCR mix nor by adjusting temperature (Figure 29). Note that all reactions of Amp3 showed an absence of amplicon.

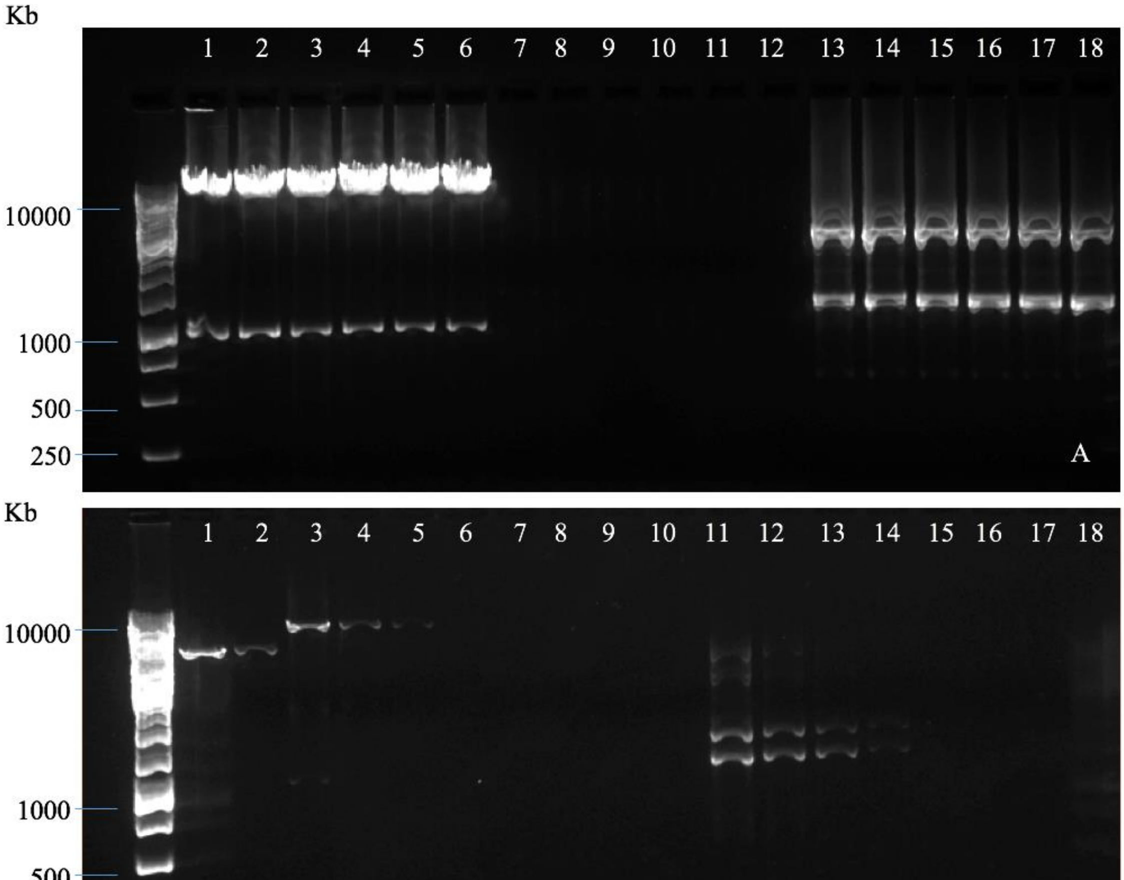

Figure 29. PCR optimization of Amp1, Amp3, Amp7 and Amp10 amplification reactions. A) Optimization with annealing temperature gradient (62°C, 63°C, 65°C, 66°C, 67°C, 68°C) for Amp1, Amp3 and Amp7 reactions. Lanes 1-6 represent amplification using Amp1 primers, 7-12 represent amplification using Amp3 primers while Lanes 13-18 represent amplification using Amp7 primers. B) Optimization of the PCR reaction by addition of X-solution in the amplification reactions of Amp1, Amp3 and Amp7 lanes 3-18, while simultaneously being run on a temperature gradient. Lanes 1-2 contained Amp10 reactions utilizing different temperatures. An overview of A) lanes 1-18 and B) lanes 1-18 was listed in table 10.

Table 10. Lane distribution in Figure 29.

Lane in Figure 29A	Amplicon	Temperature	Lane in Figure 29B	Amplicon	Temperature
1	Amp1	62°C	**1**	Amp10	65°C
2	Amp1	63°C	**2**	Amp10	66°C
3	Amp1	65°C	**3**	Amp1_X-solution	62°C
4	Amp1	66°C	**4**	Amp1_X-solution	65°C
5	Amp1	67°C	**5**	Amp1_X-solution	66°C
6	Amp1	68°C	**6**	Amp1_X-solution	68°C
7	Amp3	62°C	**7**	Amp3_X-solution	62°C
8	Amp3	63°C	**8**	Amp3_X-solution	65°C
9	Amp3	65°C	**9**	Amp3_X-solution	66°C
10	Amp3	66°C	**10**	Amp3_X-solution	68°C
11	Amp3	67°C	**11**	Amp7_X-solution	62°C
12	Amp3	68°C	**12**	Amp7_X-solution	65°C
13	Amp7	62°C	**13**	Amp7_X-solution	66°C
14	Amp7	63°C	**14**	Amp7_X-solution	68°C
15	Amp7	65°C	**15**	Amp10_X-solution	62°C
16	Amp7	66°C	**16**	Amp10_X-solution	65°C
17	Amp7	67°C	**17**	Amp10_X-solution	66°C
18	Amp7	68°C	**18**	Amp10_X-solution	68°C

The unsuccessful optimization experiments (Figure 29) led to the design of new primers targeting the same genomic regions as the original primer pairs for Amp1, Amp3, Amp7 and Amp10. Their specificity was assessed following the experimental design of the initial specificity experiment (Figure 28).

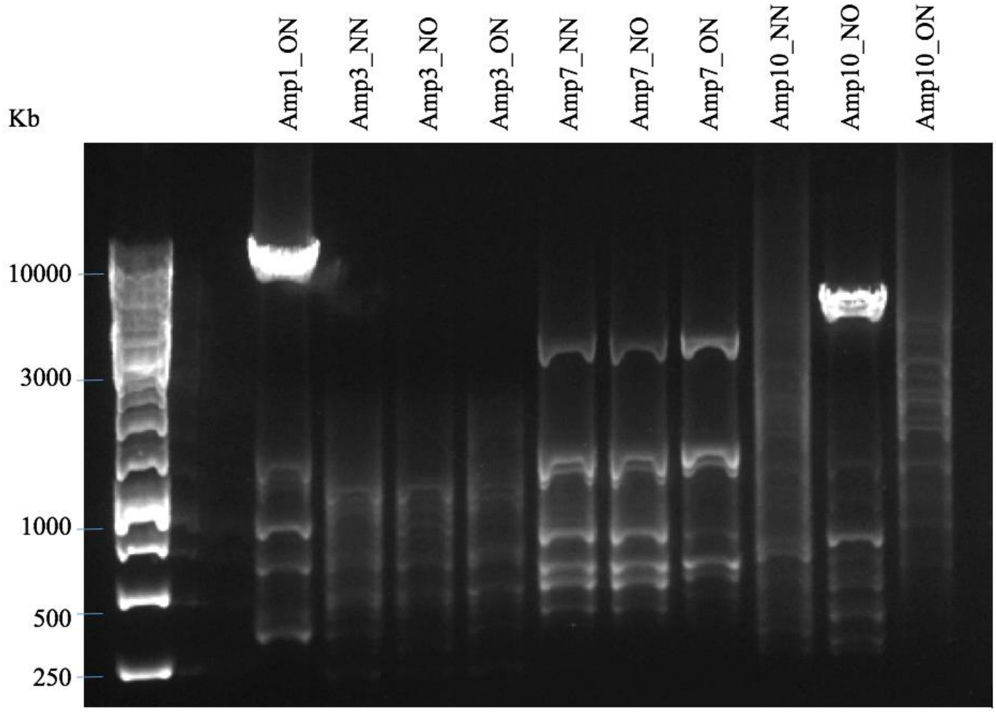

Figure 30. Specificity of combinations of original and new ERAP2 primers. Three combinations of primers expected to cover Amp3, Amp7 and Amp10 were tested; New-New (NN), New-Original (NO) and Original-New (ON). One new primer was obtained for Amp1, and specificity of the combination Original-New (ON) Amp1 primers was tested.

Unspecific bands were present in all amplification reactions utilizing combinations of the original and new primers for Amp1, Amp3, Amp7 and Amp10 (Figure 30). Expected amplicons were absent in Amp10_NN, Amp10_ON and all Amp3 and Amp 7 reactions. Amp1_ON and Amp10_NO showed bands of the expected sizes relative to one another, 6491bp and 4108bp, respectively.

During the experiment a discovery was made regarding similarity between the original and new primers. Investigation of primer similarity found both new Amp3 primers to be identical to the original Amp3 primers. Additionally, the new Amp7 reverse primer was identical to the original Amp7 reverse primer. Based on this discovery and results from testing the new primer pairs, primer picking was possible for Amp1, Amp3 and Amp10. Fewer unspecific bands were observed in the initial reactions using the original Amp1 and Amp3 primers (Figure 28), and as a result these were selected as amplification primers in the PCR reactions prior to sequencing. Amp10_NO was the only Amp10 primer combination yielding a proper amplicon amount of the expected size, based on size relative to Amp1_ON (Figure 30) and was thereby selected for application in the downstream PCR reaction. None of the Amp7 combinations generated the expected amplicons, 5906bp (Amp7_ON) and 6155bp

(Amp7_NO and Amp7_NN), which resulted in the execution of a second optimization experiment. Since Amp7_ON and Amp7_OO, and Amp7_NO and Amp7_NN were comprised of identical primers, the optimization experiment consisted of Amp7_OO and Amp7_NO primer combinations. The primers were tested on a temperature gradient, with and without X-solution (GenDx).

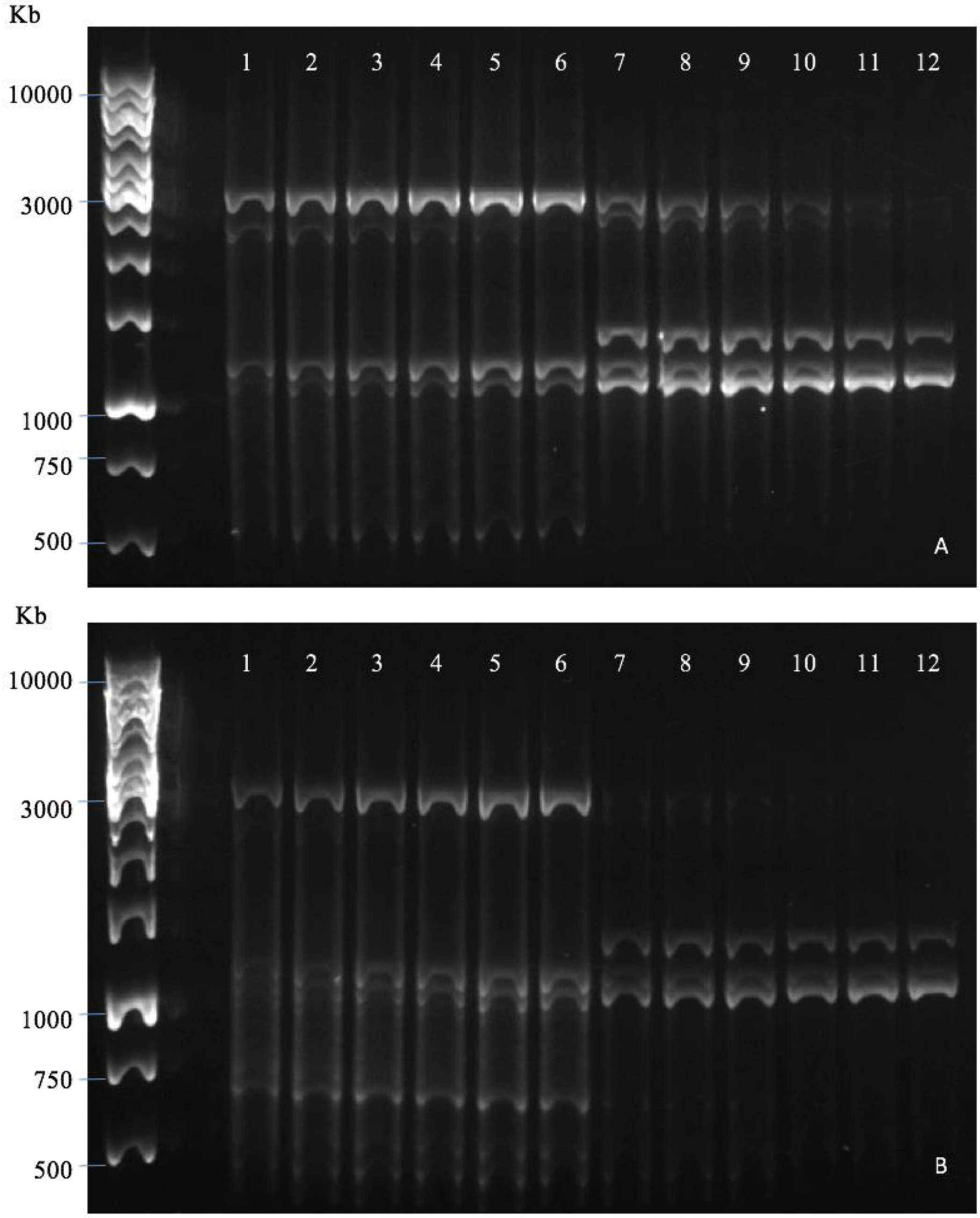

Figure 31. Optimization of PCR conditions for Amp7_OO and Amp7_NO. A) PCR products using the Amp7_OO primer combination, B) PCR products using the Amp7_NO primer combination. Lanes, 1-6 and 7-12, in A and B were amplified on a temperature gradient (58-63°C). Lanes 7-12 had X-solution added to the PCR mix.

The optimization of Amp7_OO and Amp7_NO conditions yielded unsuccessful results (Figure 31). All conditions resulted in unspecific bands, and none generated amplicons of the expected size.

Attempting to overcome the unsuccessful amplification of Amp7 and still include the region in a downstream sequencing analysis, combinations of Amp6F and Amp7R, and Amp7F and Amp8R were tested, albeit these amplicons would be very large.

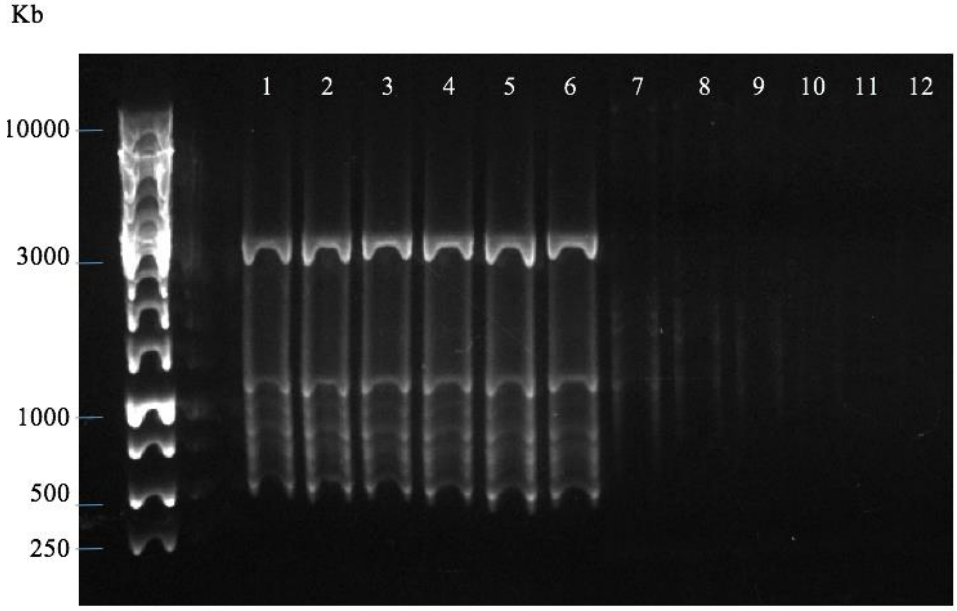

Figure 32. Extended amplicons, Amp6-7 and Amp7-8, on a temperature gradient. Gel electrophoresis results of the attempt to generate extended amplicons using combinations of the original primers Amp6F and Amp7R, and Amp7F and Amp8R. The PCR reactions were run on a temperature gradient (62°C-67°C). Lanes 1-6 represent Amp6-7 amplification products, while lanes 7-12 represent Amp7-8 amplification products.

The extended amplicons of Amp6-7 (expected size 11826bp) and Amp7-8 (expected size 11863bp) were absent in all amplification reactions (Figure 32). The unsuccessful generation of Amp7 led to an incomplete set of amplicons covering the genomic region of *ERAP2* and rs27302, hence inhibiting sequencing of these genomic regions during the time span of thesis.

5. Discussion

5.1 Variation in *ERAP2* RNA expression levels

5.1.1 *ERAP2* expression levels vary between the 42 thymic samples in the microarray data

Differing levels of gene expression in samples of the same tissue type indicate that there are underlying regulatory factors influencing the expression (Lewin, 2006). The *ERAP2* expression analysis utilizing microarray data on 42 thymic samples showed separation of samples into two main groups based on *ERAP2* expression levels. One had low or no *ERAP2* expression, while the other group showed clear expression of *ERAP2* (Figure 7, page 33*)*. As a result of differing *ERAP2* expression levels between the thymic samples, it was important to compare these results to the expression of the housekeeping genes *GAPDH* and *beta actin.* This was done to determine if the variation observed in *ERAP2* was a result of technical variations or of an underlying biological mechanism. According to the microarray data the thymic samples expressed a similar amount of the same housekeeping gene, although expression levels differed between the two genes (Figure 8). *GAPDH* and *beta actin* expression is predicted to be constant within the same tissue type (Eisenberg & Levanon, 2013), and microarray data of the two genes from BioGPS show that expression intensities of both genes are high in thymus (GeneCards, NA-a; GeneCards, NA-c). The microarray observations in this study coincided with this strengthening the idea that the differential *ERAP2* expression was not a result of technical variations. The log2 transformed expression levels ranged from 4.69-14.85 in the total microarray data, indicating the level of minimum and maximum expression of genes in the data set. These observations indicated that genes with expression levels in the lower end had low or no expression, while genes in the upper region had substantially higher levels of gene expression. Furthermore, expression levels in between the two was expected to represent intermediate expression. Hence, the statements concerning *ERAP2*, *GAPDH* and *beta actin* expression was based on this definition of gene expression levels. Gene expression can be regulated at several check points in the cell (Nature, 2014). Since the analysis showed differential expression of *ERAP2* into two distinct groups, it was suggested that a biological factor(s), was regulating the expression. Furthermore, genetic variants with genotypes correlating with *ERAP2* expression have been reported (Andres et al., 2010; Gabrielsen et al., 2016b; Kuiper et al., 2014), indicating that the differing levels of *ERAP2* were caused by a genetic factor(s).

Probe binding analysis (Figure 9, page 35) showed that only four of eleven *ERAP2* transcripts, including alternatively spliced *ERAP2-208* and full length *ERAP2*, contained the target sequence of the *ERAP2* probe (ILMN_1743145). However, it is unknown which *ERAP2* transcripts are expressed in thymus. Hence, it is uncertain if the microarray data encompasses total *ERAP2* expression in the 42 thymic samples.

It is important to notice the presence of 15 SNPs, found in >1% of samples in published reference populations, in the probe sequence, which may have affected the hybridization of the probe to transcripts, and in turn this may have decreased the measured expression levels. Microarray technology depend on the precise interaction between probe and target during hybridization to obtain the accurate expression measurements. SNPs located in the probe sequence may affect hybridization and hence alter signal intensity measurements (Benovoy et al., 2008).

Gene expression analysis is an intricate process, relying on the availability of data and the properties of the methods used to generate and analyze the data (Liu et al., 2010; Tomiuk & Hofmann, 2001). The Illumina Human WG-6 v3 Gene expression Beadchip array utilized to generate the microarray data in this study gave limited information concerning *ERAP2* expression by only containing one *ERAP2* probe (ILMN_1743145), hybridizing with four out of eleven transcripts. The utilization of another gene expression array might have increased the coverage. However, the properties of other arrays will not be discussed further as the generating of the microarray data was not a part of this thesis. Employment of quantitative PCRs could potentially have contributed to determine if samples with *ERAP2* expression levels in the lower end of the expression range of the microarray data had low expression or if expression was actually absent. This could have been achieved using primers amplifying all *ERAP2* transcripts. Moreover, it would have increased the coverage of *ERAP2* expression by including all transcripts in all samples in the analysis.

5.1.2 Strong correlation between *ERAP2* expression and rs27302 genotypes in the 42 thymic samples

The distinct pattern observed in the rs27302 eQTL plot (Figure 11b, page 36), indicated that *ERAP2* expression is perfectly correlated with the different genotypes of the SNP in the thymic data set. Individuals A/A homozygous for rs27302 showed low or no *ERAP2* expression, while the rs27302G/G homozygotes showed clear expression of *ERAP2,* with A/G heterozygotes also expressing *ERAP2* but in reduced amounts compared to the G/G homozygotes. Consequently, it was natural to assume that A/A homozygous individuals would have low or no *ERAP2* expression, while individuals carrying the rs27302 G-allele would express *ERAP2,* with levels depending on homo- and heterozygosity of the SNP.

The eQTL analysis of the splice SNP, rs2247374, illustrated that *ERAP2* expression was also clustered based on genotypes but with some deviating samples, i.e. outliers and the fact that several A/A homozygotes expression levels overlap with the expression levels observed in the A/G heterozygous group. These deviations supported the hypothesis implying that rs2248374 is not solely responsible for regulation of *ERAP2* expression, and that there are other factors involved. Additionally, several individuals A/A homozygous for rs2248374 were A/G heterozygous for rs27302, and the three outliers detected were all rs27302A/A homozygous, excluding the possibility of the two SNPs being in perfect LD in the thymic data set. Put together this indicated that rs27302, or other SNPs in high LD with it in the thymic data set, may have a greater impact on regulation of *ERAP2* expression than rs2248374. Hence, this is in accordance with the discovery of rs27302 showing stronger correlation to *ERAP2* expression than rs2248374 (Gabrielsen et al., 2016b).

It is important to remember that the analysis is partly based on microarray data from one *ERAP2* probe (ILMN_1743145) and that it is not definite that the analysis cover total *ERAP2* expression in the thymic samples. Furthermore, the eQTL analysis did not separate the different transcript variants in individual analyses, but included all *ERAP2* transcripts present in the sample capable of hybridization with the probe. Hence, preventing the investigation of expression variation between transcript variants, that may be caused by differing regulating mechanisms of transcripts. However, the observed *ERAP2* expression based on the microarray data utilized in this study gave a good indication of the correlation between rs2248374 and rs27302, and *ERAP2* expression in the thymic samples. Another limitation to the eQTL analysis was that the thymic samples were obtained from individuals with conditions that caused them to undergo corrective cardiac surgery, rendering it difficult to determine if the eQTL observations were representative to the correlation of *ERAP2* expression and rs2248374 and rs27302 genotypes in thymus in the general population. However, several studies have performed eQTL analyses on *ERAP2* expression in LCLs and identified other SNPs exhibiting stronger correlation to gene expression than rs2248374

(Cheung et al., 2005; Kuiper et al., 2014). Hence, the observation of rs27302 as an *ERAP2* eQTL showing stronger correlation with expression than rs2248374 in the thymic material coincided with previous findings.

5.1.3 Presence of *ERAP2* transcripts follows rs2248374 genotypes in both thymic samples and LCLs

According to the screening of the exon 10 splice junction, assessing expression of regular exon 10 spliced *ERAP2* transcripts and alternatively spliced *ERAP-208* transcripts, expression followed a distinct pattern with regards to rs2248374 genotypes in both thymic samples and LCLs. Individuals A/A homozygous for rs2248374 expressed regular exon 10 spliced *ERAP2* transcripts, G/G homozygotes expressed alternatively spliced *ERAP2-208*, while A/G heterozygotes showed expression of both. This expression pattern coincided with the findings from a similar analysis performed by Andrés et al. (2010), from where the primers utilized in this thesis was obtained (Andres et al., 2010). Hence, this supported the statements defining rs2248374 as the SNP regulating *ERAP2* expression, contradicting the observations from the eQTL analysis in 5.1.2, as *ERAP2* transcripts were expressed regardless of rs27302 genotypes. These findings questioned the role of rs27302 as the main *ERAP2* regulating SNP, although it is still significantly correlated with *ERAP2* expression as an eQTL, and may be involved in *ERAP2* expression in addition to other SNPs or biological factors.

Amplification of the exon 10 splice junction, utilizing the same primer pair as Andrés et al. (2010), led to the detection of bands of two sizes, presumably 136bp and 192bp. Six *ERAP2* transcripts encompasses exon 10 and exon 11, which contain the binding areas of the primers. Furthermore, it was indicated that five of the transcripts, excluding alternatively spliced *ERAP-208* (expected size 192bp), all encode regular exon 10 spliced *ERAP2* transcripts (Figure 9, page 35), proposing that the 136bp band detected in the analyses of this study and the study of Andrés et al. (2010) may theoretically represent four other *ERAP2* transcripts in addition to the full length variant. Although Andrés et al. (2010) described the 136bp band in their samples as full length *ERAP2* since it correlated with the observed protein expression, this may not be applicable to the thymic samples in this study. Full length *ERAP2* is described as one of the most expressed *ERAP2* variants, in addition to alternatively spliced *ERAP2-208*, in LCLs (Andres et al., 2010). Since the thymic samples have a different tissue of origin than LCLs, it is not certain that a similar pattern of *ERAP2* expression, with mainly two transcripts, was present in these samples. Hence, it is not definite that the 136bp band in the thymic samples mainly represented full length *ERAP2*. Considering these aspects, the overall conclusion was that full length *ERAP2* and alternatively spliced *ERAP-208* transcripts were represented in the exon 10 splice junction analysis of the LCL. Moreover, alternatively spliced *ERAP2-208* transcripts were represented in the analysis utilizing the thymic samples, but it was uncertain which transcripts were represented by the 136bp band.

5.2 Variable levels of full length ERAP2 protein expression based on rs2248374 and rs27302 genotypes

The previously discussed analyses assessing the expression of *ERAP2* transcripts yielded deviating results. The screening of the exon 10 splice junction questioned that regulation of *ERAP2* expression was contingent on rs27302 genotypes, while the eQTL analysis indicated that *ERAP2* expression follows a specific pattern with regard to rs27302 genotypes in the thymic data set. Furthermore, the analyses expressed opposite effects regarding rs2248374 genotypes. This study assessed the involvement of the two SNPs further by analyzing the expression of ERAP2 protein in the majority of samples included in the analyses discussed in section 5.1.

Western blotting of the thymic samples selected based on genotypes for rs2248374 and rs27302, corresponded with the hypothesis of this study, by implying that rs27302 or other SNPs in high LD with it may be the SNP(s) regulating *ERAP2* gene expression. The main discovery was the absence of full length ERAP2 in the thymic samples T30, being A/A homozygous for both SNPs, and T10, being A/A homozygous for rs27302 and A/G heterozygous for rs2248374 (Figure 17, page 45). The results did not coincide with current theory claiming that individuals carrying the rs2248374 A-allele should express full length ERAP2 protein, while G/G homozygotes should lack expression of the protein as the alternatively spliced *ERAP2-208* transcripts are degraded by NMD (Andres et al., 2010). On the contrary, the western blot results were in agreement with the expression pattern expected based on the eQTL analysis of rs27302. According to this it was expected that rs27302 A/A homozygote individuals would lack the expression of full length ERAP2 protein.

Further western blot analyses using protein lysate from the 18 LCLs, resulted in both expected and unexpected detection of full length ERAP2 in the various samples, based on the hypothesis of this study. Expectedly, all individuals A/G heterozygous for rs27302 carrying a rs2248374 A-allele, expressed full length ERAP2, while the protein was absent in the rs27302A/G-rs2248374G/G samples (GM18923, GM19204), that encoded the alternative spliced *ERAP-208* variant suggested to undergo NMD. Additionally, rs27302G/G-rs2248374A/A individuals showed precence of full length ERAP2, while rs27302G/G-rs2248374G/G individuals did not express the protein. Unexpectedly, the absence of full length ERAP2 that was observed in the thymic samples (T30, T10) A/A homozygous for both SNPs, was not observed in LCLs with the same genotypes (GM18646, GM19792, GM19916). The combined results for the thymic and LCL blots indicated that neither rs2248374 nor rs27302 are likely to be the sole *ERAP2* regulating SNPs, meaning that a different SNP, several SNPs or other biological factors may be the key or additional regulators of this gene. However, they do not exclude the potential involvement of the studied SNPs in *ERAP2* expression, nor that the discrepancy in ERAP2 expression between the two tissues could be a result of tissue specific regulation.

Beta actin was detected in all samples included on the different blots, indicating that protein loading was successful. However, the levels of beta actin were reduced in sample T30 compared to the remaining thymic samples (Figure 17, page 45). This may be due to T30 being the only sample, originating from an individual >12 years of age included in this study. The remaining eight samples were collected from children <2 years of age (1 individual <2 years, 7 individuals <1 year). As stated in the introduction (1.1), the thymus gland experience an involution process, where gland tissue is replaced by fat tissue (Gui et al., 2012). The T30 tissue may have consisted of more fat than gland tissue and had an overall reduced protein expression. Initially, this may be considered contradictory to the observed expression of *beta actin* and *GAPDH* mRNA in thymus, where T30 expressed these genes in similar amounts as the remaining thymic samples. This disparity may be partly explained by the fact that two separate pieces of thymic tissue were used to isolate protein and RNA. Thymus is a heterogeneous tissue (Gui et al., 2012), and the two pieces may have had different compositions and thereby different protein and RNA expression. In an attempt to prevent this situation, a sample, T27, containing protein isolated with TRIzol was included in the initial western blotting experiment (Figure 16, page 44). This protein was isolated from the same tissue as the T27 RNA used to generate the microarray data. Unfortunately, this lysate was not applicable to western blotting as ERAP2 and beta actin was absent from this sample, while it was present in T27 where protein was isolated with RIPA. Consequently, protein lysates were obtained using RIPA lysis buffer, from new pieces of thymic tissue for all nine samples included in the western blotting experiment (Figure 17, page 45).

Expression levels of ERAP2 were assessed by normalizing full length ERAP2 expression with beta actin expression, in samples expressing full length ERAP2. The objective was to evaluate if the expression levels of full length ERAP2 followed a distinct pattern based on genotypes for rs2248374 and rs27302. Expression of full length ERAP2 in the LCLs supported previous reports by indicating that the expression was regulated by rs2248374 genotypes (Andres et al., 2010), as A/A homozygotes and A/G heterozygotes expressed full length ERAP2, while it was absent in samples G/G homozygous for the SNP. Similar amounts of full length ERAP2 was observed internally on each blot, although slightly different expression levels were detected between blots. The rs27302 A/G blot showed the greatest expression while the lowest amounts were observed in the rs27302 A/A blot. When using GM12043 as reference sample in the individual blots a similar intensity distribution was observed between the blots (data not shown). Furthermore, rs2248374 A/G heterozygotes and A/A homozygotes included in the rs27302 A/G blot expressed similar amounts of full length ERAP2, although the homozygous group were expected to show the highest expression. The variations observed between blots may be due to the fact that western blotting is a technique composed of several steps rendering experiments exposed to loss of protein and inaccurate protein measurement and detection, e.g. during sample preparation, quantification of protein concentrations, protein transfer, antibody probing and image development (Murphy & Lamb, 2013). Furthermore, the similar expression levels detected between rs2248374 A/A homozygotes and A/G heterozygotes, may be explained by a biological mechanism termed allele specific expression. Alleles can be subject to allele specific expression rendering one allele to be more extensively expressed than the other allele (Ronald et al., 2005). In this case, the rs2248374 A-allele may have dominated the expression. Moreover, *ERAP2* expression may be regulated by post-transcriptional mechanisms adjusting mRNA levels e.g. by miRNA binding to a complementary mRNA either inhibiting translation or initiating mRNA degradation (Wahid et al., 2010), or by other translational regulatory factors. However, these considerations expires as western blotting is a semi-quantitative method (Murphy & Lamb, 2013) and the measurements of full length ERAP2 and beta actin expression was performed manually within the ImageQuant^{TM} TL Toolbox v8.1 software (GE Healthcare) being subject to human bias. Thus, no definite conclusions, regarding *ERAP2* expression levels, can be drawn based on the western blotting results. Hence, the western blotting analysis mainly applied to investigation of the overall expression of ERAP2 in the different samples.

The study design could have been improved by obtaining an antibody specific to detection of alternatively spliced ERAP2-208 protein. Although it is stated that this protein is not expressed, as its encoding transcript is thought to be degraded through NMD (Andres et al., 2010), an investigation of its possible expression would have been of interest. Future studies on ERAP2 expression may benefit from including such an antibody, and other ERAP2 antibodies, in their experiments.

Unspecific bands were detected when probing with anti-ERAP2 (ab69037, Abcam) in all thymic samples and in six of the LCL samples expressing ERAP2 (GM07000, GM11832, GM12043, GM18646, GM19792, GM19916). Presumably these bands are the result of the presence of proteins or amino acid residues containing an epitope able to bind anti-ERAP2 antibody. These proteins might have represented multimers, e.g. the ERAP1/ERAP2 heterodimer, but this was rather unlikely as the reducing conditions of the SDS-PAGE should have prevented their appearance on the blot. Alas, further insight into the specificities of the unknown molecules was not obtainable based on the resources available during this study. Other commercial ERAP2 antibodies are available, which could have been utilized in stead of ab69037 (Abcam). However, several of these, in addition to ab69037 (Abcam), show presence of an extra band between 37-50kDa size, with unknown identity, in addition to the 110kDa band representing full length ERAP2. Hence it is uncertain if employment of these antibodies would have improved the analysis.

5.3 Inconsistency in *ERAP2* expression between RNA and protein levels

An overall assessment of the RNA and protein observations revealed an interesting finding. The majority of samples that expressed regular exon 10 spliced *ERAP2* transcripts expressed full length ERAP2, indicating the presence of full length *ERAP2* transcripts in the samples. However, this was no true for the thymic samples T10 and T30. Regular exon 10 spliced transcripts were detected in both samples, but the full length protein was absent. This indicated that the 136 bp band did not represent full length *ERAP2* transcripts, or that something interfered with the translation of the transcript into full length protein, as there are several mechanisms regulating the pathway of RNA being translated into protein (Nature, 2014). Furthermore, the inconsistency between *ERAP2* expression on RNA and protein level in T10 and T30 suggested that a regulatory mechanism(s) may have caused full length *ERAP2* transcripts or the ERAP2 protein to evade expression. Provided that the inconsistency detected between the analysis is caused by something biological, future studies may reveal the mechanism behind this deviation.

5.4 LD patterns in the genomic region of *ERAP2* vary between the six populations represented in the study samples

LD is an important aspect to consider when executing genetic studies and especially when including samples originating from individuals of different ethnicity as the LCLs used in this thesis. The selection of LCLs was mainly based on genotypes of rs2248374 and rs27302, but an important notion is that the LCLs also were selected based on their presence in the study of Andrés et al. (2010). The researchers detected rs2248374 as the main *ERAP2* regulating eQTL without including rs27302 in their analysis. Hence, this study sought to expand the assessment of *ERAP2* expression by including rs27302 genotypes in its analyses by utilizing some of the LCLs used by Andrés et al. (2010) in addition to samples that were absent in their study, which had other genotype combinations for the two SNPs.

The LD pattern in the thymic data set and the Caucasian population (1000 Genomes Project, Phase 3) was compared by assessing LD in the genomic region of *ERAP2* and LD between nine SNPs in the *ERAP2* region, including the splice SNP, rs2248374, and the novel eQTL, rs27302 (Table 7, page 32). Evaluation of the LD patterns found that the thymic data set showed a similar LD pattern as the Caucasian population data (1000 Genomes Project, Phase 3). The thymuses were derived from Norwegian individuals, often referred to as a Caucasian population, and based on the LD results the thymic samples were deemed suitable representatives for the Caucasian population. In addition, LD was assessed between rs2248374 and rs27302, while including seven other SNPs in the *ERAP2* region, in all populations represented by the LCLs and thymic samples. The overall analysis showed that all though the majority of SNPs were in high LD, the populations had differing LD patterns. Equilibrium was absent between the nine SNPs in all populations. Several studies have found deviating associations between phenotypes and *ERAP2* polymorphisms in individuals of various populations (Hill et al., 2011; Vanhille et al., 2013). E.g. although *ERAP1/ERAP2* SNPs had been identified as polymorphisms associated with AS in European Caucasians in previous studies, Su et al. (2018) found no such association in the Han Chinese population (Su et al., 2018), indicating the influence of differing genetic architecture on association with phenotypes e.g. gene expression.

The difference in LD patterns between the six populations may be employed when proposing possible explanations for the observed correlation between rs27302 and *ERAP2* expression in thymus, which was not observed in the LCLs. The strong correlation between rs27302 and *ERAP2* expression detected in the eQTL analysis and thymic western blot may have been the result of an indirect signal caused by LD effects. This implies that a SNP, or several SNPs, in high LD with rs27302 in thymus could be the causal *ERAP2* regulating SNP(s), and that the observed influence of rs27302 on *ERAP2* expression was in fact the result of the effect of alleles in this SNP(s). The three LCL samples representing A/A-A/A individuals, which indicated that rs27302 was not the sole causal *ERAP2* SNP, were of different origin than the thymic samples (CEU), namely Mexican, African and Han Chinese. The difference in LD patterns between these four populations proposed that high LD between rs27302 and a possible causative SNP(s) in the thymic samples may not exert the same degree of LD between them in the three LCL populations represented by the A/A-A/A individuals. Consequently, resulting in the samples not exhibiting the same ERAP2 expression pattern that were expected for rs27302 genotypes (with the A/A homozygotes lacking expression of full length ERAP2), based on the initial eQTL and thymic western blot results. Moreover, the splice SNP, rs2248374, may be in higher LD with the causal SNP(s) in the Mexican, African and Han Chinese populations, than the Caucasian population. Hence, leading to the assumption of rs2248374 as the causal SNP, although the correlation between rs2248374 genotypes and *ERAP2* expression might be the result of an indirect signal.

To summarize, the deviating expression of full length ERAP2 protein between thymic samples and LCLs, indicated rs27302 as the causal *ERAP2* SNP in the thymic samples and rs2248374 as the causal SNP in the LCLs. However, the detection of varying LD patterns between the four populations suggested that there might be a causal SNP(s) in high LD with rs27302 in the Caucasian population (thymic samples), and high LD with rs2248374 in the Mexican, African and Han Chinese populations (LCLs) that mislead the identification of the two SNPs as solely causal to *ERAP2* expression. Although the results implied that the true causal SNP(s) remained undetected in part due to LD effects, it is important to remember that the contradicting results were observed between samples originating from different tissue.

Hence, the analysis cannot exclude that tissue specific effects might have influenced *ERAP2* expression resulting in the variations observed between the two tissues as well.

Furthermore, it is questionable if the LD patterns observed in the six populations are representative for the total populations. The 1000 Genomes Project populations consist of many individuals. Alas, much of the data was not included in the download of the six populations from the Ensembl website (Ensembl, 2018c), resulting in small sample sizes, with data from 15-22 individuals, representing each population in this study. With the limited sample sizes, one individual expresses a larger effect on the LD measurements. If one individual encodes a rare variant this may drastically skew the LD pattern because of the limited number of individuals, compared to the actual pattern in the population. If it is excluded it may result in an incomplete representation of the populations haplotypes. Based on the 1000 Genomes Project data included in this study it is difficult to determine the true LD patterns of the populations, indicating that the data must be used with caution. However, the LD patterns serve as indicators of different genomic architecture between the populations, in the genomic region of *ERAP2* and rs27302.

5.5 Diverging success in generating amplicons covering the genomic region of *ERAP2* and rs27302

Sequencing of the *ERAP2* gene and a genomic region containing rs27302 in all samples (18 LCLs and 48 Thymic) was considered advantageous in obtaining insight into possible causal SNPs involved in regulating *ERAP2* expression, while simultaneously determining the genotypes in rs2248374 and rs27302 to exclude potential mutations and genotyping errors. Successful targeted sequencing is dependent on proper library preparation (Head et al., 2014; Ye et al., 2012). This entails that primers amplify the genomic regions they were designed to, and that they are excluding amplification of undesirable genomic areas. Several experiments were performed to assess the primer specificity of the primer pairs designed to cover *ERAP2* and rs27302. The initial experiment found six out of ten primer pairs to fulfill the criteria of primer specificity (Figure 28, page 53). Because amplicon lengths were not in concordance with the 1kb GeneRuler^{TM} DNA ladder, the determination of amplicon length was based on the size of the PCR products relative to one another. The ladder was employed when determining the approximate size difference between samples, by assessing the relative distance between bands within the ladder. The deviation between the ladder and samples may be the result of challenges faced when performing gel electrophoresis. Successful electrophoresis depends on use of the proper gel percent, electrophoresis buffer, staining procedure, voltage applied and duration, regarding the properties of the DNA material being separated. Hence, all parameters may have influenced the gel electrophoresis. However, investigation found that preparation of the DNA ladder had not been executed as recommended by the manufacturer, hence the deviation was likely influenced by this. The method of band detection also applied in the remaining experiments implementing the 1kb or 50bp GeneRuler^{TM} DNA ladders (Thermo Fisher Scientific), including the screening of the exon 10 splice junction.

PCR optimization is necessary when the reaction fails to amplify the proper amplicons (Roux, 2009). This concerned the generation of amplicons from four of the primer pairs utilized to generate amplicons covering the genomic region of *ERAP2*, i.e. Amp1, Amp3, Amp7 and Amp10. Optimization can be achieved by varying one or several of the parameters of the reaction, e.g. Mg^{2+} concentrations, annealing temperature and the composition of the PCR solution (Roux, 2009). Alas, optimization of the PCR conditions by adjusting temperature and adding X-solution (GenDx) in the four reactions were unsuccessful. Temperature did not exert any significant effect in Amp1 or Amp7 reactions, and although addition of X-solution (GenDx) slightly increased the presence of Amp10, it reduced the amount of PCR products for the remaining amplicons. The increase observed in Amp10 was not deemed sufficient at this point in the study. The absence of Amp3 PCR products in well 7-12 (Figure 29A, page 54) was probably the result of exclusion of one or several reagents.

To resolve the unspecific PCR products generated by the four primer pairs, new primers were obtained. Unfortunately, identical primers were obtained for both Amp3 primers and the reverse primer of Amp7. This was discovered at a point in time where the order of additional primers was unfavorable due to time limitations. The lack ofAmp3 amplicon in all Amp3 reactions was unexpected. The new primers contained the identical sequence as the original primers, which generated an amplicon of the expected length (Figure 28, page 53). The main difference between the two experiments was the DNA samples used as input, which may have influenced the amplification. Although the lack of Amp 3 could be a result of insufficient reagents, this is not likely as the other reactions included in the experiment contained the exact same aliquots of reagents, and they all showed presence of PCR products. Moreover, human bias cannot be excluded from this assessment, as it may have caused the unexpected results observed for Amp3. None of the Amp7 reactions, including the reactions of the optimization experiment (Figure 31, page 57), generated the expected amplicon, and unspecific bands were present. Since the original and new reverse Amp7 primers were identical, and a slightly similar pattern was observed for the amplicon in Figures 28-31, it was likely that these unspecific bands were the result of adverse reactions caused by properties of the reverse primer.

Modification of polymerases have resulted in long range PCR being able to amplify amplicons of sizes up to >30kb depending on the polymerases utilized (Jia et al., 2014). The LongRange Enzyme (GenDx) utilized to amplify the amplicons covering the *ERAP2* and rs27302 regions can also be used to generate amplicons >10kb, however the amplification of amplicons <10kb is more likely to be successful. Hence, the original amplicons were designed to cover 2-6.5kb. However, because of the unsuccessful amplification of Amp7, an experiment was executed to generate extended amplicons covering this region. Alas, amplification of extended amplicons, Amp6-7 and Amp7-8 using forward and reverse primers from adjacent regions, were unsuccessful. The expected size of these products was ~12kb, being very large and thereby not usually optimal for most PCR amplifications. The amplicons may have been generated by application of another LongRange PCR kit, containing an enzyme able to generate the ~12kb products. However, resource and time limitations inhibited such experiments. Furthermore, a more time effective alternative would be the separation of the Amp7 region into two amplicons, only necessitating the procurement of two primer pairs covering this region.

5.6 Considerations for future studies

There are several adjustments to the study design that could have improved the overall insight into the influence of SNPs on *ERAP2* expression on RNA and protein level. The study material could have contained samples from all genotypes of rs27302 and rs2248374, by also including A/A-A/G and G/G-A/G samples. In addition, the LCL panel could have been expanded to encompass an rs27302A/A-rs2248374A/A sample representing the Caucasian population. Hence, enabling further considerations of whether the thymic western blotting results were representative to the Caucasian population or if the absence of ERAP2 in the thymic A/A-A/A samples was caused by tissue specific mechanisms, experimental errors or due to faulty samples. Moreover, a secondary analysis assessing the expression of *ERAP2* transcripts in the thymic samples would have been useful to confirm the observations based on the microarray data. Furthermore, an overall study utilizing samples with all rs27302-rs2248374 genotypes from all populations may have yielded a more complete overview of the SNPs effect on *ERAP2* expression. However, this study would be very expensive and time consuming, and all genotypes are not available for all populations in the current 1000 Genomes Project repository. Additionally, the aspect of LD in the genomic region of *ERAP2* in the various populations should be assessed further. Moreover, optimization of the western blot procedure with the inclusion of ERAP2-208 sensitive antibodies, in addition to antibodies sensitive to the remaining ERAP2 proteins, could have increased the coverage of the western blot analysis.

Future studies may benefit from completing the sequencing of the genomic region of *ERAP2* and rs27302. Depending on the resources available the researchers may gain a greater insight by sequencing an extended region of the gene, e.g. including the gap between *ERAP2* and rs27302 (located downstream *LNPEP*) and the promotor region of *ERAP2* to investigate SNPs that may influence alternative splicing and/or transcription factors. Sequencing may reveal causative SNPs involved in the regulation of *ERAP2*, expanding the current knowledge of mechanisms of *ERAP2* expression.

6. Concluding remarks

The aim of this thesis was to gain a better understanding of expression of the *ERAP2* gene on both RNA and protein level, and assess the influence of SNPs in the *ERAP2* region on gene expression. The experiments focused on genotypes of the splice SNP, rs2248374, and the novel eQTL, rs27302, with regard to the expressional pattern of *ERAP2*. This led to several interesting discoveries. The current theory stating rs2248374 as the causal SNP in *ERAP2* regulation was questioned. Observations based on the 42 thymic samples discovered outliers that deviated from the expected pattern in the rs2248374 eQTL analysis, and ERAP2 protein was absent in individuals A/A homozygous for the SNP. Furthermore, the eQTL analysis and thymic western blotting experiment indicated a strong regulatory influence of rs27302 on *ERAP2* expression. However, further analysis involving the LCL samples found results to better coincide with current theory regarding rs2248374. Hence, results obtained from the thymic samples contradicted the findings of experiments employing the LCLs. Put together the following conclusions can be made; 1) Neither rs2248374 nor rs27302 are the causal or only SNPs regulating *ERAP2* expression, 2) one or several other SNPs in high LD with rs27302 in the Caucasian population and/or other biological factor are probably involved in regulation of *ERAP2* expression.

7. References

Abcam. (2018). *Sample preparation for western blot*. Available at: http://www.abcam.com/protocols/sample-preparation-for-western-blot (accessed: 14.02.18).

Aggarwal, A. (2014). Role of autoantibody testing. *Best Pract Res Clin Rheumatol*, 28 (6): 907-20. doi: 10.1016/j.berh.2015.04.010.

Agrawal, N. & Brown, M. A. (2014). Genetic associations and functional characterization of M1 aminopeptidases and immune-mediated diseases. *Genes Immun*, 15 (8): 521-7. doi: 10.1038/gene.2014.46.

Akira, S., Uematsu, S. & Takeuchi, O. (2006). Pathogen recognition and innate immunity. *Cell*, 124 (4): 783-801. doi: 10.1016/j.cell.2006.02.015.

Alberts B, J. A., Lewis J, et al. (2002). *Molecular Biology of the Cell*. 4th ed. New York: Garland Science.

Amundsen, S. S., Viken, M. K., Sollid, L. M. & Lie, B. A. (2014). Coeliac disease-associated polymorphisms influence thymic gene expression. *Genes Immun*, 15 (6): 355-60. doi: 10.1038/gene.2014.26.

Anaya, J. M., Corena, R., Castiblanco, J., Rojas-Villarraga, A. & Shoenfeld, Y. (2007). The kaleidoscope of autoimmunity: multiple autoimmune syndromes and familial autoimmunity. *Expert Rev Clin Immunol*, 3 (4): 623-35. doi: 10.1586/1744666X.3.4.623.

Andres, A. M., Dennis, M. Y., Kretzschmar, W. W., Cannons, J. L., Lee-Lin, S. Q., Hurle, B., Program, N. C. S., Schwartzberg, P. L., Williamson, S. H., Bustamante, C. D., et al. (2010). Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. *PLoS Genet*, 6 (10): e1001157. doi: 10.1371/journal.pgen.1001157.

Aw, D. & Palmer, D. B. (2012). It's not all equal: a multiphasic theory of thymic involution. *Biogerontology*, 13 (1): 77-81. doi: 10.1007/s10522-011-9349-0.

Ayensu, W. K., Tchounwou, P. B. & McMurray, R. W. (2004). Molecular and cellular mechanisms associated with autoimmune diseases. *Int J Environ Res Public Health*, 1 (1): 39-73.

Benovoy, D., Kwan, T. & Majewski, J. (2008). Effect of polymorphisms within probe-target sequences on olignonucleotide microarray experiments. *Nucleic Acids Research*, 36 (13): 4417-4423. doi: 10.1093/nar/gkn409.

Blum, J. S., Wearsch, P. A. & Cresswell, P. (2013). Pathways of antigen processing. *Annu Rev Immunol*, 31: 443-73. doi: 10.1146/annurev-immunol-032712-095910.

Boehm, T., Takahama, Y. & SpringerLink. (2013). *Thymic Development and Selection of T Lymphocytes*. Current Topics in Microbiology and Immunology, vol. v.373. Berlin/Heidelberg: Springer.

Bolon, B. (2012). Cellular and molecular mechanisms of autoimmune disease. *Toxicol Pathol*, 40 (2): 216-29. doi: 10.1177/0192623311428481.

Celedón JC, H. G. (2017). *Principles of complex trait genetics*. In Benjamin A Raby (ed.). Available at: https://www.uptodate.com/contents/principles-of-complex-trait-genetics.

Chang, Y. F., Imam, J. S. & Wilkinson, M. F. (2007). The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem*, 76: 51-74. doi: 10.1146/annurev.biochem.76.050106.093909.

Cheung, V. G., Spielman, R. S., Ewens, K. G., Weber, T. M., Morley, M. & Burdick, J. T. (2005). Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, 437 (7063): 1365-9. doi: 10.1038/nature04244.

Cho, J. H. & Feldman, M. (2015). Heterogeneity of autoimmune diseases: pathophysiologic insights from genetics and implications for new therapies. *Nat Med*, 21 (7): 730-8. doi: 10.1038/nm.3897.

Cortes, A. & Brown, M. A. (2011). Promise and pitfalls of the Immunochip. *Arthritis Res Ther*, 13 (1): 101. doi: 10.1186/ar3204.

Coulombe-Huntington, J., Lam, K. C., Dias, C. & Majewski, J. (2009). Fine-scale variation and genetic determinants of alternative splicing across individuals. *PLoS Genet*, 5 (12): e1000766. doi: 10.1371/journal.pgen.1000766.

Dranoff, G. (2004). Cytokines in cancer pathogenesis and cancer therapy. *Nat Rev Cancer*, 4 (1): 11-22. doi: 10.1038/nrc1252.

Dueker, N. D. & Pericak-Vance, M. A. (2014). Analysis of genetic linkage data for Mendelian traits. *Curr Protoc Hum Genet*, 83: 1 4 1-31. doi: 10.1002/0471142905.hg0104s83.

Eisenberg, E. & Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends Genet*, 29 (10): 569-74. doi: 10.1016/j.tig.2013.05.010.

Elliott, D. E., Siddique, S. S. & Weinstock, J. V. (2014). Innate immunity in disease. *Clin Gastroenterol Hepatol*, 12 (5): 749-55. doi: 10.1016/j.cgh.2014.03.007.

EMBL-EBI. (2018). *What is genetic variation?* Available at: https://www.ebi.ac.uk/training/online/course/human-genetic-variation-i-introduction/what-genetic-variation (accessed: 21.04.2018).

Ensembl. (2018a). *Chromosome 5:96,875,939-96,919,716*. Available at: https://www.ensembl.org/Homo_sapiens/Location/View?r=5:96875939-96919716;tl=JuvbQTlPbzsE1NLw-3532889-663501017;db=core;g=ENSG00000164308 (accessed: 17.04.12018).

Ensembl. (2018b). *Chromosome 5: 97,037,996-97,038,096*. Available at: http://www.ensembl.org/Homo_sapiens/Location/View?db=core;r=5:97037996-97038096;source=dbSNP;v=rs27302;vdb=variation;vf=18770;time=1524482083 (accessed: 08.05.2018).

Ensembl. (2018c). *VCF to PED Converter*. Available at: http://grch37.ensembl.org/Homo_sapiens/Tools/VcftoPed?db=core;tl=RqAFHiXSL1JGHPno-3875998 (accessed: 30.01.2018).

Faber, K., Glatting, K. H., Mueller, P. J., Risch, A. & Hotz-Wagenblatt, A. (2011). Genome-wide prediction of splice-modifying SNPs in human genes using a new analysis pipeline called AASsites. *BMC Bioinformatics*, 12 Suppl 4: S2. doi: 10.1186/1471-2105-12-S4-S2.

Fierabracci, A., Milillo, A., Locatelli, F. & Fruci, D. (2012). The putative role of endoplasmic reticulum aminopeptidases in autoimmunity: insights from genomic-wide association studies. *Autoimmun Rev*, 12 (2): 281-8. doi: 10.1016/j.autrev.2012.04.007.

Fridkis-Hareli, M. (2008). Immunogenetic mechanisms for the coexistence of organ-specific and systemic autoimmune diseases. *J Autoimmune Dis*, 5: 1. doi: 10.1186/1740-2557-5-1.

Gabrielsen, I. S., Amundsen, S. S., Helgeland, H., Flam, S. T., Hatinoor, N., Holm, K., Viken, M. K. & Lie, B. A. (2016a). Genetic risk variants for autoimmune diseases that influence gene expression in thymus. *Hum Mol Genet*, 25 (14): 3117-3124. doi: 10.1093/hmg/ddw152.

Gabrielsen, I. S., Viken, M. K., Amundsen, S. S., Helgeland, H., Holm, K., Flam, S. T. & Lie, B. A. (2016b). Autoimmune risk variants in ERAP2 are associated with gene-expression levels in thymus. *Genes Immun*, 17 (7): 406-411. doi: 10.1038/gene.2016.39.

Gay, D., Saunders, T., Camper, S. & Weigert, M. (1993). Receptor editing: an approach by autoreactive B cells to escape tolerance. *J Exp Med*, 177 (4): 999-1008.

GeneCards. (NA-a). *ACTB Gene*. Available at: http://www.genecards.org/cgi-bin/carddisp.pl?gene=ACTB (accessed: 08.05.2018).

GeneCards. (NA-b). *ERAP2 Gene*. Available at: http://www.genecards.org/cgi-bin/carddisp.pl?gene=ERAP2 (accessed: 30.04.2018).

GeneCards. (NA-c). *GAPDH Gene*. Available at: http://www.genecards.org/cgi-bin/carddisp.pl?gene=GAPDH (accessed: 08.05.2018).

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

Graham, K. L., Sutherland, R. M., Mannering, S. I., Zhao, Y., Chee, J., Krishnamurthy, B., Thomas, H. E., Lew, A. M. & Kay, T. W. (2012). Pathogenic mechanisms in type 1 diabetes: the islet is both target and driver of disease. *Rev Diabet Stud*, 9 (4): 148-68. doi: 10.1900/RDS.2012.9.148.

Gregersen, P. K. & Behrens, T. W. (2006). Genetics of autoimmune diseases--disorders of immune homeostasis. *Nat Rev Genet*, 7 (12): 917-28. doi: 10.1038/nrg1944.

Gregersen, P. K. & Olsson, L. M. (2009). Recent advances in the genetics of autoimmune disease. *Annu Rev Immunol*, 27: 363-91. doi: 10.1146/annurev.immunol.021908.132653.

Groettrup, M., Kirk, C. J. & Basler, M. (2010). Proteasomes in immune cells: more than peptide producers? *Nat Rev Immunol*, 10 (1): 73-8. doi: 10.1038/nri2687.

Gui, J., Mustachio, L. M., Su, D. M. & Craig, R. W. (2012). Thymus Size and Age-related Thymic Involution: Early Programming, Sexual Dimorphism, Progenitors and Stroma. *Aging Dis*, 3 (3): 280-90.

Haroon, N. & Inman, R. D. (2010). Endoplasmic reticulum aminopeptidases: Biology and pathogenic potential. *Nat Rev Rheumatol*, 6 (8): 461-7. doi: 10.1038/nrrheum.2010.85.

Harvey, D., Pointon, J. J., Karaderi, T., Appleton, L. H., Farrar, C. & Wordsworth, B. P. (2011). A common functional variant of endoplasmic reticulum aminopeptidase 2 (ERAP2) that reduces major histocompatibility complex class I expression is not associated with ankylosing spondylitis. *Rheumatology (Oxford)*, 50 (9): 1720-1. doi: 10.1093/rheumatology/ker199.

Hassan, M. A., Butty, V., Jensen, K. D. C. & Saeij, J. P. J. (2014). The genetic basis for individual differences in mRNA splicing and APOBEC1 editing activity in murine macrophages. *Genome Research*, 24 (3): 377-389. doi: 10.1101/gr.166033.113.

Hattori, A. & Tsujimoto, M. (2004). Processing of antigenic peptides by aminopeptidases. *Biol Pharm Bull*, 27 (6): 777-80.

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R. & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*, 56 (2): 61-4, 66, 68, passim. doi: 10.2144/000114133.

Hill, L. D., Hilliard, D. D., York, T. P., Srinivas, S., Kusanovic, J. P., Gomez, R., Elovitz, M. A., Romero, R. & Strauss, J. F. (2011). Fetal ERAP2 variation is associated with preeclampsia in African Americans in a case-control study. *Bmc Medical Genetics*, 12. doi: Artn 6410.1186/1471-2350-12-64.

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.

Hrdlickova, B., Westra, H. J., Franke, L. & Wijmenga, C. (2011). Celiac disease: moving from genetic associations to causal variants. *Clin Genet*, 80 (3): 203-313. doi: 10.1111/j.1399-0004.2011.01707.x.

Janeway CA Jr, T. P., Walport M, et al. (2001). *Immunobiology: The Immune System in Health and Disease.* 5th ed. New York: Garland Science.

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.

Kenna, T. J., Robinson, P. C. & Haroon, N. (2015). Endoplasmic reticulum aminopeptidases in the pathogenesis of ankylosing spondylitis. *Rheumatology (Oxford)*, 54 (9): 1549-56. doi: 10.1093/rheumatology/kev218.

Khan Academy. (2017). *Adaptive immunity*. Available at: https://www.khanacademy.org/test-prep/mcat/organ-systems/the-immune-system/a/adaptive-immunity (accessed: 17.10.2017).

Kibbe, W. A. (2007). OligoCalc: an online oligonucleotide properties calculator. *Nucleic Acids Res*, 35 (Web Server issue): W43-6. doi: 10.1093/nar/gkm234.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, 308 (5720): 385-9. doi: 10.1126/science.1109557.

Kuiper, J. J., Van Setten, J., Ripke, S., Van, T. S. R., Mulder, F., Missotten, T., Baarsma, G. S., Francioli, L. C., Pulit, S. L., De Kovel, C. G., et al. (2014). A genome-wide association study identifies a functional ERAP2 haplotype associated with birdshot chorioretinopathy. *Hum Mol Genet*, 23 (22): 6081-7. doi: 10.1093/hmg/ddu307.

Lankat-Buttgereit, B. & Tampe, R. (2002). The transporter associated with antigen processing: function and implications in human diseases. *Physiol Rev*, 82 (1): 187-204. doi: 10.1152/physrev.00025.2001.

Lappalainen, T., Sammeth, M., Friedlander, M. R., t Hoen, P. A., Monlong, J., Rivas, M. A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501 (7468): 506-11. doi: 10.1038/nature12531.

Lea, T. (2013). *Immunologi og immunologiske teknikker*: Fagbokforlaget.

Lerner, A. & Matthias, T. (2015). Changes in intestinal tight junction permeability associated with industrial food additives explain the rising incidence of autoimmune disease. *Autoimmunity Reviews*, 14 (6): 479-489. doi: https://doi.org/10.1016/j.autrev.2015.01.009.

Lewin, B. (2006). *Essential Genes*: Pearson Education.

Liu, B., Shu, S., Kenny, T. P., Chang, C. & Leung, P. S. (2014). Stem cell therapy in autoimmune rheumatic diseases: a comprehensive review. *Clin Rev Allergy Immunol*, 47 (2): 244-57. doi: 10.1007/s12016-014-8445-8.

Liu, H., Bebu, I. & Li, X. (2010). Microarray probes and probe sets. *Front Biosci (Elite Ed)*, 2: 325-38.

Liu, L., Zhang, D., Liu, H. & Arendt, C. (2013). Robust methods for population stratification in genome wide association studies. *BMC Bioinformatics*, 14: 132. doi: 10.1186/1471-2105-14-132.

Lopez de Castro, J. A., Alvarez-Navarro, C., Brito, A., Guasp, P., Martin-Esteban, A. & Sanz-Bravo, A. (2016). Molecular and pathogenic effects of endoplasmic reticulum aminopeptidases ERAP1 and ERAP2 in MHC-I-associated inflammatory disorders: Towards a unifying view. *Mol Immunol*, 77: 193-204. doi: 10.1016/j.molimm.2016.08.005.

Maher, S., Toomey, D., Condron, C. & Bouchier-Hayes, D. (2002). Activation-induced cell death: the controversial role of Fas and Fas ligand in immune privilege and tumour counterattack. *Immunol Cell Biol*, 80 (2): 131-7. doi: 10.1046/j.1440-1711.2002.01068.x.

Marson, A., Housley, W. J. & Hafler, D. A. (2015). Genetic basis of autoimmunity. *J Clin Invest*, 125 (6): 2234-41. doi: 10.1172/JCI78086.

Meffre, E. & Wardemann, H. (2008). B-cell tolerance checkpoints in health and autoimmunity. *Curr Opin Immunol*, 20 (6): 632-8. doi: 10.1016/j.coi.2008.09.001.

Michaelson, J. J., Loguercio, S. & Beyer, A. (2009). Detection and interpretation of expression quantitative trait loci (eQTL). *Methods*, 48 (3): 265-76. doi: 10.1016/j.ymeth.2009.03.004.

Mozes, E., Lovchik, J., Zinger, H. & Singer, D. S. (2005). MHC class I expression regulates susceptibility to spontaneous autoimmune disease in (NZBxNZW)F1 mice. *Lupus*, 14 (4): 308-14. doi: 10.1191/0961203305lu2079oa.

Mueller, D. L. (2010). Mechanisms maintaining peripheral tolerance. *Nat Immunol*, 11 (1): 21-7. doi: 10.1038/ni.1817.

Murphy, R. M. & Lamb, G. D. (2013). Important considerations for protein analyses using antibody based techniques: down-sizing Western blotting up-sizes outcomes. *Journal of Physiology-London*, 591 (23): 5823-5831. doi: 10.1113/jphysiol.2013.263251.

Nacu, A., Andersen, J. B., Lisnic, V., Owe, J. F. & Gilhus, N. E. (2015). Complicating autoimmune diseases in myasthenia gravis: a review. *Autoimmunity*, 48 (6): 362-8. doi: 10.3109/08916934.2015.1030614.

Napolitano, G., Bucci, I., Giuliani, C., Massafra, C., Di Petta, C., Devangelio, E., Singer, D. S., Monaco, F. & Kohn, L. D. (2002). High glucose levels increase major histocompatibility complex class I gene expression in thyroid cells and amplify interferon-gamma action. *Endocrinology*, 143 (3): 1008-17. doi: 10.1210/endo.143.3.8674.

Nature. (2014). *Gene Expression*. Available at: https://www.nature.com/scitable/topicpage/gene-expression-14121669 (accessed: 01.05.2018).

NCBI. (2018). *ERAP2 endoplasmic reticulum aminopeptidase 2 [Homo sapiens (human)]*. Available at: https://www.ncbi.nlm.nih.gov/gene?cmd=Retrieve&dopt=full_report&list_uids=64167 (accessed: 30.04.2018).

Nemazee, D. (2017). Mechanisms of central tolerance for B cells. *Nat Rev Immunol*, 17 (5): 281-294. doi: 10.1038/nri.2017.19.

Nica, A. C. & Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci*, 368 (1620): 20120362. doi: 10.1098/rstb.2012.0362.

NIH. (2015). *All About The Human Genome Project (HGP)*. Available at: https://www.genome.gov/10001772/all-about-the--human-genome-project-hgp/)/ (accessed: 30.04.2018).

NIH. (2016). *An Overview of the Human Genome Project*. Available at: https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/ (accessed: 21.04.2018).

NIH. (2017). *1000 Genomes Project*. Available at: https://www.genome.gov/27528684/1000-genomes-project/ - al-1 (accessed: 01.05.2018).

NIH. (2018a). *What are single nucleotide polymorphisms (SNPs)?* Available at: https://ghr.nlm.nih.gov/primer/genomicresearch/snp (accessed: 03.05.18).

NIH. (2018b). *What is the International HapMap Project?* Available at: https://ghr.nlm.nih.gov/primer/genomicresearch/hapmap (accessed: 01.05.2018).

Nilsen, T. W. & Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463 (7280): 457-63. doi: 10.1038/nature08909.

Papakyriakou, A. & Stratikos, E. (2017). The Role of Conformational Dynamics in Antigen Trimming by Intracellular Aminopeptidases. *Front Immunol*, 8: 946. doi: 10.3389/fimmu.2017.00946.

Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am J Hum Genet*, 102 (1): 11-26. doi: 10.1016/j.ajhg.2017.11.002.

Parkes, M., Cortes, A., van Heel, D. A. & Brown, M. A. (2013). Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nature Reviews Genetics*, 14 (9): 661-673. doi: 10.1038/nrg3502.

Parkin, J. & Cohen, B. (2001). An overview of the immune system. *Lancet*, 357 (9270): 1777-89. doi: 10.1016/S0140-6736(00)04904-7.

Pohl, M., Bortfeldt, R. H., Grutzmann, K. & Schuster, S. (2013). Alternative splicing of mutually exclusive exons--a review. *Biosystems*, 114 (1): 31-8. doi: 10.1016/j.biosystems.2013.07.003.

Rada-Iglesias, A. (2014). Genetic variation within transcriptional regulatory elements and its implications for human disease. *Biological chemistry*, 395 (12): 1453-1460.

Ramos, P. S., Criswell, L. A., Moser, K. L., Comeau, M. E., Williams, A. H., Pajewski, N. M., Chung, S. A., Graham, R. R., Zidovetzki, R., Kelly, J. A., et al. (2011). A Comprehensive Analysis of Shared Loci between Systemic Lupus Erythematosus (SLE) and Sixteen Autoimmune Diseases Reveals Limited Genetic Overlap (Analysis of Shared Autoimmune Variants). *PLoS Genetics*, 7 (12): e1002406. doi: 10.1371/journal.pgen.1002406.

Ricano-Ponce, I. & Wijmenga, C. (2013). Mapping of immune-mediated disease genes. *Annu Rev Genomics Hum Genet*, 14: 325-53. doi: 10.1146/annurev-genom-091212-153450.

Richard-Miceli, C. & Criswell, L. A. (2012). Emerging patterns of genetic overlap across autoimmune disorders. *Genome Med*, 4 (1): 6. doi: 10.1186/gm305.

Ronald, J., Akey, J. M., Whittle, J., Smith, E. N., Yvert, G. & Kruglyak, L. (2005). Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Res*, 15 (2): 284-91. doi: 10.1101/gr.2850605.

Roux, K. H. (2009). Optimization and troubleshooting in PCR. *Cold Spring Harb Protoc*, 2009 (4): pdb ip66. doi: 10.1101/pdb.ip66.

Santori, F. R. (2015). The immune system as a self-centered network of lymphocytes. *Immunol Lett*, 166 (2): 109-16. doi: 10.1016/j.imlet.2015.06.002.

Sener, A. G. (2015). Autoantibodies in autoimmune liver diseases. *APMIS*, 123 (11): 915-9. doi: 10.1111/apm.12442.

Senolt, L., Vencovsky, J., Pavelka, K., Ospelt, C. & Gay, S. (2009). Prospective new biological therapies for rheumatoid arthritis. *Autoimmun Rev*, 9 (2): 102-7. doi: 10.1016/j.autrev.2009.03.010.

Skog, O., Korsgren, S., Wiberg, A., Danielsson, A., Edwin, B., Buanes, T., Krogvold, L., Korsgren, O. & Dahl-Jorgensen, K. (2015). Expression of human leukocyte antigen class I in endocrine and exocrine pancreatic tissue at onset of type 1 diabetes. *Am J Pathol*, 185 (1): 129-38. doi: 10.1016/j.ajpath.2014.09.004.

Slatkin, M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet*, 9 (6): 477-85. doi: 10.1038/nrg2361.

St Clair, E. W. (2009). Novel targeted therapies for autoimmunity. *Curr Opin Immunol*, 21 (6): 648-57. doi: 10.1016/j.coi.2009.09.008.

Su, W., Du, L., Liu, S., Deng, J., Cao, Q., Yuan, G., Kijlstra, A. & Yang, P. (2018). ERAP1/ERAP2 and RUNX3 polymorphisms are not associated with ankylosing spondylitis susceptibility in Chinese Han. *Clin Exp Immunol*. doi: 10.1111/cei.13121.

Tomiuk, S. & Hofmann, K. (2001). Microarray probe selection strategies. *Brief Bioinform*, 2 (4): 329-40.

Trynka, G., Hunt, K. A., Bockett, N. A., Romanos, J., Mistry, V., Szperl, A., Bakker, S. F., Bardella, M. T., Bhaw-Rosun, L., Castillejo, G., et al. (2011). Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet*, 43 (12): 1193-201. doi: 10.1038/ng.998.

Tsui, F. W., Haroon, N., Reveille, J. D., Rahman, P., Chiu, B., Tsui, H. W. & Inman, R. D. (2010). Association of an ERAP1 ERAP2 haplotype with familial ankylosing spondylitis. *Ann Rheum Dis*, 69 (4): 733-6. doi: 10.1136/ard.2008.103804.

UCSC Genome Browser. (2018a). *Human BLAT Search*. Available at: http://genome.ucsc.edu/cgi-bin/hgBlat (accessed: 10.05.2018).

UCSC Genome Browser. (2018b). *UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assebly*. Available at: http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr5%3A96875939-96919716&hgsid=668233367_9oIAPkmbmMTuicyU1OYiZnDNQ4mH (accessed: 10.05.2018).

UCSC Genome Browser. (2018c). *UCSC in silico PCR*. Available at: http://genome.ucsc.edu/cgi-bin/hgPcr (accessed: 10.05.2018).

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M. & Rozen, S. G. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Research*, 40 (15). doi: ARTN e11510.1093/nar/gks596.

Vanhille, D. L., Hill, L. D., Hilliard, D. D., Lee, E. D., Teves, M. E., Srinivas, S., Kusanovic, J. P., Gomez, R., Stratikos, E., Elovitz, M. A., et al. (2013). A Novel ERAP2 Haplotype Structure in a Chilean Population: Implications for ERAP2 Protein Expression and Preeclampsia Risk. *Mol Genet Genomic Med*, 1 (2): 98-107. doi: 10.1002/mgg3.13.

Viken, M. K., Sollid, H. D., Joner, G., Dahl-Jorgensen, K., Ronningen, K. S., Undlien, D. E., Flato, B., Selvaag, A. M., Forre, O., Kvien, T. K., et al. (2007). Polymorphisms in the cathepsin L2 (CTSL2) gene show association with type 1 diabetes and early-onset myasthenia gravis. *Hum Immunol*, 68 (9): 748-55. doi: 10.1016/j.humimm.2007.05.009.

Viken, M. K. (2008). *Genetic predisposition to autoimmune diseases with particular focus on type 1 diabetes*. Norway: University og Oslo.

Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. (2012). Five years of GWAS discovery. *Am J Hum Genet*, 90 (1): 7-24. doi: 10.1016/j.ajhg.2011.11.029.

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.

Vitulano, C., Tedeschi, V., Paladini, F., Sorrentino, R. & Fiorillo, M. T. (2017). The interplay between HLA-B27 and ERAP1/ERAP2 aminopeptidases: from anti-viral protection to spondyloarthritis. *Clin Exp Immunol*, 190 (3): 281-290. doi: 10.1111/cei.13020.

Wahid, F., Shehzad, A., Khan, T. & Kim, Y. Y. (2010). MicroRNAs: Synthesis, mechanism, function, and recent clinical trials. *Biochimica Et Biophysica Acta-Molecular Cell Research*, 1803 (11): 1231-1243. doi: 10.1016/j.bbamcr.2010.06.013.

Wang, D., Zavadil, J., Martin, L., Parisi, F., Friedman, E., Levy, D., Harding, H., Ron, D. & Gardner, L. B. (2011). Inhibition of nonsense-mediated RNA decay by the tumor microenvironment promotes tumorigenesis. *Mol Cell Biol*, 31 (17): 3670-80. doi: 10.1128/MCB.05704-11.

Williamson, K. D. & Chapman, R. W. (2015). Primary sclerosing cholangitis: a clinical update. *Br Med Bull*, 114 (1): 53-64. doi: 10.1093/bmb/ldv019.

Wong, J. J., Ritchie, W., Ebner, O. A., Selbach, M., Wong, J. W., Huang, Y., Gao, D., Pinello, N., Gonzalez, M., Baidya, K., et al. (2013). Orchestrated intron retention regulates normal granulocyte differentiation. *Cell*, 154 (3): 583-95. doi: 10.1016/j.cell.2013.06.052.

Wong, J. J., Au, A. Y., Ritchie, W. & Rasko, J. E. (2016). Intron retention in mRNA: No longer nonsense: Known and putative roles of intron retention in normal and disease biology. *Bioessays*, 38 (1): 41-9. doi: 10.1002/bies.201500117.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S. & Madden, T. L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, 13: 134. doi: 10.1186/1471-2105-13-134.

Zhang, F. & Lupski, J. R. (2015). Non-coding genetic variants in human disease. *Hum Mol Genet*, 24 (R1): R102-10. doi: 10.1093/hmg/ddv259.

8. Appendix

Section A

<u>Biological material – samples and corresponding genotypes</u>

Table S1. Overview of rs27302 and rs2248374 genotypes in the 42 human thymus tissue samples.

Thymus sample	rs27302	rs2248374
T01	A/G	A/G
T02	A/A	G/G
T05	G/G	A/A
T06	A/A	G/G
T09	A/G	A/G
T10	A/A	A/G
T11	A/G	A/G
T12	A/G	A/G
T13	A/G	A/G
T15	A/A	G/G
T16	A/A	A/G
T18	A/G	A/A
T19	A/G	A/G
T20	A/A	G/G
T22	A/G	A/G
T23	A/G	A/G
T24	G/G	A/A
T25	A/G	A/G
T27	G/G	A/A
T28	A/G	A/A
T30	A/A	A/A
T31	A/A	G/G
T32	A/G	A/G
T33	A/G	A/G
T34	G/G	A/A
T36	A/A	G/G
T37	A/A	G/G
T38	A/A	G/G
T39	A/G	A/A
T40	A/G	A/A
T43	A/A	G/G
T44	A/G	A/G
T45	A/G	A/G
T47	A/A	G/G
T48	A/G	A/A
T51	A/G	A/G
T52	A/A	G/G
T53	A/G	A/G
T55	A/G	A/G
T56	A/A	G/G
T57	A/A	G/G

T58	G/G	A/A

Table S2. Overview of ethnicity and genotypes in rs27302 and rs2248374 in the 18 LCL cell pellets obtained from Coriell Cell Repositories.

Sample	Ethnicity	rs27302	rs2248374
GM07000	CEPH Utah	G/G	A/A
GM11832	CEPH Utah	G/G	A/A
GM12043	CEPH Utah	G/G	A/A
GM18504	NHGRI Yoruba	A/G	A/A
GM18505	NHGRI Yoruba	A/G	A/A
GM18507	NHGRI Yoruba	A/A	G/G
GM18516	NHGRI Yoruba	G/G	G/G
GM18645	NHGRI Han Chinese	A/A	G/G
GM18646	NHGRI Han Chinese	A/A	A/A
GM18870	NHGRI Yoruba	A/G	A/G
GM18923	NHGRI Yoruba	A/G	G/G
GM18933	NHGRI Yoruba	G/G	G/G
GM19084	NHGRI Japanese	A/A	G/G
GM19099	NHGRI Yoruba	G/G	G/G
GM19201	NHGRI Yoruba	A/G	A/G
GM19240	NHGRI Yoruba	A/G	G/G
GM19792	NHGRI Mexican	A/A	A/A
GM19916	NHGRI African	A/A	A/A

Section B

Reagents, kits and consumables.

Table S3. Reagents, suppliers and catalog numbers.

Reagents	Supplier	Catalog number
10x TAE	Thermo Fisher Scientific, Waltham, MA, USA	15558-042
10x TBS	Bio-Rad Laboratories, Hercules, CA, USA	170-6435
10x Tris/Glycine/SDS Buffer	Bio-Rad Laboratories, Hercules, CA, USA	161-0732
100x Halt™ Protease and Phosphatase inhibitor cocktail	Thermo Fisher Scientific, Waltham, MA, USA	1861281
Agarose	Sigma-Aldrich®, Darmstadt, Germany	A9539-100G
Bovine Serum Albumin	Sigma-Aldrich®, Darmstadt, Germany	A7906-100G
Dulbecco´s Phosphate Buffered Saline	Thermo Fisher Scientific, Waltham, MA, USA	14190-094
Ethanol absolute	Avantor®, Center Valley, PA, USA	20821.310
GeneRuler™ 1kb DNA Ladder	Thermo Fisher Scientific, Waltham, MA, USA	SM0311
GeneRuler™ 50bp DNA Ladder	Thermo Fisher Scientific, Waltham, MA, USA	SM0371
Glycine	Bio-Rad Laboratories, Hercules, CA, USA	161-0718
Gel Loading Buffer II	Thermo Fisher Scientific, Waltham, MA, USA	8546G
GelRed™ Nucleic acid Gel Stain 10,000X in Water	Biotium Inc., Freemont, CA, USA	41001
Methanol	Merck, Darmstadt, Germany	1.06009.2511
Ponceau S Solution	Sigma-Aldrich®, Darmstadt, Germany	P7170-1L
Restore™ Western Blot Stripping Buffer	Thermo Fisher Scientific, Waltham, MA, USA	21059
RIPA buffer	Sigma-Aldrich®, Darmstadt, Germany	R0278-500ML
SeaKem® LE Agarose	Lonza, Basel, Switzerland	50004
TRIS-EDTA buffer solution	Sigma-Aldrich®, Darmstadt, Germany	93283
Trizma® base	Sigma-Aldrich®, Darmstadt, Germany	T1503-1KG
Tween®20	Sigma-Aldrich®, Darmstadt, Germany	P1379

Table S4. Kits, suppliers and catalog numbers.

Kit	Supplier	Catalog number
ECL^{TM} Prime Western Blotting System	GE Healthcare Life Sciences, Pittsburg, PA, USA	RPN2232
GenDX LongRange PCR Kit	GenDx, Utrecht, Netherlands	5342252
Pierce^{TM} Protein Assay Kit	Thermo Fisher Scientific, Waltham, MA, USA	23225
RNeasy® Plus Mini Kit	QIAGEN, Hilden, Germany	74134
SuperScript® III First-Strand	Thermo Fisher Scientific, Waltham, MA, USA	18080-051
QIAamp® DNA Micro Kit	QIAGEN, Hilden, Germany	56304

Table S5. Antibodies and protein ladder, suppliers and catalog numbers.

Antibody and protein ladder	Supplier	Catalog number
ß-Actin Mouse mAb	Cell Signaling Technology, Danvers, MA, USA	8H10D10
Anti-ERAP2 antibody	Abcam®, Cambridge, UK	ab69037
Mouse Immunoglobulins	Agilent, Santa Clara, CA, USA	P0260
Presicion Plus Protein^{TM} Dual Color Standards	Bio-Rad Laboratories, Hercules, CA, USA	161-0374

Table S6. Consumables, suppliers and catalog numbers.

Consumables	Supplier	Catalog number
0.2ml Non-skirted 96-well PCR Plate	Thermo Fisher Scientific, Waltham, MA, USA	AB0600
Nitrocellulose Membrane, 0.2µm	Bio-Rad Laboratories, Hercules, CA, USA	162-0112
Nunclon^{TM} Delta Surface	Thermo Fisher Scientific, Waltham, MA, USA	191093
Mini-protean TGX^{TM} Gels, 10%, 10 well comb, 50µl	Bio-Rad Laboratories, Hercules, CA, USA	456-1034
Thin Blot Paper, Filter paper	Bio-Rad Laboratories, Hercules, CA, USA	1620118
TissueRuptor® Disposable probes	QIAGEN, Hilden, Germany	NC9629296

Table S7. Instruments and suppliers.

Instrument	Supplier
2720 Thermal Cycler	Thermo Fisher Scientific, Waltham, MA, USA
ImageQuant^{TM} LAS 4000	GE Healthcare Life Sciences, Pittsburg, PA, USA
Nanodrop® ND-1000	Thermo Fisher Scientific, Waltham, MA, USA
Veriti 96 Well Thermal Cycler	Thermo Fisher Scientific, Waltham, MA, USA
VERSAmax Microplate Reader	Molecular Devices, San Jose, CA, USA
TissueRuptor ®	QIAGEN, Hilden, Germany

Table S8. Software and suppliers.

Software	Supplier
GraphPad Prism 7	GraphPad Software, Inc., La Jolla, CA, USA
Haploview 4.2	Broad Institute, Cambridge, MA, USA
ImageQuant^{TM} TL 1D v8.1	GE Healthcare Life Sciences, Pittsburg, PA, USA
ImageQuant^{TM} TL Toolbox v8.1	GE Healthcare Life Sciences, Pittsburg, PA, USA
Nanodrop 3.0.0	Thermo Fisher Scientific, Waltham, MA, USA
RStudio Version 1.1.423 –© 2009-2018	RStudio, Inc., Boston, MA, USA
SoftMax Pro 6.4	Molecular Devices, San Jose, CA, USA
Microsoft® Excel	Microsoft, Redmond, WA, USA

Section C

Primers

Table S9. Overview original primers for production of amplicons prior to sequencing the *ERAP2* and rs27302 genomic region.

Name	Target region for amplicon (GRCh38)	Direction	Sequence 5´-3´
ERAP2-amp1	Chr5 96874578 - 96881042	F	CCTGTGGTGCCCCTCTCACT
		R	AGAGCACTTGCCCCGAGATG
ERAP2-amp2	Chr5 96880854 - 96885382	F	GGCAAGGTAATTGGCCAGTGTT
		R	TCCCCAACAACCGTCTCCAT
ERAP2-amp3	Chr5 96885122 - 96890020	F	CCATTTGCCTCCGAACCATC
		R	TCCAACCTCAGTGCCCAACA
ERAP2-amp4	Chr5 96889761 96895770	F	CCTCCCATCGTGCTGCTCTT
		R	GGGAAGCCAGCAGACCCTTT
ERAP2-amp5	Chr5 96895506 - 96900422	F	CACAGAACTGGATGAAACATTGAAGG
		R	GCAGTGGGGTTGTGGGAAAG
ERAP2-amp6	Chr5 96900357 - 96906464	F	AGGGGGACAATGCTGTTGCT
		R	CAGACGCCATCTCTGCCAAA
ERAP2-amp7	Chr5 96906274 - 96912179	F	CAAGGGCTTTGCTCTTCTTCTTCA
		R	TCTCACTGTCGCCCAAGCTG
ERAP2-amp8	Chr5 96911781 - 96918136	F	GAGGATTGTGCAGGAGGCTGA
		R	TTCCCCCACCACTCCATCAT
ERAP2-amp9	Chr5 96917962 - 96920213	F	TCAGACATGACTGCCTGCATGA
		R	CTCACCTCAGCCTCCCGTGT
ERAP2-amp10	Chr5 97036876 - 97040690	F	AAGGGCACCTTTTTGGAGTT
		R	CCCTTTGGTGGTGAGTGTCT

F = Forward primer, R = Reverse primer

Table S10. Primers covering the *ERAP2* exon 10 splice variant.

Name	Location	Direction	Sequence 5’-3´
Exon 10 splice variant	*ERAP2* exon 10-exon 11	F	CATTCGGATCCCAAGATGAC
		R	GGAGTGAACACCCGTCTTGT

F = Forward primer, R = Reverse primer

Table S11. New primers obtained to cover Amp1, Amp2, Amp7 and Amp10.

Name	Target region for amplification	Direction	Sequence 5´-3´
New_ERAP2-Amp1R	Chr5 97036583 - 97,041,753	R	TGGCCTTCCTCAACCAATCC
New_ERAP2-Amp3	Chr5 96885122 - 96890020	F	CCATTTGCCTCCGAACCATC
		R	TCCAACCTCAGTGCCCAACA
New_ERAP2-Amp7	Chr5 96906025 - 96912179	F	TCAGCCTCGCAAATTGCTCA
		R	TCTCACTGTCGCCCAAGCTG
New_ERAP2-Amp10	Chr5 97036583 - 97041753	F	TCCCTCCCTCCCTCTCTGCT
		R	TTTGAGCCCAGGAGGTCGAG

F = Forward primer, R = Reverse primer

Section D

Quantified protein concentrations using the Pierce BCA Protein Assay Kit (Thermo Fisher Scientific).

Table S12. Protein concentrations in the nine thymic samples used in the western blot analysis.

Sample	Protein concentration (µg/µl)
T01	2.648
T10	2.289
T18	1.662
T27	1.508
T30	0.789
T40	1.939
T56	1.508
T57	0.764
T58	1.311

Table S13. Protein concentrations in the LCL samples used in the rs27302-A/A and rs27302-G/G western blot analyses.

Sample	Protein concentration (µg/µl)
GM07000	0.179
GM11832	0.202
GM12043	0.209
GM18507	0.207
GM18516	0.151
GM18645	0.190
GM18646	0.190
GM18933	0.204
GM19084	0.152
GM19099	0.138
GM19792	0.129
GM19916	0.141

Table S14. Protein concentrations in the LCL samples used in the rs27302-A/G western blot analysis.

Sample	Protein concentration (µg/µl)
GM12043	0.205
GM18504	0.187
GM18505	0.205
GM18870	0.207
GM18923	0.117
GM19201	0.144
GM19240	0.173

Section E

Solutions and gels used during western blotting and gel electrophoresis.

Western blotting

1x TBS-T:

- 100ml 10x TBS buffer
- 900ml MQ-H_2O
- 1ml Tween® 20

Blotting buffer:

- 3g Trizma base
- 14.4g Glycine
- 900ml MQ-H_2O
- 100ml Methanol

Running buffer:

- 100ml 10x Tris/Glycine/SDS
- 900ml MQ-H_2O

5% BSA:

- 2.5g Bovine Serum Albumin
- 50ml TBS-T

Gel electrophoresis

1x TAE buffer:

- 100ml 10x TAE Buffer
- 900ml MQ-H_2O

0.7% agarose gel:

- 0.7g Agarose
- 100ml 1xTAE buffer
- 10µl GelRed

1% agarose gel:

- 1g Agarose
- 100ml 1xTAE buffer
- 10µl GelRed

2% agarose gel:

- 2g Agarose
- 100ml 1xTAE buffer
- 10µl GelRed

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

Postboks 5003
NO-1432 Ås
Norway