RAET1/ULBP alleles and haplotypes among Kolla South American Indians

Steven T. Cox[1], Esteban Arrieta-Bolaños[1,2,3], Susanna Pesoa[4], Carlos Vullo[4], J. Alejandro Madrigal[1,2] and Aurore Saudemont[1,2]

[1]The Anthony Nolan Research Institute, The Royal Free Hospital, Hampstead, London, UK.
[2]UCL Cancer Institute, Royal Free Campus, London, UK.
[3]Centro de Investigaciones en Hematología y Trastornos Afines (CIHATA), Universidad de Costa Rica, San José, Costa Rica.
[4]HLA Laboratory, Hospital Nacional de Clinicas, Cordoba, Argentina.

Corresponding author: Steven Cox (steven.cox@anthonynolan.org.uk). Tel: +44 (0)20 7284 8324. Fax: +44 (0)20 7284 8331.

Abstract

NK cell cytolysis of infected or transformed cells can be mediated by engagement of the activating immunoreceptor NKG2D with one of eight known ligands (MICA, MICB and RAET1E-N) and is essential for innate immunity. As well as diversity of NKG2D ligands having the same function, allelic polymorphism and ethnic diversity has been reported. We previously determined HLA class I allele and haplotype frequencies in Kolla South American Indians who inhabit the northwest provinces of Argentina, and were found to have a similar restricted allelic profile to other South American Indians and novel alleles not seen in other tribes. In our current study, we characterized retinoic acid early transcription-1 (RAET1) alleles by sequencing 58 unrelated Kolla people. Only three of six RAET1 ligands were polymorphic. RAET1E was most polymorphic with five alleles in the Kolla including an allele we previously described, RAET1E*009 (allele frequency (AF) 5.2%). Four alleles of RAET1L were also found and RAET1E*002 was most frequent (AF = 78%). Potential functional diversity only affected RAET1E and RAET1L, which were in linkage disequilibrium indicating a selective advantage. The results suggest that limited RAET1 polymorphism in the Kolla was not detrimental to human survival but still necessary and may affect disease susceptibility or severity.

Keywords: RAET1, innate immunity, Natural Killer cells, Kolla American Indian, polymorphism.

Abbreviations

HLA – human leukocyte antigen, MICA/B – MHC class I-related chain A/B, AF - Allele Frequency, RAET1 - Retinoic Acid Early Transcription 1, ULBP – UL-16 Binding Protein, NKG2D - Natural Killer Group 2 member D, TM - Transmembrane, GPI - Glycophosphatidylinositol, HLA - Human Leukocyte Antigen, LD - Linkage Disequilibrium, MLE - Maximum Likelihood Estimation, HWE - Hardy Weinberg Equilibrium, NET - North East Thai, SNP - Single Nucleotide Polymorphism, CMV - Human Cytomegalovirus, HPV - Human Papillomavirus.

Conflicts of Interest

The authors have declared no conflicts of interest.

1. Introduction

The Retinoic Acid Early Transcription 1 genes (RAET1) are situated on the long arm of chromosome 6 at 6q24.2-25.3, encoding ten gene loci named RAET1E-N, of which six are potentially functional glycoproteins [1-5]. These ligands may also be referred to as UL-16 binding proteins (ULBP) as several RAET1 molecules bind human cytomegalovirus (CMV) glycoprotein, UL-16 [6]. RAET1/ULBP molecules are cell bound activating ligands for the Natural Killer Group 2 member D (NKG2D) receptor expressed by Natural Killer (NK), NKT, $\gamma\delta^+$ and $CD8^+$ T cells [7]. RAET1/ULBP molecules are MHC class I-related with similar function to MHC-encoded molecules MICA and MICB. All NKG2D ligands are markers of stress and become upregulated in infected or transformed cells, inducing cytolysis by NK cells through NKG2D [8, 9]. RAET1 molecules are expressed in diverse tissues such as skin, testes, bone marrow, heart and lung [1] whereas MICA/B expression is more limited and restricted to most epithelial cells and fibroblasts [8].

Expression of different NKG2D ligands in different tissues may relate to specialized functions. In addition, RAET1 genes encode only $\alpha1$ and $\alpha2$ domains with a transmembrane (TM) and cytoplasmic domain (RAET1E/ULBP4 and RAET1G/ULBP5) or glycophosphatidylinositol (GPI) anchors for cell surface expression (RAET1I/ULBP1, RAET1H/ULBP2, RAET1N/ULBP3 and RAET1L/ULBP6) whereas all MICA and MICB molecules have $\alpha3$, TM and cytoplasmic domains. The functional significance for these differences is unknown. Moreover, allelic polymorphism exists within all the genes encoding NKG2D ligands and may indicate that functional variants have arisen, particularly as differences in MICA/B and RAET1/ULBP allelic frequencies are observed between ethnic groups [5, 10-12]. As with human leukocyte antigen (HLA) polymorphism in adaptive immunity, NKG2D ligand polymorphism may confer innate immune resistance to pathogens endemic in different regions of the world. Alleles of NKG2D ligands encoding molecules that are perhaps not as efficient for clearing an infection may require an adaptive immune response and if this is also impaired through restricted HLA polymorphism, selective pressure could occur. However, whilst allelic polymorphism of MICA and MICB loci has been well characterized, little is known of allelic diversity that exists among the RAET1/ULBP family of genes and a formal allelic nomenclature system has yet to be established.

Archaeology and studies of genetic diversity suggest that human habitation of the Americas first occurred around 13,000 years ago, towards the end of the last ice age. It is believed that small groups entered North America from Siberia across the Bering Strait and migrated southwards, taking 1000 years to reach the southernmost tip of South America [13-15]. We refer to descendents of these people as American Indians and they have restricted genetic diversity, likely due to a genetic bottleneck created by small founder populations and reduced further by disease outbreak during their migration as they encountered new pathogens in unfamiliar habitats. American Indians have been extensively studied for polymorphism of HLA loci as they have evolved in relative isolation compared with most populations worldwide. A common genetic feature of American Indians is their restricted HLA polymorphism and high frequency of novel alleles not seen in other populations, particularly HLA-B, which offers an insight into the selective pressures giving rise to HLA diversity and function [13, 16]. Thus, by studying the polymorphism of other genes with immune function, such as RAET1/ULBP, in American Indians it may be possible to avoid some of the complexity derived from population admixture and reveal polymorphisms that may be essential for innate resistance to pathogens.

We have previously studied HLA class I allele and haplotype diversity in Kolla South American Indians [17]. The Kolla (or Colla, Qulla) are an indigenous people of Western Bolivia, Chile, and Argentina and currently situated in Jujuy and Salta Provinces of northwest Argentina. They moved freely between the borders of Argentina and Bolivia and admixture between Argentinean and Bolivian Kolla is known. Their lands are part of the yungas, or high altitude forests, at the edge of the Amazon rainforest. The 2004 Complementary Indigenous Survey reported 53,019 Kolla households living in Argentina [18]. Our previous studies found that HLA-A, -B and –C alleles characteristic of other South American Indian tribes also predominate in the Kolla and one novel allele was present for each of the class I loci [19, 20]. Generation of novel alleles by point mutation of common founder alleles is a feature of American Indian HLA polymorphism and these new variants often increase in frequency to become predominant. Although this has not occurred yet in the Kolla, all the novel variants appeared to be derived from putative founder alleles. The data suggests that diversification of HLA alleles occurred after separation of the founding population into localized tribes, supporting the hypothesis that diversity is driven by selective pressure on new alleles that may enable a more efficient adaptive immune response against local

pathogens. In innate immunity, ligands for the NK cell activatory receptor, NKG2D, also have varying levels of polymorphism encoded by MICA, MICB and RAET1E-N genes. It is unknown why there are numerous ligands for NKG2D or why they display such polymorphism, but it is probable that innate resistance to pathogens is maintained by variability in NKG2D ligands. Thus, the current study is an analysis of allelic polymorphism and haplotypes of RAET1 gene loci in the same cohort of Kolla South American Indians, to further understanding of the relevance of NKG2D ligand diversity and polymorphism in innate NK cell immunity.

2. Materials and methods

2.1 Sample collection

Peripheral blood lymphocytes were collected from 70 unrelated Kolla individuals who were mother and father of the family where possible. Efforts were made to ensure siblings of those already sampled were excluded. DNA was extracted using in-house salting-out protocols. These DNA samples had been used in previous studies [17, 19-21] and for this study, 58 of the original 70 DNA samples could be utilized.

2.2 Sequence-Based typing of RAET1 genes

Direct sequencing of polymorphic loci RAET1E, G, H, I, L and N (ULBP4, 5, 2, 1, 6, 3 respectively) was performed by amplification of exons 2-3 and bi-directional sequencing of the exons where necessary, as previously described [2, 11] and recently updated [11]. A summary of amplification and sequencing primers used in this study are provided in supplementary Table S1. The Heterozygous combinations of alleles were resolved using MATCHTOOLS 1.0 (Applied Biosystems, CA, USA) with RAET1 allele libraries. An additional cohort of 42 European-Caucasoid cell-line DNA samples were identified from DNA available in our laboratory using the IMGT/HLA database (http://www.ebi.ac.uk/ipd/imgt/hla/) [22] and typed for RAET1E/ULBP4 polymorphism. Details of cell-line DNA and RAET1E/ULBP4 typing results are given in Table S1.

2.3 Frequency analysis

Allele frequencies and haplotypes were analyzed using Arlequin 3.5 [23]. RAET1 two-locus haplotype frequencies and linkage disequilibrium (LD) were determined by ELB (Bayesian) algorithm for best gametic phases to obtain D' and R^2. Three-locus haplotypes were generated using Maximum Likelihood Estimation (MLE). Hardy-Weinberg equilibrium (HWE) was assessed by exact test using Markov chain Monte Carlo method.

3. Results

3.1 RAET1 nomenclature and allelic polymorphism

The genomic organization of expressed RAET1 genes is provided in Figure 1. In this study, four of the six expressed RAET1 genes were found to be polymorphic. Nucleotide polymorphism giving rise to alleles within RAET1E, G, H and L gene loci (ULBP4, 5, 2, 6 respectively) are shown by alignment of polymorphic positions in Table 2. All allele sequences are available from DNA sequence repositories and their associated accession numbers are shown. New alleles were determined by BLASTn searches of potentially novel sequences. The nomenclature is based on the naming system proposed by Romphruk and colleagues [11] and new alleles were named numerically in order of their discovery after cloning and sequencing in isolation where necessary, as previously reported [24].

Polymorphism within exons 2-3 of RAET1 loci in Kolla individuals was observed for RAET1E, H and L. No variation was seen for RAET1I as polymorphism of this locus is rare [11, 12] and similarly, we did not detect any variation of RAET1G or RAET1N (polymorphism is in exon 1) in Kolla samples. Currently, the most polymorphic locus is RAET1E with ten alleles including three alleles, RAET1E*008, 009 and 010, we described recently [24]. RAET1E*008 was sequenced from homozygous Italian-Caucasoid cell line CALEGERO (IHW09084). RAET1E*009 was unique among Kolla individuals and RAET1E*010 was detected among 8.3% of European-Caucasoid cell line DNA samples (n=42) listed in Table 1. All three alleles had ‘A’ at nucleotide position 383, only previously seen in the RAET1E*002 allele, with other nucleotides arising from point mutation or recombination with other RAET1E alleles. The next most polymorphic locus was RAET1L with seven alleles, followed by RAET1H with six alleles and RAET1G with three alleles.

3.2 RAET1 allele frequencies

Allele frequencies in Kolla samples are shown in Table 3. For RAET1E, the most frequent allele was RAET1E*002 with a frequency of 0.56035 followed by RAET1E*005 (0.3707) and RAET1E*009 (0.0517). Two other alleles were also detected with low frequency: RATE1E*001 (0.0086) and RAET1E*004 (0.0086). This compares with six alleles detected among 42 Euro-Caucasoid cell line DNA samples with frequencies above 1%:

RAET1E*001 (0.143), 002 (0.238), 003 (0.369), 005 (0.143), 008 (0.024) and 010 (0.083). RAET1G*001 was the only allele within this locus detected in Kolla samples and there was also no variation of exons 2-3 of RAET1N. RAET1H*002 was most prevalent, having a frequency of 0.862 with the remainder carrying the RAET1H*001 allele (0.138). Analysis of RAET1L revealed four alleles among the Kolla. RAET1L*002 was most frequent (0.776) followed by 001 (0.138), 003 (0.078) and 004 (0.0086). All polymorphic loci found in the Kolla had an allele distribution that was in agreement with HWE principle (supplementary Table S2).

3.3 Two-locus RAET1 haplotypes and linkage disequilibrium

Two-locus RAET1 haplotypes having significant LD are given in Table 4. The two loci in strongest LD were RAET1E and RAET1H and the alleles of these loci having highest frequency and positive LD were RAET1E*002 – RAET1H*002 with a frequency, D', R^2, and P-value of 0.5431, 0.77692, 0.12309 and 0.0002 respectively. The next most prevalent haplotype for this pair of loci was RAET1E*005 – RAET1H*002 found among 25% of the Kolla (D'=-0.80137, R^2 = 0.17444, P<0.0001) but these alleles had a negative LD, indicating a tendency to not occur together. Moderate LD at high frequency was found between RAET1E and RAET1L. RAET1E*002 – RAET1L*002 had the highest frequency of 0.38793 but had a negative LD (D', R^2, and P-value of -0.47511, 0.05117, 0.0148, respectively). The next most frequent was RAET1E*005 – RAET1L*002 and had moderate, positive LD (frequency = 0.33621, D'=0.58497, R^2 = 0.05823, P=0.0094). RAET1E*002 was also found in positive LD with RAET1L*001 (frequency = 0.11207, D'=0.57353, R^2 = 0.04129, P=0.0286). Low to moderate positive LD also existed between RAET1H*001 and RAET1L*001 at relatively low frequency (frequency = 0.04310, D'=0.58497, R^2 = 0.05823, P=0.0094).

3.4 Three-locus RAET1E, RAET1H and RAET1L haplotypes

Common three-locus RAET1 haplotypes are listed in Table 5. Two haplotypes were found with a combined frequency of 0.6379, both having RAET1H*002 and RAET1L*002 but differing by RAET1E*002 or 005 (38.8% and 25% respectively). Interestingly, although the strongest associations were generally between RAET1E and RAET1H (Table 4), these highest frequency haplotypes differed in their associations. RAET1E*002 was more

associated RAET1H*002 and RAET1E*005 was associated with RAET1L*002. Six other haplotypes were also present among 1-10% of the Kolla people. Notably, the Kolla allele RAET1E*009 was only found in combination with RAET1H*002 and RAET1L*002, with a haplotype frequency that was the same as the allele frequency (0.05172).

3.5 Comparison of RAET1E and RAET1L frequencies in the Kolla with other populations

Figure 2a compares RAET1E allele frequencies between Euro-Caucasoid cell-line DNA (n=42), Kolla American Indians (n=58) and northeast Thai (NET; n=176) [11]. Clear differences are seen between the populations, for example RAET1E*001 has a very high frequency of 52.8% in NET but is virtually absent in the Kolla. RAET1E*003 has highest prevalence in Euro-Caucasoid cell-lines (0.369) but was undetected in the Kolla and at a much lower frequency in NET (0.074). The allele with highest frequency in all three groups was RAET1E*002.

Figure 2b compares RAET1L allele frequencies in Caucasoid (n=32) [5], Kolla (n=58) and NET (n=176). RAET1L*002 was most prevalent in the Kolla (0.776) and was also frequent in Caucasoid (0.420) and NET (0.253). The most prevalent allele in Caucasoid individuals was RAET1L*003 (0.480) and was second most frequent among NET (0.267) but was much lower in the Kolla people (0.078).

4. Discussion

The present study demonstrated that RAET1 allele polymorphism was restricted in the Kolla compared to other populations. Three RAET1E/ULBP4 alleles with a combined frequency of 98.6% were detected compared with six alleles detected in the majority of European-Caucasoid cell lines or five alleles in NET individuals [11]. Two other alleles, RAET1E*001 and *004 were found with frequencies <1% and were likely to have been introduced by gene flow. One of the three frequent alleles detected in the Kolla was the recently described RAET1E*009 [24] and to our knowledge, has not been detected in any other population. This allele was almost identical to RAET1E*002, the most frequent RAET1E allele (56%), and appears to have been generated by a novel non-synonymous point mutation. The haplotype data shows that RAET1E*009 shares the same haplotype as RAET1E*002 and it would seem probable that it has derived from the putative founder allele RAET1E*002. The

other allele, RAET1E*005, had an allele frequency of 37% in Kolla people and is structurally disparate from RAET1E*002 with one residue difference in α1 (exon 2) and three in the α2 domain (exon 3). The positions where residues differ among RAET1E alleles are not predicted to interact directly with NKG2D but, as noted by others [5, 11, 12], could alter the conformation and orientation of residues and allow tighter interaction with NKG2D and enhanced NK cell activation. –RAET1E is currently the most polymorphic RAET1 ligand and the presence of a novel allele at relatively high frequency in the Kolla may indicate that positive selection is generating diversity.

Studies of RAET1/ULBP gene polymorphism and ethnic diversity are rare and limited compared to those for MICA/B, but frequencies also differ by ethnicity [2, 5, 11, 12]. The study by Romphruk and colleagues of NET RAET1 polymorphism was the first to provide frequencies of alleles in polymorphic RAET1 loci in a non-Caucasoid population [11]. The results, obtained by direct sequencing, showed differences in frequencies compared with Caucasoids and several new allelic variants were also present at low frequency. Antoun and co-workers used single nucleotide polymorphism (SNP) analysis of RAET1 promoter and coding regions to study polymorphism of alleles and haplotypes in Euro-Caucasoid, African-Caribbean and Indo-Asian populations [12]. They also concluded that the distribution of SNPs and haplotypes varied considerably within and between different ethnic groups. It is also clear that some loci have low or no polymorphism (RAET1I/ULBP1 and RAET1N/ULBP3) whilst other loci, in particular RAET1E/ULBP4 and also RAET1L/ULBP6 are very polymorphic, although the functional significance of this is not clear. Modern urban populations are the product of extensive genetic admixture with neighboring populations and more recent migration has resulted in further admixture. As a consequence, the ethnic diversity revealed from genetic studies does not accurately reflect polymorphisms that have arisen and been maintained by positive selection. In this respect, genetic studies of American Indian populations have proven illuminating due to their relative isolation from other ethnic groups.

There are a total of eight known ligands for NKG2D, including MICA/B and the six RAET1/ULBP molecules, and individual ligands are also polymorphic. Differences in affinity for NKG2D have been observed between the different ligands, which affect NK cell activation, the weakest affinity was with RAET1G [5]. The relevance of RAET1

polymorphism is largely unknown, however MICA transcription is upregulated by CMV but surface expression is prevented by binding of MICA to CMV glycoprotein UL42, affecting all MICA types except MICA*008, which has a truncated TM/cytoplasmic tail [25]. In addition, RAET1N (ULBP3) can also be bound by UL142, preventing expression, but RAET1H (ULBP2) was found to be unaffected [25]. Similarly RAET1 ligands are upregulated by CMV infection, but all RAET1 molecules except RAET1N (ULBP3) and RAET1E (ULBP4) bind CMV glycoprotein UL16 and are retained within the cytoplasm [1, 6]. UL16 also binds and prevents expression of MICB, but not MICA [3]. Thus, ligand and allelic diversity may have arisen from selective pressure by pathogens driving evolution of structural variants that can inhibit such mechanisms.

The small sample size of 58 Kolla individuals is a limitation of this study, which could not be avoided and may affect haplotype estimation, although low polymorphism of RAET1 alleles may have helped to limit this problem. Negative LD values observed in this study may also be a consequence of the nature of American Indian genetic diversity such as founder effects, extensive differentiation of American Indian populations, genetic bottlenecks and gene flow. In addition, genetic drift, which is often negligible in large populations, may be exaggerated in small or isolated tribes such as the Kolla. The highest R^2 values, in this study, were observed between RAET1E and 1H and are in agreement with Rareongjai and colleagues in Thai subjects [26]. The two most frequent RAET1E-RAET1H-RAET1L haplotypes in the Kolla were 002-002-002 (38.8%) and 005-002-002 (25%) respectively. The polymorphism distinguishing RAET1H*001 and 002 at residue 123 is synonymous, therefore potential functional diversity (within the α1 and α2 domains) only affects RAET1E and RAET1L and these are in positive LD, albeit with low R^2 values. Specifically, RAET1E*005 and RAET1L*002 were in positive LD, suggesting a selective advantage in having these two common Kolla alleles together. RAET1E*002 and 005 are structurally dissimilar with four amino acid differences (residues 82, 128, 141 and 142) at positions predicted to influence interaction with NKG2D [12] that may allow differential activation of NK cells. Similarly, RAET1L*001 and *002 differ at residue 85, also predicted to interact with NKG2D. Alternatively these polymorphisms may have been selected because they are not recognized by pathogens. As we have only examined the extracellular domains, differences may exist in the transmembrane region, which could allow or prevent shedding by metalloproteinases possibly leading to NK cell inhibition. This mechanism of NKG2D ligand shedding has been

discussed at length as a tumor escape mechanism [27-29] and may also be induced by viral infection in order to avoid immune recognition by NK cells. NKG2D ligands with short cytoplasmic tails (such as MICA*008) or with GPI-anchorage (RAET1E and 1G) are resistant to degradation within lysosomes as they lack a lysine motif in the cytoplasmic tail, enabling them to escape an immune evasion strategy employed by some viruses, such as human herpesvirus-7 [30, 31]. Differences may also exist in the promoter regions, known to be polymorphic, which could alter the rate of transcription of RAET1 genes. Future, larger studies would benefit from analysis of the entire promoter and coding regions to enable a more clear insight into RAET1 function.

In conclusion, investigation of RAET1 alleles and haplotypes in Kolla South American Indians revealed limited polymorphism of RAET1E, H and L compared to other populations, and no diversity of RAET1I, N and G. However, LD between some alleles of different loci may indicate a selective advantage relating to innate immunity against viruses and/or tumors. More studies are required to understand the relevance of polymorphism of MIC and RAET1 genes in different populations and the implications for innate NK cell immunity. A better understanding of the biological roles may lead to targeted strategies for immunotherapy.

Figure legends

Figure 1: Genomic organization of the RAET1 gene family in humans. RAET1 genes are encoded on the long arm of chromosome 6 (6q24.2-25.3) towards the telomeric end. Closed arrows indicate the relative positions of the six expressed genes. Shown alongside are the relative lengths of the genes and organization of the coding regions shown by closed boxes. The leader peptide sequence is indicated by the letter 'L', glycophosphatidylinositol anchorage sites are indicated by GPI and transmembrane/cytoplasmic regions by TM.

Figure 2: Panel A shows RAET1E allele frequencies in European-Caucasoid cell line DNA, Kolla South American Indian and NET populations. Highest frequencies in Kolla samples were RAET1E*002 and *005. Panel B shows RAET1L allele frequencies in Caucasoid [5], Kolla South American Indian and NET populations [11]. Highest frequency RAET1L allele in Kolla samples was RAET1L*002. The frequency distributions for both genes are different between the various populations.

References

[1] Cosman D, Mullberg J, Sutherland CL, Chin W, Armitage R, Fanslow W et al.: ULBPs, novel MHC class I-related molecules, bind to CMV glycoprotein UL16 and stimulate NK cytotoxicity through the NKG2D receptor. Immunity 2001;14(2):123-33.

[2] Radosavljevic M, Cuillerier B, Wilson MJ, Clement O, Wicker S, Gilfillan S et al.: A cluster of ten novel MHC class I related genes on human chromosome 6q24.2-q25.3. Genomics 2002;79(1):114-23.

[3] Chalupny NJ, Sutherland CL, Lawrence WA, Rein-Weston A, Cosman D: ULBP4 is a novel ligand for human NKG2D. Biochem Biophys Res Commun 2003;305(1):129-35.

[4] Bacon L, Eagle RA, Meyer M, Easom N, Young NT, Trowsdale J: Two human ULBP/RAET1 molecules with transmembrane regions are ligands for NKG2D. J Immunol 2004;173(2):1078-84.

[5] Eagle RA, Traherne JA, Hair JR, Jafferji I, Trowsdale J: ULBP6/RAET1L is an additional human NKG2D ligand. Eur J Immunol 2009;39(11):3207-16.

[6] Kubin M, Cassiano L, Chalupny J, Chin W, Cosman D, Fanslow W et al.: ULBP1, 2, 3: novel MHC class I-related molecules that bind to human cytomegalovirus glycoprotein UL16, activate NK cells. Eur J Immunol 2001;31(5):1428-37.

[7] Bahram S, Inoko H, Shiina T, Radosavljevic M: MIC and other NKG2D ligands: from none to too many. Curr Opin Immunol 2005;17(5):505-9.

[8] Groh V, Bahram S, Bauer S, Herman A, Beauchamp M, Spies T: Cell stress-regulated human major histocompatibility complex class I gene expressed in gastrointestinal epithelium. Proc Natl Acad Sci U S A 1996;93(22):12445-50.

[9] Eleme K, Taner SB, Onfelt B, Collinson LM, McCann FE, Chalupny NJ et al.: Cell surface organization of stress-inducible proteins ULBP and MICA that stimulate human NK cells and T cells via NKG2D. J Exp Med 2004;199(7):1005-10.

[10] Collins RW: Human MHC class I chain related (MIC) genes: their biological function and relevance to disease and transplantation. Eur J Immunogenet 2004;31(3):105-14.

[11] Romphruk AV, Romphruk A, Naruse TK, Raroengjai S, Puapairoj C, Inoko H et al.: Polymorphisms of NKG2D ligands: diverse RAET1/ULBP genes in northeastern Thais. Immunogenetics 2009;61(9):611-7.

[12] Antoun A, Jobson S, Cook M, O'Callaghan CA, Moss P, Briggs DC: Single nucleotide polymorphism analysis of the NKG2D ligand cluster on the long arm of chromosome 6: Extensive polymorphisms and evidence of diversity between human populations. Hum Immunol 2010;71(6):610-20.

[13] Gibbons A: Geneticists trace the DNA trail of the first Americans. Science 1993;259 (5093):312-3.

[14] Neel JV, Biggar RJ, Sukernik RI: Virologic and genetic studies relate Amerind origins to the indigenous people of the Mongolia/Manchuria/southeastern Siberia region. Proc Natl Acad Sci U S A 1994;91(22):10737-41.

[15] Parham P, Ohta T: Population biology of antigen presentation by MHC class I molecules. Science 1996;272(5258):67-74.

[16] Little AM, Parham P: Polymorphism and evolution of HLA class I and II genes and molecules. Rev Immunogenet 1999;1(1):105-23.

[17] Little AM, Scott I, Pesoa S, Marsh SG, Arguello R, Cox ST et al.: HLA class I diversity in Kolla Amerindians. Hum Immunol 2001;62(2):170-9.

[18] (INDEC). INdEyC: Results of the Additional Survey of Indigenous Peoples — ECPI— conducted from 2004. http://www.indec.mecon.ar/webcenso/ECPI/pueblos/ampliada_index.asp?mode=06, 2004.

[19] Scott I, Dunn PP, Day S, Pesoa S, Little AM, Madrigal JA, Vullo C: A novel HLA allele, HLA-B*5113, identified in the Kolla Amerindians of North-West Argentina. Tissue Antigens 1999;53(2):194-7.

[20] Ramon D, Scott I, Cox ST, Pesoa S, Vullo C, Little AM, Madrigal JA: HLA-A*6817, identified in the Kolla Amerindians of North-West Argentina possesses a novel nucleotide substitution. Tissue Antigens 2000;55(5):453-4.

[21] Perez-Rodriguez M, Raimondi E, Marsh SG, Madrigal JA: Identification of a new MICA allele, MICA*047. Tissue Antigens 2002;59(3):216-8.

[22] Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SG: The IMGT/HLA database. Nucleic Acids Res 2013;41(D1):D1222-D1227.

[23] Excoffier L, Laval G, Schneider S: Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform 2005;Online 1:47-50.

[24] Cox ST, Madrigal JA, Saudemont A: Three novel allelic variants of the RAET1E/ULBP4 gene in humans. Tissue Antigens 2012;80(4):390-2.

[25] Chalupny NJ, Rein-Weston A, Dosch S, Cosman D: Down-regulation of the NKG2D ligand MICA by the human cytomegalovirus glycoprotein UL142. Biochem Biophys Res Commun 2006;346(1):175-81.

[26] Rareongjai S, Romphruk A, Romphruk AV, Sakuntabhai A, Leelayuwat C: Linkage disequilibrium of polymorphic RAET1 genes in Thais. Tissue Antigens 2010;76(3): 230-5.

[27] Salih HR, Rammensee HG, Steinle A: Cutting edge: down-regulation of MICA on human tumors by proteolytic shedding. J Immunol 2002;169(8):4098-102.

[28] Groh V, Wu J, Yee C, Spies T: Tumour-derived soluble MIC ligands impair expression of NKG2D and T-cell activation. Nature 2002;419(6908):734-8.

[29] Fernandez-Messina L, Ashiru O, Boutet P, Aguera-Gonzalez S, Skepper JN, Reyburn HT, Vales-Gomez M: Differential mechanisms of shedding of the glycosylphosphatidylinositol (GPI)-anchored NKG2D ligands. J Biol Chem 2010;285(12):8543-51.

[30] Schneider CL, Hudson AW: The human herpesvirus-7 (HHV-7) U21 immunoevasin subverts NK-mediated cytoxicity through modulation of MICA and MICB. PLoS Pathog 2011;7(11):epub e1002362.

[31] Fernandez-Messina L, Reyburn HT, Vales-Gomez M: Human NKG2D-ligands: cell biology strategies to ensure immune recognition. Front Immunol 2012;3, article 299:1-8.

Table 1. European-Caucasoid cell line DNA details and RAET1E (ULBP4) genotypes

Cell Line Name[a]	Alias[a]	Ethnicity	RAET1E* type
AM	IHW09206	Caucasoid - England, Europe	001, 005
AMI005AN	TER202	Caucasoid - Unknown, Europe	003
APD	IHW09291	Caucasoid - Netherlands, Europe	003, 010
ARBO	IHW09102	Caucasoid - Netherlands, Europe	003, 010
BM21	IHW09043	Caucasoid - Germany/Italy, Europe	002, 003
BM92	IHW09092	Caucasoid - Italy, Europe	002
BOB	IHW09089	Caucasoid - Germany, Europe	001, 003
BOLETH	IHW09031	Caucasoid - Sweden, Europe	003, 005
BTB	IHW09067	Caucasoid - Scandinavia, Europe	001, 005
C047	IHW09265	Caucasoid - Spain, Europe	003
C073	IHW09261	Caucasoid - Spain, Europe	001, 003
CALOGERO	IHW09084	Caucasoid - Italy, Europe	008
CEK008AN	HC12081	Caucasoid - Sweden/Cyprus, Europe	003
CMD004AN	HC10424	Caucasoid - England, Europe	003, 010
DEM	IHW09007	Caucasoid - Germany, Europe	002, 003
DKB	IHW09075	Caucasoid - Netherlands, Europe	002, 003
DUCAF	IHW09019	Caucasoid - France, Europe	003, 010
EK	IHW09054	Caucasoid - Scandinavia, Europe	001, 005
HERLUF	IHW09299	Caucasoid - Denmark, Europe	001, 002
JHAF	IHW09030	Caucasoid - England, Europe	003, 010
KAS116	IHW09003	Caucasoid - Yugoslavia, Europe	001, 002
LB	IHW09289	Caucasoid - Sweden, Europe	002, 005
LUY	IHW09070	Caucasoid - Netherlands, Europe	002, 005
MOV002AN	EXT-0497	Caucasoid - England, Europe	001, 003
PGF	IHW09318	Caucasoid - England, Europe	003
PLH	IHW09047	Caucasoid - Scandinavia, Europe	002, 005
QBL	IHW09020	Caucasoid - Netherlands, Europe	003
REN	CEK-ND	Caucasoid - Wales, Europe	002, 003
SAVC	IHW09034	Caucasoid - France, Europe	001, 003
SCHU	IHW09013	Caucasoid - France, Europe	002
SPO010	IHW09036	Caucasoid - Italy, Europe	002
STEINLIN	IHW09087	Caucasoid - France, Europe	005, 010
TL	IHW09270	Caucasoid - Unknown, Europe	003, 005
VAVY	IHW09023	Caucasoid - France, Europe	002, 005
VEN	IHW09238	Caucasoid - Unknown, Europe	002
WIN	IHW09095	Caucasoid - Germany, Europe	003, 010
WT24	IHW09015	Caucasoid - Italy, Europe	003
WT47	IHW09063	Caucasoid - Italy, Europe	003
WT49	IHW09285	Caucasoid - Italy, Europe	002, 003
WT51	IHW09029	Caucasoid - Italy, Europe	001
WT52	IHW09306	Caucasoid - Italy, Europe	001, 005
WVB	IHW09062	Caucasoid - Netherlands, Europe	002, 005

[a] Further details of all cell lines utilised can be obtained from the IMGT/HLA database (http://www.ebi.ac.uk/ipd/imgt/hla/).

Table 2. Known RAET1 nucleotide and amino acid polymorphism within exons 2 to 3 of alleles revealed by sequence-based typing.

Allele details	Polymorphic nucleotides and amino acids								Accession number
	Exon 2 (89-345)				Exon 3 (346-621)				
Nucleotide no.[a]	134	244	267	296	382	383	421	425	
RAET1E	C	A	A	G	C	G	A	T	
001	–	–	–	–	–	–	–	–	JX051312
002	–	–	G	–	–	A	–	–	JX051313
003	–	–	–	–	–	–	G	C	JX051314
004	–	–	–	–	–	–	G	–	JX051315
005	–	T	–	–	–	–	G	C	JX051316
006	–	–	–	–	T	–	–	–	JX051317
007	T	–	–	–	–	–	–	–	JX051318
008	–	–	–	–	–	A	G	–	HE804773
009	–	–	G	A	–	A	–	–	HE804774
010	–	–	–	–	–	A	–	–	HE806193
Consensus aa	Pro	Asn	Glu	Arg	Arg	Arg	Thr	Iso	
Variant aa	Lys	Tyr	-	Gln	Cys	His	Ala	Thr	
Codon no.[a]	45	82	89	99	128	128	141	142	
	Exon 2 (86-349)								
Nucleotide no.[a]	209	220							
RAET1G	C	G							
001	–	–							EAW47779, JX162560
002	G	–							JX162561
003	G	A							JX162562
Consensus aa	Thr	Val							
Variant aa	Arg	Ile							
Codon no.[a]	69	73							
	Exon 2 (86-349)			Exon 3 (350-671)					
Nucleotide no.[a]	114	203	219	369	487				
RAET1H	C	A	T	A	C				
001	–	–	–	–	–				JX162563, JX162569
002	–	–	–	C	–				JX162564, JX162570
003	–	–	–	C	G				JX162565, JX162571
004	T	–	–	C	–				JX162566, JX162572
005	–	–	C	–	–				JX162567, JX162573
006	–	G	C	C	–				JX162568, JX162574
Consensus aa	Thr	Asn	Pro	Ala	Pro				
Variant aa	-	Ser	-	-	Ala				
Codon no.[a]	38	68	73	123	163				
	Exon 2 (86-349)			Exon 3 (350-671)					
Nucleotide no.[a]	254	317		417	440				
RAET1L	T	T		T	C				
001	–	–		–	–				JX162575, JX162582
002	C	–		–	–				JX162576, JX162583
003	C	G		–	T				JX162577, JX162584
004	C	G		–	–				JX162578, JX162585
005	C	–		–	T				JX162579, JX162586
006	–	–		–	T				JX162580, JX162587

007	C	G	C	–	JX162581, JX162588
Consensus aa	Met	Lys	Arg	Thr	
Variant aa	Thr	Arg	Trp	Iso	
Codon no.[a]	85	77	139	147	

Alternative nomenclature: RAET1E = ULBP4, RAET1G = ULBP5, RAET1H = ULBP2, RAET1L = ULBP6.
[a] Nucleotide numbering is based on the coding sequence (CDS) of exons 1-3. Codons/amino acids are numbered from exon 1.

Table 3. RAET1 allele frequencies among the Kolla

RAET1E	Frequency	RAET1G	Frequency	RAET1H	Frequency	RAET1N [a]	Frequency	RAET1L	Frequency
002	0.56035	001	1.00000	002	0.86207	001/002	1.00000	002	0.77586
005	0.37069			001	0.13793			001	0.13793
009	0.05172							003	0.07759
001	0.00862							004	0.00862
004	0.00862								

[a] RAET1N allelic polymorphism is located in exon 1 and was not included in the analysis.

RAET1 allele frequencies were determined by maximum likelihood estimation using Arlequin 3.5 from the phenotypes of 58 unrelated Kolla individuals. Frequencies above 1% are shown.

Abbreviation: RAET1 – retinoic acid early transcript 1.

Alternative nomenclature: RAET1E = ULBP4, RAET1G = ULBP5, RAET1H = ULBP2, RAET1N = ULBP3, RAET1L = ULBP6.

Table 4. Two-locus RAET1 haplotypes exhibiting significant linkage disequilibrium among the Kolla

Haplotype		Frequency	D'	R^2	P value
RAET1E	RAET1H				
002	002	0.54310	0.77692	0.12309	0.0002
005	002	0.25000	-0.80137	0.17444	<0.0001
005	001	0.12069	0.80137	0.17444	<0.0001
RAET1H	RAET1L				
002	001	0.09483	-0.20250	0.04101	0.0292
001	001	0.04310	0.20250	0.04101	0.0292
RAET1E	RAET1L				
002	002	0.38793	-0.47511	0.05117	0.0148
005	002	0.33621	0.58497	0.05823	0.0094
002	001	0.11207	0.57353	0.04129	0.0286

RAET1 two-locus haplotype frequencies were determined by ELB algorithm (Bayesian) using Arlequin 3.5 from
the phenotypes of 58 unrelated Kolla individuals. Frequencies above 1% are shown.
Abbreviation: RAET1 – retinoic acid early transcript 1.
Alternative nomenclature: RAET1E = ULBP4, RAET1H = ULBP2, RAET1L = ULBP6.

Table 5. RAET1E, RAET1H and RAET1L haplotype frequencies derived from maximum likelihood estimation

RAET1E	RAET1H	RAET1L	Frequency [a]	SD
002	002	002	0.38793	0.04544
005	002	002	0.25000	0.04038
002	002	001	0.09483	0.02732
005	001	002	0.08621	0.02617
002	002	003	0.06034	0.02221
009	002	002	0.05172	0.02065
005	001	001	0.02586	0.01480
002	001	001	0.01724	0.01214

[a] Frequencies greater than 1% are shown.

Alternative nomenclature: RAET1E = ULBP4, RAET1H = ULBP2, RAET1L = ULBP6.

Abbreviation: SD – standard

Table S1: Amplification and sequencing primers for analysis of RAET1 genes

Locus	Primers	Location	Reference
Amplification Primers			
RAET1E/ULBP4	5'-TCACCATAAGTGGGAGGAGG-3'	Intron 2 F	[2]
	5'-CTGAACTCGAGACAGCTTCC-3'	Intron 4 R	
RAET1G/ULBP5	5'-CTTCCTCTTATTGTCACAGTG-3'	Intron 2 F	[11]
	5'-CCCCATTTCTGATCTCATTGG-3'	Intron 4 R	
RAET1H/ULBP2 and RAET1L/ULBP6[a]	5'-CTTATTGACACACAGCGTGGAG-3'	Intron 2 F	[2]
	5'-TTGAAGCTGAACTCAAGAAC-3'	Intron 4 R	
RAET1I/ULBP1	5'-TCACCATAAGTGGGAGGAGG-3'	Intron 2 F	[2]
	5'-CTGAACTCGAGACAGCTTCC-3'	Intron 4 R	
RAET1N/ULBP3	5'-CACAGTGTTTGGGGTCTTTC-3'	Intron 2 F	[2]
	5'-GACAGACAAGGGTGTAACTC-3'	Intron 4 R	
RAET1H/ULBP2 specific[b]	5'-TCTTATTGACACAGCGTGGA-3'	Intron 2 F	[11]
	5'-TTCACCATGTTGGTCAGGCT-3'	Intron 3 R	
	5'-AATACAAATGGGAAGGTCATCA-3'	Intron 3 F	
	5'-ATTGAAGCTGAACTCAAGAAC-3'	Intron 4 R	
RAET1L/ULBP6 specific[c]	5'-TCTTATTGACACAGCGTGGAG-3'	Intron 2 F	[11]
	5'-TCACTATGTTGGTCAGGCG-3'	Intron 3 R	
	5'-AATACAAATGGGAAGGTCATCT-3'	Intron 3 F	
	5'-ATACAAATGGGAAGGTCATCG-3'	Intron 3 F	
	5'-ATTGAAGCTGAACTCAAGAAC-3'	Intron 4 R	
Sequencing primers			
RAET1E/ULBP4	5'-GTCAGGGAGAGATGGGAACA-3'	Intron 2 F	[2]
	5'-TCTCTTACTGCCTGCCTCTG-3'	Intron 3 R	
	5'-TGGAGGATGATGGACTTCTC-3'	Intron 3 F	
	5'-CACAGGGAAGGCTTTTGACC-3'	Intron 4 R	
RAET1G/ULBP5	5'-CTTCCTCTTATTGTCACAGTG-3'	Intron 2 F	[11]
	5'-AACCTACCACTGTATCTGCTC-3'	Intron 3 R	
	5'-AATTGCAAGTGGAAGAGGATG-3'	Intron 3 F	
RAET1H/ULBP2	5'-TCTTATTGACACAGCGTGGAG-3'	Intron 2 F	[2]
	5'TTCACCATGTTGGTCAGGCT-3'	Intron 3 R	[11]
	5'-AATACAAATGGGAAGGTCATCA-3'	Intron 3 F	[11]
	5'-ATTGAAGCTGAACTCAAGAAC-3'	Intron 4 R	[2]
RAET1I/ULBP1	5'-GTTACTACTGTGTCTGCTCC-3'	Intron 3 R	[2]
	5'-CAGCAGAGAGAGCAAGTCC-3'	Intron 3 F	[11]
RAET1L/ULBP6	5'-TCTTATTGACACAGCGTGGAG-3'	Intron 2 F	[2]
	5'-TCACTATGTTGGTCAGGCG-3'	Intron 3 R	[11]
	5'-ATACAAATGGGAAGGTCATCK-3'	Intron 3 F	This study
	5'-ATTGAAGCTGAACTCAAGAAC-3'	Intron 4 R	[2]
RAET1N/ULBP3	5'-ACACAGAACAGCCCCTGAG-3'	Intron 3 R	[2]
	5'-CATAGGAGGATGTGGGACAG-3'	Intron 3 F	[2]

[a] These primers co-amplify RAET1H/ULBP2 and RAET1L/ULBP6. A 1:100 dilution of the purified amplicon was prepared as a template for nested PCR using locus specific primers as described by Romphruk *et al*. [11].

[b,c] Locus specific amplification primers generating separate products for exons 2 and 3 to serve as template for sequencing primers.

Abbreviations: F – forward primer; R – reverse primer.

Table S2. Hardy Weinberg equilibrium exact test using Markov chain Monte Carlo method

Locus	P value	Observed heterozygosity	Expected heterozygosity	SD
RAET1E	0.92440	0.60345	0.55052	0.00025
RAET1H	0.27741	0.20690	0.23988	0.00043
RAET1L	0.21296	0.34483	0.37616	0.00045

Alternative nomenclature: RAET1E = ULBP4, RAET1H = ULBP2, RAET1L = ULBP6.
Abbreviation: SD – standard deviation.

Figure 1

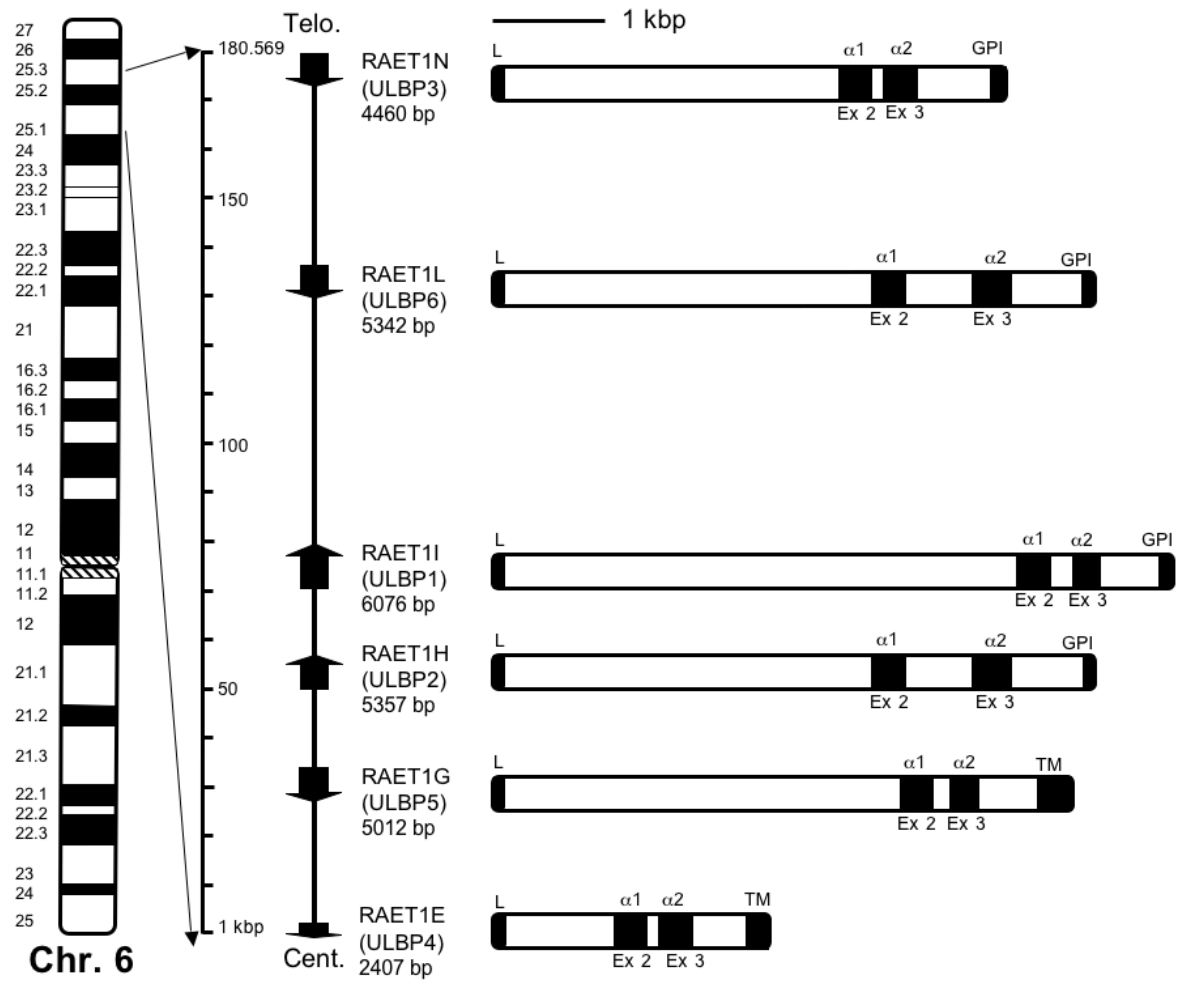
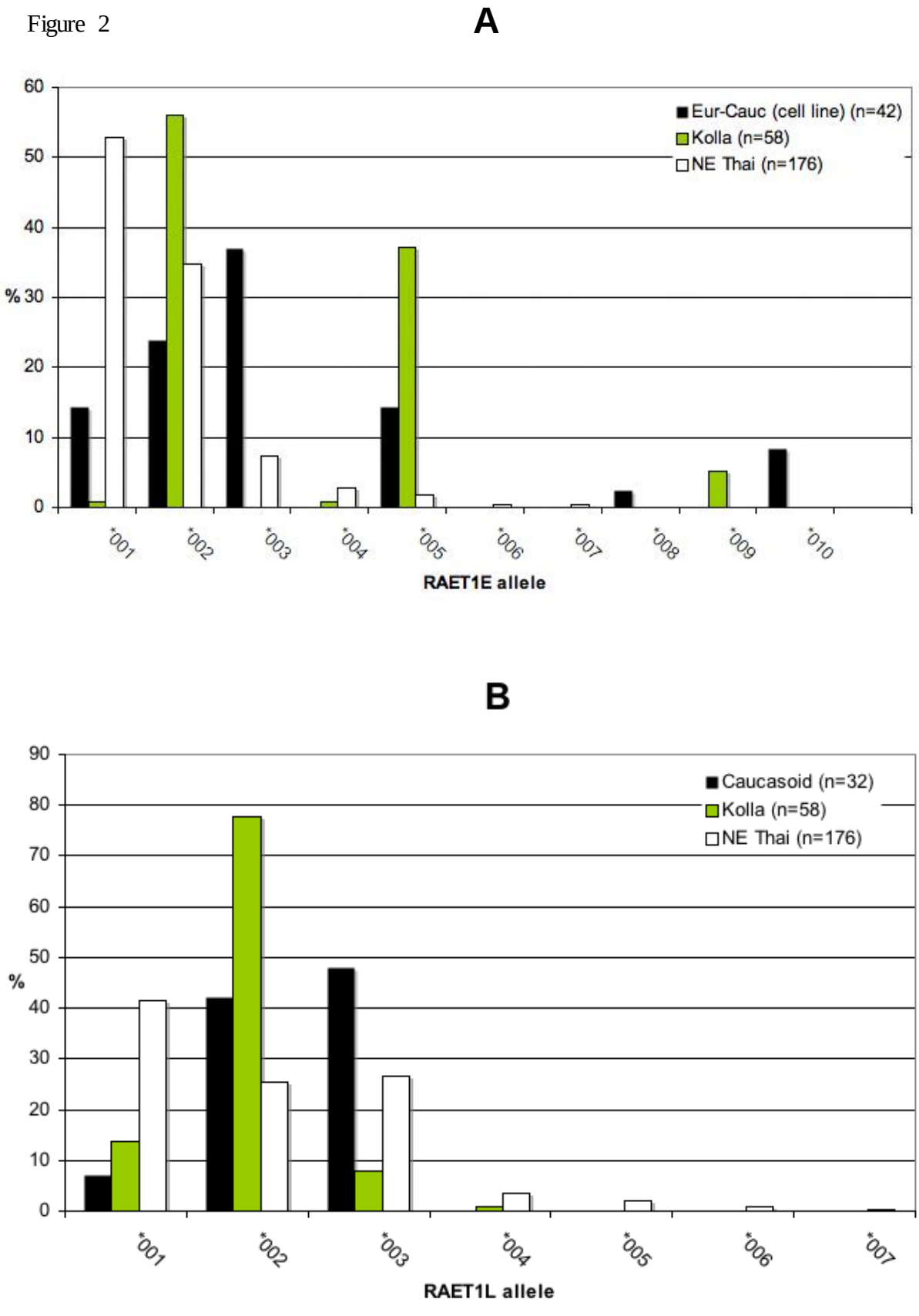

Telo.
1 kbp
180.569
150
100
50
1 kbp
Cent.
Chr. 6
RAET1N
(ULBP3)
4460 bp
RAET1L
(ULBP6)
5342 bp
RAET1I
(ULBP1)
6076 bp
RAET1H
(ULBP2)
5357 bp
RAET1G
(ULBP5)
5012 bp
RAET1E
(ULBP4)
2407 bp
L
α1
α2
GPI
TM
Ex 2
Ex 3

Figure 2


A
60
50
40
% 30
20
10
0
Eur-Cauc (cell line) (n=42)
Kolla (n=58)
NE Thai (n=176)
*001
*002
*003
*004
*005
*006
*007
*008
*009
*010
RAET1E allele
B
90
80
70
60
50
%
40
30
20
10
0
Caucasoid (n=32)
Kolla (n=58)
NE Thai (n=176)
*001
*002
*003
*004
*005
*006
*007
RAET1L allele