Mapping genes for human face shape: exploration of univariate phenotyping strategies

Meng Yuan[1,2,3*]✉, Seppe Goovaerts[2,3*]✉, Michiel Vanneste[2,3], Harold Matthews[2,3,4], Hanne Hoskens[1,2,3,5], Stephen Richmond[6], Ophir D Klein[7,8], Richard A Spritz[9], Benedikt Hallgrimsson[5], Susan Walsh[10], Mark D Shriver[11], John R Shaffer[12,13], Seth M Weinberg[12,13,14], Hilde Peeters[2], Peter Claes[1,2,3,4]✉

[1]Department of Electrical Engineering, ESAT/PSI, KU Leuven, Leuven, Belgium.
[2]Department of Human Genetics, KU Leuven, Leuven, Belgium.
[3]Medical Imaging Research Center, University Hospitals Leuven, Leuven, Belgium.
[4]Murdoch Children's Research Institute, Melbourne, Victoria, Australia.
[5]Department of Cell Biology & Anatomy, Cumming School of Medicine, Alberta Children's Hospital Research Institute, University of Calgary, Calgary, Alberta, Canada
[6]Applied Clinical Research and Public Health, School of Dentistry, Cardiff University, Cardiff, United Kingdom
[7]Departments of Orofacial Sciences and Pediatrics, and Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA.
[8]Department of Pediatrics, Cedars-Sinai Guerin Children's, Los Angeles, USA.
[9]Human Medical Genetics and Genomics Program, University of Colorado School of Medicine, Aurora, CO, USA.
[10]Department of Biology, Indiana University Indianapolis, Indianapolis, IN, USA.
[11]Department of Anthropology, Pennsylvania State University, State College, PA, USA.
[12]Center for Craniofacial and Dental Genetics, Department of Oral and Craniofacial Sciences, University of Pittsburgh, Pittsburgh, PA, USA
[13]Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA, USA.
[14]Department of Anthropology, University of Pittsburgh, Pittsburgh, PA, USA.

*These authors contributed equally.

✉Corresponding authors
Meng Yuan (meng.yuan@kuleuven.be)
Seppe Goovaerts (seppe.goovaerts@kuleuven.be)
Peter Claes (peter.claes@kuleuven.be)

Abstract

Human facial shape, while strongly heritable, involves both genetic and structural complexity, necessitating precise phenotyping for accurate assessment. Common phenotyping strategies include simplifying 3D facial features into univariate traits such as anthropometric measurements (e.g., inter-landmark distances), unsupervised dimensionality reductions (e.g., principal component analysis (PCA) and auto-encoder (AE) approaches), and assessing resemblance to particular facial gestalts (e.g., syndromic facial archetypes). This study provides a comparative assessment of these strategies in genome-wide association studies (GWASs) of 3D facial shape. Specifically, we investigated inter-landmark distances, PCA and AE-derived latent dimensions, and facial resemblance to random, extreme, and syndromic gestalts within a GWAS of 8,426 individuals of recent European ancestry. Inter-landmark distances exhibit the highest SNP-based heritability as estimated via LD score regression, followed by AE dimensions. Conversely, resemblance scores to extreme and syndromic facial gestalts display the lowest heritability, in line with expectations. Notably, the aggregation of multiple GWASs on facial resemblance to random gestalts reveals the highest number of independent genetic loci. This novel, easy-to-implement phenotyping approach holds significant promise for capturing genetically relevant morphological traits derived from complex biomedical imaging datasets, and its applications extend beyond faces. Nevertheless, these different phenotyping strategies capture different genetic influences on craniofacial shape. Thus, it remains valuable to explore these strategies individually and in combination to gain a more comprehensive understanding of the genetic factors underlying craniofacial shape and related traits.

Keywords: Phenotyping, 3D facial shape, GWAS, heritability

Author Summary

Advancements linking variation in the human genome to phenotypes have rapidly evolved in recent decades and have revealed that most human traits are influenced by genetic variants to at least some degree. While many traits, such as stature, are straightforward to acquire and investigate, the multivariate and multipartite nature of facial shape makes quantification more challenging. In this study, we compared the impact of different facial phenotyping approaches on gene mapping outcomes. Our findings suggest that the choice of facial phenotyping method has an impact on apparent trait heritability and the ability to detect genetic association signals. These results offer valuable insights into the importance of phenotyping in genetic investigations, especially when dealing with highly complex morphological traits.

Introduction

Human facial development is highly complex, resulting in a rich diversity of facial appearances both within and among populations. Furthermore, facial features have a strong genetic basis, readily apparent within families. The genome-wide association scan (GWAS) is an agnostic approach designed to investigate the statistical relationship between phenotypic traits and genetic variants. A typical GWAS involves individually testing millions of single nucleotide polymorphisms (SNPs) or other common variants dispersed across the genome. Because the precise location of SNPs and genes is known, GWAS signals showing strong evidence of association can point to genes of interest. While many human traits are relatively straightforward

to acquire, capturing facial variation is considerably less so, due to the multivariate and multipartite nature of faces.

Since the initial two GWASs on components of typical-range facial shape variation in 2012 [1,2], more than 300 genome-wide significant signals have been identified in over 20 different studies [3]. Several recent studies [4–8] have embraced a multivariate GWAS framework, regressing multiple univariate traits simultaneously onto each SNP genotype, and have thereby outperformed univariate GWAS in terms of genetic discovery. Nevertheless, several compelling arguments favor univariate GWAS. First, univariate GWAS results can be easily combined across studies via meta-analysis, thereby enhancing statistical power while obviating the need to share highly sensitive facial and genomic data. Second, several important follow-up analyses and GWAS applications, such as linkage disequilibrium score regression (LDSC) [9] and polygenic risk score calculations, require signed effect size and error estimates, which are not readily provided by multivariate techniques. Finally, univariate GWAS is simpler to execute and demands fewer computational resources than multivariate GWAS.

In a traditional anthropometric approach to facial phenotyping, researchers collect a set of univariate measurements such as the distances between pairs of well recognizable, sparsely distributed facial landmarks [1,2,10–18]. Newer approaches have used geometric morphometrics [14,16,19] and expanded sparse landmarks into spatially dense quasi-landmark representations of the face [4,5,7,8,20]. Then, starting from complete landmark configurations (sparse or dense), a popular feature extraction or phenotyping method is principal component analysis (PCA) to extract a set of orthogonal features that represent facial variation. More recently, alternative deep-learning networks, such as auto-encoders (AE), have emerged as non-linear counterparts

to PCA. Despite the current trend favoring neural networks, to the best of our knowledge, these have not yet been applied in facial GWAS.

Apart from methods involving facial anthropometrics or unsupervised learning, supervised approaches have also been used to extract specific univariate facial features. For instance, it is feasible to extract facial characteristics expected to exhibit high heritability, such as facial traits shared among siblings [21]. Another illustration is GWASs conducted using resemblance scores guided by patient facial archetype associated with Achondroplasia [22] or Pierre Robin Sequence [23]. Similarly, resemblance scores to the distinctive facial endophenotype in unaffected relatives of individuals with non-syndromic cleft lip was successfully used in GWAS, which helped to further elucidate the genetic susceptibility to non-syndromic cleft lip [24].

Here, we provide a comprehensive comparison of univariate facial phenotyping approaches in GWAS of facial shape based on a cohort of 8,246 healthy European individuals. We evaluated phenotyping approaches based on two criteria: (1) GWAS discovery rate, defined as the number of independent association signals identified in aggregate across phenotypes in the same category (e.g., all principal components), and (2) SNP-based heritability determined by LDSC [9]. Additionally, this work offers secondary contributions by (1) exploring the latent dimensions of an AE as facial traits in GWAS, and by (2) introducing two additional supervised phenotyping schemes, one by extreme facial gestalts and another by randomly selected facial gestalts.

Results

As illustrated in Fig 1, this study explored three distinct facial phenotyping strategies or categories. The first category, known as anthropometric techniques, focused on inter-landmark measurements. These measurements were defined as the Euclidean distances in 3D space between pairs of sparse facial landmarks. The second category, referred to as unsupervised techniques, involved deriving latent representations obtained through PCA and AE. These techniques generated up to 200 latent dimensions from spatially dense configurations of quasi-landmarks (n=7,610), as established using MeshMonk [25]. The third category, termed supervised techniques, centered around resemblance-based facial traits, comparing each individual in the cohort to specific facial gestalts ranging from random to extreme to syndrome-related facial examples. Each face in the cohort received a resemblance score by measuring its cosine distance in multivariate face space against the provided facial examples (random, extreme, and syndromic). All phenotyping methods were applied to the complete facial shape and, separately, to nasal shape. The focus on nasal shape was due to its high heritability, making it a particularly noteworthy facial region for detailed examination [26].

SNP-based heritability

Fig 2 illustrates the distribution of SNP-based heritability, computed using LDSC [9], for facial traits extracted by various phenotyping methods. For full facial shape, inter-landmark distances demonstrated the highest mean heritability, followed closely, without significant difference (Fig S1 in supplementary file 1), by traits extracted through an AE. PCs and resemblance scores to randomly selected facial gestalts were both ranked as the second most heritable traits, although

PCs displayed greater variation in heritability scores. Notably, the mean heritability for resemblance scores to both extreme and syndromic facial examples was the lowest, implying a reduced influence of common genetic variants. Similar trends were observed for nasal shape, except that inter-landmark distances, in this scenario, displayed significantly higher heritability than all other categories of nasal phenotypes (Fig S1-2 in supplementary file 1).

Identification of trait-associated genetic loci

We assessed the GWAS discovery rate for various categories of facial traits by counting the number of independent genetic loci associated with a set of traits of the same type. We gradually increased the numbers of traits submitted for GWAS in each phenotype category, for example, the first N PCs, with N varying between 1 and the total number of PCs. Combining multiple univariate GWASs was achieved by taking the lowest P-value for each SNP across all the univariate traits considered. To appropriately control for the multiple testing burden, we estimated a group-wide significance threshold as $P < 5e\text{-}8$ divided by the effective number of traits (Methods).

The effective number of traits within a single group is shown in Fig 3.A. As expected, PCs are uncorrelated, so the number of effective traits equals the number of PCs used in a group. In contrast, inter-landmark distances exhibited a high degree of correlation, shown as a flattened curve. A lower degree of correlation was observed for resemblance-based traits (random/extreme/syndromic) and AE latent dimensions.

For each category of traits, the discovery rate generally increased when including more effective traits in GWAS (Fig 3.B). This is most strongly observed for inter-landmark distances. For nasal

shape, the limited number of 10 inter-landmark distances resulted in the poorest discovery rate overall. In contrast, 276 inter-landmark distances were extracted from full facial shape, leading to the best discovery rate across all tested measures.

For nasal shape, the findings for the unsupervised techniques of PCA and AE exhibited similar trends. Specifically, as more effective traits were included, the number of identified genetic loci initially increased until it reached a maximum, after which a decline in the discovery rate was observed. This decline can be attributed to the tradeoff between adding less genetically interesting traits and a more significant threshold that is required to adjust for multiple testing. Particularly in the case of PCA, it is well-established that later PCs primarily model noise in the data and are not expected to contribute to further genetic discoveries. The same was observed for the latent dimensions of AE, despite their lack of a specific order in terms of phenotypic variance explained, unlike PCs. For full facial shape, a similar pattern of initial increase and subsequent decline was observed for AE and PCs, but the AE latent dimensions failed to reach the same discovery rate as PCs.

For the supervised techniques, the relatively small number of syndromes (n=25) may have impacted the overall GWAS discovery rate for this group when compared to all the other phenotyping strategies. Nonetheless, in the case of nasal shape, the maximum discovery rate for syndrome archetypes is high compared to the number of effective traits used. Conversely, this was not the case for full facial shape. This finding highlights that syndrome archetypes are valuable, particularly in nasal regions, but may not be as effective in characterizing full facial variation. The outcomes obtained by extreme facial gestalts initially showed a lower identification rate of associated common variants, but gradually converged with other

techniques as the number of effective traits increased. It is important to note that this convergence is essentially a result of treating more faces as “extreme”, even though they may actually be less or no longer extreme (as explained in the Methods). Lastly, in the case of resemblance to random facial gestalts, a steady increase in GWAS discovery rate is observed as the number of effective traits increases. Notably, when further expanding the number of random facial gestalts used (Fig S3), this approach outperforms all other methods. In other words, the benefits of adding more traits outweigh the multiple testing burden in this scenario. However, due to the randomness involved, the GWAS discovery rate showed greater variation when repeating the experiment over consecutive runs, as indicated by the error bars in Fig 3.B and Fig S3.

Fig 3.C illustrates the GWAS discovery rate plotted against the cumulative phenotypic variance explained by each phenotyping method. The variance explained for a group of facial traits was measured using partial least-squares (PLS) regression (using the ‘plsregress’ function from MATLAB R2022b) with the original images (3D quasi-landmark configurations) as responses and the grouped univariate facial traits as predictors. The cumulative variance of all PLS components reflects the explained phenotypic variance. Interestingly, the first PC, while explaining 31.22% of the phenotypic facial variation, did not yield any significant genetic loci. Furthermore, the first 10 PCs captured 80.75% of total facial variation but resulted in the identification of only 4 independent genetic loci. The same was observed for AE dimensions. This suggests that, while a substantial amount of geometric phenotypic variance is captured by the first few PCs and AE dimensions, they do not necessarily correspond to genetically relevant information. In contrast to both dimensionality reduction techniques, the number of identified genetic loci based on

inter-landmark distances and resemblance-based scores increased rapidly with even a limited number of traits, explaining only a few percent of the complete facial variation. This indicates that, while these traits capture less geometric facial variation, they result in a greater number of discoveries in GWAS, suggesting that these traits are enriched for genetically determined aspects of shape variation.

Sharing of genomic signals

We tested whether various types of traits resulted in overlapping or distinct sets of identified independent genetic loci and annotated genes (Fig 4). For each group of traits, we evaluated genetic loci under the "best-case scenario", i.e., when the maximal number of independent genetic loci was reached. Genetic loci were considered shared between two different methods if their respective lead SNPs were located within 250kb of each other. Considering that AE latent dimensions and randomly selected facial gestalts are inherently stochastic phenotyping strategies, we conducted multiple runs for these approaches to assess the impact of randomness on the results.

Surprisingly, the extent of overlap in terms of genetic loci between different methods was relatively limited. When taking the union of all independent genetic loci identified across different approaches, we found 60 loci associated with the nose and 58 loci associated with the face. This suggests that each of the phenotyping strategies capture distinct aspects of facial shape variation and, as a result, they strongly complement each other in pinpointing genetic factors that influence facial shape. Similarly, for 10 replicates of generating multiple AE latent dimensions and resemblance to random gestalts based on nasal shape, the combined set of

identified genetic loci across all 10 randomizations yielded 46 and 33 genetic loci, respectively. For full facial shape, the union set included 31 genetic loci using AE latent dimensions and 33 genetic loci using resemblance to random gestalts, respectively. This underscores the importance of conducting multiple runs, as the inherent randomness in the process proves advantageous in thoroughly exploring the entire spectrum of facial shape variation.

The number of pairwise overlapping genes followed a similar pattern to the number of pairwise overlapping genetic loci, as expected. Several key craniofacial transcription factors, including *ALX1*, *PAX3*, *TBX15*, and *SOX9*, were consistently identified, regardless of the category of traits used. The complete list of genes detected by at least four different categories of traits (out of the total of six groups) can be found in supplementary file 3 Table S2-3. When considering a single trait, the identification of genes was relatively constrained, resulting in a corresponding limitation in detecting Gene Ontology (GO) biological processes. However, based on the union set of lead SNPs from all groups of phenotypes, the top terms (based on lowest binomial P values) in the GO biological processes category were all highly relevant to craniofacial shape (full lists can be found in source data). This again indicates the idea that different phenotyping strategies are indeed complementary in capturing the diverse genetic influences on craniofacial shape.

Discussion

In this study, we evaluated and compared different techniques for extracting univariate facial phenotypes in humans, quantified from 3D facial images. Traditional anthropometric traits, such as inter-landmark distances, demonstrated the highest mean heritability suggesting that they are well focused towards genetically determined aspects of shape variation. While the set of inter-

landmark distances yielded a relatively high number of GWAS loci compared to a similarly sized set of traits from a different phenotyping category, the total number of loci identified was ultimately limited by the number of available landmarks. This became especially apparent for nasal shape, where only 5 landmarks were available to extract pairwise distances, such that all other phenotyping categories identified a greater number of GWAS loci. Even though the absolute number of inter-landmark distances rapidly increases with each additional landmark, the number of effective phenotypes lags behind due to the high degree of correlation between these measurements. Therefore, the scalability of this phenotyping approach is limited at a computational cost. This may partly be alleviated by selecting the most accurate and distinctive measures based on prior knowledge of anatomy and biology [27,28]. Altogether, measuring inter-landmark distances, already used extensively in facial GWAS [1,2,10–18], is a viable univariate phenotyping method with a good yield in GWAS on the condition that enough landmarks are available and computational cost is considered. However, in comparison to the other techniques, they are highly correlated and are likely to identify only a specific set of genetic influences to facial shape. Therefore, it is ideal for this approach to be supplemented with another strategy to cover the full spectrum of genetic factors underlying facial shape.

A more complete description of facial shape can be obtained by modeling the set of dense 3D quasi-landmark coordinates, which constitutes a highly correlated set of facial features. Unsupervised dimension reduction techniques offer a means to compress this set into a reduced set of morphological variables that can be used as traits in GWAS analysis thereby using dramatically fewer computational resources compared to using the individual landmarks.

Among the unsupervised dimension reduction methods for facial shape analysis, PCA has seen the most use in the literature, including in GWAS analysis [3]. PCA is deterministic, conceptually simple, and available in most data analysis platforms. One advantage that PCA offers is the ordering of its PCs according to their contribution to phenotypic variance. It is well-established that noise from the original images is modeled by the later PCs, which makes it straightforward to determine how many PCs to retain post hoc. However, we observed that the amount of phenotypic variance explained by a single PC does not necessarily indicate its utility for discovering genetic associations. For example, a GWAS on the first PC of facial shape failed to identify a single locus, despite this PC explaining 31.22% of overall shape variation. In fact, when looking at the combined GWAS results across all the facial shape PCs (Fig 3.B and C), we observed that the majority of independently identified genomic loci were contributed by PCs 10–40. Earlier PCs explained more phenotypic variation but did not identify as many genetic associations. Later PCs (>40) did not contribute many additional loci but did exacerbate the multiple testing burden, resulting in an optimal number of loci identified at around 70 facial PCs followed by a drop-off. Furthermore, we found that PCs exhibit a lower mean heritability compared to inter-landmark distances with a wide range in heritability values across the PCs. This may suggest that, while some components have a strong genetic basis, others may not. This may be attributed to by the fact that PCs are essentially mathematical constructs constrained to be mutually orthogonal, whereas inter-landmark distances have the freedom to be correlated, capturing slightly different yet overlapping information. Altogether, PCs derived from dense landmark configurations almost fully capture the available 3D shape information and are straightforward to acquire. However,

we have shown that the order of features/PCs based on phenotypic variance explained does not necessarily indicate their relevance for genetic findings.

Another dimension reduction technique considered in our study was an AE. These deep learning-based networks have surfaced as a popular non-linear alternative to PCA in many fields of research including image analysis [29,30]. However, the latent dimensions of an AE are currently underexplored as a phenotyping strategy, and have never, as far as we are aware at the time of writing, been used in facial GWAS analysis. In contrast to PCA, setting up and training an AE network requires far more time and expertise due to its complexity and the extensive parameter tuning required. For example, the number of latent variables needs to be set prior to model training, and creating more compact or elaborate models requires re-training. Simply excluding latent dimensions leads to poor reconstruction performance [31], hence determining the optimal latent dimensionality becomes a process of trial and error. Furthermore, latent variables of an AE are unordered, explain similar amounts of overall phenotypic variation, can encode for non-linear data interactions, and are not subject to any orthogonality constraints. These properties have likely contributed to their high SNP-based heritability, only second to inter-landmark distances and significantly higher than PCs. However, despite their high expected SNP-based heritability, AE latent dimensions identified a similar number of independent genomic loci in GWAS on nasal shape compared to PCs, and fewer in GWAS on facial shape. These results suggest that although individual AE dimensions may have a strong genetic basis, properties such as their cross-correlations and redundancy make them no better than PCs for genetic discovery. These observations challenge the increasing preference for machine learning-based algorithms in facial analysis, where PCA is criticized for relying on linear transformations and therefore likely

struggling with non-linearity in facial data. However, non-linearity might not be as abundant as one might expect in static facial shapes, or alternatively, the added value of this ability is only minimal in the context of GWAS. This is unlike situations where machine learning algorithms have outperformed PCA by learning the nonlinear variations associated with different facial expressions or pose conditions [32].

While dimension reduction methods are powerful for extracting features from high-dimensional correlated datasets, the biological meaning of their resulting features and the validity of the results reported in the field of genetics have been questioned [33,34]. To ensure biological relevancy of the obtained morphological variables, some studies [19,35] have first derived phenotypes through a dimension reduction method and subsequently selected a subset of traits for downstream analysis based on heritability estimations. A more sophisticated approach adopted by some recent studies is to rely on prior biological knowledge to derive likely heritable facial traits in a supervised manner. Focusing on heritability directly, researchers have extracted highly heritable facial traits by considering familial resemblance and family-based heritability estimations from which they derived measures such as the principal component of heritability [36–38] and siblings-shared facial traits [21]. Furthermore, to investigate both typical-range and disease-associated variation in facial morphology, some studies employed a phenotyping method supervised by genetic conditions characterized by distinct facial features. Examples include resemblance scores to the facial archetype associated with Achondroplasia [22] and Pierre Robin Sequence [23] as well as resemblance scores to the distinctive facial endophenotype in unaffected family members of individuals with non-syndromic cleft lip [24]. The general idea of this approach is to directly measure the facial features that result from subtle variations within

the same physiological pathways, which when disrupted result in distinct (sub-)clinical facial characteristics. Substantially expanding on this approach, our comparative study included resemblance scores supervised by the facial archetypes derived from 25 syndromes associated with distinct facial characteristics. Lastly, we further generalized this approach to supervise the facial phenotyping by either extreme (but non-clinical) or randomly selected facial examples.

Extreme phenotypes are often associated with strong genetic signals, such as large-effect single gene variants, as initially explored for facial shape by Crouch et al. [19]. Building on this insight, we recognize that multidimensional facial variations allow for the identification of extreme faces, which can be used to supervise resemblance scores. Similarly, a randomly selected actual face is expected to reflect genetic signals as it is a product of inheritance. Therefore, we used randomly selected actual faces to supervise resemblance scores as facial traits in GWAS. Resemblance scores supervised by syndromic facial archetypes exhibited lower mean heritability and resulted in fewer genetic loci compared to other groups of traits. This may be explained by the limited number of syndrome groups and the role of low frequency genetic variants. To illustrate, the limited number of syndrome groups resulted in a limited number of syndrome-derived traits, further leading to a lower statistical power. In addition, as GWASs focus on common genetic variants, they overlook low-frequency and rare genetic variants that could potentially underpin these traits. Similar findings were observed for resemblance scores to extreme facial gestalts. While eventually achieving a comparable GWAS discovery rate to PCA, this convergence primarily resulted from the inclusion of more extreme facial examples, which were progressively less extreme. Nevertheless, while resemblance scores derived from syndromic and extreme facial examples may not yield the greatest number of loci in GWAS, studies [22–24] have demonstrated

that a targeted facial phenotyping resulted in GWAS loci that displayed a stronger link with disease etiology versus non-targeted phenotyping approaches. Therefore, facial traits derived from genetic conditions may facilitate the discovery of disease-related genes and pathways in future investigations. This could be especially interesting in the context of uncommon and rare genetic variants available from whole-exome or whole-genome datasets.

Resemblance scores to random facial gestalts surpassed all the other phenotyping approaches in terms of the number of identified genetic loci in GWAS, on the condition that enough of such traits were considered. Measuring the resemblance to a specific randomly selected facial gestalt can be thought of as measuring the extent to which a specific person's set of facial features is present in the faces of the other individuals within the cohort. Therefore, the total number of extractable traits is equal to the cohort size, usually in the thousands. Mathematically, each randomly selected facial gestalt, under the absence of identical twins, represents a unique direction in the face space, thus allowing one to sample that space in a brute-force-like way. Compared to other phenotyping approaches, these traits displayed a high mean SNP-based heritability and yielded a high number of significant genetic loci relative to their explained phenotypic variance. Together, this suggests that a measure of resemblance to a random facial gestalt captures genetically determined aspects of facial shape variation. A possible explanation could be that this approach intentionally focusses on facial features that are observed within a cohort as a result of inheritance, rather than on purely mathematical decompositions of facial shape. In summary, the ability to generate many facial phenotypes with a high expected heritability and that yield a set of complementary loci in GWAS, make resemblance to randomly selected facial gestalts a great option for those willing to accept the computational burden.

The limited overlap observed in identified genetic loci across different methods suggests that each phenotyping strategy captures distinct genetic factors influencing facial shape. This observation may reflect the Beavis effect [39,40], where each method samples from a larger, underlying but truncated distribution of biologically real signals, and the detected loci are subsample specific. The more underpowered a study is to capture the full range of effects, the more pronounced the Beavis effect becomes, increasing the probability of non-replication of genuine signals. In other words, with unlimited and continuously growing sample sizes, it might become possible that the different phenotyping strategies converge onto each other, and that genetic loci identified by one strategy are replicated by another strategy. However, with the current sample sizes of today, that remains to be investigated.

When using resemblance scores for random gestalts and AE latent scores, the sets of identified genetic loci varied substantially across multiple replicates of GWAS due to different random initializations. While this presents challenges for interpretation and replication, the larger union set of significant genetic loci offers opportunities for comprehensively exploring the genetic underpinnings of the entire spectrum of facial shape variation. These observations suggest the possibility of optimization. For example, it could be valuable for future studies to investigate how to generate a minimal set of facial traits that maximizes genetic findings thereby alleviating some of the computational burden. Nonetheless, regardless of the category of phenotypes used, key craniofacial transcription factors were consistently identified, and the combined set of loci across all phenotyping categories yielded GO biological processes that were highly relevant to craniofacial shape. This underscores that different phenotyping approaches complement each other in the identification of genetic factors influencing facial shape.

In this comprehensive study, we conducted a thorough evaluation of various univariate phenotyping methods for the characterization of human facial shape. These methods were categorized into three groups, which encompassed anthropometric traits, traits derived through unsupervised dimension reduction techniques, and supervised resemblance-based traits. Our findings expand the current understanding of the genetic relevance of various univariate traits, including their SNP-based heritability and GWAS discovery rates. Traditional anthropometric traits, which are derived from a set of landmarks with clear anatomical meaning, exhibit high SNP-based heritability, making them suitable traits for genetic investigations. Though, their limitation mainly lies in their fundamentally incomplete morphological description, especially when the number of landmarks is limited. On the other hand, dimension reduction methods, which despite lacking a clear biological meaning, can more fully capture morphological variation and subsequently identify a good number of genomic loci in GWAS. However, our analyses have shown that for the purpose of GWAS analysis, training an AE network is likely not worth the hefty time investment as it identified fewer independent genomic loci compared to PCA. As an alternative, our study has expanded on the idea of supervised resemblance-based phenotypes by using facial gestalts from 25 genetic conditions as well as randomly selected and extreme, non-clinical facial gestalts. While resemblance scores to randomly selected facial gestalts are easy to acquire and have demonstrated their potential to capture genetically relevant facial shape variations in GWAS, resemblance scores to extreme and syndromic facial gestalts may be useful in the search of rare genetic variants in future studies. Overall, this work investigated various types of univariate phenotyping strategies for facial shape, which could potentially be extended

418 to other morphological structures, such as brain shape, providing valuable references for future
419 research.

420 Materials and methods

421 Dataset and preprocessing

422 The analysis included participants with typical-range facial shape of European descent from
423 independent population-based cohort studies conducted in the United States (US, $n_{US} = 4{,}680$)
424 and the United Kingdom (UK, $n_{UK} = 3{,}566$). In our previous work [5], this dataset (referred to as
425 the EURO dataset) was used for a multivariate GWAS study on facial morphology. The US samples
426 originated from three independent data collections: the 3D Facial Norms cohort [41] (3DFN) and
427 from studies at the Pennsylvania State University (PSU) and Indiana University-Perdue University
428 Indianapolis (IUPUI). Institutional review board approval was obtained at each recruitment site,
429 and all participants gave their written informed consent before participation. The UK samples
430 were part of the Avon Longitudinal Study of Parents and their Children [42,43] (ALSPAC). Ethical
431 approval for the study (Project B2261: "Exploring distinctive facial features and their association
432 with known candidate variants") was obtained from the ALSPAC Ethics and Law Committee and
433 the Local Research Ethics Committees. Information on the different genotyping platforms,
434 imputation, and quality control can be found in [5]. Intersection of imputed and quality-
435 controlled SNPs across the US and UK datasets yielded 7,417,619 SNPs for analysis. The 3D facial
436 surface images were registered using the MeshMonk [25] registration framework in MATLAB

(R2017b) as described in [5]. In total, 8,246 unrelated participants with recent European ancestry passed genotyping, imaging, and covariate quality control, and were used for analysis.

We used a subset from the syndromic face dataset in our previous work [44], where it was originally applied for a syndrome classification task. This subset was obtained from two databases: 1) the FaceBase repository "Developing 3D Craniofacial Morphometry Data and Tools to Transform Dysmorphology, FB00000861" [45]; 2) Peter Hammond's legacy 3D dysmorphology dataset hosted at the KU Leuven, Belgium [46]. Syndromes can be categorized based on whether the underlying genetic conditions can be diagnosed based on typical facial characteristics [44]. In this study, we focused on syndromes with typical facial features falling into category A and B as defined in [44], including 25 out of the total 51 syndromes (details in supplementary file 3 Table S1). Overall, there were 1,784 3D syndromic facial images and a control group of 54 individuals unrelated to patients with known genetic syndromes. These control images were used to determine whether the average syndromic images were significantly different from those of the healthy controls for each syndrome group.

The 3D facial surface meshes, comprising 7,160 dense quasi-landmarks were aligned using generalized procrustes analysis (GPA), symmetrized, and subsequently adjusted for age, age squared, sex, weight, height, facial size, camera system, and the first 4 genomic ancestry PCs using PLS regression (function 'plsregress' from MATLAB R2022b). The same procedure was performed independently for the nose, which was obtained by applying the data-driven hierarchical facial segmentation method described in [4,5]. Essentially, facial segments were defined by grouping strongly correlated vertices using hierarchical spectral clustering [4,47]. The strength of correlation between quasi-landmarks was measured using Escoufier's RV coefficient

[48,49]. Subsequently, the RV coefficient was used to construct a similarity matrix that defined the formation of facial segments. As shown in Fig 1.A, the highlighted nose module consists of 758 vertices.

Facial phenotyping strategies

In this study, we explored three categories of phenotyping methods: the first category involved anthropometrics traits, exemplified by inter-landmark distances; the second category encompassed latent scores derived through dimensionality reduction methods such as PCA and AE; and finally, resemblance-based traits were defined as the 1 - cosine of the Mahalanobis angle between the vectors of the target sample (extreme/syndromic/random gestalts) and each sample in the EURO cohort.

Inter-landmark distances

Since the images were symmetrized, we focused on 24 anatomical facial landmarks on the right half of the face, including the facial midline (Fig 1.A). Most landmarks have been used in previous GWASs of facial variation and have shown relatively high heritability [10,17]. The phenotypes were computed as inter-landmark Euclidean distances between landmarks (in total 276 for face, 10 for nose). We followed a semi-automatic landmarking procedure as described in [25] using MeshMonk to position the landmarks onto all samples. First, a set of randomly selected facial scans (N=5) was manually landmarked three times by two observers. Subsequently, the average positions among iterations were calculated for each landmark, and the resulting placements were transferred to the template through barycentric coordinate conversion. These average placements on the template served as the foundation for the automated landmark placements.

Finally, since the faces are in the same coordinate system as the original template, the averaged landmark positions could be automatically transferred to the entire dataset. The facial template in Wavefront (.obj) format, the coordinates of 24 facial landmarks and 5 nasal landmarks on this template can be found in source data.

Unsupervised dimensionality reduction of dense quasi landmarks

Principal component analysis

Principal component analysis (PCA) simplifies complex facial variation by transforming high-dimensional mesh configurations into a small number of uncorrelated features, i.e., principal components (PCs). The original dense landmark configurations were structured into a three-dimensional matrix with dimensions N (number of shapes), L (7,160 quasi-landmarks), and 3 (x-, y-, and z-coordinates of each landmark). To perform PCA, we first mean-centered the data and reshaped it into a two-dimensional matrix with dimensions $N \times 3L$. Subsequently, we applied low-rank singular value decomposition (SVD) to the mean-centered reshaped data matrix $X \in \mathbb{R}^{N\times 3L}$, defined as $\mathrm{X} = \mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}}$ (Fig 1.B). The diagonal matrix Σ contained the singular values and the columns of U and V consisted of the left and right singular vectors, respectively. The right singular vectors in V represented the PCs. Additionally, PCA was performed in combination with parallel analysis [50,51] to capture the major shape variance with the optimal number of variables. This resulted in 32 PCs explaining 99.21% of nasal shape variation and 70 PCs explaining 98.08% of facial shape variation.

Auto-encoder

An auto-encoder (AE) works as a non-linear generalization of PCA, comprising two main parts: an encoder and a decoder. The encoder compresses the data into a small number of variables and the decoder aims to reconstruct the original data from that compact representation. The advantage of using an AE is that it can model non-linear relationships that may be present in the data. However, as opposed to PCA, the disadvantage of an AE is that the latent variables are not necessarily uncorrelated.

Fig 1.C shows the structure of the auto-encoder network used to extract features based on 3D facial meshes as previously used in [52]. The first several layers of the encoder consist of spiral convolutional layers, which reduce the size of the input. Each spiral convolutional layer consists of a spiral convolution operator and a mesh simplification step. Spiral convolution operators [53,54] are analogous to the grid-based convolutional filters in traditional convolutional neural networks and are designed as spirals starting at a center point and proceeding outwards from a random adjacent point. The mesh simplification step reduces the input size based on a predefined fixed scheme, achieved by performing quadric edge collapse on the template using MeshLab software [55]. The three spiral convolutional layers consist of 64, 64, and 64 learned filters, respectively, followed by the addition of two fully connected layers to further compress the data into the desired number of latent variables. The decoder architecture mirrors the encoder architecture. The model is trained to minimize the reconstruction error. Training strategy and implementation details can be found in supplementary file 2.

Supervised resemblance measurements

Individual faces can be represented as single points or vectors situated in a multidimensional "face space", where each dimension reflects a continuous axis of morphological variation [56,57]. To construct such a face space, we applied PCA to the symmetrized and GPA aligned quasi-landmarks of the 8,246 samples, as mentioned above, and retained an equal number of PCs for consistency, i.e., 32 PCs explaining 99.21% of nasal shape variation and 70 PCs explaining 98.08% of facial shape variation. Note that, in principle, a shape space can alternatively be obtained using a different dimension reduction method. In our space, each face could be represented as a vector encoding the scores along each PC. In other words, the vector representation of a single face represented the extent to which the facial features encoded by each PC were present within that face. Following the idea that the resemblance between two faces can be measured by the correlation between their features, we quantified the facial resemblance of one face to another as the cosine distance derived from the angle enclosed by their feature vectors in a Mahalanobis standardized space (Fig 1.D) [58]. To obtain resemblance-based scores for GWAS analysis, we calculated facial resemblance scores between each face from the cohort and a specific facial example, whereby we considered different possibilities for the choice of facial example.

In a first scenario, we considered the facial example to be a randomly selected face from the cohort and calculated resemblance-based facial phenotypes for GWAS as the cosine distance between the vector of the EURO cohort faces and the vector of the selected random facial example. We gathered additional resemblance-based facial phenotypes by selecting additional randomly selected facial examples. A second category includes the resemblance of the EURO cohort to an extreme facial example. To do so, we first ranked all the individuals based on their

Mahalanobis distance from the estimated mean face, which could be represented as the origin of the face space. Subsequently, we selected the top k (desired dimension) individuals that were located most peripherally in the face space. Each sample from the EURO cohort was then scored by computing the cosine distance between its vector and the vector of each individual extreme facial example. A third category included resemblance to syndromic faces. We projected 1,784 syndromic faces from 25 distinct syndromes into the learned PCA space based on the EURO cohort and computed the average shape from each syndrome group. Using a permutation testing framework as described in [23], we tested which of the average syndromic faces were significantly different from the healthy controls and subsequently removed any syndromes without any distinct ($P>0.05$) characteristics ($n=0$), leaving 25 for further analysis. We repeated this procedure for the nose, where 23 out of 25 syndromes were considered for further analysis (details of syndrome groups in supplementary file 3 Table S1). Resemblance-based phenotypes for GWAS were obtained by measuring the cosine distance between the EURO cohort and the syndromic facial gestalts, which were calculated as the average face per syndrome.

Genome-wide association meta-analysis

For each univariate trait, GWASs were conducted in the US and UK cohorts independently using linear regression (function 'regstats' from MATLAB 2022b) where SNPs were coded under the additive genetic model (0, 1, 2). No further adjustment for covariates was necessary since facial surface scans were already adjusted prior to the calculation of any univariate phenotype (see Dataset and preprocessing). This generated effect size and standard error estimates for the US

and UK cohort separately which were then meta-analyzed using the inverse-variance weighted method [59]. Meta P-values were computed using a two-tailed test.

Aggregation of multiple GWAS studies

To investigate the number of identified genetic loci under different numbers of traits, we gradually increased the absolute numbers of traits in each phenotype category. For nasal shape, the experiments were conducted with absolute numbers of traits equal to [1, 5, 10, 20, 30, 50, 100]. Since there were a limited number of inter-landmark distances and syndromic groups, the absolute numbers of traits were set to [1, 2, 4, 6, 8, 10] and [1, 5, 10, 23], respectively. Similarly, for facial shape, the experiments were conducted with absolute numbers of traits equal to [1, 10, 30, 70, 100, 200]. The absolute numbers of traits based on resemblance to syndrome gestalts were set to [1, 10, 20, 25].

To aggregate multiple GWASs on univariate traits within a phenotype group, we employed Tippett's minimal-p meta-approach [60]. Furthermore, for each aggregation, we controlled for the additional multiple testing burden by estimating the number of independent traits (i.e., the effective number of traits) within the group. This adjustment allowed us to correct the genome-wide significance threshold ($P < 5e\text{-}8$) to a group-wide significance threshold (5e-8 divided by the effective number of traits). Since PCA yielded mutually uncorrelated univariate features, the number of independent phenotypes was equal to the number of PCs used. For all other methods, this number was estimated using permutation testing [61]. Specifically, each of 7,417,619 SNPs was randomly permuted and the same GWASs were repeated once. This allowed to estimate the null-distribution of the minimum P-values for each SNP across the set of univariate traits. The

number of independent phenotypes was then estimated as 0.05 divided by the 5th percentile of this null distribution [61].

SNP-based heritability estimation

SNP-heritability is defined as the proportion of phenotypic variance that is explained by additive genetic effects of SNPs. First, SNPs were intersected with the HapMap3 SNPs and any SNP with non-matching alleles was removed, as well as SNPs within the major histocompatibility complex region. The SNP heritability of each univariate trait was then estimated with LDSC (published software https://github.com/bulik/ldsc/) [9] using the GWAS summary statistics of the EURO dataset. European derived LD scores were used in LDSC (downloaded from https://doi.org/10.5281/zenodo.7768714).

We conducted a two-tailed t-test to compare the mean SNP-heritability between groups of phenotypes. The results were adjusted for multiple testing using the Benjamini-Hochberg procedure [62] (Fig S1 in supplementary file 1).

Identification of genetic loci

Peak calling was performed in three steps, starting with the SNPs that reached the adjusted genome-wide significance threshold (5 x 10^{-8} divided by the effective number of traits). First, all SNPs within ±250 kb of the most significant SNP, as well as those within 1 Mb and in LD ($r^2 > 10^{-2}$) were clumped into a single locus represented by the most significant (lead) SNP. This was repeated until all SNPs were assigned a locus. Next, any two loci were merged if the representative lead SNPs were within 10 Mb and in LD ($r^2 > 10^{-2}$). This locus was then

602 represented by the SNP with the lowest P-value. Lastly, any peaks represented by a single SNP
603 below the adjusted genome-wide significance threshold were disregarded to improve
604 robustness.

605 Gene annotation

606 The most likely candidate gene per lead SNP was identified through a two-step process. First, we
607 utilized GREAT (v.4.0.4) [63] with default settings and the Table Browser of the UCSC Genome
608 Browser [64] for gene annotation. Then, we conducted literature searches to further support our
609 findings, based on the gene lists associated with facial morphology provided in [5].

610

Figures

Fig 1. Overview of phenotyping methods. (a) inter-landmark Euclidean distances computed between 24 anatomical facial landmarks. The 5 nasal landmarks in the blue nasal region are highlighted in red. (b) principal component analysis, which is based on a low-rank singular value decomposition (SVD) applied to a reshaped representation of the 3D shape data, where matrix multiplication is denoted by ·. (c) an auto-encoder network. The encoder consists of three spiral convolutional layers, followed by two fully connected layers. The decoder architecture mirrors the structure of the encoder. (d) resemblance-based measures, defined as the cosine distance operating on the angle between the target vector (e.g., a random face, an extreme face) and a sample vector.

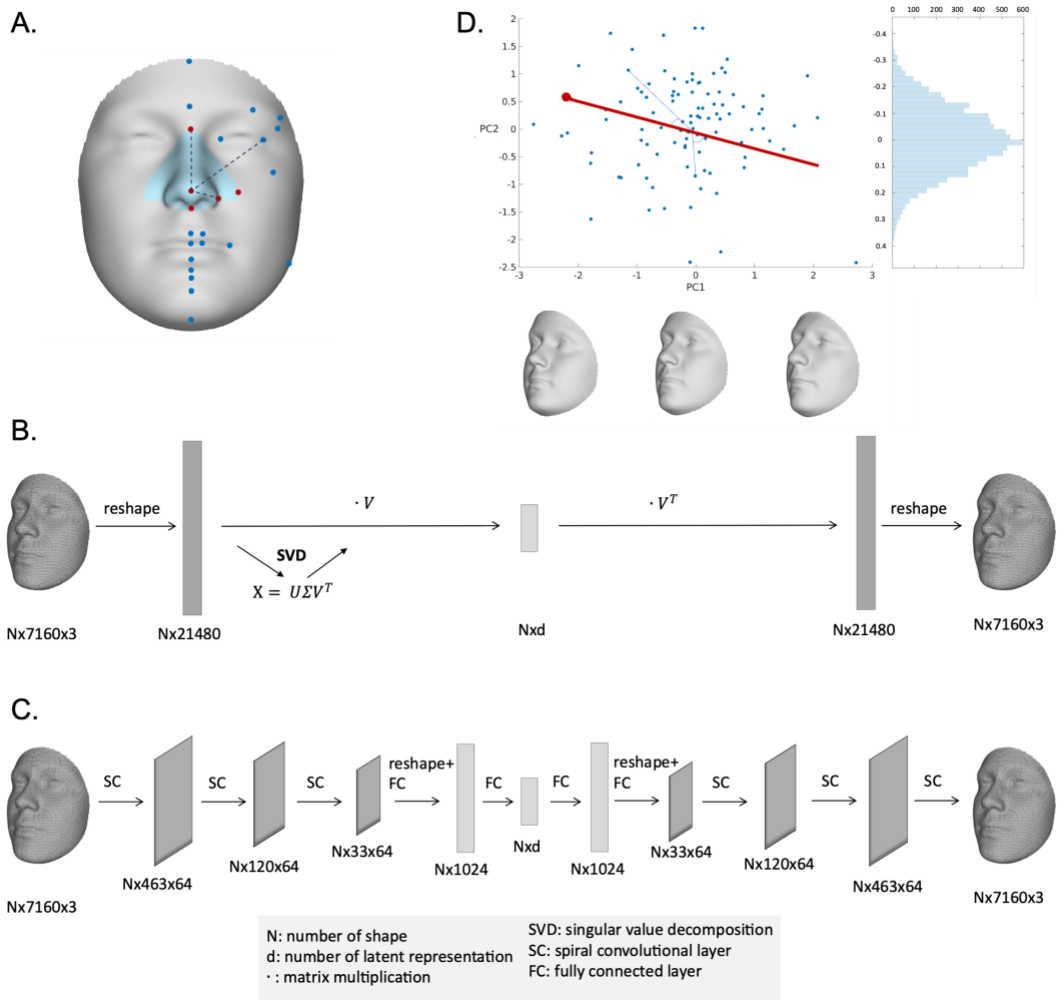


Fig 2. Comparison of SNP-based heritability between phenotyping categories. The colors represent different categories of traits: green for inter-landmark distances (DISTANCE), dark blue for traits extracted by auto-encoder (AE), light blue for traits extracted by principal component analysis (PCA), light red for resemblance scores to randomly selected facial examples (RANDOM), medium red for resemblance scores to extreme facial examples (EXTREME), and dark red for resemblance scores to syndrome facial archetypes (SYNDROME).

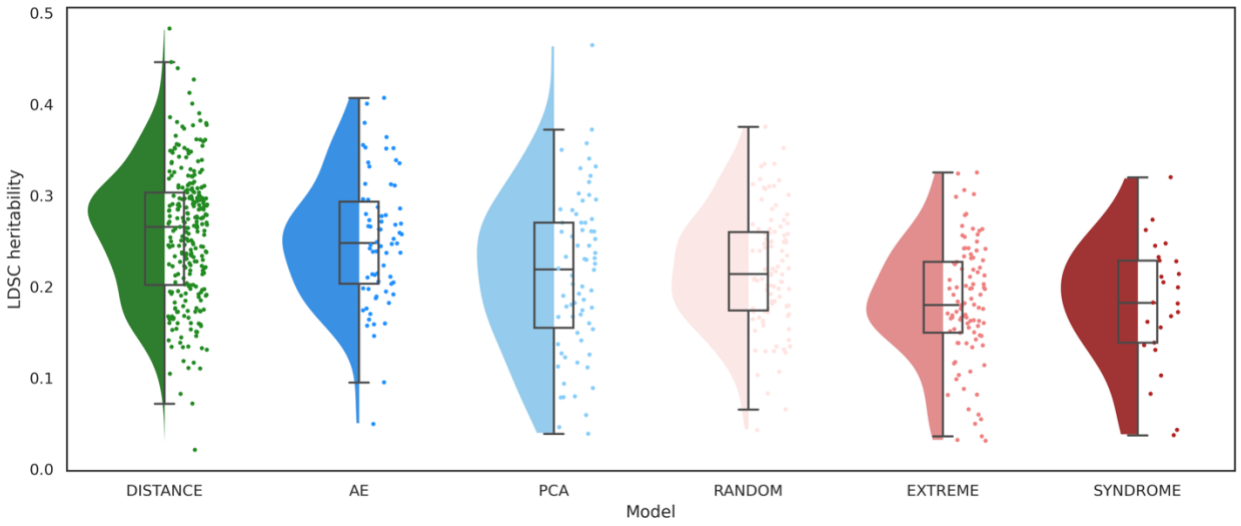


Fig 3. The interplay among the dimensionality of traits, the number of significant genetic loci, and the phenotypic variation. We compared nasal shape phenotypes (left columns) and full facial shape phenotypes (right columns) in terms of (a) the effective number of traits, (b) the effectiveness of identifying independent genetic loci through GWAS, and (c) the phenotypic variation captured by traits and their corresponding number of significant genetic loci in GWAS. The colors represent different categories of traits: green for inter-landmark distances (DISTANCE), dark blue for traits extracted by auto-encoder (AE), light blue for traits extracted by principal component analysis (PCA), light red for resemblance scores to randomly selected gestalts (RANDOM), medium red for resemblance to extreme gestalts (EXTREME), and dark red for resemblance scores to syndromic gestalts (SYNDROME). Unlike PCs, which are ordered according to descending explained variance, and resemblance scores to extreme gestalts based on the cosine distance to the mean shape, there is no specific order within other categories of traits. Therefore, given a fixed absolute number of traits, we randomly selected a subset 10 times from the full set of inter-landmark distances and resemblance to syndromic gestalts. Additionally, 10 replicates were performed for generating multiple AE latent dimensions and resemblance to random gestalts under different random initializations. The error bars represent the variation in results obtained from these 10 replicates.

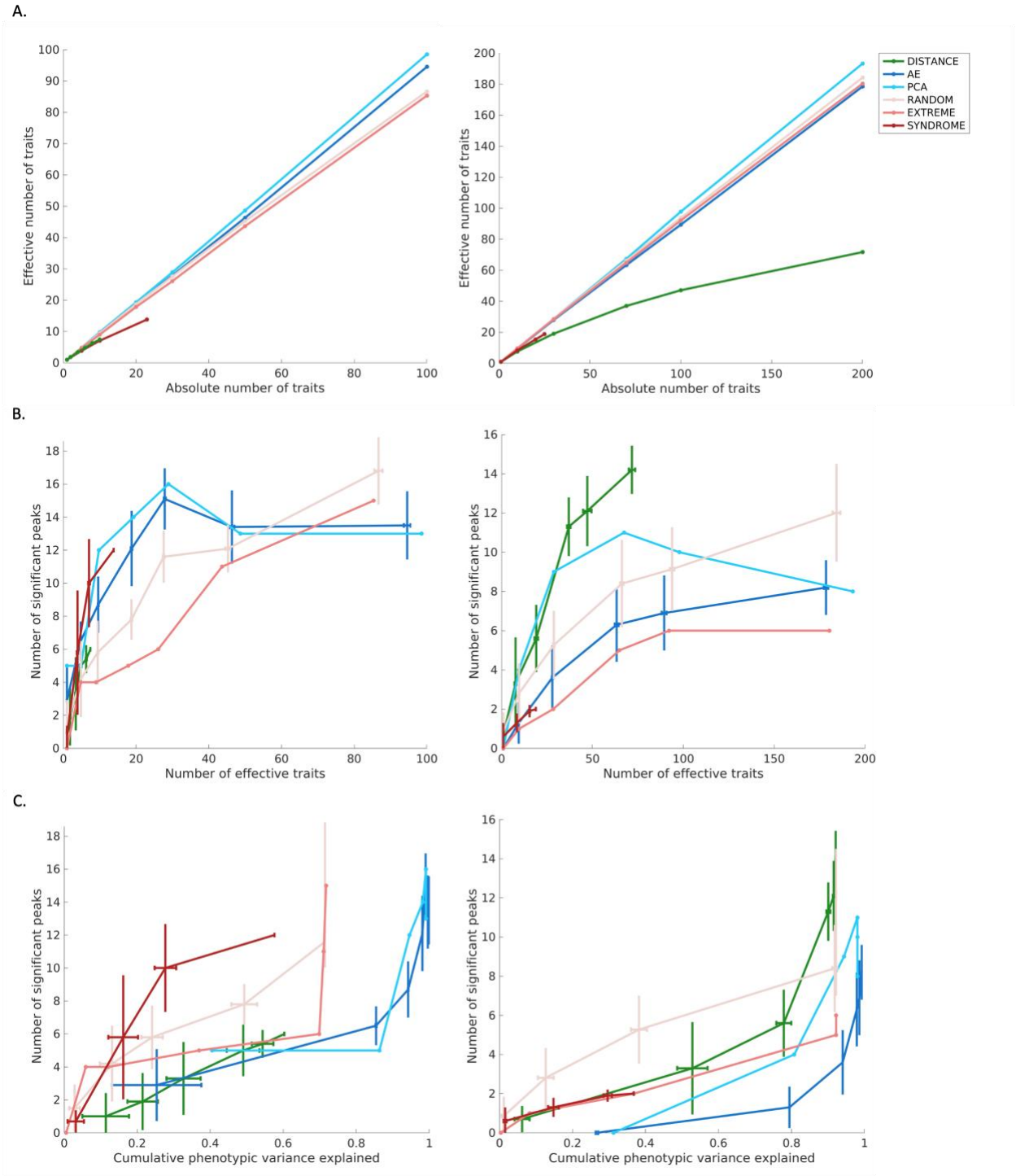
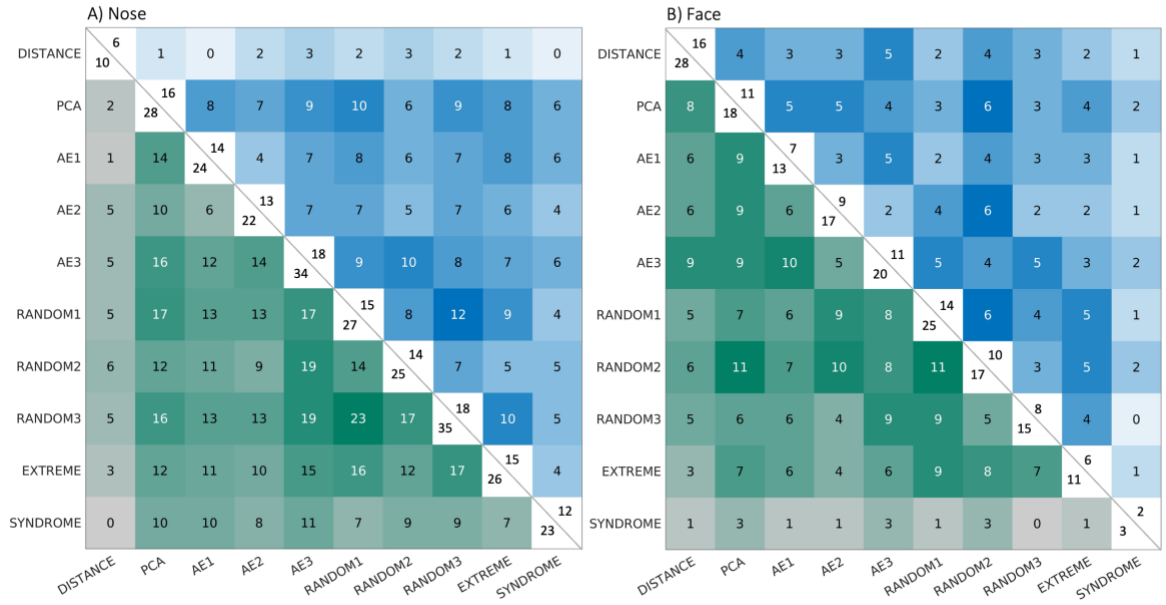


Fig 4. Comparing phenotypes in terms of overlapping genetic findings from GWAS. The number of overlapping genetic loci (within +/-250kb) is given in blue and the number of overlapping genes, annotated using GREAT, in green. The significant genetic loci were identified using the optimal number of effective traits, i.e., when the number of independently significant genetic loci after multiple testing correction was at its maximum. This maximal total number of genetic

loci per phenotyping category is displayed at the upper-right corner of the diagonal, and the corresponding number of annotated genes is displayed at the lower-left corner of the diagonal. Phenotypes include inter-landmark distances (DISTANCE), traits extracted by auto-encoder (AE), traits extracted by principal component analysis (PCA), resemblance scores to randomly selected gestalts (RANDOM), resemblance scores to extreme gestalts (EXTREME), and resemblance scores to syndromic gestalts (SYNDROME). To show the variability in the results introduced by random initializations on AE, we provide the results of three replicates denoted as AE1, AE2, and AE3. Similarly, we conducted three replicates for resemblance scores to randomly selected gestalts, denoted as RANDOM1, RANDOM2, and RANDOM3.



References

[1] Liu F, van der Lijn F, Schurmann C, Zhu G, Chakravarty MM, Hysi PG, et al. A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans. PLoS Genet 2012;8:e1002932-.

[2] Paternoster L, Zhurov A, Toma A, Kemp J, St. Pourcain B, Timpson N, et al. Genome-wide Association Study of Three-Dimensional Facial Morphology Identifies a Variant in PAX3 Associated with Nasion Position. Am J Hum Genet 2012;90:478–485. https://doi.org/10.1016/j.ajhg.2011.12.021.

[3] Naqvi S, Hoskens H, Wilke F, Weinberg SM, Shaffer JR, Walsh S, et al. Decoding the Human Face: Progress and Challenges in Understanding the Genetics of Craniofacial Morphology. Annu Rev Genomics Hum Genet 2022;23:383–412. https://doi.org/10.1146/annurev-genom-120121-102607.

[4] Claes P, Roosenboom J, White JD, Swigut T, Sero D, Li J, et al. Genome-wide mapping of global-to-local genetic effects on human facial shape. Nat Genet 2018;50:414–23. https://doi.org/10.1038/s41588-018-0057-4.

[5] White JD, Indencleef K, Naqvi S, Eller RJ, Hoskens H, Roosenboom J, et al. Insights into the genetic architecture of the human face. Nat Genet 2021;53:45–53. https://doi.org/10.1038/s41588-020-00741-7.

[6] Naqvi S, Sleyp Y, Hoskens H, Indencleef K, Spence JP, Bruffaerts R, et al. Shared heritability of human face and brain shape. Nat Genet 2021;53:830–9. https://doi.org/10.1038/s41588-021-00827-w.

[7] Zhang M, Wu S, Du S, Qian W, Chen J, Qiao L, et al. Genetic variants underlying differences in facial morphology in East Asian and European populations. Nat Genet 2022;54:403–11. https://doi.org/10.1038/s41588-022-01038-7.

[8] Liu C, Lee MK, Naqvi S, Hoskens H, Liu D, White JD, et al. Genome scans of facial features in East Africans and cross-population comparisons reveal novel associations. PLoS Genet 2021;17:e1009695-.

[9] Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 2015;47:291–5. https://doi.org/10.1038/ng.3211.

[10] Xiong Z, Dankova G, Howe LJ, Lee MK, Hysi PG, de Jong MA, et al. Novel genetic loci affecting facial shape variation in humans. Elife 2019;8:e49898. https://doi.org/10.7554/eLife.49898.

[11] Adhikari K, Fuentes-Guajardo M, Quinto-Sánchez M, Mendoza-Revilla J, Camilo Chacón-Duque J, Acuña-Alonzo V, et al. A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation. Nat Commun 2016;7:11616. https://doi.org/10.1038/ncomms11616.

[12] Bonfante B, Faux P, Navarro N, Mendoza-Revilla J, Dubied M, Montillot C, et al. A GWAS in Latin Americans identifies novel face shape loci, implicating VPS13B and a Denisovan introgressed region in facial variation. Sci Adv 2021;7:eabc6160. https://doi.org/10.1126/sciadv.abc6160.

[13] Shaffer JR, Orlova E, Lee MK, Leslie EJ, Raffensperger ZD, Heike CL, et al. Genome-Wide Association Study Reveals Multiple Loci Influencing Normal Human Facial Morphology. PLoS Genet 2016;12:e1006149-.

[14] Cole JB, Manyama M, Kimwaga E, Mathayo J, Larson JR, Liberton DK, et al. Genomewide Association Study of African Children Identifies Association of SCHIP1 and PDE8A with Facial Size and Shape. PLoS Genet 2016;12:e1006174-.

[15] Cha S, Lim JE, Park AY, Do J-H, Lee SW, Shin C, et al. Identification of five novel genetic loci related to facial morphology by genome-wide association studies. BMC Genomics 2018;19:481. https://doi.org/10.1186/s12864-018-4865-9.

719 [16] Qiao L, Yang Y, Fu P, Hu S, Zhou H, Peng S, et al. Genome-wide variants of Eurasian facial
720 shape differentiation and a prospective model of DNA based face prediction. Journal of
721 Genetics and Genomics 2018;45:419–32.
722 https://doi.org/https://doi.org/10.1016/j.jgg.2018.07.009.

723 [17] Wu W, Zhai G, Xu Z, Hou B, Liu D, Liu T, et al. Whole-exome sequencing identified four loci
724 influencing craniofacial morphology in northern Han Chinese. Hum Genet 2019;138:601–
725 11. https://doi.org/10.1007/s00439-019-02008-6.

726 [18] Li Q, Chen J, Faux P, Delgado ME, Bonfante B, Fuentes-Guajardo M, et al. Automatic
727 landmarking identifies new loci associated with face morphology and implicates
728 Neanderthal introgression in human nasal shape. Commun Biol 2023;6:481.
729 https://doi.org/10.1038/s42003-023-04838-7.

730 [19] Crouch DJM, Winney B, Koppen WP, Christmas WJ, Hutnik K, Day T, et al. Genetics of the
731 human face: Identification of large-effect single gene variants. Proceedings of the National
732 Academy of Sciences 2018;115:E676–85. https://doi.org/10.1073/pnas.1708207114.

733 [20] Liu D, Alhazmi N, Matthews H, Lee MK, Li J, Hecht JT, et al. Impact of low-frequency coding
734 variants on human facial shape. Sci Rep 2021;11:748. https://doi.org/10.1038/s41598-
735 020-80661-y.

736 [21] Hoskens H, Liu D, Naqvi S, Lee MK, Eller RJ, Indencleef K, et al. 3D facial phenotyping by
737 biometric sibling matching used in contemporary genomic methodologies. PLoS Genet
738 2021;17:e1009528-.

739 [22] Vanneste M, Hoskens H, Goovaerts S, Matthews H, Aponte JD, Cole J, et al. Syndrome-
740 informed phenotyping identifies a polygenic background for achondroplasia-like facial
741 variation in the general population. BioRxiv 2023:2023.12.07.570544.
742 https://doi.org/10.1101/2023.12.07.570544.

743 [23] Naqvi S, Kim S, Hoskens H, Matthews HS, Spritz RA, Klein OD, et al. Precise modulation of
744 transcription factor levels identifies features underlying dosage sensitivity. Nat Genet
745 2023;55:841–51. https://doi.org/10.1038/s41588-023-01366-2.

746 [24] Indencleef K, Hoskens H, Lee MK, White JD, Liu C, Eller RJ, et al. The Intersection of the
747 Genetic Architectures of Orofacial Clefts and Normal Facial Variation. Front Genet
748 2021;12.

749 [25] White JD, Ortega-Castrillón A, Matthews H, Zaidi AA, Ekrami O, Snyders J, et al. MeshMonk:
750 Open-source large-scale intensive 3D phenotyping. Sci Rep 2019;9:6085.
751 https://doi.org/10.1038/s41598-019-42533-y.

752 [26] Hoskens H, Li J, Indencleef K, Gors D, Larmuseau MHD, Richmond S, et al. Spatially Dense
753 3D Facial Heritability and Modules of Co-heritability in a Father-Offspring Design. Front
754 Genet 2018;9.

[27] Percival CJ, Devine J, Darwin BC, Liu W, van Eede M, Henkelman RM, et al. The effect of automated landmark identification on morphometric analyses. J Anat 2019;234:917–35. https://doi.org/https://doi.org/10.1111/joa.12973.

[28] Katina S, McNeil K, Ayoub A, Guilfoyle B, Khambay B, Siebert P, et al. The definitions of three-dimensional landmarks on the human face: an interdisciplinary view. J Anat 2016;228:355–65. https://doi.org/https://doi.org/10.1111/joa.12407.

[29] Guo Y, Liu Y, Oerlemans A, Lao S, Wu S, Lew MS. Deep learning for visual understanding: A review. Neurocomputing 2016;187:27–48. https://doi.org/https://doi.org/10.1016/j.neucom.2015.09.116.

[30] Chai J, Zeng H, Li A, Ngai EWT. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. Machine Learning with Applications 2021;6:100134. https://doi.org/https://doi.org/10.1016/j.mlwa.2021.100134.

[31] Nauwelaers N, Matthews H, Fan Y, Croquet B, Hoskens H, Mahdi S, et al. Exploring palatal and dental shape variation with 3D shape analysis and geometric deep learning. Orthod Craniofac Res 2021;24:134–43. https://doi.org/https://doi.org/10.1111/ocr.12521.

[32] Tran L, Liu X. Nonlinear 3D Face Morphable Model. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, p. 7346–55. https://doi.org/10.1109/CVPR.2018.00767.

[33] Björklund M. Be careful with your principal components. Evolution (N Y) 2019;73:2151–8. https://doi.org/https://doi.org/10.1111/evo.13835.

[34] Elhaik E. Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. Sci Rep 2022;12:14683. https://doi.org/10.1038/s41598-022-14395-4.

[35] Jong MA de, Wollstein A, Ruff C, Dunaway D, Hysi P, Spector T, et al. An Automatic 3D Facial Landmarking Algorithm Using 2D Gabor Wavelets. IEEE Transactions on Image Processing 2016;25:580–8. https://doi.org/10.1109/TIP.2015.2496183.

[36] Oualkacha K, Labbe A, Ciampi A, Roy M-A, Maziade M. Principal Components of Heritability for High Dimension Quantitative Traits and General Pedigrees. Stat Appl Genet Mol Biol 2012;11. https://doi.org/10.2202/1544-6115.1711.

[37] Ott J, Rabinowitz D. A Principal-Components Approach Based on Heritability for Combining Phenotype Information. Hum Hered 1999;49:106–11. https://doi.org/10.1159/000022854.

[38] Klei L, Luca D, Devlin B, Roeder K. Pleiotropy and principal components of heritability combine to increase power for association analysis. Genet Epidemiol 2008;32:9–19. https://doi.org/https://doi.org/10.1002/gepi.20257.

[39] Beavis WD. The power and deceit of QTL experiments: lessons from comparative QTL studies. Proceedings of the 49th Annual Corn and Sorghum Research Conference, American Seed Trade Association, Washington, DC, 1994, p. 250–65.

[40] Xu S. Theoretical Basis of the Beavis Effect. Genetics 2003;165:2259–68. https://doi.org/10.1093/genetics/165.4.2259.

[41] Weinberg SM, Raffensperger ZD, Kesterke MJ, Heike CL, Cunningham ML, Hecht JT, et al. The 3D Facial Norms Database: Part 1. A Web-Based Craniofacial Anthropometric and Image Repository for the Clinical and Research Community. The Cleft Palate Craniofacial Journal 2016;53:185–97. https://doi.org/10.1597/15-199.

[42] Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, et al. Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. Int J Epidemiol 2013;42:97–110. https://doi.org/10.1093/ije/dys066.

[43] Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, et al. Cohort Profile: The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. Int J Epidemiol 2013;42:111–27. https://doi.org/10.1093/ije/dys064.

[44] Mahdi SS, Matthews H, Vanneste M, Nauwelaers N, Gong S, Bouritsas G, et al. A 3D Clinical Face Phenotype Space of Genetic Syndromes using a Triplet-Based Singular Geometric Autoencoder. BioRxiv 2022:2022.12.27.521999. https://doi.org/10.1101/2022.12.27.521999.

[45] Hallgrímsson B, Aponte JD, Katz DC, Bannister JJ, Riccardi SL, Mahasuwan N, et al. Automated syndrome diagnosis by three-dimensional facial imaging. Genetics in Medicine 2020;22:1682–93. https://doi.org/10.1038/s41436-020-0845-y.

[46] Hammond P, Suttie M. Large-scale objective phenotyping of 3D facial morphology. Hum Mutat 2012;33:817–25. https://doi.org/https://doi.org/10.1002/humu.22054.

[47] Sero D, Zaidi A, Li J, White JD, Zarzar TBG, Marazita ML, et al. Facial recognition from DNA using face-to-DNA classifiers. Nat Commun 2019;10:2557. https://doi.org/10.1038/s41467-019-10617-y.

[48] Klingenberg CP. Morphometric integration and modularity in configurations of landmarks: tools for evaluating a priori hypotheses. Evol Dev 2009;11:405–21. https://doi.org/https://doi.org/10.1111/j.1525-142X.2009.00347.x.

[49] Robert P, Escoufier Y. A Unifying Tool for Linear Multivariate Statistical Methods: The RV-Coefficient. J R Stat Soc Ser C Appl Stat 1976;25:257–65. https://doi.org/10.2307/2347233.

[50] Hayton JC, Allen DG, Scarpello V. Factor Retention Decisions in Exploratory Factor Analysis: a Tutorial on Parallel Analysis. Organ Res Methods 2004;7:191–205. https://doi.org/10.1177/1094428104263675.

[51] Franklin SB, Gibson DJ, Robertson PA, Pohlmann JT, Fralish JS. Parallel Analysis: a method for determining significant principal components. Journal of Vegetation Science 1995;6:99–106. https://doi.org/https://doi.org/10.2307/3236261.

[52] Mahdi SS, Nauwelaers N, Joris P, Bouritsas G, Gong S, Bokhnyak S, et al. 3D Facial Matching by Spiral Convolutional Metric Learning and a Biometric Fusion-Net of Demographic Properties. 2020 25th International Conference on Pattern Recognition (ICPR), 2021, p. 1757–64. https://doi.org/10.1109/ICPR48806.2021.9412166.

[53] Gong S, Chen L, Bronstein M, Zafeiriou S. SpiralNet++: A Fast and Highly Efficient Mesh Convolution Operator. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, p. 4141–8. https://doi.org/10.1109/ICCVW.2019.00509.

[54] Lim I, Dielen A, Campen M, Kobbelt L. A Simple Approach to Intrinsic Correspondence Learning on Unstructured 3D Meshes. Computer Vision – ECCV 2018 Workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part III, Berlin, Heidelberg: Springer-Verlag; 2019, p. 349–62. https://doi.org/10.1007/978-3-030-11015-4_26.

[55] Cignoni P, Callieri M, Corsini M, Dellepiane M, Ganovelli F, Ranzuglia G. MeshLab: an Open-Source Mesh Processing Tool. In: Scarano V, Chiara R De, Erra U, editors. Eurographics Italian Chapter Conference, The Eurographics Association; 2008. https://doi.org/10.2312/LocalChapterEvents/ItalChap/ItalianChapConf2008/129-136.

[56] Hill H, Claes P, Corcoran M, Walters M, Johnston A, Clement J. How Different is Different? Criterion and Sensitivity in Face-Space. Front Psychol 2011;2:41. https://doi.org/10.3389/fpsyg.2011.00041.

[57] Griffin HJ, McOwan PW, Johnston A. Relative faces: Encoding of family resemblance relative to gender means in face space. J Vis 2011;11:8. https://doi.org/10.1167/11.12.8.

[58] Aeria G, Claes P, Vandermeulen D, Clement JG. Targeting specific facial variation for different identification tasks. Forensic Sci Int 2010;201:118–24. https://doi.org/https://doi.org/10.1016/j.forsciint.2010.03.005.

[59] Fleiss JL. Review papers : The statistical basis of meta-analysis. Stat Methods Med Res 1993;2:121–45. https://doi.org/10.1177/096228029300200202.

[60] Tippett LHC. Methods of Statistics. London: Williams Norgate; 1931.

[61] Kanai M, Tanaka T, Okada Y. Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set. J Hum Genet 2016;61:861–6. https://doi.org/10.1038/jhg.2016.72.

[62] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) 1995;57:289–300.

[63] McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 2010;28:495–501. https://doi.org/10.1038/nbt.1630.

[64] Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res 2004;32:D493–6. https://doi.org/10.1093/nar/gkh103.

Contributions

M.Y. - Conceptualization, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Data Curation

S.G.- Conceptualization, Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing, Data Curation

M.V.- Writing – Review & Editing, Data Curation, Resources

H.M.- Writing – Review & Editing, Software, Data Curation, Resources

H.H.- Writing – Review & Editing, Data Curation, Resources

S.R.- Writing – Review & Editing, Data Curation, Resources

O.D.K.- Funding Acquisition, Writing – Review & Editing, Data Curation, Resources

R.A.S.- Funding Acquisition, Writing – Review & Editing, Data Curation, Resources

B.H.- Funding Acquisition, Writing – Review & Editing, Data Curation, Resources

S.W.- Writing – Review & Editing, Data Curation, Resources

M.D.S.- Writing – Review & Editing, Data Curation, Resources

J.R.S.- Funding Acquisition, Writing – Review & Editing, Data Curation, Resources

S.M.W.- Funding Acquisition, Writing – Review & Editing, Data Curation, Resources

H.P.- Supervision, Writing – Review & Editing, Data Curation, Resources

P.C.-Conceptualization, Formal Analysis, Funding Acquisition, Project Administration, Supervision, Methodology, Data Curation, Resources

Acknowledgements

We are extremely grateful to all the individuals and families who took part in this study, the midwives for their help in recruiting them and the whole teams at ALSPAC, KU Leuven, and the universities of Pittsburgh, IUPUI, and Penn State which includes interviewers, computer and

laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses.

We acknowledge the use of ChatGPT v3.5 (https://chat.openai.com/) for English language editing. More specifically, ChatGPT v3.5 was used to check English spelling and grammar, without changing meaning or adding content.

Data availability

The genotype data of the 3DFN dataset are accessible via the dbGaP controlled access repository (http://www.ncbi.nlm.nih.gov/gap) at accession number phs000949.v1. p1. The phenotype data, represented as 3D facial surface in .obj format, are available through the FaceBase Consortium (https://www.facebase.org) at accession number FB00000491.01. Access to these 3D facial surface models requires proper institutional ethics approval and approval from the FaceBase data access committee.

The FaceBase repository in the syndromic face database, "Developing 3D Craniofacial Morphometry Data and Tools to Transform Dysmorphology", collected at patient support groups in the USA, Canada, and the UK. Facial images are available through FaceBase (https://www.facebase.org/chaise/record/#1/isa:data set/accession=FB00000861).

The participants making up the Peter Hammond's legacy 3D dysmorphology dataset, Penn State University (PSU) and Indiana University-Purdue University Indianapolis (IUPUI) datasets were not collected with broad data sharing consent. Given the highly identifiable nature of both facial and genomic information and unresolved issues regarding risks to participants of reidentification, participants were not consented for inclusion in public repositories or the posting of individual data. This restriction is not because of any personal or commercial interests. Further information about access to the raw 3D facial images and/or genomic data can be obtained from the respective ethics committees; the Ethics Committee Research UZ/KU Leuven (ec@uzleuven.be), the PSU IRB (IRB-ORP@psu.edu), and the IUPUI IRB (irb@iu.edu) for the Peter Hammond's legacy, PSU and IUPUI datasets, respectively.

For the ALSPAC (UK) data, please note that the study website contains details of all the data that is available through a fully searchable data dictionary and variable search tool (http://www.bristol.ac.uk/alspac/researchers/our-data/). Pregnant women resident in Avon, UK with expected dates of delivery between 1st April 1991 and 31st December 1992 were invited to take part in the study. The total sample size for analyses using any data collected after the age of seven is therefore 15,447 pregnancies, resulting in 15,658 foetuses. Of these 14,901 children were alive at 1 year of age. Consent for biological samples has been collected in accordance with the Human Tissue Act (2004). Genome wide genotyping data was generated by Sample Logistics and Genotyping Facilities at Welcome Sanger Institute and LabCorp (Laboratory Corporation of America) using support from 23andMe.

All relevant data to run future replications are provided online (https://doi.org/10.6084/m9.figshare.24867063.v1). This includes the facial template used, nasal landmark label, and mesh simplification scheme in AE models.

Code availability

KU Leuven provides the MeshMonk [25] v.0.0.6 spatially dense facial-mapping software, free to use for academic purposes (https://github.com/TheWebMonks/meshmonk). Matlab R2017b implementations of the hierarchical spectral clustering to obtain nasal segmentation are available from a previous publication [47] (https://doi.org/10.6084/m9.figshare.7649024). Code for training AE models is available at https://github.com/mm-yuan/autoencoder_3dface.

The analyses in this work were based on functions in Matlab R2022b, Python v3.7.8, MeshMonk v0.0.6, MeshLab v2020.03, LDSC v.1.0.1, GREAT v4.0.4.

Funding information

The KU Leuven research team (P.C., M.Y., S.G.) and analyses were supported by the Research Fund KU Leuven (BOF-C1, C14/20/081), and the Research Foundation-Flanders (FWO, G0D1923N). This work was funded in part by grants from the National Institute of Dental and Craniofacial Research: R01-DE027023 (S.M.W., J.R.S., P.C.) and U01DE024440 (R.A.S., O.D.K., B.H.).

The UK Medical Research Council and Wellcome (Grant ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf). Funding for the collection of 3D face shape scans was specifically provided by the MRC and Wellcome Trust (092731) and the University of Cardiff. This publication is the work of the authors, and they will serve as guarantors for the contents of this paper.

Competing interests

The authors declare no competing interests.

Supporting information

SupplementaryFile1 - Figure.docx

Fig S1. P-value matrix of pairwise differences in mean SNP-based heritability of different phenotyping categories.

Fig S2. Comparison of SNP-based heritability between phenotyping categories for nasal and facial shape.

Fig S3. Comparing facial phenotyping categories in terms of independent genetic loci identified in GWAS.

SupplementaryFile2 - Implementation.docx

SupplementaryFile3 - Table.xlsx

Supplementary Table 1: Syndrome Data.

Supplementary Table 2: Annotated genes based on the frequently identified peaks from nasal shape GWASs.

Supplementary Table 3: Annotated genes based on the frequently identified peaks from facial shape GWASs.