MeshMonk: open-source large-scale intensive 3D phenotyping

Julie D. White1\*, Alejandra Ortega-Castrillón2,3, Harold Matthews4,5,6, Arslan A. Zaidi1,7, Omid Ekrami8, Jonatan Snyders9, Yi Fan4,5,10, Tony Penington4,5,6, Stefan Van Dongen8, Mark D. Shriver1, Peter Claes2,3\*

1Department of Anthropology, The Pennsylvania State University, University Park, PA, USA.

2Department of Electrical Engineering, ESAT/PSI, KU Leuven, Leuven, Belgium

3Medical Imaging Research Center, MIRC, UZ Leuven, Leuven, Belgium

4Mudoch Children’s Research Institute, Melbourne, Australia

5Royal Children’s Hospital, Melbourne, Australia

6Department of Pediatrics, University of Melbourne, Melbourne Australia

7Department of Biology, The Pennsylvania State University, University Park, PA, USA.

8Department of Biology, University of Antwerp, Antwerp, Belgium

9WebMonks, Hasselt, Belgium

10Melbourne Dental School, University of Melbourne, Melbourne Australia

**\* Correspondence:**jdw345@psu.edu; peter.claes@kuleuven.be

Keywords: Automated landmarking1, automated phenotyping2, non-rigid registration3, phenomics4, genomics5, morphometrics6, 3D7, facial variation8.

Abstract

Introduction

In the post-genomics era, emphasis has been placed on disentangling ‘genotype-phenotype’ connections so that the biological basis of complex phenotypes can be understood. However, our ability to efficiently and comprehensively characterize phenotypes lags behind our ability to characterize genomes. Anthropometric studies of morphology have traditionally relied on sparse sets of landmarks manually placed on images, from which linear distances and angles are calculated to be used in genetic association studies. This requires the tedious placement of landmarks on many images and is error prone and sensitive to individual differences among observers. Here, we report a toolbox for fast and reproducible high-throughput phenotyping of 3D images. While we demonstrate the utility of this method using 3D facial images, the procedure can also be applied to 3D scans of other complex morphological structures, such as the human brain and skeletal bones.

Methods

Given a facial image (target) with five crude positioning landmarks, a rigid registration is first used to orient an anthropometric mask (reference) to the target scan. Then, using a weighted k-nearest neighbors and a visco-elastic transformation model, the reference is transformed to fit the specific shape of the target. For facial scans, this results in homologous spatially dense (N=7,150) quasi-landmark configurations for all 3D images. As validation, a dataset (N=41) with 19 manually-placed landmarks was superimposed onto the reference in a leave-one-out approach to identify the closest barycentric coordinate on the mask. These coordinates were then projected back onto the training faces and the manual and automatic landmark placements were compared.

Results and Conclusion

We demonstrate that this method is highly accurate, with an average root mean squared error between the manual and automatic placements of 0.6224 mm and no variation in landmark position or centroid size attributable to automatic landmarking method. Though validated using 19 landmarks, for comparison with traditional methods, this method allows for automated dense phenotyping, freeing the researcher from the use of a limited number of landmarks and allowing for more comprehensive investigations of facial shape variation. This expansion opens up an exciting avenue of study in assessing genomic and phenomic data to better understand the genetic contributions to complex morphological traits.

# Introduction

The phenotypic complement to genomics is *phenomics*, which aims to obtain high-throughput and high-dimensional phenotyping in line with our ability to characterize genomes (Houle et al., 2010). The paradigm shift is simple and similar to the one made in the Human Genome Project: instead of ‘phenotyping as usual’ or measuring a limited set of simplified features that seem relevant, why not measure it all? In contrast to genomic technologies, which successfully measure and characterize complete genomes, the scientific development of phenomics lags behind. However, with the advent of new technologies, hardware exists for extensively and intensively collecting quantitative phenotypic data. For example, 3D image surface and/or medical scanners provide the optimal means to capture information of biological morphology and appearance at the phenomic level. Today, the challenge lies in the ability to provide semantic interpretations from large scale image data that capture the phenome in the context of genetic variations (Walter et al., 2010), which is a challenge that we address with the development of the MeshMonk software.

Of interest to anthropologists, geneticists, biologists, and medical clinicians is the ability to accurately and reproducibly characterize anatomical structures, like a femur, skull, or face, such that underlying qualities about the structure can be understood. The study of variation and covariation in anatomy can provide insights into the genetic causes and evolution of the anatomical structure. In addition, comparing the anatomy of an individual patient to a control population can indicate pathology to a medical practitioner. Traditionally, this has been achieved using visual clinical assessment or by taking measurements between manually placed anatomical ‘landmarks’, traditionally defined as precise locations on biological forms that hold some developmental, functional, structural, or evolutionary significance (Richtsmeier et al., 2002) and are unambiguously defined and reliably locatable (Aldridge et al., 2005; Corner et al., 1992; Richtsmeier et al., 1995). Some examples include the endo- and exocanthi (the inner and outer corners of the eyes, respectively) and the pronasale (the tip of the nose).

However, manual landmarking is tedious to perform, difficult to standardize in practice, and prone to intra and inter-operator error (Fagertun et al., 2014; Toma et al., 2009; von Cramon-Taubadel et al., 2007; Weinberg et al., 2004; Wong et al., 2008). Furthermore, sparse landmark configurations can only quantify form at defined landmark indications that can be reliably identified and indicated by a human and lack the resolution to fully characterize shape variation in-between landmarks. An alternative is to automatically indicate quasi-landmarks across the entire surface of the structure. This is achieved by gradually warping a generic template image (anthropometric mask) composed of thousands of points into the shape of each target image through a non-rigid registration algorithm (Andresen and Nielsen, 2001; Claes, 2007; Claes et al., 2012b; Hutton et al., 2003b; Snyders et al., 2014). The coordinates of these warped templates, now in the shape of each target, can then be treated by geometric morphometric analysis. An automatic approach like this is preferable for the analysis of large datasets, avoiding the problems of manual landmarking at different sites by multiple operators. They are also more suitable for applications that require synthesis of a recognizable instance of the actual structure, such as predicting a complete shape from DNA (Claes et al., 2014a), synthetic growth and ageing of a face (Imaizumi et al., 2015; Matthews et al., 2018a), constructing 3D facial composites for forensic applications (Blanz and Vetter, 1999), and characterization of dysmorphology for clinical diagnostics (Baynam et al., 2015; Hammond et al., 2005). Here, we report the MeshMonk toolbox for fast and reproducible high-throughput phenotyping of 3D images, or quasi-landmark indication, which can be applied to 3D facial images as well as 3D scans of other complex morphological structures, such as the human brain and skeletal bones.

Surface registration, utilized in the MeshMonk software, defines a mapping of the vertices from one (template) image to their corresponding locations on another (target) and allows us to quantify and visualize both subtle and acute variation in surface form across a sample by finding the geometrical relationship (one-to-one correspondences) between 3D shapes (Andresen and Nielsen, 2001; Claes, 2007; Claes et al., 2012b; Hutton et al., 2003a; Snyders et al., 2014). The mapping strategy is akin to fitting an elastic net onto a solid facial statue through a geometry-driven mapping of anatomically corresponding features. When the template is warped onto each target, the coordinates of any anatomical landmark, manually annotated on the template, is also defined on each target, thus the complete quasi-landmark indication can also be considered a method for automatic placement of sparse anatomical landmarks (Wei et al., 2011). As a validation of the MeshMonk toolbox, we compare manual and automatic indications of a set of 19 sparse landmarks.

# Materials and Methods

## MeshMonk

MeshMonk is a free, open-source implementation of a modular surface registration framework developed in a partnership between researchers at the Medical Imaging Research Center (MIRC) at KU Leuven, Pennsylvania State University (PSU), and WebMonks ([www.webmonks.vision](http://www.webmonks.vision)), with MIRC and PSU delivering the research and intellectual property behind the methods and algorithms and WebMonks being the implementation partner.

The C++ library takes a multi-scale, iterative closest point (ICP)-based approach (Besl and McKay, 1992). Characteristic to its registration process are (1) a bi-directional, weighted K-Nearest Neighbor point matching algorithm, (2) an outlier classification step and (3) a Visco-Elastic transformation model (Claes, 2007; Snyders et al., 2014). With the library come wrappers to compile the library’s functions so that they can be used in Matlab.

## Explanation of process (Alejandra)

To accurately map the template onto the target surface, the point correspondences and the appropriate transformation need to be found. However, both are closely related and one cannot be resolved without the other. For this, several approaches have been proposed, most of which assign correspondences using the *nearest-neighbor* method (Besl and McKay, 1992; Chui and Rangarajan, 2003) and then the respective transformation is applied. This process is performed in an iterative manner, making of point-matching a joint optimization problem where the distance between the template’s pointset and the respective correspondences on the target is minimized in every loop until coverage is reached.

Snyders et al., 2014, demonstrated that the best generic registration method is a combination of symmetrical weighted k-neighbor correspondences and a viscoelastic transformation model. This pair-wise registration algorithm can be summarized in five steps:

1. **Rough alignment**

An initial alignment of the surfaces is performed by positioning few landmarks on the target surface (e.g. only five markers are required for faces).

1. **Rigid registration**

The Iterative Closest Point (ICP) approach (Besl and McKay, 1992) is then used for a preliminary pose estimation and scaling of the template onto the target surface. It is based on a binary (rigid) nearest-neighbor relationship using an iterative least-square minimization, it relies on an adequate rough alignment.

1. **Non-Rigid registration - a) Finding correspondences**

From this stage, correspondences are found by using a combination of the pull-and-push forces (symmetrical correspondences) and the robust point matching (RPM) algorithm. The symmetrical correspondences are calculated by adding two affinity matrices: 1) from template nodes to target surfaces (push forces – the typical one-to-one correspondences calculation), and 2) from target nodes to template surfaces (pull forces). This ensures that the protrusions present at the target surface will be properly registered. Meanwhile, the RPM relaxes the binary correspondence by using the weighted k-nearest neighbor rule and the softassign. The later, introduces a fuzzy-correspondence (one-to-many) which is in turn controlled by a ‘temperature’ (*T*) parameter to the Gaussian used to calculate the weights (i.e. deterministic annealing). The higher the *T,* the broader the Gaussian and thereforethe bigger the fuzziness. *T* is lowered in every iteration until coverage is reached, making the surface behave as a fluid-like material. As a result, point-matching is improved gradually and continuously, guaranteeing a one-to-one correspondence and an accurate registration of the target surface.

1. **Non-Rigid registration - b) Pruning**

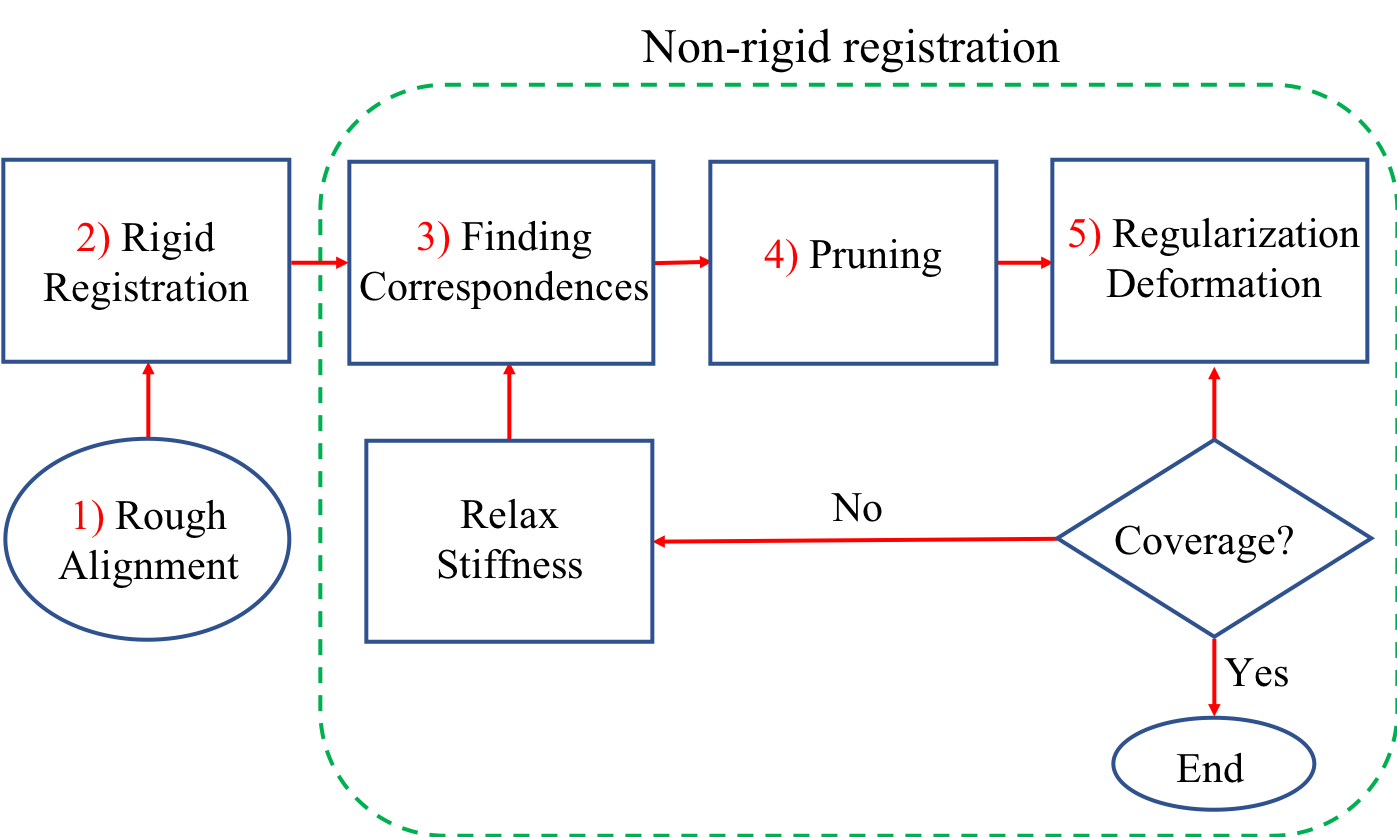
* After correspondences are found, borders and outliers need to be detected and its contribution to the transformation need to be removed. Following the statistically meaningful process described by (Claes et al., 2012a), the inlier and outlier correspondences are flagged by assigning a confidence value to each pair of points. The inlier distances are assumed to form a Gaussian distribution, and any points falling out of +/- (κ) from it is considered to be an outlier. Then, the contributions of the points flagged as outliers are fixed and the confidence values of all the points are updated.

1. **Non-Rigid registration - c) Regularization and deformation**

Before the transformation is applied, a regularization of the energy function is required to ensure that points that lie close to each other move coherently. This regularization also includes the outliers, which do not contribute to the transformation but should be consistently transformed. In this way, a smooth displacement field is established. The smoothness of the transformation is parametrized by a Thin-plate Spline (TPS), which can be thought of as convolving the estimated deformation with a Gaussian. Thus, the energy function is minimized by running the algorithm iteratively and convolving it with a narrower Gaussian every time. Hence, making the global displacement more localized in every iteration. It can be considered as a rough simulation of the elastic model.

Finally, the transformation is applied to the already deformed surface. If coverage has not been reached, the correspondences are updated and the steps from 3) to 5) are repeated. Making the global transformation an incremental approach, hence, the viscous behavior of the algorithm.

A schematic of the registration algorithm is presented in fig. xx.

****

## Parameters and tuning (Alejandra)

## Validation

### Sample and data curation

Over many years, our collaborative group has recruited study participants through several studies at the Pennsylvania State University and sampled in the following locations: State College, PA (IRB 44929 and 4320); New York, NY (IRB 45727); Urbana-Champaign, IL (IRB 13103); Dublin, Ireland; Rome, Italy; Warsaw, Poland; and Porto, Portugal (IRB 32341). Stereo photogrammetry was used to capture 3D facial surfaces of N~6,000 participants using the 3dMD Face 2-pod and 3-pod systems (3dMD, Atlanta, GA). This well-established method generates a dense 3D point cloud representing the surface geometry of the face from multiple 2D images with overlapping fields of view. During photo capture, participants were asked to adopt a neutral facial expression with their mouth closed and to gaze forward, following standard facial image acquisition protocols (Heike et al., 2010). 3D surface images were visually checked to make sure that no major holes or artifacts existed.

### Manual placement of validation landmarks

Of the larger sample, N=48 surface images were chosen at random for validation. This number was then reduced by excluding surface images from participants that reported major facial injury or surgery. This resulted in N=41 surface images for validation, which were diverse with respect to sex (NFemale = 29, NMale = 12), age (range: 18-79, *M* = 32.78), height (range: 149.86-184.00 cm, *M* = 167.13 cm), weight (range: 43.00-103.80 kg, *M* = 67.62 kg), and 3D camera system used (SI Table 1). Most participants reported being of European descent. 3dMDpatient was used to record the 3D coordinates of 19 standard landmarks (7 midline and 12 bilateral) from each unaltered surface (i.e. still containing hair and clothing) in wavefront.obj format (Table 1). Two independent observers placed landmarks three times each, with at least 24 hours in-between landmarking sessions, resulting in six total landmark indications for each facial scan. For each individual, we checked for gross landmark coordinate errors (e.g. mislabeling right and left side landmarks) before analysis. In the subsequent analysis, AML represents the average manual landmarks from observer A, BML represents the average manual landmarks from observer B, while the average of all six manual landmark indications (i.e. the combined average) is denoted as CML.

### Automatic placement of validation landmarks

To obtain automatic indications of the 19 validation landmarks, a leave-one-out approach was used to identify the placement of the validation landmark on the anthropometric mask, then indicate these landmarks on the left-out face (Figure 3). Specifically, for each surface image the manual landmark coordinates were averaged and aligned to the anthropometric mask using barycentric coordinates (Hille, 1982), giving a set of 41 total landmark placements on the anthropometric mask, which were then converted to cartesian coordinates. One by one, each face was left out while averaging the other 40 landmark placements to “train” the automatic landmarks. This average was then placed back onto the left-out (target) face, which resulted in the automatic placement of the validation landmarks using a “training” set that did not include the target face.

The placement of automatic landmarks was performed three times, changing the manual landmark data used as input: once using the average of observer A’s three manual landmark iterations (AAuto), again using the average of observer B’s three manual landmark iterations (BAuto), and a final time using the average of all six manual landmark iterations from both observers (CAuto). This process resulted in three placements of automatic landmarks for comparison.

### Validation

To validate the MeshMonk anthropometric mask registration, we compared the placement of manual and automatic landmarks and analyzed the variance structures captured by the two methods. In direct comparisons of landmark placements, we considered the manually placed landmarks to be the “gold standard” while calculating the root mean squared error (RMSE) between the manual and automatic *x*, *y*, and *z* coordinates. We also calculated Bland-Altman (Altman and Bland, 1983) and Intraclass Correlation Coefficient (Fisher, 1925) statistics to compare the manual and automatic landmark indications. The Bland-Altman method is preferred over correlation or regression as it is less influenced by the variance of the sample and the ICC is preferred because it tests both the degree of correlation and agreement between methods.

We additionally estimated the manual landmarking intra-observer error as the standard deviation between the x, y, and z coordinates of each observer’s manual landmarking indications. The inter- observer error of the manual landmark indications was calculated as the root mean squared error between each observer’s average *x*, *y*, and *z* coordinates (AML vs. BML). As an additional method to understand the variation present in the manual landmark indications only, we performed a multivariate analysis of variance (MANOVA) after aligning the six manual landmarking indications using a generalized Procrustes alignment (GPA; Rohlf and Slice, 1990). Study individual, observer, and iteration were used as predictors and landmark configuration as the response.

To determine if the automatic indication process increased the variation of landmark placements, we compared the inter-observer error calculated using only the manual landmarks (AML vs BML) to the RMSE between one observer’s manual landmarks and the automatic landmarks trained using the other observer’s manual placements (AML vs. BAuto and AAuto vs. BML), as if the automatic indications replaced the manual indications in a calculation of inter-observer error. A paired T-test was used to determine whether the “inter-observer errors” calculated using the automatic indications were significantly different than the error calculated using only the manual indications. RMSE values calculated using both automatic placements (AAuto vs. BAuto) were compared to manual landmarking inter-observer error to illustrate the variance of automatic landmark indications. Levene’s test was performed (Levene, 1960) to determine if the variances of the inter-observer errors calculated using the manual landmarks were equal to the RMSE between the automatic landmarks (the null hypothesis) or unequal (the alternative hypothesis). Levene’s test was chosen because the distribution of RMSE values was non-normal.

We utilized several methods to determine if the variance structures produced by the two methods were similar. MANOVAs were performed separately on the GPA-aligned average manual landmark indications from each observer (AML and BML) as well as on the GPA-aligned automatic landmark indications trained using the average of each observer’s three landmark placements (AAuto and BAuto), with individual and observer as predictors in both tests. By comparing the results of these two tests, we can determine how the explanation of shape variance changes given a different landmarking method. To directly determine if any variance in shape was attributable to landmarking method, we combined the average manual landmark placements of each observer with the automatic placements trained using each of these averages and aligned them using GPA (AML, BML, AAuto, and BAuto). We then tested the shape variation in this combined space as the response in a MANOVA, with individual, observer, method, and individual x observer as predictors.

The covariation structure between the manual and automatic landmarks was coPmpared using a two-block partial least squares (PLS). This test was performed using each observer’s average manual landmark indications compared against the automatic landmark placements trained using each observer’s average (AML vs. AAuto and BML vs. BAuto). We also tested the degree of association between the manual landmark placements averaged across all six manual landmark indications and the automatic landmarks trained using the combined average (CML vs. CAuto).

As a final validation that the automatic landmark indications captured the same information as the manual landmark indications, we compared centroid sizes calculated using the manual and automatic methods. We also performed an analysis of variance (ANOVA) test on the centroid size calculations, with individual, observer, method, and individual x observer as predictors to determine if variation in centroid size could be attributable to variation in landmarking method.

All analyses were performed in R using the Geomorph (Adams and Otárola-Castillo, 2013), BlandAltmanLeh (https://cran.r-project.org/web/packages/BlandAltmanLeh/BlandAltmanLeh.pdf), and ICC (https://cran.r-project.org/web/packages/ICC/ICC.pdf) packages, as well as packages for data manipulation (readxl, reshape2, plyr, car, data.table, dplyr, broom) and graphing (ggplot2, GGally, GGpubr). Centroid sizes were calculated using Geomorph and MANOVAs for shape variation were implemented using the ProcD.lm function from Geomorph (Collyer et al., 2015). The 19 manual and automatic landmark indications as well as the code used to perform this analysis are available in the following GitHub repository: https://github.com/juliedwhite/RemappingValidation/.

# Results

## Intra- and inter-observer error of manual landmarks

The quantitative study of morphology using 3D coordinates requires specific attention to measurement error and has a robust presence in the literature. For each observer, we calculated the intra-observer error of the manual landmarks as the standard deviation between the *x*, *y*, and *z* coordinates of each landmark iteration. Supplemental Table 2 reports intra-observer standard deviations for each axis, averaged only across images. The average standard deviation of observer A across all landmarks was 0.58 mm while the average standard deviation of observer B across all landmarks was 0.44 mm. The average inter-observer error, measured as the root mean squared error (RMSE) between the average *x*, *y*, and *z* coordinates of each observer’s landmark iterations was 0.56 mm. This range of deviation is considered highly precise and is similar to previously reported measures of landmark error (Aldridge et al., 2005; von Cramon-Taubadel et al., 2007).

The analysis of measurement and observer error for the manual landmarks alone, assessed using a MANOVA for shape, with individual, observer, observer x individual, and nested observer x iteration as factors showed that non-individual factors contributed significantly to variation in shape (Supp. Table 3). Individual variation contributed to most of the variation in shape (85%), as expected. Simple measurement error accounted for 3.5% of the total variation in shape. Additional to this, differences in observer accounted for 1.8% of shape variation, and drift in landmark iteration contributed an additional 1.5% of the total variation in shape. In total, non-individual effects contributed to 15% of the total shape variation, with 8.3% of this variation unexplained by our model.

## Direct comparison of manual and automatic landmark placements

As one measure of validation of the automatic landmark indications, we compared the raw coordinate values of the manual landmark indications with the raw coordinate values of the automatic landmark indications. Because of the leave-one-out nature of our approach, we can compare the manual and automatic landmark coordinates directly without fear of training bias. To compare landmark indications, we calculated the root mean squared error between the *x*, *y*, and *z* manual and automatic indications (Table 2, SI Table 4) and calculated the intraclass correlation coefficient between the *x*, *y*, and *z* coordinates produced by the two methods. When comparing the average of all six manual landmarking iterations (CML) and the automatic landmarks from this average (CAuto), the highest difference after averaging RMSE values across all axes, was 0.8472 mm for the right side exocanthion landmark. Overall, the average RMSE across all landmarks between landmarking methods was 0.6224 mm. Bland-Altman comparisons showed that the 95% confidence intervals for the landmark indication between methods are within 1.5 mm of a mean difference of 0 mm (Figure 4). Most individuals fall within these confidence limits, with only a few comparisons from each axis having differences greater than 3 mm. The intraclass correlation coefficients for each axis are around 0.99, representing very high correlation and agreement between manual and automatic landmark indications.

## Comparison of inter-observer errors

By treating the automatic landmark indications as if they were performed by a third observer, we calculated “inter-observer” errors to assess whether the automatic indication process added variation additional to that inherent in manual landmarking. In this assessment, we compared inter-observer errors calculated using only the manual landmarks (AML vs. BML) with error estimates calculated by replacing one of the observer’s manual landmark indications with the automatic indications trained using that observer’s average. This resulted in two extra estimations of inter-observer error (AML vs. BAuto and AAuto vs. BML), calculated as the root mean squared error between *x*, *y*, and *z* coordinates (Figure 5). The mean manual landmarking inter-observer error was 0.56 mm (Table 3) while both manual-automatic comparisons had mean RMSE values of 0.75 mm (SI Table 5). A paired T-test between the manual landmark error values and each of the manual-automatic comparison showed that the RMSE values for both manual-automatic comparisons were significantly different than their manual comparison counterparts at the Chelion right, Crista philtri left, Endocanthion right, both exocanthi, Glabella, and Labiale superius landmarks. The RMSE values calculated after replacing the BML landmarks with BAuto landmarks were significantly different from the AML vs. BML comparison at the endocanthion left, labiale inferius, and pronasale landmarks. The RMSE values for alar curvature left, chelion left, and subalare left landmarks were significantly different when comparing RMSE values for AAuto to BML with those of the manual indications. Overall, ten of the nineteen landmarks showed significant differences when comparing the manual landmark inter-observer error with RMSE values of AML vs. BAuto. Eleven of the nineteen landmarks showed significant differences when comparing manual landmark inter-observer error to the AAuto vs. BML RMSE values.

As an illustration of the low errors between automatic landmark indications trained using different observers, we calculated the RMSE between automatic landmark iterations trained using the average of observer A’s three landmark iterations and the average of observer B’s three landmark iterations (AAuto vs. BAuto; Table 4, Figure 6). The variance of the average RMSE values were significantly different for all landmarks except labiale superius, where we could not reject the null hypothesis that the variances of the two RMSE distributions were equal (*F* = 2.4213, *P =* 0.1236). Figure 6 shows that the variance between automatic landmarking indications is easily identified as being smaller than the manual landmark inter-observer error.

## Analysis of shape variance

A MANOVA on shape, based on the average of each observer’s manual landmark indications and automatic landmark configurations, separately, was performed to determine if the variance explained by individual and observer factors was similar in both methods (Table 5). In both methods, individual variation contributed to most of the variation in shape (R2ML = 94%; R2Auto = 97%). Differences in observer accounted for 1.9% of the variation in shape from manual landmarks and 2.6% of the variation in shape from automatic landmarks. In total, 3.9% of the variation present in manual landmark shape configurations was unexplained by our model while only 0.22% of the variation was unexplained when testing the automatic landmark configurations. A MANOVA on GPA-aligned manual and automatic configurations from each observer, with method, individual, observer, and individual x observer as predictors showed that landmarking method did not account for variation (*F* = 0.3463; *P* = 0.987; Table 6)

## Two-block partial least squares

Several two-block partial least squares tests were used to test the degree of association between the variation structure of the manual and automatic landmarking methods. In all cases, we used the manual landmarks as the response matrix and the automatic landmarks as the independent matrix. The degree of correspondence between the variance-covariance matrices of AML and AAuto (r-PLS = 0.975), BML and BAuto (r-PLS = 0.973), and the combined average (CML) and CAuto (r-PLS = 0.977) were all high and significant (P < 0.001) based on 1000 random permutations.

## Centroid size comparison

We used estimates of centroid size (CS; the cube root of the squared distances from each landmark to the geometric center of each landmark configuration) as a final assessment of the similarity between manual and automatic landmark placements. The correlation between the centroid sizes calculated using the manual and automatic landmarks were all high (RA = 0.958, RB = 0.95, RC = 0.958; Figure 7A). ANOVA by individual, observer, and method shows that individual is the only significant predictor in explaining variance in centroid size (*F* = 130.407, *P* < 2 x 10-16; Table 7). Bland-Altman comparison showed that the 95% confidence intervals for the centroid size estimates between methods are 2 mm relative to an average centroid size of about 165 mm (Figure 7B).

1. **Discussion**

Many authors have supplemented or replaced traditional analysis of inter-landmark measurements with geometric morphometric analysis (Adams et al., 2004; Rohlf and Marcus, 1993), which uses representations of the complete geometry of a configuration of landmarks, rather than simple distances. This can take the form of all pairwise distances between landmarks (Lele and Richtsmeier, 1991), but is more commonly the coordinates of the landmarks themselves. These techniques allow multivariate analysis, comparison, and prediction of the entire form (up to the resolution and density of landmark points) and have become a mainstay of taxonomic analyses (Baab, 2008; Frost et al., 2003; Havarti et al., 2004; Terhune et al., 2007), morphological evolution (Bastir et al., 2006; Klingenberg, 2010; Klingenberg and Marugan-Lobon, 2013; O’Higgins, 2000), the examination of morphological ontogeny and growth (Cole et al., 2018; Kesterke et al., 2016; Martinez-Abadias et al., 2012; Matthews et al., 2018b; Mitteroecker et al., 2004; Smith, 2006), studies of population admixture (Martinez-Abadias et al., 2006; Quinto-Sánchez et al., 2015; Schlager and Rudell, 2015), genotype-phenotype mapping studies (Claes et al., 2014b, 2018; Liu et al., 2012; Paternoster et al., 2012; Shaffer et al., 2016), and studies of dysmorphology (Hammond et al., 2014; Klingenberg et al., 2010; Richtsmeier et al., 2000; Shaner et al., 2000; Starbuck et al., 2011).

Through studies utilizing manually placed sparse landmarks, we have begun to understand the biological basis and evolution of complex phenotypes, both normative and clinical. However, there is still much to be learned. One avenue for improvement is to expand and speed up the production and analysis of data using methods derived from engineering and computer vision, which allow for the description of shapes as “big data” structures instead of sparse sets of landmarks or linear distances, thus matching our ability to describe phenotypes with our ability to describe genomes. To this end, we introduce the MeshMonk registration framework, giving researchers the opportunity to quickly and reliably establish a homologous set of positions across the entire sample. We have validated this framework using a sparse set of landmarks, though the registration framework produces thousands of landmarks (7,160 for the face) to finely characterize the structure.

Because of the relative newness of dense correspondence mapping, few studies have focused on the accuracy and reliability of the resulting registrations. Previous studies using versions of the MeshMonk framework have shown that the error associated with the registration of the anthropometric mask onto facial images is 0.2 mm (Claes et al., 2012c) and parameters of the software have been fine-tuned, as discussed in the introduction (Snyders et al., 2014). To provide some validation regarding the ability of the mask to accurately identify anatomical positions of interest, we used a set of 40 faces with manual landmark indications to “train” positions of interest on the anthropometric mask, then automatically indicate these positions on a face that was not present in the training dataset. In the comparison of manual and automatic landmark indications, the positions of the manual landmarks were considered to be the gold standard, as they have a long history of use and validation in morphological studies (Aldridge et al., 2005; Weinberg et al., 2004). By limiting ourselves to a set number of sparse landmarks, we cannot necessarily speak to the accuracy of structures not involved in our validation (i.e. the cheek bones), but we argue that the results for our comparison speak highly of the fidelity with which the MeshMonk registration framework aligns to underlying anatomical structures.

In the direct comparison of sparse landmarks placed manually and using the MeshMonk software, the average root mean squared error between manual and automatic placements was low (0.6224 mm; Table 2; Figure 4), which is well within the range of acceptable error for manual landmarks (Aldridge et al., 2005; von Cramon-Taubadel et al., 2007; Weinberg et al., 2004) and similar or below errors reported in other comparisons of manual and automatic landmarking methods (De Jong et al., 2016, 2018; Li et al., 2017; Subburaj et al., 2009). When assessing landmarking methods separately, the variance in landmark configuration attributable to individual and observer factors is similar, with considerably less variation left unexplained by an MANOVA model using automatic landmark configuration as the response (Table 5). When assessing manual and automatic landmark configurations in a single MANOVA, the landmarking method is a nonsignificant factor, indicating that variation in scans is not attributable to variation in landmarking method (Table 6). This result was also reproduced when comparing centroid sizes and variance-covariance matrices calculated using manual and automatically placed landmarks (Table 7; Figure 7), speaking highly to the high correspondence between landmark indications placed by human observers and those indicated by the MeshMonk software.

The validation results together suggest that the MeshMonk software is able to reliably reproduce information given by manual landmarking. Though the larger contribution of the MeshMonk software is the ability to quickly and densely characterize entire 3D surfaces, our illustration using a small number of manually placed landmarks as a training set could be useful for studies seeking specifically to study a sparse set of landmarks, perhaps to add more images to a dataset that is already manually landmarked or to add additional landmarks to an analysis. Utilization of the MeshMonk software also gives the opportunity to minimize variation due to different observers. Take, for example, datasets with manual landmarks indicated by two different observers. During the course of analysis, the inter-observer error of these observers would have to be calculated and taken into account when interpreting results. From our own study, the inter-observer error of the manual landmarks placed by two different observers was 0.56 mm (SI Table 2). With the automatic landmarking framework implemented during this study, we can minimize both intra-observer variance for a single scan (by averaging together all indications for that scan by a single observer) and intra-observer variance across scans by placing all indications from the training dataset on the template mesh and averaging the entire training set before carrying along these averages during the registration process to place them in an automatic fashion on the target image. This process finely tunes the position of the landmark, such that even if the training sets were indicated by two different observers, the variation in landmark indication is much smaller than the variation in manual landmark indication, averaging 0.3834 mm in our study (Table 4; Figure 6).

A visual hallmark of the ability of spatially dense surface registration to reliably represent anatomical structures is found in the crispness of “average faces,” constructed by averaging together all registered surfaces in a study sample. Because the MeshMonk registration aligns closely with the underlying anatomical structure, averages across the study samples continue to cleanly resemble the structure and detail is not lost in the averaging process. Consider, for example, the sample average of 100 mandible scans registered using only iterative closest point registration (ICP; Besl and McKay, 1992) versus the sample average of the same 100 mandibles, registered using ICP plus non-rigid transformation, the process implemented in MeshMonk (Figure 8). In the ICP-only average, fine details of the alveolar crest, mental foramen, and coronoid and condylar processes are overly smoothed compared to the level of detail present in the ICP plus non-rigid registration averages. Thus, non-expert readers can easily evaluate the quality of dense-correspondence morphometrics research by looking at the average surface shapes used, with the understanding that high quality mapping software leads to sharp average scans where anatomical positions of interest are clearly defined and identified.

As illustrated in Figure 8, MeshMonk can also be used to characterize structures other than faces.

With respect to the overarching theme of this journal issue, MeshMonk represents a step forward in our ability to describe complex structures, like the human face, for clinical and non-clinical purposes. Researchers are no longer limited to a few homologous points, chosen because they can be reliably indicated over hundreds of hours of work. Instead, fine details of the face can be identified and compared across thousands of images in a few hours, and additional images can be incorporated just as easily, regardless of the camera system with which they were captured, allowing for the incorporation of images from different sources and databases (e.g. Facebase.org). Within the dense-correspondence framework, researchers can develop algorithms to recognize fine structures indicative of a specific dysmorphology, aiding clinicians in diagnoses which are typically reliant upon the experience of the examiner (Hopman et al., 2016; Ibrahim et al., 2016; Klingenberg et al., 2010; Suttie et al., 2013, 2017). Our own recent work is an example of the potential of the MeshMonk software to contribute to our understanding of the underlying genetic contributions to normal-range 3D facial variation (Claes et al., 2018). With the increase in resolution offered by MeshMonk, we were able to utilize two different datasets and identify more loci than had previously been reported in a single GWAS of facial variation, even those with a larger sample size (Liu et al., 2012; Paternoster et al., 2012; Shaffer et al., 2016). Other works in this issue using MeshMonk highlight our ability to finely localize facial variation and genetic effects associated with a common dysmorphology (Indencleef, 2018, this issue) and to push forward our understanding of the heritability of the face in a family-based study (Hoskens, 2018, this issue).

In this study, we present MeshMonk, an open-source resource for intensive 3D phenotyping on a large scale. Compared to a sparse set of manual landmarks, MeshMonk is able to accurately place the same set of landmarks with an average indication error of less than 1 mm. Previous studies have also reported that the repeatability error associated with the template registration invoked in the software is 0.2 mm (Claes et al., 2012c). Through dense-correspondence registration algorithms, like MeshMonk, we can advance our ability to integrate genomic and phenomic data to explore variation in complex morphological traits and answer evolutionary and clinical questions about normal-range variation, growth and development, dysmorphology, and taxonomic classification.

# Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

# Author Contributions

JW performed all analyses and landmarked the 3D scans used for validation with AZ. JW, AOC, and HM wrote the first draft of the manuscript under supervision of PC. HM, YF, and TP provided input and images using mandible scans. PC and JW conceptualized the design of the study. OE, SVD, and MS provided input throughout the analyses and writing process. JS developed the MeshMonk code.

# Funding

# Acknowledgments

We thank the many participants who have volunteered their time over the years, without which we would have never been able to develop this software or perform the various research that this software has contributed to. We are also grateful to the all members of the Shriver and Claes Labs, as well as all the members of WebMonks, for their dedication and contributions to all aspects of this software, research, and manuscript.

# Ethics statement

Institutional review board (IRB) approval was obtained at all locations and all participants signed a written consent form before participation. The Pennsylvania State University IRB board approved the collection of the participants recruited at the following locations: State College, PA (IRB 44929 and 4320); New York, NY (IRB 45727); Urbana-Champaign, IL (IRB 13103); Dublin, Ireland; Rome, Italy; Warsaw, Poland; and Porto, Portugal (IRB 32341).

# References

Adams, D. C., and Otárola-Castillo, E. (2013). Geomorph: An r package for the collection and analysis of geometric morphometric shape data. *Methods Ecol. Evol.* 4, 393–399. doi:10.1111/2041-210X.12035.

Adams, D. C., Rohlf, F. J., and Slice, D. E. (2004). Geometric morphometrics: Ten years of progress following the ‘revolution.’ *Ital. J. Zool.* 71, 5–16. doi:10.1080/11250000409356545.

Aldridge, K., Boyadjiev, S. A., Capone, G. T., DeLeon, V. B., and Richtsmeier, J. T. (2005). Precision and error of three-dimensional phenotypic measures acquired from 3dMD photogrammetric images. *Am. J. Med. Genet.* 138A, 247–253. doi:10.1002/ajmg.a.30959.

Altman, D. G., and Bland, J. M. (1983). Measurement in Medicine: The Analysis of Method Comparison Studies. *Stat.* 32, 307–317. doi:10.2307/2987937.

Andresen, P. R., and Nielsen, M. (2001). Non-rigid registration by geometry-constrained diffusion. *Med. Image Anal.* 5, 81–88. doi:10.1016/S1361-8415(00)00036-0.

Baab, K. L. (2008). The taxonomic implications of cranial shape variation in Homo erectus. *J. Hum. Evol.* 54, 827–847. doi:10.1016/j.jhevol.2007.11.003.

Bastir, M., Rosas, A., and O’Higgins, P. (2006). Craniofacial levels and the morphological maturation of the human skull. *J. Anat.* 209, 637–654. doi:10.1111/j.1469-7580.2006.00644.x.

Baynam, G., Walters, M., Claes, P., Kung, S., LeSouef, P., Dawkins, H., et al. (2015). Phenotyping: Targeting genotype’s rich cousin for diagnosis. *J. Paediatr. Child Health* 51, 381–386. doi:10.1111/jpc.12705.

Besl, P. J., and McKay, N. D. (1992). A method for registration of 3-D shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* 14, 239–256. doi:10.1109/34.121791.

Blanz, V., and Vetter, T. (1999). A morphable model for the synthesis of 3D faces. in *Proceedings of the 26th annual conference on computer graphics and interactive techniques* (New York, NY, USA: ACM Press/Addison-Wesley Publishing Co), 187–194. doi:10.1145/311535.311556.

Chui, H., and Rangarajan, A. (2003). A new point matching algorithm for non-rigid registration. *Comput. Vis. Image Underst.* 89, 114–141. doi:10.1016/S1077-3142(03)00009-2.

Claes, P. (2007). A robust statistical surface registration framework using implicit function representations: application in craniofacial reconstruction. Available at: http://mic.uzleuven.be/download/public/MIC/publications/2967/PHD\_ pclaes.pdf.

Claes, P., Daniels, K., Walters, M., Clement, J. G., Vandermeulen, D., and Suetens, P. (2012a). Dysmorphometrics: the modelling of morphological abnormalities. *Theor. Biol. Med. Model.* 9, 5. doi:10.1186/1742-4682-9-5.

Claes, P., Hill, H., and Shriver, M. (2014a). Towards DNA-based facial composites: preliminary results and validation. *Forensic Sci. Int.* 13, 208–216. doi:10.1016/j.fsigen.2014.08.008.

Claes, P., Liberton, D., Daniels, K., Matthes Rosana, K., Quillen, E., Pearson, L., et al. (2014b). Modeling 3D Facial Shape from DNA. *PLoS Genet.* 10, 1–14. doi:10.1371/journal.pgen.1004224.

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

Claes, P., Walters, M., and Clement, J. (2012b). Improved facial outcome assessment using a 3D anthropometric mask. *Int. J. Oral Maxillofac. Surg.* 41, 324–330. doi:10.1016/j.ijom.2011.10.019.

Claes, P., Walters, M., Shriver, M., Puts, D., Gibson, G., Clement, J. G., et al. (2012c). Sexual dimorphism in multiple aspects of 3D facial symmetry and asymmetry defined by spatially dense geometric morphometrics. *J. Anat.* 221, 97–114. doi:10.1111/j.1469-7580.2012.01528.x.

Cole, J. B., Manyama, M. F., Nikitovic, D., Gonzalez, P. N., Liberton, D. K., Wilson, W. M., et al. (2018). Facial shape manifestations of growth faltering in Tanzanian children. *J. Anat.* 232, 250–262. doi:10.1111/joa.12748.

Collyer, M. L., Sekora, D. J., and Adams, D. C. (2015). A method for analysis of phenotypic change for phenotypes described by high-dimensional data. *Heredity (Edinb).* 115, 357–365. doi:10.1038/hdy.2014.75.

Corner, B. D., Lele, S., and Richtsmeier, J. T. (1992). Measuring Precision of Three-Dimensional Landmark Data. *J. Quantative Anthropol.* 3, 347–359.

De Jong, M. A., Hysi, P., Spector, T., Niessen, W., Koudstaal, M. J., Wolvius, E. B., et al. (2018). Ensemble landmarking of 3D facial surface scans. *Sci. Rep.* 8, 1–11. doi:10.1038/s41598-017-18294-x.

De Jong, M. A., Wollstein, A., Ruff, C., Dunaway, D., Hysi, P., Spector, T., et al. (2016). An automatic 3D facial landmarking algorithm using 2D gabor wavelets. *IEEE Trans. Image Process.* 25, 580–588. doi:10.1109/TIP.2015.2496183.

Fagertun, J., Harder, S., Rosengren, A., Moeller, C., Werge, T., Paulsen, R. R., et al. (2014). 3D facial landmarks: Inter-operator variability of manual annotation. *BMC Med. Imaging* 14, 35. doi:10.1186/1471-2342-14-35.

Fisher, R. A. (1925). *Statistical methods for research workers*. 5th editio. , eds. F. A. E. Crew and D. W. Cutler Edinburgh: Oliver & Boyd doi:10.1056/NEJMc061160.

Frost, S., Marcus, L., Bookstein, F. L., Reddy, D., and Delson, E. (2003). Cranial allometry, phylogeography, and systematics of large-bodied papionins (Primates: Cercopithecinae) inferred from geometric morphometric analysis of landmark data. *Anat. Rec. Part A* 275A, 1048–1072. doi:10.1002/ar.a.10112.

Hammond, P., Hutton, T. J., Allanson, J. E., Buxton, B. F., Campbell, L. E., Clayton-Smith, J., et al. (2005). Discriminating Power of Localized Three-Dimensional Facial Morphology. *Am. J. Hum. Genet.* 77, 999–1010. doi:10.1086/498396.

Hammond, P., McKee, S., Suttie, M., Allanson, J. E., Cobben, J.-M., Maas, S. M., et al. (2014). Opposite effects on facial morphology due to gene dosage sensitivity. *Hum. Genet.* 133, 1117–25. doi:10.1007/s00439-014-1455-z.

Havarti, K., Frost, S. R., and McNulty, K. P. (2004). Neanderthal taxonomy reconsidered: implications of 3D primate models of intra- and interspecific differences. *Proc. Natl. Acad. Sci. USA* 101, 1147–1152. doi:10.1046/j.1469-7580.2000.19710103.x.

Heike, C. L., Upson, K., Stuhaug, E., and Weinberg, S. M. (2010). 3D digital stereophotogrammetry: a practical guide to facial image acquisition. *Head Face Med.* 6, 18. doi:10.1186/1746-160X-6-18.

Hille, E. (1982). *Analytic Function Theory, Volume I*. 2nd editio. Providence, RI: AMS Chelea Publishing Company.

Hopman, S. M. J., Merks, J. H. M., Suttie, M., Hennekam, R. C. M., and Hammond, P. (2016). 3D morphometry aids facial analysis of individuals with a childhood cancer. *Am. J. Med. Genet. Part A* 170, 2905–2915. doi:10.1002/ajmg.a.37850.

Houle, D., Govindaraju, D. R., and Omholt, S. (2010). Phenomics: The next challenge. *Nat. Rev. Genet.* 11, 855–866. doi:10.1038/nrg2897.

Hutton, T. J., Buxton, B. F., and Hammond, P. (2003a). Automated registration of 3D faces using dense surface models. in *British Machine Vision Conference*, eds. R. Harvey and A. Bangham (Norwich: Citeseer), 1–10. doi:10.5244/C.17.45.

Hutton, T. J., Buxton, B. F., Hammond, P., and Potts, H. W. W. (2003b). Estimating Average Growth Trajectories in Shape-Space Using Kernel Smoothing. *IEEE Trans. Med. Imaging* 22, 747–753. doi:10.1109/TMI.2003.814784.

Ibrahim, A., Suttie, M., Bulstrode, N. W., Britto, J. A., Dunaway, D., Hammond, P., et al. (2016). Combined soft and skeletal tissue modelling of normal and dysmorphic midface postnatal development. *J. Craniomaxillofacial Surg.* 44, 1777–1785. doi:10.1016/j.jcms.2016.08.020.

Imaizumi, K., Taniguchi, K., Ogawa, Y., Matsuzaki, K., Nagata, T., Mochimaru, M., et al. (2015). Three-dimensional analyses of aging-induced alterations in facial shape: a longitudinal study of 171 Japanese males. *Int. J. Legal Med.* 129, 385–393. doi:10.1007/s00414-014-1114-x.

Kesterke, M. J., Raffensperger, Z. D., Heike, C. L., Cunningham, M. L., Hecht, J. T., Kau, C. H., et al. (2016). Using the 3D Facial Norms Database to investigate craniofacial sexual dimorphism in healthy children, adolescents, and adults. *Biol. Sex Differ.* 7, 23. doi:10.1186/s13293-016-0076-8.

Klingenberg, C. P. (2010). Evolution and development of shape: integrating quantitative approaches. *Nat. Rev. Genet.* 11, 623–635. doi:10.1038/nrg2829.

Klingenberg, C. P., and Marugan-Lobon, J. (2013). Evolutionary covariation in geometric morphometric data: analyzing integration, modularity, and allometry in a phylogenetic context. *Syst. Biol.* 62, 591–610. doi:10.1093/sysbio/syt025.

Klingenberg, C. P., Wetherill, L., Rogers, J., Moore, E., Ward, R., Autti-Rämö, I., et al. (2010). Prenatal alcohol exposure alters the patterns of facial asymmetry. *Alcohol* 44, 649–657. doi:10.1016/j.alcohol.2009.10.016.

Lele, S., and Richtsmeier, J. T. (1991). Euclidean distance matrix analysis: A coordinate‐free approach for comparing biological shapes using landmark data. *Am. J. Phys. Anthropol.* 86, 415–427. doi:10.1002/ajpa.1330860307.

Levene, H. (1960). “Robust tests for equality of variances,” in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, eds. I. Olkin and H. Hotelling (Stanford: Stanford University Press), 278–292.

Li, M., Cole, J. B., Manyama, M., Larson, J. R., Liberton, D. K., Riccardi, S. L., et al. (2017). Rapid automated landmarking for morphometric analysis of three-dimensional facial scans. *J. Anat.* 230, 1–12. doi:10.1111/joa.12576.

Liu, F., van der Lijn, F., Schurmann, C., Zhu, G., Chakravarty, M. M., Hysi, P. G., et al. (2012). A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans. *PLoS Genet.* 8, e1002932. doi:10.1371/journal.pgen.1002932.

Martinez-Abadias, N., Gonzalez-Jose, R., Gonzalez-Martin, A., Van der Molen, S., Talavera, A., Hernandez, P., et al. (2006). Phenotypic evolution of human craniofacial morphology after admixture: a geometric morphometrics approach. *Am. J. Phys. Anthropol.* 129, 387–98. doi:10.1002/ajpa.20291.

Martinez-Abadias, N., Mitteroecker, P., Parsons, T. E., Esparza, M., Sjovold, T., Rolian, C., et al. (2012). The developmental basis of quantitative craniofacial variation in humans and mice. *Evol. Biol.* 39, 554–567. doi:10.1007/s11692-012-9210-7.

Matthews, H., Penington, A., Clement, J., Kilpatrick, N., Fan, Y., and Claes, P. (2018a). Estimating age and synthesising growth in children and adolescents using 3D facial prototypes. *Forensic Sci. Int.* 286, 61–69. doi:10.1016/j.forsciint.2018.02.024.

Matthews, H. S., Penington, A. J., Hardiman, R., Fan, Y., Clement, J. G., Kilpatrick, N. M., et al. (2018b). Modelling 3D craniofacial growth trajectories for population comparison and classification illustrated using sex-differences. *Sci. Rep.* 8, 4771. doi:10.1038/s41598-018-22752-5.

Mitteroecker, P., Gunz, P., Bernhard, M., Schaefer, K., and Bookstein, F. L. (2004). Comparison of cranial ontogenetic trajectories among great apes and humans. *J. Hum. Evol.* 46, 679–698. doi:10.1016/j.jhevol.2004.03.006.

O’Higgins, P. (2000). The study of morphological variation in the hominid fossil record: biology, landmarks and geometry. *J. Anat.* 197, 103–120. doi:10.1046/j.1469-7580.2000.19710103.x.

Paternoster, L., Zhurov, A. I., Toma, A. M., Kemp, J. P., St. Pourcain, B., Timpson, N. J., et al. (2012). Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position. *Am. J. Hum. Genet.* 90, 478–485. doi:10.1016/j.ajhg.2011.12.021.

Quinto-Sánchez, M., Adhikari, K., Acuña-Alonzo, V., Cintas, C., Silva de Cerqueira, C. C., Ramallo, V., et al. (2015). Facial asymmetry and genetic ancestry in Latin American admixed populations. *Am. J. Phys. Anthropol.* 157, 58–70. doi:10.1002/ajpa.22688.

Richtsmeier, J. T., Baxter, L. L., and Reeves, R. H. (2000). Parallels of craniofacial maldevelopment in Down syndrome and Ts65Dn mice. *Dev. Dyn.* 217, 137–45. doi:10.1002/(SICI)1097-0177(200002)217:2<137::AID-DVDY1>3.0.CO;2-N.

Richtsmeier, J. T., Burke Deleon, V., and Lele, S. R. (2002). The promise of geometric morphometrics. *Am. J. Phys. Anthropol.* 119, 63–91. doi:10.1002/ajpa.10174.

Richtsmeier, J. T., Paik, C. H., Elfert, P. C., Cole III, T. M., and Dahlman, H. R. (1995). Precision, repeatibility, and validation of the localization of cranial landmarks using computed tomography scans. *Cleft Palate-Craniofacial J.* 32, 217–227. doi:10.1597/1545-1569\_1995\_032\_0217\_pravot\_2.3.co\_2.

Rohlf, F. J., and Marcus, L. F. (1993). A revolution morphometrics. *Trends Ecol. Evol.* 8, 129–132. doi:10.1016/0169-5347(93)90024-J.

Rohlf, F. J., and Slice, D. (1990). Extensions of the procrustes method for the optimal superimposition of landmarks. *Syst. Zool.* 39, 40–59. doi:10.2307/2992207.

Schlager, S., and Rudell, A. (2015). Analysis of the human osseous nasal shape — population differences and sexual dimorphism. *Am. J. Phys. Anthropol.* 157, 571–581. doi:10.1002/ajpa.22749.

Shaffer, J. R., Orlova, E., Lee, M. K., Leslie, E. J., Raffensperger, Z. D., Heike, C. L., et al. (2016). Genome-Wide Association Study Reveals Multiple Loci Influencing Normal Human Facial Morphology. *PLoS Genet.* 12, 1–21. doi:10.1371/journal.pgen.1006149.

Shaner, D. J., Peterson, A. E., Beattie, O. B., and Bamforth, J. S. (2000). Assessment of soft tissue facial asymmetry in medically normal and syndrome-affected individuals by analysis of landmarks and measurements. *Am. J. Med. Genet.* 93, 143–154. doi:10.1002/1096-8628(20000717)93:2<143::AID-AJMG12>3.0.CO;2-Q.

Smith, K. K. (2006). Craniofacial development in marsupial mammals: developmental origins of evolutionary change. *Dev. Dyn.* 235, 1181–93. doi:10.1002/dvdy.20676.

Snyders, J., Claes, P., Vandermeulen, D., and Suetens, P. (2014). Development and comparison of non-rigid surface registraion and extensions, Technical report KUL/ESAT/PSI/1401. Leuven, Belgium.

Starbuck, J. M., Reeves, R. H., and Richtsmeier, J. T. (2011). Morphological integration of soft-tissue facial morphology in Down syndrome and siblings. *Am. J. Phys. Anthropol.* 146, 560–568. doi:10.1002/ajpa.21583.

Subburaj, K., Ravi, B., and Agarwal, M. (2009). Automated identification of anatomical landmarks on 3D bone models reconstructed from CT scan images. *Comput. Med. Imaging Graph.* 33, 359–368. doi:10.1016/j.compmedimag.2009.03.001.

Suttie, M., Foroud, T., Wetherill, L., Jacobson, J., Molteno, C., Meintjes, E., et al. (2013). Facial Dysmorphism Across the Fetal Alcohol Spectrum. *Pediatrics* 131, e779–e788. doi:10.1542/peds.2012-1371.

Suttie, M., Wetherill, L., Jacobson, S. W., Jacobson, J. L., Hoyme, H. E., Sowell, E. R., et al. (2017). Facial curvature detects and explicates ethnic differences in effects of prenatal alcohol exposure. *Alcohol. Clin. Exp. Res.* 41, 1471–1483. doi:10.1111/acer.13429.

Terhune, C. E., Kimbel, W. H., and Lockwood, C. A. (2007). Variation and diversity in Homo erectus: a 3D geometric morphometric analysis of the temporal bone. *J. Hum. Evol.* 53, 41–60. doi:10.1016/j.jhevol.2007.01.006.

Toma, A. M., Zhurov, A. I., Playle, R., Ong, E., and Richmond, S. (2009). Reproducibility of facial soft tissue landmarks on 3D laser-scanned facial images. *Orthod. Craniofacial Res.* 12, 33–42. doi:10.1111/j.1601-6343.2008.01435.x.

von Cramon-Taubadel, N., Frazier, B. C., and Mirazon-Lahr, M. (2007). The problem of assessing landmark error in geometric morphometrics: Theory, methods, and modifications. *Am. J. Phys. Anthropol.* 134, 24–35. doi:10.1002/ajpa.

Walter, T., Shattuck, D. W., Baldock, R., Bastin, M. E., Carpenter, A. E., Duce, S., et al. (2010). Visualization of image data from cells to organisms. *Nat. Methods* 7, S26–S41. doi:10.1038/nmeth.1431.

Wei, R., Claes, P., Walters, M., Wholley, C., and Clement, J. G. (2011). Augmentation of linear facial anthropometrics through modern morphometrics: A facial convexity example. *Aust. Dent. J.* 56, 141–147. doi:10.1111/j.1834-7819.2011.01315.x.

Weinberg, S. M., Scott, N. M., Neiswanger, K., Brandon, C. A., and Marazita, M. L. (2004). Digital three-dimensional photogrammetry: Evaluation of anthropometric precision and accuracy using a Genex 3D camera system. *Cleft Palate-Craniofacial J.* 41, 507–518. doi:10.1597/03-066.1.

Wong, J. Y., Oh, A. K., Ohta, E., Hunt, A. T., Rogers, G. F., Mulliken, J. B., et al. (2008). Validity and reliability of craniofacial anthropometric measurement of 3D digital photogrammetric images. *Cleft Palate-Craniofacial J.* 45, 232–239. doi:10.1597/06-175.

# Data Availability Statement

The informed consent with which the data were collected does not allow for dissemination of identifiable data to persons not listed as researchers on the IRB protocol. Thus, the full surface 3D facial images used for validation cannot be made publicly available. In the interest of reproducibility, we have provided the 19 manual and automatic landmarks used for validation as well as the code used to analyze them. These data are available in the following GitHub repository: https://github.com/juliedwhite/RemappingValidation/. The MeshMonk code and tutorials are available at https://github.com/TheWebMonks/meshmonk.