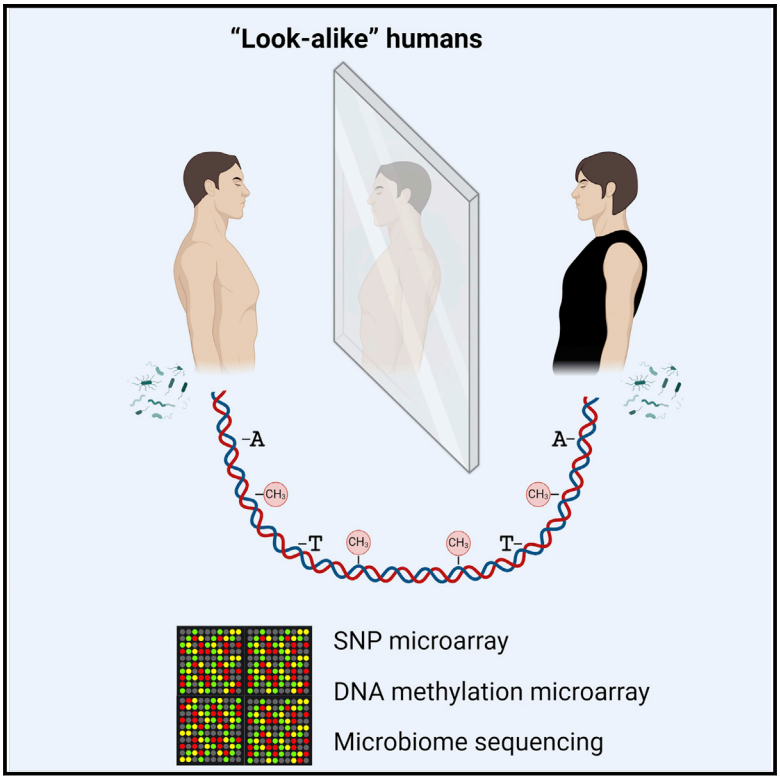
Cell Reports

Look-alike humans identified by facial recognition algorithms show genetic similarities

Graphical abstract



Highlights

- Facial recognition algorithms identify "look-alike" humans for multiomics studies
- Intrapair look-alikes share common genetic sequences such as face trait variants
- DNA methylation and microbiome profiles only contribute modestly to human likeness
- The identified SNPs impact physical and behavioral phenotypes beyond facial features

Authors

Ricky S. Joshi, Maria Rigau, Carlos A. García-Prieto, ..., Xavier Binefa, Alfonso Valencia, Manel Esteller

Correspondence

mesteller@carrerasresearch.org

In brief

We recognize each other by relying on our face uniqueness. However, there are humans with uncanny resemblance. Joshi et al. reported that look-alike pairs identified by facial recognition algorithms share genotypes but not DNA methylomes and microbiomes. The identified SNPs also provide a readout of other anthropomorphic and behavioral characteristics.

Joshi et al., 2022, Cell Reports *40*, 111257
August 23, 2022
https://doi.org/10.1016/j.celrep.2022.111257

Cell Reports

CellPress
OPEN ACCESS

Report

Look-alike humans identified by facial recognition algorithms show genetic similarities

Ricky S. Joshi,[1,10] Maria Rigau,[2,10] Carlos A. García-Prieto,[1,2,10] Manuel Castro de Moura,[1] David Piñeyro,[1,3] Sebastian Moran,[1] Veronica Davalos,[1] Pablo Carrión,[4] Manuel Ferrando-Bernal,[4] Iñigo Olalde,[4] Carles Lalueza-Fox,[4] Arcadi Navarro,[4,5,6] Carles Fernández-Tena,[7] Decky Aspandi,[8] Federico M. Sukno,[8] Xavier Binefa,[8] Alfonso Valencia,[2,6] and Manel Esteller[1,3,6,9,11,*]

[1]Josep Carreras Leukaemia Research Institute (IJC), Badalona, 08916 Barcelona, Spain
[2]Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain
[3]Centro de Investigacion Biomedica en Red Cancer (CIBERONC), 28029 Madrid, Spain
[4]Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), 08003 Barcelona, Spain
[5]Centre for Genomic Regulation (CNAG-CRG), 08003 Barcelona, Catalonia, Spain
[6]Institucio Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain
[7]Herta Security, S.L., 08037 Barcelona, Spain
[8]Departament de Tecnologies de la Informació i les Comunicaciones (DTIC), Universitat Pompeu Fabra (UPF), 08018 Barcelona, Spain
[9]Physiological Sciences Department, School of Medicine and Health Sciences, University of Barcelona (UB), L'Hospitalet, 08907 Barcelona, Spain
[10]These authors contributed equally
[11]Lead contact
*Correspondence: mesteller@carrerasresearch.org
https://doi.org/10.1016/j.celrep.2022.111257

SUMMARY

The human face is one of the most visible features of our unique identity as individuals. Interestingly, monozygotic twins share almost identical facial traits and the same DNA sequence but could exhibit differences in other biometrical parameters. The expansion of the world wide web and the possibility to exchange pictures of humans across the planet has increased the number of people identified online as virtual twins or doubles that are not family related. Herein, we have characterized in detail a set of "look-alike" humans, defined by facial recognition algorithms, for their multiomics landscape. We report that these individuals share similar genotypes and differ in their DNA methylation and microbiome landscape. These results not only provide insights about the genetics that determine our face but also might have implications for the establishment of other human anthropometric properties and even personality characteristics.

INTRODUCTION

The discussion about the relevance of "nature versus nurture," or, in a similar manner, of "genotype versus phenotype," in human biology and medicine is a long-standing issue that still remains largely unsolved. Relevant studies in this area include our original observation that monozygotic twins show epigenetic differences (Fraga et al., 2005), understood as the chemical marks such as DNA methylation and histone modifications that regulate gene expression, that might explain different population traits and distinct penetrance of diseases in these people, a finding supported in later studies (Kaminsky et al., 2009), including The NASA Twins Study (Garrett-Bakelman et al., 2019). These questions can be more easily addressed in experimental models where the researcher can intervene, such as the Agouti mice (Wolff et al., 1998) and cloned animals (Rideout et al., 2001), whereas in humans, the investigator has a more passive role, waiting for the right sample to appear. In this regard, one of the most documented cases is the Dutch famine at the end of WWII that was associated with less DNA methylation of the imprinted *IGF2* gene compared with their unexposed, same-sex siblings (Heijmans et al., 2008).

Human individual identity also relates to biological properties and environment. In this regard, the way we initially recognize each other relies often on our unique face, and there is a sophisticated brain code to distinguish facial identities (Tsao et al., 2006; Chang and Tsao, 2017; Quian Quiroga, 2017). This explains why so commonly twins catch our attention and are used to understand how the balance between nature and nurture generates a phenotype. Here, we present a study that, on a molecular level, aims to characterize random human beings that objectively share facial features. This extraordinary set of individuals, characterized by their high likeliness, are what are called, in lay-language, look-alike humans, unknown twins, twin strangers, doubles, or doppelgänger, in German. This unique set of samples has allowed us to study how genomics,

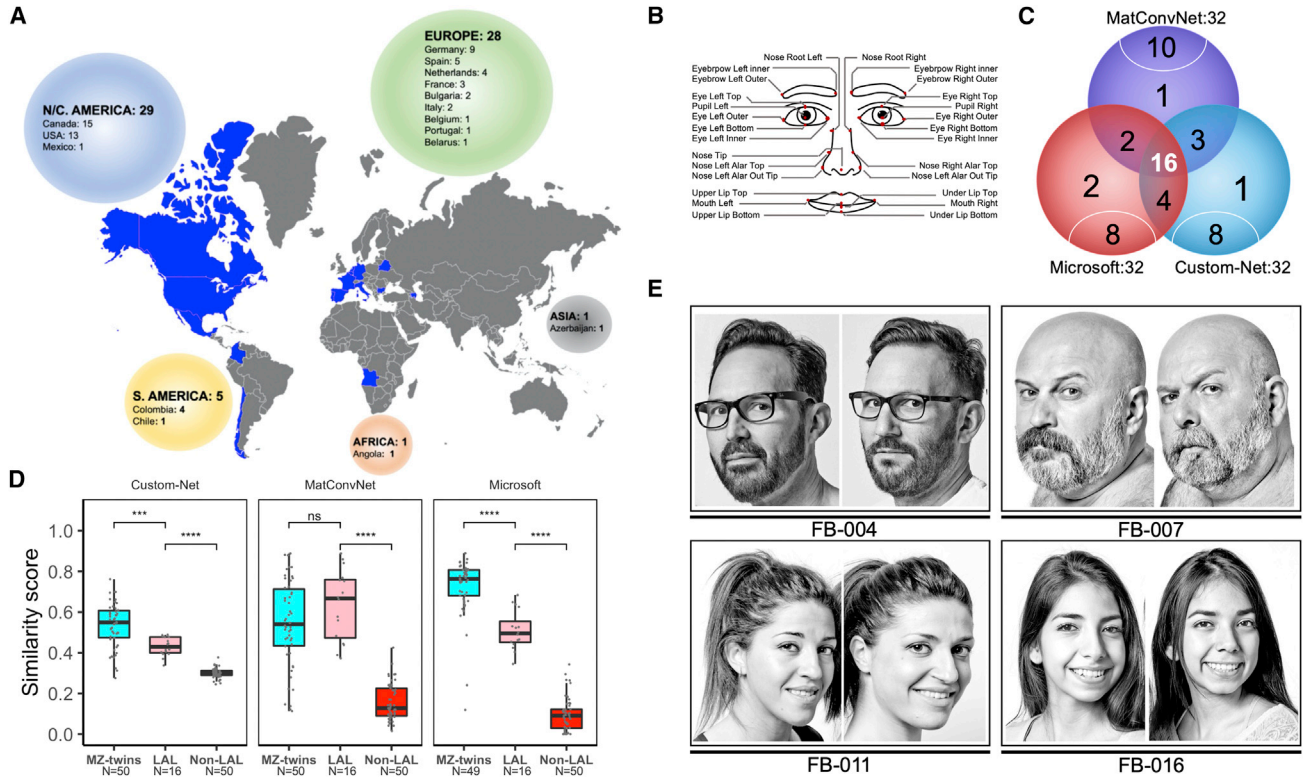


Figure 1. Recruitment and objective determination of look-alike human pairs

(A) Representation of the global worldwide distribution of 32 look-alike pairs (n = 64) in this study.

(B) 27 facial parameters by which the Microsoft Oxford Project face API (Microsoft) objectively performs face detection.

(C) Venn diagram showing the number of look-alike pairs discerned and jointly identified in the three facial recognition programs: MatConvNet, Custom-Net, and Microsoft. Numbers within the semi-circle present the pairs that did not cluster in each software.

(D) Boxplots showing unbiased quantitative similarity scores comparing each facial recognition software (MatConvNet, Custom-Net, Microsoft) for monozygotic twins (MZs; blue), look-alike pairs (LALs; rose), and random non-LALs (red). The x axis represents the different cohorts analyzed. The y axis exhibits similarity scores measured between 0 and 1. 1 represents identical facial image, and 0 represents two totally different photographic entities. "N" indicates the number of couples. Differences calculated using two-sided Mann-Whitney-Wilcoxon test: ****$p < 0.0001$; ***$p < 0.001$; ns, non-significant.

(E) Photographic examples of LALs used in this study.

epigenomics, and microbiomics can contribute to human resemblance. Our study provides a rare insight into human likeness by showing that people with extreme look-alike faces share common genotypes, whereas they are discordant at their epigenome and microbiome. Genomics clusters them together, and the rest set them apart. These findings do not only provide clues about the genetic setting associated with our facial aspect, and probably other traits of our body and personality, but also highlight how much of what we are, and what defines us, is really inherited or instead is acquired during our lifetime.

RESULTS

Facial recognition algorithms and multiomics approaches for look-alike humans

Human doubles were recruited from the photographic work of François Brunelle, a Canadian artist who has been obtaining worldwide pictures of look-alikes since 1999 (http://www.francoisbrunelle.com/webn/e-project.html). We obtained headshot pictures of thirty-two candidate look-alike couples. All participants completed a comprehensive biometric and lifestyle questionnaire in their native language (English, Spanish, and French) (Methods S1). Their geographic locations are shown in Figure 1A. We first determined an objective measure of "likeness" for the candidate double pairs. We used three different methods of facial recognition: the custom deep convolutional neural network Custom-Net, (www.hertasecurity.com), the MatConvNet algorithm (Vedaldi and Lenc 2015), and the Microsoft Oxford Project face API (https://azure.microsoft.com/es-es/services/cognitive-services/face/) (STAR Methods). We used three methods because each system can yield variable results, and we selected those systems to reflect the diversity of possible outcomes. MatConvNet was designed for facial classification, Custom-Net for surveillance, and Microsoft API for generalized facial analysis. These models have millions of learned parameters and have been trained with millions of facial images from thousands of subjects, in a variety of unconstrained situations: differences in pose, hairstyle, expression, age, and accessories within a subject. Thus, the impact of these attributes is likely minimal. Each software provides a facial similarity score between

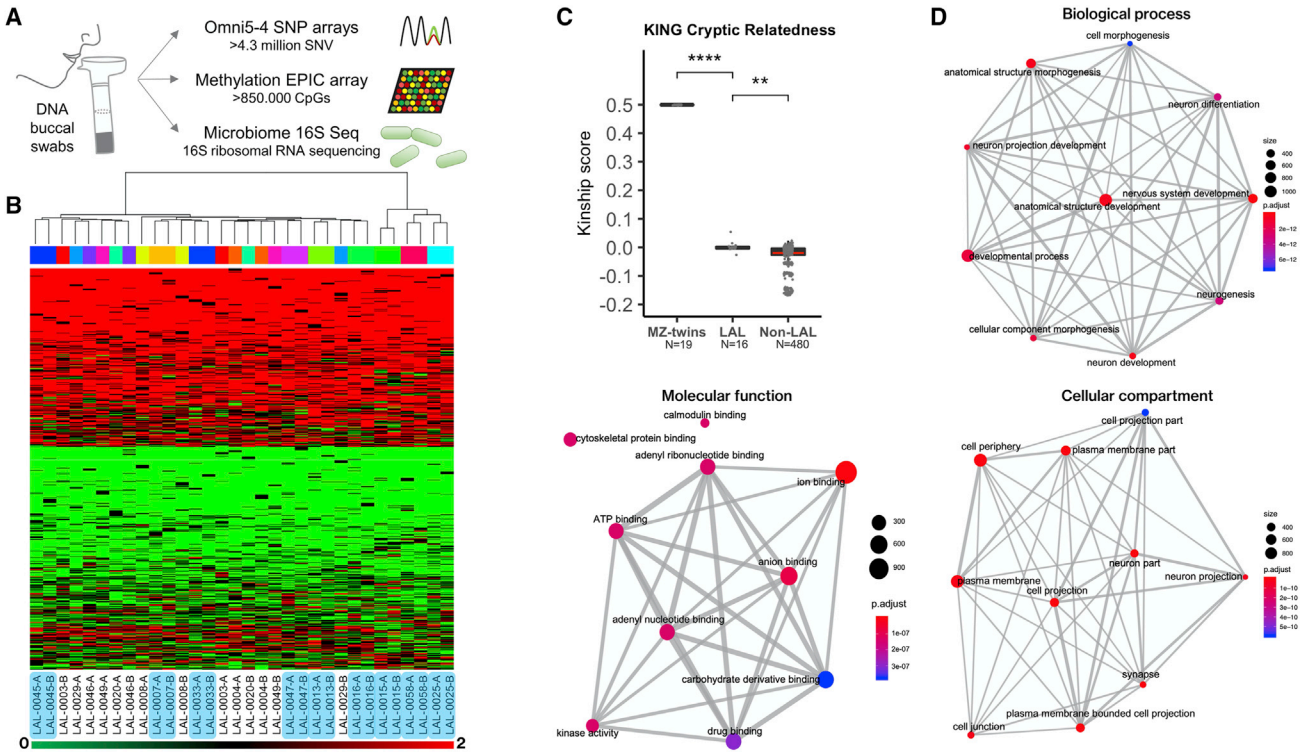


Figure 2. Genetic analysis of look-alike human pairs

(A) Saliva DNA was obtained from 32 LALs recruited to this study. DNA was subjected to genotyping (Omni5-4 SNP arrays Illumina), DNA methylation (Infinium MethylationEPIC arrays, Illumina), and microbiome analysis (16S Metagenomics sequencing, Illumina).

(B) Heatmap of hierarchical genetic clustering with bootstrap of genome-wide SNP genotyping arrays in the 16 LALs. Genotype clustering was performed using Euclidean distances and Ward.D2 cluster method. Blue rectangles represent 9 LALs that unbiasedly clustered. 0 = homozygous reference SNPs (green), 1 = heterozygous SNPs (black), and 2 = homozygous alternate SNPs (red).

(C) Boxplot showing Kinship scores between MZs, LALs, and random non-LALs. Kinship scores range between −0.2 (it represents two unrelated individuals) and 0.5 (it represents duplicated genotypes and MZs). "N" indicates the number of couples. Differences calculated using two-sided Student's t test: ****$p < 0.0001$; **$p < 0.01$.

(D) Gene Ontology (GO) analysis performed using all SNPs found to be shared in all LALs (19,277 SNPs in 3,730 genes). GO enrichments were ran using EnrichGO R package for the 3,730 genes, and the top 10 most significant hits are plotted in network graphs. GO terms are presented with circles. The size and color of each circle represents numbers of genes in each GO term and its statistical significance, respectively. The gray lines represent the interaction of genes, and the thickness is proportional to the number of genes interacting in each GO term. GO subcategories are presented: Biological Process, Cellular Component, and Molecular Function.

0 and 1, where 1 is the same facial image and 0 is two different entities. Comparisons are pairwise, with every image compared with every other image. As an example of the parameters computed, the 27 face landmarks of the Microsoft algorithm are shown in Figure 1B. The results obtained from the different combinations of each approach are shown in a Venn diagram in Figure 1C. Interestingly, the number of pairs that were considered to be correlated by at least two of the facial models was very high (25 out of total 32, >75%), closer to the human ability to recognize identical twins (Biswas et al., 2011). Most importantly, we found that 16 of the original 32 (50%) look-alike pairs were matched by all three facial recognition systems. As an internal positive control for high similarity score, we ran the three facial recognition software in monozygotic twin photograph images from the University of Notre Dame twins database 2009/2010 (https://cvrl.nd.edu/projects/data/). Importantly, similarity scor es from the 16 look-alike couples were similar to those obtained from monozygotic twins according to MatConvNet and significantly higher than those observed in random non-look-alike pairs (Figure 1D). Thus, these highly look-alike humans were the focus of our further research. Illustrative examples of these "double" individuals are shown in Figure 1E.

Saliva DNA for these cases was analyzed by multiomics at three levels of biological information: genome, by means of an SNP microarray that interrogates 4,327,108 genetic variants selected from the International HapMap and 1,000 Genomes Projects, which target genetic variation down to 1% minor allele frequency (MAF) (Xing et al., 2016); epigenome, using a DNA methylation microarray that studies over 850,000 CpG sites (Moran et al., 2016); and microbiome, by ribosomal RNA direct sequencing (Klindworth et al., 2013) (Figure 2A; STAR Methods).

Genomic characterization of look-alike humans

Genomic analyses of these 16 couples provided a striking result: more than half (9 of 16, 56.2%) of these look-alike pairs clustered together in the unsupervised clustering heatmap with bootstrap

(Figure 2B). These nine couples were denominated as "ultra" look-alike. K-means algorithm represented by principal-component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) also showed that the look-alike couples that clustered by the unsupervised clustering heatmap analysis were in close proximity (Figure S1), indicating a likely genotyping resemblance of the studied pairs. In contrast, the 16 candidate look-alike cases that did not cluster by the three facial recognition (FR) networks (Figure 1C) showed that only one pair clustered together (1 of 16, 6.2%) (Figure S1).

We studied two possible confounding factors: population stratification (ancestry) and kinship. Using KING Relationship Inference (Manichaikul et al., 2010) to determine kinship scores, we discarded the possibility of unknown familial relationships (first and second degree) between look-alike pairs (Figure 2C). We observed that look-alike pairs were more similar to non-look-alike pairs than to monozygotic twins (Figure 2C); supporting that look-alike pairing in the SNP clustering is not related to familyhood genotype but instead to a distinct subset of genetic similarity. Using PLINK (Purcell et al., 2007) (STAR Methods), close kinship could be excluded in almost all cases: only one pair share SNPs in proportions that could be compatible with up to third-degree relatives and only one pair share a long (>10 cM) identity by descent (IBD) segment that could suggest co-ancestry in the last few hundreds of years. Interestingly, the latter is a French-Canadian pair, a population known to have experienced a dramatic founder effect in the 17th century. Importantly, when we conducted all the downstream analyses without this French-Canadian pair, the remaining eight ultra-look-alike pairs clustered together (Figure S2). The detailed kinship assessment data are provided in Table S1.

Related to population stratification, among the 16 look-alike pairs, 13 were of European ancestry, 1 Hispanic, 1 East Asian, and 1 Central-South Asian. Although background genetic ancestry is a principal determinant for genetic variance between human populations, we observed that of the 13 White look-alike pairs, 7 (54%) did not cluster genetically, suggesting alternative purposes for shared genetic variation between look-alike pairs. To further determine ancestry, genotyping of the 16 look-alike cases was performed using GenomeStudio v.2.0.5 to create PACKPED Plink files (STAR Methods). Their genomic data were merged with 1,980 West Eurasian, Asian, and Native American individuals genotyped in the Affymetrix Human Origins (HO) array (Lazaridis et al., 2014), where the remaining dataset held 175,469 common SNPs. PCA was generated with the HO individuals (Figure S3) and look-alike individuals (Figure S3B for West Eurasia and Figure S3C for West Eurasia, Asia, and America) (Price et al., 2006; Patterson et al., 2006) (STAR Methods). We observed that almost all the look-alike pairs cluster close to each other according to their countries of origin (or self-attributed ethnic background) (Figure S3). However, they are not more closely related than other pairs of individuals from the same populations taken at random. The detailed population stratification data are provided in Table S1.

Among the 9 couples of ultra-look-alikes, 19,277 SNP positions annotated for 3,730 genes (Table S2) were defined as SNPs with shared genotypes in each look-alike pair. These SNPs correspond to non-monomorphic positions in which every pair of ultra-look-alikes shared the genotype. For example, where one individual in a pair was heterozygous for a given SNP, the corresponding individual in the pair was also heterozygous. This genotype match must be consistent across all pairs for an SNP to be considered shared and therefore represented indicative SNPs relevant for look-alike resemblance. The number of shared SNP positions was significantly higher compared with random non-look-alike pairs in the studied population ($p < 2.2 \times 10^{-16}$, Pearson's chi-squared test). Taking into account ethnicity, shared SNP positions by the European ultra-look-alike pairs was significantly higher compared with random non-look-alike pairs in the studied population ($p = 0.03$, Pearson's chi-squared test). For the remaining three ethnicities, only one individual from each group was available in our dataset. Thus, we interrogated the individuals genotyped in the 1000 Genomes database (https://www.internationalgenome.org/). The number of shared SNP positions by the Hispanic ultra-look-alike pair was significantly higher compared with random individual pairs from the same ethnicity ($p < 2.2 \times 10^{-16}$, Pearson's chi-squared test). No significant enrichment was observed for the remaining two couples, one East Asian and one Central-South Asian. Importantly, only 16 variants of the 19,277 SNPs (0.08%) selected from the ultra-look-alikes presented a linkage disequilibrium detected by iterative pruning analysis (Weir et al., 2014).

The identified genetic variants might have a profound impact on the degree of similitude between the phenotype of humans. Using the clusterProfiler R package (Yu et al., 2012), we performed gene enrichment analyses using the list of look-alike SNPs compared with the background of all genes annotated in the SNP microarray. We observed an enrichment for Gene Ontology (GO) Biological Processes related to anatomical, developmental, and adhesion terms (Figure 2D; Table S3), in addition to ion and anion binding for GO-Molecular function (gene subsets related to bone and skin properties) and many cellular compartments. Enrichment analysis using the DAVID signature database collection noted that the most significantly enhanced ontology was "cell junction," a critical determinant of tissue morphology (Table S4). To evaluate the face genes enrichment in our selected 19,277 SNPs corresponding to 3,730 genes (Table S2), we gather all the genes related with face traits from recent data (Claes et al., 2018; Xiong et al., 2019; White et al., 2021), Facebase dataset (https://www.facebase.org/), and Genome-wide Association Study (GWAS) Central (study HGVST1841, http://www.gwascentral.org) and applied a hypergeometric test and a Monte Carlo simulation using 10,000 iterations (STAR Methods). In no iteration of random set of genes did we observe a number equal to or higher than the face genes represented in our 19,277 SNP selection ($p < 1e{-}4$). We observed a total of 1,794 face genes in our 19,277 SNP selection, constituting 26% of all the face genes present in the array (hypergeometric test p: 6.31e−172; Monte Carlo empirical $p < 1e{-}4$). When we added the reported face associated SNPs to our 19,277 SNPs, we observed that 11 of the 16 (68.7%) look-alike pairs clustered together (Figure S4), therefore adding two new couples.

The study of the functional nature of the SNPs loci shared by the ultra-look-alikes showed that 171 caused amino acid

changes, affecting 158 genes (Table S5). GOrilla analysis for GO-Molecular function found an enrichment in anion transport descriptors (Table S3). Using the GWAS catalog database (https://www.ebi.ac.uk/gwas/), we found that 113 SNPs corresponded to 130 GWAS associations and 84 traits (Table S6). These last traits included many related to facial determinants or physical features such as cleft palate/lip, eye color, hip circumference, body height, waist-hip ratio, balding measurement, and alopecia (Table S6) with an enrichment for lip and forehead morphology, body mass index, bone mineral density, and attached earlobe (Table S6). We observed an enrichment of traits that included the word morphologytagged to the terms nose, lip, mouth, facial, cranial vault, forehead, hair, and cheekbone (Fisher's exact test, odds ratio [OR] = 4.2, p = 0.04). Using the GWAS Central database (http://www.gwascentral.org), we found an enrichment (OR = 1.2782, p = 0.0007364) for SNPs associated with human facial variation (Adhikari et al., 2016). The analyses of the look-alike SNPs according to trait in GWAS Central showed an enrichment for the phenotype names "lip" (OR = 1.8321, p = 0.000327) and "forehead" (OR = 1.886, p = 0.010389). The identified look-alike SNPs were also enriched (OR = 2.201156, p = 0.04884) for genes included in the FaceBase dataset (https://www.facebase.org/). Finally, we studied the overlap between the herein discovered look-alike SNPs and expression quantitative trait loci (eQTLs). Using the Genotype-Tissue Expression (GTEx) Portal (https://www.gtexportal.org/home/), we observed that look-alike SNPs were more frequently associated with gene-expression changes than expected by random chance (Fisher's exact test, OR = 1.1, p = 0.0001). The enrichment was observed among different morphological structures and organs (Table S6). We also used the stratified linkage disequilibrium score regression (S-LDSC) (Finucane et al., 2015) to determine the enrichment of GWAS signals from the GWAS catalog for our SNPs. We observed that these SNPs were overrepresented for the pronasale-right chelion (enrichment score [ES] = 13.84, p = 0.018) and pronasale-left chelion (ES = 12.26, p = 0.04) face traits (Figure S4) (Xiong et al., 2019). The SNPs were also overrepresented for features that define 63 facial segments (Hoskens et al., 2021) considering the entire, mid, and outer face (p < 0.05) (Figure S4). These data indicate that the 19,277 characterized SNPs exert a major impact in the way the face of humans is defined.

The SNP microarray can also be used to determine copy-number variations (CNVs) (Feber et al., 2014). Unsupervised clustering heatmap with bootstrap clustered only one couple together of the 16 look-alikes according to CNVs (Figure 3A). Interestingly, three CNVs were shared by three look-alike pairs (Table S6), including a locus in chromosome 11 that targets genes involved in craniofacial dysmorphic features such as HYLS1 (Mee et al., 2005).

Other multiomics views of look-alike humans

Similar "identities" of look-alikes could also reside in other "omic" components such as the DNA methylome and the microbiome. According to DNA methylation patterns, only one of the sixteen (6.25%) look-alike pairs matched both individuals together, as shown in the unsupervised clustering heatmap (Figure 3B). This couple also clustered together according to SNP genotyping (Figure 2B). The comparison of DNA methylation patterns among the nine look-alike couples with the observed genetic overlap (Figure 2B) only clustered one additional pair (Figure S4). K-means algorithm represented using PCA and the t-SNE plot did not show significant clustering (Figure S5). Thus, overall, human look-alikes are diverse in their epigenome settings.

However, two avenues might provide a role for DNA methylation in facial morphology: epigenetic age and methylation QTL (meQTLs). The aging process changes facial morphology, and DNA methylation is used as a proxy for "biological age" that can or can not be directly related to the "chronological age." One example is the premature epigenetic aging observed in carriers of viral infections (Esteban-Cantos et al., 2021; Cao et al., 2022). We have calculated the intrapair absolute age differences in our 16 look-alike cohort according to chronological age (date of birth) or epigenetic age (DNA methylation clock) (Hannum et al., 2013). We found no differences in intrapair chronological age between the ultra-look-alike group and the non-ultra-look-alike group. In contrast, intrapair "epigenetic" age differences were smaller among ultra-look-alike pairs compared with the non-ultra-look-alike group (two-sided Mann-Whitney-Wilcoxon test, p = 0.0052) (Figure S6). DNA methylation is also associated with genetic variation (Villicaña and Bell, 2021) and could contribute to individual similarity acting as meQTLs. Using the methylation status of 1,379 CpG sites located within a window of +100 bp from the identified 19,277 SNPs, we observed that 3 of the 16 (18.7%) look-alike pairs clustered together (Figure S6). All three of these pairs were among the 9 ultra-look-alike couples (Figure 2B). Thus, DNA methylation, as a marker of biological age and meQTL, can also provide phenotypic commonality for ultra-look-alikes.

A similar scenario was found for the microbiome. From a qualitative standpoint (alpha diversity), according to the type of bacteria present in the studied oral sample (STAR Methods), only one look-alike pair clustered together (Figure 3C). This couple did not cluster together according to SNP genotyping (Figure 2B). From a quantitative standpoint, according to the amount of each bacteria strand present (STAR Methods), we found clustering of one look-alike pair (6.25%, 1 of 16) (Figure 3D). This couple also paired together by unsupervised SNP clustering (Figure 2B). The study of the nine couples with SNP similarity did not provide further pairing of look-alikes (Figure S6). K-means algorithm illustrated by PCA and t-SNE did not demonstrate clustering (Figure S7). Thus, look-alikes do not mostly share a microbiome. However, oral microbiome relates to obesity (Yang et al., 2019), and fat in the face could relate to similarities. We found that intrapair weight differences were smaller among ultra-look-alike pairs compared with non-ultra-look-alike pairs (two-sided Student's t test test, p = 0.035) (Figure S7). Thus, it is possible that the oral microbiome, through its relation to fat content, contributes to look-alike phenotypes.

Traits of look-alike humans beyond facial features

The likeness between the identified human pairs is not limited to the shared facial traits. All the recruited participants in the study completed a comprehensive biometric and lifestyle questionnaire (Methods S1), and the collected information is summarized in

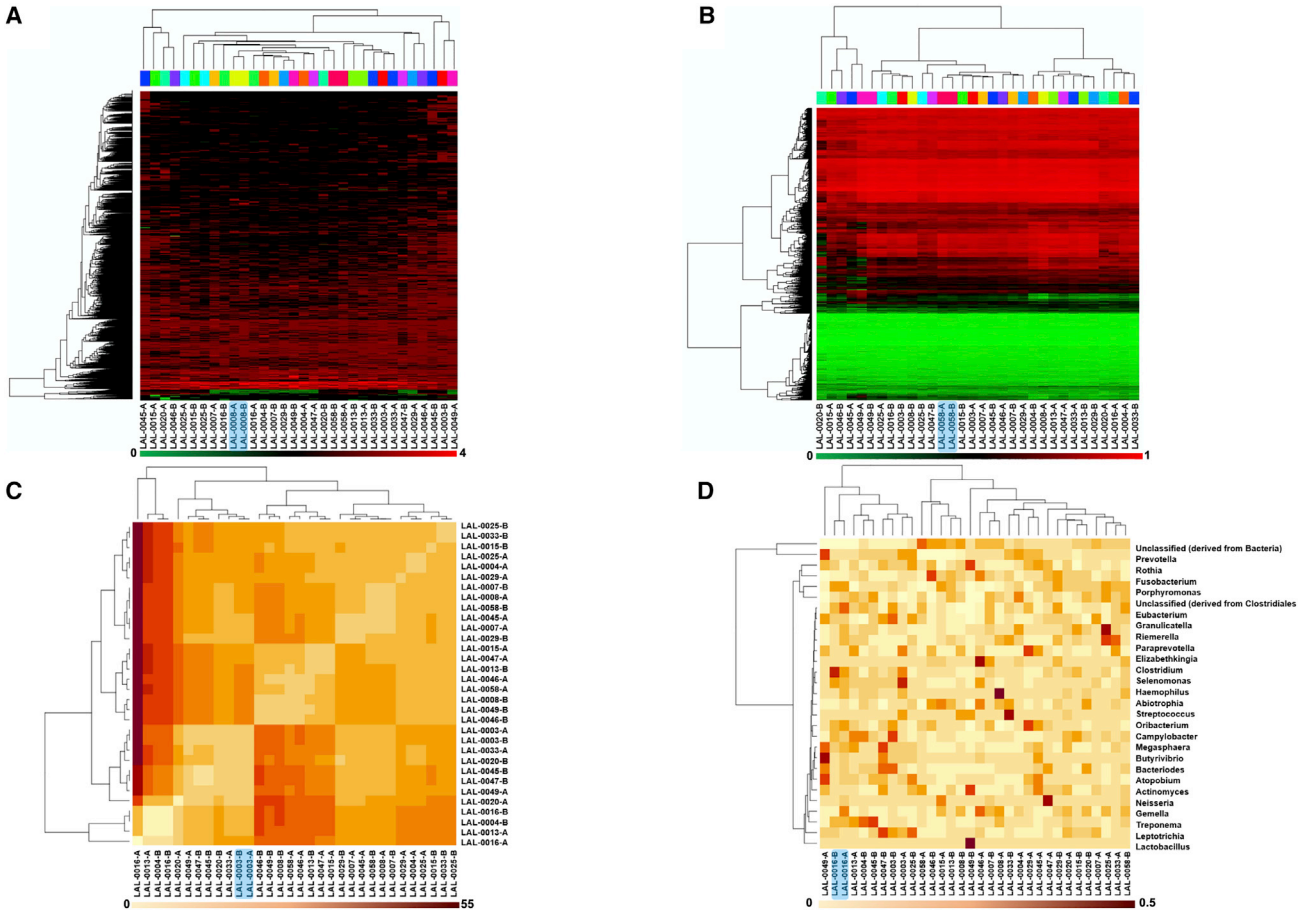


Figure 3. Copy-number variation, DNA methylation, and microbiome analysis of LALs

(A) Heatmap shows the hierarchical clustering of the samples based on the copy number (scale of 0–4) of all copy-number variation (CNV) regions, defined as regions in which at least one individual carried a different copy number. A random selection of one-fifth of such CNV regions is represented in this plot, but the clustering of samples had been obtained considering all CNV regions. The blue rectangle represents a LAL that clusters together.

(B) Heatmap shows unsupervised genome-wide DNA methylation hierarchical clustering with bootstrap of the 16 LALs, using the methylation β-values obtained from MethylationEPIC arrays. A random selection of 5000 CpGs is represented. Colors represent a continuous quantification of methylation beta values at each CpG site, where green highlights unmethylated CpGs (0), black, 50% methylated CpGs (0.5), and red, fully methylated CpGs (1). Clustered look-alikes are shown in a blue rectangle.

(C and D) Microbiome analysis of 16 LALs. Heatmaps show the distances from differences in pairwise bacterial counts of species found in the microbiome of each LAL (variation in alpha diversity scores) of counts from 0–55 (3C) and relative proportions of the taxonomic profiles at the genus level (3D) for each sample calculated on a scale of 0–0.5. Only the most represented genera are shown. Meta-genomic clustering of each look-alike sample was constructed using Euclidean distances and Ward.D2 hierarchical cluster method. Blue rectangle represents LALs whose microbiome is closely related.

Figure 4A. Overall, 68 parameters (Table S7) were included and converted to numerical or logical (0/1) variables (STAR Methods, (custom scripts GitHub: https://github.com/mesteller-bioinfolab/lookalike). The input curated questionnaire is shown in Table S7. We used a cosine similarity method (STAR Methods) to calculate likeness between the studied individuals according to the questionnaire answers. Studying the original 32 look-alike couples, we observed that the 16 look-alike pairs that matched together by all three facial recognition software showed shorter Euclidean distances within pairs ($p = 0.03475$) and higher cosine similarity scores ($p = 0.00321$) than those pairs that did not match by the facial algorithms (Figure 4B). According to their SNPs, the 16 look-alike pairs showed shorter Euclidean distances compared with those pairs that did not match by the three facial algorithms ($p = 0.00006$) (Figure 4B). Examples of independent questionnaire variables (such as height, weight, smoking habit, or level of education) further demonstrate that look-alike pairs are closer than non-look-alike pairs (Figure 4C). Thus, humans with a similar face might also share a more comprehensive physical, and probably behavioral, phenotype that relates to their shared genetic variants. Our study supports the concept of heritability estimation that individuals correlated at the phenotype level share a significant number of genotypic correlations (Visscher et al., 2008). Our results are germane to the ongoing efforts to predict biometric traits from genomic data (Lippert et al., 2017) and the diagnosis of genetic disorders using facial analysis technologies (Gripp et al., 2016; Hadj-Rabia et al., 2017; Hsieh et al., 2019; Gurovich et al., 2019).

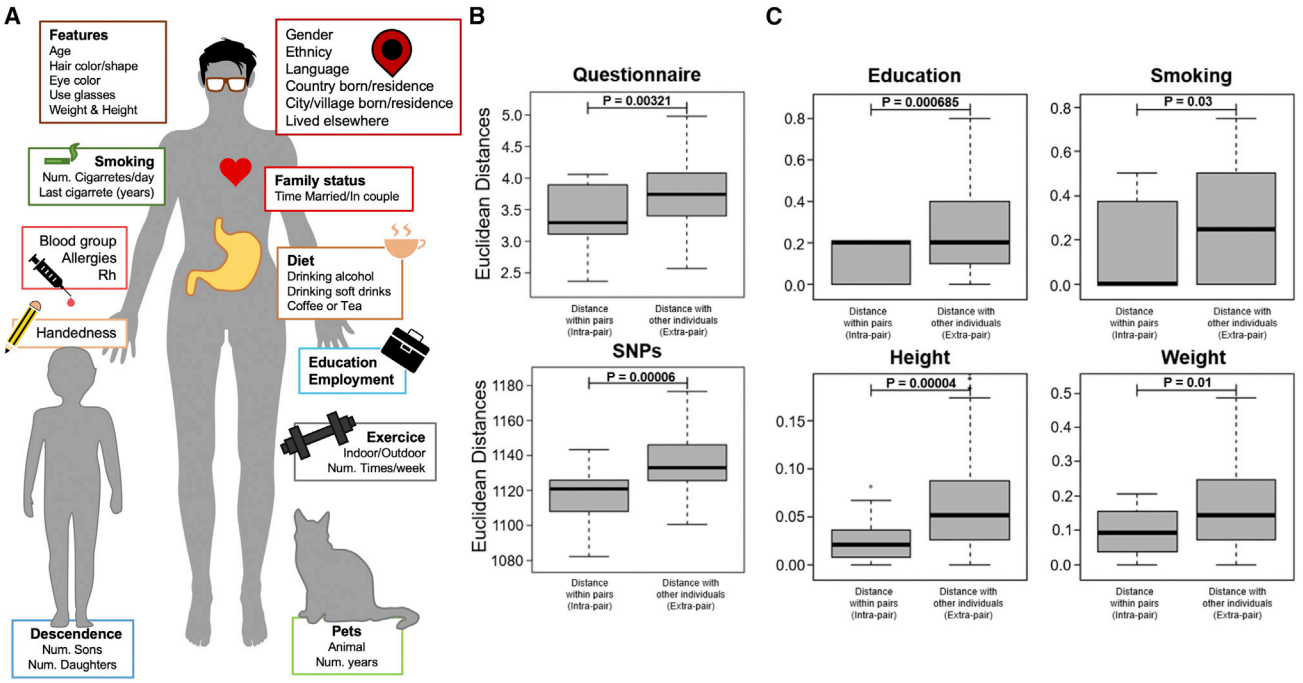


Figure 4. Biometric and lifestyle analysis of LALs using cosine similarity scores

(A) Representation of the biometric and lifestyle parameters considered to calculate cosine similarity scores.

(B) Euclidean distances between the individuals from a pair (intra-pair distance) compared with the distance between individuals from different pairs (extra-pair distance). Distances were calculated for questionnaire (top) and SNP data (below). Statistics by Student's t test.

(C) Distance boxplots for independent questionnaire variables generated by calculating, for all possible pairs of samples, their absolute differences for each variable. We then classified all pairs between pairs of look-alikes and pairs of non-look-alikes. Statistics by Wilcoxon rank sum tests.

DISCUSSION

Our study deciphers molecular components associated with facial construction by applying a multiomics approach in a unique cohort of look-alike humans that are genetically unrelated. Saliva DNA was subjected to genome-wide analyses of common genetic variation, DNA methylation, and microbiome analysis. We also performed a biometric and lifestyle analysis for all look-alike pairs. We found that 16 of the 32 look-alike pairs clustered in all three facial recognition software. Genetic analysis revealed that 9 of these 16 look-alike pairs (Figure 2B) clustered, identifying 19,277 common SNPs. Furthermore, analyses of these shared variants in GWAS and GTEx databases revealed enrichment for phenotypes related to body and face structures and an association with gene-expression changes. Together, this suggests that shared genetic variation in humans that look alike likely contribute to the common phenotype.

Historically, research into face morphology was heavily centered on craniofacial anomalies (Richmond et al., 2018). However, there is a recent growing interest into normal-range face variation, attributable to the necessity for facial recognition software for everyday life (smartphones, CCTV cameras, etc.). Easy access to low-cost, high-resolution pictures and advances in genotyping technology has ignited an age-old question: what makes humans look as they do? Association studies revealed low-frequency genetic variants with relatively small penetrance in facial features, suggesting a far more complex genetic role. Non-genetic factors can affect the expression of genes that form the face. Many epigenetic or imprinting disorders present craniofacial anomalies, such as patients with Prader-Willi or Angelman syndrome (Girardot et al., 2013), and microbial disruption is associated with developmental defects (Robertson et al., 2019). Despite evidence for epigenetic variation in human populations (Heyn et al., 2013) and development (Garg et al., 2018), only one look-alike pair clustered by DNA methylation. This pair also clustered together by SNPs, suggesting that the shared epigenetic profile is likely due to their underlining shared genetics (Lienert et al., 2011), as it was also supported by analyzing CpGs in the vicinity of the SNPs. In addition, ultra-look-alike pairs showed similar epigeneticages. Similarly, only one look-alike pair clustered by microbiome analysis, but ultra-look-alike pairs displayed similar weights, and microbiome composition could relate to obesity (Yang et al., 2019). These findings support a modest role for these biological components to determine facial shape; however, more evidence is required to discard a greater impact.

Finally, 68 biometric and lifestyle attributes from the look-alike pairs were studied. Physical traits such as weight and height as well as behavioral traits such as smoking and education were correlated in look-alike pairs, suggesting that shared genetic variation not only relates to shared physical appearance but may also influence common habits and behavior.

Overall, we provided a unique insight into the molecular characteristics that potentially influence the construction of the

human face. We suggest that these same determinants correlate with both physical and behavioral attributes that constitute human beings. These findings provide a molecular basis for future applications in various fields such as biomedicine, evolution, and forensics. Through collaborative efforts, the ultimate challenge would be to predict the human face structure based on the individual's multiomics landscape.

Limitations of the study

Due to the difficulty to obtain look-alike data and biomaterial, the sample size is small, restricting our ability to perform large-scale statistical analyses. Thus, some partially negative results, such as those derived from the non-genetic data, could relate to an underpowered study. The used headshots were two-dimensional, black and white images, and valuable information regarding three-dimensional constructs, subtle skin tones, and unique facial features are lacking. In addition, the used SNP array does not allow for the analysis of other genetic components such as structural variations and shared rare events. Another limitation is that our samples were mostly from European origin. Thus, the study could not effectively address the impact of the used multiomics in other human populations.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Recruitment of look-alikes
- METHOD DETAILS
  - Facial recognition algorithms
  - Facial similarity
  - Sample preparation
  - HumanOmni5-Quad BeadChip
  - Infinium MethylationEPIC BeadChip
  - 16S meta-genomics sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Population-level vs shared SNPs in look-alike pairs
  - Copy number variant (CNV) calling and functional annotation
  - CNV clustering and heatmap
  - Genome-wide SNP arrays from monozygotic twins
  - Cryptic relatedness
  - Ancestry assessment
  - Kinship assessment
  - Functional enrichment of shared SNPs using Gene Ontology
  - Face gene enrichment in the identified SNPs
  - GWAS analysis
  - GWAS functional enrichment of shared SNPs using S-LDSC
  - DNA methylation age estimation
  - Multiomics clustering analyses
  - Questionnaires processing and similarity analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2022.111257.

ACKNOWLEDGMENTS

We thank François Brunelle for providing the look-alike images. We thank CERCA Programme/Generalitat de Catalunya and the Josep Carreras Foundation for institutional support. This work was funded by the governments of Catalonia (2017SGR1080) and Spain (RTI2018-094049-B-I00, SAF2014-55000, and TIN2017-90124-P) and the Cellex Foundation.

AUTHOR CONTRIBUTIONS

M.E. conceived and designed the study; R.S.J., M.R., C.A.G.-P., M.C.d.M., D.P., S.M., V.D., P.C., M.F.-B., I.O., C.L.-F., A.N., C.F.-T., D.A., F.M.S., X.B., A.V., and M.E. analyzed multiomics and questionnaire data; R.J. and M.E. wrote the manuscript with contributions and approval from all authors.

DECLARATION OF INTERESTS

M.E. is a consultant of Ferrer International and Quimatryx. S.M. is an employee of Ferrer International. C.F.-T. is chief technical officer of Herta Security.

Received: July 16, 2021
Revised: June 5, 2022
Accepted: August 1, 2022
Published: August 23, 2022

REFERENCES

Adhikari, K., Fuentes-Guajardo, M., Quinto-Sánchez, M., Mendoza-Revilla, J., Camilo Chacón-Duque, J., Acuña-Alonzo, V., Jaramillo, C., Arias, W., Lozano, R.B., Pérez, G.M., et al. (2016). A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation. Nat. Commun. *7*, 11616. https://doi.org/10.1038/ncomms11616.

Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., and Irizarry, R.A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics *30*, 1363–1369. https://doi.org/10.1093/bioinformatics/btu049.

Beck, T., Shorter, T., and Brookes, A.J. (2020). GWAS Central: a comprehensive resource for the discovery and comparison of genotype and phenotype data from genome-wide association studies. Nucleic Acids Res. *48*, D933–D940. https://doi.org/10.1093/nar/gkz895.

Biswas, S., Bowyer, K.W., and Flynn, P.J. (2011). A study of face recognition of identical twins by humans. In IEEE International Workshop on Information Forensics and Security, Iguacu Falls, Brazil, pp. 1–6. https://doi.org/10.1109/WIFS.2011.6123126.

Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics. Nucleic Acids Res. *47*, D1005–D1012. https://doi.org/10.1093/nar/gky1120.

Cao, X., Li, W., Wang, T., Ran, D., Davalos, V., Planas-Serra, L., Pujol, A., Esteller, M., Wang, X., and Yu, H. (2022). Accelerated biological aging in COVID-19 patients. Nat. Commun. *13*, 2135. https://doi.org/10.1038/s41467-022-29801-8.

Chang, L., and Tsao, D.Y. (2017). The code for facial identity in the primate brain. Cell *169*, 1013–1028.e14. https://doi.org/10.1016/j.cell.2017.05.011.

Claes, P., Roosenboom, J., White, J.D., Swigut, T., Sero, D., Li, J., Lee, M.K., Zaidi, A., Mattern, B.C., Liebowitz, C., et al. (2018). Genome-wide mapping of global-to-local genetic effects on human facial shape. Nat. Genet. *50*, 414–423. https://doi.org/10.1038/s41588-018-0057-4.

Eden, E., Lipson, D., Yogev, S., and Yakhini, Z. (2007). Discovering motifs in ranked lists of DNA sequences. PLoS Comput Biol. *3*, e39.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. *10*, 48.

Esteban-Cantos, A., Rodríguez-Centeno, J., Barruz, P., Alejos, B., Saiz-Medrano, G., Nevado, J., Martin, A., Gayá, F., De Miguel, R., Bernardino, J.I., et al. (2021). Epigenetic age acceleration changes 2 years after antiretroviral therapy initiation in adults with HIV: a substudy of the NEAT001/ANRS143 randomised trial. Lancet. HIV *8*, e197–e205. https://doi.org/10.1016/S2352-3018(21)00006-0.

Feber, A., Guilhamon, P., Lechner, M., Fenton, T., Wilson, G.A., Thirlwell, C., Morris, T.J., Flanagan, A.M., Teschendorff, A.E., Kelly, J.D., et al. (2014). Using high-density DNA methylation arrays to profile copy number alterations. Genome Biol. *15*, R30. https://doi.org/10.1186/gb-2014-15-2-r30.

Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235. https://doi.org/10.1038/ng.3404.

Fortin, J.-P., Triche, T.J., Jr., and Hansen, K.D. (2017). Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. Bioinformatics *33*, 558–560. https://doi.org/10.1093/bioinformatics/btw691.

Fraga, M.F., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballestar, M.L., Heine-Suñer, D., Cigudosa, J.C., Urioste, M., Benitez, J., et al. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. Proc. Natl. Acad. Sci. USA *102*, 10604–10609. https://doi.org/10.1073/pnas.0500398102.

Garrett-Bakelman, F.E., Darshi, M., Green, S.J., Gur, R.C., Lin, L., Macias, B.R., McKenna, M.J., Meydan, C., Mishra, T., Nasrini, J., et al. (2019). The NASA Twins Study: a multidimensional analysis of a year-long human spaceflight. Science *364*, eaau8650. https://doi.org/10.1126/science.aau8650.

Garg, P., Joshi, R.S., Watson, C., and Sharp, A.J. (2018). A survey of inter-individual variation in DNA methylation identifies environmentally responsive co-regulated networks of epigenetic variation in the human genome. PLoS Genet. *14*, e1007707. https://doi.org/10.1371/journal.pgen.1007707.

Girardot, M., Feil, R., and Llères, D. (2013). Epigenetic deregulation of genomic imprinting in humans: causal mechanisms and clinical implications. Epigenomics *5*, 715–728. https://doi.org/10.2217/epi.13.66.

Gripp, K.W., Baker, L., Telegrafi, A., and Monaghan, K.G. (2016). The role of objective facial analysis using FDNA in making diagnoses following whole exome analysis. Report of two patients with mutations in the BAF complex genes. Am. J. Med. Genet. *170*, 1754–1762. https://doi.org/10.1002/ajmg.a.37672.

Gurovich, Y., Hanani, Y., Bar, O., Nadav, G., Fleischer, N., Gelbman, D., Basel-Salmon, L., Krawitz, P.M., Kamphausen, S.B., Zenker, M., et al. (2019). Identifying facial phenotypes of genetic disorders using deep learning. Nat. Med. *25*, 60–64. https://doi.org/10.1038/s41591-018-0279-0.

Hadj-Rabia, S., Schneider, H., Navarro, E., Klein, O., Kirby, N., Huttner, K., Wolf, L., Orin, M., Wohlfart, S., Bodemer, C., et al. (2017). Automatic recognition of the XLHED phenotype from facial images. Am. J. Med. Genet. *173*, 2408–2414. https://doi.org/10.1002/ajmg.a.38343.

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.B., Gao, Y., et al. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol. Cell *49*, 359–367. https://doi.org/10.1016/j.molcel.2012.10.016.

Heijmans, B.T., Tobi, E.W., Stein, A.D., Putter, H., Blauw, G.J., Susser, E.S., Slagboom, P.E., and Lumey, L.H. (2008). Persistent epigenetic differences associated with prenatal exposure to famine in humans. Proc. Natl. Acad. Sci. USA *105*, 17046–17049. https://doi.org/10.1073/pnas.0806560105.

Heyn, H., Moran, S., Hernando-Herraez, I., Sayols, S., Gomez, A., Sandoval, J., Monk, D., Hata, K., Marques-Bonet, T., Wang, L., et al. (2013). DNA methylation contributes to natural human variation. Genome Res. *23*, 1363–1372. https://doi.org/10.1101/gr.154187.112.

Hoskens, H., Liu, D., Naqvi, S., Lee, M.K., Eller, R.J., Indencleef, K., White, J.D., Li, J., Larmuseau, M.H.D., Hens, G., et al. (2021). 3D facial phenotyping by biometric sibling matching used in contemporary genomic methodologies. PLoS Genet. *17*, e1009528. https://doi.org/10.1371/journal.pgen.

Hsieh, T.C., Mensah, M.A., Pantel, J.T., Aguilar, D., Bar, O., Bayat, A., Becerra-Solano, L., Bentzen, H.B., Biskup, S., Borisov, O., et al. (2019). PEDIA: prioritization of exome data by image analysis. Genet. Med. *21*, 2807–2814. https://doi.org/10.1038/s41436-019-0566-2.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57. https://doi.org/10.1038/nprot.2008.211.

Kaminsky, Z.A., Tang, T., Wang, S.C., Ptak, C., Oh, G.H.T., Wong, A.H.C., Feldcamp, L.A., Virtanen, C., Halfvarson, J., Tysk, C., et al. (2009). DNA methylation profiles in monozygotic and dizygotic twins. Nat. Genet. *41*, 240–245. https://doi.org/10.1038/ng.286.

Keegan, K.P., Glass, E.M., and Meyer, F. (2016). MG-RAST, a metagenomics service for analysis of microbial community structure and function. Methods Mol. Biol. *1399*, 207–233. https://doi.org/10.1007/978-1-4939-3369-3_13.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., and Glöckner, F.O. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. *41*, e1. https://doi.org/10.1093/nar/gks808.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature *513*, 409–413. https://doi.org/10.1038/nature13673.

Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F., and Schübeler, D. (2011). Identification of genetic elements that autonomously determine DNA methylation states. Nat. Genet. *43*, 1091–1097. https://doi.org/10.1038/ng.946.

Lippert, C., Sabatini, R., Maher, M.C., Kang, E.Y., Lee, S., Arikan, O., Harley, A., Bernal, A., Garst, P., Lavrenko, V., et al. (2017). Identification of individuals by trait prediction using whole-genome sequencing data. Proc. Natl. Acad. Sci. USA *114*, 10166–10171. https://doi.org/10.1073/pnas.1711125114.

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873. https://doi.org/10.1093/bioinformatics/btq559.

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R., and Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. *6*, 610–618. https://doi.org/10.1038/ismej.2011.139.

Mee, L., Honkala, H., Kopra, O., Vesa, J., Finnilä, S., Visapää, I., Sang, T.K., Jackson, G.R., Salonen, R., Kestilä, M., and Peltonen, L. (2005). Hydrolethalus syndrome is caused by a missense mutation in a novel gene HYLS1. Hum. Mol. Genet. *14*, 1475–1488. https://doi.org/10.1093/hmg/ddi157.

Moran, S., Arribas, C., and Esteller, M. (2016). Validation of a DNA methylation microarray for 850, 000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics *8*, 389–399. https://doi.org/10.2217/epi.15.114.

Müllner, D. (2013). Fastcluster: fast hierarchical, agglomerative clustering routines for R and Python. J. Stat. Softw. *53*, 1–18. https://doi.org/10.18637/jss.v053.i09.

Parkhi, O.M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In Proceedings of the British Machine Vision Conference (BMVA Press), pp. 1–12. https://doi.org/10.5244/C.29.41.

Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190. https://doi.org/10.1371/journal.pgen.0020190.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909. https://doi.org/10.1038/ng1847.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

Quian Quiroga, R. (2017). How do we recognize a face? Cell *169*, 975–977. https://doi.org/10.1016/j.cell.2017.05.012.

R Core Team (2019). In R: A Language and Environment for Statistical Computing Computer Program, version 3.6. 1.

Ralph, P., and Coop, G. (2013). The geography of recent genetic ancestry across Europe. PLoS Biol. *11*, e1001555. https://doi.org/10.1371/journal.pbio.1001555.

Richmond, S., Howe, L.J., Lewis, S., Stergiakouli, E., and Zhurov, A. (2018). Facial genetics: a brief overview. Front. Genet. *9*, 462. https://doi.org/10.3389/fgene.2018.00462.

Rideout, W.M., 3rd, Eggan, K., and Jaenisch, R. (2001). Nuclear cloning and epigenetic reprogramming of the genome. Science *293*, 1093–1098. https://doi.org/10.1126/science.1063206.

Robertson, R.C., Manges, A.R., Finlay, B.B., and Prendergast, A.J. (2019). The human microbiome and child growth - first 1000 Days and beyond. Trends Microbiol. *27*, 131–147. https://doi.org/10.1016/j.tim.2018.09.008.

Suzuki, R., and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics *22*, 1540–1542. https://doi.org/10.1093/bioinformatics/btl117.

Tsao, D.Y., Freiwald, W.A., Tootell, R.B.H., and Livingstone, M.S. (2006). A cortical region consisting entirely of face-selective cells. Science *311*, 670–674. https://doi.org/10.1126/science.1119983.

Vedaldi, A., and Lenc, K. (2015). MatConvNet in Proceedings of the 23rd ACM International Conference on Multimedia (MM '15) (ACM Press), pp. 689–692. https://doi.org/10.1145/2733373.2807412.

Villicaña, S., and Bell, J.T. (2021). Genetic impacts on DNA methylation: research findings and future perspectives. Genome Biol. *22*, 127. https://doi.org/10.1186/s13059-021-02347-6.

Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era–concepts and misconceptions. Nat. Rev. Genet. *9*, 255–266. https://doi.org/10.1038/nrg2322.

Weir, B., Cockerham, C., and Feldman, M. (2014). Complete characterization of disequilibrium at two loci. In Mathematical Evolutionary Theory (Princeton: Princeton University Press), pp. 86–110. https://doi.org/10.1515/9781400859832-007.

White, J.D., Indencleef, K., Naqvi, S., Eller, R.J., Hoskens, H., Roosenboom, J., Lee, M.K., Li, J., Mohammed, J., Richmond, S., et al. (2021). Insights into the genetic architecture of the human face. Nat. Genet. *53*, 45–53. https://doi.org/10.1038/s41588-020-00741-7.

Wolff, G.L., Kodell, R.L., Moore, S.R., and Cooney, C.A. (1998). Maternal epigenetics and methyl supplements affect agouti gene expression in Avy/a mice. FASEB J. *12*, 949–957. https://doi.org/10.1096/fasebj.12.11.949.

Xing, C., Huang, J., Hsu, Y.H., DeStefano, A.L., Heard-Costa, N.L., Wolf, P.A., Seshadri, S., Kiel, D.P., Cupples, L.A., and Dupuis, J. (2016). Evaluation of power of the Illumina HumanOmni5M-4v1 BeadChip to detect risk variants for human complex diseases. Eur. J. Hum. Genet. *24*, 1029–1034. https://doi.org/10.1038/ejhg.2015.244.

Xiong, Z., Dankova, G., Howe, L.J., Lee, M.K., Hysi, P.G., de Jong, M.A., Zhu, G., Adhikari, K., Li, D., Li, Y., et al. (2019). Novel genetic loci affecting facial shape variation in humans. Elife *8*, e49898. https://doi.org/10.7554/eLife.49898.

Yang, Y., Cai, Q., Zheng, W., Steinwandel, M., Blot, W.J., Shu, X.O., and Long, J. (2019). Oral microbiome and obesity in a large study of low-income and African-American populations. J. Oral Microbiol. *11*, 1650597. https://doi.org/10.1080/20002297.2019.1650597.

Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.

Zhang, C., and Zhang, Z. (2010). A survey of recent advances in face detection. Microsoft Research.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical commercial assays		
Oragene DNA tubes	DNA Genotex	OG-500
Pico Green fluorescence kit	Life technologies/thermos	P7589
EZ DNA Methylation Kit	Zymo Research	D5003
Deposited data		
HumanOmni5-Quad BeadChip	This paper	GEO: GSE142304
Infinium MethylationEPIC BeadChip	This paper	GEO: GSE142304
16S metagenomics sequencing	This paper	BioProject: PRJNA596439
Custom scripts	This paper	https://github.com/mesteller-bioinfolab/lookalike
Look-alike photographs	www.francoisbrunelle.com/webn/e-project.html	https://github.com/mesteller-bioinfolab/lookalike/blob/master/FB_LAL_images.zip
Experimental models: Organisms/strains		
Humans (Homo sapiens)	Look-alike individuals upon consent.	N/A
Software and algorithms		
R	R Core team., 2019	www.r-project.org/
MatConvNet	VLFeat	http://www.vlfeat.org/matconvnet
Microsoft Oxford Project face API	Microsoft Azure	https://azure.microsoft.com/en-us/services/cognitive-services/face/
Herta CNN algorithm	Herta Security	www.hertasecurity.com
GenomeStudio (v2.0.4)	Illumina	https://support.illumina.com/downloads/genomestudio-2-0.html
pvclust	Suzuki and Shimodaira, 2006	http://stat.sys.i.kyoto-u.ac.jp/prog/pvclust/
hclust	Müllner, 2013	https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html
Kinship-based INference for GWAS (KING v2.2.3)	Manichaikul et al., 2010	http://people.virginia.edu/~wc9c/KING/
Minfi (v1.32.0)	Aryee et al., 2014 Fortin et al., 2017	https://bioconductor.org/packages/release/bioc/html/minfi.html
clusterProfiler	Yu et al., 2012	https://guangchuangyu.github.io/2016/01/go-analysis-using-clusterprofiler/
Database for Annotation, Visualization and Integrated Discovery (DAVID v6.8)	Huang et al., 2009	https://david.ncifcrf.gov/
GOrilla	Eden et al., 2007, 2009	http://cbl-gorilla.cs.technion.ac.il/
GTEx portal (v7)	https://gtexportal.org/	N/A
GWAS catalog	Buniello et al., 2019	https://www.ebi.ac.uk/gwas/
GWAS central	Beck et al., 2020	https://www.gwascentral.org/
MG-RAST	Keegan et al., 2016	https://www.mg-rast.org/
Greengenes rRNA database	McDonald et al., 2012	https://greengenes.secondgenome.com/
Other		
François Brunelle website	www.francoisbrunelle.com/webn/e-project.html	N/A
University of Notre Dame twins database 2009/2010	https://cvrl.nd.edu/projects/data/	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for reagents and resource may be directed to and will be fulfilled by the lead contact, Dr. Manel Esteller (mesteller@carrerasresearch.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- SNP and DNA methylation data have been deposited at GEO and are publicly available as of the date of publication. Accession numbers are listed in the key resources table. Microbiome data have been deposited on the BioProject repository and are publicly available as of date of publication. Photographs of the look-alike pairs that were matched together for all three different independent facial recognition softwares have been deposited at GitHub and is publicly available as of the date of publication. The DOI is listed in the key resources table.
- Original code has been deposited at GitHub and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Recruitment of look-alikes

32 Look-alike pairs (n = 64 individuals) that were initially recruited and photographed by François Brunelle (http://www.francoisbrunelle.com/webn/e-project.html) were enrolled to this study. All 64 individuals [42 females (65.6%) and 22 males (34.4%) with a median age of 40 years (range from 21 to 78 years), Table S7] were required to complete an extensive biometric and life-style questionnaire (Methods S1: Data collection questionnaire, related to STAR Methods) as well as provide legally signed consent forms approved by our bioethics committee for usage of both their facial images and DNA samples for this study. The study protocol was approved by the Clinical Research Ethics Committee of the Bellvitge University Hospital with the reference number PR348/16. To compliment this study, we were also provided with access to 100 monozygotic twin photos from the University of Notre Dame twins database 2009/2010 (https://cvrl.nd.edu/projects/data/). License agreements for data access were reviewed and signed by legal representatives of all entities involved in this study. 50 monozygotic twin pairs (n = 100) photographs were subsequently downloaded and analysed with the facial recognition algorithms detailed below.

METHOD DETAILS

Facial recognition algorithms

Three facial recognition algorithms were used to objectively analyze look-alike pairs: MatConvNet CNN algorithm, provided by the University of Pompeu i Fabra, Barcelona (Vedaldi and Lenc 2015); Microsoft Oxford Project face API by Microsoft; and the custom deep convolutional neural network Custom-Net (www.hertasecurity.com). The quantitative assessment of pairwise similarity between face photographs was calculated as follows. For the MatConvNet algorithm, the face biometric template from each photo was extracted from each processed face by means of a deep convolutional neural network (CNN) built into MatConvNet software. The resulting templates are represented as integer sparse descriptors of 8,192 values, which effectively encode the identity features of a face image (Vedaldi and Lenc 2015). Final pairwise similarity scores were set on a scale of 0–1 where 1 represents identical faces.

The custom deep convolutional neural network Custom-Net was developed by a leader in facial recognition platforms (www.hertasecurity.com). Firstly, a generic face detector optimized for unconstrained video surveillance scenarios was used to obtain the locations of all faces in each image (Zhang and Zhang, 2010). The threshold was adjusted to find all targeted faces in each photo, and a subsequent manual exploration was conducted to ensure that no false positives were included. Each face was cropped with a 25% extra margin from the original bounding box, converted to grayscale and resized to 250 × 250 pixels. Next, a face biometric template was extracted from each processed face by means of a deep convolutional neural network of 32 layers. The resulting templates were represented as integer sparse descriptors of 4,096 values, which effectively encode the identity features of a face image. Finally, the similarity score between a pair of images was computed as a negative mean square deviation between their template values. The final scores were mapped to a range 0–1, where 1 indicated identical faces, according to landmarks taken from the histogram of imposter pairs extracted from the well-known database (http://vis-www.cs.umass.edu/lfw/).

In the case of the custom deep convolutional neural network, the models have tens of millions of learned parameters and have been trained with more than 10 million facial images from over a hundred thousand subjects from different human populations, in a variety of unconstrained situations: differences of pose, expression, age and accessories within a subject. Moreover, the training process of a face recognition algorithm typically involves "data augmentation" operations, in which input images are randomly modified, e.g. by

artificially synthesizing glasses, adding facial occlusions, mirroring faces, etc. in order to add intraclass variability to the images and confer robustness to the resulting model. As a consequence, modern face verification algorithms have recently achieved near-perfect accuracy, as high as 99.97% on NIST's Facial Recognition Vendor Test (https://pages.nist.gov/frvt/html/frvt11.html#overview), for passport photo or mugshot scenarios, to the point that banks worldwide have widely adopted such systems for user verification. Particularly, these algorithms have become extremely reliable on controllable, almost ideal scenarios such as those captured by the photographer: 1:1 verification between large resolution images with good illumination, non-lateral poses (less than 60°) and without heavy occlusions; despite circumstancial similaritiy in interclass appearance like that given by glasses, facial expression or hairstyle. Thus, the impacts of these attributes, such as pose, hairstyle etc can be considered minimum, because the incorporated models have been exposed to these variations, in addition to additional features aspects such as colour styles, image degradations etc. The VGG dataset (https://www.robots.ox.ac.uk/~vgg/data/vgg_face/) shows examples of facial data used to train Matconvnet (Parkhi et al., 2015) and CustomNet (http://vis-www.cs.umass.edu/lfw/).

The Microsoft Oxford Project face API by Microsoft operates on a number of attributes that affect facial features such as age, gender, pose, smile, and facial hair along with 27 other landmarks for each face. These landmarks are left pupil, right pupil, nose tip, left mouth, right mouth, outer left eyebrow, inner left eyebrow, outer left eye, top left eye, bottom left eye, inner left eye, inner right eye, outer right eyebrow, inner right eye, top right eye, bottom right eye, outer right eye, left nose root, right nose root, top left nose alar, top right nose alar, left outer tip of nose alar, right outer tip of nose alar, top upper lip, bottom upper lip, top under lip and bottom under lip (https://azure.microsoft.com/en-us/services/cognitive-services/face/). The final similarity scores were also set on a scale of 0–1.

Facial similarity

Pair-wise facial similarity matrices were provided as an output for all three facial recognition software. Similarity scores were assigned as numerical values ranging between 0 – 1 where 1 represents identical images and 0, two opposed images. To obtain objective look-alike pairs, we performed unsupervised hierarchical clustering with bootstrap using the pvclust (Suzuki and Shimodaira 2006) in R statistical environment (v3.6.1) (https://www.R-project.org/).

Sample preparation

Genomic DNA from look-alike pairs in this study were isolated from saliva and self-collected into Oragene 500 DNA tubes and extracted according to the manufacturers instructions (DNA genotek). >10% of the extracted DNA corresponded to microbial DNA. DNA was quantified using Pico Green fluorescence kit/Qubit® 2.0 Fluorometer (life technologies). Bisulfite modification of genomic DNA was carried out with the EZ DNA Methylation Kit (Zymo Research) following the manufacturer's protocol.

HumanOmni5-Quad BeadChip

Comprehensive cross-examination of genome-wide single nucleotide variation of 4.3 million SNVs across all Look-alike pairs was performed using HumanOmni5-Quad BeadChip (Illumina). 400 ng of genomic DNA was applied to HumanOmni5-Quad BeadChip and scanned using HiScan SQ system (Illumina). The signal raw intensities for each array were assessed and analyzed with GenomeStudio Software (v2.0.4) (Illumina) using default normalization to generate X and Y intensity values for A and B alleles (generic labels for two alternative SNP alleles), respectively. Genotype calling were performed by using GenomeStudio GenCall method and only genotypes with high GenCall scores (GC) were selected (according to Illumina standards). The positions corresponding to Illumina internal controls were also removed from the analysis. In order to remove the positions shared between look-alike pairs by chance, a bootstrap look-alike control analysis was performed. Briefly, we generated 100 datasets of 16 random pairs extracted from the initial 32 pairs (64 individuals) used in the study and the complete SNP set from the Omni5 array (4M SNPs). The only requirement was that none of the generated random pairs in the 100 datasets included a candidate look-alike pair from the initial 32 couples. We applied to each of these new 100 "non-look-alike" datasets the same SNP selection protocol used in the look-alike datasets, i.e. removing monomorphic and non-autosomal positions and selecting the shared inter-look-alike genotypes for each of the 16 pairs. This iterative process produced 100 independent SNP datasets that represented shared genotypes between non-look-alike pairs. Each of the SNP lists obtained contained an average of 5000 SNPs. The plot of the cumulative distribution of these shared SNPs after 100 iterations shows that the number of observed SNPs tends to plateau, indicating that we are reaching a maximum number of SNPs shared by the non-look-alike pairs is being reached. Next, we pooled all 100 SNP datasets into one table removing all redundant variants. This table of unique SNPs was considered as the SNP positions shared between pairs independent of their look-alike status (by chance) and were subsequently removed from our analysis of the look-alike pairs. Then the XY and monoallelic positions for the 16 original pairs were removed. Finally, the SNPs with identical genotypes in each of the 16 pairs and located in genes were selected for further analysis. CNV calling was performed by using PennCNV plugin in GenomeStudio with default parameters.

Infinium MethylationEPIC BeadChip

Genome-wide DNA methylation interrogation of >850,000 CpG sites was performed using the Infinium MethylationEPIC BeadChip (Illumina) according to manufacturer's recommended protocol, as previously described (Moran et al., 2016). Briefly, 600 ng of DNA was used to hybridize to the EPIC BeadChip and scanned using HiScan SQ system (Illumina). Raw signal intensity data were initially QC'd and pre-processed from resulting idat files in R statistical environment (v3.6.1) using minfi Bioconductor package (v1.32.0).

A number of quality control steps were applied to minimize errors and remove erratic probe signals. Firstly, interrogation of sex chromosomes was performed to identify potential labeling errors. Next, the removal of problematic probes was carried out, such as failed probes (detection p value > 0.01), cross-reacting probes and probes that overlapped single nucleotide variants within +/− 1bp of CpG sites followed by background correction and dye-based normalization using ssNoob algorithm (single-sample normal-exponential out-of-band). Lastly, we removed all sex chromosomes. Final DNA methylation scores for each CpG were represented as a β-values ranging between standard 0 and 1 where 1 represents fully methylated CpGs and 0, fully unmethylated. All downstream analyses were performed under R statistical environment (v3.6.1).

16S meta-genomics sequencing

We identified and compared bacterial populations from diverse microbiomes from all look-alike pairs using 16S metagenomics sequencing (Illumina) (Klindworth et al., 2013). Salival DNA was extracted and bacterial libraries prepared following the Illumina 16S Library preparation protocol. The variable V3 and V4 regions of 16S rRNA was amplified in order to obtain a single amplicon of approximately 460 bp that underwent paired-end sequencing using MiSeqDx (Illumina). Resulting fastq files were analysed using MG-RAST. The counts corresponding to taxonomic abundance profiles for each sample were retrieved by using MG-RAST tools. Particularly, we retrieved the bacterial counts from sequences aligned to Genus taxonomic categories in the Greengenes rRNA database with the following cutoffs: an alignment length of 15 bp, a percent identity of 60% and an e-value equal or lower to 1×10^{-5}. The relative proportions for each genus and sample were calculated and only the most represented genus were used.

QUANTIFICATION AND STATISTICAL ANALYSIS

Population-level vs shared SNPs in look-alike pairs

In order to define the number of SNPs shared between non look-alike pairs by chance we generated 55 random combinations of the 9 ultra look-alike pairs avoiding in each dataset the presence of a look-alike pair. We selected the SNP positions with the same genotype for each of the 9 non look-alike pairs in any of the 55 control datasets, obtaining the percent of randomly shared variants in a data set of 9 non look-alikes. Finally, we calculated the statistical significance of the comparison between SNPs shared in look-alike and non look-alike pairs by a Pearson's chi-squared test (p value $<2.2\ 10^{-16}$). However, since different pairs of look-alikes were from multiple different ethnicities, but individuals in the same look-alike pair shared the same ethnicity, we also performed the enrichment analysis to determine if the number of shared SNPs was more than expected by chance accounting to ethnicity. Thus, we tested pairs of European ancestry individuals with other Europeans and repeated the same for each of the different ethnicities. To this end, we downloaded the most recent set of Omni genotypes from 1000 Genomes available in the phase 3 release directory (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/hd_genotype_chip/). The downloaded 1000 Genomes phase 3 vcf file was transformed to Genomic Data Structure (GDS) format using the function seqVCF2GDS from SeqArray R package (version 1.36.0). Look-alike PLINK PED files were also transformed to GDS format using the fucntion snpgdsPED2GDS from SNPRelate R package (version 1.30.1). The 1000 Genomes genotyping data was merged with the "ultra" look-alikes genotyping data and the remaining dataset held 67,312 common SNPs. Finally, for each ethnicity we generated 55 random combinations of non look-alike pairs to test if the number of shared SNPs in our "ultra" look-alike population was more than expected by chance. Considering the European ancestry of the majority of "ultra" look-alike (6 out of 9) and non-"ultra" look-alike (7 out of 7) pairs in our study, we used the 7 non-"ultra" look-alike pairs with European ancestry to create 55 random combinations of 6 random non look-alike pairs to compute the number of shared SNPs with the same genotype as a proxy for the European population. For East Asia, Central-South Asia and Hispanic populations, we generated 55 random combinations of 1 random non look-alike 1000 Genomes pair to compute the number of shared SNPs in each of the aforementioned populations. Finally, the number of SNPs shared by "ultra" look-alike pairs in each population was tested for statistical significance enrichment against the background number of shared SNPs in each non look-alike population by means of the Pearson's chi-squared test.

Copy number variant (CNV) calling and functional annotation

The impact of CNVs on genes was calculated in two different ways. First, we looked at whole-gene CNVs, and then partially-overlapping CNVs. Copy number of all genes in the genome was calculated by first establishing CNV breakpoints. Breakpoints were assigned to the outermost SNP positions of regions with the same copy number. The breakpoints were calculated separately for each sample. Using these coordinates, the copy number of whole protein-coding and RNA genes was calculated for all individuals. Gene coordinates were obtained from Ensembl v75 (build GRCh37). We took the genes that had a shared copy number in all pairs of look-alikes (both individuals within the pair had the same number of copies), and we selected those genes for which at least one pair of look-alikes had a different number of copies than the rest of the pairs. For example, to look for partially-overlapping CNVs, we selected all positions in the genome in which the copy number matched within all pairs, but for which at least 2 pairs of lookalikes had a different copy number to the rest of the pairs. We then looked for overlaps with partial overlaps with coding or non-coding genes. As an example, region chr11:125778219-125780253, which overlaps with a lncRNA that has a regulatory relationship with the HYLS1 gene, there are three pairs of look-alikes that carry three copies of this lncRNA, while the remaining pairs have two copies of it. All custom R scripts for CNV analysis are deposited in GitHub repository: https://github.com/mesteller-bioinfolab/lookalike.

CNV clustering and heatmap

Clustering of CNVs was done after filtering out all positions with the same copy number in all samples and merging all contiguous positions with the same copy number. Positions from the X and Y chromosomes that showed the same copy number in all males and the same copy number in all females were also filtered out. The clustering of the samples was calculated using pvclust (Suzuki and Shimodaira 2006). Variants represented in the heatmap are a random selection of one fifth of the total number of variants.

Genome-wide SNP arrays from monozygotic twins

We obtained single nucleotide polymorphism (SNP) data for 38 monozygotic twins from two publicly available studies. Both were downloaded from NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession No. GSE33598 and GSE9608. The signal raw intensities for each array were assessed and analyzed with GenomeStudio Software (v2.0.4) (Illumina) using default normalization to generate X and Y intensity values for A and B alleles (generic labels for two alternative SNP alleles), respectively. All downstream analyses were performed in the R statistical environment (v3.6.1) (https://www.R-project.org/).

Cryptic relatedness

Robust relatedness inference and genetic correlation estimates between monozygotic twins, look-alike pairs and random non look-alikes were calculated using the software KING (Kinship-based INference for GWAS) (version 2.2.3). Student's t-test was applied to calculate statistical significance between populations.

Ancestry assessment

Genotyping was performed using GenomeStudio v2.0.5; PACKPED Plink files were created using the software PLINK Input Report Plug-in v2.1.4 (https://emea.support.illumina.com/downloads/genomestudio-2-0-plugins.html). To analyze the look-alike pairs in the context of world-wide genetic diversity, their genomic data was merged using with 1,980 West-Eurasian, Asian and Native American individuals genotyped in the Affimetrix HO array (Lazaridis et al., 2014); the remaining dataset held 175,469 common SNPs. Principal Component Analysis (PCA) was generated with the HO individuals. Look-Alike individuals were then projected onto the first two components (PC1 and PC2) using options 'lsqproject: YES' and 'shrinkmode: YES' of smartpca built-in module of EIGENSOFT (v. 7.2.1) (Patterson et al., 2006; Weir et al., 2014) (https://www.hsph.harvard.edu/alkes-price/software/).

Kinship assessment

Kinship coefficients between look-alike pairs was first estimated with PLINK. PLINK uses a method-of-moments approach where the total proportion of shared SNPs IBD is calculated based on the estimated allele frequency of all SNPs in a dataset assumed to be homogeneous (Purcell et al., 2007). PLINK-indep-pairwise option was used with parameters 50 5 1.5. to generate a pruned subset of genotypes in low linkage disequilibrium of 282,122 SNPs in comparisons with 1000G dataset and 103,256 in comparisons with HO dataset; pairwise relatedness between individuals of each pair was calculated with the –genome–min-0.05 command to detect pairs with levels of IBD sharing compatible with up to a 3rd degree relationship (Manichaikul et al., 2010). Potential relatedness between pairs was subsequently explored by estimating long (>10 cM) IBD blocks that might be indicative of co-ancestry among individuals occurring in the last few hundreds or years (Ralph and Coop, 2013).

Functional enrichment of shared SNPs using Gene Ontology

Enrichment analysis was done with the enrichGO function from the clusterProfiler R package (Yu et al., 2012), using the org.Hs.eg.db genome annotation. The tested 3,730 genes annotated to the 19,277 SNPs with a matching genotype in all pairs of look-alikes. The background list of genes were all genes annotated to SNPs detected in HumanOmni5-Quad BeadChip analysis. Parameters minGSSize and maxGSSize from the enrichGO function were set to 1 and 22000, respectively, in order to capture all gene ontologies. Additional enrichment analyses were done using DAVID v6.8 and GOrilla.

Enrichment of eQTLs in the look-alike SNPs set was calculated using data from the GTEx portal, release v7 (GTEx_Analysis_v7.metasoft.txt.gz). eQTLs with a fixed effect model p-values < 0.05 were selected for the analysis. A Fisher's test was performed to calculate if the overlap between look-alike SNPs and eQTLs was bigger than expected by chance. The same enrichment analysis was done with each tissue independently, considering the eQTLs with a tissue-specific p-value <0.05. Gene ontology analysis was performed using GOrilla.

Face gene enrichment in the identified SNPs

In order to statistically evaluate the face genes enrichment in our selected 19,277 SNPs corresponding to 3,730 genes shared by all "ultra" look-alike pairs, we gather all the genes related with face traits (face genes) from recent comprehensive genomic screenings related to facial shape (Claes et al., 2018; Xiong et al., 2019; White et al., 2021), the Facebase dataset (https://www.facebase.org/) and GWAS central (study HGVST1841, http://www.gwascentral.org) and applied two different approaches. In the first approach, we applied a hypergeometric test, as it is implemented in the R "phyper" function, from the package "stats". In the second, we also performed a Monte Carlo simulation using 10,000 iterations. In each iteration, we selected a random set of 3,730 genes (the same number of genes in our 19,277 SNPs) from the total genes represented in the array (23,774 genes) and we counted the number of face genes found in this random selection. All the analyses were performed in R statistical programming language v.4.0.3.

GWAS analysis

The overlap between matching sets of SNPs called from look-alike pairs and GWAS SNPs was performed using data from two GWAS databases: GWAS Catalog and GWAS Central. In GWAS Catalog v1.0.2, all GWAS SNPs were retrieved and lifted over from GRCh38 to GRCh37 using the R package liftOver. To calculate trait enrichment, we performed Fisher's exact tests, computing matching genotypes from look-alike pairs against all SNPs detected in the HumanOmni5-Quad BeadChip. For GWAS Central analysis, studies related to facial morphology (HGVST1044, HGVST1625, HGVST1841, HGVST1892, HGVST1933, HGVST2265, HGVST2325, HGVST2359, HGVST2363 and HGVST2597) were selected. Fisher's exact tests were performed to calculate significant overlaps in the different studies and correction for multiple testing was done with Benjamini and Hochberg's adjustment method ($\alpha = 0.05$). All custom R scripts for SNP functional analysis are deposited in GitHub repository: https://github.com/mesteller-bioinfolab/lookalike.

GWAS functional enrichment of shared SNPs using S-LDSC

In order to determine the enrichment of GWAS signals for specific annotations we used the stratified LD score regression (S-LDSC) tool (github.com/bulik/ldsc). S-LDSC is a method to estimate heritability enrichment for selected functional annotations. To this end, we followed the partitioned heritability analysis tutorial (github.com/bulik/ldsc/wiki/Patitioned-Heritability) using the last and recommended version of the baseline-LD model (version 2.2) with 97 annotations. To asses the heritability enrichment of our 19,277 SNPs, we included a ''look-alike'' custom functional annotation, defined by the set of 19277 SNPs, on top of the baseline-LD model v2.2. Since S-LDSC is typically applied to large annotations, we included a 500-bp window around the set of 19,277 SNPs to define our custom ''look-alike'' functional annotation category, following the annotation format of the baseline-LD model v2.2. Considering the European ancestry of the majority of samples in our study, we performed the S-LDSC analysis using European LD scores and allele frequencies from the 1000 Genomes Phase 3 project. Full summary statistics available for ''facial morphology'' trait in European ancestry individuals were downloaded from GWAS Catalog, corresponding to two studies (Xiong et al., 2019; Hoskens et al., 2021). Finally, partition heritability analysis was performed with default parameters and facial traits with ES >1 and enrichment p value < 0.05 were considered.

DNA methylation age estimation

Epigenetic age estimation was computed using the Hannum method using the function methyAge from the ENmix R package (version 1.32.0).

Multiomics clustering analyses

To genetically, epigenetically and metagenomically categorize inherent similarities between all look-alike pairs, shared SNV, CNV, DNA methylation and microbiota profiles, robust correlations and unsupervised hierarchical clustering with bootstrapping were performed with R function packages pvclust (Suzuki and Shimodaira 2006). Euclidean distance scores and ward.d2 minimum variance method were applied to attain hierarchical clustering represented as heatmaps using R statistical environment (v3.6.1). K-means clustering was also performed and represented using the first two dimensions of a Principal Component Analysis (PCA). To perform k-means clustering, 16 ''centers'' (clusters) were indicated. The SNP set was also visualized using t-SNE representation, selecting 2 dimensions and adjusting ''perplexity'' parameter to 6 and ''max_iter'' to 5,000. All the analysis were performed in R statistical programming language v.4.0.3 using the packages ''SNPRelate'', ''gdsfmt'',''stats'', ''Rtsne'', ''ggfortify'' and ''ggplot2''.

Questionnaires processing and similarity analysis

Data obtained through questionnaires was transformed into a table, which was processed and transformed into numerical format with a custom script (deposited in GitHub; https://github.com/mesteller-bioinfolab/lookalike). In this script, all logical variables were transformed to 0 (False/No) and 1 (True/Yes). When the variables could be ordered (e.g. Never - Sometimes - Often), they were assigned numbers (0–1 - 2 in the example) that were afterwards normalized to 1. For non-sortable variables, the categories were split into logical columns (e.g. Employment category was split into three logical variables - Executive, Salaried and Own business). Finally, empty boxes were filled with the mode for each variable. Cosine similarity was calculated using the numerical matrix between all individuals. The look-alike intra and extra-pair distance analysis were defined and calculated as follows. Intra-pairs were defined as look-alike pairs that clustered in all three facial recognition software (n = 16). The extra-pairs were defined as all other combination pairs of non look-alikes in the initial 16 pairs. For 32 individuals, pairs of same individuals and their look-alike pair counterpart were removed, leaving 30 possible combinations per 16 pair (n = 480). The euclidean distances between each individual and all other samples were calculated using the dist function from the R package pvclust (Suzuki and Shimodaira 2006). Distances were calculated on SNP, CNV, methylome, quantitative and qualitative microbiome and questionnaire data. Intra-pair distances were compared to extra-pair distances using Student's T test. Distance boxplots for independent variables were generated by calculating, for all possible pairs of samples, their absolute differences for each variable. We then classified all pairs between pairs of look-alikes and pairs of non-look-alikes. Finally, we calculated if the differences were significant with Wilcoxon rank sum tests.