Polymorphisms Influencing the Shovel-Shaped Incisors Phenotype in Puerto Rico

by

Joseline Serrano González

Thesis submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

in

BIOLOGY

UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS
2017

Approved by:

Juan C. Martínez Cruzado, PhD
President, Graduate Committee

Taras K. Oleksyk, PhD
Member, Graduate Committee

Carlos A. Acevedo Suárez, PhD
Member, Graduate Committee

Matías Cafaro, PhD
Chairperson of the Department

Rocío Zapata, MA
Graduate Studies Representative

Abstract

Puerto Ricans phenotypic traits are a combination of the admixture of their ancestors Taínos with Europeans and Africans. Shovel-shaped incisors are a common trait among Taínos' skeletons and are still present in the Puerto Rican population. It is a phenotypic trait derived from its ancestors in Asia, and has been almost exclusive to that population for thousands of years. It has been reported that a SNP in *Ectodysplasin A-receptor (EDAR)*, rs3827760, is a genetic determinant for the phenotype but it might not be the only one. In the present study we aimed to find other SNPs that are also determinant for shovel-shaped incisors, and to examine if the *EDAR* region has undergone positive selection in the Puerto Rican population. We collected maxillary plaster casts from Puerto Ricans to determine shoveling grade and saliva samples to genotype candidate SNPs. We also genotyped the SNP rs3827760 in 452 representative samples of the Puerto Rican population to determine its allelic frequency. We found that the SNP rs3827760 explains 39% of the variance and that for each copy of the allele 1540C the shoveling grade increases 0.9 in the shoveling grade scale. The allelic frequency for this SNP was 13%, and this allowed us to estimate the Native American frequency in the *EDAR* locus at 16%. By comparing this estimated frequency to the 15.2% frequency of the Native American throughout the genome as a whole for the same sample set, we conclude that the *EDAR* region has not been positively selected in Puerto Ricans.

Resumen

Los rasgos fenotípicos de los puertorriqueños son una mezcla de sus ancestros Taínos con europeos y africanos. Los incisivos diente de pala son un rasgo común entre las osamentas de Taínos, y están aún presentes en la población puertorriqueña. Es un rasgo fenotípico derivado de sus ancestros en Asia y ha sido casi exclusivo de esa población por miles de años. Ha sido reportado que un polimorfismo de un nucleótido en el gen *Ectodysplasin A-receptor (EDAR),* rs3827760, es un determinante genético para este fenotipo pero podría no ser el único. Uno de los objetivos en el presente estudio fue encontrar otros polimorfismos que también sean determinantes para los incisivos diente de pala, y examinar si la región para *EDAR* ha pasado por una selección positiva en la población puertorriqueña. Colectamos moldes de yeso maxilares de puertorriqueños para determinar el grado de diente de pala y muestras de saliva para genotipar polimorfismos candidatos. También genotipamos el polimorfismo rs3827760 en 452 muestras representativas de la población puertorriqueña para determinar su frecuencia alélica. Encontramos que el polimorfismo rs3827760 explica el 39% de la varianza y por cada copia del alelo 1540C el grado de diente de pala aumenta en un 0.9 en la escala de gradación de pala. La frecuencia alélica para este polimorfismo fue 13%, lo que permite estimar la frecuencia indígena en el locus de EDAR en 16%. Al comparar este porciento con la frecuencia indígena global para la misma muestra, que fue 15.2%, concluimos que la región de *EDAR* no ha sido seleccionada positivamente entre los puertorriqueños.

To graduate students around the world, give your best and persevere.

Acknowledgements

I would like to thank my advisor, Dr. Martínez-Cruzado, for his mentorship throughout this project, for his dedication, and support. Thank you for being a remarkable mentor and for addressing us students with patience, kind, and humor. A positive attitude is a game changer during grad school. Special thanks to my committee members: Dr. Acevedo-Suárez, and Dr. Oleksyk for their support and mentorship, for their time spent advising me. Thank you for every given advice because each of them made me a better student.

I am deeply grateful to the *Sociedad de Especialistas en Ortodoncia de Puerto Rico (SEO)* for helping us collecting the samples for this project, specially to: Dr. Lynette García, Dr. Héctor L. Joy-Sobrino, Dr. Edna C. Galarza Escobar, Dr. José E. Fossas-Marxuach, Dr. Luis Toro-Lloveras, Dr. Carlos M. Muñoz, Dr. Bruni M. Ortiz-Giuliani. Additional to the members of SEO, I am very grateful to the dentists who also helped us with the collection of samples: Dr. Roberto Colón-Blanco, Dr. José Mercado-Ghigliotty, and Dr. Wanda Cruzado. Thank you to all of you and to the assistants and secretaries whom were kind enough to me and help me with this project. I would also thanks the forensic anthropologist Dr. Edwin Crespo for his collaboration in this study. Thank you for the dedication and time spent in this project. Your expertise in the matter was key to the completion of the study, thank you. Many thanks to Mr. Jorge Fonseca from Thermo Fisher Scientific who went beyond expectations to help us finish this project. Thank you Mr. Fonseca for making arrangements and taking time to ensure us a possibility to complete the project, it could not have been made possible without your help.

I would also like to thanks to all the people that in even a small way help me during graduate student life. Starting with Mrs. Gladys Toro-Labrador, my teaching assistant supervisor, thank you for really making me give my best and beyond. The experience being a

Genetics TA have had a great impact in my professional life. I can say, I learned from the best. Thanks to the people who work with me during some short period of time but from who I learned the most important things in a lab, the basic skills, Ms. Mónica Fernández, Dr. Audrey Majeskey, and Ms. Yashira Afanador. My most sincere acknowledgment to you guys, your help and mentorship help me get where I am today, I will not forget. Special thanks to the best secretaries in the University of Puerto Rico at Mayagüez, Mrs. Mary L. Jiménez, and Mrs. Vilmarie Rivera from the Graduate Studies Office at Biology Department. Thank you for your dedication with us graduate students, thank you for each and every time you reminded us important dates, seminars, paperwork. Thank you Mary, and thank you Vilmarie, because you truly helped me get through grad school. Special thanks to Wilfried Guiblet and Walter Wolfsberger for their help in bioinformatics.

Finally, I would like to thanks my family and my husband that during this time have given me their support, their comfort, and have encouraged me to persevere. Thanks for believing in me.

Table of contents

List of Tables

List of Figures

Introduction

Single Nucleotide Polymorphisms (SNPs) are variants of one nucleotide in the DNA. Such variants can have serious effects at the molecular level by altering protein function and therefore affecting signaling pathways. These effects sometimes result in observable traits in an organism. For example, in the human c-Src gene, *Src,* a SNP changes the residue in the polypeptide chain in the amino acid position 260 from tryptophan to alanine, a mutation known as W260A (Meng & Roux, 2016). This change is located in the N-terminal of the kinase domain, and even though the substitution changes one hydrophobic residue for another it alters the c-Src conformation and subsequently causes its dysregulated activation (Meng & Roux, 2016). At a molecular level, c-Src is part of a signaling pathway that activates transcription factors responsible of cell growth. Over-expression of c-Src has been linked with alterations in growth, development, progression, and metastasis of cancer (Irby & Yeatman, 2000).

For this study we analyzed several SNPs in the ectodysplasin (EDA) pathway in the Puerto Rican population. Among the proteins involved in this pathway are the Tumor Necrosis Factor (TNF) Ectodysplasin-A, Tumor Necrosis Factor Receptor (TNFR) EDAR, and its adaptor Ectodysplasin-A Receptor Associated Death Domain (EDARADD). Mutations found in the genes coding for these proteins may lead to defective protein interactions affecting the nuclear factor-κB (NF-κB) signaling pathway and therefore having catastrophic results in cell development. The NF-κB pathway regulates cell growth, development, apoptosis, and several immunological responses. Studies in mice have found that the EDA pathway stimulates tooth development by transcribing *Fgf20* and therefore activating its signaling (Häärä et al., 2012). Transcription of EDA also regulates other signaling pathways like Wnt signaling in hair follicles and salivary glands branching (Häärä et al., 2011).

The SNP known as rs3827760, or T1540C for its nucleotide position, located in the *Ectodysplasin-A* receptor gene *(EDAR)*, has been linked to hair thickness in East-Asians (Fujimoto et al., 2008). It is also responsible for one-fourth of the heritability of shovel-shaped incisors (Kimura et al., 2009), given that previous studies have found the heritability of shovel-shaped tooth to be 0.75 in Asians and Native Americans (Scott & Turner, 1997; Hanihara et al., 1974; Blanco & Chakraborty, 1976; cited by Kimura et al., 2009). This mutation is located in the death domain of the receptor and may affect protein interaction thereby reducing NF-κB signaling leading to a decrease in cell apoptosis, which explains thicker hair and shovel-shaped teeth which are indications of over cell growth.

Shovel-shaped incisors are described by Lee and Goose in 1972 as 'a combination of a concave lingual surface and elevated marginal ridges enclosing a central fossa in incisor teeth'. Shovel-shaped incisors are a common trait in Asian and Native American populations, but they go back to Neanderthals, thus providing a hypothesis for their origin. The "Anterior Dental Loading Hypothesis" suggests that anterior upper and lower teeth characteristics in Neanderthals were an adaptive response to the force applied to the teeth for domestic uses (Clement, Hillson, & Aiello, 2012). The region around *EDAR* was positively selected in East Asians when populations diverged into Europe and Asia thousands of years ago (Carlson et al., 2005) (Fujimoto et al., 2008). It is suspected that *EDAR* allele T1540C was positively selected due to its phenotypic advantages, either hair morphology against cold or stronger teeth for domestic uses (Kimura et al., 2009). Hair and teeth are derived from the ectoderm and also nails, sweat, mammary, and sebaceous glands. Therefore, variants affecting one of these could indirectly affect the others. Shovel-shaped incisors phenotype is unlikely the cause for the allele's positive selection but could have hitchhiked with other phenotypes that protect against the cold such as hair thickness or enlarged sebaceous glands (Mustonen et al., 2003). This interpretation could explain why shoveling grade increases from south to north in Asian populations (Mizoguchi, 1985, cited by Kimura et al., 2009). The reason for the *EDAR* allele T1540C to be positively selected in an ancestral Asian population after the divergence from ancestral European populations remains unknown.

Bones of Taínos in Puerto Rico have shown shovel-shaped incisors, which is why we find them in modern Puerto Ricans. Dental morphology is believed to be mostly influenced by genetic factors. As explained above, variants of *EDAR*, its ligand *Ectodysplasin-A* (*EDA),* or its adapter *Ectodysplasin-A receptor associated death domain (EDARADD)*, affect the Ectodysplasin pathway which is involved in the development of teeth and other ectodermal organs (hair, sweat glands) (S. Deshmukh, 2012). In this study we examined if the shovel-shaped incisors phenotype was selected in the Puerto Rican population. Selection tests often need the selective sweep to have taken place thousands of years ago. Population differentiation, a measure known as F_{ST}, detects positive and balancing selection that took place less than 80,000 years, while extended linkage disequilibrium (LD) tests detects positive selection from less than 30,000, which is a considerably difference (Oleksyk, Smith, & O'Brien, 2010). A novel method called mapping by admixture linkage disequilibrium (MALD) can detect positive selection in admixed population from less than 500 years (Oleksyk et al., 2010). In the Puerto Rican population admixture started only about 500 years ago. Therefore, MALD will be the most adequate method for detecting positive selection in our population. In admixed populations positive selection would increase the percentage of admixture contribution by the ancestral population that originally carried the selected variant, in comparison with the admixture contribution along the genome. At the same time, we would expect a decrease of admixture contribution of the other populations in the selected genomic region (Oleksyk et al., 2010). To examine if there is positive selection for this allele in the Puerto Rican population, we compared the frequency of “taíno” ancestry in the *EDAR* region to the frequency of “taíno” ancestry throughout the genome. In addition, we aimed to search for other SNPs in the Ectodysplasin pathway also responsible for the heritability of shovel-shaped incisors among Puerto Ricans.

Literature Review

The most upstream components of the Ectodysplasin pathway are EDA, EDAR, and EDARADD. The Tumor Necrosis Factor family of signaling molecules was discovered while cloning mutated genes in Hypohydrotic Ectodermal Dysplasia (HED) syndromes (Mikkola 2002, cited by Mustonen et al., 2003). Ectodermal Dysplasias refer to different syndromes that present defects in ectodermal appendages such as hair, nails, teeth, and sweat glands (Deshmukh & Prashanth, 2012). EDAR, a member of the Tumor Necrosis Factor receptor family, is a transmembrane protein that recognizes one of the isoforms of Ectodysplasin, EDA-A1. Upon ligand binding, these trimeric receptor proteins recruit EDARADD which interacts with EDAR via its death domain. Together with Traf6, Tab2, and Tak1, they activate the IKK complex, leading to the ubiquitination and degradation of IκB, and thus releasing the transcription factor NF-κB into the nucleus. NF-κB regulates many immunological and inflammatory responses, cell growth, developmental processes, and apoptosis (Sadier, Viriot, Pantalacci, & Laudet, 2014).

The EDA pathway is active during embryonic development of ectodermal organs. Although there are other recognized signaling pathways involved in the development of ectodermal organs such as Wnt and Fgf BMP, they are not exclusive of the ectoderm (Sadier et al., 2014). Therefore, mutations affecting their signaling can have severe consequences in the development of non-ectodermal organs (Sadier et al., 2014). In contrast, EDA pathway disruptions have been proven to exclusively affect the development of ectodermal organs (Sadier et al., 2014).

The Ectodysplasin pathway is conserved in most vertebrates providing species adaptations (Sadier et al., 2014). In the stickleback fish (*Gasterosteus aculeatus)* a mutant variant in *EDA* provided adaptation in the defensive armor of 30-36 plates. High plate phenotype is favored in marine environments but low armor was developed in freshwaters presumably because it allows better swimming, and rapid growth which favors survival on winter or because bone mineralization is costly due to low ion concentration. Reverse evolution was seen in an urban freshwater lake where the stickleback with high plate armor was increased in frequency probably because of the higher amount of predators (Sadier et al., 2014). The EDA pathway has been shown to not only serve for the development of ectodermal organs, but also to fine-tune their growth (Sadier et al., 2014). This provides an advantage in the evolution of ectodermal organs because given the specificity of the EDA pathway no pleiotropic lethal effects can be expected (Sadier et al., 2014). In humans, an adaptation involving EDAR has been found in Asians, where a mutation changing the amino acid valine for alanine was found to be positively selected in the East Asian population around 30,000 years ago (Sadier et al., 2014). The mutation produces increased hair thickness and shovel-shaped incisors. A knock-in mouse for this mutation showed increased hair thickness, increased mammary gland branch density, and increased eccrine sweat gland number (Kamberov et al., 2013). This result suggests that the mutation may have pleiotropic effects in humans (Sadier et al., 2014).

The following is a review of the genes in the EDAR pathway that highlights the reasons why we chose some of these genes in our search for mutations affecting shovel-shaped incisor phenotypes.

Ectodysplasin A-receptor

The ectodysplasin A-receptor, is encoded by the *EDAR* gene located on chromosome 2q13 from bases 109,510,927-109,605,828 (National Center for Biotechnology Information [NCBI], Gene, 2012, GRCh37 assembly). The *EDAR* gene is transcribed in the reverse direction.

It is conserved in mammals, birds, and fish (NCBI, HomoloGene, n.d.). Other names for this gene are DL; ED3; ED5; ED1R; EDA3; HRM1; EDA1R; EDA-A1R (NCBI, Gene, 2012). Surrounding genes in chromosome 2q are two protein coding genes, CCDC138 and SH3FR3, two miRNAs, SH3FR3-AS1and MIR4265, and one pseudogene, RPL39P16. The EDAR gene has three isoforms, one producing a 448 residue protein and two each producing a 480 residue protein. The most common type of EDAR isoform, EDAR-001, codes for the 448 amino acid protein and is 91% identical to the downless ('dl') mouse mutant (Mou et al., 2008). This transcript is encoded by a 12 exon gene, of which only 10 are translated into protein. The other isoforms, 002 and 201, although they share the same residues, are structured quite differently. EDAR-002 contains a larger 5'UT region than EDAR-201. Furthermore, exon 8 in 002 and 201 has 32 extra residues not found in the primary isoform, explaining the difference in protein product length compared to EDAR-001.

According to the 1000 Genomes Project phase 1 browser (Abecasis et al., 2012), the protein product of EDAR-002 and EDAR- 201 has a charge of -15.0 with an isoelectric point of 4.7369, and a molecular weight of 51,689.51 g/mol. EDAR-001 codes for a protein product with a charge of -12.5, isoelectric point of 4.8258, and a molecular weight of 48,582.02 g/mol. All *EDAR* products have two conserved domains: a death domain and a TNFR (Tumor Necrosis Factor Receptor) superfamily domain, and also have two low complexity regions. Death domains (DD) are protein-protein interaction domains that homodimerize by self-association or heterodimerize by associating with other members of the DD superfamily. They serve as adaptors in signaling pathways and can recruit other proteins into signaling complexes. DDs play a major role in apoptosis pathways and are found in a number of other signaling pathways including inflammation, and differentiation (NCBI, Conserved Domains). Receptors containing

TNFR superfamily domains when binding TNF-like cytokines, trigger signal transduction pathways. These include inflammatory responses, apoptosis, and organogenesis.

The secondary structure for the EDAR-001 protein product was predicted using the Predict Protein website tool. It is composed of 23.0% alpha helixes, 3.3% beta strands, and 73.7% loops. One transmembrane helix was detected for the best model. The protein contains a single transmembrane domain with a type 1 membrane topology. Tucker A.S. (2000) described EDAR as having the characteristic extracellular cysteine rich fold, and a single transmembrane region.

Protein secondary structure prediction also revealed the presence of Non- Regular Secondary Structure (NORS) regions in EDAR-001. NORS are regions of segments of >70 consecutive residues with <12% of the residues in helix, strand or coiled-coil regions and with at least one segment of 10 adjacent residues exposed to solvent (Liu J., 2003). NORS are also known as floppy, natively disordered, natively unfolded or loopy. These are regions in proteins that adopt regular structure when binding to substrates or other proteins. NORS regions are particularly abundant in eukaryotic proteomes, conserved during evolution, over-represented in regulatory function category and important in protein–protein interactions. The presence of NORS regions in EDAR-001 could help explain why its 3D structure has not been defined yet.

EDAR epigenetics was evaluated using the UCSC Genome Browser. We looked for CpG islands, histone marks, and DNA methylation. CpG islands (CGIs) are regions of repeated non-methylated cytosines preceding guanines. These islands are generally located in promoter regions since high GC content attracts transcription factors. Methylation on CGIs is associated with gene silencing because methyl groups activate heterochromatization (Deaton and Bird, 2011). The UCSC Genome Browser revealed that in the CGI located in the 5’UTR of the EDAR gene, 19% of the 221 bp sequence constitutes CGs. The UCSC Genome Browser also revealed the presence of trimethylated H3K9 and H3K27, meaning that the EDAR gene was silenced. These results are to be expected because two of the three cell lines used to collect these data came from the mesoderm lineage and EDAR is active only in the ectoderm lineage.

The main function of EDAR is the development of ectodermal organs such as hair, teeth, and others (Thesleff & Mikkola, 2002; cited by Mustonen et al., 2003). Studies in mice revealed EDAR expression through skin development. According to Yan et al. (2000), by embryonic day 14 EDAR was present in basal cells of developing epidermis and prominently in placodes. By days 16-17 EDAR is expressed in high amounts in mature follicles, group of cells forming a cavity where some structures grow. By postnatal day 1 EDAR expression is exclusively shown at hair follicles. EDAR binds only to Ectodysplasin-A isoform A1 (EDA-A1). EDA- A1 encodes a 391 residue protein with a domain similar to TNF at the C- terminus (Yan et al. 2000). EDARADD is coexpressed with its receptor in epithelial cells during hair follicle and teeth development (NCBI, Gene, 2012). Previous studies indicate that EDAR plays a role in the activation of NF-κB, JNK, and caspase pathways (NCBI, Gene, 2012). We chose *EDAR* as a candidate for finding mutations responsible for shovel-shaped incisors because its relationship with ectodermal organ development, and because previous evidence has suggested it is partly responsible for the phenotype.

Ectodysplasin-A

Ectodysplasin-A (EDA) is a 391 residue ligand encoded in chromosome X: 68,835,911-69,259,319 (Assembly GRCh37). EDA is a type II membrane protein that belongs to the TNF family, acts as a homotrimer, and is involved in the activation of NF-κB, JNK, and caspase pathways. The *EDA* gene is composed of eight exons with a large size first intron of 340,328 bp (Abecasis et al., 2012).

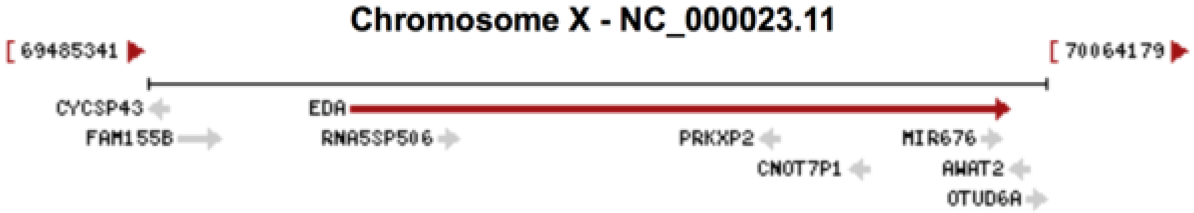


Figure 1. ***EDA* location in chromosome X.** Red arrow indicates *EDA* gene and the transcription direction. Gray arrows indicate other genes also encoded in the region.

In the first intron of *EDA*, location 69,098,181- 69,098,580, *PRKXP2* is encoded. Following *CN0T7P1*, transcribed in reverse orientation from 69,157,867 to 69,157,013, *MIR676* is transcribed from 69,242,707 to 69,242,773 (Figure 1). Both *PRKXP2* and *CNOT7P1* are pseudogenes transcribed in reverse orientation and located within *EDA*'s first intron and first exon respectively. *PRKXP2* is a protein kinase pseudogene, while *CNOT7P1* is a CCR4-NOT transcription complex, subunit 7 pseudogene 1 (NCBI, Gene, 2016). *MIR676* is a microRNA located within *EDA*'s second intron, which is known to be involved in post-transcription regulation. *EDA* is conserved in six species of mammals: humans, mouse, rats, dogs, cattle, and goats; four species of fish: Atlantic salmon, zebrafish, Nile tilapia, and channel catfish; one species of bird: chicken; and one species of amphibians: western clawed frog (“HomoloGene - NCBI,” n.d.).

The EDA gene undergoes several alternative splicing possibilities producing transcripts for seven isoforms. Four of them lack the TNF domain, and their function is unknown. Another isoform of EDA, with 386 residues annotated in the assembly GRCh38, was found but its molecular function is not clear. The other two EDA isoforms are the longest products, EDA-A1 of 391 residues and EDA-A2 of 389. Splicing events in exon eight of *EDA* result in a change in residues, protein structure, and charge, which together make two almost identical proteins specific to different receptors (Hymowitz et al., 2003); (Bayes et al., 1998). EDA-A2 results from the deletion of the residues Glu 308 and Val 309 located in the receptor binding region of the protein. Both isoforms are receptor-specific; EDA-A1 binds to EDAR, and EDA-A2 binds to XEDAR (Ectodysplasin A2-receptor) (Yan, 2000). According to the 1000 Genomes project browser, EDA-A1 has a molecular weight of 41,293.72 g/mol while EDA-A2 has a molecular weight of 41,065.48 g/mol.

The *EDA* gene has a CpG island near its transcription start site, with a size of 1415 bp which indicates that this region has been selected under evolutionary pressure (Gardiner-Garden & Frommer, 1987). The UCSC Genome Browser showed that histones H3K9 and H3K27 were tri-methylated, indicating that the gene was inactive in the cell lines used for these analyses (O ’geen, Echipare, & Farnham, 2011).

EDA is responsible for all X-linked forms of Hypohidrotic Ectodermal Dysplasia (Cluzeau et al., 2011). Ectodermal Dysplasias are categorized according to the location of the mutation, given the fact that there are several genes involved located in different chromosomes (Deshmukh, 2012). One of the phenotypic characteristics of this genetic disease is teeth malformation. It is common to find patients with oligodontia and abnormal crown form (Deshmukh, 2012). In a study conducted in mice, overexpression of EDA-A1 resulted in alterations of ectodermal appendages, extra teeth, supernumerary mammary glands with extra nipples, and abnormal hair composition (Mustonen et al., 2003). In contrast, overexpression of EDA-A2 resulted in no phenotypic abnormalities (Mustonen et al., 2003).   This could be due to the fact that EDA-A2 binds to XEDAR, a different receptor than EDA-A1. The function of XEDAR is unknown and therefore its phenotypic expression, if any, might be overlooked. Due to the relationship of mutations in *EDA* and phenotypic abnormalities in teeth, we chose this gene as a candidate for finding mutations responsible for shovel-shaped incisors.

Ectodysplasin-A receptor associated death-domain

Ectodysplasin-A receptor associated death-domain (EDARADD), an intracellular protein, serves as an adaptor to EDAR in signaling pathways leading to the activation of different transcription factors including NF-κB. EDARADD binds to the intracellular domain of EDAR, and also self-associates as it is common of death-domain proteins (Headon et al., 2001). EDARADD is encoded in human chromosome 1: 236,557,680-236,648,008 (Assembly GRCh37). Three different transcripts have been found, while in non-mammals there is only one isoform (Sadier et al., 2015).   EDARADD-001, contains six exons, and encodes a protein of 215 residues (Isoform B). EDARADD-002 has also six exons but 2,239 extra nucleotides and encodes a protein of 205 residues (Isoform A). The third isoform, transcript EDARADD-003, produces a small protein of 84 residues. A previous study stated that EDARADD Isoform B is the ancestral and conserves core functions, whereas the Isoform A is found only in mammals except in the rodent’s lineage (Pantalacci et al., 2008a). Isoforms A and B differ only in their N-terminal region, which is implicated in protein folding and stability, and use different promoters (Pantalacci et al., 2008a). The N-terminal region of EDARADD contains the binding region for TRAF6 which could affect the downstream signaling of the pathway (Sadier et al., 2015). EDARADD B Isoform has been found in all tissues and cell lines whereas the A Isoform has not, suggesting Isoform A expression is tissue-specific (Sadier et al., 2015). These findings suggest that the EDAR pathway may be regulated by the different protein isoforms therefore affecting NF-κB expression in different tissues having diverse phenotypic consequences (Sadier et al., 2015).

EDARADD plays a major role in the EDA pathway, a deletion of its N-terminal abolishes the activation of NF-κB, thus the activator EDARADD, if manipulated *in vitro,* could serve as an inhibitor (Headon et al., 2001). EDARADD overexpression activates NF-κB signaling in a dose-dependent manner, illustrating EDARADD's central role in the EDA pathway (Headon et al., 2001). A consanguineous family with a missense mutation in *EDARADD's* death domain showed decreased signaling of NF-κB. All members of the family were homozygous for the mutation and expressed phenotypic characteristics of ectodermal dysplasia (Headon et al., 2001).

X-linked Ectodysplasin-A2 receptor

X-linked Ectodysplasin-A2 receptor (*XEDAR*), with five coding exons, encodes a Tumor Necrosis Factor Receptor protein. It is located in chromosome X: 66,594,384-66,639,303 (Assembly GRCh37), and is also known as EDA2R. The amino acid sequence revealed the protein contains three-cysteine rich repeats and a trans-membrane domain (Yan, 2000). The receptor, XEDAR, is specific to the EDA-A2 isoform, failing to bind to the two-extra amino acid isoform EDA-A1. Therefore the insertion of two amino acids gives receptor binding specificity (Yan, 2000). Unlike its homolog EDAR, XEDAR does not use an adaptor for recruitment of TRAF proteins to activate NF-κB. Interestingly, in birds XEDAR was found to have a death-domain but it is unlikely that EDARADD interacts with it given the fact that the gene *EDARADD* in birds has no trace of an evolutionary shift for it to acquire this function (Pantalacci et al., 2008b). One hypothesis postulates the acquisition of a death domain could be related to feather evolution in birds, but it remains to be tested (Pantalacci et al., 2008b). XEDAR was found to bind TRAF1, TRAF3, and TRAF6, being the last one the key molecule for NF-κB signaling (Yan, 2000). XEDAR also activates the JNK pathway by TRAF3 and TRAF6, but is also dependent on ASK1, Apoptosis signal-regulating kinase 1 (Sinha, Zachariah, Quiñones, Shindo, & Chaudhary, 2002). Recently, *XEDAR* expression was found to be downregulated in breast cancer cells due to promoter methylation, but with a demethylating agent expression can be restored inducing cell-death via EDA-A2, providing a promising treatment for breast cancer (Punj, Matta, & Chaudhary, 2010). In addition, EDA-A2/XEDAR have been found to interact with the p53 signaling pathway as a barrier in tumor development (Tanikawa, Ri, Kumar, Nakamura, & Matsuda, 2010). Studies in mice have reported that XEDAR expression in development of ectodermal organs is dispensable, and mutations in XEDAR related to HED have not been reported (Newton, French, Yan, Frantz, & Dixit, 2004).

EDA pathway and shovel-shaped incisors

In 2007 a group of scientists from the University of Tokyo identified *EDAR* as a candidate gene to have undergone positive selection (Kimura, Fujimoto, Tokunaga, & Ohashi, 2007). In the subsequent year it was reported that *EDAR* was a major determinant of Asian hair thickness, and that a discovered variant, T1540 C, was positively selected recently in that population (Fujimoto et al., 2008). These findings led this group to explore the relationship of *EDAR* allele T1540C with tooth morphology, specifically the shovel-shaped incisors phenotype. It was found that the *EDAR* T1540C allele was responsible for 18.9% of the total variance of the shovel-shaped incisors phenotype, and it was also found affecting overall tooth size and mesiodistal and buccolingual diameter (Kimura et al., 2009). The forces responsible for the 81.1% of the total variance for shovel-shaped incisors phenotype are at this time unknown.

In the present study we aimed to find other variants in the main components genes of EDA pathway that might address part of the variance for shovel-shaped incisors phenotype. We also have reason to believe that *EDAR* allele T1540C might have been recently positively selected among the Puerto Rican population. The Native American ancestry in the1000 Genomes Puerto Rican data is 12.8%, while around *EDAR* it is 26%. Furthermore, the allelic frequency for *EDAR* T1540C is 21%. Therefore, we hypothesize that the *EDAR* allele T1540C has been positively selected in the Puerto Rican population.

Methodology

Sample selection and Genotyping of the *EDAR* allele 1540 C

Samples were obtained from a previous study in which they were selected using a census-based sampling frame to be representative of the Puerto Rican population. The ancestry has been already estimated through the use of 93 ancestry informative markers (AIMs) dispersed throughout the genome (Via et al., 2011). We genotyped the SNP rs3827760 in 452 of these samples using TaqMan Allelic Discrimination assay performed on Applied Biosystems (AB) Real-Time PCR platform (RT-PCR). Genomic DNA final concentration ranged between 2-20 ng/uL. Primers and probes were ordered from TaqMan SNP Genotyping assays. The probes for the reference allele 1540T were labeled with Vic dye, and for the alternate allele 1540C with FAM dye. The parameters for the assay were: 95 °C for 10 minutes, 50 cycles of 92 °C for 15 seconds followed by 60 °C for 1 minute and 30 seconds. Final extension step of 60 °C for 30 seconds was used to complete the unfinished amplicons.

Identification of one or more SNPs in the genes involved in EDAR pathway

SNPs selection

Using Puerto Rican population data from the 1000 Genomes project phase 1 (Abecasis et al., 2012) we constructed median joining networks for the genes involved in the EDAR pathway: *EDAR, XEDAR, EDARADD,* and *EDA.* The networks were made using the Free Phylogenetic Software from website www.fluxus-engineering.com (Bandelt H-J, Foster P, 1999). SNPs that distinguished haplotype families and altered the amino acid sequence were analyzed by

PolyPhen (Adzhubei et al., 2010) to predict their phenotypic impact. In addition, because the shovel shaped incisors phenotype is more prevalent in East Asia we hypothesized that other shovel shaped incisors associated SNPs might be more common in the Amerindian ancestry of Puerto Rico. Therefore, further consideration as candidates SNPs was given to SNPs that were common in Puerto Ricans but not so among Europeans or Africans, according to 1000 Genomes Project phase 1 (Abecasis et al., 2012). A total of 11 SNPs were selected for analysis by TaqMan genotyping assay.

EDAR Networks

i. *EDAR* 5'UTR region

The EDAR gene is located on chromosome 2: 109,510,927-109,605,828 (assembly CRCh37/hg19), where its first intron contains 57,927 bp. We selected the 5'UTR region of the gene to construct a network to find candidate SNPs involved in the regulation process. As shown in Figure 2 the network is divided into two major groups separated by several polymorphisms. We chose a SNP that was not repeated in any other part of the network, thereby being exclusive to separate the two major groups. The selected SNPs were rs365060, circled in red in Figure 2, and rs9967821.

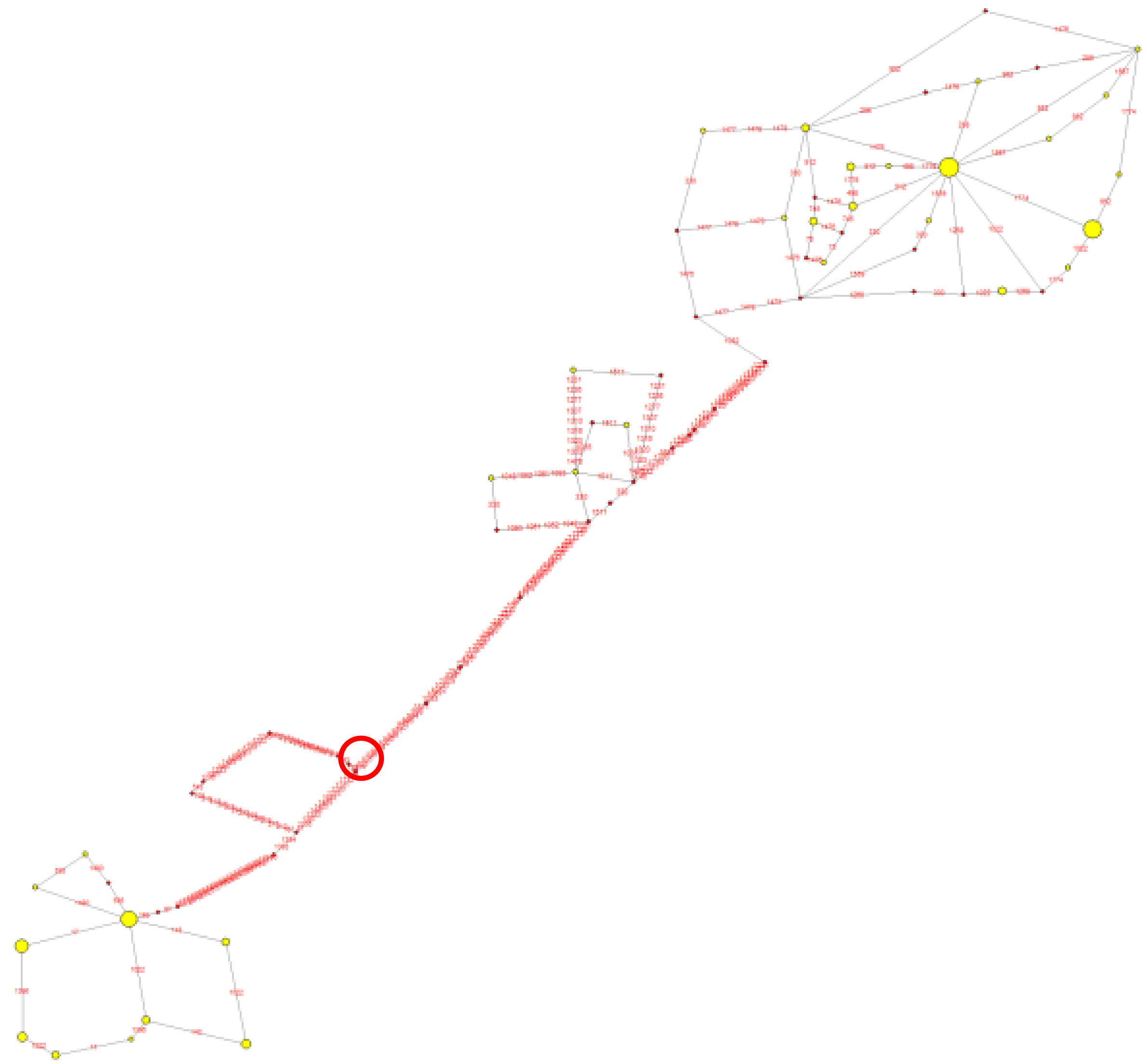


Figure 2. Network for *EDAR* 5'UTR region. Red numbers represent SNPs. Circled in red is the SNP selected for analysis. Yellow circles represent group of haplotypes.

ii. *EDAR* complete region

The *EDAR* gene is 94,901bases long. In order to make a haplotype analysis of the entire gene in the Puerto Rican population, the gene was divided into segments because of its large size; the division was made according to linkage disequilibrium data from the HapMap Project1 (data source: HapMap Data PhaseIII/Rel#2, Feb 09, on NCBI B36 assembly, dbSNP b126).

There were three populations chosen, Japanese, European, and Yoruban, to represent Puerto Rican ancestry, Taínos, Spaniards, and Africans, respectively. After haplotype analysis we chose to construct networks with European HapMap LD data because European is the major ancestry in Puerto Rico (Via et al., 2011).

Linkage disequilibrium (LD) is when alleles at two or more loci are inherited together more than expected based on their respective frequencies and random association assumption. Hence, LD plots provide recombination hotspot positions that are useful to define haplotype transition. The LD plot of European population that was used for haplotype estimation is presented in the figure below.

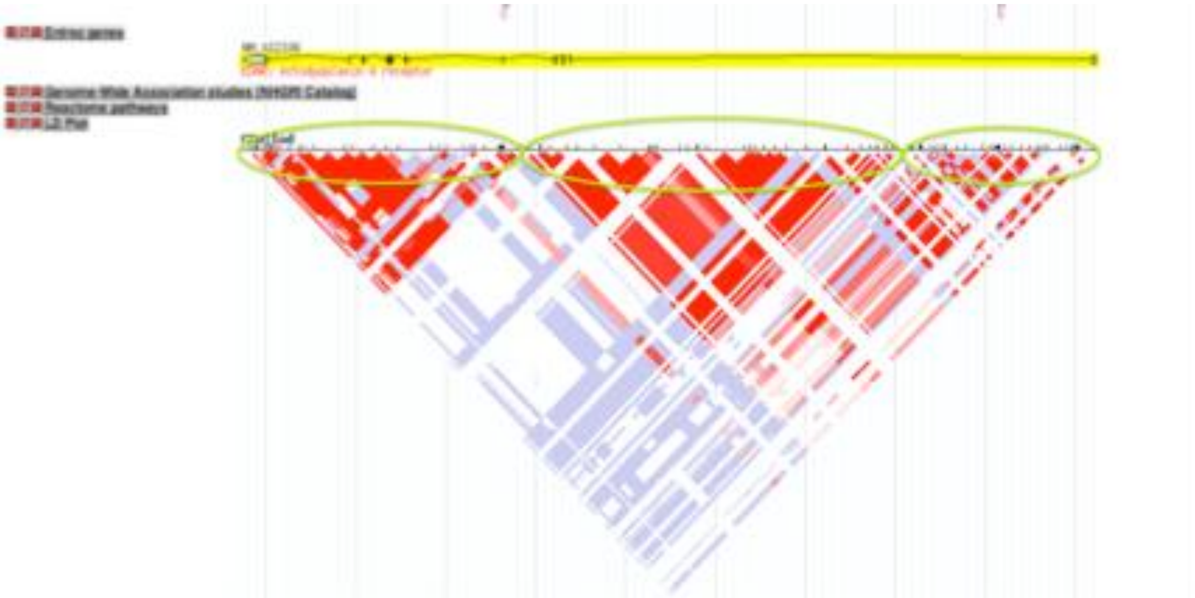


Figure 3. LD Plot for European Population. The alleles with highest LD are colored in red. White color indicates random probability, blue color indicates slightly closer than random, and light red strong albeit non highest LD. Surrounded in green are the three segments in which the gene was divided to calculate haplotype estimates.

Haplotypes estimates were analyzed by PHASE v.2.1, a software for haplotype reconstruction and recombination rate estimation from population data (Stephens, Smith, & Donnelly, 2001) (Stephens & Scheet, 2005). From the three segments into which the gene was divided (see Figure 3) we constructed three networks using only haplotypes with a frequency higher than 7 according to PHASE analysis.

The SNP for the *EDAR* allele 1540C, rs3827760, was found in segment A between two haplotypes, marked as number 23 in Figure 4 and circled in red. From segment B we selected the SNP rs10174266 given its location in the first intron which is 57,927 bp long, which could have a big impact in transcription processes (Figure 5). For segment C there were only two haplotypes, for which we made an alignment to look for the points in which they differentiated. From segment C we selected the SNP rs3749110 which is located in the 5' UTR, and therefore could have implications in both transcription and mRNA translation regulation. The SNP has a frequency of 34% in Puerto Ricans and 85% in Asians (Figure 6).

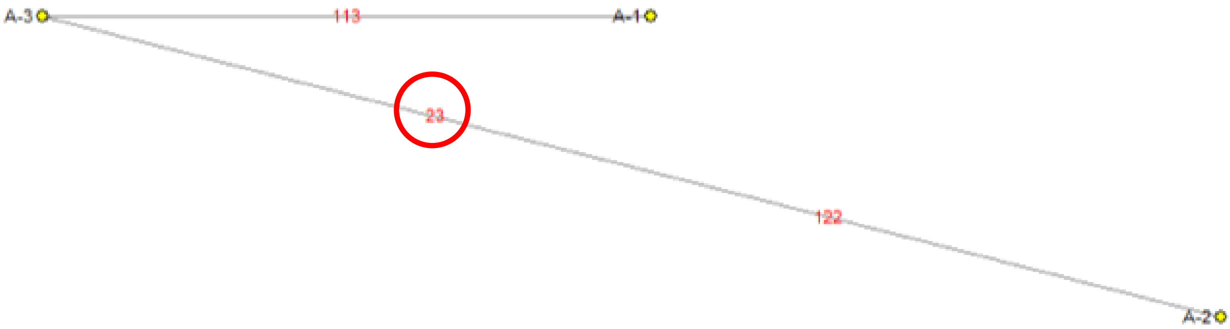


Figure 4. Network for Segment A of *EDAR*. 128 polymorphisms were used to construct this phased abbreviated network. Red numbers represent SNPs. Circled in red is the SNP selected for analysis. Yellow circles represent group of haplotypes.

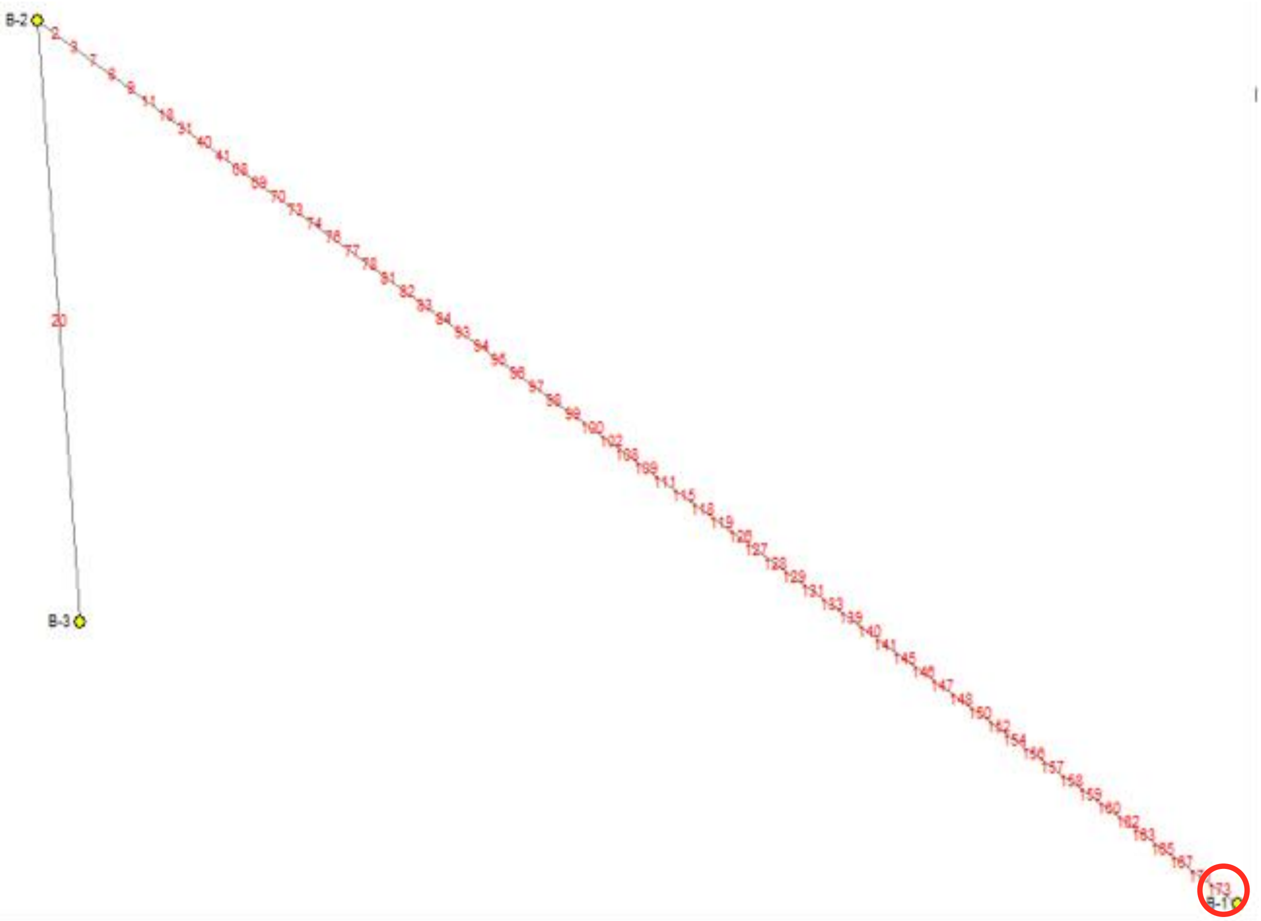


Figure 5. Network for Segment B of *EDAR*. 175 polymorphisms were analyzed when constructing this phased abbreviated network. Red numbers represent SNPs that change allele between haplotype groups. Yellow circles represent haplotype groups. Circled in red is the SNP selected for analysis.

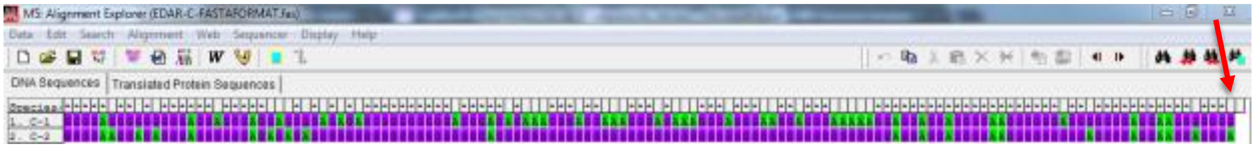


Figure 6. Screenshot of FASTA alignment for Segment C of *EDAR*. 138 polymorphisms were aligned between two haplotypes using MEGA7. Marked by a red arrow is the position number 138 which correspond to the SNP rs3749110.

XEDAR Networks

i. 5'UTR

XEDAR is located on chromosome X: 65,815,479-65,859,108 (assembly GRCh37/hg19) and has seven exons but only five are protein coding. A network was produced using the 5' UTR of XEDAR. The region selected for constructing the network includes the first non-coding exon, and the first intron (Figure 7). The network shows two major haplotype groups separated by just one SNP; number one, circled in red (Figure 8). This SNP is rs150948525 and is located in the first intron of the gene. rs150948525 has two alleles T/C where the alternate C has a frequency of 11% among Puerto Ricans.

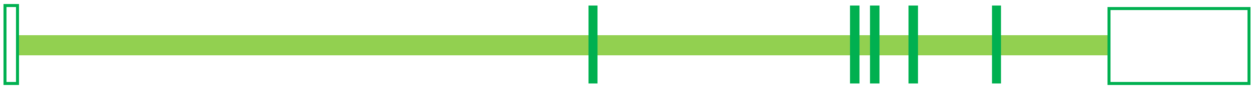


Figure 7. XEDAR gene structure. *XEDAR* is transcribed in reversed. Vertical green lines represent protein coding exons. Horizontal green line represents introns. Un-colored box represents the first non-coding exon.

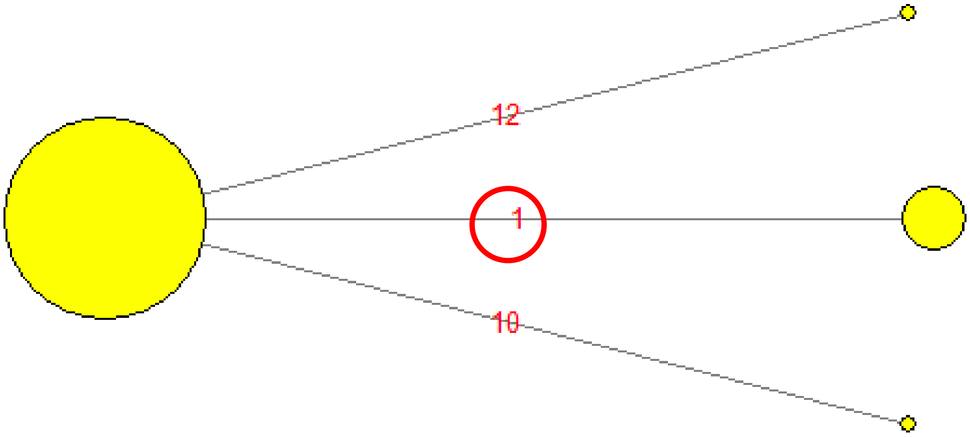


Figure 8. Network for *XEDAR* 5'UTR. 25 polymorphisms were used for the construction of this network. Red numbers represent SNPs. Circled in red is the SNP selected for analysis. Yellow circles represent group of haplotypes.

ii. Entire gene region

The network of the rest of the XEDAR gene was constructed using the 1000 Genomes phase 1 Puerto Rican population data (Abecasis et al., 2012). Figure 9 shows that the network is divided into three major groups that can be interpreted as the three major paternal populations found in Puerto Ricans: European, African, and Native American. We looked for SNPs that separated these three groups and selected a SNP (circle marked as number 93 pointed in red in Figure 9) which separated one big group from the rest of the network. This SNP is rs1385699, which is located in the third exon of the gene. The mutation is categorized as non-synonymous since the change in alleles C/T produces an amino acid change of arginine to lysine. An analysis of the functional effects of this mutation was performed by PolyPhen (Adzhubei et al., 2010) which predicted it to be a benign mutation with a score of 0.427. The SNP frequency among Puerto Ricans is 65% for the alternate allele, compared to 79% among American populations and 100% in Asians. American populations (AMR) is a category of populations used on 1000 Genomes phase 1 (Abecasis et al., 2012) that include Puerto Ricans, Mexicans from Los Angeles California, and Colombians in Medellin.

As shown in Figure 9 another SNP of the network was examined (marked as number 63 and circled). This mutation is non-synonymous and changes the amino acid threonine for an alanine. The reason this SNP was not selected for further analysis is due to the fact that although it is a non-synonymous mutation, it appears only in one of the chromosomes of the 1000 Genomes phase 1 (Abecasis et al., 2012) Puerto Rican data. Therefore, we decided this SNP was not as informative for population analysis.

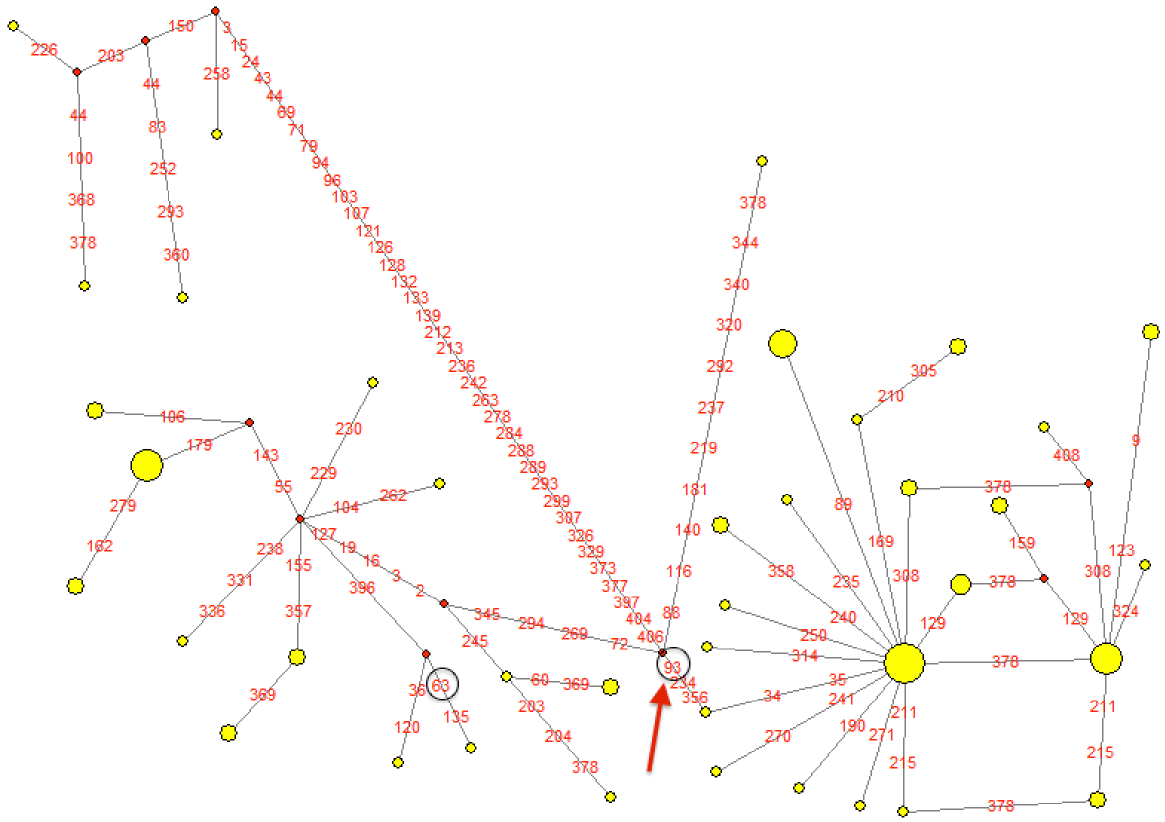


Figure 9. Network for *XEDAR* entire gene region. 409 polymorphisms along *XEDAR* gene were used to construct this network. Red numbers represent SNPs. Circled and pointed by the arrow are the SNPs selected for analysis. Yellow circles represent group of haplotypes.

EDARADD Networks

i. Entire gene region

The *EDARADD* gene is located in chromosome 1:236,511,562-236,648,214 (assembly CRGh37/hg19). Figure 10 shows the network for the entire region of *EDARADD*. In this network there are no distinguishable haplotype groups, since there are too many differences among individuals. Thereby, we did not select this network for future analysis.

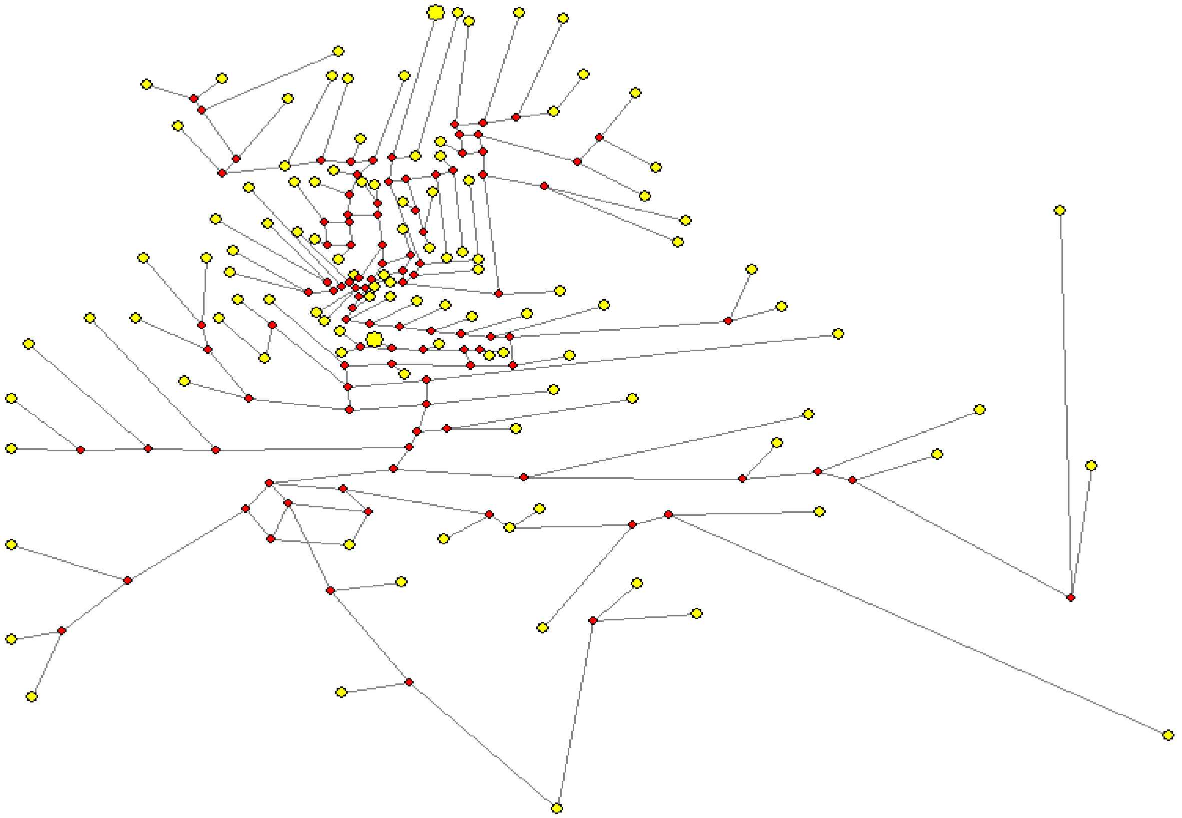


Figure 10. Network for *EDARADD*. 1,462 polymorphisms along EDARADD gene were used to construct this network. Yellow circles represent group of haplotypes. Red dots represent hypothetical ancestors.

ii. EDARADD 5'UTR

We selected the *EDARADD* 5'UT region from 1000 Genomes phase 1 Puerto Rican data (Abecasis et al., 2012) to construct a network (Figure 11). The network showed two major haplotype groups separated by one SNP marked as number 6 and circled in red on Figure 11. The SNP is rs79233817. According to the 1000 Genomes phase 1 browser (Abecasis et al., 2012), the frequency of rs79233817 among Puerto Ricans is 4%, in Africans 17%, Americans 2% and Europeans 0%. Because the alternate allele is mostly found in African populations and was thus unlikely to code for shovel-shaped incisors, the SNP was not selected for purposes of this study. The other haplotype separated by another SNP was not taken into account because it was found in only one individual.

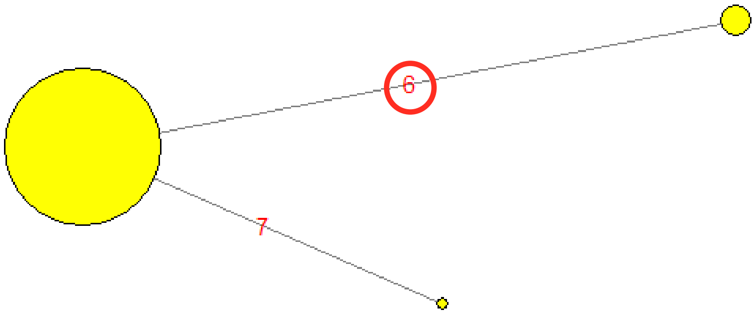


Figure 11. Network for *EDARADD* 5'UTR. A total of 7 polymorphisms from *EDARADD* 5'UTR were used to construct this network. Red numbers represent SNPs. Circled in red are is the SNP selected for analysis. Yellow circles represent group of haplotypes.

iii. EDARADD exons

For a more informative analysis we chose to use only exome data for constructing this network. As shown in Figure 12, the network has three major haplotype groups. Two are separated by number 6 that represents the SNP rs966365 but since both alleles have high frequency in the African population it would not be a determinant for shovel-shaped incisors. Number 11 separates two large haplotype groups but the SNP was not selected since the allelic frequencies were higher in the European population. We chose two SNPs in the networks marked by a red circle in Figure 12. The SNP marked as number 10 refers to SNP rs114632254, a non-synonymous mutation located in the last exon that changes the amino acid serine to phenylalanine. The SNP marked as number 4 refers to the SNP rs79233817 located in the 5' UTR of the gene. Although it is not coding, the UTR is known to have important regulatory functions that can affect the signal transduction pathway.

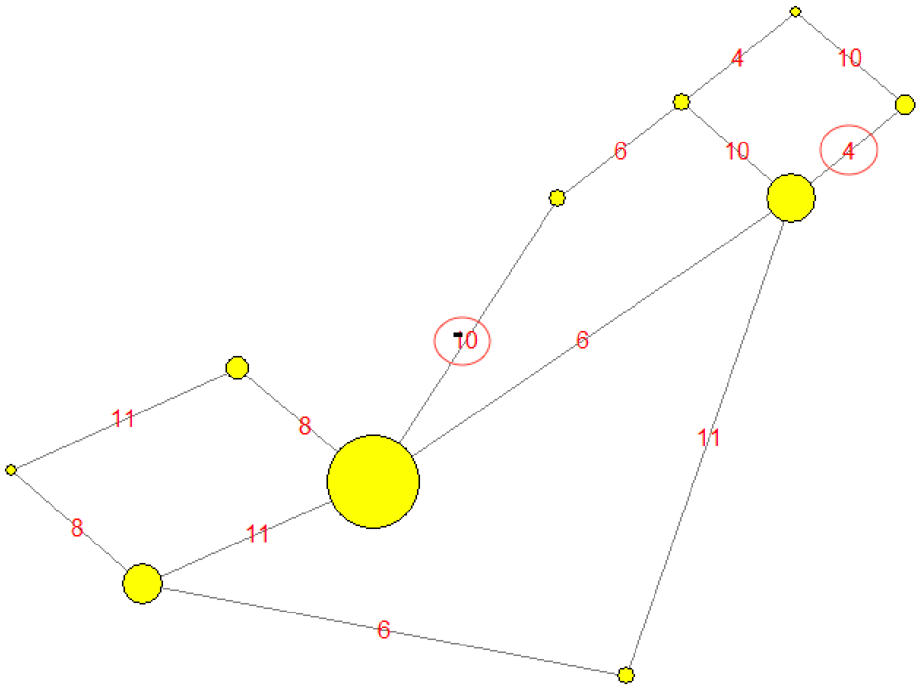


Figure 12. Network for *EDARADD* exome. 18 polymorphisms from *EDARADD* exome were used to construct this network. Red numbers represent SNPs. Circled in red are the SNPs selected for analysis. Yellow circles represent group of haplotypes.

EDA Networks

i. *EDA* gene segment 1

The *EDA* gene has more than 400,000 bp. Therefore, we had to divide the data into 5 segments of around 90,000 bp each. The network for the first segment is shown in Figure 13. As expected, the network showed three major groups of haplotypes. Since this was network 1 of 5 we decided to construct networks of exome data and promoter region to get a more informative analysis.

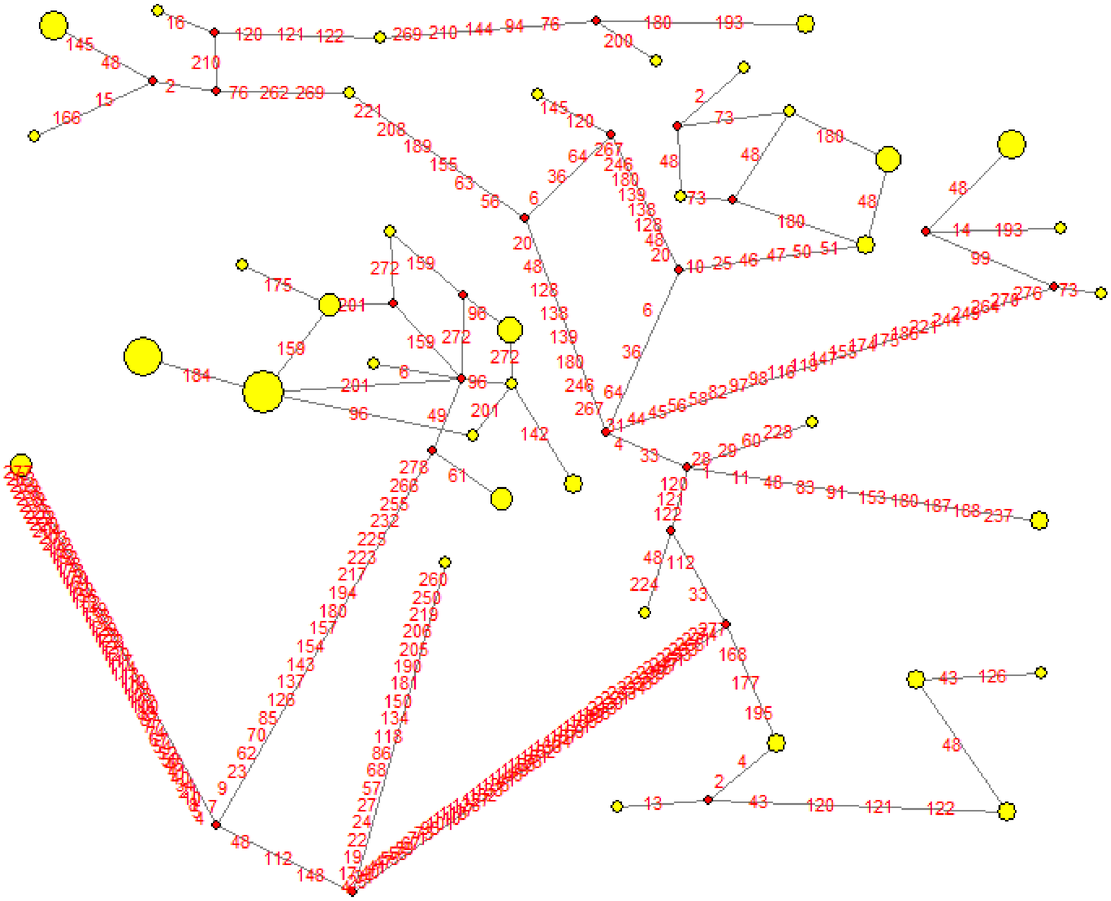


Figure 13. Network for *EDA* first segment. 278 polymorphisms from *EDA* first 90,000 bp were used to construct this network. Red numbers represent SNPs. Yellow circles represent group of haplotypes.

ii. EDA exome

A median joining network using only exome data of *EDA* was constructed (Figure 14). Two major haplotype groups were found separated by one SNP, in the network circled as red and marked as number 43. The point refers to the SNP rs3795170 which according to the 1000 Genomes phase 1data (Abecasis et al., 2012), its alternate allele G is found among 46% of Puerto

Ricans, 38% Europeans, 56% Asian populations, and 15% Africans. This SNP was selected for genotyping analysis in relation to the shovel shaped incisors phenotype.

We can see in Figure 14 that there are two haplotypes separated from the second largest group. We evaluated each point that separated the haplotypes and found that the SNP circled in red and marked as number 35 was a good candidate for further analysis. SNP number 35 in the network refers to variant rs3764746 which is found in 11% of Puerto Ricans, and in 23% of Asians, 15% in Africans, and only 2% in Europeans. Given these frequencies we assumed that the allele could have been inherited mostly by our Native American ascendency.

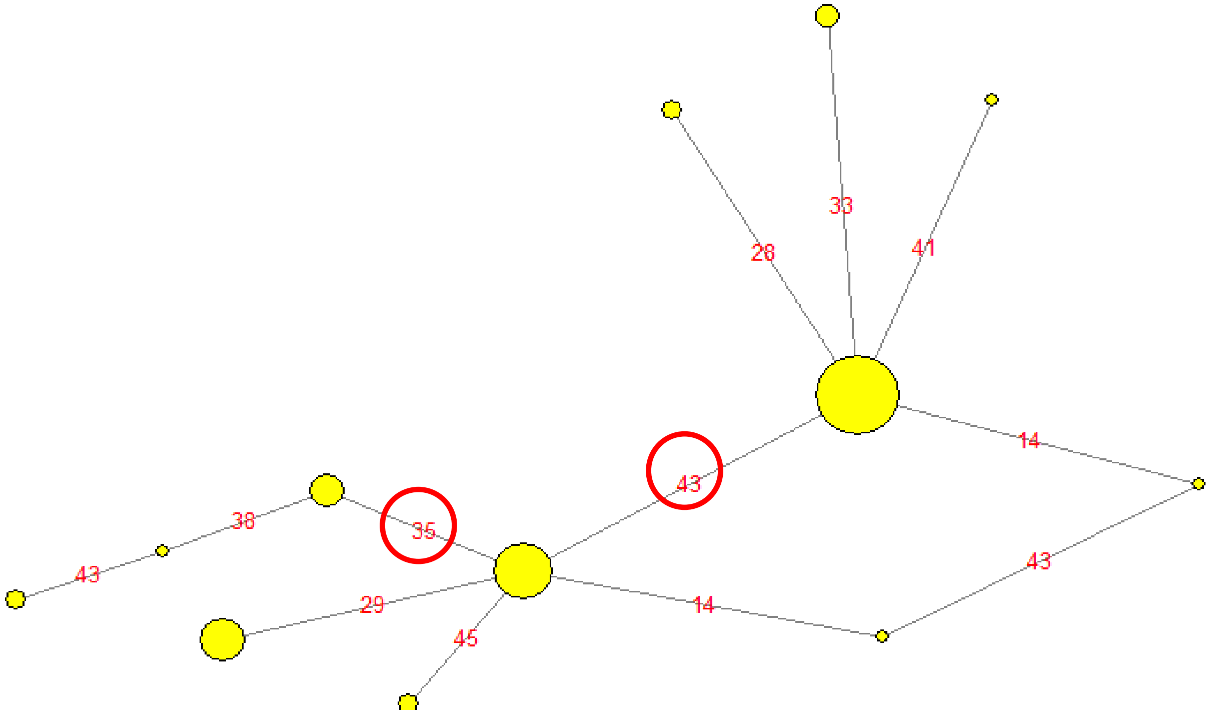


Figure 14. Network for *EDA* exome. 45 polymorphisms from *EDA* exome were used to construct this network. Red numbers represent SNPs. Circled in red are the SNPs selected for analysis. Yellow circles represent group of haplotypes.

iii. EDA intron 1

The first intron of *EDA* is 340,328 bases long. We constructed one network using only this region of the gene but it was not informative enough for selecting SNPs that separated large groups of haplotypes (Figure 15). The variation among individuals is higher due to its larger size.

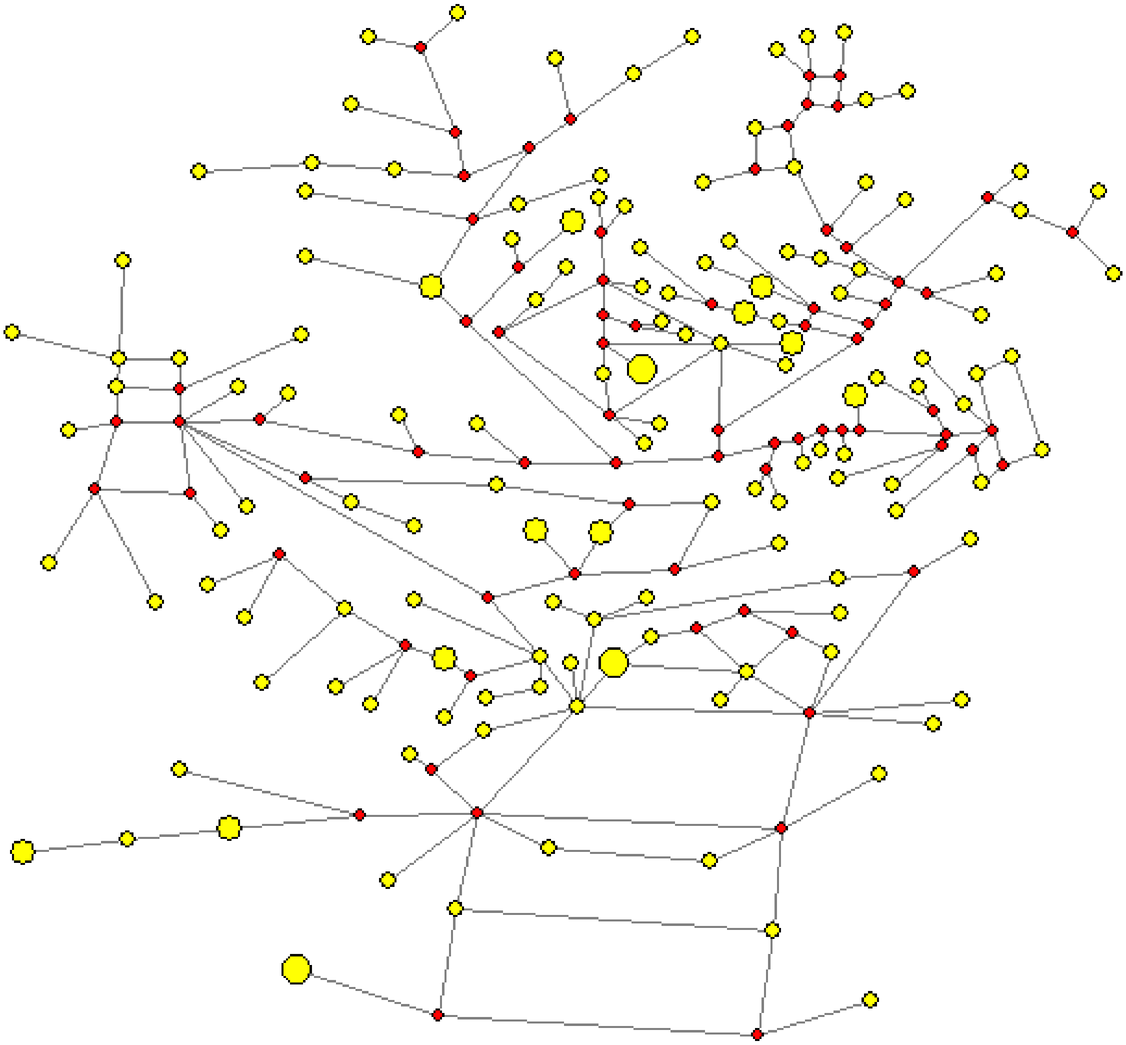


Figure 15. Network for *EDA* first intron. 1,520 polymorphisms from the first intron of *EDA* were used to construct this network. Yellow circles represent group of haplotypes. Red dots indicate points of recombination.

Sample collection and preparation

DNA samples were obtained from saliva of 56 Puerto Ricans with or without shovel-shaped incisors as determined by a dental hygienist. Dental hygienists made an effort to collect samples from patients with the phenotype. Given the low frequency of *EDAR* allele 1540C by targeting patients with the phenotype the collected samples would be enough to give our tests the statistical power necessary. Permission to collect DNA from dental patients was granted by the Institutional Review Board of the University of Puerto Rico at Mayagüez. Each patient/guardian participated voluntarily and provided informed consent for this study. Males were preferably chosen because some genes in the EDAR pathway are located in the X chromosome. Therefore, how these genes interact with other gene variants will depend on the gender of the individual. Previous studies have shown a higher degree of shovel-shaped incisors in females than in males (ref.), and we thus preferred to make comparisons only among members of the same gender. We also decided to select only males because of the more straightforward interpretation of the genotype/phenotype correlation, because males have only one X chromosome. In addition, taking women could introduce an unknown variant to our study because we would not know which of the alleles in women heterozygous for X-linked genes would be inactivated in developing tissue critical for incisor morphology. Only one clinic provided female samples and these were not taken into consideration. Finally, we preferred young males to reduce shovel-shape smothering due to tooth use.

For collecting the saliva, we used Oragene DNA (OG-500) and for DNA extraction we used PrepIT.L2P as instructed by the manufacturer (DNA Genotek). Genomic DNA final concentration was 10 ng/uL. Maxillary plaster casts were obtained also by the dental hygienist and classified for shovel-grade by forensic anthropologist Edwin Crespo, PhD, using the Arizona

State University dental anthropology system. The classification for shoveling grade was made without knowledge of the genotypes for each sample. From the 56 collected samples, only 46 were used in the experiments because of sample quality and sex gender (Table 1).

Table 1. Samples collected by Municipality

Municipality	Males	Age mean
Aguadilla	6	23
Arecibo	5	22
Bayamón	6	24
Fajardo	4	27
Manatí	3	30
Mayagüez	6	23
Ponce	3	19
Río Grande	11	24
Yauco	2	24

Genotyping

The samples for the selected 11 SNPs (Table 2) were genotyped by the Hussman Institute for Human Genomics at the University of Miami. We provided 64 DNA samples, 56 collected experimental samples and 8 control samples from the 1000 Genomes Project. All DNA concentrations were normalized to 40-70 ng/uL prior to plating in Open Array format. DNA quality was assessed via agarose gel electrophoresis on a 0.8% agarose gel. Genotyping was carried out using a custom OA32 panel on the QuantStudio 12K Flex Real Time PCR System.

Table 2. SNPs selected for genotyping.

SNP rs	Gene	Location[a]
rs3827760	EDAR	2:109513601
rs365060	EDAR	2:109575736
rs9967821	EDAR	2:109554314
rs10174266	EDAR	2:109583313
rs3749110	EDAR	2:109605767
rs150948525	XEDAR	X:65834674
rs1385699	XEDAR	X:65824986
rs114632254	EDARADD	1:236645609
rs79233817	EDARADD	1:236557742
rs3795170	EDA	X:69258515
rs3764746	EDA	X:69257972

[a] Chromosome: position, according to assembly Gh37

Results

Determination of the *EDAR* allele 1540C frequency among Puerto Ricans

The SNP rs3827760 was genotyped in 452 Puerto Rican samples obtained from a sample set representative of the Puerto Rican population by TaqMan allelic discrimination assay, as previously described. We found 8 individuals homozygotes for the *EDAR* allele 1540C, 87 heterozygotes and 305 homozygotes for the common allele 1540T (Figure 16). The genotype of 52 samples was undetermined due to bad amplification or ambiguous clustering. The allelic frequencies were 13% for 1540C and 87% for 1540T. A correlation between genotype and demographic location was not found.

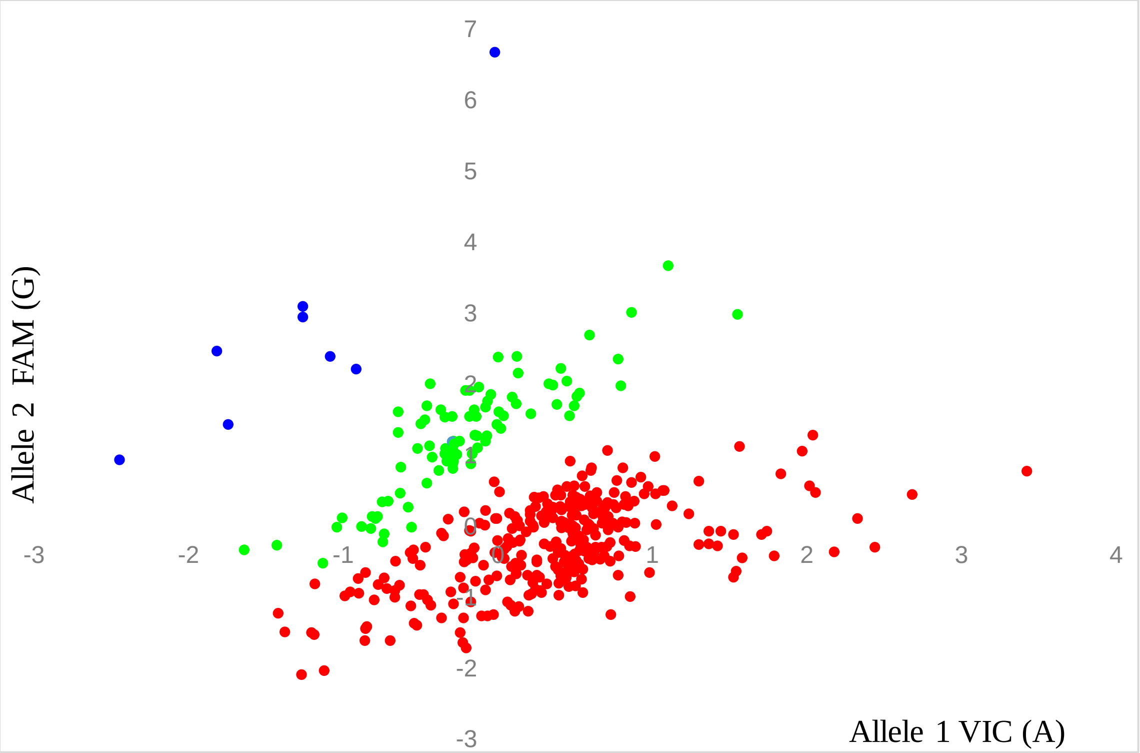


Figure 16. Genotyping *EDAR* rs3827760 in 406 samples representative of the Puerto Rican population. Legend: red = AA, green = AG, blue = GG. X axis represents VIC fluorescent dye intensity for allele 1 (A) after data normalization. Y axis represents FAM fluorescent dye intensity for allele 2 (G) after data normalization.

Identification of one or more SNPs in the genes involved in EDAR pathway also responsible for shovel-shaped incisors

After several analyses described in the Methodology, we found 11 SNPs in the genes involved in the EDA pathway that could be related to the shovel-shaped incisors phenotype. Five of the eleven SNPs were not able to be genotyped due to poor clustering. Of the 56 saliva samples collected, only 46 were useful for analysis. Only the upper incisor 1 right side shoveling grade was used for the analyses. A histogram of the distribution of the shoveling grades among our samples is presented in Figure 17. Mean shoveling grade for the samples was 1.37, with a standard deviation of 0.9859 and median of 1.0. Distribution of shoveling grade for each SNP is shown in Figures 18 to 23.

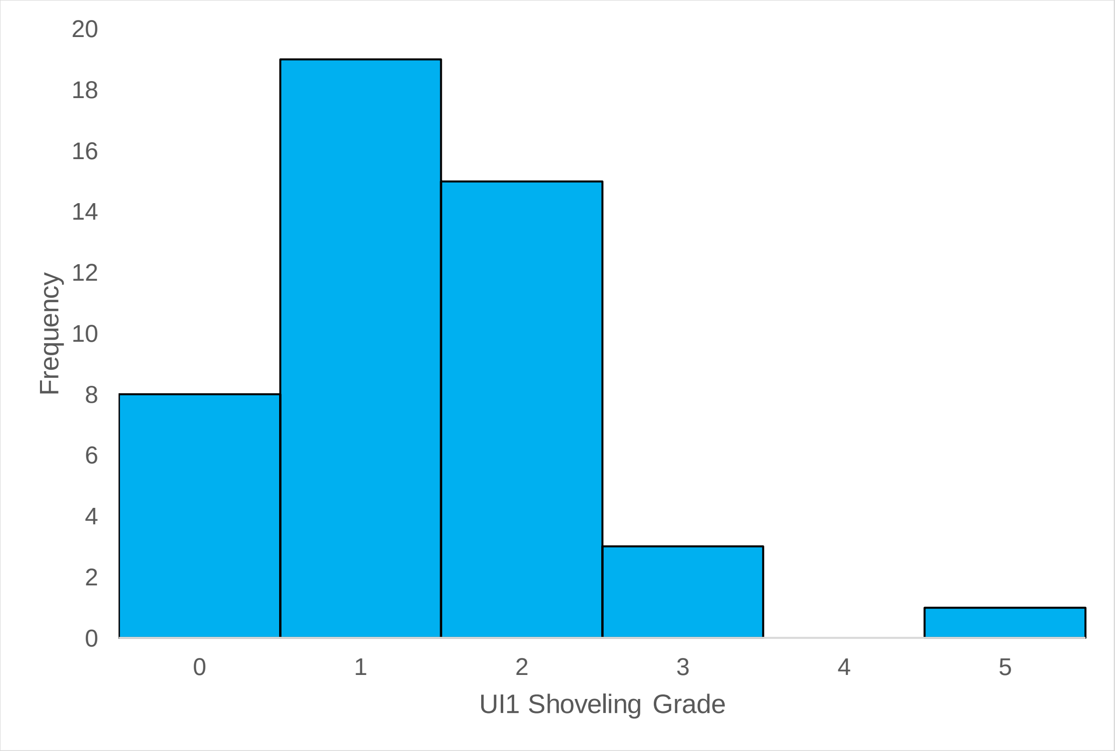


Figure 17. Histogram of shoveling grade distribution among the samples (*n* = 46).

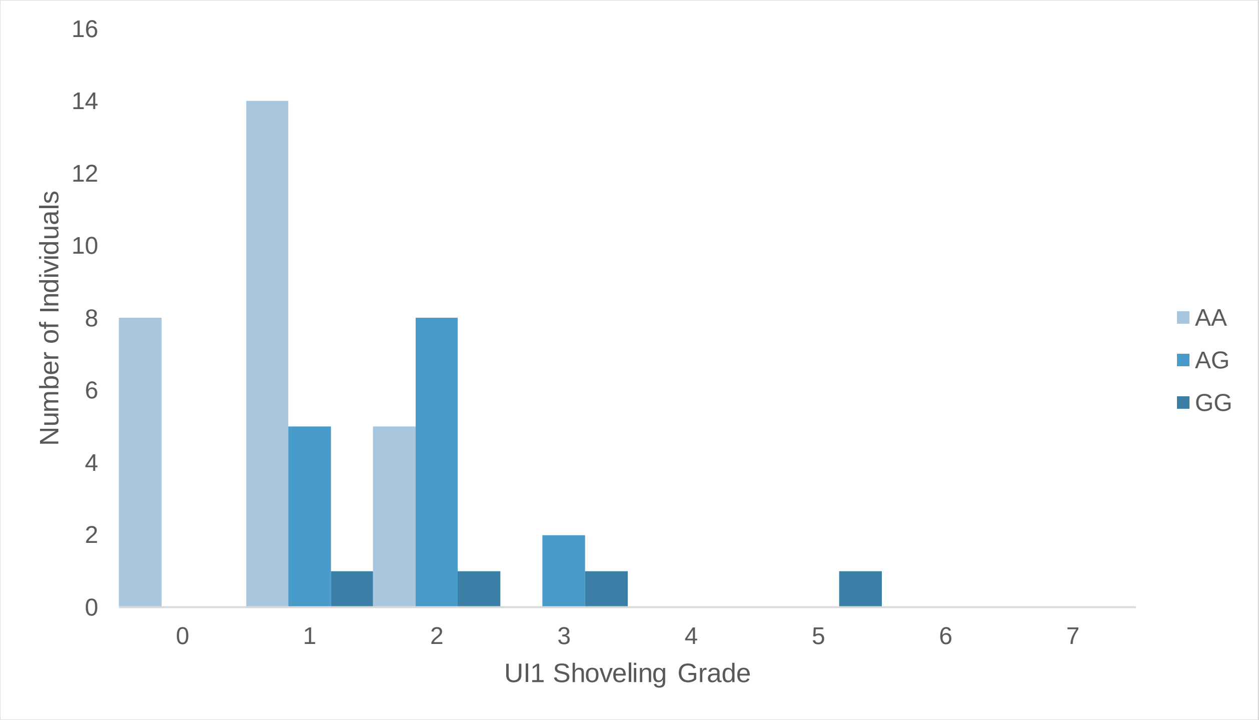


Figure 18. rs3827760 genotype distribution in dental-morphology-graded samples (*n* = 46).

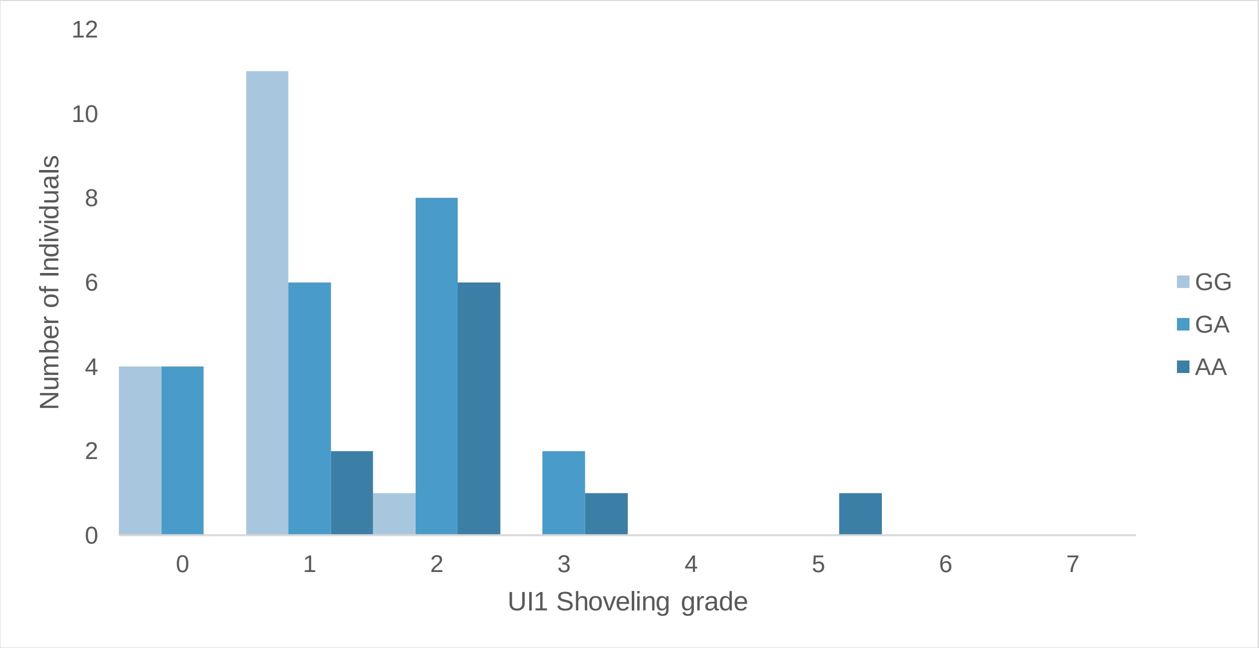


Figure 19. rs3749110 genotyping distribution in dental-morphology-graded samples (*n* = 46).

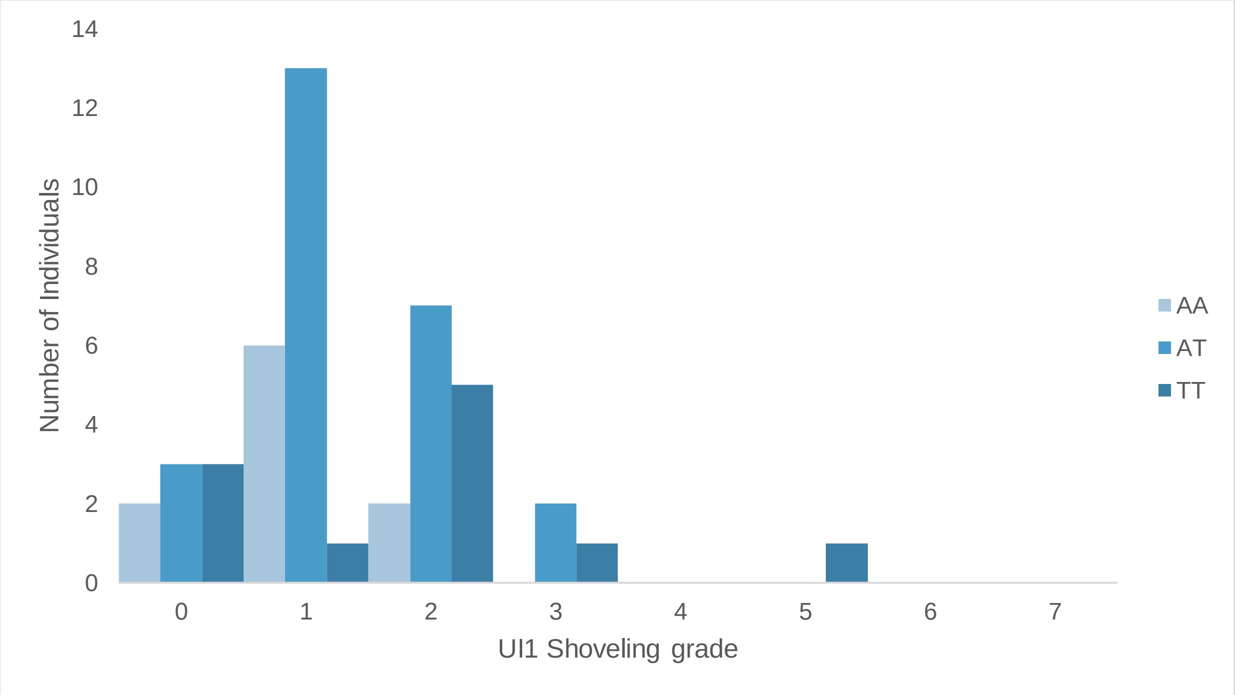


Figure 20. rs10174266 genotyping distribution in dental-morphology-graded samples (n = 46).

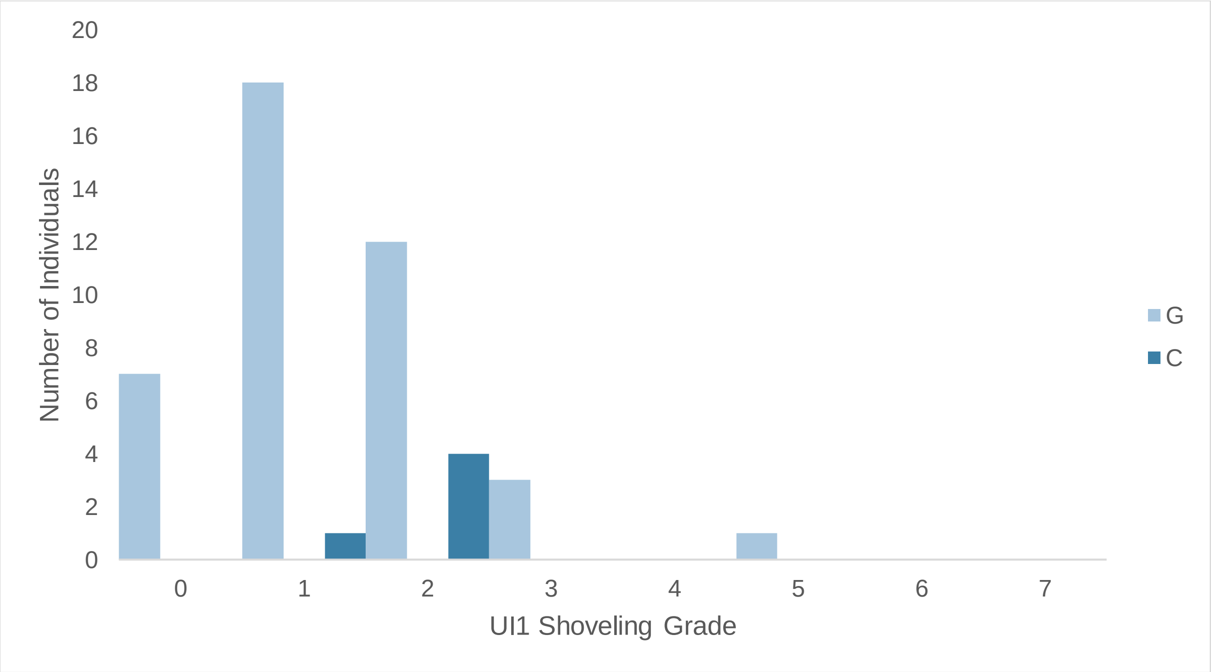


Figure 21. rs3764746 genotyping distribution in dental-morphology-graded samples (*n* = 46).

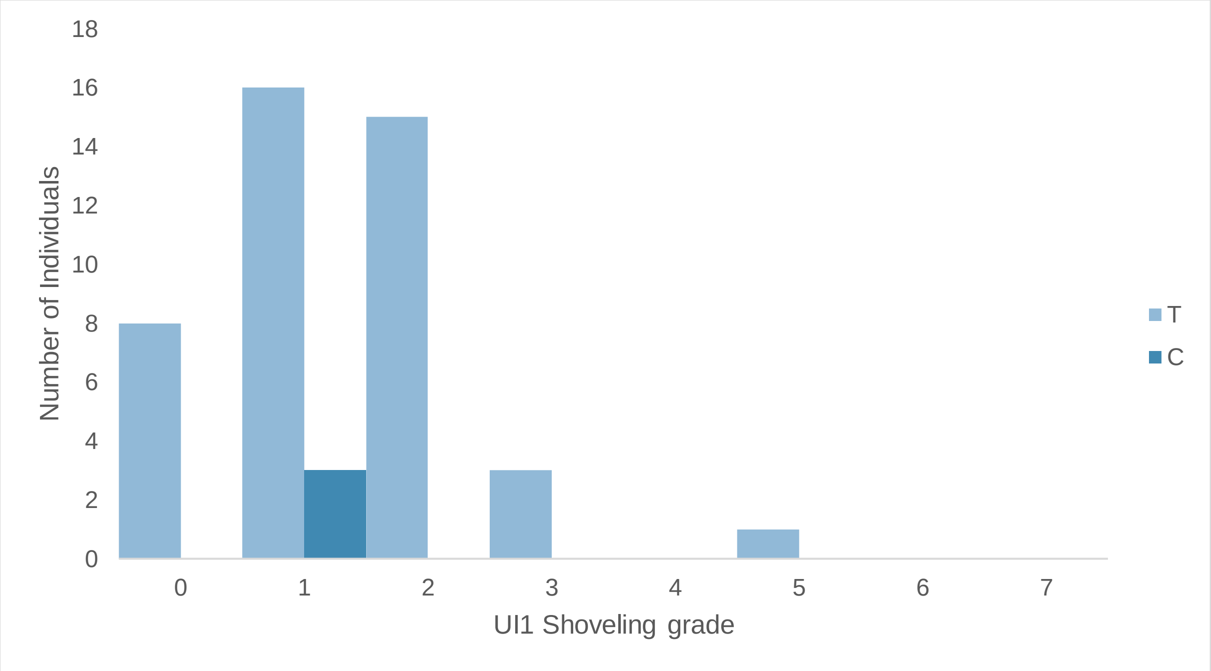


Figure 22. rs150948525 genotyping distribution in dental-morphology-graded samples (*n* = 46).

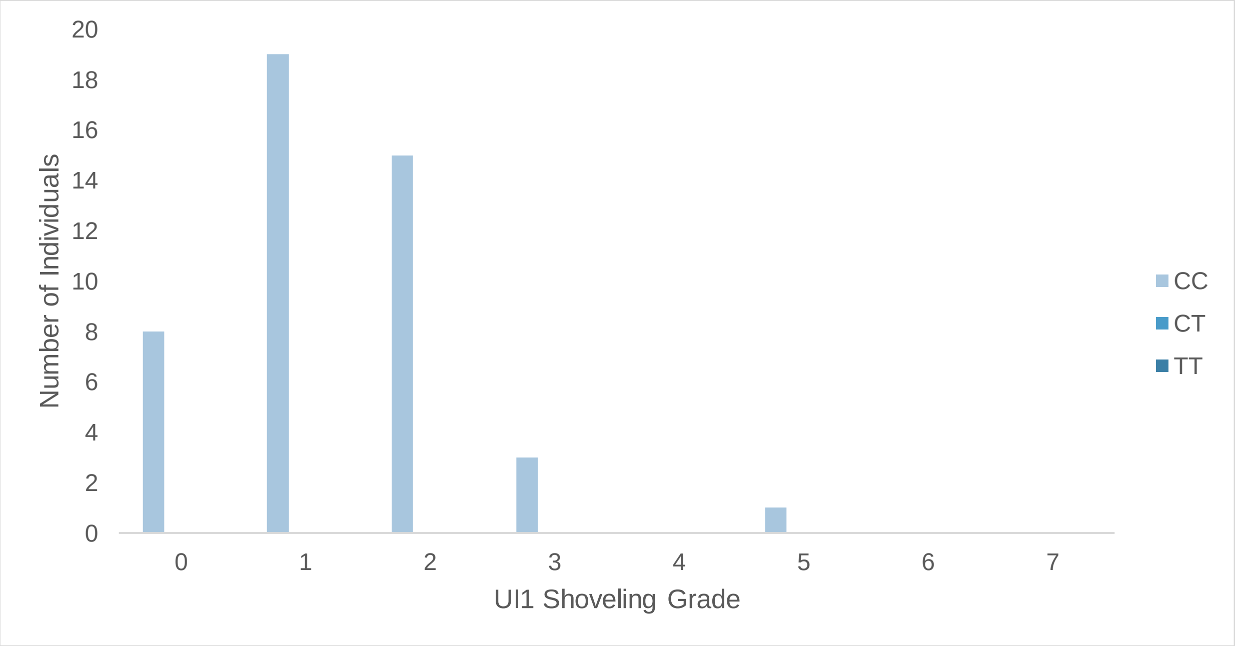


Figure 23. rs114632254 genotyping distribution in dental-morphology-graded samples (*n* = 46).

Results of the regression analyses using a codominant model for the six genotyped SNPs are summarized in Table 3. SNPs rs3827760 and rs3749110, located in *EDAR*, were found to have the most correlation with the phenotype, according to their p-values. LD was calculated between rs3827760 and rs3749110. D' was 0.9186 and r^2 was 0.38 (Figure 24) (Machiela & Chanock, 2015). Haplotype frequencies for both alleles are shown in Figure 25. The most common haplotype in the Puerto Rican population is the homozygous for the ancestral alleles, while having both rare alleles is the least frequent. No correlation was found between the shovel-shaped incisors and the SNPs located in *EDA*, *XEDAR*, and *EDARADD*. The regression coefficient shows that the *EDAR* allele 1540C increases the shoveling grade by 0.9. Coefficient of determination (R^2) shows the *EDAR* allele 1540C explains 39% of the variance.

We also tested other models for a more informative analysis between the SNP rs3827760 and shovel-shaped incisors. The Dominant Model favors the rare allele as dominant towards the ancestral allele with a p-value of 2.2855e-05. In this model the variation determined by R^2 was 0.3377 and the regression coefficient was 1.1637. As expected, in the Allelic Model, also known as No Model assumption, there was a significance relationship between allele and phenotype with a p-value of 2.1549e-06. Haplotype regression analysis for rs3827760 showed no special haplotype is correlated with shovel-shaped incisors.

Table 3. Regression Analysis for genotyped SNPs.

SNP rs	Gene	P	b_{yx}	R^2
rs3827760	EDAR	3.2015e-06	0.9487	0.3925
rs3749110	EDAR	0.00027	0.6828	0.2629
rs10174266	EDAR	0.09666	0.3644	0.06287
rs3764746	EDA	0.43058	0.4167	0.01418
rs150948525	XEDAR	0.5127	-0.3953	0.0098
rs114632254	EDARADD	0.5872	0	0.00675

Abbreviations: P, p-value; b_{yx}, regression coefficient; R^2, coefficient of determination

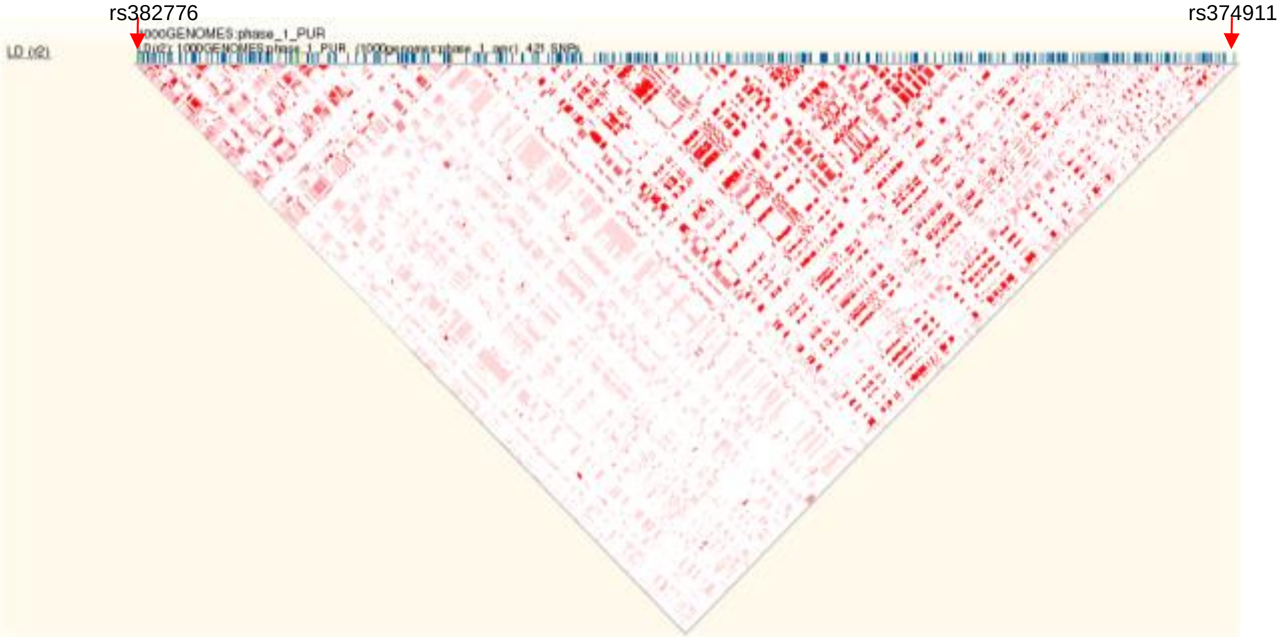
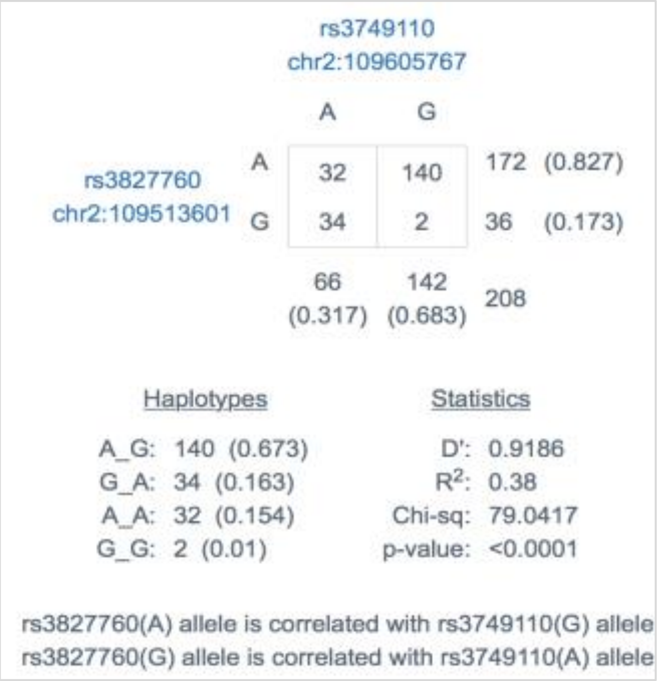


Figure 24. Linkage Disequilibrium (LD) plot of EDAR chr 2:109513601-109605767 for Puerto Rican population 1000 Genomes data. Deep red means high LD; light red means less LD; white means no LD. Indicated by the arrows are the loci for rs3827760 and rs3749110 respectively.



		rs3749110 chr2:109605767 A	G		
rs3827760 chr2:109513601	A	32	140	172	(0.827)
	G	34	2	36	(0.173)
		66 (0.317)	142 (0.683)	208	

Haplotypes		Statistics	
A_G:	140 (0.673)	D':	0.9186
G_A:	34 (0.163)	R^2:	0.38
A_A:	32 (0.154)	Chi-sq:	79.0417
G_G:	2 (0.01)	p-value:	<0.0001

rs3827760(A) allele is correlated with rs3749110(G) allele
rs3827760(G) allele is correlated with rs3749110(A) allele

Figure 25. **Haplotypes frequencies for rs3827760 and rs3749110, and LD calculations for both SNPs.**

A qualitative analysis between *EDAR* T1540C genotypes and shoveling grade was made by a One-Way ANOVA and TukeyHSD test. With this analysis we confirmed a significant relationship between genotypes and shoveling grade with a p-value of 2.19e-05. Mean shoveling grade for the genotypes was 0.89, 1.87, and 2.75, for AA, AG, and GG respectively. TukeyHSD test revealed that the genotype AA is significantly different from AG and GG, with the latter two genotypes not producing significantly different phenotypes (Table 4).

Table 4. Tukey HSD test for rs3827760 genotypes. 95% family-wise confidence level

Genotypes	diff	lwr	upr	p adj
AG-AA	0.9778	0.3566	1.5989	0.001207
GG-AA	1.8611	0.8277	2.8945	0.0002230
GG-AG	0.8833	-0.2021	1.9688	0.1305

Abbreviations: diff, difference in the observed means; lwr, lower end point of the interval; upr, upper end point of the interval; p adj, p-value after adjustment for the multiple comparisons.

For the SNP rs3749110 we also conducted a qualitative analysis between genotypes and shovel-shaped incisors phenotype. There is a significant relationship between genotypes and phenotype with a p-value of 0.00131. Mean shoveling grade for the genotypes was 0.81, 1.40, and 2.20, for GG, GA, and AA respectively. TukeyHSD test revealed that the genotype AA is significantly different from GG, but genotype AG was not shown to be different to genotype AA or GG (Table 5).

Table 5. Tukey HSD test for rs3749110 genotypes. 95% family-wise confidence level

Genotypes	diff	lwr	upr	p adj
GA-GG	0.5875	-0.1240	1.2990	0.1233
AA-GG	1.3875	0.5324	2.2426	0.0008481
AA-GA	0.8000	-0.02157	1.6216	0.05776

Abbreviations: diff, difference in the observed means; lwr, lower end point of the interval; upr, upper end point of the interval; p adj, p-value after adjustment for the multiple comparisons.

Discussion

Determination of the *EDAR* allele 1540 C frequency among Puerto Ricans

Puerto Rican ancestry deconvolution data from 1000 Genome Project phase 1 was analyzed in the *EDAR* region. Out of 110 Puerto Rican chromosomes we found 25 Native American chromosomes, of which 20 of these had the 1540C allele. Therefore, we assumed that only 80% of the Taíno chromosomes had the allele. Of the non-Native American chromosomes only 1 had the allele 1540C. In populations outside Asia and Native Americans, the *EDAR* allele 1540C frequency is below 1%. We deduced that if the allelic frequency for this SNP is 21% in Puerto Ricans based on data from 1000 Genomes, and 80% of the Native American chromosomes had the SNP, then the *EDAR* region should be 26% Native American, while overall Native American ancestry for these samples is 12.8%. These observations showed us that the region around allele 1540C has a higher percentage of Native American ancestry than the overall genome, possibly due to selection of the allele in Puerto Rico after admixture.

After genotyping rs3827760 in the representative Puerto Rican samples, the allelic frequency of 1540C was calculated at 13%. Overall Native American ancestry for these samples is 15.2% (Via et al., 2011). If the allelic frequency for 1540C is 13% and assuming only 80% of taínos carried the allele, then the Native American ancestry for the *EDAR* region in the representative samples should be 16%. If the overall Native American ancestry for the representative Puerto Rican samples is 15.2% (Via et al., 2011) and in the *EDAR* region is 16%, there is not a significant difference to suggest selection in the admixed population for the 1540C allele.

We conducted a chi-square test for the 1540C allelic frequency calculated with the Puerto Rican samples representative of the population. The result was 0.3416, for a critical value of 3.84 at 5% probability, meaning the Puerto Rican population is in Hardy-Weinberg equilibrium for this allele. For a significance level of 0.05 the p-value for this test was 0.84299. We conducted a chi-square test to evaluate if the direct comparison of 1000 genomes data and our samples is significant. The result for this test was 60.98 with a p-value < 0.00001 meaning the direct comparison has a statistically significant difference. Using 1000 genomes phase 1 data we could only observe 110 chromosomes, enough to make assumptions but also too small to be certain.

Identification of one or more SNPs in the genes involved in EDAR pathway, also responsible for shovel-shaped incisors

In a study with the Japanese population, *EDAR* 1540C was associated with shovel-shaped incisors (Kimura et al., 2009). The *EDAR* allele was shown to increase the shoveling grade by a factor of 0.7 per C allele and to be responsible for 18.9% of the variance (Kimura et al., 2009). These results indicate that there is the possibility that other alleles, even in other genes, may be also responsible for the shovel-shaped incisors phenotype. EDAR is a member of the ectodysplasin pathway, which has a major role in the development of ectodermal organs such as hair, teeth, nails, and sweat glands. Therefore, if other alleles might be involved in shovel-shaped incisors it would be wise to search for them in the ectodysplasin pathway. We selected 11 SNPs total among the genes *EDAR*, *EDA, XEDAR*, and *EDARADD*, from which only 6 were able to be genotyped (Table 2).

Of the 6 genotyped alleles, 2 were found to have an additive effect in the shoveling grade (Figures 18 and 19), and both corresponded to *EDAR*. As expected, regression analysis showed that the previously described *EDAR* 1540C was strongly correlated to the shovel-shaped incisors phenotype with a p-value of 2.19e-05. The mean shoveling grade per genotype was 0.89, 1.87, and 2.75, for AA, AG, and GG respectively. TukeyHSD test revealed that a significant difference is found between the genotype AA and AG or GG (Table 4). Regression analysis indicates that the allele increases the shoveling grade by 0.9 and that, according to the determination coefficient, it is responsible for 39% of the variance (Table 3). In comparison, among the Japanese population the *EDAR* allele 1540C increased the shoveling grade by 0.7 and was responsible for only 18.9% of the variance (Kimura et al., 2009). These differences can be explained by our admixture history. There is strong evidence for positive selection for the 1540C allele in East Asia. It is thus possible that because of selection, the Japanese population carries other alleles that have strong influence in the shovel-shaped incisor phenotype. But, not only our Taíno ancestors are a result from a bottleneck effect, Puerto Ricans are an admixed population having a higher proportion of their genome originating from Europeans and Africans. Therefore, it is possible that the phenotypic signal produced by the *EDAR* allele 1540C in Puerto Ricans can be more easily detected because of the lower frequency of other alleles with accumulative and partially redundant, albeit smaller, effects. This also explains why among Taíno skulls, the most abundant shoveling grades were 3 or higher, but among our samples the grades 3 or higher were rarely seen (Crespo, 1994). Our theory could explain why our results show a higher percent of variance and a higher additive score.

Another SNP, rs3749110, located in the first exon of *EDAR*, showed a significant correlation with the shovel-shaped incisors phenotype. Nevertheless, D’ was 0.9186 between rs3749110 and rs3827760 in the Puerto Rican population, which indicates the two SNPs are coinherited 92% of the time. They are in strong linkage disequilibrium. R^2 was 0.38, which is low for a disequilibrium but it is explained by the presence of a rare allele. Haplotypes analysis shows there is one allele combination almost absent, G_G with a frequency of 0.01 due to the low frequency of the rare allele. Therefore, rs3749110 and rs3827760 are in linkage disequilibrium. The correlation between rs3749110 and shovel-shaped incisors phenotype could be due to the effect of the *EDAR* 1540C allele and not rs3749110 by itself. This analysis between SNPs was not taken into account prior to selection of the SNPs for genotyping because the SNPs were chosen from different networks. The very strong LD found is evidence for how recent is admixture in Puerto Ricans. A correlation between shovel-shaped incisors phenotype and the rest of the genotyped SNPs was not found (Table 3).

Conclusion

The objectives of this project were to find if there are other SNPs in the ectodysplasin pathway that are responsible for the shovel-shaped incisors phenotype, and to find if the *EDAR* allele 1540C has undergone positive selection in Puerto Rican population. From a SNPs selection among *EDAR, EDA, XEDAR* and *EDARADD*, we only found a correlation between two linked SNPs, rs3827760 and rs3749110 and shovel-shaped incisors. The *EDAR* allele 1540C explains 39% of the variance for the shovel-shaped incisors, and produces an additive effect in the shoveling grade of 0.9. We believe that admixture is responsible for having the higher phenotypic effect of the *EDAR* allele 1540C for shovel-shaped incisors in Puerto Ricans than in the Japanese population. In conclusion, rs3827760 has been confirmed to be associated with shovel-shaped incisors phenotype. In Puerto Ricans, shovel-shaped incisors phenotype is influenced by *EDAR* allele 1540C more strongly that in the Japanese population. In addition, we found no evidence of selection for the rs3827760 allele among Puerto Ricans. We strongly encourage future studies among genes associated with dental morphology that could reveal the presence of other SNPs related to the shovel-shaped incisors phenotype.

Recommendations

Increasing the sample size will add more power to our tests and validity to our findings. A power test for a significance level of 0.05 and with a power of 0.8 determined sample size should be a minimum of 49. We also recommend to construct networks for genes that are not part of EDA pathway but that have been associated with dental morphology. These changes could reveal novel SNPs that are also associated with the variance missing for shovel-shaped incisors phenotype.

Bibliography

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., … McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*(7422), 56–65. http://doi.org/10.1038/nature11632

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., … Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–9. http://doi.org/10.1038/nmeth0410-248

Bandelt H-J, Foster P, R. A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, *16*, 37–48.

Bayes, M., Hartung, A. J., Ezer, S., Pispa, J., Thesleff, I., Srivastava, A. K., & Kere, J. (1998). The Anhidrotic Ectodermal Dysplasia Gene (EDA) Undergoes Alternative Splicing and Encodes Ectodysplasin-A with Deletion Mutations in Collagenous Repeats. *Human Molecular Genetics*, *7*(11), 1661–1669. http://doi.org/10.1093/hmg/7.11.1661

Carlson, C. S., Thomas, D. J., Eberle, M. A., Swanson, J. E., Livingston, R. J., Rieder, M. J., & Nickerson, D. A. (2005). Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Research*, *15*(11), 1553–65. http://doi.org/10.1101/gr.4326505

Clement, A. F., Hillson, S. W., & Aiello, L. C. (2012). Tooth wear, Neanderthal facial morphology and the anterior dental loading hypothesis. *Journal of Human Evolution*, *62*(3), 367–76. http://doi.org/10.1016/j.jhevol.2011.11.014

Cluzeau, C., Hadj-Rabia, S., Jambou, M., Mansour, S., Guigue, P., Masmoudi, S., … Smahi, A. (2011). Only four genes (EDA1, EDAR, EDARADD, and WNT10A) account for 90% of hypohidrotic/anhidrotic ectodermal dysplasia cases. *Human Mutation*, *32*(1), 70–2.

http://doi.org/10.1002/humu.21384

Crespo, E. F. (1994). *Dental Analysis of Human Burials Recovered From Punta Candelero: A Prehistoric Site on the Southeast Coast of Puerto Rico*. Arizona State University.

Deshmukh, S., & Prashanth, S. (2012). Ectodermal dysplasia: a genetic review. *International Journal of Clinical Pediatric Dentistry*, *5*(3), 197–202. http://doi.org/10.5005/jp-journals-10005-1165

Fujimoto, A., Kimura, R., Ohashi, J., Omi, K., Yuliwulandari, R., Batubara, L., … Tokunaga, K. (2008). A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Human Molecular Genetics*, *17*(6), 835–843. http://doi.org/10.1093/hmg/ddm355

Gardiner-Garden, M., & Frommer, M. (1987). CpG Islands in vertebrate genomes. *Journal of Molecular Biology*, *196*(2), 261–282. http://doi.org/10.1016/0022-2836(87)90689-9

Häärä, O., Fujimori, S., Schmidt-Ullrich, R., Hartmann, C., Thesleff, I., & Mikkola, M. L. (2011). Ectodysplasin and Wnt pathways are required for salivary gland branching morphogenesis. *Development*, *138*(13).

Häärä, O., Harjunmaa, E., Lindfors, P. H., Huh, S.-H., Fliniaux, I., Åberg, T., … Thesleff, I. (2012). Ectodysplasin regulates activator-inhibitor balance in murine tooth development through Fgf20 signaling. *Development (Cambridge, England)*, *139*(17), 3189–99. http://doi.org/10.1242/dev.079558

Headon, D. J., Emmal, S. A., Ferguson, B. M., Tucker, A. S., Justice, M. J., Sharpe, P. T., … Overbeek, P. A. (2001). Gene defect in ectodermal dysplasia implicates a death domain adapter in development. *Nature*, *414*(6866), 913–916. http://doi.org/10.1038/414913a

HomoloGene - NCBI. (n.d.). Retrieved May 17, 2016, from https://www.ncbi.nlm.nih.gov/homologene/68180

Hymowitz, S. G., Compaan, D. M., Yan, M., Wallweber, H. J. A., Dixit, V. M., Starovasnik, M. A., & de Vos, A. M. (2003). The crystal structures of EDA-A1 and EDA-A2: splice variants with distinct receptor specificity. *Structure (London, England : 1993)*, *11*(12), 1513–20. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14656435

Irby, R. B., & Yeatman, T. J. (2000). Role of Src expression and activation in human cancer. *Oncogene*, *19*(49), 5636–5642. http://doi.org/10.1038/sj.onc.1203912

Kamberov, Y. G., Wang, S., Tan, J., Gerbault, P., Wark, A., Tan, L., … Sabeti, P. C. (2013). Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell*, *152*(4), 691–702. http://doi.org/10.1016/j.cell.2013.01.016

Kimura, R., Fujimoto, A., Tokunaga, K., & Ohashi, J. (2007). A practical genome scan for population-specific strong selective sweeps that have reached fixation. *PloS One*, *2*(3), e286. http://doi.org/10.1371/journal.pone.0000286

Kimura, R., Yamaguchi, T., Takeda, M., Kondo, O., Toma, T., Haneji, K., … Oota, H. (2009). A Common Variation in EDAR Is a Genetic Determinant of Shovel-Shaped Incisors. *American Journal of Human Genetics*, *85*(4), 528–535. http://doi.org/10.1016/j.ajhg.2009.09.006

Machiela, M. J., & Chanock, S. J. (2015). LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics (Oxford, England)*, *31*(21), 3555–7. http://doi.org/10.1093/bioinformatics/btv402

Meng, Y., & Roux, B. (2016). Computational study of the W260A activating mutant of Src tyrosine kinase. *Protein Science : A Publication of the Protein Society*, *25*(1), 219–30.

http://doi.org/10.1002/pro.2731

Mizoguchi, Y. (1985). *Shovelling, a Statistical Analysis of Its Morphology*.

Mou, C., Thomason, H. A., Willan, P. M., Clowes, C., Harris, W. E., Drew, C. F., … Headon, D. J. (2008). Enhanced ectodysplasin-A receptor (EDAR) signaling alters multiple fiber characteristics to produce the East Asian hair form. *Human Mutation*, *29*(12), 1405–1411. http://doi.org/10.1002/humu.20795

Mustonen, T., Pispa, J., Mikkola, M. L., Pummila, M., Kangas, A. T., Pakkasjärvi, L., … Thesleff, I. (2003). Stimulation of ectodermal organ development by Ectodysplasin-A1. *Developmental Biology*, *259*(1), 123–36. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12812793

Newton, K., French, D. M., Yan, M., Frantz, G. D., & Dixit, V. M. (2004). Myodegeneration in EDA-A2 transgenic mice is prevented by XEDAR deficiency. *Molecular and Cellular Biology*, *24*(4), 1608–13. http://doi.org/10.1128/mcb.24.4.1608-1613.2004

O ’geen, H., Echipare, L., & Farnham, P. J. (2011). Using ChIP-Seq Technology to Generate High-Resolution Profiles of Histone Modifications. *Methods Molecular Biology*, *791*, 265–286. http://doi.org/10.1007/978-1-61779-316-5_20

Oleksyk, T. K., Smith, M. W., & O’Brien, S. J. (2010). Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *365*(1537), 185–205. http://doi.org/10.1098/rstb.2009.0219

Pantalacci, S., Chaumot, A., Benoît, G., Sadier, A., Delsuc, F., Douzery, E. J. P., & Laudet, V. (2008a). Conserved features and evolutionary shifts of the eda signaling pathway involved in vertebrate skin appendage development. *Molecular Biology and Evolution*, *25*(5), 912–928. http://doi.org/10.1093/molbev/msn038

Pantalacci, S., Chaumot, A., Benoît, G., Sadier, A., Delsuc, F., Douzery, E. J. P., & Laudet, V. (2008b). Conserved features and evolutionary shifts of the EDA signaling pathway involved in vertebrate skin appendage development. *Molecular Biology and Evolution*, *25*(5), 912–28. http://doi.org/10.1093/molbev/msn038

Punj, V., Matta, H., & Chaudhary, P. M. (2010). X-linked Ectodermal Dysplasia Receptor (XEDAR) is Down-regulated in Breast Cancer via Promoter Methylation. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *16*(4), 1140. http://doi.org/10.1158/1078-0432.ccr-09-2463

Sadier, A., Lambert, E., Chevret, P., Décimo, D., Sémon, M., Tohmé, M., … Laudet, V. (2015). Tinkering signaling pathways by gain and loss of protein isoforms: the case of the EDA pathway regulator EDARADD. *BMC Evolutionary Biology*, *15*, 129. http://doi.org/10.1186/s12862-015-0395-0

Sadier, A., Viriot, L., Pantalacci, S., & Laudet, V. (2014). The ectodysplasin pathway: From diseases to adaptations. *Trends in Genetics*, *30*(1), 24–31. http://doi.org/10.1016/j.tig.2013.08.006

Sinha, S. K., Zachariah, S., Quiñones, H. I., Shindo, M., & Chaudhary, P. M. (2002). Role of TRAF3 and -6 in the activation of the NF-kappa B and JNK pathways by X-linked ectodermal dysplasia receptor. *The Journal of Biological Chemistry*, *277*(47), 44953–61. http://doi.org/10.1074/jbc.M207923200

Stephens, M., & Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics*, *76*(3), 449–62. http://doi.org/10.1086/428594

Stephens, M., Smith, N. J., & Donnelly, P. (2001). A New Statistical Method for Haplotype

Reconstruction from Population Data. *The American Journal of Human Genetics*, *68*(4), 978–989. http://doi.org/10.1086/319501

Tanikawa, C., Ri, C., Kumar, V., Nakamura, Y., & Matsuda, K. (2010). Crosstalk of EDA-A2/XEDAR in the p53 Signaling Pathway. *Molecular Cancer Research*, *8*(6).

Via, M., Gignoux, C. R., Roth, L. A., Fejerman, L., Galanter, J., Choudhry, S., … Martínez-Cruzado, J. C. (2011). History shaped the geographic distribution of genomic admixture on the island of Puerto Rico. *PloS One*, *6*(1), e16513. http://doi.org/10.1371/journal.pone.0016513

Yan, M. (2000). Two-Amino Acid Molecular Switch in an Epithelial Morphogen That Regulates Binding to Two Distinct Receptors. *Science*, *290*(5491), 523–527. http://doi.org/10.1126/science.290.5491.523

Appendix

Table 1. List of samples collected from Puerto Rico. *n*=46

Sample Code	Sample Name[a]	Age	Municipality of residence	UI1 right[b]	Genotype[c]
AR141118-01	AR101	29	Arecibo	2	AG
AR150119-01	AR102	15	San Juan	2	GG
AR150120-01	AR103	34	Arecibo	1	AG
AR150120-02	AR104	16	Arecibo	2	AA
AR150224-01	AR105	15	Arecibo	2	AG
AG141009-01	AG101	19	Aguadilla	1	GG
AG141016-01	AG102	25	Aguadilla	1	AG
AG141212-01	AG103	40	Yauco	1	GG
AG160307-01	AG104	15	Aguadilla	1	GG
AG160317-01	AG105	23	Mayagüez	1	GG
AG160328-01	AG106	17	Aguada	1	GG
BY160712-01	BY101	16	Vega Alta	0	AG
BY160712-02	BY102	21	Bayamón	0	AG
BY160712-03	BY103	21	Vega Alta	0	GG
BY160713-01	BY104	33	Bayamón	0	AG
BY160714-01	BY105	35	Bayamón	1	GG
BY160716-01	BY106	21	Vega Baja	1	AG
FA140716-01	FA102	33	Luquillo	1	AG
FA140805-01	FA103	36	San Juan	2	AG
FA140805-02	FA104	21	Luquillo	2	AG
FA140805-03	FA105	18	Fajardo	1	AA
MY150311-01	MA101	22	Mayagüez	0	GG
MY150311-02	MA102	28	Guánica	1	GG
MY150318-01	MA103	18	Mayagüez	2	AA
MY150331-01	MA104	19	Hormigueros	1	GG
MY150709-01	MA105	22	Aguadilla	2	AG
MY150709-02	MA106	31	Isabela	1	GG
MN150220-01	MN101	22	Florida	0	GG
MN150318-01	MN106	35	Vega Baja	2	AA
MN150321-01	MN107	33	Barranquitas	2	AG
PO150715-01	PO102	20	Juana Díaz	2	AA
PO150717-01	PO104	19	Juana Díaz	3	AA
PO150717-02	PO105	17	Ponce	2	AA

RG150615-02	RG103	38	Río Grande	1	GG
RG150615-03	RG104	43	Canóvanas	3	AG
RG150618-02	RG105	24	Carolina	3	AG
RG150618-03	RG106	14	Luquillo	2	AA
RG150630-01	RG107	15	Río Grande	5	AA
RG150630-02	RG108	28	Canóvanas	1	AG
RG150805-01	RG109	24	Río Grande	0	AG
RG150615-04	RG110	16	Fajardo	1	GG
RG150716-01	RG111	20	San Juan	2	AG
RG150616-02	RG112	20	Luquillo	0	GG
RG150618-01	RG113	22	Río Grande	1	AG
YA150619-01	YA101	21	Yauco	2	AG
YA150619-02	YA102	27	Yauco	1	AA

[a] Sample name used for sample storage

[b] Shoveling grade for the upper incisor 1 right side

[c] Genotype for rs3827760