bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genome-wide Variants of Eurasian Facial Shape Differentiation and DNA based Face Prediction

Lu Qiao[1,2], Yajun Yang[3,4], Pengcheng Fu[5], Sile Hu[1,2], Hang Zhou[1,2], Shouneng Peng[1,2], Jingze Tan[3,4], Yan Lu[1], Haiyi Lou[1], Dongsheng Lu[1,2], Sijie Wu[1,2], Jing Guo[1], Li Jin[1,2,3,4,7], Yaqun Guan[6], Sijia Wang[1,2,7,*], Shuhua Xu[1,2,7,8,*] & Kun Tang[1,2,*]

[1] Chinese Academy of Sciences (CAS) Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Science, Shanghai 200031, China;
[2]University of Chinese Academy of Sciences, Beijing 100049, China;
[3]State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China;
[3]Fudan-Taizhou Institute of Health Sciences, 1 Yaocheng Road, Taizhou, Jiangsu 225300, China;
[5]Department of neurology, the First People's Hospital of Chenzhou, Hunan, 423000, China;
[6]Department of Biochemistry and Molecular Biology, Preclinical Medicine College, Xinjiang Medical University, Urumqi 830011, P. R China;
[7]Collaborative Innovation Center of Genetics and Development, Shanghai 200438, China;
[8]School of Life Science and Technology, ShanghaiTech University, Shanghai 200031, China.
*Correspondence could be addressed to K.T. (tangkun@picb.ac.cn), S.X. (xushua@picb.ac.cn), S.W. (wangsijia@picb.ac.cn),

Abstract

It is a long standing question as to which genes define the characteristic facial features among different ethnic groups. In this study, we use Uyghurs, an ancient admixed population to query the genetic bases why Europeans and Han Chinese look different. Facial traits were analyzed based on high-dense 3D facial images; numerous biometric spaces were examined for divergent facial features between European and Han Chinese, ranging from inter-landmark distances to dense shape geometrics. Genome-wide association analyses were conducted on a discovery panel of Uyghurs. Six significant loci were identified four of which, rs1868752, rs118078182, rs60159418 at or near *UBASH3B*, *COL23A1*, *PCDH7* and rs17868256 were replicated in independent cohorts of Uyghurs or Southern Han Chinese. A quantitative model was developed to predict 3D faces based on 277 top GWAS SNPs. In hypothetic forensic scenarios, this model was found to significantly enhance the verification rate, suggesting a practical potential of related research.

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

The human face plays a pivotal role in daily life. Communication, mutual identification, sexual attraction, etc... all strongly depends on face. It has been long noted that faces bear characteristic features that may surrogate one's ancestry, even in highly admixed populations[1]. Our recent investigation[2] also revealed strong morphological divergence on multiple facial features, including nose, brow ridges, cheeks and jaw, between Europeans and Han Chinese. Suggesting that facial shapes have been strongly shaped by natural selection after sub-populations split from a common ancestor. It is therefore a fundamental and intriguing question to ask: Which genetic variants contribute to the substantial morphological differences among continental populations?

Normal facial shape is known to be highly heritable[3, 4, 5]. However, until recently, very little was known about the genetic basis of common variation of facial morphology. In the last few years, several genome-wide association studies (GWAS) were carried out and multiple face shape associated loci were identified[6, 7, 8]. These studies all based their phenotyping on conventional scalar measurements involving limited number of landmarks. On the other hand, efforts have been paid to use dense 3D face modeling (3dDFM), a novel high-dimensional data format, to fully represent the complex facial shape phenome[2, 9, 10]. Peng *et al.* first applied 3dDFM to identify the association between common mouth shape variation and a cleft-lip related genetic locus[11]. Claes *et al*. showed that numerous genes are associated with complex normal facial shape variation based on 3dDFM[12], and proposed the potential of modeling the 3D face based on genotype and its use in forensic practice[12, 13].

In this study, we aimed at identifying loci on a genome-wide scale that contribute to the divergent facial morphological features between Europeans and Han Chinese. In brief, GWAS were conducted on the polarized face phenotypes along the European-Han dimensions, and Uyghur was used as the study cohort to dissect the genotype-phenotype association. Uyghur is a minority group living in Xinjiang province in China, and was found to have arisen from ancient admixtures between East-Asian and European ancestries at a roughly equal ratio, followed by a long period of isolation[14, 15, 16]. Furthermore, Uyghur facial traits demonstrated a wide range of shape gradients between the characteristic Europeans and Han-Chinese faces[2]. These properties made Uyghur an ideal group to study the genetic variants of divergent facial features across Eurasia. We performed GWAS in 694 Uyghurs using both landmark based and 3dDFM based phenotypes. Significant loci were replicated in an independent Uyghur sample and a Han Chinese cohort. Next, we investigated whether 3D faces could be predicted to certain degree by using the top associated SNPs. A quantitative model was established to summarize the phenotypic effects of multiple loci and to simulate realistic 3D face models. The predicted faces were compared to the actual data to test the performance. In the end, the 3D face prediction approach was formally tested in hypothetic forensic scenarios to evaluate its practical potential.

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

The studied cohorts included two independent Uyghur panels (694 and 171 individuals as UIG-D and UIG-R respectively) from Xinjiang China, and 1504 Han Chinese from Chenzhou China (HAN-CZ) (Table 1). Furthermore, a Han Chinese cohort from Taizhou China (HAN-TZ, 929) and 86 Shanghai residents of European ancestry (EUR) were used as the phenotype reference groups (Table 1). The participants were peer group with 20.02 +/- 2.16 (SD) years old. We collected their three-dimensional facial images and mapped them to a common 32,251 points' spatial dense mesh automatically[10]. Based on these, we first defined the candidate phenotypes of study. Briefly, the face images were jointly analyzed among EUR, UIG-D and HAN-TZ, and complex face data was decomposed to various phenotype measurements (see Methods). Candidate phenotypes were chosen if EUR and HAN-TZ exhibited strong divergence and UIG-D covered a wide range in-between (Fig. 1). Three types of phenotypes were used: First, ten inter-landmark distances were selected for the substantial Eurasian divergence based on fifteen salient landmarks (Supplementary Fig. 1a). The other two types of phenotypes were based on decomposing the high dimensional 3dDFM data. Specifically, we first extracted six facial features, namely, brow ridge, eyes, side faces, cheeks, nose and mouth, grossly based on the reported among-population differentiation[2] (Supplementary Fig. 1b). The corresponding 3dDFM data was decomposed by either partial least square (PLS) or principle component (PC) analysis, and the PLS model (sPLS) and PC model (sPC) that defined the strongest segregation between EUR and HAN-TZ were selected as candidate phenotypes in each feature (see Methods, Supplementary Fig. 2, Supplementary Table 1 and 2).

Table 1 | Characteristics of study samples

Location	Abb.	For	Ethnic	Total N	Male N	Mean Age	Range (s.d.)
Urumchi, China	UIG-D	discovery	Uyghur	694	270	20.09	17-25 (1.24)
Kashi, China	UIG-R	replication	Uyghur	171	63	20.55	17-25 (1.42)
ChenZhou, China	HAN-CZ	replication	Han Chinese	1,504	424	19.69	17-32 (1.67)
TaiZhou, China	HAN-TZ	discovery	Han Chinese	929	363	19.81	17-25 (0.95)
Europeans living Shanghai	EUR	discovery	European	86	57	27.59	16-42 (5.59)

N, sample size

Genome-wide association studies of facial shape. Given the candidate phenotypes, we carried out sex-stratified and sex-mixed GWAS in UYG-D on 847,046 SNPs after quality control (see Methods). In general, the Quantile-Quantile (Q-Q) plot[17] analyses revealed little subpopulation stratification (Supplementary Fig. 3). Six independent SNPs met genome-wide significance[18] ($P<5 \times 10^{-8}$) (Fig. 2, Table 2), including rs1868752 (at 11q24.1) associated with distance between external canthus and internal canthus (ExtCan-IntCan) in mixed genders, rs118078182 (on *COL23A1* at 5q35.3) associated with distance of Nasion point-Pronasale-Subnasale (Nsn-Prn-Sbn) in

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mixed genders, rs60159418 (on *PCDH7* at 4p15.1) associated with mouth sPC in males, rs17868256 (at 2p16.3) associated with cheek sPLS in females, rs3920540 (near *BMP2* at 20p12.3) associated with nasal sPLS in females, and rs61672954 (at 3p12.2) associated with the sPLS of side-faces in mixed genders. In order to control for potential confounding effects from varying ancestry makeup, we inferred the ancestry proportions for each UIG-D individual (see Methods). The six signals remained after accounting for the inferred ancestry in the association model (Supplementary Table 3).

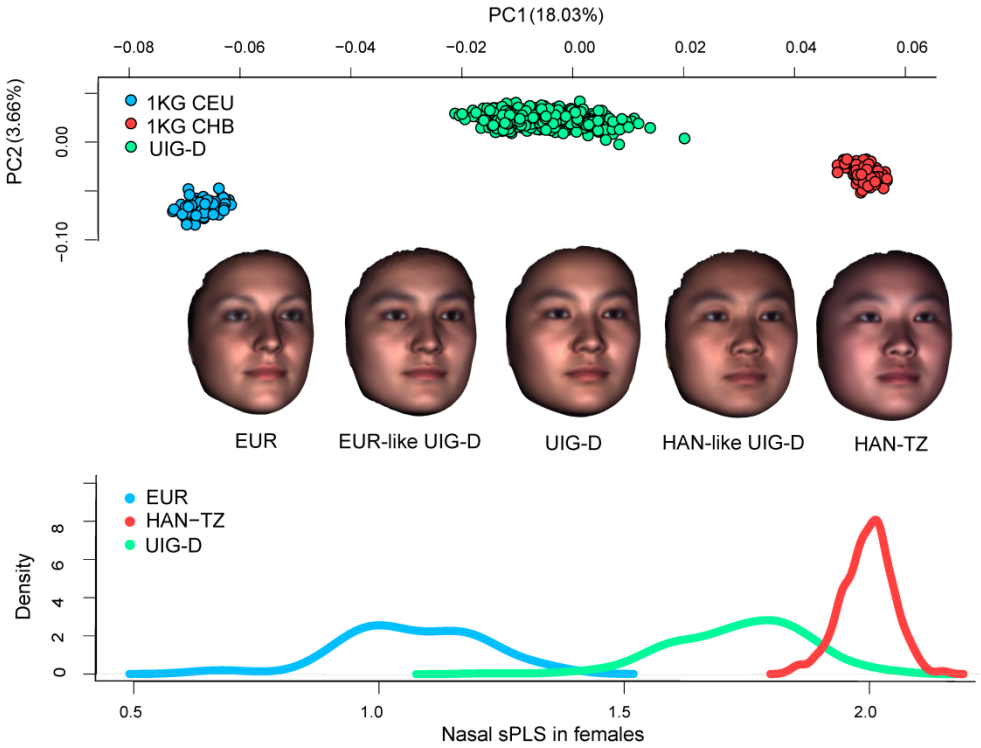

Figure 1 | Overall scheme of the study design. In the top panel, the genetic structure of three Eurasian populations was analyzed by PCA based on the 1KG genome data of 97 CHB (red), 85 CEU (blue) and the whole-genome sequencing of 694 UYG-D (green). Clear clustering can be seen based on the ethnic backgrounds. In particular, UYG-D individuals clearly lie in the half way between CEU and CHB along PC1, all consisted of a roughly equal ratio of CHB and CEU ancestries. Compared to the genetic composition, Uyghur individuals exhibit broad gradients of admixture in the facial phenotypes. The middle panel shows the average face models for EUR, European-like Uyghurs (EUR-like UIG-D), UIG-D, Han-like Uyghurs (HAN-like UIG-D) and HAN-TZ from left to right. EUR-like UIG-D and HAN-like UIG-D were obtained by averaging over 20 UIG-D individuals visually accessed to resemble Europeans or Han Chinese. The bottom panel shows the distribution of sPLS in nose, revealing a distinct segregation between EUR and HAN-TZ and a wide spread of UIG-D stretching between EUR and HAN-TZ along this phenotype dimension. In this study, the highly divergent phenotypes as shown above were selected and tested for association loci genome-widely in UIG-D.

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

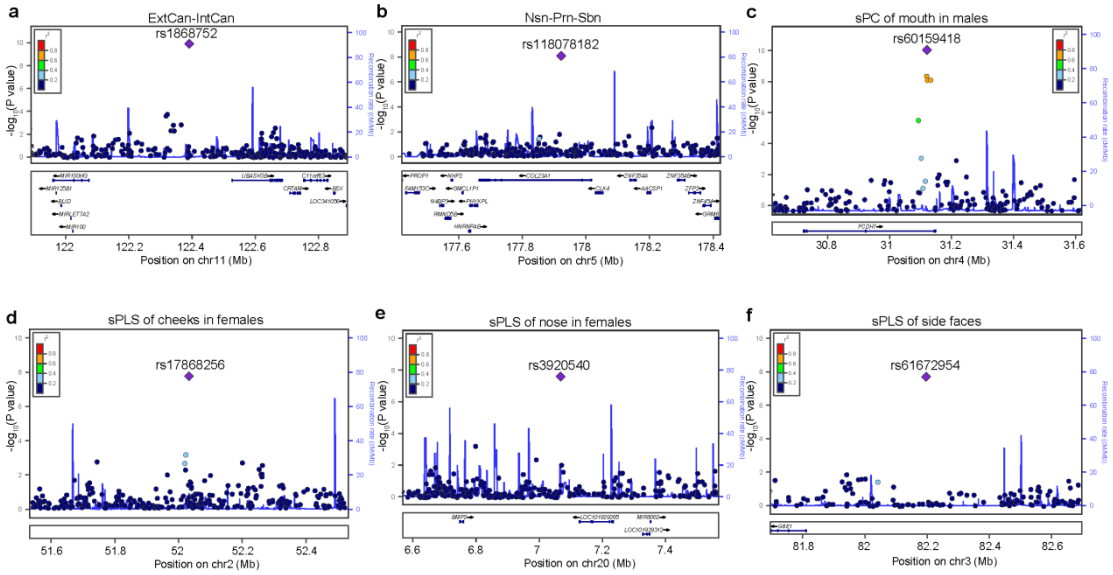

Figure 2 | Six genomic regions harboring SNPs of genome-wide significant associations with facial shape. (a) 11q24, (b) 5q35, (c) 4q15, (d) 2q16, (e) 20q12, (f) 3q12. The LocusZoom plots are given for the 500kb flanking region centered by the most significant loci.

Ideally, the effects of shape related loci should be directly visualized on facial images. We modeled the effects of candidate SNPs using heat plots as well as extrapolated faces[19] (Fig. 3, Supplementary Movie 1~6). As can be seen in Fig. 3a (Supplementary Movie 1), the extrapolated face towards the effect of rs1868752T had narrower eyes (smaller ExtCan-IntCan distance) compared to the G allele, resulting in a substantial displacement on the X axis; G also seemed to be associated with elevated nose ridge. SNP rs118078182 showed an obvious impact on the nasal shape along the Y and Z axes. Compared to rs118078182A, rs118078182G seemed to make the nose longer and more protrusive (taller) from the face, consistent with the association with Nsn-Prn-Sbn distance (Fig. 3b, Supplementary Movie 2). For rs60159418 in males (Fig. 3c, Supplementary Movie 3), the main shape changed the mouth along the Y axis, followed by Z axis. Allele G seemed to make the whole mouth area recessed from the face plane; in comparison, the mouth-chin curve bended convexly from the facial plane in the extrapolated face of the A allele. rs60159418 also seemed to influence other facial features: for G allele, nose and chin looked relatively protrusive outwards, and eye brow ridges seemed to elevate. The SNP rs17868256 mainly affected the shape of cheeks (Fig. 3d, Supplementary Movie 4), with G allele associated to laterally expanded cheeks, making the face look wider on the X axis. On the Y axis, rs17868256G also seemed to lift the cheek protrusion upwards. For SNP rs3920540 in females (Fig. 3e, Supplementary Movie 5), G allele was mainly associated with repressed nasal bridge and nasal tip along Z axis compared to T allele; G allele also seemed to link to more protrusive chin on the extrapolated face. The most notable effect of rs61672954 occurred around the jaw lines (Fig. 3f, Supplementary Movie 6), with the A allele associated with stronger jawlines and therefore comparatively wider lower face than the extrapolated face of G allele.

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2 | SNPs with GWAS signals and their narrow-sense replications

					Ancestral Allele Frequency							Discovery in additive model[E]		Replication in additive model						Replication in dominant Model[F]				Meta-analysis UIG-D+UIG-R in additive model	
												UYG-D		UIG-R		HAN-CZ		UIG-R+HAN-CZ		aa+ab:bb		aa:ab+bb			
																				UIG-R	HAN-CZ	UIG-R	HAN-CZ		
SNP	Chr.	BP[A]	at or near Gene	AA[B]	UIG[C]	CHB[D]	CEU[D]	Fst[G]	Ratio[H]	gender group	Feature	Beta	P	Beta	P	Beta	P	Beta	P	P	P	P	P	Beta	P
rs1868752	11	122391442	UBASH3B	T	0.985	0.961	0.985	0.000238	0.954	mixed	ExtCan-IntCan	3.804	**1.22×10^{-10}**	-0.666	0.604	0.126	0.579	0.103	0.647	-	**0.0449**	0.604	0.752	3.19	**7.70×10^{-9}**
rs118078182	5	177922198	COL23A1	G	0.898	0.791	0.995	0.151	0.112	mixed	Nsn-Prn-Sbn	-2.06	**8.19×10^{-9}**	-2.3	**0.00422**	-0.33	**0.0713**	-0.4	**0.0236**	0.27	**0.0167**	**0.00309**	0.247	-2.07	**2.64×10^{-10}**
rs60159418	4	31120752	PCDH7	A	0.356	0.631	0.04	0.466	0.00297	male	Mouth sPC	-19.8	**8.96×10^{-11}**	-5.73	0.363	-4.43	**0.0854**	-4.62	**0.0523**	0.683	0.124	0.195	0.175	-17	**9.78×10^{-10}**
rs17868256	2	52032773	-	A	0.643	0.476	0.753	0.101	0.2	female	Cheek sPLS	0.053	**7.22×10^{-9}**	0.0492	**0.0374**	0.00299	0.537	0.005	0.282	**0.0163**	0.22	0.22	0.872	0.0522	**1.07×10^{-9}**
rs3920540	20	7067738	BMP2	T	0.848	0.883	0.859	**1.71×10^{-5}**	0.988	female	Nose sPLS	0.071	**3.31×10^{-8}**	0.0255	0.432	-0.000669	0.872	-0。000822	0.852	-	0.386	0.432	0.649	0.0652	**4.69×10^{-8}**
rs61672954	3	82196528	-	G	0.967	0.879	1	0.0816	0.253	mixed	side-faces sPLS	0.119	**2.00×10^{-8}**	-0.02	0.704	0.00157	0.823	0.001	0.86	-	0.392	0.704	0.992	0.101	**2.61×10^{-7}**

[A]NCBI build 37

[B]Ancestral allele

[C]in meta-UIGhurs (UIG-D + UIG-R)

[D]1KG phase 3 database

[E]in additive model, homozygote ancestral allele as 0, heterozygote as 1, homozygote derived allele as 2

[F] a allele stands for ancestral allele

[G]Fst calculated between CHB and CEU in 1KG phase1 release 2 database

[H]Ratio means Rank/809084

Table 3 | Broad-sense replications of SNPs with GWAS signals

	PSD-based Permutation in HAN-CZ[A]		PLS-based Permutation	
SNP	between genotypes	HAN-CZ	UYG-R	HAN-CZ
rs1868752	TG:TT	**0.0283**	0.681	-
rs118078182	AA+GA:GG	**0.0278**	**0.003**	0.27
rs60159418	AA:GG	0.188	0.194	**$<1 \times 10^{-3}$**
rs17868256	AA:GG	**0.0085**	0.956	**$<1 \times 10^{-3}$**
rs3920540	TG:TT	**0.0139**	0.717	**0.03**
rs61672954	AA:GG	0.432	0.228	0.25

[A]replication only in HAN-CZ as sample sizes in different genotype groups are too small in UIG-R

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

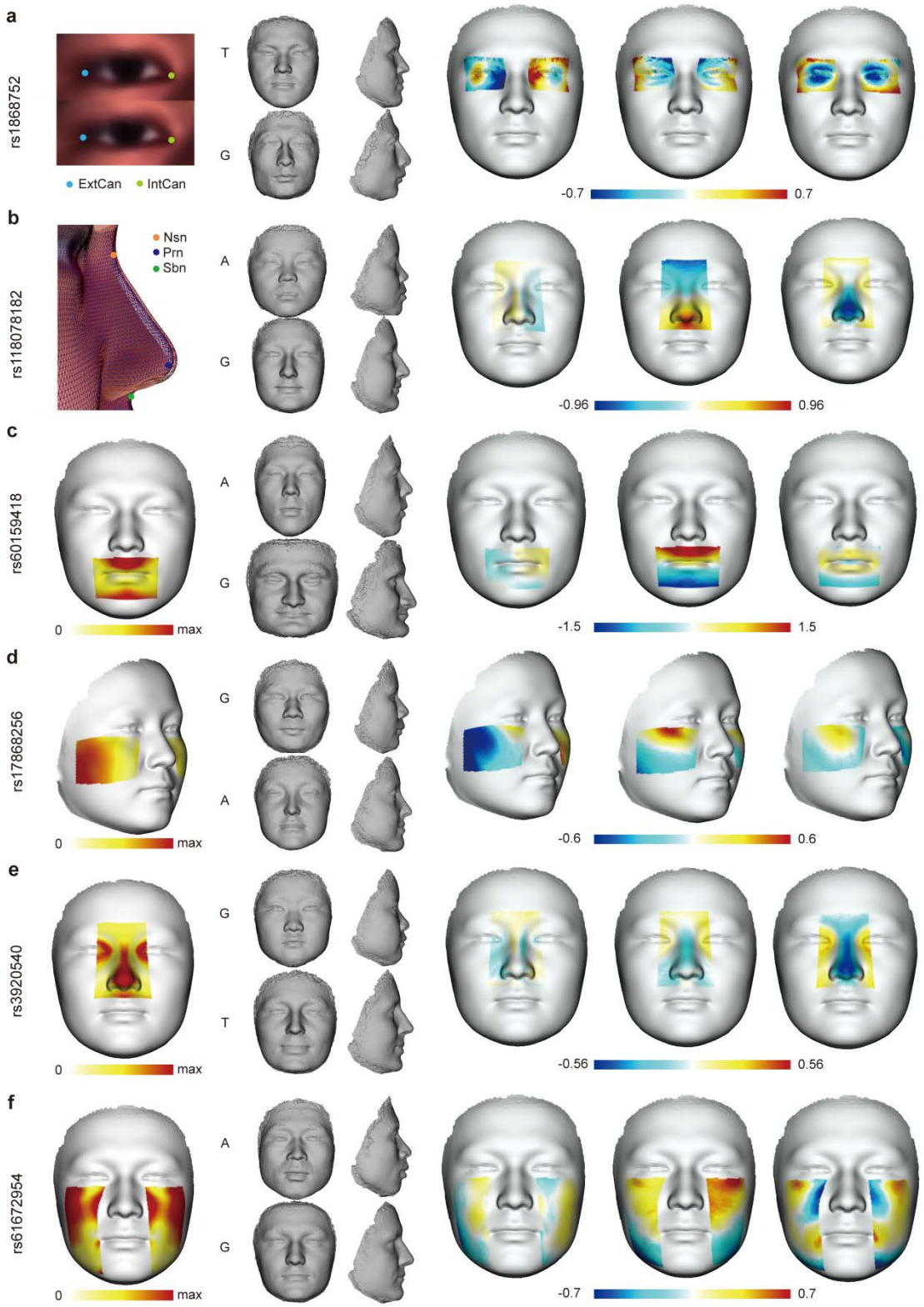

Figure 3 | Heat plots and extrapolated faces of six significant SNPs affected on responding partial shape. (a) the association of rs1868752 with the distance between ExtCan and IntCan, (b) the association of rs118078182 with the distance of Nsn-Prn-Sbn, (c) the association of rs60159418 with the sPC of mouth in males, (d) the association of rs17868256 with the sPLS of cheeks in females, (e) the association of rs3920540 with nasal sPLS in females, (f) the association of rs61672954 with the sPLS of side faces. For each SNP, the first face shows the general effect on the corresponding feature as the displacement of landmarks or meshes. The mid-panel of four

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

miniature faces gives the extrapolations towards the Han trend on the top, or the European trend on the bottom, with the associated allele labeled at the left side. The extrapolated faces were morphed by exemplifying the difference between the average faces of the opposite homozygotes if both are more than 10% frequent in UIG-D, or the major homozygote and the heterozygote if otherwise. The last three faces depict the signed displacement of the average faces of the fore-mentioned genotypes in X, Y and Z axes; obtained by subtracting the average face of European-trend from that of the Han-trend.

Replication studies and meta-analysis. We replicated the six GWAS significant loci in an independent Uyghur cohort (UIG-R) and a Han Chinese cohort (HAN-CZ). The former has the same ethnic background with UIG-D[14, 15, 16] and the latter represents a pool of Han Chinese ancestry from southern China[20]. Face is highly complex, and the effects of genetic variants on face can be subtle and strongly depend on other factors such as ethnicity and gender[12, 21, 22]. On the other hand, face related genetic loci may be pleiotropic, e.g., that a single variant may influence facial morphology on different parts and/or in different ways[6, 7, 12, 23, 24]. In view of this, we defined two types of association replications: the narrow-sense and broad-sense replications. Narrow-sense replication stood for the association signals replicated on exactly the same phenotypic measurement; whereas broad-sense replication required the candidate loci to show evidences of association with any shape changes in the same facial feature. In this study, both the narrow- and broad-sense replications were conditioned in the same gender group as for the discovery panel. As a result (Table 2), the association of rs1868752 with ExtCan-IntCan exhibited narrow-sense replication (P=0.0449) in a dominant model in HAN-CZ. The SNP rs118078182 showed narrow-sense replications in an additive (P=0.00422) and a dominant model (P=0.00309) in UIG-R, in a dominant model (P=0.0167) in HAN-CZ, as well as in the additive model of UIG-R and HAN-CZ combined (P=0.0236). For rs60159418, the narrow-sense replication was tested by projecting the UIG-R and HAN-CZ 3dDFM data to the sPC of mouth where the GWAS signal was found in UIG-D males, and revealed marginal significance in HAN-CZ (P=0.0854) and combined group of UIG-R and HAN-CZ (P=0.0523). For rs17868256, the narrow-sense association in females was successfully replicated for cheek sPLS in UIG-R for both the additive model (P=0.0374) and a dominant model (P=0.0163). The other two candidate loci didn't show evidences of narrow-sense replication.

To systematically test the broad-sense replication, we carried out pair-wise shape distance (PSD) permutation as previous proposed[11] and PLS-based permutation for the 3dDFM data (see Methods). Table 3 summarized the results of broad-sense replications. In general, these tests confirmed the results of narrow-sense replications, showing that rs1868752, rs118078182, rs60159418 and rs17868256 affect the overall shapes of the corresponding features. Furthermore, rs3920540, the nose related locus that failed to replicate in the narrow-sense test, turned out to significantly affect the overall nasal shape in PSD permutation test (P=0.0139) and PLS-based permutation test (P=0.03) in HAN-CZ.

Visualization in UIG-R and HAN-CZ revealed highly consistent effects of the candidate variants as in UIG-D[19] (Fig. 4), despite the distinct ethnicity of HAN-CZ.

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

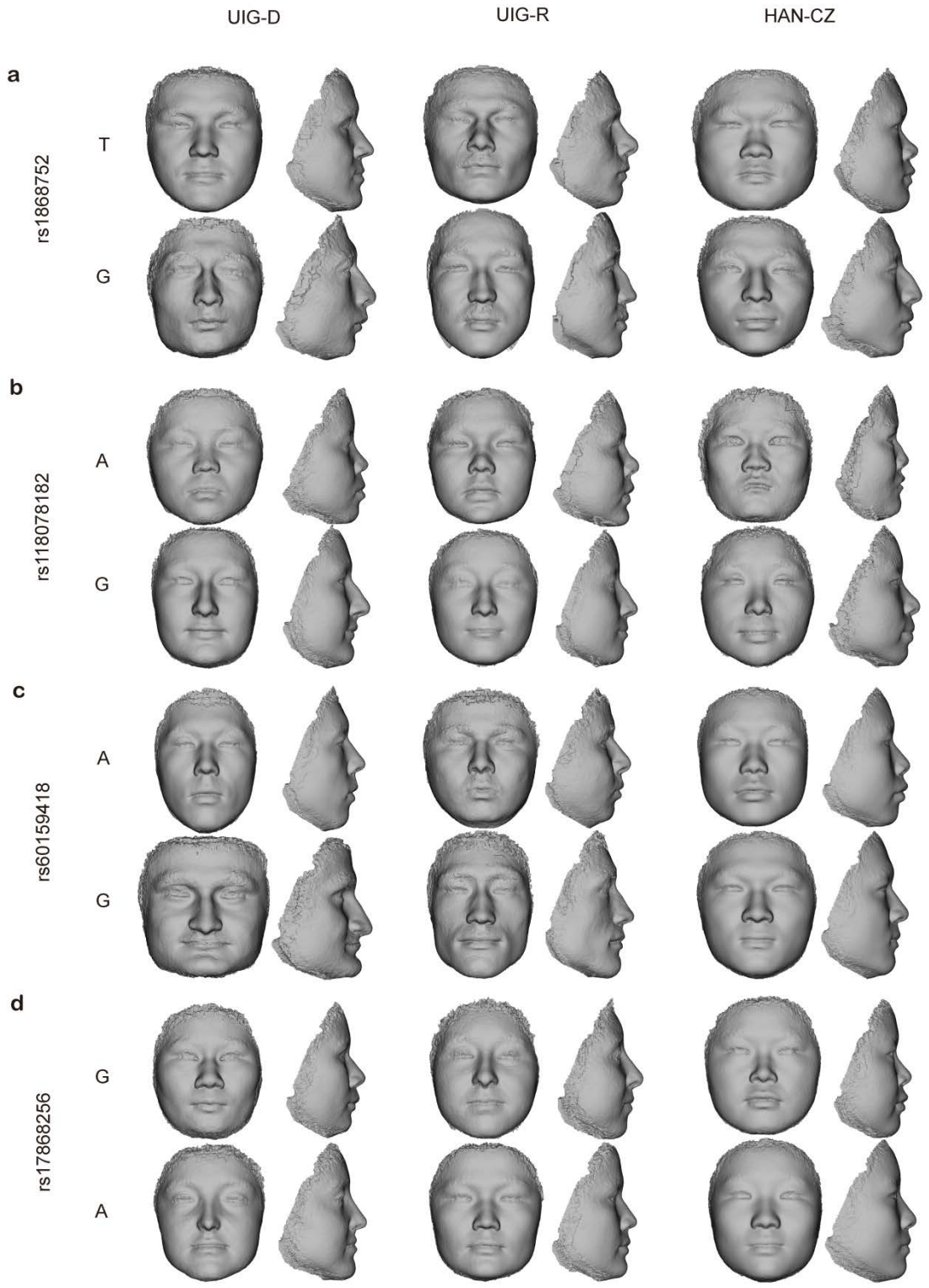

Figure 4 | Visualization in UIG-R and HAN-CZ revealed largely consistent effects of the candidate variants as in UIG-D. For the four significant loci, (a) rs1868752, (b) rs118078182, (c) rs60159418 and (d) rs17868256, we compared the extrapolated faces in UIG-D, UIG-R and HAN-CZ from left to right. For each locus, the top faces in the trend of Han Chinese and the bottom ones are in European trend for the corresponding feature.

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Intriguingly in UIG-R, the effects of some SNPs were strongly persistent not only within the facial features of GWAS signals, but also across the whole face. In HAN-CZ, similar influence on the whole face could also be observed. In particular, rs1868752T was involved in recessive eye-sockets and repressed nasal bridge in all three cohorts (Fig. 4a); rs118078182A seemed to make the whole face more flat and round in addition to its effects on nasal shape (Fig. 4b); Other than affecting the mouth shape (Fig. 4c), rs60159418G was also associated with stronger brow ridges and more protrusive chin among the three groups. These indicated that the identified association signals were authentic, and the candidate variants were in general highly pleiotropic.

Replication of reported variants of facial variation. We explored whether the candidate SNPs that previously reported to affect normal facial variation[6, 7, 8] also showed signals of association in our combined Uyghur cohort (UIG-D + UIG-R). Replications were carried out either on the original or related measurements for 12 SNPs (Supplementary Table 4). Notably, numerous candidate loci were re-validated to varying degrees. Briefly, rs4648379 in *PRDM16*, rs7559271 in *PAX3*, rs2045323 in *DCHS2*, rs17640804 in *GLI3*, rs805722 in *COL17A1* and rs927833 in *PAX1* reported to affect nasal phenotypes in different ethnic groups[6, 7, 8], were found to also modulate normal nasal shape in Uyghurs (Supplementary Table 4). The SNPs rs3827760 in *EDAR* and rs6184 in *GHR* that were previously linked to mandibular shape variation turned out to be significant or marginally significant in our study[8, 24, 25] (Supplementary Table 4). Interestingly, the SNP rs642961 in *IRF6*, previously found to be associated with the mouth shape in Han Chinese females[11, 26], also showed marginal significance (P=0.05197) in Uyghur females but not in males or mixed gender group, implying that the dependence of the genetic effect of rs642961 on gender was shared among different populations.

3D face prediction based on genome-wide SNPs. Hypothetically, if a quantitative trait is highly heritable, a proper model featuring major genetic factors should lead to true prediction[27, 28, 29]. We assumed in advance that SNPs reaching a suggestive genome-wide threshold of P value $< 1 \times 10^{-6}$ were enriched for facial shape related loci. GWA analyses were carried out in UIG-D on a comprehensive list of phenotypes of various perspectives including both divergent traits across Eurasia as well as common shape variations within Uyghurs (see Methods). All SNPs that satisfy the threshold were combined into a panel of 277 top SNPs (Supplementary Table 5, Supplementary Fig. 4). Based on these top SNPs, a simple quantitative model was constructed using UIG-D data. Briefly, for each SNP, a residual face was obtained for each genotype by subtracting the genotype average face by the global average. To compose a predicted face, the 277 residual faces were scaled by a global "effect coefficient" α (see Methods), and then added to the base face (the average face stratified by gender) according to the specific genotypes of the individual of interest. The coefficient α was a scaling factor that minimized the Euclidean distance between predicted and actual faces in the discovery panel (see Methods). The coefficient was

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

determined to be 0.312 in females (α_f) and 0.252 in males (α_m) (Supplementary Fig. 5).

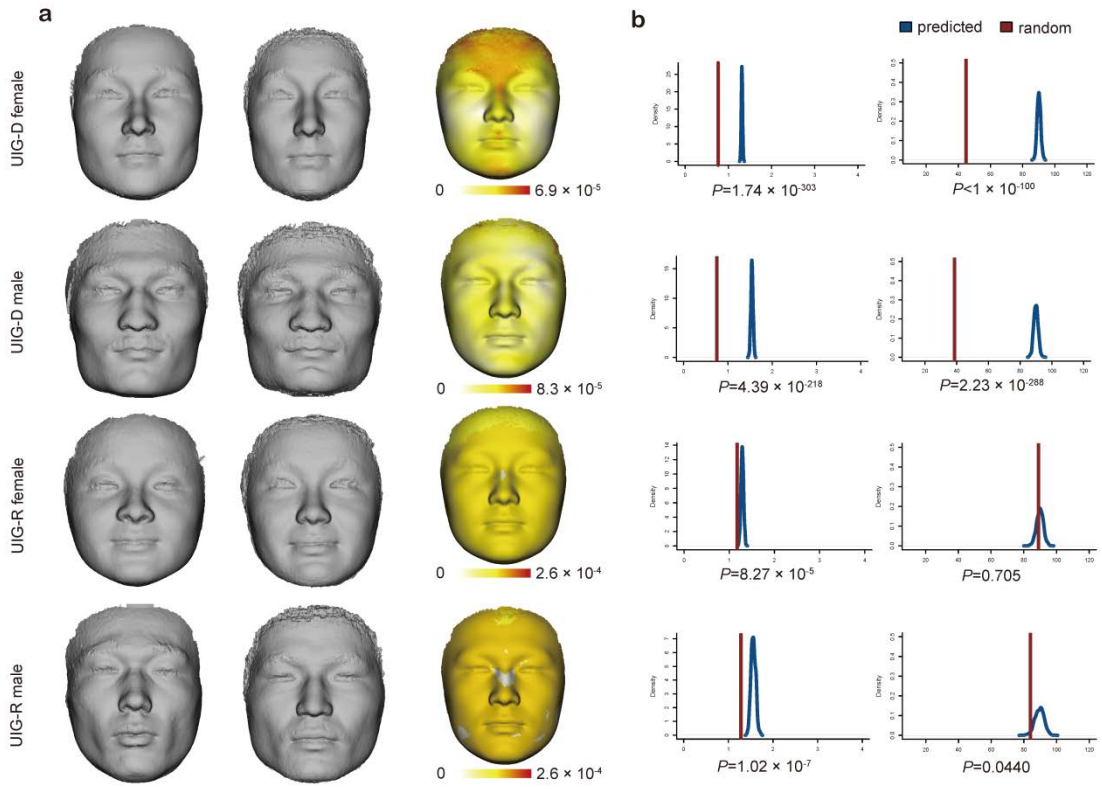

Figure 5 | Test of the prediction model. (a) cases of visualization of actual face (left column), the predicted face (middle column) and the displacement between the pair in heat plot (right column). (b) The average PSD (left column) and SSA (right column) determined for the cohorts (in red) were compared to the random distributions under null hypothesis (in blue). *P* values are the probability of predicted statistic distributed on the relative random normal distribution calculated like normal one-side *P* value.

The prediction model was applied to Uyghur individuals from the discovery panel (UIG-D) as well as the independent replication panel (UIG-R). The similarity between predicted and actual faces for individuals in UIG-D was visually obvious, but ambiguous in UIG-R (Fig. 5a, Supplementary Movie 7). To formally access the resemblance between predicted and actual faces we applied two similarity statistics: the PSD distance was the squared average Euclidean distance for the 3D vertex points, whereas the shape space angle (SSA) was defined as the angle between two shapes in the 3dDFM data space (see Methods). PSD and SSA would both achieve 0 if the two shapes were the same. It was easy to understand that SSA was independent from α, but was determined by the differentiation of "trends" along which two faces transformed from the base face. A SSA of 90 degree stood for statistical independence, whereas a SSA greater than 90 indicated that two faces deviated from the mean face in reverse directions. Tests of shape similarity were carried out in UIG-D and UIG-R separately stratified by gender. In each test, the similarity score was calculated for the pair of prediction/observation of each individual and averaged across all individuals.

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The average score was compared to that of 1000 iterations using randomly generated faces as prediction (see Methods). In UIG-D, the similarity scores for prediction data were greatly lower than that of the random datasets (Fig. 5b). The average PSD scores were approximately half of that for the random data ($P<1 \times 10^{-100}$ in females and males), whereas the average SSA were 44.85 ($P<1 \times 10^{-100}$) and 38.59 ($P<1 \times 10^{-100}$) in females and males respectively, compared to average SSA of around 90 in the random data. Most importantly in the independent panel of UIG-R (Fig. 5b), the average PSD of prediction were significantly lower than the null expectation in both females ($P=8.27 \times 10^{-5}$) and males ($P=1.02 \times 10^{-7}$); the average SSA of prediction ($SSA_{avg}=89.07$) was not different ($P=0.705$) from the null expectation in females, but was slightly and significantly lower ($SSA_{avg}=83.97$, $P=0.0440$) than the null expectation in males. To account for the potential linkage disequilibrium (LD) within the top SNPs, we further trimmed the top SNPs set either by pairwise LD ($r^2<0.8$, 240 top SNPs set) or inter-marker physical distance (<400kb, 209 top SNPs set), the prediction became slightly better and the general trends remained unchanged (Supplementary Fig. 5 and 6).

Intriguingly, face prediction based on genotypes would promote the suspect identification in pragmatic forensic scenarios. We simulated toy forensic scenarios where a single true suspect should be picked up from a group of N (N=2, 3, 5, 8, 10) candidates (Fig. 6a). We supposed that the 3D facial data and genotypes could be obtained from each candidate to allow mathematical comparison between the observed and predicted faces. In all cases with UIG-D individuals, the power of correct identification ranged from 69%~97%, much higher than null expectations regardless of the similarity statistic or gender (Fig. 6b, Supplementary Fig. 7). When UIG-R was used, male consistently showed a highly significant power increase of 4.02% to 5.86% for PSD ($P<1 \times 10^{-48}$) and 4.74% to 7.26% for SAA ($P<1 \times 10^{-73}$), compared to the random expectations; in females, the prediction did not substantially enhance the power of identification, with scenario of N = 10 showing a mild enhancement (PSD: +0.23%, $P=0.268$, SSA: +1.36%, $P=1.45 \times 10^{-10}$) (Fig. 6b, Supplementary Fig. 7). Using the SNP panels controlled of the inter-SNP LD, the power of identification was still much higher, 72% ~ 98% for the panel of 240 SNPs (Supplementary Fig. 8) and 72% ~ 98% for the panel of 209 SNPs (Supplementary Fig. 9), compared to the null expectations in UIG-D. In UIG-R males, the LD controlled SNP panels showed weaker performance in males (PSD power increase 1.37% ~ 6.41% for 240 SNPs and 0.12% ~ 3.09% for 209 SNPs; SAA power increase 2.74%~ 3.7% for 240 SNPs and 2.86% ~ 6.117% for 209 SNPs). In females, neither PSD nor SSA showed increased identification power in the two smaller SNP panels (Supplementary Fig. 8 and 9).

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

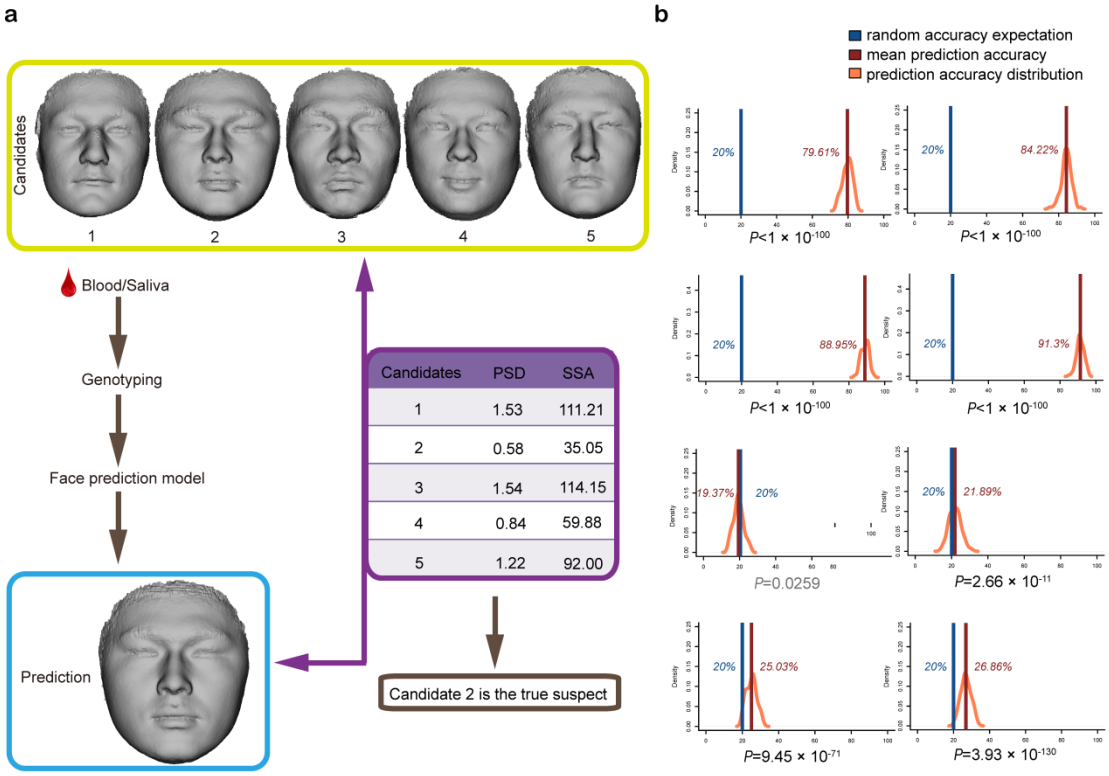

Figure 6 | Evaluation of the face prediction in a hypothetic forensic scenario. (a) the overall scheme of the forensic test. As can be seen, a number of candidates were given and one true suspect is to be selected. A predicted face is determined based on the DNA information and compared to the actual 3dDFM data of the candidates and PSD and SSA are calculated. The face with minimum PSD/SSA is called as the true suspect. (b) PSD and SSA as accuracy statistics are evaluated in UIG-D females, UIG-D males, UIG-R females and UIG-R males from top row to bottom. The accuracy rate based on PSD (left column) and SSA (right column) were determined by examining how many cases of successful identification were achieved in 100 iterations. This process was repeated 200 times to obtain a distribution of accuracy rate (in red) and compared to the expected value of accuracy (in blue) under null hypothesis. *P* values are the probability that whether the number of correct identification using face prediction model is significantly different from the random accuracy expectation (1/N) via Chi-squared test. *P* values in black indicate that the prediction perform significantly better than random draws.

Discussion

To our knowledge, this was the first GWAS aimed at identify genetic loci associated with normal facial variations based on complex 3dDFM data that revealed multiple genetic determinants underlying the European-East Asian facial trait divergence. The genome-wide significant loci were located on independent regions and respectively associated with shape of eyes, nose, mouth, cheeks and side faces. We successfully replicated four loci, rs1868752, rs118078182, rs60159418 and rs17868256 in independent cohorts on the same phenotype measurements. In addition, the association signal of rs3920540 was replicated in broad-sense.

SNP rs1868752 is not located in any gene region. The nearest protein coding gene is the ubiquitin associated and SH3 domain containing B (*UBASH3B*) about 140kb

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

distal, which regulates epidermal growth factor receptor (*EGFR*) and platelet-derived growth factor receptor (*PDGFRA*). The longer eye length, consistent to European traits is associated with the derived allele rs1868752G. However, rs1868752G has a low global frequency (< 10%), and the population frequency in CEU is even lower (~1.5%) than in CHB (~4%), suggesting that this SNP does not play a major role in the Eurasian face differentiation. The SNP rs118078182 has the most consistent association signals across different sample panels and test models. This SNP is an intronic variant in collagen type 23 alpha 1 (*COL23A1*). *COL23A1* codes a non-abundant trans-membrane collagen, primarily found in head, skin, tendon, and kidney[30]. A possible role of spatial/temporal regulation in facial morphogenesis was noted for *COL23A1*[31]. Interestingly the G allele of rs118078182, associated with the European trait of longer and taller nose is almost fixed (99.5%) in CEU compared to the sequentially lower frequencies of ~90% in Uyghur and ~79% in CHB, suggesting that rs118078182 plays a major role in the nasal shape divergence across Eurasia. The other nasal shape related SNP, rs3920540 is approximately 300kb away from the nearest protein coding gene *BMP2*, which is a member of the bone morphogenetic proteins involved in the development of bone and cartilage[32, 33, 34, 35]. The T allele of rs3920540 pertaining to the European trait of taller nose does not seem to differ in frequency (0.86, 0.85 and 0.88 in CEU, UIG and CHB respectively) among the three populations, suggesting that this SNP mainly contribute to the within-group variation of nasal shape. The mouth shape related SNP rs60159418 is situated in an intron of the protocadherin 7 gene (*PCDH7*). *PCDH7* codes an integral membrane protein functioning in cell-cell recognition and adhesion. Previous studies showed that *PCDH7* played a key role in osteoclastogenesis[36], and its homologue gene is a pivotal regulator in the head formation of the mouse embryo[37]. Consistently, the derived G allele that co-occurs with the concaved European mouth shape is almost fixed in CEU (0.96, Table 2), much lower (0.644) in UIG, whereas in CHB the ancestral A allele is the major allele (allele frequency 0.631). These suggest that rs60159418 contributes to the mouth shape differentiation across Eurasia. Notably, rs60159418 is among the most divergent SNPs between CEU and CHB ($Fst_{CEU\text{-}CHB}$=0.446), suggesting an involvement of local adaptations in this region. For the SNP rs17868256, the derived G allele is associated with the Han Chinese trait of higher zygomatic arches and more cambered outwards and backwards zygomatics, and the corresponding allele frequency is also the highest in CHB (0.524), followed by UIG (0.357) and CEU (0.247), as indicates that rs17868256 is involved in the phenotypic divergence in cheeks between Europeans and Han Chinese. The SNP rs61672954, associated with side face shapes was not replicated statistically in HAN-CZ (sample sizes for AA, GA genotypes were too small to analysis in UIG-R), but had the same facial variation patterns among different genotypes as in the discovery UIG-D cohort (Supplementary Fig. 10). So we cannot remove their potential effect on facial shape.

One evident limitation of this study is the relative small sample sizes for the UIG cohorts. A power analysis revealed that tests would be constrained and the association signals would not be highly significant (e.g. all GWAS *P* values>1×10^{-10}). And yet, even given the limited sample sizes we were still able to detect 6 genome-wide

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

association signals, among which five were replicated to various degrees. This may be attributed mainly to the specific study designs: first, as the Uyghur was examined on the phenotypic dimensions where the ancestral groups EUR and HAN differ the most, the search was thus focused on genetic variants of large effect size, rendering higher test power for a given sample size. Second, the 3dDFM data densely annotates each face by over 30,000 vertices, resulting in virtually face phenome data. Based on this, the association signals were scanned both phenome-widely and genome-widely, as would greatly enhance the power of detecting the phenotype-genotype associations.

Several trends are notable involving the genetic architecture of facial morphology. First, no GWAS loci of "major effects" were identified that account for a large portion of phenotypic variance in spite of the strong overall divergence, e.g., in nasal shape across Eurasia[2]. This is in contrast to the case of skin pigmentation whose major genetic factors explain substantial phenotypic variance[38, 39, 40]. This suggests that the human face should be best described by a typical polygenic model of complex trait, characterized by a large number of variants of small effects. Second, most facial shape related variants seem to be pleiotropic. All the candidate loci in our study seem to be associated to the complex shape changes of whole face, not limited to the features of GWAS signals, in similar trends across the three sample cohorts, implying that such dispersive facial changes were induced by genetic variants rather than stochasticity (Fig. 3, Fig. 4). At the individual gene level, *PCDH7* was also known for its versatile functions, related to not only mouth shape but also musical aptitude[41], waist-to-hip ratio[42] and many diseases[36, 37, 43, 44, 45]. Another SNP rs3827760 in *EDAR*, replicated in this study was also known for its broad effects in hair morphology, incisor shape and sweat gland density[23, 24, 46, 47]. Third, the genetic effects of face related loci seem to be shared among different ethnicities[48, 49, 50]. This is evident given that the association signals and facial patterns are in general consistent between Uyghur and Han Chinese (Table 2 and 3, Fig. 4). It may be thus hypothesized that the stereotypic faces of ethnicities are merely result of population stratification of the face-related allele frequencies. For example, a European nose is "big" probably due to the co-segregation of a higher proportion of "big" nose alleles compared to that of an average Han Chinese. Indeed, four (rs118078182, rs60159418, rs61672954 and rs17868256) out of the six candidate SNPs in this study have moderate ($Fst_{CEU\text{-}CHB}$>0.08) to strong (for rs60159418, $Fst_{CEU\text{-}CHB}$=0.466, top 0.003% genome-widely) population differentiation, each of which seems to contribute a gradient to the continuous transition from European to Han Chinese faces.

In the end, we showed that an additive genetic model of whole face shape, based on a set of SNPs of top association signals, would lead to measurable predictive power. In the discovery panel, the prediction is surprisingly accurate: PSD and SSA distances were folds lower than random expectations, and the evaluation in the simulated forensic test rendered accuracy rates of close to 100% (Fig. 5 and 6, Supplementary Fig. 7). It should be noted that this is not a true prediction as the samples used for model construction were again used for test. Nonetheless, it is obvious that our model was able to account for the vast majority of the facial shape variance. Most importantly, we showed that for independent individuals (UIG-R), the

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

prediction model can also construct realistic 3D faces significantly closer to the actual face than random expectation (Fig. 5); tests in the hypothetic forensic scenarios also revealed a robust enhancement of the identification rate (about 4~7%) in males, which is an non-negligible ~0.5 fold increase (6.36%) than random draw (12.5%) for the case of 8 suspects (Supplementary Fig. 7c). To the best of our knowledge, this was the first study that achieved true face predictive power based on pure genetic information[13]. There may be redundant markers within the 277 top SNPs, as the SNP sets filtered for strong LD gave very similar performance (Supplementary Fig. 5 for prediction, Supplementary Fig. 8 and 9 for forensic). Fundamentally, it can be argued that this prediction model is not really additive, as the best-fitting effect coefficient α is far lower than 1 (Supplementary Fig. 5), even for the SNP set (209) well controlled for the physical LD (Supplementary Fig. 5c). This may implicate epi-static interactions between the causal SNPs, or is probably mainly due to a proportion of false-positive signals in the top SNPs set. Further studies specifically designed for examination of the genetic architecture is needed to address this question. Furthermore, we noticed that increasing the number of top SNPs in the prediction model (either for P value $< 1 \times 10^{-5}$ or whole genome SNPs) could not improve the predictive power (data not shown). Such saturation analysis suggests a finite number of loci affecting the normal facial shape. After all, this study clearly demonstrates that face prediction based on DNA is possible.

Methods

Study cohorts. EUR was a resident cohort living in Shanghai with self-reported European ancestry between 16 and 42 years old. The HAN-TZ participants were self-reported Han Chinese samples collected from Taizhou, Jiangsu province. College students of self-reported Han ethnicity from Xiangnan University in Chenzhou, Hunan province were collected as HAN-CZ. The UIG-D and UIG-R were composed of college students of self-reported Uyghurs collected from Xinjiang Medical University in Urumchi, Xinjiang province. The self-reported ancestry information was requested for the last three generations, and individuals with mixed ancestry or missing information were excluded from further analyses. For EUR, a participant was used only if his/her ancestries of the last three generations were all from EU countries (as for 2015) plus Switzerland, Norway and Iceland. Individuals with obvious health problems or any history of facial surgery were ruled out. All sample collection in this study was carried out in accordance with the ethical standards of the ethics committee of the Shanghai Institutes for Biological Sciences (SIBS) and the Declaration of Helsinki, and has been specifically surveyed and approved by SIBS. For publication of samples' materials, the individuals were conformed about the published information and the informed consent forms were signed by themselves. All methods were carried out in accordance with the corresponding guidelines and regulations. A written statement of informed consent was obtained from every participant, with his/her authorizing signature. The participants, whose transformed facial images were used in this study as necessary illustrations of our methodology, have been shown the

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

manuscript and corresponding figures. Aside from the informed consent for data sampling, a consent of publication was shown and explained to each participant and their authorizing signature was obtained as well.

High-density 3D facial images alignment. The 3dMDface® system (www.3dmd.com/3dMDface) was used to collect high-resolution 3D facial images. We first established dense anatomical correspondence across dense surfaces of 3D facial images automatically as described previously[10]. Briefly, 15 salient facial landmarks were annotated automatically based on the principal component analysis (PCA) projection of texture and shape information. A reference face was selected for high image quality and smooth surface, and its mesh was resampled to achieve an even density of one vertex in each 1mm × 1mm grid. There were 32,251 vertices in total for the reference mesh. Afterwards, the reference face was warped to register to each objective face to ensure the complete overlapping of the 15 landmarks via a thin-plate spline (TPS) transformation. The vertices on the reference face then found their closest projections on the sample face to define the samples' new vertices, resulting in a point-to-point correspondence. At last, the Generalized Procrustes analysis (GPA) was used to align the sample grids into a common coordinate system. As a result, we obtained a set of 32,251 3D points to represent each participant's face. Samples with defective images were removed from the study.

Genotyping, quality control and imputation. Genomic DNA extracted from blood samples of UIG-D and UIG-R were genotyped respectively on Illumina Omni ZhongHua-8 and Affymetrix Genome-Wide Human SNP Array 6.0. Quality control were performed using PLINK v1.07[51]. Furthermore, 92 individuals from UIG-D were whole genome sequenced at high-coverage (30×). We didn't consider SNPs on mitochondria. SNPs with MAF <0.01, genotyping rate <90%, or rejection in the Hardy-Weinberg Equilibrium test with $P<1 \times 10^{-6}$ were omitted from the study. Genomic ancestry was detected using EIGENSTRAT 5.0.2[52, 53] with CHB and CEU from 1000 Genomes Project[54] (1KG phase1 release v2) to remove samples who were not Uyghurs ancestry. Samples with genotype missing rate >0.1 were removed. Data were further examined by pairwise IBD estimation, inbreeding coefficients and sex status to remove individuals of close genetic relationships or wrong sex information. After quality control, a total of 847,046 SNPs were confirmed in 694 UIG-D and 758,453 SNPs were confirmed in the 171 UIG-R. Genotypes in UIG-R were pre-phased with SHAPEIT v2.r790[55]. Imputation of UIG-D and UIG-R were then carried out by combining the UIG-D and UIG-R genotype data as well as all 1,092 individuals from 1000 Genomes Projects[54] and the 90 sequencing samples from UIG-D using IMPUTE2[56]. Genomic DNA of HAN-CZ was extracted from saliva referring to a modified Phenol-chloroform protocol[57]. Targeted genotyping for genome-wide significant SNPs were performed by SNaPshot multiplex system on an ABI3130xl genetic analyzer by Genesky Biotech, Shanghai, China.

Extraction of candidate phenotypes. We carried out principle component analysis

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(PCA) based and partial least square (PLS) based dimension reduction in sex-stratified groups. The 3dDFM data from EUR, UIG-D and HAN-TZ cohorts was decomposed into perpendicular dimensions using PCA. It was straight-forward to extract the PCs that maximized the segregation between EUR and HAN-TZ by Student's Test, which were then named as the segregating PC (sPC) for different facial features. On the other hand, PLS was used to generate a regression model that maximized the covariance between facial shape and ancestry[58, 59, 60, 61]. Briefly, in the PLS equation we labeled the individuals from EUR as 1 and Han-TZ as 2 in the response. The predictor featured the $3 \times n$ matrix of the 3dDFM data, where n stood for the number of vertices in the corresponding facial feature. Cross-validation (CV) was used to determine how many of the top PLS components should be used in the PLS regression model. The PLS regression model achieved the highest prediction when the root mean squared error of prediction (RMSEP) showed the turning point in CV. We increased the number c of PLS components in a stepwise manner until the adjusted RMSEP reached the inflexion. The segregating PLS model (sPLS) was then fixed with c_{adjcv} components that corresponded to the turning adjusted RMSEP. The PCA and PLS analyses were carried out using the prcomp[62] and plsr functions[63] in R packages.

Replications. For narrow-sense replication, the 3dDFM data from UIG-R and HAN-CZ were either projected to the sPC spaces or input to the sPLS models defined in UIG-D, to obtain the corresponding phenotype scores in narrow-sense replication.

Significant SNPs were also tested for broad-sense replication. For PLS-based permutation, genotypes (0, 1, 2 in additive model and 0, 1 in dominant model) of each SNP were taken as response. Principle components (PCs) that explained 99% of the total variance of each facial feature were adopted as predictor. A scheme of Leave One Out (LOO) was used in which *N-1* individuals were used as training and the left-out was used as test. This was repeated until every individual was used as a test sample. The training set was used to build a PLS regression model whose optimal number of components was determined by RMSEP under cross-validation, as described before. Based on this model, a predicted genotype was given to the test sample. By such analogy, every sample obtained a predicted genotype. Afterwards, the correlation between actual genotype and predicted genotype can be calculated. Next, we performed a permutation procedure to control the supervised effect of PLS[58, 59, 60, 61]. Briefly, the genotypes were reshuffled randomly among the samples for 1,000 times, followed by the same procedure as above to calculate correlations between the actual and predicted genotypes. The null distribution of correlations can be thus established based on the permutation sets. The corrected PLS-based permutation P value was generated by ranking the empirical raw correlation against the null distribution.

For the pair-wise shape distance (PSD) permutation among genotypes[11], we calculated the Euclidean distances between the mean shapes of any two genotype groups. The mean shape can be denoted as a vector,

$$s = [x_1, y_1, z_1, \dots, x_n, y_n, z_n] \qquad (1)$$

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where x_i, y_i, z_i was the X, Y, Z coordinate values of the ith points, n was the number of points.

For each two genotypes mean shape s and s', the PSD was defined as

$$PSD = \frac{\sum_{i=1}^{n} d_i^2}{3n} \qquad (2)$$

where $d_i^2 = (x_i - x_i')^2 + (y_i - y_i')^2 + (z_i - z_i')^2$.

We also randomly permuted the genotypes among the samples and then calculated the PSD between the peudo-genotype groups. The PSD scores resulted from permutation formed the null distribution. The one-side *P* value was calculated by the proportion of permuted PSD smaller than or equal to the observed PSD.

Face prediction

GWA analyses on various phenotypes: To capture the majority of the facial shape variance for prediction, numerous biometric spaces including inter-landmark features, PCA, PLS were examined for both ancestry-divergent traits and within-Uyghur variations. For each phenotype, GWA analyses were carried out in males and females separately as well as in the sex-combined group. Inter-landmark measurements were made up of 36 distances or angles: namely the distances of LAla-Prn-RAla, Nsn-Prn-Sbn, LIntCan-Nsn-RIntCan, Sbn-ULipP, LLipCn-RLipCn, ULipP-Stm-LLipP, ExtCan-IntCan, LIntCan-RIntCan, Sbn-ULipP-ChiP, height of Nsn to eyes, Nsn-Prn, LAla-RAla, Sbn-ULipP, Nsn-Sbn, LExtCan-RExtCan, LIntCan-RIntCan, LLipCn-RLipCn, ULipP-LLipP, ULipP-LLipCn-LLipP-RLipCn, ULipP-Stm-LLipP, LLipP-ChiP, height of Prn to Nsn-Sbn, height of Prn to LAla-RAla, Z-axis score of Prn, Z-axis score of Ala, Z-axis displacement of Prn to Ala; and the angles of LAla-Prn-RAla, Nsn-Prn-Sbn, LIntCan-Nsn-RIntCan, Prn-Sbn-ULipP, ULipP-Stm-LLipP, LLipCn-Stm-RLipCn, Sbn-ULipP-Stm, Stm-LLipP-ChiP, LLipCn-LLipP-RLipCn and eyes slope. Furthermore, the high-dimensional 3dDFM data of the six facial features (Supplementary Fig. 1) were examined by PCA and PLS. Aside from sPC and sPLS traits mentioned in the main text, the within-Uyghur variations were analyzed by applying PCA on the UIG-D 3dDFM data alone, and the top three PCs (PC1, PC2, PC3) were extracted as individual phenotypes that together explained ~60% variance. For the PLS regression, other than sPLS, the model when the adjusted RMSEP reached the inflexion as well as the model that uses the top two PLS components were also used as phenotypes for each feature. In the end, 234 phenotypes were defined and analyzed by GWA, and any autosomal SNPs that passed the threshold of $P<1 \times 10^{-6}$ in any test were combined into the panel of top SNPs.

Prediction face composition: For each of the top SNPs, we divided the 3dDFM images into three genotype groups, and the within-group mean shapes were obtained as $\mu_g'$where g is the genotype. A residual face can be obtained for each genotype by subtracting the genotype mean by the global cohort mean shape as,

$$r_g = \mu_g' - \mu \qquad (3)$$

where μ was the global mean. We superimposed all the re-scaled residual faces of each SNP according to the observed genotypes, onto the global mean to construct a

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

face of prediction:

$$f = \mu + \alpha \sum r_{g(i)} \qquad (4)$$

where α was the effect coefficient and i is the SNP index. The coefficient α was proposed to maximize the similarity of the predicted and actual faces. Briefly, a predicted face was calculated for every individual in the discovery panel as described by equation (4), where α varies. PSD was used to calculate the distance between the predicted face and actual face for each individual. The value of α was determined when the minimum average PSD was achieved within a gender group. This same α was also used for the independent cohort as well as the randomly simulated faces.

To generate a random face, the genotypes were randomly sampled from the known frequencies of genotypes in the combined UIG cohort, the predicted face was then obtained as indicated by equation (4), using the randomly generated genotype set.

Evaluation of the face prediction: For each sample, we described the degrees of similarity between prediction and actual face in the respects of PSD and shape space angle (SSA). PSD described the squared average Euclidean distance for 32,251 vertices from the predicted face to actual face. The SSA was the angle between the predicted and actual faces in the 3 × 32,251 face-space. For a group of N individuals, the average PSD score (PSD_{avg}) and average SSA score (SSA_{avg}) was calculated and used as the test statistic. To test whether the face prediction within this group was significantly better than random, we generated the same number (N) of random faces as described above and calculated the corresponding PSD_{avg} and SSA_{avg}. This was repeated 1,000 times and the distributions of PSD_{avg} and SSA_{avg} under the null hypothesis were compared to the observed PSD_{avg} and SSA_{avg}. Assuming normality of the null distributions, the probability of the observed statistics was either given as the percentage of random cases scored higher than the observed if any, or the standard difference of the observation from the null expectation $t = (s - m)/\sigma$ was queried against the normal distribution function (pnorm function in the R statistic package) to give the estimated P value, where s was the observed value, and m and σ were the mean and standard deviation of the null distribution.

Forensic scenarios simulation: In a simulative forensic scenario, we randomly set N (sample size) individuals as the hypothetical candidates, one of them being the true suspect. The PSD and SSA scores were obtained between the actual and predicted faces for each individual as depicted in Fig. 6a. If the true suspect happened to have the lowest PSD or SSA score among the N candidates, the identification was called successful, otherwise failure. We repeated such test 20,000 times, each assuming N individuals by random. The prediction accuracy rate can be summarized as the proportion of successful identification. Assuming no predictive power, it's unambiguous that the rate of picking up the correct suspect purely by chance was $1/N$. The observed accuracy rate among the 20,000 cases was formally tested against the expected value of $1/N$ by Chi-squared test and the corresponding P value was given.

Supplementary information

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Information includes nine figures, five tables and seven movies.

Acknowledgements

We are grateful to all the volunteers taking part in this study. We thank Chen Liu, Nianhao Cheng for the assistance in sample collection; thank Wei Qian for PLS-based Methods discussion; thank Bertram Müller-Myhsok, Benno Pütz for non-linear methods discussion; thank David Allen Hughes Jr. for manuscript polishing and comments. This work was supported by the Max-Planck-Gesellschaft Partner Group Grant, the National Science Foundation of China (31371267, 31322030, 91331108). S.X. acknowledges financial support from the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) (XDB13040100) and the National Natural Science Foundation of China (NSFC) grant (91331204), the National Science Fund for Distinguished Young Scholars (31525014), and the Program of Shanghai Academic Research Leader (16XD1404700); S.W. acknowledges the support of a National Thousand Young Talents Award and a Max Planck-CAS Paul Gerson Unna Independent Research Group Leadership Award; Y.L. acknowledges support from NSFC grant (31501011) and Science and Technology Commission of Shanghai Municipality (STCSM) (14YF1406800); H.L. acknowledges support from STCSM grant (16YF1413900); Y.G. acknowledges support from NSFC grant (31260263).

Author Contributions

Conceived and designed the study: K.T. Performed the study, summarized results and wrote the manuscript: L.Q., K.T. Advise the study and modified the manuscript: L.Q., S.W., S.X., K.T. Contributed reagents/material/software: Y.Y., P.F., S.H., H.Z., J.T., Y.L., H.L., D.L., S.W., J.G., S.P., L.J., Y.G., S.W., S.X.

Competing financial interests

The authors declare no competing financial interests.

References

1. Klimentidis YC, Shriver MD. Estimating genetic ancestry proportions from faces. *PLoS One* **4**, e4460 (2009).

2. Guo J*, et al.* Variation and signatures of selection on the human face. *J Hum Evol* **75**, 143-152 (2014).

3. Weinberg SM, Parsons TE, Marazita ML, Maher BS. Heritability of Face Shape in Twins: A Preliminary Study using 3D Stereophotogrammetry and Geometric Morphometrics. *Dent 3000* **1**, (2013).

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4. Johannsdottir B, Thorarinsson F, Thordarson A, Magnusson TE. Heritability of craniofacial characteristics between parents and offspring estimated from lateral cephalograms. *American journal of orthodontics and dentofacial orthopedics* **127**, 200-207; quiz 260-201 (2005).

5. Martinez-Abadias N, Esparza M, Sjovold T, Gonzalez-Jose R, Santos M, Hernandez M. Heritability of human cranial dimensions: comparing the evolvability of different cranial regions. *J Anat* **214**, 19-35 (2009).

6. Liu F*, et al.* A genome-wide association study identifies five loci influencing facial morphology in Europeans. *PLoS genetics* **8**, e1002932 (2012).

7. Paternoster L*, et al.* Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position. *American journal of human genetics* **90**, 478-485 (2012).

8. Adhikari K*, et al.* A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation. *Nat Commun* **7**, 11616 (2016).

9. TJ H, BF B, P H, HW P. Estimating average growth trajectories in shape-space using kernel smoothing. *IEEE Transactions on Medical Imaging* **22**, 6 (2003).

10. Guo J, Mei X, Tang K. Automatic landmark annotation and dense correspondence registration for 3D human facial images. *BMC bioinformatics* **14**, 232 (2013).

11. Peng S*, et al.* Detecting genetic association of common human facial morphological variation using high density 3D image registration. *PLoS computational biology* **9**, e1003375 (2013).

12. Claes P*, et al.* Modeling 3D Facial Shape from DNA. *PLoS genetics* **10**, (2014).

13. Claes P, Hill H, Shriver MD. Toward DNA-based facial composites: preliminary results and validation. *Forensic science international Genetics* **13**, 208-216 (2014).

14. Xu S, Jin L. A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. *American journal of human genetics* **83**, 322-336 (2008).

15. Xu S, Huang W, Qian J, Jin L. Analysis of genomic admixture in Uyghur and its implication in mapping strategy. *American journal of human genetics* **82**, 883-894 (2008).

16. Xu S, Jin W, Jin L. Haplotype-sharing analysis showing Uyghurs are unlikely genetic donors. *Molecular biology and evolution* **26**, 2197-2206 (2009).

17. Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv*, (2014).

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

18. Pruim RJ*, et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336-2337 (2010).

19. Hu S*, et al.* Signatures of personality on dense 3D facial images. *BioRxiv*, (2016).

20. Xu S*, et al.* Genomic dissection of population substructure of Han Chinese and its implication in association studies. *American journal of human genetics* **85**, 762-774 (2009).

21. Talbert L, Kau CH, Christou T, Vlachos C, Souccar N. A 3D analysis of Caucasian and African American facial morphologies in a US population. *J Orthod* **41**, 19-29 (2014).

22. Hopman SM, Merks JH, Suttie M, Hennekam RC, Hammond P. Face shape differs in phylogenetically related populations. *European journal of human genetics* **22**, 1268-1271 (2014).

23. Kamberov YG*, et al.* Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* **152**, 691-702 (2013).

24. Peng Q*, et al.* EDARV370A associated facial characteristics in Uyghur population revealing further pleiotropic effects. *Hum Genet* **135**, 99-108 (2016).

25. Zhou J*, et al.* The growth hormone receptor gene is associated with mandibular height in a Chinese population. *J Dent Res* **84**, (2005).

26. Rahimov F*, et al.* Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. *Nat Genet* **40**, 1341-1347 (2008).

27. Yang J*, et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-569 (2010).

28. Yang J*, et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet* **43**, 519-525 (2011).

29. Wood AR*, et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173-1186 (2014).

30. Koch M*, et al.* Expression of type XXIII collagen mRNA and protein. *The Journal of biological chemistry* **281**, 21546-21557 (2006).

31. Weiguo Feng*, et al.* Spatial and Temporal Analysis of Gene Expression during Growth and Fusion of the Mouse Facial Prominences. *PLoS One* **4**, (2009).

32. D C, M Z, GR M. Bone morphogenetic proteins. *Growth Factors* **22**, 8 (2004).

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

33. PJ M, F D, E H. Regulation of human cranial osteoblast phenotype by FGF-2, FGFR-2 and BMP-2 signaling. *Histol Histopathol* **17**, 8 (2002).

34. MR U. Bone: formation by autoinduction. *Science* **150**, 6 (1965).

35. Geiger M. Collagen sponges for bone regeneration with rhBMP-2. *Advanced Drug Delivery Reviews* **55**, 1613-1629 (2003).

36. Nakamura H*, et al.* Global epigenomic analysis indicates protocadherin-7 activates osteoclastogenesis by promoting cell-cell fusion. *Biochemical and biophysical research communications* **455**, 305-311 (2014).

37. Fossat N*, et al.* Context-specific function of the LIM homeobox 1 transcription factor in head formation of the mouse embryo. *Development* **142**, 2069-2079 (2015).

38. Relethford JH. Apportionment of global human genetic diversity based on craniometrics and skin color. *American Journal of Physical Anthropology* **118**, (2002).

39. JH R. Global patterns of isolation by distance based on genetic and morphological data. *Hum Biol* **76**, 14 (2004).

40. Myles S, Somel M, Tang K, Kelso J, Stoneking M. Identifying genes underlying skin pigmentation differences among human populations. *Hum Genet* **120**, 613-621 (2007).

41. Oikkonen J*, et al.* A genome-wide linkage and association study of musical aptitude identifies loci containing genes related to inner ear development and neurocognitive functions. *Molecular Psychiatry* **20**, 7 (2015).

42. Surakka I*, et al.* A genome-wide screen for interactions reveals a new locus on 4p15 modifying the effect of waist-to-hip ratio on total cholesterol. *PLoS genetics* **7**, e1002333 (2011).

43. Scannell Bryan M*, et al.* Genome-wide association studies and heritability estimates of body mass index related phenotypes in Bangladeshi adults. *PLoS One* **9**, e105062 (2014).

44. Lemaitre RN*, et al.* Genetic loci associated with circulating levels of very long-chain saturated fatty acids. *Journal of lipid research* **56**, 176-184 (2015).

45. Comuzzie AG*, et al.* Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. *PLoS One* **7**, e51954 (2012).

46. Tan J, Yang Y, Tang K, Sabeti PC, Jin L, Wang S. The adaptive variant EDARV370A is associated with straight hair in East Asians. *Hum Genet* **132**, 1187-1191 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

47. Adhikari K*, et al.* A genome-wide association study identifies multiple loci for variation in human ear morphology. *Nat Commun* **6**, 7500 (2015).

48. Bhatia G*, et al.* Genome-wide scan of 29,141 African Americans finds no evidence of directional selection since admixture. *American journal of human genetics* **95**, 437-444 (2014).

49. Baynam G*, et al.* The facial evolution: looking backward and moving forward. *Hum Mutat* **34**, 14-22 (2013).

50. Sheehan MJ, Nachman MW. Morphological and population genomic evidence that human faces have evolved to signal individual identity. *Nat Commun* **5**, 4800 (2014).

51. Purcell S*, et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559-575 (2007).

52. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909 (2006).

53. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS genetics* **2**, e190 (2006).

54. Genomes Project C*, et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).

55. Delaneau O, Marchini J, Genomes Project C, Genomes Project C. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun* **5**, 3934 (2014).

56. Bryan N. Howie, Peter Donnelly, Marchini J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS genetics* **5**, (2009).

57. Hoff-Olsen P, Meva°g B, Staalstrøm E, Hovde B, Egeland T, Olaisen B. Extraction of DNA from decomposed human tissue: An evaluation of five extraction methods for short tandem repeat typing. *Forensic science international* **105**, 12 (1999).

58. Indahl U. A twist to partial least squares regression. *Journal of Chemometrics* **19**, 32-44 (2005).

59. Liland KH, Indahl UG. Powered partial least squares discriminant analysis. *Journal of Chemometrics* **23**, 7-18 (2009).

bioRxiv preprint doi: https://doi.org/10.1101/062950; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

60. Indahl UG, Liland KH, Naes T. Canonical partial least squares-a unified PLS approach to classification and regression problems. *Journal of Chemometrics* **23**, 495-504 (2009).

61. Krishnan A, Williams LJ, McIntosh AR, Abdi H. Partial Least Squares (PLS) methods for neuroimaging: a tutorial and review. *Neuroimage* **56**, 455-475 (2011).

62. Venables, N. W, Ripley BD. Modern Applied Statistics with S. *Springer-Verlag*, (2002).

63. B-H M, R W. The pls package: Principal component and partial least squares regression in R. *Journal of Statistical Software* **18**, 24 (2007).