Genetic variants at chromosome 5q15 associated with immune-mediated diseases influence ERAP1 and ERAP2 function, expression and isoform profile

Biological Sciences

Genetics

Aimee Hanson[a], Thomas Cuddihy[a], Benoît Gautier[a], Katelin Haynes[a], Dorothy Loo[a], Craig Morton[b], Paul Leo[c], Gethin P. Thomas[a], Kim-Anh Lê Cao[a], Tony J. Kenna[c], Matthew A. Brown[c]

[a]*Diamantina Institute, The University of Queensland, Translational Research Institute, Princess Alexandra Hospital, QLD 4102, Australia*
[b]*St Vincent's Institute, St Vincent's Private Hospital, VIC 3065, Australia*
[c]*Institute of Health and Biomedical Innovation, Queensland University of Technology, Translational Research Institute, Princess Alexandra Hospital, QLD 4102, Australia*

Address correspondence to:
Matthew A. Brown
Institute of Health and Biomedical Innovation
Queensland University of Technology
Translational Research Institute
37 Kent Street
Woolloongabba
QLD 4102
Australia
Email: matt.brown@qut.edu.au

Key words: ankylosing spondylitis, aminopeptidase, isoform, eQTL, ERAP

Character count:

ABSTRACT

The endoplasmic reticulum aminopeptidases genes *ERAP1* and *ERAP2* function in the trimming of endogenously derived peptide precursors for human leucocyte antigen (HLA) mediated presentation to the immune system. The HLA presented peptidome intricately shapes the repertoire of lymphocytic cell populations within the body, educating the immune system to recognise foreign peptides while upholding tolerance to self. Concordantly, mutations in *ERAP1* and *ERAP2* have been observed in several immune mediated diseases in association with specific HLA alleles, implicating altered peptide handling and presentation as a prerequisite for inappropriate immune reactivity. Here we have demonstrated that chromosome 5q15 locus polymorphisms associated with the chronic inflammatory arthritis ankylosing spondylitis (AS) significantly influence the expression of these aminopeptidases on both the transcript and protein level. Disease-risk associated mutations in and around both genes consistently serve to increase gene expression. Furthermore, key risk-associated variants in *ERAP1* alter transcript splicing, leading to alternate expression of two distinct isoforms of the gene, and significant differences in the type and level of ERAP1 protein within the cell. These variants appear to underlie an overall elevation in ERAP1 protein expression by skewing protein production towards a more readily translated or stable isoform. In accordance with previous studies demonstrating that mutations that increase aminopeptidase activity predispose to immune disease, the elevated disease risk attributed to increased expression of ERAP1 supports the therapeutic notion of aminopeptidase inhibition to treat AS and other ERAP1 associated conditions.

SIGNIFICANACE STATEMENT

The immune system is able to differentiate between infected and healthy cells through recognising short peptides derived from endogenous or exogenously sourced proteins. The endoplasmic reticulum aminopeptidase enzymes ERAP1 and ERAP2 are responsible for trimming protein down to an appropriate size for cell-surface presentation on human leucocyte antigens (HLA) in a configuration immune-cells can recognise. Mutations in these enzymes have been associated with several immune-mediated diseases, potentially through altering peptide processing, causing the immune system to react adversely against peptides it should not. We have demonstrated that mutations predisposing to immune disease increase the amount of these enzymes in the cell. Understanding the molecular contribution of such mutations helps inform the development of novel therapeutic to treat individuals with these chronic conditions.

INTRODUCTION

Human leucocyte antigen (HLA) mediated presentation of endogenous peptides to $CD8^+$ T-cells is crucial in enabling the immune system to differentiate between self-derived and foreign antigens. Thus, aberrations in peptide presentation pathways may evoke misinformed immune cell reactivity against healthy host tissue, leading to disease. Aminopeptidases function in the processing of antigenic peptide precursors, cleaving peptides to generate molecules of optimal length and composition for loading onto the HLA. Disease risk associated polymorphisms in the endoplasmic reticulum aminopeptidase genes *ERAP1* and/or *ERAP2* are seen in several immune-mediated diseases, including ankylosing spondylitis (AS) (1, 2), psoriasis (3), Crohn's disease (CD) (4), Behçet's disease (5) and birdshot retinopathy (6). The potential implications of functional variants in the *ERAP* genes, encoding enzymes important in shaping the T-cell repertoire, make these enzymes promising targets for a therapeutic treatment that halts immune system retaliation against antigenic stimuli in patients with these conditions.

Robust genetic interactions exist between *ERAP1* and the HLA class 1 allele *HLA-B27* and *HLA-B40* in AS (1, 7), *HLA-Cw6* in psoriasis (3), and *HLA-B51* in Behçet's disease (5). Birdshot retinopathy is essentially restricted to HLA-A29 carriers, and thus whilst a strong association with *ERAP1/2* and *HLA-A29* exists with this disease, it is not statistically testable whether the two loci interact to cause disease. The synergism between HLA Class I loci and *ERAP1* is most pronounced in AS, with risk-associated polymorphisms in the *ERAP1* gene contributing to strong disease risk ($P = 4.4 \times 10^{-45}$ in the case of lead *ERAP1* SNP, rs30187) in *HLA-B27* carriers alone (1). AS is a highly heritable chronic arthritic condition that manifests primarily with inflammation involving joints of the pelvis and spine, ultimately leading to new bone formation and joint fusion. Between 80 and 95% of patients carry the major genetic risk factor *HLA-B27*, and over 100 further genetic loci have been identified with a confirmed disease association (1, 2, 8-11). Evidence points to the involvement of altered peptide handling in disease pathogenesis (12-14) potentially driving the production of an immunogenic HLA-B27 specific peptide repertoire that is

misinforming the immune response and/or impeding the folding, stability and function of the B27 molecule.

In *ERAP1* the association signal includes two independent AS-associated haplotypes (1). The strongest-associated haplotype is tagged by the SNP rs30187, which is thought to be the key AS-associated SNP on this haplotype (K528R; OR = 1.29, $P = 4.4 \times 10^{-45}$), protective alleles of which confer a ~40% decrease in enzymatic activity (1). The rs30187 risk allele associated with increased enzyme activity has been shown to alter the B27-bound peptidome by causing both elevated production and destruction of HLA-B27 epitopes (12, 15, 16). The second *ERAP1* haplotype is tagged by rs10050860 ($P = 9.28 \times 10^{-36}$). Although this and other SNPs in this region have also been shown to alter enzyme activity (1) the true functional mechanism driving disease association at this haplotype has not been confirmed. Evidence suggests that disease-associated polymorphisms in *ERAP1* also exert an expression effect (13), with risk genotype correlated with increased gene expression (17).

The genetic association at *ERAP2* is also complex, and extensive linkage disequilibrium (LD) across the locus has limited the resolution of fine-mapping. The AS-associated *ERAP2* nonsense mutation, rs2248374, is an expression quantitative trait loci (eQTL) at which the protective G allele results in complete loss of gene expression due to nonsense-mediated decay of an alternatively spliced transcript (18). This has been shown to influence the HLA-B27 peptide pool, leading to presentation of peptides with fewer basic N-terminal amino acids, and increasing the percentage of 9-mer peptides being presented (19), thought to be due to increased destruction of shorter peptides which otherwise would inhibit ERAP1 function (20). Functional studies of another AS-associated *ERAP2* SNP, rs2549782, have shown that the AS-protective allele results in a loss of catalytic activity (15). Given the role of these aminopeptidases in shaping the HLA-B27 peptidome, it is possible that variants increasing aminopeptidase expression may exacerbate the pathogenic effect of aberrant trimming by a highly active enzyme, such that a strong AS-risk haplotype results from the concordant effect of activity and expression-altering mutations in these genes.

In this study we used sequencing of mRNA transcripts (RNASeq) derived from *ERAP1* and

ERAP2 to identify mutations associated with varied gene and isoform expression in a cohort of 54 genotyped AS cases and 70 healthy controls. In addition to facilitating eQTL investigations, RNASeq allows interrogation of the transcriptional landscape of a gene, and enables the identification of splice-altering variants (sQTLs) that may drive the production of structurally and functionally distinct isoforms with potential phenotypic consequences. Despite extensive characterisation of alternate *ERAP2* transcripts that radically transform the global expression of this gene (18), the relevance of alternate isoforms of *ERAP1* has been infrequently studied in the broader literature, and not at all in the context of immune-mediated diseases exhibiting *ERAP1* associations. Elucidating how mutations in aminopeptidase genes contribute to immune system activation will be an important step in understanding the genetic origin of diseases with pathology linked to altered antigen presentation. Here we demonstrate that AS risk associated SNPs in *ERAP1* and *ERAP2* exert a substantial eQTL effect at both the transcript and protein level. Furthermore, risk variants in *ERAP1* appear to act by altering splicing of the encoded gene transcript, suggesting that alterations in the transcriptional profile of this gene may play a far greater role in disease than previously acknowledged. These findings contribute greatly to understanding of the molecular mechanisms underlying the association of this locus with disease.

RESULTS

Disease risk SNPs in *ERAP1* and *ERAP2* are associated with increased total gene expression. Total gene expression information (normalised transcript counts from RNA sequencing) was correlated with genotype at 1,221 genotyped and imputed SNPs across the chromosome 5q15 locus genes *ERAP1* and *ERAP2*. There were 113 disease associated SNPs identified as eQTLs significantly influencing total *ERAP1* expression (Table S1). At every SNP the disease risk allele was also associated with an increase in gene expression over the corresponding protective allele. Pairwise conditioning, conducted in order to dissect the variant/s exerting the greatest expression effect, identified 38 SNPs as variants controlling the expression effect at this locus (Table S1) (conditioning with any of these 38 abolished the significance of all 113 original eQTLs). The 38 variants span a region from between the 19th intron of the gene to

approximately 50kb upstream of the first exon (Fig. 1A). All 38 SNPs lost their disease association but retained their expression having conditioned on rs30187, which itself displayed a modest eQTL effect (IRR^2 = 1.185, P = 2.35 x 10^{-3}). This indicates that these SNPs are not directly AS-associated, but do themselves have, or tag, variants other than rs30187 which do have effects on ERAP1 transcription. Imputed SNP rs39840 (linked with genotyped SNPs rs27038 and rs27041, AS disease associations P = 5.70 x 10^{-19} and P = 4.93 x 10^{-19} respectively) was the most significant eQTL identified (IRR^2 = 1.343, P = 1.46 x 10^{-7}), at which risk homozygous genotype conferred a 34.3% increase in *ERAP1* expression over the protective homozygous genotype (Fig. 1B, Table S1). There was no significant difference in mean *ERAP1* expression between AS cases and healthy controls (Wilcoxon test).

ERAP2 expression has a bimodal distribution due to the effects of the rs2248374 SNP found in approximately 50% frequency in the population, and which leads to nonsense mediated decay of the transcribed RNA (18), complicating statistical analysis of expression from this gene. For the purpose of generating an expression distribution amendable to modelling, 38 rs2248374-GG individuals, expressing very low to no *ERAP2* transcript, were removed, and in subsequent analyses investigating potential eQTL effects at ERP2, rs2248374 genotype was corrected. There were 156 SNPs that exhibited a significant association with *ERAP2* expression. Further analysis identified 94 SNPs, conditioning on any of which controlled for the *ERAP2* expression association of all 155 other SNPs (Table S2). Tight linkage disequilibrium between SNPs across the entirety of *ERAP2* locus (Fig. 1C) made it difficult to pinpoint an approximate position for the second eQTL signal at this position. As at *ERAP1*, risk genotype was consistently correlated with increased gene expression at all expression-controlling SNPs within and about the *ERAP2* gene, with risk genotype at the top eQTLs resulting in a 148% elevation in expression (IRR^2 = 2.476, P = 1.7 x 10^{-6}; Fig. 1D). There was no significant difference in mean *ERAP2* expression between AS cases and healthy controls (Wilcoxon test). Given that the expression effect of the key disease associated SNP in *ERAP2* (rs2248374) has been thoroughly discussed in the literature, the remainder of this study focused on elucidating the SNP(s) responsible for variability in the expression of *ERAP1*, the gene most strongly implicated in AS after *HLA-B27*.

Alternate expression of two *ERAP1* isoforms is governed by genotype and associated with disease. Eleven unique *ERAP1* isoforms were assembled from the RNASeq data (Fig. S1A). Only two transcripts, differing in the inclusion of the C-terminal exon 20 (RM[STOP] → HDPEADATG[STOP]), Ensembl ID ENST00000443439 (941 amino acids, 19 exons; 19E) and ENST00000296754 (948 amino acids, 20 exons; 20E) (Fig. 2, Table S3), were consistently highly expressed in all individuals (Fig. S1B). We identified 158 disease-associated SNPs as significant eQTLs for both isoforms when tested independently for an expression effect on each (Table S4). At every SNP, the disease risk variant correlated with a significant increase in the expression of the isoform 19E and a significant decrease in the expression of the isoform 20E. The proportion of isoform 19E expressed by an individual was significantly associated with disease status (P = 0.047, T-test), with cases expressing a significantly greater proportion of the 19-exon isoform (19E proportion = 0.661) than controls (19E proportion = 0.619).

Genotype at the same 158 disease-associated SNPs (henceforth called sQTLs; splice-altering quantitative trait loci) had a significant effect on *ERAP1* isoform proportion (Table S4). The SNP rs7063 (disease association OR = 1.34, $P = 1.3 \times 10^{-41}$; (8)), situated between exon 19 and 20 (Fig. 2), exhibited the most significant effect on isoform proportion (OR = 1.16, $P = 1.2 \times 10^{-23}$), at which the risk (major allele) homozygotes expressed 105% more (IRR^2 = 2.05, $P = 8.7 \times 10^{-12}$) and 47% less (IRR^2 = *0.53*, $P = 1.0 \times 10^{-17}$) of the 19-exon and 20-exon isoforms respectively than protective homozygotes (Fig. 3A, 3B). rs7063 heterozygotes expressed similar amounts of the two transcripts (56% isoform 19E, 44% isoform 20E), whereas risk allele homozygotes expressed predominantly isoform 19E (71% on average) and protective allele homozygotes expressed predominantly isoform 20E (59% on average; Fig. 4A). None of the 58 most significant sQTLs exhibited a significant eQTL effect on total *ERAP1* expression, and all retained a significant disease association upon correction for rs30187 (Table S4) but lost this association upon correction for both rs30187 and rs10050860. These SNPs clustered within and about the C-terminus encoding exons 19 and 20 (Fig. 3C). Upon correction for rs10050860 alone, the lead sQTL rs7063 retained a significant disease association ($P = 1.4 \times 10^{-10}$). However, upon correction for rs7063, the

rs10050860 disease association was lost (P = 2.6 x 10^{-5}). LD between rs7036 and rs30187 was calculated as R^2 = 0.21, and between rs7063 and rs10050860 as R^2 = 0.55.

A strong risk haplotype results from co-occurrence of the risk genotype at rs30187 and the sQTL rs7063. The disease-protective haplotype, rs30187-C/rs7063-T (low ERAP1 enzyme activity/decreased expression of *ERAP1* isoform 19E, increased expression of isoform 20E), showed a disease association of P = 1.07 x 10^{-61} (Table 1) in 9069 genotyped AS cases and 13578 healthy controls (chi-squared test). This haplotype association exceeded the individual disease association of both SNPs in isolation (rs30187: P = 4.4 x 10^{-45}; rs7063: P = 1.3 x 10^{-41}).

Genotype at *ERAP1* sQTL rs7063 is strongly associated with alternate expression of two ERAP1 isoforms at the protein level. Protein derived from both the 19E and 20E isoforms of the *ERAP1* transcript was detected using mass spectrometry. SNP genotype at rs7063 was significantly associated with the protein expression of both isoforms independently, and total ERAP1 expression (taken as the sum of the expression of both isoforms; Fig. 5). Individuals homozygous for the risk allele at rs7063 expressed 3.52 times more isoform 19E (P = 3.08 x 10^{-6}) and 0.29 times less isoform 20E (P = 2.13 x 10^{-5}), and expressed 2.37 times more total ERAP1 protein overall (P = 5.61 x 10^{-5}) than those homozygous for the protective allele. At the protein level, isoform 19E was the predominantly expressed form of ERAP1 in all three genotype groups (Fig. 4B), expressed significantly higher on average than isoform 20E (P = 1.4 x 10^{-10}). There was no significant difference in the mean expression of either isoform at the protein level, or total ERAP1 protein levels, between AS cases and healthy controls.

***In silico* analysis of *ERAP1* isoform 20E predicts folding nature of the alternate protein.**

Ab initio protein structure modelling of the ERAP1 20E isoform suggests that the unique additional residues on the C-terminal end of the protein fold back across the surface of domain 4 (D4) of ERAP1, potentially forming salt-bridge interactions with residues Arg750 and Arg708 (Fig. 6).

DISCUSSION

The genetic association of variants in the chromosome 5q15 locus genes, *ERAP1* and *ERAP2*, with several immune mediated diseases make these enzymes important targets for functional investigation, particularly given their role in shaping the MHC-presented peptidome that instructs the immune system. This study showed that AS risk-associated variants in these genes consistently demonstrate an eQTL effect that serves to increase aminopeptidase expression. Indeed, the protective influence of the nonsense mutation rs2248374 (18) in *ERAP2* implies that some aspect of the enzyme's functional role in peptide presentation is linked to pathology in a way that can be ablated with loss of expression. It is expected that this is also the case at *ERAP1*, at which increased expression may be exacerbating the pathogenic effect of co-occurring missense mutations shown to alter enzymatic activity and peptide handling (12, 13). *ERAP1* mRNA and protein expression is elevated in lymphoid cell lines derived from individuals carrying disease susceptibility variants across *ERAP1* haplotypes (16, 17). The major finding of the current study suggests complex regulation of *ERAP1* expression by two independently acting variants; those tagged by rs30187 that influence global *ERAP1* expression, and those within the rs10050860 haplotype, influencing alternate splicing of two distinct transcripts. Our results suggest that the differences between these two forms of the *ERAP1* transcript, and encoded protein, are of critical importance, and influence the levels of functional enzyme in the cell. Given the cross-disease concordant action of *ERAP1* and *ERAP2* mutations in conditions such as psoriasis and Crohn's disease (21), these finding have relevance far beyond AS alone.

In a study published by Harvey *et al.* (13) in 2009, a strong positive correlation was noted between the strength of AS disease association for variants in *ERAP1* and their effect on *ERAP1* expression, previously reported in a genome wide microarray eQTL study (22). Similarly, we noted that the most significant *ERAP1* eQTLs are strongly associated with AS. Although linked within the haplotype containing the key functional variant, rs30187, these SNPs retained a significant expression effect upon rs30187 correction. This implied that, additional to the effects of rs30187 on ERAP1 activity,

there are variants driving altered enzyme expression also present within this haplotype. Joint inheritance of these alleles would mean that any alterations in the HLA-B27 peptidome due to ERAP1 hyperactivity would be exacerbated by ERAP1 overexpression. This could potentially either flood the HLA-B27 molecule with an altered peptide repertoire with the potential to trigger a CD8-mediated immune response (23), or lead to destruction of peptides that protect in some way against the immunological processes that lead to AS (24). Concordantly, increased transcript expression in association with AS risk genotype at SNPs in *ERAP2,* believed to act in concert with *ERAP1* to facilitate peptide trimming (25), supports the significance of increased aminopeptidase expression in disease pathology. The identification of disease-associated variants influencing *ERAP2* expression, independent of the rs2248374 null mutant, implies degrees of expression variation at this locus, such that the portion of the population that do express this aminopeptidase have varied expression phenotypes and thus disease susceptibility.

Of great interest was the identification that two highly expressed isoforms of *ERAP1*, the alternate expression of which, at both the transcript and protein level, correlated with genotype at disease risk SNPs, may play a substantial role in disease. These two predominant transcripts encoding different versions of the ERAP1 protein have been previously identified (26, 27) and annotated as ENST00000443439 (19 exons, 18 coding) and ENST00000296754 (20 exons, 19 coding) in the Ensembl database (28) (herein called isoform 19E and 20E). Isoform quantification on the transcript level revealed that 19E and 20E are both predominant forms *ERAP1*, expressed in all 124 individuals, and at high enough levels to potentially contribute to the functional output of the enzyme. We demonstrate here that alternate expression of these two transcripts is modulated by a genetic splice-interfering variant rather than by the common mechanism of alternate splicing, and that this variant is strongly AS associated. The observation that the disease risk allele at all significant *ERAP1* sQTLs correlated with increased expression of the 19 transcript and decreased expression of the 20E transcript suggests there may be some disease protective feature of 20E isoform specifically. SNPs exhibiting the most significant sQTL effect were localised around the exon/intron 19 boundary, around the first point of sequence variation between the two isoforms, and it has been previously noted that a number of the strongest disease associated

variants in *ERAP1* fall in this location with a potential involvement in splicing (13). The most significant sQTLs retained a strong disease association upon correction for rs30187, implying they are situated within the rs10050860 tagged haplotype; perhaps a splice site variant governing *ERAP1* isoform expression is of greatest functional importance at this position.

Mass spectrometric quantification of isoform expression showed that both the 19E and 20E forms of the ERAP1 protein were present at detectable levels in all 39 assayed samples. Likewise, observed differences in the molecular mass of ERAP1 isolated from various human cell lines confirms the co-occurrence of the two forms of the full-length protein (29). The finding that genotype at putative splice site SNPs significantly influences protein isoform expression, as at the transcript level, confirms that splice site mutations in *ERAP1* contribute to marked variability in the type of *ERAP1* protein expressed by an individual. The splice site variant rs7063 has an AS association of $P = 1.3 \times 10^{-41}$ (8) and 2.48×10^{-17} upon correction for rs30187. The observation that rs7063 retains association with disease upon correction for rs10050860 ($P = 1.4 \times 10^{-10}$), with which it is in LD ($R^2 = 0.55$), indicates it is the more important variant at the second *ERAP1* disease associated haplotype. The lead variant in the first associated haplotype, rs30187, likely contributes independent and additional functional effects. Of interest was the observation that rs7063 falls in the middle of the conserved transcription termination sequence (AA**T**AAA) in the 3′UTR of isoform 19E, a motif recognised by cleavage and polyadenylation specific factors involved in 3′-end transcription termination in mammals (30). The risk allele (major allele T on the reverse strand in the motif AA**T**AAA) would be expected to promote correct termination of the 19-exon form of the transcript, whereas the protective allele could potentially result in loss of termination sequence recognition, producing the 20-exon form of the transcript. Variant rs111774449, falling in the splice donor sequence at the 3′ end of exon 19 (31), is also a likely candidate for isoform switching but was not genotyped or imputed in this study. The strong disease-associated haplotype that arises from co-expression of rs30187 and rs7063 risk alleles is evidence that *ERAP1* expression dynamics contribute to the pathogenicity afforded by increased enzyme activity. The enhanced genetic contribution of SNP haplotypes at the chromosome 5q15 locus has been previously noted, with carriage of both the two

disease-associated haplotypes increasing risk by ~4 fold, a far greater risk than the additive effects of either variant alone (1.2-1.3 fold for either variant in isolation) (1). The current study provides a potential functional mechanism for this association.

The observation that average isoform 20E protein expression was significantly lower than isoform 19E expression across all individuals implies differences in the cellular handing of the two forms of the ERAP1 protein. Despite approximately equal levels of the 19E and 20E transcripts in rs7063 heterozygotes, 81% of the ERAP1 protein detected in these individuals was of the 19E form. This raises the question, can loss of isoform 20E post transcription be attributed to the sequence or structural variation between the two isoforms, perhaps contributing to less efficient translation of isoform 20E, or misfolding and degradation of the protein, with subsequent lack of functionality? If this is the case, the mechanism of protection conferred by the disease protective genotype at splice site SNPs may arise due to skewed expression towards the 20E transcript, subsequent loss of this isoform at the protein level, and thus a decrease in the overall level of functional ERAP1 available to the cell.

To date, all bar one of the published ERAP1 structures have been of the 19E isoform sequence. In the one study of the ERAP1 20E isoform sequence (Protein Data Bank entry 3MDJ; (32)) residues beyond Gln934 were not resolved in the experimental electron density and are absent from the structure. As this includes the C-terminal modification generated by the addition of exon 20, *ab initio* modelling of these residues was required. The area of the D4 with which the exon-20 derived residues interact is effectively invariant structurally between the open and closed forms of ERAP1, implying that the presence of the isoform 20E C-terminal extension is unlikely to alter the open to closed form dynamic of ERAP1 believed to be required for substrate binding and release (32, 33). Tertiary structure modelling, however, is unable to predict whether residues of the C-terminal extension interfere with folding of the mature protein. The UCSC 100 vertebrates conservation track (34) shows very low sequence conservation in exon 20 relative to the other 19 exons of *ERAP1*, and, indeed, other coding exons in general. This suggests that the 19-exon ERAP1 is the predominant functional form of the enzyme and that the appended seven amino acids in isoform 20E are either not of any relevance to the mature enzyme and

thus tolerant of substituting mutations, or that the protein itself is being removed by the cell before it can actively contribute to peptide trimming.

To date, differential gene expression studies in AS have found no evidence for significant changes in the expression of genes encoding disease-associated aminopeptidases in patients relative to healthy controls (35, 36). However, it has become apparent in this study that genetics is an important driver of expression variability in these genes. Here, we demonstrate the ability of eQTL studies to identify important information about the significance of genotype-driven altered gene expression in disease, which would go unnoticed by differential expression studies that pool samples of different genotypes. Foremost, it must be acknowledged that dynamic changes in the isoforms derived from a transcriptional unit can be far more insightful and relevant than measures of total gene expression, adding layers of complexity to the functional output of a gene. It appears that increased aminopeptidase expression governed by altered transcript dynamics at both *ERAP1* and *ERAP2*, as previously acknowledged, is a key mechanism driving the degree to which these enzymes contribute to immune-mediated disease. Given such findings, it is becoming increasingly apparent that ERAP inhibition may be a promising therapeutic approach for treating diseases harbouring *ERAP* associations, particularly given the likelihood that alteration in some aspect of normal peptide-MHC presentation is the stimulus for harmful malfunction of the immune response.

Materials and Methods

Ethics

Human ethics approval was granted by the Princess Alexandra Hospital and the The University of Queensland Ethics Committees (ethic no. Metro South HREC/05/QPAH/221 and UQ 2006000102). Written informed consent was received from all participants prior to inclusion in the study.

Sample Selection

54 patients diagnosed with AS according to the modified New York criteria (37) and 70 healthy subjects

were included in this study (Table S5). Peripheral venous blood was collected from AS patients at the Princess Alexandra Hospital Ankylosing Spondylitis Clinic and peripheral blood mononuclear cells (PBMCs) extracted using a standard density gradient centrifugation over Ficoll-Paque PLUS (GE Healthcare, Uppsala, Sweden). Cell viability was determined by 0.4% Trypan blue (Gibco, Mulgrave, VIC) exclusion. Cells were cryopreserved at 10^7 per mL in freezing media, consisting of heat-inactivated fetal calf serum (Gibco, Mulgrave, VIC) with 10% DMSO (D8418, Sigma), until required.

Genotyping, Imputation and Disease Association Analysis

Sample genotyping was conducted as part of the 2013 Immunochip study of 9049 AS cases and 13607 healthy controls (8). SNP disease association *P*-values and risk alleles reported in this study were derived from the Immunochip published data, calculated as per the outlined methodology, with $P = 5 \times 10^{-8}$ used as the threshold for genome wide significance. *P*-values for SNP disease associations upon conditioning with rs30187 were derived using logistic regression on the complete Immunochip data set in PLINK, with inclusion of rs30187 genotype as an additional covariate. Imputation of SNPs within an R^2 window of 0.1 with rs30187 (*ERAP1*) or rs2910686 (*ERAP2*), determined using PLINK (http://pngu.mgh.harvard.edu/purcell/plink/; (38), was conducted using the 1000 Genomes Phase 1 reference panel in the subset of 124 individuals used in this study. Phasing was conducted using SHAPEIT (39) and imputation using IMPUTE2 (40) with the 'info' metric used to remove poorly imputed SNPs (info < 0.5). A total of 1221 SNPs spanning *ERAP1* and *ERAP2* were used in the final analysis.

RNA Sequencing

Sample RNA was extracted from PBMC samples, reverse transcribed, prepared for sequencing using Illumina TruSeq Standard Total RNA Library Prep Kit, and sequenced on an Illumina platform. Sequence reads were mapped to the human genome NCBI Build 37 (Hg19) using TopHat version 2.0.6 employing the Bowtie 2 version 2.0.2 aligner (41, 42). Aligned reads were supplied to HTSeq (43) to generate read

counts per gene. Isoform-specific counts were generated using the Cufflinks suit (44) and assembled into a reference transcriptome to which RSEM (45) was used to align reads to generate isoform counts. Gene and isoform counts were normalised using DESeq2 (46) to correct for variability in sequence depth between individuals.

Statistical Regression for eQTL Detection

All statistical analyses were performed using the statistical software, R (47). A generalised linear mixed effects (GLME) model was applied using the lme4 package (48) to test for the effect of genotype on gene or isoform expression under a negative binomial distribution (logarithmic link), correcting for the fixed effect of patient sex and random effect of sequencing batch:

```
glmer(expression ~ genotype + sex + (1|batch. no.), data = data, family = neg.bin)
```

The model AIC (Akaike information criterion) value, residual versus fitted and QQ plots were used to test the suitability of the chosen distribution relative to others tested. Disease and HLA-B27 status exerted no significant expression effect when tested during model fitting and so were excluded from the final GLME. The change in gene or isoform expression attributed to the addition of a minor allele was quantified as the exponential of the genotype coefficient returned from the model (referred to as the incident rate ratio (IRR)), with IRR^2 comparable to the fold change in expression between individuals with minor allele count 2 over count 0. *P*-values were adjusted for multiple testing using a Benjamini and Hochberg false discovery rate correction (49). Significant SNPs were cross-referenced to the set of significant disease associated SNPs identified by Cortes *et al.* (8) at this locus, and the direction of effect of the disease risk allele inferred.

Pairwise conditioning was conducted by adding each significant eQTL individually as a covariate into the statistical model and re-testing the expression effect of all remaining SNPs:

```
glmer(expression ~ genotype + sex + conditional SNP genotype + (1|batch. no.), data
                                = data, family = neg.bin)
```

A Wilcoxon test was used to test for any difference in gene expression between cases and controls.

Statistical Regression for sQTL Detection

For two highly expressed isoforms of *ERAP1*, isoform 19E proportion was calculated for each individual based on isoform counts (isoform 19E / (isoform 19E + 20E)), and modelled using a normal distribution as confirmed using a Shapiro-Wilk test. A two-sample T-tailed was used to assess the difference in transcript proportion between cases and controls. A linear model was applied to test for the effect of genotype on transcript proportion while correcting for any disease status effect:

```
lm(transcript proportion ~ genotype + status, data = data)
```

SNPs that were determined to be significant eQTLs for both transcripts in isolation, and significant sQTLs with an effect on isoform proportion, were cross-referenced to the set of significant disease-associated SNPs identified by Cortes *et al.* (8) at this locus, and the direction of the effect of the disease risk allele inferred. A Wilcoxon test was used to test for any difference in isoform expression between cases and controls.

Haplotype Analysis

A chi-squared test assessed the disease association of a haplotype containing the ERAP1 activity altering SNP rs30187 and rs7063, the top sQTL identified in this analysis, using the European cohort of 9,069 AS cases and 13,578 healthy controls genotyped in the Immunochip investigation (8).

Mass Spectrometry for Quantification of ERAP1 Isoform and Total Protein Expression

PBMCs (approximately 10^7 cells) from 39 samples (mixture of cases and controls; 15 homozygous for the risk (major) allele at rs7063, 15 heterozygous and 9 homozygous for the protective allele; Table S5) were spun at 700xg for three minutes and washed twice with PBS. Cells were lysed with SDT-lysis buffer (4% (w/v) SDS, 100mM Tris/HCL pH 7.6, 0.1M DDT and protease inhibitor *cOmplete ULTRA* mini protease tablets, EDTA-free, prepared as specified). Samples were sonicated to shear DNA, heated at 95°C for 5 minutes and spun at 16000xg for 5 minutes prior to collection of cell lysate. Protein concentration in lysate was measured using the Thermo Scientific™ Pierce™ 660nm Protein Assay. 50µL of protein extract was digested using the FASP™ Protein Digestion Kit (Expedeon) as per the

recommended protocol, substituting ammonium chloride with 0.5mM Tris/HCL pH7.6. FASP digestion solution was substituted with 1µL activated ArgC enzyme (Promega), prepared as per the supplier's protocol in the specified buffers. The following peptides were targeted for detection and quantification in samples via HPLC-QQQ mass spectrometry (Table S6 for details):

ERAP1 isoform 19E: VW**L**QSEKLER

ERAP1 isoform 20E: VWLQSEK**L**EHDPEADATG

ERAP1 common peptide: NP**V**GYPLAWQFLR

Samples were spiked with 0.5fmol weighted peptide mix standard for quantification, and 8µL was subjected to mass spectrometry. Standard curve was generated by injecting 1µL peptide mix standards at 0.1fmol/µ 0.2fmol/µL, 0.4fmol/µL, 0.6fmol/µL, 0.8fmol/µL.

Statistical analysis of Mass Spectrometry Data

Expression of both ERAP1 isoforms and total ERAP1 expression (isoform 19E + isoform 20E) was confirmed to follow a normal distribution with a Shapiro-Wilk test. Subsequently, a linear model was used to test for the effect of rs7063 genotype on protein expression for each isoform and total protein, correcting for the effect of disease status and HLA-B27 status. A *t*-test was used to test for any difference in isoform or total ERAP1 expression between cases and controls.

Protein modelling of ERAP1 Isoform 20E

The sequence of human ERAP1 isoform 20E was taken from the Uniprot sequence entry Q9NZ08-2. The sequence was submitted to the online modelling server Phyre2 (50) in intensive mode. A set of four templates were used by the server to generate the final model, with 92% of the sequence modelled at >90% confidence. A total of 80 residues (residues 1-45, 487-513 and 941-948) were built *ab initio*, including residues in the C-terminal extension caused by exon 20.

Acknowledgements

TO ADD

1. Evans DM, *et al.* (2011) Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nature genetics* 43(8):761-767.
2. Wellcome Trust Case Control C, *et al.* (2007) Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nature genetics* 39(11):1329-1337.
3. Strange A, *et al.* (2010) A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nature genetics* 42(11):985-990.
4. Franke A, *et al.* (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature genetics* 42(12):1118-1125.
5. Kirino Y, *et al.* (2013) Genome-wide association analysis identifies new susceptibility loci for Behcet's disease and epistasis between HLA-B*51 and ERAP1. *Nature genetics* 45(2):202-207.
6. Kuiper JJ, *et al.* (2014) A genome-wide association study identifies a functional ERAP2 haplotype associated with birdshot chorioretinopathy. *Human molecular genetics* 23(22):6081-6087.
7. Cortes A, *et al.* (2015) Major histocompatibility complex associations of ankylosing spondylitis are complex and involve further epistasis with ERAP1. *Nat Commun* 6:7146.
8. International Genetics of Ankylosing Spondylitis C, *et al.* (2013) Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nature genetics* 45(7):730-738.
9. Australo-Anglo-American Spondyloarthritis C, *et al.* (2010) Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. *Nature genetics* 42(2):123-127.
10. Ellinghaus D, *et al.* (2016) Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nature genetics*.
11. Danoy P, *et al.* (2010) Association of variants at 1q32 and STAT3 with ankylosing spondylitis suggests genetic overlap with Crohn's disease. *PLoS genetics* 6(12):e1001195.

12. Seregin SS, *et al.* (2013) Endoplasmic reticulum aminopeptidase-1 alleles associated with increased risk of ankylosing spondylitis reduce HLA-B27 mediated presentation of multiple antigens. *Autoimmunity* 46(8):497-508.

13. Harvey D, *et al.* (2009) Investigating the genetic association between ERAP1 and ankylosing spondylitis. *Human molecular genetics* 18(21):4204-4212.

14. Benjamin RJ, Abrams JR, Parnes JR, Madrigal JA, & Parham P (1992) Polymorphic specificity of Q1/28, a monoclonal antibody that preferentially reacts with free class I heavy chains. *Immunogenetics* 37(1):73-76.

15. Evnouchidou I, *et al.* (2012) A common single nucleotide polymorphism in endoplasmic reticulum aminopeptidase 2 induces a specificity switch that leads to altered antigen processing. *Journal of immunology* 189(5):2383-2392.

16. Sanz-Bravo A, Campos J, Mazariegos MS, & Lopez de Castro JA (2015) Dominant role of the ERAP1 polymorphism R528K in shaping the HLA-B27 Peptidome through differential processing determined by multiple peptide residues. *Arthritis & rheumatology* 67(3):692-701.

17. Costantino F, *et al.* (2015) ERAP1 gene expression is influenced by non-synonymous polymorphisms associated with predisposition to spondyloarthritis. *Arthritis & rheumatology*.

18. Andres AM, *et al.* (2010) Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. *PLoS genetics* 6(10):e1001157.

19. Martin-Esteban A, Guasp P, Barnea E, Admon A, & Lopez de Castro JA (2016) Functional interaction of the ankylosing spondylitis associated endoplasmic reticulum aminopeptidase 2 with the HLA-B*27 peptidome in human cells. *Arthritis and rheumatism* 68(10):2468-2475.

20. Garcia-Medel N, *et al.* (2014) Peptide handling by HLA-B27 subtypes influences their biological behavior, association with ankylosing spondylitis and susceptibility to endoplasmic reticulum aminopeptidase 1 (ERAP1). *Mol Cell Proteomics* 13(12):3367-3380.

21. Parkes M, Cortes A, van Heel DA, & Brown MA (2013) Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nature reviews. Genetics* 14(9):661-673.

22. Dixon AL, *et al.* (2007) A genome-wide association study of global gene expression. *Nature genetics* 39(10):1202-1207.

23. Benjamin R & Parham P (1992) HLA-B27 and disease: a consequence of inadvertent antigen presentation? *Rheum Dis Clin North Am* 18(1):11-21.

24. Kenna TJ & Brown MA (2013) Immunopathogenesis of ankylosing spondylitis. *Int J Clin Rheumatol* 8(2):265-274.

25. Saveanu L, *et al.* (2005) Concerted peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum. *Nature immunology* 6(7):689-697.

26. Hattori A, Matsumoto H, Mizutani S, & Tsujimoto M (1999) Molecular Cloning of A-LAP highly related to placental leucine aminopeptidase/oxytocinase. *Journal of Biochemistry* 125:931-938.

27. Hattori A, Matsumoto K, Mizutani S, & Tsujimoto M (2001) Genomic Organisation of A-LAP gene and its relationship to the placental leucine aminopeptidease:ozytocinase. *Journal of Biochemistry* 130(2):235-241.

28. Yates A, *et al.* (2016) Ensembl 2016. *Nucleic acids research* 44(D1):D710-716.

29. Yousaf N, *et al.* (2015) Differences between disease-associated endoplasmic reticulum aminopeptidase 1 (ERAP1) isoforms in cellular expression, interactions with tumour necrosis factor receptor 1 (TNF-R1) and regulation by cytokines. *Clinical and experimental immunology* 180(2):289-304.

30. Richard P & Manley JL (2009) Transcription termination by nuclear RNA polymerases. *Genes & development* 23(11):1247-1269.

31. National Center for Biotechnoligy Information (2015) NCBI Human Genome Browser. (U.S National Library of Medicine).

32. Nguyen TT, *et al.* (2011) Structural basis for antigenic peptide precursor processing by the endoplasmic reticulum aminopeptidase ERAP1. *Nature structural & molecular biology* 18(5):604-613.

33. Kochan G, *et al.* (2011) Crystal structures of the endoplasmic reticulum aminopeptidase-1 (ERAP1) reveal the molecular basis for N-terminal peptide trimming. *Proceedings of the National Academy of Sciences of the United States of America* 108(19):7745-7750.

34. University of California Santa Cruz (UCSC) (2015) UCSC Genome Browser: Vertebrate Multiz Alignment & Conservation (100 Species).

35. Duan R, Leo P, Bradbury L, Brown MA, & Thomas G (2010) Gene expression profiling reveals a downregulation in immune-associated genes in patients with AS. *Annals of the rheumatic diseases* 69(9):1724-1729.

36. Assassi S, *et al.* (2011) Whole-blood gene expression profiling in ankylosing spondylitis shows upregulation of toll-like receptor 4 and 5. *The Journal of rheumatology* 38(1):87-98.

37. van der Linden S, Valkenberg HA, & Cats A (1984) Evaluation of Diagnostic Criteria for Akylosing Spondylitis. *Arthritis and rheumatism* 27(4):361-368.

38. Purcell S, *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 81(3):559-575.

39. Delaneau O, Marchini J, & Zagury JF (2012) A linear complexity phasing method for thousands of genomes. *Nature methods* 9(2):179-181.

40. Howie BN, Donnelly P, & Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* 5(6):e1000529.

41. Kim D, *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* 14(4):R36.

42. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods* 9(4):357-359.

43. Anders S, Pyl PT, & Huber W (2014) HTSeq - A Python framework to work with high-throughput sequencing data. *bioRxiv*.

44. Trapnell C, *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* 28(5):511-515.

45. Li B & Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* 12:323.

46. Love MI, Huber W, & Anders S (2014) Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv*.

47. Team RC (2015) R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria).

48. Bates D, Maechler M, Bolker B, & Walker S (2015) lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1.-8.

49. Beniamini Y & Hochberg Y (1995) Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Satistical Society* 57(1):289-300.

50. Kelley LA, Mezulis S, Yates CM, Wass MN, & Sternberg MJ (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols* 10(6):845-858.

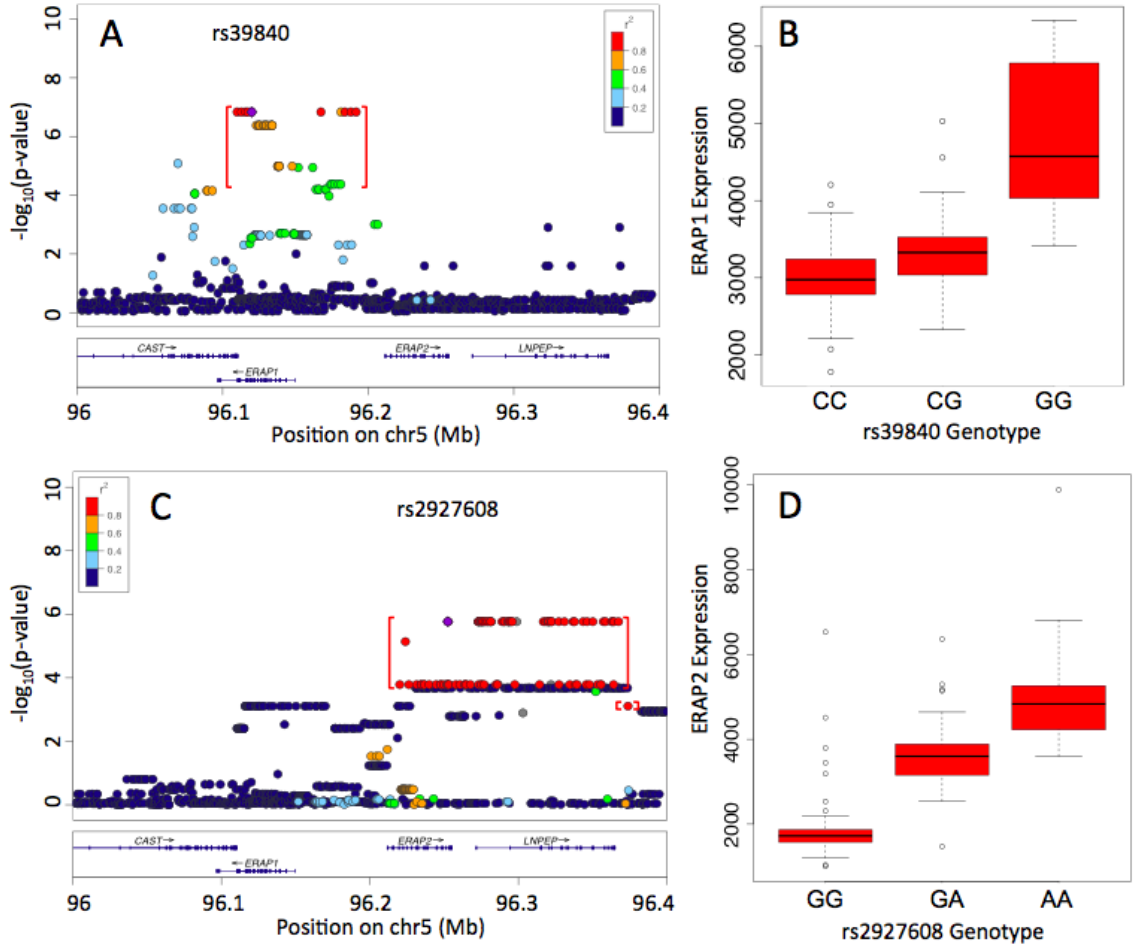

Fig. 1: Genotypes across chromosome 5q15 locus SNPs correlate with *ERAP1* and *ERAP2* gene expression. **(A)** Locus plot showing P-values for the *ERAP1* expression effect of chromosome 5q15 locus SNPs plotted against genomic region, with SNPs controlling for the eQTL effect bracketed in red. Lead *ERAP1* eQTL rs39840 is marked in purple and SNPs are coloured according to LD with this variant as per the colour key at top right. **(B)** Lead *ERAP1* eQTL SNP (rs39840) genotype verses *ERAP1* total gene transcript expression derived from normalised RNASeq counts; the AS risk allele is G. **(C)** Locus plot showing P-values for the *ERAP2* expression effect of chromosome 5q15 locus SNPs plotted against genomic region, with SNPs controlling for the eQTL effect bracketed in red. Representative lead *ERAP2* eQTL rs2927608 is marked in purple and SNPs are coloured according to LD with this variant as per the colour key at top right. **(D)** Representative top *ERAP2* eQTL SNP (rs2927608) genotype verses *ERAP2* gene expression derived from normlised RNASeq counts; the AS risk allele is A.

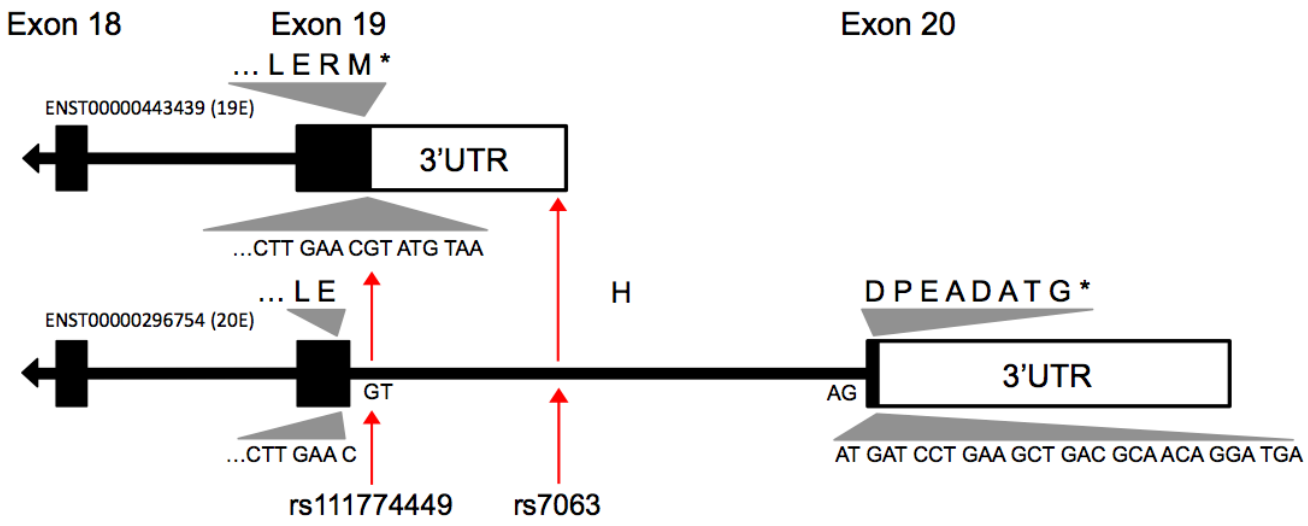

Fig. 2: Graphical representation (not to scale) of sequence and structural variation at the 3′ end of two *ERAP1* transcripts found to be highly expressed in PBMCs. Isoform ENST00000443439 has 19 exons (19E) and isoform ENST00000296754 20 exons (20E) (an additional C-terminal encoding exon) with an alternate 3′UTR sequence. Encoded amino acids are indicated above exons 19 and 20 (black boxes, numbered at top of figure) and encoding codons beneath, beginning two amino acids prior to the first point of variation between the transcripts (RM* → HDPEADATG*). The codon for amino acid H in isoform 20E spans a splice junction. Red arrows indicate the location of two putative splice interfering SNPs; rs111774449 (G/A; Arg/His) at the exon-intron 19 interface, and rs7063 (A/T) falling within the transcription termination motif in the 3′UTR of isoform 19E, identified as the most significant *ERAP1* sQTL from the statistical analysis conducted on RNASeq data.

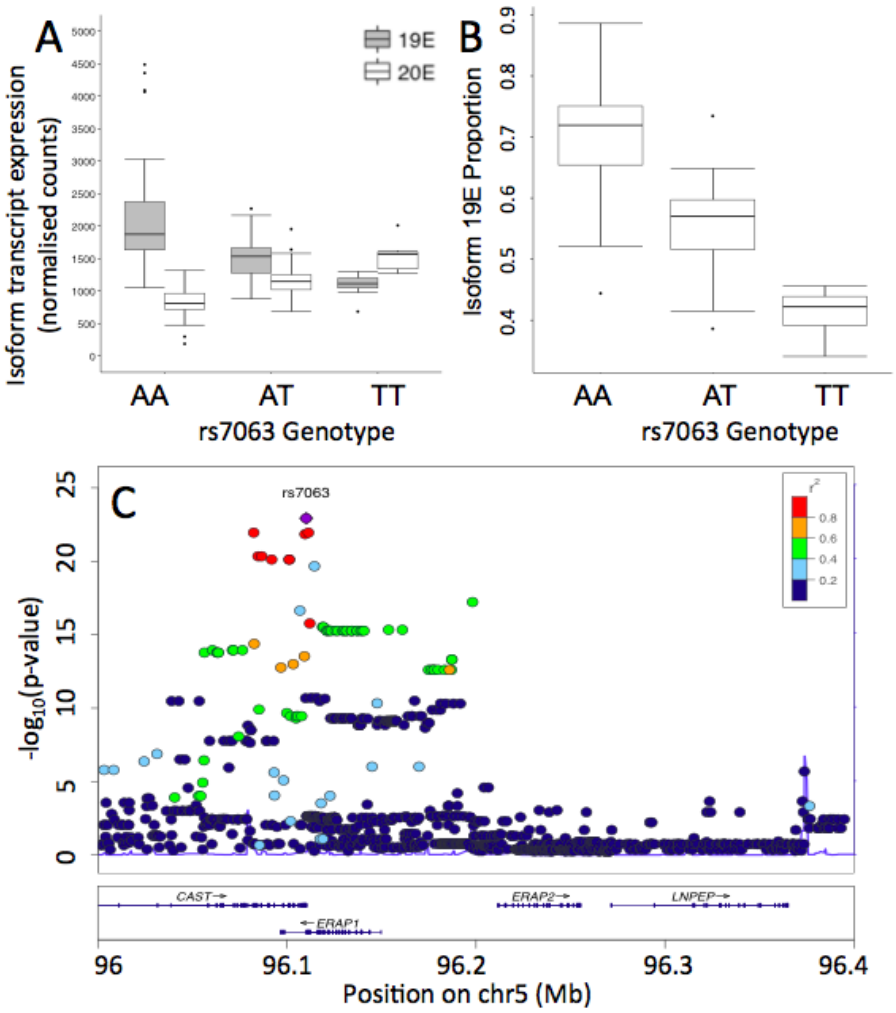

Fig. 3: Genotypes across chromosome 5q15 locus SNPs correlate with expression of two isoforms of the *ERAP1* gene. **(A)** Transcript expression of *ERAP1* isoforms 19E (gray) and 20E (white) split by genotype at the top *ERAP1* sQTL (rs7063). Individuals homozygous for the AS-risk allele (A) express significantly more isoform 19E and significantly less of isoform 20E than individuals homozygous for the protective allele. **(B)** *ERAP1* isoform 19E proportion split by genotype at the top *ERAP1* sQTL (rs7063). **(C)** Locus plot showing *P*-value for the effect of chromosome 5q15 locus SNPs on *ERAP1* isoform 19E proportion (19E expression over the sum of 19E and 20E expression) plotted against genomic region. The lead sQTL, rs7063 is marked in purple and SNPs are coloured according to LD with this variant as per the colour key at top right.

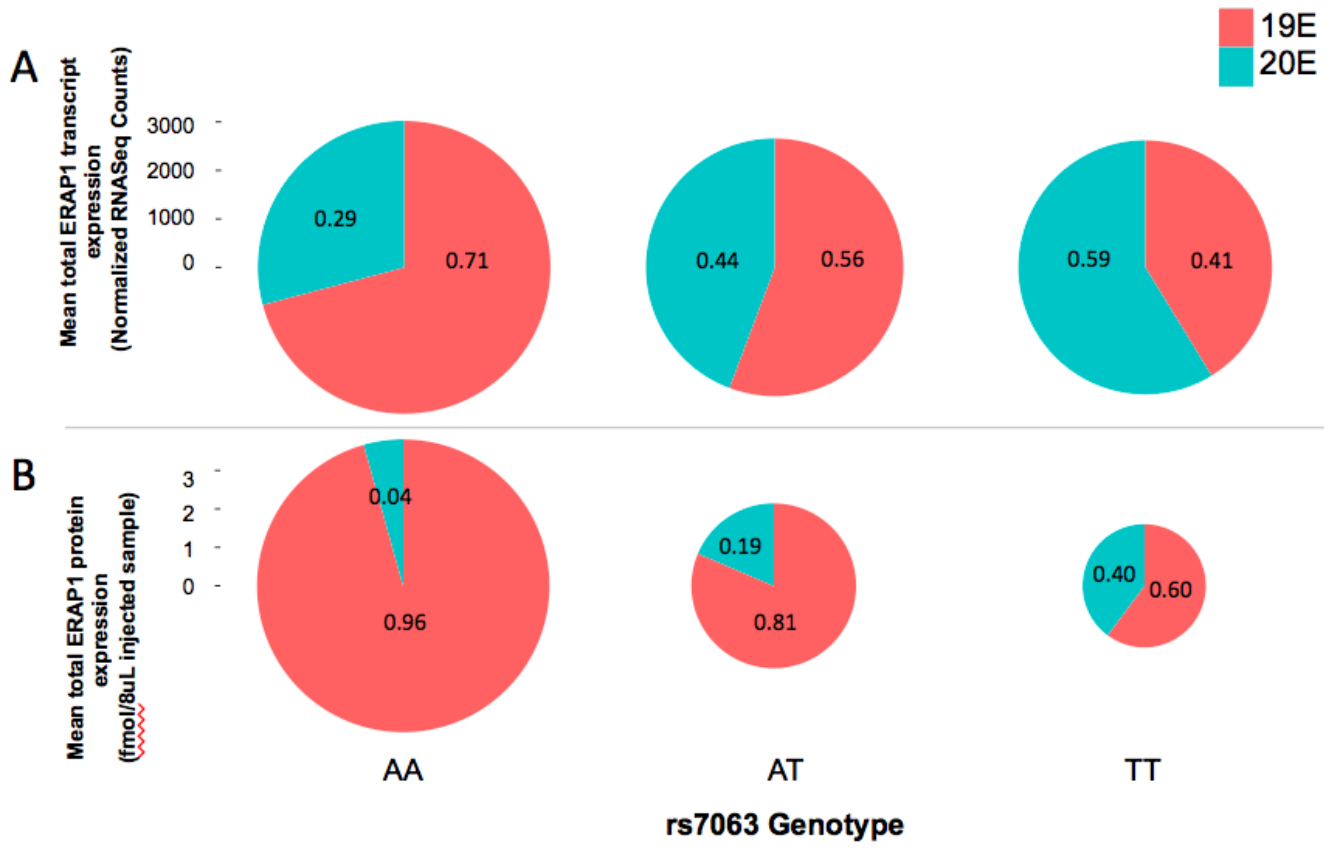

Fig. 4: Pie graph representing the proportion of total *ERAP1* expression derived from isoform 19E (red) and 20E (green) on the transcript level **(A)** in comparison to the protein level **(B)** for individuals split by genotype at the lead sQTL rs7063 (risk allele A). The radius of each graph denotes mean total *ERAP1* expression on the transcript or protein level, quantified on the axis to the left of the figure. There are 124 individuals in the RNASeq cohort and 39 in the mass spectrometry (protein quantification) cohort.

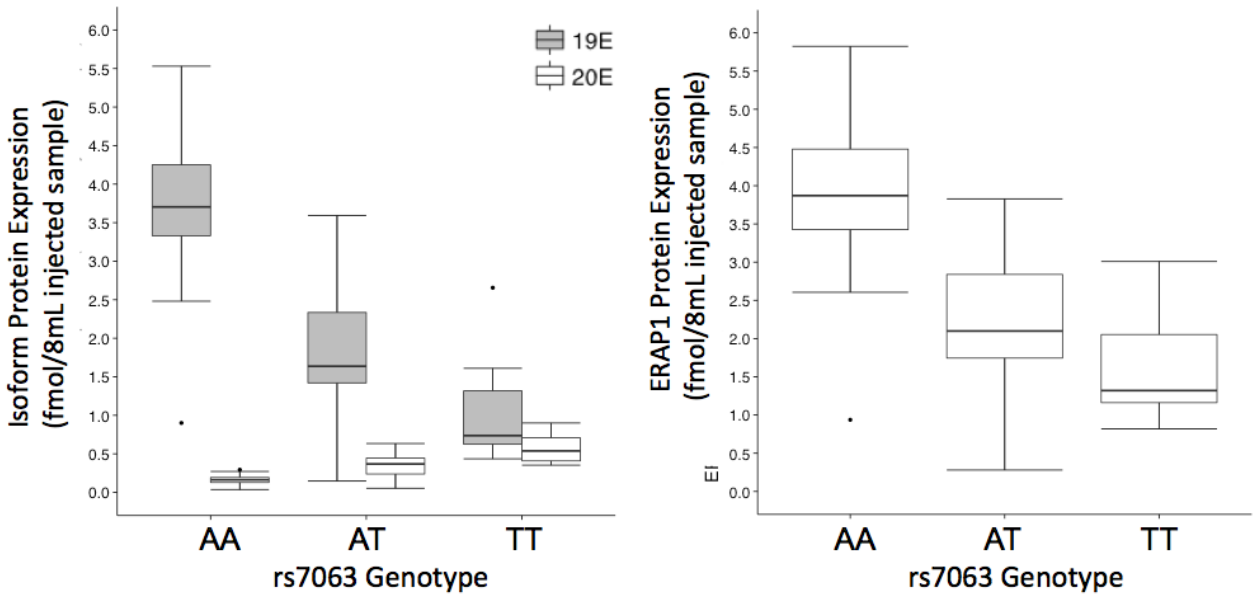

Fig. 5: Genotype at sQTL rs7063 correlates with expression of two ERAP1 protein isoforms and total ERAP1 expression. **(A)** Protein expression of ERAP1 isoforms 19E (gray) and 20E (white) split by genotype at the top *ERAP1* sQTL (rs7063). Individuals homozygous for the AS-risk allele (allele A) express significantly more of the 19E isoform and significantly less of the 20E isoform on the protein level than individuals homozygous for the protective allele. **(B)** Total ERAP1 protein expression (given as the sum of isoform 19E and 20E expression) split by genotype at rs7063.

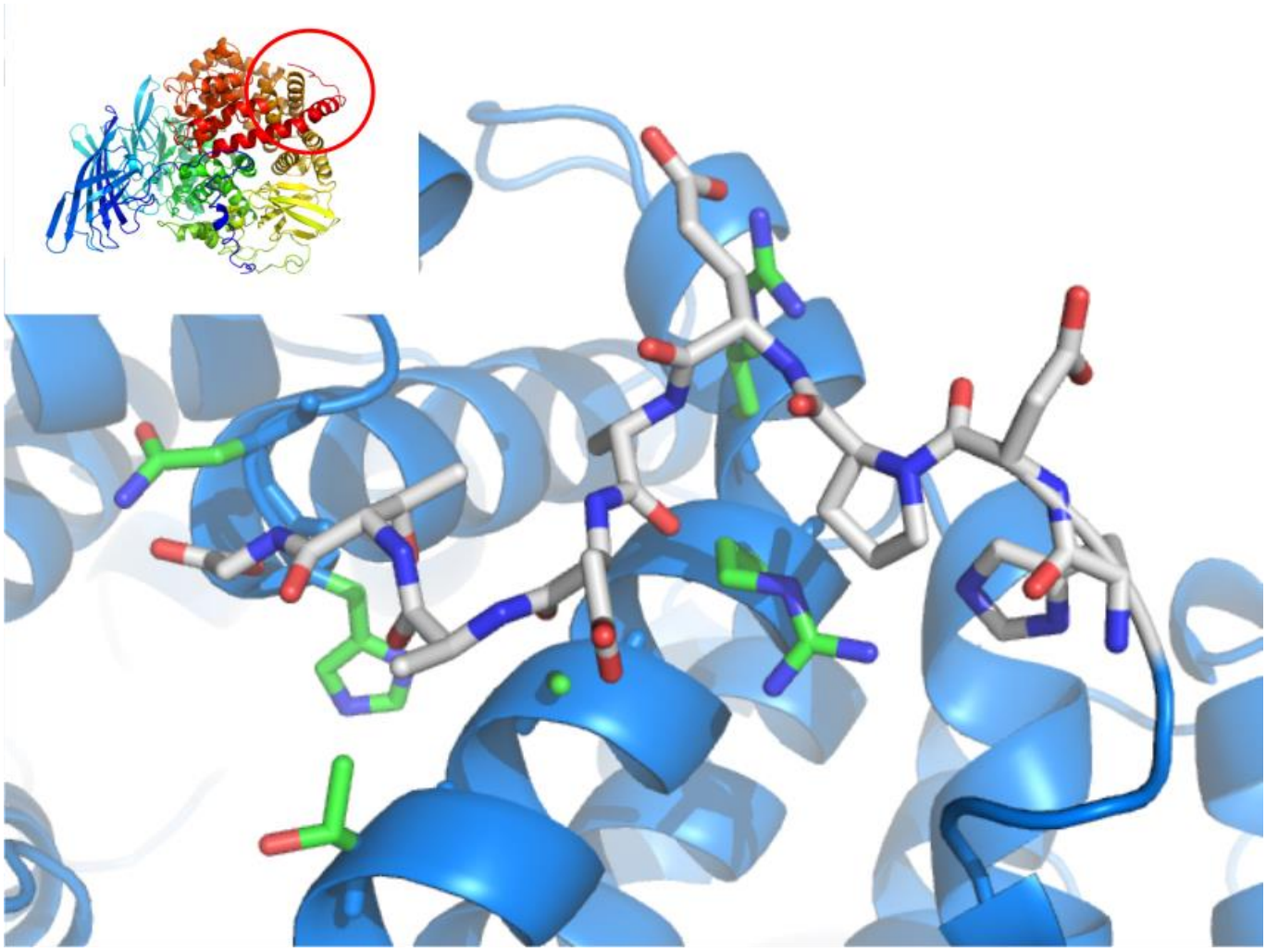

Fig. 6: Protein model of ERAP1 isoform 20E (top left). Zoom to additional C-terminal amino acid residues of the 20th exon with side chains depicted.