Characterisation of a second gain of function *EDAR* variant, encoding EDAR380R, in East Asia

Jon Riddell[1], Chandana Basu Mallick[1,*], Guy S. Jacobs[2], Jeffrey J. Schoenebeck[1], Denis J. Headon[1]

[1]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, United Kingdom

[2]Complexity Institute, Nanyang Technological University, Singapore

[*]Current affiliation: Centre for Genetic Disorders, Institute of Science, Banaras Hindu University, Varanasi, India

Abstract

EDAR is a TNF receptor family member with roles in the development and growth of hair, teeth and glands. A derived allele of *EDAR*, single nucleotide variant rs3827760, encodes EDAR370A, a receptor with more potent signalling effects than the ancestral EDAR370V. This allele of rs3827760 is at very high frequency in modern East Asian and Native American populations as a result of ancient positive selection and has been associated with straighter, thicker hair fibres, alteration of tooth and ear shape, reduced chin protrusion and increased fingertip sweat gland density. Here we report the characterisation of another SNV in *EDAR*, rs146567337, encoding EDAR380R. The derived allele of this SNV is at its highest global frequency, of up to 5%, in populations of southern China, Vietnam, the Philippines, Malaysia and Indonesia. Using haplotype analyses, we find that the rs3827760 and rs146567337 SNVs arose on distinct haplotypes and that rs146567337 does not show the same signs of positive selection as rs3827760. From functional studies in cultured cells, we find that EDAR380R displays increased EDAR signalling output, at a similar level to that of EDAR370A. The existence of a second SNV with partly overlapping geographic distribution, the same *in vitro* functional effect and similar evolutionary age as the derived allele of rs3827760, but of independent origin and not exhibiting the same signs of strong selection, suggests a northern focus of positive selection on EDAR function in East Asia.

Introduction

Ectodysplasin A1 receptor (EDAR) is a cell surface receptor involved in the development of ectodermal structures including hair, teeth and glands (Headon and Overbeek, 1999). Upon activation by its ligand Ectodysplasin (EDA), EDAR signals through its cytoplasmic adapter protein EDARADD to trigger the activation of the transcription factor NF-κB (Headon et al., 2001), this signalling sequence being essential for its developmental function (Schmidt-Ullrich *et al.*, 2001). Disruption of this highly conserved signalling pathway via loss-of-function mutations of any of *EDA*, *EDAR* or *EDARADD* causes hypohidrotic ectodermal dysplasia (HED) (Kere *et al.*, 1996; Monreal *et al.*, 1999; Headon *et al.*, 2001), a condition characterised by sparseness of hair, the loss or reduction of many skin-associated glands, and tooth agenesis (Reyes-Reali *et al.*, 2018). Selective absence of teeth commonly occurs as a result of milder function-reducing mutations in the EDA-EDAR pathway, without eliciting the complete set of clinical HED phenotypes (Arte *et al.*, 2013).

The death domain of the EDAR protein is essential for the recruitment of the EDARADD protein and thus for EDAR function (Headon *et al.*, 2001). This domain is present in many proteins, the majority of which are involved in cell death and inflammation (Park *et al.*, 2007). The death domain is approximately 80 amino acids in length and is composed of six alpha helices, these forming a surface that is capable of self-association and of binding to other specific death domain containing proteins (Park *et al.*, 2007; Ferrao and Wu, 2012).

A non-synonymous single nucleotide variant (SNV), rs3827760 (*EDAR:*c.1109T>C), encodes a valine to alanine substitution within the death domain of EDAR at amino acid position 370. The derived allele is at very high frequency in northern East Asian and Native American populations, with allele frequencies of up to 90% in some groups (The 1000 Genomes Project Consortium, 2015). The *EDAR:*c.1109T>C allele displays clear evidence of positive selection both from haplotype and allele frequency spectrum based analyses (Carlson *et al.*, 2005; Sabeti *et al.*, 2007; Bryk *et al.*, 2008; Myles *et al.*, 2008; Grossman *et al.*, 2010; Cheng, Xu and DeGiorgio, 2017); at least some of this selection presumably occurred in the common ancestors of modern East Asian and Native American populations. EDAR370A has been shown to increase the activation of NF-κB compared to that of the protein encoded by the ancestral allele (EDAR370V) *in vitro* using reporter assays (Bryk et al. 2008; Mou et al. 2008), and ameliorate the clinical signs of HED caused by hypomorphic *EDA* mutations in heterozygous carriers of EDAR370A (Cluzeau *et al.*, 2012), strongly indicating that the derived allele is a gain-of-function. The physiological consequences of this increased signalling have been assessed in mouse models, either with multiple copies of *EDAR* to increase expression level and signalling, or through engineering of the *EDAR:*c.1109T>C variant in mice (Mou *et al.*, 2008; Kamberov *et al.*, 2013). Both of these models were observed to have thicker hair fibres and, complementing these findings, human association studies have shown that *EDAR:*c.1109T>C is associated with thicker, straighter scalp hair, along with other traits such as shovelling of incisors, altered ear and chin shape, and increased fingertip sweat gland density (Fujimoto *et al.*, 2008; Kimura *et al.*, 2009; Kamberov *et al.*, 2013; Tan *et al.*, 2013; Adhikari *et al.*, 2015; Adhikari, Fontanil, *et al.*, 2016; Adhikari, Fuentes-Guajardo, *et al.*, 2016; Wu *et al.*, 2016; Shaffer *et al.*, 2017).

Here we identify another SNV in *EDAR* (rs146567337, *EDAR*:c.1138A>C) which causes a serine to arginine substitution at amino acid position 380 (EDAR380R). The geographic distribution of the derived allele of this SNV partly overlaps that of the previously characterised *EDAR:*c.1109T>C (encoding EDAR370A), though at lower frequency and with a more southerly prevalence. The EDAR380R substitution increases the signalling function of EDAR to a similar degree as the EDAR370A substitution, but its genomic context does not show the same signs of strong positive selection in human populations, despite both alleles having approximately the same age (Albers and McVean, 2019). These findings suggest that *EDAR*:c.1138A>C (EDAR380R) may influence the same human traits as those associated with *EDAR:*c.1109T>C (EDAR370A), and that these traits may have been under different selective pressures in different regions of Asia.

Results

The death domain of EDAR is a highly conserved region of the protein, with mutations altering this domain commonly leading to a loss-of-function and thus clinically diagnosed hypohidrotic ectodermal dysplasia (Cluzeau *et al.*, 2011), presumably due to altered or abrogated EDAR interaction with EDARADD (Okita *et al.*, 2019). We identified SNV rs146567337

(*EDAR*:c.1138A>C) in *EDAR* in the gnomAD database (https://gnomad.broadinstitute.org/) (Karczewski *et al.*, 2019). The derived allele encodes a serine to arginine substitution at the highly conserved amino acid 380 (EDAR380R), only 10 amino acids from the alteration in the well-characterised EDAR370A variant (Fig 1A). In the gnomAD database (Karczewski *et al.*, 2019), which is not a representative sampling of worldwide populations, the frequency of the derived allele at rs146567337 was 1.85%. Using publicly available datasets (The 1000 Genomes Project Consortium, 2015; Mallick *et al.*, 2016; Pagani *et al.*, 2016; Jacobs *et al.*, 2019), we found *EDAR*:c.1138A>C only in East and Southeast Asian populations, at highest frequency in southern China, Vietnam, the Philippines, Malaysia, and Indonesia. However, the distribution of this allele did not extend further south and east into New Guinean populations (Fig 1B and 1C). Since *EDAR:*c.1109T>C is at very high frequency in many of the populations with appreciable frequencies of *EDAR*:c.1138A>C, we assessed whether *EDAR*:c.1138A>C and *EDAR:*c.1109T>C appear on the same haplotype. Using the same datasets, we analysed haplotypes spanning a 20 kb window surrounding rs146567337, on which *EDAR*:c.1138A>C is present and found only one occurrence, out of 33 assessed *EDAR*:c.1138A>C haplotypes from 5,608 evaluated chromosomes, where *EDAR:*c.1109T>C and *EDAR*:c.1138A>C co-existed on the same haplotype. This singular occurrence, possibly a genotyping or phasing error, suggests that the derived allele of rs146567337 arose on a different haplotype to that of *EDAR:*c.1109T>C (Fig 1D). We also found in modern human populations the entire haplotype context of the *EDAR*:c.1138A>C allele, but with the ancestral allele at this SNV, likely representing the immediately ancestral haplotype to *EDAR*:c.1138A>C on which this mutation occurred. This immediately ancestral haplotype is geographically widely dispersed, in East Asian, South Asian, African, and American populations. The variant was also found to be ancestral in both the Altai Neanderthal (Prüfer *et al.*, 2014) and Altai Denisovan (Meyer *et al.*, 2012) genomes, and was not inferred to be in archaic introgressed haplotypes identified in either the Simons Genome Diversity Project (Mallick *et al.*, 2016) or Indonesian Genome Diversity Project data (Jacobs *et al.*, 2019). Hence, we infer that the *EDAR*:c.1138A>C variant arose in modern humans rather than through introgression from an archaic human population.

To further define the distribution of *EDAR* haplotypes, we constructed a median-joining haplotype network consisting of 142 SNPs spanning about ±10 kb around SNVs of interest (rs3827760 and rs146567337) using a publicly available dataset (Pagani *et al.*, 2016) (Fig 2). We used the HapMap combined genetic map (Frazer *et al.*, 2007) to confirm that the window did not have especially fast recombination likely to disrupt the network reconstruction (0.0797cM; 44th percentile of total genetic map distance in non-overlapping genome wide 114.9kb windows). The network identified 88 haplotypes and further supports the independent origins of *EDAR*:c.1138A>C and *EDAR*:c.1109T>C. The haplotype associated with *EDAR*:c.1109T>C is mainly composed of individuals from East Asia, Siberia, Southeast Asia Island and mainland populations, and the Americas (Fig 2). Individuals from Siberia (denoted by cyan) represent almost 50% of this haplotype in this population sample. The dataset used for the construction of this network sampled few East Asian individuals (n=11) than Siberians (n=108), thus explaining the greater proportion of the latter with the associated haplotype (Pagani *et al.*, 2016). We also observed that the *EDAR*:c.1109T>C associated haplotype demonstrates a star-like pattern, suggestive of a historic demographic expansion and corroborating earlier evidence of positive selection at this locus. In contrast, the haplotype associated with *EDAR*:c.1138A>C was found to be distant from *EDAR*:c.1109T>C and showed more restricted geographic distribution, confined to individuals mainly from the islands of Southeast Asia and one individual from South Asia.

As *EDAR:*c.1109T>C is one of the most well-supported examples of a positively selected locus in the human genome (Carlson *et al.*, 2005; Kelley *et al.*, 2006; Sabeti *et al.*, 2007; Xue *et al.*, 2009; Hider *et al.*, 2013), we tested for indications of selection on *EDAR*:c.1138A>C. Using a large-scale whole genome sequence dataset of the Han Chinese population (Cai *et al.*, 2017), we constructed extended haplotype homozygosity (EHH) plots of 433 *EDAR*:c.1138A>C haplotypes (derived rs146567337, ancestral rs3827760) and 20,293 *EDAR:*c.1109T>C haplotypes (derived rs3827760, ancestral rs146567337) against 554 double ancestral haplotypes (haplotypes bearing the ancestral alleles for both rs146567337 and rs3827760) (Fig 3A). No double derived allele haplotypes were found in this Han Chinese dataset. As expected for loci that underwent selection, and as demonstrated previously (Sabeti *et al.*, 2002, 2007), *EDAR:*c.1109T>C shows a broad region of haplotype homozygosity compared to the double ancestral haplotype. *EDAR*:c.1138A>C exhibits much less extended haplotype homozygosity than *EDAR:*c.1109T>C, despite being at a lower derived allele frequency, suggesting that *EDAR*:c.1138A>C has not been subjected to the same pressures or degree of selection as *EDAR:*c.1109T>C. The Han Chinese dataset included 433 *EDAR*:c.1138A>C (EDAR380R) haplotypes, therefore we constructed EHH bifurcation plots by random subsampling of 433 haplotypes from double ancestral allele and *EDAR:*c.1109T>C (EDAR370A) haplotypes. The bifurcation plots confirmed that the *EDAR*:c.1138A>C haplotype had been reduced by recombination less frequently than the double ancestral haplotype, but more frequently than the *EDAR:*c.1109T>C haplotype (Fig 3B).

After determining the global distribution and genomic context of *EDAR*:c.1138A>C, we next investigated the effect of the substitution on the encoded protein. To map the position of EDAR380R within the death domain and identify any predicted structural effects of this amino acid substitution, we modelled the variant EDAR death domain structures using the intensive mode of the Phyre2 server. This program uses multiple homologous templates to predict the structure of a given input protein sequence (Kelley *et al.*, 2015). The resulting predicted protein structure positioned amino acid EDAR380 within an alpha helix (Fig 4A), a structural feature known to be important for the protein–protein interactions mediated by death domains (Ferrao and Wu, 2012). However, the alternate amino acid variants did not alter the predicted structure of this helix or any other part of the death domain. The protein structure also remained unaltered when we modelled the EDAR370A substitution (Fig 4A). Taken together, based on the conservation of the serine residue at position EDAR380 among vertebrates (Fig 1A), introduction of a positive charge through its substitution to arginine and strong evidence of functional alteration reflected from SIFT (score 0) (Ng and Henikoff, 2003) and PolyPhen (score 0.999) (Adzhubei *et al.*, 2010), we predicted that the EDAR380R substitution would alter EDAR protein function. To test this, we transfected HEK293T cells, a human cell line derived from embryonic kidney, with *EDAR* cDNAs encoding either the ancestral EDAR, EDAR370A, EDAR380R, or the double substituted EDAR370A+EDAR380R protein. We also included the known loss-of-function variant EDAR379K, which is dominant for selective tooth agenesis in humans and autosomal recessive for HED in mice (Headon and Overbeek, 1999; Arte *et al.*, 2013), as a control in these experiments. Each form was assayed for its ability to activate a co-transfected NF-κB luciferase reporter. We found that EDAR380R activated NF-κB in these cells to a greater degree than ancestral EDAR, and to the same extent as EDAR370A, and that the greatest activation of NF-κB was observed when EDAR carried both the 370A and 380R amino acid substitutions (Fig 4C). These effects on signalling activity were broadly confirmed in the human HaCaT cell line, derived from the skin's epidermis (Fig 4D).

Discussion

We identified and characterised a novel functional variant in *EDAR* through haplotype analyses and cell-based experiments. We find that *EDAR*:c.1138A>C has its highest allele frequency in Southeast Asia and appears to have arisen on a different haplotype to that of the more common and previously characterised *EDAR:*c.1109T>C substitution. We find that *EDAR*:c.1138A>C does not show the same signs of having been under strong positive selection as *EDAR:*c.1109T>C based on extended haplotype homozygosity analyses, but that the encoded protein increases NF-κB activation *in vitro*, to approximately the same extent as the *EDAR:*c.1109T>C substitution.

Several theories as to what the selective advantage conferred by EDAR370A was have been advanced. Chang et al. suggested that EDAR370A was positively selected in the ancestors of East Asians and Native Americans for adaptation to a cold and dry climate, in which increased skin-associated glands and resulting glandular secretions, perhaps together with straighter hair, could be advantageous in producing a functional barrier to the environment (Chang *et al.*, 2009). Hlusko et al. suggested a latitude-based adaptive scenario, in which altered transfer of nutrients, particularly vitamin D, through breast milk in far northeast Asia (Hlusko *et al.*, 2018) was caused by the mammary gland alterations enacted by enhanced EDAR signalling (Chang *et al.*, 2009; Kamberov *et al.*, 2013). Kamberov et al. placed the origin of the EDAR370A encoding allele in central China at greater than 30,000 years ago, and suggested that increased eccrine sweat gland number, associated with the *EDAR:*c.1109T>C variant in mouse and human in their study, as the potential selective force that would have been advantageous in hot and humid climates due to increased ability to perspire (Kamberov *et al.*, 2013).

A recent genealogical estimation of allele ages in humans assessed the derived alleles *EDAR:*c.1109T>C and *EDAR*:c.1138A>C as having a very similar date of origin, at approximately 1,400 generations ago (Albers and McVean, 2019). The geographic distribution of these alleles is somewhat similar, and, though at much lower frequency than *EDAR:*c.1109T>C in all regions, it is notable that the highest frequencies of *EDAR*:c.1138A>C overlap the more southerly regions in which *EDAR:*c.1109T>C is prevalent. The *EDAR*:c.1138A>C variant is notably absent from the Americas, where in native populations *EDAR:*c.1109T>C is essentially at fixation (Bryk et al. 2008). We found that *EDAR*:c.1138A>C does not show as strong a signal of positive selection in human populations as *EDAR:*c.1109T>C, despite cell culture experiments predicting similar outcomes resulting from the EDAR370A and EDAR380R substitutions. The frequency of the *EDAR*:c.1138A>C variant peaks in Southeast Asia and it is thus most likely to have arisen in that region. The *EDAR:*c.1109T>C variant appears to have arisen further north, based on its present-day population distribution and ancient DNA analyses (Mathieson *et al.*, 2015; Siska *et al.*, 2017; Lamnidis *et al.*, 2018), suggesting that phenotypes associated with an EDAR-dependent increase in NF-κB activation have been preferentially selected for in more northern regions of Asia. The overlap between these alleles could be influenced by frequency dependent selection, perhaps on a phenotype directly perceptible by others.

The possibility that EDAR370A and EDAR380R may have similar phenotypic effects should be considered in future gene or genome-wide association studies, particularly in populations in which the derived allele of rs3827760 is at high frequency. In these populations, only a small fraction of ancestral rs3827760 alleles exist and a sizeable proportion of these haplotype will carry the derived rs146567337 allele, which could obscure the phenotypic associations with the derived allele of rs3827760. The *EDAR*:c.1138A>C allele has been identified in people exhibiting tooth agenesis in East Asian populations (He *et al.*, 2013; Yamaguchi *et al.*, 2017). However, our data suggest that this allele is unlikely to be causative for the condition due to its increased, rather than decreased, activity, as observed for *EDAR*:c.1109T>C.

The discovery of a second SNV in *EDAR* that increases NF-κB activation to the same extent as *EDAR*:c.1109T>C raises questions as to how many routes there are to achieving the same molecular effect of increased EDAR activity. Multiple mutations have also been identified in an enhancer region upstream of the *LCT* gene that have the same molecular effect of increasing *LCT* transcription (Tishkoff *et al.*, 2007). However, these *LCT* SNVs exhibited clear EHH, suggesting that this molecular effect was selected for in each of the populations containing these variants. In this case, *EDAR*:c.1138A>C does not show the same extent of EHH as *EDAR*:c.1109T>C, even though the alleles are predicted to be of a similar age, and should therefore both have had the opportunity to be selected. This indicates that either the molecular consequences of *EDAR*:c.1109T>C are more complex *in vivo* than reflected in cell signalling assays, or that the phenotypic consequences of enhanced *EDAR* signalling have only been strongly selected for in northern East Asian populations. This work highlights that exploring the population genetics of variants with similar molecular phenotypes as known selected variants could prove beneficial in the future for refining the features of those variants, and the relevant environments, that led to their selection.

Materials and Methods

Generation of phylogeographic maps

Maps of the world and of Southeast Asia were generated using MapChart (https://mapchart.net/). The rs3827760 and rs146567337 allele frequencies were gathered from publicly available datasets (The 1000 Genomes Project Consortium, 2015; Pagani *et al.*, 2016; Jacobs *et al.*, 2019) and plotted on to pie charts for each population. The pie charts were used to annotate the maps based on the co-ordinates that samples originated from.

Determination of archaic human genotypes

We used the high coverage Altai Neanderthal (Prufer et al. 2014; downloaded from http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal/VCF/) and Altai Denisovan (Meyer et al. 2012 et al; downloaded from http://cdna.eva.mpg.de/denisova/VCF/hg19_1000g/) to determine the state of rs146567337 in archaic hominins. As the archaic genomes are sampled from populations that are divergent from introgressing archaic populations, we additionally confirmed a non-archaic origin using introgressing archaic haplotypes inferred by Jacobs *et al*. (2019) for samples in the Simons Genome Diversity Project and Indonesian Genome Diversity Project.

Haplotype analysis

VCF files from publicly available datasets (The 1000 Genomes Project Consortium, 2015; Mallick *et al.*, 2016) were narrowed down to a 20 kb window surrounding rs146567337. These files were viewed using inPHAP software (v1.1) (Jäger, Peltzer and Nieselt, 2014) and the variants that disagreed with the human reference genome (GRCh37) were mapped for samples containing *EDAR*:c.1138A>C. This process was repeated for samples containing *EDAR:*c.1109T>C and the list of variants mapped were combined to give a total of 50 SNVs that were used for the final haplotype constructs.

Construction of the median-joining haplotype network of *EDAR*

A median-joining haplotype network was constructed using NETWORK 5.0 (http://www.fluxus-engineering.com) to study the phylogenetic relationship between the haplotypes. For this, we used the region around ± 10 kb of the SNVs of interest (rs3827760 and rs146567337) among the individuals included in the Pagani et al. 2016 dataset.

Generation of extended haplotype homozygosity (EHH) and bifurcation plots

EHH plots and bifurcation plots were constructed from data generated in a large scale whole-genome sequencing dataset of 11,670 individuals from the Han Chinese population (Cai *et al.*, 2017) using R package rehh (v2.0.2) (Sabeti *et al.*, 2002). The EHH plot was generated by separating the haplotypes into three categories: Ancestral – where the ancestral alleles of both rs3827760 and rs146567337 are present, *EDAR*:c.1138A>C – where the derived allele of rs146567337 and the ancestral allele of rs3827760 are present, *EDAR:*c.1109T>C - where the ancestral allele of rs146567337 and the derived allele of rs3827760 are present. In total there were 554 double ancestral haplotypes, 433 *EDAR*:c.1138A>C haplotypes and 20,293 *EDAR:*c.1109T>C haplotypes. In order to aid visualisation in the bifurcation plot, 433 haplotypes from each category were randomly selected and plotted.

Sequence alignment and protein structure modelling

The EDAR death domain sequence alignment was generated with the T-coffee alignment tool (Notredame, Higgins and Heringa, 2000; Di Tommaso *et al.*, 2011) using peptide sequences gathered from the NCBI protein database (GenInfo Identifiers: Human - 11641231, Mouse -

6753714, Chicken - 60302666, Zebrafish - 924859488, Xenopus - 55742031). The generated fasta_aln file was then entered into BOXSHADE v3.21 to obtain a shaded output indicating amino acid sequence conservation.

The EDAR death domain protein structure was generated using the intensive mode of Phyre2 v2.0 (Kelley *et al.*, 2015) by inputting amino acid positions 345-431 of the human EDAR peptide sequence gathered from the NCBI protein database (GenInfo Identifier: 11641231).

Transfection of cells and luciferase assays

HEK293T and HaCaT cells were maintained at 37°C in 5% CO_2 in high glucose Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% foetal bovine serum (FBS) and 50 µg/ml streptomycin and 100 U/ml penicillin (Gibco). Transfections of HEK293T and HaCaT cells were performed using Lipofectamine 3000 (Invitrogen) in 24-well plates (well surface area: 1.9 cm^2). Both cell lines were seeded at a density of 5 x 10^4 24 hours prior to transfection. Each well was transfected with a DNA mix diluted in opti-MEM (Gibco) consisting of 125 ng pNFκB-luc, 62.5 ng pRLTK, 10 ng pCR3::EDAR, pCR3::EDAR370A, pCR3::EDAR380R, pCR3::EDAR370A/EDAR380R or pCR3::EDAR379K, and made up to a total amount of 500 ng with empty pCR3.1 vector. Transfections were performed according to manufacturer's instructions in DMEM supplemented with 10% FBS, 50 µg/ml streptomycin and 100 U/ml penicillin.

Luciferase assays were performed 18 hours post-transfection using the Dual-Luciferase Assay System (Promega) according to manufacturer's instructions, using a MicroLumatPlus LB96V Microplate Luminometer.

References

Adhikari, K. *et al.* (2015) 'A genome-wide association study identifies multiple loci for variation in human ear morphology', *Nature Communications*, 6, p. 7500. doi: 10.1038/ncomms8500.

Adhikari, K., Fuentes-Guajardo, M., *et al.* (2016) 'A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation', *Nature Communications*, 7, p. 11616. doi: 10.1038/ncomms11616.

Adhikari, K., Fontanil, T., *et al.* (2016) 'A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features', *Nature Communications*, 7, p. 10815. doi: 10.1038/ncomms10815.

Adzhubei, I. A. *et al.* (2010) 'A method and server for predicting damaging missense mutations', *Nature Methods*, 7(4), pp. 248–249. doi: 10.1038/nmeth0410-248.

Albers, P. K. and McVean, G. (2019) 'Dating genomic variants and shared ancestry in population-scale sequencing data', *bioRxiv*, p. 416610. doi: 10.1101/416610.

Arte, S. *et al.* (2013) 'Candidate Gene Analysis of Tooth Agenesis Identifies Novel Mutations in Six Genes and Suggests Significant Role for WNT and EDA Signaling and Allele Combinations', *PLoS One*, 8(8), pp. 1–12. doi: 10.1371/journal.pone.0073705.

Bryk, J. *et al.* (2008) 'Positive selection in East Asians for an EDAR allele that enhances NF-kappaB activation', *PLoS ONE*, 3(5). doi: 10.1371/journal.pone.0002209.

Cai, N. *et al.* (2017) '11,670 whole-genome sequences representative of the Han Chinese population from the CONVERGE project', *Scientific Data*, 4. doi: 10.1038/sdata.2017.11.

Carlson, C. S. *et al.* (2005) 'Genomic regions exhibiting positive selection identified from dense genotype data', *Genome Research*, 15(11), pp. 1553–1565. doi: 10.1101/gr.4326505.

Chang, S. H. *et al.* (2009) 'Enhanced Edar Signalling Has Pleiotropic Effects on Craniofacial and Cutaneous Glands', *PLoS ONE*, 4(10). doi: 10.1371/journal.pone.0007591.

Cheng, X., Xu, C. and DeGiorgio, M. (2017) 'Fast and robust detection of ancestral selective sweeps', *Molecular Ecology*, 26(24), pp. 6871–6891. doi: 10.1111/mec.14416.

Cluzeau, C. *et al.* (2011) 'Only four genes (EDA1, EDAR, EDARADD, and WNT10A) account for 90% of hypohidrotic/anhidrotic ectodermal dysplasia cases', *Human Mutation*, 32(1), pp. 70–77. doi: 10.1002/humu.21384.

Cluzeau, C. *et al.* (2012) 'The EDAR370A allele attenuates the severity of hypohidrotic ectodermal dysplasia caused by EDA gene mutation', *British Journal of Dermatology*, 166(3), pp. 678–681. doi: 10.1111/j.1365-2133.2011.10620.x.

Ferrao, R. and Wu, H. (2012) 'Helical assembly in the death domain (DD) superfamily', *Current Opinion in Structural Biology*, 22(2), pp. 241–247. doi: 10.1016/j.sbi.2012.02.006.

Frazer, K. A. *et al.* (2007) 'A second generation human haplotype map of over 3.1 million SNPs', *Nature*, 449(7164), pp. 851–861. doi: 10.1038/nature06258.

Fujimoto, A. *et al.* (2008) 'A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness', *Human Molecular Genetics*, 17(6), pp. 835–843. doi: 10.1093/hmg/ddm355.

Grossman, S. R. *et al.* (2010) 'A composite of multiple signals distinguishes causal variants in regions of positive selection', *Science*, 327(5967), pp. 883–886. doi: 10.1126/science.1183863.

He, H. *et al.* (2013) 'Involvement of and interaction between WNT10A and EDA mutations in tooth agenesis cases in the Chinese population', *PLoS ONE*, 8(11). doi: 10.1371/journal.pone.0080393.

Headon, D. J. *et al.* (2001) 'Gene defect in ectodermal dysplasia implicates a death domain adapter in development', *Nature*, 414(6866), pp. 913–916. doi: 10.1038/414913a.

Headon, D. J. and Overbeek, P. a (1999) 'Involvement of a novel Tnf receptor homologue in hair follicle induction.', *Nature Genetics*, 22(4), pp. 370–4. doi: 10.1038/11943.

Hider, J. L. *et al.* (2013) 'Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry', *BMC Evolutionary Biology*, 13(1). doi: 10.1186/1471-2148-13-150.

Hlusko, L. J. *et al.* (2018) 'Environmental selection during the last ice age on the mother-to-infant transmission of vitamin D and fatty acids through breast milk', *Proceedings of the National Academy of Sciences*, 115(19), pp. E4426–E4432. doi: 10.1073/pnas.1711788115.

Jacobs, G. S. *et al.* (2019) 'Multiple Deeply Divergent Denisovan Ancestries in Papuans', *Cell*, 177(4), pp. 1010–1021.e32. doi: 10.1016/j.cell.2019.02.035.

Jäger, G., Peltzer, A. and Nieselt, K. (2014) 'InPHAP: Interactive visualization of genotype and phased haplotype data', *BMC Bioinformatics*, 15(1). doi: 10.1186/1471-2105-15-200.

Kamberov, Y. G. *et al.* (2013) 'Modeling recent human evolution in mice by expression of a selected EDAR variant', *Cell*, 152(4), pp. 691–702. doi: 10.1016/j.cell.2013.01.016.

Karczewski, K. J. *et al.* (2019) 'Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes', *bioRxiv*, p. 531210. doi: 10.1101/531210.

Kelley, J. L. *et al.* (2006) 'Genomic signatures of positive selection in humans and the limits of outlier approaches', *Genome Research*, 16(8), pp. 980–989. doi: 10.1101/gr.5157306.

Kelley, L. A. *et al.* (2015) 'The Phyre2 web portal for protein modeling, prediction and analysis', *Nature Protocols*, 10(6), pp. 845–858. doi: 10.1038/nprot.2015.053.

Kere, J. *et al.* (1996) 'X-linked anhidrotic (hypohidrotic) ectodermal dysplasia is caused by mutation in a novel transmembrane protein', *Nature Genetics*, 13(4), pp. 409–416. doi: 10.1038/ng0895-409.

Kimura, R. *et al.* (2009) 'A Common Variation in EDAR Is a Genetic Determinant of Shovel-Shaped Incisors', *American Journal of Human Genetics*, 85(4), pp. 528–535. doi: 10.1016/j.ajhg.2009.09.006.

Lamnidis, T. C. *et al.* (2018) 'Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe', *Nature Communications*, 9(1), p. 5018. doi: 10.1038/s41467-018-07483-5.

Mallick, S. *et al.* (2016) 'The Simons Genome Diversity Project: 300 genomes from 142 diverse populations', *Nature*, 538(7624), pp. 201–206. doi: 10.1038/nature18964.

Mathieson, I. *et al.* (2015) 'Genome-wide patterns of selection in 230 ancient Eurasians', *Nature*, 528(7583), pp. 499–503. doi: 10.1038/nature16152.

Meyer, M. *et al.* (2012) 'A high-coverage genome sequence from an archaic Denisovan individual', *Science*, 338(6104), pp. 222–226. doi: 10.1126/science.1224344.

Monreal, a W. *et al.* (1999) ‘Mutations in the human homologue of mouse dl cause autosomal recessive and dominant hypohidrotic ectodermal dysplasia.’, *Nature Genetics*, 22(4), pp. 366–369. doi: 10.1038/11937.

Mou, C. *et al.* (2008) ‘Enhanced Ectodysplasin-A receptor (EDAR) signaling alters multiple fiber characteristics to produce the east Asian hair form’, *Human Mutation*, 29(12), pp. 1405–1411. doi: 10.1002/humu.20795.

Myles, S. *et al.* (2008) ‘Identification and analysis of genomic regions with large between-population differentiation in humans’, *Annals of Human Genetics*, 72(1), pp. 99–110. doi: 10.1111/j.1469-1809.2007.00390.x.

Ng, P. C. and Henikoff, S. (2003) ‘SIFT: Predicting amino acid changes that affect protein function’, *Nucleic Acids Research*, 31(13), pp. 3812–3814. doi: 10.1093/nar/gkg509.

Notredame, C., Higgins, D. G. and Heringa, J. (2000) ‘T-coffee: a novel method for fast and accurate multiple sequence alignment’, *Journal of Molecular Biology*, 302(1), pp. 205–217. doi: 10.1006/jmbi.2000.4042.

Okita, T. *et al.* (2019) ‘Functional studies for a dominant mutation in the EDAR gene responsible for hypohidrotic ectodermal dysplasia’, *Journal of Dermatology*. doi: 10.1111/1346-8138.14983.

Pagani, L. *et al.* (2016) ‘Genomic analyses inform on migration events during the peopling of Eurasia’, *Nature*, 538(7624), pp. 238–242. doi: 10.1038/nature19792.

Park, H. H. *et al.* (2007) ‘The Death Domain Superfamily in Intracellular Signaling of Apoptosis and Inflammation’, *Annual Review of Immunology*, 25(1), pp. 561–586. doi: 10.1146/annurev.immunol.25.022106.141656.

Prüfer, K. *et al.* (2014) ‘The complete genome sequence of a Neanderthal from the Altai Mountains’, *Nature*, 505(7481), pp. 43–49. doi: 10.1038/nature12886.

Reyes-Reali, J. *et al.* (2018) ‘Hypohidrotic ectodermal dysplasia: clinical and molecular review’, *International Journal of Dermatology*, 57(8), pp. 965–972. doi: 10.1111/ijd.14048.

Sabeti, P. C. *et al.* (2002) ‘Detecting recent positive selection in the human genome from haplotype structure’, *Nature*, 419(6909), pp. 832–837. doi: 10.1038/nature01140.

Sabeti, P. C. *et al.* (2007) ‘Genome-wide detection and characterization of positive selection in human populations’, *Nature*, 449(7164), pp. 913–918. doi: 10.1038/nature06250.

Schmidt-Ullrich, R. *et al.* (2001) ‘Requirement of NF-kappaB/Rel for the development of hair follicles and other epidermal appendices.’, *Development (Cambridge, England)*, 128, pp. 3843–3853.

Shaffer, J. R. *et al.* (2017) ‘Multiethnic GWAS Reveals Polygenic Architecture of Earlobe Attachment’, *American Journal of Human Genetics*, 101(6), pp. 913–924. doi: 10.1016/j.ajhg.2017.10.001.

Siska, V. *et al.* (2017) ‘Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago’, *Science Advances*, 3(2), p. e1601877. doi: 10.1126/sciadv.1601877.

Tan, J. *et al.* (2013) 'The adaptive variant EDARV370A is associated with straight hair in East Asians', *Human Genetics*, 132(10), pp. 1187–1191. doi: 10.1007/s00439-013-1324-1.

The 1000 Genomes Project Consortium (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), pp. 68–74. doi: 10.1038/nature15393.

Tishkoff, S. A. *et al.* (2007) 'Convergent adaptation of human lactase persistence in Africa and Europe', *Nature Genetics*, 39(1), pp. 31–40. doi: 10.1038/ng1946.

Di Tommaso, P. *et al.* (2011) 'T-Coffee: A web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension', *Nucleic Acids Research*, 39(SUPPL. 2). doi: 10.1093/nar/gkr245.

Wu, S. *et al.* (2016) 'Genome-wide scans reveal variants at EDAR predominantly affecting hair straightness in Han Chinese and Uyghur populations', *Human Genetics*, 135(11), pp. 1279–1286. doi: 10.1007/s00439-016-1718-y.

Xue, Y. *et al.* (2009) 'Population differentiation as an indicator of recent positive selection in humans: An empirical evaluation', *Genetics*, 183(3), pp. 1065–1077. doi: 10.1534/genetics.109.107722.

Yamaguchi, T. *et al.* (2017) 'Comprehensive genetic exploration of selective tooth agenesis of mandibular incisors by exome sequencing', *Human Genome Variation*, 4(1). doi: 10.1038/hgv.2017.5.

Figures

Figure 1: Conservation, distribution and haplotype structure of *EDAR* variants

(A) Multiple sequence alignment of vertebrate EDAR death domains. Amino acid positions within the EDAR protein are numbered at the start and end of the sequence for each species. The position of 370 is indicated by a blue triangle, the position of 380 by a red triangle. The positions of known recessive and dominant mutations causing hypohidrotic ectodermal dysplasia in humans are indicated by black and orange squares, respectively, above the alignment. Purple bars below the alignment indicate the positions of the predicted alpha helices. **(B)** Worldwide allele frequencies for *EDAR:*c.1109T>C and *EDAR*:c.1138A>C in the 1000 Genomes dataset plotted as pie charts for each population. The remaining allele frequency was depicted as ancestral. **(C)** *EDAR:*c.1109T>C and *EDAR*:c.1138A>C allele frequencies in the Southeast Asian Island populations were gathered from publicly available datasets (Pagani *et al.*, 2016; Jacobs *et al.*, 2019) and plotted on a map of Southeast Asia as in **(B)**. The area of each chart is proportional to the sample number of each population. **(D)** Diagram of the *EDAR*:c.1138A>C haplotypes from the 1000 Genomes Project and Simons Genome Diversity Project datasets plotted for a region 10 kb upstream and 10 kb downstream of rs146567337. White boxes indicate alleles matching the reference human genome (GRCh37) and grey boxes indicate the presence of the alternate allele. Red shading indicates the *EDAR* 1138C allele and blue shading indicates *EDAR* 1109C. In total, 33 *EDAR*:c.1138A>C haplotypes were present in these datasets, with five unique haplotype structures identified. These were ranked in order of frequency, as shown by the percentages to the right of each haplotype, with the total number of each individual haplotype in the dataset indicated in brackets. The scale bar indicates position on chromosome 2.

Figure 2: Median-joining haplotype network of *EDAR*

Median-joining haplotype network spanning ± 10 kb around rs3827760 and rs146567337 showing the relationship of the haplotypes. The network is based on 446 individuals included in the Pagani *et al.*, 2016 dataset. Each pie chart represents a unique haplotype and the size of the chart is proportional to the number of chromosomes carrying it. Colours represent the geographic location of populations where each haplotype was found. Lines represent mutations, with greater branch length indicating a greater number of distinguishing mutations. The associated haplotypes of interest (for *EDAR:*c.1109T>C and *EDAR*:c.1138A>C) have been labelled. The black arrows represent locations where an *EDAR:*c.1109T>C mutational event was inferred. The multiple arrows likely reflect ambiguity in the network reconstruction.

Figure 3: Extended haplotype homozygosity (EHH) and EHH bifurcation plots surrounding *EDAR* variants

(A) EHH plot showing the length of conserved haplotype on either side of rs146567337. An EHH value of 1 indicates that haplotypes are identical at this position. Double ancestral haplotypes are represented by the black dotted line, *EDAR*:c.1138A>C (EDAR380R) haplotypes are represented by the red line, and *EDAR:*c.1109T>C (EDAR370A) haplotypes are represented by the blue line. **(B)** Bifurcation plot showing the branching of each haplotype. Thicker lines indicate more common haplotypes. Double ancestral haplotypes are represented by the black line, *EDAR*:c.1138A>C haplotypes are represented by the red line, and *EDAR:*c.1109T>C haplotypes are represented by the blue line.

Figure 4: Functional effects of *EDAR* variants

(A) The predicted protein structures of the ancestral, EDAR370A and EDAR380R death domains, modelled using Phyre2. The location of amino acid positions 370 and 380 are indicated by arrows. Amino acid position 380 is located towards the end of an alpha helix. **(C)** HEK293T and **(D)** HaCaT cells were transfected to express *EDAR* variants and resulting NF-κB luciferase reporter activity determined. Error bars represent the standard error of the mean from experiments performed in quadruplicate and repeated independently 6 times. Statistical significance was calculated using a Student *t* test (** P < 0.005, *** P <0.0005).

Figure 1

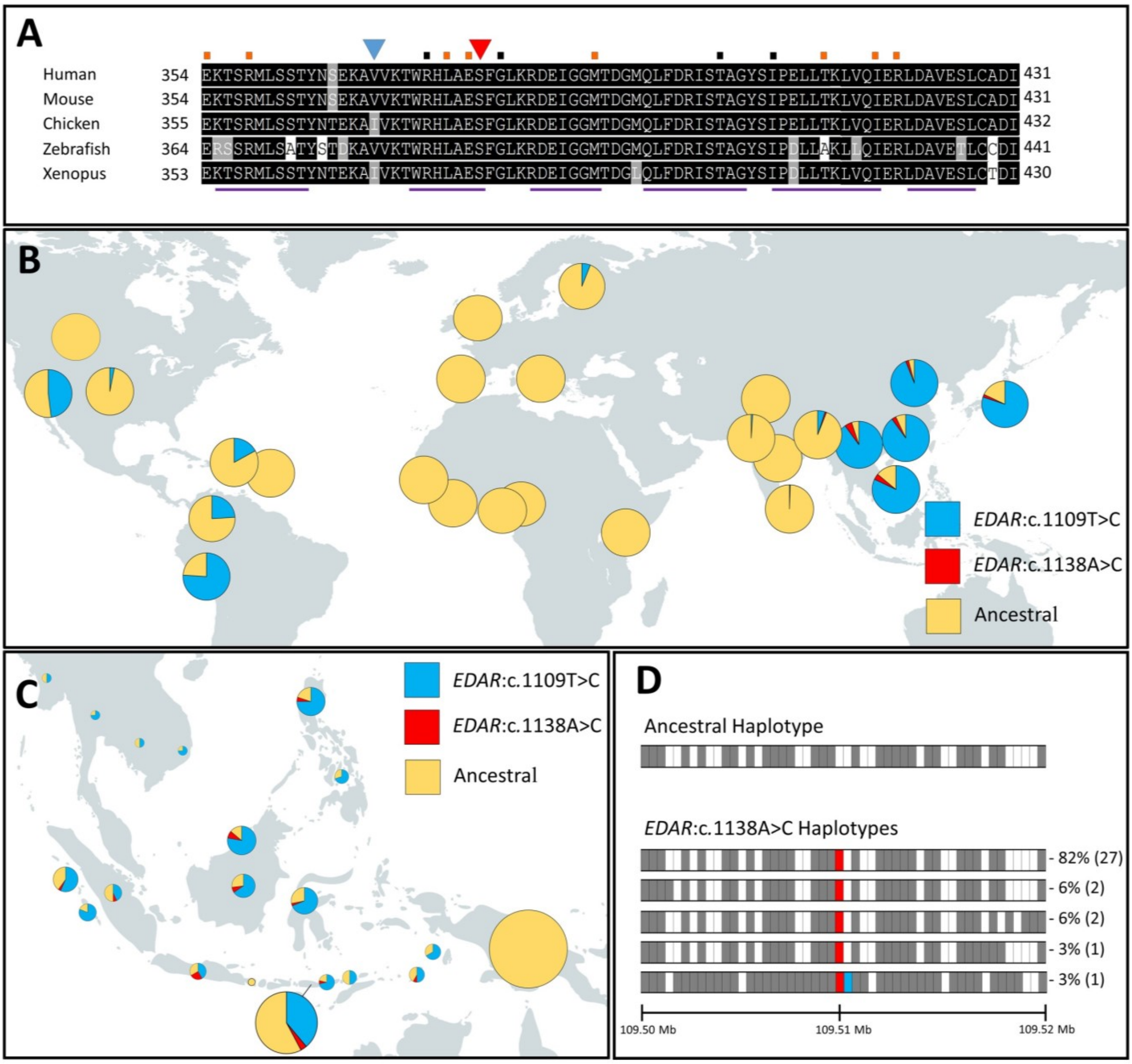

A
Human 354 EKTSRMLSSTYNSEKAVVKTWRHLAESFGLKRDEIGGMTDGMQLFDRISTAGYSIPELLTKLVQIERLDAVESLCADI 431
Mouse 354 EKTSRMLSSTYNSEKAVVKTWRHLAESFGLKRDEIGGMTDGMQLFDRISTAGYSIPELLTKLVQIERLDAVESLCADI 431
Chicken 355 EKTSRMLSSTYNTEKAIVKTWRHLAESFGLKRDEIGGMTDGMQLFDRISTAGYSIPELLTKLVQIERLDAVESLCADI 432
Zebrafish 364 ERSSRMLSATYSTDKAVVKTWRHLAESFGLKRDEIGGMTDGMQLFDRISTAGYSIPDLLAKLLQIERLDAVETLCCDI 441
Xenopus 353 EKTSRMLSSTYNTEKAIVKTWRHLAESFGLKRDEIGGMTDGLQLFDRISTAGYSIPDLLTKLVQIERLDAVESLCTDI 430
B
EDAR:c.1109T>C
EDAR:c.1138A>C
Ancestral
C
EDAR:c.1109T>C
EDAR:c.1138A>C
Ancestral
D
Ancestral Haplotype
EDAR:c.1138A>C Haplotypes
- 82% (27)
- 6% (2)
- 6% (2)
- 3% (1)
- 3% (1)
109.50 Mb
109.51 Mb
109.52 Mb

Figure 2

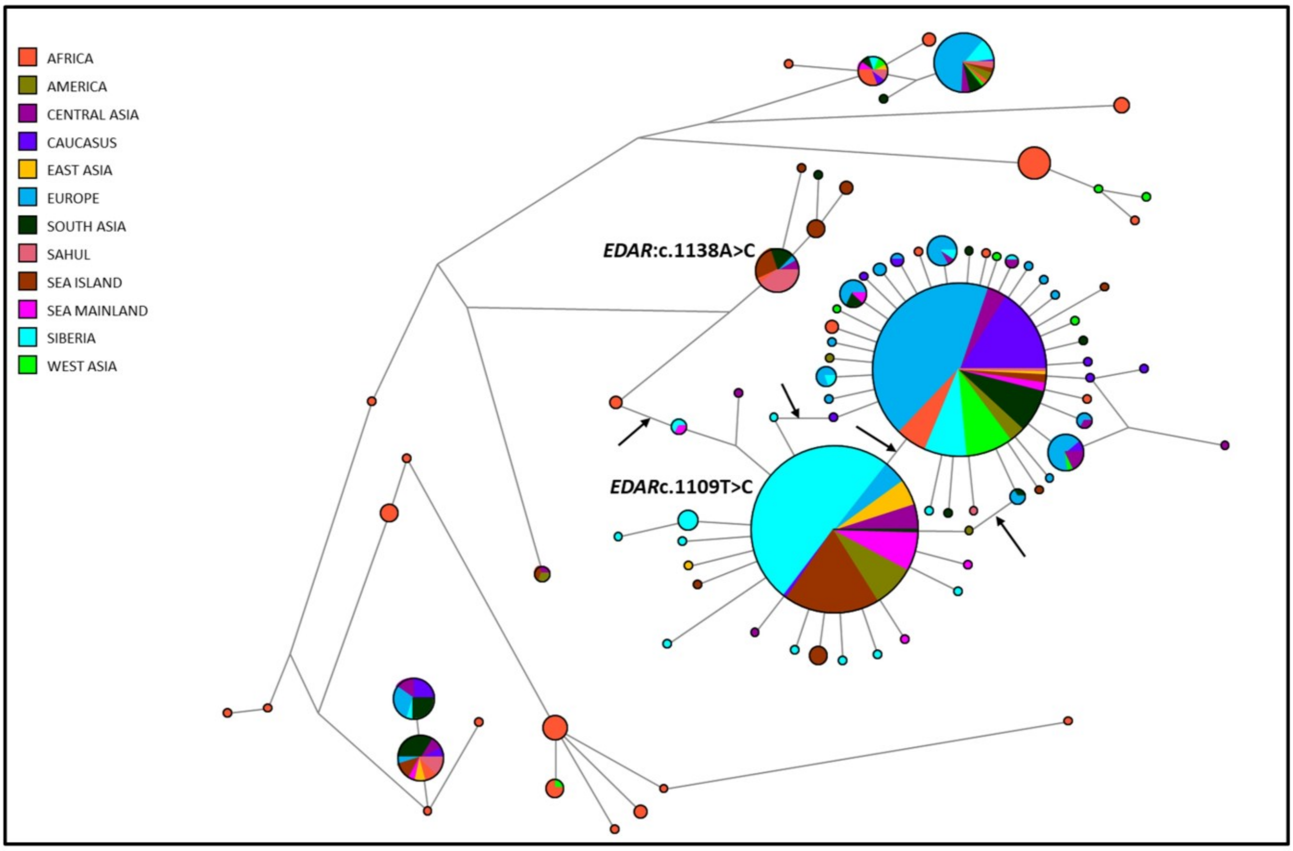

AFRICA
AMERICA
CENTRAL ASIA
CAUCASUS
EAST ASIA
EUROPE
SOUTH ASIA
SAHUL
SEA ISLAND
SEA MAINLAND
SIBERIA
WEST ASIA
EDAR:c.1138A>C
*EDAR*c.1109T>C

Figure 3

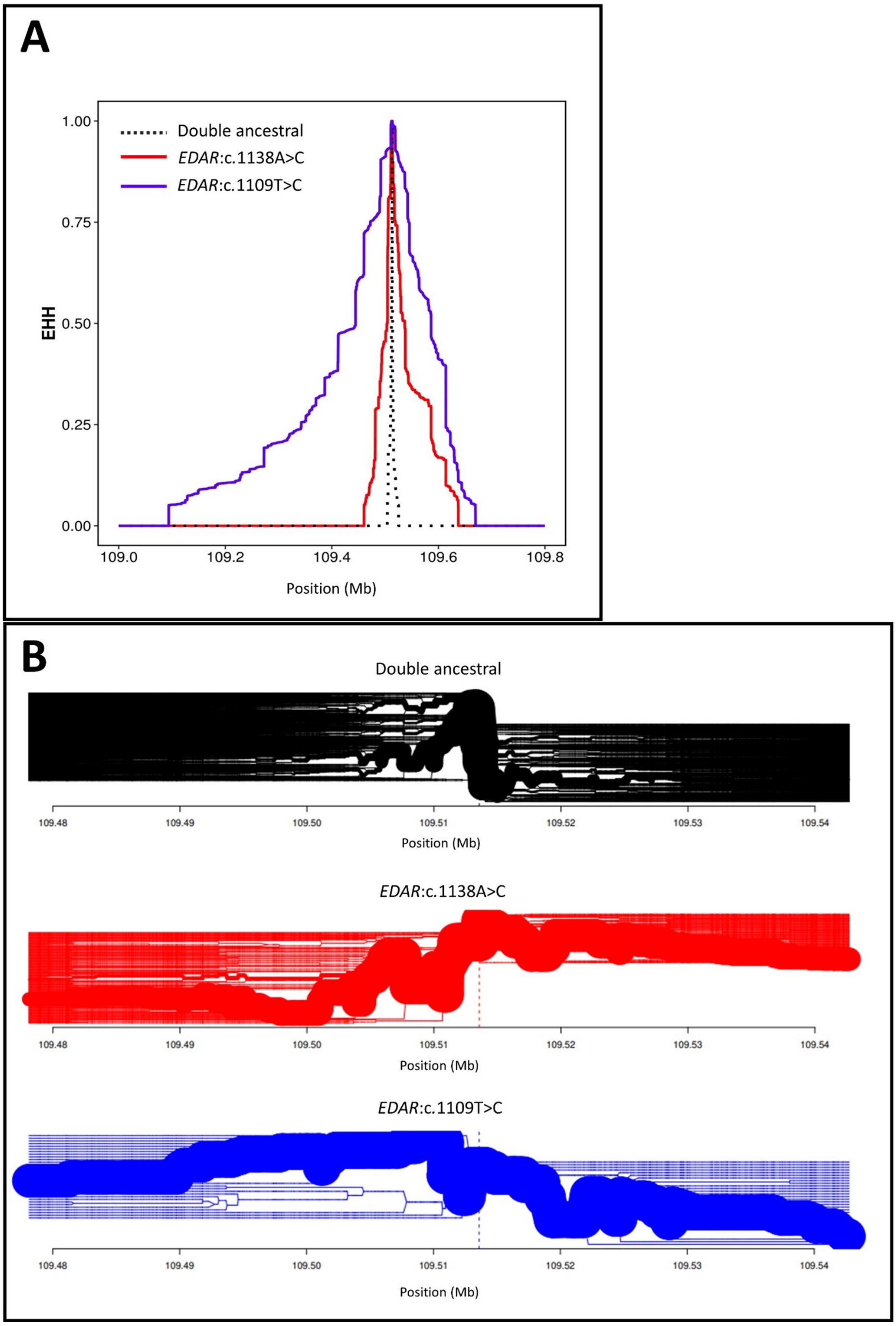
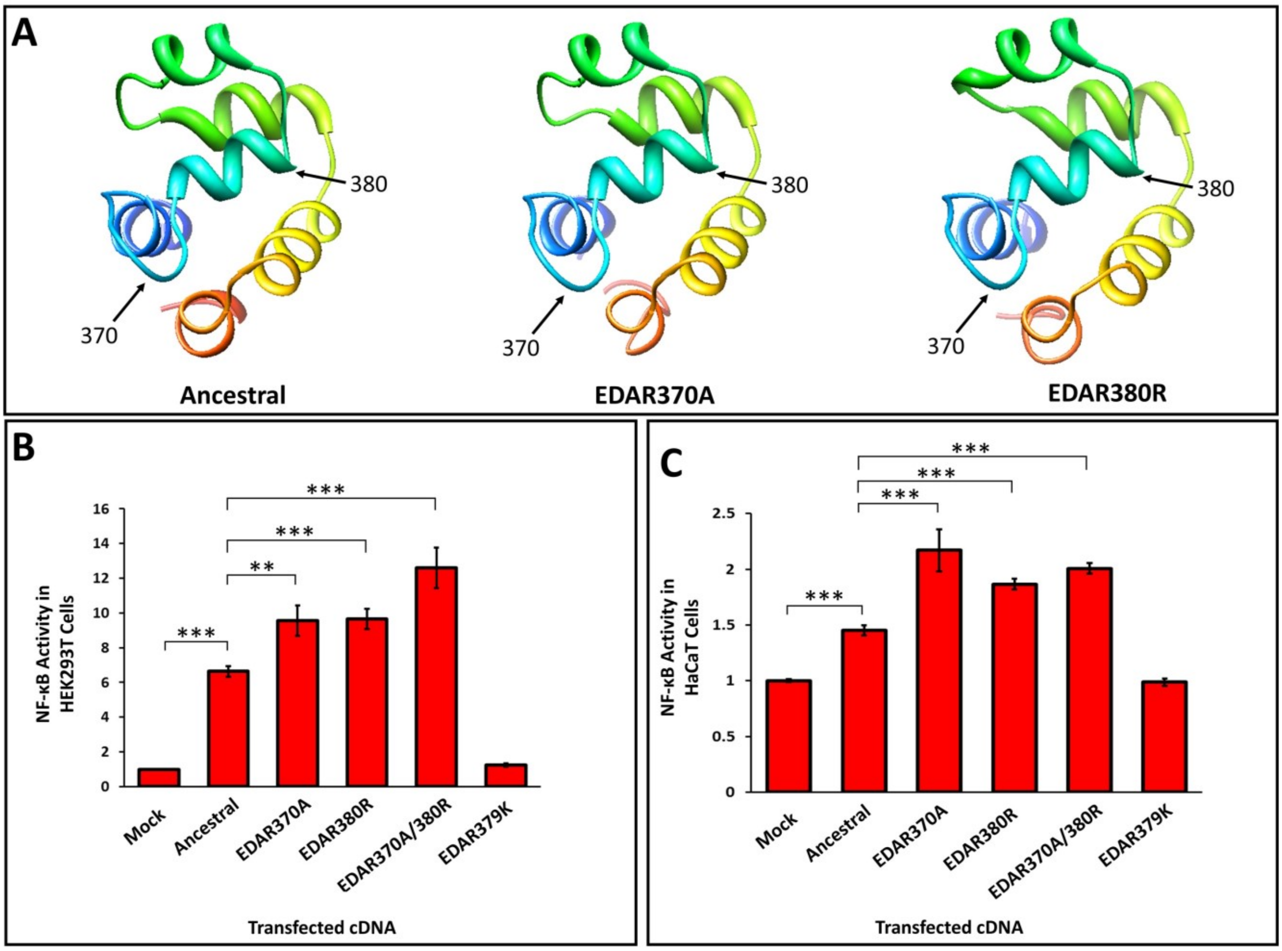

A
Double ancestral
EDAR:c.1138A>C
EDAR:c.1109T>C
EHH
1.00
0.75
0.50
0.25
0.00
109.0
109.2
109.4
109.6
109.8
Position (Mb)
B
Double ancestral
109.48
109.49
109.50
109.51
109.52
109.53
109.54
Position (Mb)
EDAR:c.1138A>C
109.48
109.49
109.50
109.51
109.52
109.53
109.54
Position (Mb)
EDAR:c.1109T>C
109.48
109.49
109.50
109.51
109.52
109.53
109.54
Position (Mb)

Figure 4


A
380
370
Ancestral
380
370
EDAR370A
380
370
EDAR380R
B
NF-κB Activity in HEK293T Cells
16
14
12
10
8
6
4
2
0
***
**
***
***
Mock
Ancestral
EDAR370A
EDAR380R
EDAR370A/380R
EDAR379K
Transfected cDNA
C
NF-κB Activity in HaCaT Cells
2.5
2
1.5
1
0.5
0
***
***
***
***
Mock
Ancestral
EDAR370A
EDAR380R
EDAR370A/380R
EDAR379K
Transfected cDNA