Supporting Information

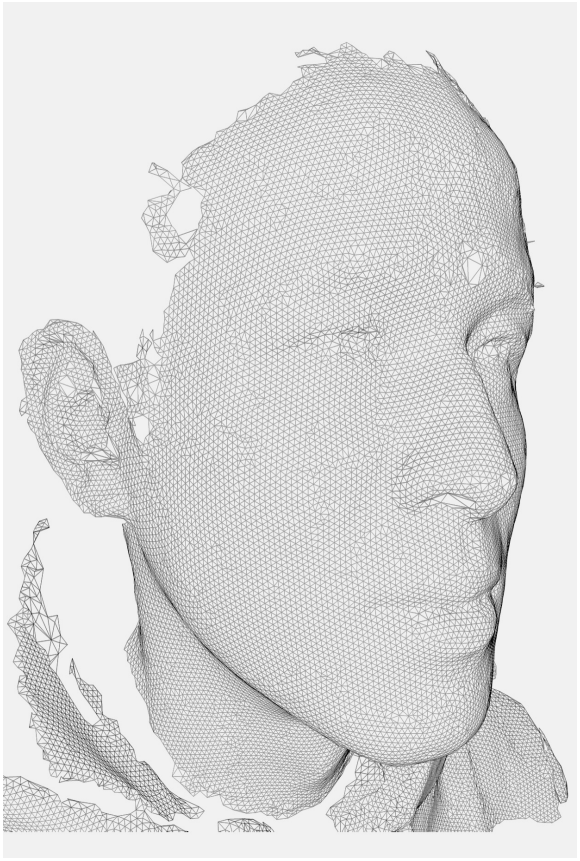
Facial registration

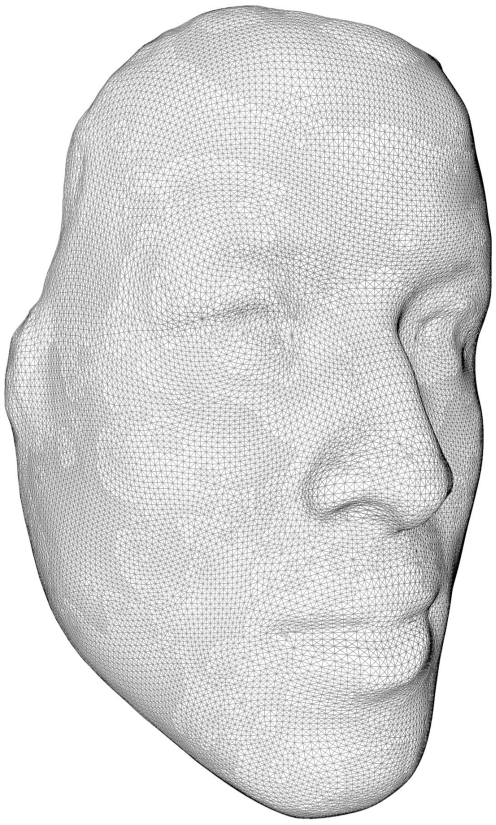
The 3dMD system uses two stereo camera units (each containing two stereo cameras, one above the other), mounted at approximately 45^{o} to the left and right of the participant, roughly a metre away. Stereo triangulation algorithms match surface features recorded by each unit, yielding a single 3D surface. Additional cameras on each unit produce colour photographs that are merged and matched to the 3D surface by another algorithm, producing the 'texture' map, which, for example, portrays skin and eye colour.

The 3D face images generated by the 3dMD 3D camera system are provided in the form of a triangulated mesh (Fig. S1A). Each vertex has associated with it a 3D location and an RGB appearance value. However, the identity of each mesh vertex at this stage is unknown, and the number of vertices varies from image to image. Here we describe the process of fitting the mesh of a generic face model to this mesh; the result is a standardised triangulated mesh in which the identity of each node in the mesh is known (Fig. S1B). This process is known as mesh registration. A complete description of the process can be found in Tena et. al. [1]; more detail can be found in Tena Rodriguez [2]. We summarise the process here.

Fig. S1
Example of triangulated mesh prior to (A) and after (B) registration, and locations of 14 manual landmarks (C) (with left corner of mouth and left corner of nose obscured by the orientation).

A) B)





C)

The generic face model consists of a triangulated mesh describing the surface of a human face. Each vertex of the model mesh has an initial 3D location and a predefined ID that identifies it as belonging to a specific part of the face. The model has an associated set of tools that can warp it to progressively align its surface with that of an input face mesh.

There are four main steps to the registration process:

1. Landmarking, in which 14 predefined salient landmarks are identified in the 3D input image. In order to match faces so that measurements at particular points correspond to one another in a meaningful way, each 3D photograph was manually 'annotated' at 14 landmarks. This involved placing a visual marker with a mouse cursor, for each photograph, on the nose tip, chin tip, labiomental crease, nasion, both corners of the mouth, the top and bottom of the lip, both sides of the nostrils, and each corner of each eye (Fig. S1C). This process was done manually by 3 different people for each photograph, and the mean position and corresponding model vertex ID recorded.

2. Global fitting, in which landmarks identified on the input mesh are used to warp the model so that corresponding landmarks in the model are brought into exact correspondence with them. The Thin Plate Spline algorithm [3] was used for the

warping; this minimises the local deformation of the model as the landmarks are being brought into correspondence.

3. Local matching, in which, for each model vertex, the most similar vertex in the input mesh is identified. Similarity is defined here as the negative of the Euclidean distance between the vertices.

4. Energy minimisation, in which the model mesh surface is warped to align more closely with that of the input mesh, guided by the correspondences found in the previous stage. The energy E_{tot} that is minimised is the weighted sum of external and internal energy terms:

$$E_{tot} = E_{ext} + \varepsilon E_{int} \ ,$$

where ε is a weighting parameter, E_{tot} denotes the distance between the n input and model mesh vertices, $\{\mathbf{x}_i, i = 1 \ldots n\}$ and $\{\tilde{\mathbf{x}}_i, i = 1 \ldots n\}$ respectively:

$$E_{ext} = \sum_{i=1}^{n} \left(\tilde{\mathbf{x}}_i - \mathbf{x}_i \right)^2 \ ,$$

and E_{int} is a smoothness constraint that minimises the deformation of the model:

$$E_{int} = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\left(\mathbf{x}_i - \mathbf{x}_j \right) - \left(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_j \right) \right)^2 \ ,$$

where $\{\overline{x}_i, i = 1 \ldots n\}$ denotes the original positions of the model mesh vertices, and $j = 1 \ldots m$ are the neighbours of vertex i. The weighting parameter ε was set to 0.25, and the conjugate gradient method was used for the energy minimisation.

Steps 3 and 4 of the above steps are combined in an iterative coarse-to-fine process. In the first two iterations, a reduced-resolution model (845 vertices) is used, while in the last two, the full-resolution model (3,300 vertices) is used. Hence the complete algorithm can be summarised as:

- Landmarking
- Global fitting
- Iterate 2 times with reduced resolution model:
    - Local matching

    - Energy minimization
- Iterate 2 times with full resolution model:
    - Local matching
    - Energy minimisation

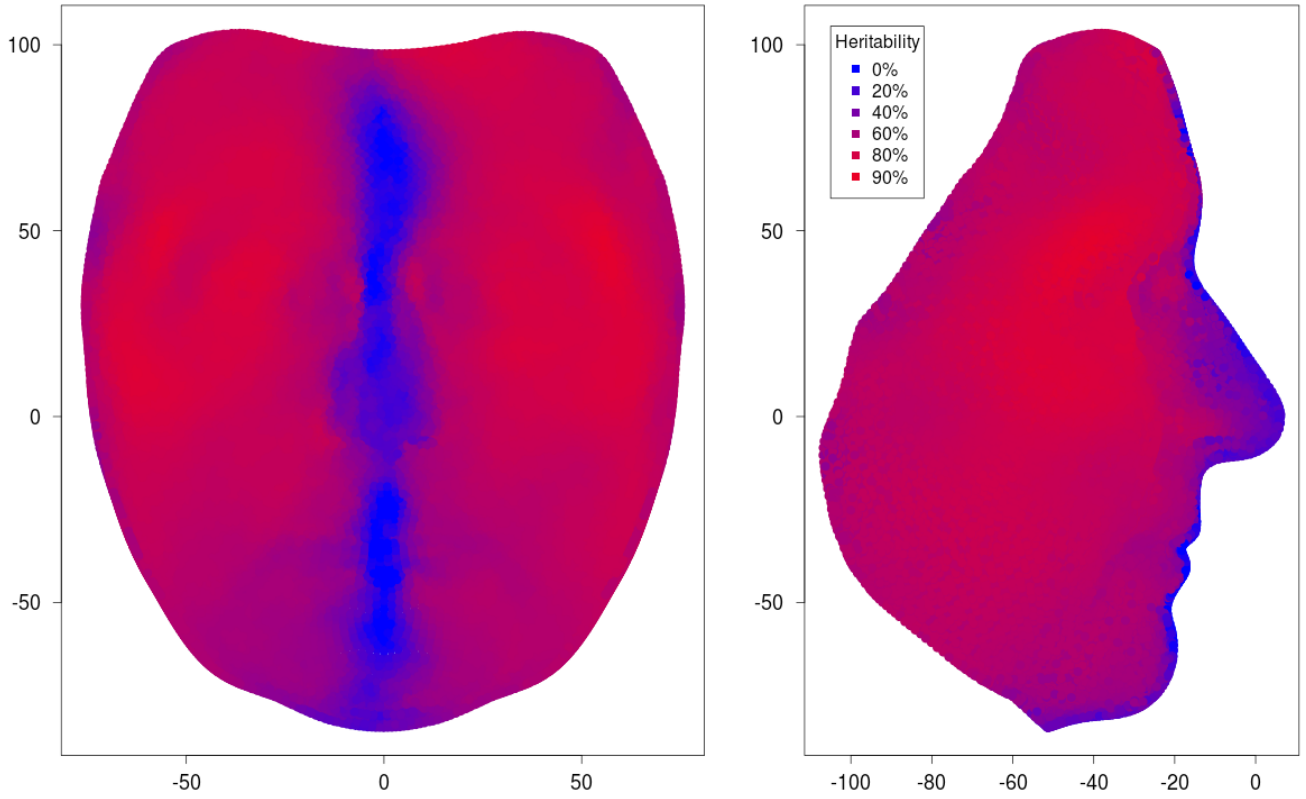
Heritability of registered data

Heritabilities of the registered face data were estimated using the 3dMD data collected from the TwinsUK sample described in the main text. For each facial variable in turn (total 3x29,658=88,974), the variables V_{MZ} and V_{DZ} were obtained by taking half the mean squared difference in phenotypic measurements between members of twin pairs, using the available 357 MZ and 394 DZ pairs. Heritabilities were then calculated as $(V_{DZ} - V_{MZ})/(V_{DZ} - \frac{1}{2}V_{MZ})$ [4]. This estimator provides similar results to Falconer's Formula [5] when the extent of dominance is low, but has more desirable properties when there is substantial dominance, which is a possibility in this case. These were plotted as a heat map on the average face calculated from the PoBI data, assigning colours using the colorRampPalette function in R (Fig. S2). The x variables are notable for having a strip of very low heritability down the center of the face, which is likely to be due to the situation of 6 landmarks in this region.

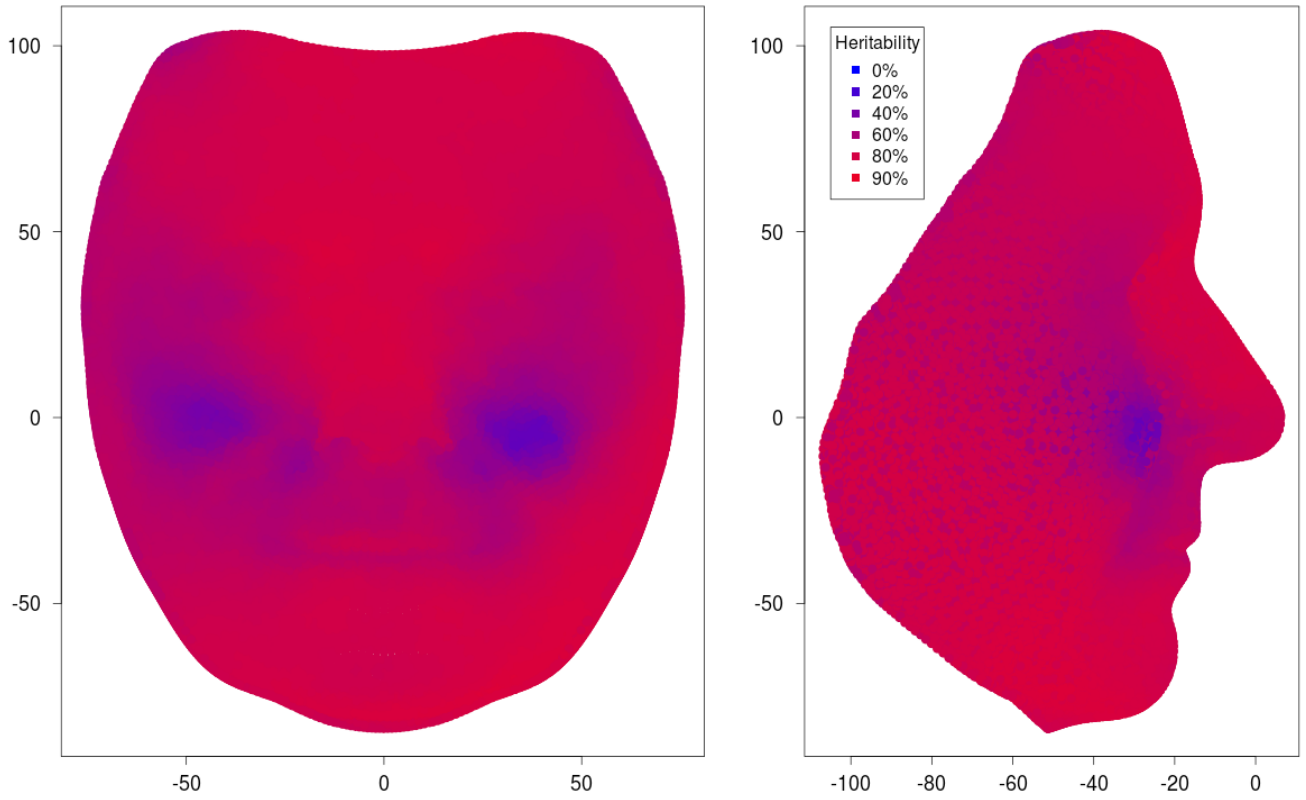
Fig. S2

Heritabilities of original vertex data, after registration but before transformation into additive genetic values, plotted on the average face.

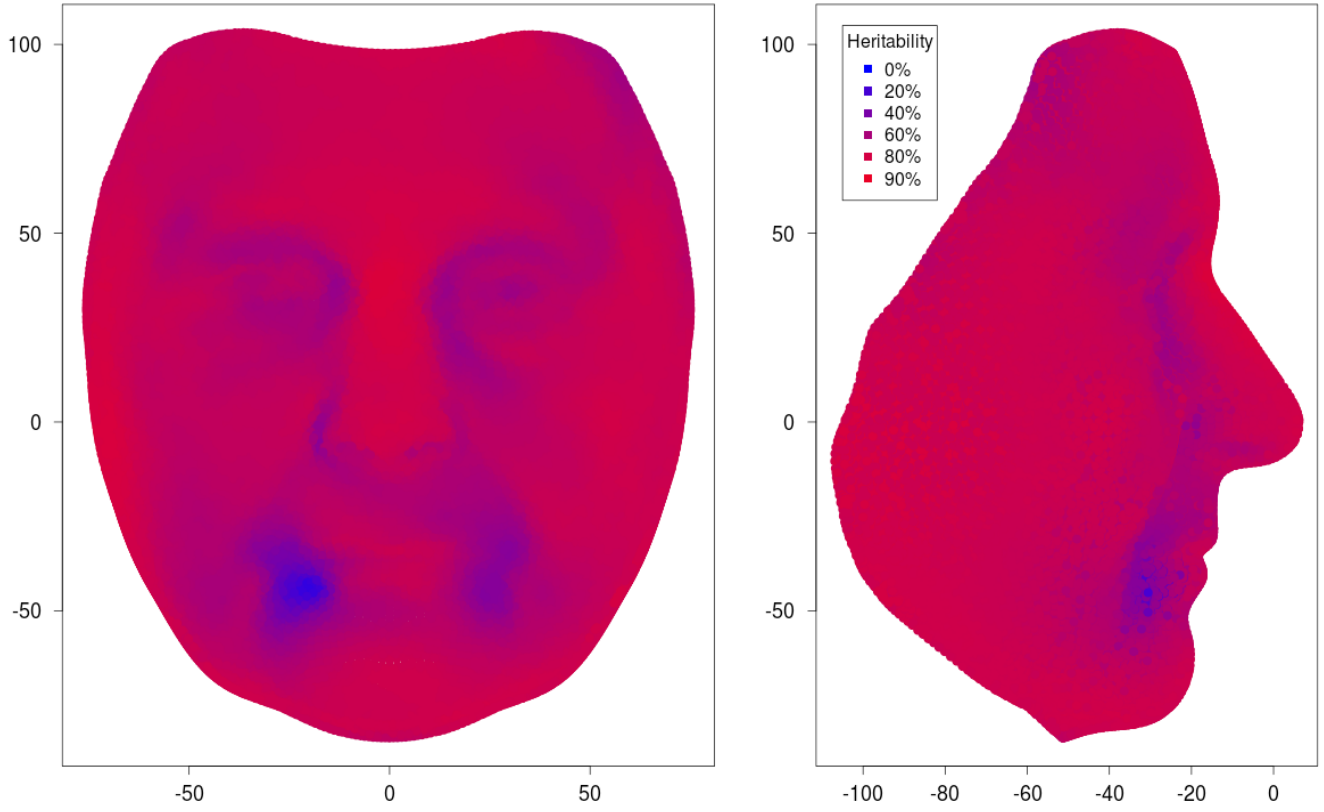
A) Heritabilities of X (left-right) coordinates



B) Heritabilities of Y (height or superior/inferior) coordinates



C) Heritabilities of Z (depth or dorsal/ventral) coordinates



Additive genetic value prediction (AGVP)

We present a method (denoted AGVP for additive genetic value prediction) for increasing the genetic signal present in a multidimensional phenotypic dataset where there are genetic correlations between measurements (i.e., where individual genetic variants affect a group of facial measurements). This is done without any reference to molecular genetic data. Rather, we interrogate the covariances between relatives' (here, twins') facial measurements. The resulting data, enriched for genetic signal, should be more amenable to subsequent genetic analysis, and increase the statistical power of genotype to phenotype comparisons.

The data consist of the original facial surface measurements x_{ij}, which are continuous random variables describing a position in one-dimensional space, where *i* and *j* are indices for individuals (total *n*) and variables (total *m*) respectively. These have already been registered using the procedure described above. In the present application $m \approx 90{,}000$, as there are approximately 30,000 registered surface 'points' each with positions in 3 dimensions.

To increase the heritability of facial surface measurements, we aim to estimate the unobserved values $E_e[X_j \mid g]$, where the expectation is taken over the stochastic environmental effects (E_e), for a given individual and for each measurement *j*, where X_j is the random variable of which x_{ij} is a realisation for the *i*th individual of the *j*th measurement, and g represents a random vector of genotypes. Henceforth we denote $Y_j = E_e[X_j \mid g]$ as a random variable with respect to genetic effects. In the quantitative genetics literature, the departure of Y_j from its population mean would be termed the *genetic value*. Under purely additive genetic effects (no dominance or epistasis), the assumption we rely on below, it is known as the *breeding value* or *additive genetic value* (AGV).

The objective is to predict y_{ij}, the realised AGV for individual *i* at measurement *j*, which cannot be directly observed, using the set of facial surface measurements taken on the same individual: $\{x_{ik} : k \in 1,2,3 \ldots m\}$. This is performed for each *j* in turn. Each X_k is modelled as

$$X_k = Y_k + \varepsilon_k \qquad (1)$$

for all *k* in 1,2,3...*m*, where ε_k represents the effects of the environment. We assume that $E_e[\varepsilon_k \mid g] = 0$, i.e. that there are no gene-environment interactions.

We minimise the expected least squares error,

$$E_g E_e[(Y_j - \sum_{k=1}^{m} \theta_{jk} X_k)^2 \mid g] \,, \qquad (2)$$

with respect to each coefficient θ_{jk} , which represents the predictive influence of variable X_k on variable X_j. This double expectation is taken with respect to the stochastic environmental and genetic effects (E_e and E_g respectively).

Unbiasedness Constraint

We apply the constraint

$$\bar{x}_j = \sum_{k=1}^{m} \theta_{jk} \bar{x}_k \,, \qquad (3)$$

where $\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$. Put another way, the predictor, when run on the 'average individual', must return the values for the 'average individual'. Under the proposed model (Equation 1) $E_g E_e[X_k \mid g] = E_g E_e[Y_k \mid g]$, so, the constraint implies that

$$E_g E_e\left[\sum_{k=1}^{m} \theta_{jk} \bar{X}_k \mid g\right] = E_g E_e\left[\bar{X}_j \mid g\right] = E_g E_e\left[X_j \mid g\right] = E_g E_e\left[Y_j \mid g\right] \qquad (4)$$

where $\bar{X}_k = \frac{1}{n} \sum_{i=1}^{n} X_{ik}$, meaning that the estimator $\sum_{k=1}^{m} \theta_{jk} x_k$ is unbiased, as $E_g E_e\left[\sum_{k=1}^{m} \theta_{jk} X_k\right] = E_g E_e\left[\sum_{k=1}^{m} \theta_{jk} \bar{X}_k\right]$. However, without applying this constraint, $E_g E_e\left[\sum_{k=1}^{m} \theta_{jk} X_k \mid g\right]$ does not necessarily equal $E_g E_e[Y_j \mid g]$, and so the estimator is not necessarily unbiased.

With the constraint applied, the quantity to be minimised becomes

$$E_g E_e[(Y_j - \sum_{k=1}^{m} \theta_{jk} X_k)^2 \mid g] + 2\lambda_j(\bar{x}_j - \sum_{k=1}^{m} \theta_{jk} \bar{x}_k) \,, \qquad (5)$$

where λ_j is a Lagrangian parameter.

Solution using additive genetic covariance

As it is assumed that the expectation of the environmental effects, conditioned on genotype, is zero (i.e. that there are no gene-environment interactions) then, using Equation 1,

$$E_g E_e[Y_j \sum_{k=1}^{m} \theta_{jk} X_k \mid g] = \sum_{k=1}^{m} \theta_{jk} E_g E_e[Y_j (Y_k + \varepsilon_k) \mid g] = \sum_{k=1}^{m} \theta_{jk} E_g[Y_j Y_k \mid g] \ , \qquad (6)$$

so Expression 5 expands to

$$E_g[Y_j^2 \mid g] + \sum_{k=1}^{m} \sum_{l=1}^{m} \theta_{jk} \theta_{jl} E_g E_e[X_k X_l \mid g] - 2 \sum_{k=1}^{m} \theta_{jk} E_g[Y_j Y_k \mid g] + 2\lambda_j (\overline{x}_j - \sum_{k=1}^{m} \theta_{jk} \overline{x}_k) \ . \qquad (7)$$

After cancellation of terms that are equal due to the Lagrangian constraint, Expression 7 becomes

$$\mathrm{var}_g(X_j) + \sum_{k=1}^{m} \sum_{l=1}^{m} \theta_{jk} \theta_{jl} \mathrm{cov}_t(X_k, X_l) - 2 \sum_{k=1}^{m} \theta_{jk} \mathrm{cov}_g(X_j, X_k) + 2\lambda_j (\overline{x}_j - \sum_{k=1}^{m} \theta_{jk} \overline{x}_k) \qquad (8)$$

where $\mathrm{cov}_t(X_k, X_l)$ represents the total covariance between variables *k* and *l,* $\mathrm{var}_g(X_j)$ represents the genetic variance component for variable *j* and $\mathrm{cov}_g(X_j, X_k)$ represents the genetic covariance component between variables *j* and *k*. To proceed, we assume that the full effects of alleles can be well-approximated by their additive effects, and so substitute $\mathrm{cov}_g(X_j, X_k)$ with $\mathrm{cov}_a(X_j, X_k)$, the additive genetic covariance, which can be estimated by taking the covariance between twins with respect to the population mean, devalued by inverse relatedness:

$$\sum_{p=1}^{n_{pairs}} (x_{p1j} - \overline{x}_j)(x_{p2k} - \overline{x}_k) / r_p (n_{pairs} - 1), \qquad (9)$$

where *p* is a twin pair index, n_{pairs} is the number of pairs, x_{p1j} is the measurement on the first of the pair for variable *j*, and x_{p2k} the measurement on the second of the pair for variable *k.* We represent the coefficient of relationship between the twins in the pair as r_p, which is 1 for monozygotic (identical) twins and 0.5 for dizygous (non-identical) twins. Expression 9 assumes that effects attributable to pairs' shared family environments are negligible, though it would be possible to accommodate for this using standard techniques [5]. Shared family environment components are typically found to be very small [6], and this is especially likely to be the case for facial phenotypes, so we take this assumption to be reasonable. It is possible, using

Expression 9, to estimate the additive genetic covariance using pairs of more distantly related individuals, e.g. those from a population sample, so long as sufficient molecular genetic data are available for calculating their coefficients of relationship.

To minimise Expression 8 we differentiate with respect to each θ_{jk}. Setting the resulting partial derivatives zero, we obtain the solution

$$\begin{bmatrix} \theta_{j1} \\ \theta_{j2} \\ \theta_{j3} \\ \dots \\ \theta_{jm} \\ \lambda_j \end{bmatrix} = \begin{bmatrix} \mathrm{var}_t(X_1) & \mathrm{cov}_t(X_1,X_2) & \mathrm{cov}_t(X_1,X_3) & \dots & \mathrm{cov}_t(X_1,X_m) & \bar{x}_1 \\ \mathrm{cov}_t(X_2,X_1) & \mathrm{var}_t(X_2) & \mathrm{cov}_t(X_2,X_3) & \dots & \mathrm{cov}_t(X_2,X_m) & \bar{x}_2 \\ \mathrm{cov}_t(X_3,X_1) & \mathrm{cov}_t(X_3,X_2) & \mathrm{var}_t(X_3) & \dots & \mathrm{cov}_t(X_3,X_m) & \bar{x}_3 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \mathrm{cov}_t(X_m,X_1) & \mathrm{cov}_t(X_m,X_2) & \mathrm{cov}_t(X_m,X_3) & \dots & \mathrm{var}_t(X_m) & \bar{x}_m \\ \bar{x}_1 & \bar{x}_2 & \bar{x}_3 & \dots & \bar{x}_m & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{cov}_a(X_1,X_j) \\ \mathrm{cov}_a(X_2,X_j) \\ \mathrm{cov}_a(X_3,X_j) \\ \dots \\ \mathrm{cov}_a(X_m,X_j) \\ \bar{x}_j \end{bmatrix}. \quad (10)$$

The predicted AGVs, for each *i* and *j*, are then $\hat{y}_{ij} = \sum_{k=1}^{m} \theta_{jk} x_{ik}$. Equation 10 is analogous to the *Universal Kriging* estimator used to predict quantities for variables of interest, e.g. mineral levels, at particular geographic locations, based on measurements of the same variable taken at other locations nearby [7].

As each variable *j* is treated independently, all of their coefficients can be represented in the single equation $\mathbf{P} = \mathbf{T}^{-1}\mathbf{A}$, where

$$\mathbf{P} = \begin{bmatrix} \theta_{11} & \theta_{21} & \dots & \theta_{m1} \\ \theta_{12} & \theta_{22} & \dots & \theta_{m2} \\ \theta_{13} & \theta_{23} & \dots & \theta_{m3} \\ \dots & \dots & \dots & \dots \\ \theta_{1m} & \theta_{2m} & \dots & \theta_{mm} \\ \lambda_1 & \lambda_2 & \dots & \lambda_m \end{bmatrix}, \quad \mathbf{T} = \begin{bmatrix} \mathrm{var}_t(X_1) & \mathrm{cov}_t(X_1,X_2) & \mathrm{cov}_t(X_1,X_3) & \dots & \mathrm{cov}_t(X_1,X_m) & \bar{x}_1 \\ \mathrm{cov}_t(X_2,X_1) & \mathrm{var}_t(X_2) & \mathrm{cov}_t(X_2,X_3) & \dots & \mathrm{cov}_t(X_2,X_m) & \bar{x}_2 \\ \mathrm{cov}_t(X_3,X_1) & \mathrm{cov}_t(X_3,X_2) & \mathrm{var}_t(X_3) & \dots & \mathrm{cov}_t(X_3,X_m) & \bar{x}_3 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \mathrm{cov}_t(X_m,X_1) & \mathrm{cov}_t(X_m,X_2) & \mathrm{cov}_t(X_m,X_3) & \dots & \mathrm{var}_t(X_m) & \bar{x}_m \\ \bar{x}_1 & \bar{x}_2 & \bar{x}_3 & \dots & \bar{x}_m & 0 \end{bmatrix}, \text{ and}$$

$$\mathbf{A} = \begin{bmatrix} \mathrm{var}_a(X_1) & \mathrm{cov}_a(X_1,X_2) & \dots & \mathrm{cov}_a(X_1,X_m) \\ \mathrm{cov}_a(X_2,X_1) & \mathrm{var}_a(X_2) & \dots & \mathrm{cov}_a(X_2,X_m) \\ \mathrm{cov}_a(X_3,X_1) & \mathrm{cov}_a(X_3,X_2) & \dots & \mathrm{cov}_a(X_3,X_m) \\ \dots & \dots & \dots & \dots \\ \mathrm{cov}_a(X_m,X_1) & \mathrm{cov}_a(X_m,X_2) & \dots & \mathrm{var}_a(X_m) \\ \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_m \end{bmatrix}.$$

Fitting process using facial subsets

Ideally the linear predictor uses all facial variables $\{x_{ik} : k \in 1,2,3 \ldots m\}$ as described above, where *m* is the total number processed and registered from the camera system (approximately 3x30,000=90,000). However, computational limitations prevent this, so a workaround is proposed whereby analysis is restricted to rectangular subregions of the face. These are selected based on visual inspection of an image of the average face (Fig. 1). Vertices that fall within the subregion on the average face are taken forward for AGVP analysis in all individuals, so that the number of variables remains constant across individuals.

Two subregions, henceforth referred to as 'profile' and 'eyes', were defined for AGVP analysis and subsequent genetic association mapping. The constituent vertices of these sub-regions are highlighted according to their positions on the average face in Fig. 1. These two particular regions were chosen, fairly informally, on the basis that they are among the most strongly identifying aspects of the face. The eyes subregion contains 2763 vertices, each in 3 dimensions, giving 3 x 2763 = 8289 variables for analysis, whereas the profile subregion contains 1646 vertices. Only the Y (height) and Z (depth) dimensions were used, giving 2 x 1646 = 3292 variables for analysis, as the X (width) dimension of the profile is both less characteristic of appearance and less heritable (Fig. S2A).

5-fold cross validation for protection against overfitting

The method above is susceptible to overfitting due to the involvement of a large number of variables. To guard against this, a ridge penalty is applied to the parameters $\{\theta_k : k \in 1,2,3 \ldots m\}$ so that $\mathbf{T}^{-1}$ is now

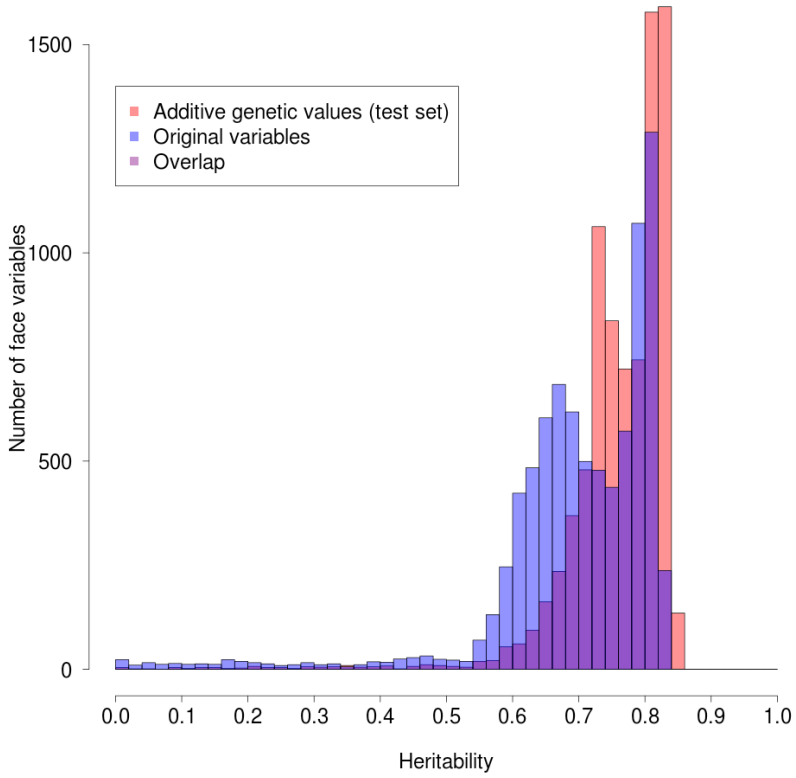
$$\begin{bmatrix} \mathrm{var}_t(X_1)+\Lambda & \mathrm{cov}_t(X_1,X_2) & \mathrm{cov}_t(X_1,X_3) & \ldots & \mathrm{cov}_t(X_1,X_m) & \bar{x}_1 \\ \mathrm{cov}_t(X_2,X_1) & \mathrm{var}_t(X_2)+\Lambda & \mathrm{cov}_t(X_2,X_3) & \ldots & \mathrm{cov}_t(X_2,X_m) & \bar{x}_2 \\ \mathrm{cov}_t(X_3,X_1) & \mathrm{cov}_t(X_3,X_2) & \mathrm{var}_t(X_3)+\Lambda & \ldots & \mathrm{cov}_t(X_3,X_m) & \bar{x}_3 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \mathrm{cov}_t(X_m,X_1) & \mathrm{cov}_t(X_m,X_2) & \mathrm{cov}_t(X_m,X_3) & \ldots & \mathrm{var}_t(X_m)+\Lambda & \bar{x}_m \\ \bar{x}_1 & \bar{x}_2 & \bar{x}_3 & \ldots & \bar{x}_m & 0 \end{bmatrix}^{-1}.$$

Predicted AGVs cannot be compared with any measured values in order to calculate an error rate for the fitted model, as it is not possible to observe AGVs directly. However, AGVs, representing purer genetic effects rather than the original variables, should have higher heritabilities. Therefore, 5-fold cross-validation was used to find a value of Λ that gave high-quality AGV predictions as measured by the mean heritability taken across AGV variables. 1567 TwinsUK individuals were split into 5 validation sets, each consisting of 50 MZ and 50 DZ pairs (200 individuals), and a test

set consisting of 107 MZ and 144 DZ pairs (502 individuals). Leaving out a single validation set, and for a fixed value of Λ, AGVP analysis was performed using the remaining 865 individuals (200 MZ and 200 DZ pairs, plus 65 unrelated individuals), and the fitted model applied to each validation set in turn. Using the excluded validation set alone, heritability was calculated for the predicted AGVs as previously described [4], and the mean heritability taken across all facial variables under analysis, giving a heritability score. This was repeated in turn for each validation set, and the mean and standard error of the heritability score taken over the 5 sets, giving an overall score with associated standard error for the starting value of Λ. This process was repeated for different values of Λ, using identical validation and test sets, thus obtaining a mean and standard error for the heritability score for each Λ. The range of values over which to increment Λ was set to 1,2,3...10, for both subregions. The optimal value of Λ was chosen by subtracting one standard error from each mean heritability score, and picking the Λ giving the highest resulting value. For both the profile and eyes subregions, the optimal value was $\Lambda = 10$. Finally, this optimal value was used to fit an AGVP model using all the data apart from the test set, which was used to assess the heritabilities in the same way; taking heritabilities of predicted AGVs for each phenotypic variable, then taking the mean across variables. The mean heritability was 76.1% for the eyes sub-region and 81.5% for the profile sub-region, compared to 69.8% and 76.6% for the original variables respectively. Perhaps more importantly, the variance in heritabilities was also reduced, as original variables with very low heritabilities had much improved heritabilities in the predicted AGVs (Figs. 2 and S3).

Fig. S3

Comparison of heritabilities in original variables versus additive genetic values (calculated using an independent test set) for the eyes subregion. Fig. 2 in the main text shows the equivalent data for the profile subregion.



Selecting a subset of facial Principle Components

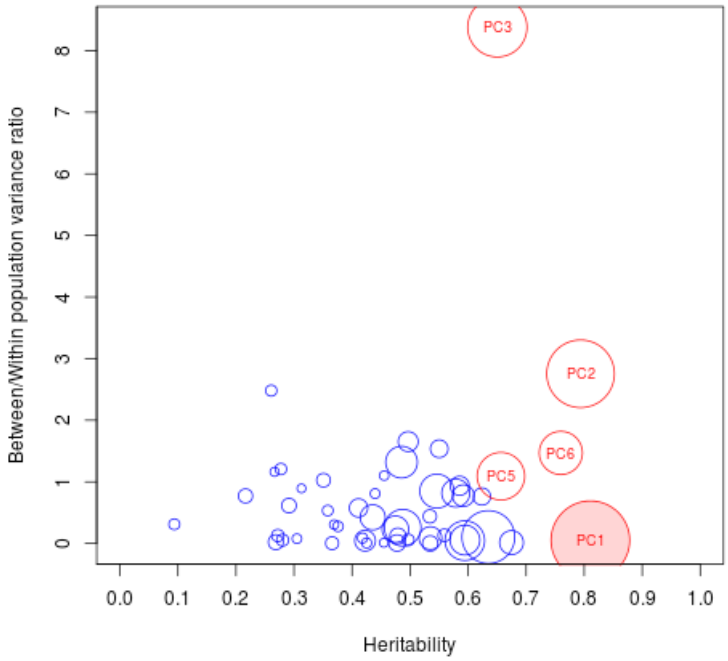
Shape translation and PCA were performed as described in the main text (Materials and Methods). The largest 50 PC axes then were inspected for promising phenotypes using two criteria:

a) Heritability of each PC axis as determined using the PC scores of the 1567 TwinsUK individuals and b) The squared difference between the mean PoBI (European) PC score and the mean East Asian score, taken as a ratio against the within-population variance. Plots of these statistics are shown in Fig. S4. PCs with heritability>0.75 or heritability>0.65 and a population variance ratio of >1 were taken forward for genetic association analysis.

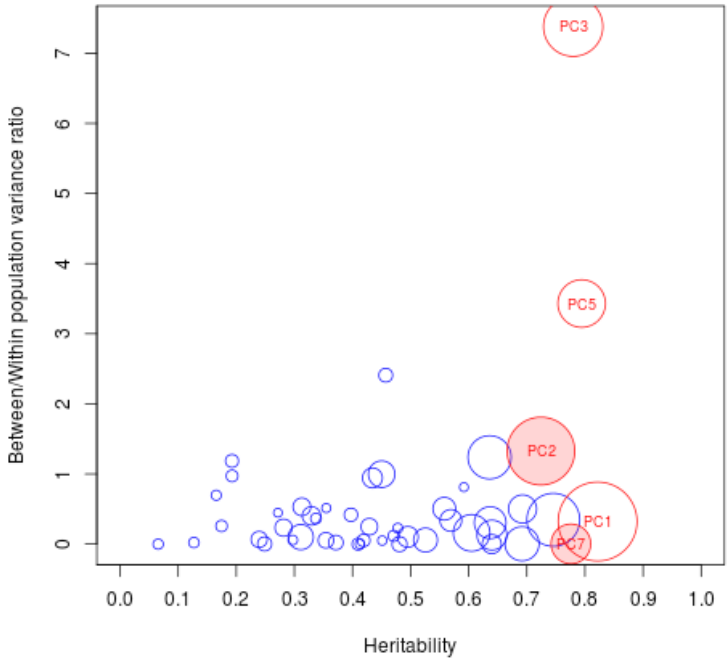
Fig. S4

European/East Asian differences (between/within population variance ratios) and heritabilities for largest 50 PCs performed on facial additive genetic values. Size of circles indicate the rank of the corresponding PC, and red circles denote PCs taken forward for genetic association analysis. Filled in circles denote phenotypes for which SNP associations were replicated. The x-axis has been limited to the interval [0,1]. Some small PCs had heritabilities below zero.

A) Eyes subregion



B) Profile subregion



Genotype quality control

Discovery data

The intersection between the two platforms (Illumina Human 1.2M-Duo and Illumina Infinium OmniExpress-24 BeadChip 750K) was 547,863 SNPs. Prior to QC there were 3,735 genotyped DNA samples, constituting 3,616 unique individuals.

151 individuals were removed due to unusual genotyping on the sex chromosomes that did not correspond with their reported sex. Self-reported females were removed if Y chromosome missingness was less than 50% and excess homozygosity on the X chromosome, as determined using PLINK's F-statistic, was greater than 0.3. Similarly, self-reported males were removed if Y chromosome missingness was greater than 50% and excess homozygosity on the X chromosome was less than 0.8. 14,987 SNPs and 63 individuals with genotyping rates less than 1% were discarded. 52 individuals were removed for having a genomewide F statistic greater than 3 standard deviations from the population mean; all 52 showing an excess rather than deficit in homozygosity, probably due to parental relatedness. One individual with an extreme F statistic of 0.13 was confirmed from our paperwork as having parents who were probably first cousins. Markers were removed due to showing evidence for a departure from Hardy-Weinberg Equilibrium (3276 SNPs), using a cutoff of P less than or equal to 10^{-4}. 24 individuals showed some remaining evidence of a mismatch between self-reported and SNP-determined sex (based on PLINK's X-chromosome homozygosity scores with $F<0.2$ suggesting female and $F>0.8$ male) and were removed due to probable mislabelling or contamination. Of these, 21 were in the WTCCC2 exclusion list (Genetic Analysis of Psoriasis et al., 2010) and 3 were genotyped after the WTCCC2 study. 5 samples' sex discrepancies were best explained by ID mismatches, as they showed full relatedness with other individuals. Relatedness was assessed using a subset of 74,615 SNPs with pairwise $r^2<0.1$. A number of samples otherwise passing QC were removed due to close relatedness or due to being duplicates of the same individual. 64 known duplicated samples and 187 individuals with an identity-by-descent statistic greater than 0.125 (equivalent to first cousins) with at least one other individual were removed. 30 of these samples were removed due to full relatedness with a different individual's sample. All apart from 2 could be attributed to either a) genotyping, contamination or misidentification issues meaning that they were present in the WTCCC2 exclusion list, b) multiple sample collection events visited by the same person, identified by examining paperwork, or c) in one instance, identical twins.

Principal Components Analysis was performed using the same 74,615 SNPs with $r^2<0.1$ and the largest 5 axes inspected visually for outlying samples. The largest Principal Component contained a cluster of 33 outlying individuals (> 3 standard deviations from the mean) that could not be attributed to population structure or batch effects. A large proportion of these individuals were found to be nearby to one another on the same genotyping plate, suggesting contamination of DNA. All 33 samples were removed.

As genotyping was performed in 7 different batches, 2x3 chi-square tests were performed in turn between each batch and all other combined batches, for each SNP, to detect possible batch effects. Using a P-value threshold of 10^{-6}, 5024 SNPs were found significantly associated with genotyping batch, and were consequently removed. After QC, genotypes for 3161 individuals (1532 male, 1629 female) and 524,576 SNPs were retained for association analysis.

Replication data

TwinsUK genotype data were available on two platforms; 1278 samples represented by 2,287,998 array-typed SNPs and 612 whole-genome sequenced samples (19,725,734 autosomal variants). Separate QC procedures were performed for the two platforms before merging. For the array data, 81 samples with a genotype call rate less than 5% were removed. Examining the distribution of F-statistics for individuals with extreme levels of homozygosity, 3 individuals stood out as clear outliers with F-statistics 4 standard deviations above the mean, and were removed. For sequenced samples, 326,065 variants with genotyping call rates of less than 1% were removed. Distributions of heterozygosity F-statistics were examined and 12 outlying individuals greater than 3.25 standard deviations from the mean removed (7 with excess and 5 with deficient heterozygosity). After merging the array and sequencing data and retaining variants common to both sets, there were 1,887,250 variants and 1794 samples, 1275 of whom were unique individuals. Duplicate samples due to individuals being both array-typed and sequenced were eliminated by discarding the array sample. The largest PC axes derived from the genotype data were inspected for any evidence of batch effects between array-typed and sequence data, with none found.

Candidate SNP list

Candidate genes were chosen by reviewing the literature on human facial dysmorphias and facial morphologies in other species, in order to produce a list of regions with prior evidence for involvement in facial features, which can be therefore be subjected to less-stringent multiple comparisons correction. Based on the literature findings, additional searches were performed in the Online Mendelian Inheritance in Man (OMIM) and Ensembl databases. There are 381 genes/regions in the candidate database, representing 168 different dysmorphic facial features. Large candidate regions (approximately 1Mb or larger) were broken down so that only the genic subregions were retained. SNPs were designated candidates if they were located within 75Kb of a candidate region, allowing for LD tagging of functional variants by non-functional SNPs physically close to the relevant gene. This yielded a list of 66,769 candidate SNPs.

Genetic association analysis (discovery)

In order to investigate the hypothesis that there are genetic variants conferring large effects on facial morphology, we focus on individuals with facial features that can, in some sense, be considered 'extreme' relative to the general population. We therefore dichotomised our PCA based facial phenotypes into subsets of upper and lower extremes, which constitute the top and bottom 10% of individuals when ranked according to their scores for each PC. Ranking was performed within 1423 individuals (652 males and 771 females) for whom both genotype and phenotypic data were available. Ranking was performed separately within each sex. For visualisation purposes, average faces (Figs. 4, 5 and 6) within each extreme were produced from the original values using a Matlab script, by plotting the arithmetic mean for each coordinate measurement of each vertex, among all individuals falling into the designated extreme, and overlaid with a surface texture. Only females were used to produce average faces, so as to facilitate comparison with the TwinsUK data, which are almost entirely female. Average faces were produced using all East Asian females and all British (PoBI) females separately.

For each PC selected for further analysis (5 in each subregion), upper extremes were tested against all remaining individuals (including both the lower extremes, individuals in neither extreme plus the 1738 non-phenotyped PoBI individuals remaining after genotype QC) by analysing the 3x2 tables of genotypes (*aa/Aa/AA* where *a* represents the minor allele) versus extreme/control status, for all 512,181 autosomal SNPs. This procedure was repeated for the lower extremes, combining non-phenotyped PoBI individuals with individuals falling outside of the lower extreme as the control sample. All association analyses were performed twice; for female extremes (incorporating all males into the control set - approx. 77 extremes, 3084 controls), and combined male and female extremes (approx. 142 extremes, 3019 controls). Males were not analysed alone as the TwinsUK sample used for replication is almost entirely female, so there would be no appropriate means of re-testing associations. In total, there were 4 association analyses performed for each PC under consideration, due to separate testing of both upper and lower extremes.

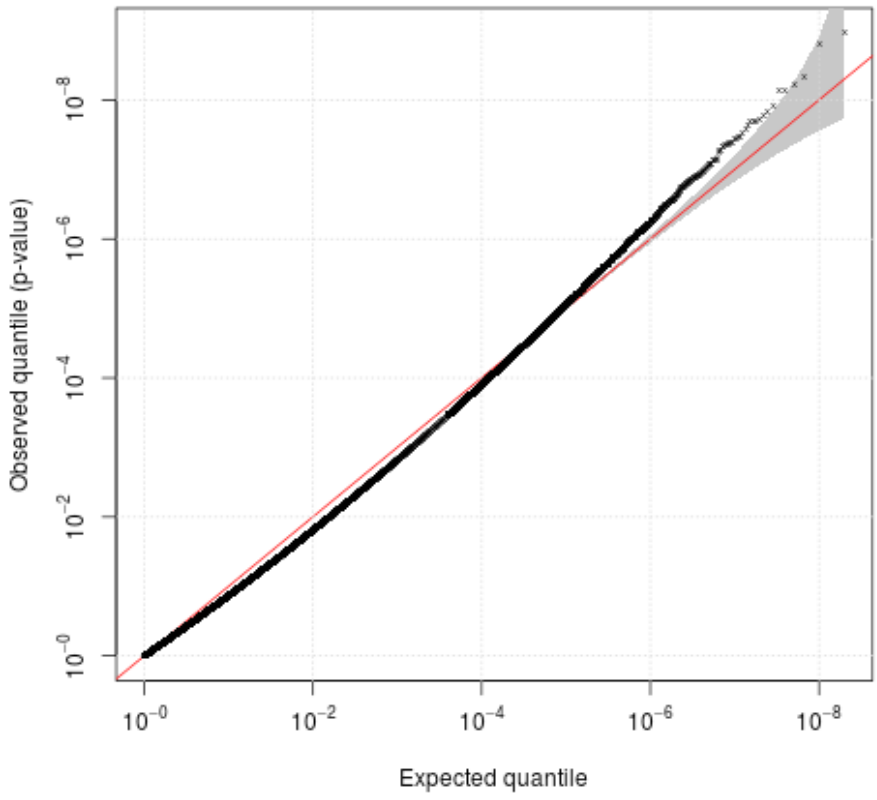
Selection of appropriate statistical test

The relatively small number of extreme individuals motivated a careful selection of the appropriate statistical test for contingency table significance. This is especially pertinent when examining models of inheritance involving the effects of minor allele homozygotes, which can be quite rare even when an allele is common. In order to establish the most appropriate test, null hypothesis simulations of 3x2 tables were performed, and the type-I error rates quantified. Previous work [8] has demonstrated that, when the numbers of observations are small and very low P-values are required, e.g. less than $5x10^{-8}$ for genomewide significance, a variety of traditional frequentist tests are conservative, either due to asymptotic assumptions breaking down, or due to inherent conservativeness in the case of Fisher's Exact test. Simulations performed

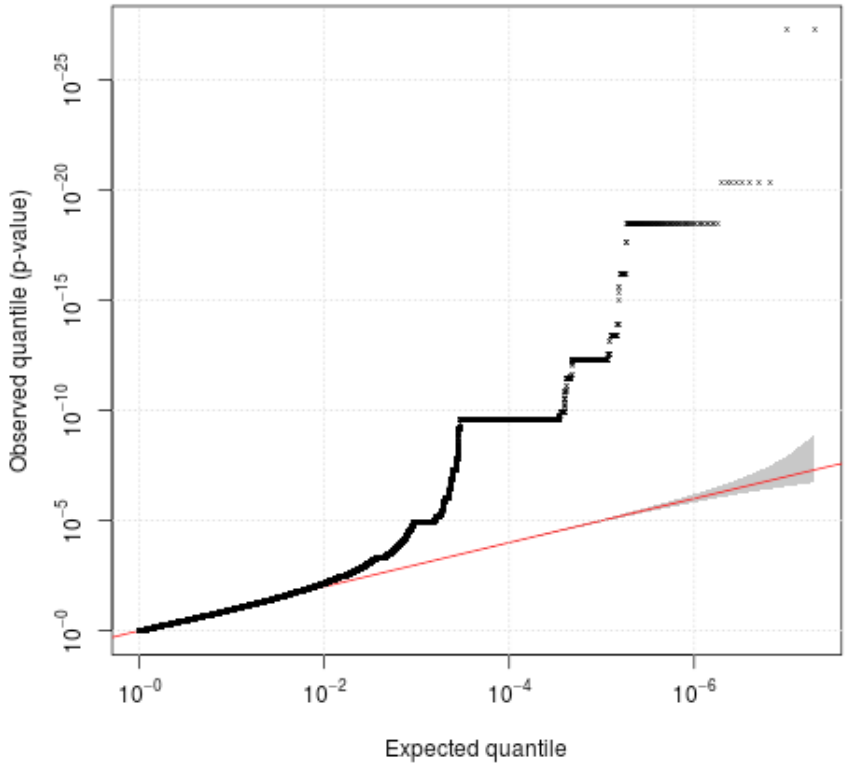
under our study design, which in contrast with Bigdeli et. al. [8], had unequal proportions of extremes versus controls, found that most of these tests are anti-conservative. Fisher's exact test controlled the type-I error well in simulations, but was not chosen as it is often conservative and may reduce power. The best performing approximate test, showing only mild deflation of P-values within the borderline genomewide significance range (10^{-5} to 10^{-6}) (Fig. S5A), was an implementation of a Wald test, in which standardised log odds ratios under recessive and dominant models were tested against an N(0,1) distribution for significant departures from 0. In order to ensure null simulated P-values were close to their expected distribution at high levels of significance (approx. $P<10^{-4}$), it was necessary to apply a standard correction to the estimated OR, adding 0.5 to each cell in the contingency table [9], and to account for this transformation in the expected value of the test statistic when evaluating against the Null distribution (Details in caption of Fig. S5). As in the discovery analysis, the best fitting inheritance model of the two (lowest P-value out of recessive and dominant) was taken as the P-value for the table in question, after multiplying by 2 (setting to 1 if it exceeds 1) to account for two tests being performed. Variants with an observed minor allele frequency equal to or less than 10% were excluded to further ensure accurate control of type-I error, and to improve the appearance of volcano plots (effect sizes versus significance).

Fig. S5

A) Distribution of Wald test P-values from 10^8 simulations of genotypes for 77 cases and 3084 controls, assuming no true effect of genotypes on case/control status. P-values are plotted for both recessive and dominant inheritance models together ($2x10^8$ P-values total). For each replicate, an allele frequency was drawn from a beta distribution with both shape parameters set to 0.8, roughly equivalent to the distribution seen on a SNP chip. Individual genotypes were then drawn from a binomial distribution based on this allele frequency, assuming Hardy-Weinberg equilibrium (HWE). Wald statistics were calculated as $\left(\ln\hat{\theta}_{obs} - \ln\hat{\theta}_{\exp}\right)^2 / \hat{\sigma}^2$ and evaluated against a chi-square distribution with 1 degree of freedom (equivalent to comparing the square root of this value to both tails of an N(0,1) distribution), where θ_{obs} is the observed OR from the appropriate 2x2 table (dominant or recessive inheritance model) after adding 0.5 to each cell, and $\theta_{\exp} = (e_{11}+0.5)(e_{22}+0.5)/(e_{12}+0.5)(e_{21}+0.5)$, where $\{e_{11}, e_{12}, e_{21}, e_{22}\}$ is the set of expected genotype counts. Expected counts are produced using the number of cases and controls together with the minor allele frequency, $\hat{p}$, which is estimated from the full set of genotypes (e.g. for a recessive 2x2 table $e_{11} = \hat{p}^2 \times n_{cases}$). The denominator of the Wald test statistic is $\hat{\sigma}^2 = \frac{1}{o_{11}+0.5} + \frac{1}{o_{12}+0.5} + \frac{1}{o_{21}+0.5} + \frac{1}{o_{22}+0.5}$ where $\{o_{11}, o_{12}, o_{21}, o_{22}\}$ is the observed set of genotype counts. The 95% confidence regions are shown in grey.



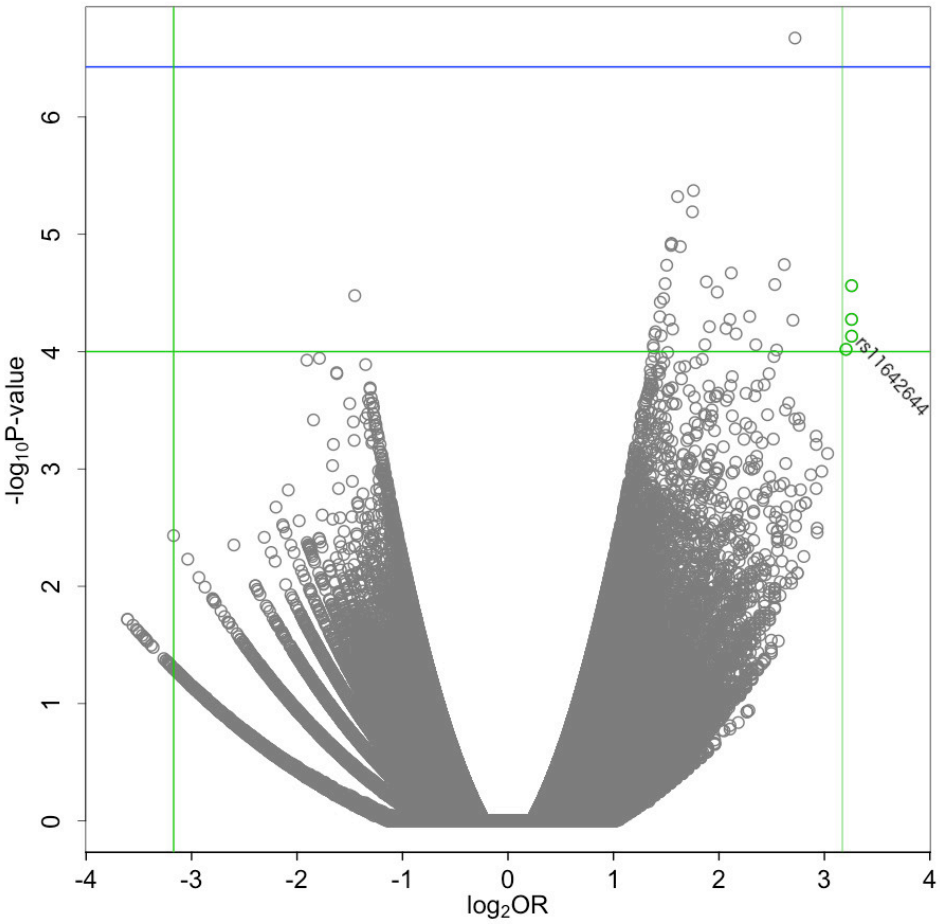
B) Distribution of 10^7 P-values from Pearson's chi-square test, under the same Null hypothesis simulations, showing serious deflation. Similar results are obtained when using Yates' correction.



The Wald-test was applied to the PoBI extreme/control PC-based phenotypes for all SNPs passing QC. Figs. 3, S6 and S7 display results from the discovery analysis relating to SNPs that subsequently replicated in our follow up analysis. Fig. S5B shows the serious departure from the expected linear relationship between the observed and expected P-value quantiles that is obtained for the Pearson's Chi square test as compared to the Wald test.

Fig. S6

A) PC7, profile, females, upper extreme volcano plot (see main text for details). The rs11642644 association is highlighted in green along with 3 other discovery associations, and the green lines denote the P-value (10^{-4}) and OR (9) thresholds for following up associations with large effects. The blue line is the follow-up threshold for candidate SNPs of any OR.



B) PC7, profile. Belonging to the upper 10% is associated with polymorphism at rs11642644 in females. The size of the East Asian histogram has been magnified by 10 times for visualisation purposes. Dotted lines show the upper and lower 10% quantiles

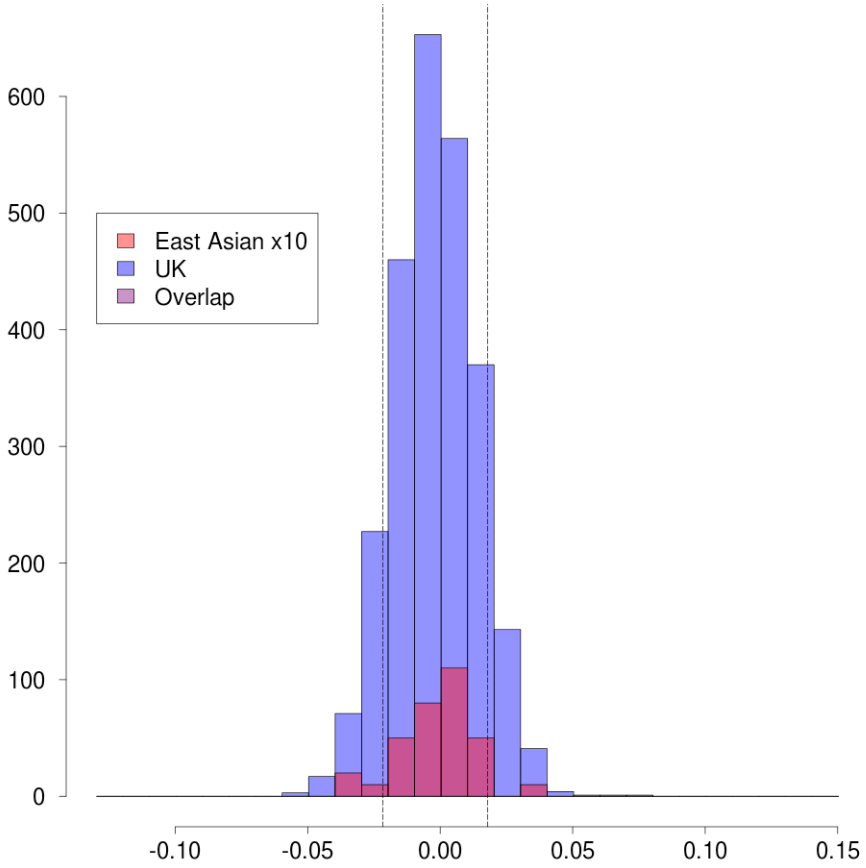
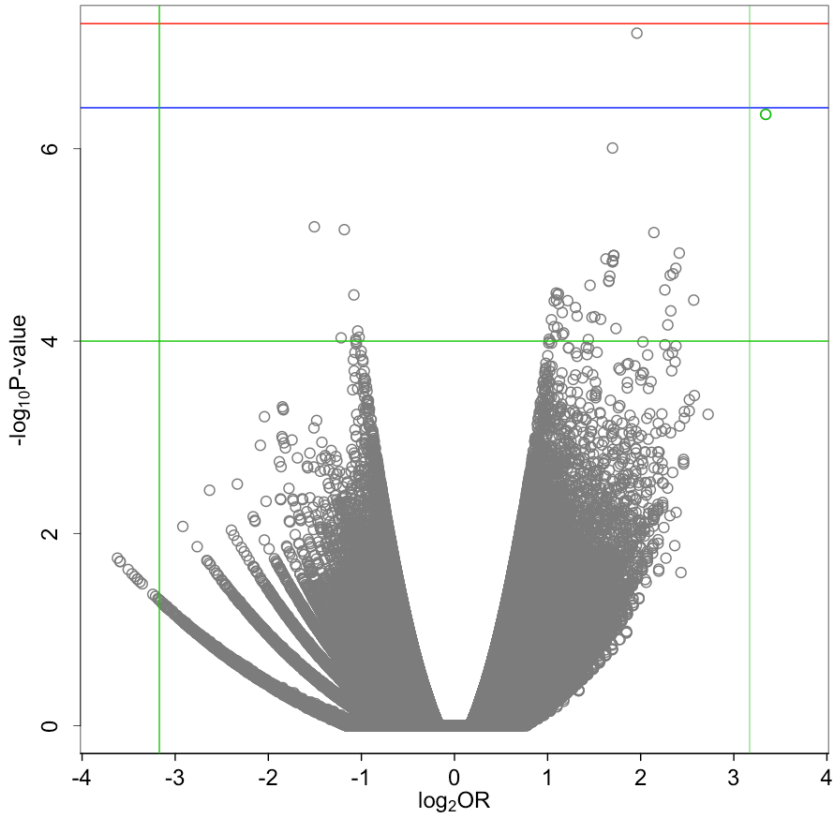
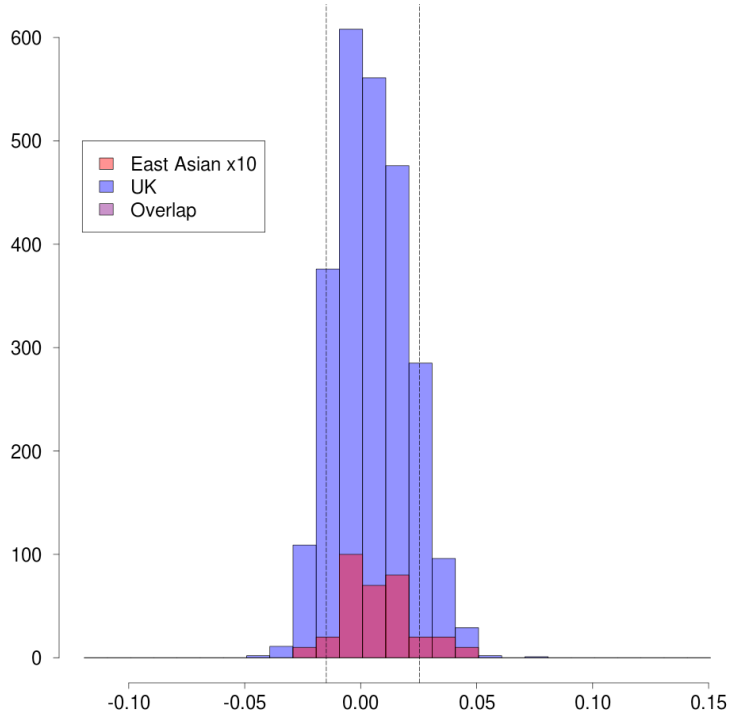


Fig. S7

A) PC1, eyes, combined sexes, upper extreme volcano plot. The rs7560738 association is highlighted in green. The red line indicates the genomewide significance threshold and the green lines denote the P-value (10^{-4}) and OR (9) thresholds for following up associations with large effects.



B) PC1, eyes. Belonging to the upper 10% is associated with polymorphism at rs7560738. The size of the East Asian histogram has been magnified by 10 times for visualisation purposes. Dotted lines show the upper and lower 10% quantiles



Replication analysis

Variants passing any of the three criteria for follow-up (see main text) were analysed in the TwinsUK cohort (see the 'genotype quality control' section for QC procedures). Although this dataset was used previously to produce estimates of the AGVs, which were the foundation of the discovery analysis above, it nevertheless constitutes a valid and unbiased replication set, as the AGVP method makes no reference to DNA data.

1271 of the 1275 TwinsUK individuals with genotype data after QC had available facial phenotypic data, 599 of which were sequenced and 1190 array-typed (with 518 being both sequenced and array-typed). One complication of replication analysis was the presence of many related individuals in the TwinUKs data. Among the 1275, there were 246 MZ twins, 327 DZ twins, and 129 unrelated individuals. The median age was 61 (mean=59.50, sd=9.70) on the date of photographic phenotyping (between 2009-2012 with most photographed in 2010/2011). Just 4 of these individuals were male.

As in the discovery analysis, individuals were ranked according to their PC scores and categorised into upper and lower extremes (top and bottom 10%). Average faces were calculated, within each extreme, as in in the discovery analysis. All profile- and eyes-associated discovery SNPs were tested for association in the replication dataset using Wald tests as in the discovery analysis. A notable difference between the two analyses

is that a high proportion (1271/1275) of the TwinsUK genotyped and QC samples have phenotypic data available, whereas 1738 of 3161 PoBI individuals were not phenotyped and treated as 'unscreened' control samples. The correction for discrete distributions (adding 0.5 to each contingency table cell) was not applied to the replication Wald tests, as permutation was used to ensure accurate control of Type-I error, and was computationally feasible due to the relatively small number of required tests.

In total, 17 profile-associated and 12 eyes-associated SNPs were taken forward for replication analysis. Of the 12 eye-associated discovery SNPs, 1 (rs2039473, eyes PC3 associated in females) was not present on the TwinsUK genotyping platform and not pursued further. After removing a single SNP having $r^2 > 0.1$ with another discovery hit (retaining the SNP with the highest OR in the discovery analysis), 16 remained associated with profile phenotypes, with all 11 remaining eyes-associated SNPs retained as independent associations. For each of the 27 SNPs, association was only tested for the extreme, either upper or lower, PC with which it was associated in the discovery analysis. To increase power, control samples from the appropriate discovery analyses were incorporated into the replication contingency tables before testing. For each SNP, a Wald test was performed under whichever inheritance model was found most significant in the discovery analysis (recessive in all cases), and the total number of tests performed was 27, equal to the number of SNPs tested for replication.

Procedure for dealing with the relatedness between MZ and DZ twins

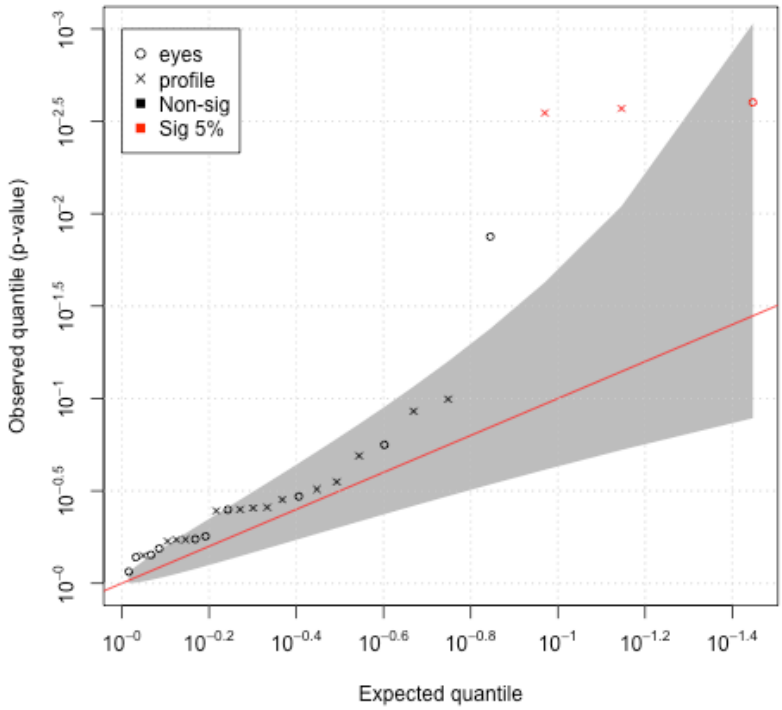
A complication in using the TwinsUK cohort as a replication dataset is the high degree of relatedness between MZ and DZ twins. Removing all related individuals satisfies the requirements for independent samples required by standard statistical analyses, but also loses information and results in an arbitrarily chosen set of unrelated samples. To address these issues, we first treated MZ twins as a single composite individual by averaging over their phenotypic measurements; for each PC phenotype (before assignment of extreme/control status) we took the mean of each MZ pair and assigned this phenotypic value to their shared genotype. Averaging over each member of the MZ pair reduces the influence of environmental variation in the facial phenotype associated. After removing one member of each MZ pair, leaving a sequenced over an array-typed member where possible, 1029 individuals remained, 1025 of which had image data. Of the 1029, 654 were DZ individuals (327 pairs), 129 were unrelateds and 246 MZ twins remained without any of their related pair members.

We dealt with the presence of DZ twins by performing 10,000 random selections of a list of unrelated DZ twins plus, for a random half of those individuals, their twin relative. The most significant FDR-adjusted P-value over the 10,000 selections was taken as the overall observed P-value. Each random selection was again permuted to obtain 10,000 P-values distributed under the Null Hypothesis of no association.

Related DZ twins had their genotypes drawn randomly according to Mendelian inheritance laws from 2 simulated parents (each with randomly drawn genotypes under Hardy-Weinberg Equilibrium using the observed allele frequency). Each related pair's PC phenotypes were simulated as two N(0,1) distributed variables with 50% correlation. This is equivalent to the quite conservative assumption that the phenotype is 100% heritable. Then, individuals in the top theoretical decile of this distribution were designated as extremes before testing. Unrelated individuals were permuted by randomly shuffling their case/control statuses. Empirical P-values were obtained by comparing the overall observed P-value with the permuted P-values. For each SNP separately, the observed P-values were adjusted for 10,000 tests being performed using the Benjamini-Hochberg method [10], and the most significant P-value after adjustment taken as the overall single P-value for that SNP. This was then compared with the 10,000 permutated P-values to obtain an empirical P-value ($P=(M+1)/(N+1)$ where M is the number of permuted P-values lower than the observed P-value and N the total number). These were converted to 1-tail tests by dividing by 2 if the effect size was in the correct direction (based on the expectation from the discovery analysis) or dividing by 2 and subtracting from 1 otherwise. The distribution of these empirical, 1-tailed P-values, as compared to the expected values based on their ranks, is shown in Fig. S8, with those passing an FDR of 5% highlighted, and the deviation from the expected line under no associations is clear.

Fig. S8

Quantile-quantile plot of the 27 replication P-values from permutation analysis of the TwinsUK data. The 95% confidence region is shown in grey, and SNPs passing an FDR of 5% are highlighted in red. Circles and crosses denote eyes and profile associations respectively.



Analysis of combined discovery and replication data

Contingency tables for replication analyses were produced using the rounded mean cell counts over random selections, and these were used to estimate the ORs. The combination of these tables together with the PoBI discovery data was used to perform a Wald test (applying the 0.5 correction factor) for significance and for the estimation of the overall OR.

Assessment of evolutionary conservation

Species comparisons were made between humans and primates using ensembl's online tools (www.ensembl.org), as shown in Figure S9.

Fig. S9

Conserved gene sequences around the discovery SNPs. In each case the base underlined in bold denotes the discovery SNP (showing the minor allele in the British Population), and those on a red background differ from the major allele in the human sequence.

A) rs2045145 in *PCDH15*

The European Human minor allele (facial extreme associated) is A, and the major allele is G.

```
Human         TATATGAATTATATAGGCAGA
Chimpanzee    TATATGAATTGTATAGGCAGA
Gorilla       TATATGAATTGTAAAGGCAGA
Orangutan     TATATGAATTGTATAGGCGGA
Vervet-AGM    TATATGAATTGTATAGGCAAA
Macaque       TATATGAATTGTATAGGCAAA
Olive baboon  TATATGAATTGTATAGGCAAA
Marmoset      TATATGAATTGTATAGGCAAA
```

B) rs11642644 in *MBTPS1*

The European Human minor allele (facial extreme associated) is C, and the major allele is T.

```
Human         AGAAACGCCACGTGGCCGACC
Chimpanzee    AGAAACGCCATGTGGCCGACC
Gorilla       AGAAACACCATGTGGCCGACC
Orangutan     AGAAACGCCATGTGGCCGACC
Vervet-AGM    AGAAACACCACGTGGCCAACC
Macaque       AGAAACACCACGTGGCCGACC
Olive baboon  AGAAACACCACGTGGCCGACC
Marmoset      GTCAACACCATGTGGCCACCC
```

C) rs7560738 in *TMEM163*

The European Human minor allele (facial extreme associated) is A, and the major allele is G.

```
Human         TTTTTCAGGTAG-ACACCTGCT
Chimpanzee    TTTTTCAGGTGG-ACACCTGCT
Gorilla       TTTTTCAGGTGG-ACACCTGCT
Orangutan     TTTTTCAGGTGGAACATCTGCT
Vervet-AGM    TTTTTCAGGTGGAACATCTGCT
Macaque       TTTTTCAGGTGAAACATCTGCT
Olive baboon  TTTTTCAGGTGAAACATCTGCT
Marmoset      TTTTTCAGGTGGAACA-CAGCA
```

Genetic analysis of extremes in TwinsUK sequence data

Variants that were identified as putatively causal were further analysed in sequenced members of the TwinsUK cohort (n=600). Dominant or recessive ORs for putatively functional SNPs that had similarly high OR as the discovery SNP were taken to be further evidence for the causal involvement of the putatively functional variant in the phenotype. Dominant or recessive ORs were calculated using extreme and control statuses assigned in the same way as before, using the reduced set of 600 individuals.

References

1. Tena, J.R., et al., *A validated method for dense non-rigid 3D face registration.* Proc. Video and Signal Based Surveillance, 2006.
2. Tena Rodriguez, J.R., *3D Face Modelling for 2D+3D Face Recognition. .* PhD thesis, University of Surrey, 2007.
3. Bookstein, J., *Principal warps: Thin-plate splines and the decomposition of deformation.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989. **11**(6): p. 567–585.
4. Cavalli-Sforza, L.L. and W.F. Bodmer, *The genetics of human populations*. Series of biology books. 1971, San Francisco: W.H. Freeman. xvi, 965 p.
5. Falconer, D.S. and T.F.C. Mackay, *Introduction to quantitative genetics*. 4th ed ed. 1996, Harlow: Longman. xv, 464 p.
6. Plomin, R., *Commentary: Why are children in the same family so different? Non-shared environment three decades later.* Int J Epidemiol, 2011. **40**(3): p. 582-92.
7. Cressie, N.A.C., *Statistics for spatial data*. Wiley series in probability and mathematical statistics Applied probability and statistics. 1991, New York ; Chichester: Wiley. xx, 900 p.
8. Bigdeli, T.B., B.M. Neale, and M.C. Neale, *Statistical properties of single-marker tests for rare variants.* Twin Res Hum Genet, 2014. **17**(3): p. 143-50.
9. Haldane, J.B., *The estimation and significance of the logarithm of a ratio of frequencies.* Ann Hum Genet, 1956. **20**(4): p. 309-11.
10. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.