Marine and Freshwater Research
https://doi.org/10.1071/MF21169

Supplementary material

Genetic diversity and restricted genetic connectivity in an endangered marine fish (*Brachionichthys hirsutus*) provides a model for conservation management in related and data-deficient species

Sharon A. Appleyard[A,D], *Tim P. Lynch*[B], *Mark A. Green*[B] *and Francisco Encinas-Viso*[C]

[A]CSIRO Australian National Fish Collection, National Research Collections Australia, Castray Esplanade, Hobart, Tas. 7004, Australia.

[B]CSIRO Ocean and Atmosphere, Castray Esplanade, Hobart, Tas. 7004, Australia.

[C]CSIRO Australian National Herbarium, National Research Collections Australia, Clunies Ross Street, Canberra, ACT 2601, Australia.

[D]Corresponding author. Email: sharon.appleyard@csiro.au

Supplementary material for spotted handfish sampling, SNP processing, filtering and population analyses

Owing to the endangered nature of spotted handfish, and as the sampling was opportunistic, the sample numbers for each 'collection' did not reach a preferred sampling size of $N = 30$ per collection (7/8 collections where $N < 40$). Fig. S1 below shows the range in total length of individuals from each collection, with the HW2006 individuals the smallest in size.

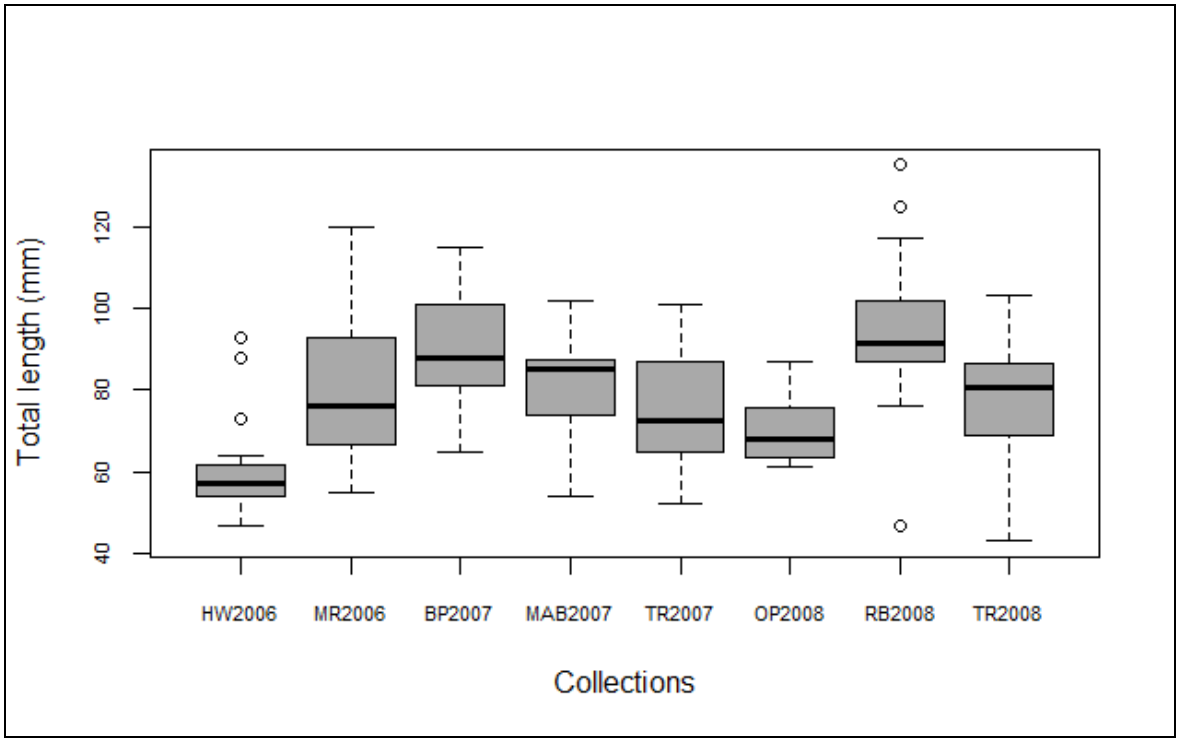

Fig. S1. Average total lengths of spotted handfish sampled for genomic analyses.

SNP processing

Genomic DNA from the spotted handfish individuals was sent at room temperature to AGRF in Melbourne, Australia for SNP genotyping (by their in-house Genotype-by-Sequencing (GBS) service). The AGRF in-house library preparation and GBS included (and as outlined in Lynch *et al*. 2020):

- DNA digestions with two restriction enzymes (*Eco*RI and *MspI*; determined from a GBS establishment service);
- ligation of barcoded adapters;
- size selection of pooled digested-ligated fragments;
- amplification of libraries by PCR using indexed primers;
- sequencing on an Illumina NextSeq platform flow cell (Illumina Inc., USA) with 150 cycles in MID-output mode according to their in-house GBS methodology

AGRF then processed the raw reads using their in-house bioinformatic pipeline which included:

- raw sequences were demultiplexed, checked for read quality and restriction site presence and trimmed; RAD-tags were analysed with Stacks software (ver. 1.47, http://catchenlab.life.illinois.edu/stacks/; Catchen *et al*. 2011, 2013) resulting in a separate FASTQ file for each sample) (using 'process_radtags' in Stacks);
- sequence reads were aligned into matching stacks/tags from which loci were formed and SNPs are detected (using 'ustacks','cstacks' 'sstacks' in Stacks);
- parameters used to define a 'stack' and resulting subsequent SNPs for each individual from the catalogue included: a minimum depth coverage of two to create a stack; disabling haplotype calls from secondary reads; one mismatch allowed between sample tags when generating the catalogue; a minimum of five reads to call a homozygous genotype and a heterozygote was called when the frequency of the minor allele in a stack was <0.1 across the entire dataset and
- AGRF provided the post processed SNPs as raw and unfiltered SNP output in a variant call format (VCF) file.

High performance computing SNP filtering

Individuals in the VCF file were renamed using bcftools reheader (ver. 1.10, http://samtools.github.io/bcftools/bcftools.html#reheader; Li *et al*. 2009) and filtered initially using VCFtools (ver. 0.1.14, https://github.com/vcftools/vcftools; Danecek *et al*. 2011) using the CSIRO high performance computing platform, with an initial high-level filtering undertaken by treating all individuals as belonging to one group. This filtering removed:

- sites whose minor allele frequency was too low (as a result of sequencing or alignment errors);
- kept loci that were genotyped in at least 50% of individuals and
- where multiple SNPS were detected on the same fragment, a single SNP was randomly chosen for the analyses to avoid linkage disequilibrium between SNPs

R filtering

The VCF file was further filtered and converted (i.e. before population genomic analyses) in R (ver. 3.5.1, R Foundation for Statistical Computing, Vienna, Austria, see https://www.R-project.org/) using R-Studio (ver. 1.1.463, RStudio, Inc., Boston, MA, USA, see http://www.rstudio.com/) with the following R packages: vcfR (ver. 1.8.0, see https://cran.r-project.org/web/packages/vcfR/index.html) and dartR (ver. 0.91, see https://CRAN.R-project.org/package=dartR) (Knaus and Grünwald 2017; Gruber *et al*. 2018). This consisted of:

• filtering out loci that were monomorphic in all collections;

• filtering on a call rate per individual and collection of > 0.85;

• ensuring loci with a maf > 0.025 were used and utilising loci that were polymorphic in at least one collection and

• using loci in Hardy–Weinberg Equilibrium (HWE) across all collections individually

Population genetic analyses – diversity, proximity and structure

In R (ver. 3.5.1) and RStudio (ver. 1.1.463), the following population genetics packages were used to analyse the SNP loci per collection:

• adegenet (ver. 2.1.1, see https://cran.r-project.org/web/packages/adegenet/index.html; Jombart 2008; Jombart and Ahmed 2011);

• diveRsity (ver. 1.9.90, see https://cran.r-project.org/web/packages/diveRsity/index.html; Keenan *et al*. 2013);

• hierfstat (ver. 0.04.22, see https://cran.r-project.org/web/packages/hierfstat/index.html; Goudet 2005) and

• genepop (ver. 1.1.2, see https://rdrr.io/cran/genepop/; Raymond and Rousset 1995; Rousset 2008).

Stand-alone versions of PGDSpider (ver. 2.1.1.5, see http://www.cmpg.unibe.ch/software/PGDSpider/; Lischer and Excoffier 2012) and Arlequin (ver. 3.5, see http://cmpg.unibe.ch/software/arlequin35; Excoffier *et al*. 2005; Excoffier and Lischer 2010) were used for additional file conversions, genetic diversity analyses and AMOVA (Analysis of Molecular Variance). AMOVA, based on genotypes and pairwise squared Euclidean distances between individuals, was undertaken to assess the overall structure of the spotted handfish populations. This method detected collection differentiation based on covariance components (leading to Φ statistics) corresponding to different hierarchical levels. We considered these Φ statistics analogous to *F*-statistics (i.e. a measure of the correlation between genes drawn at different hierarchical levels in collections (Wright 1949). Pair-wise collection differentiation estimates (based on F_{ST} (Wright 1949) and based on the Weir and Cockerham 1984 implementation) were also undertaken in Arlequin. F_{ST} values range from 0 to 1, with high F_{ST} values implying considerable differentiation among collections. Significance for all tests was assessed following 10 000 permutations and *P*-values for each pairwise comparison were corrected following the conservative Benjamini and Yekutieli (2001) (false discovery rate correction) approach.

We tested for Isolation by Distance (IBD) by a redundancy analysis (RDA) (as implemented in vegan, ver. 2.5–7, J. Oksanen, F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs, and H. Wagner, see https://cran.r-project.org/web/packages/vegan/index.html). The latitude and longitude (in decimal degrees, from the Transverse Mercator projection in the GDA_1994 MGA Zone 55 coordinate system) for the mid-points of polygons representing each sampling site were used as inputs for calculating distances (km) between the sites (over water) (these geographic distances were the environmental or explanatory variables), while genetic distance was based on pairwise F_{ST} values as calculated in Arlequin. Significance of the RDA model was based on 10 000 permutations This was a linear multivariate constrained ordination analysis with the RDA model showing 100% of the variation in genetic distance among the collections was explained by geographic distance (with adjusted $R^2 = 1.000$, $P < 0.05$) (see Fig. S2. and S3).

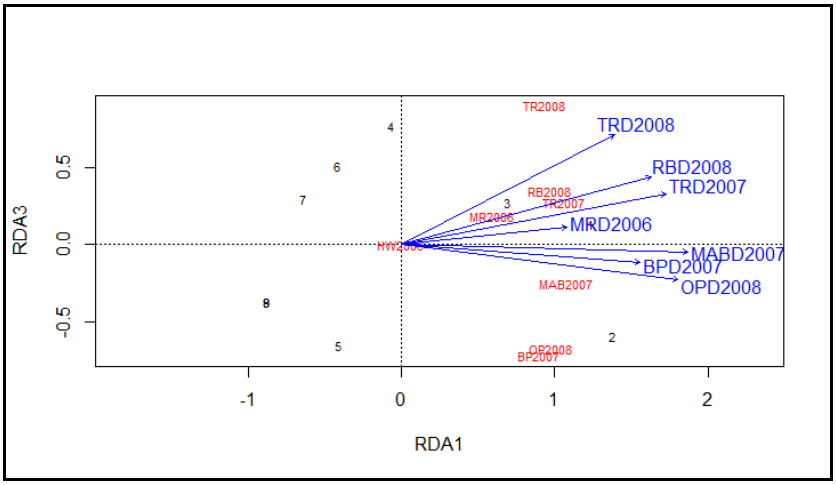

Fig. S2. Isolation by Distance testing for spotted handfish based on RDA analyses in vegan (ver. 2.5–7), where the linear model tested the regression between genetic distance (F_{ST} values; the response variable) and environmental data (i.e. the geographic distance; the explanatory variable). Symmetrical scaling (which scales the F_{ST} distance by the square root of the eigenvalues) was used.

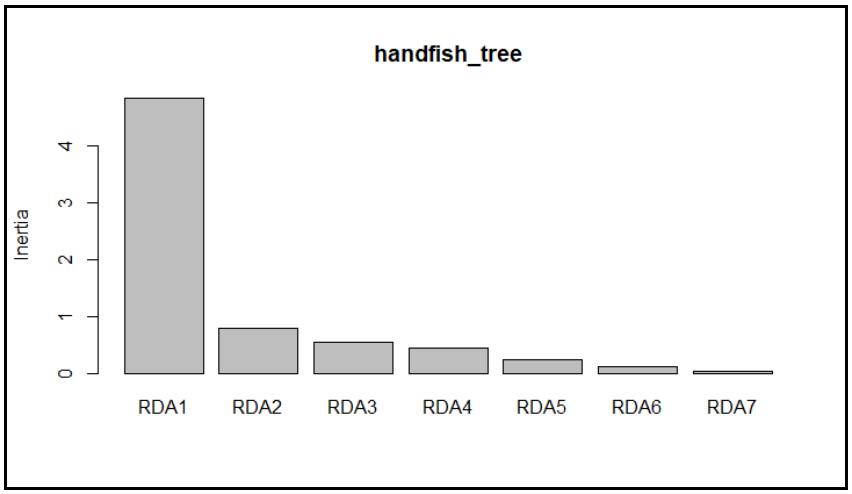

Fig. S3. RDA (model testing regression between spotted handfish genetic distance and geographic distances) screeplot from vegan (ver. 2.5–7) (eigenvalues for the constrained axes).

A Mantel (Mantel 1967) test (implemented in Arlequin, ver. 3.5; Excoffier *et al*. 2005; Excoffier and Lischer 2010) which also tested the proportional increase in genetic distance as geographic distance between sampling sites increased (as recommended by Legendre and Fortin 2010; Diniz-Filho *et al*. 2013; Legendre *et al*. 2015; Meirmans 2020) also confirmed the RDA-IBD outcomes (r = 0.449, *P* = 0.017).

Additionally, the number of genetic groups in the spotted handfish SNP dataset was estimated using Bayesian clustering algorithms implemented in STRUCTURE (ver. 2.3.4, see https://web.stanford.edu/group/pritchardlab/structure_software/release_versions/v2.3.4/html/structure.html) (Pritchard *et al*. 2000) run using admixture models (without *a priori* knowledge of location) with correlated allele frequencies with K (the number of clusters) set between 2 and 10; ten independent runs per K value were undertaken. The STRUCTURE algorithm inferred the proportion of ancestry from each cluster. Each run had a burn-in of 200 000 followed by 1 000 000 iterations. The simulations were implemented for computing parallelisation using the StrAuto python script (Chhatre and Emerson 2017).

We then used Structureharvester (ver. 0.6.94, see http://taylor0.biology.ucla.edu/structureHarvester/; Earl and Vonholdt 2012) to examine the STRUCTURE result and used the ΔK method to determine the optimal clustering level (Evanno *et al*. 2005). We processed the ten STRUCTURE runs using the Greedy algorithm in clumpp (ver. 1.1.2, see https://rosenberglab.stanford.edu/clumpp.html; Jakobsson and Rosenberg 2007) testing 1000 random input order repeats per K. We did not detect evidence of multimodality for the most likely K (2) considering native and introduced invaded ranges using clumpak (Kopelman *et al*. 2015) although additional likely sub-structuring was detected by Structureharvester, as indicated by the Delta K values > 2.5 (see Fig. S4). Finally, clustering analysis was visualised with distruct (ver. 1.1, see https://rosenberglab.stanford.edu/distruct.html; Rosenberg 2004).

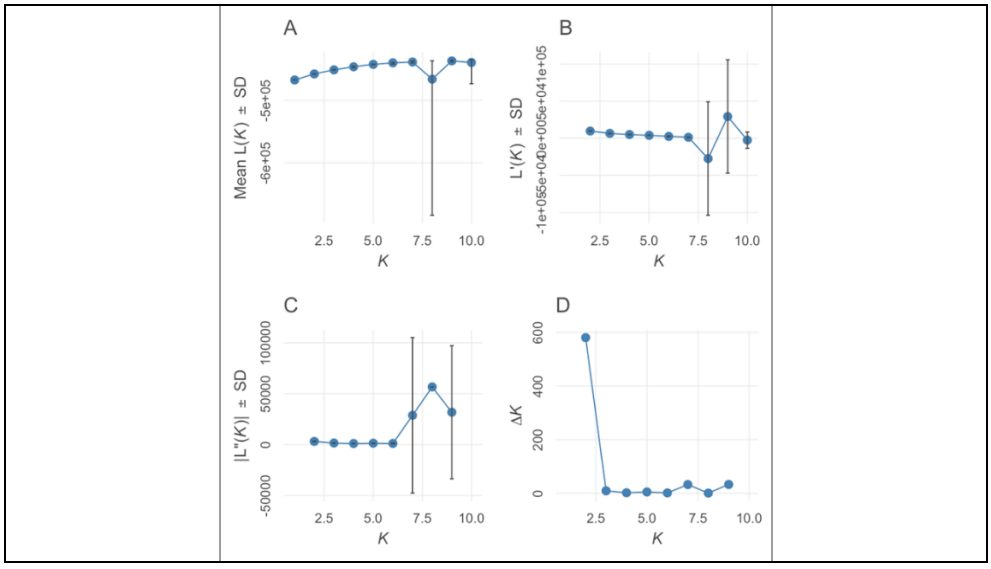

Fig. S4. Evanno output from Structureharvester, analyses based on 4172 SNPs in the eight spotted handfish (*B. hirsutus*) collections. Delta K is shown in panel D.

DAPC (Discriminant Analysis of Principal Components), a multivariate, non-model sequential method that identifies clusters of genetic variation maximised between clusters of individuals and minimised within clusters

(Jombart *et al*. 2010; Pritchard *et al*. 2000; Grünwald and Goss 2011) was also used. DAPC provided the determination of genetic clusters using synthetic variables (i.e. discriminant functions) and derived probabilities of membership (i.e. the genetic proximity of individuals to the different clusters) into different groups. Data were first transformed using principal components, with clusters then identified with discriminant analysis (without making assumptions of panmixia). As outlined by Jombart *et al*. (2010), one third of the principle components were retained in the current DAPC analysis so that discriminant functions were not overfitted.

We also estimated contemporary gene flow using Bayesian methods in the program BayesAss (ver. 3.0.4, see https://github.com/brannala/BA3/releases; Wilson and Rannala 2003) (Table S1). This analysis estimated the rates and direction of recent gene flow between the sampled locations. An initial run was conducted using the default parameters for allelic frequency (a), gene flow rate (m) and inbreeding (f). Delta values were modified in subsequent runs to ensure that proposed changes between chains at the end of the run were 20–40% (Wilson and Rannala 2003). Mixing parameters for subsequent runs were: $\Delta a = 0.5$, $\Delta m = 0.3$, $\Delta f = 0.7$. We performed five runs using different random seeds. Diagnostics of the MCMC output and convergence were analysed in the software Tracer (ver. 1.7.1, A. Rambaut, M. A. Suchard, D. Xie, and A. J. Drummond, see http://beast.bio.ed.ac.uk/Tracer).

Table S1. Contemporary gene flow estimations among eight collections of spotted handfish (*B. hirsutus*)

Mean (s.d.) dispersal rates (from five independent runs) estimated with BayesAss 3.0.4, based on 153 genotyped individuals and 4172 SNPs. The eight collections were coded as: [0] BP2007; [1] HW2006; [2] MAB2007; [3] MR2006; [4] OP2008; [5] RB2008; [6] TR2007; [7] TR2008. The coefficient m represents the proportion of immigrants in the collection

Collection	Direction	Mean (s.d.)	Collection	Direction	Mean (s.d.)
BP2007	m[0][0]:	0.733 (0.033)	OP2008	m[4][0]:	0.012 (0.011)
	m[0][1]:	0.083 (0.035)		m[4][1]:	0.012 (0.011)
	m[0][2]:	0.041 (0.026)		m[4][2]:	0.073 (0.026)
	m[0][3]:	0.041 (0.026)		m[4][3]:	0.012 (0.011)
	m[0][4]:	0.020 (0.019)		m[4][4]:	0.833 (0.033)
	m[0][5]:	0.037 (0.027)		m[4][5]:	0.031 (0.020)
	m[0][6]:	0.020 (0.019)		m[4][6]:	0.012 (0.011)
	m[0][7]:	0.020 (0.019)		m[4][7]:	0.012 (0.011)
HW2006	m[1][0]:	0.017 (0.016)	RB2008	m[5][0]:	0.006 (0.006)
	m[1][1]:	0.860 (0.036)		m[5][1]:	0.006 (0.006)
	m[1][2]:	0.017 (0.016)		m[5][2]:	0.027 (0.013)
	m[1][3]:	0.017 (0.016)		m[5][3]:	0.006 (0.006)
	m[1][4]:	0.017 (0.016)		m[5][4]:	0.013 (0.009)
	m[1][5]:	0.034 (0.023)		m[5][5]:	0.923 (0.020)
	m[1][6]:	0.017 (0.016)		m[5][6]:	0.007 (0.006)
	m[1][7]:	0.017 (0.016)		m[5][7]:	0.007 (0.006)
MAB2007	m[2][0]:	0.014 (0.013)	TR2007	m[6][0]:	0.009 (0.009)
	m[2][1]:	0.014 (0.013)		m[6][1]:	0.009 (0.008)
	m[2][2]:	0.803 (0.034)		m[6][2]:	0.018 (0.012)
	m[2][3]:	0.014 (0.013)		m[6][3]:	0.009 (0.009)

Collection	Direction	Mean (s.d.)	Collection	Direction	Mean (s.d.)
	m[2][4]:	0.101 (0.031)		m[6][4]:	0.009 (0.009)
	m[2][5]:	0.022 (0.019)		m[6][5]:	0.074 (0.022)
	m[2][6]:	0.014 (0.013)		m[6][6]:	0.789 (0.026)
	m[2][7]:	0.014 (0.013)		m[6][7]:	0.081 (0.023)
MR2006	m[3][0]:	0.015 (0.014)	TR2008	m[7][0]:	0.012 (0.012)
	m[3][1]:	0.015 (0.014)		m[7][1]:	0.013 (0.012)
	m[3][2]:	0.015 (0.014)		m[7][2]:	0.012 (0.012)
	m[3][3]:	0.879 (0.033)		m[7][3]:	0.012 (0.012)
	m[3][4]:	0.029 (0.019)		m[7][4]:	0.021 (0.016)
	m[3][5]:	0.015 (0.014)		m[7][5]:	0.051 (0.023)
	m[3][6]:	0.015 (0.014)		m[7][6]:	0.102 (0.029)
	m[3][7]:	0.015 (0.014)		m[7][7]:	0.773 (0.030)

References

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29**, 1165–1188 . doi:10.1214/aos/1013699998

Catchen, J. M., Amores, A., Hohenlohe, W. C., and Postlethwait, J. H. (2011). Stacks: building and genotyping loci *de novo* from short read sequences. *G3 – Genes Genomes Genetics* **1**, 171–182 doi:10.1534/g3.111.000240.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., and Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124–3140 . doi:10.1111/mec.12354

Chhatre, V. E., and Emerson, K. J. (2017). StrAuto: automation and parallelization of STRUCTURE analysis. *BMC Bioinformatics* **18**, 192 . doi:10.1186/s12859-017-1593-0

Danecek, P., Auton, A., Abecasis, G., Albers, C., Cornelis, A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 . doi:10.1093/bioinformatics/btr330

Diniz-Filho, J. A. F., Soares, T. N., Lima, J. S., Dobrovolski, R., Landeiro, V. L., Pires de Campos Telles, M., Rangel, T. F., and Bini, L. M. (2013). Mantel test in population genetics. *Genetics and Molecular Biology* **36**, 475–485 . doi:10.1590/S1415-47572013000400002

Earl, D. A., and Vonholdt, B. M. (2012). Structure Harvester: a website and program for visualising STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**, 359–361 . doi:10.1007/s12686-011-9548-7

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611–2620 . doi:10.1111/j.1365-294X.2005.02553.x

Excoffier, L., and Lischer, H. E. L. (2010). Arlequin suite ver. 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564–567 . doi:10.1111/j.1755-0998.2010.02847.x

Excoffier, L., Laval, G., and Schneider, S. (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* **1**, 47–50 . doi:10.1177/117693430500100003

Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical *F*-statistics. *Molecular Ecology Notes* **5**, 184–186 . doi:10.1111/j.1471-8286.2004.00828.x

Gruber, B., Unmack, P. J., Berry, O. F., and Georges, A. (2018). dartR: an R package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular Ecology Resources* **18**, 691–699 . doi:10.1111/1755-0998.12745

Grünwald, N. J., and Goss, E. M. (2011). Evolution and population genetics of exotic and re-emerging pathogens: novel tools and approaches. *Annual Review of Phytopathology* **49**, 249–267 . doi:10.1146/annurev-phyto-072910-095246

Jakobsson, M., and Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801–1806 . doi:10.1093/bioinformatics/btm233

Jombart, T. (2008). ADEGENET: an R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405 . doi:10.1093/bioinformatics/btn129

Jombart, T., and Ahmed, I. (2011). Adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071 . doi:10.1093/bioinformatics/btr521

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* **11**, 94 . doi:10.1186/1471-2156-11-94

Keenan, K., McGinnity, P., Cross, T. F., Crozier, W. W., and Prodohl, P. A. (2013). diveRsity: an R package for the estimation and exploration of population genetics parameters and their associated errors. *Methods in Ecology and Evolution* **4**, 782–788 . doi:10.1111/2041-210X.12067

Knaus, B. J., and Grünwald, N. J. (2017). VCFR: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources* **17**, 44–53 . doi:10.1111/1755-0998.12549

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., and Mayrose, I. (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources* **15**, 1179–1191 . doi:10.1111/1755-0998.12387

Legendre, P., and Fortin, M.-J. (2010). Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular Ecology Resources* **10**, 831–844 . doi:10.1111/j.1755-0998.2010.02866.x

Legendre, P., Fortin, M.-J., and Borcard, D. (2015). Should the Mantel test be used in spatial analysis? *Methods in Ecology and Evolution* **6**, 1239–1247 . doi:10.1111/2041-210X.12425

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 . doi:10.1093/bioinformatics/btp352

Lischer, H. E. L., and Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298–299 . doi:10.1093/bioinformatics/btr642

Lynch, T., Appleyard, S. A., Wong, L., Martini, A., Bessesll, T., Stuart-Smith, J., Soo, L., Faber, S., and Devine, C. (2020). Conservation of handfish and their habitats – annual report. Report to the National Environmental Science Program, Marine Biodiversity Hub, Canberra, ACT, Australia.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**, 209–220.

Meirmans, P. G. (2020). GENODIVE version 3.0: easy-to-use software for the analysis of genetic data of diploids and polyploids. *Molecular Ecology Resources* **20**, 1126–1131 . doi:10.1111/1755-0998.13145

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 . doi:10.1093/genetics/155.2.945

Raymond, M., and Rousset, F. (1995). GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *The Journal of Heredity* **86**, 248–249 . doi:10.1093/oxfordjournals.jhered.a111573

Rosenberg, N. A. (2004). Distruct: a program for the graphical display of population structure. *Molecular Ecology Notes* **4**, 137–138 . doi:10.1046/j.1471-8286.2003.00566.x

Rousset, F. (2008). GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources* **8**, 103–106 . doi:10.1111/j.1471-8286.2007.01931.x

Weir, B. S., and Cockerham, C. C. (1984). Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370.

Wilson, G. A., and Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* **163**, 1177–1191 . doi:10.1093/genetics/163.3.1177

Wright, S. (1949). The genetical structure of natural populations. *Annals of Eugenics* **15**, 323–354 . doi:10.1111/j.1469-1809.1949.tb02451.x