Fully automatic landmarking of 2D photographs identifies novel genetic loci influencing facial features

Qing Li
Jieyi Chen
Pierre Faux
Betty Bonfante
Macarena Fuentes-Guajardo
Javier Mendoza-Revilla
Juan Chacón-Duque
Malena Hurtado
Valeria Villegas
Vanessa Granja
Claudia Jaramillo
William Arias
Rodrigo Barquera
Max Planck Institute for the Science of Human History, Kahlaische Strasse 10, 07745 Jena https://orcid.org/0000-0003-0518-4518
Paola Everardo-Martínez
Autonomous University of Mexico City
Mirsha Sánchez-Quinto
Jorge Gómez-Valdés
Hugo Villamil-Ramírez
Caio C. Silva de Cerqueira
Tábita Hünemeier
Virginia Ramallo
Sijie Wu
Siyuan Du
CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences https://orcid.org/0000-0003-1602-1669
Rolando Gonzalez-José
Lavinia Schüler-Faccini
Maria-Cátira Bortolini
Victor Acuña-Alonzo
Samuel Canizales-Quinteros
Carla Gallo
Giovanni Poletti
Winston Rojas
Francisco Rothhammer
Nicolas Navarro
Biogéosciences, UMR 6282 CNRS, EPHE, Université Bourgogne Franche-Comté, Dijon 21078, France. https://orcid.org/0000-0001-5694-4201
Sijia Wang
CAS Key Laboratory of Computational Biology, Shanghai Institutes of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences https://orcid.org/0000-0001-6961-7867
Kaustubh Adhikari
Andrés Ruiz-Linares (✉ **andresruiz@fudan.edu.cn**)
Fudan University

Article

Keywords:

Posted Date: March 30th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1479643/v1

Abstract

We report a genome-wide association study for facial features in > 6,000 Latin Americans. We placed 106 landmarks on 2D frontal photographs using the cloud service platform Face++. After Procrustes superposition, genome-wide association testing was performed for 301 inter-landmark distances. We detected nominally significant association (P-value < 5×10^{-8}) for 42 genome regions. Of these, 9 regions have been previously reported in GWAS of facial features. In follow-up analyses, we replicated 26 of the 33 novel regions (in East Asians or Europeans). The replicated regions include 1q32.3, 3q21.1, 8p11.21, 10p11.1, and 22q12.1, all comprising strong candidate genes involved in craniofacial development. Furthermore, the 1q32.3 region shows evidence of introgression from archaic humans. These results provide novel biological insights into facial variation and establish that automatic landmarking of standard 2D photographs is a simple and informative approach for the genetic analysis of facial variation, suitable for the rapid analysis of large population samples.

One Sentence Lay Summary

We used a fully automatic landmarking approach on frontal 2D face photographs of >6,000 Latin Americans and conducted a genome-wide association study, identifying novel genomic regions influencing facial features.

Introduction

Recent years have seen a flurry of genome-wide association studies seeking to identify loci impacting on facial appearance in the general human population[1–10]. These studies have reported significant associations for over a hundred loci with various facial features, including morphology and facial hair. These studies are providing insights into the genetic architecture of human facial appearance, including the overlap between common variation in the general population and rarer alterations in craniofacial development[1, 2, 10].

GWASs of facial shape variation have used a range of phenotyping methods, from qualitative assessment of morphological features on 2D photographs, to measurements based on landmarking of 2D photographs, to semi-automatic analyses of 3D facial images. These approaches vary greatly in their informativity and ease of application. Although 3D images fully represent facial morphology, acquisition of such data requires specialized equipment, complicating their widespread application. Although less informative than 3D imaging, plain 2D photographs can facilitate the collection of large study samples, as they only require conventional photographic equipment. However, manual landmarking of 2D photographs is a very labor-intensive task. This has fostered an interest in the development of automatic landmarking approaches, but so far these have enjoyed limited success[11–13]. Most studies based on 2D photographs have therefore been based on entirely manual[1] or, at times, semi-automatic landmarking[14] (i.e. combining some automatic landmarking with manual editing).

We have previously reported the identification of genetic loci impacting on facial features in a sample of > 6,000 Latin Americans of mixed Native, European, and African ancestry. Initially, we conducted a GWAS based on a qualitative assessment of facial features on 2D photographs[15, 16] and subsequently identified additional associated loci based on measurements derived from manual landmarking of the profile photographs[1]. Here report a GWAS of measures derived from fully automatic landmarking of the frontal 2D photographs. The signals detected here show a large overlap with previous GWAS findings, providing confidence in the automatic landmarking approach we used. In addition, we also identify dozens of novel associations with facial traits. For the majority of these novel loci, we find evidence of statistical replication in European or East Asian GWAS data. The replicated loci we detected include genes implicated in human dysmorphology or associated with craniofacial phenotypes in animal models.

Results And Discussion

Study sample and phenotyping

The individuals examined here are part of the CANDELA cohort, collected in five Latin American countries[17], which has been previously examined in several GWASs of physical appearance[1, 15, 16, 18, 19]. We have reported two facial morphology GWASs based on 2D photographs: one based on categorical phenotyping and one based on measurements derived from manual landmarking of the profile photographs[1, 16, 18]. Individuals included in these studies were genotyped on Illumina's OmniExpress chip (including >700,000 SNPs) and characterized for a set of standard covariates (age, sex, BMI, and genetic ancestry estimated from the chip data)[1, 16, 18].

Here we used the Face++ cloud service platform (https://www.faceplusplus.com) to automatically place 106 landmarks on the frontal 2D photographs from CANDELA individuals (Supplementary Figure S1). We examined the robustness of 16 of these landmarks, which had previously been placed manually on a subset of the photographs (Supplementary Figure S2)[16]. ICCs between Face++ and manually placed landmarks were >0.90 (Supplementary Table S1) and the median Euclidean distance between these two sets of landmarks was <25 pixels (relative to an image size of 2,136 × 3,126 pixels, Supplementary Table S1).

After Procrustes superposition, we calculated inter-landmark distances (ILDs) for 34 Face++ landmarks, most of which corresponding to well-defined anatomical landmarks (Figure 1A, Supplementary Table S2). Accounting for face symmetry, we obtained 301 distances. Some of the Face++ landmarks are placed on the eyebrow edges, and distances based on these landmarks reflect eyebrow size, rather than facial morphology (Figure 1B).

Most of the inter-landmark distances obtained show considerable variation and are approximately normally distributed in the CANDELA sample (Supplementary Figure S3). Many distances showed a significant correlation with three head angles estimated by Face++ (pitch, roll, and yaw angle), reflecting the effect of head pose (Supplementary Table S3). We therefore included these head angles as covariates in the genetic association tests.

Trait/covariate correlation and heritability

A low to moderate (significant) correlation was detected for certain inter-landmark distances with sex, BMI or genetic ancestry. Considering the large number of distances examined, here we refer only to the most significant correlations (full results are presented in Supplementary Table S3). Strongest correlation with sex was seen for the distance between right frontotemporale and right nasal root contour ($r = 0.62$, $p < 0.0001$), the distance between right frontotemporale and right palpebrale inferious ($r = 0.59$, $p < 0.0001$), and the distance between nasion and right eyebrow upper left corner ($r = 0.58$, $p < 0.0001$) (Supplementary Table S3, Supplementary Figure S4). All of these distances are greater in female than in male, probably reflecting eyebrow shaping in women. The strongest correlation with age was seen for the distance between frontozygomathic suture and palpebrale superious ($r = -0.36$, $p < 0.0001$). Lip thickness measures also show a strong negative correlation with age ($r = -0.30$, $p < 0.0001$, Supplementary Table S3), consistent with our previous analyses of the CANDELA sample[1, 16].

Strongest correlation ($r = -0.44$, $p < 0.0001$) with BMI was seen for distances related to right frontotemporale (Supplementary Table S3).

Strong correlation with European/Native American ancestry was seen for nasion position ($r = -0.25$, $p < 0.0001$), lip thickness ($r = -0.19$, $p < 0.0001$), nasal root breadth ($r = -0.23$, $p < 0.0001$) and nose wing breadth ($r = -0.21$, $p < 0.0001$) (Supplementary Table S3, Supplementary Figure S4). These correlations agree with those we calculated previously in the CANDELA sample through manual phenotyping[1, 16]. Our previous finding of a correlation between European/Native American ancestry and nose and chin protrusion could not be properly examined here, as the distances extracted from the automatic landmarking of frontal photographs have little connection to facial protrusion features.

We used LDAK5[20] to estimate narrow-sense heritability (h^2) for all used facial traits based on the kindship matrix derived from SNP data. Except for a small number of low heritability values, we obtained moderate values for most of the traits. The highest value (0.56) was observed for the distance between the right alare and eyebrow upper corner (Supplementary Table S3).

Overview of GWAS results and integration with the literature

After quality control filters, we evaluated association for 301 inter-landmark distances with up to 11,532,785 SNPs and 5,988 individuals. Associations that were nominal genome-wide significant (P-value $< 5\times10^{-8}$) and survived the multiple testing correction using the Benjamini-Hochberg procedure for the false-discovery rate (FDR) were studied. A total of 42 genomic regions showed significant association with at least one distance, and 148 distances show significant association with at least one of these genomic regions (Figure 1C, Supplementary Table S4). Among these 42 regions, 9 have been previously reported in other GWASs of facial features. Furthermore, most of these regions have strong candidate genes for which various lines of evidence point to them having an important role in craniofacial development. Table 1 provides key information on these 9 regions (association plots for them are shown in Supplementary Note 1).

The most robustly replicated region involves SNPs in strong LD over a segment of >10Mb in 2q12.3 centered around the *EDAR* (Ectodysplasin A Receptor) gene (Supplementary Note 1). SNPs in this region have been previously associated with facial morphology in Latin Americans and East Asians, particularly in relation to facial flatness, jaw protrusion, and distance between ectocanthion and otobasion inferius[1, 5, 16]. Overall, we observe that 1,245 SNPs in 2q12.3 are associated with 75 inter-landmark distances. The strongest association was detected for D198 (involving landmarks 9-23, P-value 5.6×10^{-34}, Figure 2). The associated SNPs include rs3827760, which encodes a functional amino acid substitution in *EDAR*, the derived allele being absent in Europeans, but having a high frequency in East Asians and being fixed in Native Americans (resulting in a frequency in the CANDELA sample of 0.41). Most of the distances showing association with the *EDAR* region are strongly correlated, and involve landmarks around the eye and eyebrow (Figure 2). We evaluated statistical interaction between *EDAR* and other associated regions but found no significance for any of them (Supplementary Table S5). A previous facial hair GWAS in CANDELA showed a suggestive association between *EDAR* and eyebrow thickness[15], thus some of the associations seen here could be influenced by variation in eyebrow size. In addition, some associated distances involve landmarks around the nose or mouth, consistent with a previous study in Eurasians[5]. Overall, the derived allele (G) of rs3827760 is associated with an increase of eyebrow and eye size, and a smaller distance between the eyes.

Associated SNPs in 7p14.1 overlap the *GLI3* (GLI Family Zinc Finger 3) (Supplementary Note 1). We previously reported an association of SNPs in this region with nose wing breadth assessed qualitatively and through manual landmarking[16]. Consistent with our previous findings, the automatic landmarking performed here revealed an association of SNPs in this region with the distance between right and left alare, and the distance between left and right nasal root contour. The minor allele of the index SNP (rs846315), an intron variant, is associated with a wider nose. *GLI3* has been shown to play a key role in embryogenesis[21], mutations in *GLI3* causing Greig cephalopolysyndactyly and Pallister-Hall syndromes[22, 23].

PAX3 (Paired box 3) (2q36.1, $P=2.04\times10^{-10}$), is the most frequently replicated locus across studies of facial morphology, so far having been associated in eight independent GWAS[1, 2, 4, 5, 8, 14, 24, 25], always in relation to measures sensitive to nasion position. We previously detected a *PAX3* association in GWAS of the CANDELA 2D photographs, using both categorical phenotyping and measures derived from manual landmarking[1, 16]. In addition, SNPs in *PAX3* have been associated with the distance between eyes, and with brow ridge protrusion[1, 14]. Here we observe association of SNPs in this region with 18 distances, mostly involving the nasion and eyebrow area. In addition, some associated distances are sensitive to nose height (Figure 2). The 18 distances associated with *PAX3* are strongly correlated. The strongest association is seen for rs13022712 with D99 (landmarks 9-16), and this index SNP is located within an intron of *PAX3*. The minor allele of the index SNP in this region (rs13022712) is associated with a shortening of the nose/mid-face. *PAX3* has been shown to play a key role in fetal development, particularly in relation to neural development and myogenesis[26].

SNPs in 5q13.2 located 182 Kb away from *FOXD1* (forkhead box D1) show association with 10 ILDs in the eye/eyebrow area (Supplementary Note 1). Six of these distances are sensitive to eyebrow size (thickness and length), with the other four being sensitive to either eye size, or spacing distance between eye and eyebrows. *FOXD1* has been associated with eyebrow thickness in a GWAS in a Chinese sample[27] and we previously observed suggestive association with the same trait evaluated qualitatively in the CANDELA sample[15, 27]. *FOXD1* has been shown to be involved in hair growth[28]. The minor allele (C) of the index SNP in this region (rs7341037) is associated with larger eyebrows.

We previously reported association of SNP in 6p21.1 overlapping *RUNX2* (RUNX Family Transcription Factor 2) and *SUPT3H* (SPT3 Homolog) with nose bridge breadth[16], forehead protrusion, brow ridge protrusion and upper face flatness[1]. Subsequently, GWAS in other study samples have reported association of SNPs in this region with nose morphology, chin dimples, and distance between right entocanthion and left otobasion inferius[5, 8, 25]. Here we observe association of *RUNX2*/ *SUPT3H* with 15 distances, involving landmarks on the eyebrows and mouth, thus generally sensitive to variation in midface height. The strongest association is seen for rs141680515 with D99 (landmarks 9-16), and rs141680515 is intronic to *SUPT3H*. Previous studies have shown an evolutionary effect of *RUNX2* variation on facial length in carnivores[29]. Furthermore, *RUNX2* is known to be involved in osteoblastic differentiation and skeletal morphogenesis[30-32], and *RUNX2*/ *SUPT3H* has been associated with Cleidocranial Dysplasia and familial Hepatic Adenomas[33, 34].

Based on the manual landmarking of CANDELA profile photographs, we previously reported an association signal in 15q21.1, overlapping the *SLC24A5* (Solute Carrier Family 24 Member 5) gene, with nose tip roundness (and suggestive significantly associated with columella inclination and lower lip thickness)[1]. The automatic landmarking of the frontal CANDELA photographs performed here revealed an association of this region with upper lip thickness (D527: landmarks 31-33). The index SNP (rs1426654) showed the strongest association in this and in our previous study. This SNP encodes a nonsynonymous amino-acid change in exon3 of *SLC24A5*. The alternative allele is absent in Europeans, but has an extremely high frequency in East Asians (0.97), Native Americans (0.98), and Africans (0.99), which results in an intermediate frequency in the CANDELA sample of 0.43. This index SNP has a positive effect on both ILDs, which would cause lip thickness increases.

SNPs in 4q31.3 in the vicinity (76 Kb away) of the *SFRP2* (secreted frizzled related protein 2) and (234 Kb away) *DCHS2* (dachsous cadherin-related 2) genes (Supplementary note 1), have been reported to be associated with endocanthion-alare distance[4], columella inclination[16] and nose morphology[8]. Here we find association of this region with 8 distances, including endocanthion-alare (D205: landmarks 16-23). We also find a strong association of SNPs in this region with nose morphology. Furthermore, two previously reported index SNPs (rs6535972 and rs9995821), show genome-wide significant association with at least one distance obtained here with Face++ landmarks. The minor (A) allele of the SNP (rs2045323) with the smallest P-value obtained here (P-value 9.35×10^{-10}), is associated with a narrower and longer nose (mainly resulting from a lower position of the alare landmark, Figure 2). *SFRP2* has been shown to have important roles in craniofacial development in mice[35]. Furthermore, cartilage defects have been observed in zebrafish embryos deficient in *Dchs2*[36].

A region of 5p12 about 190 Kb away from *FGF10* (Fibroblast Growth Factor 10) (Supplementary note 1), has been reported to be associated with variation in a temporal facial segment, using a 3D multivariate approach[2]. Here, we detected an association of SNPs in this genomic region with three inter-landmark distances (the most significant one is the distance between right alare and left eyebrow lower right corner). Rare mutations in *FGF10* have been shown to cause Lacrimo-auriculo-dento-digital syndrome (LADDS)[37], an autosomal dominant ectodermal dysplasia associated with facial dysmorphology.

Similarly, SNPs in 15q25.2 overlapping *ADAMTSL3* (ADAMTS like 3) have been reported to be associated with variation on a GWAS of 3D of facial variation using a multivariate phenotyping approach[2]. Here we observe association of this region with distances between landmarks 8 and 23, reflecting variation between eye and eyebrow. The alternative allele of the index SNP (rs62027787) is absent in East Asians and Africans, and has a relatively low frequency in Native Americans (0.002) and relatively high frequency in Europeans (0.14), which results in a frequency in CANDELA samples of 0.076. One of the SNPs associated here (rs34047645), encodes a nonsynonymous substitution in exon18 of *ADAMTSL3*. This gene has been shown to impact on the proliferation of hepatocellular carcinoma cells[38].

Follow-up of genomic regions newly associated with facial features: Replication in two human cohorts

Amongst the 42 regions with nominally significant association, 33 have not been previously reported in GWAS of facial features. We evaluated replication for the index SNP (and SNPs with r=0.1) that also showed significant P-value in each of these 33 regions in two independent datasets. On one hand, we examined an East Asian dataset consisting of 5,078 2D frontal facial photographs[27, 39]. These photographs were processed, and association P-values were obtained, following the same protocol for the CANDELA cohort. On the other hand, we examined P-values from a GWAS performed in Europeans for 78 inter-landmark distances[4]. For six regions there were no polymorphic SNPs shared between the CANDELA and replication samples, thus preventing evaluation of replication. We defined the significance threshold for the replication cohorts using the Benjamini-Hochberg procedure for the false-discovery rate (FDR), accounting for 33 regions and two replication cohorts. Altogether, 26 index SNPs show association P-values below the FDR threshold for at least one distance, in either the East Asians (22 in EA), Europeans (21 in EU) or both (17 replicate in both) (Supplementary Table S4). We found all candidate genes of these 33 regions expressed significantly higher in the cranial neural crest cells (CNCCs) than background genes (Supplementary Figure S5). For several of the replicating regions, there is additional evidence pointing to them playing a role in craniofacial morphology/development. Below we highlight five of these regions (comments on the other associated regions are provided in the Supplementary Note 2-3):

1q32.3 SNPs in this region are associated with distances between landmarks: pronasal, exocanthion, endocanthion or palpebrale inferious (strongest association being observed for rs12564392, with the pronasal-exocanthion distance, P-value 2×10^{-8}, Figure 3). Associated SNPs in the region overlap the *ATF3* (Activating Transcription Factor 3) and *FAM71A* (Family with sequence similarity 71member A) genes. Interestingly, evidence for archaic human introgression in this region has been previously reported[40]. We therefore screened a 1Mb window around the association signal for the presence of archaic introgression in CANDELA samples using a hidden-Markov model developed for that purpose[41]. Considering only tracts called at >99% confidence and >10 Kb long, we found that 31.3% of the chromosomes carry a Neanderthal introgressed tract that overlaps the *ATF3*/ *FAM71A* gene region (Figure 4). By crossing these results with continental ancestry tracts, we found that ~80% of the archaic tracts were located on Native American ancestry tracts. As the tract lengths and locations were quite variable from a sampled chromosome to another, we used an archaic admixture mapping scheme to test the association between archaic introgression and phenotypes significantly associated to this region in the initial GWAS. The region was chunked into 140 admixture mapping segments, of which 103 have a MAF greater than 1% and were therefore tested. The Bonferroni-corrected threshold was thus equal to 4.9×10^{-4}. We found significant associations of archaic ancestry with D203 (landmarks 13-23), D166 (landmarks 13-19), D223 (landmarks 13-25), illustrating the distance between pronasal to right endocanthion, right exocanthion and right palpebrale inferious, respectively.

3q21.1 SNPs in this region are associated with 8 ILDs (strongest association for rs820360), with D213, P-value 5.24×10^{-10}, Figure 3). The distances showing association in the CANDELA sample reflect variation mainly in the width of the upper face (Figure 3). SNP rs820360 is located within the *MYLK* (Myosin light chain kinase) gene. A study in mice indicates that *MYLK* is involved in the actomyosin contractility pathway, which drives cellular extrusion and intercalation during palate fusion[42].

8p11.21 SNPs in this region are associated with 7 ILDs (strongest association for rs59547557 with D96, P-value 2.24×10^{-9}, Figure 3). All associated ILDs are sensitive to the position (height) of the right cheilion (Figure 3). The minor allele of rs59547557 (C) is associated with a shorter distance between the right cheilion and landmarks in eye/eyebrow region. This region includes a cluster of disintegrin and metalloproteinase (ADAM) domain genes. One gene in this cluster, *ADAM3A*, has been found to be suggestively associated with non-syndromic cleft lip/palate[43].

10p11.1 SNPs in this region are associated with 8 ILDs (strongest association seen for rs58831446 with D511, P-value 1×10^{-10}, Figure 3). These 8 ILDs all reflect philtrum height (Figure 3). The region of association overlaps *MTRNR2L7* and a cluster of Zinc Finger proteins genes. This cluster includes *ZNF25*, which has been shown to be involved in osteoblast differentiation of human skeletal stem cells[44], a process in which *RUNX2* (discussed above) plays a major role[32, 45, 46].

22q12.1 Association was detected for SNPs in this region (smallest P-value of 1.8×10^{-8} for rs9608473 with D437, Figure 5). This region has been strongly associated with height[47], and suggestively associated with facial variation[48, 49] as well as with cleft lip/palate[50].

Follow-up of genomic regions newly associated with facial features: effects in the mouse

To evaluate the potential effect of the novel facial feature regions detected here in the mouse, we reanalyzed a GWAS of craniofacial shape in 692 outbred mice[51], applying a powerful multivariate mixed model not used in the original study[51]. Out of the 33 regions for which we found significant association in the CANDELA GWAS, 30 could be successfully mapped to the mouse genomic data. A region on mouse chromosome 5 (homologous to human 22q12.1), showed strong association (maximum P-value: 2×10^{-34}), far exceeding a multiple testing-adjusted significance threshold. The index SNP in this region impacts on multiple aspects of mouse craniofacial morphology (Figure 5, Supplementary Movie). This region overlaps the *Ttc28* and *Mn1* genes, which experimental studies have been shown to play an important role in mouse craniofacial development[51].

Genome annotations at associated loci

We used FUMA[52] to examine genome annotations of the 1,865 SNPs significantly associated across the 42 independent genomic regions detected here. Altogether, 10 SNPs are exonic, 636 are intergenic, 1,001 are intron variants, 154 are non-coding transcript variants, 12 variants overlap a 5’ untranslated region, 12 variants overlap a 3’ untranslated region, 24 are upstream gene variants, and 16 are downstream gene variants. Of the 10 coding SNPs, 8 are in the well-established *EDAR* region, but are in different genes (4 in *RGPD4*, 1 in *SULT1C3*, 1 in *GCC2*, 1 in *EDAR*, and 1 in *TMEM87B*). The other two exon variants are in *SUPT3H* and *SLC24A5*, both gene regions having been previously associated with facial features. By contrast, all the significant SNPs in the 33 regions we newly associated with facial features are non-coding. In agreement with previous analyses showing an enrichment of facial features SNPs in regulatory elements active during craniofacial development[4, 8] we observe that these novel SNPs are often near or within craniofacial enhancers/promoters (e.g. 1q32.3 and 12q21.31) (Figure 4, Supplementary Figure S6).

Conclusion

The results presented here demonstrate that fully automated landmarking of 2D photographs, as implemented in Face++, is a reliable approach for the genetic analysis of facial features. For a set of well-defined anatomical landmarks, we find that Face + + landmarking is reliable, compared with manual landmarking. More importantly, we find strong statistical evidence for replication of a large number of previously reported GWAS signals, most of which are also supported by experimental data. Finally, the majority of the novel association signals detected here replicate in independent datasets, including an Asian sample also characterized using Face++.

Given the wide-spread availability of facial photographs, automated landmarking could facilitate future genetic studies based on sample sizes much larger than hitherto examined and comparable to those included in the largest GWAS performed to date (now including over a million individuals)[53–55]. The study of such large samples will enable high-powered analyses of the genetic architecture of facial variation in humans. Importantly, the use of plain 2D photography should also facilitate a more comprehensive world-wide sampling of human facial variation than hereto possible. The importance of studying non-European populations is illustrated by our analysis of Latin Americans, which in addition to European ancestry also have Native American and African ancestry. The informativity of Latin Americans is illustrated here by the novel case of archaic introgression we detected here. The introgressed haplotype around the *ATF3* gene region has a high frequency in East Asians and Native Americans, but is essentially absent in Europeans. Although 3D image analysis provides a fuller description of morphological features than 2D photographs, the need for specialized equipment for the acquisition of 3D images complicates its widespread use across worldwide populations. In addition, although sophisticated multivariate methods for 3D image analysis have been developed[8], composite phenotypes constructed in such analysis are specific to the samples studied, complicating the comparison of traits across independent study samples. Simple metrics, such as inter-landmark distances do not suffer from this difficulty and could also be more sensitive to specific genetic effects on narrow facial features[8]. Finally, automatic analysis of 2D facial photographs could facilitate the simultaneous study of facial features other than morphology, as illustrated here by the effects detected on eyebrow size. This could be of special relevance in forensics, which makes extensive use of 2D photographs. It is unclear that the prediction of facial features from genetic data will be of sufficient accuracy to be of practical utility in forensics[56, 57]. However, considering the much wider availability of 2D, relative to 3D data, progress in this area of research is more likely to stem from the analysis of 2D than 3D images (at least in the short term). Appropriate development of such forensic research will need to consider a wide range of ethical and legal issues associated with the use of highly-sensitive personal data[58–60].

Methods

Study subjects

Discovery sample: 6,486 (Colombia, N=1,407; Brazil, N=674; Chile, N=2,003; Mexico, N=1,203 and Peru, N=1,199) individuals from the Consortium for the Analysis of the Diversity and Evolution of Latin America (CANDELA consortium) were included in frontal photographs collection. CANDELA consortium (https://www.ucl.ac.uk/biosciences/gee/candela/) has been used to study physical appearance in Latin American for multiple studies, and details could be seen in Ruiz-Linares et al. 2014[17]. Ethical approval was obtained from the Universidad Nacional Autónoma de México (México), Universidad de Antioquia (Colombia), Universidad Perúana Cayetano Heredia (Perú), Universidad de Tarapacá (Chile), Universidade Federal do Rio Grande do Sul (Brazil) and University College London (UK). All participants provided written informed consent.

Replication samples: We examined replication in two independent data samples: one Chinese and one European.

The Chinese sample includes 5,298 individuals[39]. This sample stems from the National Survey of Physical Traits (NSPT) cohort (n = 2,628) and the Taizhou Longitudinal Study (TZL) cohort (n = 2,670)[27, 61]. The Taizhou Longitudinal Study (TZL) was approved by the Ethics Committee of Human Genetic Resources at the Shanghai Institute of Life Sciences, Chinese Academy of Sciences (ER-SIBS-261410). The National Survey of Physical Traits (NSPT) is the sub-project of The National Science & Technology Basic Research Project which was approved by the Ethics Committee of Human Genetic Resources of School of Life Sciences, Fudan University, Shanghai (14117). All participants provided written informed consent.

The European replication sample is the discovery cohort examined in the GWAS of Xiong et al. 2019[4]. This sample includes 10,115 individuals of European ancestry recruited in three countries (Netherlands, N=3,193; United Kingdom, N=4,727 and United States, N=2,195). The summary statistics are publicly available at https://doi.org/10.6084/m9.figshare.10298396

Genotype data

The genotype data examined here are those analyzed in previous GWAS of the CANDELA sample[1, 15, 16, 18, 19]. Briefly, a blood sample was collected from each volunteer and DNA extracted following standard laboratory procedures. DNA samples were genotyped on the Illumina HumanOmniExpress chip including 730,525 SNPs. PLINK v1.90 was used for quality control. Individuals and SNPs with >5% missing genotypes, SNPs with<1% minor allele frequency, and individuals who failed the X- or Y- chromosome sex checks were excluded. After these QC filters, ~650,000 SNPs and 5,500 individuals were retained for further analyses. SHAPEIT2[62] was used for pre-phasing the chip genotype data, and IMPUTE2[63] was then used to impute variants using the 1000 Genomes Phase 3 reference panel. Imputation led to 11,532,785 SNPs being available for association testing.

Phenotyping

Frontal digital photographs were taken for each CANDELA volunteer, at eye level, 1.5m away, using a Nikon D90 camera (12,3 Megapixels resolution) fitted with a Nikkor 50 mm fixed-focal-length lens[17]. The photographs were anonymized for confidentiality, and stored on a secure cluster, where an API script available from Face++ (https://www.faceplusplus.com), implementing a pre-trained deep learning model was run. Face++ placed 106 landmarks on each photograph (Supplementary Figure S1), and provided a set of attribute values. Face images with attribute values indicative of poor quality (e.g. blurriness, head pose) were excluded. We focused on 34 landmarks corresponding mostly to well-defined anatomical landmarks of common usage (Figure 1A, Supplementary Table S2). Procrustes superimposition was performed using MorphoJ[64] and pairwise inter-landmark distance (ILD) calculation was carried out using R[65]. Since Procrustes-adjusted landmarks coordinates were symmetrized, some inter-landmark distances were identical. After removing these duplicates, 301 distances (labeled as 'D' followed by a number, Supplementary Table S3) were retained for the GWAS.

To evaluate the robustness of the Face++ landmarking, we examined the accuracy of 16 landmarks which were placed manually on a subset of 1,610 photographs in a previous study[16] (Supplementary Figure S2). Median Euclidean distances and intraclass correlation coefficients (ICCs) between Face++ and manual landmark coordinates were obtained using Matlab[66].

Statistical genetic analysis

To estimate the narrow-sense heritability (h^2) for each trait, we firstly computed a genomic relationship matrix (GRM) after gathering all individuals who appeared in at least one trait, using LDAK5[20] with default parameters. For each trait, h^2 was estimated by fitting an additive linear model with a random effect term whose variance was obtained from the GRM, and added age, sex, BMI, 6 genetic PCs and head angles as covariates.

Genome-wide association study (GWAS) was conducted on the ILD phenotypes using PLINK v1.9. Sex, age, BMI, three head angles (yaw, pitch, roll) and the first 6 genetic PCs were included as the covariates.

Multiple testing was corrected by estimating the false discovery rate (FDR) threshold with the Benjamini-Hochberg procedure. The adjusted genome-wide significance threshold was 1.82×10^{-6}, which is larger than the commonly used GWAS genome-wide significance threshold (5×10^{-8}). Therefore, we eventually used 5×10^{-8} as the threshold.

After correction of multiple testing with FDR, the combined significance threshold in the two replication cohorts was 0.0253.

To group SNP-based GWA results across all analyses based on linkage disequilibrium (LD) between SNPs, we conducted clumping in PLINK v1.9 on the combined output file of all GWA analyses. We used 0.1 for LD threshold, and 1,000 Kb for the physical distance threshold, which in total resulted in 62 clumps. To further determine if each clump is independent, we conducted conditioned analyses on the signals physically close to each other. All covariates used in the original GWA analysis were also added in the conditional GWAS. All signals with a conditioned P-value greater than 5×10^{-8} were merged with their neighboring signals.

Conditional GWAS was also carried out to determine if a signal has been reported or not. We first picked out signals that fall on the chromosome bands that have been reported. Then we gathered all reported SNPs in each chromosome band and added those reported SNPs into the regression models of corresponding SNPs of the same chromosome band in our results. If P-value obtained was above the suggestive significant threshold (1×10^{-5}), this signal would be regarded as a reported signal, and conversely, it would be regarded as a new signal.

Interaction of *EDAR* with other genes

We examined the interaction between rs3827760 (*EDAR* causal SNP) and other index SNPs by testing regression models, adjusting for age, sex, BMI, three head angles, and the first six genetics PCs, as in our primary association analysis.

Index SNPs' independence of rs3827760 were determined based on the results of this conditional GWAS, with the suggestive significant threshold (1×10^{-5}) applied.

Multiple-testing corrected P-value threshold of 9.6×10^{-4} was used.

Expression analysis for significant SNPs

To further study the functions in facial development of the genome-wide significantly associated SNP markers and candidate genes nearby, we downloaded the RNA-seq data of CNCCs from the study of Prescott et al. (2015)[67]. Among all 118 genes annotated to 42 significant genomic regions, we found that 103 genes could express in CNCCs. We divided all the genes expressed in CNCCs into three groups. The first group includes 30 genes annotated to nine reported genomic regions in Table 1. The second group includes 73 genes annotated to the rest of 33 novel genomic regions. The third group includes the other 55,779 genes expressed in CNCCs as background genes. We performed comparisons of average expression levels among three groups by t-test.

Detection of archaic introgression near *ATF3* and association with facial features

Imputed genotypes of all samples and on 4,311 SNPs located in a 1Mb window around ATF3 were firstly phased by SHAPEIT4 (with default parameters). We then used a dedicated strategy[68] (relying on SHAPEIT2 with option "–call") to align these initial phases to those estimated from chip genotypes in this region. This low-density phasing was obtained by running RFMix (v1)[69]; it is expected to be more accurate and is aligned with local ancestry estimates resulting from that RFMix run. The rephased data was then filtered following the procedure described in Bonfante et al. (2021)[1], with the "Altai" Neanderthal sample and the 108 YRI samples from 1000GP3 respectively as archaic and modern references. A total of 3,231 SNPs were eventually retained to perform introgression scans with *admixtureHMM*[41], a hidden-Markov model for the purpose of detecting archaic admixture. For subsequent analyses, we considered only tracts that were called with a minimal confidence of 99% and that were longer than 10Kb (to avoid spurious ancestry calls). As the phases were aligned with RFMix ancestry calls, we were able to assign most archaic tracts to a continental ancestry background.

In order to test the association between archaic tracts and traits associated to that region, we counted the number of alleles that were carried by an introgression tract for each imputed SNP in the region and for each individual, which would be used as a coding of genotypes in modern and archaic counts. We defined 140 admixture mapping segments by merging consecutive SNPs with similar codings (that is, allowing a maximum of 1% change in coding from a SNP to the next one). A total of 103 segments had sufficient MAF (>1%) and were tested for association using the same linear model as for the initial GWAS.

Annotation of SNPs in FUMA

A subset of GWAS summary statistics including only significant SNPs ($P < 5×10^{-8}$) and a pre-defined lead SNP list obtained after clumping in Plink v1.9 were loaded to functional mapping and annotation (FUMA)[52]. SNP2GENE was processed to identify independent SNPs (r^2<0.6) and candidate SNPs. Candidate SNPs are the SNPs in LD of one of the independent significant SNPs, which includes non-GWAS tagged SNPs extracted from 1000 genomes reference panel. Implemented tool ANNOVAR was used to annotate the functional consequences on gene function on the independent SNPs and candidate SNPs.

Shape GWAS in outbred mice

We reanalyzed GWAS data on craniofacial shape variation in outbred mice[53]. Coordinates for 44 landmarks (17 pairs of symmetric landmarks and 10 landmarks on the median plane), along with allele dosage at 70k SNPs for 692 mice were kindly provided by Luisa Pallares. We performed a full generalized Procrustes analysis with object symmetry[70], and the phenotypic variation was modeled on the basis of the 67 non-null principal components (PC). A multivariate mixed model was fitted to evaluate the association between allele dosage and craniofacial shape (including skull centroid size as covariate).

Environmental and genetic covariances were modeled only from PC1 to PC10 (explaining 62.3 % of the total variance), and only variances were modeled for higher PCs. The variance components estimated for the model without SNP from the genomic relatedness matrix were used to correct both gene dosage and phenotypes[71]. Association was tested based on Pillai trace statistics obtained from the multivariate regression between the corrected genotypes and PC scores. A False Discovery Rate (FDR) was computed based on 100 permutations of corrected PC scores following the approach of Nicod et al (2016)[71] and used to identify SNPs exceeding a FDR threshold of 5%.

Declarations

Acknowledgments

General: Professor Gabriel Bedoya led the CANDELA team in Colombia but passed away during preparation of this manuscript. We thank the volunteers for their enthusiastic support for this research. We also thank Alvaro Alvarado, Mónica Ballesteros Romero, Ricardo Cebrecos, Miguel Ángel Contreras Sieck, Francisco de Ávila Becerril, Joyce De la Piedra, María Teresa Del Solar, William Flores, Martha Granados Riveros, Rosilene Paim, Ricardo Gunski, Sergeant João Felisberto Menezes   Cavalheiro, Major Eugênio Correa de Souza Junior, Wendy Hart, Ilich Jafet Moreno, Paola León-Mimila, Francisco Quispealaya, Diana Rogel Diaz, Ruth Rojas, and Vanessa Sarabia, for assistance with volunteer recruitment, sample processing and data entry. We are very grateful to the institutions that allowed the use of their facilities for the assessment of volunteers, including: Escuela Nacional de Antropología e Historia and Universidad Nacional Autónoma de México (México); Universidade Federal do Rio Grande do Sul (Brazil); 13° Companhia de Comunicações Mecanizada do Exército Brasileiro (Brazil); Pontificia Universidad Católica del Perú, Universidad de Lima and Universidad Nacional Mayor de San Marcos (Perú). We acknowledge M Arfan Ikram, Tamar EC Nijsten, Markus A de Jong, Stefan Boehringer, Myoung Keun Lee, Eleanor Feingold, Mary L Marazita, Lavinia Paternoster, Holly Thompson, Gemma C Sharp, Sarah Lewis, Stephen Richmond, Alexei Zhurov and Luisa Pallares for facilitating access to published datasets. We thank Luisa Pallares and Abraham Palmer for kindly sharing the GWA mouse data. Centre de Calcul Intensif d'Aix-Marseille is acknowledged for granting access to its high performance computing resources.

Funding: Work leading to this publication was funded by grants from: the National Natural Science Founddation of China (#31771393), the Scientific and Technology Committee of Shanghai Municipality (18490750300), Ministry of Science and Technology of China (2020YFE0201600),   Shanghai Municipal Science and Technology Major Project (2017SHZDZX01) and the 111 Project (B13016), the Leverhulme Trust (F/07 134/DF), BBSRC (BB/I021213/1), the Excellence Initiative of Aix-Marseille University - A*MIDEX (a French “Investissements d'Avenir” programme), the National Natural Science Founddation of China (#31771393), Universidad de Antioquia (CODI sostenibilidad de grupos 2013- 2014 and MASO 2013-2014), Conselho Nacional de Desenvolvimento Científico e Tecnológico, Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (Apoio a Núcleos de Excelência Program), Fundação de Aperfeiçoamento de Pessoal de Nível Superior and the National Institute of Dental and Craniofacial Research (R01-DE027023; U01-DE020078; R01-DE016148; X01-HG007821), Santander Research & Scholarship Award. B.B. is supported by a doctoral scholarship from Ecole Doctorale 251 Aix-Marseille Université.

Author contributions: B.B., C.C.S.C., C.G., C.J., C.M., E.W., F.M.S., F.R., G.B., G.P., H.V.R., J.C., J.G.V., L.P., L.S.F., M.C.B., M.D., M.F.G, M.H., P.E., P.J., R.B.L., R.G.J., S.C.Q., T.H., V.A.A., V.G., V.V., V.R. and W.A. contributed to volunteer recruitment or data collection. Q.L., J.C., K.A., N.N, and P.F. performed analyses.

B.Y., D.B, K.A, L.D., M.K., N.N., R.G.J., and T.C. provided guidance on aspects of study design. A.R.L., C.C., C.T., D.B., L.F., F.L., S.M.W., J. R. S., E. S., L. J. H., P. G. H., T. D. S. and M.K. obtained funding or provided access to resources. A.R.L. and K.A. designed the project. A.R.L., Q.L., K.A., and C.J. wrote the paper with input from co-authors. A.R.L. coordinated the study.

Competing interests: The authors declare that they have no competing interests.

Data availability: Raw genotype or phenotype data cannot be made available due to restrictions imposed by the ethics approval. Summary statistics obtained here have been deposited at GWAS central and will be available from the Autumn 2022 release.

URLs. The Altai Neanderthal genome was downloaded from the website of the Max Planck Institute for Evolutionary Anthropology at http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal/VCF/.

European cohort summary statistics: https://doi.org/10.6084/m9.figshare.10298396

References

1. Bonfante, B., et al., *A GWAS in Latin Americans identifies novel face shape loci, implicating VPS13B and a Denisovan introgressed region in facial variation*. Science Advances, 2021. 7(6)
2. White, J.D., et al., *Insights into the genetic architecture of the human face*. Nature Genetics, 2021. 53(1)
3. Wu, W., et al.: Whole-exome sequencing identified four loci influencing craniofacial morphology in northern Han Chinese. Hum. Genet. **138**(6), 601–611 (2019)
4. Xiong, Z.Y., et al., *Novel genetic loci affecting facial shape variation in humans*. Elife, 2019. 8
5. Li, Y., et al.: EDAR, LYPLAL1, PRDM16, PAX3, DKK1, TNFSF12, CACNA2D3, and SUPT3H gene variants influence facial morphology in a Eurasian population. Hum. Genet. **138**(6), 681–689 (2019)
6. Weinberg, S.M., et al.: Hunting for genes that shape human faces: Initial successes and challenges for the future, 22, pp. 207–212. Orthodontics & Craniofacial Research (2019)

7. Qiao, L., et al.: Genome-wide variants of Eurasian facial shape differentiation and a prospective model of DNA based face prediction. J. Genet. Genomics **45**(8), 419–432 (2018)
8. Claes, P., et al.: Genome-wide mapping of global-to-local genetic effects on human facial shape. Nat. Genet. **50**(3), 414–414+ (2018)
9. Cha, S., et al., *Identification of five novel genetic loci related to facial morphology by genome-wide association studies*. Bmc Genomics, 2018. 19
10. Richmond, S., et al., *Facial Genetics: A Brief Overview*. Frontiers in Genetics, 2018. 9
11. de Jong, M.A., et al.: Ensemble landmarking of 3D facial surface scans. Sci. Rep. **8**(1), 12 (2018)
12. Bannister, J.J., et al., *Fully Automatic Landmarking of Syndromic 3D Facial Surface Scans Using 2D Images*. Sensors (Basel), 2020. 20(11)
13. Quinto-Sanchez, M., et al.: Socioeconomic Status Is Not Related with Facial Fluctuating Asymmetry: Evidence from Latin-American Populations. PLoS One **12**(1), e0169287 (2017)
14. Liu, F., et al.: A genome-wide association study identifies five loci influencing facial morphology in Europeans. PLoS Genet. **8**(9), e1002932 (2012)
15. Adhikari, K., et al., *A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features*. Nature Communications, 2016. 7
16. Adhikari, K., et al., *A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation*. Nature Communications, 2016. 7
17. Ruiz-Linares, A., et al., *Admixture in Latin America: Geographic Structure, Phenotypic Diversity and Self-Perception of Ancestry Based on 7,342 Individuals*. Plos Genetics, 2014. 10(9)
18. Adhikari, K., et al.: A genome-wide association study identifies multiple loci for variation in human ear morphology. Nat. Commun. **6**, 7500 (2015)
19. Adhikari, K., et al.: A GWAS in Latin Americans highlights the convergent evolution of lighter skin pigmentation in Eurasia. Nat. Commun. **10**(1), 358 (2019)
20. Speed, D., et al.: Reevaluation of SNP heritability in complex human traits. Nat. Genet. **49**(7), 986–986+ (2017)
21. Marigo, V., et al.: Sonic hedgehog differentially regulates expression of GLI and GLI3 during limb development. Dev. Biol. **180**(1), 273–283 (1996)
22. Abdullah, et al.: Variants in GLI3 Cause Greig Cephalopolysyndactyly Syndrome. Genetic Test. Mol. Biomarkers **23**(10), 744–750 (2019)
23. Roscioli, T., et al., *Pallister-Hall syndrome: Unreported skeletal features of a GLI3 mutation.* American Journal of Medical Genetics Part A, 2005. **136a**(4): p. 390–394
24. Paternoster, L., et al.: Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position. Am. J. Hum. Genet. **90**(3), 478–485 (2012)
25. Pickrell, J.K., et al., *Detection and interpretation of shared genetic influences on 42 human traits (vol 48, pg 709*, 2016). Nature Genetics, 2016. **48**(10): p. 1296–1296
26. Boudjadi, S., et al.: The expression and function of PAX3 in development and disease. Gene **666**, 145–157 (2018)
27. Wu, S.J., et al., *Genome-wide association studies and CRISPR/Cas9-mediated gene editing identify regulatory variants influencing eyebrow thickness in humans*. Plos Genetics, 2018. 14(9)
28. Sennett, R., et al.: An Integrated Transcriptome Atlas of Embryonic Hair Follicle Progenitors, Their Niche, and the Developing Skin. Dev. Cell **34**(5), 577–591 (2015)
29. Sears, K.E., et al., The correlated evolution of Runx2 tandem repeats, transcriptional activity, and facial length in Carnivora. Evolution & Development, 2007. 9(6): pp. 555–565
30. Komori, T.: Roles of Runx2 in Skeletal Development. Adv. Exp. Med. Biol. **962**, 83–93 (2017)
31. Otto, F., et al.: Cbfa1, a candidate gene for cleidocranial dysplasia syndrome, is essential for osteoblast differentiation and bone development. Cell **89**(5), 765–771 (1997)
32. Komori, T., et al.: Targeted disruption of Cbfa1 results in a complete lack of bone formation owing to maturational arrest of osteoblasts. Cell **89**(5), 755–764 (1997)
33. Purandare, S.M., et al., *De novo three-way chromosome translocation 46,XY,t(4;6;21)(p16;p21.1;q21) in a male with cleidocranial dysplasia.* American Journal of Medical Genetics Part A, 2008. **146a**(4): p. 453–458
34. Ritter, D.I., et al.: Identifying gene disruptions in novel balanced de novo constitutional translocations in childhood cancer patients by whole-genome sequencing. Genet. Sci. **17**(10), 831–835 (2015)
35. Kurosaka, H., et al.: Disrupting hedgehog and WNT signaling interactions promotes cleft lip pathogenesis. J. Clin. Invest. **124**(4), 1660–1671 (2014)
36. Le Pabic, P., Ng, C., Schilling, T.F.: Fat-Dachsous signaling coordinates cartilage differentiation and polarity during craniofacial development. PLoS Genet. **10**(10), e1004726 (2014)
37. Milunsky, J.M., et al.: LADD syndrome is caused by FGF10 mutations. Clin. Genet. **69**(4), 349–354 (2006)
38. Zhou, X., et al.: Genome-wide CRISPR knockout screens identify ADAMTSL3 and PTEN genes as suppressors of HCC proliferation and metastasis, respectively. J. Cancer Res. Clin. Oncol. **146**(6), 1509–1521 (2020)
39. Zhang, M., et al., *Genetic mechanisms underlying East Asian and European Facial differentiation.* 2021
40. Chintalapati, M., Dannemann, M., Prufer, K.: *Using the Neandertal genome to study the evolution of small insertions and deletions in modern humans*. Bmc Evolutionary Biology, 2017. 17
41. Racimo, F., et al.: Archaic Adaptive Introgression in TBX15/WARS2. Mol. Biol. Evol. **34**(3), 509–524 (2017)

42. Bush, J., Kim, S.: *Convergence and extrusion driven by non-muscle myosin II are required for normal fusion of the mammalian secondary palate*. Faseb Journal, 2015. 29
43. Gajera, M., et al., *MicroRNA-655-3p and microRNA-497-5p inhibit cell proliferation in cultured human lip cells through the regulation of genes related to human cleft lip*. Bmc Medical Genomics, 2019. 12
44. Twine, N.A., et al., *Transcription factor ZNF25 is associated with osteoblast differentiation of human skeletal stem cells*. Bmc Genomics, 2016. 17
45. Lian, J.B., et al.: Regulatory controls for osteoblast growth and differentiation: Role of Runx/Cbfa/AML factors. Crit. Rev. Eukaryot. Gene Expr. **14**(1–2), 1–41 (2004)
46. Marie, P.J.: Transcription factors controlling osteoblastogenesis. Arch. Biochem. Biophys. **473**(2), 98–105 (2008)
47. Kichaev, G., et al.: Leveraging Polygenic Functional Enrichment to Improve GWAS Power. Am. J. Hum. Genet. **104**(1), 65–75 (2019)
48. Lee, M.K., et al.: Genome-wide association study of facial morphology reveals novel associations with FREM1 and PARK2. PLoS One **12**(4), e0176566 (2017)
49. Shaffer, J.R., et al.: Genome-Wide Association Study Reveals Multiple Loci Influencing Normal Human Facial Morphology. PLoS Genet. **12**(8), e1006149 (2016)
50. Curtis, S.W., et al., *The PAX1 locus at 20p11 is a potential genetic modifier for bilateral cleft lip*. HGG Adv, 2021. 2(2)
51. Pallares, L.F., et al., *Mapping of Craniofacial Traits in Outbred Mice Identifies Major Developmental Genes Involved in Shape Determination*. Plos Genetics, 2015. 11(11)
52. Watanabe, K., et al.: Functional mapping and annotation of genetic associations with FUMA. Nat. Commun. **8**(1), 1826 (2017)
53. Lee, J.J., et al.: Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat. Genet. **50**(8), 1112–1112+ (2018)
54. Liu, M.Z., et al.: Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat. Genet. **51**(2), 237–237+ (2019)
55. Nalls, M.A., et al.: Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet Neurol. **18**(12), 1091–1102 (2019)
56. Kayser, M., de Knijff, P., *Improving human forensics through advances in genetics, genomics and molecular biology (vol 12, pg 179*,: 2011). Nature Reviews Genetics, 2012. **13**(10): p. 754–754
57. Kayser, M.: Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. Forensic Sci. International-Genetics **18**, 33–48 (2015)
58. Wright, E.A., et al.: Practical and Ethical Considerations of Using Personal DNA Tests with Middle-School-Aged Learners. Am. J. Hum. Genet. **104**(2), 197–202 (2019)
59. Erlich, Y., Narayanan, A.: Routes for breaching and protecting genetic privacy. Nat. Rev. Genet. **15**(6), 409–421 (2014)
60. Wan, Z., et al., *Sociotechnical safeguards for genomic data privacy*. Nat Rev Genet, 2022
61. Wang, F., et al.: A Genome-Wide Scan on Individual Typology Angle Found Variants at SLC24A2 Associated with Skin Color Variation in Chinese Populations. J Invest Dermatol (2021)
62. O'Connell, J., et al., *A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness*. Plos Genetics, 2014. 10(4)
63. Howie, B., et al.: Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. **44**(8), 955–955+ (2012)
64. Klingenberg, C.P.: MorphoJ: an integrated software package for geometric morphometrics. Mol. Ecol. Resour. **11**(2), 353–357 (2011)
65. Core Team, R.: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria (2021)
66. MATLAB: version 7.10.0 (R2010a). The MathWorks Inc (2010)
67. Prescott, S.L., et al.: Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. Cell **163**(1), 68–83 (2015)
68. Faux, P., Druet, T.: *A strategy to improve phasing of whole-genome sequenced individuals through integration of familial information from dense genotype panels*. Genetics Selection Evolution, 2017. 49
69. Maples, B.K., et al.: RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. Am. J. Hum. Genet. **93**(2), 278–288 (2013)
70. Klingenberg, C.P., Barluenga, M., Meyer, A.: Shape analysis of symmetric structures: Quantifying variation among individuals and asymmetry. Evolution **56**(10), 1909–1920 (2002)
71. Nicod, J., et al.: Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing. Nat. Genet. **48**(8), 912–912+ (2016)

Tables

Table 1
Features of nine genome regions reported in previous facial morphology GWASs for which genome-wide significant association is also observed here.

Chromosomal region	Index SNP	Candidate gene[a]	# significant traits	# significant SNPs	Strongest P-value	Refs.
2q12.3	rs72627476	EDAR	75	1,245	5.60×10^{-34}	1, 5, 16
7p14.1	rs846315	GLI3	5	31	1.20×10^{-12}	2, 16
2q36.1	rs13022712	PAX3	18	37	6.42×10^{-12}	1, 2, 4, 5, 8, 14, 24, 25
5q13.2	rs7341037	*FOXD1*	10	18	1.49×10^{-10}	27
6p21.1	rs141680515	SUPT3H/RUNX2	15	275	1.68×10^{-10}	1, 2, 5, 8, 16, 25
15q21.1	rs1426654	SLC24A5	2	4	1.74×10^{-10}	1
4q31.3	rs2045323	*DCHS2/SFRP2*	8	64	9.35×10^{-10}	1, 2, 8, 16
5p12	rs4505960	*FGF10*	3	1	2.43×10^{-8}	2
15q25.2	rs62027787	ADAMTSL3	1	3	2.65×10^{-8}	2

[a] Genes including significantly associated SNPs are underlined (bold indicates that an associated SNP leads to an amino-acid substitution).

Table 2
Features of five genomic regions newly associated with facial variation.

Region	Index SNP	Alleles (Ref/Alt)	Association P-value	Candidate genes	# Significant traits	# Significant SNPs	Replication P-Value[a]	CAN AF	NAM AF	EUR AF	AFR AF
1q32.3	rs12564392	C/A	2×10^{-08}	*ATF3,FAM71A,BATF3,NSL1*	3	3	0.0105	0.28	0.61	0.01	0.00
3q21.1	rs820360	A/G	5×10^{-10}	*PTPLB,MYLK-AS1,MYLK*	8	3	0.0014	0.35	0.32	0.37	0.06
8p11.21	rs59547557	T/C	7×10^{-10}	*ADAM18,ADAM2,IDO1,IDO2*	7	39	0.0025	0.48	0.31	0.55	0.81
10p11.1	rs58831446	A/AT	1×10^{-10}	*ZNF248,ZNF33BP1,ZNF25*	8	25	0.0002	0.51	0.66	0.41	0.44
22q12.1	rs9608473	G/A	1×10^{-08}	*SEZ6L*	6	11	0.0011	0.29	0.49	0.10	0.07

(a) Smallest P-value observed in the Chinese and European replication samples.

AF, alternative allele frequency; CAN, CANDELA; AFR, Sub-Saharan African; EUR, European (from 1000 genomes); NAM, Native American.

Figures

Figure 1

Overall summary of the GWAS performed here. (A) Illustration of the location of the 34 landmarks used for calculation of the 301 inter-landmark distances examined in the GWAS. **(B)** Illustration of 148 inter-landmark distances showing significant association with at least one genomic region. Distances significantly associated with the 9 genomic regions reported in previous GWAS of facial features are in black (genomic regions that are included in Table 1). Distances significantly associated with the 5 novel highlighted genomic regions are in red (genomic regions that are included in Table 2). Distances significantly associated with the rest of genomic regions are in light blue. **(C)** A combined Manhattan plot illustrating the significant GWAS signals across all 301 inter-landmark distances. 1,865 SNPs (dots) across 42 genomic regions exceed the nominal significance threshold of -log(P)>7.3 (red line). For visibility, the y-axis has been truncated at a -log(P) of 13. Red labels indicate novel candidate genes highlighted in this study (Table 2). Black labels indicate candidate genes reported in previous GWAS of facial features (Table 1).

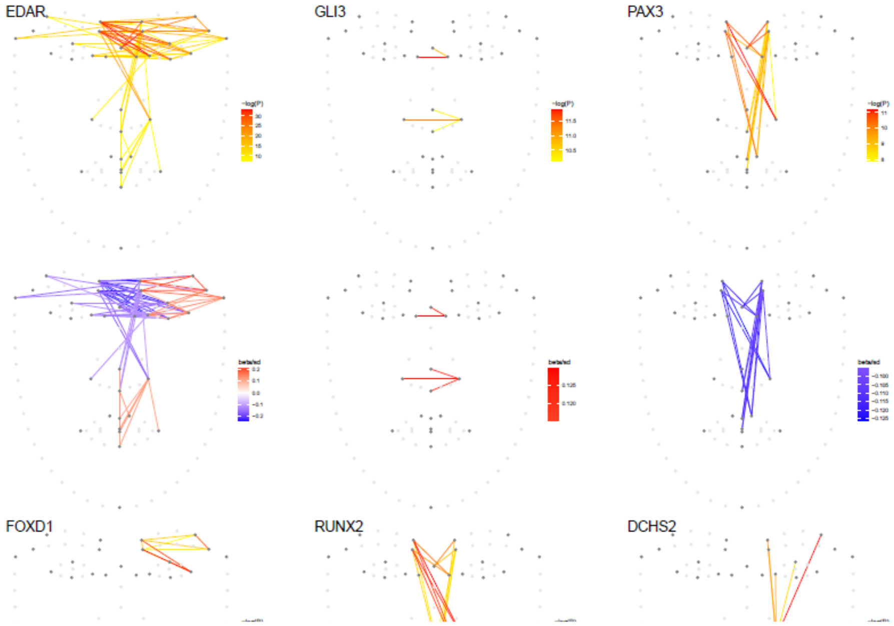

Figure 2

Illustration of inter-landmarks distances associated with the 6 most robustly replicated signals (*EDAR, GLI3, PAX3, FOXD1, RUNX2, DCHS2*). Red/yellow color gradient reflects the strength of SNP association with a facial distance. Red/blue gradient indicates the direction of the SNP effect.

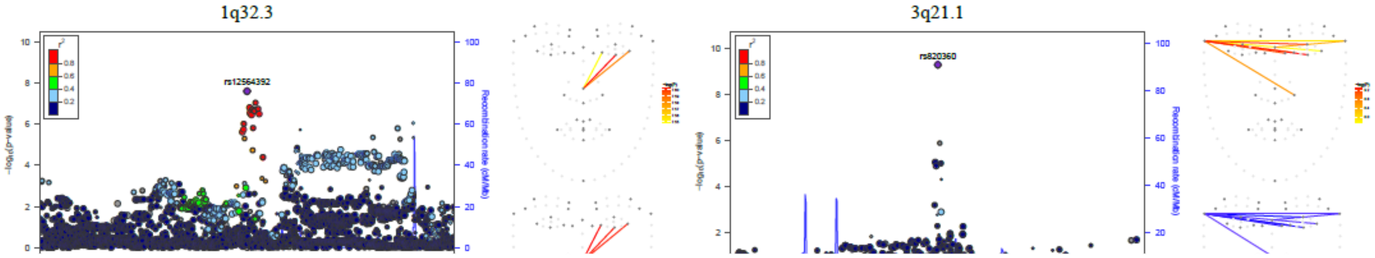

Figure 3

Regional association plots for 4 novel genomic regions associated with facial features (1q32.3, 3q21.1, 8p11.21, 10p11.1). Each panel shows on the left the association P-values for SNPs in a region (with the index SNP labeled, and annotated genes underneath). On the right of each panel are displayed the associated inter-landmark distances (Top: color reflects P-value. Bottom: color reflects the direction of the effect)

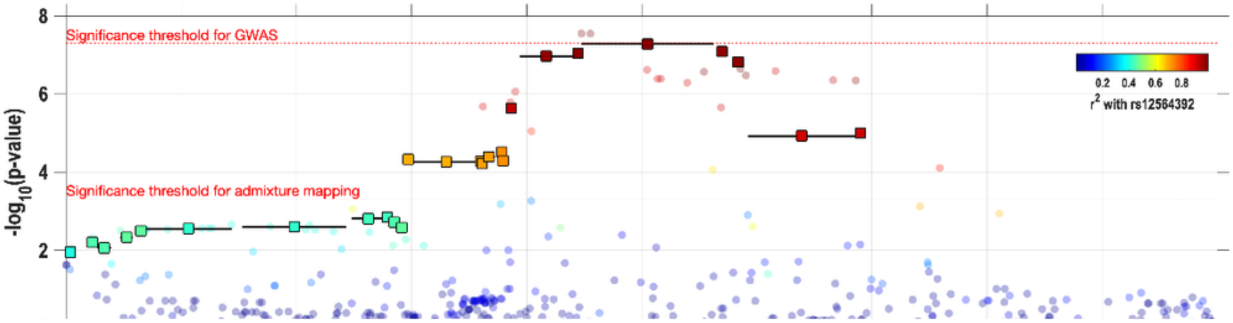

Figure 4

Regional association for SNPs in 1q32.3 with the distance between landmarks 13-23 (D203). Top panel: -log P-values obtained in the initial GWAS (circles in the background) and with archaic (Neanderthal) admixture mapping analyses (squares in the foreground marking the center of an admixture mapping segment and whiskers marking its extent). For the two types of association tests, SNPs have been colored to reflect LD with the most significant SNP in the initial GWAS (rs12564392). Middle line: Combined craniofacial annotations, including craniofacial-specific super-enhancers, around the index SNP from the Roadmap Epigenomics project for a model learned with ChromHMM using 25 states and 12 marks. Color codes follow the key defined by the Roadmap Epigenomics project (https://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp). Middle panel: archaic introgression tracts called in any CANDELA sample have been aggregated, yielding, at each position, the frequency of archaic introgression. Bottom panel: candidate genes in the vicinity. A width of 100 KB on each side of the index SNP rs12564392 is shown.

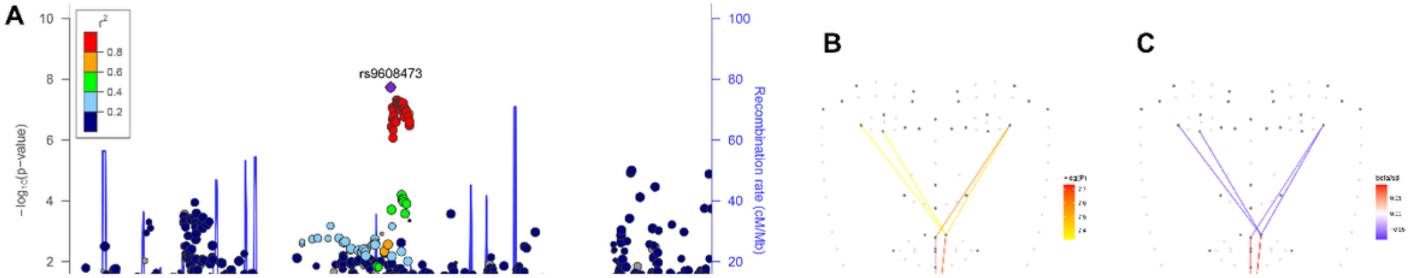

Figure 5

Regional association plots for the 22q12.1 region in human (top) and mice (bottom). (A) LocusZoom plot for the 22q12.1 region: Top panel shows regional association P-values (index SNP is labelled); bottom panel shows the genes in vicinity. **(B-C)** Associated inter-landmark distances; (B): color reflects P-value. (C): color reflects the direction of the effect. **(D)** Regional association plot for the Tct28/Mn1 homologous region in mouse chromosome 5. Dot colours correspond to the LD (r2) with rs32069343. Brown genes are annotated for craniofacial, skeleton, bone, etc. **(E)** Phenotypic effect on the cranial shape associated with allele dosage at the rs32069343. Expansion/contraction relative to the mean shape is shown in blue/brown. To facilitate visualization, the effect has been amplified by a factor of 20 (see also Supplementary Movie).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementarytables.xlsx
- supplementarymovie.mp4
- Supplement.docx