OXFORD

Briefings in Bioinformatics, 2025, **26(2)**, bbaf090
https://doi.org/10.1093/bib/bbaf090
Problem Solving Protocol

Optimized phenotyping of complex morphological traits: enhancing discovery of common and rare genetic variants

Meng Yuan[1,2,3,*], Seppe Goovaerts[2,3], Myoung K. Lee[4], Jay Devine[1,2,3], Stephen Richmond[5], Susan Walsh[6], Mark D. Shriver[7], John R. Shaffer[4,8], Mary L. Marazita[4,8], Hilde Peeters[2], Seth M. Weinberg[4,8], Peter Claes[1,2,3,9,*]

[1]Department of Electrical Engineering, ESAT/PSI, KU Leuven, Oude Markt 13, 3000 Leuven, Belgium
[2]Department of Human Genetics, KU Leuven, Oude Markt 13, 3000 Leuven, Belgium
[3]Medical Imaging Research Center, University Hospitals Leuven, Herestraat 49, 3000 Leuven, Belgium
[4]Center for Craniofacial and Dental Genetics, Department of Oral and Craniofacial Sciences, University of Pittsburgh, Pittsburgh, 4200 Fifth Ave, Pittsburgh, PA 15260, United States
[5]Applied Clinical Research and Public Health, School of Dentistry, Cardiff University, Cardiff CF10 3AT, United Kingdom
[6]Department of Biology, Indiana University Indianapolis, 420 University Blvd, Indianapolis 46202, IN, United States
[7]Department of Anthropology, Pennsylvania State University, 201 Old Main, University Park, PA 16802, United States
[8]Department of Human Genetics, University of Pittsburgh, 4200 Fifth Ave, Pittsburgh, PA 15260, United States
[9]Murdoch Children's Research Institute, 50 Flemington Rd, Parkville VIC 3052, Australia

*Corresponding authors. Meng Yuan, Department of Electrical Engineering, ESAT/PSI, KU Leuven, Leuven, Belgium. E-mail: meng.yuan@kuleuven.be; Peter Claes, Department of Electrical Engineering, ESAT/PSI, KU Leuven, Leuven, Belgium. E-mail: peter.claes@kuleuven.be

Abstract

Genotype–phenotype (G-P) analyses for complex morphological traits typically utilize simple, predetermined anatomical measures or features derived via unsupervised dimension reduction techniques (e.g. principal component analysis (PCA) or eigen-shapes). Despite the popularity of these approaches, they do not necessarily reveal axes of phenotypic variation that are genetically relevant. Therefore, we introduce a framework to optimize phenotyping for G-P analyses, such as genome-wide association studies (GWAS) of common variants or rare variant association studies (RVAS) of rare variants. Our strategy is two-fold: (i) we construct a multidimensional feature space spanning a wide range of phenotypic variation, and (ii) within this feature space, we use an optimization algorithm to search for directions or feature combinations that are genetically enriched. To test our approach, we examine human facial shape in the context of GWAS and RVAS. In GWAS, we optimize for phenotypes exhibiting high heritability, estimated from either family data or genomic relatedness measured in unrelated individuals. In RVAS, we optimize for the skewness of phenotype distributions, aiming to detect commingled distributions that suggest single or few genomic loci with major effects. We compare our approach with eigen-shapes as baseline in GWAS involving 8246 individuals of European ancestry and in gene-based tests of rare variants with a subset of 1906 individuals. After applying linkage disequilibrium score regression to our GWAS results, heritability-enriched phenotypes yielded the highest SNP heritability, followed by eigen-shapes, while commingling-based traits displayed the lowest SNP heritability. Heritability-enriched phenotypes also exhibited higher discovery rates, identifying the same number of independent genomic loci as eigen-shapes with a smaller effective number of traits. For RVAS, commingling-based traits resulted in more genes passing the exome-wide significance threshold than eigen-shapes, while heritability-enriched phenotypes lead to only a few associations. Overall, our results demonstrate that optimized phenotyping allows for the extraction of genetically relevant traits that can specifically enhance discovery efforts of common and rare variants, as evidenced by their increased power in facial GWAS and RVAS.

Keywords: genotype–phenotype association; complex trait; heritability; facial phenotyping; 3D shape

Introduction

Determining and measuring phenotypes is crucial for the success of genotype–phenotype (G-P) analyses. Extracting complex morphological traits is particularly challenging due to their multivariate and multipartite nature. Traditional morphometric approaches are widely used in G-P studies to analyze various structures, such as the human face [1–11], limb, and skeleton [12, 13]. These methods often rely on linear distances, distance ratios, and angles between key points or anatomical landmarks. Unfortunately, landmarks are often criticized for oversimplifying shape variation, reducing it to a limited set of sparse, predefined metrics. They can also be difficult to acquire, as extensive clinical or biological knowledge is needed to identify.

Alternative computational techniques like principal component analysis (PCA) have been applied to structures such as the human face [14–20], cranium, and brain [21–24] to address these limitations in phenotype simplification. These data-driven methods can extract comprehensive representations of three-dimensional (3D) morphology without supervision or predefined traits. PCA, for instance, involves decomposing high-dimensional correlated facial data into fewer dimensions or principal

Received: October 31, 2024. **Revised:** February 3, 2025. **Accepted:** February 18, 2025

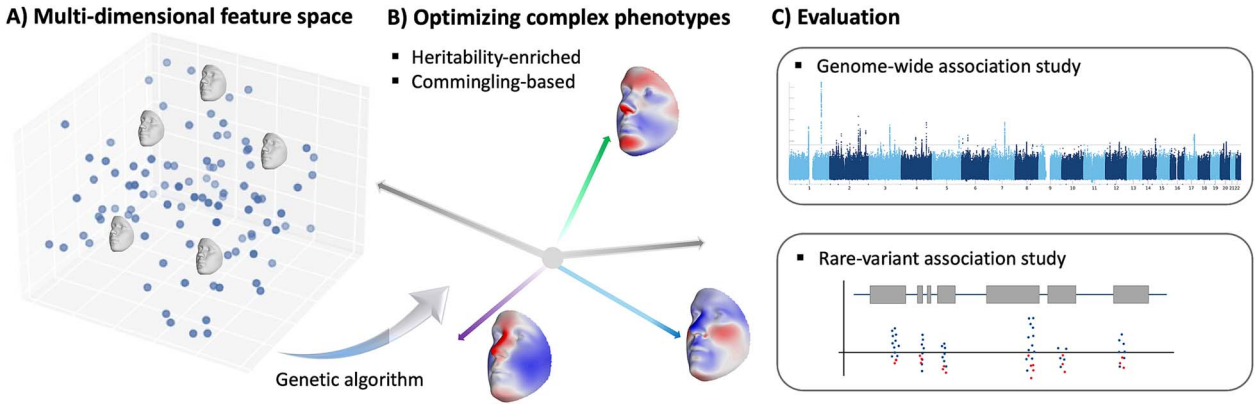

Figure 1. Schematic of the phenotyping framework.

components (PCs), often referred to as "eigen-faces" or "eigen-shapes" [25, 26]. These PCA-based features can capture nearly all geometric shape variations and are highly effective when analyzed in multivariate genome-wide association studies (GWAS) [15]. However, the biological relevance of individual components becomes questionable when treated independently, as these features do not necessarily reflect the most genetically relevant aspects of variation [20]. Some studies [14, 15, 24] have used a fixed number of eigen-shapes as phenotypes for GWAS based on their cumulative variance explained, whereas others [19, 27] have prioritized traits via heritability, only retaining the most heritable eigen-shapes. Likewise, another study [28] introduced a family-based design to generate sibling-shared phenotypes at an individual level, as close phenotypic similarity among sibling pairs suggests high heritability.

Moving beyond predefined measurements, unsupervised feature extraction, and phenotype selection, we introduce a novel, genetically informed phenotyping method that optimizes phenotypes at the population level. This approach was first explored in our previous work [29], where its application was limited to studies with family data availability and focused solely on identifying common genetic variants in GWAS. In the current work, we have expanded this framework to be applicable not only to family data but also to datasets that do not contain closely related individuals. Moreover, this method can now be used for a wider range of G-P analyses, enabling the detection of both common and rare genetic variant associations. Our strategy, illustrated in Fig. 1, is two-fold. First, we employ PCA to construct a lower-dimensional feature space composed of a wide range of complex and multi-dimensional shape variations. Second, we utilize a genetic algorithm (GA), an optimization approach that mimics evolution [30], to search for directions or traits in this feature space that are genetically relevant. This approach is flexible and can be adapted for various purposes, such as finding common genetic variants in GWAS or rare variants in rare variant association studies (RVAS).

For GWAS, our method involves training a GA model to obtain heritability-enriched phenotypes by identifying directions in feature space with high heritability. We explore two possibilities: (i) *GA-Family optimized traits*: When family data is available, the objective function is based on family-based trait heritability, which can be estimated using twin pairs, parent-offspring trios, or sibling pairs. In this work, more explicitly, we use the dataset and a regression of offspring traits on parental traits from the work of Hoskens et al. [28]; (ii) *GA-GREML optimized traits*: This approach leverages SNP genotypes from unrelated individuals, applying the Genomic Relatedness Matrix Restricted Maximum Likelihood (GREML) [31, 32] method to estimate SNP heritability. The GA model evolves according to these estimates. The rationale behind this approach is that variation with a strong genetic basis is likely to span directions in the feature space with high heritability. Previous studies [33–35] reveal varying heritability in PCs, suggesting the untapped potential to identify more heritable axes.

For RVAS, the GA model optimizes the skewness of trait distributions as its objective, resulting in *GA-Commingling optimized traits*. This is inspired by a commingling analysis, which fits different genetic models to observed phenotypic distributions. This technique was commonly used in the pregenomic era to provide preliminary evidence for a single or few genomic loci with a major effect [36, 37]. Large-effect variants, often causative in rare monogenic disorders or traits following Mendelian inheritance patterns, are typically rare and identified through RVAS. In contrast, common variants with small individual effects cumulatively contribute to polygenic traits and are identified through GWAS. Moreover, large-effect variants tend to produce commingled (skewed) phenotypic distributions, as evidenced in traits such as Lipoprotein (a) levels [38, 39]. Based on these insights, traits with skewed distributions are expected to be enriched for less common to rare variants with larger effects. While similar ideas have been explored previously, especially in the pre-genomic era working from phenotype data only, they have never been applied in this specific manner, making our proposed commingling-based optimization a particularly novel and promising contribution.

To validate our approach, we focus on human facial shape, a complex trait strongly influenced by genetic factors [10, 18, 19, 33–35, 40–42] and for which we have extensive resources and data available. We compare our proposed optimized phenotypes against eigen-shapes in a facial GWAS involving 8246 individuals of European ancestry and in exome-wide gene-based tests of rare variants in a subset of 1906 individuals. Specifically, we evaluate heritability and number of genetic signals identified by each phenotyping method. Our findings demonstrate that the data-driven optimized phenotyping approach enhances genetic discovery for facial shape in both GWAS and RVAS by extracting traits that capture a richer set of genetically relevant information. More broadly, we demonstrate how optimized phenotypes are a compelling alternative to conventional phenotypes and underscore that this framework can be generalized to other G-P association analyses of complex morphological traits.

Results

Trait optimization and testing

To derive GA-Family optimized traits, we used two family-based datasets and one larger dataset of unrelated individuals. The first

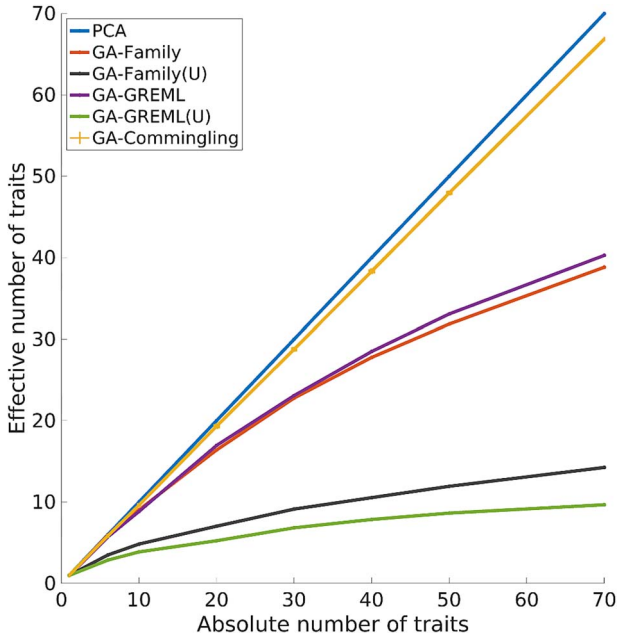

Figure 2. Correlation of phenotypes.

was the Avon Longitudinal Study of Parents and their Children (ALSPAC) dataset [43, 44], which includes 770 father-offspring pairs with 3D facial images. The second was the Technopolis dataset [28], consisting of 163 parent-offspring trios. Both family datasets were aligned and integrated into the facial feature space of a larger European cohort (EURO dataset) from [15], which includes 8246 unrelated individuals. A summary of the datasets and their respective analyses is provided in Supplementary File 2 and Supplementary Table S1.

In the first step, the ALSPAC dataset was used as a training dataset to optimize traits. Multiple rounds of GA optimization resulted in an increasing number of traits. Fig. 2 shows the relationship between the absolute number of traits (i.e. the number of GA rounds) and the effective number of traits (see Supplementary File 1, Section 3.1). This comparison highlights the correlation among the optimized traits, and it is useful to contrast this with the null correlation of principal components. Since PCA, by design, constructs uncorrelated dimensions, the absolute and effective number of traits are equal, represented by a diagonal line in Fig. 2. The high correlation observed across multiple rounds of optimization suggests that the process consistently reaches the same global minimum or few local minima, producing very similar traits each time. However, this repetition of the same traits is not particularly useful. Therefore, to increase trait diversity, an additional constraint is introduced during consecutive runs (Methods and Supplementary File 1, Implementation). This constraint monitors the correlation with previously optimized traits and forces new traits to exhibit low correlation with the others (see Fig. 2). The GA-Family optimized traits from the unconstrained model exhibited higher correlation, reducing the 70 absolute dimensions to 14, while the constrained model resulted in 39 independent dimensions. Therefore, in subsequent analyses, we consistently use the GA-Family optimized traits from constrained models only.

Subsequently, we assessed the generalizability of the trained model with increased trait diversity in the independent Technopolis dataset. Within both the training and test sets, the median trait heritability was higher for GA-Family optimized traits than for eigen-shapes (Fig. 3a). While the median heritability was statistically significantly different between the GA-Family optimized traits and eigen-shapes in the test data ($P = .0366$ from the Wilcoxon rank sum test for equal medians), the increase in trait heritability appeared less pronounced than in the training set. Detailed summary statistics and p-values are provided in Supplementary File 2 and Supplementary Table S2. This modest increase in the test data was not surprising, as heritability estimates vary by population and sample, making generalization difficult.

To derive GA-GREML optimized traits, we focused on the EURO dataset only, for which both genotype and phenotype data were available. More specifically this dataset consists of a US cohort ($n = 4680$), which was used for training, and a UK cohort ($n = 3566$), which served as an independent test set. Much like GA-Family optimized traits, multiple GA runs generated highly correlated traits, imposing the need for an additional constraint to increase trait diversity (Fig. 2). Compared to eigen-shapes, the median SNP-heritability of GA-GREML optimized traits with increased diversity was significantly higher in both the training and test sets, as estimated using a GREML-based approach [45] ($P = 1.7922e\text{-}24$ for training, $P = .0157$ for testing from the Wilcoxon rank-sum test for equal medians). However, the relative increase was more modest in the test set (Fig.3b and Supplementary File 2 and Supplementary Table S2).

To derive GA-Commingling optimized traits, we again focused on the EURO dataset and analyzed the phenotype distribution along various directions in the feature space using a commingling analysis. Interestingly, compared to previous cases, multiple GA runs produced a more diverse range of traits without the need to impose additional constraints on trait correlation. GA-Commingling optimized traits exhibited lower correlation, resulting in 67 independent dimensions out of the original 70 (Fig. 2). This suggests that traits likely influenced by one or a few loci with major effects result in the generation of multiple local minima during optimization. As a result, random initializations tend to converge on different local minima, enhancing trait diversity. Thus, GA-Commingling traits are derived from unconstrained models only in subsequent analyses. Similarly, across both the training and test sets, the median skewness was statistically significantly higher for GA-Commingling optimized traits than for eigen-shapes (Fig. 3c and Supplementary File 2; Supplementary Table S2).

Genome-wide association studies

We then evaluated our optimization pipeline in a genome-wide association meta-analysis using the entire EURO dataset. In the first instance, we performed linkage disequilibrium score regression (LDSC) [46] on GWAS summary statistics to estimate GWAS-based SNP heritability. Figure 4a illustrates how the median LDSC-based heritability is significantly higher for heritability-enriched phenotypes than for eigen-shapes (*P*-values between 5.1421e-21 and 2.4826e-11 from the Wilcoxon rank-sum test for equal medians). GA-GREML optimized traits exhibited the highest heritability, followed by GA-Family optimized traits, with eigen-shapes ranking third. Notably, GA-Commingling optimized traits had the lowest median heritability, suggesting a diminished contribution of common genetic effects. Detailed summary statistics and p-values are available in Supplementary File 2 and Supplementary Tables S3 and S4.

In a second instance, we investigated genetic discovery rates across methods by enumerating the genomic loci associated with different numbers of traits. Specifically, the number of traits within the same phenotype category was incrementally enlarged

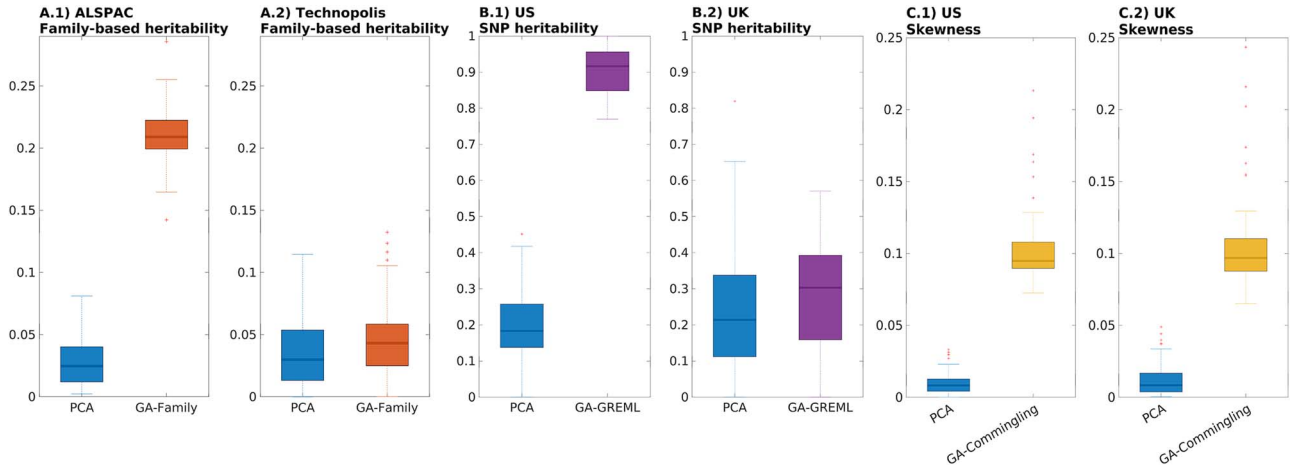

Figure 3. Evaluation of model development and generalizability.

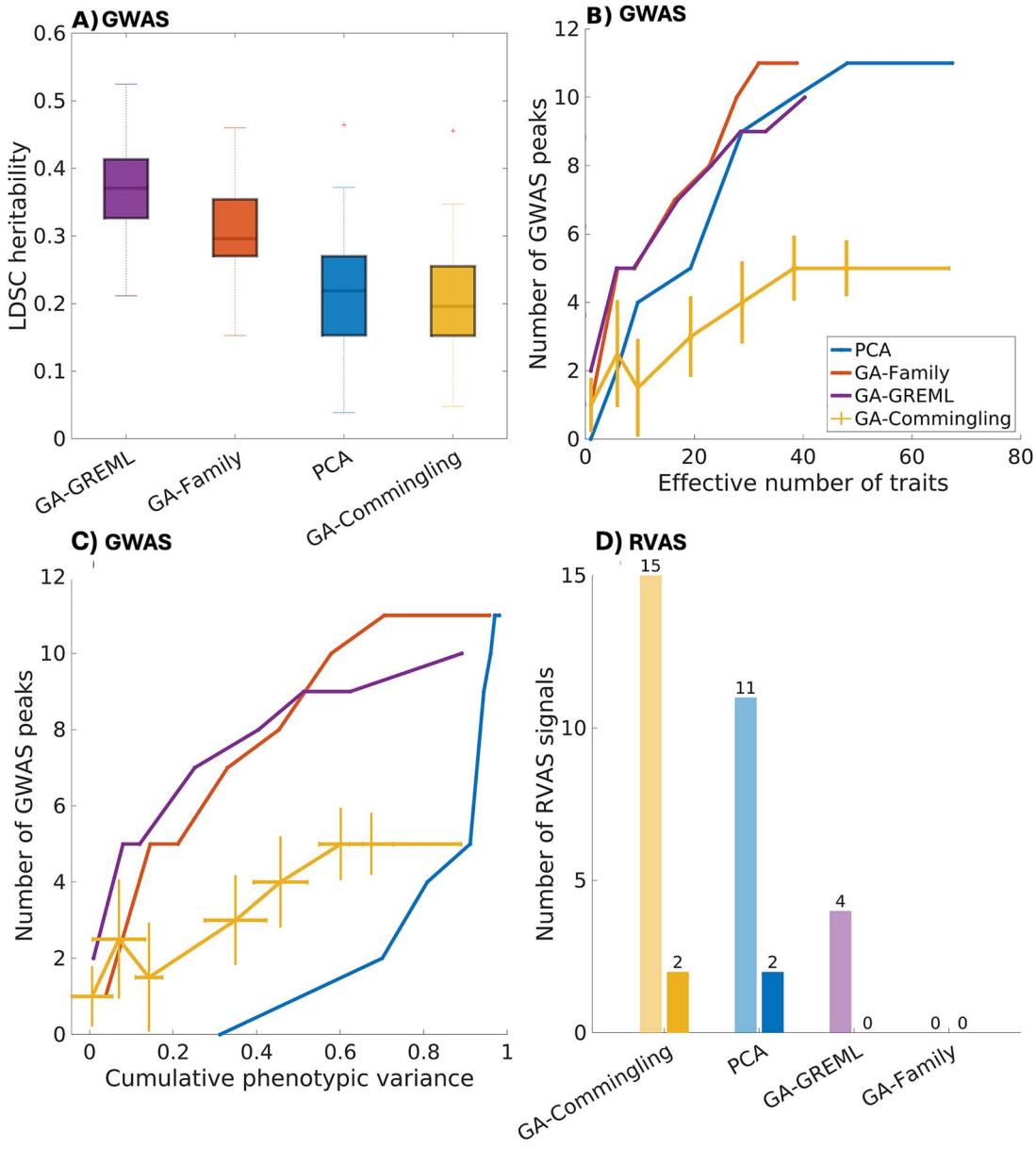

Figure 4. Comparison of genetic relevance across phenotypes.

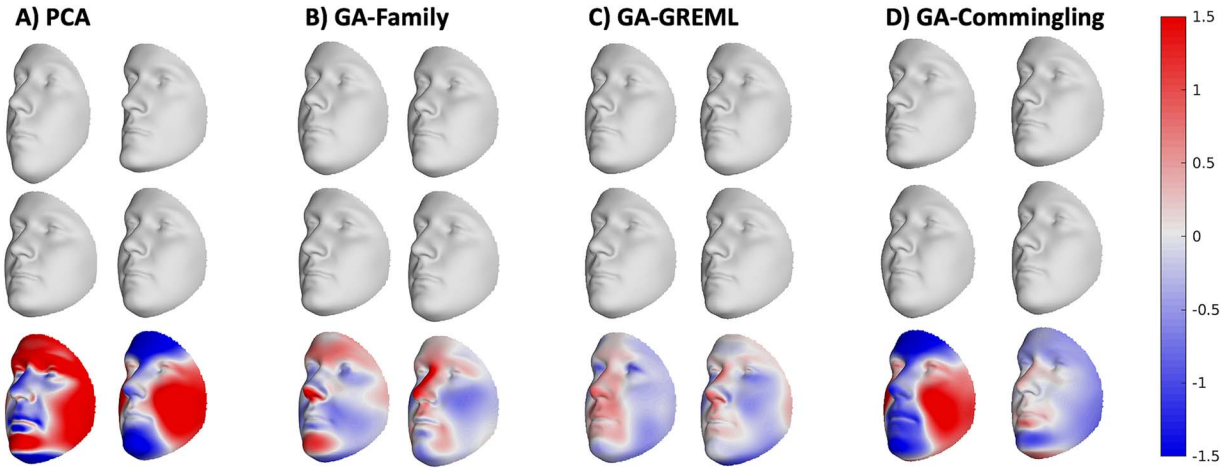

Figure 5. Visualization of facial phenotypes.

and subjected to GWAS. After aggregating the number of identified loci across multiple univariate GWASs, we observed that the number of identified loci generally increased with the effective number of traits (Fig. 4b). However, more importantly, the heritability-enriched phenotypes exhibited the steepest slopes, suggesting that fewer independent traits were needed to yield more genomic loci. For example, the maximum number of significant loci ($n = 11$) was obtained for GA-Family optimized traits using only 39 independent traits, whereas eigen-shapes required 70 independent traits to reach the same number. GA-GREML optimized traits identified a slightly lower maximum of significant loci ($n = 10$) with 40 independent traits. In contrast, the curve resulting from GA-Commingling optimized traits situated lowest, identifying the least genomic loci ($n = 5$) when 67 independent traits were considered.

Figure 4c shows the relationship between the total number of genomic loci identified by GWAS and the amount of phenotypic variance explained by different trait categories. Detailed numerical data underlying this figure can be found in the Supplementary File 2 and Supplementary Table S5. The highest order GA-Family and GA-GREML optimized traits clearly explained less phenotypic variance than eigen-shapes, but again their slopes were much steeper, indicating better GWAS discovery rates. For example, the first GA-Family and GA-GREML optimized traits captured 3.8% and 0.9% of shape variation, respectively, yet still identified one and two significant genomic loci, respectively. By contrast, the first eigen-shape captured 31.22% of shape variation but was not significantly associated with any genomic loci. Only after considering the first six eigen-shapes, which in total explained 70.05% of shape variation, were two independent genomic loci identified. Whereas the discovery rate for GA-Commingling optimized traits fell between those of the eigen-shapes and the heritability-enriched traits, the genetic discovery for eigen-shapes rapidly increased after six traits and ultimately reached its limit at 50 eigen-shapes. In short, although the majority of the phenotypic variance is explained by the highest order eigen-shapes, this variation is not necessarily correlated with the underlying genetics. Figure 5 supports this finding through visualization of the facial features, with all facial traits available online (Supplementary data). While the phenotypes associated with the first few eigen-shapes spanned the entire face, the heritability-enriched phenotypes were more focused and localized to highly heritable regions, like the nose and supraorbital ridge. Interestingly, despite being optimized for different heritability measures, GA-Family and GA-GREML optimized traits resulted in similar facial features.

Rare variant association studies

In a last experiment intended to identify rare variant associations with facial phenotypes, we performed exome-wide gene-based tests on the PITT cohort ($n = 1906$) using SKAT-O [47] implemented in [48]. In Fig. 4d, GA-Commingling optimized traits resulted in the highest number of significant genes at the exome-wide level ($n = 15$, $P < 3.16\text{e-}06$), with two genes also meeting the more stringent group-wide threshold adjusted for the effective number of traits ($P < 4.72\text{e-}08$, see *Methods* section). Following this, eigen-shapes identified 11 genes at the exome-wide threshold and 2 genes at the group-wide threshold. In contrast, phenotypes derived from GA-Family and GA-GREML identified 4 and 0 genes at the exome-wide threshold, respectively.

Among the genes identified using GA-Commingling optimized traits, two were previously associated with human syndromes involving facial dysmorphic features, whereas none of the genes identified by eigen-shapes were linked to any syndrome. One of these genes, *PTPN11* is implicated in LEOPARD syndrome [49–51] and Noonan syndrome [52–54]. The associated phenotype highlighted the nasal, philtrum, and chin regions (Fig. 5d), which correspond to typical facial features seen in Noonan syndrome [55]. Additionally, *TCF12* is linked to syndromic forms of craniosynostosis [56, 57] (corresponding trait shown in Fig. 5d). The complete list of identified genes is provided in Supplementary File 2 and Supplementary Table S6.

Discussion

In this work, we introduce an optimization-based phenotyping framework designed to identify both common and rare genetic influences on complex morphological traits, using facial shape as a case example. Our methodology first uses PCA to construct a low-dimensional feature space from 3D image data in which shape variations are heterogeneous and multidimensional. Afterward, we apply a genetic algorithm to scan this space for directions whose corresponding trait exhibits either high heritability or skewness. As anticipated, heritability-enriched phenotypes exhibited higher trait- and SNP-heritability than eigen-shapes. These heritability-enriched phenotypes also identified the same number of independent genomic loci with a lower effective number of

traits than eigen-shapes. In contrast, when analyzing rare variants, commingling-based skewed traits yielded more genes passing the exome-wide significance threshold than eigen-shapes, while heritability-enriched phenotypes resulted in very few associations. In summary, our proposed method represents a shift away from subjective trait selection towards data-driven trait optimization, where phenotypes are genetically informed and tailored specifically for G-P association analyses.

While PCA remains a key technique for decomposing high-dimensional data into fewer, more manageable latent variables, the extracted features are based on statistics, not biology. This brings into question their utility for GWAS, RVAS, and other genetic investigations [58, 59]. Compared to the first few eigen-shapes, heritability-enriched phenotypes identified more significant independent genomic loci, despite explaining less phenotypic variation (Fig. 4c). In addition, the first few eigen-shapes are also not the eigen-shapes resulting in rare variant associations as shown in Supplementary Table S6. Because PCA seeks to maximize variance, and facial shape is a highly integrated structure, eigen-shapes are constrained to a feature space dominated by a few major directions with large eigenvalues. Importantly, the construction of eigen-shapes is agnostic to both common and rare variants, and when working with limited sample sizes, as in this work, it is clear that optimized traits yield higher genetic discovery. Future work and additional data, including larger sample sizes, are needed to verify the genetic value of the initial eigen-shapes and whether they are mainly environmentally determined. In contrast, GA-optimized traits are not designed to maximize variance but instead prioritize localized phenotypes that are genetically relevant. For instance, the first few eigen-shapes encompassed facial regions known to have stronger environmental contributions (Fig. 5), such as the cheeks, mandible, and mouth, which are influenced by factors like nutrition, aging, and oral function [35, 40, 42]. By comparison, heritability-enriched traits targeted smaller, highly heritable facial regions. Many heritability studies have identified strong genetic influences predominantly in central midfacial parameters, such as the prominence and height of the nose and the upper lip philtrum length [35, 42, 60]. By optimizing for heritability, these traits demonstrate substantial genomic associations, leading to more effective identification of genetic factors on facial shape.

Previous studies have explored the idea of scoring traits along directions in PCA space, including GWASs on sibling-shared traits [28] and syndrome-informed phenotypes [61]. Another study [20] demonstrated that traits generated from random directions in feature space offer a promising alternative to eigen-shapes in GWAS, showing comparable median heritability and identifying more genomic loci than the first few eigen-shapes. Building on these findings, our previous work [29] took this further by implementing an optimization approach. We initialized this approach with a collection of random directions and later deployed a GA to search this space for as many genetically enriched directions as possible. In that prior work, the method was limited to studies with family data and focused solely on identifying common genetic variants in GWAS. In the present study, we expanded the framework to accommodate studies with both family data and unrelated individuals, broadening its application to detect both common and rare genetic variants. As anticipated, this method reorganized the feature space and emphasized traits with phenotypic distributions characteristic of either common or rare genetic effects. Given a limited number of dimensions, the heritability-enriched traits were associated with more independent significant loci than eigen-shapes. However, as the effective number of traits increased, the GA-Family optimized traits and eigen-shapes ultimately identified the same number of loci. Given that the GA optimization relies on a predefined PCA space, future work could investigate whether simultaneously constructing the feature space and optimizing feature extraction enhances performance.

Determining heritable traits through phenotypes alone (e.g. twin and parent-offspring study designs) is a long-standing strategy that predates the genomic era. These studies subjectively select phenotypes and only use trait heritability *post hoc* to rank traits [19, 27]. More biologically inspired approaches, like principal components of heritability (PCH) [62, 63], have managed to capture family structure across traits by calculating linear combinations of traits that maximize heritability. However, the ability of PCH to generalize to independent cohorts remains uncertain. The fact that heritability estimates may not necessarily apply across different populations in space or time [35, 64] is a general concern across phenotype-driven heritability research. Indeed, selecting phenotypes with a high trait heritability has been a convenient approach over the past few decades, but it has not necessarily resulted in convincing added value compared to other phenotypes. A major advantage of the heritability-driven phenotype optimization pipeline presented here is the transferability of heritable traits between populations or cohorts. We found that both the GA-Family and GA-GREML optimized traits exhibited high heritability in two independent cohorts and believe this is due to the optimization instead of a simple selection of traits.

Because the availability of large-scale family-based data is limited, we expanded on our approach by investigating the use of GREML-based SNP-heritability estimates derived from unrelated individuals. Like GA-Family optimized traits, our results demonstrate that GA-GREML optimized traits exhibit higher GREML-based heritability in both the training and an independent testing cohort. Additionally, higher GREML-based heritability, computed from population-level genetic similarity estimates (i.e. GRM), translates to higher LDSC-based heritability measures obtained from GWAS summary statistics. This clearly demonstrates that optimizing for GREML-based heritability has a favorable effect on GWAS outcomes. This extension also broadens the applicability of our approach, as it eliminates the need for family data, allowing researchers to use the same dataset for both GWAS and phenotype extraction.

To expand our phenotyping approach for rare variant identification, we proposed commingling-based skewed phenotypes. This method is inspired by a commingling analysis, where the skewed distribution of phenotypes provides initial evidence for a single or few genomic loci with a major effect [36, 37]. Despite our limited sample size, which may not be adequately powered for rare variant analysis, our results show that commingling-based skewed phenotypes can detect more associations in gene-based tests compared to eigen-shapes. Interestingly, even with the small sample size, we identified two genes, *PTPN11* and *TCF12*, which are plausible given their associations with syndromes characterized by craniofacial dysmorphism [49–54, 56, 57]. In future studies, leveraging whole exome or whole genome sequencing in larger samples may further validate the advantages of these commingling-based traits.

In summary, our phenotype optimization framework offers a way to *a priori* extract traits with high genetic and biological relevance, as demonstrated by their increased discovery rates in GWAS and RVAS. While we concentrated on individuals of European ancestry, this methodology can be flexibly modified for diverse populations and ancestries. Future research could

investigate phenotypic heritability, GWAS, and RVAS for the traits optimized in this work in more heterogeneous or admixed cohorts. Our method also holds promise for investigating high-dimensional representations (e.g. shape, form) of other anatomy, as well as complex traits more generally. Additionally, the optimization can be adapted for various purposes by altering the GA objective function. For example, it could be customized for trait extraction in family-based genomic designs using heritability estimates derived from kinship matrices. Moreover, the approach can be easily extended to twin-based studies, where heritability estimates are typically higher than those observed in our study [60], though not directly comparable due to differences in facial traits (e.g. sparse anthropometric landmarks). Given the flexibility of our method, we hope that future studies will apply it across different datasets to derive traits and compare the genetic value of these traits to advance the field's overall understanding of trait optimization.

Materials and Methods

Datasets

We employed three distinct datasets in this study (Supplementary File 2; Supplementary Table S1; Supplementary File 3; Supplementary Fig. S1):

(1) The ALSPAC father-offspring dataset is part of a UK-based family cohort study [43, 44] and consists of 3D facial surface scans, anthropometric measurements (e.g. weight, height), self-reported demographic information (e.g. age, sex, ethnicity), and family relationships. We only included individuals of self-reported European ancestry with complete covariate data and high-quality images. This resulted in 770 father-offspring pairs. We obtained ethical approval for the current study (B2409) from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. All participants provided written informed consent.

(2) The Technopolis dataset [28] is from a Belgium-based family study and contains 3D facial images for children, anthropometric measurements (e.g. weight, height), and demographic information (e.g. age, sex, self-reported ancestry) from surveys taken by parents. These data were processed using the ALSPAC quality control pipeline, resulting in a sample of 163 parent-offspring trios. We received ethical approval from the Ethics Committee Research UZ/KU Leuven (S56392: ML10285). Every participant provided written informed consent. Parents also provided consent for children under 18 years of age.

(3) The EURO dataset [15] ($n = 8246$) includes unrelated individuals of European ancestry from population-based cohort studies across the US ($n = 4680$) and UK ($n = 3566$). The US sample is a combination of independent datasets from the Pennsylvania State University (PSU, $n = 1990$), Indiana University Indianapolis (IUI, $n = 784$), and the 3D Facial Norms cohort [65] at the University of Pittsburgh (PITT, $n = 1906$). All recruitment sites were granted approval by an institutional ethics committee, and every participant provided written informed consent. The UK sample is again data from the ALSPAC cohort [43, 44] acquired under a different project number (B2261). We received ethical approval from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Along with the 3D images and covariate information, we retrieved genotype data on individual SNPs to complete GWASs of facial shape. Exome data were available for PITT samples, enabling RVAS. Details on genotyping platforms, imputation, and quality control are in [15]. Following imputation and quality control, the merged dataset contained 7,417,619 SNPs for analysis.

3D facial image processing

We utilized the MeshMonk toolbox [66] to obtain spatially dense quasi-landmark configurations for each facial scan. After quality control, as in [15], we superimposed the 7160 quasi-landmarks into a common shape space via generalized Procrustes analysis (GPA). Next, the shapes were symmetrized and corrected for facial size, weight, height, sex, age, age squared, and scanner using partial least-squares regression (PLSR, function "plsregress" from Matlab R2022b). We also adjusted for population structure in the EURO datasets by including the first four genomic ancestry components in the PLSR model. These ancestry components were calculated using PCA on genome-wide SNP data (see Supplementary Fig. S2 and Supplementary File 1). Note this differs from eigen-shapes, which were derived from PCA on facial shape data.

Constructing a facial shape feature space

An individual face can be represented as a point or a vector in a multivariate feature space (Fig. 1a). Shapes vary continuously along the dimensions or axes of this space [67, 68]. While various approaches can be used to construct a feature space, we employed PCA, a common dimension reduction method for 3D facial shapes [25, 69]. Initially, the landmark configurations were composed as a 3D array of dimensions $\mathbf{N}$ (number of shapes) $\times$ $\mathbf{L}$ (7,160 quasi-landmarks) $\times$ 3 (x-, y-, and z-coordinates). We then mean-centered and reshaped the data into a 2D matrix and applied low-rank singular value decomposition (SVD). To retain components whose encoded information was distinguishable from random noise, we combined PCA with parallel analysis [70, 71], yielding 70 PCs (eigen-shapes) that explained 98.08% of facial shape variation, representing each face as a 70-dimensional vector. This approach was first applied to the EURO dataset due to its large sample size, ensuring a feature space that captured diverse shape variations. Both family datasets were then aligned and integrated into the same space.

A genetic algorithm for phenotype optimization

In machine learning, a genetic algorithm (GA) is a popular optimization method inspired by evolutionary processes. It mimics Darwin's principle of the survival of the fittest across a population of individuals (i.e. a set of candidate solutions) that evolve from one generation to the next [30]. Given a $\mathbf{d}$-dimensional PCA space as the feature space, single directions (i.e. linear combinations of PCs), define distinct facial traits or multidimensional shape transformations at the quasi-landmark level, onto which individual faces can be projected and scored. The objective is to scan the $\mathbf{d}$-dimensional PCA space for $\mathbf{p}$ directions maximizing heritability or skewness ($\mathbf{p} \leq \mathbf{d} = 70$). After simulating $\mathbf{p}$ populations representing potential directions or candidate solutions, we trained a GA model on each population until convergence. The algorithm follows a series of steps (pseudocode in Supplementary File 1):

(1) Initialization: An initial population $\mathbf{W}_{\mathbf{m}\times\mathbf{d}}^{(\mathbf{p},0)}$ containing $\mathbf{m}$ random directions in PCA space is created. Every direction is represented by a $\mathbf{d}$-dimensional vector, or a random linear combination of PCs, that encodes unique features of the face. Individuals are scored along these directions using the cosine angle, which is the angle between the direction vector and

their vector. This angle is a univariate trait score indicating the absence or presence of a particular trait. Individual PCs are also retained for comparison with the optimized phenotypes because they define their own directions or facial traits (i.e. eigen-shapes) in PCA space.

(2) Evaluation: Directions are evaluated with an objective function (i.e. fitness in the context of evolutionary biology), defined as heritability or skewness (details in the next section). This "fitness" function determines which directions are selected and reproduced in each generation, thereby guiding the evolution of the GA model.

(3) Selection: The elites, or top $\mathbf{t}$% of the population, are defined according to their "fitness" and retained for the next generation.

(4) Reproduction: Parent directions are mutated to generate new, diverse directions for further exploration. A total of $(1-\mathbf{t}) \times \mathbf{m}$ mutated directions are kept, with directions of higher "fitness" having a greater probability of being selected for the next generation. By the $\mathbf{g}^{\text{th}}$ generation, the population $\mathbf{W}_{\mathbf{m}\times\mathbf{d}}^{(\mathbf{p},\mathbf{g})}$ comprises both elite and mutated directions, while the total population size remains the same.

(5) Constraint: A correlation constraint is applied to current directions to ensure that subsequent directions are sufficiently distinct from those already optimized, thereby avoiding redundancy across multiple GA iterations. More specifically, directions are retained based on their mean correlations with the best directions $\mathbf{B}_{(\mathbf{p}-1)\times\mathbf{d}}$ from the previous generation, $(\mathbf{W}_{\mathbf{m}\times\mathbf{d}}^{(\mathbf{p},\mathbf{g})}, \mathbf{B}_{(\mathbf{p}-1)\times\mathbf{d}}) \leq \mathbf{threshold}$, ensuring the correlations are below a predetermined threshold. While constraints are atypical for GAs, this helped introduce trait diversity.

(6) Termination: The algorithm continues until the maximum number of user-defined generations is reached.

By iteratively applying the above steps, the GA methodically searches the facial feature space and progressively evolves directions or traits toward higher "fitness" values. We developed two distinct phenotyping approaches based on distinct "fitness" values to more effectively identify common genetic variants and rare variants associated with facial shape phenotypes. In the first approach, we hypothesize that heritability-enriched phenotypes reflect the contributions of common variants. In the second approach, we explore commingling-based skewed phenotypes suited to capture the contribution of rare variants. Further implementation details are provided in Supplementary File 1.

Heritability-enriched phenotypes

In the first approach, we trained a GA model to identify directions in feature space that maximize heritability. Heritability refers to the proportion of phenotypic variation explained by genetic variation [72]. Common methods for estimating heritability include (i) family-based designs without genetic data and (ii) molecular genomic designs, which can be either family-based or based on unrelated individuals [64, 73, 74]. Accordingly, we developed two variants of GA models to obtain heritability-enriched phenotypes.

With family image data available, the objective function was defined as family-based trait heritability. This heritability quantifies the degree to which genetic factors influence variation in a specific trait among individuals within a family, and it can be estimated via the regression of offspring on parents [28, 75]. Specifically, we linearly regressed (function "regstats" from Matlab 2022b) the trait scores of children on the corresponding trait scores of their parents and used the R-squared (R^2), or heritability estimate, to guide the evolution of the model (referred to as GA-Family). For one-parent one-offspring designs, the regression coefficients can be multiplied by two to estimate the heritability [75].

The SNP-based heritability-enriched phenotypes followed a similar logic, but instead of leveraging family phenotypic data to estimate trait heritability, we used SNP genotypes from unrelated individuals to estimate SNP heritability. The method employed here, GREML [31, 32], is based on the principle that genetically more related individuals exhibit greater phenotypic similarity within a population of unrelated people. This approach involves building a genetic relationship matrix (GRM), which captures the magnitude of relatedness between all pairs of individuals. The extent to which the GRM can predict similar phenotypes reflects the level of heritability. Several software tools are available to implement GREML [45, 76, 77]. We applied the approach described in [45], specifically the SNPlib toolbox, which is an open-source, in-house Matlab library that was easily integrated with the Matlab-based GA model training (referred to as the GA-GREML). Note that in practice such approaches also allow the inclusion of related individuals, but this was not the case in this work.

Commingling-based skewed phenotypes

In the second approach, we proposed a commingling-based strategy to account for a complementary portion of heritability attributed to rare variant effects. A commingling analysis provides preliminary evidence for the presence of one or a few genomic loci with a large effect by evaluating whether the observed distribution of a quantitative trait is better characterized by a single normal distribution or a combination of normal distributions [36, 37]. Moreover, the distribution skewness is an indicator of commingled distributions [36] that is fast and easy to compute. This makes it an interesting objective function in GA optimization. At each iteration, we evaluated the Pearson median skewness [78, 79], defined as $\frac{3\times(\text{mean}-\text{median})}{\text{standard deviation}}$, and used it to direct the search of the GA-Commingling optimized traits.

Genome-wide association meta-analysis

We performed GWASs on our univariate traits in the UK and US cohorts separately using linear regression (function "regstats" from Matlab 2022b) and additive genotype coding for SNPs (0, 1, 2). Prior to GWAS, we corrected both the independent (SNP) and dependent (facial shape) variables for sex, age, age squared, height, weight, facial size, and camera system with PLSR. We also included the first four genomic ancestry components in the PLSR model to adjust for population structure (details in Supplementary File 1). Next, effect size and standard error estimates were obtained from the US and UK cohort regressions, then meta-analyzed with the inverse-variance weighted method [80]. Finally, we computed meta p-values with a two-tailed test.

To evaluate the GWAS discovery rates across all facial trait categories, we summed the number of genomic loci associated with traits of the same type (e.g. a group of traits optimized with the same objective function). This involved gradually increasing the absolute number of traits ($\mathbf{p}$) included in the GWAS, then aggregating the univariate GWASs within each category using Tippet's minimal-p meta-approach [81]. The lowest P-value for each SNP was retained. We sampled a wide range of traits (e.g. from 1 to the total number of PCs, namely 1, 6, 10, 20, 30, 40, 50, and 70) to interrogate the discovery rate under different scenarios.

To account for multiple testing, we estimated the number of independent traits in each group, defined here as the effective number of traits. While in PCA this was equivalent to the number of PCs because the derived univariate features were completely uncorrelated, the effective number of traits across all other approaches was calculated via permutation, according to the protocol of Kanai et al. [82] (see Supplementary File 1). We implemented a group-wide significance threshold ($P < 5e\text{-}8$ divided by the effect number of traits) as opposed to a genome-wide significance threshold [83] ($P < 5e\text{-}8$) after this correction.

We utilized LDSC [46] to estimate SNP heritability and measure potential confounding bias. To identify genomic loci, we adhered to the procedures described in [20] (details in Supplementary File 1). While the PCA-based traits (eigen-shapes) were ordered according to their variance explained, the constrained GA traits were ordered according to the correlation constraint across multiple optimization rounds. The unconstrained GA traits, by contrast, are not presented in any order. Therefore, we generated results from random samples of traits by selecting a random number of traits, identifying loci associated with these traits, and repeating this procedure 10 times.

Exome-wide rare variant association analysis

We conducted RVAS in the PITT sample ($n = 1906$), which contained both facial images and exome data, to investigate rare variant discovery rates. Variants with a minor allele frequency below 0.01 were assigned to their respective genes, and those containing two or more variants were subjected to association scans. Here, we used the rvtests software (https://github.com/zhanxw/rvtests) [48] to implement SKAT-O [47], a method that combines sequence kernel association tests and burden tests. We analyzed 79 605 variants distributed across 15 804 genes, with 12 841 genes containing at least two variants. The exome-wide significance threshold was established at $P < 3.16e\text{-}06$ (i.e. $P < 0.05/15{,}804$). The total absolute number of traits was fixed at 70 for all methods. The group-wide significance thresholds were calculated as $P < 3.16e\text{-}06$ divided by the effective number of traits within a group of traits. Specifically, for PCA with 70 traits, the threshold was $P < 4.52e\text{-}08$; for GA-Commingling with 67 traits, the threshold was $P < 4.72e\text{-}08$; for GA-GREML with 40 traits, the threshold was $P < 7.91e\text{-}08$; and for GA-Family with 39 traits, the threshold was $P < 8.11e\text{-}08$.

Phenotypic variation explained by facial traits

We regressed the original 3D quasi-landmark configurations on our facial traits, either selected or optimized, using PLSR to quantify the amount of phenotypic variation explained by these derived traits. The sum of the variances across every PLS component is the total variance explained.

Key Points

- Accurate phenotype measurement is crucial for successful genotype–phenotype analysis.
- Machine learning advancements offer potential for optimizing phenotyping workflows.
- GA-optimized phenotyping enables the extraction of genetically relevant traits in a supervised manner.
- Extensive evaluations demonstrated that optimized phenotyping enhances the discovery of both common and rare genetic variants, and the results are biologically meaningful.

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, IUI, and Penn State, which include 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.

Author contributions

Meng Yuan (Conceptualization, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing—Original Draft Preparation, Data Curation), Seppe Goovaerts (Conceptualization, Formal Analysis, Investigation, Methodology, Software, Writing—Review & Editing, Data Curation), Myoung Keun Lee (Formal Analysis, Investigation, Writing—Review & Editing, Data Curation), Jay Devine (Writing—Review & Editing, Data Curation), Stephen Richmond (Writing—Review & Editing, Data Curation, Resources), Susan Walsh (Writing—Review & Editing, Data Curation, Resources), Mark D. Shriver (Writing—Review & Editing, Data Curation, Resources), John R. Shaffer (Funding Acquisition, Writing—Review & Editing, Data Curation, Resources), Mary L. Marazita (Writing—Review & Editing, Data Curation, Resources), Hilde Peeters (Writing—Review & Editing, Data Curation, Resources), Seth M. Weinberg (Funding Acquisition, Writing—Review & Editing, Data Curation, Resources), and Peter Claes (Conceptualization, Funding Acquisition, Project Administration, Supervision, Investigation, Methodology, Writing—Review & Editing, Data Curation, Resources)

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

The KU Leuven research team (P.C., M.Y., S.G., J.D.) and analyses were supported by the Research Fund KU Leuven (BOF-C1, C14/20/081) and the Research Foundation-Flanders (FWO G0D1923N, FWO G017225N). This work was funded in part by grants from NIH, US, 2R01DE027023-04 (S.M.W., J.R.S., P.C.). 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.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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 participants making up the Technopolis dataset, Penn State University (PSU) and Indiana University Indianapolis (IUI) 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 IUI IRB (irb@iu.edu) for the Technopolis, PSU and IUI 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/). 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.

Code availability

The analyses in this work were based on functions in MALAB R2022b, MeshMonk v0.0.6, LDSC v.1.0.1. KU Leuven provides the MeshMonk v.0.0.6 spatially dense facial-mapping software, free to use for academic purposes (https://github.com/TheWebMonks/meshmonk). Scripts for replicating and training the GA model are available online (https://doi.org/10.6084/m9.figshare.27175998).

References

1. Liu F, van der Lijn F, Schurmann C. *et al.* A genome-wide association study identifies five loci influencing facial morphology in Europeans. *PLoS Genet* 2012;**8**:e1002932. https://doi.org/10.1371/journal.pgen.1002932
2. Paternoster L, Zhurov A, Toma A. *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**:478–85. https://doi.org/10.1016/j.ajhg.2011.12.021
3. Xiong Z, Dankova G, Howe LJ. *et al.* Novel genetic loci affecting facial shape variation in humans. *Elife* 2019;**8**:e49898. https://doi.org/10.7554/eLife.49898
4. Adhikari K, Fuentes-Guajardo M, Quinto-Sánchez M. *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
5. Bonfante B, Faux P, Navarro N. *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
6. Shaffer JR, Orlova E, Lee MK. *et al.* Genome-wide association study reveals multiple loci influencing normal human facial morphology. *PLoS Genet* 2016;**12**:e1006149. https://doi.org/10.1371/journal.pgen.1006149
7. Cole JB, Manyama M, Kimwaga E. *et al.* Genomewide association study of African children identifies association of SCHIP1 and PDE8A with facial size and shape. *PLoS Genet* 2016;**12**:e1006174. https://doi.org/10.1371/journal.pgen.1006174
8. Cha S, Lim JE, Park AY. *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
9. Qiao L, Yang Y, Fu P. *et al.* Genome-wide variants of Eurasian facial shape differentiation and a prospective model of DNA based face prediction. *J Genet Genomics* 2018;**45**:419–32. https://doi.org/10.1016/j.jgg.2018.07.009
10. Wu W, Zhai G, Xu Z. *et al.* Whole-exome sequencing identified four loci influencing craniofacial morphology in northern Han Chinese. *Hum Genet* 2019;**138**:601–11. https://doi.org/10.1007/s00439-019-02008-6
11. Li Q, Chen J, Faux P. *et al.* Automatic landmarking identifies new loci associated with face morphology and implicates Neanderthal introgression in human nasal shape. *Commun Biol* 2023;**6**:481. https://doi.org/10.1038/s42003-023-04838-7
12. Li J, Glover JD, Zhang H. *et al.* Limb development genes underlie variation in human fingerprint patterns. *Cell* 2022;**185**:95–112.e18. https://doi.org/10.1016/j.cell.2021.12.008
13. Kun E, Javan EM, Smith O. *et al.* The genetic architecture and evolution of the human skeletal form. *Science*1979 2024;**381**:eadf8009. https://doi.org/10.1126/science.adf8009
14. Claes P, Roosenboom J, White JD. *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
15. White JD, Indencleef K, Naqvi S. *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
16. Zhang M, Wu S, Du S. *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
17. Liu C, Lee MK, Naqvi S. *et al.* Genome scans of facial features in east Africans and cross-population comparisons reveal novel associations. *PLoS Genet* 2021;**17**:e1009695. https://doi.org/10.1371/journal.pgen.1009695
18. Liu D, Alhazmi N, Matthews H. *et al.* Impact of low-frequency coding variants on human facial shape. *Sci Rep* 2021;**11**:748. https://doi.org/10.1038/s41598-020-80661-y
19. Crouch DJM, Winney B, Koppen WP. *et al.* Genetics of the human face: identification of large-effect single gene variants. *Proc Natl Acad Sci* 2018;**115**:E676–85. https://doi.org/10.1073/pnas.1708207114
20. Yuan M, Goovaerts S, Vanneste M. *et al.* Mapping genes for human face shape: exploration of univariate phenotyping strategies. *PLoS Comput Biol* 2024;**20**:e1012617. https://doi.org/10.1371/journal.pcbi.1012617
21. Goovaerts S, Hoskens H, Eller RJ. *et al.* Joint multi-ancestry and admixed GWAS reveals the complex genetics behind human cranial vault shape. *Nat Commun* 2023;**14**:7436. https://doi.org/10.1038/s41467-023-43237-8
22. Gunz P, Tilot AK, Wittfeld K. *et al.* Neandertal Introgression sheds light on modern human Endocranial globularity. *Curr Biol* 2019;**29**:120–127.e5. https://doi.org/10.1016/j.cub.2018.10.065

23. Gregory MD, Kippenhan JS, Eisenberg DP. *et al.* Neanderthal-derived genetic variation shapes modern human cranium and brain. *Sci Rep* 2017;**7**:6308. https://doi.org/10.1038/s41598-017-06587-0
24. Naqvi S, Sleyp Y, Hoskens H. *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
25. Sirovich L, Kirby M. Low-dimensional procedure for the characterization of human faces. *J Opt Soc Am A* 1987;**4**:519–24. https://doi.org/10.1364/JOSAA.4.000519
26. Turk MA, Pentland AP. Face recognition using eigenfaces. In: *Proceedings of the 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 586–91. HI, USA: Maui, 1991.
27. de Jong MA, Wollstein A, Ruff C. *et al.* An automatic 3D facial Landmarking algorithm using 2D Gabor wavelets. *IEEE Trans Image Process* 2016;**25**:580–8. https://doi.org/10.1109/TIP.2015.2496183
28. Hoskens H, Liu D, Naqvi S. *et al.* 3D facial phenotyping by biometric sibling matching used in contemporary genomic methodologies. *PLoS Genet* 2021;**17**:e1009528. https://doi.org/10.1371/journal.pgen.1009528
29. Yuan M, Goovaerts S, Hoskens H. *et al.* Data-driven trait heritability-based extraction of human facial phenotypes. *IEEE Int Conf Bioinform Biomed* 2023;**2023**:312–9.
30. Del Ser J, Osaba E, Molina D. *et al.* Bio-inspired computation: where we stand and what's next. *Swarm Evol Comput* 2019;**48**:220–50. https://doi.org/10.1016/j.swevo.2019.04.008
31. Yang J, Benyamin B, McEvoy BP. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 2010;**42**:565–9. https://doi.org/10.1038/ng.608
32. Yang J, Zeng J, Goddard ME. *et al.* Concepts, estimation and interpretation of SNP-based heritability. *Nat Genet* 2017;**49**:1304–10. https://doi.org/10.1038/ng.3941
33. Cole JB, Manyama M, Larson JR. *et al.* Human facial shape and size heritability and genetic correlations. *Genetics* 2017;**205**:967–78. https://doi.org/10.1534/genetics.116.193185
34. Tsagkrasoulis D, Hysi P, Spector T. *et al.* Heritability maps of human face morphology through large-scale automated three-dimensional phenotyping. *Sci Rep* 2017;**7**:45885. https://doi.org/10.1038/srep45885
35. Hoskens H, Li J, Indencleef K. *et al.* Spatially dense 3D facial heritability and modules of co-heritability in a father-offspring design. *Front Genet* 2018;**9**:9. https://doi.org/10.3389/fgene.2018.00554
36. Maclean CJ, Morton NE, Elston RC. *et al.* Skewness in commingled distributions. *Biometrics* 1976;**32**:695–9. https://doi.org/10.2307/2529760
37. Kwon JM, Boehnke M, Burns TL. *et al.* Commingling and segregation analyses: comparison of results from a simulation study of a quantitative trait. *Genet Epidemiol* 1990;**7**:57–68. https://doi.org/10.1002/gepi.1370070113
38. Nordestgaard BG, Chapman MJ, Ray K. *et al.* Lipoprotein(a) as a cardiovascular risk factor: current status. *Eur Heart J* 2010;**31**:2844–53. https://doi.org/10.1093/eurheartj/ehq386
39. de Boer LM, Hutten BA, Zwinderman AH. *et al.* Lipoprotein(a) levels in children with suspected familial hypercholesterolaemia: a cross-sectional study. *Eur Heart J* 2023;**44**:1421–8. https://doi.org/10.1093/eurheartj/ehac660
40. Weinberg SM, Parsons TE, Marazita ML. *et al.* Heritability of face shape in twins: a preliminary study using 3D stereophotogrammetry and geometric morphometrics. *Dent* 2013;**1**:7–11. https://doi.org/10.5195/d3000.2013.14
41. Naini FB, Moss JP. Three-dimensional assessment of the relative contribution of genetics and environment to various facial parameters with the twin method. *Am J Orthod Dentofacial Orthop* 2004;**126**:655–65. https://doi.org/10.1016/j.ajodo.2003.08.034
42. Naqvi S, Hoskens H, Wilke F. *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
43. Fraser A, Macdonald-Wallis C, Tilling K. *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
44. Boyd A, Golding J, Macleod 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
45. Li J, Zarzar TG, White JD. *et al.* Robust genome-wide ancestry inference for heterogeneous datasets: illustrated using the 1,000 genome project with 3D facial images. *Sci Rep* 2020;**10**:11850. https://doi.org/10.1038/s41598-020-68259-w
46. Bulik-Sullivan BK, Loh P-R, Finucane HK. *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
47. Lee S, Emond MJ, Bamshad MJ. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 2012;**91**:224–37. https://doi.org/10.1016/j.ajhg.2012.06.007
48. Zhan X, Hu Y, Li B. *et al.* RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* 2016;**32**:1423–6. https://doi.org/10.1093/bioinformatics/btw079
49. Fahad A, Nouf A, Leena Z. *et al.* LEOPARD syndrome with PTPN11 gene mutation in three family members presenting with different phenotypes. *J Pediatr Genet* 2020;**09**:246–51. https://doi.org/10.1055/s-0039-3400226
50. Legius E, Schrander-Stumpel C, Schollen E. *et al.* PTPN11 mutations in LEOPARD syndrome. *J Med Genet* 2002;**39**:571–4. https://doi.org/10.1136/jmg.39.8.571
51. Kalidas K, Shaw AC, Crosby AH. *et al.* Genetic heterogeneity in LEOPARD syndrome: two families with no mutations in PTPN11. *J Hum Genet* 2005;**50**:21–5. https://doi.org/10.1007/s10038-004-0212-x
52. Ko JM, Kim J-M, Kim G-H. *et al.* PTPN11, SOS1, KRAS, and RAF1 gene analysis, and genotype–phenotype correlation in Korean patients with Noonan syndrome. *J Hum Genet* 2008;**53**:999–1006. https://doi.org/10.1007/s10038-008-0343-6
53. Tartaglia M, Kalidas K, Shaw A. *et al.* PTPN11 mutations in Noonan syndrome: molecular spectrum, genotype-phenotype correlation, and phenotypic heterogeneity. *The American Journal of Human Genetics* 2002;**70**:1555–63. https://doi.org/10.1086/340847
54. Ferreira LV, Souza SCAL, Montenegro LR. *et al.* Analysis of the PTPN11 gene in idiopathic short stature children and Noonan syndrome patients. *Clin Endocrinol (Oxf)* 2008;**69**:426–31. https://doi.org/10.1111/j.1365-2265.2008.03234.x
55. Allanson JE, Bohring A, Dörr H-G. *et al.* The face of Noonan syndrome: does phenotype predict genotype. *Am J Med Genet A* 2010;**152A**:1960–6. https://doi.org/10.1002/ajmg.a.33518
56. Sharma VP, Fenwick AL, Brockop MS. *et al.* Mutations in TCF12, encoding a basic helix-loop-helix partner of TWIST1, are a

frequent cause of coronal craniosynostosis. *Nat Genet* 2013;**45**:304–7. https://doi.org/10.1038/ng.2531
57. Paumard-Hernández B, Berges-Soria J, Barroso E. *et al.* Expanding the mutation spectrum in 182 Spanish probands with craniosynostosis: identification and characterization of novel TCF12 variants. *Eur J Hum Genet* 2015;**23**:907–14. https://doi.org/10.1038/ejhg.2014.205
58. Björklund M. Be careful with your principal components. *Evolution (N Y)* 2019;**73**:2151–8. https://doi.org/10.1111/evo.13835
59. 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
60. Djordjevic J, Zhurov AI, Richmond S. *et al.* Genetic and environmental contributions to facial morphological variation: a 3D population-based twin study. *PloS One* 2016;**11**:e0162250. https://doi.org/10.1371/journal.pone.0162250
61. Vanneste M, Hoskens H, Goovaerts S. *et al.* Syndrome-informed phenotyping identifies a polygenic background for achondroplasia-like facial variation in the general population. *Nat Commun* 2024;**15**:10458. https://doi.org/10.1038/s41467-024-54839-1
62. Oualkacha K, Labbe A, Ciampi A. *et al.* 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
63. 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
64. Barry C-JS, Walker VM, Cheesman R. *et al.* How to estimate heritability: a guide for genetic epidemiologists. *Int J Epidemiol* 2023;**52**:624–32. https://doi.org/10.1093/ije/dyac224
65. Weinberg SM, Raffensperger ZD, Kesterke MJ. *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
66. White JD, Ortega-Castrillón A, Matthews H. *et al.* MeshMonk: open-source large-scale intensive 3D phenotyping. *Sci Rep* 2019;**9**:6085. https://doi.org/10.1038/s41598-019-42533-y
67. Hill H, Claes P, Corcoran M. *et al.* How different is different? Criterion and sensitivity in face-space. *Front Psychol* 2011;**2**:41. https://doi.org/10.3389/fpsyg.2011.00041
68. 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
69. Blanz V, Vetter T. A morphable model for the synthesis of 3D faces. In: *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 187–94, 1999.
70. 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
71. Franklin SB, Gibson DJ, Robertson PA. *et al.* Parallel analysis: a method for determining significant principal components. *J Veg Sci* 1995;**6**:99–106. https://doi.org/10.2307/3236261
72. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era — concepts and misconceptions. *Nat Rev Genet* 2008;**9**:255–66. https://doi.org/10.1038/nrg2322
73. Tenesa A, Haley CS. The heritability of human disease: estimation, uses and abuses. *Nat Rev Genet* 2013;**14**:139–49. https://doi.org/10.1038/nrg3377
74. Zhu H, Zhou X. Statistical methods for SNP heritability estimation and partition: a review. *Comput Struct Biotechnol J* 2020;**18**:1557–68. https://doi.org/10.1016/j.csbj.2020.06.011
75. Monteiro LR, JAF D-F, dos Reis SF. *et al.* Geometric estimates of heritability in biological shape. *Evolution (N Y)* 2002;**56**:563–72.
76. Yang J, Bakshi A, Zhu Z. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* 2015;**47**:1114–20. https://doi.org/10.1038/ng.3390
77. Yang J, Lee SH, Goddard M. *et al.* GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011;**88**:76–82. https://doi.org/10.1016/j.ajhg.2010.11.011
78. Yule GU, Kendall MG. *An Introduction to the Theory of Statistics*. 3rd Edition. Harper Publishing Company, 1950;162–3.
79. Doane DP, Seward LE. Measuring skewness: a forgotten statistic? *Journal of Statistics Education* 2011;**19**. https://doi.org/10.1080/10691898.2011.11889611
80. Fleiss JL. Review papers: the statistical basis of meta-analysis. *Stat Methods Med Res* 1993;**2**:121–45. https://doi.org/10.1177/096228029300200202
81. Tippett LHC. *The Methods of Statistics*. London: Williams and Norgate, 1931.
82. 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
83. Uffelmann E, Huang QQ, Munung NS. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* 2021;**1**:59. https://doi.org/10.1038/s43586-021-00056-9

Briefings in Bioinformatics, 2025, **26(2)**, bbaf090
https://doi.org/10.1093/bib/bbaf090
Problem Solving Protocol