UNIVERSITY OF SOUTHAMPTON

FACULTY OF MEDICINE

Cancer Sciences Academic Unit

Developing long read sequencing and its application and understanding the role of ERAP1 in cervical carcinoma

By

Michaela Eleni Christodoulaki

30522544

Doctor of Philosophy PhD thesis

May 2023

ABSTRACT

Long read sequencing of components of the antigen processing and presentation (APP) pathway, specifically endoplasmic reticulum aminopeptidase 1 (ERAP1), could provide a new tool for identifying those women that are at high risk of poor cervical cancer prognosis. ERAP1 edits the peptide repertoire presented to cytotoxic T lymphocytes (CTLs) through N-terminal trimming of peptide precursors to the optimal length for stable MHC I binding prior to presentation. Single nucleotide polymorphisms (SNPs) in ERAP1 exist in multiple combinations that form distinct haplotypes that when expressed as allotypes, alter ERAP1 trimming function. Individual ERAP1 SNPs have been associated with increased cervical cancer risk in GWAS studies, however a cause and effect relationship between ERAP1 allotypes and cervical cancer progression has not been established. In this study, the identification of ERAP1 allotypes from HeLa and 293T cells using long read sequencing by MinION enabled the establishment of a methodological pipeline, including optimisation of each step of the protocol, identification of the limitations of this technology and the development of a robust bioinformatics analysis pipeline. Subsequently, ERAP1 allotypes from a total of 81 patients at varying stages of cervical carcinoma were identified using long read sequencing. In this cohort, a total of 14 ERAP1 allotypes were identified and those were found in 28 distinct ERAP1 allotype combinations affecting the enzyme's trimming function given that both chromosomal copies are codominantly expressed. Functional analysis showed that the majority of patients with a high CD8+/TIL number had ERAP1 allotypes with efficient trimming ability and better overall prognosis. ERAP1 allotype identification and CD8+/TILs could be a useful tool for identifying those women at high risk of poor prognosis so that women can receive early treatment.

Contents

List of figures and tables

RESEARCH THESIS: DECLARATION OF AUTHORSHIP

MICHAELA ELENI CHRISTODOULAKI

Developing long read sequencing and its application and understanding the role of ERAP1 in cervical carcinoma

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;

2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;

3. Where I have consulted the published work of others, this is always clearly attributed;

4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;

5. I have acknowledged all main sources of help;

6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

Signature:

Date:

Acknowledgements

I would like to thank those who have helped me throughout the duration of my PhD. First, I would like to thank my supervisors, Professor Edd James, Dr Emma Reeves and Professor Tim Elliott for their continuous support and help, especially for their comments during the time that I was writing my thesis. Many thanks to Dr Jane Gibson for her kind assistance in developing a bioinformatics pipeline for the analysis of my data and to Professor Marco de Bruyn and the rest of the researchers at the University of Groningen for providing the patient samples that were used for this study.

I would also like to thank my fellow colleagues, Dr Denise Boulanger, Dr Elisa Lori, Dr David Arcia-Anaya and Dr Luke Hill for helping me out in the lab, especially when panic mode kicked in. I would particularly like to thank Nasia Kontouli, who I can now call a good friend, for listening to me when I needed to talk and offering advice when my brain was just not working.

Last but not least, I would like to thank my family for always being there for me, especially at times that I needed them the most.

List of abbreviations and definitions

A-LAP	**Adipocyte-derived leucine aminopeptidase**
APC	Antigen presenting cell
ARTS1	Aminopeptidase regulator of TNFR type I shedding
AS	Ankylosing spondylitis
AS-US	Atypical squamous cells of unknown significance
BD	Behçet's disease
CIN	Cervical intraepithelial neoplasia
CLIP	Class II-associated li peptide
CRT	Calreticulin
CSCC	Cervical squamous cell carcinoma
CTL	Cytotoxic T lymphocyte
DC	Dendritic cell
DRiPs	Defective ribosomal products
ER	Endoplasmic reticulum
ERAAP	Endoplasmic reticulum aminopeptidase associated with antigen processing
ERAP1	Endoplasmic reticulum aminopeptidase 1
GWAS	Genome wide association studies
HEK	Human embyronic kidney
HLA	Human leucocyte antigen
HPV	Human Papillomavirus
HSIL	High grade squamous intraepithelial lesions
Hsp90a	Heat shock protein 90 a
LAP	Leucine aminopeptidase
LMP2	Low molecular mass peptide
LSIL	Low grade squamous intraepithelial lesions
MHC	Major Histocompatibility complex
MIIC	MHC II compartment
NK	Natural killer
OPSCC	Oropharyngeal squamous cell carcinoma
PD-1	programmed death-1 receptor

PDI	Protein disulfide isomerase
PILSAP	Puromycin-insensitive leucyl-specific aminopeptidase
P-LAP	Placental leucine aminopeptidase
PLC	Peptide loading complex
pMHC I	Peptide:MHC I complex
SNP	Single nucleotide polymorphism
TAPBPR	TAP-binding protein related
TCR	T cell receptor
TGN	Trans golgi network
TNFR1	Tumour necrosis factor receptor 1
TPPII	Tripeptidyl peptidase II
TRiC	TCP-1 ring complex
UGGT	UDP-glycosyl glucose transferase
WT	Wild type

1 Introduction

1.1 The immune system

The body is protected from potentially harmful infectious agents and the damage they cause by different classes of effector cells and molecules that together constitute the human immune system, an essential part of the body's defence mechanism [1]. In cancer, the immune system plays a fundamental role in the suppression of tumour growth by eliminating cancer cells or inhibiting outgrowth as well as in the progression and maintenance of cancer by selecting those tumour cells that could survive in a healthy host [2]. The innate and the adaptive immune response are responsible for immunosurveillance to eliminate invading pathogens and eradicate tumour cells. Innate immunity is the first line of defence against pathogens, however it does not lead to long lasting immunity. Adaptive immunity is mediated by T- and B-lymphocytes which lead to the generation of a cell-mediated or an antibody-mediated response, respectively. Adaptive immune responses are antigen specific and can be generated against antigens that have been encountered before due to immunological memory.

1.2 T-lymphocytes

T cells express receptors on their surface called T cell receptors (TCR) which recognise an antigenic peptide when it is bound to a cell surface glycoprotein, major histocompatibility complex I (MHC I). Co-receptors function together with TCR; CD4 and CD8 are the co-receptors on T helper cells, Tregs and cytotoxic T lymphocytes (CTLs), respectively. CD4+ T cells recognise antigens bound to MHC class II (MHC II) and CD8+ T cells recognise antigens bound to MHC class I (MHC I). Co-stimulatory receptors, most prominently CD28, are key to the activation of a naïve T cell into an effector cell. There are also co-inhibitory receptors such as programmed death-1 (PD-1) receptor which can inhibit an immune response from being initiated or downregulate it following initiation. In recent years, therapies address co-regulatory receptors, such as PD-1, for the treatment of cancer as well as autoimmune disease [3-5].

1.3 Antigen processing and presentation

T cells are dependent on presentation of peptide antigens by MHC molecules for their activation. MHC are polygenic with a high degree of allelic polymorphism, and are the most polymorphic genes found in humans. The MHC gene region is divided in three sub-regions, MHC I, II and III. In humans, MHC is located on chromosome 6 and in mice it is located on chromosome 17. MHC I encodes the human leukocyte antigens (HLA)-A, -B and –C in humans and histocompatibility 2 (H-2)-D, -K, -L in mice. MHC II encodes HLA-DP, -DQ, -DR and –DO in humans and H-2-A and –E in mice. The MHC III region was found to be the most gene-dense region of the human genome (>14% of the sequence is coding) containing several genes that encode proteins which play an important role in the innate immune system, including members of the complement fixation cascade Bf, C2 and C4 [6, 7]. Antigen processing and presentation occurs through three distinct pathways: i) The endogenous pathway involves formation of peptide:MHC I (pMHC I) from peptide fragments originating in the cytosol, with the complexes being presented at the cell surface to CTLs; ii) the exogenous antigen presentation pathway involves processing of proteins entering the cell through endocytosis in acidified endocytic vesicles, with peptides loaded on to MHC II molecules for presentation to CD4+ T cells; and iii) cross-presentation pathway is one of several alternative pathways involving formation of pMHC I complexes from peptides derived from extracellular proteins which entered the cell through endocytosis or phagocytosis, or pMHC II complexes from cytosolic antigens [8].

1.3.1 Exogenous antigen presentation pathway

MHC II molecules are expressed by professional antigen presenting cells (APCs), most notably dendritic cells (DCs), macrophages and B cells, as well as other cells such as fibroblasts and endothelial cells when stimulated by IFNγ [8]. MHC II molecules are loaded with peptides that have been generated from endocytosis of extracellular proteins. The following endocytosis methods including micropinocytosis by immature DCs, clathrin-mediated endocytosis exhibited by B cells only and

phagocytosis by phagocytes, promote efficient internalisation of antigens into endosomal vesicles [8]. The exogenous antigen presentation pathway consists initially of early endosomes with a neutral pH, followed by acidification of vesicles which activates degradation of proteins into peptide fragments by proteases. These vesicles will later fuse with other vesicles containing MHC II molecules [9]. Peptides loaded onto MHC II are then presented to CD4+ T cells at the cell surface [8, 10].

The MHC II molecule is composed of two transmembrane glycoprotein chains, α (light chain, 34 kDa) and β (heavy chain, 29 kDa), both of which are encoded within the MHC locus (Figure 1.1) [11]. Each chain consists of two domains (α1, α2 and β1, β2) together forming a four-domain structure similar to MHC I. In MHC II, the two domains forming the peptide binding cleft are the α1 and the β1, and since these two domains belong to a different chain, they are not joined covalently as in the case of MHC I (Figure 1.1). MHC II differs from MHC I in the peptide-binding cleft, the ends of which are more open in MHC II and therefore the ends of the peptide binding to MHC II are not buried within it as in the case of MHC I. The optimal length of peptides binding to MHC II is 15-20 amino acids [12]. As the MHC II binding cleft accommodates a greater variety of side chains than MHC I, defining anchor residues and predicting the nature of the peptide binding to a specific MHC II is less accurate.

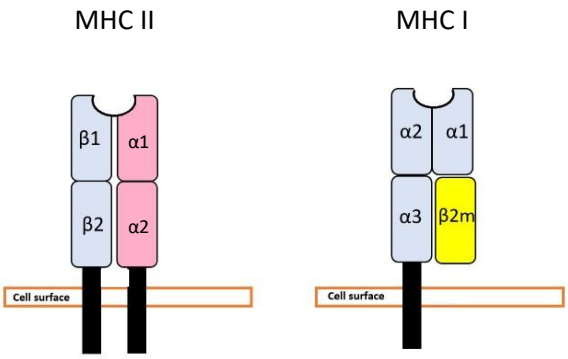


Figure 1.1. Structure of the MHC molecules

The MHC II molecule is composed of two transmembrane glycoprotein chains, α (light chain, 34kDa, shown in pink) and β (heavy chain, 29kDa, shown in blue). Each chain consists of two domains (α1, α2, β1, β2) forming a four-domain structure. The peptide binding cleft is formed by domains α1 and β1. The MHC I molecule is a heterodimer composed of two polypeptide chains: a membrane-spanning α chain (43kDa) bound non-covalently to β_2-microglobulin (12kDa). The α chain folds into three domains: α1, α2, and α3. The peptide binding cleft is formed by domains α1 and α2.

MHC class II chains assemble in the ER and are stabilised by association with a chaperone called invariant (Ii) chain or CD74 forming nonameric structures (Figure 1.2). A nonamer consists of three MHC II α chains, three MHC II β chains and three Ii chains [13]. The binding of antigens on the MHC II is prevented by the occupation of the MHC II binding groove by a portion of the Ii chain known as class II-associated Ii peptide (CLIP), spanning roughly 20 residues [13] [14]. The MHC II-Ii chain complex leaves the ER and is oriented towards the endocytic pathway by the di-leucine motifs found on the Ii chain. This can be achieved either by direct targeting from the Trans Golgi Network (TGN) or indirectly by endocytosis from the plasma membrane and then the complex is found in the late endosome (MVB) where antigen will be loaded on to MHC II [15-17]. Due to acidic pH in endosomes, the proteases cathepsin S and L are activated and digest Ii chain leaving only CLIP in the MHC II binding groove [18]. Next, exchange of CLIP for an antigen that has been processed from an extracellular protein in the

endosomal pathway is mediated by HLA-DM [19, 20]. The function of DM is modulated by HLA-DO in specialised cells such as B cells and dendiritc cells and has a role in inhibiting HLA-DM function [21]. Research in HLA-DO$^{-/-}$ mice revealed decreased peptide repertoire diversity and activation of CD4+ T cells upon immunisation with wild type (WT) APCs, indicating that HLA-DO expression is required for presentation of self-antigens to generate peripheral tolerance, hence preventing development of autoimmune disease [22]. Following stable binding of antigen on the MHC II binding groove, the complex can be expressed on the cell surface and presented to CD4+ T cells (Figure 1.2).

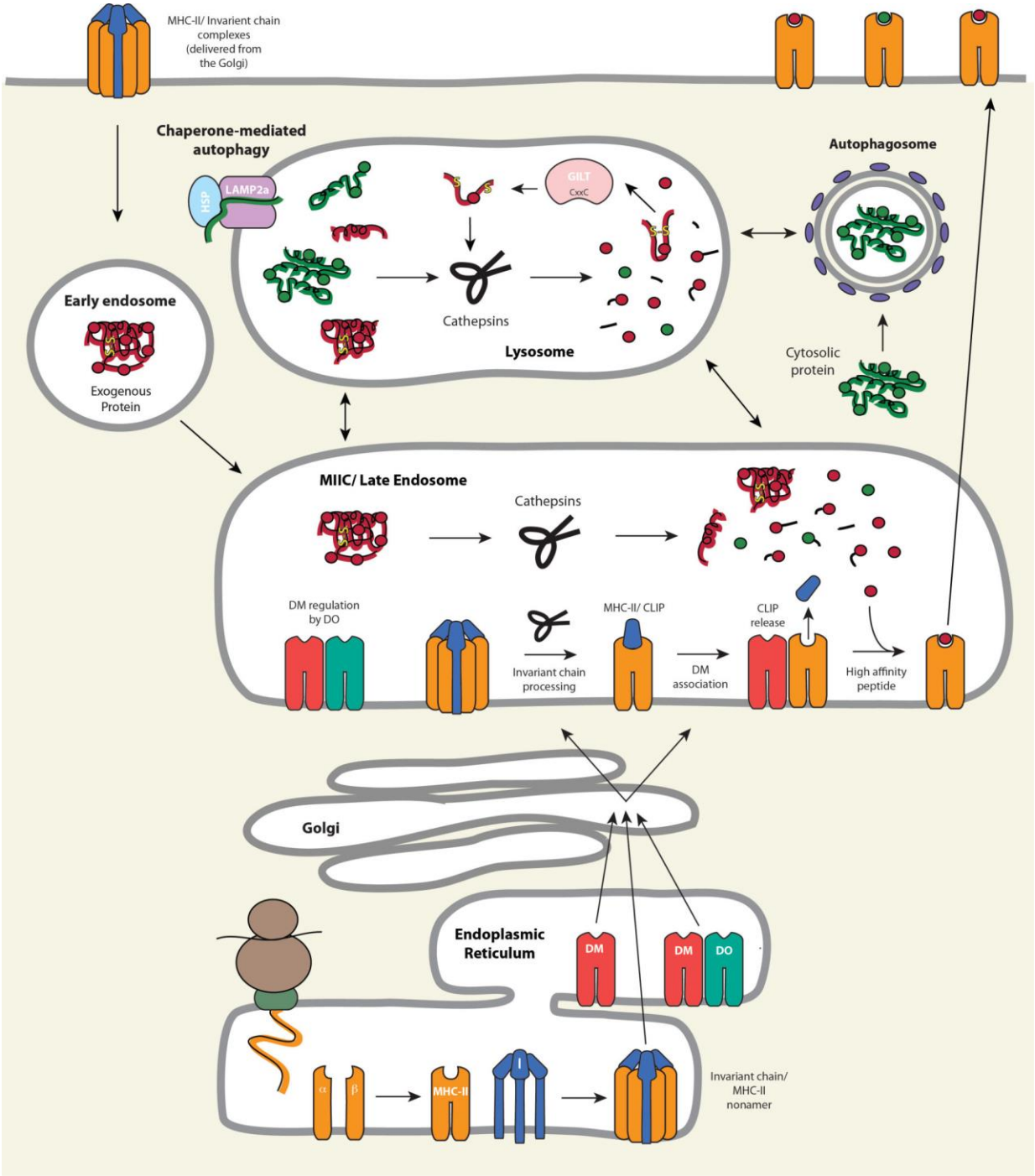


Figure 1.2. The exogenous antigen presentation pathway

Following assembly of MHC II, the molecule binds invariant chain to form a nonamer, a trimer of trimers. The complexes are transferred to endosomes via Trans Golgi Network (TGN) or by recycling from the cell surface. Inside endosomes, digestion of invariant chain leads to only CLIP remaining in the binding groove of MHC II which prevents binding of low affinity peptides. CLIP is removed from the groove by HLA-DM which is negatively regulated by HLA-DO. Antigens are delivered to endosomes through endocytosis, pinocytosis, autophagy,

phagocytosis and they are digested by cathepsins and loaded on to MHC II using HLA-DM. MHC II-peptide complexes are then transferred to the cell surface for presentation to CD4+ T cells [8].

1.3.2 Cross-presentation of exogenous antigens

The predominant cross-presentation pathway involves internalisation of exogenous antigens by endocytosis or phagocytosis, followed by processing and loading on to MHC I for presentation to CTLs [8]. Cross-presentation is a crucial pathway of antigen processing and presentation in the induction of adaptive immune responses against tumours and viral infections [23, 24]. There are two main pathways of cross-presentation, the vacuolar pathway and the endosome-to-cytosol pathway [25]. Different subpopulations of DCs are involved in cross-presentation and activation of naïve CTLs depending on the experimental setting [26]. Resident CD8+ DCs were shown to be more efficient at cross-presentation than CD8- DC and among migratory DCs, CD103+ DCs are the most efficient for the cross-presentation in the lymph nodes [26, 27]. In the vacuolar pathway, internalised exogenous antigens are degraded by lysosomal proteases, especially cathepsin S, and antigen-derived peptides are loaded on to the MHC I in the lysosome [28]. Research has shown that macroautophagy in tumour or target cells can enhance their internalisation and MHC I cross-presentation to CTLs [29]. In the endosome-to-cytosol pathway which is thought to be the dominant pathway, internalised antigens are transferred from endosomes to the cytosol where the proteasome degrades them into peptide fragments [25]. For loading on to MHC I, the peptides require transportation into the ER or back into the antigen-containing endosomes, a process undertaken by the transporter associated with antigen processing (TAP) [30].

1.3.3 Endogenous antigen presentation pathway

Peptide antigens stably bind to MHC I and the resulting pMHC I is presented on the surface of cells to CTLs which may initiate an appropriate immune response upon recognition of the pMHC I [31].

Elimination of malignant or transformed cells requires an effective antigen processing and presentation pathway for the N-terminal trimming of tumour antigens that will be presented to CTLs to initiate the appropriate anti-tumour immune response [31]. The MHC I molecule is composed of two polypeptide chains: a membrane spanning α chain (43kDa) which folds into domains α1, α2 and α3, and β_2-microglobulin (12kDa) [32]. The α chain is non-covalently bound to β_2m and the peptide binding cleft is formed by domains α1 and α2 (Figure 1.1). Cotranslational oxidative folding of the MHC I α chain is carried out in the lumen of the endoplasmic reticulum (ER) and is facilitated by the chaperone calnexin and the oxidoreductase ERp57 [1]. The association of MHC I α chain with β_2m results in dissociation of calnexin [1]. The binding of peptide onto MHC I molecules is coordinated by the peptide-loading complex (PLC), which is composed of the heterodimer TAP, the chaperones calreticulin and tapasin, and ERp57 [33]. In the cytosol, the proteasome degrades defective or old proteins into peptides which enter the ER via TAP [8]. Many of the peptides entering the ER have elongated N-termini, as TAP preferentially transports peptides of 11-14 amino acids in length, which are not of optimal length for stable binding to MHC I and presentation to CTLs [34-37]. Endoplasmic reticulum aminopeptidase 1 (ERAP1) is one of the enzymes responsible for the trimming of peptides to the optimal length of 8-11 amino acids, so that a stable pMHC I is generated for transport to the cell surface and presentation to CTLs (Figure 1.3) [34-37].

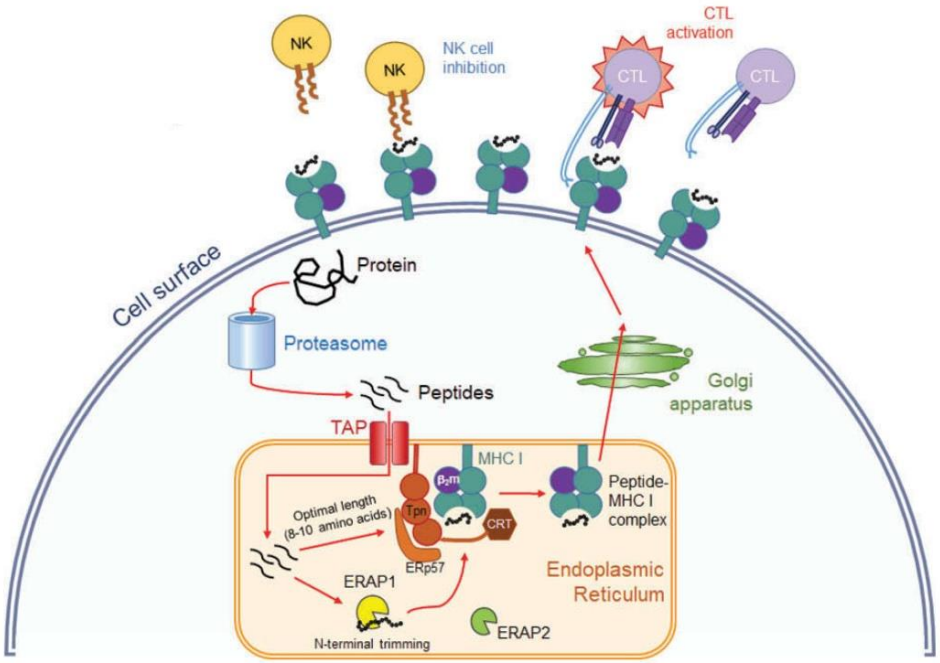


Figure 1.3. The endogenous antigen processing and presentation pathway

In the cytosol, the proteasome or the immunoproteasome degrades proteins into peptides that enter the lumen of the ER via TAP. Binding of peptides onto MHC I is facilitated by the peptide loading complex composed of the heterodimer heavy chain (HC):β2m (MHC I), calreticulin, ERp57, tapasin and TAP. The peptides entering the ER are often not of optimal length and ERAAP in mice or ERAP1 and ERAP2 in humans trim the peptides to the optimal length for the generation of a stable MHC I-peptide complex which is transported to the cell surface, via the Golgi apparatus, to be presented to CTLs which exert their effector function. Taken from *Reeves et al* [38].

1.4 Origin and generation of antigenic peptides for MHC I presentation

The immune system utilises proteolytic pathways for the turnover of intracellular proteins, with the ubiquitin-proteasome system (UPS) being the major non-lysosomal pathway through which 80% of normal and abnormal intracellular proteins are degraded [39, 40]. The proteasome, a multi-catalytic enzyme located in the cytosol, is responsible for the degradation of the majority of intracellular proteins. Proteosomal degradation prevents accumulation of misfolded or incorrectly folded proteins and regulates protein concentration in order to maintain a normal homeostatic cellular environment

[39, 40]. Protein degradation by the proteasome is important for a number of biological functions occurring within the cell such as cell cycle regulation, responses to oxidative stress and generation of antigenic peptides for stable binding on the MHC I [39, 40] .

1.4.1 Generation of antigenic peptides by the proteasome

The degradation of polyubiquitinated proteins is carried out by the proteasome through the UPS [39]. The 26S proteasome comprises one 20S catalytic complex and two 19S regulatory complexes located at either end of the catalytic complex [40]. The 20S catalytic core of the 26S proteasome is composed of four rings stacked one on top of each other, two inner β rings containing seven β subunits and two outer α rings containing seven α subunits (Figure 1.4) [40]. The proteolytic signal involves interaction between the polyubiquitin chain attached to the proteins targeted to the proteasome for degradation, and the S6' ATPase subunit of the base of the 19S complex [41]. Following binding of ATP, possibly to the ATP binding site of a different ATPase of the 19S complex, 19S associates with the 20S proteasome and enables substrate unfolding and entry into the proteolytic core [41]. Once in the core, proteins are hydrolysed into oligopeptides.

Regarding substrate specificity of the proteasome, the 20S core has six catalytically active sites: i) two "chymotrypsin-like" sites that preferentially cleave after hydrophobic amino acids, ii) two "trypsin-like" sites that preferentially cleave after basic amino acids and iii) two "caspase-like" sites that preferentially cleave after acidic amino acids [39]. The proteasome generates 7-15 amino acid long peptides, whereas aminopeptidases in the cytosol or the endoplasmic reticulum have a trimming function of N-terminally extended epitope precursors [42-44]. Treatment with proteasome inhibitors peptide aldehyde and lactacystin led to antigenic peptide generation being blocked, preventing assembly of MHC I heavy and light chains into a stable complex [42]. In contrast, when eight or nine-residue long antigenic peptides are introduced into cells which have been previously been treated with proteasome inhibitors, their presentation is not blocked, enabling binding to MHC I and its

stabilisation [45]. If the antigenic peptides contained even a single C-terminal flanking residue, proteasome inhibitors blocked their presentation [42]. In contrast, proteasome inhibitors did not block the presentation of antigenic peptides that were extended by up to 25 N-terminal flanking residues [42]. In vitro studies showed that the proteasome makes the appropriate cleavage relatively infrequently (less than 20% of the time) [46]. Interestingly, studies have shown that approximately 70% of the proteasome products are too short to bind to the groove in MHC I molecules, approximately 15% are of appropriate length and 15% are too long, however they could be trimmed to the appropriate length for binding on MHC I molecules if trimmed by cytosolic peptidases and the ER aminopeptidases [39, 46, 47]. A subset of NH_2-extended MHC I epitopes (10-16 amino acids) are translocated to the ER because they have higher affinities for the peptide transporter TAP than mature epitopes where they require further trimming by aminopeptidases to an appropriate length for stable MHC I binding [39, 48, 49].

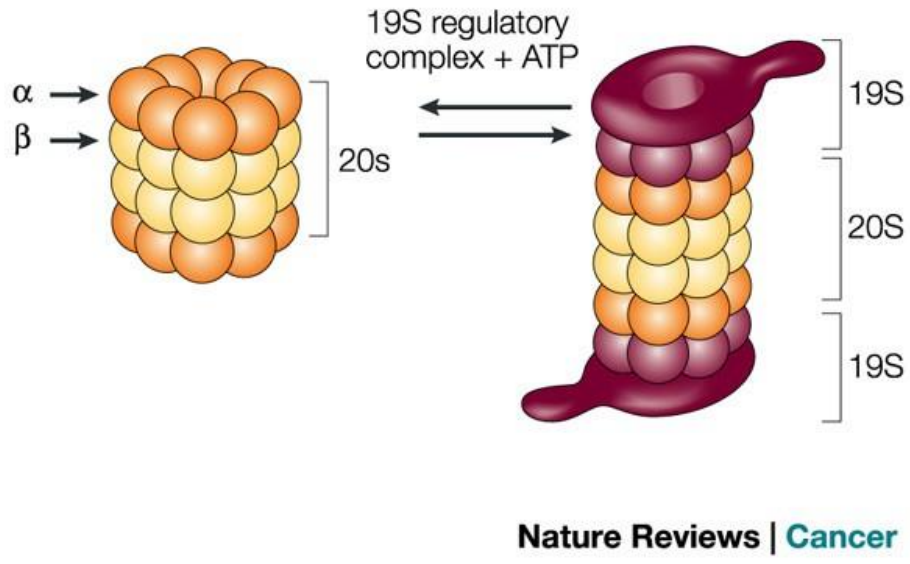


Figure 1.4. The ATP-dependent 26S proteasome

The 26S proteasome consists of a catalytic core (20S) and two regulatory subunits at each end of the 20S core (19S). The 20S catalytic core is composed of four rings stacked one on top of each other; two inner β rings containing seven β subunits and two outer α rings containing seven α subunits. Adapted from Adams 2004 [50].

1.4.2 Generation of antigenic peptides by the immunoproteasome

When cells were stimulated with IFNγ, their proteasomes were shown to contain two catalytic subunits, low molecular mass polypeptide (LMP)-2 and LMP-7, which are homologous to the β subunits of the 26S proteasome and are encoded within the MHC II region of chromosome 6 [51-53]. In the presence of IFNγ, LMP-2, LMP-7 and MECL1 (encoded outside the MHC region) are incorporated in the proteasome, replacing the subunits β1, β5 and β2 respectively [39, 51-54]. This proteasome is referred to as the immunoproteasome because of its enhanced capability to generate antigenic peptides presented to CTLs (Figure 1.5) [40].

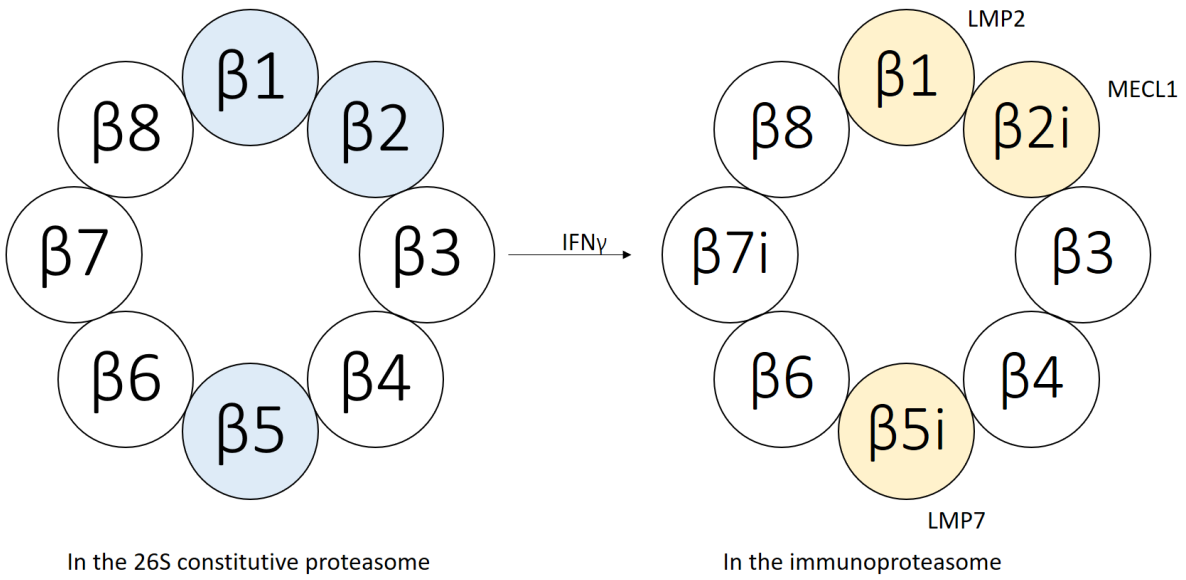


Figure 1.5. Constitutive proteasome and immunoproteasome subunits.

The β ring subunits of the constitutive proteasome and of the immunoproteasome. Upon stimulation of cells with IFNγ, proteasomes were shown to contain three catalytic subunits in the β ring of the 20S core LMP2, LMP7 and MECL1 (encoded outside the MHC region), replacing the 26S proteosomal subunits β1, β5 and β2 respectively. This proteasome is referred to as the immunoproteasome because of its enhanced capability to generate antigenic peptides presented to CTLs [39, 40].

PA28 (11S REG), a ring-shaped multiple subunit complex of molecular weight of 180kDa, binds to the two ends of the 20S proteasome forming the immunoproteasome [55]. In mammals, PA28 is composed of two IFNγ-induced, homologous subunits (α and β) and a non-IFNγ induced subunit, PA28γ (Ki antigen) [56]. PA28 stimulates degradation of short peptides by the catalytic core of the immunoproteasome, enhancing the epitope diversity for antigen presentation [39].

While the proteolytic function of the immunoproteasome remains unchanged, the cleavage preferences differ from those of the 26S proteasome, as the IFNγ-induced enzyme preferentially hydrolyses proteins after non-polar amino acids [57-59]. LMP-2 and LMP-7 knockout mice were found to have lower MHC I expression on the surface of cells and did not efficiently present certain antigenic peptides [60, 61]. It was also shown that LMP-2 and LMP-7 are not essential for antigen presentation, hinting the important role of the 26S proteasome in antigen processing [62]. Research on the generation of the immunodominant peptide SIINFEKL from the degradation of ovalbumin, showed that the 26S proteasome primarily generates N-terminally-extended forms of SIINFEKL (1-7 additional N-terminal residues) [46]. While immunoproteasomes generated the same levels of SIINFEKL as proteasomes, the concentration of SIINFEKL extended at the NH_2 terminus by one or more residues was significantly higher (2-4 fold higher) through ovalbumin degradation carried out by the immunoproteasome [46]. Even though the immunoproteasome has an important role in the generation of a variable pool of antigenic peptides with a precise C-terminus, N-terminal trimming by aminopeptidases is also required to convert these precursor peptides into mature epitopes for stable binding to MHC I [42, 46, 63]. ERAP1 is an additional peptide editor affecting the epitope repertoire, with absence of ERAP1 being associated with reduced peptide presentation and an increase in abundance of peptides with N-terminal extensions [64-66].

1.4.3 Antigen sources for MHC I-restricted presentation

Peptides for MHC I binding are derived from the proteolysis of functional and properly folded cytosolic proteins, but also from the rapid degradation of truncated or misfolded proteins after their synthesis, Defective Ribosomal Products (DRiPs) [67]. The first MHC I-restricted epitope identified was an Influenza peptide that binds to H2-D^b and originated from the expression of truncated nucleoproteins [68]. Upon stimulation of cells with IFNγ, there is formation of oxidant-damaged proteins, DRiPs, which can result in aggresome-induced structures and cell apoptosis [69]. The immunoproteasome is known for its role in antigen processing and presentation by MHC I, but it also has a function in ensuring cell viability under IFNγ-induced stress conditions through degradation of accumulated DRiPs following their ubiquitylation [69]. Precursors of antigenic peptides have been found to be associated with high molecular weight chaperones in the cytosol [70]. Heat shock protein 90 α (hsp90α) was shown to associate with large degradation fragments of ovalbumin, with knockdown of hsp90 expression preventing accumulation of fragments in cell extracts [70]. The group II chaperonin TRiC (TCP-1 ring complex) was revealed to have a non-redundant role in the protection of certain N-terminally extended proteolytic intermediates from degradation in the cytosol by proteases such as tripeptidyl peptidase II (TPPII) [70]. Therefore, the following pathway is suggested; hsp90 binds to proteins which are then targeted to the proteasome for degradation through ubiquitination by the E3-ubiquitin ligase CHIP that binds to hsp90α [8]. The resulting proteosomal degradation fragments associate with TRiC and can be further trimmed by cytosolic proteases to a suitable length for entry into ER through TAP [8].

1.4.4 Non-proteasomal proteolysis

Craiu et al revealed that proteosomal inhibitors did not affect the trimming of peptides containing 2-25 additional residues at the N-terminus of the immunodominant ovalbumin epitope SIINFEKL, and that the final SIINFEKL epitope was efficiently presented by the MHC I allele, H2-K^b, at the cell surface [42]. This finding along with the fact that the proteasome does not possess aminopeptidase activity, indicated that other proteolytic enzymes are responsible for the N-terminal trimming of peptides and

since TAP preferentially transports peptides up to 16 residues long, most of the peptides must have been trimmed in the cytosol by aminopeptidases [42, 71, 72].

The cytosolic protease TPPII is a eukaryotic serine-peptidase of the subtilisin-type, it is larger than the proteasome but one-tenth as abundant and has a rod-shaped structure. TPPII efficiently generates antigenic peptides through exo- and endoproteolytic cleavage by trimming substrates longer than 15 residues with unblocked NH_2-termini [73].

Leucine aminopeptidase (LAP) is a cytosolic aminopeptidase thought to play an important role in the generation of epitopes for MHC I presentation [72]. In vivo studies revealed that the epitope SIINFEKL can be generated from precursors through a non-proteasomal pathway and stimulation with IFNγ in HeLa cell extracts resulted in N-terminal trimming of the 11mer QLESIINFEKL to 8-10mer peptides at different rates by the IFNγ-induced LAP [42, 72].

1.4.5 Assembly of the peptide loading complex

The folding and assembly of the MHC I molecule involves the association of the MHC I α chain with β_2m forming a heterodimer. The domains α1 and α2 of the MHC I heavy chains form the peptide binding groove, with the absence of peptide being associated with MHC I instability [74-77]. The peptides are entering the ER from the cytosol and are edited and loaded onto the assembled MHC I molecule with the aid of the peptide loading complex (PLC). This complex is composed of seven subunits; the MHC I heterodimers (HC:β2m), the peptide transporter heterodimer TAP (TAP1/TAP2), the MHC-I-specific chaperone tapasin, the oxidoreductase ERp57 and the lectin-like chaperone calreticulin (CRT) [8]. A structural model of the human native PLC was determined through the isolation of an endogenous, transient PLC with a total height of 240Å from human Burkitt's lymphoma using a viral inhibitor as bait [78]. It was shown that the ER-lumenal PLC consists of two editing modules, each of them containing one of each: tapasin, ERp57, CRT and MHC I [78]. However, one of the modules was lacking either CRT or MHC I or both, whereas the other editing module was always

fully assembled, a finding that is supported by immunoprecipitation evidence generated by Panter et al [79]. The difference in the assembly status between the two editing modules could be explained by the faster rate at which disassembly of the PLC occurs following peptide binding [78]. Tapasin and MHC I form the catalytic, rigid core of PLC which is surrounded by a belt composed of ERp57 and the highly flexible CRT and this is supported by the model created by Fisette et al which incorporates data from the cryo-EM structure by Blees et al [74, 78]. Upon binding of the peptide on the MHC I molecule, one of the two editing modules undergoes structural changes and the pMHC I dissociates from the PLC followed by translocation to Golgi where the tapasin homologue, TAP-binding protein related (TAPBPR) and UDP-glucosyl glucose transferase (UGGT) further edit the complex [80]. The pMHC I transits to the cell surface for presentation to CTLs and Natural Killer (NK) cells (Figure 1.6) [81].

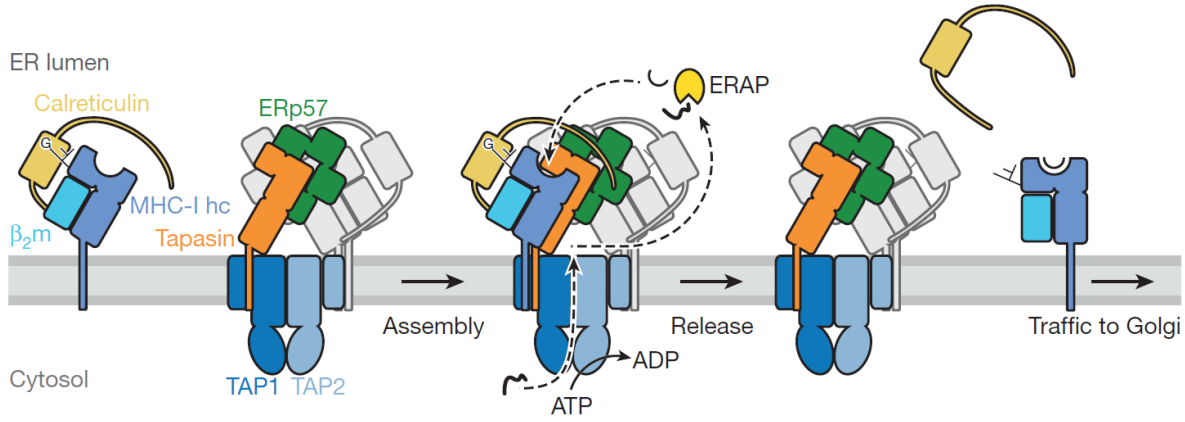


Figure 1.6. Peptide loading complex (PLC) assembly in the ER lumen.

The asymmetric PLC becomes symmetric upon recruitment of the empty HC:β2m heterodimers (MHC I) by the chaperon calreticulin and consists of two editing modules. Each of the two modules is composed of tapasin, CRT, ERp57 and MHC I located around the TAP heterodimer. Antigenic peptide precursors enter the ER lumen through the heterodimer TAP and are subsequently N-terminally trimmed by ERAP1 for optimal binding to the peptide-receptive MHC I molecules. Following optimisation of the peptide cargo by tapasin, the stable MHC I-peptide complexes are translocated to the Golgi and from there to the cell surface for presentation to CTLs which will then exert their immune function. Adapted from Blees et al 2017 [78].

1.5 Endoplasmic reticulum aminopeptidase 1 (ERAP1)

1.5.1 ERAP1 in the generation of antigenic peptides

In 1996, a novel human placental leucine aminopeptidase/oxytocinase (P-LAP), a member of the type II membrane-spanning zinc metalloprotease family, was cloned and its expression was found to be increased in maternal serum during pregnancy [82]. A few years later in 1999, researchers discovered a novel leucine aminopeptidase that had a significant homology to P-LAP [83]. The aminopeptidase was termed adipocyte-derived leucine aminopeptidase (A-LAP) because the cDNA encoding the enzyme was found in an adipose tissue cDNA library [83]. In 2002, Saric et al isolated this enzyme from HeLa S cells and proposed the name endoplasmic reticulum aminopeptidase 1 (ERAP1) as the most appropriate for this aminopeptidase in rats and humans [35]. In the same year, the Shastri lab identified the relevant aminopeptidase in mice, termed endoplasmic reticulum aminopeptidase associated with antigen processing (ERAAP), which shares 86% homology with human ERAP1 [35, 84, 85]. ERAP1 is also designated puromycin-insensitive leucyl-specific aminopeptidase (PILSAP) and aminopeptidase regulator of tumour necrosis factor receptor (TNFR) type 1 shedding (ARTS1) based on its alternative roles. Studies in ERAP1-knockout mice revealed the role of ERAP1 in peptide trimming as part of the MHC I antigen processing and presentation pathway [35, 36, 85]. ERAP1, induced by IFNγ, is responsible for the final processing of N-terminally extended precursors entering the ER from the cytosol through TAP, generating a number of high affinity, 8-9 amino acid-long peptides and two trimming mechanisms have been proposed [86-90]. MHC I expression at the cell surface depends on the quality and quantity of peptides in the ER which are affected by the trimming function of ERAP1 [64-66]. ERAAP-deficient mice had reduced cell surface expression of pMHC I [64, 65]. In these mice, ERAAP deficiency resulted in significant reduction in the levels of MHC I molecules, specifically H-2K^{b} and H-2D^{b} [91, 92]. Another study confirmed this by showing that ERAAP knockdown by small interfering RNA (siRNA) in murine L cells lead to decreased expression of the MHC I molecules

H-2K^{k} and H-2L^{d} [84]. The expression of MHC II was the same in ERAP1-deficient cells and in WT cells [91, 92]. Interestingly, ERAP1 silencing in HeLa.B27 cells revealed a reduction in the number of 9mer HLA-B27-bound peptides, while the percentage of longer peptides (11 to 13 amino acids) was increased [93]. Furthermore, ERAP1 silencing resulted in detection of C-terminally extended peptides which were binding to HLA-B*27 more than N-terminally extended peptides, likely due to the presence of arginine at position 2, a common feature of the HLA-B*27 peptide binding motif [93, 94]. These studies show that in the absence of ERAP1, the reduction in pMHC I is dependent on the cell surface expression of MHC I, which is reduced up to 70% in mice and 30% in humans depending on the allele [64, 65, 89, 91]. The lack of ERAP1 activity depleted the pMHC I repertoire of many peptides and the ERAP1-deficient cells expressed a large number of unstable but highly immunogenic, structurally unique pMHC I which elicited potent CD8+ and B-cell responses [64, 65, 91]. Similarly to the effects of ERAP1 silencing in HeLa.27 cells, a significant number of peptides were one to four amino acids longer than the canonical 8 (for binding to H-2K^{b}) or 9 (for binding to H2-D^{b}) amino acids in ERAAP$^{-/-}$ deficient cells, confirming that ERAAP has a significant role in the generation of optimal length peptides for presentation by MHC I [66]. Certain peptides do not require trimming by ERAP1 and consequently are unaffected by the loss of ERAP1, whereas other antigenic peptides such as the tumour antigen GSW11 in the colorectal carcinoma CT26 and MART-1 in melanoma, are destroyed by ERAP1 over-trimming [64, 66, 81, 95].

A second ER-resident aminopeptidase expressed in humans, ERAP2, shares 49% homology with ERAP1 and it is also thought to be involved in the N-terminal trimming of extended peptide precursors [96]. It has been suggested that ERAP1 and ERAP2 can form heterodimers that may alter enzymatic activity and increase efficiency of the generation of peptides such as the HIV-derived epitope GPGRAFVTI, indicating that certain peptides could require trimming to optimal length for MHC I binding through the cooperation of ERAP1 and ERAP2 [34, 96]. Approximately 25% of the population are lacking ERAP2

expression due to the presence of a single nucleotide polymorphism (SNP, rs2248374) in the gene sequence that targets the protein for nonsense-mediated decay [97].

1.5.2 ERAP1 expression

ERAP1 expression is higher in tissues with high expression of MHC I and it is regulated at genetic, transcriptional and post-transcriptional levels [8, 35, 36, 98]. Tumour cell lines have been shown to both increase and decrease ERAP1 expression at pre- and post- transcriptional level to evade immune recognition by CD8+ T cells [99]. In vivo findings show that ERAP1 and ERAP2 expression is lowest in HLA class I-low tumour lesions, and may be low or imbalanced in HLA class I-high tumours [100]. Downregulation of ERAP1 and/or ERAP2 was shown in breast, ovary and lung carcinomas deriving from tissues co-expressing the two aminopeptidases [100]. Moreover, increase of ERAP1 and ERAP2 expression by transfection in HeLa cells lead to increased MHC I expression, directly showing that MHC I expression is affected by the expression of the two ER aminopeptidases [101].

1.5.3 ERAP1 structure and localisation in the ER

ERAP1 is a 2.8Kb-long gene located on chromosome 5q15, with alternative splicing of the gene giving rise to two isoforms of 941 and 948 amino acids with an active site spanning 375 amino acids. The crystal structures of ERAP1 have provided insight into the trimming mechanism and the substrate specificity of the aminopeptidase [102-104]. The previous crystal structures of ERAP1 were of medium-to-low resolution (2.7-3.0Å), however recently a high-resolution crystal structure of ERAP1 at 1.60Å was determined with a high affinity inhibitor bound to the active site of ERAP1 [104]. ERAP1 is a member of the M1 zinc-metalloprotease family (oxytocinase subfamily) and is characterised by the presence of two important motifs, the five amino acid motif GAMEN (306-310 amino acids in mouse and 317-321 in humans) for substrate binding and the $HEXXH(X)_{18}E$ Zn^{2+}-binding motif [83, 105, 106]. Substitution of glutamic acid (E) at residue 320 with alanine (A) (E320A, GAMAN motif) leads to a mutant ERAP1 that when transfected into ERAP1-deficient cells, failed to restore the defective

antigen presentation, showing the importance of the GAMEN motif in enzymatic activity [107]. The two key motifs are found in domain II of ERAP1 which consists of four globular domains in total forming a concave-like structure with an internal cavity that contains the catalytic active site where substrate peptides bind [102, 103]. The four-domain architecture can adopt either of two conformations: open or closed, with the active site being reconfigured during transition [102, 103]. The crystal structure suggests that ERAP1 binds peptides in open conformation, switching to closed conformation upon peptide binding. This mechanism of action is also used by other M1 metalloproteases [103]. Domain I (residues 46-254) is an all-β sandwich domain that packs against domain II, capping off the active site found in domain II and providing sites for the binding of the NH_2 terminus of the substrate peptides [102, 103]. Domain II (residues 255-529) is the catalytic domain, domain III (residues 530-614) is a small β-sandwich domain (2 β sheets) between domains II and IV that acts as a hinge enabling interconversion between the two ERAP1 conformations and domain IV (residues 615-940) consists of 16 α helices of various sizes and extends away from the catalytic domain, forming a large internal cavity of ~10974 $Å^3$ extending from the active site to the interior part of domain IV [104]. The large cavity fully occludes the catalytic site and it is plausible that it provides access to long peptides [86, 102, 103, 108]. Domain IV is thought to have a regulatory region binding the carboxy terminus of the substrate encouraging longer peptide trimming [103, 109, 110]. Two published high-resolution crystal structures of ERAP1 with bound peptide analogues (1.68Å and 1.72Å) along with biochemical data suggest that because ERAP1 is not as polymorphic as MHC I, in order to trim every antigenic peptide it has a wide peptide-binding site that can accommodate long peptides entering the ER (10mer and 15mer found in internal cavity in closed conformation) [104, 108]. In the same study, the three residues Tyr684, Lys685 and Arg807 in domain IV, were shown to form hydrogen-bonding interactions with the C-terminal residue of a 15mer peptide [108]. This interaction was able to stabilise the peptide which was tethered by both ends, suggesting that for long peptides up to 16 amino acids in length, ERAP1 can recognise both the C- and the N-terminus of peptides and enable accommodation in a large

internal cavity [86, 108]. Interestingly, the substitution of Gln181 with Asp resulted in a mutant ERAP1 with altered specificity from hydrophobic towards basic amino acids, showing the importance of this residue in substrate specificity (Figure 1.7) [111].

ERAP1 has a molecular localisation in the ER lumen as a soluble monomeric protein and despite the lack of any obvious ER retention sequence or KDEL motif, it localises with other ER-resident proteins that contain such motifs [35, 84, 107]. The exon 10 coding sequence in domain II of ERAP1 has been identified as potential contributor to the ERAP1 retention in the ER through the construction of expression plasmids encoding a chimeric protein of ERAP1 and P-LAP and transfecting them into 293T cells [112]. In the chimeric proteins, the C-terminal ERAP1 sequence was replaced with that of P-LAP and secretion into culture medium was investigated [112]. It was shown that ERAP1, ERAP2 and the chimeric protein containing the N-terminal sequence 1 to 615 amino acids of ERAP1 were co-localised with the ER-retained protein PDI. Interestingly, the chimeric proteins containing the N-terminal sequence from 482 to 615 amino acids of ERAP1 (domain II and domain III)were also retained in the ER, showing the significance of these residues for ER retention [112]. It was suggested that exon 10, encoding the sequence between 485-508, is responsible for ERAP1 retention in the ER, as ERAP1 that lacked exon 10 was secreted in the culture medium [112]. Another study showed that ERAP1 was secreted in the culture medium when ectopically overexpressed in COS-7 cells [83].

1.5.4 ERAP1 substrate specificity

The substrate specificity of ERAP1 includes preference for aromatic and hydrophobic amino acids, such as leucine and valine, with a length preference of 9-13 amino acids [90, 96, 113]. While ERAP2 is also involved in N-terminal trimming of peptides, it preferentially trims 9mers or shorter peptides, and has a trimming specificity for polar and charged amino acids, such as arginine [90, 96, 107, 114, 115]. Both ERAP1 and ERAP2 are unable to process X-pro-X bonds and TAP cannot transport peptides with proline at position two. ERAP1 is able to generate MHC I peptides containing X-pro-X, but from their

precursors [85, 113, 116]. Research revealed that peptides entering the ER through TAP fall into three categories based on whether they are affected by the trimming ability of ERAP1 before binding on MHC I: independent, dependent and sensitive [65]. ERAP1-independent peptides are not affected by ERAP1 trimming and are likely to already be of optimal length upon entering the ER, ERAP1-dependent peptides are N-terminally extended and require ERAP1 trimming to transform into high affinity peptides and ERAP1-sensitive peptides are destroyed by ERAP1, therefore they can only bind to MHC I in the absence of the enzyme [36, 65]. ERAAP-deficient mice express different pMHC I than WT mice and when peptide-H-2K^{b} or peptide-H-2D^{b} complexes from WT mice are presented to CD8+ T cells, potent immune responses are raised [65]. Similar CD8+ T cell responses were observed when WT mice were immunised with ERAAP-deficient splenocytes [65]. An example of an ERAP1-sensitive peptide is the aforementioned CT26 tumour-derived H2-D^{d} specific antigen, GGPESFYCASW (GSW11) which is normally destroyed by ERAP1 activity. Using CT26, a murine colorectal carcinoma model, it was revealed that downregulation of ERAP1 expression enabled sufficient generation of GWS11 (~75-fold increase) for presentation to CTLs which exerted their anti-tumour function and lead to tumour growth arrest and enhanced survival [81]. These experiments reveal the role of ERAP1 in the generation of optimal peptides for MHC I binding and the ability of tumours to take advantage of ERAP1 activity with the aim to escape immune recognition.

1.5.5 ERAP1 trimming mechanism

Currently, two non-mutually exclusive hypotheses have been proposed regarding the trimming mechanism of ERAP1. The first hypothesis suggests that ERAP1 itself acts as a 'molecular ruler' trimming peptides to the optimal length for MHC I binding [36, 86, 117]. In this proposed trimming mechanism, substrates with a hydrophobic side chain at the C-terminal residue bind with high affinity within the hydrophobic pocket of ERAP1 and at the catalytic site [86]. Chang et al investigated the trimming of a number of peptides of different length by recombinant ERAP1 and it was shown that 8- to 9- to residue mature epitopes were not affected by ERAP1 trimming or they were only affected at

some degree [86]. Hydrolysis of peptides longer than 14 residues was reduced and ERAP1 did not exert its function on peptides shorter than 8 residues or longer than 18 residues [86]. The maximum ERAP1 activity was measured for peptides with length between 10 and 14 amino acids. Hydrolytic activity is affected by the length of the peptide and by both the N- and C-terminal residues, with ERAP1 having a preference for hydrophobic C-terminal amino acids [118]. It was proposed that the reason why peptides shorter than 8- to 9- residues are not trimmed by ERAP1, is that these peptides cannot extend from the binding site, the hydrophobic pocket, to the catalytic site of ERAP1 [86]. The recently published structure of ERAP1 supports the 'molecular ruler' theory, by showing that 15mer peptides were able to interact with both the C-terminal docking site in domain IV and the catalytic site, while 10mer peptides did not extend to the C-terminal binding site [108]. Alternatively, it is possible that the 10-mer did not interact with the C-terminal binding site because the C-terminal residue was a positively charged amino acid (Lys, K), while the relevant residue of the 15-mer was a non-polar amino acid (Ile, I) and could be recognised by the C-terminal docking site which carries a positive electrostatic potential [108]. Although the findings of Giastas et al support the 'molecular ruler' theory by showing that longer peptides with a hydrophobic C-terminal residue extend into a carboxypeptidase-like C-terminus binding site while also being accommodated in the catalytic site of domain II, this might apply only to epitopes with hydrophobic C-termini [108].

Regarding the second proposed trimming mechanism, the 'MHC I-template' mechanism, it has been suggested that MHC I binds N-terminally extended precursors in the ER to protect them from degradation and this pMHC I could act as a template for ERAP1 to trim only the additional N-terminal residues to optimal length [90, 107]. Generation of the peptide epitope QLSPFPFDL (QL9) was carried out in the presence of both ERAP1 and H-2L^{d} [107]. The absence of the MHC I molecule resulted in elimination of the epitope from the ER, indicating that the MHC I allele is required to protect the peptide from further trimming by ERAP1 [107]. The immunoprecipitation of H2-L^{d} with an antibody specific for its α3 domain showed association with the 12mer antigenic precursor of QL9 peptide [107].

It was suggested that since most antibodies and T cells recognise the final pMHC I, complexes consisting of a long precursor peptide bound to MHC I such as the above, are not detected [107]. Chen et al showed that the ERAP1-ERAP2 heterodimer also trims peptide precursors bound to MHC I [87]. A plausible scenario is that both of the hypotheses regarding the ERAP1 trimming mechanism are used by the enzyme in different cases. ERAP1 can trim free peptides as well as peptides bound to MHC which was shown by the ability of ERAP1 to trim a single chain trimer construct despite the lack of a free C-terminus [90, 102, 103]. Trimming of peptides tethered to MHC I is likely dependent on access of ERAP1 to the N-terminus and associated with MHC I affinity indicating a balance between ERAP1 and MHC I binding of peptides based on which molecule has higher affinity for the peptide [90].

1.5.6 ERAP1 polymorphism

Human ERAP1 is polymorphic and is thought to contain at least seventy exonic, non-synonymous, biallelic single nucleotide polymorphisms (SNPs) [89]. Ombrello et al investigated the 1000 Genomes dataset and identified 10 missense SNPs in ERAP1 present with >5% frequency in at least one population (European, Asian or African). The relevant SNPs include: rs72773968 (T12I), rs3734016 (E56K), rs26653 (R127P), rs26618 (I276M), rs27895 (G346D), rs2287987 (M349V), rs30187 (K528R), rs10050860 (D575N), rs17482078 (R725Q) and rs27044 (Q730E), where the first letter reflects the ancestral amino acid, followed by the amino acid position and the non-ancestral amino acid (Table 1.1) [119, 120].

Table 1.1. ERAP1 single nucleotide polymorphisms

rs number	Nucleotide change	Amino acid change	Base position	ERAP1 domain
rs72773968	C/T	T12I	36	I
rs3734016	G/A	E56K	166	I
rs26653	G/C	R127P	380	I
rs26618	A/G	I276M	828	I/II

rs27895	G/A	G346D	1037	II
rs2287987	A/G	M349V	1045	II
rs30187	A/G	K528R	1583	II/III
rs10050860	G/A	D575N	1723	II/III
rs17482078	G/A	R725Q	2174	IV
rs27044	C/G	Q730E	2188	IV

The SNPs commonly associated with disease result in amino acid changes in the ERAP1 protein and they exist in multiple combinations forming haplotypes that encode multiple distinct ERAP1 protein variants, referred to as 'allotypes' and alter the enzyme's trimming function (Table 1.2) [90]. ERAP1 allotype nomenclature was standardised using the coding sequence of ERAP1 allotypes from a cohort of autoimmune disease, ankylosing spondylitis (AS), to reflect the polymorphic nature of ERAP1 [89, 120]. Ombrello et al assigned nine SNPs in ten ERAP1 haplotypes, indicated as Hap1 to Hap10 referring to the genetic sequence of ERAP1 [120]. Reeves et al published thirteen allotypes referring to the translated protein expressed ERAP1 and the Stratikos group published six more later [89, 90, 121-123]. Since both chromosomal copies of ERAP1 are co-dominantly expressed in an individual, both ERAP1 allotypes affect overall ERAP1 function in the generation of the peptide repertoire presented to CTL [89, 124, 125]. A well characterised assay was used by the James lab for investigating the combined effect of the two ERAP1 allotypes in an individual on the enzyme's trimming function using an N-terminally extended antigenic peptide (Results section) [89]. ERAP1 allotypes are classified based on their overall trimming function compared to the WT ERAP1 allotype containing only ancestral amino acids at the aforementioned positions mentioned: hypoactive, hyperactive and efficient in their ability to generate antigens [89, 90]. These differences in trimming activity are likely to alter the peptide repertoire, which may affect the ability to generate antigenic peptides and induce protective CD8+ T cell responses [90]. Polymorphic amino acids have been identified near the catalytic site (rs27895 G346D, rs2287987 M349V), on the inner surface of domain IV in the proposed regulatory site and likely involved in altering conformation from open to closed upon peptide binding (rs17482078 R725Q,

rs27044 Q730E) and in interdomain regions (rs30187 K528R, rs10050860 D575N) or in other parts of the enzyme that have an effect on conformation and consequently the trimming activity of ERAP1 [103]. The amino acid change from G to the acidic, negatively charged D at position 346 may affect substrate recognition [108]. R127P and K528R are found at domain junctions, close to the catalytic site and these SNPs may result in formation of hydrogen bonds or polar interactions with peptide side chains, leading to a problematic ERAP1 transition from open to closed state [102, 103]. The change of the polar uncharged Q for the negatively charged E at position 730 in domain IV, may influence the electrostatic potential of the substrate-binding cavity, preventing the binding of the negatively charged C-terminus of longer peptides and increasing in this way the trimming of shorter peptides [122]. A similar amino acid change occurs at position 725 in ERAP1 which also involves change in charge; from the positively charged R to the polar Q [108]. The most widely studied disease-associated SNP effect on the ERAP1 trimming function is rs30187, encoding K528R, has been shown to reduce protein expression of ERAP1 as well as a ~30% reduction in the ability to generate the final peptide, suggesting a hypo-trimming phenotype [89, 90]. On the contrary, presence of the amino acid change Q730E affects ERAP1 trimming activity based on substrate length, it increases ERAP1 substrate preference for shorter peptides, likely because of its localisation deep in the internal cavity of ERAP1 and the negative charge carried by glutamic acid which many alter the local electrostatic potential, affecting the binding of the peptide C-terminus [104, 122]. ERAP1 allotypes containing M349V or M349V/D575N/R725Q result in the same trimming as the wild type ERAP1 and are thought to be efficient towards trimming a model peptide substrate (AIVMK-SIINFEHL, SHL8) [90]. However, there are also haplotypes that result in expression of a hyper-trimming ERAP1 allotype; R725Q/Q730E and K528R/R725Q, shown by trimming a peptide epitope to shorter products than the optimal for MHC I binding [90]. Analysis of ERAP1 and ERAP2 expression in B-LCLs, revealed that a SNP in the ERAP2 promoter region, rs75862629 A/G, correlates with a double effect on ERAP2 mRNA, lower transcription and higher degradation, as well as increased ERAP1 transcription [126]. Therefore, this

allelic variant in the intergenic region between ERAP1 and ERAP2 regulates differential expression of ERAP1 and ERAP2.

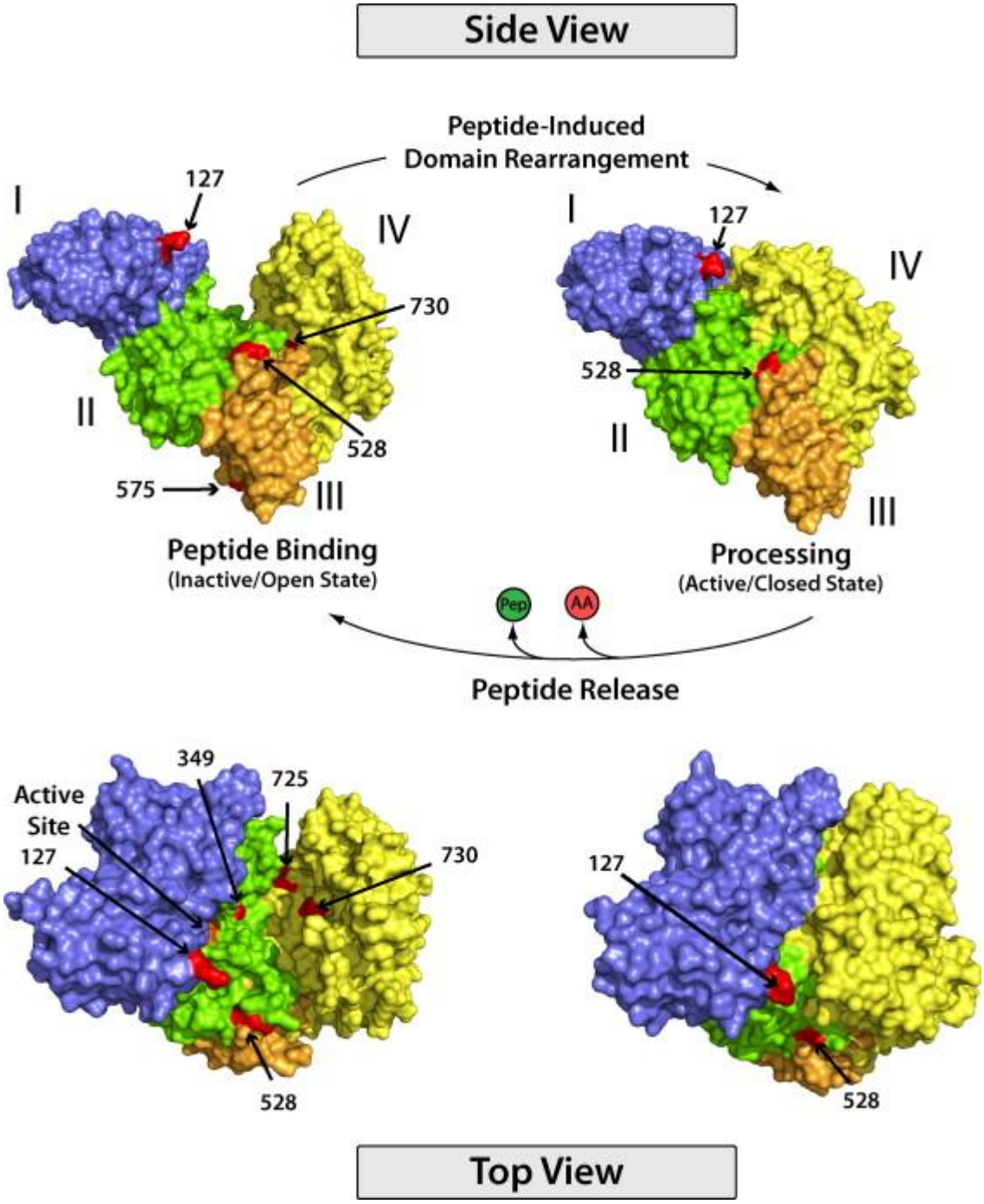


Figure 1.7. ERAP1 structure. Top: side view of ERAP1 and transition from open to closed state upon binding of peptide. Bottom: front view. Figure adapted from **[**127**].**

Table 1.2: Effect of SNPs on ERAP1 trimming function

Amino acid changes in ERAP1 affect the enzyme's trimming function [88-90, 102, 125, 128]. No functional effects on the ERAP1 trimming function have been reported in literature for E56K and G346D. N/A indicates that the trimming effect has not been shown. Positions 12 and 276 have not been explored in the trimming assays used in published literature. Assay used to investigated trimming of the N-terminally extended epitope AIVMK-SIINFEHL.

SNPs	**ERAP1 trimming of AIVMK-SIINFEHL**	**ERAP1 domain (single SNPs)**
E56K	N/A	I
R127P	efficient	I
G346D	N/A	II
M349V	efficient	II
K528R	hypotrimming	II/III
D575N	efficient	III
R725Q	hypotrimming	IV
Q730E	hypotrimming	IV
K528R/Q730E	hypotrimming	
M349V/K528R	hypotrimming	
R725Q/Q730E	hypertrimming	
K528R/R725Q	hypertrimming	
M349V/D575N/R725Q	efficient	
M349V/K528R/R725Q/Q730E/D575N	hypotrimming	

1.5.7 Polymorphic ERAP1 associated conditions

Genome wide association studies (GWAS) have identified ERAP1 SNPs associated with predisposition to a number of autoimmune and autoinflammatory diseases including Behçet's disease (BD), psoriasis, type I diabetes, multiple sclerosis and birdshot chorioretinopathy (reviewed in [117]) as well as ankylosing spondylitis (AS) [129-133], in which the link between AS and ERAP1 was only observed in AS patients who were HLA-B*27 positive [93, 128, 134] (Table 1.3). Strong linkage disequilibrium between SNPs in ERAP1 prevents associating disease risk to a single SNP [89]. GWAS studies have indicated the polygenic nature of AS, and shown strong associations in both ERAP1 and IL-23R which

contribute 26 and 9% to the risk of developing AS (reviewed in [129]). Reduction in ERAP1 expression affects the generation of peptides for binding to HLA-B*27 by increasing their peptide length as well as increasing expression of free heavy chains, affecting stability of HLA-B*27 inside the cell and at the cell surface as longer peptides have a higher rate of dissociation from HLA-B*27 [93, 135]. The following haplotypes, K528/D575/R725, K528/D575/E730 as well as simultaneous occurrence of ERAP1 haplotype Q730/K528 and ERAP2 SNP K392N have been associated with AS [89]. Conversely, the ERAP1 haplotype R528/N575 has been shown to be protective, with HLA-B*27 positive individuals who are homozygous for R528/N575 having the lowest risk of AS [128].

The major risk factor for susceptibility to the autoinflammatory condition BD is the presence of the HLA-B*51 allele. Other risk genes include IL-10 and IL-23R as in the case of AS. However, contrary to AS, the ERAP1 SNPs encoding N575 and Q725 are associated with increased risk of BD [128]. A study on the role of ERAP1 in BD revealed that individuals homozygous for both the ERAP1 allotype *001 (P127/V349/R528/N575/Q725/E730) and for HLA-B*51, had approximately 11-fold higher risk of suffering from BD [136]. An association between ERAP1 SNP that encodes K528R and type 1 diabetes was identified with a significance of p=0.003895 (reviewed in [117]). An epistatic interaction was observed between rs27044 (E730) and HLA-C*06 which may be involved in early onset of psoriasis, while another study showed association of rs26653 (R127P) with onset of disease in puberty and independently of HLA-C*06 [137, 138]. Furthermore, the ERAP1 haplotype M349/K528/Q730 along with ERAP2 expression was found to offer higher susceptibility to psoriasis, while the haplotype M349/R528/E730 without ERAP2 expression was protective against psoriasis (reviewed in [117]).

Table 1.3: Allelic variation in ERAP1 and disease linkage

Disease	HLA association	ERAP1 SNP association
Ankylosing spondylitis [89, 128, 139]	HLA-B*27	rs26653 (R127P) rs2287987 (M349V) rs30187 (K528R) rs10050860 (D575N) rs17482078 (R725Q) rs27044 (Q730E)
Behçet's disease [136, 140]	HLA-B*51	rs10050860 (D575N)
		rs2287987 (M349V)
		rs17482078 (R725Q)
Type I diabetes [141]	HLA-DR3	rs30187 (K528R)
Psoriasis [137, 138, 142]	HLA-Cw*06:02	rs26653 (R127P)
		rs30187 (K528R)
		rs27044 (Q730E)
Multiple Sclerosis (MS) [143]	HLA-DR15	rs30187 (K528R)
	HLA-C*05	
	(protective)	
Inflammatory Bowel	HLA-C*07	rs30187 (K528R)
Disease (IBD) [144]		

In cancer, defects in the expression and function of both ERAP1 and ERAP2 genes have been detected in solid and haematological cancers (reviewed in [2]). Comparison of ERAP distribution between neoplastic and normal cells revealed frequent low expression in tumours regardless of the tumour histotype, downregulation of either ERAP protein in breast, ovary and lung carcinomas deriving from tissues that express both proteins, upregulation of ERAP expression in colon and thyroid carcinomas deriving from tissues lacking expression of both proteins and ERAP1/ERAP2 imbalance in all tumour histotypes [98, 100].

1.6 Cervical cancer

Cervical cancer is the result of the uncontrolled development of abnormal cells at the lining of the cervix, the anatomical region found between the vagina and the uterus [145]. According to Cancer Research UK (CRUK), cervical cancer is the 14th most common cancer in women in the UK with the incidence rates reaching 3,197 cases (2016-2018 average) while the mortality rates from cervical cancer have reached almost 24,000 cases per year in Europe [145, 146]. From 2000-2007, the five-year survival rates for cervical cancer ranges from 67% in Northern Europe to the low 57% for Eastern Europe [147]. It is noteworthy that there has been a 25% decrease in cervical cancer incidence rates since the early 1990s due to regular screening appointments and HPV vaccination [145].

The most common type of cervical cancer is cervical squamous cell carcinoma (CSCC), accounting for 70-80% of all cervical cancer cases, which involves the uncontrolled proliferation of abnormal epithelial cells covering the outer part of the cervix, the ectocervix [145]. Cervical adenocarcinoma, accounting for 10% of all cervical cancer cases, arises from the mucinous endocervical epithelium and usually involves accumulation of atypical mucus-producing glandular cells [145, 148]. The remaining 10% of cervical cancer cases are adenosquamous carcinoma (5-6%) which affects both squamous and glandular cells, and small cell cancer (3%) [145] . A very small number of cancers in the cervix are lymphomas or sarcomas, but these women receive different treatment than the other cervical cancer types [145]. More than 95% of cervical cancer cases are caused by infection with human papillomavirus (HPV), however not all women with an HPV infection will develop cervical cancer, as a significant proportion of women infected with HPV are able to clear the virus without intervention [149]. Given that 80% of sexually active individuals are exposed to HPV in their lifetime, and the significant contribution of HPV to the development of cervical carcinoma, it is vital to be able to identify those women who are at high risk of progressing to cervical cancer following persistent HPV infection and treat them earlier than we do now [150].

1.7 Cervical carcinogenesis

Approximately 90% of women with a functional immune system are able to clear the asymptomatic HPV infection, which can take as long as two years [151]. However, it remains unclear whether the virus is truly cleared or if it remains in a dormant stage known as 'viral latency', which can later be re-activated [152-155]. HPV type-specific antibodies have been detected in the serum of 60% of individuals who cleared the HPV infection, but these might not offer protection in case of re-infection [155, 156]. The 10% of infected women that are unable to clear the HPV infection, and have viral persistence, may progress to cervical dysplasia and ultimately cervical cancer if left undetected or untreated (Figure 1.9) [151].

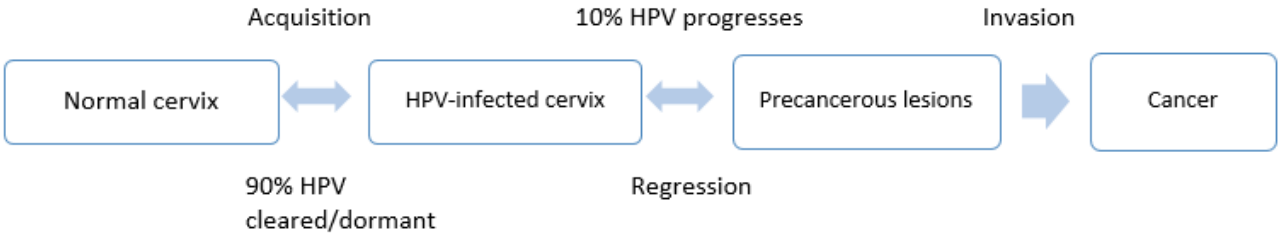


Figure 1.8. From HPV infection to cervical cancer development

Women with normal cervix can become infected with HPV, with most cases (90%) clearing the viral infection within two years. In 10% of cases, HPV infection progresses to the formation of pre-cancerous lesions which can either regress or progress to invasive cervical cancer. Modified from Gravitt and Winer 2017 [155].

The precursor of CSCC is termed squamous intraepithelial lesion (SIL) previously known as cervical intraepithelial neoplasia (CIN) (Figure 1.10). CIN grades 1 to 3 are still used in biopsy reports today referring to lesion severity. Below is the primary classification of cervical lesions [157]:

- CIN1 (low grade squamous intraepithelial lesions, LSIL): affect one third of cervical epithelium and they result from a complete HPV life cycle.

-CIN2 and CIN3 (high grade squamous intraepithelial lesions, HSIL): characterised by presence of immature basaloid cells found in the upper third cervical epithelial layer. The most severe HSIL case is the carcinoma in situ which has not invaded the cervical stroma as opposed to invasive cervical cancer.

-Atypical squamous cells of unknown significance, AS-US: neither completely LSIL nor completely HSIL.

-Atypical squamous cells 'cannot exclude high-grade', ASC-H.

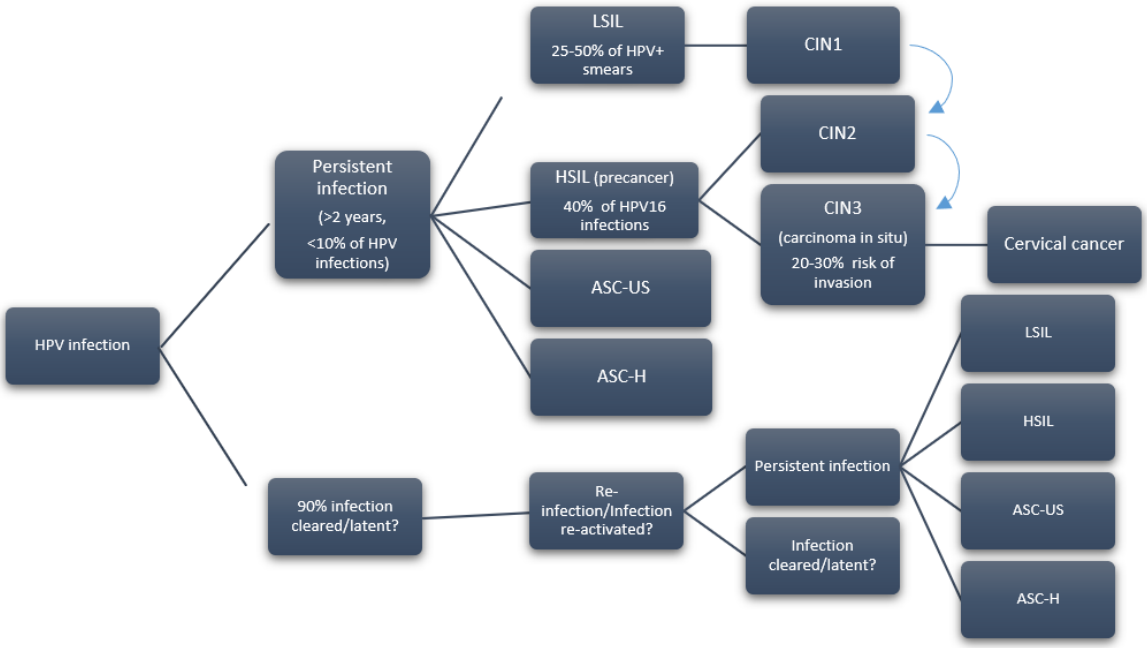


Figure 1.9. Pathway from HPV infection to dysplastic lesions and cervical cancer

HPV infection may persist and lead to development of squamous intraepithelial lesions (SIL) which can be either low (LSIL) or high grade (HSIL). LSIL (CIN1) can progress to HSIL (CIN2/3) with 20-30% risk of invasion and cervical cancer development.

Cellular transformation occurs in those sites most prone to tumorigenesis and in the case of cervical carcinoma, the transformation zone is the area of the cervix where intraepithelial lesions occur. HPV can modify host cell genes along with the expression of viral genes to evade immune system recognition, important for tumour progression and invasion [158]. In 'high-risk' HPV types, the HPV genome integrates into the host genome, resulting in a growth advantage of cells compared to those carrying episomal HPV genome [158].

The minimum carcinogenic event for cell transformation resulting in dysplasia lesions, is the expression of HPV E6 and E7 oncoproteins [158]. In the 'low-risk' HPV types, E6 and E7 have little or no activity. However, upon integration of 'high-risk' HPV genome into the host genome, most viral genes are lost except for oncogenes E6 and E7, which are expressed as a result of the disruption of the E2 gene, alleviating transcriptional repression of both E6 and E7 [159]. The absence of E2 amplification is thought to be an indication of HPV genome integration and/or progression to cervical cancer [158]. E6 and E7 oncoproteins are vital for the virus as they contribute to tumorigenesis through their ability to interact with tumour suppressors p53 and pRB, respectively (Figure 1.11) [160, 161]. Chromosomal abnormalities including gain at chromosome 3q contribute to the progression into invasive cervical cancer [158]. Centrosome abnormalities such as duplication errors, mitotic defects and multiplications, aneuploidy can be generated by HPV-16 E7 [158].

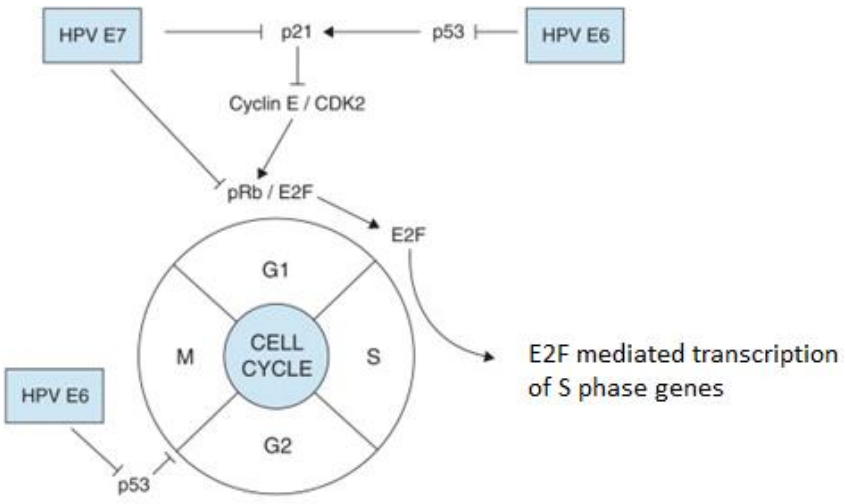


Figure 1.10. HPV E6 and E7 oncoproteins in the dysregulation of the cell cycle

The HPV oncoproteins E6 and E7 contribute to tumorigenesis through the ability to inhibit tumour suppressors p53 and pRB, respectively. E7 also inhibits p21. The inhibition of E2F results in arrest of E2F-mediated transcription of S phase genes, preventing cell apoptosis. Modified from Gustafson 2009 **[**162**].**

1.8 The role of the immune response in HPV infection and cervical cancer

1.8.1 T-cell responses in the tumour microenvironment (TME)

The TME results from the interactions between tumour cells, cells found in the stroma surrounding the tumour which supports tumour development and progression to malignancy, as well as immune cells [31, 163]. Tumours can be classified into "hot" and "cold" tumours based on their immune phenotype. "Hot tumours" have a higher number of immune cells infiltrating the TME and/or interferon (IFN) signature, characteristic of a T-cell-inflamed phenotype, compared to "cold tumours" [164]. It has been suggested that human cancers should be classified into "hot" and "cold" based on the number of detected T cells as well as the tumour-recognition potential of the T cell receptors (TCRs) they express [165].

A high percentage of epithelial tumour infiltrating lymphocytes (TILs) is associated with increased overall survival in cervical cancer patients [31, 166]. However, it is the infiltration of intraepithelial CD3+CD8+ T cells exerting anti-tumour immune responses that contributes to better prognosis [167]. Progression of CIN to invasive cancer is associated with a more immunosuppressive environment involving loss of pro-inflammatory cytokines and increased numbers of IL-10-producing cells [168]. In addition, regression of dysplastic lesions in the cervix requires strong tumour-specific CTL responses as well as an immune stimulating environment [169]. CD103 is a marker for intraepithelial CD8+ T cells and expression of the marker was correlated with expression of T cell markers (CD3, CD2), exhaustion

molecules (PD-1, TIGIT) and activated T cell markers (HLA-DR, HLA-DQ) indicating that high CD103 expression characterises a group of 'hot' tumours in the cervical cancer cohort studied [170]. 'Hot' tumours with a high number of CD103+CD8+ were found in early stage cervical cancer patients, while 'cold' tumours were found in patients with a more advanced stage of disease [170].

Studies of HPV persistence in infected individuals identified an association with the lack of demonstrable, circulating HPV-specific T cells, while spontaneous regression of HPV-induced premalignant lesions was associated with presence of circulating HPV early antigen-specific CD4+ and CD8+ T cells [167, 171, 172]. In cervical cancer patients with tumours integrated deep in cervical tissue, presence of HPV-specific T cells was associated with better disease-free survival (DFS) after radiotherapy [173].

1.8.2 Immunotherapy for cervical cancer

Currently, most research papers on the use of immunotherapy as treatment approach for HPV-associated cancer are open label meaning that both patients and investigators are aware of the treatment used, with historic controls, in advanced cancer patients. Results from clinical trials involving immunotherapy as the sole treatment for pre-malignant disease have shown either none or moderate efficacy [174]. Regarding advanced HPV-positive metastatic cancer, immunotherapy as a sole treatment has not resulted in cure [174]. There are only two FDA-approved immunotherapy approaches for treatment of cervical cancer: an immune checkpoint inhibitor and a targeted antibody. Pembrolizumab is an IgG4 antibody against PD-1 traded under the name Keytruda (Merck & Co) and in 2018, FDA approved its use for treatment of recurrent or metastatic cervical cancer during or after chemotherapy in the USA. Liu et al reported that blockading cervical tumour cells with soluble PD-1 and then co-culturing with PBMCs, resulted in increased PBMC proliferation and CTL activity (Figure 1.12) [175]. Recently, a study indicated that depletion of ERAP1 in a mouse transplantable tumour model increased the efficacy of immune checkpoint inhibitor, specifically anti-PD1 immunotherapy

[176]. Women with high grade cervical dysplasia and advanced cervical cancer have increased expression of the vascular endothelial factor (VEGF) which is responsible for angiogenesis. Cancer cells rely on blood supply to grow and inhibition of VEGF would consequently inhibit cancer growth. An anti-VEGF antibody has been developed, bevacizumab traded under the name Avastin and has now been approved as an immunotherapy drug for treating persistent, recurrent or metastatic cervical cancer in Europe.

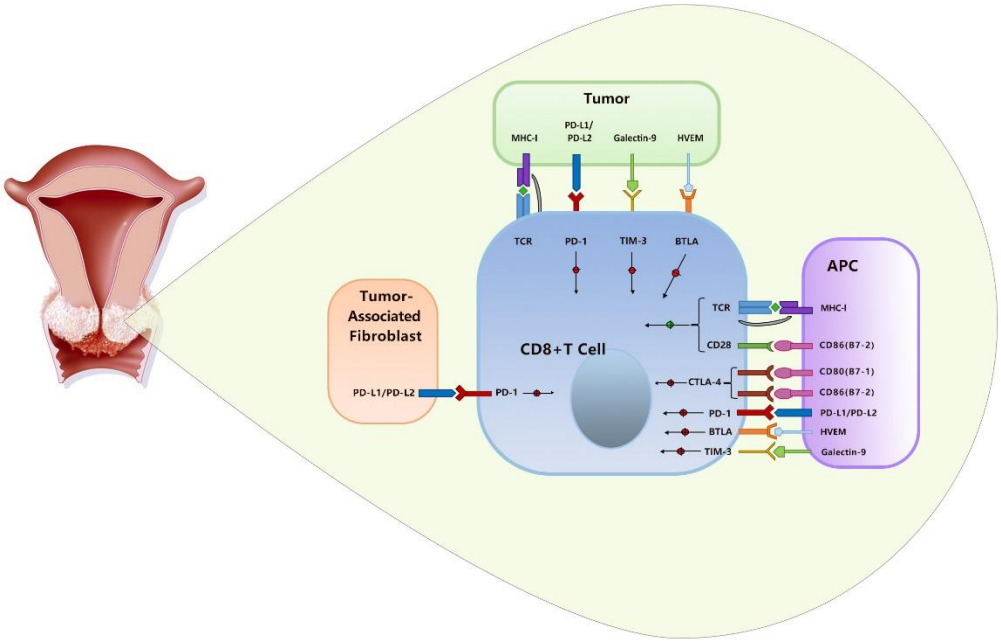


Figure 1.11. T cell interactions in cervical tumour microenvironment

Interaction between CD8+ T cell and APCs, tumour cells and tumour associated fibroblasts at the transformation zone of the cervix. The specific cell surface markers expressed on each cell type and their interacting ligand are shown, including the positive or negative response to each interaction on the T cell activation. Taken from Liu et al [4].

1.9 The role of ERAP1 in cervical cancer

Downregulation of ERAP1 and partial loss of HLA I was also associated with worse survival with ERAP1 downregulation (in 15% of samples) being an independent predictor of worse survival [177]. The

downregulation of ERAP1 may have served to reduce the tumour antigen pool preventing efficient activation of anti-tumour CTLs [169, 178] [121]. However, there were important factors that were not addressed in these studies, such as the effect of individual/combinations of SNPs on ERAP1 allotype trimming function and whether there is a relationship between trimming functions of distinct ERAP1 allotypes/allotype combinations and cervical carcinoma.

Since the majority of cervical cancer cases are caused by persistent HPV infections and increased infiltration of TILs is associated with increased survival, it is likely that HPV-derived antigens activate HPV-specific, peripheral and tumour infiltrating CTLs, which then exert their anti-tumour effector function [121, 166, 179]. The hypothesis that distinct ERAP1 allotypes in CIN and CSCC patients process HPV E6/E7 epitopes for presentation to CTLs, which can infiltrate the TME, is supported by the data generated in HPV+ oropharyngeal squamous cell carcinoma (OPSCC) [121]. In patients with low infiltration of CD8+/TILs, ERAP1 allotype pairs were shown to be poor at generating HPV-derived epitopes thus correlated with decreased CD8+/TIL infiltration and poor anti-tumour immune response, further associated with poor prognosis. The SNPs identified in the ERAP1 allotypes of HPV+ OPSCC patients were shown to affect the trimming function of the ERAP1 allotype combinations, however the trimming function of the ERAP1 allotypes identified in cervical cancer have not been studied [121]. OPSCC patients with a high number of tumour infiltrating CD8+/TILs were more efficient at trimming of an HPV-16 E7-derived epitope than the pairs in patients with a low number of CD8+/TILs [121]. The impact of the ERAP1 allotypes ERAP1*001 and ERAP1*014 (Hap7) on virus-specific CD8+ T cell epitope repertoire in an HLA-B*27:05 positive individual with acute hepatitis C virus infection was investigated [180]. It was shown that the ERAP1 allotypes, which are hypoactive, did not trim the N-terminally extended precursor of the immunodominant 9mer peptide to optimal length, while C-terminally extended 11mer peptide avoided destruction and was able to prime and activate atypical HCV-specific CD8+ T cell responses, thus virus clearance was prevented [180].

Homozygosity for a major allele at amino acid position 56 in ERAP1 and a minor allele at amino acid position 127 in ERAP1, was associated with decreased overall survival in cervical cancer patients [181]. The simultaneous presence of SNPs in ERAP1 at positions 127 and 730, in LMP7 at position 145 and in TAP2 at position 651, has been associated with 3-fold increase in cervical cancer risk [181]. The haplotype sequence that was significantly associated with increased cervical risk ($p < 0.001$) is (C-G)-(C-C), listed in the following order: ERAP1-127, ERAP1-730, TAP2-651, LMP7-145 [181]. The haplotype represents the presence of a minor allele at both ERAP1-127 and ERAP1-730 loci (amino acid changes R127P and Q730E, respectively) and the presence of a major allele at both TAP2-651 and LMP7-145 [181]. Interestingly, a research study investigating correlation between ERAP1 SNPs and cervical cancer risk, found that homozygosity for rs26618 (C/C genotype, I276M, odds ratio=1.53) may be associated with increased cervical cancer risk, contrary to the findings of Mehta et al [181, 182]. The reasons behind this discrepancy in findings include population genetic background (Chinese Han vs Dutch) and samples size (2890 vs 251). However, the effects of these SNPs on the function of the proteins produced, has not been explored.

1.10 Use of long read sequencing for ERAP1 allotyping from cervical cancer patient samples

Long read sequencing is an example of third generation sequencing using Nanopore technology and was patented in 2007 by Oxford Nanopore Technologies (ONT). Sequencing attempts began three years later and the first sequencer, the MinION was presented for the first time at the Advances in Genome Biology and Technology Conference in 2012. MinION became available to researchers through the MinION Access Program (MAP) in 2014 [183].

1.10.1 Nanopore principle

The MinION technology utilises nanopores which are proteins specifically designed with a wide hollow tube in the core, spanning a few nanometers in diameter. Nanopores are inserted in an electrically resistant polymer membrane bathed in an electrophysiological solution. By applying a potential across the membrane, an ionic current is generated through the nanopores. Single stranded molecules are sequenced as they pass through the nanopore base by base causing characteristic changes in the ionic current, generating a 'nanopore signal' (Figure 1.13, Figure 1.14). These changes in electrical current, resulting from DNA strands passing through the nanopore as well as around it, are measured and used to determine the base. Every nucleotide, while passing through the nanopore, obstructs the nanopore to a different, characteristic degree and the amount of change is characteristic for each different base. The electrical signature generated is measured at each pore several thousand times per second, with every single event being described by mean and variance of the current as well as event duration.

The process of converting raw data into nucleotide sequences is termed basecalling and a target DNA nucleotide sequence is called a read (Figure 1.13). The basecaller generates one read per target DNA strand, thus millions of reads can be generated from a single sequencing reaction. Sequencing coverage refers to the average number of reads that align to known reference bases. At high coverage levels, each base is covered by a high number of aligned reads therefore basecalls can be made with a high degree of confidence.

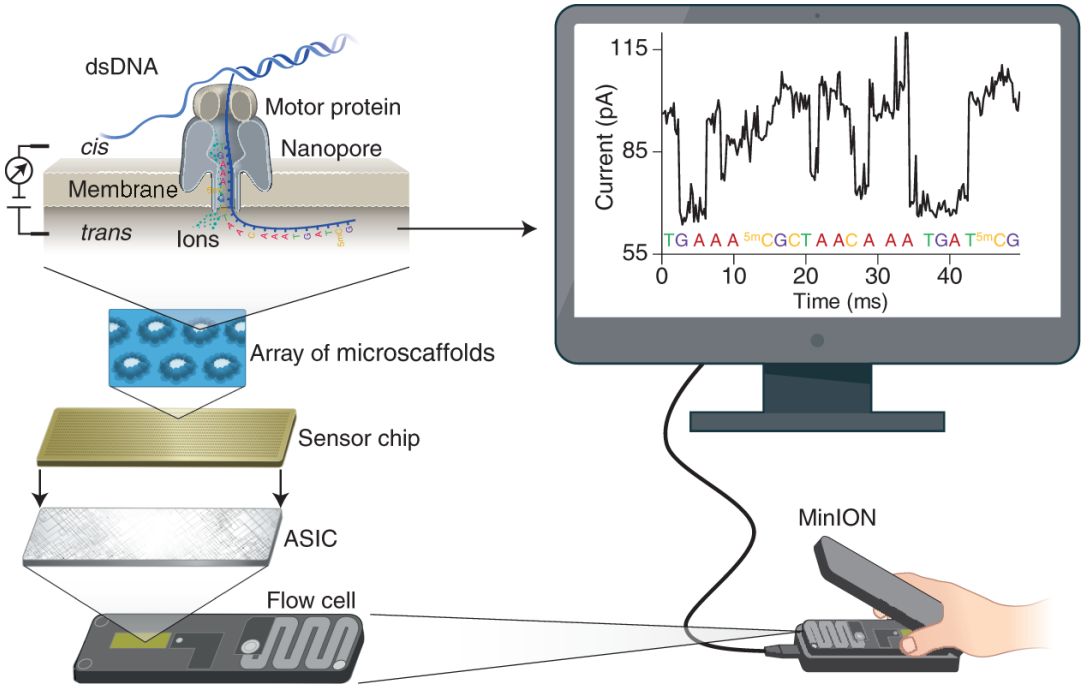


Figure 1.12: Long read sequencing of ERAP1 with MinION

MinION is a device developed by Oxford Nanopore Technologies used to carry out long read sequencing of genes such as ERAP1 here. The most important part of the MinION is the flow cell which includes the sensor chip bearing 512 individually addressed electronic channels, each connected to 4 sensor wells (total of 2048 wells). A non-conducting membrane with 2048 nanopores (proteins with a hollow core spanning a few nanometers in diameter) inserted into it, lies on top of the 2048 sensor wells. The nanopore signal is measured by an application specific integrated circuit (ASIC), which is capable of recording through all channels simultaneously, with the 2048 active well electrodes organised in groups of four. Nanopores are inserted in the membrane bathed in electrophysiological solution and by applying a potential across the membrane, an ionic current is generated through the nanopores. Single strands of DNA are sequenced base by base (bidirectionally) as they go through the nanopore causing characteristics changes in the ionic current, that is the nanopore signal. The process of converting raw signal from the disruptions that bases cause to the current as converted to a nucleotide sequence through a process called basecalling and is executed using the MinKNOW software. Taken from Wang et al 2021 [184].

1.10.2 The choice of long read sequencing for ERAP1 identification

The MinION is a portable device that can be used in various environments, ranging from field to the international space station. As it enables data acquisition in real time, with sequencing starting as soon

as the experiment begins and duration of sequencing at the user's discretion, it is a significantly time and effort saving sequencing method compared to traditional Sanger and Illumina sequencing, especially when there is a considerably large cohort of patient samples to be sequenced, as in this study (103 cDNA samples in the cervical cancer patient cohort). Long read sequencing has been developed to provide the user with full length reads of long fragments of DNA and even whole genome sequencing of microorganisms to viruses and sequencing of the reference genome of a human cell line has been performed successfully [185, 186]. Recently, ONT has developed an assay (LamPORE) to detect SARS-CoV-2, the virus that causes COVID-19, through the sequencing of three viral genes and a control target (available at https://community.nanoporetech.com/ accessed on 15/09/20). After sample collection, RNA extraction is carried out followed by isothermal amplification of SARS-CoV-2 E, N and ORF1a genes and actin control. After library preparation, the sample(s) are used for nanopore sequencing on MinION Direct sequencing of SARS-CoV-2 RNA has also been successfully carried out [187]. This specification of the technology is particularly advantageous for obtaining a vast number of reads of the 2.7kb ERAP1 allotypes and also reads that are not fragmented, therefore do not require 'stitching together', as in the case of Illumina sequencing, to form contiguous sequence. As the focus of this study is to identify SNPs in ERAP1 that result in non-synonymous amino acid changes related to ERAP1 trimming function, long read sequencing was chosen as the preferred sequencing method because it enables acquisition of full-length reads for identifying the two parental copies of ERAP1, both of which are important in determining the combined function of the enzyme in the cervical cancer patient cohort. Some of the advantages of long read sequencing constitute its ability to detect sequence deletions, translocations and inversions at dilutions as low as 1:100 with as few as 500 reads/sample. It was actually shown that HLA typing using long read sequencing was possible even with acquisition of poorly represented data (only 80 reads) [188].

1.10.3 Long read sequencing use in literature relevant to this research project

Since its release to early access users in 2014, long read sequencing has already been used in multiple research fields ranging from cancer research to immunology and was even used in space for the sequencing of bacterial, viral and animal libraries. Notably, Jain et al, used long read sequencing for the sequencing and assembly of a cell line genome [186]. The results of this study revealed the generation of ultra-long reads approximately 6.3Mb long and with an N50 882Kb long. Half the genome is covered by contigs equal or larger than N50 [186]. The theoretical genome coverage was 30x i.e the average number of reads that align or cover known reference base positions that constitute the cell line genome. If only the ultra-long reads were taken into consideration, the genome coverage was 5x. Interestingly, the assembly included a single 3Mb-long contig that contained all the HLA I genes from the MHC region on chromosome 6. However, the more repetitive HLA II gene locus was shown to be fragmented, but most genes were present in a single locus [186]. In a study of the HLA-B locus in Māori and Pacific Island samples, long read sequencing was used to sequence 943bp-long PCR amplicons from 49 DNA samples; long read sequencing data were compared with those obtained with either Sanger sequencing or data developed from the 1000 genomes project. Results indicated that approximately 5,854 reads were generated per sample and the average read length was indeed close to the one expected for the region spanning exon 2 and 3 of the HLA-B locus. Interestingly, even though the manufacturer recommended the use of 2nM DNA for barcoding, amplification and acquisition of reads was possible even at 0.19 nM. Moreover, it was shown that just 80 reads were used to generate sufficient data for HLA-typing [188]. Consistency of long read sequencing data were confirmed when compared to previously acquired data using other commonly-used sequencing methods [188]. These results regarding i) DNA concentration loading on the flow cell, ii) minimum number of reads that leads to successful HLA-typing, iii) reads generated per sample given that simultaneous sequencing of multiple samples was also used in the present research, as well as iv) confirmation of long read sequencing accuracy with Sanger sequencing data, were of great interest for the methodology developed in the present study.

Interestingly, long read sequencing was also used for the detection and analysis of HPV and microbiome status of a single cervical liquid-based cytology sample [189]. This study involved combination of multiplex PCR targeting HPV-16 E6 and E7 as well as full-length 16S ribosomal RNA and sequencing with MinION to investigate HPV infection and vaginal microbe composition. Data were confirmed by comparison with Illumina data, without obtaining an error rate, showing a similar microbiome. This methodology could potentially be used in the future to detect HPV detection and analyse the microbes present, guiding treatment approaches [189].

1.11 Aims and hypothesis

Individual ERAP1 SNPs are associated with increased cervical cancer risk as well as overall survival in GWAS studies [181, 190], however a cause and effect relationship between ERAP1 allotypes and cervical cancer prognosis has not been established. This study aimed to identify ERAP1 allotypes from 103 patients at varying stages of cervical cancer. The hypothesis tested is that ERAP1 allotypes are associated with cervical cancer prognosis and this is achieved through the editing of peptides which are presented to CTL that infiltrate the tumour microenvironment and exert anti-tumour immune responses. The function of the ERAP1 allotypes on the generation of peptides will also be investigated to determine whether an association exists between disease outcome and ERAP1 function.

The aims of this research were:

- To identify ERAP1 allotypes/combinations in a cohort of 103 cervical cancer patient samples to examine a correlation between ERAP1 identity and disease progression and/or clinical phenotype.

-To examine the impact of ERAP1 allotype/combination trimming properties on the generation of a well-characterised OVA-derived epitope and their effect on stimulating antigen-specific CTLs.

-To identify HLA-A*0201 positive patients and examine the impact of their ERAP1 allotype combinations trimming properties on the generation of an HPV-derived epitope and their effect on stimulating antigen-specific CTLs.

2 Materials and methods

2.1 Cervical cancer patient cohort

The cervical cancer cohort used in this study was acquired from the University of Groningen which has obtained the relevant ethical approvals (Marco de Bruyn, Joyce Lubbers, University Medical Center Groningen – UMCG Obstetrics & Gynaecology). cDNA samples were generated from RNA extracted from cervical scrapings by researchers at the University of Groningen from a cohort of 103 cervical cancer patients consisting of squamous cell carcinoma, adenocarcinoma, adenosquamous carcinoma and other non-disclosed types of carcinoma. On average, a total of 12µl of sample (cDNA) was provided for each patient.

2.2 ERAP1 cloning and sequencing

2.2.1 RNA isolation

HeLa and 293T cells were harvested, counted and washed in phosphate buffer saline (PBS) before RNA purification was carried out using Zymo Quick RNA miniprep kit (Zymo Research Corp, California USA) according to manufacturer's instructions, eluting in 30µl RNase free H_2O. The concentration of purified RNA was measured using Qubit 4 fluorometer (Invitrogen) with the Qubit RNA BR (Broad-Range) assay kit (Thermo Scientific).

2.2.2 cDNA synthesis

Transcriptor High Fidelity cDNA synthesis kit (Roche) was used to synthesise cDNA using 100-500ng purified RNA according to manufacturer's instructions. Briefly, random hexamer primers (60µM final) were added to RNA, and the template-primer mixture was denatured by heating the tube for 10 minutes at 65^{o}C before cooling at 4^{o}C. A final 20µl reaction containing of 8mM $MgCl_2$, 1x reaction buffer, 20U protector RNase inhibitor, 1mM deoxynucleotide mix, 5mM dithiothreitol, and 22U reverse transcriptase was heated at 29^{o}C for 10 minutes followed by 60 minutes at 48^{o}C, then 5

minutes at 85°C before transferring to 4°C. cDNA yield was determined using Nanodrop spectrophotometer 1000 (Thermo Scientific) and was used for polymerase chain reaction (PCR) or stored at -20°C for later use.

2.2.3 ERAP1 amplification by PCR

PCR was carried out with 1µl cDNA as template using KOD Hot Start DNA polymerase (Merck). PCR primers to incorporate a 5' EcoRI restriction site and a 3' XhoI site in ERAP1 were used in a 50µl reaction (Table 2.1, Table 2.2, Table 2.3). PCR conditions were optimised to successfully produce a single, discrete PCR amplified ERAP1 product of 2.7kb length, confirmed by running PCR reactions on a 1% agarose electrophoresis gel. Unsuccessful PCR reactions were repeated to ensure issues are not of technical nature, along with a negative control (no template reaction) and a positive control (ERAP1 expression previously verified). If a single, discrete PCR band was not detected on the agarose gel, a second PCR reaction was performed using 1µg amplified ERAP1 from the first PCR reaction as template and primer set 4 (Table 2.1, Table 2.2, Table 2.3).

Table 2.1: Primer sets used for ERAP1 amplification by PCR

Primer Pair Number	Restriction site	Sequence
#1 5'	EcoRI. 2.7Kb	GACGAATTCATGGTGTTTCTGCCCCTCAAATG
#1 3'	XhoI. 2.7Kb.	GACCTCGAGCATACGTTCAAGCTTTTCAC
#2 5'	EcoRI. Internal.	GACGAATTCGCCCCTCAAATGGTCCCTTGCAAC
#2 3'	XhoI. Internal	GACCTCGAGCACTTTGCAGCCACACTCTG
#3 5'	XhoI. 2.7Kb.	GACCTCGAGATGGTGTTTCTGCCCCTCAAATGGTC
#3 3'	BamHI. 2.7Kb	GACGGATCCCATACGTTCAAGCTTTTCACTTTG
#4 5'	EcoRI. 2.6Kb	GACGAATTCCTAACTTTCCTCACTGTTGGCTC
#4 3'	XhoI. 2.6Kb	GACCTCGAGCTTATCCATCCAACCGATG
#5 5'	Blunt. 2.7Kb	ATGGTGTTTCTGCCCCTCAAATGGTCCCTTGC
#5 3'	Blunt. 2.7Kb	CATACGTTCAAGCTTTTCACTTTGCAGCC

Table 2.2: PCR components used for amplification of ERAP1

PCR components	Final concentration	Volume
10x buffer	1x	5µl
25mM MgSO4	1.5mM	4µl
dNTPs 2mM each	0.2mM each dNTP	5µl
10µM sense 5' primer	0.3mM	1.5µl
10µM anti-sense 3' primer	0.3mM	1.5µl
Template cDNA	100-250ng	1µl
KOD hot start DNA polymerase	0.02U/µl	1µl
PCR grade H_2O		31µl
Total		50µl

Table 2.3: PCR cycling conditions

Step	Cycling conditions
1. KOD Hot Start polymerase activation	95°C for 2 minutes
2. Denaturation	95°C for 20 seconds
3. Annealing	55°C for 10 seconds
4. Extension	70°C for 80 seconds
Repeat steps 2-4	35 PCR cycles

2.2.4 Cloning ERAP1 into pCR-Blunt II-TOPO vector

Amplified ERAP1 DNA products from 293T and HeLa were ligated into pCR-Blunt II-TOPO vector (Life Technologies) using Zero Blunt TOPO PCR Cloning Kit (Life Technologies) with a molar ratio of 3:1 ERAP1 insert to vector, along with 1µl salt solution (final concentration: 200mM NaCl, 10mM $MgCl_2$) and PCR grade water, in a final reaction volume of 6µl. The ligation reaction was incubated for 5 minutes at 22°C and placed on ice before bacterial transformation.

2.2.5 Bacterial transformation

Chemically competent *E.coli* cells (TOP10 One Shot Competent cells, Life Technologies) were transformed using ligated ERAP1 in the TOPO vector. Site directed mutagenesis (SDM) was used to introduce mutations into the wild type ERAP1, *002-WT, previously cloned into pcDNA3 vector (section 2.5.3). JM109 cells were then transformed using the constructs. 2µl of the ligation reaction (section 2.2.4)/1µl of the SDM product was added to bacterial cells (TOP10/JM109) and incubated on ice for 30 minutes. Next, the bacteria were heat-shocked at 42°C for 30 seconds and immediately transferred to ice before the addition of 250µl SOC media (Table 2.4), and incubation at 37°C for 1 hour with shaking at 220rpm. Approximately 50µl of transformed bacteria were plated onto agar plates (kanamycin-containing agar plates for TOP10 cells, 50µg/ml and ampicillin-containing agar plates for JM109 cells, 100µg/ml), before being incubated at 37°C overnight.

Table 2.4: LB and SOC medium components

Medium	Component
LB	0.5% yeast extract
	2% Tryptone
	10mM NaCl
SOC	2.5mM KCl

10mM $MgCl_2$
10mM $MgSO_4$
20mM Glucose

2.2.6 Screening of bacterial colonies

Two to six bacterial colonies were selected and incubated in 2ml LB medium containing kanamycin (TOP10, 50µg/ml) or ampicillin (JM109, 100µg/ml) at 37°C with 220rpm shaking for 16-18 hours. Plasmid DNA was isolated from 1.5ml bacterial culture using QIAprep Spin Miniprep kit (Qiagen) according to manufacturer's instructions and eluted in 30µl dH_2O.

2.2.7 Digestion with restriction enzymes

Single digest of plasmid DNA with the restriction enzyme EcoRI was used to determine whether ERAP1 was successfully cloned into the TOPO vector. Standard 20µl reactions consisting of 13µl dH_2O, 1µl of 1x 2.1 buffer (New England Biolabs), 5µl DNA, 1µl EcoRI (12units/µl; Promega) were set up and incubated at 37°C for 1 hour. To confirm the presence of ERAP1 in the TOPO vector, 10µl of each reaction were subsequently run on 1% agarose electrophoresis gel followed by Sanger sequencing.

2.2.8 Sanger sequencing

Miniprep DNA samples containing the ERAP1 insert were identified and sent for sequencing at SourceBioscience Sequencing Ltd (Nottingham site, UK) in order to determine the ERAP1 sequence using specific sequencing primers. DNA samples containing the ERAP1 insert (2.8Kb) were sent for sequencing for the identification of the two ERAP1 allotypes as in section 2.2.8 but using different forward and reverse primers (Table 2.5, Table 2.6). The translate tool 'ExPasy – Bioinformatics Resource Portal' (available online at https://web.expasy.org/translate/) was first used to translate the nucleotide sequence into the corresponding amino sequence of ERAP1, and then the Basic Local Alignment Search Tool (BLAST, available online at https://blast.ncbi.nlm.nih.gov/Blast.cgi U.S National

Library of Medicine) was used to compare the ERAP1 amino acid sequences obtained from cell lines with the wild type amino acid sequence of ERAP1 [89].

Table 2.5: Primers for sequencing of ERAP1 in pcDNA3 plasmid vector

Name	Target sequence	Tm (°C)
T7 forward	TAATACGACTCACTATAGGG	48
SP6 reverse	ATTTAGGTGACACTATAG	41
hERseq2	GGATGCTGCGGTGACTCTTCTAG	68.2

Table 2.6: Primers used for sequencing of ERAP1 in pCR-Blunt II-TOPO vector

Name	Target sequence	Tm (°C)
M13 forward	TGTAAAACGACGGCCAGT	48
M13 reverse	CAGGAAACAGCTATGACC	48
hERseq2	GGATGCTGCGGTGACTCTTCTAG	68.2

2.3 Long read sequencing using MinION by Oxford Nanopore Technologies

2.3.1 Computer requirements for long read sequencing

Long read sequencing using MinION (ONT, Oxford) and analysis required the following: a Dell Precision 3520 laptop (Operating system: Windows 10 Pro for workstations, 64 bit, Memory/RAM: 16GB, CPU: i5 8th generation), a MinION Mk1B device, a USB 3.0 cable, a configuration flow cell and R9.4.1 flow cells (ONT). MinION was plugged directly into a standard USB 3.0 port on the laptop computer. The device is controlled by MinKNOW^{TM} software (current version 20.06.5 available at https://nanoporetech.com) with a graphical user interface (GUI). Following raw data acquisition and basecalling by MinKNOW^{TM}, the data analysis platform Epi2me (current version 2020.2.10, available at https://nanoporetech.com) also downloaded as a desktop agent with a GUI, enables real time analysis through bioinformatics workflows such as alignment of acquired reads to a reference sequence and Fastq barcoding for simultaneous sequencing of multiple samples.

2.3.1 MinION compartments

The MinION has a *priming port* where the library is loaded before samples. The most important part of the MinION is the *flow cell* which includes the sensor chip bearing 512 individually addressed electronic channels, each connected to 4 sensor wells (total of 2048 wells). The non-conducting membrane with 2048 nanopores inserted into it, lies on top of the 2048 sensor wells. The nanopore signal is measured by an application specific integrated circuit (ASIC), which is capable of recording through all channels simultaneously, with the 2048 active well electrodes organised in groups of four. Multiplexing is defined as the choice of a single well for each of the channels and the 'MUX selection' which involves selecting the best 512 wells chosen as the first group is performed as part of the sequencing script. Other wells can be switched to during sequencing to increase yield. A *waste port* is also found on the MinION for discarding the samples after sequencing (Figure 2.1).

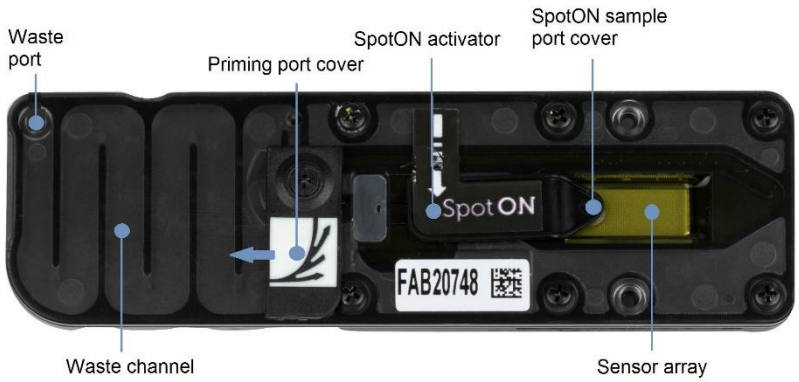


Figure 2.1. The compartments of the MinION

The flow cell containing a sensor chip, a sample port, a priming port and a waste port. A USB3.0 cable connects the sequencing device to a computer for data acquisition (Acquired through the Oxford Nanopore Technologies website available at: **https://community.nanoporetech.com/**)

2.3.2 MinION configuration test and flow cell check

Prior to the first sequencing experiment using MinION, a configuration test cell was used in the MinION device to ensure suitable data exchange between the device and the laptop. Prior to all sequencing runs, a flow cell check was undertaken to verify the quality of the flow cell used by determining the number of available and active pores for sequencing. The SpotON R9.4.1 flow cell

(FLO-MIN106, ONT >800 pores), was equilibrated at room temperature for 5 minutes to allow stable increase of the flow cell temperature. Next, the flow cell was inserted into the MinION device and connected to the laptop. After launching the MinKNOW GUI, the flow cell type FLO-MIN106 was selected from the panel, before the option "Check flow cells" was selected and run as soon as the temperature of 37^{o}C was reached (for current version the set temperature to be reached was 34^{o}C). Once completed, the quality of the flow check was indicated, either showing i) the check was not completed during the session, ii) that the number of available pores was above the warranty of 800 pores (out of possible 2048) and was ready for a sequencing run, or iii) the number of available pores was below warranty and consequently the flow cell was deemed not suitable for sequencing.

2.3.3 Sequencing amplicons with MinION

2.3.3.1 ERAP1 amplification by PCR using normal or tailed primers

ERAP1 was amplified from HeLa and 293T cDNA as explained in section 2.2.3. The ERAP1 amplicons were then used in experiments involving the sequencing of a single ERAP1 amplicon to develop a methodological pipeline.

Regarding sequencing of multiple samples with MinION, a different protocol was used for preparing the library before DNA repair and clean-up in section (PCR barcoding amplicons vAugust 2019), ERAP1 was amplified from cDNA from a cohort of ninety-four cervical cancer patients by PCR using KOD Hot Start DNA polymerase (Merck). PCR primers specific for ERAP1 were used (primer set #5) (Table 2.1, Table 2.2,Table 2.3) tailed with the universal sequences indicated below by Oxford Nanopore Technologies in a 50ul reaction. In brackets is the sequence of the ERAP1 primers normally used for amplification as in section 2.2.3.

5' TTTCTGTTGGTGCTGATATTGC-[ATGGTGTTTCTGCCCCTCAAATGGTCCCTTGC] 3'

5’ ACTTGCCTGTCGCTCTATCTTC-[CATACGTTCAAGCTTTTCACTTTGCAGCC] 3’

PCR conditions were optimised (annealing temperature increased to 60^{o}C) to successfully produce a single, discrete ERAP1 amplicon of 2.7kb length, confirmed by visualising amplicons on a 1% agarose electrophoresis gel.

ERAP1 amplicons were quantified using Nanodrop and, for later experiments, a Qubit^{TM} dsDNA HS Assay kit (ThermoFisher Scientific) on Qubit was used to measure DNA concentration.

2.3.3.2 Barcoding PCR

To sequence multiple patient ERAP1 in the same sequencing run, ERAP1 amplicons required attachment of barcodes to each individual patient ERAP1 amplicon. Approximately 100 fmol of each patient ERAP1 amplicon was used in a barcoding PCR reaction with cycling conditions optimised based on ERAP1 length (Table 2.7, Table 2.8). DNA was quantified using Qubit and all barcoded amplicons were pooled in the desired ratios to achieve 1µg of pooled barcoded libraries was prepared in 47µl nuclease-free water.

Table 2.7: Barcoding PCR components

Barcoding PCR components	Concentration	Volume	Manufacturer
PCR barcode	10µM	2µl	ONT
First-PCR product	100-200fmol	48µl	ONT
LongAmp Taq 2x master mix		50µl	NEB
Total		100µl	

Table 2.8: Barcoding PCR cycling conditions

Barcoding PCR steps	Cycling conditions
1. Initial denaturation	95°C for 3 minutes
2. Denaturation	95°C for 15 seconds
3. Annealing	62°C for 15 seconds
4. Extension	65°C for 2 minutes and 42 seconds
5. Final extension	65°C for 2 minutes and 42 seconds
Repeat steps 2-4	15 PCR cycles

2.3.3.3 DNA repair and end-prep

Amplicon DNA was transferred into a 1.5ml Eppendorf DNA LoBind tube, adjusting the volume to 47µl with nuclease-free water. In a 0.2 ml thin-walled PCR tube, a standard 60µl DNA repair and end-prep reaction was set up and incubated on a thermal cycler at 20°C for 5 minutes followed by 65°C for 5 minutes (Table 2.9). The reaction was stored on ice before clean-up with AMPure XP beads.

Table 2.9: DNA repair and end prep components

Components	Volume
DNA control strand (DCS; SQK-LSK109 kit, ONT)	1µl
Amplicon DNA	47µl
NEBNext Formalin-Fixed, Paraffin-Embedded (FFPE) DNA repair buffer (NEB)	3.5µl
NEBNext FFPE DNA repair mix (NEB)	2µl
Ultra II End-prep reaction buffer (NEB)	3.5µl
Ultra II End-prep enzyme mix (NEB)	3µl
Total	60µl

2.3.3.4 AMPure XP bead clean-up

The Agencourt AMPure XP magnetic bead clean-up (A63880, Beckman Coulter Genomics, Brea, CA) was used to purify the amplified end prep reaction. 200µl AMPure XP beads were incubated at room temperature for 30 minutes and vortexed briefly. The end-prepped DNA sample was transferred into a 1.5 ml Eppendorf tube and purified by incubation with 60µl (1x) AMPure XP beads on a Hula mixer for 5 minutes, followed by incubation on a magnetic rack for 5 minutes. The supernatant was discarded and the remaining bead-DNA complexes were washed twice with 250µl of freshly prepared 70% ethanol. Beads were air-dried for ~45 seconds before re-suspending them in 61µl nuclease-free water and incubating for 2 minutes at room temperature. The tube was placed back on the magnet until the eluate was clear and colourless. The eluate was transferred in a clean 1.5ml Eppendorf tube that could be stored at 4°C overnight. Concentration of DNA was determined using Qubit and the dsDNA High Sensitivity kit.

2.3.3.5 Adapter ligation and clean-up

To ligate the ONT-specific sequencing adapters to the end-prepped DNA, a 100µl reaction consisting of 60µl DNA sample, 25µl Ligation Buffer (LNB, SQK-LSK109 kit, ONT), 10µl NEBNext Quick T4 DNA ligase (NEB) and 5µl adapter mix (AMX, SQK-LSK109 kit, ONT) was set up and incubated at room temperature for 10 minutes. A total of 40µl of re-suspended AMPure beads were added to the reaction and incubated on a Hula mixer at for 5 minutes at room temperature. The tube was then transferred on a magnet and the supernatant was discarded without disturbing the pellet, which was washed twice with 250µl long fragment buffer (LFB, SQK-LSK109, ONT). The pellet was air-dried for ~30 seconds before re-suspending in 15µl elution buffer (EB, SQK-LSK109, ONT) and incubating at 37°C for 10 minutes. The tube was placed back on the magnet and the eluate was transferred in a clean 1.5ml Eppendorf tube and the concentration of the prepared library was determined by Qubit. The prepared library was stored on ice until ready to load into the flow cell.

2.3.3.6 Priming and loading the SpotON flow cell

Using a P1000 Gibson pipette, approximately 20-30µl of storage buffer was removed from the priming port of the flow cell to prevent bubbles from entering the sensor chip and damaging the pores, while ensuring that the wells are covered with buffer at all times. The flow cell priming mix was prepared by mixing 30µl flush tether (FLT, SQK-LSK109, ONT) with a tube of flush buffer (FB, SQK-LSK109, ONT). 800µl of priming mix was gradually loaded into the flow cell via the priming port, followed by a 5 minute incubation at room temperature to equilibrate. The DNA library was prepared in a clean Eppendorf by adding 37.5µl sequencing buffer (SQB, SQK-LSK109, ONT), 25.5µl loading beads (LB) and 12µl DNA library (recommended loading between 5-50fmol). 200µl of the priming mix was loaded into the flow cell via the priming port while the sample port remained open, followed by loading the prepared library dropwise via the sample port. Both sample and priming ports were closed and a new experiment was selected from the MinKNOW GUI.

2.3.3.7 Wash protocol

For the sequencing of the cervical cancer patient samples, the wash protocol was carried out with the relevant kit (Flow cell wash kit EXP-WSH003, protocol version WFC_9088_v1_revE_18Sep2019). Solution A was placed on ice and one tube of solution B was thawed at room temperature, vortexed, mixed briefly and placed on ice too. In a clean 1.5ml Eppendorf tube the washing mix consisting of 20µl wash solution A and 380µl wash solution B was prepared and mixed by pipetting. The sequencing run was stopped from the MinKNOW GUI, the priming port was opened and all waste was removed from waste port 1. 400µl of washing mix was loaded to the flow cell through the priming port and a 30-minute incubation at room temperature followed. If a second library was to be loaded immediately, the flow cell was primed and loaded as in section 2.3.3. If the flow cell was to be stored for later use, one tube of storage buffer (S) was thawed, mixed by pipetting and centrifuged. 500µl S

buffer were added through the priming port which was then closed to remove all fluid through waste port 1. The flow cell was stored at 4^{o}C for later use.

2.3.4 Basecalling

To sequence the prepared library, MinKNOW^{TM} software consisting of a repository of Python scripts to control the MinION device during the sequencing run based on user's input, was launched and the relevant ONT library preparation kit was selected. The basecall model selected was 'Fast basecalling' and the barcoding option was turned on to enable barcoding of basecalled data. The voltage, measured in mV, was adjusted based on the voltage reading at the end of the previous run with the starting voltage of a new flow cell being -180mV. The duration of the sequencing run was set to 12 hours, however sequencing was manually stopped once the number of reads reached between 30K and 100K. Active channel selection was turned on for the software to switch to a new channel if a channel was deemed to be in a "saturated" or "multiple" state. The time between multiplex scans (mux scans) was set as 1.5 hours. The raw data generated by the MinION device were processed by MinKNOW which performed live basecalling with read files being generated in real time, as soon as the sequencing run began using a Recurrent Neural Network (RNN). The output settings retrieved data in .fast5 format, with 4000 reads to be written out per file, and in .fastq format with 4000 files written into a single folder. Fast5 format, based on .hdf5 file type, contained raw read data used for re-basecalling while fastq files contained both nucleotide sequences and quality scores. Following completion of basecalling, read files were organised 3 categories: *'pass'* for reads produced and basecalled, *'skip'* for reads generated but not basecalled yet and *'fail'* for produced and basecalled reads but with basecalling not successfully completed. Each barcode demultiplexed by the basecaller Guppy used by MinKNOW had its own folder containing the relevant fast5 and fastq files, which were used for further data analysis in Epi2me. The report generated by MinKNOW at the end of the sequencing run was exported in PDF format.

2.4 Long read sequencing analysis pipeline

2.4.1 Epi2me

Epi2me is a cloud-based data analysis platform offering a number of bioinformatics workflows for analysis of MinION sequencing data. Basecalled read files were uploaded to the Epi2me platform via the desktop agent. In the case of single amplicon sequencing, the workflow 'Fasta reference upload' was selected and the ERAP1 reference sequence (KM357887.1, Homo sapiens endoplasmic reticulum aminopeptidase 1 (ERAP1) mRNA, ERAP1-002:01 allele available at: https://www.ncbi.nlm.nih.gov/nuccore/KM357877.1, [89]) was uploaded. The workflow 'Custom human alignment' was used to map reads to ERAP1 and the output file was .BAM format. Workflow parameters included: min qscore of 7, no detection of barcode, no splitting of barcodes into separate folders, disabled Epi2me storage and confirmation of the data uploaded were of human origin. In general settings, the type of input file was .fastq and the input directory was set as the file location where data generated by MinKNOW for that sample were stored. Accelerated transfer was turned off and download mode was set as 'data and telemetry'. In attributes, the project name was indicated as 'ddmmyy' (date of the sequencing run). In the case of multiple amplicon sequencing, the Epi2me workflow run was 'Fastq barcoding'. Workflow parameters remained as in single amplicon sequencing but with automatic detection of barcoding and splitting of barcodes into separate folders. Output files were in .fastq format with barcodes trimmed from the reads which were demultiplexed by barcode. The Epi2me report generated with information including read length and number was stored in PDF format.

2.4.2 Bioinformatics analysis pipeline

In order to run the script for identifying the ERAP1 allotypes from each patient sample, the input data found on the table below were required (Table 2.10).

Table 2.10: Input files required for running the custom script for the identification of ERAP1 allotypes of the cervical cancer patient samples.

File	Format	Details
Reference genome	fasta	2,823bp-long ERAP1, coding/no introns
File with fast5s/sample	fast5	generated by MinKNOW™
File with fastqs/sample	fastq	generated by Epi2me platform
Sites	vcf	formatted file with details of the 10 sites over which
		the haplotypes/allotypes are defined
Sites	bed	same as vcf file, different format
Haplotypes	txt	list of haplotype/allotypes
		and corresponding SNPs

The software required included: NanoPack (version 1.13.0, https://github.com/wdecoster/nanopack) specifically NanoPlot and NanoFilt, was used to assess the quality of basecalled and demultiplexed fastq files generated by Epi2me for each sample based on length and plot [191, 192]. The length filter deemed best for data analysis was between 2,500 and 3,000 base pairs (based on the ERAP1 reference sequence length: 2,823bp). Minimap2 (version 2.17, https://github.com/lh3/minimap2) was used to map reads to Homo sapiens endoplasmic reticulum aminopeptidase 1 (ERAP1) mRNA, ERAP1-002:01 allele, KM357887.1 of a length of 2823 base pairs (https://www.ncbi.nlm.nih.gov/nuccore/KM357877.1) [89, 193]. Samtools (version 1.6, using htslib 1.6, Genome research Ltd) was used to convert the file generated by minimap2 into a bam file, sorted and indexed which contained all the reads mapping to ERAP1 and could be visualised using IGV (version 2.3.88 and Java version 1.7.0_21).

Nanopolish (version 0.11.1, https://github.com/jts/nanopolish) was used for variant calling, first by calling all variants present and second by calling the variants identified only across the 10 sites of interest which correspond to the 10 SNPs that have been identified in the ERAP1 sequence (Sean Eddy public domain code, code originally part of hmmer3, ONT scrappie basecaller, licenced under MPL).

Picard was used to replace groups in the BAM file generated earlier by Samtools to match the relevant sample name. Whatshap (version 0.17, https://whatshap.readthedocs.io/en/latest/) was used for phasing on the file that contained all the variants identified (vcf format) and on the file that contained only the variants identified across the 10 sites [194]. Whatshap was then used to tag haplotypes in the new bam files based on phasing. Bcftools (version 1.6 using htslib 1.6) was used to create consensus fasta versions for each haplotype identified and as SNPs might not be in the same locations in these fasta files, Bedtools (version 2.26.0) was used to extract the base for each of the 10 sites over which SNPs occur. Finally, an output file containing the two allotypes/haplotypes for each sample as well as a sequence of all the bases identified at the 10 sites over which SNPs occur for the two allotypes, was generated in .txt format. The output files generated using the custom script for each sample, can be found below (Table 2.11).

All the above software was run in a python environment (version 3.5.5). Conda (version 4.4.0) was loaded for the use of NanoPack, singularity container (version 3.1.0) loaded for NanoPolish and biobuilds (2017.11) loaded for running the rest of the analysis software.

Table 2.11: Custom script output files

File	Format
Combined fastq	FQ
Report on fastq files and associated txt, log and png files	HTML
Combined fastq after filtering on length	FQ
Report on fastq after filtering on read length	HTML
Reads mapped to reference and sorted	BAM and BAI
Reads mapped to reference and sorted with haplotype tag	BAM and BAI
Index files for cross referencing fastq and fast5 files	INDEX
Variant calls across whole region, containing only	VCF
heterozygotes and alternative homozygotes	
Variant calls only for the 10 sites of interest, and reference	VCF
homozygote calls where coverage allows it.	
Phased variant calls across whole regions	GZ and TBI
Phased variant calls only across the 10 sites with minimum 2 heterozygotes	GZ and TBI
File giving locations of haplotype blocks called across whole regions:	GTF
1 or empty if not phased.	
File giving locations of haplotype blocks called only across the 10 sites:	GTF
1 or empty if not phased.	
Files incorporating all variants into reference sequence for each	FASTA and FAI
haplotype/allotype	
Files incorporating the 10 site variants into reference sequence for	FASTA and FAI
each haplotype/allotype	
Output file containing genotypes, phased (if >1 heterozygote),	TXT
haplotype/allotype and combination of nucleotides identified across the 10 sites	
Standard error output from the PBS queue	PBS
Standard error output from the PBS queue with job info such as walltime	PBS

2.5 Generation of identified ERAP1 allotype plasmid constructs

2.5.1 Cloning ERAP1 allotypes from cell lines into pcDNA3 plasmid vector

Upon identification of the ERAP1 allotypes in the cell lines, PCR was carried out using KOD start DNA polymerase, miniprep of cloned cDNA amplicon as the template (section 2.2.4) and ERAP1 specific primer set #1 (Table 2.1), with thermocycling conditions as described in section 2.2.3. The PCR product was run on a 1% agarose electrophoresis gel at 100V for 1 hour. The pcDNA3 vector (Life Technologies)

was digested with EcoRI and XhoI at 37°C and run on 1% agarose gel (section 2.2.7). Next, resulting bands corresponding to the pcDNA3 vector (5.4Kb) and the ERAP1 insert (2.5Kb) were excised, extracted using QIAquick Gel Extraction Kit (Qiagen) according to manufacturer's instructions, and eluted in 30µl dH_2O. Single digests of pcDNA3 with EcoRI or XhoI were carried out as control reactions and run on 1% agarose electrophoresis gel at 100V for 1 hour for comparison.

2.5.2 DNA ligation

The concentration of purified pcDNA3 and ERAP1 DNA were measured using Nanodrop. Ligation of ERAP1 PCR product into the digested pcDNA3 vector was carried out by using a molar ratio of 3:1 of DNA to vector, with 150ng vector used in all reactions. 1µl T4 DNA ligase and 1.5µl 10x ligase buffer (New England Biolabs) were used in a 15µl ligation reaction which was incubated overnight at 16°C. Bacterial transformation, screening of bacterial colonies and sequencing were completed as explained in sections 2.2.5, 2.2.8).

2.5.3 Site directed mutagenesis (SDM)

The WT allotype identified from the patient cohort using long read sequencing was the non-mutated ERAP1 allotype *002-WT that following the trimming of N-terminally extended antigenic peptide precursors, leads to generation of an optimal response by both the B3Z and the BE7A2Z T cell hybridoma. This allotype was cloned in the pcDNA3 vector for use in transfection into 293TE1KO cells (section 2.2.7 and section 2.7). To investigate the functional role of the ERAP1 allotype combinations identified from the patient cohort, the SNPs that result in amino acid changes in the allotypes were introduced into the wild type allotype using site directed mutagenesis. Primers were designed specifically for a PCR to incorporate single base changes into the wild type allotype plasmid DNA. All primers were designed to have a GC content of 55-60% and a melting temperature of 65-88°C and were of approximate length range of 25-45 bases long (Table 2.10). The primers used for sequencing were the previously mentioned hERAAPseq2, SP6 and T7F which were provided by SourceBiosciences.

SDM PCR was carried out using the KOD Hot Start DNA Polymerase kit, with both the components and the cycling conditions being presented in Table 2.13 and Table 2.14. Following completion of the SDM PCR, PCR products were digested with 1µl Dpn1 (10units/µl, Promega) for 1 hour at 37°C in order to digest methylated adenine residues in the parental DNA.

Table 2.12: Primers designed for site directed mutagenesis

Primer Pair Number	Sequence	Tm°C
E320A 5'	CTGGTGCTATGGCAAACTGGGGACTG	75.2
E320A 3'	CAGTCCCCAGTTTGCCATAGCACCAG	75.2
E56K 5'	AATACGACTTCCTAAGTACGTCATCCC	
E56K 3'	GGGATGACGTACTTAGGAAGTCGTATT	
R127P 5'	CTGGAACACCCCCCTCAGGAGCAAATT	78.4
R127P 3'	AATTTGCTCCTGAGGGGGGTGTTCCAG	78.4
I276M 5'	TTTATGCTGTGCCAGACAAGATGAATCAAGCAGATTATGCACTGGA	83
I276M 3'	TCCAGTGCATAATCTGCTTGATTCATCTTGTCTGGCACAGCATAAA	83
G346D 5'	TCTTCTGCATCAAGTAAGCTTGACATCACAATGACTGTGGCCCATG	84.8
G346D 3'	CATGGGCCACAGTCATTGTGATGTCAAGCTTACTTGATGCAGAAGA	84.8
M349V 5'	CTTGGCATCACAGTGACTGTGG	67.9
M349V 3'	CCACAGTCACTGTGATGCCAAG	67.9
K528R 5'	GGACACTGCAGAGGGGCTTTCCTCTG	75.9
K528R 3'	CAGAGGAAAGCCCCTCTGCAGTGTCC	75.9
D575N 5'	CAGCAAATCCAACATGGTCCATC	69.3
D575N 3'	GATGGACCATGTTGGATTTGCTG	69.3
R725Q 5'	GCTCAGTCTCAGAGCAAATGCTGCGGAG	77.3
R725Q 3'	CTCCGCAGCATTTGCTCTGAGACTGAGC	77.3
Q730E 5'	CTCAGAGCGAATGCTGCGGAGTGAACTACTACTCCTCGCCTGTGCG	88.1
Q730E 3'	CGCACAGGCGAGGAGTAGTAGTTCACTCCGCAGCATTCGCTCTGAG	88.1

Table 2.13: Site directed mutagenesis PCR components for insertion of mutations to ERAP1 plasmid DNA

PCR components	Final concentration	Volume
10x buffer	1x	5µl
25mM $MgSO_4$	2mM	4µl
dNTPs 2mM each	0.2mM each dNTP	5µl
10µM sense 5' primer	0.3mM	1.5µl
10µM anti-sense 3' primer	0.3mM	1.5µl
0.2-0.5µg template plasmid DNA (pcDNA3 vector)		
KOD hot start DNA polymerase	0.02U/µl	1µl
PCR grade H_2O		
Total		50µl

Table 2.14: Site directed mutagenesis PCR cycling conditions

Step		Cycling conditions
1.	**KOD Hot Start polymerase activation**	95°C for 2 minutes
2.	**Denaturation**	95°C for 20 seconds
3.	**Annealing**	65°C for 10 seconds
4.	**Extension**	70°C for 3 minutes 40 seconds
Repeat steps 2-4		18 PCR cycles

2.5.4 Ethanol precipitation

Following completion of dpn1 digestion, 50µl of distilled H_2O (dH_2O) was added to the tube containing the PCR product to a total of 100µl. Next, 0.1x of 3M sodium acetate and 2.5x of 100% ethanol (-20°C) were added to the tube containing the SDM PCR product in water. The reaction was then incubated at -20°C for 20 minutes, followed by a 20-minute centrifugation step at 13,000rpm. The supernatant was discarded, and DNA pellet was washed in 250µl of 70% ethanol (-20°C) by centrifuging for 5 minutes at 13,000rpm and at 4°C. The supernatant was discarded, and the DNA pellet was air-dried for 10 minutes, followed by resuspension in 10µl dH_2O.

2.5.5 Maxiprep

Those bacterial cultures that contained the plasmids carrying the correct sequences of the distinct ERAP1 allotypes identified from the patient cohort were grown overnight in 100ml LB medium containing ampicillin (100ug/ml) at 37°C at 220rpm shaking. Plasmid DNA was isolated using QIAprep Spin Maxiprep kit (Qiagen) according to manufacturer's instructions, and DNA was eluted in 100µl TE buffer, and the concentration was measured with Nanodrop.

2.6 Cell line culture and maintenance

Human embryonic kidney (HEK) 293T and 293TE1KO ERAP1 knockout cells

Cells were cultured in 10cm culture dish in Dulbecco's modified Eagle's medium (DMEM; Sigma-Aldrich, Darmstadt, Germany) supplemented with 10% foetal calf serum (FCS, GE Healthcare, Chicago, IL), 2mM L-glutamine (Sigma-Aldrich), 1mM sodium pyruvate (Life Technologies-BRL, Rockville, MD), 1mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES buffer, Lonza, Basel, Switzerland) and 2mM each streptomycin and penicillin (Sigma-Aldrich, DMEM complete medium). Both 293T and 293TE1KO cells were cultured in 10cm culture dishes at 37°C and at 5% CO_2.

HeLa, K89, B3Z and BE7A2Z T cell hybridoma

HeLa, K89 and the T cell hybridoma cell lines, B3Z and BE7A2Z, were cultured in Roswell Park Memorial Institute (RPMI) 1640 medium (Lonza) supplemented with 10% FCS (GE Healthcare), 2mM L-glutamine (Sigma-Aldrich), 1mM sodium pyruvate (Life Technologies), 1mM HEPES buffer (Lonza), 2mM streptomycin and penicillin (Sigma-Aldrich) and 50 mM 2-ME (Sigma-Aldrich, RPMI complete medium). HeLa and K89 cells were cultured in 10cm culture dishes at 37°C/5% CO_2, while both T cell hybridomas were cultured in T25 flasks ($2x10^6$ thawed cells) and later transferred to T75 flasks also at 37°C and at 5% CO_2.

2.6.1 Subcloning of LLMGTLGIV(LV9)/HLA-A2 specific BE7A2Z T cell hybridoma

The BE7A2Z T cell hybridoma was previously generated to recognise the HPV-16 $E7_{82-90}$ epitope LLMGTLGIV (LV9) presented by the human MHC I allele HLA-A*0201 [195]. BEZA2Z T cells were cultured in a T25 flask in complete RPMI medium, counted using a hemacytometer and seeded at a concentration of 1 cell/ well in 200µl complete RPMI in a 96-well flat-bottomed culture plate. Cells were observed for growth daily and additional media was replenished when necessary. Approximately 7-11 days after plating, wells that contained BE7A2Z growth arising from a single cell, were harvested and their sensitivity to LV9 peptide was tested. To test the sensitivity of BE7A2Z cells towards LV9-HLA-A*0201 complexes at the cell surface, $1x10^5$ 293TE1KO cells were pulsed with 10µM/well LV9 peptide before the addition of 100µl BE7A2Z cells co-cultured overnight before assessing for activation through T cell activation assay (section 2.8).

2.7 Transfection of human ERAP1 and plasmid minigene constructs

To determine the function of the ERAP1 allotypes and combinations identified through long read sequencing, ERAP1 variants and minigene constructs, either AIVMK-SIINFEHL or ED-LLMGTLGIV, were introduced into 293TE1KO cells using Fugene 6 transfection agent (Roche, Germany). 293TE1KO cells were seeded at a concentration of 1.5×10^5 cells/ml in 2ml DMEM supplemented as in section 2.6 per well in a six well tissue culture plate and incubated overnight at 37°C and at 5% CO_2 to achieve a 50-70% confluency. Fugene 6 transfection reagent (Promega, UK) was used at a previously optimised at 3:1 ratio of Fugene 6:DNA (i.e. 3µl Fugene 6: 1µg total DNA). For transfection, 97µl serum free DMEM was added to 1.5ml Eppendorf tube followed by the addition of 3µl Fugene 6 transfection reagent and incubated for 5minutes at room temperature before the addition of 1µg total plasmid DNA. For all transfections, 0.5µg of ERAP1 or 0.25µg per ERAP1 allotype was used. For assessment of AIVMK-SIINFEHL (X5-SHL8) trimming, 0.25µg H-$2K^b$ and 0.25µg X5-SHL8 plasmid DNA were transfected into 293TE1KO cells alongside the relevant ERAP1. For assessment of ED-LV9 trimming, 0.25µg of ED-LV9 and 0.25µg empty pcDNA3 vector plasmid DNA were transfected into 293TE1KO cells alongside the

relevant ERAP1. The Fugene6/plasmid DNA was incubated for 15 minutes at room temperature. The transfection mix was then added drop-wise to a single well containing 293TE1KO cells and incubated for 24 hours. The transfection reaction was repeated for every ERAP1 allotype combination, the trimming of which was investigated.

2.8 T cell activation assay

Upon completion of the 24-hour incubation, transfected 293TE1KO cells were harvested, counted and centrifuged at 1200 rpm for 5 minutes. The starting cell number was $1x10^5$. Cells were then titrated in 96-well flat-bottomed culture plates. Depending on the peptide being assessed (SHL8 vs LV9) either of the two LacZ inducible T cell hybridomas, B3Z or BE7A2Z, were harvested, counted and co-cultured with transfected 293TE1KO cells at $1x10^5$ cells per well. B3Z T cell hybridoma recognises SIINFEHL bound to H2-Kb and the BE7A2Z T cell hybridoma recognises LLMGTLGIV bound to HLA-A*0201 on the cell surface of 293TE1KO cells. The co-culture was incubated overnight at 37^{o}C and at 5% CO_2. The next day, the plates containing the cells were centrifuged at 1500rpm for 2 minutes. The supernatant was discarded and replaced with 100µl per well of chlorophenolred-β-D-galactopyranoside (CPRG, Roche) solution (45.5 mg CPRG, 625µl Nonidet-P40 (US Biological), 4.5ml of 1M $MgCl_2$ (Sigma) made to 500ml with freshly prepared PBS. The mix was stored at 4^{o}C to be used for the colorimetric assays. Upon recognition of the MHC I:peptide complexes at the surface of the antigen presenting cells, 293TE1KO, T cell activation leads to the transcription of the *lacZ* gene and hence the subsequent production of β-galactosidase. CPRG is a substrate for β-galactosidase and it cleaved by the enzyme which results in a substrate that causes colour change from yellow to varying intensities of red. The colour change corresponds to the number of either HLA-A*0201:LV9 or H-$2K^b$:SHL8 complexes at the surface of 293TE1KO cells and is an indirect measure of ERAP1 trimming activity. Quantification was completed using a Biorad 680 plate reader. Readings were taken every hour for six hours with

absorbance at 595nm and 655nm as a reference. Data analysis was carried out using GraphPad Prism software version 9.1.2.

2.9 Immunoblot

2.9.1 Preparation of cell lysates

To examine endogenous ERAP1 expression, cell lysates were prepared from 293T and HeLa cells, $3x10^6$ cells were harvested and counted using a haemocytometer. Immunoblots investigating the ERAP1 expression in 293TE1KO cells that were transfected with the identified ERAP1 allotype combinations were also carried out. Cells were harvested and counted 24h after the transfection. Cell pellets were washed in 500µl PBS before being transferred to a 1.5ml Eppendorf tube and centrifuged at 2,500rpm for 3 minutes. The PBS supernatant was discarded and the cell pellet was re-suspended in NP40 lysis buffer containing; 1% Nonidet-P40 (NP40, US Biological USA), 150mM NaCl (Sigma), 5mM ethylenediaminetetraacetic acid (EDTA, ThermoFisher) and 20mM Tris-HCl pH7.4 (Sigma) and the addition of protease inhibitors phenylmethylsulfonyl fluoride (5% stock PMSF, Sigma) and iodoacetamide (5% stock IAA, Sigma) to achieve a final concentration of $2x10^7$/ml. Cells were incubated in the lysis buffer for 30 minutes on ice and centrifuged at 4°C at 13,000rpm for 10 minutes to pellet the cell debris. The supernatant was transferred to clean 1.5ml tubes (Eppendorf, Germany) and the cell debris was discarded. Approximately 20µl of cell lysates, corresponding to approximately $0.5x10^6$ cells, were generated for each transfection condition. Aliquots were stored in single use tubes at -20°C until use.

2.9.2 SDS-PAGE gel

A 10% sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) gel was prepared in 1mm gel cassettes (Invitrogen). Firstly, preparation of a 10% resolving gel (Table 2.15) was added to the cassette and topped with water saturated-butanol to create an even surface. Once the resolving

gel had set, the butanol was discarded and 5% stacking gel (Table 2.15) was prepared and added on top of the resolving gel, along with a 10- or 12- well comb. 1.5µl of GeneFlow BLUeye protein ladder (Sigma) and 1.5µl of Supersignal invisible ladder (Thermo Fisher) were added in the first and last well of the gel respectively. 3x Non reducing sample buffer (NRSB) consisting of 50% glycerol, 1M Tris pH 6, 10% SDS and dH_2O was added to each sample by using a ratio of 2:1 of lysate to 3x NRSB before samples were loaded onto the gel. The gel was run in 1x SDS running buffer, first at 100V until samples had run through the stacking gel and then at 250V until samples reached the bottom of the resolving gel. Transfer of protein onto a nitrocellulose membrane Hybond C (GE healthcare) was carried out by stacking two layers of sponges, filter paper (Whatman Cytiva, UK), the resolving gel, a previously wetted nitrocellulose membrane and then two more layers of filter paper and sponges. The components of the transfer cassette were soaked in 1x transfer buffer made by diluting 5x transfer buffer (72.05g glycine, 15.15g Tris base in 1L dH_2O) in dH_2O and ethanol. 1x transfer buffer was also added into the tank along with ice packed around the transfer cell and run at 26V for 75 minutes.

Table 2.15: Components of the 10% resolving gel and 5% stacking gel.

Components	10% resolving gel (ml)	5% stacking gel (ml)
dH_2O	4.0	2.1
30% Acrylamide/Bis (37:5:1, BioRad, UK)	3.3	0.5
1.5M Tris HCl (pH 8.8)	2.5	-
1.0M Tris HCl (pH 6.8)	-	0.38
10% SDS	0.1	0.03
10% APS	0.1	0.03
TEMED (Sigma, UK)	0.004	0.004

2.9.3 Blocking and immunodetection

After protein transfer, the nitrocellulose membrane was blocked overnight in blocking buffer (5% milk (Marvel) in PBS with 0.1% Tween 20 (Sigma)) at 4^{o}C with rocking. After overnight incubation, the membrane was divided horizontally in two; the top membrane for ERAP1 protein expression and the

bottom membrane for expression of the loading control glyceraldehyde 3-phosphate dehydrogenase (GAPDH). The membranes were incubated with the respective primary antibodies (Table 2.16, all diluted in blocking buffer) for 1h at room temperature with gentle rocking. The membranes were washed (3x 10 minute washes) with 10ml wash buffer (PBS with 0.1% Tween 20) before incubation with secondary antibodies for 1 hour at room temperature with rocking. Membranes were washed as before. For detection of ERAP1 and GAPDH from cell lysates from transfected 293TE1KO cells, different antibody concentrations were used (Table 2.16).

Membranes were developed by mixing equal volumes of super signal enhancer and super signal stable peroxide, SuperSignal femto for ERAP1 and SuperSignal pico for GAPDH, (SuperSignal West Femto and SuperSignal West Pico chemiluminescence substrate kits, Thermo Scientific) and incubating on the membrane for 5 minutes at room temperature in the dark. Chemiluminescence was visualised using BioRad FluoroS Multi Imager and the VisionWorks software and ERAP1 levels were compared with GAPDH expression. Quantitative data of ERAP1 and GAPDH protein expression were generated using Image J (National Institute of Health, NIH).

Table 2.16. Antibodies used for immunodetection

The anti-goat IgG-HRP antibody used for the western blots using transfected 293TE1KO cells was provided from Life Technologies company and used at the same concentration due to discontinuation of the previously used antibody.

Antibody	Specificity	Species	Dilution	Provider/Stock concentration
ARTS-1	ERAP1	Goat	1:500	R&D systems 1mg/ml
Anti-goat IgG-HRP	Fc specific	Donkey	1:5,000	R&D systems/ Life Technologies 1mg/ml
GAPDH	GAPDH	Mouse mAb	1:5,000/1:1,000	Abcam 1mg/ml
Anti-mouse IgG-HRP	Fc specific	Goat	1:20,000	Abcam 1mg/ml

2.10 HLA-A*02:01 amplification by PCR

In order to complete the functional assessment of ERAP1 allotypes/combinations towards an HLA-A*02:01 restricted HPV E7 peptide, it was necessary to identify those patients who were HLA-A*02:01 positive. BE7A2Z T cell hybridoma was specifically designed to recognise the LV9 epitope bound to HLA-A*02:01 at the surface of antigen presenting cells, 293T E1KO cells. To identify the HLA-A*02:01 positive patients, amplification of this gene was completed by PCR.

To optimise the PCR protocol used for amplification of the HLA-A2 gene by PCR [196], cDNA prepared using RNA isolated from the 293T cell line that naturally expresses HLA-A*02:01 was used along with the relevant primers presented in the research paper and the KOD Hot Start DNA polymerase kit. Primers were elongated to 22 bases to increase specific binding to ERAP1 (Merck). Amplification of HLA-A2*02:01 by PCR using 293T cDNA was used as a positive control and cDNA prepared from RNA isolated from HeLa cells was used as the negative control as HeLa cells do not express HLA-A*02:01. PCR was carried out with 2µl patient cDNA as template and KOD Hot Start DNA polymerase. The primers used, PCR components and cycling conditions are shown below (Table 2.17, Table 2.18, Table 2.19). PCR conditions were optimised to successfully produce a single, discrete PCR amplified HLA-A*02:01 product of 206bp length, confirmed by running PCR reactions on a 2% agarose electrophoresis gel. PCR reactions were carried out using cDNA from all patients whose ERAP1 allotypes were successfully identified by long read sequencing along with a negative control (HeLa cell line) and a positive control (293T cell line).

Table 2.17: Primer sets used for HLA-A*0201 amplified by PCR

Primer Pair Number	Sequence
5' AP31	GCTCTCACTCCATGAGGTATTT
3' AP2	CCTTCACTTTCCGTGTCTCCCC

Table 2.18: PCR components used for amplification of HLA-A*0201 by PCR

PCR components	Final concentration	Volume

10x buffer	1x	5µl
25mM $MgSO_4$	1.5mM	3µl
dNTPs 2mM each	0.2mM each dNTP	5µl
10µM sense 5' primer	0.3mM	1.5µl
10µM anti-sense 3' primer	0.3mM	1.5µl
Cervical cancer patient cDNA		2µl
KOD hot start DNA polymerase	0.02U/µl	1µl
PCR grade H_2O		32µl
Total		50µl

Table 2.19: PCR cycling conditions

Step		**Cycling conditions**
1.	**KOD Hot Start polymerase activation**	95°C for 2 minutes
2.	**Denaturation**	95°C for 20 seconds
3.	**Annealing**	58.1°C for 10 seconds
4.	**Extension**	70°C for 10 seconds
Repeat steps 2-4		35 PCR cycles

2.11 Flow cytometry

Flow cytometry was used to investigate whether 293TE1KO cells would require transfection with HLA-A*02:01 to increase surface presentation of the HLA-A*0201:LV9 complexes that are recognised by the BE7A2Z T cell hybridoma [195]. First, a 6-well flat-bottom culture plate was used to culture $2x10^5$ 293TE1KO cells per well; untransfected 293TE1KO cells were seeded in the one well, cells transfected with 0.5µg of HLA-A2 plasmid DNA and cells transfected with 1µg of HLA-A2 plasmid DNA in two separate wells, respectively. Another well contained 293TE1KO cells that would be used as the untreated, unstained negative control. After 24 hours, cells from the 6-well plate were harvested and 200µl of the cells harvested from each well were seeded in a 96-well U-bottom plate. Cells were then centrifuged at 1,500 rpm for 2 minutes at 4°C. The supernatant was discarded and the cell pellet was resuspended in 50µl of FACS wash buffer (PBS and 5% FCS) was added to each well followed by 1:100

dilution of BB7.2 antibody-conjugated to green fluorescent protein (GFP, made in house) specific for HLA-A*02:01 per well, except for the well containing the cells that were used as a negative control. The plate were incubated on ice for 30 minutes before addition of 100µl FACS wash buffer and centrifuging at 4^{o}C for 2 minutes at 1,500rpm. The supernatant was discarded and the cell pellet was resuspended in 200µl FACS wash buffer and transferred to three separate FACS tubes for analysis at the BD FACs Canto II (BD Bioscience) machine and using the FlowJo software (TreeStar Inc).

2.12 Statistical analysis

One-way ANOVA with Dunnet's post hoc test was performed for analysis of differences between multiple groups and control (*002-WT; Graphpad prism, Dotmatrics). Chi-square test was used for evaluation of associations between ERAP1 SNPs/allotypes/combinations and either patient or control cohorts. $P<0.01$ and $p<0.05$ indicated statistical significance.

3 Results: part 1

Clinical features of the cervical cancer patient cohort

3.1 Introduction

In this chapter, clinical information about the cervical cancer patient cohort investigated in this study is presented. The aim was to investigate associations between clinical information, such as lymph node metastasis with disease outcome as well as association of recurrence of disease, cervical cancer prognosis with infiltration of CD8+ T cells in the tumour microenvironment (CD8+/TILs) from each of these patients. Given that >95% of cervical cancer cases are caused by HPV and prognostic benefit has been shown to be associated with strong CD8+ T cell tumour infiltration, it is likely that HPV positive tumours are infiltrated by HPV-specific CD8+ TILs [166, 170, 179]. The ERAP1 identity and effect on the trimming function of the enzyme, described in later chapters, may play a role in cervical cancer prognosis through the trimming of HPV-derived epitopes presented to HPV-specific CD8+/TILs that exert anti-tumour immune responses. This was explored in a study by Reeves et al, investigating ERAP1 identity and the effect on trimming an HPV-16 E7-derived epitope (LLMGTLGIV) in a cohort of HPV positive OPSCC patients [121].

3.2 The cervical cancer patient clinical cohort

A total of 103 cervical cancer patient samples (cDNA were provided by researchers at the University of Groningen (Methods and Materials, section 2.1). A double-blind approach for this study was used, i.e clinical information for the patients was made available following the completion of ERAP1 allotyping with long read sequencing to avoid researcher bias. For 96/103 cervical cancer patients, information regarding type of carcinoma, disease stage, distant and lymph node metastasis as well as disease outcome in the form of classifying patients as disease-free or deceased. Recurrence is presented in Table 3.1. No clinical information was available for S39, S57, S105, S112, S114, S117 and S119. cDNA samples from S7, S8, S13, S16, S90, S91, S92, S93, S94, S95, S96, S97, S98, S99 and S100 were not provided by the researchers at the University of Groningen and therefore their clinical

information was not included in this cohort analysis. Consequently, the total number of patients for whom clinical information is presented in this chapter was 96/103 (Table 3.1).

Table 3.1. Clinical information for the cervical cancer patient cohort

SCC= Squamous cell carcinoma, AC= Adenocarcinoma, ASC= Adenosquamous cell carcinoma, LA= locally advanced, LN= Lymph node, cc= cervical carcinoma, O=other cervical carcinoma type

Patient	Type	FIGO stage	Distant metastasis	LN metastasis	Recurrence	Last follow-up
S1	AC	early stage	No	Yes	No	Disease-free
S2	AC	Early stage	No	No	No	Disease-free
S3	O	LA	Yes	No	No	Died due to cc
S4	SCC	LA	Yes	Yes	Yes	Died due to cc
S5	AC	early stage	No	No	No	Disease-free
S6	SCC	early stage	Yes	Yes	Yes	Died due to other causes
S9	SCC	early stage	Yes	No	No	Disease-free
S10	AC	LA	No	Yes	Yes	Died due to cc
S11	SCC	LA	No	No	No	Disease-free
S12	AC	early stage	No	No	No	Disease-free
S14	SCC	LA	No	No	No	Disease-free
S15	O	early stage	No	No	Yes	Died due to cc
S17	SCC	early stage	No	No	No	Disease-free
S18	AC	early stage	No	No	Yes	Died due to cc
S19	SCC	early stage	No	No	No	Disease-free
S20	ASC	early stage	No	No	No	Disease-free
S21	SCC	early stage	No	No	No	Disease-free
S22	AC	early stage	No	No	No	Disease-free
S23	SCC	early stage	No	No	No	Disease-free
S24	AC	early stage	No	No	Yes	Disease-free
S26	O	LA	No	No	No	Disease-free
S27	SCC	LA	Yes	Yes	No	Died due to cc
S28	SCC	early stage	No	No	No	Disease-free
S29	SCC	LA	Yes	Yes	Yes	Died due to cc
S30	SCC	LA	Yes	Yes	No	Died due to cc
S31	AC	early stage	No	No	No	Disease-free
S32	AC	early stage	No	No	No	Disease-free
S33	AC	LA	Yes	Yes	Yes	Died due to cc
S34	SCC	early stage	No	Yes	No	Disease-free
S35	SCC	early stage	No	Yes	No	Disease-free
S36	SCC	LA	No	No	No	Disease-free
S37	SCC	early stage	No	No	No	Disease-free
S38	SCC	LA	No	No	No	Disease-free
S40	AC	early stage	No	Yes	No	Disease-free

S41	SCC	LA	No	Yes	No	Disease-free
S42	SCC	early stage	No	No	No	Disease-free
S43	AC	early stage	Yes	Yes	No	Died due to cc
S44	AC	LA	No	No	No	Disease-free
S45	AC	LA	No	No	No	Disease-free
S46	AC	LA	No	No	Yes	Disease-free
S47	SCC	early stage	No	No	No	Disease-free
S48	AC	early stage	No	No	No	Disease-free
S49	SCC	early stage	No	No	No	Disease-free
S50	SCC	early stage	No	No	No	Disease-free
S51	SCC	early stage	No	No	Yes	Disease-free
S52	SCC	LA	No	No	No	Disease-free
S53	SCC	early stage	No	Yes	No	Disease-free
S54	AC	early stage	No	No	No	Disease-free
S55	SCC	early stage	No	No	No	Disease-free
S56	SCC	early stage	No	No	No	Disease-free
S58	ASC	early stage	No	No	No	Disease-free
S59	SCC	LA	No	No	No	Disease-free
S60	SCC	early stage	No	No	No	Died due to other causes
S61	SCC	LA	No	Yes	Yes	Died due to cc
S62	SCC	early stage	No	Yes	No	Disease-free
S63	SCC	LA	No		No	Died due to cc
S64	AC	early stage	No	No	No	Disease-free
S65	AC	early stage	No	No	No	Disease-free
S66	SCC	early stage	No	Yes	No	Disease-free
S67	SCC	early stage	No	No	No	Disease-free
S68	SCC	early stage	No	No	No	Disease-free
S69	SCC	early stage	No	No	No	Disease-free
S70	O	early stage	No	No	No	Disease-free
S71	SCC	early stage	No	No	No	Disease-free
S72	SCC	LA	No	No	No	Disease-free
S73	SCC	LA	No	Yes	No	Died due to cc
S74	SCC	early stage	No	Yes	No	Disease-free
S75	AC	early stage	No	No	No	Disease-free
S76	SCC	LA	No	Yes	No	Died due to cc
S77	SCC	early stage	No	No	No	Disease-free
S78	AC	LA	No	No	No	Disease-free
S79	SCC	early stage	No	No	No	Disease-free
S80	AC	LA	No	No	No	Died due to cc
S81	AC	early stage	No	Yes	Yes	Died due to cc
S82	AC	LA	No	Yes	No	Disease-free
S83	SCC	LA	Yes	No	No	Died due to cc
S84	SCC	LA	No	No	No	Disease-free
S85	SCC	early stage	No	Yes	No	Disease-free
S86	SCC	LA	No	No	No	Died due to other causes

S87	AC	early stage	No	No	No	Disease-free
S88	AC	early stage	No	No	No	Disease-free
S89	SCC	LA	No	Yes	No	Died due to cc
S101	SCC	LA	No	No	No	Disease-free
S102	SCC	LA	No	No	No	Disease-free
S103	SCC	LA	No	No	No	Disease-free
S104	SCC	LA	No	No	No	Disease-free
S106	SCC	LA	No	Yes	Yes	Died due to cc
S107	SCC	early stage	No	No	No	Disease-free
S108	AC	LA	No	No	No	Disease-free
S109	AC	early stage	No	No	No	Disease-free
S110	SCC	early stage	Yes	Yes	Yes	Died due to cc
S111	SCC	LA	Yes	Yes	No	Died due to cc
S113	AC	early stage	No	No	No	Disease-free
S115	AC	early stage	No	No	No	Disease-free
S116	AC	early stage	No	No	No	Disease-free
S118	SCC	LA	No	No	No	Disease-free

Frequency distributions for type of carcinoma, clinical FIGO stage (International Federation of Gynecology and Obstetrics system for classification of disease stage), presence of either distant or lymph node metastasis (or both), recurrence and disease outcome at the last follow-up are shown in Table 3.2. Results revealed that the majority of patients suffered from squamous cell carcinoma of the cervix (60.4%), which is the most frequent type of cervical carcinoma observed in other studies investigating cervical carcinoma, as well as studies investigating ERAP1 specifically in cervical carcinoma that include cohorts of different ethnic origin (CEU, Utah residents with Northern and Western European ancestry from the CEPH collection Metha et al and Asian Li et al) [181, 182, 197]. The percentage of patients that experienced lymph node and/or distant metastasis is 28.2 and 12.5%, respectively. Lymph node metastasis was significantly associated with survival as 35% of patients that experienced lymph node metastasis were reported as 'died due to complications' associated with their disease or other causes ($p<0.01$; Fisher's exact test, SPSS). Distant metastasis was also associated with death due to cervical carcinoma as only 1/11 patients with distant metastasis were noted as disease-free at the last follow-up ($p<0.01$; Fisher's exact test, SPSS). Recurrence of disease was significantly

associated with death due to cervical carcinoma complications ($p<0.01$; Fisher's exact test, SPSS). In addition, 60.4% patients were diagnosed at an early stage, and only 12% were reported as having died due to complications associated with cervical carcinoma or other causes. An association was observed between diagnosis at an early stage and patient survival at the last follow-up ($p<0.01$; Fisher's exact test, SPSS). The HPV type(s) that 28 patients were infected with are shown in Table 3.3. Infection with HPV-16 was the most prevalent with 16 patients infected with only that HPV type (64.3%), 3 patients infected with HPV-16 + HPV-18 (10.7%), 5 patients infected with HPV-16 and another undisclosed type (17.9%) and 2 patients infected with HPV-16, HPV-18 and another undisclosed type (7.14%).

Table 3.2. Frequency distributions of clinical information in the patient cohort expressed in percentages

SCC= Squamous cell carcinoma, AC= Adenocarcinoma, ASC= Adenosquamous carcinoma, ES= Early Stage, LA= Locally Advanced, LN= Lymph Node, DF= Disease-free, CC= cervical carcinoma, O=other cervical carcinoma type, Individual= This patient's cervical carcinoma was diagnosed at the most advanced disease stage (IVb).

Type of carcinoma (%)		FIGO clinical stage (%)	
SCC	60.4	ES	60.4
AC	33.3	LA	38.5
ASC	2.1	Individual	1
O	4.2		
Distant metastasis (%)		LN metastasis (%)	
No	87.5	No	70.8
Yes	12.5	Yes	28.2
Recurrence (%)		Last follow-up (%)	
No	85.4	DF	75
Yes	14.6	Died (CC)	21.9
		Died (other)	3.1

Table 3.3. HPV types patients have been infected with

Patient	HPV type
S2	HPV-16
S4	HPV-16
S35	HPV-16
S40	HPV-16
S41	HPV-16
S42	HPV-16
S43	HPV-16
S48	HPV-16
S49	HPV-16
S52	HPV-16
S54	HPV-16
S55	HPV-16
S58	HPV-16
S59	HPV-16
S62	HPV-16
S63	HPV-16
S69	HPV-16
S103	HPV-16
S47	HPV-16 and HPV-18
S61	HPV-16 and HPV-18
S70	HPV-16 and HPV-18
S38	HPV-16 and other
S44	HPV-16 and other
S60	HPV-16 and other
S101	HPV-16 and other
S102	HPV-16 and other
S53	HPV-16, HPV-18 and other
S46	HPV16, HPV-18 and other

3.3 CD8+/TIL status of cervical cancer patients

High CD8+/TILs have been associated with better overall clinical outcome in HPV-driven cervical and oropharyngeal squamous cell carcinoma of the head and neck [121, 166, 179, 190, 198]. Since >95% cervical carcinoma is driven by the same mechanism of HPV infection, CD8+/TILs from the cohort in this study were determined and correlated with disease survival. It is possible that HPV E6/E7 derived epitopes are recognised by HPV-specific T cells infiltrating the TME, with levels of infiltration affecting

disease prognosis. ERAP1 identity and trimming function is thought to play a role in processing of HPV-derived epitopes for presentation to HPV-specific CD8+/TILs which exert their effector function contributing to better disease survival. To address this, the number of available CD8+/TILs were classified into three categories, similar to the previous study for HPV positive OPSCC patients [121], to determine if there was any correlation between CD8+/TIL status and ERAP1 function. CD8+/TILs were classified into" low", "moderate" and "high" based on the median that was calculated for the cohort using the available CD8+/TILs per tumour mm^2. From here and after, the categories are referred to as CD8+/TIL^{low}, CD8+/TIL^{mod} and CD8+/TIL^{high}, respectively. The median was calculated to be 465.865 cells per tumour mm^2. The patients whose CD8+/TILs lie within 25% above or below the median were assigned to the "moderate" group, while those patients whose CD8+/TIL numbers were found to be higher or lower than 25% of the median were assigned to the CD8+/TIL^{high} or CD8+/TIL^{low} group, respectively (Table 3.4, Table 3.5). In a study by Lubbers et al, it was shown that after dichotomising CD8+/TILs based on the median, cervical cancer patients with different alleles of the gene STING1 had similar CD8+/TIL infiltration, but it is likely these cells were 'bystanders' and do not actually contribute to anti-tumour immunity [199]. As in the study by Lubbers et al, CD8+/TIL groups were generated based on the median and the findings of that study may apply here; it is possible that CD8+/TILs reported for cervical cancer patients in this study are also 'bystanders', however due to lack of tumour material this could not be investigated further. CD8+/TIL numbers were available for a total of 65/96 patients for whom other clinical data were provided. Analysis revealed that CD8+/TIL^{high} and CD8+/TIL^{low} groups had the same number of patients (n=23), with slightly fewer being in the CD8+/TIL^{mod} category.

Table 3.4. CD8+/TIL status from the cohort of 96 cervical cancer patients

N/A indicate that CD8+/TILs were not available for these patients.

Patient	CD8+/TIL	Patient	CD8+/TIL	Patient	CD8+/TIL	Patient	CD8+/TIL
S1	low	S24	Low	S45	N/A	S65	N/A
S2	mod	S26	N/A	S46	High	S66	High
S3	low	S27	Low	S47	Low	S67	N/A
S4	N/A	S28	High	S48	High	S68	N/A
S5	Mod	S29	High	S49	N/A	S69	N/A
S6	Low	S30	Low	S50	High	S70	Low
S9	Low	S31	N/A	S51	Mod	S71	High
S10	Mod	S32	N/A	S52	Mod	S72	Low
S11	High	S33	N/A	S53	N/A	S73	Mod
S12	Mod	S34	High	S54	Mod	S74	Low
S14	Low	S35	Low	S55	High	S75	Low
S15	Low	S36	N/A	S56	High	S76	N/A
S17	Mod	S37	N/A	S58	High	S77	Mod
S18	Low	S38	N/A	S59	Low	S78	Mod
S19	High	S40	High	S60	High	S79	Mod
S20	N/A	S41	Low	S61	N/A	S80	N/A
S21	N/A	S42	High	S62	Low	S81	Mod
S22	Low	S43	High	S63	N/A	S82	High
S23	Low	S44	N/A	S64	Mod	S83	N/A
Patient	**CD8+/TIL**	**Patient**	**CD8+/TIL**				
S84	N/A	S106	Mod				
S85	Mod	S107	High				
S86	N/A	S108	N/A				
S87	Mod	S109	Mod				
S88	Mod	S110	High				
S89	Low	S111	High				
S101	N/A	S113	High				
S102	High	S115	Low				
S103	N/A	S116	N/A				
S104	N/A	S118	N/A				

Table 3.5. Classifying cervical cancer patients in three categories based on the CD8+/TIL number identified from tumour samples

The number of CD8+ T cells was available for a total of 65/96 patients for whom other clinical information was provided. Every category was assigned approximately a third of the cohort patients, meaning that patients were almost equally divided into the three categories using their CD8+/TILs. No data available for 31 patients. Table generated using SPSS.

	Frequency	Percent	Valid Percent
low	23	24.0	35.4
mod	19	19.8	29.2
high	23	24.0	35.4
Total	65	67.7	100.0
System	31	32.3	
Total	96	100.0	

3.4 Patients in the CD8+/TIL^{high} group survived for longer

As mentioned above, studies have shown that there is a strong link between CD8+/TIL status and overall cervical cancer survival [166, 170, 179]. Kaplan-Meier analysis of overall survival of patients in the three CD8+/TIL groups indicated that patients classified in either the CD8+/TIL^{mod} or CD8+/TIL^{high} category had higher cumulative survival than patients in the CD8+/TIL^{low} category. Censored results represent those patients that died due to complications associated with cervical carcinoma or other causes in the CD8+/TIL^{high} group (high-censored), CD8+/TIL^{low} group (low-censored) and CD8+/TIL^{mod} group (mod-censored). These findings confirm those of previous studies revealing that a higher CD8+/TIL number is associated with better overall prognosis in cervical carcinoma [166, 170, 179] (Figure 3.1).

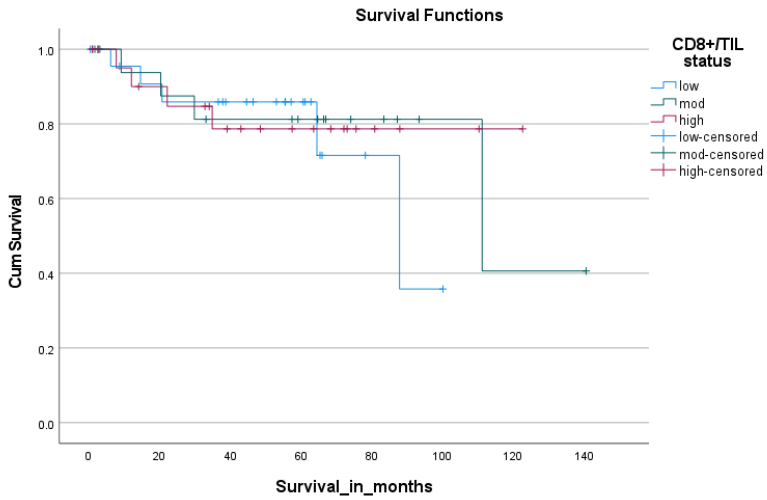


Figure 3.1. Kaplan-Meier graph showing that a higher CD8+/TIL number is associated with better prognosis in the form of survival measured in months

Patients were categorised based on their CD8+/TIL status and survival was determined. The CD8+/TIL^{mod} and CD8+/TIL^{high} group had higher survival than those patients in the CD8+/TIL^{low} group confirming previous study findings [121, 166, 170, 179]. There was a total of 23, 19 and 23 patients in the CD8+/TIL^{high}, CD8+/TIL^{mod} and CD8+/TIL^{low} group, respectively. Notably, there is a difference of 40 months of survival between the longest-surviving patient in the CD8+/TIL^{low} and the CD8+/TIL^{mod} group. Censored refers to the death of the patient due to cervical cancer-related complications or other causes.

The overall survival for patients with and without distant metastasis as well as how survival varies for patients suffering from different types of cervical carcinoma, is shown in Figure 3.2. Patients without distant metastasis survived for longer compared to those with the presence of distant metastasis (average 51 months of survival for patients without distant metastasis vs 26 months of survival for patients with distant metastasis) (Figure 3.2). The exception was patient S9 who survived 52 more months than the average survival for the patients with distant metastasis. Interestingly, that patient was classified in the CD8+/TIL^{low} group (Table 3.4).

The range of survival in months was more variable for squamous cell carcinomas compared to adenocarcinoma patients (Interquartile range 61 months vs 35 months, respectively, Figure 3.2). Two adenocarcinoma patients were shown to be outliers in the boxplot of carcinoma type vs survival. These patients, S82 and S88, were found to have survived for longer compared to other adenocarcinoma patients. This is likely due to the fact that both of these patients had a high number of CD8+/TILs, associated with better survival (Table 3.4). Even though patient S88 was classified in the CD8+/TIL^{mod} group, the CD8+/TIL number was close to the upper boundary and can therefore be assumed that the patient's CD8+/TIL status is in between the two groups (CD8+/TIL$^{mod/high}$, Appendix C). Also, S82 was deemed disease-free at the last follow-up even though there was a previous diagnosis of lymph node metastasis (Table 3.1).

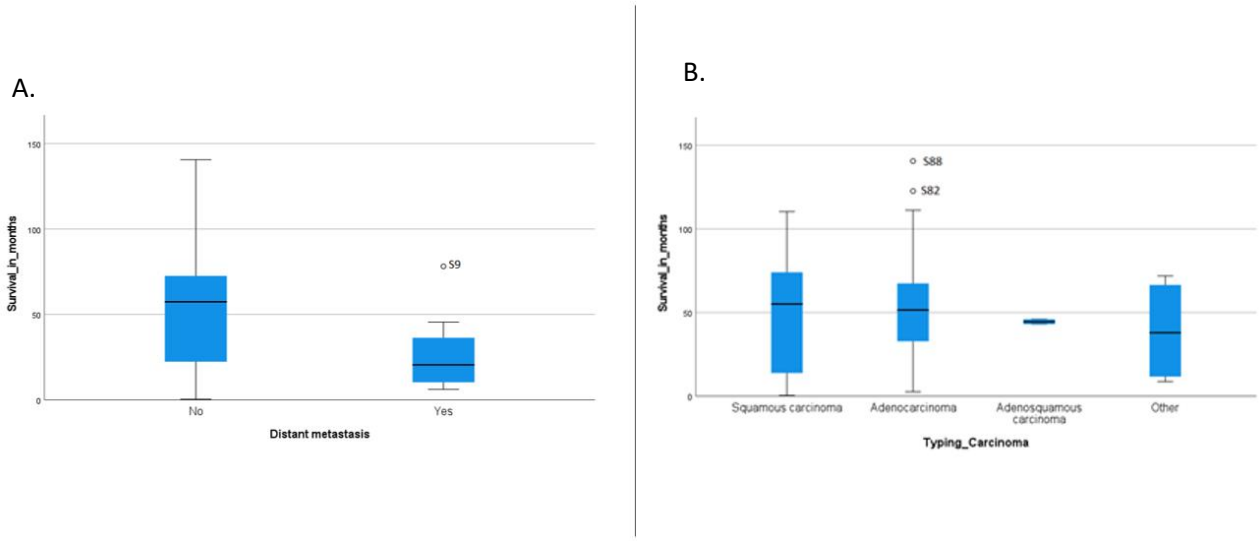


Figure 3.2. Longer survival in patients that did not progress to distant metastasis and two adenocarcinoma patients in the CD8+/TIL^{high} group had longer survival compared to the other adenocarcinoma patients

There was a total of 23, 19 and 23 patients in the CD8+/TIL^{high}, CD8+/TIL^{mod} and CD8+/TIL^{low} group, respectively. Boxplots generated using the SPSS software indicated that patients without distant metastasis survived for a higher number of months compared to those that were diagnosed with distant metastasis (A). When determining the level of survival of the different cc types, increased survival was observed for two patients suffering from adenocarcinoma and had a high CD8+ T cell numbers associated with better prognosis (S82 and S88, B) [166, 170, 179].

Recurrences after treatment of cervical carcinoma are usually located in the true pelvis and other distant recurrences can be observed which highlight the prognostic heterogeneity of cervical cancer [200]. A study by Rapiti et al showed that there was 7.5-fold increase in cervical cancer risk up to 9 years after diagnosis of CIN III and the risk was higher for patients who were treated with cryotherapy and other ablative methods compared to the risk for patients who were treated with excision methods [201]. These studies indicate that new tools are needed to identify those patients at risk of poor disease prognosis, especially by identifying those patients at high risk of disease recurrence.

The timeframe from locally advanced disease to recurrence was shorter in patients that were diagnosed at an advanced disease stage compared to those diagnosed at an early stage (average time

to recurrence 39 vs 23 months, respectively) and confirmed the findings of a previous study showing that for FIGO stages Ib-IIa and IIb-IVa, the relapse rates were 11-22% and 28-64%, respectively (Figure 3.3, A) [202]. The exception was patient S10 who died due to cervical adenocarcinoma after experiencing distant metastasis, even though there was a longer time between recurrence of disease and being classified in the CD8+/TIL^{mod} category (Figure 3.3). Time to recurrence was also investigated for patients who experienced lymph node metastasis, however no correlation was found (Figure 3.3, C). Interestingly, both adenocarcinoma patients S10 and S81 had longer time to recurrence compared to the other patients with lymph node metastasis (81 months and 108 months, respectively, Figure 3.3, C). Clinical data of patient S81 were similar to those of patient S10 and both patients experienced distant metastasis and succumbed to their disease even though they were in the CD8+/TIL^{mod} group (Table 3.4).

The time to recurrence was more variable for adenocarcinoma patients compared to squamous cell carcinoma patients (range of time to recurrence for squamous cell carcinoma 95% confidence interval lower bound=5.97 months and upper bound=36.6 and range of time to recurrence for adenocarcinoma, 95% confidence interval lower bound=2.55 and upper bound=90.1). Patient S51 suffered from squamous cell carcinoma and had a longer time frame to recurrent disease compared to the rest of the patients with the same type of carcinoma (53 months). This patient was alive at the last follow-up and a likely explanation may be that ERAP1 allotype combinations efficiently trimmed HPV-derived epitopes presented to CD8+ T cells which exert anti-tumour immune responses, as other patients in the same group succumbed to their disease. A study by Piersma et al showed that tumour-specific T cells against HPV oncoproteins were detected in 50% of the patients, albeit at reduced levels [166]. In addition, it is possible that innate immune cells, such as CD3-CD57+ NK-like cells and other cells present within the TME contributed to the patient's survival [166, 203]. Treatment might have played a role as well, as it has been shown that the number of HPV-specific T cells was increased after radiotherapy [175]. CD8+/TILs have been associated with better prognosis in cervical carcinoma and

this could explain the recurrent disease survival compared to the rest of the patients in the group [166, 170, 179].

Recurrence survival was compared between patients who experienced recurring disease but at a different status recorded at the last follow-up compared to previous follow-ups (Figure 3.3, D). As expected, the survival range for patients that died due to cervical carcinoma was shorter compared to that for the disease-free patients (95% confidence interval, lower bound=12.5 and upper bound=39.8). Patient S88, found in the CD8+/TIL^{high} group, had the longest survival after recurrent disease and was deemed to be disease-free at the last follow-up. Patient S89 was noted as 'died due to cervical carcinoma complications' and was classified in the CD8+/TIL^{low} group which may be a dominant factor in the negative disease outcome (Figure 3.3, D and Table 3.4).

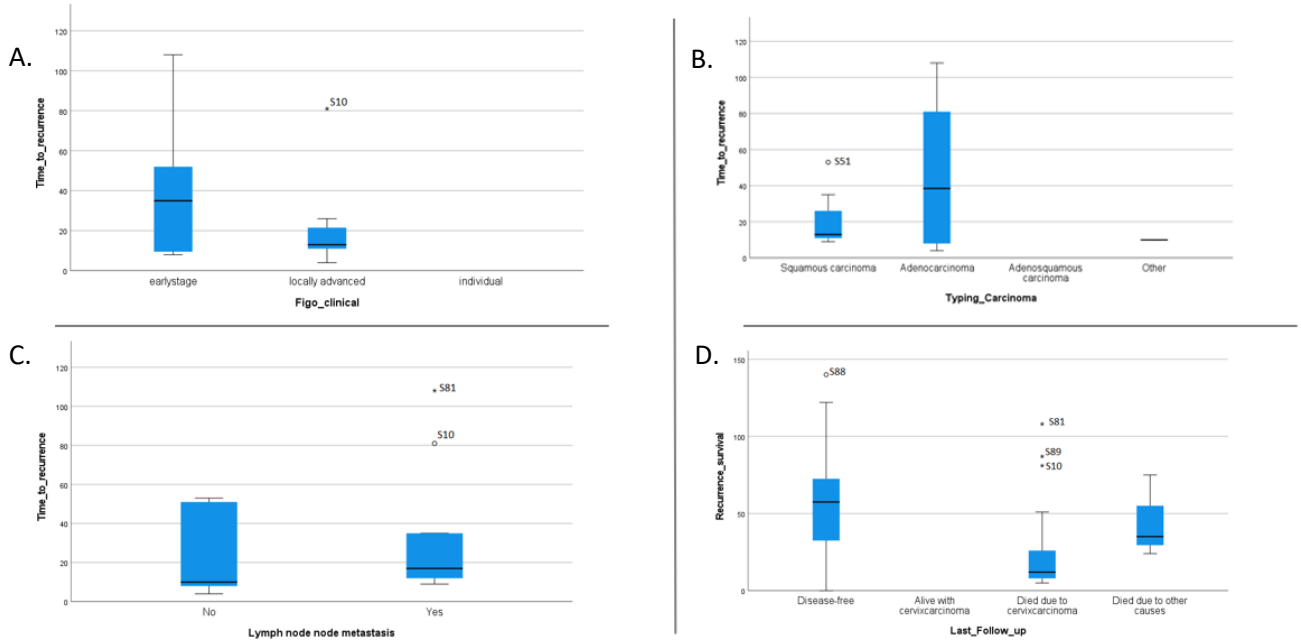


Figure 3.3. Time to recurrence was higher among patients without lymph node metastasis and at an earlier disease stage and recurrence survival was lowest in patients that succumbed to disease

Boxplots generated using SPSS revealed that patients at an early disease stage had longer period of time occurring between remission and recurrence of disease with the exception of patient S10 (A). Recurrence time was variable among adenocarcinoma patients compared to squamous cell carcinoma patients (B). Time to

recurrence was lower in patients with lymph node metastasis with the exception of patient S10 and S81 (C). Survival after recurrence was the lowest in the patients that succumbed to their disease (D).

Another interesting observation was made regarding the clinical stage of the patients in the CD8+/TIL^{high} group. Five patients of this group, three at an early disease stage (S28, S56 and S71) and two at a locally advanced disease stage (S102 and S44), experienced higher CD8+/TILs per tumour mm^2 than the rest of the patients found in either the early or locally advanced stage. All five patients experienced neither lymph node nor distant metastasis, they were classified as disease-free at the last follow-up and showed no signs of recurrence. S44 suffered from adenocarcinoma, while S28, S44, S56, S71 and S102 suffered from squamous cervical cell carcinoma. These data confirm that high CD8+/TIL levels are indeed associated with better cervical carcinoma prognosis, regardless of the clinical stage the patient is at (Figure 3.4).

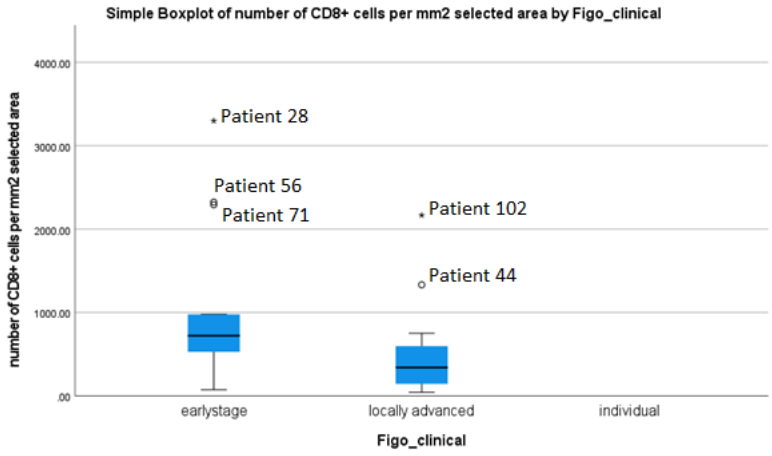


Figure 3.4: High CD8+/TIL numbers are associated with better overall prognosis regardless of the clinical stage they are at.

S28, S56 and S71 were at an early clinical stage and experienced higher CD8+/TIL numbers than the rest of the patients also at an early clinical stage. S102 and S44 were at a locally advanced clinical stage and also had higher CD8+/TIL numbers than the rest of the patients at the same clinical stage. Boxplot, SPSS.

3.5 Identification of HLA-A*0201-positive patient samples

HLA-A*02 is the most prevalent MHC I allele in Northern Europeans and HLA-A2-restrited CTL peptide epitopes derived from the HPV oncoproteins E6 and E7 was identifying those patients who were HLA-A*0201 positive [204, 205]. In the study by Reeves et al, anti-HPV responses in 5 HPV positive OPSCC patient tumours directed to two HLA-A*0201-restricted epitopes were prevalent [121]. Therefore, we sought to identify the HLA-A*0201 positive patients from the cohort and investigate the trimming of an N-terminally extended peptide that was also investigated in the study by Reeves et al to allow for meaningful comparisons of ERAP1 function in the two HPV-driven cancers.

Wang et al developed a PCR-based method for HLA-A*0201 typing to investigate the effects of HLA-A*0201 antigen loss in two melanoma cell lines, SK-MEL-29.1.22 and SK-MEL-29.1.29 [196]. The primers for HLA-A*0201 typing can be found in the Materials and Methods chapter, section 2.10. The reverse primer was elongated by four bases and the forward primer sequence was partly modified to bind within an exon of the HLA-A*0201 allele (NCBI, OM255389.1, 504-707 bases) [196]. 293T cells express endogenous HLA-A*0201 and were used to optimise the protocol for HLA-A*0201 amplification by PCR with amplicons visualised on an agarose gel. Visible bands of a length of 203 base pairs on agarose gels representing HLA-A*0201 DNA enabled determining those patients from the cohort who were HLA-A*0201 positive. HeLa cells do not express HLA-A*0201 and cDNA was used as a negative control. Optimisation of the protocol involved adjusting the amount of $MgSO_4$ to 3µl in a 50µl PCR (final concentration 1.5mM), as well as the annealing temperature based on the primer length (Figure 3.5). Once the optimal PCR conditions were determined, the identification of HLA-A*0201 positive patients from the cohort using the original cDNA samples provided by the University of Groningen was undertaken.

Results revealed that 48% of patients (39/81) were HLA-A*0201 positive (Figure 3.6). For S30, S60, S61 and S65, HLA-A*0201 data were not generated due to the lack of sufficient cDNA material. For the HLA-A*0201 positive S11, S114 and S117, there were no data available on either CD8+/TILs or on

specifics of the disease such as those mentioned above. For eleven patients (S31, S32, S33, S36, S38, S45, S49, S53, S76, S103 and S108), there were data available disease specifics, however no data regarding CD8+/TIL numbers were available and therefore the ERAP1 allotype combinations of these patients were excluded from the functional assays. (Results, Chapter 5). Out of the 39 HLA-A*0201 positive patients, there were data available for both CD8+/TILs and disease specifics for 25 of them. Only those 25 patient samples were tested for presence of HLA-A2 cell surface expression (chapter 4.2).

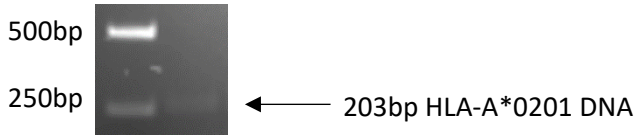


Figure 3.5: HLA-A*0201 amplification from 293T cDNA.

Amplification of HLA-A*0201 from 293T cDNA using HLA-A*0201 specific primers. These PCR products are the result of 35 cycles. All PCR products were run on a 2% agarose electrophoresis gel. Band represents 203bp HLA-A*0201 DNA indicated by an arrow.

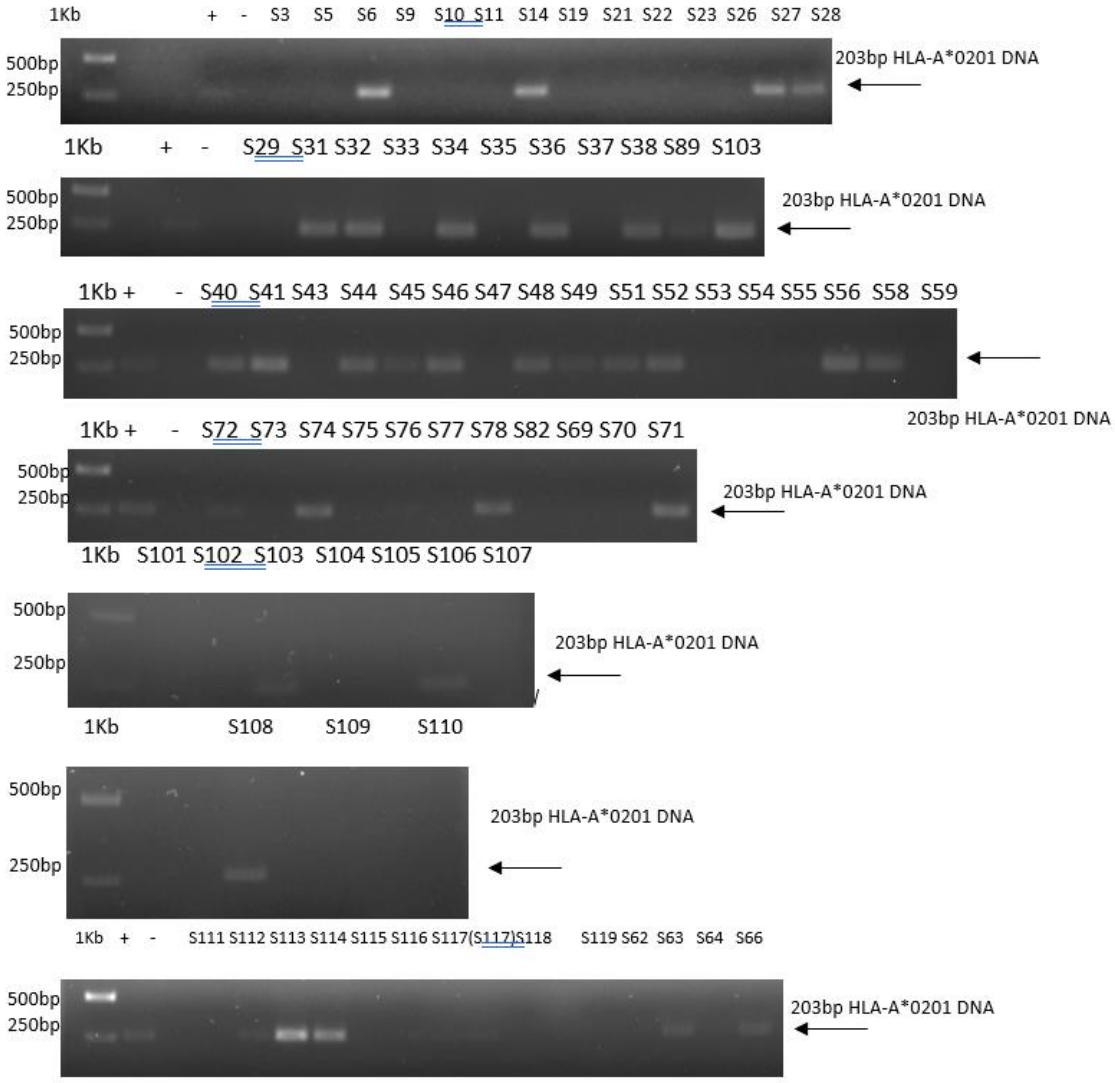


Figure 3.6: HLA-A*0201 amplification from cDNA from the cervical cancer patient cohort.

HLA-A*0201 was amplified by PCR using 35 cycles, 3ul $MgSO_4$ and 58.1°C annealing temperature. PCR products were run on a 2% agarose electrophoresis gel. Band represents 203bp HLA-A*0201 DNA indicated by an arrow.

3.6 Discussion

Clinical information was available for 96/103 cervical cancer patients in this study. Squamous cell carcinoma was the most prevalent type and lymph node and distant metastases as well as recurrence were correlated with disease survival. Cervical cancer patients with a high CD8+/TIL number have

better disease prognosis, therefore patients were classified into three CD8+/TIL groups to investigate whether this was the case here as well. However, as it was already mentioned there were patients in the CD8+/TIL^{low} group that had good disease outcome (alive at the last follow-up) and this may be explained by the presence of other immune cell types in TME. In addition, some patients were infected with more than one HPV type and perhaps treatment has affected immune response; in advanced melanoma patients radiotherapy after immunotherapy increased immune responses leading to prolonged survival [206]. In an HPV-16 positive cervical cancer patient cohort, HPV-related T-cell clones were identified and were shown to selectively increase in numbers following treatment with chemoradiotherapy, however no functional analysis was undertaken and these cells could have exhausted or dysfunctional post-treatment [207]. As expected, it was shown that absence of distant metastasis was associated with longer survival with the exception of S9 who even though had distant metastasis, survived for a longer time compared to the rest. The range of survival in months was higher for SCC patients than AC patients with the exception of the two adenocarcinoma patients S82 and S88 who survived longer likely due to the fact they were in the CD8+/TIL^{high} and CD8+/TIL^{mod} group, respectively. Recurrence time was shorter for patients who were diagnosed at a locally advanced disease stage and these findings are confirmed in literature [208]. Early cervical carcinoma stages have been associated with a higher number of CD103+CD8+T cells (smaller tumours) than late stages that have escaped immune control [170].

4 Results: part 2

Developing the long read pipeline and identification of ERAP1 allotypes

4.1 Long read sequencing for the identification of ERAP1 allotype combinations from a cohort of cervical cancer patients

MinION is the device developed by Oxford Nanopore Technologies that was used for the long read sequencing of the ERAP1 gene from a cohort of cervical cancer patients in this study. It utilises nanopores, proteins specifically designed with a hollow core through which strands of DNA are sequenced bidirectionally as they pass through base by base. The raw signal is converted into nucleotide sequences, known as reads, through a process called basecalling and this is completed with MinKNOW software provided by ONT. Reads can be aligned to a reference sequence using the cloud platform Epi2me.

Third generation long read sequencing was chosen for the identification of ERAP1 from the patient cohort as it allowed sequencing of the entire 2.7Kb coding region of ERAP1 in one continuous sequence, thereby allowing phasing and identification of both chromosomal copies of ERAP1 expressed within each individual within the cohort. This is a considerable advantage for identifying the SNPs since the requirement to stitch together many short reads as in the case of other sequencing methods such as Illumina, could potentially lead to loss of data or errors in sequence. With Sanger sequencing, a sufficient number of plasmids containing the ERAP1 gene need to be isolated from bacterial cells to identify both ERAP1 allotypes from one individual. As this proved to be a considerably costly and time-consuming procedure compared to that proposed for long read sequencing, optimising sequencing using MinION was chosen as the method preferred for sequencing ERAP1 from cervical cancer patients.

4.1.1 Establishing the long read sequencing methodological pipeline

Long read sequencing is a relatively new technology (commercially available in 2014) and although there is evidence on its efficacy, investigation into the specifics of its use for different research fields

has so far been limited [186, 188, 209, 210]. Therefore, before this technology was used for the identification of ERAP1 allotypes from the cervical cancer patient cohort, a methodological pipeline had to be established, including optimising each step of the protocol, identification of the limitations of the technology as well as a robust analysis pipeline. In order to identify ERAP1 allotypes from the output of long read sequencing of ERAP1 amplicons for each patient, analysis of sequencing, including ERAP1 allotyping, was completed using a bioinformatics analysis pipeline employing independently developed software (not provided by ONT) as part of a custom script that was developed with the kind assistance of Dr Jane Gibson (Cancer Sciences, University of Southampton). Data analysis was undertaken using IRIDIS4 (and later IRIDIS5), the University of Southampton High Performance Computing System which enables users to run multiple resource-intensive sequential jobs. The custom script (detailed in Materials and Methods, section 2.3) utilised several software packages; Nanopack, minimap2, samtools and picard, as well as input files containing the ERAP1 reference sequence, SNP sites and the list of previously identified haplotypes. The long read sequencing analysis pipeline for this study is detailed in Methods and Materials.

The first step in the pipeline was extraction of RNA from cell lines, cDNA synthesis and ERAP1 amplification by PCR followed by identification of ERAP1 allotypes through Sanger sequencing [90]. The amplified ERAP1 from the two cell lines, 293T human embryonic kidney cell line and HeLa cervical cancer cell line, was used to identify their ERAP1 allotype combinations using MinION. This enabled the development of a MinION user protocol tailored to the needs of this research, as well as testing of the MinION sequencing accuracy for the project by comparing ERAP1 sequences obtained with MinION with those obtained with traditional Sanger sequencing.

HeLa and HEK293T (293T) cells were the two cell lines used to develop the methodological pipeline for the technology. As HeLa is a cervical carcinoma cell line, it was of interest to compare ERAP1 allotypes in this cell line with those identified from the cervical cancer patient cohort. 293T cell line is an immortalised normal human embryonic kidney cell line used in experiments to assess ERAP1

trimming function [90]. ERAP1 protein expression was also compared between the two aforementioned cell lines.

4.1.2 ERAP1 amplification from cell line cDNA by PCR and allotyping through Sanger sequencing

4.1.2.1 Distinct ERAP1 allotype combinations identified from HeLa and 293T cells

The rationale behind this experiment was to identify the ERAP1 allotype combinations from two cell lines, 293T and HeLa, through Sanger sequencing and later compare the identified allotypes with those obtained with long read sequencing to test the device's suitability for this project. RNA was extracted from HeLa and 293T cells and, following quantification with Qubit, cDNA was synthesised through reverse transcription. ERAP1 specific primers for full length 2.7Kb ERAP1 were used to amplify ERAP1 from HeLa and 293T cDNA before being cloned into the pCR-Blunt II-TOPO vector construct (Figure 4.1). Cloning ERAP1 into the vector was carried out in order to be able to transfect the plasmid into bacterial TOP10 cells. Successful plasmids containing ERAP1 from 293T and HeLa were sent for Sanger sequencing to identify the sequence of the two ERAP1 allotypes.

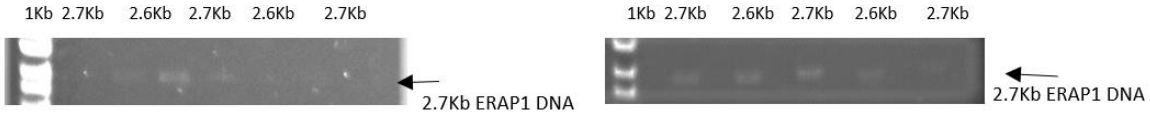


Figure 4.1. Amplification of ERAP1 from HeLa and 293T cell lines

ERAP1 was amplified from 1µl cDNA from either 293T (left) or HeLa (right), by PCR using different sets of ERAP1 primers that have been previously designed within the James lab, either full length ERAP1 (2.7Kb) or to bind within ERAP1 (2.6Kb). The latter set have previously been used to amplify ERAP1 in a second PCR if there was no visible ERAP1 amplification on the agarose gel from the first PCR. ERAP1 DNA (2.7Kb or 2.6Kb) indicated by arrow on 1% agarose gel.

Sequencing results confirmed HeLa expressed two distinct ERAP1 allotypes, the first containing the single SNP R127P (ERAP1 *013) and the second the SNPs R127P/G346D/K528R/Q730E (ERAP1 *020) (Table 4.1). These identified allotypes matched those previously identified from HeLa cells (Dr Emma Reeves, unpublished data). The second ERAP1 allotype, which was later given the nomenclature *020, has not been documented in published studies investigating ERAP1 polymorphism in healthy controls, ankylosing spondylitis or in HPV+ OPSCC patients [121]. Conversely, sequencing of the 293T ERAP1 revealed only a single allotype from the conventional Sanger sequencing containing SNP combinations R127P/K528R/Q730E (ERAP1 *015) which was previously identified by our lab (Dr Emma Reeves, unpublished data) (Table 4.1). While these results suggest the possibility that 293T cells might have been homozygous for ERAP1 *015, there was an insufficient number of plasmids (total of five clones) containing the ERAP1 insert to confirm this hypothesis. This issue confirms the fact that there was a need to switch sequencing of ERAP1 to a less time and effort-consuming method, i.e. long read sequencing with MinION.

Table 4.1: ERAP1 allotype identity in HeLa and 293T cell lines identified from Sanger sequencing.

Bold type indicates variants at the indicated amino acid position.

Cell line	ERAP1 allotype		Amino acid at indicated position								
		12	56	127	276	346	349	528	575	725	730
		T/I	E/K	R/P	I/M	G/D	M/V	K/R	D/N	R/Q	Q/E
HeLa	*013	T	E	P	I	G	M	K	D	R	Q
	*020	T	E	P	I	D	M	R	D	R	E
293T	*015	T	E	P	I	G	M	R	D	R	E

4.1.3 Developing methodology for library preparation and testing the bioinformatics analysis pipeline using ERAP1 amplicons from cell lines

4.1.3.1 ERAP1 allotypes were successfully identified from HeLa and 293T cells in two trial long read sequencing runs

For the first trial long read sequencing run, a single ERAP1 amplicon prepared from 293T cDNA was used to test the suitability of MinION at successfully identifying ERAP1 allotypes through comparison with Sanger sequencing data as well as optimise the part of the protocol involving preparation of the sequencing library (Table 2.2). Following successful amplification of ERAP1 from 293T cells with the blunt primer set that allows amplification of full length 2.7Kb ERAP1, the resulting 293T ERAP1 amplicon (Figure 4.1) was prepared for long read sequencing using a library protocol that included size selection of fragments <3Kb, close to the expected length of the ERAP1 gene. The flow cell check carried out before the sequencing run began indicated a total of 800 active pores available for sequencing, which is also the minimum available pore number recommended by ONT for sequencing (Figure 4.2).

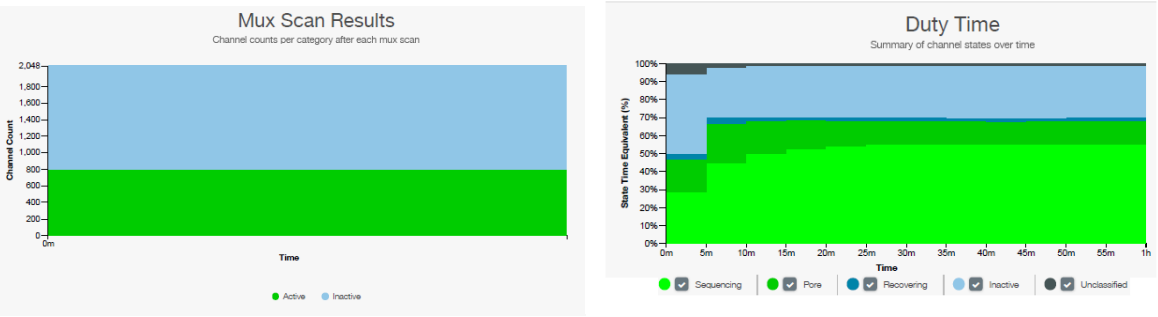


Figure 4.2. Mux scan showing the number of available pores as part of the flow cell check and duty time plot showing percentage of active pores during sequencing.

A flow cell check indicated by the Mux scan (left) was carried out before the sequencing run which showed that the total number of active, available pores for sequencing was 800 (minimum pore number recommended by ONT). The summary of current channels is shown by the duty plot (right) which shows the percentage of pores 1) currently sequencing (sequencing, light green), 2) available but not currently sequencing (pore, dark green), 3) pores that could become later available for sequencing (recovering, dark blue), 4) unavailable pores (inactive, light blue) and 5) unclassified pores (unclassified, black).

A total of 100fmol 293T ERAP1 DNA library was sequenced over one hour on a R9.4.1 flow cell of a MinION device generated a total of 125K reads which were basecalled using MinKNOW and a read length histogram indicates that the majority of reads are close to the expected length of the ERAP1 gene (2.7Kb, Figure 4.3). The first analysis of MinKNOW generated sequencing data is using the Epi2me platform with the reference alignment workflow in the Epi2me desktop agent (v2.2.8). The MinKNOW generated fastq files were aligned to the ERAP1 reference sequence in Epi2me and a total of 96,708 reads aligned to the reference sequence with an average read length of 2,848bp, matching the expected length of ERAP1 and the total yield of sequencing data was 275.5 Mbases with an average quality score of 9.24 (Figure 4.3). Quality scores are calculated as Phred scores and calibration is based on alignment accuracy which also takes into account matches, mismatches, insertions and deletions [211, 212]. The quality score of 9.24 indicates that the read has accuracy of approximately 90.8% (Figure 4.3).

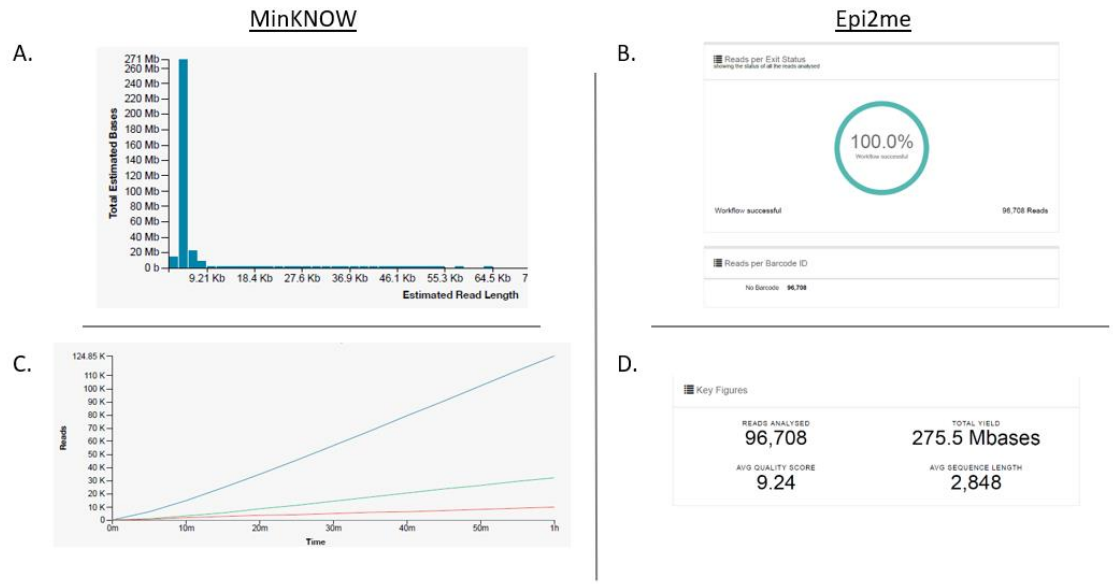
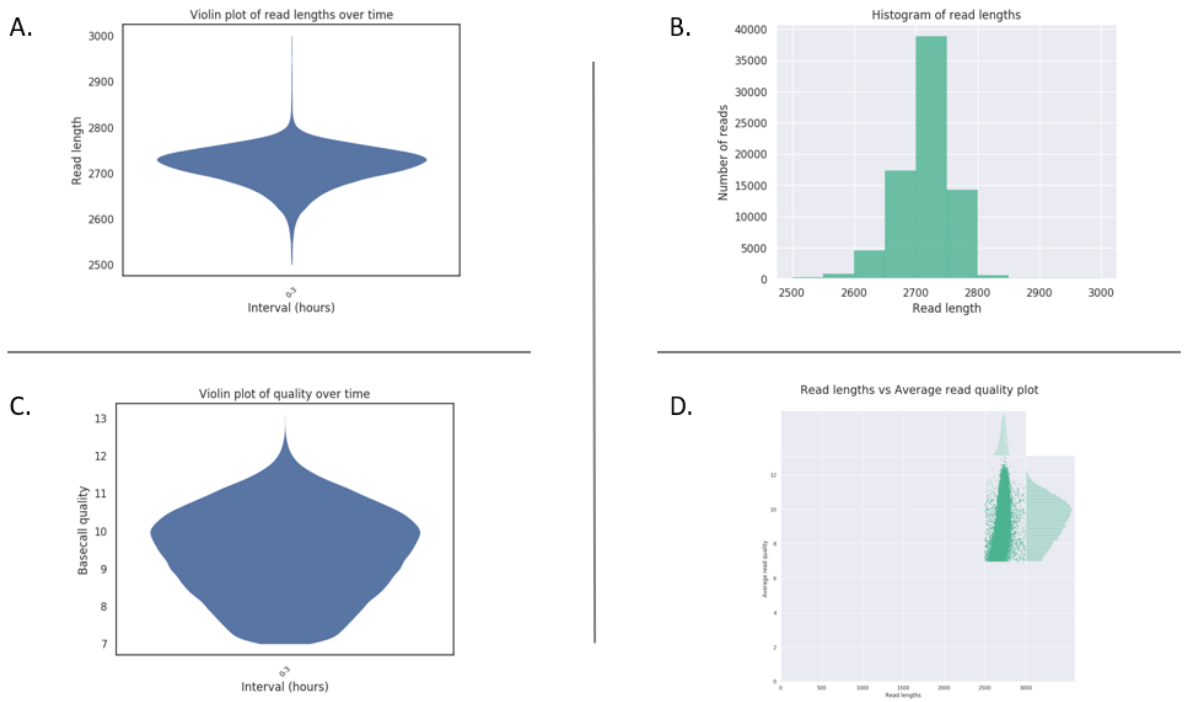


Figure 4.3. Data generated with MinKNOW and Epi2me from the trial sequencing run of ERAP1 from 293T

293T derived ERAP1 amplicon generated from 35 cycles PCR was sequenced using the MinION device and analysed with MinKNOW and Epi2me analysis platform. MinKNOW indicated a total of 125K reads were generated (C, total reads in blue, called passed reads in green and called failed reads in red) with the majority of reads generated from the first trial were close to the expected length of ERAP1 and consistent with the observed amplicon size (2.7Kb, A). The reference alignment workflow in Epi2me was carried out and the report showed that a total of 96,708 reads aligned to the reference ERAP1 sequence (Homo sapiens endoplasmic reticulum aminopeptidase 1 (ERAP1) mRNA, ERAP1-002:01 allele, KM357887.1 of a length of 2823 base pairs (https://www.ncbi.nlm.nih.gov/nuccore/KM357877.1) (B, C). The average sequence length was 2,848 which is close to the expected ERAP1 length, with an average quality score of 9.24 (D).

Data analysis performed using NanoPlot and NanoFilt (downloaded as part of the software NanoPack), indicated that of the total number of reads, 76,860 reads (approximately 79.5% of all reads generated) passed the filtering based on a length range between 2,500 and 3,000 bp and the quality score of 7 (Figure 4.4). Statistical analysis undertaken on the aligned reads indicated a mean read length of 2,715.9bp, a mean read quality of 9.5 and a read length N50 of 2,721 bp, which is also close to the value indicated by both MinKNOW and Epi2me (Figure 4.4). Following variant calling and phasing, two distinct ERAP1 allotypes were identified from the 293T cell line: *015 (Hap6) and *021 (Hap8), the

latter containing all the SNPs encoding the amino acid changes existing in allotype *015 but also an additional SNP that encodes the amino acid change I276M. The nomenclature Hap was first used by Ombrello et al to refer to the ERAP1 gene sequence [120]. The identified allotypes from the 293T cell line match the ERAP1 allotypes identified previously (Dr Emma Reeves, unpublished data).



E.

General summary:	
Active channels:	348.0
Mean read length:	2,715.9
Mean read quality:	9.4
Median read length:	2,721.0
Median read quality:	9.5
Number of reads:	76,860.0
Read length N50:	2,721.0
Total bases:	208,745,015.0

Figure 4.4. Data generated with NanoPlot and NanoFilt for the trial run involving sequencing of ERAP1 from 293T

293T derived ERAP1 amplicon generated from 35 cycles PCR was sequenced using the MinION device and analysed through the bioinformatics pipeline. Nanoplot filtered reads (2500-3000bp) analysis shows the

majority of reads that passed the filtering were approximately 2.7Kb-long, matching the ERAP1 length (B). A violin plot shows the read lengths generated over time (A) and the quality over time (C). A second histogram shows read lengths vs average read quality plot with minimum q score over 7 (D). The average number of active channels along with data on read length and quality scores are shown on (E).

Table 4.2. Identification of ERAP1 allotypes from 293T using long read sequencing in a trial sequencing run

The output generated using the analysis pipeline. *Top,* **Line 1**: all the nucleotide positions at which SNPs encoding known amino acid changes in ERAP1. **Line 2**: wild type haplotype (*002, Hap2): nucleotide sequence of the reference ERAP1 sequence at the nucleotide positions indicated in Line 1. **Line 3**: alternative haplotype: alternative nucleotides that occur at the nucleotide positions indicated in Line 1. **Line 4:** the nucleotides detected at the known nucleotide positions in line 1 for the two ERAP1 allotypes of 293T cells, separated by a forward slash (/), with 0= nucleotide detected to be the same with the reference and 1= nucleotide detected was different to that of the reference. Allotypes include ERAP1 allotype *015 which is the same allotype identified with Sanger sequencing and a second ERAP1 allotype, allotype *021, which contains all the SNPs encoding the amino acid changes in *015, but also the additional SNP encoding the amino acid change I276M. The bottom table shows the amino acid changes at indicated SNP positions.

1	Nucleotide	36	166	380	828	1037	1045	1583	1723	2174	2188		
2	ref	C	G	G	A	G	A	A	G	G	C		
3	alt	**T**	**A**	**C**	**G**	**A**	**G**	**G**	**A**	**A**	**G**		
4	293T	0/0	0/0	1/1	0/1	0/0	0/0	1/1	0/0	0/0	1/1	Hap6/Hap8	*015/*021
1	Amino acid	12	56	127	276	346	349	528	575	725	730		
2	ref	T	E	R	I	G	M	K	D	R	Q		
3	alt	**I**	**K**	**P**	**M**	**D**	**V**	**R**	**N**	**Q**	**E**		
4	293T	0/0	0/0	1/1	0/1	0/0	0/0	1/1	0/0	0/0	1/1	Hap6/Hap8	*015/*021

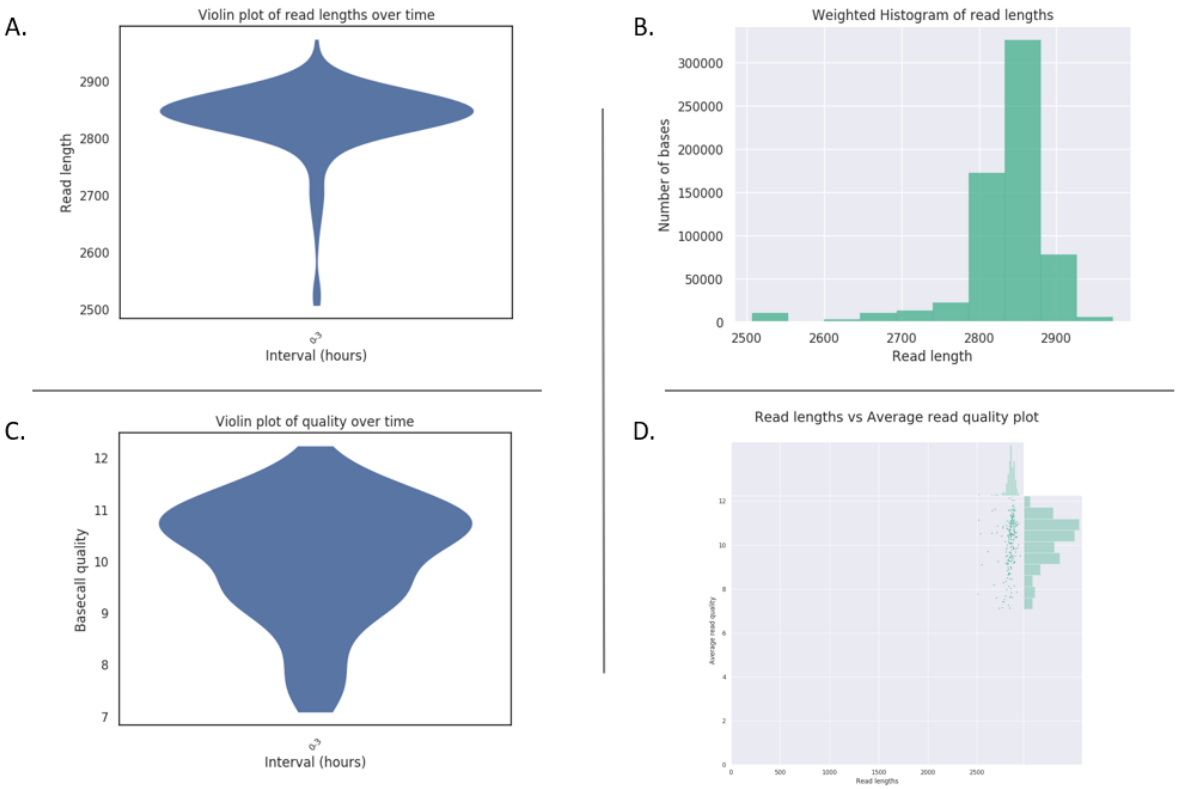
To confirm the validity of the long read sequencing methodology, an ERAP1 amplicon from HeLa cDNA was sequenced using the same methodology and analysis. Due to a technical difficulty, a new flow cell was used for the sequencing of ERAP1 from HeLa cells and the mux indicated a total of 1,200 pores available for sequencing (Figure 4.5). Sequencing was carried out in an hour and the majority of pores was actively sequencing throughout. Alignment of reads to the reference sequence using the relevant

Epi2me workflow as for 293T cells indicated a total of 1,147 reads with the average length of 2,269. Although this length was not close to the expected length of ERAP1 (2.7Kb), filtering completed with NanoPlot and NanoFilt, revealed a total of 226 reads with an average length of 2,833.9 bp that enabled identification of the ERAP1 allotypes from HeLa cells (Figure 4.6). The allotypes identified were *020 and *013, matching those identified previously with Sanger sequencing and those identified by the James lab (Dr Emma Reeves, unpublished data, Table 4.3).

With these two experiments, the first step in establishing a methodological pipeline for long read sequencing with MinION, which included practicing library preparation, investigating the accuracy of the generated data as well as developing an analysis pipeline was accomplished.

Figure 4.5. Data generated with MinKNOW and Epi2me from the trial sequencing run of ERAP1 from HeLa

A flow cell check was carried out before sequencing and the mux scan showed that the total number of active, available pores for sequencing was 1,200. The duty plot shows the status of the pores in the flow cell (see Figure 4.2). The Epi2me report revealed a total of 1,147 reads were generated through the long read sequencing of the ERAP1 amplicon prepared from HeLa cells using 35 PCR cycles that aligned to the reference ERAP1 sequence (Homo sapiens endoplasmic reticulum aminopeptidase 1 (ERAP1) mRNA, ERAP1-002:01 allele, KM357887.1 of a length of 2823 base pairs (https://www.ncbi.nlm.nih.gov/nuccore/KM357877.1).



E.

General summary:

Active channels:	170.0
Mean read length:	2,833.9
Mean read quality:	10.1
Median read length:	2,845.0
Median read quality:	10.4
Number of reads:	226.0
Read length N50:	2,845.0
Total bases:	640,462.0

Figure 4.6: Data generated with NanoPlot and NanoFilt for the trial run involving sequencing of ERAP1 from HeLa

HeLa derived ERAP1 amplicon generated from 35 cycles PCR was sequenced using the MinION device and analysed through the bioinformatics pipeline. Nanoplot filtered reads (2500-3000bp) analysis shows the total number of reads that passed the filtering were approximately 2.7Kb-long, matching the ERAP1 length was 226 (B). A violin plot shows the read lengths generated over time (A) and another one shows the quality over time (C). A second histogram shows read lengths vs average read quality plot with minimum q score over 7 (D). The average number of active channels along with data on read length and quality scores are shown on (E).

Table 4.3. ERAP1 allotypes detected following data analysis from HeLa cells

Allotypes identified from HeLa cells were ERAP1 *013 and ERAP1 *020, matching the allotypes identified using Sanger sequencing. Upper and lower table formatting identical to that displayed in Table 4.2. N/A indicates that haplotype nomenclature was given to that ERAP1 allotype.

1	Nucleotide	36	166	380	828	1037	1045	1583	1723	2174	2188		
2	ref	C	G	G	A	G	A	A	G	G	C		
3	alt	**T**	**A**	**C**	**G**	**A**	**G**	**G**	**A**	**A**	**G**		
4	HeLa	0/0	0/0	1/1	0/0	0/1	0/0	0/1	0/0	0/0	0/1	Hap1/ N/A	*013/*020
1	Amino acid	12	56	127	276	346	349	528	575	725	730		
2	ref	T	E	R	I	G	M	K	D	R	Q		
3	alt	**I**	**K**	**P**	**M**	**D**	**V**	**R**	**N**	**Q**	**E**		
4	HeLa	0/0	0/0	1/1	0/0	0/1	0/0	0/1	0/0	0/0	0/1	Hap1/ N/A	*013/*020

4.1.3.2 Successful ERAP1 allotyping was possible when two barcoded ERAP1 amplicons were sequenced in the same run

The next step in the pipeline involved simultaneous sequencing of the ERAP1 amplicons from the two cell lines mentioned above to investigate whether their distinct ERAP1 allotype combinations could be successfully identified. The aim of this trial sequencing run was to verify that multiple patient samples can be sequenced together in the same run and their respective ERAP1 allotypes can be identified with MinION. ERAP1 was amplified from both HeLa and 293T by PCR using ERAP1-specific tailed primers that contained specific nucleotide sequences developed by ONT for attachment of barcodes to the amplicon. PCR amplicons were confirmed using agarose gel (Figure 4.7). The protocol for simultaneous sequencing of more than one sample, required attachment of barcoding adapters to the tailed primers. The ERAP1 amplicons from HeLa and 293T cell lines received barcode 1 (BRC01) and barcode 2 (BRC02), respectively, before library preparation as before (section 4.1.2.1). The library was sequenced on a single R9.4.1 flow cell on a MinION device over one hour, generating a total of 207,227 reads which were basecalled using MinKNOW and de-multiplexed using the Fastq barcoding

workflow in Epi2me. The Epi2me report indicated that 117,741 reads contained BRC01, and 84,464 contained BRC02 (Figure 4.8). After filtering the reads based on their length ranging between 2,500 and 3,000 bp using NanoPlot and NanoFilt, followed by visualisation of reads in IGV, as well as variant calling and phasing, the ERAP1 allotypes of both 293T and HeLa cells were successfully identified, and they matched those identified in the previous experiments involving single sample sequencing with MinION (Figure 4.9, Figure 4.10, Table 4.2, Table 4.3).

The focus of this study was to investigate the ERAP1 SNPs that have previously been documented in the literature and that result in functional alterations in the ERAP1 allotype; T12I, E56K, R127P, I276M, G346D, M349V, K528R, D575N, R725Q and Q730E.

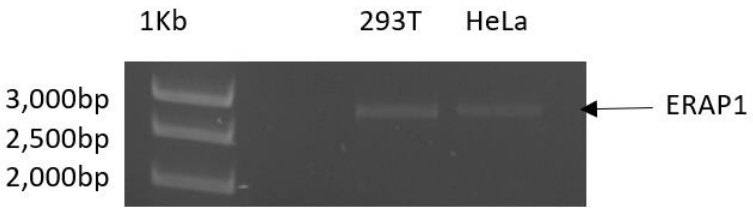


Figure 4.7. Amplification of ERAP1 from 293T and HeLa

Amplification of ERAP1 from 293T and HeLa cDNA using ERAP1 specific tailed primers for 2.7Kb ERAP1. Amplicons were run on a 1% gel and band indicates presence of 2.7Kb ERAP1 DNA.

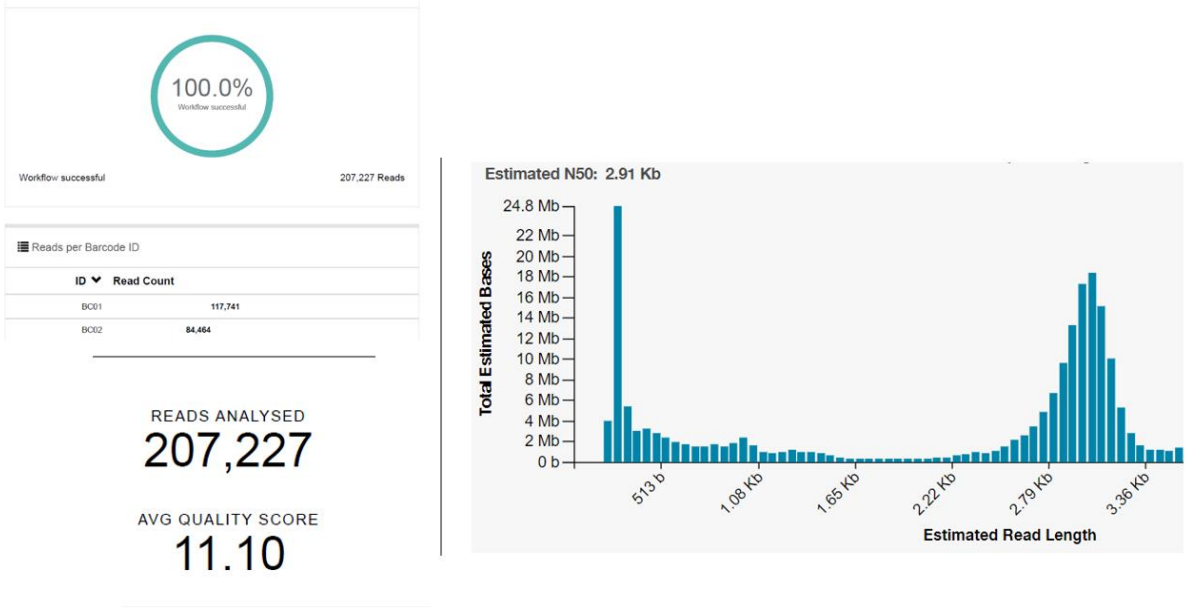


Figure 4.8. Read counts generated for two barcoded ERAP1 amplicons from HeLa (BRC01) and 293T (BRC02) using MinKNOW and demultiplexed using Epi2me

The Epi2me report generated following the sequencing of ERAP1 from HeLa (BRC01) and 293T cells (BRC02) indicated a total of 207,227 reads (117,741 reads for HeLa and 84, 464 reads for 293T). The quality score was 11.10. Reads were sufficient for successful ERAP1 allotyping.

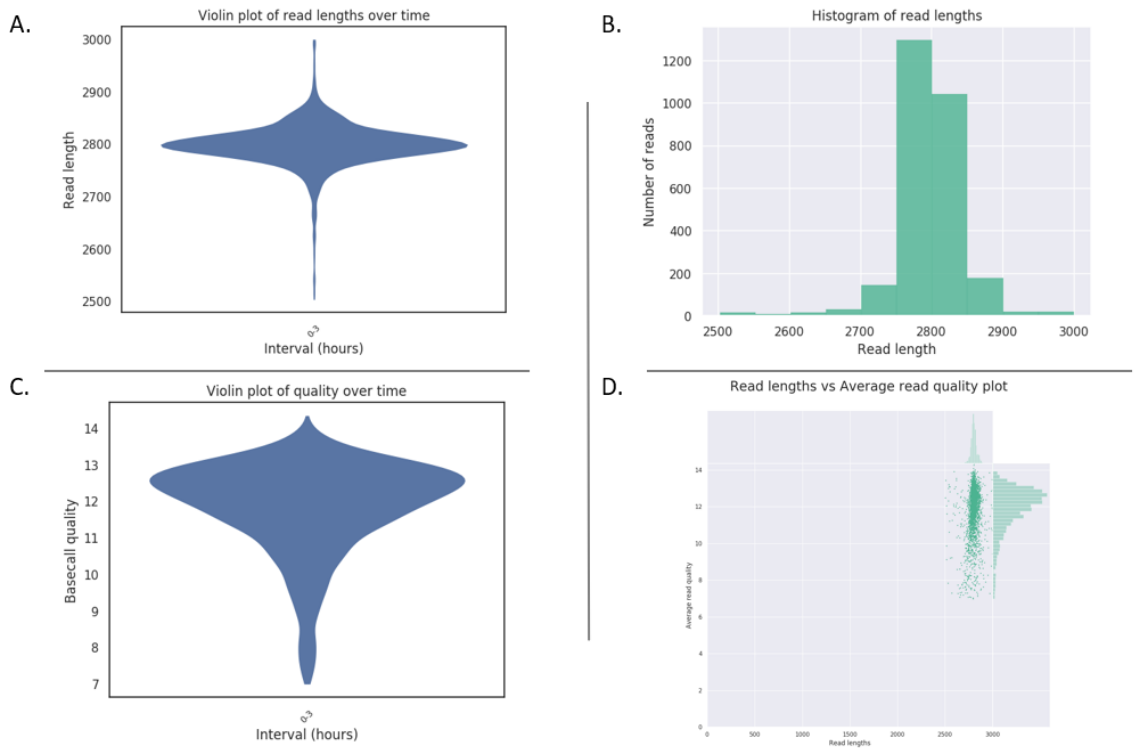


Figure 4.9. Data generated with NanoPlot and NanoFilt for the trial run sequencing of ERAP1 from 293T (BRC01)

293T (BRC01) derived ERAP1 amplicon generated from 35 cycles PCR was sequenced using the MinION device and analysed through the bioinformatics pipeline. Nanoplot filtered reads (2500-3000bp) analysis shows the total number of reads that passed the filtering were approximately 2.7Kb-long, matching the ERAP1 length was 226 (C). A violin plot shows the read lengths generated over time (A) and another one shows the quality over time (B). A second histogram shows read lengths vs average read quality plot with minimum q score over 7 (D).

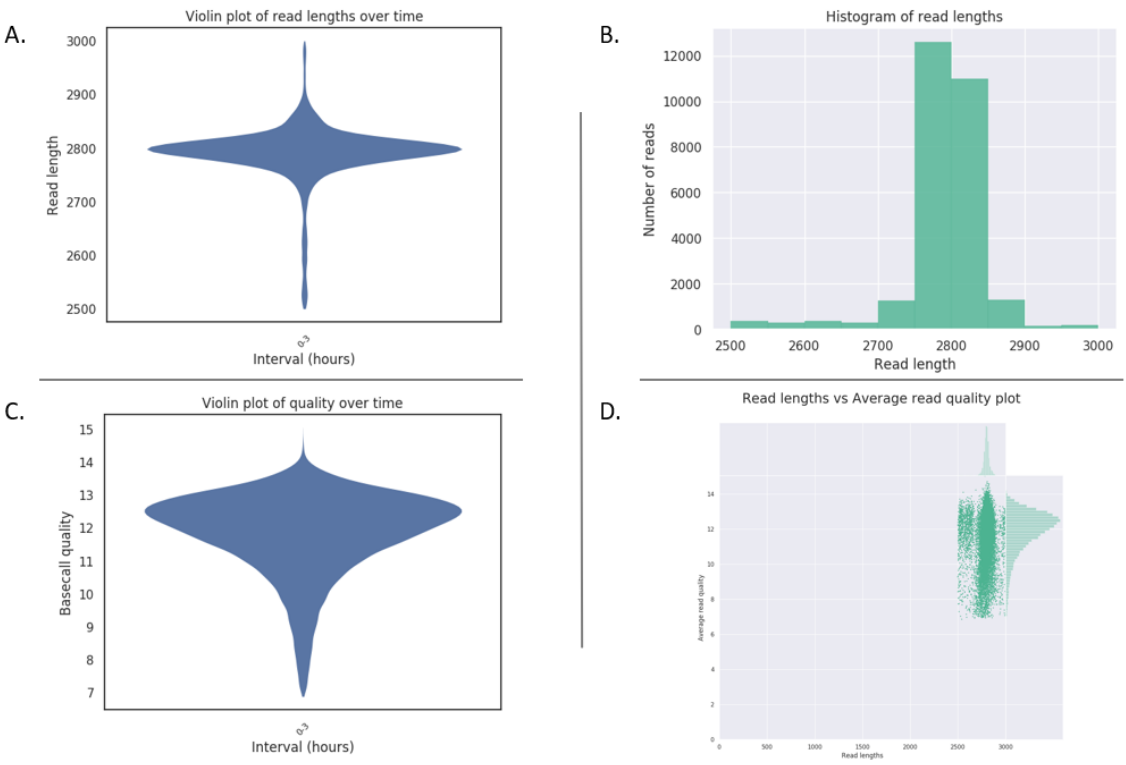


Figure 4.10. Data generated with NanoPlot and NanoFilt for the trial run sequencing of ERAP1 from HeLa (BRC02)

HeLa (BRC02) derived ERAP1 amplicon generated from 35 cycles PCR was sequenced using the MinION device and analysed through the bioinformatics pipeline. Nanoplot filtered reads (2500-3000bp) analysis shows the total number of reads that passed the filtering were approximately 2.7Kb-long, matching the ERAP1 length was 226 (C). A violin plot shows the read lengths generated over time (A) and another one shows the quality over time (B). A second histogram shows read lengths vs average read quality plot with minimum q score over 7 (D).

4.1.3.3 The number of cycles used to amplify ERAP1 by PCR affected accurate DNA quantification and hence ERAP1 allotyping

To investigate whether long read sequencing could accurately identify the ERAP1 allotype combinations from more than two barcoded samples, four ERAP1 amplicons were barcoded and sequenced. During the same experiment, the minimum number of PCR cycles, and hence amplicon concentration, required for successful ERAP1 allotyping was also investigated. Previous experiments on ERAP1 amplification from patient cohorts have generated low ERAP1 yield, therefore since ONT

long read sequencing requires low concentration of DNA in the 5-50fmol range, it was of interest to investigate whether successful ERAP1 sequencing and allotype identification could be successful from low yield PCR. ERAP1 from HeLa and 293T was amplified by PCR using 5, 15, 25 and 35 cycles. All PCR products were run on agarose gel which revealed that ERAP1 DNA was only visible if it had been amplified using 35 PCR cycles (Figure 4.11).

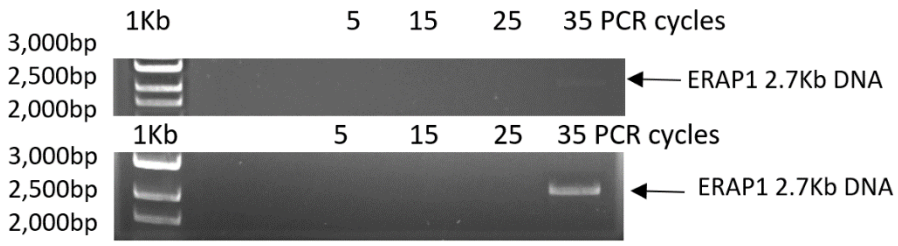


Figure 4.11. ERAP1 amplification by PCR from 293T (top) and HeLa (bottom) cells

Top: ERAP1 amplification from 293T cDNA using ERAP1 specific tailed primers for 2.7Kb ERAP1. *Bottom*: ERAP1 amplification from HeLa cDNA using ERAP1 specific tailed primers for 2.7Kb ERAP1. The amplicons that resulted from 5, 15 and 25 PCR cycles are not visible while arrow indicates presence of ERAP1 2.7Kb DNA for the amplicon that resulted from 35 PCR cycles. Amplicons were run on a 1% agarose gel.

The concentration of the amplicons was measured using Nanodrop, however the concentration measured reflected neither the expectation that ERAP1 concentration increases exponentially with increasing number of amplification cycles, nor the concentrations required to visualise DNA on an agarose gel. In later experiments, concentration was measured with Qubit, as it generated more accurate data on the concentration of the double stranded DNA used in the sequencing experiments. The four amplicons prepared from HeLa cDNA and using 5, 15, 25 and 35 PCR cycles received barcodes BRC03, BRC04, BRC05 and BRC06, respectively. As BRC01 and BRC02 had already been used on the same flow cell in the previous run, we were also able to measure the level of cross-contamination from previous sequencing runs. Libraries were pooled together in equal DNA concentrations. The library was sequenced on the same flow cell as the previous run for an hour generating 217,985 reads

which were basecalled using MinKNOW and de-multiplexed with Epi2me (Figure 4.12). According to the Epi2me report, of the total number of reads, 35 reads contained BRC03 (no reads passed length filtering), 186 contained BRC04 (only 1 read passed filtering), 125 contained BRC05 (95 passed filtering) and 24,733 contained BRC06 (8,127 passed filtering) (Figure 4.12). Interestingly, the majority of the total reads generated contained either BRC01 or BRC02, leading to a staggering 85% cross-contamination between the two experiments. The most likely explanation behind this result could be the potential saturation of pores with DNA from the previous run that were not adequately removed during flow cell wash protocol following completion of the sequencing run. Data analysis showed successful identification of the ERAP1 allotypes from the HeLa amplicons resulting from 25 and 35 cycles of PCR with allotypes matching those identified in the previous run as well as those identified through Sanger sequencing. Interestingly, even though only 95 reads (NanoPlot) were generated that contained BRC05 (amplicon from 25 PCR cycles), they were adequate for successfully identifying ERAP1 allotypes (Figure 4.12).

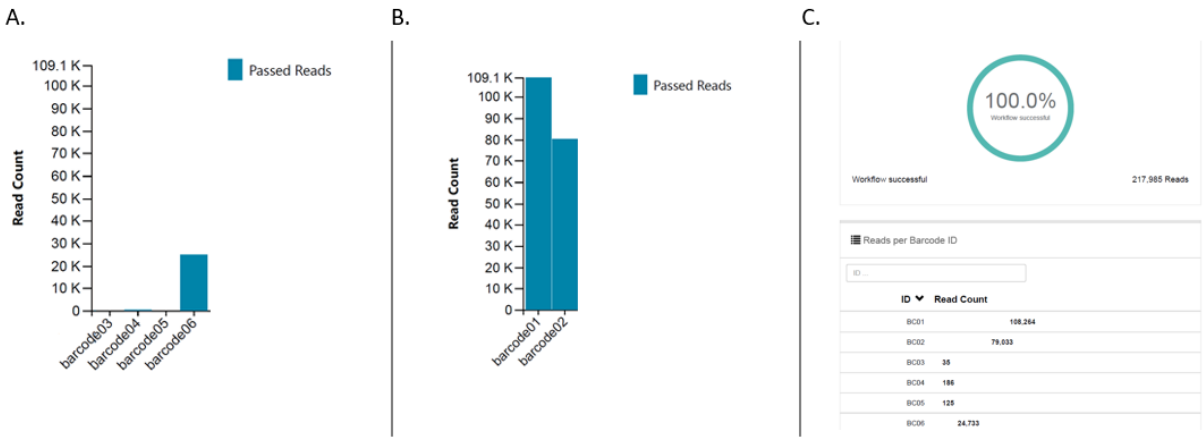


Figure 4.12. Read counts generated with MinKNOW for each barcoded sample in the sequencing run (BRC03-BRC06) and demultiplexed with Epi2me

ERAP1 was amplified from 293T and HeLa cDNA as follows; BRC03= HeLa ERAP1 amplified using 5 PCR cycles, BRC04= HeLa ERAP1 amplified using 15 PCR cycles, BRC05=HeLa ERAP1 amplified using 25 PCR cycles, BRC06=HeLa ERAP1 amplified using 35 PCR cycles (BRC03-BRC06 used in the same sequencing run mentioned in section) (A). There was considerable cross-contamination shown by high numbers of read counts generated

for the barcoded ERAP1 amplicons sequenced together in the previous run, BRC01= HeLa amplicon after 35 PCR cycles and BRC02= 293T amplicon after 35 PCR cycles. Cross-contamination was minimised first with more washes of the flow cell and later with a new version of the wash kit (B). Epi2me was used for demultiplexing the barcoded amplicons (C). Epi2me confirms the high levels of cross-contamination that were seen on the histogram generated with MinKNOW regarding barcode read counts.

The ERAP1 allotypes of the HeLa amplicons resulting from 5 and 15 cycles of PCR amplification were not identified. This could be related to saturation of pores with DNA from the amplicons of 25 and 35 PCR cycles and that is likely because of the inherent accuracy of DNA quantification after amplification by a higher number of PCR cycles compared to a lower number which could explain the acquisition of a greater number of reads for the amplicons from 25 and 35 PCR cycles (Figure 4.13).

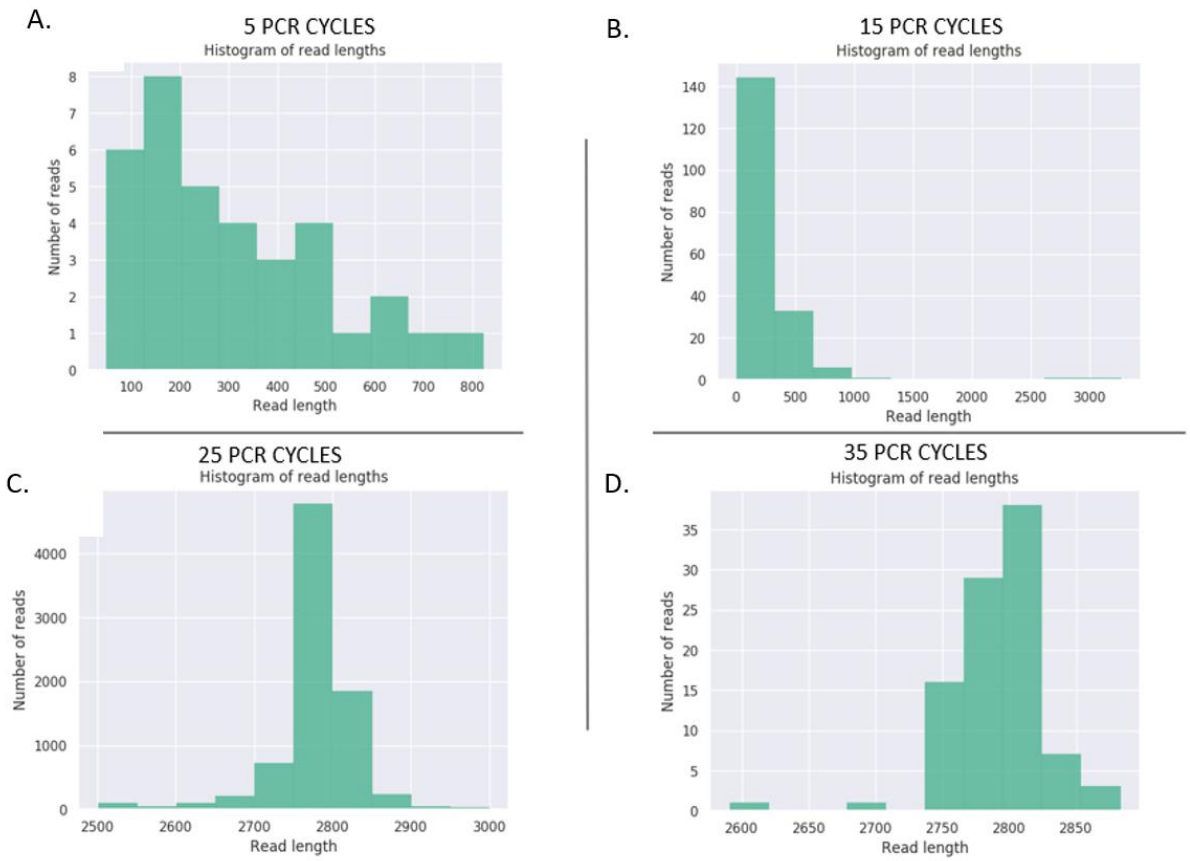


Figure 4.13. NanoPlot histograms of read length vs number of reads for HeLa

HeLa derived ERAP1 was amplified using 5, 15, 25 and 35 PCR cycles sequenced using the MinION device and analysed through the bioinformatics pipeline. NanoPlot filtered reads (2500-3000bp) analysis generated a histogram of read lengths for the HeLa amplicon of 5 (A), 15 (B), 25 (C) and 35 (D) PCR cycles, showing the read

lengths consistent with ERAP1 in (C) and (D), correlating with the successful ERAP1 allotype identification in these samples but not with (A) and (B).

The same experiment was conducted using 293T cells. RNA was isolated from 293T cells, cDNA was synthesised and used for ERAP1 amplification by 5, 15, 25 and 35 PCR cycles using the tailed primers. The library generated 26,504 reads which were basecalled using MinKNOW and de-multiplexed using Epi2me. The majority of these reads were found to contain the barcode that was attached to the 293T amplicon that resulted from 35 PCR cycles, and hence the amplicon with the highest ERAP1 concentration (Supplementary data, Appendix A). The number of reads generated for amplicons resulting from both 5 and 15 PCR cycles were too low for efficient data analysis also pointing towards saturation of pores with amplicons from 25 and 35 PCR cycles, based on frequency of DNA sequence loaded as a possible explanation. The ERAP1 allotypes matched those identified in earlier experiments as well as those identified with Sanger sequencing (section 4.1.2).

4.1.3.4 A dominating effect of ERAP1 amplified by PCR using 35 cycles was observed during long read sequencing

To address the issue of cross-contamination between experiments on the same flow cell, it was of interest to investigate whether a lower amount of DNA could prevent saturation of pores in the flow cell and minimise the percentage of cross-contamination between experiments. Quantification of DNA by Nanodrop seemed to be more accurate when measuring the concentration of ERAP1 amplified using 35 PCR cycles. An interesting observation was that even though all barcoded amplicons were pooled together at equal ratios of DNA concentration determined by Nanodrop before library preparation, a greater number of reads was generated for ERAP1 amplified by PCR using 35 PCR cycles (Figure 4.14, C). For the next two sequencing runs, only ERAP1 amplified by PCR using 15, 25 and 35 PCR cycles from HeLa cells were used. The ERAP1 amplified by PCR using 5 PCR cycles was not added

in this run because the previous run indicated that the number of cycles was too low, and ERAP1 DNA concentration, was too low for successful ERAP1 allotyping. The aim of this experiment was to eliminate the amplicon that had the lowest ERAP1 concentration as this was likely not accurately measured with Nanodrop. In addition, it is likely that the other three samples that were amplified by PCR using a higher number of cycles (15, 25 and 35) could have dominated and been preferably sequenced by MinION over the lowest cycle amplicon. The reason behind likely domination of some samples over others could have lied in ERAP1 DNA quantification issues with Nanodrop and it can be observed also through the read count discrepancy between the different amplicons depending on the number of cycles they had been amplified for (highest read count for amplicon using 25 PCR cycles, Figure 4.14).

In the first run, the three amplicons received BRC07, BRC08 and BRC09 (Figure 4.14). The library was sequenced over one hour on the same flow cell used before and 5,168 reads were generated, basecalled using MinKNOW and de-multiplexed using Epi2me. Despite acquiring a lower number of reads, data analysis resulted in successful identification of the ERAP1 allotype combinations from the ERAP1 products amplified using 25 and 35 PCR cycles. It is likely that the library concentration loaded was too low for generation of a higher number of reads for ERAP1 amplified using 25 and 35 PCR cycles (<5fmol) and this was confirmed by the duty time plot which showed less than 10% of pores sequencing. The percentage of cross-contamination from previous runs was reduced to 8.78% (454 reads out of 5,168 total reads), indicating that the amendment made to the protocol as part of its optimisation for the research purposes before the sequencing run, involving an increase in the number of flow cell washes after every sequencing run, reduced cross-contamination below 10% (Figure 4.14). Interestingly, while also analysing the reads that were found to contain one of the barcodes that were used in previous runs, it was observed that only 18 reads that had passed the filtering based on read length (2,500-3,000bp) originating from the HeLa amplicon that underwent 35 PCR cycles (BRC06, Figure 4.14 B) led to successful identification of ERAP1 allotypes. This experiment revealed the high

sensitivity of MinION at sequencing the ERAP1 gene. It has to be noted though, that 18 reads were not sufficient for identifying the ERAP1 allotypes from amplicons that had undergone less than 35 cycles of amplification, showing that amplification of ERAP1 by 35 PCR cycles results in a DNA concentration that can be measured more accurately and also lead to successful ERAP1 allotyping even with a relatively low number of reads. Allotypes were also successfully identified for ERAP1 amplicon from HeLa cells that underwent amplification using 35 PCR cycles from the first run with only 27 reads passing filtering (section 4.1.3.2, Figure 4.14, B).

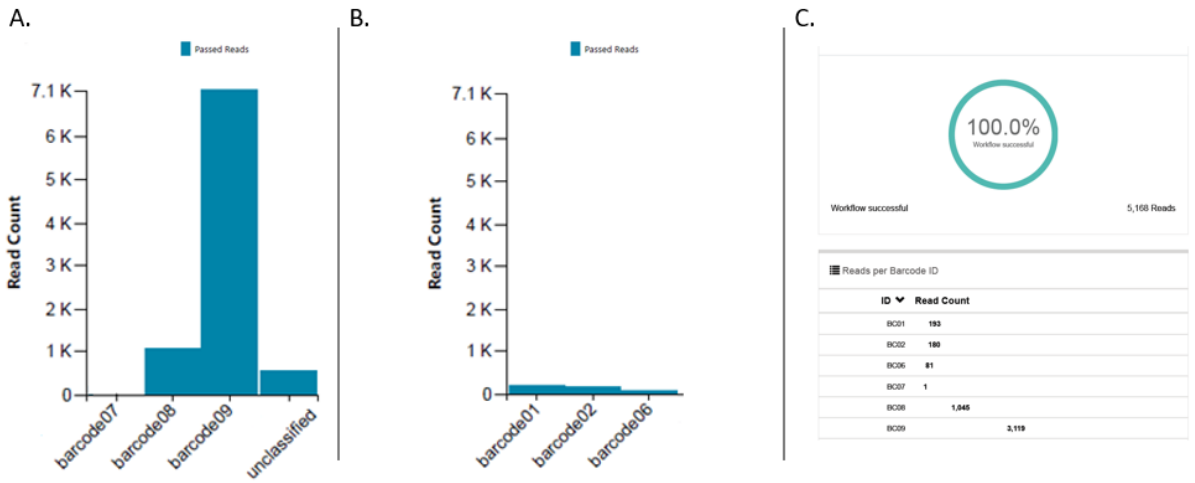


Figure 4.14. Read counts generated with MinKNOW for each barcoded sample prepared from HeLa using 5, 15 and 25 PCR cycles in the sequencing run (BRC07-BRC09) and demultiplexed with Epi2me

ERAP1 was amplified from HeLa cDNA as follows; BRC07= HeLa amplicon after 15 PCR cycles, BRC08=HeLa amplicon after 25 PCR cycles and BRC09= HeLa amplicon after 35 PCR cycles and unfiltered read count histograms were generated using MinKNOW (A). The read counts generated for BRC01=HeLa amplicon after 35 PCR cycles, section 4.1.3.2 , BRC02= 293T amplicon after 35 PCR cycles, section 4.1.3.2 and BRC06= HeLa amplicon after 35 PCR cycles, section 4.1.3.3 are part of cross-contamination reads from previous sequencing runs (B). Amplicons were pooled together at a concentration of 250ng per sample. The amplicon from the lowest number of PCR cycles (5 cycles), was eliminated as the concentration of this was likely not accurately measured with Nanodrop. Reads were demultiplexed using Epi2me and read counts (unfiltered) are shown above (C).

In a second sequencing run, another hypothesis was tested to see whether a higher concentration of ERAP1 product amplified using only 15 PCR cycles (section 4.1.3.3) would result in accurate ERAP1 allotyping. The three amplicons used above received barcodes BRC10, BRC11 and BRC12 in the barcoding PCR and pooled together at a higher concentration of 300ng per barcoded amplicon (50ng increase in the concentration of each amplicon compared to section 4.1.3.3). The library was sequenced on the same flow cell as the previous runs over one hour and generated a total of 17,106 reads were generated. Again, the ERAP1 allotypes from the HeLa sample that was amplified by PCR using 35 PCR cycles were identified but not for the amplicon with the lowest number of PCR cycles. It would have been of interest to have conducted another sequencing run during which only the concentration of the amplicon from the lowest PCR cycles would be increased compared to that of the other two amplicons.

These results, which were consistent across all runs, indicate that the quantification of DNA was biased towards the higher number of PCR cycles. It is likely that the amplicon resulting from 35 PCR cycles was dominant over the other amplicons because its concentration was more accurately measured with Nanodrop compared to the other amplicons and MinION was detecting disturbances in the ionic current caused by the most abundant bases, in this case those belonging to the ERAP1 amplicon with the highest ERAP1 amplification (35 PCR cycles). The percentage of cross-contamination this time was again below 10%, confirming that multiple washes did eliminate most of the old library DNA, however cross-contamination was not completely absent until a new version of a flow cell wash kit became available (Figure 4.15).

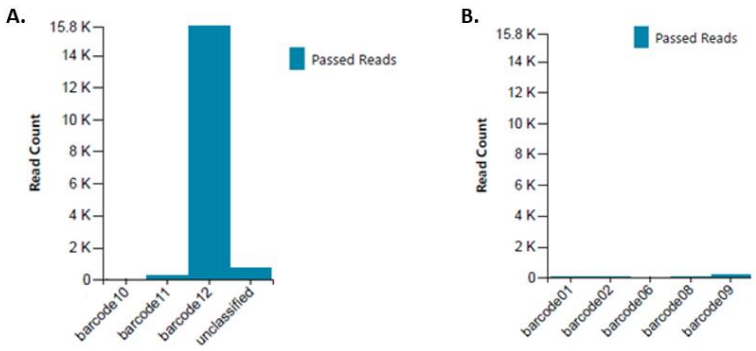


Figure 4.15. Read counts generated with MinKNOW for each barcoded sample prepared from HeLa using 15, 25 and 35 PCR cycles in the sequencing run (BRC07-BRC09)

ERAP1 was amplified from HeLa cDNA using 15, 25 and 35 PCR cycles as follows; BRC010= HeLa amplicon using 15 PCR cycles, BRC11=HeLa amplicon using 25 PCR cycles and BRC12= HeLa amplicon using 35 PCR cycles and unfiltered read count histograms were generated using MinKNOW (A). Barcoded amplicons were pooled together at a concentration of 300ng per sample. The read counts generated for BRC01=HeLa amplicon after 35 PCR cycles, section 4.1.3.2 , BRC02= 293T amplicon after 35 PCR cycles, section 4.1.3.2 and BRC06= HeLa amplicon after 35 PCR cycles, section 4.1.3.3 are part of cross-contamination reads from previous sequencing runs (B). The amplicon from the lowest number of PCR cycles (5 cycles), was eliminated as the concentration of this was likely not accurately measured with Nanodrop.

4.1.3.5 Re-using barcodes on the same flow cell was deemed possible

Once the 12 unique barcodes provided in the ONT PCR expansion kit were used once on the flow cell for sequencing, successful re-use of the same 12 barcodes on the same flow cell with different samples was investigated to identify ERAP1 allotypes. To test this hypothesis, the sequencing run in section 4.1.3.2 was repeated, but this time the barcodes the ERAP1 amplicons from the two cell lines received were swapped around (HeLa amplicon received BRC02 and 293T amplicon received BRC01). The library was sequenced over one hour and thirty minutes with a total of 973 pores available at the start of sequencing following completion of the flow cell check (Figure 4.16). The relevant available pore number following completion of the flow cell check before the flow cell was used for the first sequencing run in section 4.1.3.2 indicated a total of 1,500 available pores (Figure 4.16). A total of

80,185 reads were generated according to Epi2me, with 84% of these containing either BRC01 or BRC02 (Figure 4.17). Following data analysis, it was confirmed that indeed the set of 12 barcodes can be re-used on the same flow cell to accurately identify the ERAP1 allotype combinations of multiple amplicons sequenced in the same run after an effective wash of the flow cell to remove the old library DNA. This can be confirmed by the number of reads generated for samples that received barcodes BRC03-BRC12 as these were used for barcoding samples in previous sequencing runs (110 reads out of 80,185 or 0.13% cross-contamination percentage) (Figure 4.17).

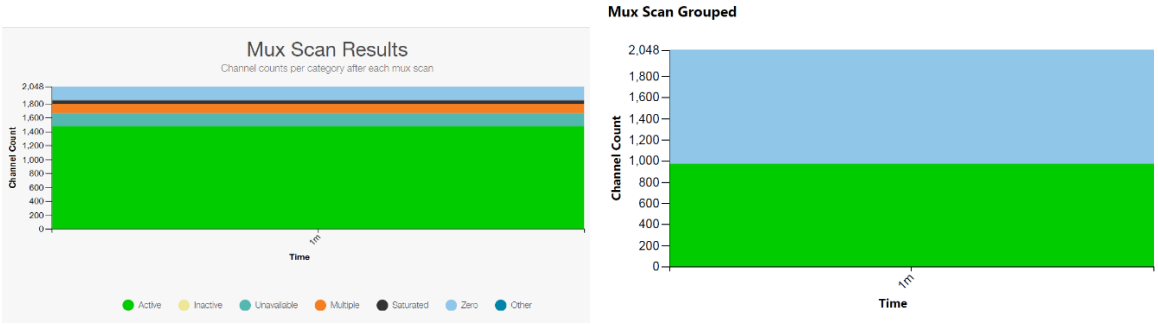


Figure 4.16. Pore availability before the first sequencing run on the flow cell (left) and before the fifth run (right) as reported by MinKNOW

The number of available pores following completion of the flow cell check before the flow cell was used for the first sequencing run was 1,500 (left). The flow cell check carried out before the fifth sequencing run indicated there was a total of 973 pores available for sequencing, hence there is a decrease of 527 pores (or 35.2%). This confirms the ability to use the same flow cell multiple times to sequence ERAP1 consistently and accurately.

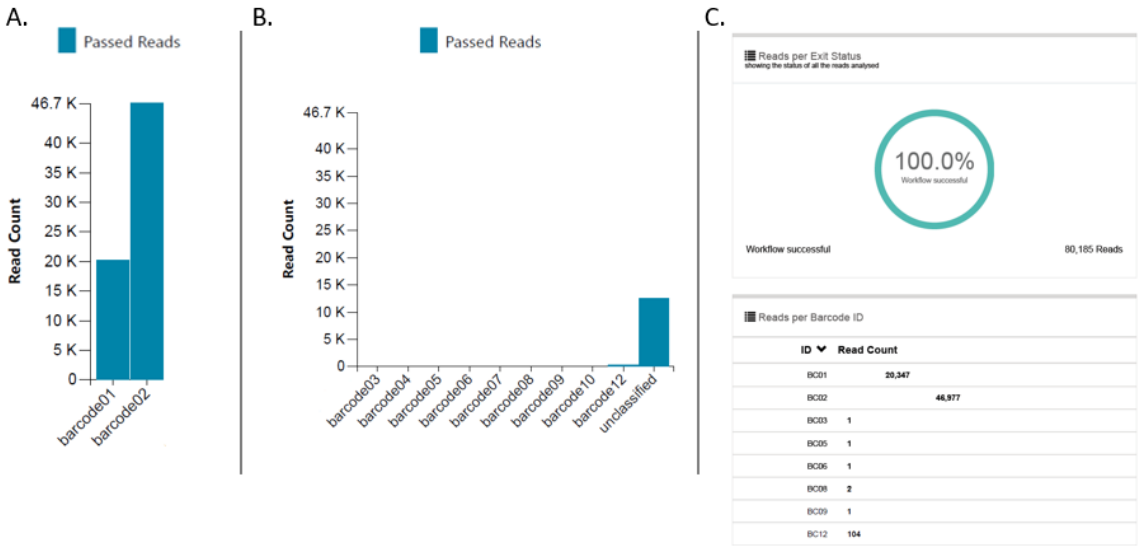


Figure 4.17. Read counts generated with MinKNOW for each barcoded sample prepared from 293T (BRC01) and HeLa (BRC02) using 35 PCR cycles in the sequencing run and demultiplexed with Epi2me

HeLa and 293T ERAP1 amplified using 35 PCR cycles sequenced in the same run and received barcodes as follows; BRC01= 293T, BRC02= HeLa generated with MinKNOW (A). The percentage of cross contamination was minimal as seen by the read counts generated for previously used barcoded amplicons (B). The Epi2me report revealed that a sufficient number of reads (unfiltered) was generated for both ERAP1 amplicons from the cell lines and subsequently ERAP1 allotypes were successfully identified (C). The experiment also confirmed that barcodes can be successfully re-used on the same flow cell as when barcodes were swapped around for the two samples, their respective allotype combinations were identified.

4.1.3.6 Incorporating only ERAP1 amplified by PCR using 5, 15 and 25 PCR cycles in a single sequencing run did not systematically lead to successful identification of ERAP1 allotypes in an already-used flow cell.

At this phase of establishing the methodological pipeline for library preparation, it was investigated whether the sequencing of just the three amplicons that had resulted from 5, 15 and 25 PCR cycles would lead to successful ERAP1 allotype identification. As the ERAP1 amplicon from 35 cycles could have potentially dominated earlier runs and thus have been preferentially sequenced over other amplicons it was eliminated from this run. This experiment was conducted for both HeLa and 293T

cell lines. The 293T library was sequenced on the same flow cell as the previous experiments (Figure 4.16) and the run length was 1 hour and 34 minutes. It is noteworthy, that read acquisition could take longer time given that the lower number of available pores (629 pores, Figure 4.18). According to ONT, the minimum number of pores required at the beginning of every sequencing run should be 800, however sequencing with less than the optimal number of pores was investigated in this experiment as this project focuses on the sequencing of a single gene rather than whole genome sequencing, for which a higher number of pores would be required. The run generated 11,680 reads, the majority of which (over 80%) were shown to correctly contain only the barcodes used during this library preparation (Figure 4.19). Of these reads, only 28 correctly aligned to the ERAP1 reference sequence and only allotype *021 was identified (Figure 4.19).

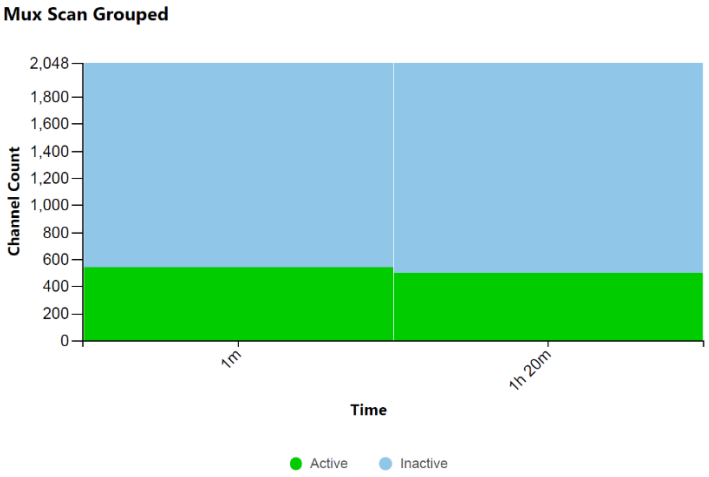


Figure 4.18. Mux scan showing the number of active, available pores for sequencing before the sequencing run begins.

This mux scan reveals a total of 629 pores available for sequencing (green) and the inactive pores that are no longer available for sequencing (blue). Even though ONT recommends a minimum of 800 pores available before sequencing begins, data revealed that sequencing of a single gene, in this case ERAP1, is possible with a lower number of available pores. The Mux scan showing the number of available pores for sequencing is repeated every one hour and a half. Data generated using MinKNOW.

A. B. C.

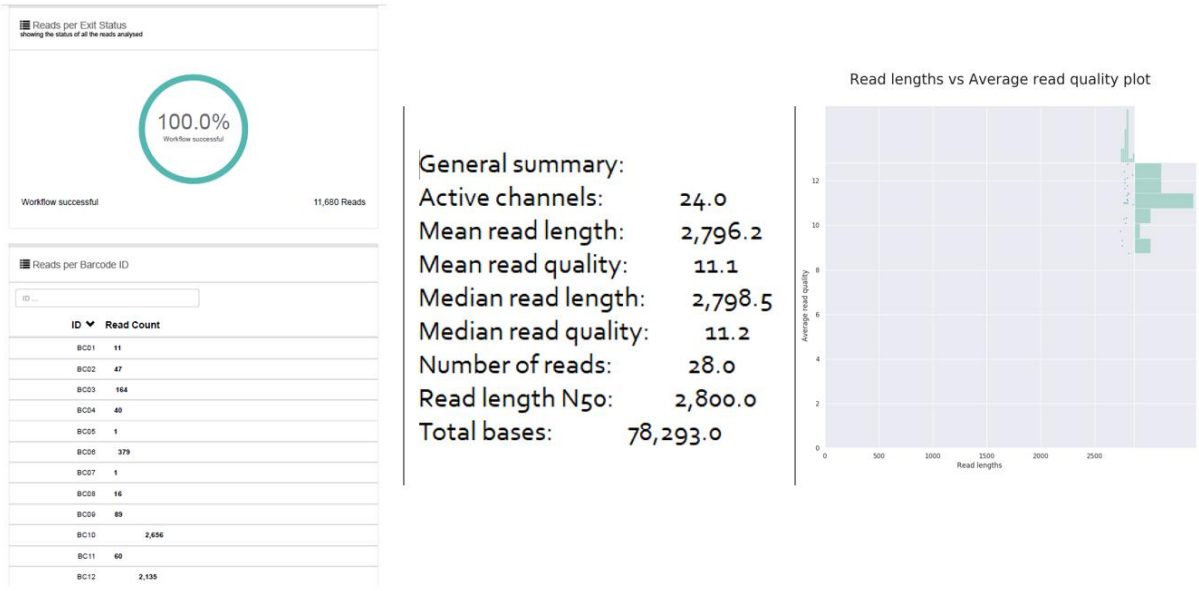


Figure 4.19. Read counts per barcoded amplicons as these were reported with Epi2me and read filtering with Nanoplot

ERAP1 amplified using 5, 15 and 25 PCR cycles from 293T cDNA. Amplicons were sequenced in the same run and Epi2me reported a total of 11,680 demultiplexed reads for the three amplicons (A). Filtering with Nanoplot revealed that only 28 reads of mean read length 2,796.2bp aligned to the ERAP1 reference sequence (B). A histogram of read length vs average quality was also generated following read filtering (2,500-3,000bp) (C).

Even though it was possible to identify both ERAP1 allotypes from samples amplified using just 25 PCR cycles in earlier experiments and with as few as 18 reads, this was not the case here (Figure 4.20). Results point towards two possible explanations; either that the duration of sequencing was too short to generate an adequate number of reads that would enable ERAP1 allotyping, especially given the decrease of pore number with increasing sequencing runs on the same flow cell (Figure 4.18). Or, it is possible that the concentration of ERAP1 used for library preparation was lower than the concentration of ERAP1 used in previous experiments as it is likely that the quantification of the DNA concentration of amplicons from lower than 35 PCR cycles was not accurately carried out before. This could be attributed to the method used for measuring concentration (Nanodrop, Methods and

Materials). When DNA concentration was measured with Qubit, it was shown that it was more sensitive at measuring double stranded DNA concentrations as low as 10pg/µl. The effectiveness of Qubit at measuring concentration using the high sensitivity kit was confirmed by measuring the concentrations of empty vectors, including pCR4Blunt-TOPO vector (Thermo Fisher Scientific Inc.) with the concentration reading matching the one indicated by the manufacturer. Also, the number of available pores was sufficient for ERAP1 sequencing even though it was below the recommended 800.

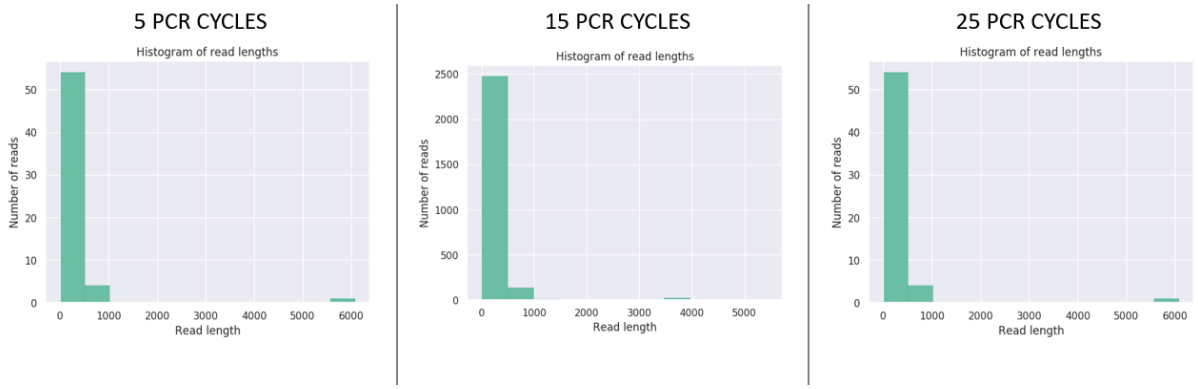


Figure 4.20. NanoPlot reads generated from the sequencing of ERAP1 from 293T cells amplified using 5, 15 and 25 PCR cycles

Histograms generated with NanoPlot using the reads generated for the 293T samples amplified using 5, 15 and 25 PCR cycles. All plots show an insufficient read count for successful ERAP1 allotyping as most reads for these amplicons were below the expected length of ERAP1 (2.7Kb).

The above experiment was repeated for HeLa cells and results matched those observed for 293T cells above. The length of the sequencing run was increased to 2 hours and 48 minutes and the number of available pores after the flow cell check was 320, which, even though it was well below the number of optimal pores recommended by ONT, it still resulted in successful ERAP1 allotype identification for the amplicon of 25 PCR cycles with as low as 279 reads (passed filtering 2,500-3,000bp) (Figure 4.21,

Figure 4.22). Therefore, it was assumed that the life of the flow cell can be extended by using it with <800 available pores for ERAP1 sequencing.

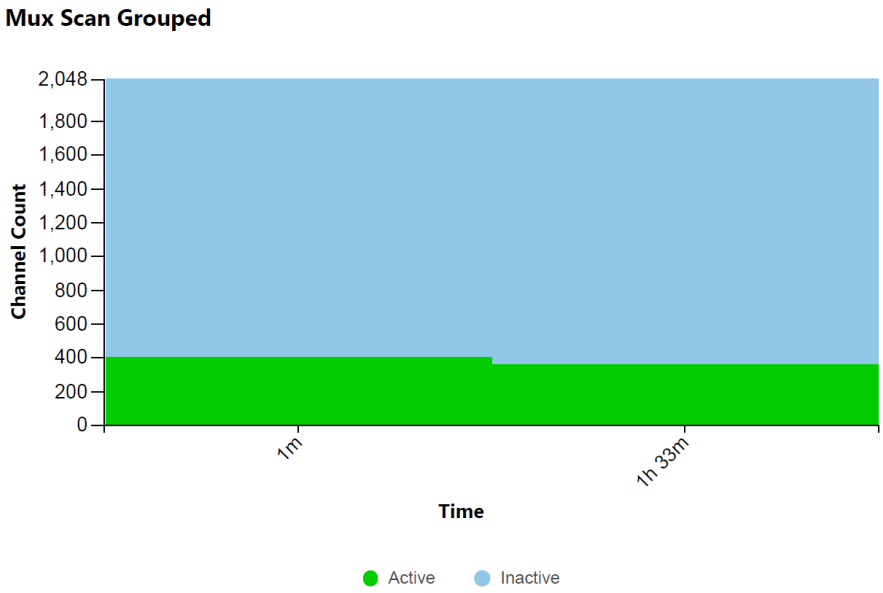


Figure 4.21. Mux scan showing the number of available pores before the sequencing run

Mux scan reports the number of active and available pores for sequencing (green) as well as the pores that are no longer available (blue). This experiment confirmed that long read sequencing of ERAP1 with MinION is possible even with less than 800 pores before the sequencing run began (minimum pore recommendation by ONT).

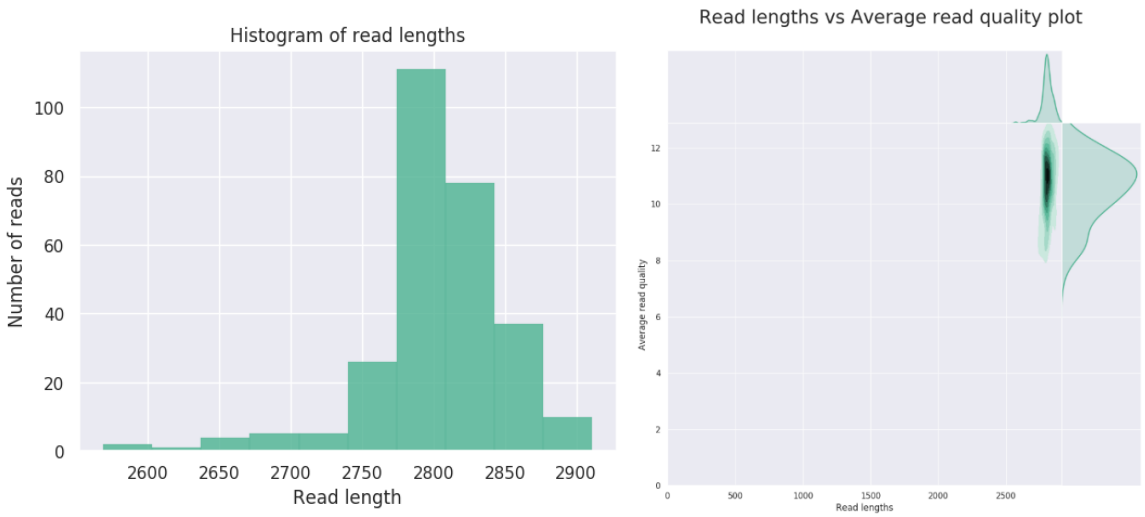


Figure 4.22. Histogram of read lengths and quality plot for ERAP1 amplicon from HeLa cells

NanoPlot analysis of sequencing data generated by MinKNOW and Epi2me. Graphs indicate read lengths vs average read quality for the reads generated for the HeLa amplicon of 25 PCR cycles that have passed the filtering based on length ranging from 2,500 to 3,000bp. The majority of filtered reads were observed to be 2800bp (left panel), with all reads having an average read quality score >8 (right panel), with the successful identification of ERAP1 allotypes from this sample.

One of the possible explanations behind the non-systematic identification of ERAP1 allotypes from amplicons that had been amplified using less than 25 PCR cycles lies in the quantification of the amplicons for library preparation. The measurements for the amplicons from 5 and 15 PCR cycles were very similar in every experiment they were measured, either from 293T or HeLa cells; however a 10-cycle difference should have resulted in an exponential increase in ERAP1 concentration. This observation indicated that quantifications of concentration with both Nanodrop and Qubit are biased towards 35 PCR cycles; both devices are able to measure more accurately the concentration of the more amplified amount of DNA. Consequently, neither of these instruments is able to accurately measure the concentration of low-cycle amplicons which could explain the differences in the number of reads acquired for the amplicons using various PCR cycles, and the inability to generate an adequate number of reads for ERAP1 allotyping for the amplicons of the lower PCR cycles. Nonetheless, the measurements taken with Qubit for the amplicons from the 35 PCR cycles more accurate following

verification on an agarose gel as it was the ERAP1 amplified for 35 PCR cycles that was visible on the gel compared to ERAP1 amplified for 5, 10 and 15 cycles.

To eliminate the factor of the extensive use of the flow cell that resulted in a reduced number of pores and the inability to identify allotypes from the amplicons of the lowest degree of ERAP1 amplification, the previous experiment was repeated on a new flow cell and sample number was also increased from four to six; three amplicons from HeLa and three amplicons from 293T cells corresponding to 5, 15 and 25 PCR cycles were used for library preparation. Interestingly, previous experiments had indicated at least a 10-fold loss of DNA (compared to initial concentration measurements) after AMPure bead clean-up of DNA. Following contact with the bead manufacturer, it was recommended to alter the ratio to a 0.5 sample/bead, as opposed to a 1:1 ratio recommended in the original protocol provided by ONT. This would ensure that the 2.7Kb-long ERAP1 reads would bind to the beads, while lower length fragments are eluted.

The number of available pores at the beginning of the sequencing run were 1699 with sequencing starting with 506 pores, the highest quality pores chosen by MinION which aims to maximise output in the beginning of sequencing. A total of 100K reads were generated in ten minutes. Basecalling was completed with MinKNOW and de-multiplexing with Epi2me, which indicated that 94% of reads contained the barcodes attached to the HeLa and 293T amplicons resulting from 25 PCR cycles and ERAP1 allotypes were successfully identified (Figure 4.23). The reads that were acquired for this amplicon were the highest compared to all other sequencing runs during which this amplicon was used for library preparation. It is possible that, that specific amplicon (25 PCR cycles) was now dominating the run and consequently, it was sequenced at higher levels than the rest of the amplicons by MinION.

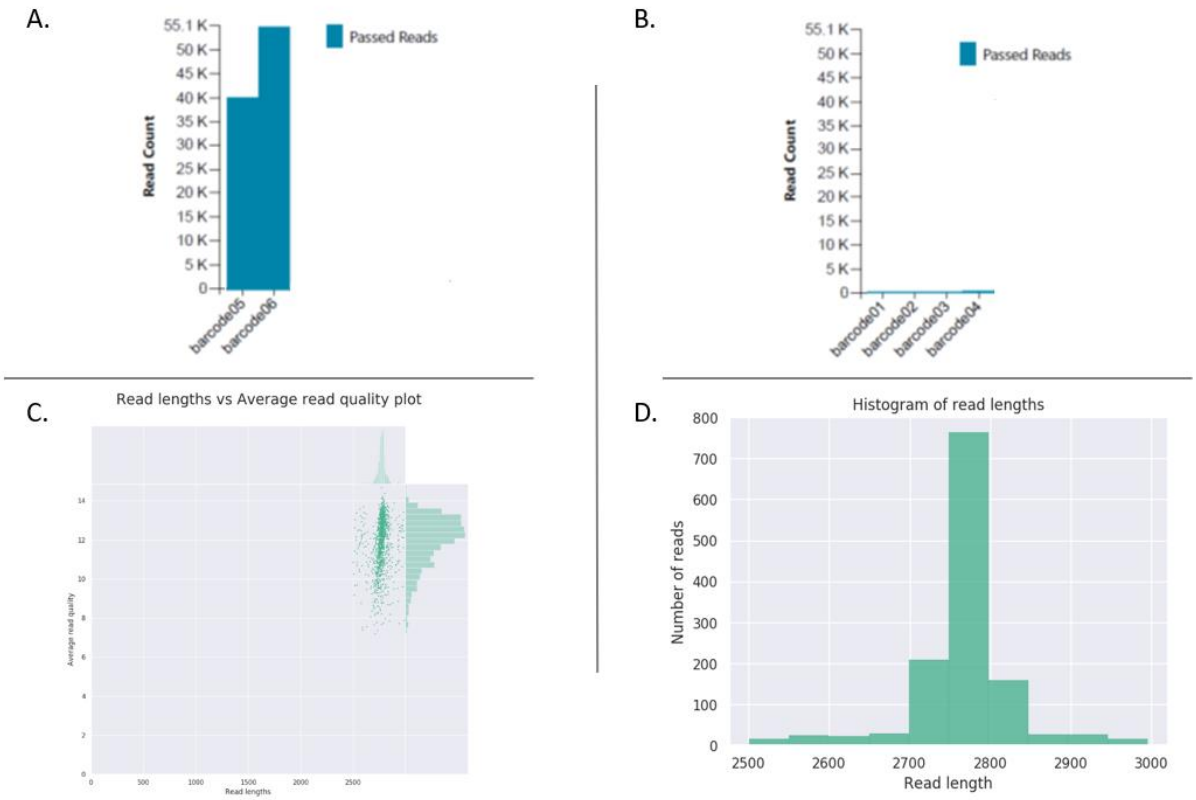


Figure 4.23. MinKNOW and NanoPlot read counts and quality for HeLa and 293T ERAP1 amplicons

Long read sequencing analysis of ERAP1 amplification; BRC5= HeLa amplicon of 25 PCR cycles, BRC06=293T amplicon of 25 PCR cycles (A). Reads generated for 293T and HeLa ERAP1 amplified for 5 and 15 cycles (B). BRC01=HeLa amplicon of 5 cycles, BRC02=293T amplicon of 5 cycles, BRC03=HeLa amplicon of 15 cycles, BRC04=293T amplicon of 15 cycles. ERAP1 allotype identification was not successful for those 4 amplicons. Read lengths vs average read quality plot for HeLa amplicon of 25 PCR cycles (C). Histogram of read lengths for the HeLa amplicon of 25 PCR cycles, reads have passed the filtering based on length ranging from 2,500 to 3,000bp (D). 293T data not included as they were similar with those generated for HeLa cells.

The above experiment was repeated using the rest of the six barcodes contained in the PCR expansion kit (BRC07-BRC12). The number of available pores at the beginning of the sequencing run was 1579, indicating a decrease of 200 pores within 24 hours, with sequencing starting with the best 477 pores. Sequencing of the prepared library generated a total of 94,000 reads over approximately 12 minutes which showed that a higher number of reads could be generated on a relatively new flow cell over a short period of time. Approximately 73% of acquired reads contained the barcodes that were attached

to the HeLa and 293T amplicons from 25 PCR cycles, matching the results obtained from the previous run. It should be noted that the Epi2me report showed a staggering 21% of cross-contamination even after the wash protocol was completed. The percentage of cross-contamination was later <5% with the novel wash kit developed by ONT which enabled the sequencing of up to 12 amplicons in a single run.

Only the ERAP1 allotypes of the amplicons resulting from 25 cycles from both HeLa and 293T cells were identified. All experiments, including these two, confirmed that the lowest number of PCR cycles that could be used for successful identification of ERAP1 allotypes using MinION was 25, but this was not the case for the patient samples as shown later in the chapter (Figure 4.24).

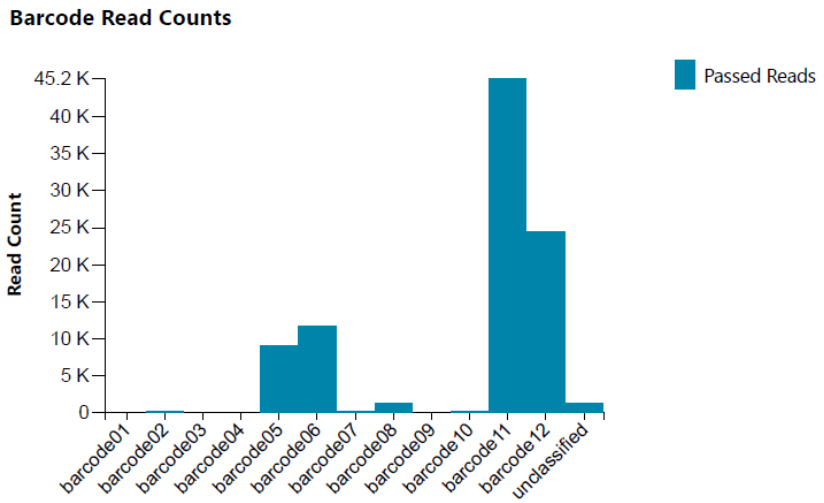


Figure 4.24. MinKNOW analysis of barcode read counts for ERAP1 amplified from HeLa and 293T cells

Long read sequencing analysis of ERAP1 amplification by MinION. BRC07= HeLa amplicon of 5 PCR cycles, BRC08= 293T amplicon of 5 PCR cycles, BRC09= HeLa amplicon of 15 PCR cycles, BRC10= 293T amplicon of 15 cycles, BRC11= HeLa amplicon of 25 PCR cycles, BRC12= 293T amplicon of 25 cycles. There was an increase in percentage of cross-contamination seen from the number of reads containing BRC01-BRC06 from the previous barcoding experiment (Figure 4.23).

4.2 Identification of ERAP1 allotype combinations from a cervical cancer patient cohort using long read sequencing

The previous long sequencing runs with MinION using ERAP1 amplicons from 293T and HeLa cells to identify their ERAP1 allotypes enabled the establishment and hence optimisation of a methodological pipeline, including a bioinformatics analysis pipeline, that could be used for ERAP1 allotyping from the cervical cancer patient cohort. To test the accuracy of the methodological pipeline for sequencing with MinION that was previously established using HeLa and 293T cell lines, a sequencing run was carried out using four cDNA samples from the cervical cancer patient cohort. This experiment enabled determining whether the methodological pipeline established using cell line amplicons, specifically minimum cDNA concentration (as RNA concentrations were unavailable) and number of PCR cycles that resulted in an ERAP1 amplicon as well as pore availability for ERAP1 allotyping (Figure 4.25).

The total number of cervical cancer patient cDNA samples was 103, and the cohort is detailed in Chapter 3. Samples are numbered from S1 to S119, however 16 samples were not available for sequence analysis; S7, S8, S13, S16, S25, S90, S91, S92, S93, S94, S95, S96, S97, S98, S99 and S100. Overall, ERAP1 allotypes were identified for a total of 81/103 cervical cancer patients, and for three patient samples, S70, S71 and S101, ERAP1 allotypes were verified through Sanger sequencing as well as long read sequencing as a method to verify the long read sequencing data. ERAP1 allotypes were not identified for the remaining 22 patients and despite multiple efforts with ERAP1 amplification and sequencing, these runs failed to generate a sufficient read count for successful ERAP1 allotyping. These were S1, S4, S12, S15, S17, S18, S20, S24, S39, S42, S50, S57, S65, S67, S79, S80, S81, S83, S84, S85, S86, S87 and S88. For a number of samples out of the total of 103, the cDNA material available was not sufficient to prepare ERAP1 amplicons by PCR that would be used for sequencing. All the identified ERAP1 allotypes can be found in Table 4.4. Read counts per allotype can be found on the same table. For samples S10, S45, S53, S59, S68, S74, S101 and S110, there were no phased

haplotypes. That is because whatshap only phases heterozygotes, and phasing is not possible when only a single heterozygotic set of the 10 SNPs was identified. However, Sanger sequencing confirmed the allotypes identified for S101 (Table 3.5).

For the first experiment using samples S1, S2, S3 and S4, only 25 PCR cycles were completed, and amplicons were not visible on the agarose gel. It was of interest to investigate whether this number of cycles would result in a sufficient amount of DNA from patient material for successful allotype identification before attempting to use 35 PCR cycles since allotyping was possible for ERAP1 that was amplified from HeLa cDNA by PCR using 25 PCR cycles in one of the trial sequencing runs above (section 4.1.3.3). As the RNA concentration of the patient samples was not available and cDNA concentration cannot be accurately measured, it was decided to attempt amplification of ERAP1 using different two-fold dilutions of the S2 template cDNA (Figure 4.25). These were 1:2, 1:4 and 1:6 dilutions and a total of 1µl template cDNA was used for three separate ERAP1 amplification by PCR using 25 PCR cycles. A total of 1µl of undiluted S2 template cDNA was used for ERAP1 amplification by PCR using 25 PCR cycles and 1ul of plasmid DNA (ERAP1 ligated in pcDNA3) was used as a positive control resulting in a visible band representing 2.7Kb ERAP1 DNA on the agarose gel (Figure 4.25). ERAP1 amplicons prepared from varying dilutions of S2 template cDNA were run on an agarose gel (Figure 4.25). When the concentration of the four amplicons prepared from different dilutions of the S2 template cDNA was determined, the one with the highest concentration was the undiluted S2 with concentration of 2.06 ng/µl and hence the rest of the amplicons were disregarded. The flow cell check performed before sequencing indicated that there were 1539 available pores and sequencing began with 459 pores (Figure 4.26). A total of 5.54fmol of library was loaded onto the flow cell and the length of the sequencing run was approximately 18 minutes, generating 93,148 reads. However, the average length of sequencing reads was 244 bases, which was well below the expected size for ERAP1 (2.7Kb) and reads were basecalled with MinKNOW and de-multiplexed with Epi2me (Figure 4.27). Of the total

read count, a range of 13,000-30,000 reads were generated for each barcoded amplicon of the run (BRC01-04, Figure 4.27).

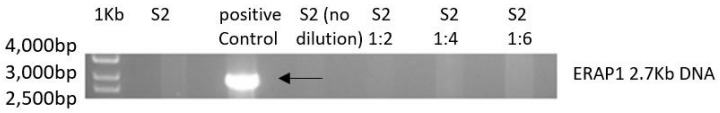


Figure 4.25. ERAP1 amplification by 25 PCR cycles using cervical cancer patient cDNA and ERAP1 specific primers

ERAP1 was amplified from patient cDNA using ERAP1-ONT specific primers for 25 PCR cycles. **Lane 1:** amplification of ERAP1 from 1µl patient 2 (S2) cDNA using ERAP1 specific primers for full length 2.7Kb. **Lane 2:** positive control from ERAP1 in pcDNA3.1 vector. **Lanes 3-6**: amplification of ERAP1 from patient 2 (S2) cDNA using ERAP1 specific primers for 2.7Kb ERAP1. No dilution, 1:2, 1:4, 1:6 dilutions of S2 cDNA correspond to lanes 3-6 respectively. PCR products were run on a 1% agarose electrophoresis gel. Band represents 2.7Kb ERAP1 DNA.

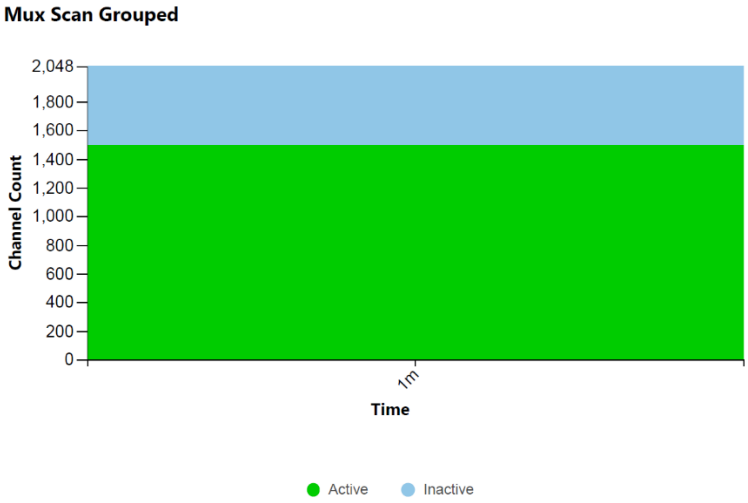


Figure 4.26. Mux scan generated using MinKNOW shows the active pores available for the sequencing of amplicons prepared from S1, S2, S3 and S4 using 25 PCR cycles

A flow cell check was carried out before sequencing and the mux scan showed that the total number of active, available pores for sequencing was 1,500. Active pores are shown in green and inactive pores that are unavailable for sequencing are shown in blue. ONT recommendation is minimum of 800 pores but ERAP1

allotyping was previously shown to be completed successfully with a lower pore number after the flow cell check.

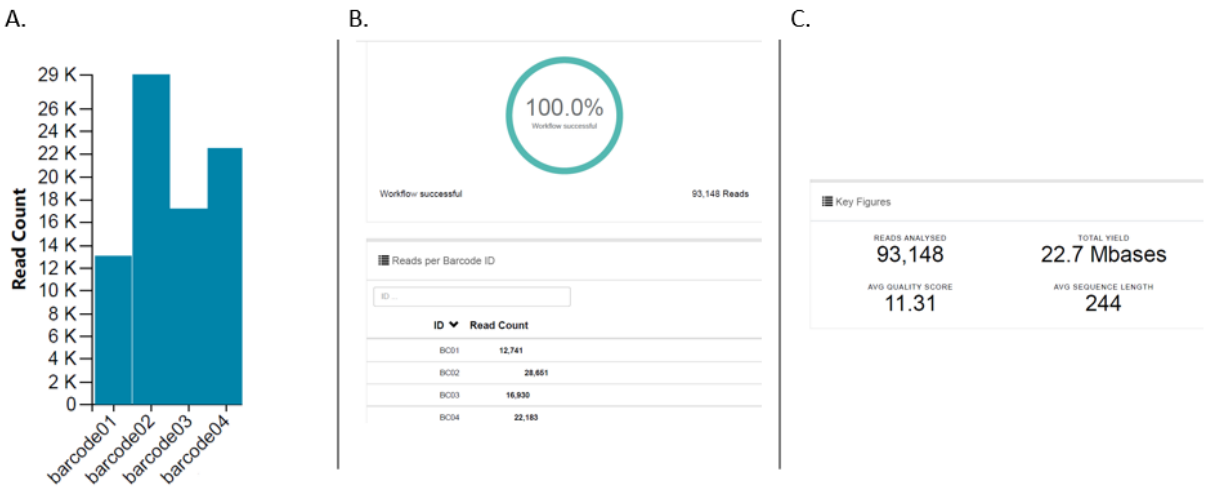


Figure 4.27. Sequencing run data generated with MinKNOW and Epi2me showing the read count generated for S1, S2, S3 and S4

ERAP1 was amplified from S1 (BRC01), S2 (BRC02), S3 (BRC03) and S4 (BRC04) using 25 PCR cycles and resulting amplicons were sequenced together in the same run. Although a high number of reads was generated for all barcoded amplicons, they did not pass the read filtering by NanoPlot for successful ERAP1 allotyping indicating that ERAP1 amplification was too low for sequencing and for accurately measuring ERAP1 DNA concentration.

It is noteworthy that following filtering of reads based on length (2500-3000bp) using NanoPlot and NanoFlit as well as mapping of those reads to the ERAP1 reference sequence using Minimap2, it was shown that there were either none or one read close to the expected ERAP1 length that also mapped to the reference sequence (Figure 4.28).

When S3 was amplified using 35 PCR cycles and ERAP1 was sequenced in a second run with it being the sole amplicon of the run, two novel ERAP1 allotypes were identified, *023 and *024 which contain the amino acid changes R127P/I276M/Q730E and R127P/K528R, respectively. Since a total of more than 14,000 read counts was generated that passed the length and quality score filtering set in

NanoPlot and NanoFilt, further verification of the new allotypes was not deemed necessary (Supplementary data, Appendix B). It is noteworthy, that sequencing of ERAP1 from S3 was completed with less than 50 available pores according to the mux scan showing that for the sequencing of a single gene, a sufficient read count can be generated with less than the recommended 800 pores but sequencing should be run for a longer amount of time.

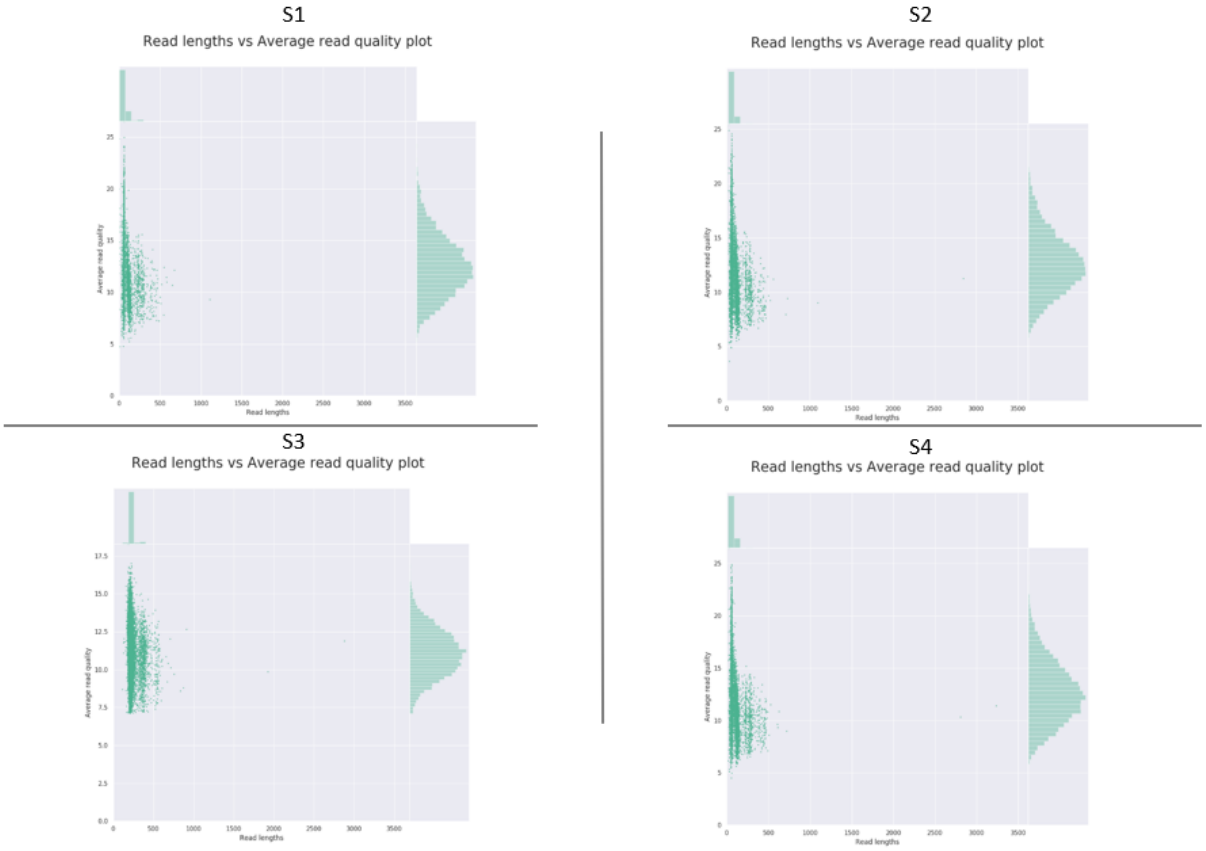


Figure 4.28. Read lengths vs average read quality plots generated for ERAP1 amplified from S1, S2, S3 and S4 using 25 PCR cycles

NanoPlot was used as part of the bioinformatics analysis pipeline to filter the generated sequencing reads between 2500-3000bp for alignment to ERAP1. Analysis for S1-S4 revealed an inadequate number of reads generated for S1, S2, S3 and S4 amplified using 25 PCR cycles for successful ERAP1 allotyping. A high number of short reads below 500bp can be observed in each of these graphs that were later shown to be primer dimers.

From the analysis above it could be presumed that barcodes were ligating to i) DNA fragments that were below the expected size of ERAP1, ii) to sequences of DNA forming concatemers or iii) to primer dimers. Analysis of reads using UGENE (Unipro) and BLAST (NCBI), showed formation of primer dimers to which the barcodes were bound. To try and address this issue, the sequence of the primers was elongated within the ERAP1 region to enable specific binding to ERAP1 and prevent primer dimer formation.

For S5, S6, S9 and S10, ERAP1 was amplified for 35 PCR cycles as the previous experiment involving sequencing of ERAP1 amplified from samples S1 to S4 by PCR using 25 cycles did not lead to successful ERAP1 allotype identification. The agarose gel showed presence of ERAP1 DNA for S9 and S10 after ERAP1 amplification by PCR using 35 PCR cycles (Figure 4.29).

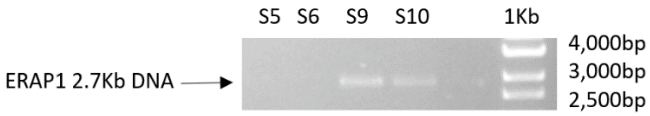


Figure 4.29. ERAP1 amplification from cervical cancer patient cDNA

ERAP1 amplification using ERAP1-ONT specific primers and 35 PCR cycles. **Lane 1, 2, 3, 4**: amplification of ERAP1 from patients 5, 6, 9, 10 (S5, S6, S9, S10) using tailed primers for 2.7Kb ERAP1. All PCR products were run on a 1% agarose electrophoresis gel. Band represents 2.7Kb ERAP1 DNA indicated by an arrow.

Qubit was used to measure the concentration of the amplicons from samples S5, S6, S9 and S10 (*note: cDNA from S7 and S8 was not available*). The library was sequenced over 8 minutes resulting in generation of 61,172 reads. Even though the number of reads generated per sample as reported by Epi2me ranged between 8,000 and 25,000 (Figure 4.30, Figure 4.31), when reads were visualised in IGV, the read counts for S5, S6, S9 and S10 were 47, 1362, 1523 and 193, respectively.

Figure 4.30. Total read counts generated for ERAP1 amplified from S5, S6, S9 and S10 using 35 PCR cycles

Data generated with MinKNOW and Epi2me reveal a high number of reads generated for ERAP1 amplified from S5 (BRC08), S6 (BRC10), S9 (BRC11) and S10 (BRC12). The average read length was just 386bp and these short reads were later shown to be primer dimers. Interestingly, the two amplicons with the highest read count close to the ERAP1 expected length were S9 and S10 which were also the only amplicons with visible ERAP1 amplification on the agarose gel indicating that their concentration was more accurately measured compared to the other two amplicons and hence been preferentially sequenced.

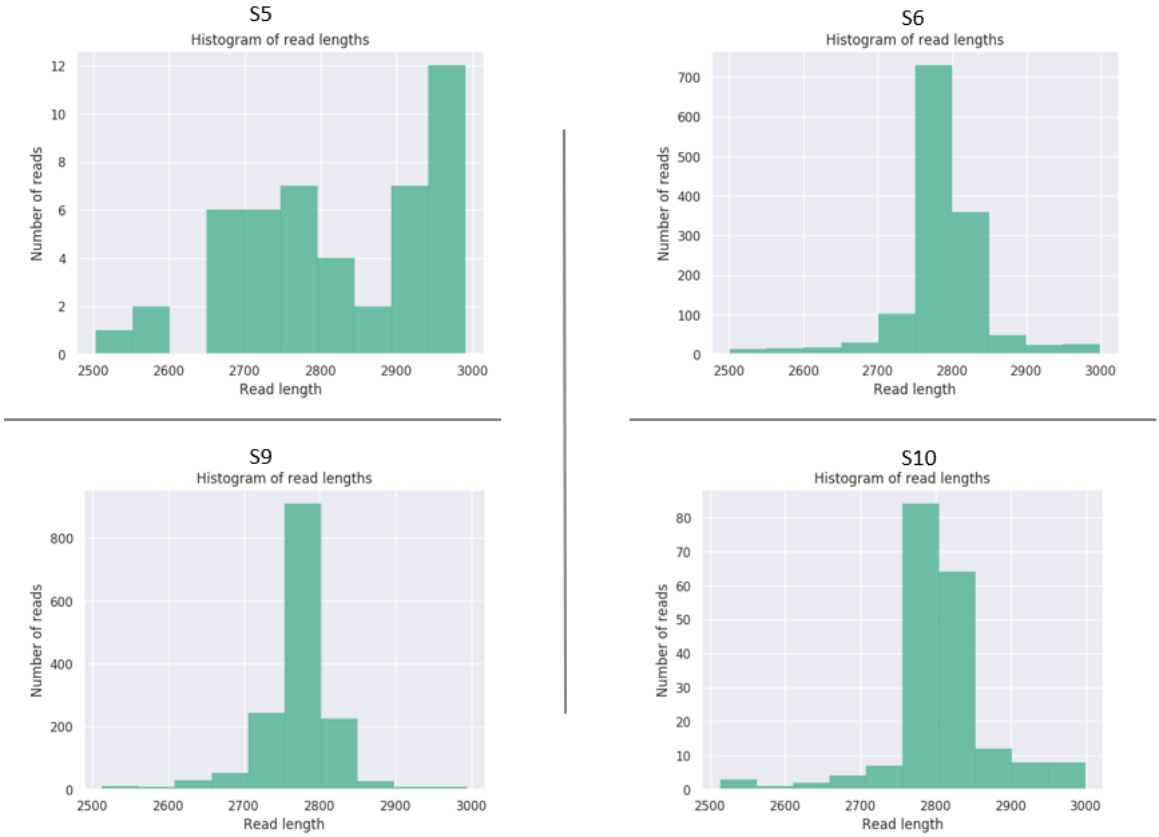


Figure 4.31. Histograms showing read count and relative length for S5, S6, S9 and S10 amplicons

NanoPlot was used as part of the bioinformatics analysis pipeline and it revealed that the reads that passed the length and quality score filtering are close to the expected length of ERAP1 (2.7Kb) and could therefore be used for successful ERAP1 allotyping. A relatively small number of reads close to the expected ERAP1 length was generated for S5 but ERAP1 allotypes were nonetheless identified.

S6 was shown to be homozygotic for allotype *013. The unavailability of genomic DNA for the patient samples made confirmation of true homozygotes difficult as amplification of each exon region of ERAP1 containing the SNPs would be used to verify true homozygosity. However, a second sequencing experiment undertaken for S51 that was shown to be homozygotic for *002-WT with a read count of just 20 the first experiment, still revealed only one allotype with 2,340 reads in the second sequencing run, increasing confidence in ERAP1 allotyping for homozygotic samples by confirming the original findings (Supplementary data, Appendix B2). This was also an important outcome as it showed that ERAP1 allotyping is possible even with a read count of just 20 (Table 4.4).

The completion of the above experiments as well as that of the trial sequencing runs involving the ERAP1 amplicons prepared using cDNA from the two cell lines, 293T and HeLa, lead to the decision to carry out ERAP1 amplification for 35 PCR cycles using patient cDNA as template, initially in groups of four. Later the number of barcoded samples increased to that of the maximum available unique barcodes in the PBC001 barcoding kit, BRC01-BRC12.

S11 to S24 were sequenced in three separate sequencing runs in groups of four. 2.7Kb ERAP1 DNA was visible for samples 11, 21 and 23 and the allotypes from all three samples were identified (Table 4.4). It was shown that the majority of the generated reads for each of the twelve samples were less than 300 bases long which turned out to be primer dimers that were shown to be reduced upon modification of the primer set used for ERAP1 amplification. A sufficient number of reads for allotyping was not successfully generated for S12, S15, S17, S18 and S20.

ERAP1 amplicons that were visible on the agarose gel were sequenced in the same run. When S21, S22, S23 and S24 were sequenced together, only the ERAP1 allotypes of S21 and S23 were identified, and these amplicons were visible when run on the agarose gel (Table 4.4). The inherent accuracy of Qubit was confirmed through the sequencing of ERAP1 amplified by a different number of PCR cycles; a higher number of reads close to the expected length of ERAP1 were generated for ERAP1 that had visible ERAP1 amplification on the agarose gel and a higher number of pores actively sequencing ERAP1 with visible amplification on the gel (Figure 4.34).

For sequencing of S19, only 18 reads passed the filtering and thus it was re-sequenced, generating a total number of 35 reads that were used for allotyping. Despite the fact that only 15 reads were generated for S14, two distinct ERAP1 allotypes were identified and this sample was included in the analysis, yet with caution as ERAP1 allotyping has not been confirmed from the trial sequencing runs for read count below 18 (Section 4.1.3.4 and Table 4.4).

Table 4.4. ERAP1 allotypes and combinations identified from 81 cervical cancer patients at varying stages of disease and read counts per allotype

ERAP1 allotypes were identified from 81 patients using long read sequencing. The read counts generated per allotype are also shown below. For the homozygotic combinations a single read count is shown. For S10, S45, S53, S59, S68, S74, S101 and S110 there were no phased haplotypes because phasing was not possible with a single heterozygotic set of the 10 SNPs investigated. Novel ERAP1 allotypes identified from the patient cohort are shown in red and bold letters.

Samples	ERAP1	ERAP	Amino acid changes at indicated positions										Run 1	Run2	Unphased haplotypes
			12	56	127	276	346	349	528	575	725	730	Number of r	Number of reads	
			T/I	E/K	R/P	I/M	G/D	M/V	K/R	D/N	R/Q	Q/E	per allotype	per allotype	

S2	*018	Hap3	T	E	R	I	G	M	K	D	R	E		38	
S3	*023	N/A												7340	
	*024	N/A	T	E	P	I	G	M	R	D	R	Q		4801	
S5	*025	N/A	T	E	P	M	D	M	R	D	R	E	17		
	*013	Hap1	T	E	P	I	G	M	K	D	R	Q	25		
S6	*013	Hap1	T	E	P	I	G	M	K	D	R	Q	1362		
S9	*013	Hap1	T	E	P	I	G	M	K	D	R	Q	825		
	*021	Hap8	T	E	P	M	G	M	R	D	R	E	434		
S10	*013	Hap1	T	E	P	I	G	M	K	D	R	Q			193
	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q			
S11	*018	Hap3	T	E	R	I	G	M	K	D	R	E	52		
	*021	Hap8	T	E	P	M	G	M	R	D	R	E	36		
S14	*018	Hap3	T	E	R	I	G	M	K	D	R	E	9		
	*013	Hap1	T	E	P	I	G	M	K	D	R	Q	3		
S19	*021	Hap8	T	E	P	M	G	M	R	D	R	E	18	10	
	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q		15	
S20	*013	Hap1	T	E	P	I	G	M	K	D	R	Q		11	
														7	
S21	*018	Hap3	T	E	R	I	G	M	K	D	R	E	15		
	*013	Hap1	T	E	P	I	G	M	K	D	R	Q	61		
S22	*021	Hap8	T	E	P	M	G	M	R	D	R	E		11	
	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q		35	
S23	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	44		
	*015	Hap6	T	E	P	I	G	M	R	D	R	E	33		
S26	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q	387		
	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	644		
S27	*021	Hap8	T	E	P	M	G	M	R	D	R	E	61		
	*013	Hap1	T	E	P	I	G	M	K	D	R	Q	119		
S28	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q	22	6048	
	*021	Hap8	T	E	P	M	G	M	R	D	R	E		2369	
S29	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	2489		
	*021	Hap8	T	E	P	M	G	M	R	D	R	E	307		
S30	*023	N/A	T	E	P	M	G	M	K	D	R	E	38	2037	
	*024	N/A	T	E	P	I	G	M	R	D	R	Q		1669	
S31	*013	Hap1	T	E	P	I	G	M	K	D	R	Q	75		
	*015	Hap6	T	E	P	I	G	M	R	D	R	E	37		
S32	*019	Hap5	T	E	R	I	D	M	R	D	R	E	550		
	*015	Hap6	T	E	P	I	G	M	R	D	R	E	484		
S33	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	12	44	
	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q		9	

S34	*019	Hap5	T	E	R	I	D	M	R	D	R	E	495		
	*015	Hap6	T	E	P	I	G	M	R	D	R	E	94		
S35	*013	Hap1	T	E	P	I	G	M	K	D	R	Q	103		
	*015	Hap6	T	E	P	I	G	M	R	D	R	E	611		
S36	*013	Hap1	T	E	P	I	G	M	K	D	R	Q	305		
	*018	Hap3	T	E	R	I	G	M	K	D	R	E	655		
S37	*021	Hap8	T	E	P	M	G	M	R	D	R	E	405		
	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	220		
S38	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	3895		
	*021	Hap8	T	E	P	M	G	M	R	D	R	E	647		
S40	*018	Hap3	T	E	R	I	G	M	K	D	R	E	38		
	*021	Hap8	T	E	P	M	G	M	R	D	R	E	107		
S41	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q	809		
	*015	Hap6	T	E	P	I	G	M	R	D	R	E	507		
S43	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	164		
	*013	Hap1	T	E	P	I	G	M	K	D	R	Q	66		
S44	*013	Hap1	T	E	P	I	G	M	K	D	R	Q	143		
S45	*015	Hap6	T	E	P	I	G	M	R	D	R	E			276
	*021	Hap8	T	E	P	M	G	M	R	D	R	E			
S46	*018	Hap3	T	E	R	I	G	M	K	D	R	E	389		
S47	*014	Hap7	T	K	P	I	G	M	R	D	R	E	130		
	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	220		
S48	*015	Hap6	T	E	P	I	G	M	R	D	R	E	109		
	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	58		
S49	*015	Hap6	T	E	P	I	G	M	R	D	R	E		683	
S51	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q	20	2340	
S52	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	199		
S53	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q			383
	*013	Hap1	T	E	P	I	G	M	K	D	R	Q			
S54	*018	Hap3	T	E	R	I	G	M	K	D	R	E	198		
S55	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	2	267	
S56	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q	352		
	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	148		
S58	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q	18	589	
	*021	Hap8	T	E	P	M	G	M	R	D	R	E		166	
S59	*015	Hap6	T	E	P	I	G	M	R	D	R	E			2418
	*021	Hap8	T	E	P	M	G	M	R	D	R	E			

S60	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	17		
	*021	Hap8	T	E	P	M	G	M	R	D	R	E	89		
S61	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q	148		
	*015	Hap6	T	E	P	I	G	M	R	D	R	E	63		
S62	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q	1108		
S63	*013	Hap1	T	E	P	I	G	M	K	D	R	Q	10	1059	
S64	*021	Hap8	T	E	P	M	G	M	R	D	R	E	700		
S66	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	513		
	*021	Hap8	T	E	P	M	G	M	R	D	R	E	83		
S68	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q			34/1426
	*013	Hap1	T	E	P	I	G	M	K	D	R	Q			
S69	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	10524		
S70	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	4509		
	*015	Hap6	T	E	P	I	G	M	R	D	R	E	464		
S71	*018	Hap3	T	E	R	I	G	M	K	D	R	E	5985		
	*013	Hap1	T	E	P	I	G	M	K	D	R	Q	3184		
S72	*018	Hap3	T	E	R	I	G	M	K	D	R	E	2	171	
	*013	Hap1	T	E	P	I	G	M	K	D	R	Q		444	
S73	*018	Hap3	T	E	R	I	G	M	K	D	R	E	185		
	*021	Hap8	T	E	P	M	G	M	R	D	R	E	123		
S74	*015	Hap6	T	E	P	I	G	M	R	D	R	E			1159
	*021	Hap8	T	E	P	M	G	M	R	D	R	E			
S75	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q		26	
S76	*021	Hap8	T	E	P	M	G	M	R	D	R	E	233		
S77	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q	95		
	*021	Hap8	T	E	P	M	G	M	R	D	R	E	65		
S78	*018	Hap3	T	E	R	I	G	M	K	D	R	E	139		
	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	61		
S82	*026	N/A	T	E	P	I	G	V	K	N	Q	E		282	
	*024	N/A	T	E	P	I	G	M	R	D	R	Q		169	
S89	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q	128		
S101	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q			1053
	*018	Hap3	T	E	R	I	G	M	K	D	R	E			
S102	*018	Hap3	T	E	R	I	G	M	K	D	R	E	438		
	*021	Hap8	T	E	P	M	G	M	R	D	R	E	198		
S103	*018	Hap3	T	E	R	I	G	M	K	D	R	E	85		
	*027	N/A	T	E	P	M	G	M	R	D	R	Q	23		

S104	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q	240		
	*028	N/A	T	E	P	I	G	M	K	D	R	E	52		
S105	*018	Hap3	T	E	R	I	G	M	K	D	R	E	714		
	*015	Hap6	T	E	P	I	G	M	R	D	R	E	388		
S106	*013	Hap1	T	E	P	I	G	M	K	D	R	Q	3752		
S107	*021	Hap8	T	E	P	M	G	M	R	D	R	E	1136		
S108	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	99		
	*014	Hap7	T	K	P	I	G	M	R	D	R	E	32		
S109	*013	Hap1	T	E	P	I	G	M	K	D	R	Q	26	1	
	*021	Hap8	T	E	P	M	G	M	R	D	R	E	22		
S110	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q			163
	*013	Hap1	T	E	P	I	G	M	K	D	R	Q			
S111	*018	Hap3	T	E	R	I	G	M	K	D	R	E	577		
	*021	Hap8	T	E	P	M	G	M	R	D	R	E	156		
S112	*002-WT	Hap2	T	E	R	I	G	M	K	D	R	Q	429		
S113	*018	Hap3	T	E	R	I	G	M	K	D	R	E	153		
	*013	Hap1	T	E	P	I	G	M	K	D	R	Q	120		
S114	*021	Hap8	T	E	P	M	G	M	R	D	R	E	993		
S115	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	564		
	*021	Hap8	T	E	P	M	G	M	R	D	R	E	243		
S116	*021	Hap8	T	E	P	M	G	M	R	D	R	E	678		
S117	*015	Hap6	T	E	P	I	G	M	R	D	R	E	621		
S118	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	266		
S119	*001	Hap10	T	E	P	I	G	V	R	N	Q	E	23		
	*021	Hap8	T	E	**P**	**M**	G	M	**R**	D	R	**E**	12		

As it was shown in previous experiments, a number of short reads were generated from the sequencing runs involving both cell line and patient ERAP1 amplicons. It was later understood that these are primer dimers and hence a new primer set of ERAP1 specific tailed primers was designed which was elongated to enable specific binding to the ERAP1 gene. The new set of primers can be found in Methods and Materials chapter, section 2.3.3.1.

S29, S32, S35 and S36 ERAP1 amplicons were generated using the new set of primers and all of them were visible on the agarose gel and therefore sequenced together in the same run. This was the first run that involved sequencing of four ERAP1 amplicons all of which were visible on the agarose gel and also the first amplicons that were generated using the new set of primers. They were quantified with Qubit and measurements before and after the barcoding PCR are presented in Table 4.5.

Table 4.5. Concentrations of ERAP1 amplicons S29, S32, S35 and S36 using the new primer set measured with Qubit

Samples	PCR (ng/µl)	Barcoding PCR (1:2 dilution, ng/µl)
S29	12.5	56
S32	7.95	59
S35	4.82	57
S36	4.61	43.9

The samples received barcodes BRC09, BRC10, BRC11 and BRC12, respectively. The run was completed in twelve minutes on a flow cell that had 1,302 pores available for sequencing according to the mux scan completed before sequencing (Figure 4.32). The Epi2me report showed that a total of 6,665 reads were generated for these samples with the specific read count for each barcoded sample of this run shown on Table 4.4. The average sequence length was 2,856, close to the expected length of the ERAP1 gene (2.7Kb).

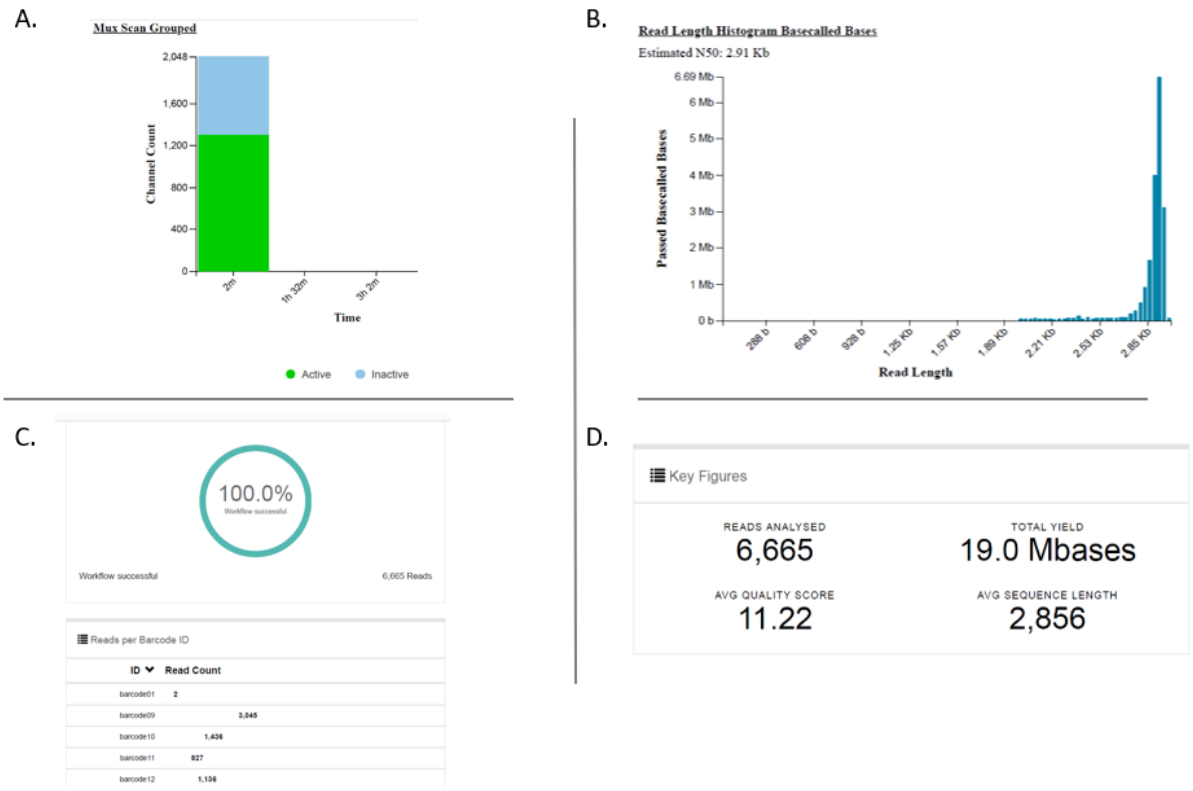


Figure 4.32. Data generated with MinKNOW, Epi2me and NanoPlot for the sequencing of amplicons S29, S32, S35 and S36.

The mux scan revealed the total number of available pores for sequencing before the run took place (A). The N50 was 2.9Kb which is close to the expected length of ERAP1 (2.7Kb, B) and this was confirmed by the average read length reported with Epi2me (D). Data generated with MinKNOW. The Epi2me report showed the number of demultiplexed read counts generated for amplicons S29, S32, S35 and S36 which had visible ERAP1 amplification on the agarose gel (C).

Bioinformatics data analysis that involved further filtering of the reads based on their length (2,500-3,000b) and read quality, indicated that 2,836 reads passed the filtering for S29 leading to successful ERAP1 allotyping. The relevant number for S32 was 1,254 and for S35 and S36 it was 766 and 1,047 reads, respectively as indicated by NanoPlot (Figure 4.33). These data indicate that when ERAP1 amplicons are visible on an agarose gel, hundreds of reads of length close to that of the ERAP1 gene are generated upon sequencing. It is noteworthy, that all four ERAP1 allotype combinations were

different from each other with only allotypes *013 and *015 seen in two patients, yet these were found in three distinct combinations (*019 + *015, *013 + *015 and *013 + *018).

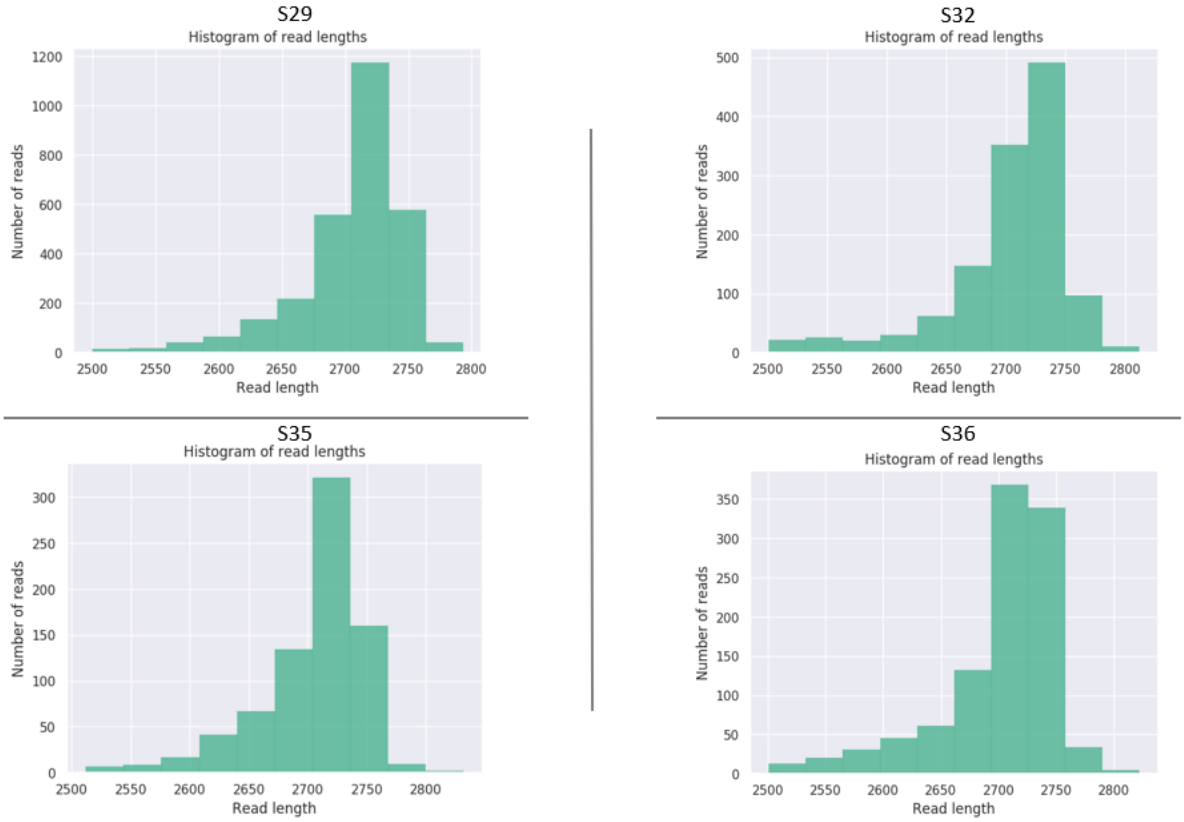


Figure 4.33. Histogram of read lengths generated for S29, S32, S35 and S36 using NanoPlot.

The histograms generated for S29, S32, S35 and S36 depict the number of reads generated for each amplicon that have passed the filtering settings. The majority of reads are close to the expected length of ERAP1 (2.7Kb).

S26, S27, S28 and S30 did not show visible ERAP1 amplification and were pooled together and sequenced in the same run (Supplementary data, Appendix B3). A considerably lower number of reads that passed both read length and quality filtering was generated for these four amplicons for which ERAP1 amplification was not visible on the agarose gel. The allotypes of S28 and S30 were identified in a second run and reads for both runs can be found on Table 4.4. It is noteworthy that the two

allotypes identified from this amplicon matched those identified for S3, both of which were novel (*023 and *024).

A reason behind the fact that only the allotypes of two amplicons were identified in the first run is that Qubit likely underestimates the concentration of the amplicons that do not show visible ERAP1 amplification on the agarose gel. This probably lead to the preparation of a library that was below the recommended minimum 5fmol for loading onto the flow cell, which can be confirmed by the number of pores that were actively sequencing during that run (<50%) compared to over 60% of pores actively sequencing S29, S32, S35 and S36 (Figure 4.34).

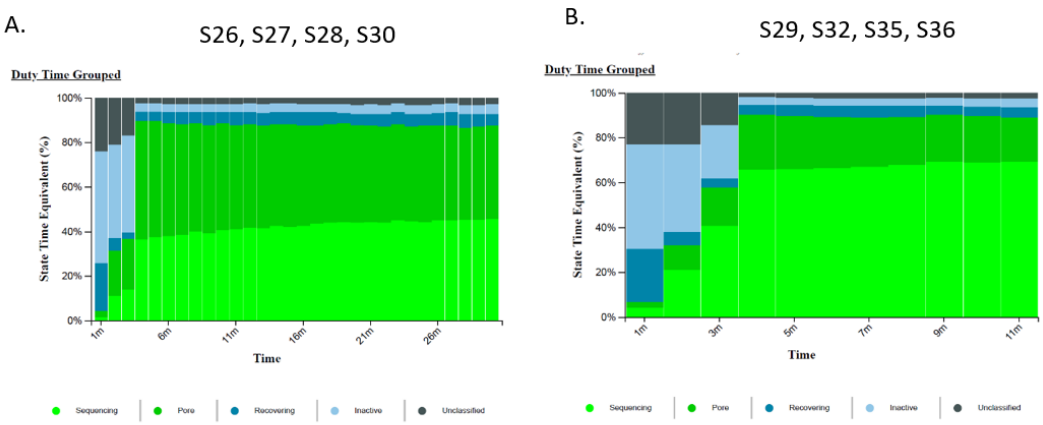


Figure 4.34. Duty time plot showing the state of the nanopores during sequencing for two different runs

The MinKNOW duty time plot for run 1 S26, S27, S28 and S30 (A) shows that less than 50% of available pores are actively sequencing during the entire time that the run was executed (light green). In contrast to this, the duty time plot generated for run 2 S29, S32, S35 and S36 (B) shows that more than 60% of available pores are actively sequencing ERAP1. The difference could lie in the inability of Qubit to accurately measure the concentration of ERAP1 amplicons that did not show visible ERAP1 amplification on the agarose gel and hence that of the resulting library to be loaded on the flow cell.

S42, S43, S44 and S45 did not reveal visible presence of ERAP1 2.7Kb DNA and were sequenced together which was also the case for samples S44, S47, S48 and S49. Of note, the ERAP1 allotypes

from samples S44 and S49 were only identified upon a second sequencing run (Table 4.4). Following discussions and method recommendations by Dr Jade Forster (Cancer Sciences Genomics Lab), a sequencing run involving ERAP1 allotyping from more than four amplicons was completed.

There was an insufficient number of reads for ERAP1 identification from the first sequencing run for samples S33, S54 and S55, however following a second experiment, the ERAP1 allotype combination of these 3 samples were identified (Table 4.4). Data showed that S54 and S55 were homozygotes for allotypes *018 and *001, respectively (Table 4.4).

Interestingly, when S58 was sequenced, only 18 reads passed filtering by NanoPlot (2500-3000bp), however these 18 reads allowed the identification of two distinct ERAP1 sequences, *021 and a potentially novel ERAP1 allotype. To confirm the presence of the novel ERAP1 allotype, S58 was sequenced in a second run, generating 769 reads, a notably higher read count than the first run (Table 4.4). This time, the two ERAP1 allotypes identified were *021, which was one of the allotypes identified from the first run, and allotype *002-WT. It was therefore decided that for amplicons identified from a relatively low read count was generated (less than 50), and also for which a novel allotype was observed, the sequencing of this amplicon sample would be repeated to confirm the ERAP1 allotype novelty (Table 4.4, read counts).

S68, S69, S70 and S71 were sequenced together as all of these resulted in a visible band on the agarose gel representing 2.7Kb ERAP1 DNA (Table 4.4). When the concentration of amplicons was measured with Qubit and the dsDNA HS kit, the range of quantification was between 4-7ng/µl, which even though it is higher than the concentration measured for the amplicons without ERAP1 DNA presence on the agarose gel, it is possible that this was the actual concentration of ERAP1 measured by Qubit and not the concentration of other double stranded DNA present, such as primer dimers. The final library concentration loaded on the flow cell was 50fmol (Supplementary data, Appendix B4). It is noteworthy, that S70 and S71 were two of the amplicons that were randomly chosen to be sequenced

with Sanger sequencing as well to verify the ERAP1 allotypes with long read sequencing. Sanger sequencing was able to confirm one of the two allotypes for both S70 and S71; *001 and *018, respectively. The other two allotypes, *015 and *013 were identified from samples S70 and S71 using long read sequencing, respectively. These allotypes were not identified with Sanger sequencing due to the limited number of plasmids that were sent for sequencing. As both allotypes were verified with Sanger sequencing for S101, *002-WT and *018, the other amplicon that was sent for sequencing, it is possible that a higher number of plasmids isolated from TOP10 cells would lead to verification for S70 and S71 as well (Figure 4.35).

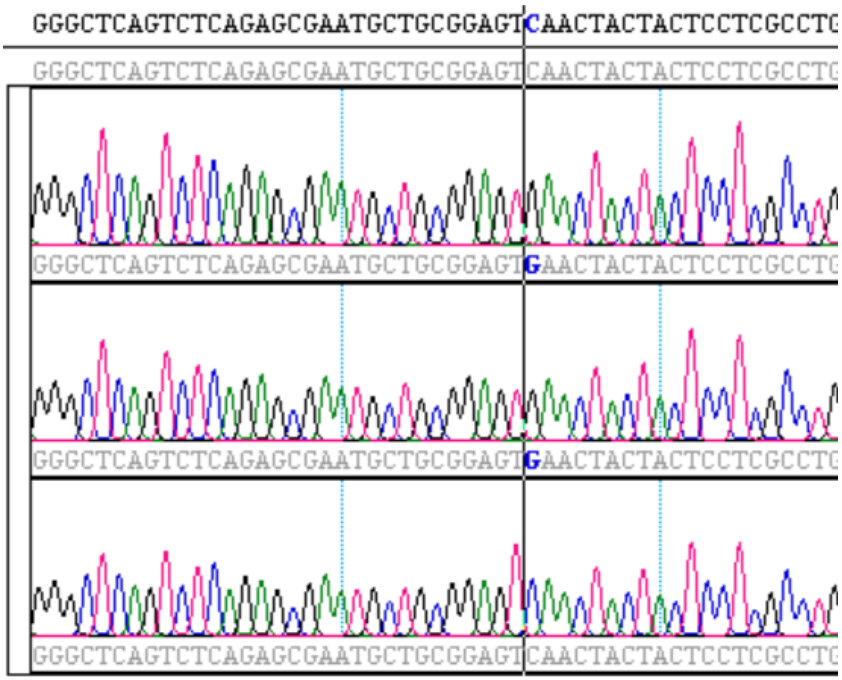


Figure 4.35. Sanger sequencing chromatographs for S101 showing the SNP at position 2188 (Q730E) of ERAP1

The chromatograph generated with Seqman pro (DNASTAR) shows the SNP at position 2188 of ERAP1 where the amino acid change Q730E occurs. The top two chromatographs are from two separate plasmids isolated from TOP10 cells which revealed that one of the two allotypes for S101 is *018 containing the single amino acid change Q730E. The bottom one did not contain any amino acid changes and that was hence the second *002-WT allotype for S101. Both of these were confirmed when ERAP1 from this patient was sequenced using long read sequencing.

S61, S62, S66 and S74 that were sequenced were also visible on the agarose gel with a band representing 2.7Kb ERAP1 DNA, except for S61. This run was carried out to investigate whether the ERAP1 allotypes of an amplicon that was not shown to have presence of ERAP1 DNA on the agarose would be identified when sequenced with another three amplicons that had presence of ERAP1 DNA. The run was carried out over one hour and twenty-one minutes and the mux scan showed that 929 pores were available for sequencing (Figure 4.36). Interestingly, the read count that passed filtering following demultiplexing with Epi2me, was between 600 and 1200 reads for the amplicons that had a visible ERAP1 band on the gel. The relevant read count for S61 was 226 (Figure 4.37). This indicates that Qubit was more accurate at quantifying the ERAP1 concentration of those samples with visible ERAP1 DNA on the agarose gel and therefore the read count generated for these samples was similar compared to that generated for S61. The four distinct ERAP1 allotype combinations from all four amplicons were identified; these were *002-WT + *015 (S61), *002-WT (S62), *001 + *021 (S66) and *015 + *021 (S74) shown in Table 4.4.

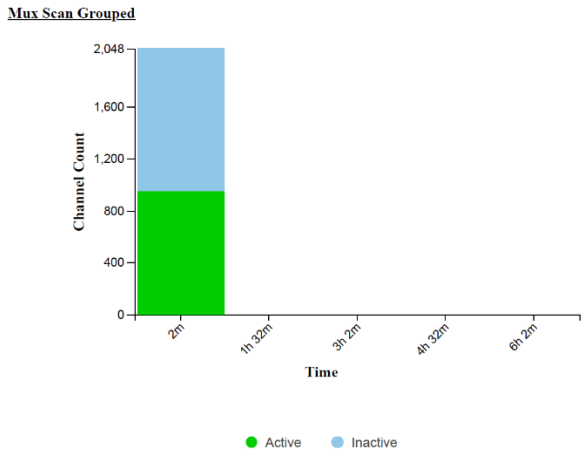


Figure 4.36. Mux scan showing the number of active pores available for sequencing

A flow cell check (Mux scan) was undertaken using MinKNOW before the sequencing run which showed the number of active pores available for sequencing (green) were 929. The Mux scan will repeat every 90 minutes during the sequencing run. Unavailable pores are shown in blue.

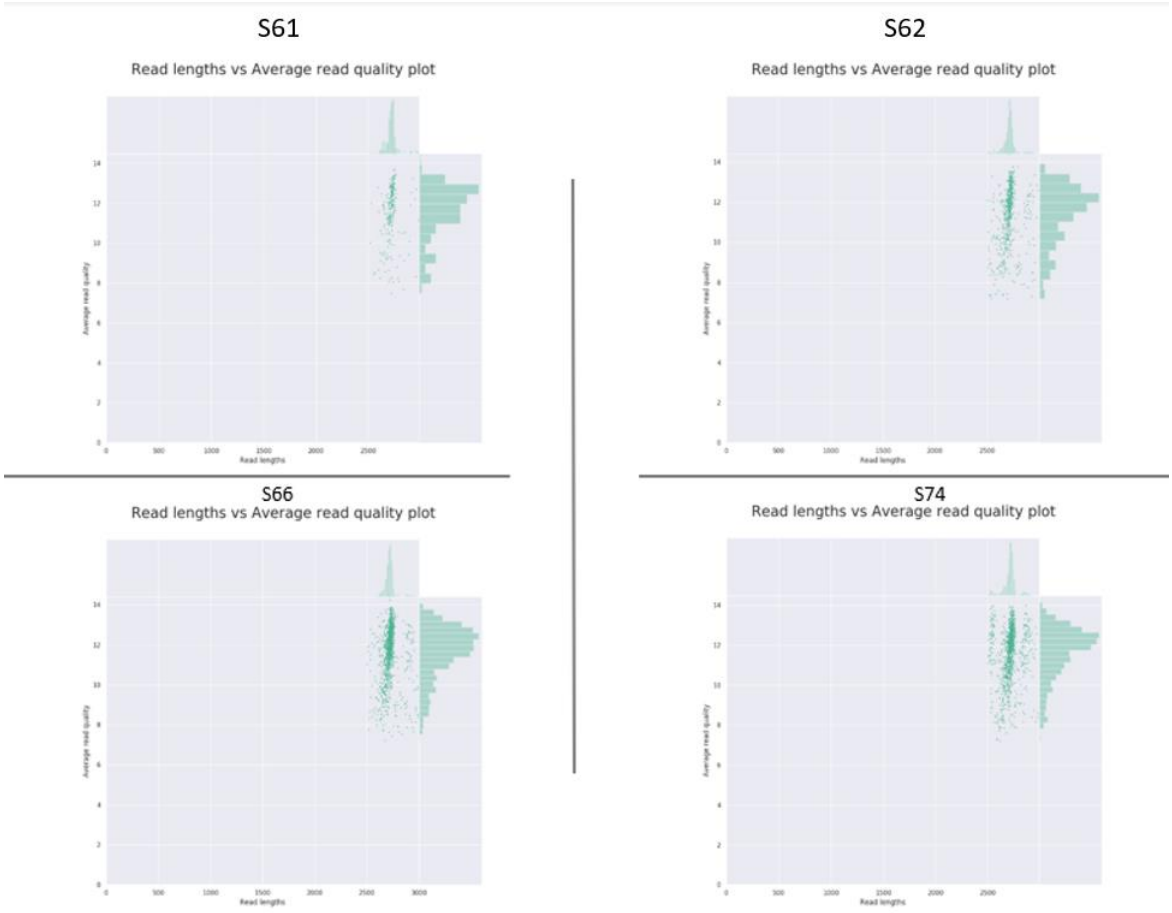


Figure 4.37. Plots showing read lengths vs average read quality scores for S61, S62, S66 and S74 generated with NanoPlot

Plots showing read lengths vs average read quality scores for S62, S66 and S74 that had visible ERAP1 amplification on the agarose gel and for S61 that did not have visible ERAP1 amplification on the agarose gel. This experiment confirmed that allotypes can be successfully identified when a single amplicon without visible ERAP1 amplification is sequenced in the same run with three other amplicons with visible ERAP1 amplification, however a higher number of read counts were generated for S62, S66 and S74 compared to S61 and the reason lies in concentration quantification with Qubit for the amplicons without a band representing ERAP1 2.7Kb DNA on the agarose gel. Data generated with NanoPlot.

The sequencing of S72 and S75 was repeated and ERAP1 allotypes were then successfully identified. From the sequencing run including S63, S64, S65, S67, a sufficient count of 700 reads was used for ERAP1 allotyping from S64, showing that the sample was homozygotic for allotype *021. For S63 and S65, sequencing was repeated but not for S67 due to lack of sufficient amount of cDNA. The allotypes

of S63 were identified from a total of around a thousand reads showing a homozygotic combination for allotype *013. The allotypes of S65 were not identified (Table 4.4).

S77, S101, S102 and S103 had visible ERAP1 amplification on the agarose gel and were sequenced in the same run. Sequencing was carried out using the same flow cell as four previous sequencing runs, revealing that a single flow cell can have enough pores for sequencing even after multiple runs, following the wash protocol to remove the old libraries (Figure 4.38). There was a total of 558 pores available which even though it was below the recommended minimum pore number of 800, ERAP1 allotyping was successful for all four patients (Table 4.4). A total of 2,465 reads were generated with a quality score of 11.12 (Figure 4.38). A novel ERAP1 allotype was identified from S103 and it was assigned number *027 containing the amino acid changes R127P/I276M/K528R and this was found in combination with allotype *018 containing the single amino acid change Q730E. Interestingly, this allotype was also identified from a study by Stratikos et al (allotype 16 in that study) [123]. As mentioned above, both ERAP1 allotypes of S101 were verified with Sanger sequencing as this was one of the amplicons chosen at random to be verified through a second sequencing method, thus confirming the accuracy of ERAP1 allotyping with MinION.

Figure 4.38. Mux scan showing the number of available pores for sequencing and total demultiplexed read counts generated for ERAP1 amplicons.

Mux scan showing the number of active pores available for sequencing (green) before sequencing took place (A). Unavailable pores are shown in blue. This experiment verified that sequencing of ERAP1 is possible with less than 800 pores. Data generated with MinKNOW. Epi2me report showing total number of reads generated for every barcoded sample in that sequencing run (B, C).

S104, S105, S106 and S107 also with visible ERAP1 amplification on the agarose gel were sequenced on the same flow cell as the run above with 439 pores available for sequencing. This showed that within 48 hours between the first to the second run there was a pore reduction of approximately 22% (Figure 4.39). As there was a smaller pool of pores available, sequencing was set to run for a longer period of time. The ERAP1 allotype combinations from all four patients were identified (Table 4.4). A novel ERAP1 allotype was identified from S104 and was assigned the number *028 as it was identified after the last novel allotype from S103. It contained the amino acid changes R127P/Q730E.

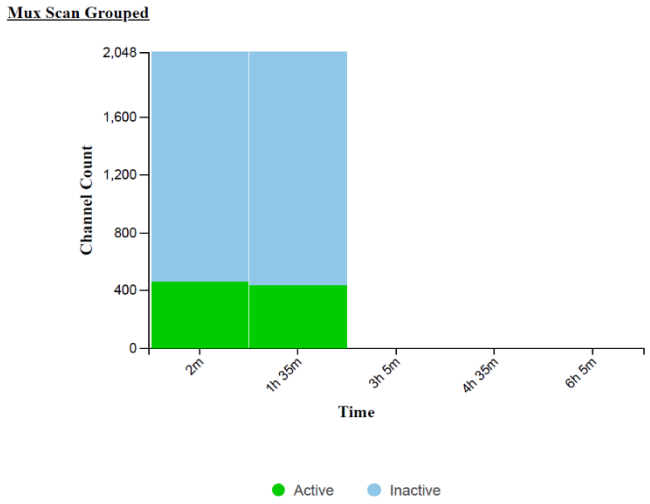


Figure 4.39. Mux scan showing the number of active pores available for sequencing

Mux scan showing the number of active pores available for sequencing (green) before sequencing took place and then every one hour and a half. Unavailable pores are shown in blue. This experiment verified that sequencing of ERAP1 is possible with less than 800 pores. Data generated with MinKNOW.

S78, S79, S80, S81 (run 1), S82, S83, S84, S85 (run 2) and S86, S87, S88, S89 (run 3) did not show presence of ERAP1 DNA on the agarose gel and were sequenced in three separate runs. The ERAP1 allotypes *001 + *018 were identified from S78 and this was the only sample from which allotypes were successfully identified (Table 4.4). Amplification of ERAP1 by PCR was repeated for S79, S82, S83, S84, S85, S86, S87, S88 and S89 and again the agarose gel did not reveal presence of ERAP1 DNA. When these samples were sequenced in the same run along with three more amplicons (total of twelve amplicons in the same run), a novel allotype was identified from S82 from a total of 462 reads and it was assigned the number *026 (R127P/M349V/D575N/R725Q/Q730E). Interestingly, another novel allotype was identified from S82 that was also identified previously from S3 and S30 and had already been assigned the number *024. Allotype *026 is similar to *001 but it lacks the K528R mutation. Functionally, the allotype was expected to have a hypertrimming phenotype due to the simultaneous presence of amino acid changes R725Q/Q730E (Table 4.4).

Even though S119 did not produce a visible band when run on the agarose gel, it was sequenced along with S115, S116 and S118, the ERAP1 amplification of which could be seen on the gel. It seemed like a reasonable experiment as when a single amplicon was sequenced in the same run as three amplicons with visible ERAP1 DNA on the agarose gel, the ERAP1 allotype combinations of all four samples were identified. This was also the case for S119. Only 36 reads passed filtering for this amplicon but ERAP1 allotypes were identified, nonetheless (Table 4.4). Amplicons prepared using cDNA from patients 111, 112, 114 and 117 were part of the same run as amplicons S68, S72 and S75 and all ERAP1 allotype combinations were successfully identified (Table 4.4).

4.2.1 SNPs and allotypes associated with decreased ERAP1 trimming of peptide antigenic precursors seen with higher frequency in the cervical cancer patient cohort

In total, 14 ERAP1 allotypes were identified from the patient cohort which formed 28 distinct combinations, of which 6 were homozygotes and 22 heterozygous (Table 4.7). The most common ERAP1 allotype and ERAP1 allotype combination were *021 (28 samples) and *001 + *021 (7 samples), respectively (Table 4.7). SNP frequency from the patient cohort is shown on Table 4.9 which also shows the relative frequency distributions of SNPs in the CEU cohort as reported in a review by Ombrello et al [120]. Interestingly, 6 novel ERAP1 allotypes were identified from the patient cohort (Table 4.6) and the combinations they were found in are shown on Table 4.4. It is noteworthy that the combination of *023 + *024 was identified from two separate patients, S3 and S30, the ERAP1 amplicons of which were not sequenced in the same run, therefore they can be considered true novel allotypes. Additionally, ERAP1 allotype *024 was identified from a third patient, S82, in combination with another novel allotype *026 (Table 4.7).

Table 4.6. Novel ERAP1 allotypes identified from the cervical cancer patient cohort

Allotype	Amino acid change at indicated position									Frequency in cohort n=162 (%)
	56	127	276	346	349	528	575	725	730	
	E/K	R/P	I/M	G/D	M/V	K/R	D/N	R/Q	Q/E	
*023	E	P	M	G	M	K	D	R	E	2 (1.2)
*024	E	P	I	G	M	R	D	R	Q	3 (1.9)
*025	E	P	M	D	M	R	D	R	E	1 (0.62)
*026	E	P	I	G	V	K	N	Q	E	1 (0.62)
*027	E	P	M	G	M	R	D	R	Q	1 (0.62)
*028	E	P	I	G	M	K	D	R	E	1 (0.62)

Table 4.7. ERAP1 allotypes and combinations identified with long read sequencing and their frequency in the cervical cancer patient cohort

ERAP1 allotypes	Cohort frequency n=81 (%)
*013 + *021	3 (3.7)
*023 + *024	2 (2.5)
*013 + *025	1 (1.2)
*024 + *026	1 (1.2)
*018 + *027	1 (1.2)
*002-WT + *028	1 (1.2)
*013	5 (6.2)
*001	4 (4.9)
*018	3 (2.5)
*002-WT	5 (6.2)
*021	5 (6.2)
*015	2 (2.5)

*013 + *002-WT	4 (4.9)
*018 + *021	5 (6.2)
*013 + *018	6 (7.4)
*021 + *002-WT	5 (6.2)
*001 + *015	3 (3.7)
*002-WT + *001	3 (3.7)
*001 + *021	7 (8.6)
*013 + *015	2 (2.5)
*019 + *015	2 (2.5)
*002-WT + *015	2 (2.5)
*001 + *013	1 (1.2)
*015 + *021	3 (3.7)
*014 + *001	2 (2.5)
*001 + *018	1 (1.2)
*002-WT + *018	1 (1.2)
*015 + *018	1 (1.2)
TOTAL	n=81 (100%)

Upon obtaining the ERAP1 allotypes from the cervical cancer patient cohort using long read sequencing, it was of interest to carry out a comparison of allotype frequencies between the European cohort (CEU) as these were reported in a review by Ombrello et al and the patient cohort [120]. Allotype frequencies are presented in percentages on Table 4.8. The most frequent allotype in the CEU cohort was *001, while the most frequent allotype in the cervical cancer patient cohort of the present study was *021 (Table 4.8). Allotype*022 was not observed in the cervical cohort, but it was identified in 1.2% of the CEU cohort. Chi-square revealed that the differences in frequency of allotypes between the cervical cancer patient cohort and the CEU cohort were statistically significant (p=0.015). Due to the fact that allotype combinations were not taken under consideration in the study by

Ombrello et al, it was not possible to accurately compare individual allotype frequencies in the two cohorts.

Table 4.8. Comparison of ERAP1 allotype frequencies between this cervical cancer patient cohort and the European population (CEU) [120].

The number of individuals that had each allotype was not presented in the Ombrello et al study. Frequency of allotypes was only expressed in percentage.

Allotype	Haplotype	CEU frequency n=80 (%)	Cohort frequency n=129 (%)
*013	Hap1	(12.0)	22 (17.1)
*002-WT	Hap2	(13.7)	21 (16.3)
*018	Hap3	(5.6)	18 (14)
*008	Hap4	(1.3)	0
*019	Hap5	(7.5)	2 (0.02)
*015	Hap6	(7.4)	15 (11.6)
*014	Hap7	(2.5)	2 (0.02)
*021	Hap8	(21.9)	28 (21.7)
*022	Hap9	(1.2)	0
*001	Hap10	(26.2)	21 (16.3)

A study by Mehta et al is the only study investigating ERAP1 SNP frequency in a cervical cancer patient cohort compared to healthy controls [181]. Controls were randomly selected, unrelated Dutch women, with ethnic pool similar to the cervical cancer patient cohort used in the same study [181]. Despite the informative nature of that study, especially given the relatively larger cohort sizes (124 healthy controls and 127 cervical cancer patients) and the geographic identity of the two patient cohorts, it only took under consideration SNPs encoding amino acid changes at positions 56, 127, 276, 528, 575 and 730, disregarding linkage disequilibrium which is actually considered in the present study, reporting data in the form of ERAP1 allotypes. Additionally, in the study by Mehta et al, the frequencies of only 6 SNPs were compared between the two cohorts (healthy vs disease), while in this study 9 distinct SNPs were found in 14 ERAP1 allotypes and 28 separate ERAP1 allotype combinations

identified from 81 cervical cancer patients. The function of the ERAP1 SNPs was not investigated by researchers who reported their frequencies.

It was hence decided that it would be of interest to compare frequencies of the six SNPs on an individual basis reported by Mehta et al for both the healthy controls and the cervical cancer patients with the frequencies for this cervical cancer cohort. Frequencies are reported for both the major and minor alleles for SNPs that lead to the amino acid changes E56K, R127P, I276M, K528R, D575N and Q730E expressed in percentages (Table 4.9) [181]. Statistical analysis revealed that there is a significant difference in the frequency of K56 between the Mehta control study and the cervical cancer patient cohort ($p<0.05$) [182]. This was also the case for R528, N575 and E730, with p-values of $p<0.01$, $p<0.005$ and $p<0.01$, respectively. These data showed that there is an association between these four amino acid changes (K56, R528, N575 and E730) and cervical cancer. Comparison of the frequency of amino acid changes at indicated positions between the cervical cancer patient cohort of this study and the Mehta patient cohort revealed that there was a statistically significant difference in the frequency of R528 and E730 in the two cohorts ($p<0.01$ for both of them), with both of these amino acid changes identified with higher frequency in the cervical cancer patient cohort of this study (R528: 51.9 vs 36.7% and E730: 64.8 vs. 33.5%).

Table 4.9. Frequency of SNPs in this cervical cancer patient cohort vs healthy controls/cervical cancer cases from a study by Mehta et al [181]

Amino acid changes at indicated positions in the cervical cancer patient cohort of this study are indicated in bold.

	Mehta control cohort, n=248 **n (%)**		*Study patient cohort, n=162* **n (%)**		*Mehta patient cohort, n=254* **n (%)**	
SNP	Major allele, n	Minor allele, n	Major allele, n	Minor allele, n	Major allele, n	Minor allele, n
56	E, 235 (94.8)	K, 13 (5.2)	E, 160 (98.8)	**K**, 2 (1.2)	E, 242 (95.3)	K, 12 (4.7)
127	P, 194 (78.2)	R, 54 (21.8)	**P**, 113 (69.8)	R, 49 (30.2)	P, 174 (68.5)	R, 80 (31.5)
276	I, 188 (75.8)	M, 60 (24.2)	I, 125 (77.2)	**M**, 37 (22.8)	I, 199 (78.3)	M, 55 (21.7)
528	K, 175 (70.6)	R, 73 (29.4)	K, 78 (48.1)	**R,** 84 (51.9)	K, 157 (63.3)	R, 91 (36.7)
575	D, 184 (74.8)	N, 62 (25.2)	D, 136 (84)	**N**, 26 (16)	D, 197 (79.4)	N, 51 (20.6)
730	Q, 191 (77.0)	E, 57 (23.0)	Q, 57 (35.2)	**E**, 105 (64.8)	Q, 169 (66.5)	E, 85 (33.5)

The only other study investigating ERAP1 allotypes (and hence individual SNPs) in an HPV-induced cancer, was the study undertaken by Reeves et al in HPV+ OPSCC patients [121]. It was therefore of interest to compare both allotype and SNP frequencies between the two HPV-driven cancer patient cohorts. The HPV+ OPSCC cohort size was considerably smaller than that of the cervical cancer patient cohort (n=25 and n=81, respectively) and although both of the cohorts include patients of European decent, patients were of different country origin (UK and Netherlands, respectively). Frequencies are reported in percentages on Table 4.10.

The frequency distributions of D346, V349, N575, Q725 and E730 were similar in the two patient cohorts (Figure 4.10). M276 was not identified in OPSCC patients but it was found in 39.5% CC patients. P127 was identified in more than double the number of CC patients than OPSCC patients (69.8 vs. 32%). The Chi-square test revealed there is a significant association between the amino acid change P at position 127 of ERAP1 and the HPV-positive cancer patients suffer from (Chi-square test, $p<0.01$).

Table 4.10. Comparison of ERAP1 SNP frequencies between an HPV+ OPSCC cohort and the cc cohort in this study

SNP distribution taking into consideration both ERAP1 allotypes from the two cohorts. Amino acid changes are shown in bold. N/A indicates that this SNP was not identified in that cohort. CC=cervical carcinoma, OPSCC= oropharyngeal squamous cell carcinoma of the head and neck.

ERAP1 SNP	**OPSCC SNP frequency n=50 (%)**	**CC cohort SNP frequency n=162 (%)**
T12**I**	N/A	0

E56**K**	1 (2)	2 (1.2)
R127**P**	16 (32)	113 (69.8)
I276**M**	N/A	37 (22.8)
G346**D**	2 (4)	3 (1.9)
M349**V**	11 (22)	26 (16)
K528**R**	31 (62)	84 (51.9)
D575**N**	11 (22)	26 (16)
R725**Q**	13 (26)	26 (16)
Q730**E**	35 (70)	105 (64.8)

Next, the frequencies of individual ERAP1 allotypes were compared between the two HPV-driven cancer cohorts (Table 4.12). As expected from the SNP frequency data, the frequency of allotype *018 that only contains E730 was similar in the two cohorts. Four allotypes that were identified in the OPSCC cohort were not identified in the CC cohort (*011, *016, *017, *007, Table 4.12), while the six novel ERAP1 allotypes that were identified in the CC cohort, were not identified in the OPSCC cohort (Table 4.6). A total of 24% of OPSCC patients had the *002 allotype, while the relevant percentage of CC patients that had the *002 allotype was 16 (chi-square, $p<0.05$). Allotype *011 was not identified in the CC patient cohort, however approximately a third of OPSCC patients had this allotype (chi-square, $p<0.01$). Allotype *021 was only identified in the CC patient cohort (chi square, $p<0.01$).

Table 4.11. Comparison of ERAP1 allotype frequencies between an HPV+ OPSCC cohort and the present cervical cancer patient cohort (*novel not included*)

Allotype	**Haplotype**	**HPV+ OPSCC cohort frequency n=50 (%)**	**HPV+ CC cohort frequency n=162 (%)**
*013	Hap1	1 (2)	27 (16.7)
*002-WT[a]	Hap2	12 (24)	26 (16)
*018	Hap3	3 (6)	21 (13)
*019	Hap5	2 (4)	2 (1.2)
*015	Hap6	3 (6)	17 (10.5)
*014	Hap7	1 (2)	2 (1.2)
*021[a]	Hap8	0	33 (20.4)
*001	Hap10	10 (20)	25 (15.4)
*011[a]	N/A	14 (28)	0
*016	N/A	2 (4)	0
*007	N/A	1 (2)	0
*017	N/A	1 (2)	0

[a]indicates statistical significance, chi-square.

The comparison of allotype combination frequencies between the OPSCC and the cervical cancer patient cohort revealed that five distinct ERAP1 allotype combinations are shared between the two cohorts; *001, *002-WT and *015 as homozygotic combinations, *002-WT + *001 and *001 + *018 (Table 4.12). It is noteworthy that the chi-square did not reveal a statistically significant difference in the allotype combinations between the OPSCC and CC patient cohorts [121].

Table 4.12. Comparison of ERAP1 allotype combination frequencies between an HPV+ OPSCC patient cohort and the present HPV+ cervical cancer patient cohort

N/A= not applicable, these allotype pairs were not identified from the patient cohort.

ERAP1 allotype combinations	CC patient frequency n=81 (%)	OPSCC patient frequency n=25 (%)
*013 + *021	3 (3.7)	N/A
*023 + *024	2 (2.5)	N/A
*013 + *025	1 (1.2)	N/A
*024 + *026	1 (1.2)	N/A
*018 + *027	1 (1.2)	N/A
*002-WT + *028	1 (1.2)	N/A
*013	5 (6.2)	N/A
*001	4 (4.9)	1 (4)
*018	3 (2.5)	N/A
*002-WT	5 (6.2)	2 (8)
*021	5 (6.2)	N/A
*015	2 (2.5)	1 (4)
*013 + *002-WT	4 (4.9)	N/A
*018 + *021	5 (6.2)	N/A
*013 + *018	6 (7.4)	N/A
*021 + *002-WT	5 (6.2)	N/A
*001 + *015	3 (3.7)	N/A
*002-WT + *001	3 (3.7)	4 (16)
*001 + *021	7 (8.6)	N/A
*013 + *015	2 (2.5)	N/A
*019 + *015	2 (2.5)	N/A

*002-WT + *015	2 (2.5)	N/A
*001 + *013	1 (1.2)	N/A
*015 + *021	3 (3.7)	N/A
*014 + *001	2 (2.5)	N/A
*001 + *018	1 (1.2)	1 (4)
*002-WT + *018	1 (1.2)	N/A
*015 + *018	1 (1.2)	N/A

4.2.2 Summary and conclusions drawn from the long read sequencing of ERAP1 from the cervical cancer patient cohort

Long read sequencing was successfully used to identify the ERAP1 allotype combinations from two cell lines, 293T and HeLa. These trial sequencing runs enabled the establishment of a methodological pipeline for sequencing ERAP1 from a total of 81 cervical cancer patients (originally 103 patient samples available), as well as use bioinformatics tool for data analysis. A total of 28 distinct ERAP1 allotypes were identified from the patient cohort, 22 heterozygotic and 6 homozygotic combinations. Due to the lack of genomic DNA, confirmation of true homozygotes was not possible. Despite the lack of genomic DNA that would enable confirmation of true homozygotes, repeated sequencing of ERAP1 amplicons such as S51 revealed that the combination was the same as the one identified in the first sequencing run.

Statistical analysis revealed that there is a significant difference in the frequency of allotypes between the CEU cohort and the cervical cancer patient cohort of this study; however due to the fact that allotype combinations were not taken under consideration in the Ombrello study, it was not possible to identify significant differences in the frequency of individual ERAP1 allotypes between the two cohorts [120]. There was a significant difference in the frequency of K56, R528, N575 and E730 between the Mehta control study and the cervical cancer patient cohort in thie present study. In addition, there was a significant difference in the frequency of R528 and E730 between the Mehta cervical cancer patient cohort and the cervical cancer patient cohort in the present study. P127 was

seen with high frequency in both the cohort of this study and in the cervical cancer patient cohort in the Mehta study, pointing towards the likely role of this amino acid change with increased cervical cancer risk and poor prognosis. E730 was seen with higher frequency in the present cohort compared to the Mehta patient cohort, but a total of 8 out of the 14 allotypes identified from the cervical cancer patient cohort of the current study contained both P127 and E730 indicating an association with cervical cancer as Mehta found that the simultaneous presence of P127 and E730 along with SNPs in other APM components increased cervical cancer risk 3-fold [181]. Chi-square test showed that there was a significant difference in the frequency of P127 in the OPSCC and the cervical cancer patient cohort indicating an association of this amino acid change with cervical cancer. The frequency of allotypes *011 and *002 was significantly different between the two patient cohorts (OPSCC vs CC), indicating that these allotypes are associated with OPSCC. *021 was not identified in OPSCC cohort, but it was identified at high frequency in the CC patient cohort ($p<0.01$).

5 Results: part 3

Functional assessment of the identified ERAP1 allotype combinations

5.1 The peptide trimming function of ERAP1 allotypes and combinations identified in cervical cancer

ERAP1 is part of the MHC I antigen processing and presentation pathway. It has a role in the processing of N-terminally extended peptides entering the ER through TAP into the appropriate length for stable loading on to MHC I molecules for subsequent presentation to CD8+ T cells expressing the relevant TCR [34-37]. TAP preferentially transports N-terminally extended peptides 11-14 amino acids long [213]. ERAP1 is the enzyme responsible for the N-terminal processing of these peptide precursors to a length of 8 to 10 amino acids for binding on to MHC I. After peptide loading on to the MHC I, the resulting pMHC I is translocated to the cell surface after passing through the Golgi apparatus. Upon recognition of the MHC I:peptide complex, CD8+ T cells become activated and exert anti-tumour immune responses [34-37]. The role of ERAP1 in antigen processing was first investigated about 20 years ago as upon upregulation of IFNγ, there was increase in the trimming of N-terminally extended peptides by ERAP1 [35].

Regarding trimming of N-terminally extended peptides, ERAP1 has a preference for hydrophobic and aromatic amino acids, such as leucine and valine and it is unable to process X-pro-X bonds [85, 90, 96, 113]. Peptides fall into three categories based on whether they are affected by the ERAP1 trimming activity; ERAP1-independent peptides are unaffected by ERAP1 trimming, ERAP1-dependent peptides are those that require N-terminal trimming by ERAP1 to generate the final epitope and ERAP1-sensitive peptides are normally destroyed by ERAP1 such as GSW11, a CT26 murine colorectal carcinoma model H2-D^d specific antigen [36, 64, 81].

SNPs in ERAP1 affect trimming function and therefore the peptide repertoire at the cell surface and may be crucial in anti-tumour immune responses [121]. In this cohort, the amino acid change T12I was not seen in any of the patient ERAP1 amplicons that were successfully sequenced. P127 is found in domain I of ERAP1, away from the enzyme's active site, and has been shown not to have a direct effect

on the trimming function of ERAP1, yet it is possible that is affects the transition of ERAP1 from open to closed conformation upon peptide binding [127]. *In vitro* investigation of R528 and E730 revealed reduced trimming activity but this was not the case for the amino acid change D575N [88, 102, 128]. Q725 has been shown to be associated with a hyperactive ERAP1 enzyme, meaning that it affects ERAP1 function in a way that the enzyme is overtrimming antigenic peptide precursors [89]. As SNPs are found in ERAP1 in linkage disequilibrium, appointing the functional effect in ERAP1 to a single SNP is difficult. Moreover, as ERAP1 chromosomal copies are co-dominantly expressed, it is important to investigate the combined effect that SNPs found in both ERAP1 allotypes have on the overall function of the enzyme. One of the first allotypes the trimming of which was investigated, *001 (Hap10 [120]; P127/V349/R528/N575/Q725/E730), has been shown to have reduced trimming function compared to wild type ERAP1, *002-WT [89]. In a study by Mehta et al in 2015, minor alleles at ERAP1-127 and ERAP1-730 were shown to be associated with increased cancer risk, and when occurring simultaneously with two additional mutations in TAP2 and LMP7, cervical cancer risk is increased three-fold (TAP2-651, LMP7-145) [177]. Homozygosity for the major allele at position 56 and the minor allele at position 127 of ERAP1, were also associated with decreased overall survival. However, the effects of these SNPs on the function of the enzyme were not investigated and linkage disequilibrium was not taken under consideration.

In 2019, our lab published a study on the investigation of the functional activity of identified ERAP1 allotype combinations from a cohort of HPV-positive patients suffering from oropharyngeal squamous cell carcinoma of the head and neck (OPSCC) [121]. The generation of a well-characterised OVA-derived epitope as well as that of HPV-16 E7 derived epitopes from N-terminally extended precursor peptides by the allotypes was investigated. Results showed that ERAP1 function correlates with the CD8/TIL status of the patients and hence, cancer prognosis. The aim of this chapter was to determine the trimming function of those ERAP1 allotypes and combinations identified in the cervical cancer cohort (Chapter 3, 4). This investigation was carried out through the trimming of the same N-

terminally extended precursors as the previous studies by our lab [121, 124]. Since fresh tumour material was not available for this cohort, the CD8+/TIL status from a subset of patients was determined using CD8+/TIL numbers per tumour mm^2 provided by the University of Groningen (Chapter 3).

5.1.1 Investigating ERAP1 function in the trimming of N-terminally extended peptides using a model system

To investigate the functional activity of those ERAP1 allotypes, and combinations, identified in the cervical cancer cohort (Chapter 4), a model system utilising the T cell hybridoma, B3Z, was used. In 1993, Shastri and Gonzalez revealed that the OVA-derived peptide SIINFEKL (SL8, $OVA_{257-264}$), presented on antigen presenting cells by H-$2K^b$, is required for the stimulation of the B3Z T cell hybridoma [214]. This model system is well characterised within the James Lab and has been previously used to study the function of ERAP1 in AS and head and neck cancer patients [121, 124]. The response raised by the B3Z T cells upon recognition of the complex consisting of H-$2K^b$ and a modified form of SL8, SIINFEHL (SHL8), is measured using the lacZ reporter gene product, β-galactosidase, which parallels the production of IL-2 by T cells upon recognition of the pMHC I. β-galactosidase cleaves CPRG, a colorimetric substrate, with change from yellow to varying shades of red being indicative of the lacZ activity in B3Z T cells. This can be used as an indirect measure of ERAP1 trimming activity to generate the final SHL8 peptide, and therefore H2-K^b:SHL8 complexes, from N-terminally extended precursors.

ERAP1 knock-out 293T human embryonic kidney cells, 293TE1KO, were transfected with the ERAP1 allotypes previously identified from the cervical cancer patient cohort using long read sequencing, the minigene construct AIVMK-SHL8 and H2-K^b to investigate the trimming ability of the ERAP1 allotypes (Chapter 4.2). The ability of B3Z T cells to recognise the H2-K^b:SHL8 peptide complexes, that is the sensitivity of B3Z T cells to the presentation of the complexes at the cell surface, was measured by

adding 100pM of SHL8 peptide into cultures with K89 cells which are adherent mouse L-cells stably expressing H2-K^b at the cell surface, and co-culturing them with B3Z T cells, before measuring T cell activation by CPRG colour change to determine the sensitivity of B3Z. Figure 5.1 shows the B3Z T cell response at the recognition of the pMHC I presented at the cell surface of K89 cells following the addition of 100pM peptide in the APCs.

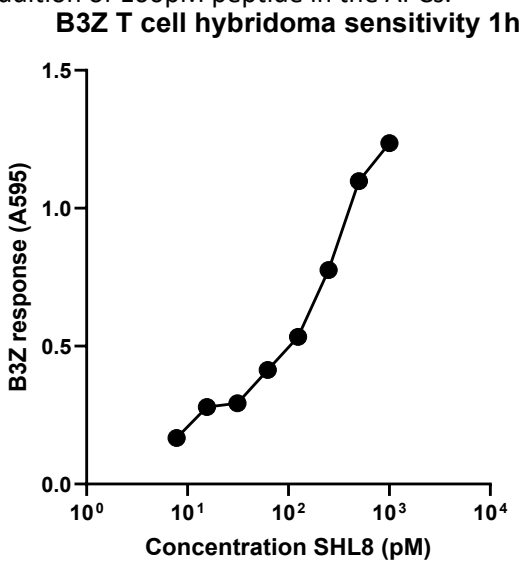


Figure 5.1. Assessing the B3Z T cell hybridoma sensitivity towards cell surface H2-K^b:SHL8.

K89 cells were pulsed with a starting concentration of 100pM SHL8 peptide titrated at 1:2 dilution. The B3Z T cell hybridoma cells were co-cultured with SHL8-pulsed K89 cells overnight before assessing T cell activation using CPRG colour change measured at 595nm.

Historically, the ability of ERAP1 allotypes to generate SHL8 from the five amino acid extended precursor, AIVMK-SHL8 (X5-SHL8), has been compared to the trimming ability of the wild type allotype, *002-WT, and has been shown to result in maximal response by B3Z T cells [89]. For the T cell activation assay used to investigate the ability of the different ERAP1 allotypes to trim antigenic peptide precursors, 293TE1KO cells previously generated in the James lab, were transfected with a total of 1µg of plasmid DNA including ERAP1, H2-K^b and the minigene construct X5-SHL8.

First, a control T cell activation assay was carried out which involved transfection of 293TE1KO cells with either *002-WT or non-functional ERAP1 (E320A), H2-K^b and X5-SHL8 plasmid DNA. The non-

functional ERAP1 contains the amino acid change from aspartic acid to alanine at position 320 (E320A) and leads to abrogation of ERAP1 trimming function as this amino acid is part of the GAMEN active site motif of ERAP1 which is vital for the activity of the enzyme and found in all the M1 metalloproteases [107]. ERAP1 containing E320A, from here and after referred to as non-functional ERAP1, was used as a negative control in all the assays. As expected, the control T cell activation assay revealed that *002-WT was able to successfully trim X5-SHL8 to SHL8 which bound to H2-K^b was recognised by the B3Z T cells, while non-functional ERAP1 was unable to generate SHL8 from X5-SHL8 (Figure 5.2).

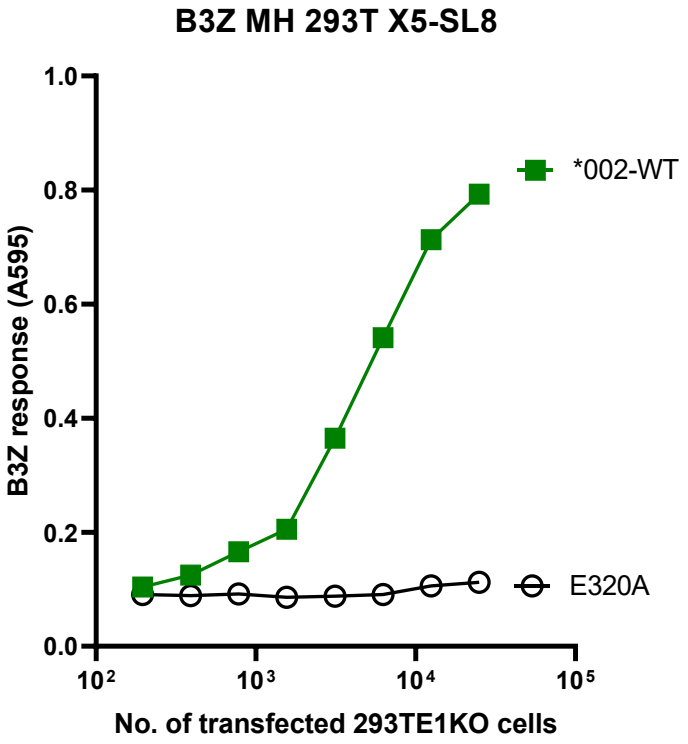


Figure 5.2: Control T cell activation assay.

293TE1KO cells were transfected with a total of 1µg of DNA consisting of the following minigene constructs: either 0.5µg of ERAP1 allotype *002-WT or 0.5µg of non-functional E320A, used as negative control in this assay, alongside 0.25µg of H2-K^b and 0.25µg of X5-SHL8. As shown on the graph, the B3Z response is optimal following transfection of 293TE1KO cells with either of the two allotypes and abrogated for the non-functional ERAP1 that serves as the negative control.

T cell activation assays were used to investigate the trimming of the ERAP1 allotype combinations identified from the patient cohort. The plasmids consisting of the ERAP1 allotypes identified were generated through site directed mutagenesis of existing ERAP1 allotypes in pcDNA3 vectors. The trimming of the individual ERAP1 allotypes was also investigated along with the combination that they were found to be part of from the patient cohort to enable understanding of how each allotype affects the overall trimming activity of the enzyme. Western blots were then used to verify that functional effects were indeed attributed to the ERAP1 sequence containing SNPs that produce amino acid changes and not related to variabilities in ERAP1 protein expression in cells.

5.1.2 Investigating the trimming of individual ERAP1 allotypes and the combinations found in the patient cohort

One of the first T cell activation assays and western blots carried out to determine the functional relevance of those ERAP1 allotypes and combinations identified with long read sequencing from the cervical cancer patient cohort are shown in Figure 5.3. . The highest absorbance at 595nm, indicating the highest B3Z response, measured for each of the allotypes/combination of allotypes was multiplied by the ERAP1/GAPDH ratio for that individual allotype or combination in an effort to normalise trimming activity and be able to attribute trimming ability to SNPs instead of ERAP1 expression. Both *001 and *021 ERAP1 allotypes have reduced trimming ability compared to *002-WT as shown by the reduction in B3Z response. Allotype *021 contains the amino acid changes R127P/I276M/K528R/Q730E and allotype *001 contains the amino acid changes R127P/M349V/K528R/D575N/R725Q/Q730. It was shown that *001 and *021 have <30% and <40% trimming activity compared to *002-WT, respectively. Normalisation using both T cell activation assays and the western blots was carried out in order to attribute the functional activity of the allotypes down to the amino acid changes they contain and not to altered ERAP1 expression or

technical issues associated with ERAP1 allotype transfection of 293TE1KO cells. In order to complete the normalisation of the trimming ability of the ERAP1 allotypes identified from the cervical cancer patient cohort, the highest B3Z response (A595) generated for each allotype was multiplied by the ERAP1 protein expression quantified using Image J divided by GAPDH expression.

Six different western blots were carried out showing ERAP1 protein expression in 293TE1KO cells previously transfected with *001 ERAP1 DNA. Allotype *001 was included in each trimming assay that was used to determine the combination trimming effect. This was the case for the majority of the single ERAP1 allotypes that were found in more than one ERAP1 allotype combination; these were *001, *013, *015 and *021. We could hence verify that the normalisation carried out for the ERAP1 allotype combinations was accurate through normalising the trimming of the most frequently identified single ERAP1 allotypes comprising the combinations in at least three separate experiments. This would allow understanding of how each allotype in the combination affects the generation of SHL8 from the antigenic precursor was enabled. The combination of *001 + *021 led to a similar trimming compared to the individual allotypes either *001 or *021, the trimming of which was significantly reduced compared to that of *002-WT ($p<0.0017$ for *001 + *021 vs *002-WT and $p<0.0001$; Figure 5.3).

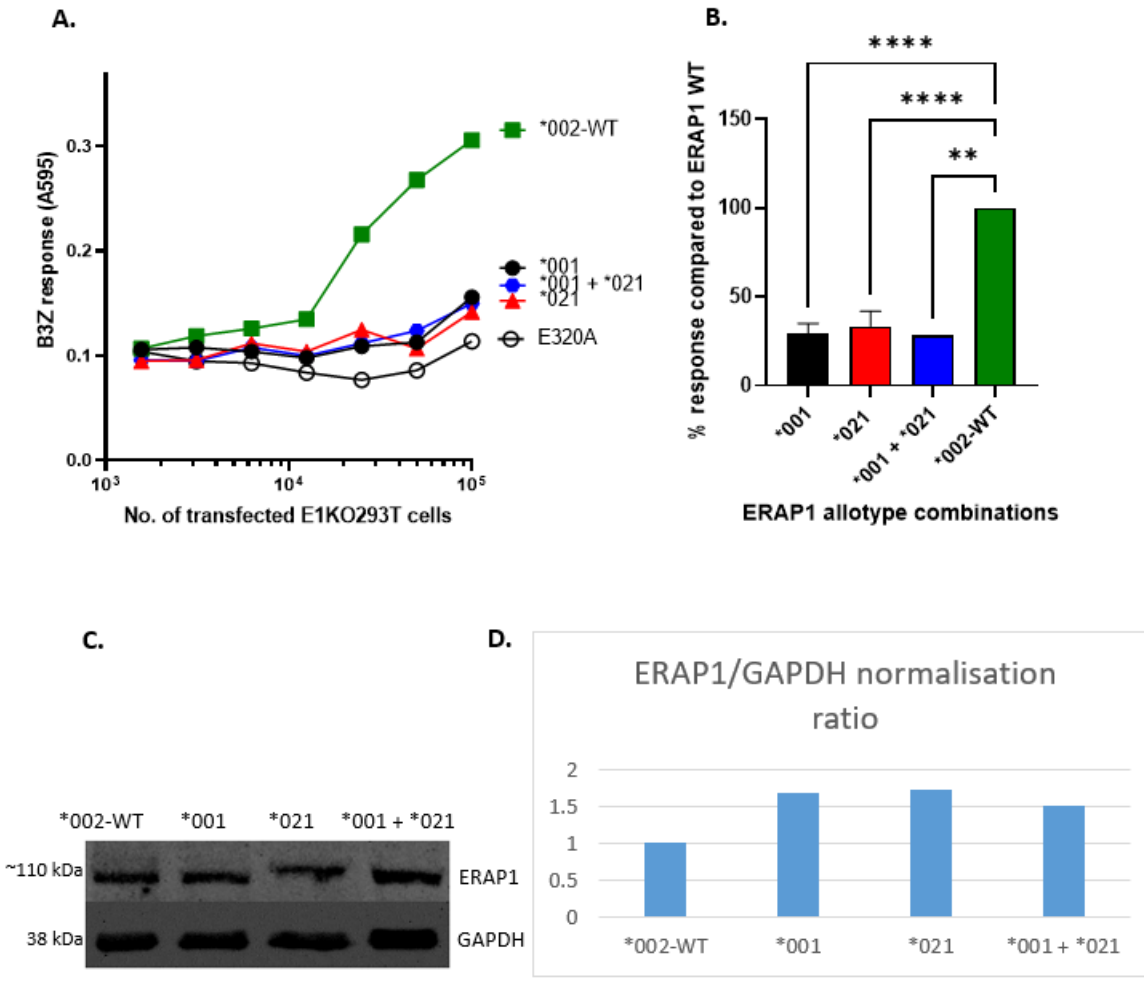


Figure 5.3: Functional assays investigating the trimming of ERAP1 allotype combinations compared to *002-WT

293TE1KO cells were transfected with X5-SHL8, H2-K^b and the relevant ERAP1 allotype or allotype combination. The data show both ERAP1 allotypes *001 and *021 have a hypotrimming phenotype compared to the wild type allotype *002-WT. T cell activation assay (A) and normalisation using T cell activation assay and western blot for each of the two allotypes and the combination with B3Z response expressed in percentage, compared to *002-WT presented on the histogram (B). Representation of the western blot carried out to investigate ERAP1 protein expression in 293TE1KO cells that have previously been transfected with 0.25µg X5-SHL8, 0.25µg H2-K^b and 0.5µg of either ERAP1 allotype *001 or *021 or a combination of both to a total of 1µg (C). Normalisation for allotypes *001 and *021 was repeated in seven and six separate experiments, respectively with error bars for mean ± SEM. Investigating the individual trimming of ERAP1 allotypes through three or more repeats enabled accurate predictions of the trimming of ERAP1 allotype combinations. ERAP1/GAPDH ratio (D) used along with the T cell activation assay (A) for the normalisation of ERAP1 allotype trimming ability. **p<0.0017 and ***p<0.0001.

The next allotypes that were investigated for overall trimming function were those in combination with *002-WT or *013, identified from a total of 39 patients. Four patients were positive for the *002-WT + *013 combination. ERAP1 allotype *013 contains a single amino acid change, R127P, and the T cell activation assays reveal that the trimming of this ERAP1 allotype is similar to the one of *002-WT allotype, highlighted by comparable B3Z responses (Figure 5.4). Consequently, it can be deduced from these results that the amino acid change at position 127 from arginine to proline does not seem to affect the function of the enzyme with the epitope SHL8 being successfully generated from the antigenic precursor X5-SHL8. The allotype combinations of *013 together with either *001 or *021 were assessed for function and protein expression (experiments repeated three times, Figure 5.4 B). ERAP1 pair *013 + *001 show similar activity compared with *002-WT, confirming previous findings that an allotype with normal trimming phenotype, in this case *013, is able to rescue the overall trimming of the enzyme when found in combination with a hypotrimming allotype, in this case *001 [89]. The normal trimming allotype *002-WT was able to increase the overall trimming ability of the enzyme as it was shown in the past by the James lab [89]. The allotype combination *002-WT + *013 was among the most efficient trimmers in the cohort with approximately 82% B3Z response.

As mentioned in Chapter 4, a total of six novel ERAP1 allotypes were identified from the patient cohort. In Figure 5.4, the trimming of two of those was investigated, *025 (R127P/I276M/G346D/K528R/Q730E) and *028 (R127P/Q730E). These allotypes were found in combination with either of the two efficient trimming ERAP1 allotypes, *002-WT or *013. The prediction of trimming based on the SNP combinations within these allotypes, was that *025 allotype would be a hypotrimmer, mainly due to the amino acid changes at positions 528 and 730 as the simultaneous presence of these two has been shown to be associated with reduced trimming [90]. It was of interest to understand whether G346D would further reduce trimming as the effect of this amino acid change has not been investigated, and it is located in close proximity to the active site

region. The trimming of *028 (R127P/Q730E) was predicted to be similar to that of allotype *018 that only contains the amino acid change Q730E as P127 was shown not to have an effect on function.

It can be seen on Figure 5.4 that indeed ERAP1 allotype *025 had a reduced ability to generate the final epitope from the antigenic precursor as the B3Z response is <20% compared to *002-WT ($p<0.05$). It is possible that the amino acid change G346D plays a role in this reduced phenotype of allotype *025. The amino acid position 346 is found in domain I of ERAP1 and it is close to another important position, 349, where the amino acid change M349V has been identified before. Interestingly, when the allotype *025 was found in a combination with the normal trimming allotype *013, the response was almost 100% suggesting that allotype *013 is able to rescue the overall trimming of the enzyme and the B3Z response raised is maximal.

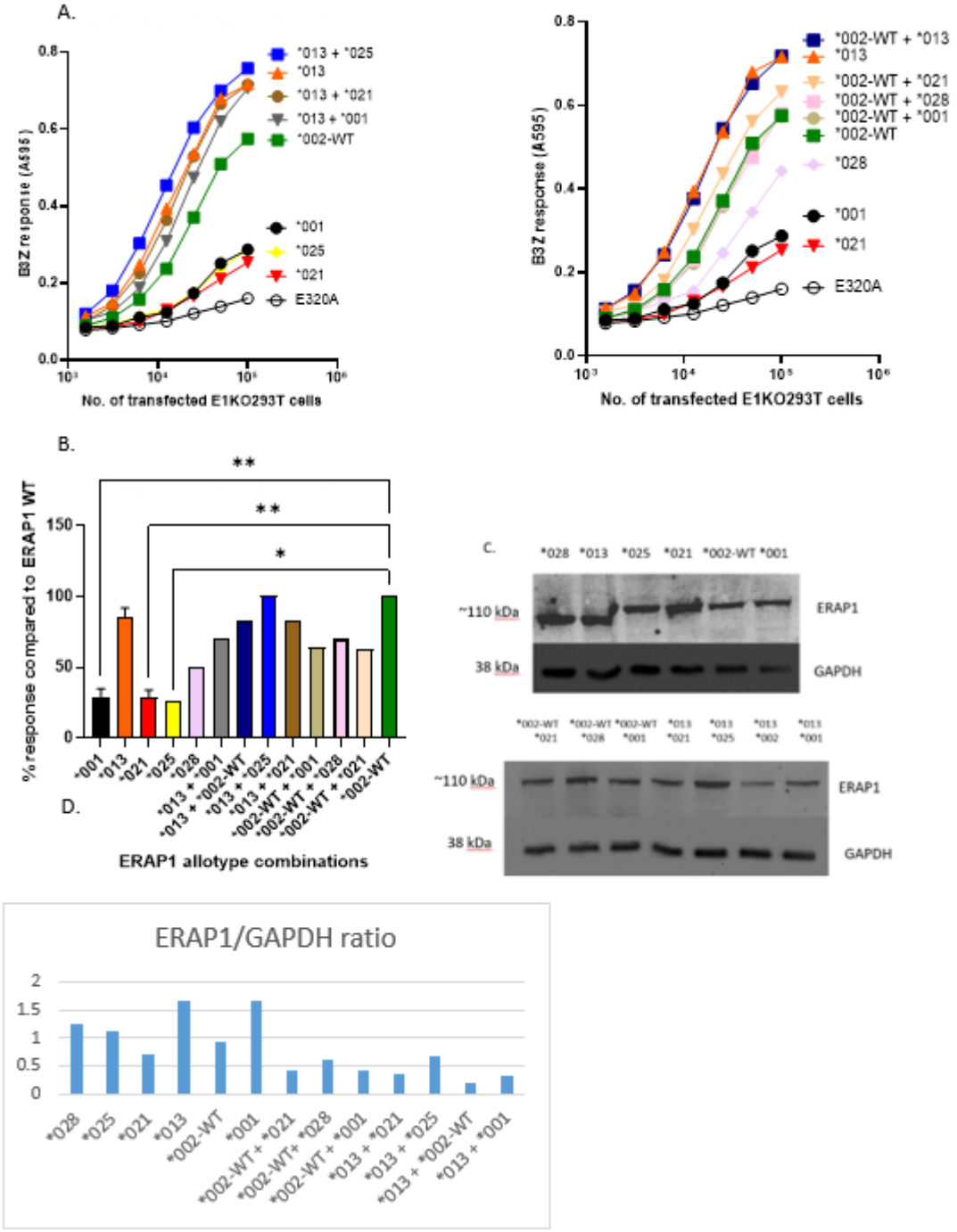


Figure 5.4: Functional assays investigating the trimming of ERAP1 allotype combinations compared to *002-WT.

293TE1KO cells were transfected with X5-SHL8, H2-K^{b} and the relevant ERAP1 allotype/combination. Data revealed that the novel allotype *025 had a hypotrimming phenotype, rescued by efficient trimmer *013 when in combination with the latter. The novel allotype *028 had the same phenotype as *018 (later section) that contains only Q730E. T cell activation assay (split into two for better visualisation) (A) and normalisation using T

cell activation assay and western blot for every allotype/combinations with B3Z response expressed in percentage, compared to *002-WT presented on the histogram (B). Representation of the western blot carried out to investigate ERAP1 protein expression in 293TE1KO cells that have previously been transfected with 0.25µg X5-SHL8, 0.25µg H2-K^b and 0.5µg of either *002-WT or *013 or an allotype combination one of which was *002-WT/*013 to a total of 1µg (C). Normalisation for allotypes *001, *013 and *021 was repeated in seven, four and six separate experiments with error bars for mean ± SEM. ERAP1/GAPDH ratio was used for normalisation of trimming activity (D). *p<0.05 and **p<0.005.

The ERAP1 allotype *015 was identified in 15 patients as well as in the 293T cell line, and contains the amino acid changes R127P/K528R/Q730E. As shown in Figure 5.3 and Figure 5.4, the simultaneous presence of K528R and Q730E leads to reduced ability to generate the final epitope from the extended peptide precursor, while the amino acid change R127P had minimal effect on the trimming function of the enzyme. This was expected to be the case as well for allotype *015.

The response generated for allotype *015 was less than 30% which was similar to the prediction given the presence of the amino acid changes K528R and Q730E and because the allotypes also containing these amino acid changes had similarly low trimming ability ($p<0.001$). The combination of *015 and either *002-WT or *013, led to a rescue of the overall trimming ability confirming findings from the James lab showing that a combination of an allotype with efficient trimming ability and a hypotrimmer can lead to similar trimming ability as the one of the efficient trimmer [89]. The combination of allotype *015 + *001 showed significantly reduced trimming compared to *002-WT ($p<0.03$).

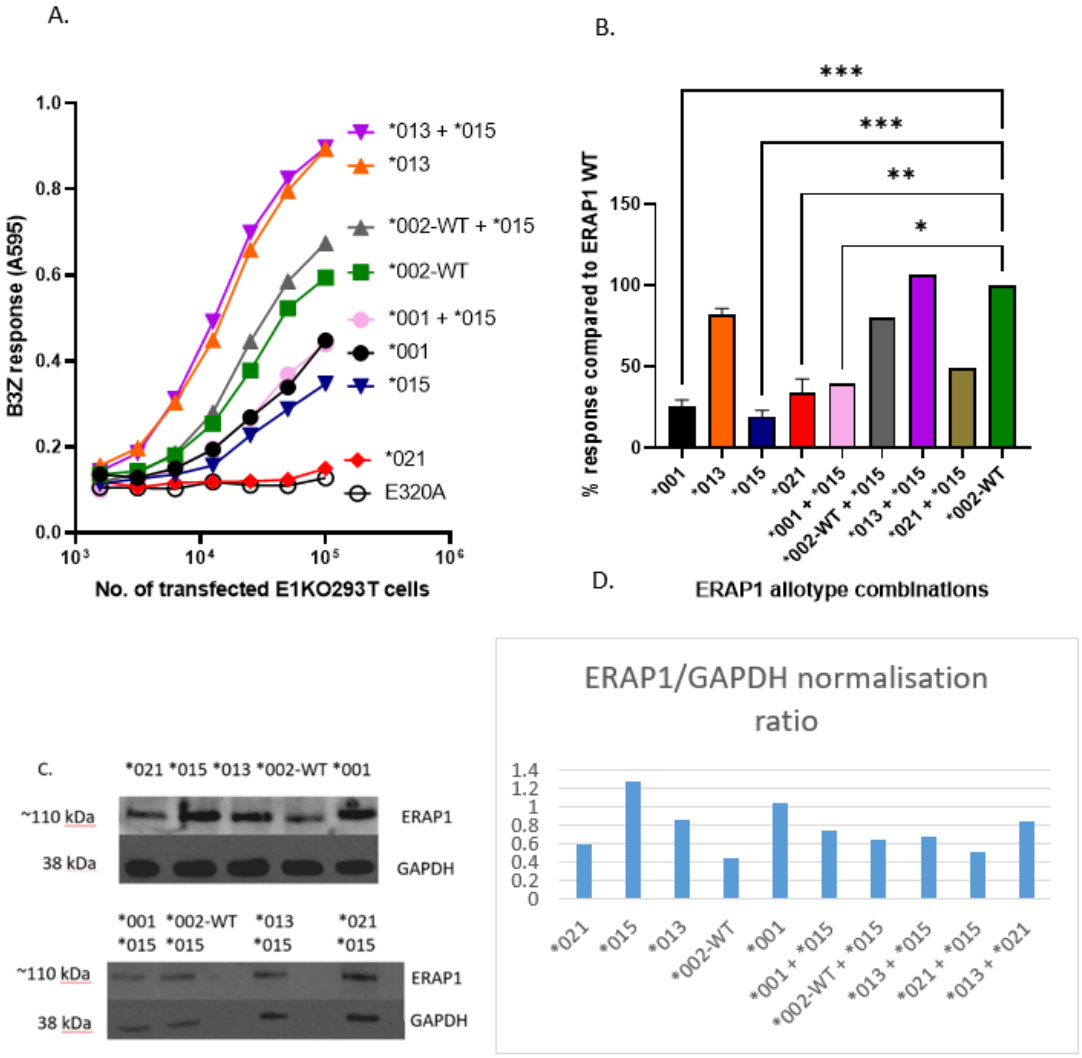


Figure 5.5: Functional assays investigating the trimming of ERAP1 allotype combinations compared to the wild type allotype *002-WT

293TE1KO cells were transfected with the minigene construct X5-SHL8, H2-K^{b} and the relevant ERAP1 allotype or allotype combination to be investigated. The data show that allotype *015 has a hypotrimming phenotype that leads to normal trimming when found in combination with an allotype with normal trimming phenotype. T cell activation assay (A) and normalisation using the T cell activation assay and western blot for every allotype/combinations with B3Z response expressed in percentage, compared to *002-WT presented on the histogram (B). Representation of the two western blots carried out to investigate ERAP1 protein expression in 293TE1KO cells that have previously been transfected with ERAP1 allotype combinations that consist of allotype *015 and either a normal or a hypotrimming allotype (C). ERAP1/GAPDH ratio was used for normalisation of trimming (D). Normalisation for allotypes *001, *013, *015 and *021 was repeated in seven, four, three and six separate experiments, respectively confirming the accuracy of the normalisation for the allotype combinations

one of which was either of these three allotypes with error bars for mean ± SEM. *p<0.05, **p<0.005 and ***p<0.001.

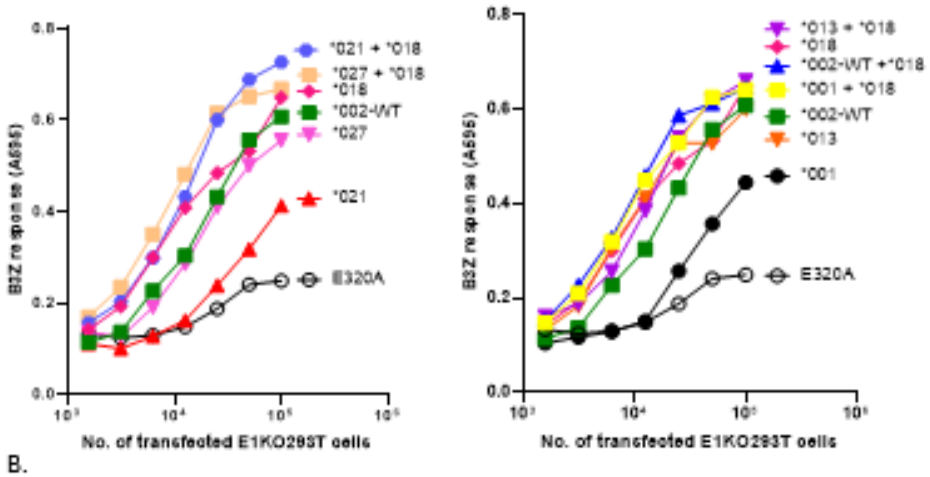
ERAP1 allotype *018 contains the amino acid change Q730E only that is in the regulatory domain of ERAP1, domain IV. In a study by Reeves et al, the ERAP1 allotype *018 was identified in a cohort of patients with HPV+ oropharyngeal squamous cell carcinoma of the head and neck (OPSCC) and it was shown that this allotype had half the ability of *002-WT to generate SHL8 from its precursor antigenic peptide [121]. In the present cervical cancer patient cohort, allotype *018 was found in the following combinations: *001 + *018, *002-WT + *018, *021 + *018, *015 + *018 and *027 + *018.

The ERAP1 allotype *027 (P127/M276/R528) was one of the six novel allotypes that were identified from the patient cohort using long read sequencing. Historically, it has been shown that the amino acid change K528R is found at the junction of domain I and II, and it is likely having an effect on the conformation of the enzyme from open to closed reduces the trimming ability of ERAP1 [88, 90, 102, 125]. However, assays have revealed that the trimming of the allotype containing the amino acid change K528R is substrate specific [88, 128]. It was therefore of interest to investigate whether the additional presence of the amino acid changes R127P and I276M would have a different effect on trimming compared to the presence of K528R only, no effect or would further contribute to the reduced ability of this allotype to generate SHL8 from X5-SHL8. I276M is found in domain I of ERAP1 and like R127P it is thought to have an indirect role in the conformation of ERAP1 from open to closed state upon substrate binding. Interestingly, P127 was associated with increased cervical cancer risk in a cohort of Chinese cervical cancer patients [182].

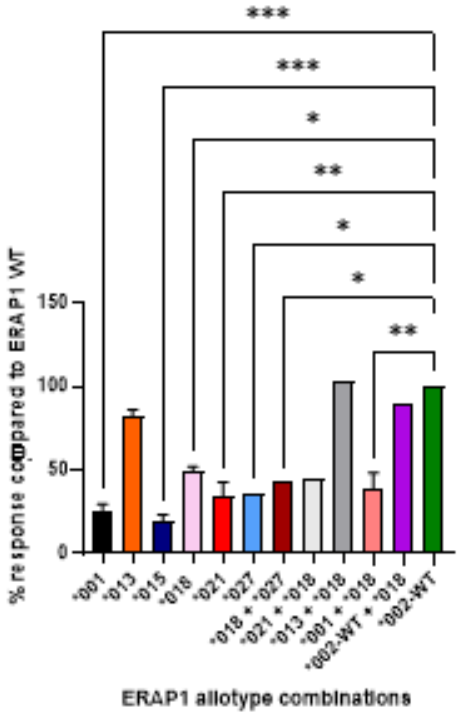
The trimming assays revealed that the ability of the ERAP1 allotype *018 to generate SHL8 from X5-SHL8 is about half that of that of *002-WT ($p<0.05$). When this allotype was found in combination with either *002-WT or *013, the overall trimming of the enzyme was rescued and returned to levels similar to that of *002-WT alone. When *018 was found in combination with *001, the B3Z response remained below 50% ($p<0.01$). When the trimming of the novel allotype *027 was investigated, it was

shown that the B3Z response generated was approximately 38% compared to that generated after the trimming of X5-SHL8 to SHL8 by *002-WT ($p<0.05$). The amino acid changes R127P and I276M did not reduce trimming further, revealing the major role of K528R on the function of the enzyme. When the novel allotype *027 was found in combination with allotype *018, trimming was still significantly reduced compared to *002-WT ($p<0.05$, Figure 5.6).

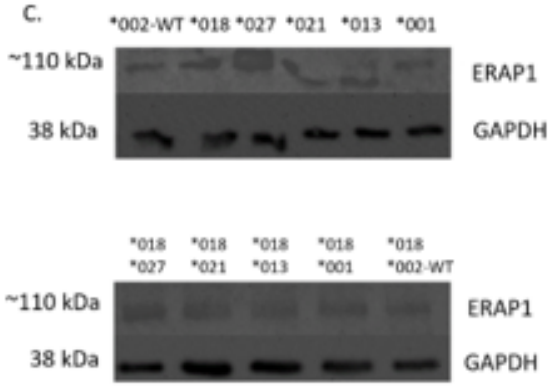
A.


B3Z response (A595)
No. of transfected E1KO293T cells
*021 + *018
*027 + *018
*018
*002-WT
*027
*021
E320A
*013 + *018
*018
*002-WT +*018
*001 + *018
*002-WT
*013
*001
E320A

B.


% response compared to ERAP1 WT
ERAP1 allotype combinations
*001
*013
*015
*018
*021
*027
*018 + *027
*021 + *018
*013 + *018
*001 + *018
*002-WT + *018
*002-WT

C.


*002-WT *018 *027 *021 *013 *001
~110 kDa
38 kDa
ERAP1
GAPDH
*018 *018 *018 *018 *018
*027 *021 *013 *001 *002-WT
~110 kDa
38 kDa
ERAP1
GAPDH

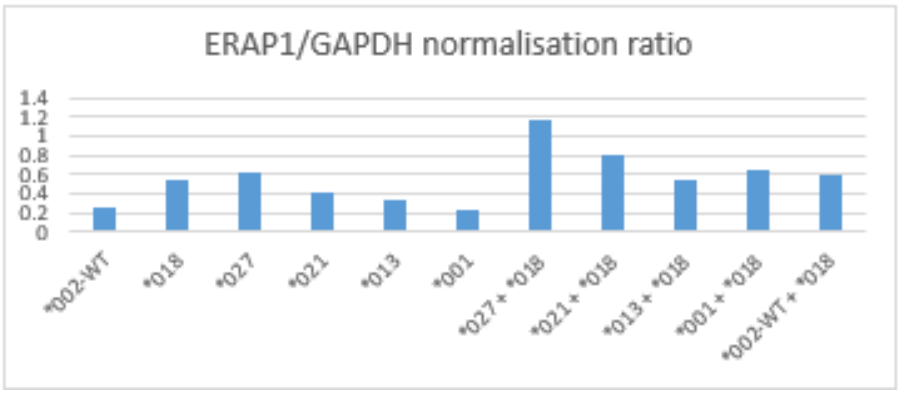

ERAP1/GAPDH normalisation ratio
*002-WT
*018
*027
*021
*013
*001
*027 + *018
*021 + *018
*013 + *018
*001 + *018
*002-WT + *018

Figure 5.6: Functional assays investigating the trimming of ERAP1 allotype combinations compared to the wild type allotype *002-WT

293TE1KO cells were transfected with the minigene construct X5-SHL8, H2-K^b and the relevant ERAP1 allotype or allotype combination to be investigated. The data show that allotype *018 has a hypotrimming phenotype, approximately half that of *002-WT. Normal trimming was observed when allotype *018 was found in combination with an allotype with normal trimming phenotype. T cell activation assay (A) and normalisation using the T cell activation assay and western blot for every allotype/combinations with B3Z response expressed in percentage, compared to *002-WT presented on the histogram (B). Representation of the two western blots carried out to investigate ERAP1 protein expression in 293TE1KO cells that have previously been transfected with ERAP1 allotype *018 or the combinations it was found to be part of (C). ERAP1/GAPDH ratio was used for normalisation (D). Normalisation for allotypes *001, *013, *015 and *021 was repeated in seven, four, three and six separate experiments, respectively (error bars for mean ± SEM). *p<0.05, **p<0.01 and ***p<0.001.

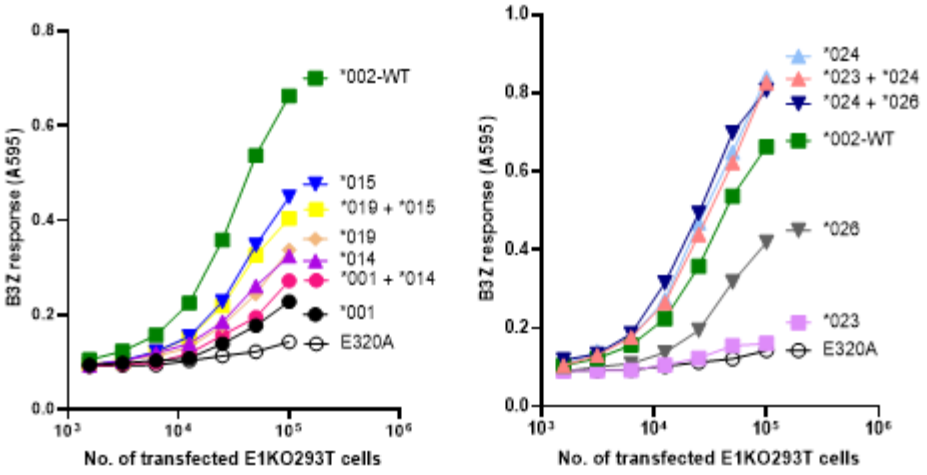
The trimming function of *019 + *015, *001 + *014, *023 + *024 and *024 + *026 was investigated next. Previous experiments included *001 and *015 as these allotypes were found to be part of multiple combinations from the patient cohort and had significantly reduced trimming compared to *002-WT (*p<0.001). The combination of *019 + *015 lead to ≤15% T cell response (P<0.001; Figure 5.7).

Regarding allotype *014 (E56I/R127P/K528R/Q730E), normalisation of response using both the T cell activation and the western blot revealed that trimming was significantly reduced compared to *002-WT and that was also the case for *001 + *014 (P<0.001). Although both ERAP1 allotypes contain amino acid changes that could explain a hypotrimming phenotype, the low percentage could perhaps be attributed to technical difficulties in the preparation of the western blots and hence having an effect on the normalisation of trimming for that allotype and combination. However, when the western blot was repeated for that allotype and combination, results were similar to this experiment. Interestingly, heterozygosity at position 56 was found to be associated with decreased overall survival in cervical cancer patients [190]. Homozygosity for either the major or minor allele of R127P (as in the

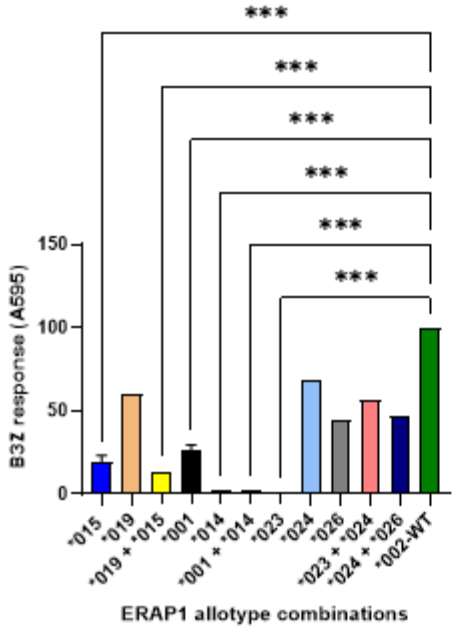
case of this combination as both *014 and *001 contain R127P), has been associated with decreased ERAP1 expression and homozygosity for the minor allele specifically was found to be associated with decreased disease-free and overall survival in cervical cancer [190]. This combination was seen in two patients (S47 and S103) and it was shown that both patients were disease-free at the last follow-up (5 and 7 years later, respectively Chapter 3). It is likely that early treatment received by these patients could have played a role in better disease outcome. Age could have played a role in the case of S103 as well given that this patient was one of the youngest in the cohort (28.4 age of diagnosis). S47 was in the CD8+/TIL^{mod} group (explained more later in this section), meaning that the number of CD8+/TILs could have also played a role in the prognosis. No CD8+/TIL data was available for S103 and neither of these patients were HLA-A*0201 positive (chapter 3) and hence not included in the functional analysis with an HPV-derived epitope (Figure 5.7).

As refers to the novel ERAP1 allotypes identified from the patient cohort, *023, *024 and *026, data revealed that these had a hypotrimming, efficient and hypotrimming phenotype respectively. Allotype *023 specifically (P127P/M276/E730), was shown to have a function close to that of non-functional ERAP1 ($p<0.001$; Figure 5.7). It is possible that the combination of I276M and Q730E reduces trimming function as R127P has not been shown to have an effect on trimming and Q730E reduced trimming by 50% (allotype *018). Allotype *024 (R127P/K528R) had unusually high trimming activity (approximately 70%, and this was shown by data normalisation as well as the T cell activation assay repeats). It is possible that the effect of K528R is not as dominant on reducing ERAP1 function when not in combination with Q730E or perhaps R127P and K528R are working in synchronisation to open and close ERAP1 leading to good overall trimming ability. As expected, *023 + *024 had a hypotrimming phenotype which shows that even though *024 has a normal trimming efficiency, it failed to restore trimming ($p<0.05$). Consequently, either there was a technical error in the investigation of *024 that would make it a hypotrimmer, as efficient allotypes were shown to restore trimming when in combination with a hypotrimmer, or homozygosity for K528R had an overall

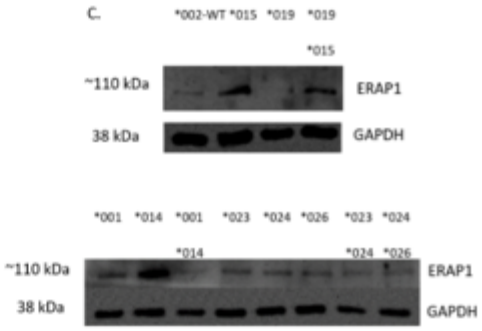
negative effect on trimming and possibly heterozygosity for Q730E as well (Figure 5.7). In addition, it may be likely that the amino acid changes in the allotype make it a hypertrimmer since trimming ability was not restored with the presence of the efficient trimmer *024. Allotype *026 contains all the amino acid changes that *001 has, except for K528R and had significantly reduced trimming compared to *002-WT ($p<0.05$). The combination of *024 + *026 also had a hypotrimming phenotype ($p<0.05$; Figure 5.7).


B3Z response (A595)
No. of transfected E1KO293T cells
*002-WT
*015
*019 + *015
*019
*014
*001 + *014
*001
E320A
*024
*023 + *024
*024 + *026
*026
*023

B.


B3Z response (A595)
ERAP1 allotype combinations
*015
*019
*019 + *015
*001
*014
*001 + *014
*023
*024
*026
*023 + *024
*024 + *026
*002-WT

C.


*002-WT *015 *019 *019 *015
~110 kDa
ERAP1
38 kDa
GAPDH
*001 *014 *001 *023 *024 *026 *023 *024 *014 *024 *026
~110 kDa
ERAP1
38 kDa
GAPDH

D.

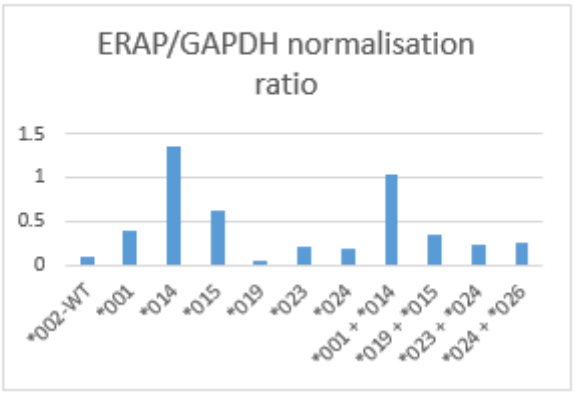

ERAP/GAPDH normalisation ratio
*002-WT
*001
*014
*015
*019
*023
*024
*001 + *014
*019 + *015
*023 + *024
*024 + *026

Figure 5.7. Functional assays investigating the trimming of ERAP1 allotype combinations compared to the wild type allotype *002-WT

293TE1KO cells were transfected with X5-SHL8, H2-K^b and the relevant ERAP1 allotype or allotype combination. The data show both ERAP1 allotypes *001 and *021 have a hypotrimming phenotype compared to the wild type allotype *002-WT. T cell activation assay (A) and normalisation using T cell activation assay and western blot for each allotype and combination with B3Z response expressed in percentage, compared to *002-WT presented on the histogram (B). Representation of the western blot carried out to investigate ERAP1 protein expression in 293TE1KO cells that have previously been transfected with 0.25µg X5-SHL8, 0.25µg H2-K^b and 0.5µg of either ERAP1 allotype *001 or *021 or a combination of both to a total of 1µg (C). ERAP1/GAPDH ratio used for normalisation (D). *002-WT was run on a separate western blot gel hence it is shown not to be part of the top gel along with *019, *015 and *019 + *015. Normalisation for allotypes *001 and *021 was repeated in six and five separate experiments, respectively confirming the accuracy of the normalisation for the allotype combination *001 and *021 (error bars for mean ± SEM). ***$p<0.001$. Also, $p<0.05$ for *026, *023 + *024 and *024 + *026 vs *002-WT.

5.1.3 Investigating the ERAP1 trimming of the HLA-A*0201 restricted N-terminally extended HPV-16 $E7_{82-90}$ epitope

As 99% of cervical cancer cases are caused by infection with HPV, it was of interest to investigate the ability of the ERAP1 allotype combinations identified within this cohort to trim an N-terminally extended antigenic peptide precursor derived from HPV-16 E7 protein. In the study by Reeves et al, ERAP1 allotype combinations from a cohort of HPV+ OPSCC patients were identified and their effect on the trimming ability of the enzyme was investigated towards both the model SHL8 peptide and the HLA-A*0201 restricted, HPV-16 $E7_{82-90}$ epitope [121]. Interestingly, the proteasomal cleavage predictions of HPV-16 E6/E7 proteins highlighted several peptide epitopes that would require N-terminal trimming by ERAP1 to generate the final peptide epitope. The epitope LLMGTLGIV (LV9) was selected over others due to its IC50 of 37.11nmol/L, the possible single amino acid extension D-LV9 as well as a two-amino acid extension, ED-LV9 and the fact that there were HPV-specific T cells in two out of five OPSCC patient tumours [121]. Upon binding of LV9 to the MHC I molecule HLA-A*020101,

the complex is presented at the cell surface of the APCs to the previously engineered T cell hybridoma cell line, BE7A2Z [121].

In this study, only the trimming of ED-LV9 to the final epitope LV9 by the ERAP1 allotype combinations identified from the cervical cancer patient cohort was investigated. ED-LV9 was chosen over the single amino acid extension (D-LV9) as the aforementioned study showed that when the trimming of either E-LV9 or D-LV9 was investigated, trimming was less efficient by ERAP1 allotypes compared to the trimming of ED-LV9 [121]. Since the LV9 epitope is presented on HLA-A*0201 alleles, those cervical cancer patients who were HLA-A*0201 positive (presented in chapter 3) needed to be determined before functional analysis could be undertaken. To further streamline the functional assays, only those patients that were both HLA-A*0201 positive and for whom there were also data available from the University of Groningen regarding CD8+/TIL numbers that were calculated from tumours removed from these patients (chapter 3) were assessed for LV9 trimming activity.

Therefore, functional assays were carried out involving the ERAP1 allotype combinations identified from these twenty-five patients (Table 5.1).

Table 5.1: A total of 39 HLA-A*0201 positive patients were identified from the cervical cancer patient cohort.

The samples for which CD8+/TIL data are available are indicated in bold. Novel ERAP1 allotypes are indicated in red. For the patients from which a homozygotic ERAP1 allotype combination was identified, only one allotype is shown on the table due to the lack of genomic DNA that would confirm true homozygosity.

Patient sample	ERAP1 allotype	Amino acid changes at indicated positions									
		12	56	127	276	346	349	528	575	725	730
		T/I	E/K	R/P	I/M	G/D	M/V	K/R	D/N	R/Q	Q/E
S6	*013	T	E	P	I	G	M	K	D	R	Q
S14	*018	T	E	R	I	G	M	K	D	R	E
	*013	T	E	P	I	G	M	K	D	R	Q
S27	*021	T	E	P	M	G	M	R	D	R	E

	*013	T	E	P	I	G	M	K	D	R	Q
S28	*002-WT	T	E	R	I	G	M	K	D	R	Q
	*021	T	E	P	M	G	M	R	D	R	E
S31	*013	T	E	P	I	G	M	K	D	R	Q
	*015	T	E	P	I	G	M	R	D	R	E
S32	*019	T	E	R	I	D	M	R	D	R	E
	*015	T	E	P	I	G	M	R	D	R	E
S33	*001	T	E	P	I	G	V	R	N	Q	E
	*002-WT	T	E	R	I	G	M	K	D	R	Q
S34	*019	T	E	R	I	D	M	R	D	R	E
	*015	T	E	P	I	G	M	R	D	R	E
S36	*013	T	E	P	I	G	M	K	D	R	Q
	*018	T	E	R	I	G	M	K	D	R	E
S38	*001	T	E	P	I	G	V	R	N	Q	E
	*021	T	E	P	M	G	M	R	D	R	E
S40	*018	T	E	R	I	G	M	K	D	R	E
	*021	T	E	P	M	G	M	R	D	R	E
S41	*002-WT	T	E	R	I	G	M	K	D	R	Q
	*015	T	E	P	I	G	M	R	D	R	E
S44	*013	T	E	P	I	G	M	K	D	R	Q
S45	*015	T	E	P	I	G	M	R	D	R	E
	*021	T	E	P	M	G	M	R	D	R	E
S46	*018	T	E	R	I	G	M	K	D	R	E
S48	*015	T	E	P	I	G	M	R	D	R	E
	*001	T	E	P	I	G	V	R	N	Q	E
S49	*015	T	E	P	I	G	M	R	D	R	E
S51	*002-WT	T	E	R	I	G	M	K	D	R	Q
S52	*001	T	E	P	I	G	V	R	N	Q	E
S53	*002-WT	T	E	R	I	G	M	K	D	R	Q
	*013	T	E	P	I	G	M	K	D	R	Q
S56	*002-WT	T	E	R	I	G	M	K	D	R	Q
	*001	T	E	P	I	G	V	R	N	Q	E
S58	*002-WT	T	E	R	I	G	M	K	D	R	Q
	*021	T	E	P	M	G	M	R	D	R	E

S59	*015	T	E	P	I	G	M	R	D	R	E
	*021	T	E	P	M	G	M	R	D	R	E
S64	*021	T	E	P	M	G	M	R	D	R	E
S66	*001	T	E	P	I	G	V	R	N	Q	E
	*021	T	E	P	M	G	M	R	D	R	E
S71	*018	T	E	R	I	G	M	K	D	R	E
	*013	T	E	P	I	G	M	K	D	R	Q
S72	*018	T	E	R	I	G	M	K	D	R	E
	*013	T	E	P	I	G	M	K	D	R	Q
S74	*015	T	E	P	I	G	M	R	D	R	E
	*021	T	E	P	M	G	M	R	D	R	E
S76	*021	T	E	P	M	G	M	R	D	R	E
S78	*018	T	E	R	I	G	M	K	D	R	E
	*001	T	E	P	I	G	V	R	N	Q	E
S89	*002-WT	T	E	R	I	G	M	K	D	R	Q
S102	*018	T	E	R	I	G	M	K	D	R	E
	*021	T	E	P	M	G	M	R	D	R	E
S103	*018	T	E	R	I	G	M	K	D	R	E
	*027	T	E	P	M	G	M	R	D	R	Q
S106	*013	T	E	P	I	G	M	K	D	R	Q
S108	*001	T	E	P	I	G	V	R	N	Q	E
	*014	T	K	P	I	G	M	R	D	R	E
S112	*002-WT	T	E	R	I	G	M	K	D	R	Q
S113	*018	T	E	R	I	G	M	K	D	R	E
	*013	T	E	P	I	G	M	K	D	R	Q
S114	*021	T	E	P	M	G	M	R	D	R	E
S117	*015	T	E	P	I	G	M	R	D	R	E

5.1.3.1 Optimising the transfection method of 293TE1KO cells to investigate the trimming of an HPV-derived N-terminally extended peptide precursor by patient ERAP1 allotypes

In order to determine the ability of those ERAP1 allotype combinations found in HLA-A*0201 patients within the cervical cancer patient cohort, the T cell activation assay utilising the BE7A2Z T cell hybridoma specific for the HLA-A*0201 restricted HPV-16 $E7_{82-90}$ epitope, LV9, was optimised.

As the antigen presenting cells used in these assays, 293TE1KO cells, naturally express HLA-A*0201, it was of interest to investigate whether HLA-A*0201 plasmid DNA should be transfected into 293TE1KO cells along with the relevant patient ERAP1 allotype combinations and the minigene construct of N-terminally extended version of LV9, ED-LV9, for the functional assays. The reason that this was investigated was because 293TE1KO cells may not express sufficient levels of HLA-A*0201 at the cell surface at sufficient levels to present the LV9 epitope to B3ZA2Z T cell hybridoma cells and hence the transfection of HLA-A*0201 might be required to bring levels to optimal for presentation of LV9.

First, 293TE1KO cells were transfected with either 0.5µg or 1µg HLA-A*0201 plasmid DNA. Untransfected 293TE1KO cells were also cultured along with the transfected cells and were used for comparing natural HLA-A*0201 expression by 293TE1KO cells to HLA-A*0201 expression following transfection with a suitable minigene construct. After twenty-four hours, cells were harvested and HLA-A*0201 surface expression was investigated by flow cytometry using HLA-A*0201 specific antibodies (Figure 5.8). Flow cytometry data revealed that a similar number of 293TE1KO cells expressed HLA-A2 at the cell surface either naturally or following transfection with either 0.5µg or 1µg HLA-A2 plasmid DNA. Consequently, transfection of 293TE1KO cells with HLA-A2 was not deemed necessary for cell surface expression of HLA-A2.

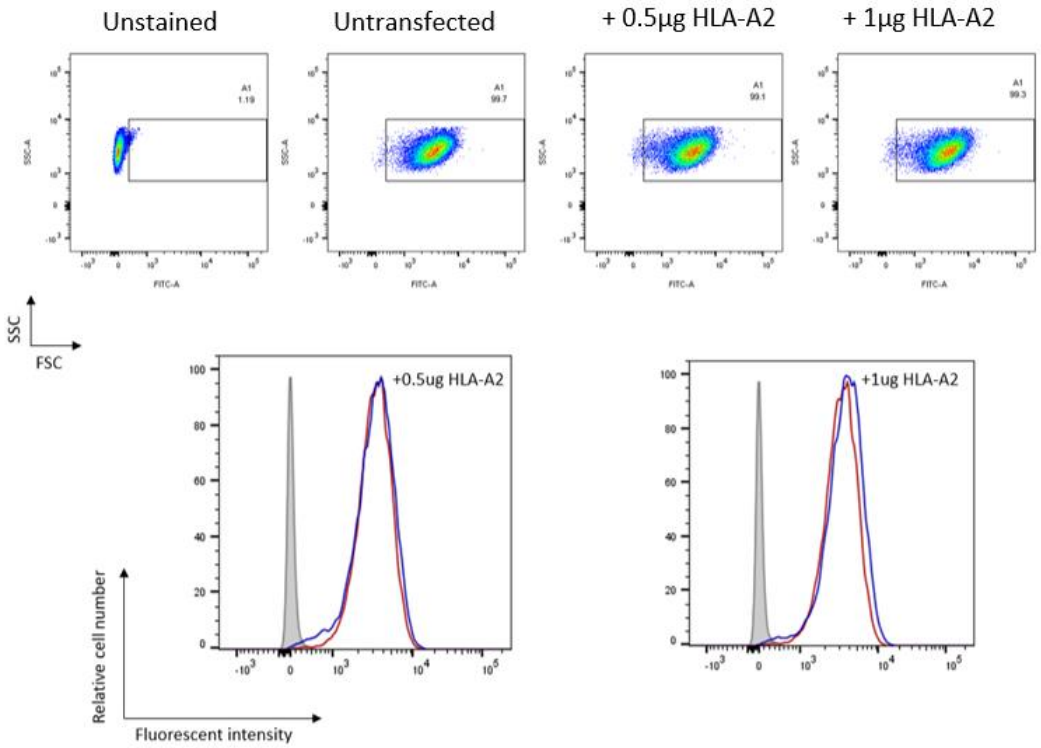


Figure 5.8: Cell surface HLA-A*0201 in 293TE1KO are similar with and without transfection with HLA-A*0201 minigene construct.

293TE1KO cells were transfected with either 0.5µg or 1µg of HLA-A*0201 plasmid DNA followed by staining with antibodies that show HLA-A2 expression by flow cytometry. Untransfected 293T1KO were also stained to investigate natural HLA-A*0201 expression by cells and unstained untransfected cells were used as a negative control. Results indicated that transfection of 293T1KO cells with HLA-A*0201 is not required for the functional assays.

In addition to examining cell surface expression of HLA-A*0201, another method that was used to investigate whether HLA-A*0201 plasmid DNA should be further transfected into 293TE1KO cells involved carrying out a T cell activation assay. In this assay, four separate transfections of 293TE1KO cells were carried out. Cells were transfected with either *002-WT ERAP1 DNA, with and without the HLA-A*0201 plasmid DNA or with non-functional ERAP1 (E320A), with and without further addition of HLA-A*0201. Cells were transfected with the final peptide, LV9, or the N-terminally extended precursor, ED-LV9, which requires functional ERAP1 activity to generate the final epitope (Figure 5.9).

The control T cell activation assay showed that the highest BE7A2Z T cell hybridoma response was generated after 293TE1KO cells were transfected with *002-WT and with either LV9 or ED-LV9 without transfection of an HLA-A*0201. This meant that the natural expression of HLA-A*0201 on the cell surface of 293TE1KO cells was sufficient to generate an optimal BE7A2Z response following the successful trimming of ED-LV9 to the final epitope LV9 by *002-WT. Consequently, as it was shown by both the flow cytometry experiment and the control T cell activation assay, BE7A2Z response was optimal without further transfection of HLA-A*0201 into 293TE1KO cells that present the complex comprised of the naturally expressed HLA-A*0201 at the cell surface and the final epitope LV9 to BE7A2Z T cells (Figure 5.9).

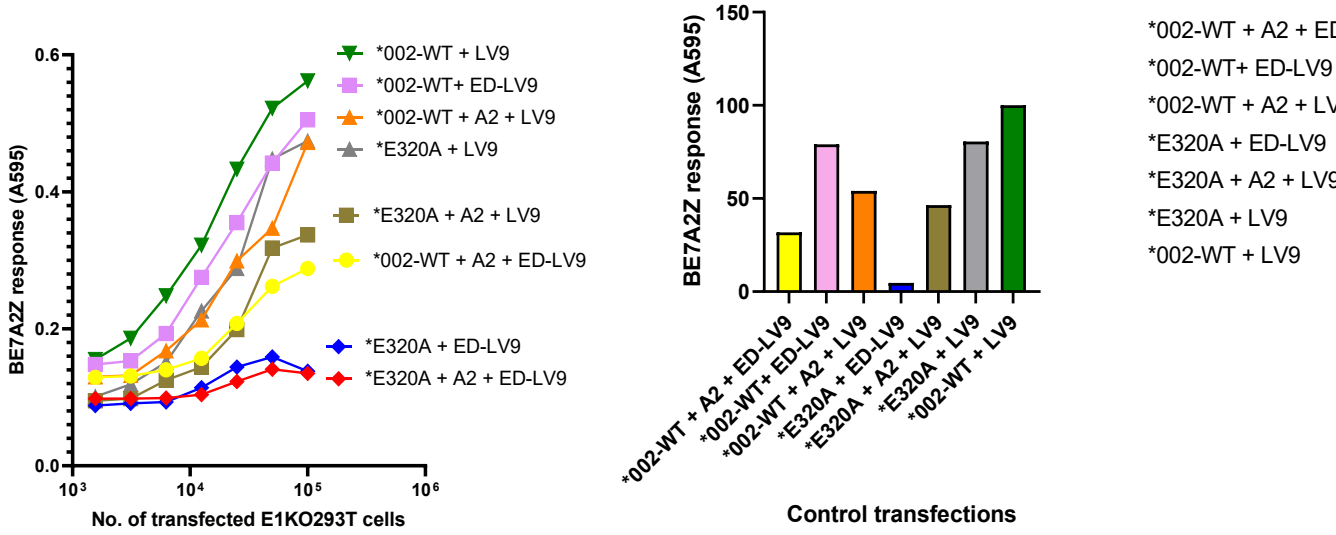


Figure 5.9: Control T cell activation assay investigating whether the transfection of HLA-A*0201 plasmid DNA into 293TE1KO cells is required for maximal BE7A2Z response.

The control activation assay was used to investigate whether addition of HLA-A*0201 into 293TE1KO cells is required to maximise BE7A2Z T cell hybridoma response. As expected, transfection of either *002-WT or the non-functional E320A and LV9 lead to the highest BE7A2Z response because the epitope did not require trimming by ERAP1. Nonetheless, ERAP1 *002-WT was able to trim ED-LV9 to the final epitope LV9 with efficient BE7A2Z response generated. Transfection with *002-WT, HLA-A2 and ED-LV9 lead to one of the lowest T cell hybridoma responses indicating that addition of HLA-A2 is not required to present the pMHC I to the hybridoma cells. Data analysed using Graphpad Prism.

5.1.4 Functional assays carried out to investigate the trimming of an HPV-derived epitope by ERAP1 allotype combinations identified from 25 HLA-A*0201 positive cervical cancer patients for which CD8+/TIL status is known

Similarly to the assessment of trimming X5-SHL8 by ERAP1, T cells activation assays were carried out using the ED-LV9 precursor alongside western blots to determine relative protein expression of the transfected cells. As mentioned earlier, it was decided to investigate the trimming function of the ERAP1 allotype combinations from those HLA-A*0201 positive patients for whom the CD8+/TIL status was known. A total of 7 distinct ERAP1 allotypes, found in 16 combinations, was identified from those HLA-A*0201 positive patients for whom the CD8+/TIL status is known, 11 heterozygotes and 5 homozygotes which could not be confirmed as true homozygotes due to lack of genomic DNA that could be used for sequencing (Table 4.4). The BE7A2Z T cell hybridoma response generated following presentation of LV9:HLA-A2 peptide complexes that were previously trimmed by the ERAP1 allotype combinations identified from the 25 patients mentioned above on the surface of 293TE1KO cells is expressed as a percentage as part of figures presented in this section. The trimming of ED-LV9 to LV9 by the allotype combinations *018 + *001 and *002-WT + *001 was not investigated here as trimming of this epitope has already been investigated in the study by Reeves et al when the combinations were identified from the HPV+ OPSCC patient cohort [121].

The combination of ERAP1 allotypes with *021 (*001, *002-WT and *015) was assessed first (Figure 5.10). In the previous experiments involving the B3Z T cell hybridoma, *021 was shown to have on average 32% activity compared to *002-WT. This experiment involved a different hybridoma T cell line and the trimming of a different epitope, LV9, was investigated. *021 was shown to have higher activity; approximately 56% compared to *002-WT ($p<0.005$). This percentage was the average of five separate T cell activation assays and western blots as it was found in five separate combinations. The trimming activity of *001 and *015 also showed differences compared to the activity towards X5-

SHL8. BE7A2Z T cell response following the trimming of ED-LV9 by allotype *001 was over 60%, while the relevant percentage for B3Z T cells was less than 30%. For allotype *015, BE7A2Z response was shown to be close to 60% which is more than 40% higher than that shown for B3Z T cells. The trimming data for both allotypes *015 and *001 match those presented by Reeves et al investigating the function of ERAP1 allotypes identified from a cohort of HPV positive OPSCC patients [121]. It has been suggested by studies that N-1 (the first amino acid of the N-terminal extension preceding the peptide) affects the ability of ERAP1 to trim the antigenic precursor to the final epitope more than the sequence of the epitope [90, 113]. This was confirmed in the study mentioned above by Reeves et al showing that trimming specificities of ERAP1 allotypes/combinations differed between X-SHL8 and X-LV9. The top three N-1 specificities that would lead to highest trimming for *001 were C, A and E for X-SHL8 and E, K and N for X-LV9. The relevant specificities for *015 were E, H and Q for X-SHL8 and E, H and K for LV9 [121]. The combination of *002-WT + *021 lead to a slight increase in trimming activity compared to *021 alone. Interestingly, the combination formed by allotypes *001 + *021 showed a significant difference in trimming ED-LV9 compared to the trimming of X5-SHL8. The average BE7A2Z response was more than 70%, while the relevant percentage for the trimming of X5-SHL8 was 50% of *002-WT. A similar trimming activity was observed for the combination composed of allotype *015 + *021, the trimming activity of which was approximately double the one shown for the trimming of X5-SHL8.

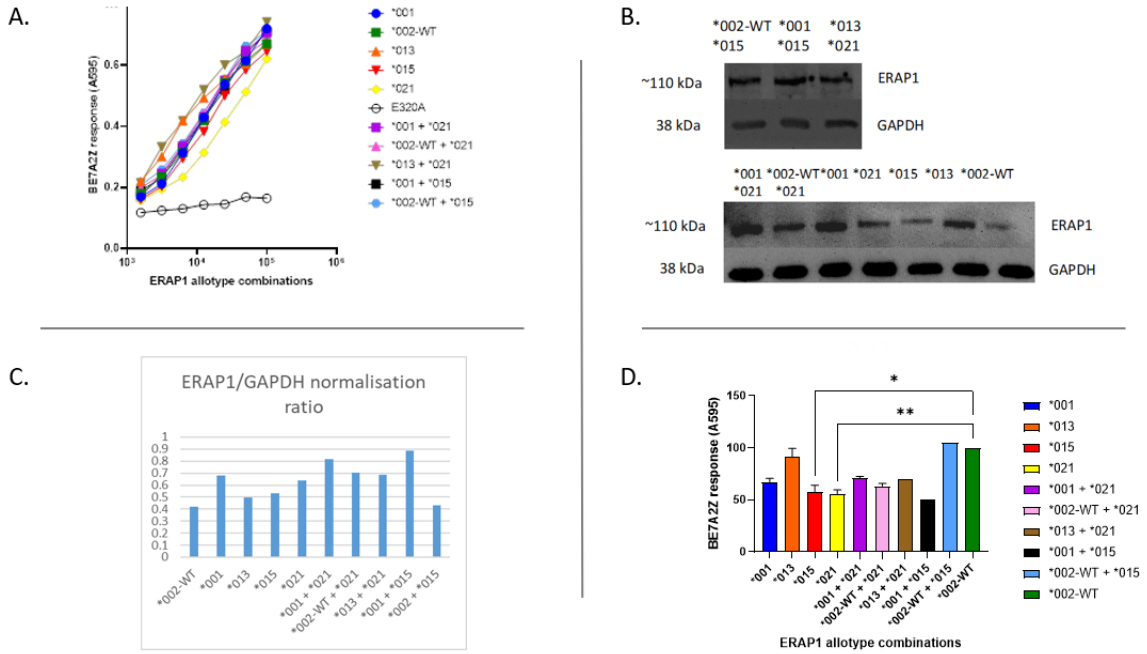


Figure 5.10. Functional assays investigating the trimming of ERAP1 allotype combinations compared to *002-WT

293TE1KO cells were transfected with ED-LV9, empty pcDNA3 vector and the relevant ERAP1 allotype or allotype combination. The data show that the BE7A2Z T cell hybridoma response following the trimming of ED-LV9 to LV9 by the allotypes shown in the T cell activation assay (A) Representation of the western blot carried out to investigate ERAP1 protein expression in 293TE1KO cells that have previously been transfected with 0.25µg ED-LV9, 0.25µg empty pcDNA3 vector and 0.5µg of ERAP1 allotype/combination (0.25µg per allotype in heterozygous combinations) (B). Normalisation bar chart using T cell activation assay and western blot for each allotype and combination with BE7A2Z response expressed in percentage, compared to *002-WT presented on the histogram (C, D). Normalisation for single allotypes *001, *015, *021 was repeated in four, four and five separate experiments, respectively confirming the accuracy of the normalisation for the allotype combinations *002-WT + *021, *015 + *021 and *001 + *021 (error bar indicates mean ± SEM). *p<0.05 and **p<0.005.

The next allotype combination that was investigated was *019 and allotype *015 (Figure 5.11). The T cell hybridoma responses after the trimming of X5-SHL8 and ED-LV9 by *019 were similar. The overall trimming of ERAP1 allotypes *019 + *015 was about 65% which is similar to the one by either allotype *019 or *015. The percentage for *019 + *015 was also about 50% higher than that seen for B3Z T

cells meaning that again, the difference in the two peptides used contributed to the differences in the trimming and hence the response shown by T cells.

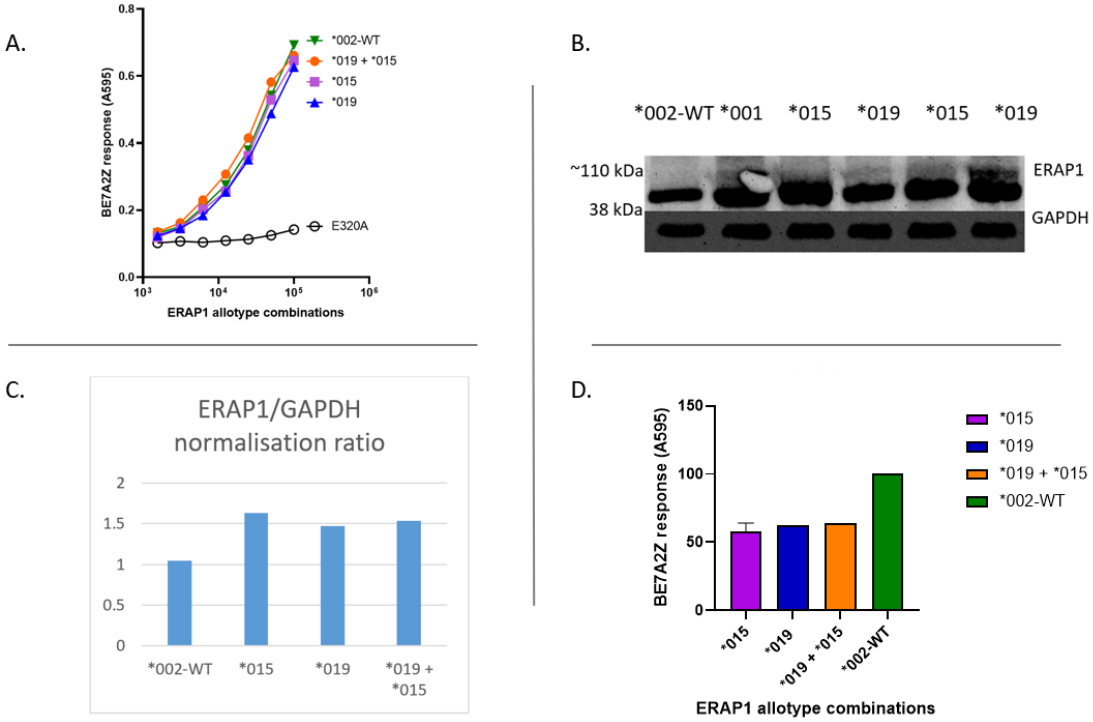


Figure 5.11. Functional assays investigating the trimming of ERAP1 allotype combinations compared to *002-WT

293TE1KO cells were transfected with ED-LV9, empty pcDNA3 vector and the relevant ERAP1 allotype or allotype combination. The data show that the allotype combination *019 + *015 leads to efficient generation of LV9 from ED-LV9. T cell activation assay (A). Representation of the western blot carried out to investigate ERAP1 protein expression in 293TE1KO cells that have previously been transfected with 0.25µg ED-LV9, 0.25µg empty pcDNA3 vector and 0.5µg of ERAP1 allotype/combination (0.25µg per allotype in heterozygous combinations) (B). Normalisation using T cell activation assay and western blot for each allotype and combination with BE7A2Z response expressed in percentage, compared to *002-WT presented on the histogram(C, D). Normalisation for single allotypes *015 was repeated in five separate experiments, error bar represents mean ± SEM. No statistically significant differences in trimming were observed compared to *002-WT.

Allotype *018 (E730) was identified with high frequency in both the cohort in this study and the study by Reeves et al in another HPV driven cancer [121]. This allotype was found in the following combinations; *018 + *013 as well as *018 + *021 (Figure 5.12). Earlier experiments involving B3Z T cells indicated that the response generated following trimming of X5-SHL8 to SHL8 by allotype *018 was approximately 50% of *002-WT. Trimming of ED-LV9 by *018 was more efficient than trimming of X5-SHL8. Allotype *018 contains the amino acid change Q730E which changes the interaction between ERAP1 (negative potential) and the C-terminal moiety of the substrate, hence reducing trimming [122]. When this allotype was found in combination with allotype *013, trimming was as efficient as for *002-WT.

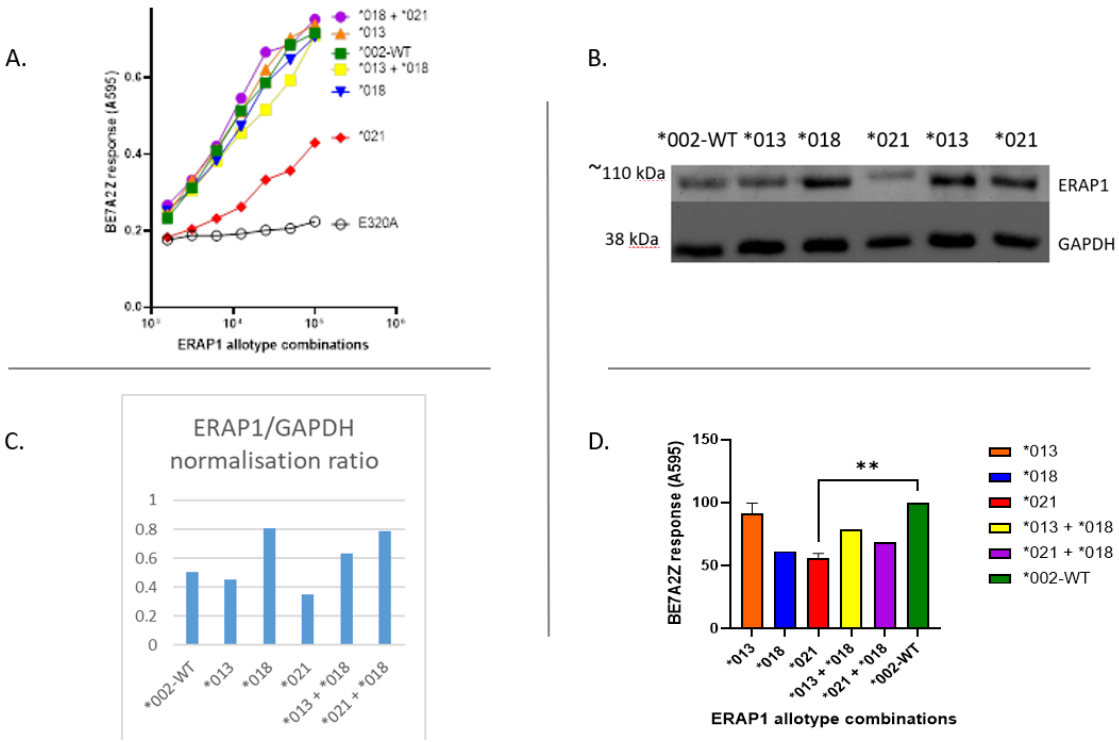


Figure 5.12. Functional assays investigating the trimming of ERAP1 allotype combinations compared to *002-WT

293TE1KO cells were transfected with ED-LV9, empty pcDNA3 vector and the relevant ERAP1 allotype or allotype combination. The data show that *018 is more efficient at trimming ED-LV9 than X5-SHL8 and trimming is restored when in combination with an efficient trimmer (*013). T cell activation assay (A). Representation of the western blot carried out to investigate ERAP1 protein expression in 293TE1KO cells that have previously been

transfected with 0.25µg ED-LV9, 0.25µg empty pcDNA3 vector and 0.5µg of ERAP1 allotype/combination (0.25µg per allotype in heterozygous combinations) (B). Normalisation using T cell activation assay and western blot for each allotype and combination with BE7A2Z response expressed in percentage, compared to *002-WT presented on the histogram (C, D). Normalisation for single allotypes *021 was repeated in five separate experiments, error bars indicate mean ± SEM. **p<0.005.

The trimming of ED-LV9 to LV9 by the majority of ERAP1 allotype combinations identified from this HLA-A*0201 positive cervical cancer patient cohort has not been investigated before. Lysates created from 293TE1KO cells that were previously transfected with either of two ERAP1 allotype combinations, *002-WT + *021 or *015 + *021, were used to investigate ERAP1 expression through western blots that were repeated in two separate experiments. The transfected cells were also used to carry out the relevant T cell activation assays. Both the T cell activation assays and the western blots were used to calculate the overall trimming activity of the ERAP1 allotype combinations. The data that were generated for the functional activity of these combinations showed that there was a difference of less than 10% in the BE7A2Z response generated following the trimming of the HPV-derived epitope by either of the two allotype combinations.

The functional activity of two other ERAP1 allotype combinations that were identified from the HLA-A*0201 positive patient cohort was investigated; *002-WT + *015 as well as *013 + *021.

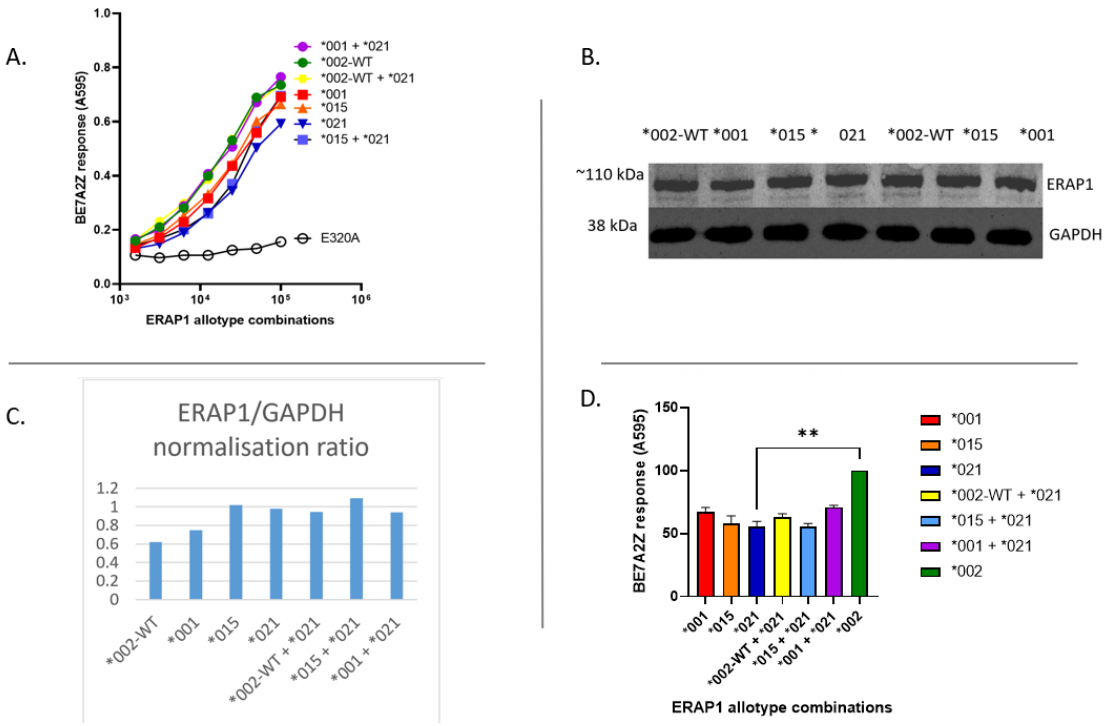


Figure 5.13. Functional assays investigating the trimming of ERAP1 allotype combinations compared to *002-WT

293TE1KO cells were transfected with ED-LV9, empty pcDNA3 vector and the relevant ERAP1 allotype or allotype combination. The data show that *002-WT restores trimming when in combination with the hypotrimmer *015 but not *021, indicating that perhaps the amino acid change I276M has a negative effect in the interaction with the substrate. T cell activation assay (A) and Representation of the western blot carried out to investigate ERAP1 protein expression in 293TE1KO cells that have previously been transfected with 0.25µg ED-LV9, 0.25µg empty pcDNA3 vector and 0.5µg of ERAP1 allotype/combination (0.25µg per allotype in heterozygous combinations) (B). Normalisation using T cell activation assay and western blot for each allotype and combination with BE7A2Z response expressed in percentage, compared to *002-WT presented on the histogram (C, D). Normalisation for single allotype *015 was repeated in five separate experiments and assisted in predicting the accuracy of the data for the trimming of allotype combinations one of which is *015 with error bars for mean ± SEM. **p<0.005.

*5.1.4.1 CD8+/TIL status of HLA-A*0201 positive cervical cancer patients*

CD8+ tumour infiltrating T cells (CD8+/TILs) from the selected HLA-A*0201 positive patients were classified into three categories as in chapter 3, section 3.3 but the median for the HLA-A*0201 positive patients with CD8+/TIL data available was calculated to be approximately 528.9 cells per tumour mm^2.

The patients whose CD8+/TIL numbers lie within 25% above or below the median were assigned to the "moderate" group, while those patients whose CD8+/TIL numbers were found to be higher or lower than 25% of the median were assigned to the CD8+/TIL^{high} or CD8+/TIL^{low} group, respectively (Figure 5.14).The ERAP1 allotype combinations identified with long read sequencing from the selected 25 HLA-A*0201 positive cervical cancer patients along with the relevant T cell group are shown in Table 5.2. There was a total of 8 HLA-A*0201 positive patients in the CD8+/TIL^{low} group, 7 patients in the CD8+/TIL^{mod} group and 10 patients in the CD8+/TIL^{high} group (Figure 5.14).

Table 5.2. ERAP1 allotype combinations identified from HLA-A*0201 positive patients and their CD8+/TIL status

Amino acid changes indicated in bold.

			Amino acid changes at indicated positions									
Patient samples	CD8+/TIL status	ERAP1 allotypes	12	56	127	276	346	349	528	575	725	730
			T/I	E/K	R/P	I/M	G/D	M/V	K/R	D/N	R/Q	Q/E
S6	Low	*013	T	E	P	I	G	M	K	D	R	Q
S14	Low	*018	T	E	R	I	G	M	K	D	R	E
		*013	T	E	P	I	G	M	K	D	R	Q
S27	Low	*021	T	E	P	M	G	M	R	D	R	E
		*013	T	E	P	I	G	M	K	D	R	Q
S28	High	*002-WT	T	E	R	I	G	M	K	D	R	Q
		*021	T	E	P	M	G	M	R	D	R	E
S34	High	*019	T	E	R	I	D	M	R	D	R	E
		*015	T	E	P	I	G	M	R	D	R	E
S40	High	*018	T	E	R	I	G	M	K	D	R	E
		*021	T	E	P	M	G	M	R	D	R	E
S41	Low	*002-WT	T	E	R	I	G	M	K	D	R	Q
		*015	T	E	P	I	G	M	R	D	R	E
S44	High	*013	T	E	P	I	G	M	K	D	R	Q
S46	High	*018	T	E	R	I	G	M	K	D	R	E

S48	Moderate	*015	T	E	P	I	G	M	R	D	R	E
		*001	T	E	P	I	G	V	R	N	Q	E
S51	Moderate	*002-WT	T	E	R	I	G	M	K	D	R	Q
S52	Moderate	*001	T	E	P	I	G	V	R	N	Q	E
S56	High	*002-WT	T	E	R	I	G	M	K	D	R	Q
		*001	T	E	P	I	G	V	R	N	Q	E
S58	High	*002-WT	T	E	R	I	G	M	K	D	R	Q
		*021	T	E	P	M	G	M	R	D	R	E
S59	Low	*015	T	E	P	I	G	M	R	D	R	E
		*021	T	E	P	M	G	M	R	D	R	E
S64	Moderate	*021	T	E	P	M	G	M	R	D	R	E
S66	Moderate	*001	T	E	P	I	G	V	R	N	Q	E
		*021	T	E	P	M	G	M	R	D	R	E
S71	High	*018	T	E	R	I	G	M	K	D	R	E
		*013	T	E	P	I	G	M	K	D	R	Q
S72	Low	*018	T	E	R	I	G	M	K	D	R	E
		*013	T	E	P	I	G	M	K	D	R	Q
S74	Low	*015	T	E	P	I	G	M	R	D	R	E
		*021	T	E	P	M	G	M	R	D	R	E
S78	Moderate	*018	T	E	R	I	G	M	K	D	R	E
		*001	T	E	P	I	G	V	R	N	Q	E
S89	Low	*002-WT	T	E	R	I	G	M	K	D	R	Q
S102	High	*018	T	E	R	I	G	M	K	D	R	E
		*021	T	E	P	M	G	M	R	D	R	E
S106	Moderate	*013	T	E	P	I	G	M	K	D	R	Q
S113	High	*018	T	E	R	I	G	M	K	D	R	E
		*013	T	E	P	I	G	M	K	D	R	Q

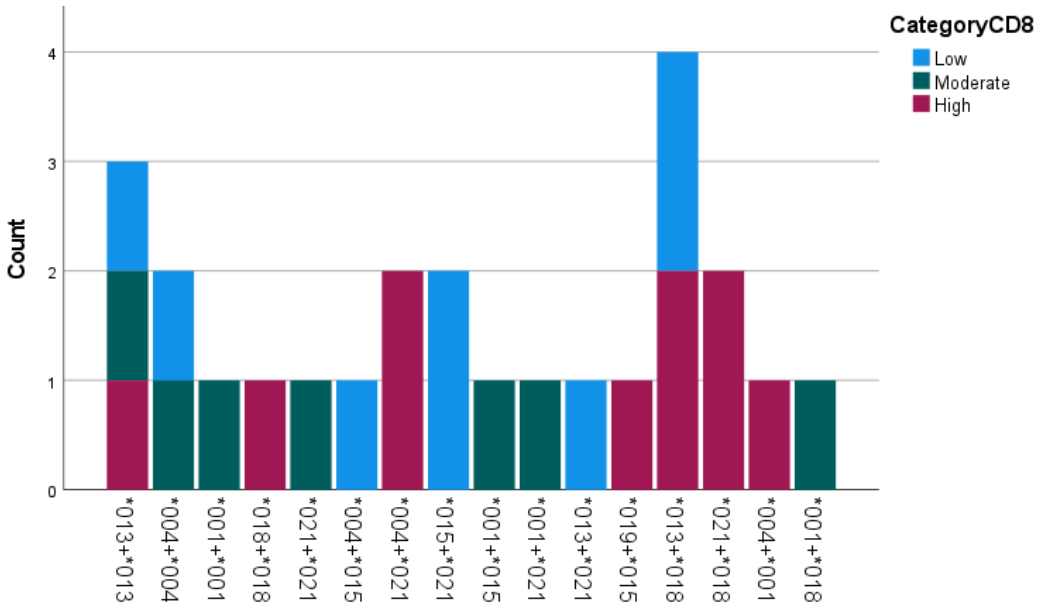


Figure 5.14: Frequency of ERAP1 allotype combinations assigned to either of three CD8+/TIL groups.

Histogram shows how ERAP1 allotype combinations are assigned to the three CD8+/TIL groups created using the median calculated from the CD8+/TILs per tumour mm^2 that was available for 25/39 HLA-A*0201 positive patients determined from the cohort of patients for which ERAP1 allotypes were identified with long read sequencing. High CD8+/TIL group (red), moderate CD8+/TIL group (green) and low CD8+/TIL group (blue).

5.1.4.2 High CD8+/TIL number and ERAP1 with normal trimming activity were associated with better disease prognosis

High CD8+/TIL numbers have been associated with better overall prognosis in cervical cancer patients [166, 170, 179]. It was therefore of interest to investigate whether this was the case for the patient cohort in the present study and identify associations with the trimming by the ERAP1 allotype combinations of these patients.

As it can be noted from Table 5.3, all 10 patients assigned to the CD8+/TIL^{high} group had an ERAP1 allotype combination with an efficient overall trimming phenotype. The two most common allotype combinations were *002-WT+*021 and *018 +*013. When the trimming of X5-SHL8 to SHL8 by these two allotype combinations was investigated, it was revealed that *002-WT was able to rescue the

overall trimming of ERAP1 as it was shown that both allotypes *021 and *018 had a hypotrimming phenotype when their trimming activity was individually investigated. With regards to the trimming of ED-LV9, the combination *002-WT + *021 revealed a BE7A2Z response of approximately 60%, while the relevant percentage for the combination *013 + *018 was similar to the one generated for *002-WT. Even though the number of HPV-specific CD8+/TILs was not available due to the lack of tumour samples from the patients, the presence of HPV-specific T cells in the CD8+/TIL^{high} group recognising HLA-A*0201:LV9 is likely (Table 5.3).

Table 5.3: ERAP1 allotype combinations identified in the CD8+/TIL^{high} group and disease progression/outcome

AC= Adenocarcinoma, DF= Disease-free.

Patients	Type	ERAP1 allotype combinations	Trimming phenotype ED-LV9	Lymph node metastasis	Distant metastasis	Last follow-up
S28	Squamous	*002-WT + *021	Efficient	No	No	DF
S34	Squamous	*019 + *015	Efficient	Yes	No	DF
S40	AC	*018 + *021	Efficient	Yes	No	DF
S102	Squamous	*018 + *021	Efficient	No	No	DF
S46	AC	*018	Efficient	No	No	DF
S56	Squamous	*002-WT + *001	Efficient	No	No	DF
S44	AC	*013	Efficient	No	No	DF
S71	Squamous	*013 + *018	Efficient	No	No	DF
S58	AC	*002-WT + *021	Efficient	No	No	DF
S113	AC	*013 + *018	Efficient	No	No	DF

As regards to the CD8+/TIL^{mod} patient group, it was observed that it contained most of the homozygotic ERAP1 allotype combinations identified with long read sequencing in the cohort of the 25 HLA-A2 positive cervical cancer patients (Table 5.2). These were *001 + *001, *021 + *021, *013 +

*013 and *002-WT + *002-WT with optimal trimming activity for both X5-SHL8 and ED-LV9. This was the only group of patients containing the homozygotic combinations of *001 and *021. Both of these homozygotic combinations were shown to have hypotrimming activity in the functional assays investigating the trimming of X5-SHL8. The BE7A2ZZ responses generated following the trimming of ED-LV9 by either *001 + *001 (S52) or *021 + *021 (S64) were shown to be approximately 70 and 60%, respectively. S52 (*001 homozygote) was recorded as disease-free during the last follow-up and did not experience any lymph node or distant metastasis. It is therefore possible that a proportion of the CD8+/TILs had a TCR that was HPV-specific for the LV9 epitope as the homozygotic combination of *001 efficiently trimmed ED-LV9. However, this cannot be confirmed as tumours removed from patients for investigating the role of LV9-specific CD8+/TILs were not available. Also, the functionality of these cells was not investigated, and it is likely that other cells infiltrating the TME might be contributing to better overall prognosis. S64 (*021 homozygote) suffering from adenocarcinoma, was also disease-free at the last follow-up and did not experience any kind of metastasis. Again, it is likely that the homozygotic allotype combination of *021 is contributing to some degree to better disease outcome through the trimming of HPV-derived epitopes that are recognised by HPV-specific T cells, but other explanations are possible such as the infiltration of other immune cells in TME or treatment approaches [175, 203] (Table 5.4). S66 experienced lymph node metastasis but was disease-free at the last follow-up pointing towards the role of the normal trimming allotype combination, *001 + *021 of the N-terminally-extended HPV epitope investigated as a likely contributor.

Table 5.4: ERAP1 allotype combinations identified in the CD8+/TIL^{mod} group and disease progression/outcome

AC= Adenocarcinoma, DF= Disease-free.

Patients	Type	ERAP1 allotype combinations	Trimming phenotype ED-LV9	Lymph node metastasis	Distant metastasis	Last follow-up
S48	AC	*015 + *001	Hypotrimmer	No	No	DF

S52	Squamous	*001	Efficient	No	No	DF
S64	AC	*021	Efficient	No	No	DF
S66	Squamous	*001 + *021	Efficient	Yes	No	DF
S78	AC	*001 + *018	Efficient	No	No	DF
S51	Squamous	*002-WT	Efficient	No	No	DF
S106	Squamous	*013	Efficient	Yes	No	Died

Regarding the ERAP1 allotype combinations identified from the CD8+/TIL^{low} group, the majority were efficient at generating LV9 from ED-LV9. Interestingly, the ERAP1 allotype combinations *013 + *021, *002-WT + *015 and *015 + *021, with the last combination identified in two patients both of which were in the CD8+/TIL^{low} group (S59 and S74), were identified only in this group (Table 5.5). Additionally, S27 whose allotypes were only identified from the CD8+/TIL^{low} group, succumbed to the disease as reported at the last follow-up even though the allotype pair of this patient had efficient trimming ability, indicating that other factors also play a role in disease prognosis (Table 5.5). Since the majority of patients in the CD8+/TIL^{low} group had allotypes with efficient trimming activity, it is possible that ERAP1 trims well other HPV epitopes in addition to ED-LV9 (S41 infected with HPV-16 only) and these epitopes could originate from other HPV types the patients might have been infected with (HPV status unknown for S27).

Table 5.5: ERAP1 allotype combinations identified in the CD8/TIL^{low} group and disease progression/outcome

DF = Disease-free.

Patients	**Type**	**ERAP1 allotype combinations**	**Trimming phenotype ED-LV9**	**Lymph node metastasis**	**Distant metastasis**	**Last follow-up**
S27	Squamous	*013 + *021	Efficient	Yes	Yes	Died
S41	Squamous	*002-WT + *015	Efficient	Yes	No	DF

S59	Squamous	*015 + *021	Hypotrimmer	No	No	DF
S74	Squamous	*015 + *021	Hypotrimmer	Yes	No	DF
S6	Squamous	*013	Efficient	Yes	Yes	Died
S14	Squamous	*013 + *018	Efficient	No	No	DF
S72	Squamous	*013 + *018	Efficient	No	No	DF
S89	Squamous	*002-WT	Efficient	Yes	No	Died

The youngest patient in the patient cohort was S40 and was classified in the CD8+/TIL^{high} group (Figure 5.15). The ERAP1 allotype combination identified from this patient was *021 + *018 which was shown to trim ED-LV9 to LV9 efficiently. It is possible that even though the combination*021 + *018 trims efficiently other HPV-derived epitopes presented to CD8+/TILs and other immune cells might be present as well to contribute to better disease prognosis.

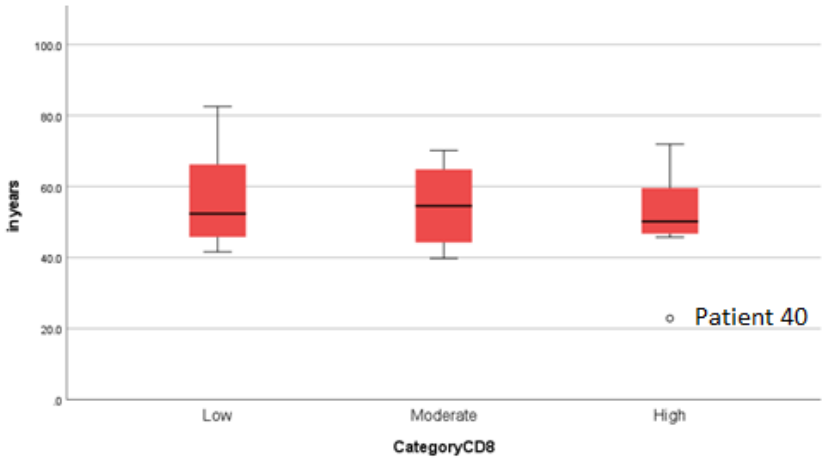


Figure 5.15: The youngest patient in the cohort at the age of diagnosis in the CD8+/TIL^{high} group had good overall prognosis

Boxplot shows the age of diagnosis of cervical carcinoma in patients categorised in three categories based on the CD8+/TIL number identified from tumours. Patient S40 was the youngest patient in the cohort and was assigned to the CD8+/TIL^{high} group. This patient had good overall prognosis. Graph generated with SPSS.

5.1.5 Summary and conclusions from the functional analysis of the ERAP1 allotype combinations

T cell activation assays and western blots were carried out to investigate the trimming phenotype of the ERAP1 allotype combinations identified from the cervical cancer patient cohort. The trimming of the OVA-derived N-terminally extended antigenic peptide precursor AIVMK-SHL8 was investigated from the allotypes identified from 81 patients and then the trimming of the HPV-derived N-terminally extended epitope ED-LV9 was investigated by the allotypes identified from 25 patients who were determined to be HLA-A*0201 positive and for whom the CD8+/TIL status is available. The trimming phenotype for ED-LV9 and X5-SHL8 epitopes is summarised on Table 5.8.

The trimming data for X5-SHL8 confirm that allotypes with a hypotrimming phenotype that were found in combination with an allotype with a normal trimming phenotype, resulted in an enzyme with an overall normal trimming ability (*013 + *001, *002-WT + *001, *002-WT + *021, *013 + *021, *013 + *025, *002-WT + *028, *002-WT + *015, *013 + *015, *013 + *018, *002-WT + *018). Interestingly, allotype *018 was not able to trim X5-SL8 as efficiently as ED-LV9 and the reason could lie in the substrate length to be trimmed by ERAP1 (13 amino acids vs 11 amino acids). These data for *018 are confirmed by the study in OPSCC patients by the James lab [121]. The combination of K528R/Q730E was shown to have a dominant effect on the trimming phenotype of the allotypes, specifically by making trimming of antigenic precursors less efficient. However, *001 was able to trim ED-LV9 efficiently which was not the case for X5-SHL8 and also it did not match the data generated by Reeves et al regarding this allotype. Four separate experiments were carried out for this allotype which increased confidence in determining the trimming phenotype of the allotype.

Regarding CD8+/TIL status and disease progression, a higher number of CD8+/TILs was associated with better overall prognosis in cervical carcinoma confirming historical data [166, 170, 179]. The median of CD8+/TILs was used to classify patients in three groups; CD8+/TIL^{high}, CD8+/TIL^{mod} and CD8+/TIL^{low}. Each group had a pair of unique ERAP1 allotypes identified only in that group and in the CD8+/TIL^{mod} the majority were homozygotic combinations. From the uniquely assigned ERAP1 allotype combinations, the patients that succumbed to the disease were from the CD8+/TIL^{low} group as

expected, but the allotype combination was efficient indicating that perhaps there are other epitopes that are not efficiently trimmed by the ERAP1 of this patient or epitopes that do not require trimming by ERAP1 explaining the disease outcome.

Table 5.6. Trimming efficiency of ERAP1 allotype combinations identified from the cervical cancer patient cohort.

N/A indicates that trimming of ED-LV9 was by the relevant pair of allotypes was not investigated as these pairs were not identified in HLA-A*0201 positive patients. The trimming of ED-LV9 by *002-WT + *001 as well as *018 + *001 was not investigated as this has previously been shown by the James lab [121]. Novel ERAP1 allotypes are indicated in red.

ERAP1 allotype combinations	Trimming of X5-SHL8 (n=81)	Trimming of ED-LV9 (n=25)
*002-WT	efficient	efficient
*001	hypotrimmer	efficient
*015	hypotrimmer	efficient
*021	hypotrimmer	efficient
*013	efficient	efficient
*018	hypotrimmer	efficient
*001 + *021	hypotrimmer	efficient
*002-WT + *013	efficient	N/A
*001 + *013	efficient	N/A
*002-WT + *001	efficient	efficient [121]
*002-WT + *021	efficient	efficient
*013 + *021	efficient	efficient
*013 + *025	efficient	N/A
*002-WT + *028	efficient	N/A
*002-WT + *015	efficient	efficient
*013 + *015	efficient	N/A
*001 + *015	hypotrimmer	hypotrimmer
*021 + *015	hypotrimmer	efficient
*013 + *018	efficient	efficient
*001 + *018	hypotrimmer	efficient [121]
*027 + *018	hypotrimmer	N/A
*021 + *018	hypotrimmer	efficient
*002-WT + *018	efficient	N/A
*019 + *015	hypotrimmer	efficient
*001 + *014	hypotrimmer	N/A

*023 + *024	hypotrimmer	N/A
*024 + *026	hypotrimmer	N/A

6 Discussion

6.1 Genetic variation in ERAP1 is associated with cervical cancer

ERAP1 is the ER-aminopeptidase responsible for the N-terminal trimming of peptides for binding to MHC I and presentation to CD8+ T cells and NK cells. ERAP1 has some unique enzymatic properties; it shows selectivity for the sequence and length of the epitope that it trims, sparing peptides 8-9 amino acids long for binding onto MHC I [86, 103, 118]. It preferentially removes large hydrophobic residues at the N-terminus of the peptide precursor, especially leucine and valine [96]. ERAP1 is polymorphic in humans and ten SNPs have been seen with >5% frequency in the population (Table 1.1) [119, 120]. Knowledge of the mechanism used by ERAP1 to trim peptides and the factors that affect trimming efficiency, for example the substrate sequence, the number and properties of the amino acids consisting the N-terminal extension, the specific HLA allele they bind to, is limited.

Genetic variation in ERAP1 has been associated with increased cervical cancer risk and decreased overall survival [181, 190, 215]. GWAS studies have investigated association of individual ERAP1 SNPs or imputed haplotypes with cervical cancer; heterozygosity of E56 and minor allele homozygosity for P127 were associated with decreased survival and heterozygosity for haplotype of E56/P127 associated with better overall prognosis [190]. P127 and E730 in a haplotype were associated with increased cervical cancer risk [181]. Genetic association studies have found ERAP1 SNPs in linkage disequilibrium with other SNPs of the same gene forming distinct haplotypes K528/D575/R725, K528/D575/E730 and R127/K528/D575 associated with increased risk of AS [130, 216]. Haplotypes P127/I276/528R and I276/R528/N575 were associated with AS protection [130]. However, these studies explored a limited number of SNPs and did not investigate function. Haplotypes encode distinct ERAP1 protein variants, referred to as 'allotypes' that alter the enzyme's trimming function. As both chromosomal copies of ERAP1 are co-dominantly expressed, both ERAP1 allotypes have an effect on the enzyme's overall trimming function. Tight linkage disequilibrium among ERAP1 SNPs potentially affects allotype diversity within the population and also increases the ability of individual

SNPs to affect function [217]. For instance, K528 increases protein expression [132, 218, 219]. SNPs in ERAP1 are in strong linkage disequilibrium and P127/R528 were shown to co-occur in the lymphoblastoid cell lines, although since P127 was shown to have little effect on ERAP1 activity, effects are more likely attributed to R528 [125, 220].

Further reinforcement of the importance of ERAP1, and the effect of polymorphic variation in ERAP1 in HPV-driven cancers was demonstrated in a cohort of HPV+ OPSCC patients where ERAP1 allotype combinations associated with a low level of tumour CD8+/TILs had reduced ability to generate HPV-16 E7 epitopes from N-terminally extended peptides, highlighting the importance of investigating ERAP1 function in disease prognosis [121]. Although the size of the HPV+ OPSCC cohort was considerably smaller than that of the CC cohort (25 vs 81 patients) and the different geographic origin (UK vs Netherlands), this study by Reeves et al aimed to investigate the frequency and function of SNP combinations in the ERAP1 sequence in an HPV-driven cancer, hence making it appropriate for comparisons with the cervical carcinoma patient cohort. The identification of genetic variation in ERAP1 and the effect on the enzyme's ability to generate epitopes that will be presented to CD8+/TILs from N-terminally extended antigenic peptide precursors could be used as a tool to help women at increased risk of poor cervical cancer prognosis.

6.2 Clinical information on the cervical cancer patient cohort

Following the establishment of the methodological pipeline for long read sequencing, identification of ERAP1 from a cohort of 103 cervical cancer patient cDNA samples was attempted. Clinical information regarding type of cervical carcinoma, clinical disease stage (FIGO), lymph node and/or distant metastasis, recurrent disease and survival at the last-follow-up was provided for a total of 96/103 patients. There was a total of 58 CSCC (60.4%), 32 adenocarcinoma (33.3%), 2 adenosquamous carcinoma (2.1%) and 4 non-disclosed carcinoma (4.2%) patients. In the study by Mehta et al mentioned earlier in this chapter, the most prevalent type of carcinoma was the squamous cell

carcinoma with 76.1% of patients from the cohort investigated suffering from that type. The relevant percentage for the study by Li et al was 55.5% [182]. Squamous cell carcinoma prevalence seems to be independent of ethnic origin as the current study and the study by Mehta et al used cases of European origin, while the study by Li et al used cases of Chinese Han origin. It is worth noting that normal ERAP1 expression was observed in tissues from both squamous cell carcinoma and adenocarcinoma patients using immunohistochemical analysis [177]. The HPV-type that patients were infected with (only available in 27 patients) revealed that most were either infected with HPV-16 only or together with HPV-18 or other types (22/27). Five patients were infected with only HPV-18, and it would be important to compare the ability to generate HPV-18 E6/E7 epitopes by the ERAP1 allotypes expressed in these patients to compare with those from HPV-16-infected individuals.

6.3 Long read sequencing for ERAP1 allotype identification from a cervical cancer patient cohort

6.3.1 Technical and optimisation discussion

To determine the ERAP1 allotype combinations in a cohort of patients with varying degrees of cervical carcinoma, long read sequencing was used (MinION, ONT). MinION was selected as the preferred sequencing platform as it can generate full length reads of ERAP1 in real time. However, given that the MinION has not previously been used for ERAP1 allotyping, a methodological pipeline, as well as a bioinformatics analysis pipeline, was first established.

A significant advantage of long read sequencing is that it enables sequencing of multiple amplicons in the same run by barcoding each sample [188]. In this study, up to 12 barcoded ERAP1 amplicons were pooled together as one library, and ERAP1 allotypes could be accurately identified, while also minimising the cost per sample at the same time. An interesting observation was that a higher number

of reads was generated for one of the two barcodes, even though both amplicons were amplified for the same number of PCR cycles (35 in total) and this finding has been confirmed by other researchers [221]. The observation could be attributed to inherent inaccuracies of instruments to accurately measure DNA concentration quantification. Long read sequencing of a HeLa ERAP1 amplicon prepared using 35 PCR cycles, lead to successful ERAP1 allotyping with as low as 18 reads.

One of the advantages of long read sequencing is the use of low amplicon concentration for library preparation without the need to carry out an additional amplification reaction. The sequencing of a library that had been prepared using four barcoded samples amplified for a different number of PCR cycles revealed that the samples with higher DNA concentration (ERAP1 amplified using 25 and 35 PCR cycles) were dominant over those with lower DNA concentration (ERAP1 amplified using 5 and 15 PCR cycles), with the allotypes of the latter two low-cycle amplicons not being identified. The most likely explanation is the problematic quantification of those lower-cycle amplicons with both Nanodrop and Qubit instruments, which could inherently fail to accurately measure the concentration of low-PCR cycle amplified samples.

Regarding the cervical cancer patient cohort, up to 12 barcoded ERAP1 amplicons from patient cDNA were sequenced together in the same run to generate a high throughput pipeline. If long read sequencing becomes the standard method of sequencing ERAP1 from large cohorts of cancer patient samples, and possibly even from samples of patients suffering from other autoimmune and inflammatory conditions in which ERAP1 plays a role such as ankylosing spondylitis, sequencing conditions should be optimised.

It was shown that Qubit did not accurately measure the concentration of ERAP1 amplicons resulting from fewer than 35 PCR cycles due to insufficient ERAP1 DNA and the concentration reported for those was that of what turned out to be primer dimers (UGENE and nucleotide BLAST, NCBI). Consequently, ERAP1 from patient cDNA was amplified for 35 PCR cycles to maximise the chances of

having sufficient amounts of DNA to sequence. Primer re-design included elongating their sequence within ERAP1 to enable more specific binding to the appropriate ERAP1 regions for increasing ERAP1 copies. DNA quantification by Nanodrop and later Qubit (after patient sample S24) was more accurate and no samples would be dominating over the others in the run due to concentration issues. Samples were sequenced in a second run if the first one did not result in successful ERAP1 allotype identification. A second PCR using the product of the first PCR would likely enable successful ERAP1 amplification, the concentration of which would be more accurately measured with Qubit. Direct RNA or cDNA sequencing would alleviate the issue of introducing PCR bias, however at the time that long read sequencing was performed in the present study, multiple cDNA samples could not be sequenced simultaneously with MinION which would increase both the cost per samples and it would also be a time- and effort- consuming procedure. In a review by Klasberg et al, PCR bias was analysed with one of them being PCR-mediated formation of crossover products depending on PCR cycles and initial template concentrations and allelic dropout due to low DNA quality and poor sample integrity [222-224]. The same review indicated that one of the issues with Sanger sequencing is the inherent phasing ambiguities which result in difficulty in putting heterozygous positions in phase.

Due to lack of genomic DNA for these patient samples, verification of true homozygotes was not possible. To try to validate the homozygosity of samples, repeated sequencing of samples revealed that both sequencing experiments identified the same homozygotic combination which increased confidence in the accuracy of long read sequencing at identifying homozygotic ERAP1 allotype combinations. Data from a study sequencing a cell line genome revealed PacBio and Illumina genotypes agreed at 94% of heterozygous alternate deletions but at only 79% of homozygous alternate deletions. Nanopore and Illumina genotypes agreed at 90% for both heterozygous and homozygous alternate deletions [186].

To further compare Sanger sequencing with long read sequencing, and to verify the same ERAP1 allotype combinations are identified, three ERAP1 amplicons from patients chosen at random (S70,

S71 and S101) were subjected to both Sanger and long read sequencing. Results confirmed the accuracy of ERAP1 allotyping with MinION. Interestingly, a study using long read sequencing for HLA-B typing from Māori and Pacific island New Zealand populations have confirmed accuracy of the technology through data comparisons with Sanger sequencing [188]. This study involving HLA-B typing is applicable because amplicons were used for long read sequencing, as in the present study, and for identifying alleles also of a highly polymorphic gene, HLA-B. In addition, in the HLA-B typing study mentioned above, long read sequencing data generated with MinION were compared with Sanger sequencing data for 9 controls and here the latter sequencing method was also used to verify the accuracy of the technology at identifying ERAP1 allotypes from two cell lines and also a number of randomly chosen patient samples. Importantly, HLA-B alleles were identified with a read count of just 80 and a low concentration of amplicons was used for barcoding (0.19nM contrary to 2nM recommended by ONT).

6.3.2 Four separate ERAP1 allotypes were identified from 293T and HeLa using long read sequencing and confirmed with Sanger sequencing

The allotypes identified in HeLa through Sanger sequencing are consistent with those previously identified (Dr Emma Reeves, unpublished data), verifying that the methodology used resulted in reproducible data. The identified allotypes were later compared with those identified with long read sequencing to confirm the instrument's accuracy at ERAP1 allotyping. The two ERAP1 allotypes identified in HeLa cells from cloning and Sanger sequencing were *013 (P127) and *020 (P127/D346/R528/E730).

Initially, in this study, the only ERAP1 allotype identified from 293T cells using Sanger sequencing was *015 (P127/R528/E730) which was consistent with one of the allotypes identified previously within the James lab (Dr Emma Reeves, unpublished data). However, ERAP1 allotype identification using long

read sequencing revealed two distinct ERAP1 allotypes in 293T cells, *015 and *021 (also referred to as haplotype 8 [120]), highlighting the ability of long read sequencing to identify ERAP1 allotypes using low ERAP1 amplicon concentrations. The *015 and *021 allotypes identified using long read sequencing match both allotypes described before for this cell line (Dr Emma Reeves, unpublished data). Allotype *021 contains 4 amino acid changes, P127/ I276M/R528/E730, three of which are also found in the *015 allotype (not M276).

6.3.3 Six novel ERAP1 allotypes were identified from the cervical cancer patient cohort

Sequencing of ERAP1 allotypes from this cohort revealed a total of 14 distinct ERAP1 allotypes, which were found in 28 distinct combinations. Of these allotype combinations, 22 were heterozygotes and 6 were found to be possible homozygotes. The homozygotic allotype pairs were *001, *002-WT, *013, *015, *018 and *021. The issues associated with verifying homozygotic allotype pairs and how these were addressed in this study was explained in section 6.3.1. Genomic sequencing would enable sequencing of intros as well as exons to verify true homozygotes.

Six novel ERAP1 allotypes were identified from the patient cohort and, consistent with previous ERAP1 allotype nomenclature published by Reeves et al [89], were given the allotype identities *023-*028. The most frequent novel allotype within the cohort was allotype *024, which contains the amino acid changes P127/R528, and was found in three different patients, S3, S30 and S82, in two combinations, including *023 + *024 from S3 and S30 and *024 + *026 from S82. Interestingly, the amino acid changes in allotype *024 were also identified from 1.3% of Javanese cervical cancer cases vs 0.6% of patients in this study, however no significant difference in frequency was observed between Javanese cases and controls [215]. This allotype was not identified in Balinese cervical cancer cases, hinting the

factor of different ethnic origin (Javanese vs Balinese). That study investigated the presence of only six amino acid changes, while in the present study nine amino acid positions were investigated.

Since allotype pair *023 + *024 was identified from both patient S3 and S30, it increased confidence in the accuracy of ERAP1 allotyping using long read sequencing as it was not a single observation and importantly, it is possible that there is an association between this allotype pair and poor cervical cancer prognosis. The ERAP1 amplicons prepared from these three patients were not sequenced in the same run and therefore it is unlikely that the allotypes were falsely identified from other samples they were sequenced with, which is also unlikely given that amplicons had different barcodes attached to them. It was predicted that the allotype pair *024 + *026 would have a hypotrimming phenotype given the amino acid changes present. However, as this was just a hypothesized trimming phenotype, it was important to undertake functional studies on this allotype and its combinations.

Allotype *025 containing the amino acid changes P127/I276M/G346D/R528/E730 is also expected to have reduced trimming function as *001 that contains all of the above amino acid changes as well as M348V and R725Q [89, 90]. Given that *025 was found in combination with allotype *013 that only contains the amino acid change P127 that has not been shown to have an effect on the enzyme's function, the overall phenotype of the enzyme was expected to be efficient as studies have shown before that an allotype with an efficient trimming function can rescue the enzyme's trimming function when in combination with a hypotrimming phenotype [89, 90]. This is also a functional investigation that was of interest for another novel ERAP1 allotype, *028 (P127/E730), that was found in combination with the wild type allotype, *002-WT. Allotype *028 was identified from a cohort of Balinese cervical cancer cases with frequency of 32.4% vs 0.6% for the cohort in this study, and the frequency was similar between Balinese cases and controls [215]. As mentioned already, the difference could be attributed to different ethnic origin of the patients (CEU vs ASN) and also that study investigated a limited set of six SNPs.

Allotype *026 contains all the amino acid changes found in *001 mentioned above but not R528. As this SNP has been shown to have a dominant effect on the enzyme's trimming function, it was of interest to investigate the trimming of the allotype in combination with *024 as identified from S82. Last but not least, *027 (P127/I276M/R528) was identified from S103 in combination with *018 that only contains the amino acid change E730. This combination is of interest because it was shown that the SNP frequency distribution for E730 and interestingly also allotype *018 was similar between the cervical cancer patient cohort and the HPV+ OPSCC patient cohort that was used for a study by Reeves et al [121]. It is noteworthy that allotype *027 was identified in another study as well published last year, although this allotype was identified in 2020 and hence the nomenclature is as presented [123].

The most common ERAP1 allotypes in the patient cohort were: *001, *002-WT, *013, *015, *018 and *021 with frequency between 10-20% in the cohort (n=162). These were found in 15 distinct combinations (combinations consisting of two of the aforementioned allotypes). Interestingly, all of these allotypes were also found in potentially homozygotic combinations but due to lack of genomic DNA this cannot be confirmed.

Comparisons of allotype frequencies between the cervical cancer patient cohort in this study, and the CEU (n=80) and ASN (n=80) Hapmap populations reported by Ombrello et al, revealed that allotype *013 (Hap1) was identified in 12% of CEU individuals and it was absent in the ASN population [120]. In contrast, allotype *013 was one of the most prevalent in this study with the allotype identified in 27.2% of cervical cancer patients. The percentage discrepancy could indicate a link between increased cervical cancer risk, and specifically cancer in patients of CEU origin, and allotype *013 as interestingly, the amino acid change P127 was also determined to be associated with increased cervical cancer risk in the relevant study by Mehta et al. P127 was also found associated with psoriasis patients after puberty (10-20 years old) and was shown to be independent of the HLA-C*06 allele [138]. This finding is a contrast to another study showing the ERAP1 association with psoriasis being dependent on the HLA-C*06 allele [225]. The difference in findings could be related to different sample sizes in the two

studies above. The study by Strange et al investigating association of ERAP and HLA-C*06 in 2,622 patients with psoriasis and 5,667 controls. P127 has also been associated with protection against ankylosing spondylitis by Harvey et al [218]. P127 is located in domain I where the end of the S1 specificity pocket of the catalytic domain borders, which indicates that this amino acid change may have an effect on the formation of the active site [102, 103]. The S1 specificity pocket is adjacent to the active site, and it is involved in determining the specificity of ERAP1 for the free N-termini of potential substrates and it is also where the GAMEN substrate-binding motif is located [226]. In a study by Mehta et al, heterozygosity of K56 and minor allele homozygosity at P127 were associated with decreased overall survival [190]. Regarding haplotypes, heterozygosity of a major allele at E56 and a minor allele at P127 was significantly associated with normal ERAP1 expression and better overall cervical cancer survival [190]. Previous findings by the same group indicated that P127 was significantly associated with cervical carcinoma risk pointing towards P127 being the major factor affecting cervical cancer risk and survival [190]. The frequency of *002-WT in the patient cohort was closer to the frequency of the allotype in the ASN population rather than the CEU one (43.7 vs 13.7%). The frequency of *018 was also considerably higher in the patient cohort compared to both the ASN and the CEU populations (22.2 vs 2.5 and 5.6%, respectively). More than double the frequency of allotype *015 in CEU cohort was identified from the present cervical cancer patient cohort (18.5 vs 7.4%) [120]. The relevant percentage in ASN cohort was just 1.2%. Allotype *021 was identified in about 34.6% of patients compared to 21.9 and 20% of CEU and ASN populations, respectively. Both allotypes *021 and *015 contain the amino acid changes P127/R528/E730 which have previously been associated with reduced ERAP1 function (detailed in section 6.4). Interestingly, the frequency of allotype *001 was similar between the patient cohort and the CEU population, while only 5% of ASN individuals expressed this allotype. The similarity for this allotype could be attributed to the fact that the cervical cancer patients in this study were of European origin. Interestingly, the frequency of allotype *014 in the ASN population was approximately 20% higher than that in both this cohort and

the CEU population. Allotype *014 is the only one of the allotypes, the frequency of which was investigated, that contains K56.

The frequency of ERAP1 SNPs in the cervical cancer patient cohort of the present study was compared to that in the two cohorts (cervical cancer and control cohorts) in the study by Mehta et al [181]. Although recent evidence suggests it is the combination of SNPs present in ERAP1 allotypes that has the overall effect on trimming function, understanding the frequency of SNPs within the cohort will allow comparison with previous studies and GWAS susceptible SNPs.

Comparison of SNP frequency distribution revealed that frequencies were similar for all investigated SNPs (E56K, P127, I276M, N575), except for R528 and E730 [182]. Mehta et al revealed that homozygosity for P127 is associated with decreased overall and disease survival in cervical cancer [190]. R528 and E730 were identified in approximately 15 and 30% more cervical cancer patients in this study than in the patient cohort of the Mehta et al study, respectively (R528: 51.9 vs 36.7%, E730: 64.8 vs 33.5%). E730 has previously been associated with increased cervical cancer risk [181]. These SNPs have been shown to affect function of ERAP1 and could therefore have an effect on the enzyme's function and hence disease outcome. The frequency of E56K, P127, I276M and N575 was almost identical between both patient cohorts and controls, indicating that perhaps these were not important in the risk of cervical cancer development. In a study by *Li et al*, M276 was found to be associated with increased cervical cancer risk in a cohort of 2890 Chinese Han individuals, 556 CIN, 1072 cervical cancer, of which 151 adenocarcinoma, 903 CSCC and 18 of other types, compared to 1262 controls [182]. Mehta et al did not report any association between this amino acid change and cervical carcinoma which was likely due to the smaller cohort size (n=248 vs n=2890) and potentially the different ethnic origin of cohort individuals (CEU vs ASN).

Although both the patient and the control cohorts were larger in size compared to the cervical cancer patient cohort of the present study, (124 healthy controls and 127 cervical cancer patients vs 81

cervical cancer patients in the current study) and the fact that both patient cohorts originated from the same geographic origin (patient samples collected in the Netherlands), the study by Mehta et al did not take into consideration linkage disequilibrium or the fact that both ERAP1 copies are co-dominantly expressed and hence both ERAP1 copies have an effect on the enzyme's overall trimming function [181, 190]. In addition to these study drawbacks, the function of the identified SNPs was not tested and also only six SNPs were reported while the present study investigated the function of the identified allotypes containing any of nine distinct amino acid changes.

Comparison of SNP frequency distribution between the cervical cancer patient cohort in this study and the OPSCC cohort [121] revealed that P127 was found in approximately 40% more CC patients than OPSCC patients (n=113/162 vs n=16/50), which confirms data by the Mehta lab that P127 is significantly associated with cervical cancer risk and homozygosity with decreased survival [190]. The study in OPSCC does not specify the gender of the patients but a study showed that the risk of progressing to cervical cancer development in HPV+ OPSCC patients was considerably higher than that of the general population of Alberta, Canada [227]. Interestingly, the amino acid M276 was absent from the OPSCC cohort, while it was one of the most frequently identified amino acid changes from the CC cohort (22.8%), again confirming data by another lab which found association between M276 and cervical cancer risk, although the study was conducted in an ASN cohort [182]. Regarding the frequency of K56, D346, V349, N575 and Q725, it was similar in the two cohorts with less than 8% difference between them. This is significant as this SNP has not been associated with HPV-driven cancers before. R528 has been associated with decreased trimming of peptides and hence pMHC I are not formed to be presented to CD8+ T cells to exert their effector function [88, 102, 125, 128, 228]. A method of immunoescape used by HPV to avoid immune detection could potentially be through the reduced trimming of HPV-derived, N-terminally extended precursors to final epitopes [121]. The frequency of E730 in combination with other amino acid changes as well as frequency of allotype *018 which only contains E730 was similar between the two patient cohorts (CC: 64.8% vs OPSCC: 70%).

This difference in frequency of the above amino acid changes could be due to the small size of the OPSCC patient cohort. Mehta et al have also identified E730 with increased cervical cancer risk and it could be generalised to HPV-driven cancers [181]. E730 may have an effect on the enzymatic activity of ERAP1 depending on the substrate as shown by *in vitro* studies. More specifically, it was hypothesized that since E730 lies >27Å from the active site, in the regulatory domain IV of ERAP1, it does not have a direct effect on substrate catalysis but potentially makes contact with the C-terminal moiety of substrates [88, 229]. It has been shown that the E730 affects trimming due to the negative potential which in turn makes interaction between the regulatory domain of ERAP1 and the C-terminal moiety of the peptide worse and hence trimming is reduced [122, 230].

Regarding ERAP1 allotype combination frequencies, the frequencies reported for *002 homozygotic combination and *002 + *021 (allotype nomenclature by James lab) in the patient cohort were close to those reported in a study by Hutchinson et al in 2021 for the populations that were pooled together in that study, including investigation of allotype frequencies in the CEU population [123]. These matched the frequencies reported by Ombrello et al [120]. The findings reported by the researchers of the studies above increase confidence in the accuracy of ERAP1 allotyping completed with long read sequencing for individuals of European origin.

6.4 Assessing the trimming function of ERAP1 allotype combinations from the cervical cancer patient cohort

Whilst the trimming ability of each ERAP1 allotype is important, the chromosomal copies of ERAP1 are codominantly expressed and the strongest association between ERAP1 and disease was shown when the combined function of both ERAP1 allotypes was investigated [89].

A well-characterised T cell activation assay was used within the James lab to investigate ERAP1 trimming function *in vivo*. This assay has historically been used to classify ERAP1 allotypes in three

categories based on their ability to generate the modified OVA-derived epitope SIINFEHL (SHL8) from the N-terminally extended precursor AIVMK-SHL8 in 293TE1KO cells, as follows; hypotrimming, efficient and hypertrimming. The SHL8:H2-K^b complexes are detected at the surface of the APCs by the B3Z T cell hybridoma cells. It still remains unknown how different ERAP1 allotype pairs affect the HPV-derived E7 epitope repertoire in cervical cancer and this gap in knowledge was addressed here to establish a linkage between ERAP1 function and disease prognosis [121]. The ability of ERAP1 allotypes/pairs to generate the HPV-16 E7-derived epitope LLMGTLGIV (LV9) from the N-terminally extended precursor, ED-LV9, was investigated as it was done in the study investigating linkage of ERAP1 allotypes and function with the HPV-driven cancer OPSCC [121, 231]. LV9 was chosen because it was the only HPV-derived epitope that had an IC50 for HLA-A*0201 of less than 100nM and contained a natural N-terminal extension (Glu, Asp; ED-LV9) [121]. For the trimming of ED-LV9, trimming was investigated only by the ERAP1 allotypes/combinations identified from HLA-A*0201 positive patients for whom the CD8+/TIL status was available in order to make predictions regarding disease prognosis. The hypothesis tested was that genetic variation in ERAP1 likely affects the infiltration of CTLs in the tumour microenvironment through the varying trimming of HPV-derived epitopes, and hence the enzyme may be used to predict prognosis in cervical cancer.

Functional data from the present study showed that the combination of an ERAP1 allotype with a hypotrimming phenotype and an allotype with an efficient trimming phenotype, restored the overall activity of the enzyme. This finding confirmed previous data by Reeves et al which identified different allotype pairs from AS cases and controls, and those from the control cohort were efficient at generating the final epitope SHL8 from the N-terminally extended antigenic precursor X5-SHL8 as seen by the B3Z T cell hybridoma response recognising the final epitope bound to MHC I [89]. More specifically, when allotype *005 (R528), which was shown to have reduced ability to generate SHL8 from X5-SHL8, was found in combination with *013 that had an efficient trimming phenotype, overall trimming of the peptide was rescued [90, 124, 218]. *In vitro* studies have shown that R528 has reduced

trimming activity compared to K528, and when HeLa cells were transfected with *005 and an HLA-B*27 peptide precursor, the cells presented a reduced number of pMHC I at the cell surface compared to K528 [88, 102, 125, 128, 228]. Presence of R528 in ERAP1 has been shown to reduce the trimming ability of the enzyme by having an effect on the kinetic process involved in transition from open to closed conformation (inactive to active state of ERAP1) [122].

In this study, trimming of X5-SHL8 remained below 40% when two allotypes with a hypotrimming phenotype were found in a combination. R528/E730 seemed to have a dominant effect on reducing the enzyme's trimming ability which has been confirmed by previous literature findings but trimming is substrate specific [88]. Reeves et al showed that R528/E730 had a hypotrimming phenotype when the trimming of X5-SHL8 was investigated (~40% trimming activity compared to *002-WT), while X6-SHL8 was trimmed efficiently, revealing that R528/E730 is substrate-specific [89, 90].

Allotype *001 (P127/349V/528R/N575/725Q/E730) was shown to be the least efficient allotype in trimming X5-SHL8 to SHL8 and this is confirmed by historical data showing that *001 was 60-fold less active than other allotypes investigated [89, 123]. However, allotype *001 was able to trim ED-LV9 to LV9 more efficiently than AIVMK-SHL8 (BE7A2Z T cell hybridoma response: 67.2% vs B3Z T cell hybridoma response: 29.7%). This was also the case for allotype *015 (P127/R528/E730, BE7A2Z T cell hybridoma response: ~60% vs B3Z T cell hybridoma response: ~20%). Comparison of trimming efficiency of X-SHL8 and X-LV9 by the ERAP1 allotypes identified from the OPSCC cohort revealed that the N-1 amino acid (first amino acid of the N-terminal extension proceeding the peptide) affects trimming at a higher degree than the substrate sequence and this was confirmed by findings in other studies as well [90, 113, 121]. The top three N-1 specificities that enable the highest trimming for *001 were C, A, and E for X-SHL8, but E, K and N for X-LV9. The top three N-1 specificities for *015 were E, H and Q for X-SHL8, but E, H and K for X-LV9. These data indicate that the sequence of the substrate (here SHL8 and LV9) may play a role in N-1 amino acid specificity by the allotypes. In addition, it appears that the polymorphism in the ERAP1 and MHC I genes, both of which are part of the antigen

processing and presentation pathway, enables a variable immune response to take place within the population [123]. Trimming of ED-LV9 by *015 (CD8+/TIL^{mod} group) suggests that this indicates a better cancer prognosis because a higher number of peptides are trimmed to optimal length and presented to HPV-specific CD8+/TILs. The BE7A2Z response generated after the trimming of ED-LV9 to LV9 by the pair *001 + *021 was approximately 40% higher than the B3Z response, which was expected given the fact that the trimming activity of the individual allotypes was also higher for ED-LV9.

Allotype *019 was more efficient at generating LV9 than SHL8, confirming the substrate-specificity of R528/E730. The combination of *019 + *013 was identified from an OPSCC patient [121]. Since amino acid position 346 is found at the catalytic site, the amino acid change G to the negatively charged D may affect trimming G346D [104, 108]. G346D has been previously identified in a cohort of 48 Caucasian of UK origin AS patients but the involvement of this amino acid in cervical cancer prognosis has not been investigated before [218].

Both *002-WT and *013 were efficient at generating both SHL8 and LV9. [181]. There have not been any studies showing specific effect of P127 on the ERAP1 function but as position 127 lies on the substrate exit and the substitution with proline likely reducing flexibility at the mouth of the substrate exit. Formation of hydrogen bonds or polar interactions of P127 with the peptide side chains is likely, leading to problematic transition from open to closed conformation upon peptide binding and this would theoretically reduce the enzyme's trimming activity [102-104, 108, 123].

Since six novel allotypes were identified in this cohort, the trimming activity of these allotypes was tested to determine the effect of the novel SNP combinations on trimming function. It is noteworthy that the novel allotypes were not identified from HLA-A*0201 positive patients and hence the trimming of ED-LV9 to LV9 was not investigated. The T cell activation assay and western blot confirmed that the novel allotype *025 (P127/M276/D346/R528/E730) had reduced activity, as predicted from the amino acid changed R728 and E730 which have been shown to be less efficient at generating the

final epitope compared to *002-WT [90]. In combination with allotype *013, trimming was restored confirming historical data [89]. Allotype *027 (P127/I276M/R528), that was also identified in a study by Hutchinson et al, had reduced trimming ability to generate SHL8 which was restored when in combination with *002-WT but not in combination with *018 [123]. It can be speculated that the presence of R528 and E730, even though as part of different allotypes, had a negative effect on the overall trimming phenotype of the enzyme. Allotypes *023 (P127/M276/E730) and *028 (P127/E730) had a hypotrimming phenotype, similarly to allotype *018 that contains only the amino acid change E730. Allotype *002-WT was able to rescue the enzyme's overall trimming ability when trimming was investigated in combination with *028. Allotype *026 (P127/V349/N575/Q725/E730) could be have a hypertrimming phenotype due to the presence of Q725/E730 as with previous study findings [90]. Trimming of X5-SHL8 by allotype *024 (P127/R528) generated an unusually high B3Z T cell hybridoma response (68.2%) given that R528 has been shown to reduce trimming ability [88, 125]. Both positions 127 and 528 in ERAP1 are at domain junctions and not within the cavity, therefore they are unlikely to participate in interactions with the substrate. They likely affect trimming or substrate specificity by altering the conformational change between 'open' and 'closed' states [103]. It is likely that P127 and R528 are working together to change the conformation of the enzyme upon substrate binding, leading to better trimming of X5-SHL8. Combination of *024 + *026 generated a B3Z response of 56% compared to *002-WT.

Regarding T12I, it was included in the data analysis, however it was not identified in the patient cohort. As mentioned in another study investigating the same ERAP1 SNPs in the ERAP1 sequence, position 12 lies in the signal peptide that once ERAP1 is translocated into the ER is excised and hence it was not included in the present study [123].

In HPV^+ CSCC, a high number of TILs has been associated with increased survival and it is likely that tumour rejection occurs through anti-tumour immune responses exerted by CTLs recognising HPV E6/E7-derived epitopes [166, 170, 179]. Regarding another HPV-driven cancer, OPSCC, it was shown

that allotypes from patients classified in the CD8+/TIL^{high} group were more efficient at generating LV9 from its precursor and this could be associated with better overall prognosis [121].

HLA binding studies have predicted epitopes presented to CTLs from HPV-16 E6 and E7 proteins and these are restricted by HLA-A molecules, importantly HLA-A*0201 [232]. Other studies have identified HLA-A*0201-restricted HPV-16/18 E5/E6/E7-derived epitopes that are presented to CTLs in in HPV-driven cancers and neoplasias [121, 205, 233, 234]. The majority of HPV-16 E6/E7-derived epitopes are restricted by HLA-A*0201, the most frequent MHC I allele in the Caucasian population. A study identified an HPV-16 E6-derived epitope that is restricted by HLA-A*2402, a frequently identified MHC I allele in the Japanese population (~60%) as well as other Asian population with the aim to develop immunotherapy approaches for HLA-A*2402 positive patients [235]. As the cervical cancer patient cohort of the present study is of CEU cohort, the trimming of an HPV-derived epitope restricted by HLA-A*0201 was investigated with HLA-A*0201 positive patients identified first and then investigate the trimming of the epitope by their allotypes/pairs. This provided an opportunity to examine how the function of ERAP1 impacts the peptide repertoire binding to HLA-A*0201 forming complexes that are presented to CTLs in cervical cancer and hence, making predictions regarding disease prognosis.

The CD8+/TIL status of patient tumours was only available in 25/39 patients. Analysis showed that the majority of the allotype combinations in the CD8+/TIL^{high} group had a normal trimming phenotype and interestingly, the two patients who experienced lymph node metastasis were classed as disease-free at the last follow-up. This could be due to the normal trimming phenotype of their ERAP1 allotype combination which was able to trim N-terminally extended HPV-derived epitopes such as ED-LV9 to final epitopes presented to HPV-specific T cells. However, the trimming of a single HPV-16-derived epitope was investigated in this study, and it is possible that there are other HPV-derived epitopes that may not be necessarily restricted by HLA-A*0201. This could be the case for patients S47 and S70 who were infected with both HPV-16 and HPV-18 and even though their allotype pairs had an overall hypotrimming phenotype towards X5-SHL8 (*001 + *014 and *001 + *015, respectively), they were

disease-free at the last follow-up. It is possible that other HPV-18 derived epitopes might have been presented to CD8+/TILs which were activated and exerted their effector function. Tumour samples were not available and therefore ERAP1 protein expression could not be investigated. It has been shown that altering the expression of an allotype with efficient trimming phenotype, *002-WT, when in combination with a hypotrimmer, *001, can negatively affect tumour epitope generation and hence reduce CD8+ T cell activation which exert their effector function [121]. In both the present study and the study in OPSCC patients, there were some ERAP1 allotype combinations that were only found in one of the three CD8+/TIL groups [121]. In the present study, one of the allotype combinations in the CD8+/TIL^{low} group had a hypotrimming phenotype when the trimming of ED-LV9 was investigated, however both patients with this combination were disease-free at the last follow-up indicating that other factors could have contributed to disease prognosis.

In conclusion, this study used long read sequencing to sequence ERAP1 from a cohort of 81 cervical cancer patients. Trial sequencing runs with MinION and Sanger sequencing confirmed the suitability of this method of sequencing for ERAP1 allotyping. There was a total of 28 combinations, 22 heterozygotes and 6 homozygotes. The function of the identified combinations was investigated using a model epitope and for those HLA-A*0201 positive patients for whom CD8+/TIL data were available, the trimming of an HPV-derived by their allotype combinations was also investigated. Comparisons of frequencies with CEU population and another cervical cancer patient cohort revealed that the patients in the present cohort have more amino acid changes than the population which could have an effect on cervical cancer risk and P127 was found with particularly high frequency in the cohort, previously associated with increased cervical cancer risk [190]. HLA-A*0201 patients with high CD8+/TILs had better overall prognosis and the majority of the patients in that group had allotype combinations that are efficient at generating an HPV-derived epitope. However, the majority of patients in the CD8+/TIL^{low} group (~70%, 17/22 in CD8+/TIL^{low} in cohort of 96 patients) were alive at the last follow-up indicating that other factors could have played a role in the good prognosis apart from the CD8+/TIL

number. Such factors include the presence of other cells apart from CD8+ T cells in the TME such as CD3-CD57+ natural killer cell-like cells and a novel T cell subset, Th9, which can inhibit cervical cancer progression and immune evasion [166, 236]. In addition, even though some patients were in either the CD8+/TIL^{high} or CD8+/TIL^{mod} category, they had poor prognosis with a possible explanation being that cells such as Tregs were present in TME inhibiting antitumour immune responses. A particular type of Tregs (HLA-DR Tregs) was found to be associated with poor prognosis in patients with squamous cell carcinoma [237]. For future studies, it would be useful to investigate the trimming of other HPV-16-derived epitopes as well as HPV-18-derived epitopes for patients that were affected with both types in the cohort, to get a better understanding of how cervical cancer prognosis can be affected by the trimming phenotype of ERAP1 allotype combinations [149].

6.5 ERAP1 expression in cervical cancer

In 15% of cervical cancer cases downregulation of ERAP1 expression was found to be an independent predictor of decreased overall and disease-free survival of cervical carcinoma [238]. The study used tissue samples from 26 cervical adenocarcinoma and 83 CSCC patients. In a study by Hasim et al, partial and total loss of ERAP1 expression at protein level was observed in 31.7% and in 20.6% of CSCC patient samples, respectively [239]. Loss of ERAP1 protein expression was statistically significant for CSCC compared to CIN patients/controls. The findings of the studies above are conflicted by another study by Steinbach et al that involved immunohistochemical analysis of ERAP1 in 10 CSCC patient tissue sections and showed high ERAP1 epithelial expression in 80% of them [240]. In the latter study, the sequence of the ERAP1 allotypes was unknown, therefore the effect of SNPs on ERAP1 expression could not be determined. Knockdown of ERAP1 expression in the HPV-16 positive cervical cancer cell line CaSki followed by cytotoxicity assays revealed enhanced killing of cells with attenuated ERAP1 expression by HPV-16 $E7_{81-91}$-specific T cells (LV9-specific T cells) [177, 240]. However, that was not the outcome for the cytotoxicity assay involving another cervical cancer cell line, 866, pointing

towards SNPs in the ERAP1 sequence as one of the likely explanations for the discrepancy between the two assays [240]. The ERAP1 expressed by CaSki cells has been shown to be an efficient trimmer but it overtrims the peptide investigated in that study revealing the role of ERAP1 in trimming of HPV-derived epitopes presented to HPV-specific T cells [240]. The difference in protein expression of ERAP1 between the Steinbach et al study and the studies by Mehta et al and Hasim et al was attributed to variable ERAP1 expression in a single tumour sample and use of different experimental methods and controls [238-240]. It is noteworthy that according to the Human Portein Atlas, ERAP1 is upregulated in cervical cancer [241]. The studies above reveal that ERAP1 may be an important target for identifying those women at 'high-risk' of poor cervical cancer prognosis.

It has been consistently reported that K528 is associated with higher ERAP1 expression in EBV-transformed cell lines [132, 218]. This finding was confirmed by another study in which two HLA-B*27:05 lymphoblastoid cell lines had ~2-fold and ~3-fold reduced expression of the ERAP1 variant carrying the SNP rs30187 encoding the amino acid change R528 [132, 220]. In tumour cells, ERAP1 expression has been shown to vary and can be both increased and decreased. ERAP1 destroys the tumour antigen GSW11 in the murine colorectal carcinoma model CT26 and inhibition of ERAP1 activity resulted in ~75-fold increase in GSW11 presentation, which led to a CTL-mediated anti-tumour immune response [81]. Similarly, an increase in ERAP1 expression in human colorectal adenocarcinoma compared with healthy tissue was observed [100]. Therefore, it is likely that by altering ERAP1 expression, tumour cells are able escape immunosurveillance through the destruction of tumour epitopes. Regarding HPV-driven cancers, downregulation of expression of ERAP1 and other components of the APP pathway (TAP1, LMP7) as well as partial loss of HLA I, has been associated with decreased OS and DFS in cervical carcinoma with ERAP1 specifically being an independent prognostic marker of shorter survival (OS and DFS) [177]. Furthermore, loss of ERAP1 expression occurred at pre-transcriptional level accompanied by a loss of heterozygosity and genetic variation at the 127 amino acid position [238]. Downregulation of ERAP1 expression may enable immune evasion

of tumour cells through prevention or reduction of generation of immunogenic epitopes. This study therefore aimed to investigate the contribution of ERAP1 in cervical carcinoma by identifying the ERAP1 allotypes present in the patient cohort, and the subsequent analysis of trimming function towards a model peptide epitope (SHL8) and the HPV-derived epitope (LV9) that has been proposed to enter the ER as an N-terminally extended precursor. This would allow assessment of the importance of antigen processing in generating HPV peptides for immune recognition in cervical carcinoma.

6.6 Future project directions

A study by Mehta et al, revealed that the simultaneous presence of the ERAP1 haplotype P127/E730 and a major allele at TAP2-651 and LMP7-145 loci was associated with three-fold increase in cervical carcinoma risk [181]. In the future, long read sequencing could be used to sequence TAP2 and LMP7 from the cervical cancer patient cohort of this study to investigate association of genetic variation at these loci with cervical cancer prognosis. In another study by the same lab, partial loss of LMP2 and/or LMP7 was significantly associated with absence of detectable lymph node metastases in cervical cancer patients [177]. Partial loss of HLA-I or total loss of TAP1, were also significantly associated with decreased overall survival in cervical cancer. Although it would be of great interest to investigate association of loss of protein expression of ERAP1 and other APP components with cervical cancer prognosis, tumour samples were not available for any of the patients studied here which limits the acquisition of further knowledge of the role of ERAP1 and the MHC I APP pathway in cervical cancer outcome.

Regarding the role of genetic variation in ERAP2 in cervical cancer, a study by Chuanyin et al established a correlation between the SNP rs2287988 in ERAP2 as well as rs26653 (P127) and rs27044 (E730) in ERAP1 with cervical cancer, indicating that both ERAP enzymes could play a role in disease susceptibility. A limitation of the study by Chuanyin et al is the lack of HPV screening which would increase confidence in the correlation between SNPs in ERAP1/2 and cervical cancer [242]. Developing

a method for ERAP2 sequencing using the cervical cancer cDNA samples available from this cohort would enable investigation of a possible combined association of ERAP1 and ERAP2 with disease prognosis. In addition, a systematic approach, that is carrying out whole genome sequencing with nanopore sequencing, could lead to identifying additional factors influencing cervical cancer prognosis.

7 References

1. Murphy K, Travers P, Walport M: **Janeway's Immunobiology**, Seventh edn. New York and London: Garland Science

Taylor & Francis Group 2008.

2. Stratikos E, Stamogiannos A, Zervoudi E, Fruci D: **A role for naturally occurring alleles of endoplasmic reticulum aminopeptidases in tumor immunity and cancer pre-disposition**. *Front Oncol* 2014, **4**:363.

3. McDermott DF, Atkins MB: **PD-1 as a potential target in cancer therapy**. *Cancer Med* 2013, **2**(5):662-673.

4. Liu Y, Wu L, Tong R, Yang F, Yin L, Li M, You L, Xue J, Lu Y: **PD-1/PD-L1 Inhibitors in Cervical Cancer**. *Front Pharmacol* 2019, **10**:65.

5. Zhao P, Wang P, Dong S, Zhou Z, Cao Y, Yagita H, He X, Zheng SG, Fisher SJ, Fujinami RS *et al*: **Depletion of PD-1-positive cells ameliorates autoimmune disease**. *Nat Biomed Eng* 2019, **3**(4):292-305.

6. Gruen JR, Weissman SM: **Human MHC class III and IV genes and disease associations**. *Front Biosci* 2001, **6**:D960-972.

7. Xie T, Rowen L, Aguado B, Ahearn ME, Madan A, Qin S, Campbell RD, Hood L: **Analysis of the gene-dense major histocompatibility complex class III region and its comparison to mouse**. *Genome Res* 2003, **13**(12):2621-2636.

8. Blum JS, Wearsch PA, Cresswell P: **Pathways of antigen processing**. *Annu Rev Immunol* 2013, **31**:443-473.

9. Murphy K, Travers P, Walport M: **Janeway's Immunobiology**. NY and London: Garland Science.

10. Nesminayov P: **Antigen processing and presentation**. In. https://www.immunology.org/public-information/bitesized-immunology/sistemas-y-procesos/antigen-processing-and-presentation: British Society of Immunology.

11. K. M, Travers P, Walport M: **Janeway's Immunobiology**, 7th edition edn. New York and London: Garland Science.

12. Rudensky A, Preston-Hurlburt P, Hong SC, Barlow A, Janeway CA, Jr.: **Sequence analysis of peptides bound to MHC class II molecules**. *Nature* 1991, **353**(6345):622-627.

13. Roche PA, Cresswell P: **Invariant chain association with HLA-DR molecules inhibits immunogenic peptide binding**. *Nature* 1990, **345**(6276):615-618.

14. Stumptner P, Benaroch P: **Interaction of MHC class II molecules with the invariant chain: role of the invariant chain (81-90) region**. *EMBO J* 1997, **16**(19):5807-5818.

15. Neefjes JJ, Stollorz V, Peters PJ, Geuze HJ, Ploegh HL: **The biosynthetic pathway of MHC class II but not class I molecules intersects the endocytic route**. *Cell* 1990, **61**(1):171-183.

16. Landsverk OJ, Bakke O, Gregers TF: **MHC II and the endocytic pathway: regulation by invariant chain**. *Scand J Immunol* 2009, **70**(3):184-193.

17. Roche PA, Teletski CL, Stang E, Bakke O, Long EO: **Cell surface HLA-DR-invariant chain complexes are targeted to endosomes by rapid internalization**. *Proc Natl Acad Sci U S A* 1993, **90**(18):8581-8585.

18. Nakagawa T, Roth W, Wong P, Nelson A, Farr A, Deussing J, Villadangos JA, Ploegh H, Peters C, Rudensky AY: **Cathepsin L: critical role in Ii degradation and CD4 T cell selection in the thymus**. *Science* 1998, **280**(5362):450-453.

19. Morris P, Shaman J, Attaya M, Amaya M, Goodman S, Bergman C, Monaco JJ, Mellins E: **An essential role for HLA-DM in antigen presentation by class II major histocompatibility molecules**. *Nature* 1994, **368**(6471):551-554.

20. Denzin LK, Cresswell P: **HLA-DM induces CLIP dissociation from MHC class II alpha beta dimers and facilitates peptide loading**. *Cell* 1995, **82**(1):155-165.

21. Denzin LK, Sant'Angelo DB, Hammond C, Surman MJ, Cresswell P: **Negative regulation by HLA-DO of MHC class II-restricted antigen processing**. *Science* 1997, **278**(5335):106-109.

22. Nanaware PP, Jurewicz MM, Leszyk JD, Shaffer SA, Stern LJ: **HLA-DO Modulates the Diversity of the MHC-II Self-peptidome**. *Mol Cell Proteomics* 2019, **18**(3):490-503.

23. Heath WR, Carbone FR: **Cross-presentation in viral immunity and self-tolerance**. *Nat Rev Immunol* 2001, **1**(2):126-134.

24. Huang AY, Golumbek P, Ahmadzadeh M, Jaffee E, Pardoll D, Levitsky H: **Role of bone marrow-derived cells in presenting MHC class I-restricted tumor antigens**. *Science* 1994, **264**(5161):961-965.

25. Embgenbroich M, Burgdorf S: **Current Concepts of Antigen Cross-Presentation**. *Front Immunol* 2018, **9**:1643.

26. Joffre OP, Segura E, Savina A, Amigorena S: **Cross-presentation by dendritic cells**. *Nat Rev Immunol* 2012, **12**(8):557-569.

27. Shortman K, Heath WR: **The CD8+ dendritic cell subset**. *Immunol Rev* 2010, **234**(1):18-31.

28. Shen L, Sigal LJ, Boes M, Rock KL: **Important role of cathepsin S in generating peptides for TAP-independent MHC class I crosspresentation in vivo**. *Immunity* 2004, **21**(2):155-165.

29. Li Y, Wang LX, Pang P, Cui Z, Aung S, Haley D, Fox BA, Urba WJ, Hu HM: **Tumor-derived autophagosome vaccine: mechanism of cross-presentation and therapeutic efficacy**. *Clin Cancer Res* 2011, **17**(22):7047-7057.

30. Guermonprez P, Saveanu L, Kleijmeer M, Davoust J, Van Endert P, Amigorena S: **ER-phagosome fusion defines an MHC class I cross-presentation compartment in dendritic cells**. *Nature* 2003, **425**(6956):397-402.

31. Reeves E, James E: **Antigen processing and immune regulation in the response to tumours**. *Immunology* 2017, **150**(1):16-24.

32. Murphy K, Weaver C: **Janeway's Immunobiology**, 9th edn. New York and London: Garland Science, Taylor and Francis group; 2017.

33. Cresswell P, Bangia N, Dick T, Diedrich G: **The nature of the MHC class I peptide loading complex**. *Immunol Rev* 1999, **172**:21-28.

34. Evnouchidou I, Weimershaus M, Saveanu L, van Endert P: **ERAP1-ERAP2 dimerization increases peptide-trimming efficiency**. *J Immunol* 2014, **193**(2):901-908.

35. Saric T, Chang SC, Hattori A, York IA, Markant S, Rock KL, Tsujimoto M, Goldberg AL: **An IFN-gamma-induced aminopeptidase in the ER, ERAP1, trims precursors to MHC class I-presented peptides**. *Nat Immunol* 2002, **3**(12):1169-1176.

36. York IA, Chang SC, Saric T, Keys JA, Favreau JM, Goldberg AL, Rock KL: **The ER aminopeptidase ERAP1 enhances or limits antigen presentation by trimming epitopes to 8-9 residues**. *Nat Immunol* 2002, **3**(12):1177-1184.

37. York IA, Brehm MA, Zendzian S, Towne CF, Rock KL: **Endoplasmic reticulum aminopeptidase 1 (ERAP1) trims MHC class I-presented peptides in vivo and plays an important role in immunodominance**. *Proc Natl Acad Sci U S A* 2006, **103**(24):9202-9207.

38. Reeves E, Islam Y, James E: **ERAP1: a potential therapeutic target for a myriad of diseases**. *Expert Opin Ther Targets* 2020, **24**(6):535-544.

39. Goldberg AL, Cascio P, Saric T, Rock KL: **The importance of the proteasome and subsequent proteolytic steps in the generation of antigenic peptides**. *Mol Immunol* 2002, **39**(3-4):147-164.

40. Wang J, Maldonado MA: **The ubiquitin-proteasome system and its role in inflammatory and autoimmune diseases**. *Cell Mol Immunol* 2006, **3**(4):255-261.

41. Lam YA, Lawson TG, Velayutham M, Zweier JL, Pickart CM: **A proteasomal ATPase subunit recognizes the polyubiquitin degradation signal**. *Nature* 2002, **416**(6882):763-767.

42. Craiu A, Akopian T, Goldberg A, Rock KL: **Two distinct proteolytic processes in the generation of a major histocompatibility complex class I-presented peptide**. *Proc Natl Acad Sci U S A* 1997, **94**(20):10850-10855.

43. Schwarz K, de Giuli R, Schmidtke G, Kostka S, van den Broek M, Kim KB, Crews CM, Kraft R, Groettrup M: **The selective proteasome inhibitors lactacystin and epoxomicin can be used to either up- or down-regulate antigen presentation at nontoxic doses**. *J Immunol* 2000, **164**(12):6147-6157.

44. Cresswell P, Ackerman AL, Giodini A, Peaper DR, Wearsch PA: **Mechanisms of MHC class I-restricted antigen processing and cross-presentation**. *Immunol Rev* 2005, **207**:145-157.

45. Rock KL, Gramm C, Rothstein L, Clark K, Stein R, Dick L, Hwang D, Goldberg AL: **Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules**. *Cell* 1994, **78**(5):761-771.

46. Cascio P, Hilton C, Kisselev AF, Rock KL, Goldberg AL: **26S proteasomes and immunoproteasomes produce mainly N-extended versions of an antigenic peptide**. *EMBO J* 2001, **20**(10):2357-2366.

47. Kisselev AF, Akopian TN, Woo KM, Goldberg AL: **The sizes of peptides generated from protein by mammalian 26 and 20 S proteasomes. Implications for understanding the degradative mechanism and antigen presentation**. *J Biol Chem* 1999, **274**(6):3363-3371.

48. Lauvau G, Kakimi K, Niedermann G, Ostankovitch M, Yotnda P, Firat H, Chisari FV, van Endert PM: **Human transporters associated with antigen processing (TAPs) select epitope precursor peptides for processing in the endoplasmic reticulum and presentation to T cells**. *J Exp Med* 1999, **190**(9):1227-1240.

49. van Endert PM, Tampe R, Meyer TH, Tisch R, Bach JF, McDevitt HO: **A sequential model for peptide binding and transport by the transporters associated with antigen processing**. *Immunity* 1994, **1**(6):491-500.

50. Adams J: **The proteasome: a suitable antineoplastic target**. *Nat Rev Cancer* 2004, **4**(5):349-360.

51. Glynne R, Powis SH, Beck S, Kelly A, Kerr LA, Trowsdale J: **A proteasome-related gene between the two ABC transporter loci in the class II region of the human MHC**. *Nature* 1991, **353**(6342):357-360.

52. Ortiz-Navarrete V, Seelig A, Gernold M, Frentzel S, Kloetzel PM, Hammerling GJ: **Subunit of the '20S' proteasome (multicatalytic proteinase) encoded by the major histocompatibility complex**. *Nature* 1991, **353**(6345):662-664.

53. Kelly A, Powis SH, Glynne R, Radley E, Beck S, Trowsdale J: **Second proteasome-related gene in the human MHC class II region**. *Nature* 1991, **353**(6345):667-668.

54. Tanaka K, Kasahara M: **The MHC class I ligand-generating system: roles of immunoproteasomes and the interferon-gamma-inducible proteasome activator PA28**. *Immunol Rev* 1998, **163**:161-176.

55. Ma CP, Slaughter CA, DeMartino GN: **Identification, purification, and characterization of a protein activator (PA28) of the 20 S proteasome (macropain)**. *J Biol Chem* 1992, **267**(15):10515-10523.

56. Rechsteiner M, Realini C, Ustrell V: **The proteasome activator 11 S REG (PA28) and class I antigen presentation**. *Biochem J* 2000, **345 Pt 1**:1-15.

57. Boes B, Hengel H, Ruppert T, Multhaup G, Koszinowski UH, Kloetzel PM: **Interferon gamma stimulation modulates the proteolytic activity and cleavage site preference of 20S mouse proteasomes**. *J Exp Med* 1994, **179**(3):901-909.

58. Groettrup M, Ruppert T, Kuehn L, Seeger M, Standera S, Koszinowski U, Kloetzel PM: **The interferon-gamma-inducible 11 S regulator (PA28) and the LMP2/LMP7 subunits govern the peptide production by the 20 S proteasome in vitro**. *J Biol Chem* 1995, **270**(40):23808-23815.

59. Gaczynska M, Rock KL, Spies T, Goldberg AL: **Peptidase activities of proteasomes are differentially regulated by the major histocompatibility complex-encoded genes for LMP2 and LMP7**. *Proc Natl Acad Sci U S A* 1994, **91**(20):9213-9217.

60. Fehling HJ, Swat W, Laplace C, Kuhn R, Rajewsky K, Muller U, von Boehmer H: **MHC class I expression in mice lacking the proteasome subunit LMP-7**. *Science* 1994, **265**(5176):1234-1237.

61. Van Kaer L, Ashton-Rickardt PG, Eichelberger M, Gaczynska M, Nagashima K, Rock KL, Goldberg AL, Doherty PC, Tonegawa S: **Altered peptidase and viral-specific T cell response in LMP2 mutant mice**. *Immunity* 1994, **1**(7):533-541.

62. Yewdell J, Lapham C, Bacik I, Spies T, Bennink J: **MHC-encoded proteasome subunits LMP2 and LMP7 are not required for efficient antigen presentation**. *J Immunol* 1994, **152**(3):1163-1170.

63. Mo XY, Cascio P, Lemerise K, Goldberg AL, Rock K: **Distinct proteolytic processes generate the C and N termini of MHC class I-binding peptides**. *J Immunol* 1999, **163**(11):5851-5859.

64. Hammer GE, Gonzalez F, Champsaur M, Cado D, Shastri N: **The aminopeptidase ERAAP shapes the peptide repertoire displayed by major histocompatibility complex class I molecules**. *Nat Immunol* 2006, **7**(1):103-112.

65. Hammer GE, Gonzalez F, James E, Nolla H, Shastri N: **In the absence of aminopeptidase ERAAP, MHC class I molecules present many unstable and highly immunogenic peptides**. *Nat Immunol* 2007, **8**(1):101-108.

66. Blanchard N, Kanaseki T, Escobar H, Delebecque F, Nagarajan NA, Reyes-Vargas E, Crockett DK, Raulet DH, Delgado JC, Shastri N: **Endoplasmic reticulum aminopeptidase associated with antigen processing defines the composition and structure of MHC class I peptide repertoire in normal and virus-infected cells**. *J Immunol* 2010, **184**(6):3033-3042.

67. Yewdell JW, Anton LC, Bennink JR: **Defective ribosomal products (DRiPs): a major source of antigenic peptides for MHC class I molecules?** *J Immunol* 1996, **157**(5):1823-1826.

68. Townsend AR, Gotch FM, Davey J: **Cytotoxic T cells recognize fragments of the influenza nucleoprotein**. *Cell* 1985, **42**(2):457-467.

69. Seifert U, Bialy LP, Ebstein F, Bech-Otschir D, Voigt A, Schroter F, Prozorovski T, Lange N, Steffen J, Rieger M *et al*: **Immunoproteasomes preserve protein homeostasis upon interferon-induced oxidative stress**. *Cell* 2010, **142**(4):613-624.

70. Kunisawa J, Shastri N: **The group II chaperonin TRiC protects proteolytic intermediates from degradation in the MHC class I antigen processing pathway**. *Mol Cell* 2003, **12**(3):565-576.

71. Momburg F, Roelse J, Hammerling GJ, Neefjes JJ: **Peptide size selection by the major histocompatibility complex-encoded peptide transporter**. *J Exp Med* 1994, **179**(5):1613-1623.

72. Beninga J, Rock KL, Goldberg AL: **Interferon-gamma can stimulate post-proteasomal trimming of the N terminus of an antigenic peptide by inducing leucine aminopeptidase**. *J Biol Chem* 1998, **273**(30):18734-18742.

73. Geier E, Pfeifer G, Wilm M, Lucchiari-Hartz M, Baumeister W, Eichmann K, Niedermann G: **A giant protease with potential to substitute for some functions of the proteasome**. *Science* 1999, **283**(5404):978-981.

74. Fisette O, Schroder GF, Schafer LV: **Atomistic structure and dynamics of the human MHC-I peptide-loading complex**. *Proc Natl Acad Sci U S A* 2020, **117**(34):20597-20606.

75. Rock KL, Reits E, Neefjes J: **Present Yourself! By MHC Class I and MHC Class II Molecules**. *Trends Immunol* 2016, **37**(11):724-737.

76. Neefjes J, Jongsma ML, Paul P, Bakke O: **Towards a systems understanding of MHC class I and MHC class II antigen presentation**. *Nat Rev Immunol* 2011, **11**(12):823-836.

77. Elliott T, van Hateren A: **Protein Plasticity and Peptide Editing in the MHC I Antigen Processing Pathway**. *Biochemistry* 2018, **57**(9):1423-1425.

78. Blees A, Januliene D, Hofmann T, Koller N, Schmidt C, Trowitzsch S, Moeller A, Tampe R: **Structure of the human MHC-I peptide-loading complex**. *Nature* 2017, **551**(7681):525-528.

79. Panter MS, Jain A, Leonhardt RM, Ha T, Cresswell P: **Dynamics of major histocompatibility complex class I association with the human peptide-loading complex**. *J Biol Chem* 2012, **287**(37):31172-31184.

80. Thomas C, Tampe R: **MHC I chaperone complexes shaping immunity**. *Curr Opin Immunol* 2019, **58**:9-15.

81. James E, Bailey I, Sugiyarto G, Elliott T: **Induction of protective antitumor immunity through attenuation of ERAAP function**. *J Immunol* 2013, **190**(11):5839-5846.

82. Rogi T, Tsujimoto M, Nakazato H, Mizutani S, Tomoda Y: **Human placental leucine aminopeptidase/oxytocinase. A new member of type II membrane-spanning zinc metallopeptidase family**. *J Biol Chem* 1996, **271**(1):56-61.

83. Hattori A, Matsumoto H, Mizutani S, Tsujimoto M: **Molecular cloning of adipocyte-derived leucine aminopeptidase highly related to placental leucine aminopeptidase/oxytocinase**. *J Biochem* 1999, **125**(5):931-938.

84. Serwold T, Gonzalez F, Kim J, Jacob R, Shastri N: **ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum**. *Nature* 2002, **419**(6906):480-483.

85. Serwold T, Gaw S, Shastri N: **ER aminopeptidases generate a unique pool of peptides for MHC class I molecules**. *Nat Immunol* 2001, **2**(7):644-651.

86. Chang SC, Momburg F, Bhutani N, Goldberg AL: **The ER aminopeptidase, ERAP1, trims precursors to lengths of MHC class I peptides by a "molecular ruler" mechanism**. *Proc Natl Acad Sci U S A* 2005, **102**(47):17107-17112.

87. Chen H, Li L, Weimershaus M, Evnouchidou I, van Endert P, Bouvier M: **ERAP1-ERAP2 dimers trim MHC I-bound precursor peptides; implications for understanding peptide editing**. *Sci Rep* 2016, **6**:28902.

88. Evnouchidou I, Kamal RP, Seregin SS, Goto Y, Tsujimoto M, Hattori A, Voulgari PV, Drosos AA, Amalfitano A, York IA *et al*: **Cutting Edge: Coding single nucleotide polymorphisms of endoplasmic reticulum aminopeptidase 1 can affect antigenic peptide generation in vitro by influencing basic enzymatic properties of the enzyme**. *J Immunol* 2011, **186**(4):1909-1913.

89. Reeves E, Colebatch-Bourn A, Elliott T, Edwards CJ, James E: **Functionally distinct ERAP1 allotype combinations distinguish individuals with Ankylosing Spondylitis**. *Proc Natl Acad Sci U S A* 2014, **111**(49):17594-17599.

90. Reeves E, Edwards CJ, Elliott T, James E: **Naturally occurring ERAP1 haplotypes encode functionally distinct alleles with fine substrate specificity**. *J Immunol* 2013, **191**(1):35-43.

91. Yan J, Parekh VV, Mendez-Fernandez Y, Olivares-Villagomez D, Dragovic S, Hill T, Roopenian DC, Joyce S, Van Kaer L: **In vivo role of ER-associated peptidase activity in tailoring peptides for presentation by MHC class Ia and class Ib molecules**. *J Exp Med* 2006, **203**(3):647-659.

92. Firat E, Saveanu L, Aichele P, Staeheli P, Huai J, Gaedicke S, Nil A, Besin G, Kanzler B, van Endert P *et al*: **The role of endoplasmic reticulum-associated aminopeptidase 1 in immunity to infection and in cross-presentation**. *J Immunol* 2007, **178**(4):2241-2248.

93. Chen L, Fischer R, Peng Y, Reeves E, McHugh K, Ternette N, Hanke T, Dong T, Elliott T, Shastri N *et al*: **Critical role of endoplasmic reticulum aminopeptidase 1 in determining the length and sequence of peptides bound and presented by HLA-B27**. *Arthritis Rheumatol* 2014, **66**(2):284-294.

94. Weiss GA, Valentekovich RJ, Collins EJ, Garboczi DN, Lane WS, Schreiber SL, Wiley DC: **Covalent HLA-B27/peptide complex induced by specific recognition of an aziridine mimic of arginine**. *Proc Natl Acad Sci U S A* 1996, **93**(20):10945-10948.

95. Keller M, Ebstein F, Burger E, Textoris-Taube K, Gorny X, Urban S, Zhao F, Dannenberg T, Sucker A, Keller C *et al*: **The proteasome immunosubunits, PA28 and ER-aminopeptidase 1 protect melanoma cells from efficient MART-126-35 -specific T-cell recognition**. *Eur J Immunol* 2015, **45**(12):3257-3268.

96. Saveanu L, Carroll O, Lindo V, Del Val M, Lopez D, Lepelletier Y, Greer F, Schomburg L, Fruci D, Niedermann G *et al*: **Concerted peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum**. *Nat Immunol* 2005, **6**(7):689-697.

97. Andres AM, Dennis MY, Kretzschmar WW, Cannons JL, Lee-Lin SQ, Hurle B, Program NCS, Schwartzberg PL, Williamson SH, Bustamante CD *et al*: **Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation**. *PLoS Genet* 2010, **6**(10):e1001157.

98. Compagnone M, Cifaldi L, Fruci D: **Regulation of ERAP1 and ERAP2 genes and their disfunction in human cancer**. *Hum Immunol* 2019, **80**(5):318-324.

99. Kamphausen E, Kellert C, Abbas T, Akkad N, Tenzer S, Pawelec G, Schild H, van Endert P, Seliger B: **Distinct molecular mechanisms leading to deficient expression of ER-resident aminopeptidases in melanoma**. *Cancer Immunol Immunother* 2010, **59**(8):1273-1284.

100. Fruci D, Giacomini P, Nicotra MR, Forloni M, Fraioli R, Saveanu L, van Endert P, Natali PG: **Altered expression of endoplasmic reticulum aminopeptidases ERAP1 and ERAP2 in transformed non-lymphoid human tissues**. *J Cell Physiol* 2008, **216**(3):742-749.

101. Fruci D, Ferracuti S, Limongi MZ, Cunsolo V, Giorda E, Fraioli R, Sibilio L, Carroll O, Hattori A, van Endert PM *et al*: **Expression of endoplasmic reticulum aminopeptidases in EBV-B cell lines from healthy donors and in leukemia/lymphoma, carcinoma, and melanoma cell lines**. *J Immunol* 2006, **176**(8):4869-4879.

102. Kochan G, Krojer T, Harvey D, Fischer R, Chen L, Vollmar M, von Delft F, Kavanagh KL, Brown MA, Bowness P *et al*: **Crystal structures of the endoplasmic reticulum aminopeptidase-1 (ERAP1) reveal the molecular basis for N-terminal peptide trimming**. *Proc Natl Acad Sci U S A* 2011, **108**(19):7745-7750.

103. Nguyen TT, Chang SC, Evnouchidou I, York IA, Zikos C, Rock KL, Goldberg AL, Stratikos E, Stern LJ: **Structural basis for antigenic peptide precursor processing by the endoplasmic reticulum aminopeptidase ERAP1**. *Nat Struct Mol Biol* 2011, **18**(5):604-613.

104. Giastas P, Neu M, Rowland P, Stratikos E: **High-Resolution Crystal Structure of Endoplasmic Reticulum Aminopeptidase 1 with Bound Phosphinic Transition-State Analogue Inhibitor**. *ACS Med Chem Lett* 2019, **10**(5):708-713.

105. Hattori A, Matsumoto K, Mizutani S, Tsujimoto M: **Genomic organization of the human adipocyte-derived leucine aminopeptidase gene and its relationship to the placental leucine aminopeptidase/oxytocinase gene**. *J Biochem* 2001, **130**(2):235-241.

106. Tanioka T, Hattori A, Masuda S, Nomura Y, Nakayama H, Mizutani S, Tsujimoto M: **Human leukocyte-derived arginine aminopeptidase. The third member of the oxytocinase subfamily of aminopeptidases**. *J Biol Chem* 2003, **278**(34):32275-32283.

107. Kanaseki T, Blanchard N, Hammer GE, Gonzalez F, Shastri N: **ERAAP synergizes with MHC class I molecules to make the final cut in the antigenic peptide precursors in the endoplasmic reticulum**. *Immunity* 2006, **25**(5):795-806.

108. Giastas P, Mpakali A, Papakyriakou A, Lelis A, Kokkala P, Neu M, Rowland P, Liddle J, Georgiadis D, Stratikos E: **Mechanism for antigenic peptide selection by endoplasmic reticulum aminopeptidase 1**. *Proc Natl Acad Sci U S A* 2019.

109. Gandhi A, Lakshminarasimhan D, Sun Y, Guo HC: **Structural insights into the molecular ruler mechanism of the endoplasmic reticulum aminopeptidase ERAP1**. *Sci Rep* 2011, **1**:186.

110. Sui L, Gandhi A, Guo HC: **Crystal structure of a polypeptide's C-terminus in complex with the regulatory domain of ER aminopeptidase 1**. *Mol Immunol* 2016, **80**:41-49.

111. Goto Y, Tanji H, Hattori A, Tsujimoto M: **Glutamine-181 is crucial in the enzymatic activity and substrate specificity of human endoplasmic-reticulum aminopeptidase-1**. *Biochem J* 2008, **416**(1):109-116.

112. Hattori A, Goto Y, Tsujimoto M: **Exon 10 coding sequence is important for endoplasmic reticulum retention of endoplasmic reticulum aminopeptidase 1**. *Biol Pharm Bull* 2012, **35**(4):601-605.

113. Hearn A, York IA, Rock KL: **The specificity of trimming of MHC class I-presented peptides in the endoplasmic reticulum**. *J Immunol* 2009, **183**(9):5526-5536.

114. Zervoudi E, Papakyriakou A, Georgiadou D, Evnouchidou I, Gajda A, Poreba M, Salvesen GS, Drag M, Hattori A, Swevers L *et al*: **Probing the S1 specificity pocket of the aminopeptidases that generate antigenic peptides**. *Biochem J* 2011, **435**(2):411-420.

115. Mpakali A, Giastas P, Mathioudakis N, Mavridis IM, Saridakis E, Stratikos E: **Structural Basis for Antigenic Peptide Recognition and Processing by Endoplasmic Reticulum (ER) Aminopeptidase 2**. *J Biol Chem* 2015, **290**(43):26021-26032.

116. Neisig A, Roelse J, Sijts AJ, Ossendorp F, Feltkamp MC, Kast WM, Melief CJ, Neefjes JJ: **Major differences in transporter associated with antigen presentation (TAP)-dependent translocation of MHC class I-presentable peptides and the effect of flanking sequences**. *J Immunol* 1995, **154**(3):1273-1279.

117. Reeves E, James E: **The role of polymorphic ERAP1 in autoinflammatory disease**. *Biosci Rep* 2018, **38**(4).

118. Evnouchidou I, Momburg F, Papakyriakou A, Chroni A, Leondiadis L, Chang SC, Goldberg AL, Stratikos E: **The internal sequence of the peptide-substrate determines its N-terminus trimming by ERAP1**. *PLoS One* 2008, **3**(11):e3658.

119. Tran TM, Colbert RA: **Endoplasmic reticulum aminopeptidase 1 and rheumatic disease: functional variation**. *Curr Opin Rheumatol* 2015, **27**(4):357-363.

120. Ombrello MJ, Kastner DL, Remmers EF: **Endoplasmic reticulum-associated amino-peptidase 1 and rheumatic disease: genetics**. *Curr Opin Rheumatol* 2015, **27**(4):349-356.

121. Reeves E, Wood O, Ottensmeier CH, King EV, Thomas GJ, Elliott T, James E: **HPV epitope processing differences correlate with ERAP1 allotype and extent of CD8+ T cell tumor infiltration in OPSCC**. *Cancer Immunol Res* 2019.

122. Stamogiannos A, Koumantou D, Papakyriakou A, Stratikos E: **Effects of polymorphic variation on the mechanism of Endoplasmic Reticulum Aminopeptidase 1**. *Mol Immunol* 2015, **67**(2 Pt B):426-435.

123. Hutchinson JP, Temponeras I, Kuiper J, Cortes A, Korczynska J, Kitchen S, Stratikos E: **Common allotypes of ER aminopeptidase 1 have substrate-dependent and highly variable enzymatic properties**. *J Biol Chem* 2021, **296**:100443.

124. Kuiper JJW, Setten JV, Devall M, Cretu-Stancu M, Hiddingh S, Ophoff RA, Missotten TOAR, Velthoven MV, Den Hollander AI, Hoyng CB *et al*: **Functionally distinct ERAP1 and ERAP2 are a hallmark of HLA-A29-(Birdshot) Uveitis**. *Hum Mol Genet* 2018, **27**(24):4333-4343.

125. Goto Y, Hattori A, Ishii Y, Tsujimoto M: **Reduced activity of the hypertension-associated Lys528Arg mutant of human adipocyte-derived leucine aminopeptidase (A-LAP)/ER-aminopeptidase-1**. *FEBS Lett* 2006, **580**(7):1833-1838.

126. Paladini F, Fiorillo MT, Vitulano C, Tedeschi V, Piga M, Cauli A, Mathieu A, Sorrentino R: **An allelic variant in the intergenic region between ERAP1 and ERAP2 correlates with an inverse expression of the two genes**. *Sci Rep* 2018, **8**(1):10398.

127. Alvarez-Navarro C, López de Castro JA: **ERAP1 structure, function and pathogenetic role in ankylosing spondylitis and other MHC-associated diseases**. *Mol Immunol* 2014, **57**(1):12-21.

128. Evans DM, Spencer CC, Pointon JJ, Su Z, Harvey D, Kochan G, Oppermann U, Opperman U, Dilthey A, Pirinen M *et al*: **Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility**. *Nat Genet* 2011, **43**(8):761-767.

129. Reeves E, Elliott T, James E, Edwards CJ: **ERAP1 in the pathogenesis of ankylosing spondylitis**. *Immunol Res* 2014, **60**(2-3):257-269.

130. Maksymowych WP, Inman RD, Gladman DD, Reeve JP, Pope A, Rahman P: **Association of a specific ERAP1/ARTS1 haplotype with disease susceptibility in ankylosing spondylitis**. *Arthritis Rheum* 2009, **60**(5):1317-1323.

131. Mahmoudi M, Jamshidi AR, Amirzargar AA, Farhadi E, Nourijelyani K, Fallahi S, Oraei M, Noori S, Nicknam MH: **Association between endoplasmic reticulum aminopeptidase-1 (ERAP-1) and susceptibility to ankylosing spondylitis in Iran**. *Iran J Allergy Asthma Immunol* 2012, **11**(4):294-300.

132. Costantino F, Talpin A, Evnouchidou I, Kadi A, Leboime A, Said-Nahal R, Bonilla N, Letourneur F, Leturcq T, Ka Z *et al*: **ERAP1 Gene Expression Is Influenced by Nonsynonymous Polymorphisms Associated With Predisposition to Spondyloarthritis**. *Arthritis Rheumatol* 2015, **67**(6):1525-1534.

133. Cortes A, Pulit SL, Leo PJ, Pointon JJ, Robinson PC, Weisman MH, Ward M, Gensler LS, Zhou X, Garchon HJ *et al*: **Major histocompatibility complex associations of ankylosing spondylitis are complex and involve further epistasis with ERAP1**. *Nat Commun* 2015, **6**:7146.

134. Seregin SS, Rastall DP, Evnouchidou I, Aylsworth CF, Quiroga D, Kamal RP, Godbehere-Roosa S, Blum CF, York IA, Stratikos E *et al*: **Endoplasmic reticulum aminopeptidase-1 alleles associated with increased risk of ankylosing spondylitis reduce HLA-B27 mediated presentation of multiple antigens**. *Autoimmunity* 2013, **46**(8):497-508.

135. Haroon N, Tsui FW, Uchanska-Ziegler B, Ziegler A, Inman RD: **Endoplasmic reticulum aminopeptidase 1 (ERAP1) exhibits functionally significant interaction with HLA-B27 and relates to subtype specificity in ankylosing spondylitis**. *Ann Rheum Dis* 2012, **71**(4):589-595.

136. Takeuchi M, Ombrello MJ, Kirino Y, Erer B, Tugal-Tutkun I, Seyahi E, Ozyazgan Y, Watts NR, Gul A, Kastner DL *et al*: **A single endoplasmic reticulum aminopeptidase-1 protein allotype is a strong risk factor for Behcet's disease in HLA-B*51 carriers**. *Ann Rheum Dis* 2016, **75**(12):2208-2211.

137. Das A, Chandra A, Chakraborty J, Chattopadhyay A, Senapati S, Chatterjee G, Chatterjee R: **Associations of ERAP1 coding variants and domain specific interaction with HLA-C *06 in the early onset psoriasis patients of India**. *Hum Immunol* 2017, **78**(11-12):724-730.

138. Lysell J, Padyukov L, Kockum I, Nikamo P, Ståhle M: **Genetic association with ERAP1 in psoriasis is confined to disease onset after puberty and not dependent on HLA-C*06**. *J Invest Dermatol* 2013, **133**(2):411-417.

139. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ *et al*: **Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants**. *Nat Genet* 2007, **39**(11):1329-1337.

140. Zhang L, Yu H, Zheng M, Li H, Liu Y, Kijlstra A, Yang P: **Association of ERAP1 Gene Polymorphisms With Behcet's Disease in Han Chinese**. *Invest Ophthalmol Vis Sci* 2015, **56**(10):6029-6035.

141. Fung EY, Smyth DJ, Howson JM, Cooper JD, Walker NM, Stevens H, Wicker LS, Todd JA: **Analysis of 17 autoimmune disease-associated variants in type 1 diabetes identifies 6q23/TNFAIP3 as a susceptibility locus**. *Genes Immun* 2009, **10**(2):188-191.

142. Wisniewski A, Matusiak L, Szczerkowska-Dobosz A, Nowak I, Luszczek W, Kusnierczyk P: **The association of ERAP1 and ERAP2 single nucleotide polymorphisms and their haplotypes with psoriasis vulgaris is dependent on the presence or absence of the HLA-C*06:02 allele and age at disease onset**. *Hum Immunol* 2018, **79**(2):109-116.

143. Guerini FR, Cagliani R, Forni D, Agliardi C, Caputo D, Cassinotti A, Galimberti D, Fenoglio C, Biasin M, Asselta R *et al*: **A functional variant in ERAP1 predisposes to multiple sclerosis**. *PLoS One* 2012, **7**(1):e29931.

144. Castro-Santos P, Moro-Garcia MA, Marcos-Fernandez R, Alonso-Arias R, Diaz-Pena R: **ERAP1 and HLA-C interaction in inflammatory bowel disease in the Spanish population**. *Innate Immun* 2017, **23**(5):476-481.

145. **Cervical cancer** [https://www.cancerresearchuk.org/about-cancer/cervical-cancer]

146. **Results are in for Europe's First Biobanking Study on HPV cervical cancer**

147. Marth C, Landoni F, Mahner S, McCormack M, Gonzalez-Martin A, Colombo N, Committee EG: **Cervical cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up**. *Ann Oncol* 2018, **29**(Supplement_4):iv262.

148. Pirog EC: **Cervical Adenocarcinoma: Diagnosis of Human Papillomavirus-Positive and Human Papillomavirus-Negative Tumors**. *Arch Pathol Lab Med* 2017, **141**(12):1653-1667.

149. de Vos van Steenwijk PJ, Heusinkveld M, Ramwadhdoebe TH, Löwik MJ, van der Hulst JM, Goedemans R, Piersma SJ, Kenter GG, van der Burg SH: **An unexpectedly large polyclonal repertoire of HPV-specific T cells is poised for action in patients with cervical cancer**. *Cancer Res* 2010, **70**(7):2707-2717.

150. Chesson HW, Dunne EF, Hariri S, Markowitz LE: **The estimated lifetime probability of acquiring human papillomavirus in the United States**. *Sex Transm Dis* 2014, **41**(11):660-664.

151. Schiffman M, Castle PE, Jeronimo J, Rodriguez AC, Wacholder S: **Human papillomavirus and cervical cancer**. *Lancet* 2007, **370**(9590):890-907.

152. Fu TC, Carter JJ, Hughes JP, Feng Q, Hawes SE, Schwartz SM, Xi LF, Lasof T, Stern JE, Galloway DA *et al*: **Re-detection vs. new acquisition of high-risk human papillomavirus in mid-adult women**. *Int J Cancer* 2016, **139**(10):2201-2212.

153. Shew ML, Ermel AC, Tong Y, Tu W, Qadadri B, Brown DR: **Episodic detection of human papillomavirus within a longitudinal cohort of young women**. *J Med Virol* 2015, **87**(12):2122-2129.

154. Liu SH, Cummings DA, Zenilman JM, Gravitt PE, Brotman RM: **Characterizing the temporal dynamics of human papillomavirus DNA detectability using short-interval sampling**. *Cancer Epidemiol Biomarkers Prev* 2014, **23**(1):200-208.

155. Gravitt PE, Winer RL: **Natural History of HPV Infection across the Lifespan: Role of Viral Latency**. *Viruses* 2017, **9**(10).

156. Gravitt PE: **Evidence and impact of human papillomavirus latency**. *Open Virol J* 2012, **6**:198-203.

157. Ibeanu OA: **Molecular pathogenesis of cervical cancer**. *Cancer Biol Ther* 2011, **11**(3):295-306.

158. de la Garza-Salazar J, Moralez-Vasquez F, Meneses-Garcia A: **Cervical cancer**. Switzerland: Sanger; 2017.

159. Eskander RN, Tewari KS: **Immunotherapy: an evolving paradigm in the treatment of advanced cervical cancer**. *Clin Ther* 2015, **37**(1):20-38.

160. Scheffner M, Huibregtse JM, Vierstra RD, Howley PM: **The HPV-16 E6 and E6-AP complex functions as a ubiquitin-protein ligase in the ubiquitination of p53**. *Cell* 1993, **75**(3):495-505.

161. Tomaić V: **Functional Roles of E6 and E7 Oncoproteins in HPV-Induced Malignancies at Diverse Anatomical Sites**. *Cancers (Basel)* 2016, **8**(10).

162. S. Gustafson K, P. Clark D: **Cell and tissue based molecular pathology**, vol. Section III, Chapter 17: Churchill Livingstone; 2009.

163. Maimela NR, Liu S, Zhang Y: **Fates of CD8+ T cells in Tumor Microenvironment**. *Comput Struct Biotechnol J* 2019, **17**:1-13.

164. Maleki Vareki S: **High and low mutational burden tumors versus immunologically hot and cold tumors and response to immune checkpoint inhibitors**. *J Immunother Cancer* 2018, **6**(1):157.

165. Scheper W, Kelderman S, Fanchi LF, Linnemann C, Bendle G, de Rooij MAJ, Hirt C, Mezzadra R, Slagter M, Dijkstra K *et al*: **Low and variable tumor reactivity of the intratumoral TCR repertoire in human cancers**. *Nat Med* 2019, **25**(1):89-94.

166. Piersma SJ, Jordanova ES, van Poelgeest MI, Kwappenberg KM, van der Hulst JM, Drijfhout JW, Melief CJ, Kenter GG, Fleuren GJ, Offringa R *et al*: **High number of intraepithelial CD8+ tumor-infiltrating lymphocytes is associated with the absence of lymph node metastases in patients with large early-stage cervical cancer**. *Cancer Res* 2007, **67**(1):354-361.

167. de Vos van Steenwijk PJ, Ramwadhdoebe TH, Goedemans R, Doorduijn EM, van Ham JJ, Gorter A, van Hall T, Kuijjer ML, van Poelgeest MI, van der Burg SH *et al*: **Tumor-infiltrating**

CD14-positive myeloid cells and CD8-positive T-cells prolong survival in patients with cervical carcinoma. *Int J Cancer* 2013, **133**(12):2884-2894.

168. Kobayashi A, Weinberg V, Darragh T, Smith-McCune K: **Evolving immunosuppressive microenvironment during human cervical carcinogenesis**. *Mucosal Immunol* 2008, **1**(5):412-420.

169. van der Burg SH, Melief CJ: **Therapeutic vaccination against human papilloma virus induced malignancies**. *Curr Opin Immunol* 2011, **23**(2):252-257.

170. Komdeur FL, Prins TM, van de Wall S, Plat A, Wisman GBA, Hollema H, Daemen T, Church DN, de Bruyn M, Nijman HW: **CD103+ tumor-infiltrating lymphocytes are tumor-reactive intraepithelial CD8+ T cells associated with prognostic benefit and therapy response in cervical cancer**. *Oncoimmunology* 2017, **6**(9):e1338230.

171. Woo YL, van den Hende M, Sterling JC, Coleman N, Crawford RA, Kwappenberg KM, Stanley MA, van der Burg SH: **A prospective study on the natural course of low-grade squamous intraepithelial lesions and the presence of HPV16 E2-, E6- and E7-specific T-cell responses**. *Int J Cancer* 2010, **126**(1):133-141.

172. Farhat S, Nakagawa M, Moscicki AB: **Cell-mediated immune responses to human papillomavirus 16 E6 and E7 antigens as measured by interferon gamma enzyme-linked immunospot in women with cleared or persistent human papillomavirus infection**. *Int J Gynecol Cancer* 2009, **19**(4):508-512.

173. Heusinkveld M, Welters MJ, van Poelgeest MI, van der Hulst JM, Melief CJ, Fleuren GJ, Kenter GG, van der Burg SH: **The detection of circulating human papillomavirus-specific T cells is associated with improved survival of patients with deeply infiltrating tumors**. *Int J Cancer* 2011, **128**(2):379-389.

174. Frazer IH, Chandra J: **Immunotherapy for HPV associated cancer**. *Papillomavirus Res* 2019, **8**:100176.

175. Liu C, Lu J, Tian H, Du W, Zhao L, Feng J, Yuan D, Li Z: **Increased expression of PD-L1 by the human papillomavirus 16 E7 oncoprotein inhibits anticancer immunity**. *Mol Med Rep* 2016.

176. Manguso RT, Pope HW, Zimmer MD, Brown FD, Yates KB, Miller BC, Collins NB, Bi K, LaFleur MW, Juneja VR *et al*: **In vivo CRISPR screening identifies Ptpn2 as a cancer immunotherapy target**. *Nature* 2017, **547**(7664):413-418.

177. Mehta AM, Jordanova ES, Kenter GG, Ferrone S, Fleuren GJ: **Association of antigen processing machinery and HLA class I defects with clinicopathological outcome in cervical carcinoma**. *Cancer Immunol Immunother* 2008, **57**(2):197-206.

178. Albers A, Abe K, Hunt J, Wang J, Lopez-Albaitero A, Schaefer C, Gooding W, Whiteside TL, Ferrone S, DeLeo A *et al*: **Antitumor activity of human papillomavirus type 16 E7-specific T cells against virally infected squamous cell carcinoma of the head and neck**. *Cancer Res* 2005, **65**(23):11146-11155.

179. Welters MJP, Ma W, Santegoets SJAM, Goedemans R, Ehsan I, Jordanova ES, van Ham VJ, van Unen V, Koning F, van Egmond SI *et al*: **Intratumoral HPV16-Specific T Cells Constitute a**

Type I-Oriented Tumor Microenvironment to Improve Survival in HPV16-Driven Oropharyngeal Cancer. *Clin Cancer Res* 2018, **24**(3):634-647.

180. Kemming J, Reeves E, Nitschke K, Widmeier V, Emmerich F, Hermle T, Gostick E, Walker A, Timm J, Price DA *et al*: **ERAP1 allotypes shape the epitope repertoire of virus-specific CD8**. *J Hepatol* 2019, **70**(6):1072-1081.

181. Mehta AM, Jordanova ES, van Wezel T, Uh HW, Corver WE, Kwappenberg KM, Verduijn W, Kenter GG, van der Burg SH, Fleuren GJ: **Genetic variation of antigen processing machinery components and association with cervical carcinoma**. *Genes Chromosomes Cancer* 2007, **46**(6):577-586.

182. Li C, Li Y, Yan Z, Dai S, Liu S, Wang X, Wang J, Zhang X, Shi L, Yao Y: **Polymorphisms in endoplasmic reticulum aminopeptidase genes are associated with cervical cancer risk in a Chinese Han population**. *BMC Cancer* 2020, **20**(1):341.

183. Jain M, Olsen HE, Paten B, Akeson M: **The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community**. *Genome Biol* 2016, **17**(1):239.

184. Wang Y, Zhao Y, Bollas A, Au KF: **Nanopore sequencing technology, bioinformatics and applications**. *Nat Biotechnol* 2021, **39**(11):1348-1365.

185. Imai K, Tamura K, Tanigaki T, Takizawa M, Nakayama E, Taniguchi T, Okamoto M, Nishiyama Y, Tarumoto N, Mitsutake K *et al*: **Whole Genome Sequencing of Influenza A and B Viruses With the MinION Sequencer in the Clinical Setting: A Pilot Study**. *Front Microbiol* 2018, **9**:2748.

186. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT *et al*: **Nanopore sequencing and assembly of a human genome with ultra-long reads**. *Nat Biotechnol* 2018, **36**(4):338-345.

187. Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H: **The Architecture of SARS-CoV-2 Transcriptome**. *Cell* 2020, **181**(4):914-921 e910.

188. Ton KNT, Cree SL, Gronert-Sum SJ, Merriman TR, Stamp LK, Kennedy MA: **Multiplexed Nanopore Sequencing of HLA-B Locus in Maori and Pacific Island Samples**. *Front Genet* 2018, **9**:152.

189. Quan L, Dong R, Yang W, Chen L, Lang J, Liu J, Song Y, Ma S, Yang J, Wang W *et al*: **Simultaneous detection and comprehensive analysis of HPV and microbiome status of a cervical liquid-based cytology sample using Nanopore MinION sequencing**. *Sci Rep* 2019, **9**(1):19337.

190. Mehta AM, Jordanova ES, Corver WE, van Wezel T, Uh HW, Kenter GG, Jan Fleuren G: **Single nucleotide polymorphisms in antigen processing machinery component ERAP1 significantly associate with clinical outcome in cervical carcinoma**. *Genes Chromosomes Cancer* 2009, **48**(5):410-418.

191. Leidenfrost RM, Pöther DC, Jäckel U, Wünschiers R: **Benchmarking the MinION: Evaluating long reads for microbial profiling**. *Sci Rep* 2020, **10**(1):5125.

192. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C: **NanoPack: visualizing and processing long-read sequencing data**. *Bioinformatics* 2018, **34**(15):2666-2669.

193. Li H: **Minimap2: pairwise alignment for nucleotide sequences**. *Bioinformatics* 2018, **34**(18):3094-3100.

194. Marcel M, Patterson M, Garg S, O.Fischer S, Pisanti N, W.Klau G, Schoenhuth A, Marschall T: **WhatsHap: fast and accurate read-based phasing**. 2016, **085050**.

195. Sanderson S, Shastri N: **LacZ inducible, antigen/MHC-specific T cell hybrids**. *Int Immunol* 1994, **6**(3):369-376.

196. Wang Z, Marincola FM, Rivoltini L, Parmiani G, Ferrone S: **Selective histocompatibility leukocyte antigen (HLA)-A2 loss caused by aberrant pre-mRNA splicing in 624MEL28 melanoma cells**. *J Exp Med* 1999, **190**(2):205-215.

197. de Sanjose S, Quint WG, Alemany L, Geraets DT, Klaustermeier JE, Lloveras B, Tous S, Felix A, Bravo LE, Shin HR *et al*: **Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study**. *Lancet Oncol* 2010, **11**(11):1048-1056.

198. Ward MJ, Thirdborough SM, Mellows T, Riley C, Harris S, Suchak K, Webb A, Hampton C, Patel NN, Randall CJ *et al*: **Tumour-infiltrating lymphocytes predict for outcome in HPV-positive oropharyngeal cancer**. *Br J Cancer* 2014, **110**(2):489-500.

199. Lubbers JM, Koopman B, de Klerk-Sluis JM, van Rooij N, Plat A, Pijper H, Koopman T, van Hemel BM, Hollema H, Wisman B *et al*: **Association of homozygous variants of STING1 with outcome in human cervical cancer**. *Cancer Sci* 2021, **112**(1):61-71.

200. Landoni F, Colombo A, Milani R, Placa F, Zanagnolo V, Mangioni C: **Randomized study between radical surgery and radiotherapy for the treatment of stage IB-IIA cervical cancer: 20-year update**. *J Gynecol Oncol* 2017, **28**(3):e34.

201. Rapiti E, Usel M, Neyroud-Caspar I, Merglen A, Verkooijen HM, Vlastos AT, Pache JC, Kumar N, Bouchardy C: **Omission of excisional therapy is associated with an increased risk of invasive cervical cancer after cervical intraepithelial neoplasia III**. *Eur J Cancer* 2012, **48**(6):845-852.

202. Quinn MA, Benedet JL, Odicino F, Maisonneuve P, Beller U, Creasman WT, Heintz AP, Ngan HY, Pecorelli S: **Carcinoma of the cervix uteri. FIGO 26th Annual Report on the Results of Treatment in Gynecological Cancer**. *Int J Gynaecol Obstet* 2006, **95 Suppl 1**:S43-103.

203. Piersma SJ: **Immunosuppressive tumor microenvironment in cervical cancer patients**. *Cancer Microenviron* 2011, **4**(3):361-375.

204. Wang H, Chen L, Ma W, Zeng Y, Qin L, Chen M, Li L: **Prediction and identification of human leukocyte antigen-A2-restricted cytotoxic T lymphocyte epitope peptides from the human papillomavirus 58 E7 protein**. *Oncol Lett* 2018, **16**(2):2003-2008.

205. Kather A, Ferrara A, Nonn M, Schinz M, Nieland J, Schneider A, Durst M, Kaufmann AM: **Identification of a naturally processed HLA-A*0201 HPV18 E7 T cell epitope by tumor cell mediated in vitro vaccination**. *Int J Cancer* 2003, **104**(3):345-353.

206. Grimaldi AM, Simeone E, Giannarelli D, Muto P, Falivene S, Borzillo V, Giugliano FM, Sandomenico F, Petrillo A, Curvietto M *et al*: **Abscopal effects of radiotherapy on advanced melanoma patients who progressed after ipilimumab immunotherapy**. *Oncoimmunology* 2014, **3**:e28780.

207. Colbert LE, El MB, Lynn EJ, Bronk J, Karpinets TV, Wu X, Chapman BV, Sims TT, Lin D, Kouzy R *et al*: **Expansion of Candidate HPV-Specific T Cells in the Tumor Microenvironment during Chemoradiotherapy Is Prognostic in HPV16**. *Cancer Immunol Res* 2022, **10**(2):259-271.

208. de Foucher T, Bendifallah S, Ouldamer L, Bricou A, Lavoue V, Varinot J, Canlorbe G, Carcopino X, Raimond E, Monnier L *et al*: **Patterns of recurrence and prognosis in locally advanced FIGO stage IB2 to IIB cervical cancer: Retrospective multicentre study from the FRANCOGYN group**. *Eur J Surg Oncol* 2019, **45**(4):659-665.

209. Cornelis S, Gansemans Y, Vander Plaetsen AS, Weymaere J, Willems S, Deforce D, Van Nieuwerburgh F: **Forensic tri-allelic SNP genotyping using nanopore sequencing**. *Forensic Sci Int Genet* 2019, **38**:204-210.

210. Norris AL, Workman RE, Fan Y, Eshleman JR, Timp W: **Nanopore sequencing detects structural variants in cancer**. *Cancer Biol Ther* 2016, **17**(3):246-253.

211. Wick RR, Judd LM, Holt KE: **Performance of neural network basecalling tools for Oxford Nanopore sequencing**. *Genome Biol* 2019, **20**(1):129.

212. Bowden R, Davies RW, Heger A, Pagnamenta AT, de Cesare M, Oikkonen LE, Parkes D, Freeman C, Dhalla F, Patel SY *et al*: **Sequencing of human genomes with nanopore technology**. *Nat Commun* 2019, **10**(1):1869.

213. Neefjes JJ, Momburg F, Hammerling GJ: **Selective and ATP-dependent translocation of peptides by the MHC-encoded transporter**. *Science* 1993, **261**(5122):769-771.

214. Shastri N, Gonzalez F: **Endogenous generation and presentation of the ovalbumin peptide/Kb complex to T cells**. *J Immunol* 1993, **150**(7):2724-2736.

215. Mehta AM, Spaans VM, Mahendra NB, Osse EM, Vet JN, Purwoto G, Surya IG, Cornian S, Peters AA, Fleuren GJ *et al*: **Differences in genetic variation in antigen-processing machinery components and association with cervical carcinoma risk in two Indonesian populations**. *Immunogenetics* 2015, **67**(5-6):267-275.

216. Kadi A, Izac B, Said-Nahal R, Leboime A, Van Praet L, de Vlam K, Elewaut D, Chiocchia G, Breban M: **Investigating the genetic association between ERAP1 and spondyloarthritis**. *Ann Rheum Dis* 2013, **72**(4):608-613.

217. Lopez de Castro JA: **How ERAP1 and ERAP2 Shape the Peptidomes of Disease-Associated MHC-I Proteins**. *Front Immunol* 2018, **9**:2463.

218. Harvey D, Pointon JJ, Evans DM, Karaderi T, Farrar C, Appleton LH, Sturrock RD, Stone MA, Oppermann U, Brown MA *et al*: **Investigating the genetic association between ERAP1 and ankylosing spondylitis**. *Hum Mol Genet* 2009, **18**(21):4204-4212.

219. Hanson AL, Cuddihy T, Haynes K, Loo D, Morton CJ, Oppermann U, Leo P, Thomas GP, Lê Cao KA, Kenna TJ *et al*: **Genetic Variants in ERAP1 and ERAP2 Associated With Immune-**

Mediated Diseases Influence Protein Expression and the Isoform Profile. *Arthritis Rheumatol* 2018, **70**(2):255-265.

220. Sanz-Bravo A, Campos J, Mazariegos MS, López de Castro JA: **Dominant role of the ERAP1 polymorphism R528K in shaping the HLA-B27 Peptidome through differential processing determined by multiple peptide residues**. *Arthritis Rheumatol* 2015, **67**(3):692-701.

221. Seah A, Lim MCW, McAloose D, Prost S, Seimon TA: **MinION-Based DNA Barcoding of Preserved and Non-Invasively Collected Wildlife Samples**. *Genes (Basel)* 2020, **11**(4).

222. Yin Y, Lan JH, Nguyen D, Valenzuela N, Takemura P, Bolon YT, Springer B, Saito K, Zheng Y, Hague T *et al*: **Application of High-Throughput Next-Generation Sequencing for HLA Typing on Buccal Extracted DNA: Results from over 10,000 Donor Recruitment Samples**. *PLoS One* 2016, **11**(10):e0165810.

223. Koboldt DC, Ding L, Mardis ER, Wilson RK: **Challenges of sequencing human genomes**. *Brief Bioinform* 2010, **11**(5):484-498.

224. Klasberg S, Surendranath V, Lange V, Schöfl G: **Bioinformatics Strategies, Challenges, and Opportunities for Next Generation Sequencing-Based HLA Genotyping**. *Transfus Med Hemother* 2019, **46**(5):312-325.

225. Strange A, Capon F, Spencer CC, Knight J, Weale ME, Allen MH, Barton A, Band G, Bellenguez C, Bergboer JG *et al*: **A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1**. *Nat Genet* 2010, **42**(11):985-990.

226. Laustsen PG, Vang S, Kristensen T: **Mutational analysis of the active site of human insulin-regulated aminopeptidase**. *Eur J Biochem* 2001, **268**(1):98-104.

227. Gazzaz MJ, Jeffery C, O'Connell D, Harris J, Seikaly H, Biron V: **Association of human papillomavirus related squamous cell carcinomas of the oropharynx and cervix**. *Papillomavirus Res* 2019, **8**:100188.

228. Martín-Esteban A, Gómez-Molina P, Sanz-Bravo A, López de Castro JA: **Combined effects of ankylosing spondylitis-associated ERAP1 polymorphisms outside the catalytic and peptide-binding sites on the processing of natural HLA-B27 ligands**. *J Biol Chem* 2014, **289**(7):3978-3990.

229. Hattori A, Tsujimoto M: **Endoplasmic reticulum aminopeptidases: biochemistry, physiology and pathology**. *J Biochem* 2013, **154**(3):219-228.

230. López de Castro JA, Alvarez-Navarro C, Brito A, Guasp P, Martín-Esteban A, Sanz-Bravo A: **Molecular and pathogenic effects of endoplasmic reticulum aminopeptidases ERAP1 and ERAP2 in MHC-I-associated inflammatory disorders: Towards a unifying view**. *Mol Immunol* 2016, **77**:193-204.

231. zur Hausen H, de Villiers EM: **Human papillomaviruses**. *Annu Rev Microbiol* 1994, **48**:427-447.

232. Kast WM, Brandt RM, Sidney J, Drijfhout JW, Kubo RT, Grey HM, Melief CJ, Sette A: **Role of HLA-A motifs in identification of potential CTL epitopes in human papillomavirus type 16 E6 and E7 proteins**. *J Immunol* 1994, **152**(8):3904-3912.

233. Davidson EJ, Davidson JA, Sterling JC, Baldwin PJ, Kitchener HC, Stern PL: **Association between human leukocyte antigen polymorphism and human papillomavirus 16-positive vulval intraepithelial neoplasia in British women**. *Cancer Res* 2003, **63**(2):400-403.

234. Liu DW, Yang YC, Lin HF, Lin MF, Cheng YW, Chu CC, Tsao YP, Chen SL: **Cytotoxic T-lymphocyte responses to human papillomavirus type 16 E5 and E7 proteins and HLA-A*0201-restricted T-cell peptides in cervical cancer patients**. *J Virol* 2007, **81**(6):2869-2879.

235. Morishima S, Akatsuka Y, Nawa A, Kondo E, Kiyono T, Torikai H, Nakanishi T, Ito Y, Tsujimura K, Iwata K *et al*: **Identification of an HLA-A24-restricted cytotoxic T lymphocyte epitope from human papillomavirus type-16 E6: the combined effects of bortezomib and interferon-gamma on the presentation of a cryptic epitope**. *Int J Cancer* 2007, **120**(3):594-604.

236. Chauhan SR, Singhal PG, Sharma U, Bandil K, Chakraborty K, Bharadwaj M: **Th9 cytokines curb cervical cancer progression and immune evasion**. *Hum Immunol* 2019, **80**(12):1020-1025.

237. Yang H, Ye S, Goswami S, Li T, Wu J, Cao C, Ma J, Lu B, Pei X, Chen Y *et al*: **Highly immunosuppressive HLADR**. *Int J Cancer* 2020, **146**(7):1993-2006.

238. Mehta AM, Osse M, Kolkman-Uljee S, Fleuren GJ, Jordanova ES: **Molecular backgrounds of ERAP1 downregulation in cervical carcinoma**. *Anal Cell Pathol (Amst)* 2015, **2015**:367837.

239. Hasim A, Abudula M, Aimiduo R, Ma JQ, Jiao Z, Akula G, Wang T, Abudula A: **Post-transcriptional and epigenetic regulation of antigen processing machinery (APM) components and HLA-I in cervical cancers from Uighur women**. *PLoS One* 2012, **7**(9):e44952.

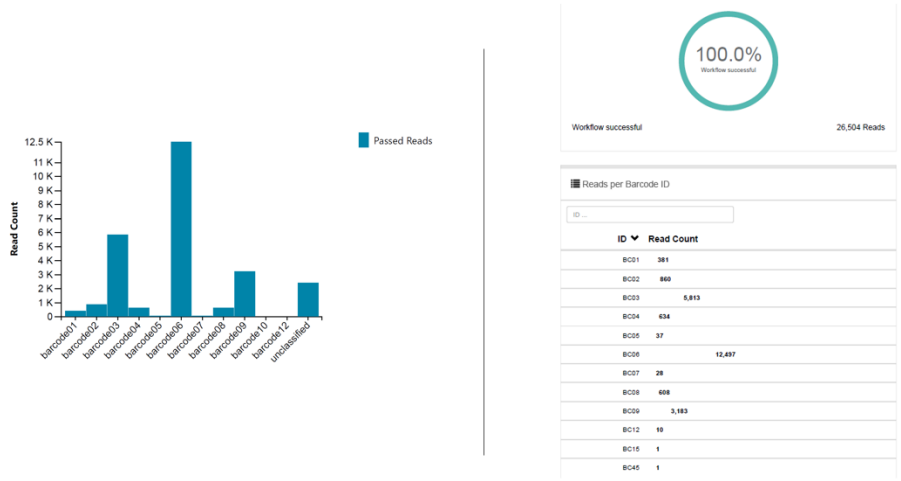
240. Steinbach A, Winter J, Reuschenbach M, Blatnik R, Klevenz A, Bertrand M, Hoppe S, von Knebel Doeberitz M, Grabowska AK, Riemer AB: **ERAP1 overexpression in HPV-induced malignancies: A possible novel immune evasion mechanism**. *Oncoimmunology* 2017, **6**(7):e1336594.

241. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A *et al*: **Proteomics. Tissue-based map of the human proteome**. *Science* 2015, **347**(6220):1260419.

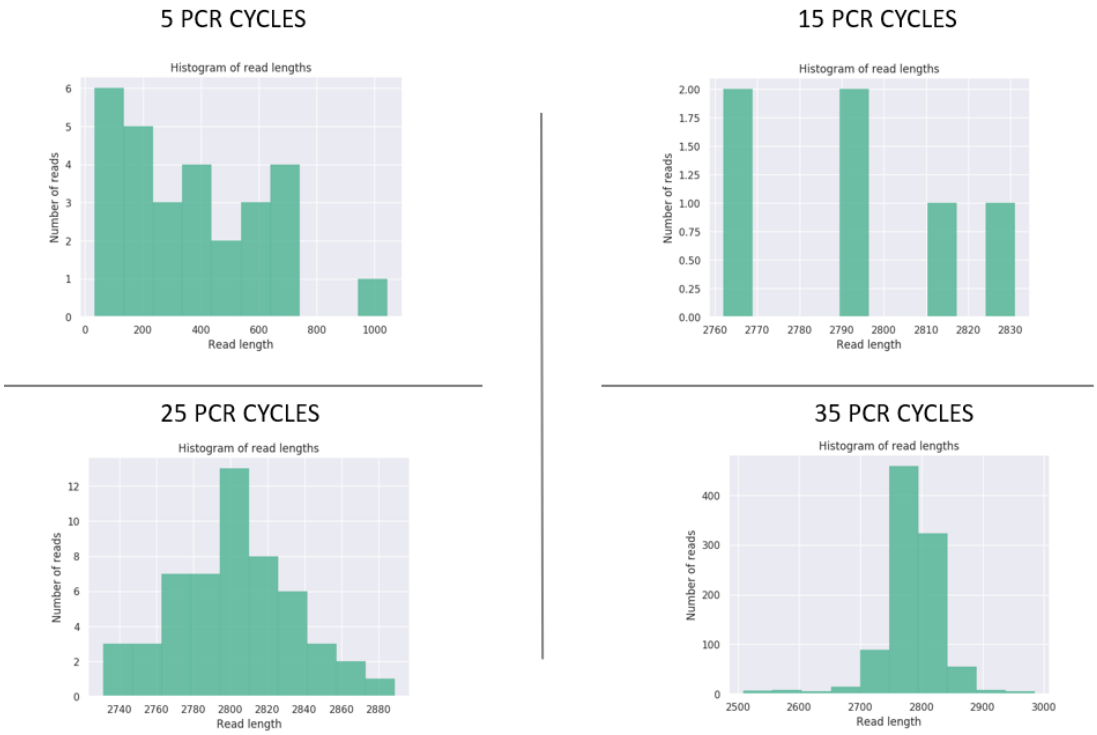
242. Chuanyin L, Xiaona W, Zhiling Y, Yu Z, Shuyuan L, Jie Y, Chao H, Li S, Hongying Y, Yufeng Y: **The association between polymorphisms in microRNA genes and cervical cancer in a Chinese Han population**. *Oncotarget* 2017, **8**(50):87914-87927.

8 Appendices

Appendix A



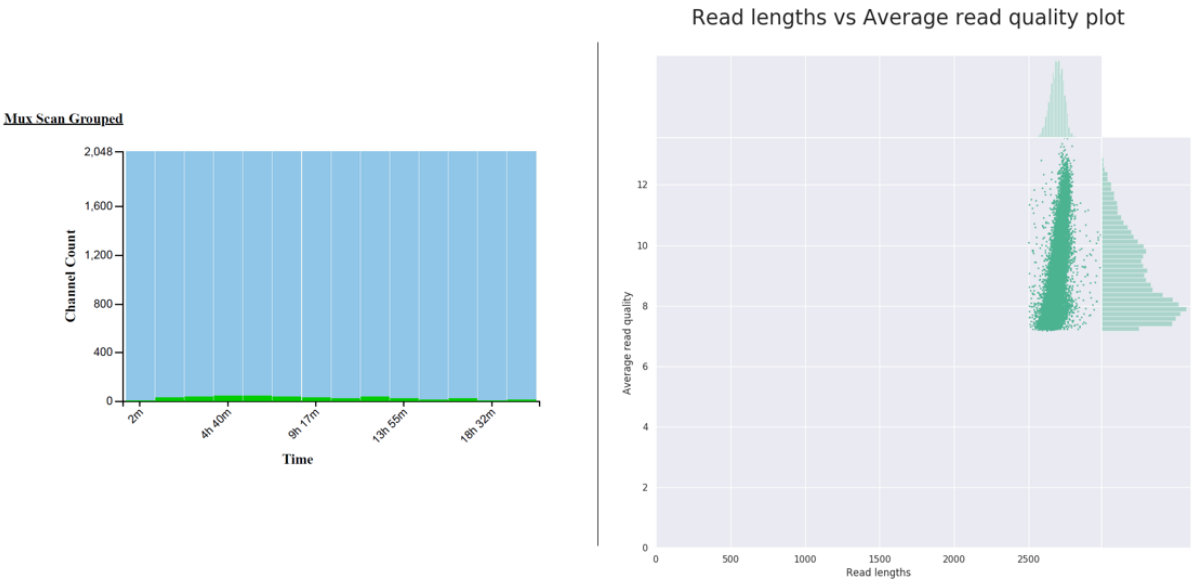
Barcoded read counts generated with MinKNOW for amplicons from 293T cells with ERAP1 amplified using 5 (BRC07), 15 (BRC08), 25 (BRC09) and 35 (BRC06) PCR cycles. The Epi2me report shows demultiplexed reads. A total of 26,504 reads were generated.



Read counts generated for ERAP1 from 293T amplified using 5, 15, 25 and 35 PCR cycles that passed the read filtering (2,500-3,000bp) with Nanoplot. A number of primer dimers of short length were identified for the

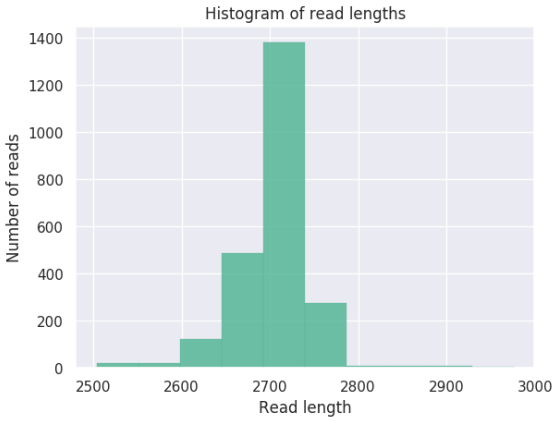
amplicon from 5 PCR cycles. Allotypes from 293T were successfully identified for ERAP1 amplified using 25 and 35 PCR cycles, as in the case of HeLa cells. The allotypes matched those generated in previous sequencing experiments for 293T and with Sanger sequencing.

Appendix B1



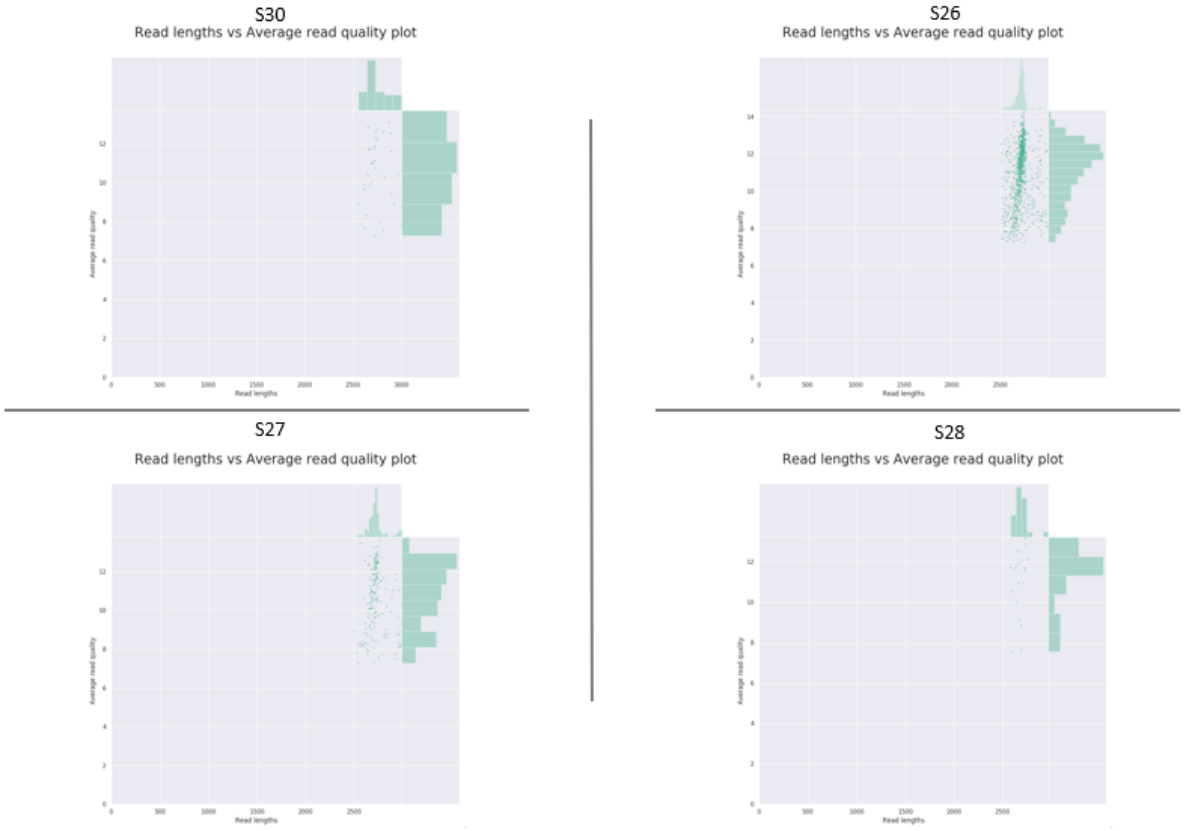
Sequencing run data generated for ERAP1 from patient S3 amplified using 35 PCR cycles. Data generated using MinKNOW (left) and Nanoplot (left). Experiment showed that the sequencing of ERAP1 (a single gene) is possible with just 50 pores available at the start of run but increasing the running time of sequencing. On the left, the read lengths and average quality scores are shown (reads passed filtering 2,500-3,000bp). The majority of reads matched the expected length of ERAP1 (2.7Kb) and allotypes were successfully identified.

Appendix B2



Read counts vs read length plot generated for S51 using Nanoplot. ERAP1 was amplified using 35 PCR cycles and the reads shown above that passed the read and score filter (2,500-3,000bp and min quality score is 7). ERAP1 allotypes were successfully identified for S51.

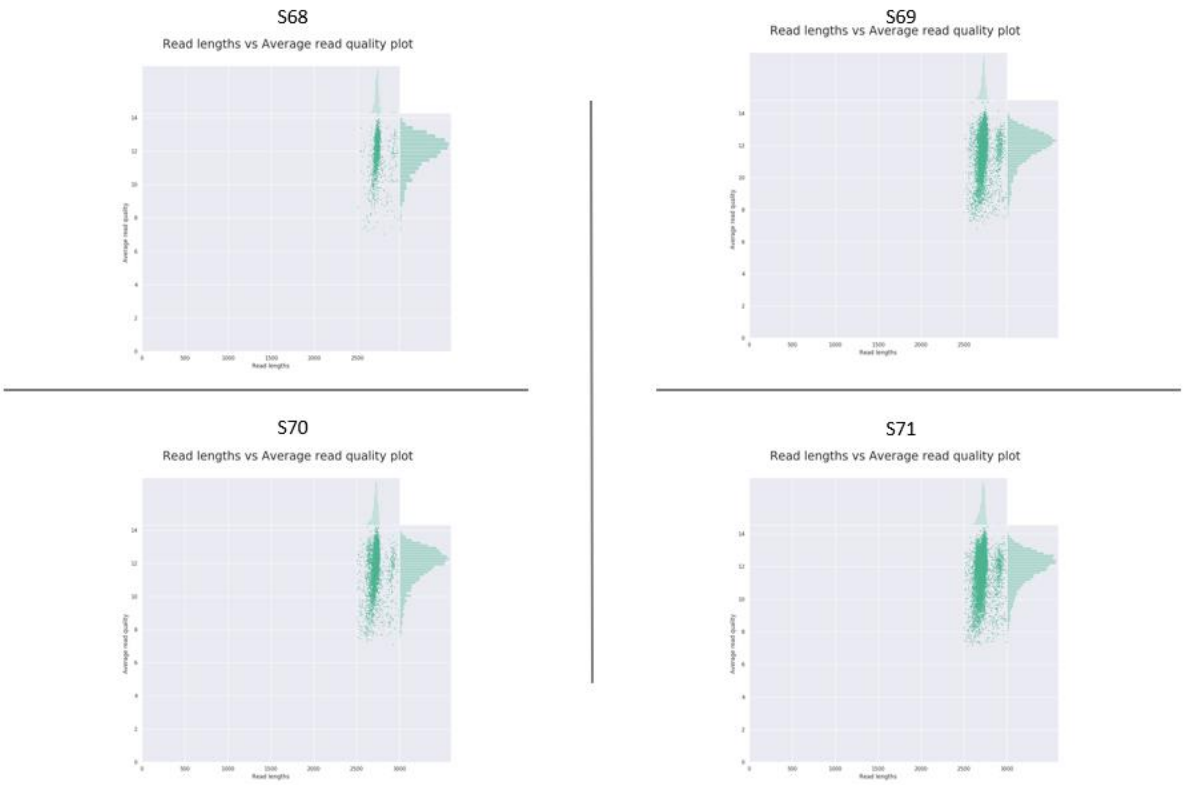
Appendix B3



Plots showing read lengths vs average read quality scores for S26, S27, S28 and S30 generated with NanoPlot

Plots showing read lengths vs average read quality scores for S26, S27, S28 and S30 that did not show visible ERAP1 amplification on the agarose gel. Allotypes were successfully identified from all four samples. Data generated with NanoPlot.

Appendix B4



Plots showing read lengths vs average read quality scores for S68, S69, S70 and S71 generated with NanoPlot

Plots showing read lengths vs average read quality scores for S26, S27, S28 and S30 that showed visible ERAP1 amplification on the agarose gel. Allotypes were successfully identified from all four samples. Data generated with NanoPlot.

Appendix C

Patient	CD8+ T cells per tumour mm2	CD8+/TIL status
S3	26.5	low
S59	44.1	low
S24	56.83	low
S15	57.29	low
S115	69.98	low

S74	74.31	low
S6	79.68	low
S72	122.51	low
S35	135.35	low
S27	144.52	low
S14	145.06	low
S18	151.35	low
S30	170.34	low
S1	170.88	low
S22	176.84	low
S89	197.04	low
S9	243.74	low
S62	254.4	low
S23	258.88	low
S75	286.21	low
S70	303.29	low
S41	313.44	low
S47	330.43	low
S17	351.86	mod
S79	361.17	mod
S85	365.74	mod
S78	368.39	mod
S52	393.47	mod
S12	403.21	mod
S81	434.31	mod
S106	441.26	mod
S51	446.82	mod
S87	450.49	mod
S77	481.24	mod
S5	491.03	mod
S2	509.28	mod
S10	520.96	mod
S64	528.89	mod
S73	529.38	mod
S109	540.2	mod
S54	563.47	mod
S88	574.06	mod
S29	635.08	high
S66	649.67	high
S82	674.46	high
S48	698.82	high
S40	722.57	high
S34	745.22	high
S46	752.58	high
S11	850.21	high

S107	887.47	high
S113	939.36	high
S110	977.07	high
S58	979.33	high
S50	1050.24	high
S60	1085.76	high
S44	1332.92	high
S111	1510.64	high
S55	1726.57	high
S102	2167.02	high
S71	2292.35	high
S56	2324	high
S43	2348.92	high
S42	3238.51	high
S28	3298.31	high
S19	3644.07	high