



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

Jon Riddell<sup>1</sup> · Chandana Basu Mallick<sup>1,4</sup> · Guy S. Jacobs<sup>2,3</sup> · Jeffrey J. Schoenebeck<sup>1</sup> · Denis J. Headon<sup>1</sup>

Received: 13 November 2019 / Revised: 7 April 2020 / Accepted: 12 May 2020 / Published online: 4 June 2020

© The Author(s), under exclusive licence to European Society of Human Genetics 2020

## Abstract

Ectodysplasin A1 receptor (*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 EDAR:p.(Val370Ala), a receptor with more potent signalling effects than the ancestral EDAR370Val. 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 EDAR:p.(Ser380Arg). 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 EDAR:p.(Ser380Arg) displays increased EDAR signalling output, at a similar level to that of EDAR:p.(Val370Ala). 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 [1]. Upon activation by its ligand ectodysplasin (*EDA*), *EDAR* signals through its cytoplasmic adapter protein *EDARADD* to trigger the activation of the transcription factor NF- $\kappa$ B [2], this signalling sequence being essential for its function [3]. Disruption of this highly conserved signalling pathway via

loss-of-function mutations of any of *EDA*, *EDAR*, or *EDARADD* causes hypohidrotic ectodermal dysplasia (HED) [2, 4, 5], a condition characterised by sparse hair, the loss or reduction of many skin-associated glands, and tooth agenesis [6]. Selective absence of teeth can occur as a result of milder function-reducing mutations in the *EDA*–*EDAR* pathway, without eliciting the complete set of clinical HED phenotypes [7].

The death domain of the *EDAR* protein is essential for the recruitment of the *EDARADD* protein and thus for *EDAR* function [2]. This domain is present in many proteins, the majority of which are involved in cell death and inflammation [8]. The death domain is ~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 [8, 9].

A non-synonymous single-nucleotide variant (SNV), rs3827760 (NM\_022336.4:c.1109T>C), hereafter referred to as *EDAR*:c.1109T>C, encodes a valine-to-alanine substitution within the death domain of *EDAR* at amino acid position 370 (NP\_071731.1:p.(Val370Ala), hereafter referred to as *EDAR*:p.(Val370Ala). The derived allele is at very high frequency in northern East Asian and Native American

✉ Denis J. Headon  
denis.headon@roslin.ed.ac.uk

<sup>1</sup> The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, UK

<sup>2</sup> Complexity Institute, Nanyang Technological University, Singapore 637723, Singapore

<sup>3</sup> Department of Archaeology, University of Cambridge, Cambridge, UK

<sup>4</sup> Present address: Centre for Genetic Disorders, Institute of Science, Banaras Hindu University, Varanasi, India

populations, with allele frequencies of up to 90% in some groups [10]. The *EDAR:c.1109T>C* allele displays clear evidence of positive selection both from haplotype and allele frequency spectrum based analyses [11–13]; at least some of this selection presumably occurred in the common ancestors of modern East Asian and Native American populations. *EDAR:p.(Val370Ala)* has been shown to increase the activation of NF- $\kappa$ B compared with that of the protein encoded by the ancestral allele (*EDAR370Val*) in vitro using reporter assays [14, 15], and ameliorate the clinical signs of HED caused by hypomorphic *EDA* mutations in heterozygous carriers of *EDAR:p.(Val370Ala)* [16], 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 [15, 17, 18]. Both of these models were observed to have thicker hair fibres and; 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 [18–23].

Here we identify another SNV in *EDAR* (rs146567337, NM\_022336.4:c.1138A>C), hereafter referred to as *EDAR:c.1138A>C*, which causes a serine-to-arginine substitution at amino acid position 380 (NP\_071731.1:p.(Ser380Arg), hereafter referred to as *EDAR:p.(Ser380Arg)*). The geographic distribution of the derived allele of this SNV partly overlaps that of the previously characterised *EDAR:c.1109T>C* (encoding *EDAR:p.(Val370Ala)*), though at lower frequency and with a more southerly prevalence. The *EDAR:p.(Ser380Arg)* substitution increases the signalling function of *EDAR* to a similar degree as the *EDAR:p.(Val370Ala)* 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 [24]. These findings suggest that *EDAR:c.1138A>C* (*EDAR:p.(Ser380Arg)*) may influence the same human traits as those associated with *EDAR:c.1109T>C* (*EDAR:p.(Val370Ala)*), and that these traits may have been under different selective pressures in different regions of Asia.

## 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 [10, 25, 26] and plotted for each population.

### Determination of archaic human genotypes

We used high coverage Altai Neanderthal (<http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal/VCF/>) [27] and Altai Denisovan ([http://cdna.eva.mpg.de/denisova/VCF/hg19\\_1000g/](http://cdna.eva.mpg.de/denisova/VCF/hg19_1000g/)) [28] genomes to determine rs146567337 state.

### Haplotype analysis

VCF files from publicly available datasets [10, 29] were reduced to a 20 kb window surrounding rs146567337. Files were viewed using inPHAP (v1.1) [30] and variants that disagreed with the human reference genome (GRCh37) were mapped for samples containing *EDAR:c.1138A>C* or *EDAR:c.1109T>C*. The variants thus mapped combined to give a total of 50 SNVs used for haplotype construction.

A median-joining haplotype network was constructed using NETWORK 5.0 (<http://www.fluxus-engineering.com>). We used the region  $\pm 10$  kb from the SNVs of interest among the individuals included in the Estonian Bio-centre Human Genome Diversity Panel dataset [25].

### Generation of extended haplotype homozygosity (EHH) and bifurcation plots

EHH plots and bifurcation plots were constructed from a dataset of 10,640 individuals from the Han Chinese population [31] using R package rehh (v2.0.2) [12, 32]. 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, and *EDAR:c.1109T>C*—where the ancestral allele of rs146567337 and the derived allele of rs3827760 are present. There were 554 double ancestral haplotypes, 433 *EDAR:c.1138A>C* haplotypes and 20,293 *EDAR:c.1109T>C* haplotypes.

Genome-wide distributions of EHH were generated from analysis of 2 Mb windows seeded using SNVs with derived allele frequencies comparable with rs146567337 and rs3827760 (i.e. transversions ranging between >0.00 and 0.04 and transitions ranging between 0.93 and 0.97, respectively). In total, 1533 transversions and 151 transitions were analysed. EHH distance was calculated from the EHH x-intercepts, with a lower limit of 0.05. The Bash and R scripts used to sample the CONVERGE data [31] are available here: <https://doi.org/10.7488/ds/2798>.

## Sequence alignment and protein structure modelling

EDAR death domain sequence alignment was generated with the T-coffee alignment tool [33] using peptide sequences gathered from the NCBI protein database (GenInfo Identifiers: human, 11641231; mouse, 6753714; chicken, 60302666; zebrafish, 924859488; *Xenopus*, 55742031). Sequence conservation was visualised using BOXSHADE v3.21. The EDAR death domain structure was generated in the intensive mode of Phyre2 v2.0 [34] using positions 345–431 of human EDAR.

## Transfection of cells and luciferase assays

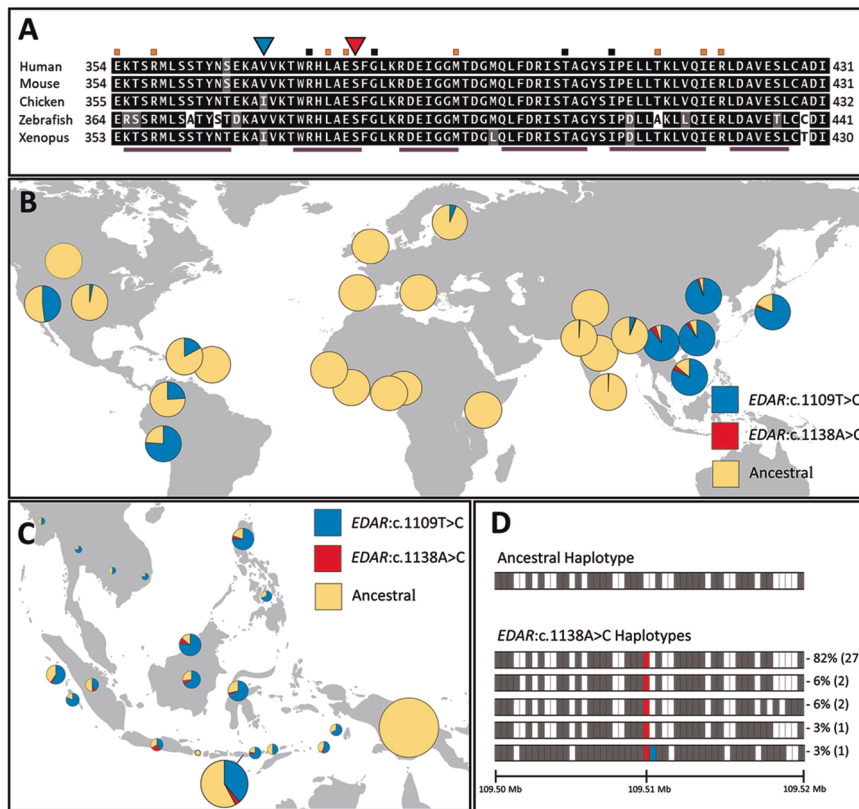
HEK293T and HaCaT cells were maintained at 37 °C in 5% CO<sub>2</sub> in high glucose Dulbecco's modified Eagle's medium (DMEM) (Sigma-Aldrich, St. Louis, MO, USA) supplemented with 10% foetal bovine serum (FBS) and 50 µg/ml streptomycin and 100 U/ml penicillin (Thermo Fisher Scientific, Waltham, MA, USA). Transfections of HEK293T and HaCaT cells were performed using Lipofectamine 3000 (Thermo Fisher Scientific) in 24-well plates (well surface area: 1.9 cm<sup>2</sup>). Cells were seeded at a density of  $5 \times 10^4$  24 h prior to transfection. Each well was transfected with plasmid DNA mix in opti-MEM (Thermo Fisher Scientific), consisting of 125 ng pNFκB-luc, 62.5 ng pRLTK, 10 ng pCR3::EDAR expression vector (different variants) and made up to a total of 500 ng with empty pCR3.1 vector. Transfections were performed according to the manufacturer's instructions in DMEM supplemented with 10% FBS, 50 µg/ml streptomycin and 100 U/ml penicillin. Luciferase assays were performed 18 h post transfection using the Dual-Luciferase Assay System (Promega, Madison, WI, USA) according to the manufacturer's instructions.

## Results

The death domain of EDAR is a highly conserved region of the protein, with variants altering this domain commonly leading to a loss of function and thus clinically diagnosed HED [35], presumably due to altered or abrogated EDAR interaction with EDARADD [36]. We identified SNV rs146567337 (*EDAR*:c.1138A>C) in *EDAR* in the gnomAD database (<https://gnomad.broadinstitute.org/>) [37]. The derived allele encodes a serine-to-arginine substitution at the highly conserved amino acid 380 (*EDAR*:p.(Ser380-Arg)), only ten amino acids from the alteration in the well-characterised *EDAR*:p.(Val370Ala) variant (Fig. 1a). In the gnomAD database [37], the frequency of the derived allele at rs146567337 was 1.85%. Using publicly available datasets [10, 25, 26, 29], 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, c). Since *EDAR*:c.1109T>C is at very high frequency in many 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 5608 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 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. This immediately ancestral haplotype has a worldwide population distribution. The variant was also found to be ancestral in both the Altai Neanderthal [27] and Altai Denisovan [28] genomes, and was not inferred to be in archaic introgressed haplotypes identified in either the Simons Genome Diversity Project [29] or Indonesian Genome Diversity Project data [26]. 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 (Fig. 2) [25]. We used the HapMap combined genetic map [38] to confirm that the window did not have especially fast recombination likely to disrupt the network reconstruction (0.0797 cM; 44th percentile of total genetic map distance in non-overlapping genome wide 114.9 kb 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 fewer East Asian individuals ( $n = 11$ ) than Siberians ( $n = 108$ ), thus explaining the greater proportion of the latter with the associated haplotype [25]. We also observed that the *EDAR*:c.1109T>C associated haplotype demonstrates a star-like pattern, suggestive of a demographic expansion



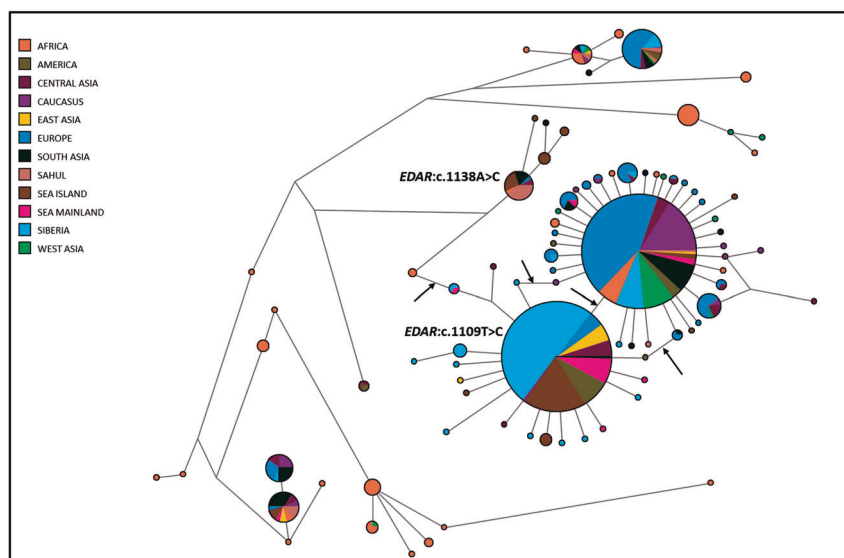
**Fig. 1 Conservation, distribution and haplotype structure of *EDAR* variants.** **a** Multiple sequence alignment of vertebrate *EDAR* death domains. Amino acid positions within the *EDAR* protein are numbered at the start and end of the sequence for each species. The position of 370 is indicated by 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 [25, 26] 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.

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 displays very strong evidence for positive selection [11–13], we tested for indications of selection on *EDAR*:c.1138A>C. Using a large-scale whole genome sequence dataset of the Han Chinese population [31], we constructed 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 [11, 12], *EDAR*:c.1109T>C shows a broad region of haplotype homozygosity compared with the double ancestral haplotype. *EDAR*:c.1138A>C exhibits much less EHH than *EDAR*:c.1109T>C, 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 (*EDAR*:p.(Ser380Arg)) haplotypes, therefore we constructed EHH bifurcation plots by random subsampling of 433 haplotypes from double ancestral allele and *EDAR*:



**Fig. 2 Median-joining haplotype network of *EDAR*.** Median-joining haplotype network spanning  $\pm 10$  kb around rs3827760 and rs146567337 showing the relationship of the haplotypes. The network is based on 446 individuals included in the Estonian Biocentre Human Genome Diversity Panel dataset [25]. 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

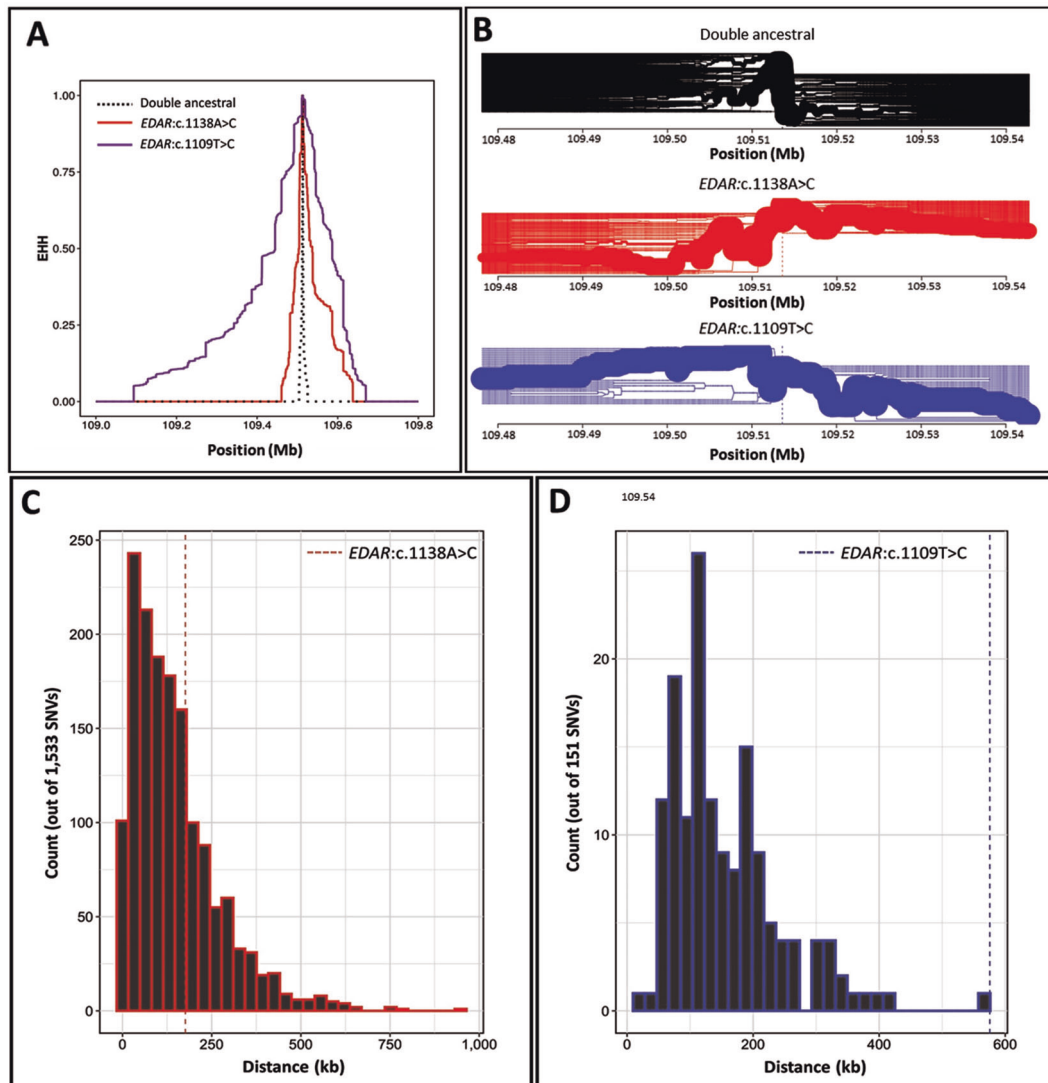
location of populations where each haplotype was found. Lines represent variants, with greater branch length indicating a greater number of distinguishing variants. The associated haplotypes of interest (for *EDAR*:c.1109T>C and *EDAR*:c.1138A>C) have been labelled. The black arrows represent locations where an event resulting in the *EDAR*:c.1109T>C variant was inferred. The multiple arrows likely reflect ambiguity in the network reconstruction.

c.1109T>C (*EDAR*:p.(Val370Ala)) 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).

We also compared the EHH scores of *EDAR*:c.1138A>C and *EDAR*:c.1109T>C to alleles elsewhere in the genome defined by derived alleles of the same frequency and mutation type. In total, 1533 transversions (allele frequency  $\geq 0.00$ – $0.04$ ) were selected to compare against *EDAR*:c.1138A>C. Similarly, 151 transitions (allele frequency =  $0.93$ – $0.97$ ) were selected to compare with *EDAR*:c.1109T>C. From this analysis, *EDAR*:c.1138A>C ranked slightly to the right of the middle of the EHH distribution (70th percentile) (Fig. 3c). In contrast, EHH associated with the *EDAR*:c.1109T>C allele ranked highest of all SNVs assessed (Fig. 3d). These findings further support the idea that *EDAR*:c.1138A>C has not been under detectable selection, while *EDAR*:c.1109T>C has undergone strong positive selection.

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 *EDAR*:p.(Ser380Arg) and identify any predicted structural effects of this amino acid substitution, we modelled the variant *EDAR* death domain structures [34]. The resulting predicted protein structure positioned amino acid *EDAR*380 within an alpha helix (Fig. 4a), a

structural feature known to be important for the protein–protein interactions mediated by death domains [9]. 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 *EDAR*:p.(Val370Ala) substitution (Fig. 4a). Based on the conservation of the serine residue at position *EDAR*380 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) [39] and PolyPhen (score 0.999) [40], we predicted that the *EDAR*:p.(Ser380Arg) 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*, *EDAR*:p.(Val370Ala), *EDAR*:p.(Ser380Arg) or the double substituted *EDAR*:p.[(Val370Ala;Ser380Arg)] protein. We also included the known loss-of-function variant *EDAR*:p.(Glu379Lys), a mutation that is dominant for selective tooth agenesis in humans and recessive for HED in mice [1, 7], as a control in these experiments. Each form was assayed for its ability to activate a co-transfected NF- $\kappa$ B luciferase reporter. We found that *EDAR*:p.(Ser380Arg) activated NF- $\kappa$ B in these cells to a greater degree than ancestral *EDAR*, and to the same extent as *EDAR*:p.(Val370Ala). Generation of an *EDAR* variant carrying both the Val370Ala and Ser380Arg amino acid substitutions led



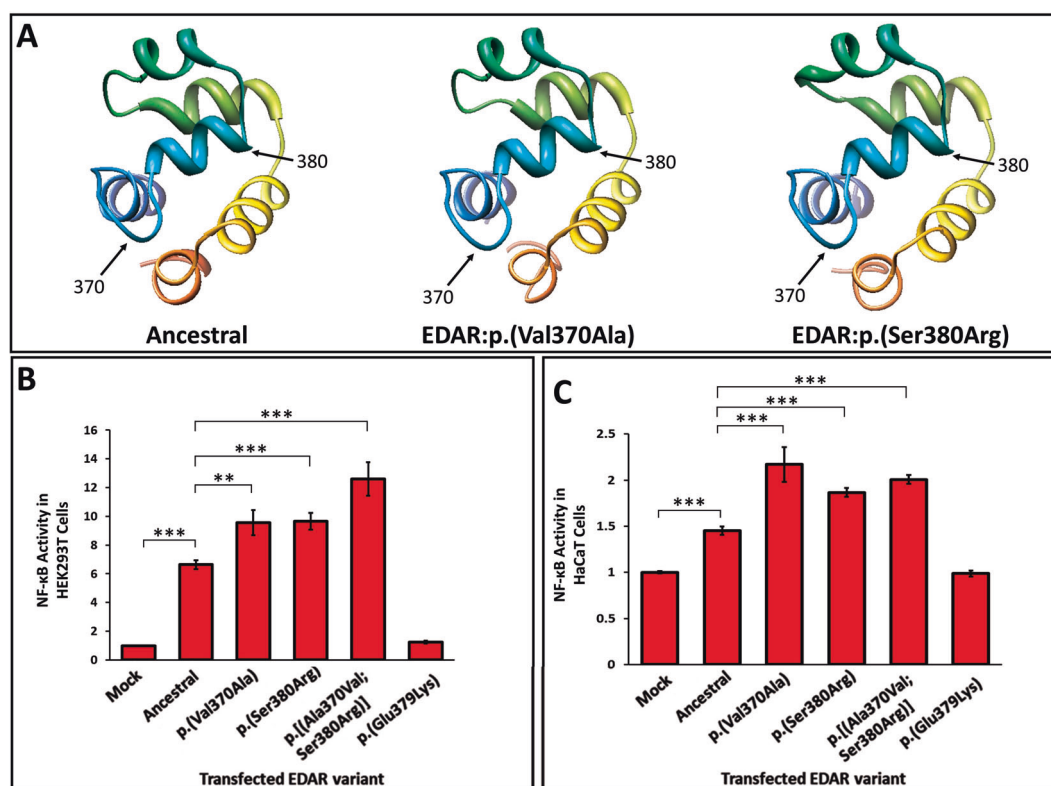
**Fig. 3** Extended haplotype homozygosity (EHH) and EHH bifurcation plots surrounding *EDAR* variants. **a** EHH plot showing the length of conserved haplotype on either side of rs146567337. An EHH value of 1 indicates that haplotypes are identical at this position. Double ancestral haplotypes are represented by the black dotted line, *EDAR*:c.1138A>C (*EDAR*:p.(Ser380Arg)) haplotypes are represented by the red line, and *EDAR*:c.1109T>C (*EDAR*:p.(Val370Ala)) 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. **c** Distribution of autosomal EHH x-intercept distances of 1533 alleles with a derived allele frequency ranging from >0.00 to 0.04 across autosomes. The red dashed line indicates the x-intercept value of *EDAR*:c.1138A>C and is located in the 70th percentile of EHH values. **d** Distribution of autosomal EHH x-intercept distances of 151 alleles with a derived allele frequency ranging from 0.93 to 0.97. The blue dashed line indicates the x-intercept value of *EDAR*:c.1109T>C and is located as the 100th percentile of EHH values.

to a level of NF- $\kappa$ B activity comparable with or slightly greater than that caused by either single substitution (Fig. 4b). These effects on signalling activity were broadly confirmed in the human HaCaT cell line, derived from the skin's epidermis (Fig. 4c). The different fold changes observed in *EDAR*-stimulated NF- $\kappa$ B activity between HEK293T cells and HaCaT cells may result from different basal levels of NF- $\kappa$ B activity in these cell lines, and from different levels of expression of components of the *EDAR* signal transduction pathway.

## Discussion

We characterised a novel functional variant in *EDAR*. We find that *EDAR*:c.1138A>C is at highest 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, but that the encoded protein increases NF- $\kappa$ B activation



**Fig. 4 Functional effects of *EDAR* variants.** **a** The predicted protein structures of the ancestral *EDAR*, *EDAR*:p.(Val370Ala) and *EDAR*:p.(Ser380Arg) death domains were 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. **b** HEK293T and **c** 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 six times. Statistical significance was calculated using a Student's unpaired *t* test (\*\* $P < 0.005$ , \*\*\*  $P < 0.0005$ ).

in vitro, to approximately the same extent as the *EDAR*:c.1109T>C substitution. As there are no known splice variants of *EDAR*, which do not include the death domain encoding exon, and as the *EDAR*:c.1138A>C variant is present within the death domain, which is essential for protein function, the change in receptor activity detected in this assay suggests that *EDAR*:c.1138A>C is likely to have a similar direction of phenotypic effects in vivo to those observed with the *EDAR*:c.1109T>C allele.

Several theories as to what the selective advantage conferred by *EDAR*:p.(Val370Ala) has been advanced. Chang et al. suggested that *EDAR*:p.(Val370Ala) was selected for 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 [17]. 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 [41] was caused by the mammary gland alterations enacted by enhanced *EDAR* signalling [17, 18].

Kamberov et al. placed the origin of the *EDAR*:p.(Val370Ala) encoding allele in central China at greater than 30,000 years ago, and suggested increased eccrine sweat gland number, associated with this variant in mouse and human in their study, as one of the potential selective forces that would have been advantageous in the hot and humid climate there [18].

A recent genealogical estimation of allele ages in the human genome assessed the derived alleles *EDAR*:c.1109T>C and *EDAR*:c.1138A>C as having a similar date of origin, at ~1400 generations ago [24]. The geographic distribution of these alleles is somewhat similar, and, though at much lower frequency than *EDAR*:c.1109T>C in all regions, 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 [14]. We found that *EDAR*:c.1138A>C does not show a strong signal of positive selection in human populations, as does *EDAR*:c.1109T>C, despite cell culture experiments predicting similar outcomes resulting from these two 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 [42, 43], suggesting that phenotypes associated with an *EDAR*-dependent increase in NF- $\kappa$ B activation have been preferentially selected for in more northern regions of Asia.

The possibility that *EDAR:p.(Val370Ala)* and *EDAR:p.(Ser380Arg)* 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 haplotypes 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 as one of many candidate alleles in people exhibiting tooth agenesis in East Asian populations [44, 45]. 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*. Tooth agenesis has also been reported in some patients carrying the *EDAR:c.1109T>C* allele [46], demonstrating that increased *EDAR* activity is not necessarily sufficient to mitigate the effects of mutations in other genes or pathways that cause this condition.

The discovery of a second SNV in *EDAR* that increases NF- $\kappa$ B 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 variants have also been identified in an enhancer region upstream of the *LCT* gene that have the same molecular effect of increasing *LCT* transcription [47]. However, some of 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 both therefore should 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.

**Acknowledgements** We thank Toomas Kivisild for his insightful comments on the manuscript.

**Funding** JR has been funded by University of Edinburgh studentship. CBM has been supported by Marie Skłodowska-Curie Actions Individual Fellowship (Grant Agreement: 706429) and Ramalingaswami Re-entry Fellowship, Department of Biotechnology, Government of India.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## References

- Headon DJ, Overbeek PA. Involvement of a novel Tnf receptor homologue in hair follicle induction. *Nat Genet.* 1999;22:370–4.
- Headon DJ, Emmal SA, Ferguson BM, Tucker AS, Justice MJ, Sharpe PT, et al. Gene defect in ectodermal dysplasia implicates a death domain adapter in development. *Nature.* 2001;414:913–6.
- Schmidt-Ullrich R, Aebischer T, Hülsken J, Birchmeier W, Klemm U, Scheidereit C. Requirement of NF- $\kappa$ B/Rel for the development of hair follicles and other epidermal appendices. *Development.* 2001;128:3843–53.
- Kere J, Srivastava AK, Montonen O, Zonana J, Thomas N, Ferguson B, et al. X-linked anhidrotic (hypohidrotic) ectodermal dysplasia is caused by mutation in a novel transmembrane protein. *Nat Genet.* 1996;13:409–16.
- Monreal AW, Ferguson BM, Headon DJ, Street SL, Overbeek PA, Zonana J. Mutations in the human homologue of mouse dl cause autosomal recessive and dominant hypohidrotic ectodermal dysplasia. *Nat Genet.* 1999;22:366–9.
- Reyes-Realí J, Mendoza-Ramos MI, Garrido-Guerrero E, Méndez-Catalá CF, Méndez-Cruz AR, Pozo-Molina G. Hypohidrotic ectodermal dysplasia: clinical and molecular review. *Int J Dermatol.* 2018;57:965–72.
- Arte S, Parmanen S, Pirinen S, Alaluusua S, Nieminen P. 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.* 2013;8:1–12.
- Park HH, Lo Y-C, Lin S-C, Wang L, Yang JK, Wu H. The death domain superfamily in intracellular signaling of apoptosis and inflammation. *Annu Rev Immunol.* 2007;25:561–86.
- Ferrao R, Wu H. Helical assembly in the death domain (DD) superfamily. *Curr Opin Struct Biol.* 2012;22:241–7.
- The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature.* 2015;526:68–74.
- Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. *Nature.* 2007;449:913–8.
- Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature.* 2002;419:832–7.
- Cheng X, Xu C, DeGiorgio M. Fast and robust detection of ancestral selective sweeps. *Mol Ecol.* 2017;26:6871–91.
- Bryk J, Hardouin E, Pugach I, Hughes D, Strotmann R, Stoneking M, et al. Positive selection in East Asians for an *EDAR* allele that enhances NF- $\kappa$ B activation. *PLoS ONE.* 2008;3:e2209.
- Mou C, Thomason HA, Willan PM, Clowes C, Edwin Harris W, Drew CF, et al. Enhanced ectodysplasin-A receptor (*EDAR*)



- signaling alters multiple fiber characteristics to produce the east Asian hair form. *Hum Mutat.* 2008;29:1405–11.
16. Cluzeau C, Hadj-Rabia S, Bal E, Clauss F, Munnich A, Bodemer C, et al. The EDAR370A allele attenuates the severity of hypohidrotic ectodermal dysplasia caused by EDA gene mutation. *Br J Dermatol.* 2012;166:678–81.
  17. Chang SH, Jobling S, Brennan K, Headon DJ. Enhanced Edar signalling has pleiotropic effects on craniofacial and cutaneous glands. *PLoS ONE.* 2009;4:e7591.
  18. Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, et al. Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell.* 2013;152:691–702.
  19. Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, Batubara L, et al. A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum Mol Genet.* 2008;17:835–43.
  20. Kimura R, Yamaguchi T, Takeda M, Kondo O, Toma T, Haneji K, et al. A common variation in EDAR is a genetic determinant of shovel-shaped incisors. *Am J Hum Genet.* 2009;85:528–35.
  21. Adhikari K, Reales G, Smith AJP, Konka E, Palmén J, Quinto-Sánchez M, et al. A genome-wide association study identifies multiple loci for variation in human ear morphology. *Nat Commun.* 2015;6:7500.
  22. Adhikari K, Fuentes-Guajardo M, Quinto-Sánchez M, Mendoza-Revilla J, Camilo Chacón-Duque J, Acuña-Alonzo V, et al. A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation. *Nat Commun.* 2016;7:11616.
  23. Adhikari K, Fontanil T, Cal S, Mendoza-Revilla J, Fuentes-Guajardo M, Chacón-Duque J-C, et al. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. *Nat Commun.* 2016;7:10815.
  24. Albers PK, McVean G. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS Biol.* 2020;18:e3000586.
  25. Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, et al. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature.* 2016;538:238–42.
  26. Jacobs GS, Hudjashov G, Saag L, Kusuma P, Darusallam CC, Lawson DJ, et al. Multiple deeply divergent denisovan ancestries in Papuans. *Cell.* 2019;177:1010–21.e32.
  27. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature.* 2014;505:43–9.
  28. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, et al. A high-coverage genome sequence from an archaic Denisovan individual. *Science.* 2012;338:222–6.
  29. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature.* 2016;538:201–6.
  30. Jäger G, Peltzer A, Nieselt K. InPHAP: interactive visualization of genotype and phased haplotype data. *BMC Bioinform.* 2014;15:200.
  31. Cai N, Bigdeli TB, Kretschmar WW, Li Y, Liang J, Hu J, et al. 11,670 whole-genome sequences representative of the Han Chinese population from the CONVERGE project. *Sci Data.* 2017;4:170011.
  32. Gautier M, Klassmann A, Vitalis R. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Mol Ecol Resour.* 2017;17:78–90.
  33. Di Tommaso P, Moretti S, Xenarios I, Orobitg M, Montanyola A, Chang JM, et al. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.* 2011;39:W13–17.
  34. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc.* 2015;10:845–58.
  35. Cluzeau C, Hadj-Rabia S, Jambou M, Mansour S, Guigue P, Masmoudi S, et al. Only four genes (EDA1, EDAR, EDARADD, and WNT10A) account for 90% of hypohidrotic/anhidrotic ectodermal dysplasia cases. *Hum Mutat.* 2011;32:70–7.
  36. Okita T, Asano N, Yasuno S, Shimomura Y. Functional studies for a dominant mutation in the EDAR gene responsible for hypohidrotic ectodermal dysplasia. *J Dermatol.* 2019;46:710–5.
  37. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. 2019. bioRxiv <https://doi.org/10.1101/531210>.
  38. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature.* 2007;449:851–61.
  39. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 2003;31:3812–4.
  40. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nat Methods.* 2010;7:248–9.
  41. Hlusko LJ, Carlson JP, Chaplin G, Elias SA, Hoffecker JF, Huffman M, et al. Environmental selection during the last ice age on the mother-to-infant transmission of vitamin D and fatty acids through breast milk. *Proc Natl Acad Sci USA.* 2018;115:E4426–32.
  42. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature.* 2015;528:499–503.
  43. Siska V, Jones ER, Jeon S, Bhak Y, Kim HMH, Cho YS, et al. Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago. *Sci Adv.* 2017;3:e1601877.
  44. He H, Han D, Feng H, Qu H, Song S, Bai B, et al. Involvement of and interaction between WNT10A and EDA mutations in tooth agenesis cases in the Chinese population. *PLoS ONE.* 2013;8:e80393.
  45. Yamaguchi T, Hosomichi K, Yano K, Kim Y-I, Nakaoka H, Kimura R, et al. Comprehensive genetic exploration of selective tooth agenesis of mandibular incisors by exome sequencing. *Hum Genome Var.* 2017;4:17005.
  46. Lee KE, Ko J, Shin TJ, Hyun HK, Lee SH, Kim JW. Oligodontia and curly hair occur with ectodysplasin-a mutations. *J Dent Res.* 2014;93:371–5.
  47. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, et al. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet.* 2007;39:31–40.