Is shape in the eye of the beholder? The reproducibility of geometric morphometric analyses on live fish

Paolo Moccetti[1,2,3], Jessica R. Rodger[4], Jonathan D. Bolland[2], Phoebe Kaiser-Wilks[5], Rowan Smith[5], Andy D. Nunn[2], Colin E. Adams[5], Jen A. Bright[6], Hannele M. Honkanen[5], Angus J. Lothian[4], Matthew Newton[5], Domino A. Joyce[1]

[1] Evolutionary and Ecological Genomics Group, School of Natural Sciences, University of Hull, UK
[2] Hull International Fisheries Institute, School of Natural Sciences, University of Hull, UK
[3] Energy and Environment Institute, University of Hull, UK
[4] Atlantic Salmon Trust Fellow, Scottish Centre for Ecology and the Natural Environment, Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, UK
[5] Scottish Centre for Ecology and the Natural Environment, Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, UK
[6] School of Natural Sciences, University of Hull, UK

Corresponding Author:
Paolo Moccetti[1,2,3]
School of Natural Sciences, University of Hull
Cottingham Road, Hull, HU6 7RX, UK
Email address: p.m.moccetti-2019@hull.ac.uk, moccetti.paolo@gmail.com

Abstract

Geometric morphometrics is widely used to quantify morphological variation between biological specimens, but the fundamental influence of operator bias on data reproducibility is rarely considered, particularly in studies using photographs of live animals taken under field conditions. We examined this using four independent operators that applied an identical landmarking scheme to replicate photographs of 291 live Atlantic salmon (*Salmo salar* L.) from two rivers. Using repeated measures tests, we found significant inter-operator differences in mean body shape, suggesting that the operators introduced a systematic error despite following the same landmarking scheme. No significant differences were detected when the landmarking process was repeated by the same operator on a random subset of photographs. Importantly, in spite of significant operator bias, small but statistically significant morphological differences between fish from the two rivers were found consistently by all operators. Pairwise tests of angles of vectors of shape change showed that these between-river differences in body shape were analogous across operator datasets,

suggesting a general reproducibility of findings obtained by geometric morphometric studies. In contrast, merging landmark data when fish from each river are digitised by different operators had a significant impact on downstream analyses, highlighting an intrinsic risk of bias. Overall, we show that, even when significant inter-operator error is introduced during digitisation, following an identical landmarking scheme can identify morphological differences between populations. This study indicates that data sharing across research groups can potentially yield reliable results provided all operators involved digitise at least a sub-set of all data groups of interest.

Introduction

Landmark-based geometric morphometrics (GM) is a quantitative approach widely used to describe the shape of biological specimens and its covariation with other biological and environmental factors (Zelditch et al., 2004; Webster & Sheets, 2010). Morphological variables are quantified using a set of Cartesian landmarks located on distinct homologous anatomical points, and observed body shape variations are then displayed through user-friendly graphical representations (Zelditch et al., 2004; Mitteroecker & Gunz, 2009; Adams, Rohlf & Slice, 2013). GM is a powerful technique capable of detecting even tiny morphological differences among groups of specimens (Mitteroecker & Gunz, 2009; Webster & Sheets, 2010), but is highly sensitive to measurement errors introduced during data acquisition, which can affect subsequent analyses and produce inaccurate results (von Cramon-Taubadel, Frazier & Lahr, 2007; Fruciano, 2016; Robinson & Terhune, 2017; Fox, Veneracion & Blois, 2020). This is particularly problematic when such morphological differences are erroneously regarded as biologically meaningful variation (Fruciano, 2016).

Surprisingly, despite GM being a widely used technique, researchers rarely consider measurement error in their study design and statistical analyses (Fruciano, 2016; Fox, Veneracion & Blois, 2020). Measurement error can be introduced at different stages of the data acquisition process, i.e. when positioning specimens in front of the imaging device (camera or scanner), during image capture and landmark digitisation (Arnqvist & Mårtensson, 1998; Muir, Vecsei & Krueger, 2012; Fruciano et al., 2020; Fox, Veneracion & Blois, 2020). Indeed, the so-called inter-operator (or inter-observer) error during landmarking has been found to be one of the most critical factors affecting GM analyses because different operators tend to position what should be homologous landmarks in slightly different locations (Ross & Williams, 2008; Dujardin, Kaba & Henry, 2010; Campomanes-Álvarez et al., 2015; Fruciano, 2016; Fruciano et al., 2020; Fox, Veneracion & Blois, 2020). Importantly, inter-operator error can be substantial and potentially obscure biological variation, making data sharing and comparisons of landmarked datasets difficult (Shearer et al., 2017).

Intra-operator (or intra-observer) error has also been shown to significantly affect GM analyses (Wilson, Cardoso & Humphrey, 2011; Fox, Veneracion & Blois, 2020). Intra-operator error is introduced when specimens are inconsistently digitised by a single operator and can be influenced by several factors, including landmarking experience or time between

landmarking sessions (Fox, Veneracion & Blois, 2020). However, the magnitude of intra-operator error is invariably modest compared to inter-operator discrepancies (Cardoso & Saunders, 2008; Dujardin, Kaba & Henry, 2010; Wilson, Cardoso & Humphrey, 2011; Robinson & Terhune, 2017; Shearer et al., 2017; Thoma et al., 2018; Fox, Veneracion & Blois, 2020), indicating a general good precision in digitisation by individual operators (but see Engelkes et al., 2019).

The degree and impacts of operator error in GM studies have been tested for a range of organisms, anatomical structures, preservation methods and image acquisition devices (Fruciano, 2016; Fruciano et al., 2020; Fox, Veneracion & Blois, 2020). Nevertheless, most studies have focussed on images of specific human, bone or plant structures acquired under identical (laboratory) conditions (e.g., Ross & Williams, 2008; Cardoso & Saunders, 2008; Gonzalez, Bernal & Perez, 2011; Wilson, Cardoso & Humphrey, 2011; Viscosi & Cardini, 2011; Shearer et al., 2017; Carayon et al., 2019; Engelkes et al., 2019; Messer et al., 2021). Few have investigated images of live animals (but see Fruciano et al., 2020), despite commonly being used when it is not possible to euthanise samples for ethical reasons or research purposes. Undeniably, such photographs, especially if taken under field conditions, are more likely to result in subsequent measurement error (relative to preserved specimens) (Muir, Vecsei & Krueger, 2012), thereby restricting the utility of such datasets (Webster & Sheets, 2010). Understanding the prevalence, magnitude and implications of inter- and intra-operator error during the landmark digitisation process for photographs of live animals could facilitate data sharing and open science practices.

With the increasing focus on reproducibility in science (Baker, 2016), and an acknowledgment that sharing data can accelerate scientific progress, assessing whether live animals digitised repeatedly by single versus multiple operators produce consistent results and conclusions is essential. Data exchange, such as crowdsourcing, is opening new frontiers in GM research, enabling large-scale studies, which use unprecedented sample sizes, to be conducted within a short time frame (Thomas, Bright & Cooney, 2016; Chang & Alfaro, 2016). Such studies, involving several operators collecting shape data, can potentially address key questions in evolutionary biology and other disciplines (Cooney et al., 2017; Hughes et al., 2022). However, pooling landmarked datasets from multiple operators can increase the degree of measurement error (Fruciano et al., 2017; Evin, Bonhomme & Claude, 2020), but the consequences of inter-individual operator error when sharing datasets remain poorly understood.

The aim of this study was therefore to determine whether GM analyses on photographs of live animals are reproducible. To accomplish this, four independent operators digitised the same photographs of sedated Atlantic salmon (*Salmo salar* L.) sampled in two rivers, following a shared landmarking scheme. The shape data and results obtained by the four operators were then compared and contrasted to assess the magnitude of inter- and intra-operator error, and infer the potential for meaningful data sharing.

Material & Methods

Study design

Salmon were captured from the River Spey (57° 24.960' N 3° 22.602' W) and River Oykel (57° 59.640' N 4° 48.282' W) in Scotland using a 1.5 m diameter Rotary Screw trap during their smolt stage, i.e. on their first migration to sea. The sampling occurred in the context of a tracking study aiming to identify areas and causes of smolt mortality during their seaward migration (see Whelan, Roberts & Gray, 2019). Fish were photographed in the field under anaesthetic before being tagged and released to the river after recovery. Photographs of the left side of each fish were taken freehand from approximately 30 cm directly above the fish, with a Fujifilm FinePix XP130 Compact Digital Camera with fish on a background reference scale. Photographs were taken by a team of eight people who met prior to field work to standardise methods as far as possible. Our study here focusses on inter-operator variation downstream of photography, but variation caused by variation between individual photographers would be worthy of future study. The care and use of experimental animals complied with the UK Home Office animal welfare laws, guidelines and policies (UK Home Office Licence PPL 70/8794) and was approved by the University of Glasgow Animal Welfare and Ethics Review Board (AWERB).

The GM analyses were based on photographs of 291 salmon (Spey n = 144, Oykel n = 147). The images were imported into tpsUtil v. 1.78 (Rohlf, 2019) and randomly shuffled using the relevant function so that operators were blinded to the river-of-origin of the specimens. Twenty-two landmarks were digitised on each image by four independent operators (Op.1, Op.2, Op.3 and Op.4) using tpsDig v. 2.31 (Rohlf, 2017) and following an identical scheme (Fig. 1). All landmarks were fixed. Landmarks (1), (2) and (3) were placed using the *curves tool* in tpsDig, with the curve starting at landmark (1) and ending at landmark (3). Landmark (2) was automatically placed as equidistant between landmarks (1) and (3) using the *by length* option and choosing *number of points* = 3. The three points along the curve were then converted to fixed landmarks using the *Append tps Curve to landmarks* function. The landmark positions chosen were those commonly used in studies on salmonids (e.g., Boulding et al., 2008; Muir, Vecsei & Krueger, 2012; Simonsen et al., 2017; Goerig et al., 2019; Dermond, Sperlich & Brodersen, 2019). In addition, the first ten fish for each river, after using the *randomly order specimens* function in tpsUtil, were consecutively landmarked a further two times (i.e. three times in total) by each operator to evaluate the intra-operator consistency in digitisation.

Landmark coordinates from all operators were imported as unique files into R (R Core Team, 2021) and analysed using the 'geomorph' and 'RRPP' v. 4.0.4 (Adams et al., 2021; Baken et al., 2021; Collyer & Adams, 2021), 'Morpho' v. 2.8 (Schlager, 2017), and 'GeometricMorphometricsMix' v. 0.0.8.4 (Fruciano, 2018) packages. Plots were produced with the 'ggplot2' package (Wickham, 2016), while projections of body shape variation between groups were generated with the *plotRefToTarget* function in 'geomorph'.

The landmark data were then used to test if: (1) similar mean body shapes were obtained by all operators; (2) any morphological differences between salmon from the two different rivers were detected by all operators; (3) identified between-river differences were consistent across all operators; (4) divergent datasets from different operators could be merged; and (5) the magnitude of intra-operator error was similar across operators.

Preliminary analyses

First, a generalised Procrustes analysis (GPA) was performed to remove effects not related to body shape through translation, scaling and rotation of the landmark configurations (Rohlf & Slice, 1990). A preliminary principal component analysis (PCA) conducted on superimposed coordinates revealed body bending as a major source of shape variation, a known issue in morphometric studies on fish (Valentin et al., 2008). To remove the bending effect, the *unbend* function in tpsUtil was used, employing landmarks 1, 21, 22 and 17, which normally lie in a straight line in salmonids (Arbour, Hardie & Hutchings, 2011; Dermond, Sperlich & Brodersen, 2019). All subsequent analyses were performed on landmarks 1-20 only. A new GPA on coordinates with the bending deformation removed was then executed and outlier specimens were investigated for each operator using the *plotOutliers* function in 'geomorph'. Two specimens digitised by one operator were found to be very different to the other individuals and were therefore removed from the dataset of all four operators, leaving 289 samples for analyses (Spey n = 144; Oykel n = 145). Another GPA using the landmark data without outliers was then implemented.

Test 1. Were similar mean body shapes obtained by all operators?

To investigate whether results produced by a single operator are accurate and reproducible, we tested differences in the mean body shapes of fish digitised by independent operators. First, a between-group PCA (Boulesteix, 2004) was computed to explore variations between the four operators. Between-group PCA is a type of discriminant analysis used to maximise segregation between known groups which, unlike canonical variate analysis (CVA), does not overestimate the degree of distinction among groups (Mitteroecker & Bookstein, 2011). The leave-one-out cross-validation operation was implemented to quantify the proportion of fish specimens correctly assigned to the operator who digitised them.

To investigate whether landmarking by multiple operators introduced bias, i.e. systematic error affecting body shape (*sensu* Fruciano, 2016), differences in the mean body shapes of the fish digitised by the four independent operators were tested using Hotelling's T^2 as implemented by the *repeated_measures_test* function in 'GeometricMorphometricsMix'. To compute the differences in mean body shapes, a PCA was performed on all Procrustes-aligned coordinates of all operators, and the scores for all the PC axes (i.e. 100% variance explained) of each operator were then used in a repeated measures test as an approximation of shape.

Test 2. Were morphological differences between salmon from different rivers detected by all operators?

We next tested whether there was a difference in body shape between rivers, and whether the operators were consistent in identifying any differences. The following analyses were performed separately for each operator. First, a GPA was computed on landmark coordinate datasets obtained by each operator with outliers removed (see end of *Preliminary analyses*). The effect of fish size on body shape was tested using Procrustes ANOVAs (*procD.lm* function in 'geomorph'), with Procrustes coordinates used as an outcome variable, the log value of centroid size and 'River' as independent variables and a randomised residual permutation procedure (10,000 iterations). A small but significant effect of size on shape was found for all operators (P-values < 0.0001, $r^2 = 0.022–0.034$). Procrustes coordinates were therefore adjusted for allometry by using residuals from a regression of shape against centroid size + 'River'. Procrustes ANOVAs were then used to compare mean body shape between rivers, while another between-group PCA was implemented to quantify the proportion of fish correctly assigned to the river of origin, for each of the four datasets.

Test 3. Were identified between-river differences consistent across all operators?

To assess if body shape differences between rivers were analogous across operators, pairwise angles (Li, 2011) of vectors of shape change between fish from the rivers Spey and Oykel were computed. The *TestOfAngle* function in 'GeometricMorphometricsMix' based on the analogous function implemented in 'Morpho' was used, as performed by Fruciano et al. (2020). Specifically, we calculated the pairwise angles among between-group principal components obtained using 'River' as the grouping factor within each operator subset of digitisations (one between-group PC axis - herein bwgPC - per operator) to test if they followed the same "direction", i.e. if the shape differences between rivers were approximately the same for all operators.
Furthermore, bwgPC1 vectors of between-river differences for each operator were compared (test of angles) with bwgPC1-3 vectors of inter-operator differences obtained in *Test 1*. In this way, it was possible to determine whether or not biological body shape differences between rivers and artefactual variation among operators were similar (following the same "direction").
Finally, the magnitude of shape differences between rivers obtained by each operator was examined with the *dist_mean_boot* function in 'GeometricMorphometricsMix'. This function was used to perform a bootstrap estimate of the shape distance between the two rivers and allowed us to test if the amount of shape difference between the rivers Spey and Oykel was consistent across different operators or, on the contrary, one or more operators detected larger or smaller between-river differences than the others.

Test 4. Can divergent datasets from different operators be merged?

The two operators producing the most dissimilar mean shapes were used to simulate a worst-case-scenario process of data pooling, in which two independent researchers perform their own GM study each on different rivers, but following the same landmarking scheme. Inter-operator analysis showed that Op.2 and Op.4 produced the most dissimilar body shapes (greatest Euclidean distance), so from these, two datasets were created: one comprising shape data from the River Oykel digitised by Op.2 (herein Op.2-Oykel) and the River Spey data digitised by Op.4 (herein Op.4-Spey) and *vice versa*, i.e. the River Spey data digitised by Op.2 (herein Op.2-Spey) and the River Oykel data digitised by Op.4 (herein Op.4-Oykel). For both datasets, differences between rivers were tested with Procrustes ANOVA, as described earlier. Then, a between-group PCA was performed and the resulting bwgPC1 separating the two rivers was used to run a test of angles to compare between-river differences detected by these two separate datasets. We also compared these latter between-river differences with those found when using the complete datasets of all four operators including both rivers (see section above). This enabled us to test whether any between-river differences as a result of different operators outweighed any biological differences between rivers found when using the complete intra-operator datasets.

Test 5. Quantifying intra-operator error

A GPA was computed separately on landmark coordinates obtained by each operator re-digitising a sub-sample of 20 fish (ten per river). Individual consistency in landmarking was then investigated using PCA and tested using repeated measures tests. To test for differences in mean body shapes between digitisation trials, a PCA was performed on the Procrustes-aligned coordinates of each operator separately and the PC scores of each trial were then used in the repeated measures tests as an approximation of shape. Repeatability among digitisation trials was also calculated for each operator using the intraclass correlation coefficient (Fisher, 1958). A one-way Procrustes ANOVA was computed using individual fish as a categorical variable (Fruciano, 2016). The resulting mean squares were used to calculate repeatability by applying equations presented in Arnqvist & Mårtensson (1998) and Fruciano (2016). Here, repeatability measured variation in the three independent digitisations of the sub-sample of 20 salmon relative to the variability among specimens, i.e. the biological variation among all fish samples. Repeatability assumes a value of between zero and one, with one indicating 100% repeatability and an absence of measurement error (Arnqvist & Mårtensson, 1998; Fruciano, 2016). Finally, a Procrustes ANOVA with individual fish specimens (‘ID’) as the main factor and ‘operator’ nested within ‘ID’ was run to test the relative contributions of biological variation (‘ID’) and variation introduced by inter-operator (‘ID:operator’) and intra-operator (residual) error.

Results

Test 1. Were similar mean body shapes obtained by all operators?

Despite digitising replicate photographs with homologous landmarks, fish specimens were correctly assigned to their operator based on body shape with 82.6% accuracy by the exploratory between-group PCA (Fig. 2, Supplementary Table 1). There was a significant operator effect on mean body shape, with all pairwise tests displaying highly significant differences between operators (P-value < 0.001 for all comparisons; Table 1), supporting the exploratory between-group PCA (Fig. 2). The Euclidean distances between means, i.e. the measure of the extent of shape change, highlighted different distances among pairs of operators, with the smallest difference (0.00992) occurring between Op.1 and Op.3 and the greatest between Op.2 and Op.4 (0.02270). The between-group PCA scatterplot (Fig. 2) broadly reflected these results along axis 1 (71.8% of variance), with Op.1 and Op.3 overlapping extensively and Op.2 and Op.4 being furthest apart. The anatomical differences among operators were concentrated mainly on the head (Fig. 3, Supplementary Fig. 1), with major areas of disagreement being the snout, eye, mouth and posterior of the cranium. Morphological dissimilarities were more or less pronounced depending on the operator comparisons (Fig. 3).

Test 2. Were morphological differences between salmon from different rivers detected by all operators?

There were significant differences in body shape between fish from different rivers (Spey and Oykel; Table 2), with between-group PCA (Supplementary Table 2) separating them for all operators (72.2% mean classification success rate). The fish from the River Oykel had a greater body depth, more pronounced caudal peduncle, larger eye, longer mouth and more pointed snout than those from the River Spey (Fig. 4).

Test 3. Were identified between-river differences consistent across all operators?

All comparisons of the "direction" of body shape variation between rivers were significant for all operators, meaning that the way in which shape differed between the rivers Spey and Oykel was approximately the same for all operators (Table 3). In contrast, pairwise comparisons of between-river and between-operator differences were mostly non-significant, with only two of 12 tests generating P-values < 0.05 (Table 3; Supplementary Table 3). This indicates that the shape variation between rivers and operators were divergent and not collinear. Estimated mean distances between rivers computed through bootstrapping were similar across operators, as shown by the widely overlapping confidence intervals (Table 4), suggesting that different operators did not influence the magnitude of shape difference detected between the rivers Spey and Oykel.

Test 4. Can divergent datasets from different operators be merged?

There were significant differences in body shape between fish from the rivers Spey and Oykel digitised separately by Op.2 and Op.4 (Table 5). Notably, shape variation explained by the rivers was markedly higher for these merged datasets compared to the between-river

differences detected by single operators (r^2 = 0.18-0.26 vs. 0.03-0.04, respectively; Tables 2 and 5). Similarly, the between-group PCA separated fish from different rivers with a higher accuracy than the analogous analysis performed on individual operator datasets (93.1% vs. 72.2% mean classification success rate, respectively; Supplementary Tables 2 and 4). The comparison of the "direction" of between-river body shape variation detected by the two merged datasets from Op.2 and Op.4 was highly significant (P-value < 0.0001; Supplementary Table 5), meaning that the way in which shape differed between the rivers Spey and Oykel was approximately the same regardless of the selected dataset. However, only half of the comparisons were found to be significant when comparing the two Op.2 and Op.4 merged datasets with the complete within-operator datasets including both rivers (Supplementary Table 5), indicating that the river differences detected by combined and individual operator datasets were only partly similar.

Test 5. Quantifying intra-operator error

There was extensive overlap among landmarking trials, suggesting a high consistency in digitisation across all operators (Fig. 5). Pairwise comparisons supported this since none of the mean body shapes differed significantly between repeated digitisations (P-values > 0.81; Table 6). All four operators achieved the highest landmarking consistency between trials 2 and 3, as indicated by the smallest Euclidean distance values (0.002-0.005). Repeatability was also high for all operators (0.892-0.975), indicating high landmarking precision (Table 7). Nested Procrustes ANOVA showed that 57.5% of the morphological variation within the subset of 20 fish was explained by individual fish ('ID'), while 34.8% and 7.8% of the variation was attributable, respectively, to inter- ('ID:operator') and intra- (residual) operator digitisation error (Supplementary Table 6).

Discussion

We show here that independent operators applying an identical landmarking scheme to replicate photographs of live Atlantic salmon taken in field conditions yielded significantly different mean body shapes (*Test 1*). However, morphological differences between salmon from different rivers were detected by all operators (*Test 2*), and these were consistent differences across all operators (*Test 3*), provided they landmarked both rivers and not one each (*Test 4*). Furthermore, intra-operator error calculated on a subset of samples was minimal, suggesting that it did not have a significant influence on the body shape results obtained by the different operators (*Test 5*).

Despite digitising replicate photographs with homologous landmarks, all the operators produced significantly different mean body shapes. The high rate (82.6%) of specimens assigned to the correct operator by the between-group PCA suggests that the operators introduced a systematic error, which created four identifiable body shapes despite following the same landmarking scheme. This digitisation bias is likely to have been introduced by operators consistently applying personal, fine-scaled landmarking rules in addition to the general scheme. The fact that the differences among operators were localised mainly in the

head region (landmarks 1-12) may be explained by the less discrete and recognizable nature of these landmarks compared to those located on well-defined anatomical loci, such as fin intersections (landmarks 13-20). This suggests that the use of unambiguous landmarks can be an effective way of reducing measurement error in GM (Fagertun et al., 2014; Campomanes-Álvarez et al., 2015, Fruciano et al., 2017). The effect of sliding semi-landmarks (i.e. placed on non-discrete points along curves and surfaces; Bookstein, 1997; Gunz & Mitteroecker, 2013) on measurement error was not tested here, but Evin, Bonhomme & Claude (2020) found that, even though they are more difficult to digitise than those located on well-defined anatomical loci, semi-landmarks can reduce the amount of inter-operator error because GPA "spreads" the error ubiquitously among all geometric coordinates. It would be useful to fully understand how inter-operator error varies with and without the incorporation of semi-landmarks.

In GM studies, digitisation is typically performed by a single operator, leaving the question of whether multiple operators digitising the same set of images would generate different results. This could undermine the reliability of findings presented by many GM investigations, particularly those using images of live animals taken in field conditions, which are potentially more prone to measurement error (Webster & Sheets, 2010; Muir, Vecsei & Krueger, 2012). In our study, however, we found that inconsistencies between operators did not mask small, but significant morphological differences between fish from the rivers Spey and Oykel, which were consistent across operators. The fact that, as shown by tests of angles and bootstrapped estimates of mean distances, all the operators detected analogous between-river differences, strongly suggests that they were biologically authentic. Similarly, Fruciano et al. (2020) found that preservation methods significantly affected the body shape of brown trout (*Salmo trutta* L.), but the subsequent between-groups classification was similar regardless of preservation method. As suggested by Fruciano et al. (2020), this could be because the shape variation detected by the operators between the rivers Spey and Oykel was not significantly affected by inter-operator differences in landmarking because they were not collinear (i.e. they followed different "directions", as shown by the angle comparisons).

Conversely, merging landmark data of fish from the rivers Spey and Oykel digitised by two distinct operators (Op.2 and Op.4, *Test 4*) had a significant impact on subsequent analyses and produced contrasting results. As shown by Procrustes ANOVA and between-group PCA classification rate, shape differences between rivers in the merged datasets were greater than those detected by single operators, suggesting they were artificially inflated by inter-operator digitisation error. Angle comparisons showed that the river differences detected by combined and individual operator datasets were partly dissimilar. Overall, these findings point towards a potential risk in pooling datasets from multiple operators when there are confounding biological factors, as highlighted by other studies (Fruciano et al., 2017; Evin, Bonhomme & Claude, 2020). Distinct operators obtained analogous results when they landmarked both rivers (and not one river each as in *Test 4*). This suggests that operators digitising at least a

sub-set of all data groups of interest (rivers in this case) may be an effective way of mitigating inter-operator error and potentially enabling data sharing.

In contrast to the inter-operator effects described in this study, we found no statistical evidence of intra-operator effects on the quantification of fish morphology. On the contrary, we found a very high level of repeatability across trials for all operators. This corroborates previous studies that showed intra-operator error to be limited (e.g. Cardoso & Saunders, 2008; Dujardin, Kaba & Henry, 2010; Wilson, Cardoso & Humphrey, 2011; Robinson & Terhune, 2017; Shearer et al., 2017; Thoma et al., 2018; Fox, Veneracion & Blois, 2020). Interestingly, for all operators, landmarking consistency was highest between their last two trials, suggesting that they 'learnt' where to place the landmarks with increasing experience of the images. However, it should be noted that the first trial was performed while digitising all specimens, whereas trials 2 and 3 were performed consecutively after digitising the full dataset, which may have artificially inflated precision, with operators "remembering" their landmarking choices in trial 2 during trial 3.
The negligible impact of intra- compared to inter-operator error was also clearly shown by the percentage of variance explaining shape variation in the sub-sample of 20 fish (*Test 5*, 7.8% vs. 34.8%, respectively). Interestingly, the percentage of variance explained by inter-operator error (34.8%) is similar to that reported by Fruciano et al. (2020) for brown trout photographed in the field (30.1%), and supports previous studies that identified inter-operator effects as the major source of error in GM analyses (Ross & Williams, 2008; Dujardin, Kaba & Henry, 2010; Campomanes-Álvarez et al., 2015; Fruciano, 2016; Shearer et al., 2017; Fox, Veneracion & Blois, 2020).

Conclusions

Overall, we show that, even when significant inter-operator error is introduced through digitisation, following an identical landmarking scheme can be an effective tool to obtain robust and reliable results, even without accounting for variation introduced by the photography process, which was not quantified here. This implies that GM studies based on common landmarking schemes are potentially reproducible, even when analyses are based on images of live specimens taken in the field, as in the current study. Nevertheless, since operator error can vary between studies and is impossible to determine *a priori*, we recommend assessing the magnitude and effects of landmarking error by using multiple operators for a sub-set of samples, as here, to improve confidence in study results. If landmark data merging is required, we recommend that all the operators involved digitise at least a sub-set of all data groups of interest.

Acknowledgements

We would like to thank the Spey Fisheries Board and the Kyle of Sutherland Fisheries Trust for their help and support with this project. We would also like to thank Jonathan Archer, Georgios Kyriakou, Fraser Brydon, Jessica Whitney and Mustafa Soganci for their help collecting the data for this study. We also thank Carmelo Fruciano, Michelle C. Gilbert and Kiran Liversage for their constructive reviews.

References

Adams DC, Rohlf FJ, Slice DE. 2013. A field comes of age: geometric morphometrics in the 21st century. *Hystrix, the Italian Journal of Mammalogy* 24:7–14. DOI: 10.4404/hystrix-24.1-6283.

Adams DC, Collyer M, Kaliontzopoulou A, Baken E. 2021. "Geomorph: Software for geometric morphometric analyses. R package version 4.0.2." https://cran.r-project.org/package=geomorph.

Arbour JH, Hardie DC, Hutchings JA. 2011. Morphometric and genetic analyses of two sympatric morphs of Arctic char (*Salvelinus alpinus*) in the Canadian high Arctic. Canadian Journal of Zoology 89:19–30. DOI: 10.1139/Z10-100.

Arnqvist G, Mårtensson T. 1998. Measurement error in geometric morphometrics: empirical strategies to assess and reduce its impact on measures of shape. *Acta Zoologica Academiae Scientiarum Hungaricae* 44:73–96.

Baken EK, Collyer ML, Kaliontzopoulou A, Adams DC. 2021. geomorph v4.0 and gmShiny: Enhanced analytics and a new graphical interface for a comprehensive morphometric experience. *Methods in Ecology and Evolution* 12:2355–2363. DOI: https://doi.org/10.1111/2041-210X.13723.

Baker M. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533:452–454. DOI: 10.1038/533452a.

Bookstein FL. 1997. Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. *Medical Image Analysis* 1:225–243. DOI: https://doi.org/10.1016/S1361-8415(97)85012-8.

Boulding EG, Culling M, Glebe B, Berg PR, Lien S, Moen T. 2008. Conservation genomics of Atlantic salmon: SNPs associated with QTLs for adaptive traits in parr from four trans-Atlantic backcrosses. *Heredity* 101:381–391. DOI: 10.1038/hdy.2008.67.

Boulesteix A. 2004. A note on between-group PCA. *International Journal of Pure and Applied Mathematics* 19:359–366. DOI: 10.5282/ubm/epub.1767.

Campomanes-Álvarez BR, Ibáñez O, Navarro F, Alemán I, Cordón O, Damas S. 2015. Dispersion assessment in the location of facial landmarks on photographs. *International Journal of Legal Medicine* 129:227–236. DOI: 10.1007/s00414-014-1002-4.

Carayon D, Adhikari K, Monsarrat P, Dumoncel J, Braga J, Duployer B, Delgado M, Fuentes-Guajardo M, de Beer F, Hoffman JW, Oettlé AC, Donat R, Pan L, Ruiz-Linares A, Tenailleau C, Vaysse F, Esclassan R, Zanolli C. 2019. A geometric morphometric approach to the study of variation of shovel-shaped incisors. *American Journal of Physical Anthropology* 168:229–241. DOI: https://doi.org/10.1002/ajpa.23709.

Cardoso HF V, Saunders SR. 2008. Two arch criteria of the ilium for sex determination of immature skeletal remains: a test of their accuracy and an assessment of intra- and inter-observer error. *Forensic Science International* 178:24–29. DOI: 10.1016/j.forsciint.2008.01.012.

Chang J, Alfaro ME. 2016. Crowdsourced geometric morphometrics enable rapid large-scale collection and analysis of phenotypic data. *Methods in Ecology and Evolution* 7:472–482. DOI: https://doi.org/10.1111/2041-210X.12508.

Collyer ML, Adams DC. 2021. “RRPP: Linear Model Evaluation with Randomized Residuals in a Permutation Procedure, R package version 1.1.2.” https://cran.r-project.org/package=RRPP.

Cooney CR, Bright JA, Capp EJR, Chira AM, Hughes EC, Moody CJA, Nouri LO, Varley ZK, Thomas GH. 2017. Mega-evolutionary dynamics of the adaptive radiation of birds. *Nature* 542:344–347. DOI: 10.1038/nature21074.

Dermond P, Sperlich N, Brodersen J. 2019. Heritable morphological differentiation in salmonids from two distinct stream types. *Journal of Fish Biology* 95:1215–1222. DOI: 10.1111/jfb.14121.

Dujardin J-PAL, Kaba D, Henry AB. 2010. The exchangeability of shape. *BMC Research Notes* 3:266. DOI: 10.1186/1756-0500-3-266.

Engelkes K, Helfsgott J, Hammel JU, Büsse S, Kleinteich T, Beerlink A, Gorb SN, Haas A. 2019. Measurement error in μCT-based three-dimensional geometric morphometrics introduced by surface generation and landmark data acquisition. *Journal of Anatomy* 235:357–378. DOI: https://doi.org/10.1111/joa.12999.

Evin A, Bonhomme V, Claude J. 2020. Optimizing digitalization effort in morphometrics. *Biology Methods and Protocols* 5:bpaa023. DOI: 10.1093/biomethods/bpaa023.

Fagertun J, Harder S, Rosengren A, Moeller C, Werge T, Paulsen RR, Hansen TF. 2014. 3D facial landmarks: Inter-operator variability of manual annotation. *BMC Medical Imaging* 14:35. DOI: 10.1186/1471-2342-14-35.

Fisher RA. 1958. *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

Fox NS, Veneracion JJ, Blois JL. 2020. Are geometric morphometric analyses replicable? Evaluating landmark measurement error and its impact on extant and fossil *Microtus* classification. *Ecology and Evolution* 10:3260–3275. DOI: https://doi.org/10.1002/ece3.6063.

Fruciano C. 2016. Measurement error in geometric morphometrics. *Development Genes and Evolution* 226:139–158. DOI: 10.1007/s00427-016-0537-4.

Fruciano C. 2018. GeometricMorphometricsMix: miscellaneous functions useful for geometric morphometrics. Version 0.0.3. https://github.com/fruciano/GeometricMorphometricsMix.

Fruciano C, Celik MA, Butler K, Dooley T, Weisbecker V, Phillips MJ. 2017. Sharing is caring? Measurement error and the issues arising from combining 3D morphometric datasets. *Ecology and Evolution* 7:7034–7046. DOI: https://doi.org/10.1002/ece3.3256.

Fruciano C, Schmidt D, Ramírez Sanchez MM, Morek W, Avila Valle Z, Talijančić I, Pecoraro C, Schermann Legionnet A. 2020. Tissue preservation can affect geometric morphometric analyses: a case study using fish body shape. *Zoological Journal of the Linnean Society* 188:148–162. DOI: 10.1093/zoolinnean/zlz069.

Goerig E, Wasserman BA, Castro-Santos T, Palkovacs EP. 2019. Body shape is related to the attempt rate and passage success of brook trout at in-stream barriers. *Journal of Applied Ecology* 57:91–100. DOI: 10.1111/1365-2664.13497.

Gonzalez PN, Bernal V, Perez SI. 2011. Analysis of sexual dimorphism of craniofacial traits using geometric morphometric techniques. *International Journal of Osteoarchaeology* 21:82–91. DOI: https://doi.org/10.1002/oa.1109.

Gunz P, Mitteroecker P. 2013. Semilandmarks: a method for quantifying curves and surfaces. *Hystrix, the Italian Journal of Mammalogy* 24:103–109. DOI: 10.4404/hystrix-24.1-6292.

Hughes EC, Edwards DP, Bright JA, Capp EJR, Cooney CR, Varley ZK, Thomas GH. 2022. Global biogeographic patterns of avian morphological diversity. *Ecology Letters* 25:598–610. DOI: https://doi.org/10.1111/ele.13905.

Li S. 2011. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics & Statistics* 4:66–70. DOI: 10.3923/ajms.2011.66.70.

Messer D, Svendsen MS, Galatius A, Olsen MT, Dahl VA, Conradsen K, Dahl AB. 2021. Measurement error using a SeeMaLab structured light 3D scanner against a Microscribe 3D digitizer. *PeerJ* 9:e11804. DOI: 10.7717/peerj.11804.

Mitteroecker P, Bookstein F. 2011. Linear discrimination, ordination, and the visualization of selection gradients in modern morphometrics. *Evolutionary Biology* 38:100–114. DOI: 10.1007/s11692-011-9109-8.

Mitteroecker P, Gunz P. 2009. Advances in geometric morphometrics. *Evolutionary Biology* 36:235–247. DOI: 10.1007/S11692-009-9055-X/FIGURES/7.

Muir AM, Vecsei P, Krueger CC. 2012. A perspective on perspectives: methods to reduce variation in shape analysis of digital images. *Transactions of the American Fisheries Society* 141:1161–1170. DOI: 10.1080/00028487.2012.685823.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Robinson C, Terhune CE. 2017. Error in geometric morphometric data collection: combining data from multiple sources. *American Journal of Physical Anthropology* 164:62–75. DOI: https://doi.org/10.1002/ajpa.23257.

Rohlf F. 2017. tpsUtil v. 1.78. Stony Brook, NY: Department of Ecology and Evolution, State University of New York.

Rohlf F. 2019. tpsDig v. 2.31, Stony Brook, NY: Department of Ecology and Evolution, State University of New York.

Rohlf FJ, Slice D. 1990. Extensions of the Procrustes method for the optimal superimposition of landmarks. *Systematic Biology* 39:40–59. DOI: 10.2307/2992207.

Ross AH, Williams S. 2008. Testing repeatability and error of coordinate landmark data acquired from crania. *Journal of Forensic Sciences* 53:782–785. DOI: https://doi.org/10.1111/j.1556-4029.2008.00751.x.

Schlager S. 2017. Morpho and Rvcg – Shape Analysis in R. In: *Statistical Shape and Deformation Analysis* (eds Zheng G, Li S, Szekely G), pp. 217–256. Academic Press, London. ISBN 9780128104934.

Shearer BM, Cooke SB, Halenar LB, Reber SL, Plummer JE, Delson E, Tallman M. 2017. Evaluating causes of error in landmark-based data collection using scanners. *PLOS ONE* 12:e0187452.

Simonsen MK, Siwertsson A, Adams CE, Amundsen PA, Præbel K, Knudsen R. 2017. Allometric trajectories of body and head morphology in three sympatric Arctic charr (*Salvelinus alpinus* (L.)) morphs. *Ecology and Evolution* 7:7277–7289. DOI: 10.1002/ece3.3224.

Thoma C, Makridou K, Bakaloudis D, Vlachos C. 2018. Evaluating the potential of three-dimensional laser surface scanning as an alternative method of obtaining morphometric data. *Annales Zoologici Fennici* 55:55–66. DOI: 10.5735/086.055.0106.

Thomas GH; Bright JA; Cooney CR. 2016. *Mark My Bird* [Data set]. Natural History Museum. DOI: 10.5519/0005413

Valentin AE, Penin X, Chanut JP, Sévigny JM, Rohlf FJ. 2008. Arching effect on fish body shape in geometric morphometric studies. *Journal of Fish Biology* 73:623–638. DOI: 10.1111/j.1095-8649.2008.01961.x.

Viscosi V, Cardini A. 2011. Leaf Morphology, Taxonomy and Geometric Morphometrics: A Simplified Protocol for Beginners. *PLOS ONE* 6:e25630.

von Cramon-Taubadel N, Frazier BC, Lahr MM. 2007. The problem of assessing landmark error in geometric morphometrics: theory, methods, and modifications. *American Journal of Physical Anthropology* 134:24–35. DOI: https://doi.org/10.1002/ajpa.20616.

Webster M, Sheets HD. 2010. A practical introduction to landmark-based geometric morphometrics. *Quantitative methods in Paleobiology* 16:168–188. DOI: 10.1017/S1089332600001868.

Whelan K, Roberts D, Gray J. 2019. The SAMARCH Project International Salmonid Coastal and Marine Telemetry Workshop. Available at https://www.samarch.org/wp-content/uploads/2020/05/SAMARCH-Tracking-Conference-Nov-2019-final_compressed.pdf (accessed 11 March 2022).

Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York. ISBN 978-3-319-24277-4.

Wilson LAB, Cardoso HFV, Humphrey LT. 2011. On the reliability of a geometric morphometric approach to sex determination: a blind test of six criteria of the juvenile ilium. *Forensic Science International* 206:35–42. DOI: 10.1016/j.forsciint.2010.06.014.

Zelditch ML, Swiderski DL, Sheets HD, Fink WL. 2004. Introduction. In: *Geometric Morphometrics for Beginners: a Primer* (eds Zelditch ML, Swiderski DL, Sheets HD, Fink WL), pp. 1–20. Academic Press, San Diego. DOI: https://doi.org/10.1016/B978-

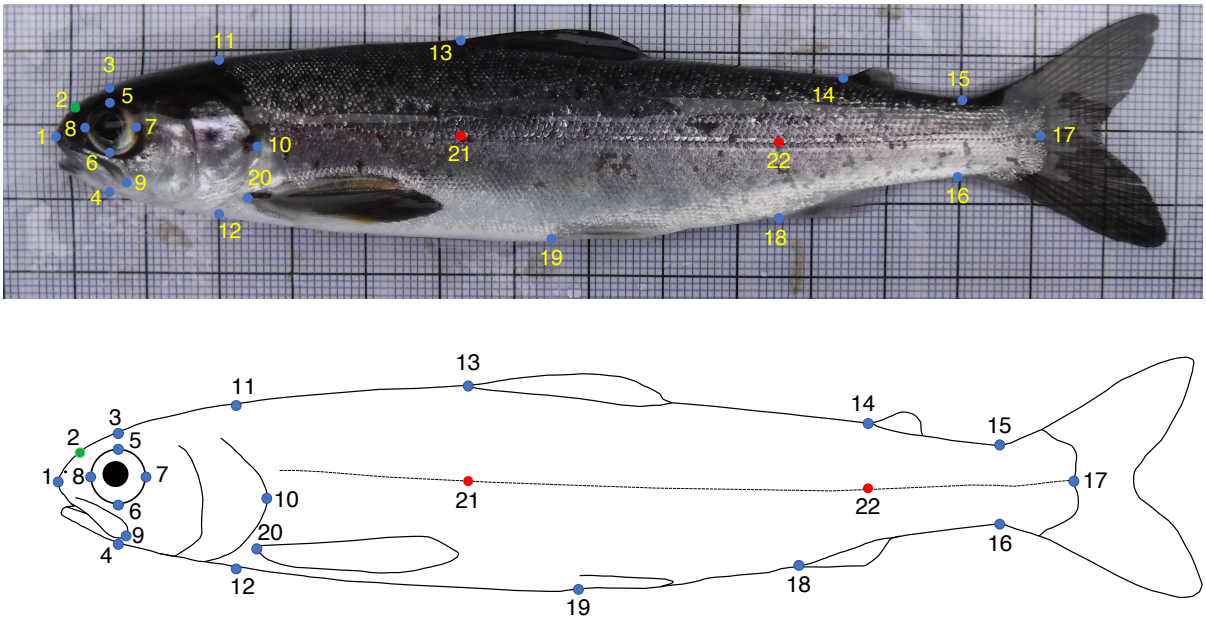

Figure 1. Landmarks used for the geometric morphometrics analyses of Atlantic salmon smolts. **(1)** Tip of snout; **(2)** Midpoint between 1 and 3; **(3)** Directly above middle of eye; **(4)** Perpendicular to 3; **(5)** Midpoint of top of eye (directly below 3); **(6)** Midpoint of bottom of eye (directly below 3); **(7)** Midpoint of posterior of eye; **(8)** Midpoint of anterior of eye; **(9)** End of maxillary bone; **(10)** Posterior of bony operculum; **(11)** Dorsal surface posterior of cranium; **(12)** Perpendicular to 11; **(13)** Anterior insertion point of dorsal fin; **(14)** Anterior insertion point of adipose fin; **(15)** Dorsal insertion point of caudal fin; **(16)** Directly below 15; **(17)** Posterior midpoint of hypural plate; **(18)** Anterior insertion point of anal fin; **(19)** Anterior insertion point of ventral fin; **(20)** Anterior insertion point of pectoral fin; **(21)** Lateral line - perpendicular to 13; **(22)** Lateral line - perpendicular to 18.

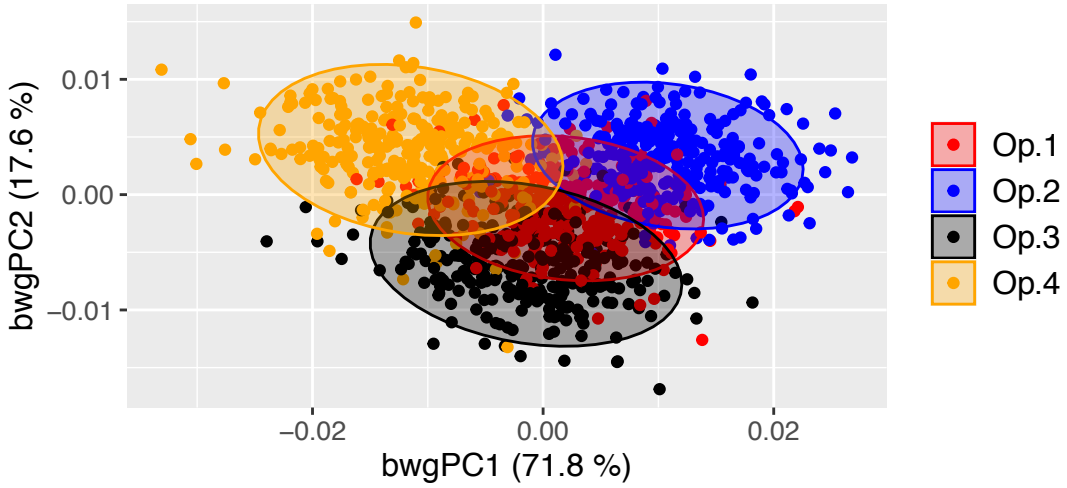

Figure 2. Between-operator PCA scatterplot showing the cross-validated scores along the first two between-group principal components (bwgPCs). Dots represent individual Atlantic salmon (n = 289) landmarked by four independent operators (different colours). Between-operator variance (%) explained by the first and second axes is shown.

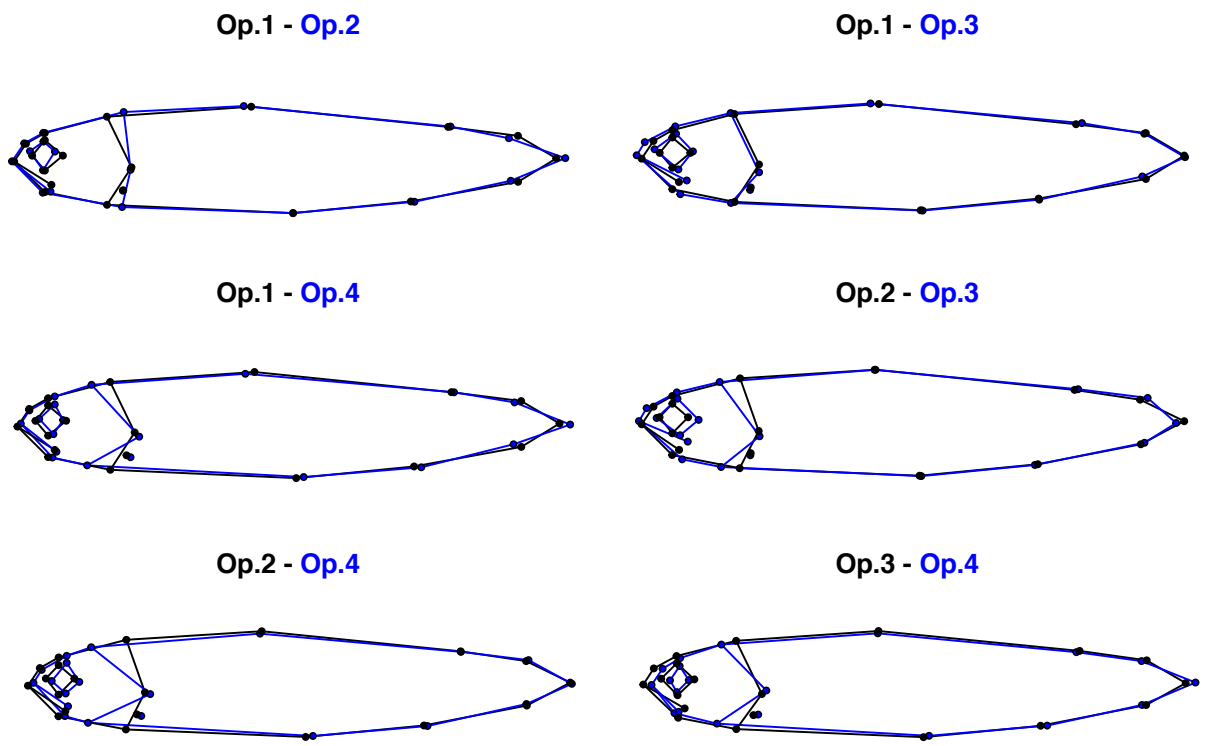

Figure 3. Pairwise comparisons of the mean body shape of 289 Atlantic salmon landmarked by four independent operators. Morphological differences were magnified five times to aid visualisation.

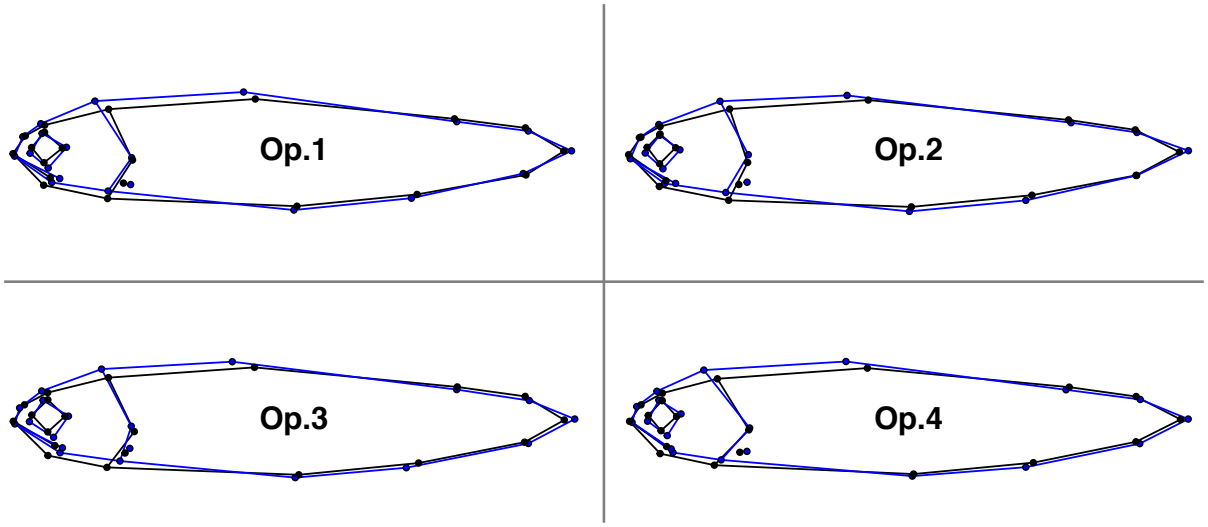

Figure 4. Comparisons of the mean body shape of 289 Atlantic salmon in the rivers Spey (black) and Oykel (blue) landmarked by four independent operators. Morphological differences were magnified six times to aid visualisation.

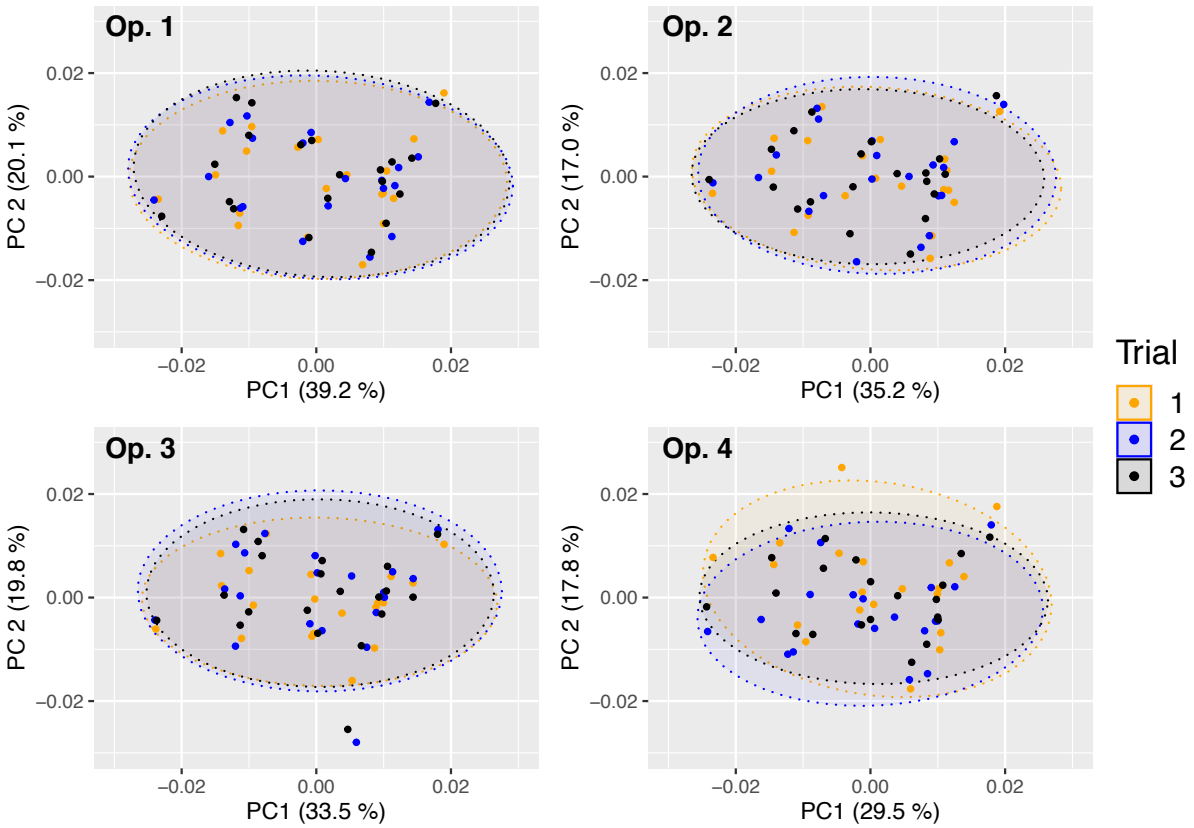

Figure 5. Principal components analysis scatterplots of Procrustes-aligned coordinates for 20 Atlantic salmon in three landmarking trials by four independent operators. Dots represent individual fish. Variance (%) explained by the first and second axes and 95% confidence ellipses are shown.

Tables

Table 1. Pairwise comparisons of the body shape of 289 Atlantic salmon landmarked by four independent operators based on Hotelling's T^2.

Comparison	Euclidean dist.	Hotelling's T^2	F	P-value
Op.1 vs. Op.2	0.01249	6687.53	163.19	$< 1 \times 10^{-6}$
Op.1 vs. Op.3	0.00992	7394.42	180.44	$< 1 \times 10^{-6}$
Op.1 vs. Op.4	0.01569	6372.29	155.50	$< 1 \times 10^{-6}$
Op.2 vs. Op.3	0.01575	12100.50	295.28	$< 1 \times 10^{-6}$
Op.2 vs. Op.4	0.02270	7484.90	182.65	$< 1 \times 10^{-6}$
Op.3 vs. Op.4	0.01425	6717.82	163.93	$< 1 \times 10^{-6}$

Table 2. Procrustes ANOVA summary statistics of effect of river-of-origin on the body shape of 289 Atlantic salmon landmarked by four independent operators.

Operator	Df	SS	r^2	*F*	*Z*	*P*-value
Op.1	1	0.002760	0.02928	8.6573	4.832	0.0001
Op.2	1	0.003487	0.03363	9.9865	5.5316	0.0001
Op.3	1	0.004730	0.04228	12.671	5.8035	0.0001
Op.4	1	0.004028	0.03465	10.302	5.8474	0.0001

Df = Degrees of freedom, SS = Sum of squares, *F* = *F* statistics, *Z* = Effect size

Table 3. Pairwise tests of angles between body shape differences among rivers and operators. Measurements of angles (degrees) between bwgPC1 vectors (below the diagonal) and *P*-values (above the diagonal) are shown. Significant *P*-values (in bold) indicate that shape change vectors are similar to each other. Additional comparisons with bwgPC2 and bwgPC3 vectors of shape difference among operators are reported in Supplementary Table 3.

	Op.1 (Spey vs. Oykel)	**Op.2** (Spey vs. Oykel)	**Op.3** (Spey vs. Oykel)	**Op.4** (Spey vs. Oykel)	**bwgPC1** (among operators)
Op.1 (Spey vs. Oykel)		**2.0×10^{-16}**	**5.5×10^{-12}**	**4.8×10^{-15}**	**0.028**
Op.2 (Spey vs. Oykel)	25.1°		**2.2×10^{-12}**	**2.5×10^{-13}**	**0.037**
Op.3 (Spey vs. Oykel)	33.4°	32.5°		**7.0×10^{-20}**	0.288
Op.4 (Spey vs. Oykel)	27.4 °	30.5°	20.2°		0.168
bwgPC1 (among operators)	72.5	73.7	84.8	81.1	

Table 4. Estimated mean and median shape distance (with confidence intervals) between the rivers Spey and Oykel obtained by each operator.

Operator	Mean distance	Median distance	Lower CI extreme	Upper CI extreme
Op.1	0.00651	0.00649	0.00526	0.00783
Op.2	0.00726	0.00722	0.00593	0.00871
Op.3	0.00834	0.00832	0.00688	0.00978
Op.4	0.00782	0.00781	0.00649	0.00921

Table 5. Procrustes ANOVA summary statistics of effect of river-of-origin on the body shape of 289 Atlantic salmon based on combined datasets of Op.2 and Op.4.

Dataset	Df	SS	r^2	F	Z	P-value
Op.2-Oykel Op.4-Spey	1	0.024607	0.18193	63.825	7.4748	0.0001
Op.2-Spey Op.4-Oykel	1	0.040397	0.26930	105.77	7.0742	0.0001

Df = Degrees of freedom, SS = Sum of squares, F = F statistics, Z = Effect size

Table 6. Pairwise comparisons of the body shape of 20 Atlantic salmon in three landmarking trials by four independent operators.

Operator	Trials	Euclidean dist.	Hotelling's T^2	F	P-value
Op.1	1 vs. 2	0.00262	69.11016	0.1914	0.97
	1 vs. 3	0.00353	129.63507	0.3591	0.89
	2 vs. 3	0.00167	20.58643	0.0570	0.99
Op.2	1 vs. 2	0.00394	42.34704	0.1173	0.99
	1 vs. 3	0.00407	60.60636	0.1679	0.98
	2 vs. 3	0.00245	51.99629	0.1440	0.98
Op.3	1 vs. 2	0.00389	50.95239	0.1411	0.98
	1 vs. 3	0.00405	59.86808	0.1658	0.98
	2 vs. 3	0.00158	17.94435	0.0497	0.99
Op.4	1 vs. 2	0.01339	197.45520	0.5470	0.81
	1 vs. 3	0.01026	166.04592	0.4600	0.84
	2 vs. 3	0.00457	71.89377	0.1992	0.96

Table 7. Repeatability values for the three landmarking trials on 20 Atlantic salmon by four independent operators.

Operator	Repeatability	Procrustes ANOVA r^2 (%)
Op.1	0.975	94.9
Op.2	0.955	91.3
Op.3	0.947	89.9
Op.4	0.892	81.5