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

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 observer 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 observers that applied an identical landmarking protocol to replicate photographs of 291 live Atlantic salmon (*Salmo salar* L.) from two rivers. Using Hotelling's T^2 analyses, we found significant inter-observer differences in mean body shape, suggesting that the observers introduced a systematic error despite following the same landmarking protocol. No significant differences were detected when the landmarking process was repeated by the same observer on a random subset of photographs. Importantly, in spite of significant observer bias, small but statistically significant morphological differences between fish from the two rivers were found consistently by all observers and combined datasets of two separate observers. Pairwise tests of angles of vectors of shape change showed that these between-river differences in body

shape were analogous across observer datasets suggesting a general reproducibility of findings obtained by geometric morphometric studies. Overall, we show that, even when significant inter-observer error is introduced during digitisation, following an identical landmarking protocol can identify biological differences between populations. This study indicates that data sharing across research groups can yield reliable results and therefore has important implications for the future development of data sharing in geometric morphometrics.

Introduction

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 during the course of landmark digitisation (Arnqvist & Mårtensson, 1998; Muir, Vecsei & Krueger, 2012; Fruciano et al., 2020; Fox, Veneracion & Blois, 2020). Indeed, the so-called inter-observer error during landmarking has been found to be one of the most critical factors affecting GM analyses because different observers 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-observer error can be substantial and potentially obscure biological variation making data sharing and comparing of landmarked datasets difficult. As an example, Shearer et al. (2017) found that morphological variation between two observers was comparable to shape differences between macaque species.

Intra-observer error has also been shown to significantly affect GM analyses (Wilson, Cardoso & Humphrey, 2011; Fox, Veneracion & Blois, 2020). Intra-observer 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-observer error is invariably modest compared to inter-observer 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 observers (but see Engelkes et al., 2019).

The degree and impacts of observer 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, animal 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). Thus further research is needed to understand the inter- and intra-observer error during the landmark digitisation process for photographs of live animals.

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 a 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 et al., 2016; Chang & Alfaro, 2016). Such studies, involving several individuals collecting shape data, can potentially address key questions of 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 consequence of inter-individual observer error when sharing datasets remains 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 observers digitised the same photographs of sedated Atlantic salmon (*Salmo salar* L.) sampled in two different rivers, following a shared landmarking protocol. The shape data and results obtained by the four observers were then compared and contrasted to assess the magnitude of inter- and intra-observer error to infer the potential for meaningful data sharing. The results will have significant implications for open science and reproducibility in GM studies.

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 anaesthesia before being tagged and released to the river after recovery. Photographs from the left side of each fish were taken with a Fujifilm 16613354 FinePix XP140 Compact Digital Camera with fish on a background reference scale. 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).

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 observers were blinded to river-of-origin of the specimens. Twenty-two landmarks were digitised on each image by four independent observers (Ob.1, Ob.2, Ob.3 and Ob.4) using tpsDig v.2.31 (Rohlf, 2017) and following an identical protocol (Fig. 1). Landmark positions were chosen according to the ones 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, 20 fish from both rivers were randomly selected from the dataset and consecutively landmarked a further two times (i.e. three times in total) by each observer to evaluate the within-individual observer consistency in digitisation.

Landmark coordinates from all observers 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 observers (2) any morphological differences between salmon from different rivers were detected by all observers; (3) identified between-river differences were consistent across all observers; (4) divergent datasets from different observers could be merged (5) the magnitude of intra-observer error was similar across observers.

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 bending of fish body 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. 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 observer using the *plotOutliers* function in 'geomorph'. Two specimens digitised by one observer were found to be very different from the other individuals and were therefore removed from the dataset of all four observers, 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 observers?

To investigate whether results produced by a single observer are accurate and reproducible, we tested differences in the mean body shapes of fish digitised by independent observers. First, a Procrustes ANOVA (*procD.lm* function in 'geomorph') using 'specimen' and 'observer' as factors and a randomised residual permutation procedure (10,000 iterations) was performed to determine the sources of morphological variation.

A between-group PCA (Boulesteix, 2004) was then computed to explore variations between the four observers. 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 observer who digitised them.

To investigate whether landmarking by multiple observers has 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 observers 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 observers, and the PC scores of each observer were then used in a Hotelling's T^2 test as an approximation of shape.

Test 2. Are there morphological differences between salmon from different rivers detected by all observers?

We next tested whether there was a difference in body shape between rivers, and whether the observers were consistent in identifying any differences. For each observer, separate PCAs were performed on Procrustes-aligned coordinates. The effect of fish size on body shape was

tested using Procrustes ANOVAs, with Procrustes coordinates used as an outcome variable and the log value of centroid size as an independent variable. A small but significant effect of size on shape was found for all observers (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. Procrustes ANOVAs were 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. Are identified between-river differences consistent across all observers?

To assess if body shape differences between rivers were analogous across observers, pairwise tests of angles (Li, 2011) of vectors of shape change between fish from the Spey and Oykel Rivers 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 angles among the PC1 vectors (used as an approximation of shape difference) obtained by the between-group PCA separating fish from Spey and Oykel Rivers to test if they followed the same "direction", i.e. if the shape differences between rivers were approximately the same for all observers.

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

The two observers showing the most dissimilar mean shape 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 protocol. Inter-observer analysis showed Ob.2 and Ob.4 produced the most different body shapes (greatest Euclidean distance), so from these, two datasets were created: one including shape data from the River Oykel digitised by Ob.2 (herein Ob.2-Oykel) and the River Spey data digitised by Ob.4 (herein Ob.4-Spey) and vice versa, i.e. the River Spey digitised by Ob.2 (herein Ob.2-Spey) and the River Oykel data digitised by Ob.4 (herein Ob.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 PC1 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 the ones found when using the complete datasets of all four observers including both rivers (see section above). This enabled us to test whether any between-river differences as a result of different observers outweighed any biological differences between rivers found when using the complete within observer datasets.

Test 5. Quantifying intra-observer error

A GPA was computed separately on landmark coordinates obtained by each observer re-digitising a sub-sample of twenty fish. Individual consistency in landmarking was then

investigated using PCA and tested using Hotelling's T^2. To test for the differences in mean body shapes between digitisation trials, a PCA was performed on the Procrustes-aligned coordinates of each observer separately and the PC scores of each trial were then used in a Hotelling's T^2 test as an approximation of shape.

Repeatability among digitisation trials was also calculated for each observer using the intraclass correlation coefficient (Fisher, 1958). A one-way Procrustes ANOVA (*procD.lm* function in 'geomorph') 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 twenty 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).

Results

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

Despite digitising replicate photographs with homologous landmarks, Procrustes ANOVA of corrected coordinates showed that 19.5% of the morphological variation within the dataset was explained by the observers, while 64.7% was attributable to variation between fish (Supplementary Table 1). Specimens were correctly assigned to their observer based on body shape with 82.6% accuracy by the exploratory between-group PCA (Fig. 2).

There was a significant observer effect on mean body shape, with all pairwise tests displaying highly significant differences between observers (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 observers, with the smallest difference (0.00992) occurring between Ob.1 and Ob.3 and the greatest between Ob.2 and Ob.4 (0.02270). The PCA scatterplot (Fig. 2) broadly reflected these results along axis 1 (71.8% of variance), with Ob.1 and Ob.3 overlapping extensively and Ob.2 and Ob.4 being furthest apart.

The anatomical differences among observers were concentrated mainly on the head (Fig. 3, Supplementary Fig. 1), with major areas of disagreement being snout, eye, mouth and posterior of cranium. Morphological dissimilarities were more or less pronounced depending on the observer comparisons (Fig. 3).

Test 2. Are there morphological differences between salmon from different rivers detected by all observers?

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 fish from different rivers for all observers (72.2% mean classification success rate). The fish from the River Oykel had a greater body depth, more pronounced caudal peduncle, larger eye and longer mouth and more pointed snout than those from the River Spey (Fig. 4).

Test 3. Are identified between-river differences consistent across all observers?

All comparisons of the "direction" of body shape variation between rivers were significant for all observers, meaning that the way in which shape differed between the rivers Spey and Oykel was approximately the same for all observers (Table 3).

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

There were significant differences in body shape between fish from the rivers Spey and Oykel digitised separately by Ob.2 and Ob.4 (Table 4), with between-group PCA separating fish from different rivers with a high accuracy (>90% classification success rate for both comparisons, Supplementary Table 3).

The comparison of the "direction" of between-river body shape variation detected by the merged datasets from Ob.2 and Ob.4 was highly significant (P-values < 0.0001; Supplementary Table 4), meaning that the way in which shape differed between the rivers Spey and Oykel was approximately the same regardless of the selected dataset. Similar shape changes were also found when comparing the two Ob.2 and Ob.4 merged datasets with the complete within observer datasets including both rivers (P-values < 0.013, Supplementary Table 4), indicating that the river differences detected by combined and individual observer datasets were sufficiently similar.

Test 5. Quantifying intra-observer error

There was extensive overlap among landmarking trials, suggesting a high consistency in digitisation across all observers (Fig. 5). Pairwise comparisons supported this since none of the average body shapes differed significantly between repeated digitisations (P-values > 0.81; Table 5). All four observers 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 observers (0.892-0.975), indicating high landmarking precision (Table 6).

Discussion

We show here that independent observers applying an identical landmarking protocol to replicate photographs of live Atlantic salmon specimens taken in field conditions yielded significantly different mean body shapes (Test 1). However, these differences were not

sufficient to mask small morphological differences between geographically separate populations (Test 2), which were consistent across all observers (Test 3). The same shape differences between the two rivers were also found when datasets from the most dissimilar observers (Ob.2 and Ob.4) were merged (Test 4). Furthermore, intra-observer error calculated on a subset of samples was minimal suggesting that it did not have a significant influence on body shape results obtained by the different observers (Test 5).

Inter-observer error explained 19.5% of the morphological variation in the dataset. This is line with previous studies that identified the inter-observer effect 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). Despite digitising replicate photographs with homologous landmarks, all the observers produced significantly different mean body shapes. The high rate (82.6%) of specimens assigned to the correct observer by the between-group PCA suggests that the observers introduced a systematic error, which created four identifiable body shapes despite following the same landmarking protocol. This digitisation bias is likely to have been introduced by observers consistently applying personal, fine-scaled landmarking rules in addition to the general protocol. The fact that the differences among observers 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 choice 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).

In GM studies, digitisation is typically performed by a single observer, and it is usually not possible to assess if an identified morphological variation is biologically meaningful or simply an artefact attributable to observer bias. This may undermine the reliability of results 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 observers did not mask small, but significant morphological differences between fish from the rivers Spey and Oykel, which were consistent across observers. The fact that, as shown by tests of angles, all the observers detected analogous between-river differences, strongly suggest that these were biologically authentic. Similarly, Fruciano et al. (2020) found preservation methods significantly affected the body shape of brown trout (*Salmo trutta* L.), but the downstream between-groups classification was similar regardless of preservation method.

Importantly, similar between-river differences in shape were also found when comparing landmark data of fish from the rivers Spey and Oykel digitised by two distinct observers (Ob.2 and Ob.4) using the same protocol, i.e. the fish shape data were analogous to the results from a single observer landmarking all fish from both rivers. This strongly suggests an exchangeability of these distinct sets of shape data to detect between-river differences. These

findings have relevant implications in terms of GM data sharing and pooling and indicate that, although the magnitude and direction of inter-observer error is likely dependent on study characteristics, data exchange based on a common landmarking protocol seems to be a promising possibility. Specifically, these results suggest that applying a public reference landmarking protocol to all GM studies investigating a specific taxon or structure would enable data exchange among researchers without affecting the robustness of results. This is also reassuring considering the growing number of online GM repositories where images and shape data are exchanged among researchers (Thomas, Bright & Cooney et al., 2016; Evin, Bonhomme & Claude, 2020)

In contrast to the inter-observer effects described in this study, we found no statistical evidence of intra-observer effects on the quantification of fish morphology. On the contrary, we show a very high level of repeatability across trials for all observers tested. This corroborates previous studies that showed intra-observer 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 observers, 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 observers 'remembering' landmarking choices of trial 2 during trial 3.

Conclusions

Overall, we show that, even when significant inter-observer error is introduced through digitisation, following an identical landmarking protocol is an effective tool to obtain robust and reliable results. This implies that GM studies based on common protocols are potentially reproducible, even when analyses are based on images of live specimens taken in the field, as in the current study. Nevertheless, since observer error can vary between studies and the magnitude and effect are impossible to determine *a priori*, we recommend assessing the effect of landmarking error by using multiple observers for a subset of samples as presented here in order to be more confident about the study results. Future works should aim to provide standardised and public taxon-specific landmarking protocols which could be used by all researchers performing GM studies in the same field and promote data exchange and large-scale studies. For example, the protocol and code [http://dx.doi.org/10.17605/OSF.IO/ZVU5P] presented here provide a starting point for ensuring consistency between researchers carrying out GM studies on Atlantic salmon between rivers. Once this has been achieved, new data can be added and compared in the knowledge that an initial standardisation process maximises chances of success. This study will provide important implications for the future development of data sharing in GM.

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.

Funding

This research was part of the Missing Salmon Project funded by the Atlantic Salmon Trust. PM was supported by the Leeds-York-Hull Natural Environment Research Council (NERC) Doctoral Training Partnership (DTP) Panorama under grant NE/S007458/1.

Author contributions

The field element of the study was executed by MN, HMH and AJL. The geometric morphometric component was designed by DAJ, JDB, PM, ADN, JRR and CEA and shape data collection was carried out by PM, JRR, PKW and RS. Data analysis was conducted by PM. All authors contributed to data interpretation, manuscript preparation and editing.

Data availability

Supplementary material, raw data and scripts are available at:
http://dx.doi.org/10.17605/OSF.IO/ZVU5P

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.

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.

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.

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/B978-012778460-1/50003-X

Figures

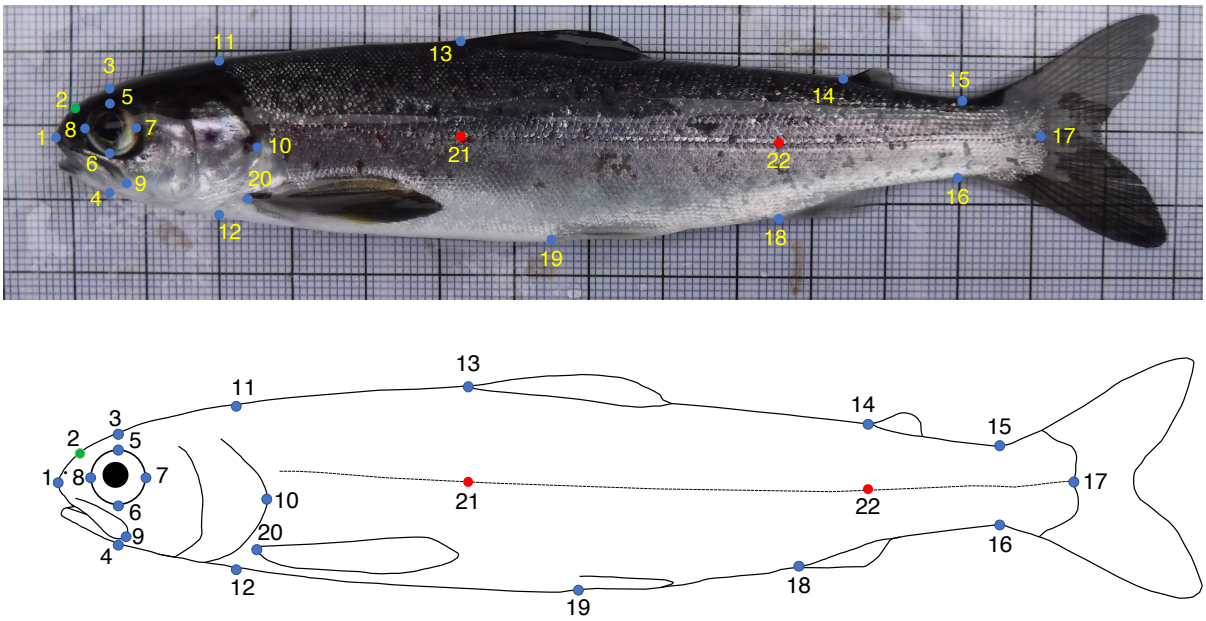

Figure 1. Landmarks used for the geometric morphometrics analyses of Atlantic salmon smolts. **(1)** Tip of snout; **(2)** Midpoint between 1 and 3 (automatic semi landmark); **(3)** Directly above middle of eye; **(4)** Perpendicular to 3; **(5)** Midpoint of top of eye (directly below 3); **(6)** Midpoint 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.

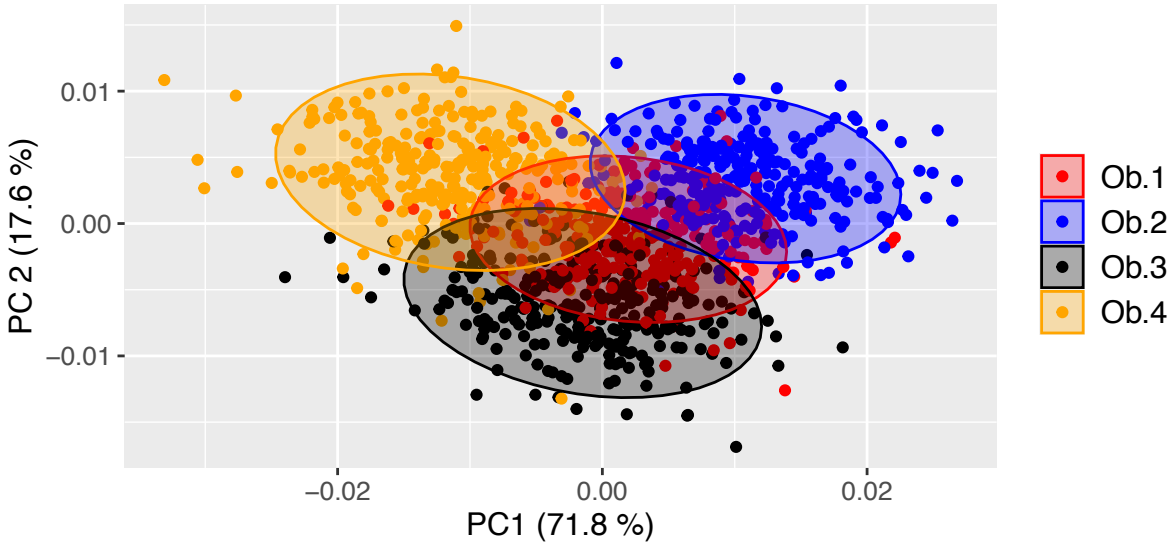

Figure 2. Between-observer PCA biplot of Procrustes-aligned coordinates for 289 Atlantic salmon (dots) landmarked by four independent observers (different colours). Variance (%) explained by the first and second axes is shown.

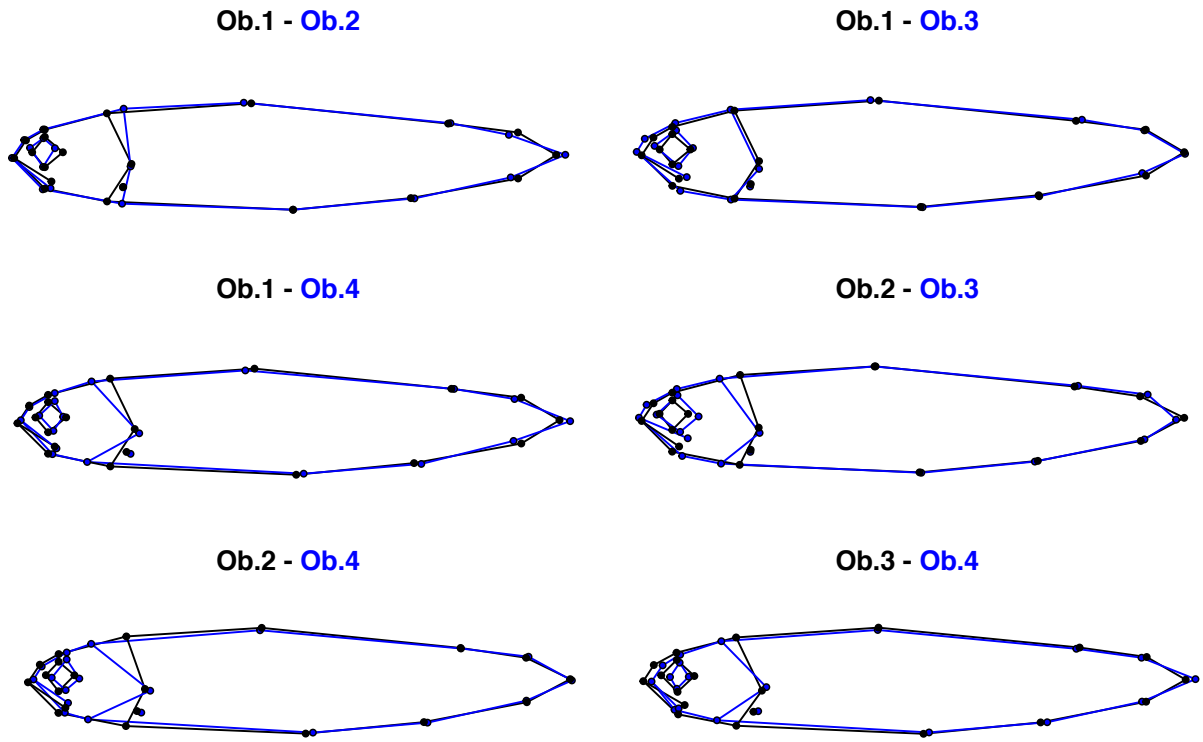

Figure 3. Pairwise comparisons of the mean body shape of 289 Atlantic salmon landmarked by four independent observers. Morphological differences were magnified five times to aid visualisation.

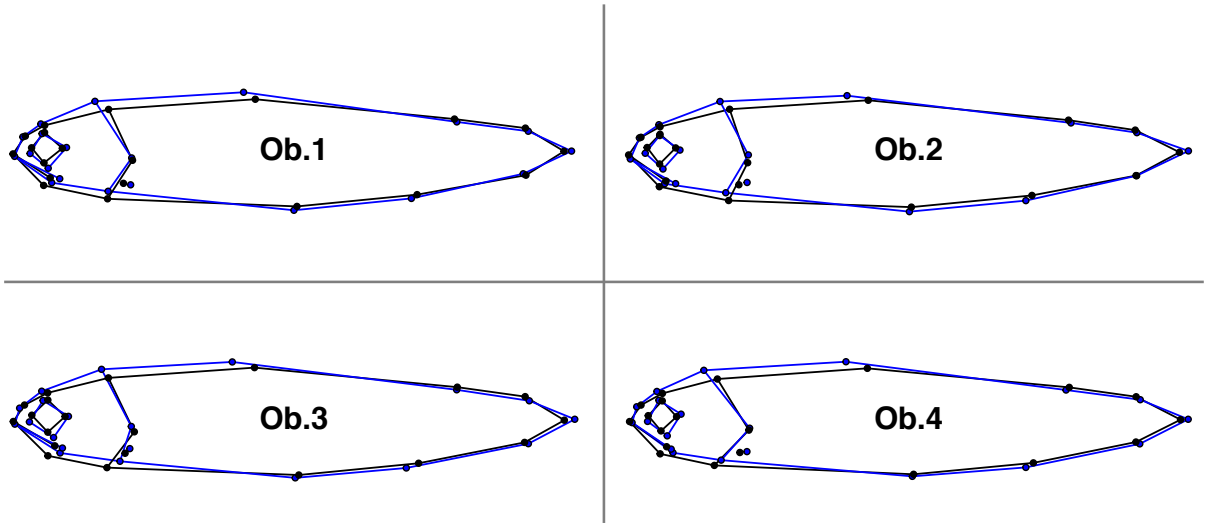

Figure 4. Comparisons of the mean body shape of 289 Atlantic salmon in the rivers Spey (black) and Oykel (blue) landmarked by four independent observers. Morphological differences were magnified six times to aid visualisation.

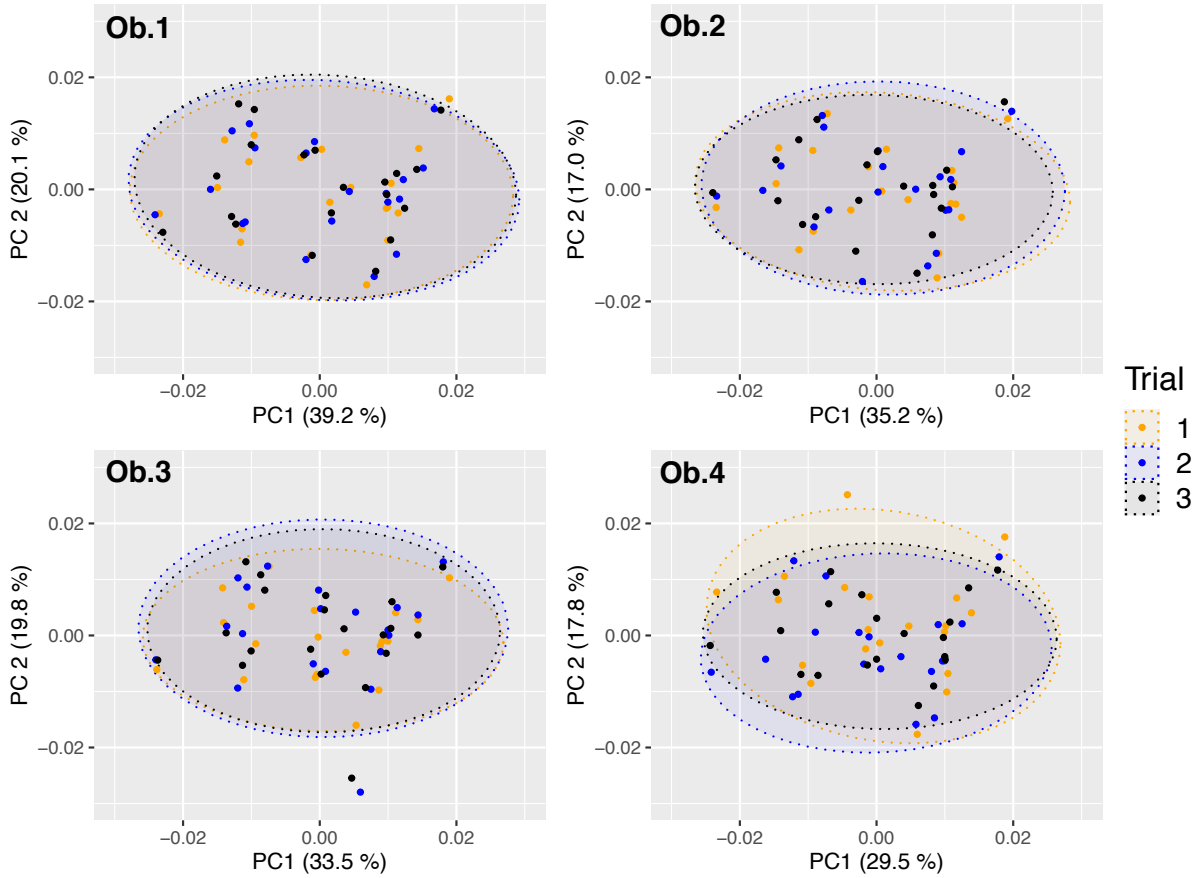

Figure 5. Principal components analysis biplots of Procrustes-aligned coordinates for 20 Atlantic salmon in three landmarking trials by four independent observers. Dots represent individual fish. Variance (%) explained by the first and second axes is shown.

Tables

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

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

Observer	Df	SS	r^2	F	Z	P-value
Ob.1	1	0.002760	0.02928	8.6573	4.832	0.0001
Ob.2	1	0.003487	0.03363	9.9865	5.5316	0.0001
Ob.3	1	0.004730	0.04228	12.671	5.8035	0.0001
Ob.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 detected by four independent observers. Measurements of angles (degrees) between PC1 vectors (below the diagonal) and P-values (above the diagonal) are shown. Significant P-values indicate that shape change vectors are similar to each other.

	Ob.1 (Spey vs. Oykel)	**Ob.2** (Spey vs. Oykel)	**Ob.3** (Spey vs. Oykel)	**Ob.4** (Spey vs. Oykel)
Ob.1 (Spey vs. Oykel)		4.1×10^{-97}	1.0×10^{-70}	7.4×10^{-72}
Ob.2 (Spey vs. Oykel)	27.9°		3.5×10^{-76}	1.0×10^{-86}
Ob.3 (Spey vs. Oykel)	35.3°	33.6°		9.7×10^{-91}
Ob.4 (Spey vs. Oykel)	35.0 °	30.6°	29.5°	

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

Dataset	Df	SS	r^2	F	Z	P-value
Ob.2-Oykel Ob.4-Spey	1	0.024607	0.18193	63.825	7.4748	0.0001
Ob.2-Spey Ob.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 5. Pairwise comparisons of the body shape of 20 Atlantic salmon in three landmarking trials by four independent observers.

Observer	Trials	Euclidean dist.	Hotelling's T^2	F	P-value
Ob.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
Ob.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
Ob.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
Ob.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 6. Repeatability values for the three landmarking trials on 20 Atlantic salmon by four independent observers.

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