Universität Zürich[UZH]

Geographisches Institut

Exploring Genomic Adaptations of Reef Fish to Global Change - A Comparative Seascape Genomics Approach

GEO 510 Master's Thesis

Author: Naroa Olivia Schweizer, 21-724-414

Supervised by: Dr. Oliver Michele Selmoni

Faculty representative: Prof. Dr. Meredith Schuman

28.08.2026

Exploring Genomic Adaptations of Reef Fish to Global Change

A Comparative Seascape Genomics Approach

Naroa Olivia Schweizer

[21-724-414]

University of Zurich

Supervisor: **Dr. Oliver Michele Selmoni**

Faculty member: **Prof. Dr. Meredith Christine Schuman**

Zurich, 28.08.2026

Abstract

Coral reefs are declining rapidly under global change, yet while the effects on corals themselves are well studied, much less is known about how reef fish respond to these changes. Most existing research has focused on individual species, geographic regions, or single environmental factors, with comparatively little attention given to anthropogenic pressures, such as boat-detection frequency and built-up area density, alongside climate-related pressures.

This Master's thesis addresses this gap using a comparative approach to seascape genomics, an emerging method that integrates genomic data with environmental variables to investigate whether environmental variation can explain patterns of genomic adaptation. Four reef fish species from different geographic regions worldwide, the twoband anemonefish (*Amphiprion bicinctus*), nassau grouper (*Epinephelus striatus*), red lionfish (*Pterois volitans*), and urchin cardinalfish (*Siphamia tubifer*), were examined for associations between genomic variation and 32 climatic and anthropogenic environmental variables. As none of the study species has a fully annotated reference genome, the well-studied zebrafish (*Danio rerio*) genome was used as a common reference for gene annotation and to enable comparisons across species. Repeated adaptive signals were identified where different species showed associations with the same environmental variable.

Species-specific potential adaptive signals were identified in all four species but varied considerably among species and environmental variables. Associations with sea-water velocity, built-up area density, and boat-detection frequency were detected in all four species. Across species, repeated potential adaptive signals were associated with six environmental variables. The strongest and most consistent signal involved the zebrafish homolog *lrp1ab* and variability of iron concentration, which was detected in all four species. Sea-surface temperature was also associated with multiple repeated adaptive signals in a genomic region, but no clear functional context could be identified. Weaker repeated adaptive signals were found for phosphate concentration, land-surface density, and cropland density, but were supported by fewer species. These results indicate that both climatic and anthropogenic environmental conditions may contribute to adaptation in reef fish under global change. The study also demonstrates the value of comparative multi-species approaches for identifying potential repeated adaptive signals that may be overlooked in single-species analyses.

Table of Contents

1 Introduction **1**
2 Research Question **2**
3 Methodology **2**
3.1 Workflow 2
3.2 Genomic Data Processing Pipeline 3
3.2.1 Exploratory Assessment of Data Quality and Population Structure 6
3.2.2 Imputation of Genotypes 6
3.2.3 Linkage Disequilibrium Decay Estimation 6
3.3 Environmental Variables and Processing 7
3.4 Genotype-Environment Association 8
3.4.1 Latent Factor Mixed Model 9
3.4.2 Correlation Analysis 9
3.5 Cross-Species Analysis 9
3.5.1 Mapping SNPs to Genes 9
3.5.2 PicMin 10
3.5.3 Gene Set Enrichment Analysis 11
3.5.4 Visualisation of Candidate SNP-Environment Associations 11
4 Results **11**
4.1 Exploratory and Preliminary Analysis 11
4.2 Genotype-Environment Analysis 13
4.3 Cross-Species Analysis 15
5 Discussion **21**
5.1 Repeated Adaptive Signal at *lrp1ab* with Iron Variability 21
5.2 Repeated Adaptive Signals for Average Sea Surface Temperature 23
5.3 Methodological Considerations 23
5.4 Implications and Future Directions 24
6 Conclusion **24**
7 AI Disclosure **25**
8 Code Availability **25**
9 Acknowledgment and Attributions **25**
10 Bibliography **26**
10.1 List of Figures 26
10.2 List of Tables 26
10.3 References 26
1 Appendix **1**
1.1 Software, R Packages, and Version Information 1
1.2 Excluded Sequencing Runs and Exclusion Criteria 2
1.3 Supplementary Figures per Species 3
1.3.1 *A. Bicinctus* - Special Supplementary Analysis 3
1.3.2 PCoA Analysis 10
1.3.3 F_{st} Analysis 13
1.3.4 Ancestral Population Analysis 15

1.3.5 Linkage Disequilibrium Decay Analysis 17
1.3.6 Latent Factor Mixed Modell 19
1.3.7 Correlation Analysis 22
1.3.8 Overall GEA summary statistics table 25
1.3.9 Gene Set Enrichment Analysis 26
1.4 List of Figures 28
1.5 List of Tables 29
Personal Declaration 30

Abbreviations

Abbreviation	Meaning	Abbreviation	Meaning
ATP	Adenosine triphosphate	LD	Linkage Disequilibrium
avg	Average	LFMM	Latent Factor Mixed Model
BATHY / Bathy	Bathymetry (depth) [m]	MF	Molecular Function (Gene Ontology category)
BCFtools	Binary Call Format tools (variant calling software)	MultiQC	Software aggregating quality-control reports across samples/tools
BLAST	Basic Local Alignment Search Tool	n_est	Number of contributing species (in PicMin)
BP	Biological Process (Gene Ontology category)	NA	Not available / not applicable
bp	Base pair(s)	NCBI	National Center for Biotechnology Information
CA	Correlation Analysis	NO3	Nitrate concentration [mmol/m^3]
CC	Cellular Component (Gene Ontology category)	O2	Oxygen (dissolved) concentration [mmol/m^3]
CHL	Chlorophyll concentration [mg/m^3]	PCoA (PC1/PC2)	Principal Coordinates Analysis (first/second principal coordinate axis)
CPU	Central Processing Unit	PDEN	(Human) population density [hab/km^2]
CROP	Cropland density [%]	PH	Seawater pH
DHW	Degree heating week(s) [°C·$week^{-1}$]	PicMin	Statistical method (order statistics) for identifying repeated adaptive signals across species
EDM	Euclidean Distance Matrix	PO4	Phosphate concentration [mmol/m^3]
ENSO	El Niño-Southern Oscillation	QQ (plot)	Quantile-Quantile plot
env.	Environmental	R^2	Coefficient of determination
FASTA	FAST-All (sequence file format)	RECIFS	Reef Environment Centralized InFormation System (database)
FASTQ	FASTA file with quality scores (sequencing read file format)	SAMtools	Sequence Alignment/Map tools
FastQC	Read-quality control software	SCV	Sea(water) surface/current velocity [m/s]
FDR	False Discovery Rate	sd	Standard deviation
FE	Iron concentration [mmol/m3]	SD	Standard deviation
F_{st}	Fixation index (measure of genetic differentiation)	SNP(s)	Single-Nucleotide Polymorphism(s)
GC (content)	Guanine-Cytosine content	SPM	Suspended matter concentration [g/m^3]
GEA	Genotype-Environment Association	SRA	Sequence Read Archive
GIF	Genomic Inflation Factor	SSS	Sea-surface salinity [10^{-3}]
GO	Gene Ontology	SST	Sea-surface temperature [°C]
GRCz11	Genome Reference Consortium Zebrafish build 11 (reference genome assembly)	Stouffer	Statistical method for combining p-values
GSEA	Gene Set Enrichment Analysis	tblastn	Translated BLAST nucleotide-protein search (BLAST command-line function)
GT	Genotype	UniProt	Universal Protein Resource (protein sequence/annotation database)
GTF	Gene Transfer Format	URBA	Built-up (urban) area density [%]
ID(s)	Identifier(s)	USD	United States dollar
IQR	Interquartile Range	VBD	Boat-detection frequency [%]
K	Number of ancestral populations / latent factors (model parameter)	VCF	Variant Call(ing) Format
kb	Kilobase(s)	LD	Linkage Disequilibrium
LAND	Land-surface density [%]	LFMM	Latent Factor Mixed Model

1 Introduction

Global change is a pervasive phenomenon that has widespread effects on ecosystems, with coral reefs being among the most severely affected (Goreau and Hayes 2021; Selmoni et al. 2023; 2020; Whiting et al. 2024). Coral reefs are the most diverse marine ecosystem and provide habitat for a high number of species (Cowman et al. 2017; Selmoni et al. 2023; 2020). It is estimated that approximately 830,000 multicellular plant and animal species inhabit coral reefs worldwide (Fisher et al. 2015), including around 7,000 species of reef associated tropical fish (Strona et al. 2021). Beyond their biodiversity, coral reefs provide essential ecosystem services. Millions of people, particularly those living in coastal regions, depend on coral reefs for seafood collection and fisheries (Eddy et al. 2021; Strona et al. 2021). Approximately six million reef fisheries exist globally, generating an estimated economic value of six billion USD (Eddy et al. 2021).

Coral reef cover is rapidly declining as they are highly sensitive to marine heatwaves and anthropogenic disturbances (Selmoni et al. 2020; 2023; Eddy et al. 2021). Between 1957 and 2007, global reef cover declined by approximately 50% (Eddy et al. 2021) with a clear trend predicting further decline (Hughes et al. 2017; Thirukanthan et al. 2023). These declines are primarily driven by global change and associated environmental pressures (Thirukanthan et al. 2023). Global change describes large-scale, human-driven changes to the Earth system and is closely linked to the concept of the Anthropocene, which describes an era in which human activities dominate changes in all Earth's spheres (Steffen et al. 2015). A major driver of global change is the increase in atmospheric greenhouse gas concentrations, which enhances atmospheric warming and, in turn, increases ocean temperatures (Steffen et al. 2015, Thirukanthan et al. 2023). Elevated seawater temperatures disrupting the obligate symbiosis between hard-skeleton corals and their microbial symbionts, which are essential for coral survival, and can lead to coral bleaching (Selmoni et al. 2024; 2020). Global change also extends to direct anthropogenic pressures, including overfishing, pollution, and the commodification of natural resources, which can influence nutrient availability and larval dispersal (Thirukanthan et al. 2023; Fraimout et al. 2026; Steffen et al. 2015). Beyond their effects on corals, these global changes also impose significant selective pressures on reef fish species that inhabit these ecosystems and play a vital role in maintaining coral reef health and biodiversity (Cowman et al. 2017; Lefcheck et al. 2019; Selmoni et al. 2020; Strona et al. 2021; Fraimout et al. 2026).

Species facing drastic environmental changes can either shift their geographical range or persist within their local habitat through phenotypic plasticity, which enables tolerance of changing conditions, or through genetic adaptation, which supports survival and evolution under these conditions (Freitas et al. 2021; Fraimout et al. 2026; Smith et al. 2013; Waldvogel et al. 2020; Whiting et al. 2024). While considerable research has investigated the responses of corals to climate-driven environmental change (Selmoni et al. 2024; 2020; Thirukanthan et al. 2023), less attention has been given to the adaptive responses of reef fish. Existing studies on fish have predominantly examined genetic adaptation to environmental variables, such as sea surface temperature, salinity, and pollution (Munday et al. 2017; Chen et al. 2018; Fraimout et al. 2026; Smith et al. 2013; Velotta et al. 2022). In comparison, the potential contribution of more direct anthropogenic pressures, *e.g.*, proximity to built-up areas or average boat traffic, to adaptive genetic responses has received less consideration. For example, tourism and boat traffic have been linked to declines in reef fish density and richness on Bora Bora (Gairin et al. 2024), while boat noise has been shown to affect the early life stages of Lusitanian toadfish (Faria et al. 2022). Multi-species studies have examined genomic differentiation associated with contrasting salinity environments (Velotta et al. 2022) and phenotypic responses to climate change (Crozier and Hutchings 2014).

To investigate how reef fish respond to these environmental changes, seascape genomics provides a valuable approach by integrating genomic data with environmental variables, which are often obtained through remote sensing (Selmoni et al. 2020). By integrating environmental variables with genomic data, seascape genomics can assess whether variation in environmental conditions helps explain genetic variation. These variables can include key biotic and abiotic factors such as temperature, salinity, chlorophyll concentration, and boat traffic. This can improve the detection of Genotype-Environment Associations (GEA) (Oleksiak and Rajora 2020; Riginos et al. 2016; Selmoni et al. 2020; Waldvogel et al. 2020). However, GEA approaches can be prone to false-positive associations, particularly with present population structure or confounding variables due to multicollinearity (Frichot et al. 2013; François et al. 2016; Whiting et al. 2024). Cross-species GEA comparisons can provide additional support for associations that are consistently detected across species. Repeated associations may indicate shared responses to the same environmental pressures and therefore provide greater confidence in the observed signals. However, cross-species comparisons remain limited (Whiting et al. 2024), with most studies focusing on individual species, specific geographic regions, or single environmental variables (Fraimout et al. 2026; Gairin et al. 2024; Velotta et al. 2022). Expanding these comparisons across geographically distributed reef fish species may provide valuable insights into how different reef fish respond to global change and whether similar environmental pressures are associated with repeated adaptive genomic signals across global reef systems.

2 Research Question

This Master's thesis uses a seascape genomic approach to explore potential genomic adaptations to global change across four globally distributed reef fish species (Figure 2). GEAs are assessed using two complementary methods across 32 global change variables, followed by cross-species analysis to identify potential repeated adaptive signals. The functional context of associated genes is also explored to support their biological interpretation. Specifically, this study aims to (i) identify potential adaptive signals associated with global change variables, (ii) assess whether similar environmental conditions are associated with repeated adaptive signals across reef fish species, and (iii) investigate the functional context of associated genes.

3 Methodology

3.1 Workflow

The overall workflow consists of four main components, summarized in Figure 1. The first comprises the genomic data processing pipeline. The second involves the preparation of the environmental data. The third performs the GEA analyses by combining the processed genomic and environmental data for each species. The fourth component consists of the cross-species analysis. Each component is described in detail in the following sections.

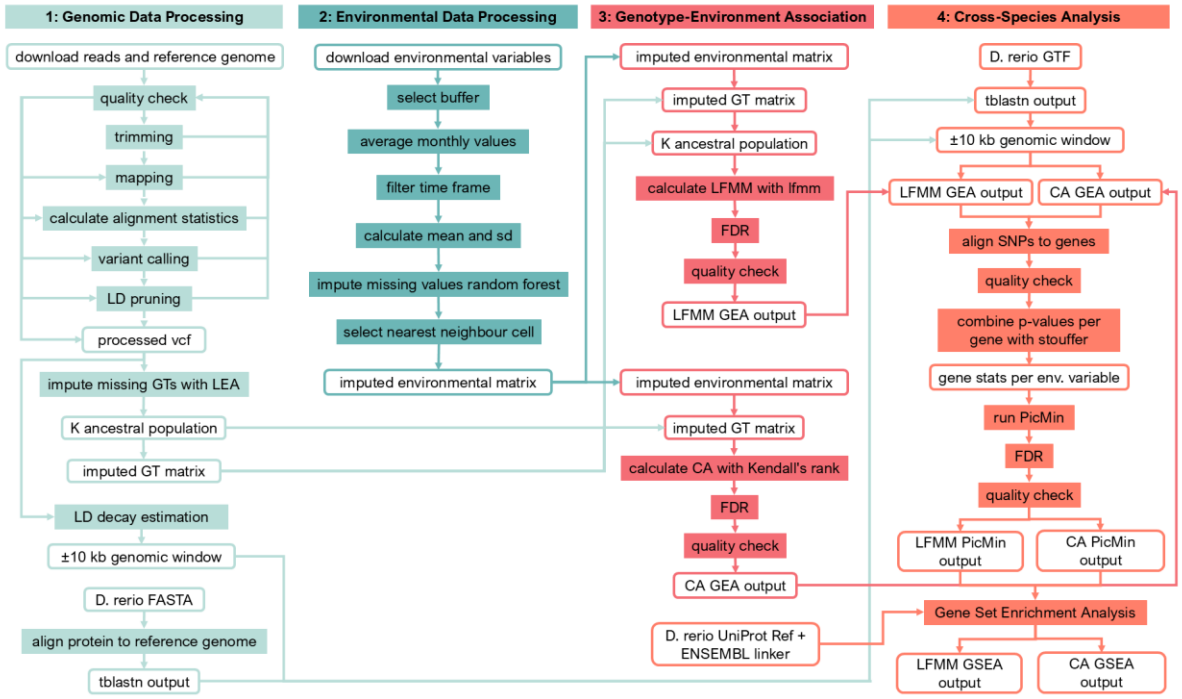


Figure 1: Detailed flowchart illustrating the overall data processing and analysis workflow. Abbreviations: VCF, Variant Calling File; LEA: R package (3.2.2 Imputation of Genotypes); GT, genotype; LD, linkage disequilibrium; D. rerio FASTA, Danio rerio FASTA-format sequence file; tblastn, BLAST command-line function; SD, standard deviation; LFMM, Latent Factor Mixed Model; FDR, false discovery rate; GEA, genotype-environment association; CA, correlation analysis; D. rerio UniProt Ref, UniProt protein reference/annotation file; D. rerio ENSEMBL linker, mapping file linking ENSEMBL gene IDs to UniProt protein IDs; D. rerio GTF, Danio rerio Gene Transfer Format; SNPs, single-nucleotide polymorphisms; env., environmental. Stouffer, statistical method for combining p-values (3.5.1 Mapping SNPs to Genes); PicMin, statistical method for identifying repeated adaptive signals across species (3.5.2 PicMin); GSEA: Gene Set Enrichment Analysis.

3.2 Genomic Data Processing Pipeline

The genomic data processing workflow was implemented as a modular pipeline consisting of wrapper scripts written in R and Slurm batch scripts executed in Bash on a high-performance computing cluster. All command line tools were set up in a Miniconda (v25.3.1, Anaconda, Inc. 2025) environment. The R scripts handle data preparation, file management, and parameter definition, while the Slurm scripts execute the individual command line tools for each processing step.

Genomic data for four reef fish species, *Amphiprion bicinctus* (twoband anemonefish) (Saenz-Agudelo et al. 2015), *Epinephelus striatus* (nassau grouper) (Sherman et al. 2020), *Pterois volitans* (red lionfish) (Bors et al. 2019) and *Siphamia tubifer* (urchin cardinalfish) (Gould and Dunlap 2017), were obtained from the publicly available National Center for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/). Species were selected based on the availability of a reference genome, a sufficient number of samples from multiple sampling locations, and geographic information for these locations, either as coordinates provided in the BioProject run table or in supplementary information in the corresponding publications. To ensure global geographic coverage, species were chosen that had been sampled in different regions around the world (Figure 2).

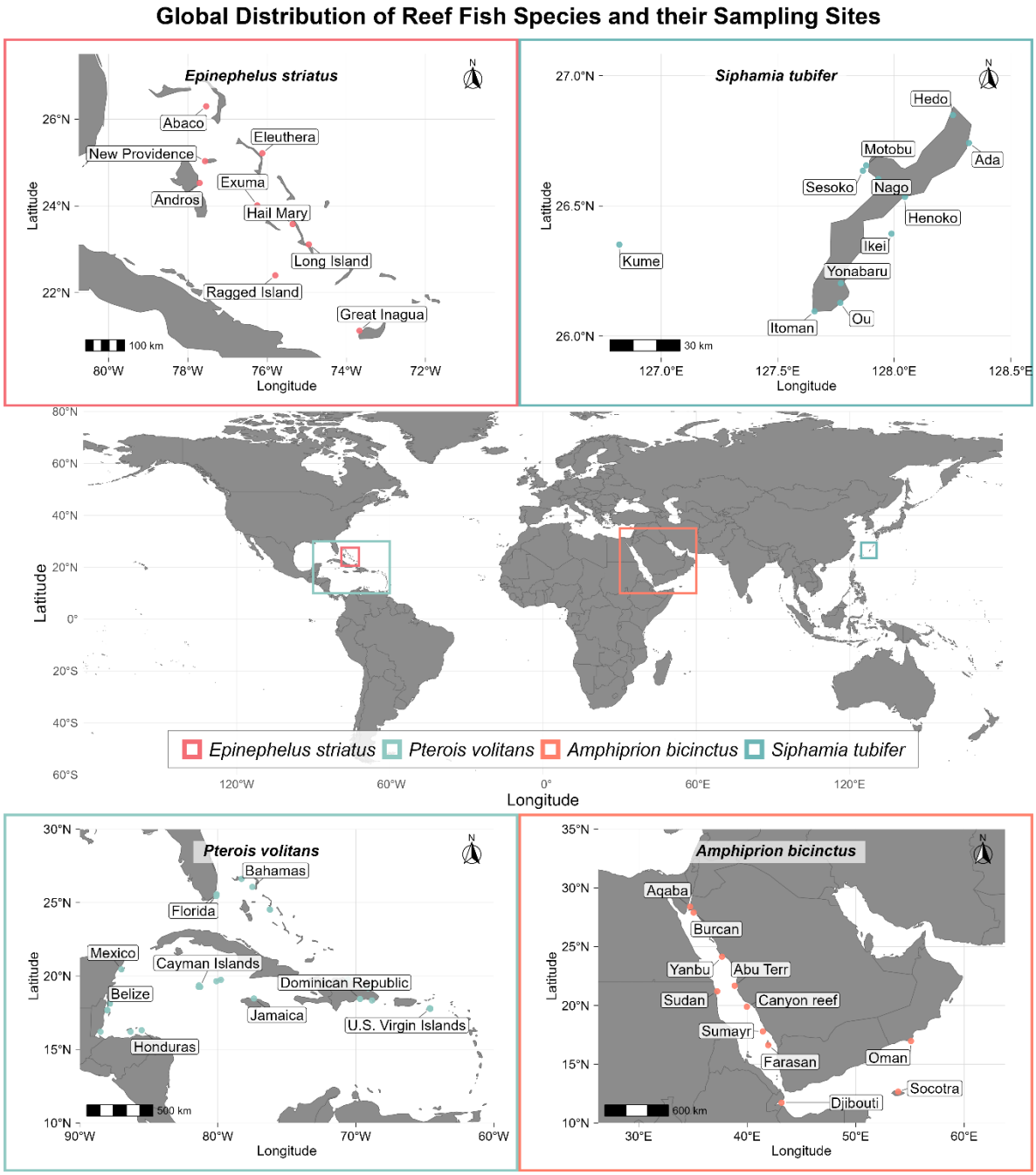


Figure 2: Global distribution of sampling sites for the four species of interest, with species specific maps showing the detailed geographic distribution of sampling locations for each species.

To download Sequence Read Archive (SRA) data from NCBI and convert it into FASTA and FASTQ files, the command line tool parallel-fastq-dump (v.0.6.7, Valieris 2025) was used. Read quality was checked using FastQC (v0.12.1, Andrews 2023) and MultiQC (v1.33, MultiQC 2025), with FastQC generating read quality metrics (Sayers et al. 2022) and MultiQC combining outputs from FastQC, SAMtools statistics, and flagstats into a single summary for quality control at each processing stage (Ewels et al. 2016). The first read quality assessment was used to define the trimming thresholds (Table 3). This involved assessing the Phred scores at the reads five prime and three prime ends, adaptor contamination and identifying barcodes. Reads were trimmed and quality filtered with Trim Galore (v0.6.10, Krueger 2019), which was selected because it integrates automated adapter detection with quality-based trimming (Krueger 2019; Utturkar et al. 2020).

For each species, the trimmed and quality-filtered reads were aligned to the reference genome using NextGenMap (v0.5.5, Rescheneder 2017), selected for its computational efficiency and robust performance on highly polymorphic genomes, making it well suited for mapping reads to congeneric reference genomes where species specific assemblies were unavailable (Sedlazeck et al. 2013). The reference genomes used were *Amphiprion ocellaris* for *A. bicinctus, Epinephelus awoara* for *E. striatus*, *Pterois miles* for *P. volitans*, and *S. tubifer* for *S. tubifer* (Table 1). Alignment statistics and variant calling were performed using SAMtools (v1.22.1, SAMtools 2025) and BCFtools (v1.22, BCFtools 2025) (Danecek et al. 2021), as bcftools mpileup function performs well on non-human species and non-model reference genomes (Lefouili and Nam 2022).

The resulting Variant Call Format (VCF) file was filtered using bcftools view function to retain only biallelic SNPs with less than ten percent missing genotypes and a minor allele frequency above five percent. Soft linkage disequilibrium (LD) pruning was applied using a window size of 1000 bp and a correlation threshold of $R^2 = 0.9$ to remove only closely linked SNPs with highly similar allele-frequency patterns, which can complicate correction for multiple testing and inflate test statistics in GEA analyses (Bulik-Sullivan et al. 2015; Yang et al. 2011; Cinar and Viechtbauer 2022). Individual samples were excluded based on low mapping rates, abnormal GC content, or missing geographic metadata required for downstream analyses. Sequencing runs for which the alignment step failed due to computational or technical errors were also excluded (S Table 3).

Table 1: Summary of genomic input data for reef fish species, including, geographical region, sampling time frame, the number of samples (N samples), number of sampling locations (N sites), their bioproject in NCBI, the reference genome and species and notes on manual adjustments.

Species	***Amphiprion bicinctus***	***Epinephelus striatus***	***Pterois volitans***	***Siphamia tubifer***
Common Name	twoband anemonefish	nassau grouper	red lionfish	urchin cardinalfish
Region	Red Sea, Gulf of Aden	Caribbean (Bahamas)	Caribbean	Okinawa
Sampling Time Frame	2014	2014-2017	2012-2013	2012-2014
N Samples	113	96	120	280
N Sites	8	9	9	11
Bioproject	PRJNA294760	PRJEB36904	PRJNA510810	PRJNA385011
Publication	Saenz-Agudelo et al. 2015	Sherman et al. 2020	Bors et al. 2019	Gould and Dunlap 2017
Reference Genome Species	*Amphiprion ocellaris*	*Epinephelus awoara*	*Pterois miles*	*Siphamia tubifer*
Reference Genome	GCF_022539595.1	GCA_035609425.1	GCA_947000775.1	GCA_020466265.1
Publication	Ryu et al. 2022	Zhang et al. 2024	Kitsoulis et al. 2023	Gould et al. 2022
Assembly Level	chromosome	chromosome	contig	chromosome
Manual Adjustments	Latitude/longitude re-formatted	Coordinates digitised from publication map	Coordinate system standardised	Coordinates added from supplementary table

3.2.1 Exploratory Assessment of Data Quality and Population Structure

Prior to the GEA, exploratory analyses were conducted to identify potential technical artefacts and assess genetic structure. For each species, a Principal Coordinates Analysis (PCoA) was performed on a Euclidean Distance Matrix (EDM) calculated from the genotype matrix after importing the filtered VCF file into R using the vcfR package (v1.15.0, Knaus 2023) and converting it into a genlight object using the adegenet package (v2.1.11, Jombart 2025). The distribution of samples along the first two principal coordinate axes provided initial insight into potential genetic differentiation and population structure. Sequencing quality metrics extracted from the MultiQC reports with the TidyMultiqc package (v1.0.3, Milton 2016) were compared with the first two principal coordinate axes and using Pearson correlation coefficients to assess whether the observed genetic variation was associated with technical rather than biological factors. In addition, sampling locations obtained from the corresponding NCBI BioProject run tables were visualised on a map to verify the geographical distribution of the samples.

Potential population structure was further assessed by calculating pairwise genetic differentiation (pairwise F_{st}) following the estimator of Weir and Cockerham using the package hierfstat (v0.5.11, Goudet et al. 2022). As geographical coordinates were available for all samples, they were used as a standardised basis for defining sampling clusters. Samples were hierarchically clustered using a distance threshold of 5 km. These clusters were then used for the pairwise F_{st} estimation. The resulting pairwise F_{st} values were visualised to assess spatial patterns of genetic differentiation.

3.2.2 Imputation of Genotypes

Missing genotypes in the filtered genlight object were imputed to prepare the dataset for GEA analysis and comply with LEA input requirements (Frichot and François 2015). Population structure was then analysed using the snmf function in the LEA package (v3.20.0, Frichot et al. 2024), which decomposes the genotype matrix into two lower-dimensional probabilistic matrices $P \approx QF$. The matrix Q represents individual admixture coefficients across K ancestral populations, while the matrix F contains locus-specific allele frequencies within these populations (Frichot and François 2015; Gain and François 2021). The model was fitted across K values, in this study from 1 to 8, with a cross-entropy criterion used to identify the optimal number of ancestral populations (Frichot and François 2015). Based on the optimal K value (*i.e.*, the lowest cross-entropy), missing genotypes were imputed using the impute function in the LEA package. Missing genotypes were replaced with the most probable genotype at each locus based on the reconstructed QF distribution, thereby accounting for the inferred population structure (Frichot and François 2015; Gain and François 2021).

3.2.3 Linkage Disequilibrium Decay Estimation

To determine the genomic distance used for SNP-to-gene assignment (3.5.1 Mapping SNPs to Genes), linkage disequilibrium (LD) decay was estimated for each species. Compressed VCF files were imported using the vcfR package and converted to genlight objects using the vcfR2genlight function. Pairwise LD was calculated for the longest assembled sequence (contig or chromosome) of each species using the gl.report.ld function in dartR (v2.9.7, Gruber 2023), and the physical distance between SNP pairs was calculated from their genomic positions. Because recombination generally reduces LD with increasing genomic distance, LD was expected to decrease with physical distance. The observed species specific LD-distance relationships were broadly consistent with an exponential decline across species; therefore, a simple exponential decay model $R^2(d) = a \times e^{-b \times d + c}$ was fitted, with c representing background LD. The genomic distance corresponding to 95% decay of excess LD above background was used as a threshold beyond which SNPs were considered largely unlinked. These estimates were then used to define a suitable common mapping window for SNP-to-gene assignment across species.

3.3 Environmental Variables and Processing

The environmental variables were obtained from the publicly available Reef Environment Centralized InFormation System (RECIFS) database (Selmoni et al. 2023). RECIFS provides data on key environmental variables that characterise the abiotic, biotic, and anthropogenic conditions of coral reefs worldwide. A total of 61,038 reef cells, represented as a 5 km raster, cover all global coral reefs. For each reef cell, raw environmental values were extracted without upscaling or downscaling and averaged within buffer distances of 2.5, 10, 25, and 50 km to capture environmental conditions across different spatial scales (Selmoni et al. 2023). The processed data are organised into summary tables, with each row representing a reef cell and each column representing a temporal observation, such as monthly or annual values, or a single value for time-invariant variables (Selmoni et al. 2023). For this thesis, the environmental data were downloaded from the RECIFS web-based interface for all reef cells and buffer distances described above, with data accessed on 15 October 2025.

The following environmental variables were prepared in an R script: chlorophyll concentration, degree heating week, iron concentration, dissolved oxygen concentration, seawater pH, nitrate concentration, phosphate concentration, seawater velocity, suspended matter concentration, sea surface salinity, sea surface temperature, bathymetry depth, cropland density, surface land density, human population density, built-up area density, and boat detection freqeuncy. A 10 km buffer was selected to account for larval dispersal. Most reef fishes exhibit larval dispersal distances of 5 to 15 km (Green et al. 2015), with reported values of up to 16.4 km in *Elacatinus lori* (D'Aloia et al. 2015), 11 km in *Amphiprion clarkii* (Pinsky et al. 2010), and 9.8 km in *Amphiprion biaculeatus* (Fitz et al. 2023), while broader estimates for reef fish larval dispersal range from 10 to 100 km (Cowen et al. 2006). However, as self-recruitment is common (Green et al. 2015; Pinsky et al. 2010), a 10 km buffer was deemed suitable. Then, monthly environmental variables were first averaged by year, while time invariant variables were retained unchanged. The data was then filtered to include all available environmental values from the beginning of each time series to the first sampling year of the respective species. The average (avg) and standard deviation (sd) across years were then calculated for each reef cell. For cropland density, human population density, built-up area density, and boat detection frequency, which were available as time series but not as monthly measurements, the same summary statistics were calculated across the available time points. To generate the environmental matrix for the GEA, each fish sample was assigned to its nearest RECIFS reef cell using the Nearest Neighbour index implemented in the sf package (v1.0.24, Pebesma and Bivand 2026). Sample identifiers and coordinates were merged with the corresponding environmental variables using the dplyr package (v1.2.0, Wickham et al. 2026). Missing environmental values were subsequently imputed with the missForest package (1.6.1, Stekhoven 2025), which applies an iterative random forest algorithm to predict missing values from relationships among the observed variables. The resulting imputed dataset was used as the environmental matrix for the GEA analyses. This R script was run for each fish species separately.

Table 2: List of RECIFs environmental variables, including descriptions, available time periods, and original data sources. Data were accessed from RECIFs on 15th October 2025.

Code	Variable name	Description	Temporal coverage	Data source
CHL	Chlorophyll concentration [mg/m^3]	Monthly averages of mass concentration of chlorophyll-a in seawater.	1997-2023	(E.U. Copernicus Marine Service Information 2023a)
DHW	Degree heating week [°C-week]	Monthly maxima of degree heating week (DHW). DHW is calculated as the accumulation of thermal stress (*i.e.* temperature >1°C above the monthly maximal mean temperature) over the previous 12 weeks.	1985-2023	(NOAA Coral Reef Watch 2018a)
FE	Iron concentration [mmol/m^3]	Monthly averages of mole concentration of dissolved iron in sea water.	1993-2022	(E. U. Copernicus Marine Service Information 2024)

Code	Variable name	Description	Temporal coverage	Data source
O2	O2 concentration[mmol/m^3]	Monthly averages of mole concentration of dissolved molecular oxygen in sea water.	1993-2022	(E. U. Copernicus Marine Service Information 2024)
PH	Seawater pH	Monthly averages of sea water pH reported on total scale.	1993-2022	(E. U. Copernicus Marine Service Information 2024)
NO3	Nitrate concentration [mmol/m^3]	Monthly averages of mole concentration of nitrate in sea water.	1993-2022	(E. U. Copernicus Marine Service Information 2024)
PO4	Phosphate concentration [mmol/m^3]	Monthly averages of mole concentration of phosphate in sea water.	1993-2022	(E. U. Copernicus Marine Service Information 2024)
SCV	Seawater velocity [m/s]	Monthly averages of sea water surface velocity, computed as the Euclidean norm of eastward and northward velocity.	1993-2021	(E.U.Copernicus Marine Service Information 2023b)
SPM	Suspended matter concentration [g/m^3]	Monthly averages of mass concentration of suspended matter in sea water.	1997-2023	(E.U. Copernicus Marine Service Information 2023a)
SSS	Sea surface salinity [1e-3]	Monthly averages of sea surface salinity.	1993-2021	(E.U. Copernicus Marine Service Information 2023b)
SST	Sea surface temperature [°C]	Monthly average of sea surface temperature.	1985-2023	(NOAA Coral Reef Watch 2018b)
BATHY	Depth [m]	Average depth calculated for area surrounding the reef of interest.	-	(Ryan et al. 2009a) (Ryan et al. 2009b)
CROP	Density of cropland [%]	Percentage of pixels corresponding to cropland land cover for the area surrounding the reef of interest.	2015, 2016, 2017, 2018, 2019	(E. U. Copernicus Land Monitoring Service Information 2019)
LAND	Density of surface land [%]	Percentage of pixels corresponding to surface land for the area surrounding the reef of interest.	-	(Ryan et al. 2009a) (Ryan et al. 2009b)
PDEN	Human population density [hab/km^2]	Average human population density calculated the area surrounding the reef of interest.	2000, 2005, 2010, 2015, 2020	(Center For International Earth Science Information Network-CIESIN-Columbia University 2018)
URBA	Density of built-up [%]	Percentage of pixels corresponding to built-up land cover for the area surrounding the reef of interest.	2015, 2016, 2017, 2018, 2019	(E. U. Copernicus Land Monitoring Service Information 2019)
VBD	Boat detection [%]	Average of the percentage of boat detected per satellite overpass coverage for the area surrounding the reef of interest.	2017, 2018, 2019, 2020, 2021	(Elvidge et al. 2015; 2018; Hsu et al. 2019)

3.4 Genotype-Environment Association

The next step, after data processing, is the GEA analysis, which statistically associates the genotype variation to an environmental variable (Lasky et al. 2023; Rellstab et al. 2015; Otte 2024; Selmoni et al. 2020). This analysis was carried out separately for each reef fish species, using two GEA methods: a Latent Factor Mixed Model (LFMM) and a Correlation Analysis (CA).

3.4.1 Latent Factor Mixed Model

LFMM was used to identify associations between genotypes and environmental variables while accounting for background population structure through latent factors (Frichot et al. 2013; Luo et al. 2021; Rellstab et al. 2015). The method was selected for its computational efficiency and ability to control for neutral population structure (Rellstab et al. 2015; Caye et al. 2019; Luo et al. 2021).

LFMM requires the number of latent factors (K) as an input parameter (Luo et al. 2021). For each species, the K value identified from the genotype imputation analysis based on the lowest cross-entropy value was used (Caye et al. 2019). Analyses were performed using the lfmm package (v1.1, Jumentier et al. 2021b). The imputed genotype matrix was analysed against the processed environmental variables, with a separate LFMM model fitted for each variable using the lfmm_ridge function. Association statistics, including z-scores, p-values, and the genomic inflation factor (GIF), were obtained using the lfmm_test function (Jumentier et al. 2021a). P-values were calibrated using the estimated GIF, and model performance was assessed using QQ plots (Appendix: 1.3.6 Latent Factor Mixed Modell), created using ggplot2 (v4.0.2, Wickham et al. 2026) and patchwork (v1.3.2, Pedersen 2025).

Following recommendations for GEA analyses, p-values were adjusted using the Benjamini-Hochberg false discovery rate (FDR) procedure to control the expected proportion of false discoveries among significant SNP-environment associations and obtain q-values (Benjamini and Hochberg 1995; Rellstab et al. 2015). SNPs with q-values below 0.05 were considered significantly associated with the respective environmental variable. All LFMM results, including SNP identifiers, p-values, GIF-corrected p-values, and q-values, were saved for downstream analyses.

3.4.2 Correlation Analysis

To compare the LFMM GEA with an alternative approach, a correlation analysis (CA) was performed using the same imputed genotype matrix and environmental dataset. Associations between each environmental variable and all SNPs were tested independently using Kendall's rank correlation coefficient. Kendall's tau is a non-parametric, rank-based measure of correlation that does not assume a specific population distribution and has been shown to be more robust than Pearson's correlation under different genotype encodings (Alcaraz et al. 2022; Kim and Lee 2020), making it well suited for GEA analyses. Q-values were obtained using the Benjamini-Hochberg procedure to control the false discovery rate (FDR), as in the LFMM analysis (Benjamini and Hochberg 1995; Rellstab et al. 2015). Again, the model's performance was assessed using QQ plots (Appendix: 1.3.7 Correlation Analysis), which were created using ggplot2 and patchwork. Significant SNPs were identified with a q-value threshold of below 0.05. All CA results, including SNP identifiers, Kendall's tau values, p-values and q-values were saved for downstream analyses.

3.5 Cross-Species Analysis

Following the species-specific GEA analyses, a cross-species analysis was implemented to determine whether repeated adaptive signals to global change variables occur in different species across different geographic locations.

3.5.1 Mapping SNPs to Genes

Homologous sequences were identified between *Danio rerio* (zebrafish) and the reference genomes of the study species to link SNPs to genes and enable functional comparisons across reef fish species. *D. rerio* was used as a functional reference due to its extensive gene function, orthology, and Gene Ontology (GO) annotations (Bradford et al. 2022).

A FASTA file containing annotated protein sequences for *D. rerio* was downloaded from the Ensembl release 115, based on the GRCz11 (GCA_000002035.4) reference genome assembly. Homologous *D. rerio* protein sequences were identified in the reference genomes of each study species using the tblastn function of the BLAST command-line tool (v.2.17.0, Camacho et al. 2009; Gertz et al. 2006), with an E-value threshold of 5E-7. BLAST hits were furthermore filtered to retain high confidence matches (percent_identity > 70) using the dplyr package. SNPs in each species were assigned to the closest homologous gene within a ±10 kb genomic window. This window size was selected to accommodate LD decay across all four species (3.2.3 Linkage Disequilibrium Decay Estimation).

The GEA results were then combined with the BLAST output and the *D. rerio* GTF annotation from Ensembl release 115 based on the GRCz11 (GCA_000002035.4). Gene, transcript, and protein identifiers were extracted from the GTF file, while SNP positions were extracted from SNP identifiers using the stringr package (v1.6.0, Wickham 2025). SNPs were assigned to nearby proteins within the ±10 kb window using a custom mapping function, and Ensembl protein version suffixes were removed before linking proteins to their corresponding gene IDs. Finally, p-values of SNPs assigned to the same gene were combined into a single gene-level p-value using the Stouffer method from the poolr package (v.1.2.0, Cinar and Viechtbauer 2026). Stouffer combines individual p-values by converting them to z-scores, aggregating them and transforming the combined z-score back into a single p-value (Cinar and Viechtbauer 2022). This method was applied separately to the p-values from LFMM and CA for each species. The resulting combined p-values served as input for the subsequent PicMin analysis.

3.5.2 PicMin

PicMin is a statistical method based on order statistics that identifies genes showing repeated signals of adaptation across multiple species using the results of GEA analyses. It compares the genome-wide ranks of orthologous genes across species to assess whether repeated signals of adaptation are more frequent than expected by chance (Booker et al. 2023). This makes PicMin well suited for identifying repeated adaptive signals in comparative datasets and is available as an R package (v.0.0.0.9000, Booker et al. 2023; Whiting et al. 2024).

Species-specific gene-level datasets containing the combined p-values were merged by environmental variable, producing a matrix with genes as rows and species as columns for each environmental variable. Following the recommended workflow, the EmpiricalPs function in the PicMin package was then used to calculate empirical p-values for each gene and species from these matrices. A null distribution was precomputed for three to four lineages using the PicMinNull function. Gene-level association signals were then analysed across species using the RunPicmin function. For each gene, PicMin compared the observed ranking of empirical p-values across species with the expected null distribution to identify genes showing stronger cross-species association signals than expected by chance (Booker et al. 2023; Whiting et al. 2024). P-values were estimated using an iterative permutation procedure, in which the number of permutations was increased only for genes with very small p-values to improve estimation accuracy while reducing computational demand. Resulting p-values were adjusted for multiple testing using the false discovery rate (FDR), producing pooled q-values across all genes. Analyses were parallelised across CPU cores using the foreach (v.1.5.2, Folashade, Hong, et al. 2022) and doParallel packages (v.1.0.17, Folashade, Weston, et al. 2022). The workflow was applied separately for each environmental variable and GEA method (LFMM and CA). With four species included in the cross-species analysis, a pooled q-value threshold of below 0.25 was used to identify repeated adaptive signals. PicMin has been shown to perform well with four or more species, with candidate signals supported across all species providing stronger evidence for repeated adaptation (Booker et al. 2023). A more stringent threshold than the q-value < 0.5, used in a previous analysis, was applied here because that study included 25 species (Whiting et al. 2024).

3.5.3 Gene Set Enrichment Analysis

To investigate the functional context of the cross-species PicMin results, Gene Set Enrichment Analysis (GSEA) was performed using SetRank (v.1.1.0, Simillion 2016). SetRank was selected because it accounts for overlap and hierarchical relationships between Gene Ontology (GO) terms, reducing potential false-positive enrichment caused by redundant gene sets (Hallal et al. 2021; Simillion et al. 2017).

GO annotations for Biological Process (BP), Molecular Function (MF), and Cellular Component (CC) were downloaded from UniProt (Boeckmann et al. 2003) for *D. rerio* on 29 April 2026 and linked to Ensembl gene identifiers. Separate GO gene set collections were generated for the LFMM- and CA-derived PicMin results. For each environmental variable, genes with a PicMin pooled q-value below 0.25 were used as the candidate gene set. Environmental variables without any genes below this threshold were excluded. SetRank was run in its unranked mode separately for BP, MF, and CC gene sets and for the LFMM- and CA-derived candidate sets for each environmental variable. Gene sets with an initial p-value above 0.01 were removed. Among the remaining gene sets, SetRank accounted for overlap between GO terms and removed terms whose significance was explained by other overlapping gene sets (Simillion et al. 2017). GO terms with a Benjamini-Hochberg FDR-corrected q-value below 0.05 were considered significantly enriched (Appendix: 1.3.9 Gene Set Enrichment Analysis).

3.5.4 Visualisation of Candidate SNP-Environment Associations

To visualise candidate signals, the SNP with the lowest p-value within the gene of interest was selected for each species and GEA method. Alternate allele frequencies were calculated for each sampling site from the imputed genotype matrix as the mean genotype dosage divided by two. Environmental values were assigned to sampling sites, and allele frequencies and environmental values were visualised in a map to assess their geographic covariation. Genotype-specific environmental variation was additionally visualised in a box plot by comparing environmental values across the three genotype classes (0, 1 and 2) for each selected SNP. Summary figures were generated in R using the ggplot2, ggspatial (v1.1.10, Dunnington et al. 2025), ggnewscale (v.0.5.2, Campitelli 2025), rnaturalearth (v.1.2.0, Massicotte and South 2026), and patchwork packages, with dplyr, tibble (v.3.3.1, Müller and Wickham 2026), and purrr (v.1.2.2, Wickham and Henry 2026) used for data processing and restructuring. Gene names corresponding to selected Ensembl gene IDs were identified by querying the Ensembl gene IDs against the UniProt database (Boeckmann et al. 2003).

4 Results

4.1 Exploratory and Preliminary Analysis

The table below summarises trimming parameters, post-QC dataset characteristics, and preliminary analyses of data quality, population structure, and linkage disequilibrium decay.

Table 3: Summary of trimming parameters, post quality control datasets, and key results from the exploratory and preliminary analyses for each reef fish species. Corresponding figures are provided in the Supplementary Material.

Species	***Amphiprion bicinctus*** twoband anemonefish	***Epinephelus striatus*** nassau grouper	***Pterois volitans*** red lionfish	***Siphamia tubifer*** urchin cardinalfish
Trimming Parameter				
Trimming 3’ [bp]	0	0	5	0
Trimming 5’ [bp]	6	10	10	10
Stringency Adaptor Contamination	3	1	3	3

Species	***Amphiprion bicinctus*** twoband anemonefish	***Epinephelus striatus*** nassau grouper	***Pterois volitans*** red lionfish	***Siphamia tubifer*** urchin cardinalfish
Post Quality Check Dataset				
Remaining Samples	113	90	99	275
Remaining SNPs	19’675	19’398	12’679	16’127
Exploratory Analyses Results				
PCoA Clusters	3	0	0	0
Technical Correlation Visible	Yes: GC content (PC1: R = −0.43; PC2: R = 0.42)	mapped reads, GC content, read length, coverage: |R| ≈ 0.1–0.6	No	No
Variation explained by PC1	5,5%	1.4%	2%	0.5%
Genetic Groups (F_{st})	3	0	0	0
F_{st} range	0 – 0.04	-0.0005 – 0.0015	-0.0025 – 0.0005	-0.0025 – 0.0005
Preliminary Analyses Results				
Ancestral Populations (K)	3	1	2	1
LD decay estimation	±9 kb	±4 kb	±11 kb	±0.5 kb

Trimming parameters differed only slightly among species. All species were trimmed by 10 bp at the five prime end, except *A. bicinctus* (6 bp), while three prime trimming was only applied to *P. volitans* (5 bp). Adapter trimming stringency was set to 3 for all species except *E. striatus*, where a value of 1 was used. Following read quality control and filtering 113 *A. bicinctus*, 90 *E. Striatus*, 99 *P. volitans* and 275 *S. tubifer* samples were retained for downstream analyses, comprising 19,675, 19’398, 12,679, and 16,127 biallelic SNPs, respectively, after filtering for genotype missingness below 10 percent, a minor allele frequency above 5 percent, and linkage disequilibrium pruning (R^2 = 0.9, 1000 bp window).

Exploratory analyses showed contrasting patterns of population structure among the three species. No distinct genetic clusters were detected for *E. striatus*, *P. volitans* or *S. tubifer* in the PCoA (Appendix: 1.3.2 PCoA Analysis), consistent with the very low pairwise F_{st} values observed in all three species (max $F_{st} \leq \pm 0.0025$), indicating little to no genetic differentiation among geographical sampling locations. In contrast, *A. bicinctus* exhibited three distinct genetic clusters that were supported by pairwise F_{st} analysis (F_{st} range 0 to 0.04), indicating the presence of two to three genetically differentiated populations (Appendix: 1.3.3 F_{st} Analysis). PC1 explained 5.5% of the genetic variation in *A. bicinctus*, compared with 1.4%, 2.0% and 0.5% in *E. striatus*, *P. volitans* and *S. tubifer*, respectively. For *A. bicinctus*, GC content showed a moderate correlation with the first two PC axes, whereas no other quality control metrics were associated with the observed genetic structure (S Figure 5). *E. striatus* showed weak to moderate correlations between the first two PC axes and mapped reads, GC content, read length, and sequencing coverage (S Figure 8). No evidence of technical bias was detected for *P. volitans* or *S. tubifer* (S Figure 9; S Figure 10).

The ancestral population analyses were largely consistent with the exploratory analyses. *A. bicinctus* was assigned to the optimal number of three ancestral populations (K = 3), corresponding to the distinct genetic clusters identified previously, *S. tubifer* and *E. striatus* were assigned a single ancestral population (K = 1), consistent with the absence of detectable population structure, and *P. volitans* was assigned two ancestral populations (K = 2) based on the lowest cross entropy, despite the limited genetic differentiation observed in the exploratory analyses (Appendix: 1.3.4 Ancestral Population Analysis). Linkage disequilibrium decay analysis estimated decay distances of approximately 9 kb for *A. bicinctus*, 4 kb for *E. striatus*, 11 kb for *P. volitans* and 0.5 kb for *S. tubifer*. As LD decay estimates ranged from 0.5 to 11 kb, a consistent window of 10 kb was selected for SNP-to-gene assignment. This encompassed the estimated 95% LD decay distance for three species (*A. bicinctus*, *E. striatus* and *S. tubifer*), while the estimate for *P. volitans* only slightly exceeded the selected window (Appendix: 1.3.5 Linkage Disequilibrium Decay Analysis).

4.2 Genotype-Environment Analysis

GEA analyses identified significant SNP-environment associations (q-value < 0.05) across all four species. Overall, CA identified more associations than LFMM, with a higher median number of associations per environmental variable across species (CA: median = 782.5, IQR = 2584; LFMM: median = 69, IQR = 173.8). The number of significant SNP-environment associations varied among species and environmental variables (Figure 3). *A. bicinctus* showed the highest number of associations, with a median of 62.5 SNPs per environmental variable (interquartile range: IQR = 168.2) for LFMM and 779.5 (IQR = 2584) for CA, followed by *E. striatus* (LFMM: median = 1, IQR = 3; CA: median = 0, IQR = 0), *P. volitans* (LFMM: median = 1, IQR = 4; CA: median = 0, IQR = 0) and *S. tubifer* (LFMM and CA: median = 0, IQR = 0) (S Table 9).

A. bicinctus

For *A. bicinctus*, CA identified significant SNP associations for average sea-surface salinity (3,835 SNPs), average pH (3,565), standard deviation of chlorophyll concentration (3,170), average chlorophyll concentration (3,133), standard deviation of oxygen concentration (3,045), standard deviation of phosphate concentration (3,006), standard deviation of sea-surface salinity (2,900), standard deviation of pH (2,865), average nitrate concentration (2,558), average phosphate concentration (2,545), average degree heating weeks (2,196), standard deviation of nitrate concentration (2,174), standard deviation of degree heating weeks (1,205), average sea-water velocity (878), standard deviation of sea-water velocity (824), average oxygen concentration (823), standard deviation of suspended matter (736), standard deviation of boat-detection frequency (625), average sea-surface temperature (617), depth (468), standard deviation of sea-surface temperature (398), average iron concentration (320), average suspended matter (111), average cropland density (55), standard deviation of population density (38), standard deviation of built-up area density (35), standard deviation of cropland density (18), land-surface density (10), standard deviation of iron concentration (6), and average built-up area density (2).

LFMM identified significant SNP associations for standard deviation of nitrate concentration (671), average nitrate concentration (557), standard deviation of oxygen concentration (540), average sea-water velocity (538), average phosphate concentration (501), average chlorophyll concentration (235), standard deviation of chlorophyll concentration (233), standard deviation of sea-water velocity (201), standard deviation of cropland density (182), average cropland density (137), standard deviation of suspended matter (115), average iron concentration (113), average sea-surface temperature (111), average population density (99), depth (70), standard deviation of built-up area density (69), standard deviation of sea-surface salinity (56), standard deviation of population density (52), average built-up area density (51), land-surface density (40), average suspended matter (38), average sea-surface salinity (27), average boat-detection frequency (22), average oxygen concentration (21), average pH (11), standard deviation of pH (8), standard deviation of boat-detection frequency (7), standard deviation of phosphate concentration (2), average degree heating weeks (1), and standard deviation of iron concentration (1).



Figure 3: Summary of GEA analyses across the four study species. For each environmental variable, bars represent the number of significant SNPs (q-value < 0.05) detected by LFMM or by CA.
Environmental variables: CHL (Chlorophyll [mg/m³]), DHW (Degree Heating Weeks [°C·week⁻¹]), FE (Iron [mmol/m³]), O_2 (Oxygen [mmol/m³]), PH (pH), NO_3 (Nitrate [mmol/m³]), PO_4 (Phosphate [mmol/m³]), SCV (Sea-water velocity [m/s]), SPM (Suspended matter [g/m³]), SSS (Sea-surface salinity [10^{-3}]), SST (Sea-surface temperature [°C]), Bathy (Depth [m]), CROP (Cropland density [%]), LAND (Land-surface density [%]), PDEN (Population density [hab km^{-2}]), URBA (Built-up area density [%]), VBD (Boat-detection frequency [%]).

E. striatus

For *E. striatus*, CA identified one significant SNP association for standard deviation of cropland density and one for land-surface density.

LFMM identified significant SNP associations for standard deviation of population density (11), average population density (9), standard deviation of sea-surface temperature (8), average built-up area density (8), average boat-detection frequency (7), standard deviation of cropland density (5), average nitrate concentration (5), average chlorophyll concentration (3), standard deviation of degree heating weeks (3), standard deviation of oxygen concentration (3), standard deviation of sea-water velocity (2), average sea-surface salinity (2), standard deviation of boat-detection frequency (2), average degree heating weeks (1), land-surface density (1), average oxygen concentration (1), standard deviation of sea-surface salinity (1), and standard deviation of built-up area density (1).

P. volitans

For *P. volitans*, CA identified one significant SNP association for average sea-surface salinity.

LFMM identified significant SNP associations for standard deviation of boat-detection frequency (9), average sea-water velocity (8), standard deviation of built-up area density (7), average population density (5), standard deviation of population density (5), average sea-surface temperature (5), average boat-detection frequency (5), average cropland density (4), standard deviation of cropland density (4), standard deviation of chlorophyll concentration (3), standard deviation of sea-water velocity (3), average phosphate concentration (2), standard deviation of phosphate concentration (2), average built-up area density (2), average iron concentration (1), land-surface density (1), standard deviation of nitrate concentration (1), and standard deviation of pH (1).

S. tubifer

For *S. tubifer*, CA identified significant SNP associations for standard deviation of suspended matter (6), average boat-detection frequency (4), and standard deviation of sea-water velocity (2).

LFMM identified significant SNP associations for standard deviation of suspended matter (4), standard deviation of built-up area density (3), average boat-detection frequency (3), standard deviation of chlorophyll concentration (1), average cropland density (1), and standard deviation of sea-surface salinity (1).

SNP-Environment Associations in All Species

Significant SNP associations were identified in all four species for standard deviation of sea-water velocity (*A. bicinctus* = 1,025; *E. striatus* = 2; *P. volitans* = 3; S. tubifer = 2), standard deviation of built-up area density (*A. bicinctus* = 104; *E. striatus* = 1; *P. volitans* = 7; *S. tubifer* = 3), and average boat-detection frequency (*A. bicinctus* = 22; *E. striatus* = 7; *P. volitans* = 5; *S. tubifer* = 7), irrespective of the GEA method.

4.3 Cross-Species Analysis

Overall, PicMin identified 14 genes (pooled q-value < 0.25) showing repeated adaptive signals in the LFMM GEA results across three environmental variables: average sea-surface temperature (10 genes), standard deviation of iron concentration (3), and standard deviation of built-up area density (1). Of these, three were supported by two species, eight by three species, and three by all four species. For the CA GEA results, PicMin identified five genes across four environmental variables: standard deviation of phosphate concentration (2), standard deviation of iron concentration (1), land-surface density (1), and standard deviation of cropland density (1). Of these, four were supported by two species and one by all four species.

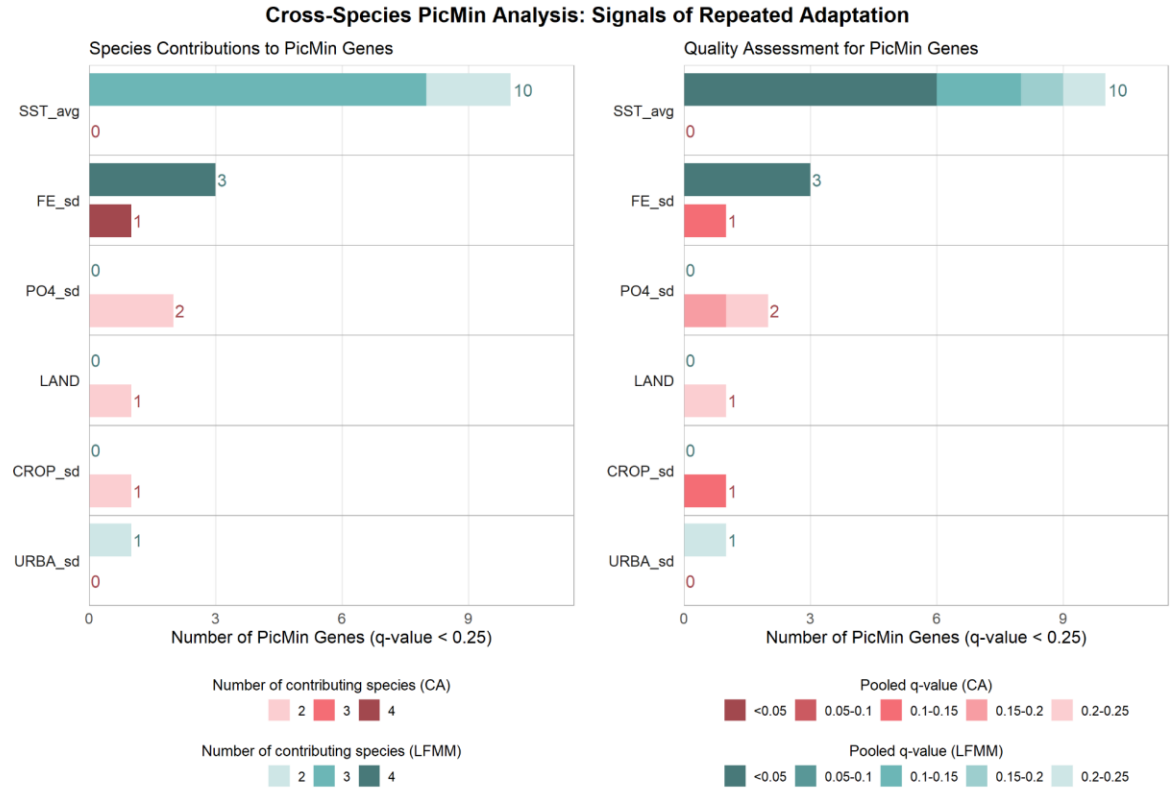


Figure 4: From left to right: (left) Number of PicMin genes indicating repeated adaptation per environmental variable based on GEA method LFMM or CA results. Bars show genes with a pooled q-value < 0.25, with colour intensity indicating the number of contributing species. (right) Quality assessment of PicMin genes indicating repeated adaptation based on pooled q-values from LFMM and CA GEA method, with darker colours indicating lower q-values.
Environmental variables: SST (Sea-surface temperature [°C]), FE (Iron [mmol/m³]), PO_4 (Phosphate [mmol/m³]), LAND (Land-surface density [%]), CROP (Cropland density [%]), URBA (Built-up area density [%]).

Average sea-surface temperature

The cross-species analysis, using the LFMM GEA inputs, identified 10 genes showing repeated signals of adaptation for average sea-surface temperature. Of these, eight were supported by three species and two by two species. Six genes had pooled q-values < 0.05, with five having q-values of 0.044 and one having a q-value of 0.046. For these six genes (ENSDARG00000092941, ENSDARG00000093323, ENSDARG00000094297, ENSDARG00000094346, ENSDARG00000098706, ENSDARG00000092294), the species-specific GEA p-values correspond to the same SNPs with the lowest p-value within these genes, and were 5.919E-6 for *A. bicinctus*, 0.00119 for *E. striatus*, 0.04413 for *P. volitans* and NA for *S. tubifer*. The remaining four genes had pooled q-values of 0.138 (two genes), 0.184, and 0.221. Gene (ENSDARG00000094608) with the pooled q-value of 0.138 supported by three species had the same species-specific p-values as the six genes mentioned above: 5.919E-6 for *A. bicinctus*, 0.00119 for *E. striatus*, 0.04413 for *P. volitans* and NA for *S. tubifer*. The other gene (ENSDARG00000005972) also with a q-value of 0.138, supported by two species, had p-values of 0.06901 for *A. bicinctus*, NA for *E. striatus*, 0.00067 for *P. volitans*, and 0.00627 for *S. tubifer*. The two genes (ENSDARG00000091026, ENSDARG00000098183) with q-values of 0.184 and 0.221 had species-specific p-values of 3.318E-9, 0.00119, 0.13881, and 0.01763, and 3.318E-9, 0.48331, 0.01332, and 0.01763, for *A. bicinctus*, *E. striatus*, *P. volitans* and *S. tubifer* respectively (Table 4).

Standard deviation of iron concentration

The cross-species analysis, using the LFMM GEA inputs, identified three genes showing repeated signals of adaptation for standard deviation of iron concentration. All three genes were supported by four species. Two genes (ENSDARG00000029146, ENSDARG00000097827) had a pooled q-value of 0.019, while one gene (ENSDARG00000004789) had a pooled q-value of 0.037. All three genes had the same species-specific GEA p-values: 0.00882 for *A. bicinctus*, 0.02575 for *E. striatus*, 0.04682 for *P. volitans*, and 0.04629 for *S. tubifer*. Using the CA GEA inputs, the cross-species analysis identified one gene (ENSDARG00000029146) supported by all four species, with a pooled q-value of 0.111. The corresponding species-specific GEA p-values were 0.00593 for *A. bicinctus*, 0.01509 for *E. striatus*, 0.12059 for *P. volitans*, and 0.08858 for *S. tubifer*.

Standard deviation of phosphate concentration

The cross-species analysis, using the CA GEA inputs, identified two genes showing repeated signals of adaptation for standard deviation of phosphate concentration. Both genes were supported by two species. ENSDARG00000075962 had a pooled q-value of 0.166, with species-specific GEA p-values of 1.098E-9 for *A. bicinctus*, 0.00257 for *E. striatus,* 0.18213 for *P. volitans*, and NA for *S. tubifer*. ENSDARG00000115482 had a pooled q-value of 0.221, with species-specific GEA p-values of 0.214777 for *A. bicinctus*, 0.00315 for *E. striatus*, 0.00129 for *P. volitans*, and NA for *S. tubifer*.

Land-surface density

For land-surface density, the cross-species analysis using the CA GEA inputs identified one gene (ENSDARG00000075962) showing a repeated signal of adaptation. The gene was supported by two species and had a pooled q-value of 0.221. The corresponding species-specific GEA p-values were 7.974E-7 for *A. bicinctus*, 0.00095 for *E. striatus*, 0.43223 for *P. volitans*, and NA for *S. tubifer*.

Standard deviation of cropland density

For standard deviation of cropland density, the cross-species analysis using the CA GEA inputs identified one gene (ENSDARG00000077656) showing a repeated signal of adaptation. The gene was supported by two species and had a pooled q-value of 0.111. The corresponding species-specific GEA p-values were NA for *A. bicinctus*, 0.00217 for *E. striatus*, 0.16675 for *P. volitans*, and 0.00945 for *S. tubifer*.

Standard deviation of built-up area density

The cross-species analysis, using the LFMM GEA inputs, identified one gene (ENSDARG00000061233) showing a repeated signal of adaptation for standard deviation of built-up area density. The gene was supported by two species and had a pooled q-value of 0.221. The corresponding species-specific GEA p-values were NA for *A. bicinctus*, 0.00039 for *E. striatus*, 0.00139 for *P. volitans*, and 0.10628 for *S. tubifer*.

Table 4: PicMin genes with pooled q-values < 0.25 identified from the LFMM and CA genotype-environment association (GEA) analyses. The table shows the environmental variable, Ensembl gene ID, PicMin pooled q-value, number of contributing species (n_est), and species-specific GEA p-values for each repeated adaptive signal.
Environmental variables: SST (Sea-surface temperature [°C]), FE (Iron [mmol/m³]), PO_4 (Phosphate [mmol/m³]), LAND (Land-surface density [%]), CROP (Cropland density [%]), URBA (Built-up area density [%]).

Environmental Variable	ENSEMBL Gene ID	PicMin q-value (pooled q-value)	contributing species (n_est)	*A. bicinctus* p-value	*E. striatus* p-value	*P. volitans* p-value	*S. tubifer* p-value
Correlation Analysis (CA)							
FE_sd	ENSDARG00000029146	0.110699	4	0.00593	0.01509	0.12059	0.08858
CROP_sd	ENSDARG00000077656	0.110699	2	NA	0.00217	0.16675	0.00945
PO4_sd	ENSDARG00000075962	0.166048	2	1.098E-9	0.00257	0.18213	NA
PO4_sd	ENSDARG00000115482	0.221398	2	0.214777	0.00315	0.00129	NA
LAND	ENSDARG00000075962	0.221398	2	7.974E-7	0.00095	0.43223	NA
Latent Factor Mixed Model (LFMM)							
FE_sd	ENSDARG00000029146	0.019372	4	0.00882	0.02575	0.04682	0.04629
FE_sd	ENSDARG00000097827	0.019372	4	0.00882	0.02575	0.04682	0.04629
FE_sd	ENSDARG00000004789	0.0369	4	0.00882	0.02575	0.04682	0.04629
SST_avg	ENSDARG00000092941	0.04428	3	5.919E-6	0.00119	0.04413	NA
SST_avg	ENSDARG00000093323	0.04428	3	5.919E-6	0.00119	0.04413	NA
SST_avg	ENSDARG00000094297	0.04428	3	5.919E-6	0.00119	0.04413	NA
SST_avg	ENSDARG00000094346	0.04428	3	5.919E-6	0.00119	0.04413	NA
SST_avg	ENSDARG00000098706	0.04428	3	5.919E-6	0.00119	0.04413	NA
SST_avg	ENSDARG00000092294	0.046125	3	5.919E-6	0.00119	0.04413	NA
SST_avg	ENSDARG00000094608	0.138361	3	5.919E-6	0.00119	0.04413	NA
SST_avg	ENSDARG00000005972	0.138361	2	0.06901	NA	0.00067	0.00627
SST_avg	ENSDARG00000091026	0.184482	2	3.318E-9	0.00119	0.13881	0.01763
SST_avg	ENSDARG00000098183	0.221378	3	3.318E-9	0.48331	0.01332	0.01763
URBA_sd	ENSDARG00000061233	0.221398	2	NA	0.00039	0.00139	0.10628

Gene Ontology enrichment analysis identified seven enriched terms in the LFMM-derived gene sets and five terms in the CA-derived gene sets (S Table 10). In both analyses, standard deviation of iron concentration was associated with receptor-mediated endocytosis (GO:0006898), regulation of vasculature development (GO:1901342), clathrin-coated pit (GO:0005905), and organelle (GO:0043226), all mapped to ENSDARG00000029146 (*lrp1ab*). Additional enrichment was identified for the standard deviation of built-up area density in the LFMM analysis, while the CA analysis identified enrichments for the standard deviation of phosphate concentration and land-surface density, both associated with the same gene and GO terms (ENSDARG00000075962) (S Table 10).

ENSDARG00000029146 (*lrp1ab*) showed a repeated adaptive signal for standard deviation of iron concentration in both LFMM (q-value = 0.019) and CA (q-value = 0.11) results, supported by all four species.

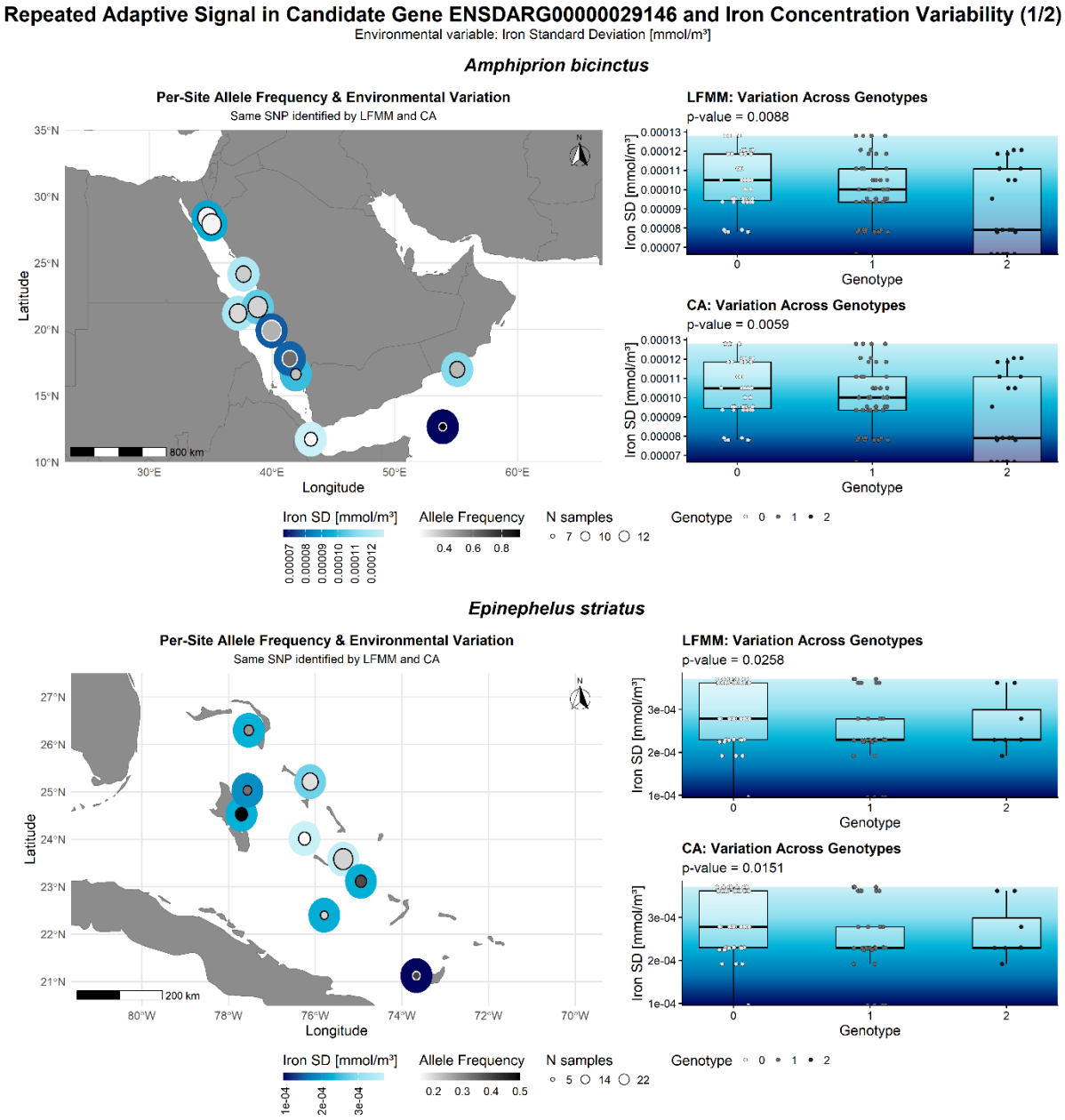


Figure 5: Potential parallel adaptive signal associated with lrp1ab (ENSDARG00000029146) and standard deviation of iron concentration [mmol/m³] across four reef fish species (A. bicinctus, E. striatus, P. volitans and S. tubifer). Results are shown for A. bicinctus and E. striatus. For each species, the left panel shows the spatial distribution of site-specific allele frequencies for the most contributing SNP identified by both methods, together with iron concentration variability; circle size indicates the number of samples per site. The top-right and bottom-right panels show environmental variation across genotypes (0, 1, 2) for LFMM and CA, respectively, including the p-values of the most strongly associated SNP within the candidate gene.

In *A. bicinctus* the same SNP (NC_072770.1_24915445) was identified for the LFMM (p-value = 0.00882) and CA (p-value = 0.00593) methods. Alternate allele frequencies varied among sampling sites, with higher allele frequency values associated with smaller variation. The genotype plots showed differences in iron concentration variability among genotypes, with the highest standard deviation values observed for genotype 0, intermediate values for genotype 1, and the lowest values for genotype 2.

The same SNP (CM069301.1_21017502) was identified for the LFMM (p-value = 0.02575) and CA (p-value = 0.01509) methods in *E. striatus*. Alternate allele frequencies varied among sampling sites, showing a similar spatial pattern to that observed in *A. bicinctus*. The genotype plots showed differences in iron concentration variability among genotypes, with slightly higher standard deviation values observed for genotype 0 and intermediate values for genotype 1 and 2.

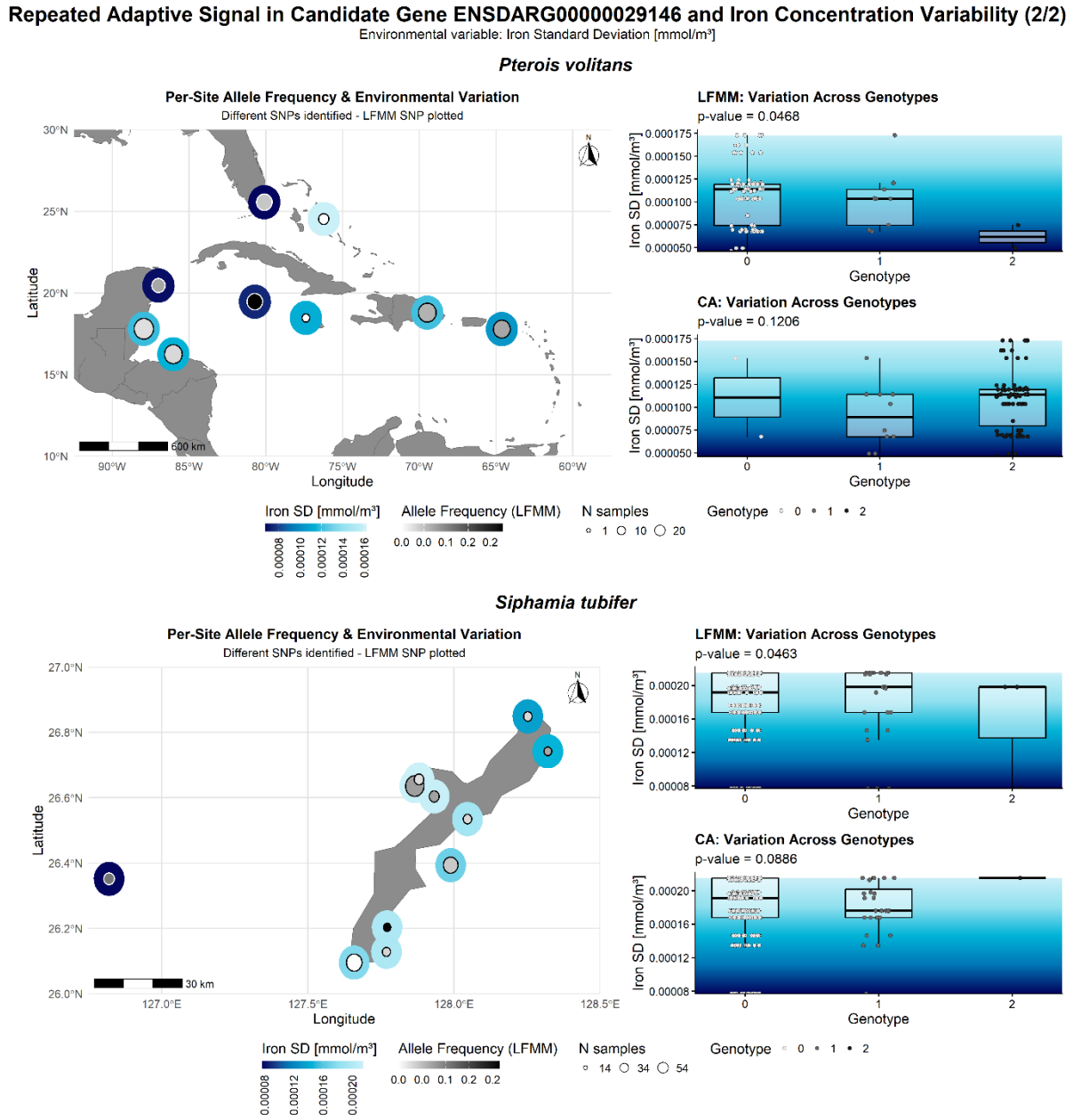


Figure 6: Potential parallel adaptive signal associated with lrp1ab (ENSDARG00000029146) and standard deviation of iron concentration [mmol/m³] across four reef fish species (A. bicinctus, E. striatus, P. volitans and S. tubifer). Results are shown for P. volitans and S. tubifer. For each species, the left panel shows the spatial distribution of site-specific allele frequencies for the most contributing SNP identified by both methods, together with iron concentration variability; circle size indicates the number of samples per site. The top-right and bottom-right panels show environmental variation across genotypes (0, 1, 2) for LFMM and CA, respectively, including the p-values of the most strongly associated SNP within the candidate gene.

For *P. volitans*, different SNPs were identified for the LFMM (CAMQGK010000637.1_21172188, p-value = 0.04682) and CA (CAMQGK010000559.1_1973554, p-value = 0.12059) methods. Alternate allele frequencies for the LFMM SNP were relatively low as mainly genotype 1 was present. The spatial pattern between allele frequency and environmental variability was less consistent. Genotype plots revealed differences in the variability of iron concentrations among the genotypes for both methods. For LFMM, the same pattern was observed, with genotype 0 showing the highest standard deviation values, genotype 1 intermediate values, and genotype 2 the lowest values. Genotype 0 was predominant, whereas genotypes 1 and 2 were represented by fewer samples. For CA, genotype 2 was most frequent and showed the highest standard deviation values, while genotype 1 showed intermediate values and genotype 0 showed the lowest values.

In *S. tubifer* different SNPs were identified for the LFMM (CM035789.1_23617679, p-value = 0.04629) and CA (CM035787.1_34801499, p-value = 0.08858) methods. Alternate allele frequencies were generally low across sampling locations and showed weaker and less consistent spatial patterns in relation to environmental variability. In both SNPs, genotypes 0 and 1 were associated with higher standard deviation values, although genotype 1 occurred at slightly lower values for CA compared with LFMM. Genotype 2 was rarely observed in either SNP and was associated with medium to higher standard deviation values.

5 Discussion

Overall, the results indicate that all four reef fish species show both species-specific and repeated adaptive associations with global change factors. GEA identified significant SNP-environment associations in all four species (*A. bicinctus, E. striatus, P. volitans* and *S. tubifer*), while the cross-species analysis revealed repeated adaptive signals for six environmental variables. Most notably, homologs of zebrafish ENSDARG00000029146 (*lrp1ab*) showed a repeated association with variability in iron concentration across all four species and both GEA methods. This repetition is very unlikely to happen by chance, so this gene-environment association was therefore further investigated and discussed.

5.1 Repeated Adaptive Signal at *lrp1ab* with Iron Variability

Iron is an essential trace element in fish, with most body iron incorporated into haemoglobin and therefore required for oxygen transport and cellular respiration (Buyinza et al. 2024; Lall and Kaushik 2021). It has important roles in metabolism, cell division and gene regulation (Lemos and Balado 2020; Buyinza et al. 2024). Marine fish acquire iron mainly through their diet via the gastrointestinal tract, with a smaller contribution from uptake through the gills (Buyinza et al. 2024; Khawar et al. 2024; Lall and Kaushik 2021). Both low and high iron availability can have negative effects. Iron deficiency can reduce haemoglobin production and cause microcytic anaemia, impair immune function and alter tissue structure, whereas excessive iron can disrupt iron homeostasis, affect immune function and tissue structure, reduce growth, and cause oxidative damage, liver dysfunction and, at high environmental concentrations, impaired gill function (Buyinza et al. 2024; Feng et al. 2025; Guo et al. 2017; Lall and Kaushik 2021). Some of these effects have been demonstrated experimentally in young grass carp (*Ctenopharyngodon idella*), where iron deficiency and excess impaired immune function and tissue structure in the head, kidney, spleen and skin, with deficiency additionally increasing skin haemorrhage and lesion severity, while excess iron reduced growth (Guo et al. 2017). Therefore, maintaining appropriate iron levels is important for normal physiological function and fish health (Lall and Kaushik 2021).

Marine iron originates from multiple sources, including glacial meltwater and dust, sediment weathering, and hydrothermal activity (Conway and John 2014; Jones et al. 2025; Shoenfelt et al. 2018). Iron distribution varies across geographic regions and throughout the water column, influenced by dust deposition, ocean circulation and mixing, upwelling of deeper waters, and the properties of different water masses (Bergquist et al. 2007; Misumi et al. 2013). Iron availability is closely linked to nutrient cycling and can limit phytoplankton growth, thereby influencing ocean primary productivity and broader marine biogeochemical cycles (Hutchins and Boyd 2016; Tagliabue et al. 2017). Climate change may further alter iron availability through changes in ocean warming, stratification, circulation, deoxygenation, and other processes affecting its sources, transport and recycling (König and Tagliabue 2023; Hutchins and Boyd 2016; Misumi et al. 2013). For example, a study using global ocean model simulations under recent and future climate scenarios, found substantial variability in dissolved iron isotope composition associated with ENSO-driven changes in circulation and productivity (König and Tagliabue 2023). These findings indicate that climate change may alter not only mean iron availability but also its temporal and spatial variability, providing a potential context for the observed association between *lrp1ab* and variation in environmental iron concentration.

lrp1ab belongs to the low-density lipoprotein receptor related protein (LRP) family, whose members function as multi ligand endocytic receptors involved in lipoprotein and micronutrient transport, as well as cell signalling. In zebrafish, LRP proteins have also been linked to developmental and tissue specific processes (Mao et al. 2017; Zhai et al. 2023). Zebrafish macrophages co-express *lrp1ab* and c1qa, which interact with calreticulin to form a phagocytic signalling complex involved in haematopoietic stem cell clearance (Dai et al. 2026). More broadly, zebrafish macrophages contribute to tissue repair and regeneration by clearing dead cells and pathogens, regulating inflammation, remodelling the extracellular matrix, and promoting cell proliferation and tissue recovery (Zhao et al. 2024). Consistent with a role in macrophage-mediated tissue maintenance, *lrp1ab* is also upregulated in Mac1 macrophages during tissue regeneration, alongside genes involved in blood vessel formation, cellular debris clearance, and extracellular matrix organisation, including *vegfaa/bb*, *elmo1*, and *fn1a* (Wei et al. 2023). Together, this evidence suggests that *lrp1ab* may link its endocytic and macrophage functions to processes relevant to iron recycling and homeostasis, providing a potential biological link to the observed association with iron concentration variability. Although this relationship remains speculative, the involvement of *lrp1ab* in macrophage-mediated clearance and tissue maintenance suggests that it could represent a potential target of selection in environments with varying iron conditions.

Another possibility is that iron variability reflects broader ecological processes rather than acting as a direct selective pressure itself. Iron availability can influence marine primary production and microbial activity, particularly in iron-limited regions (Hutchins and Boyd 2016; Misumi et al. 2013), while also playing an important role in host-pathogen interactions. Bacterial fish pathogens possess specialised iron-acquisition systems, some of which contribute to virulence and host colonisation (Lemos and Balado 2020). Thus, variation in iron availability could reflect differences in microbial or pathogen-mediated conditions that impose selection on host immune pathways. This provides an alternative explanation for the repeated adaptive signal with *lrp1ab*, given its involvement in macrophage function and immune regulation (Dai et al. 2026; Wei et al. 2023).

Two points should be considered when interpreting this result. First, the ±10 kb window used for SNP to gene assignment points to an approximate genomic region rather than a precise location of the causal variant. The *lrp1ab* association should therefore be interpreted as a signal from SNPs within this genomic window, rather than evidence that a particular SNP is involved in adaptation. Second, the two other genes (ENSDARG00000097827, q-value = 0.019; ENSDARG00000004789, q-value = 0.037) identified for the standard deviation of iron concentration using the LFMM-derived inputs were supported by the same most contributing SNPs as *lrp1ab* in all four species. This suggests that the three genes are located in close proximity and may represent a shared genomic signal rather than three independent associations. However, the functional roles of these two genes could not be determined.

5.2 Repeated Adaptive Signals for Average Sea Surface Temperature

Seven out of ten genes associated with average sea surface temperature (ENSDARG00000092941, ENSDARG00000093323, ENSDARG00000094297, ENSDARG00000094346, ENSDARG00000098706, ENSDARG00000092294, ENSDARG00000094608) were supported by the same most contributing SNPs across three species, suggesting that these genes are located in close proximity and may represent a shared genomic signal. However, no enriched GO terms were identified for these genes in the GSEA (S Figure 28). Manual search using UniProt provided functional information for only one gene (ENSDARG00000098183, si:ch211-203c5.3), which was associated with ATP binding (GO:0005524). As this term was not identified as significantly enriched in the GSEA and functional annotation for the region was otherwise limited, the signal was not further investigated.

5.3 Methodological Considerations

A methodological consideration is the use of different reference genomes. SNP calling was performed against a species-specific reference genome for each species (Table 1). Only *S. tubifer* was mapped to its own assembly, whereas *A. bicinctus, E. striatus and P. volitans* were mapped to congeneric species (*Amphiprion ocellaris, Epinephelus awoara* and *Pterois miles*, respectively). NextGenMap was used because it performs well when mapping reads to a reference genome from a closely related species (Sedlazeck et al. 2013). The *P. miles* assembly was also the only reference assembled at the contig rather than chromosome level (Kitsoulis et al. 2023).

Gene assignment for the cross-species analysis was performed separately using tblastn, by aligning homologous *D. rerio* protein sequences to the reference genome of each species. This provided a common annotation framework across all four species, independent of the reference genome used for SNP calling. However, this approach may introduce some uncertainty in gene identity. Teleosts retain duplicated genes from the teleost-specific whole-genome duplication, and sequence similarity alone may not always distinguish between orthologous and paralogous gene copies (Glasauer and Neuhauss 2014; Zhao et al. 2024). Thus, although the same *D. rerio* gene identifier was used to represent corresponding signals across species, this does not guarantee that the identified sequences represent the same orthologous locus in every species.

A. bicinctus exhibited an excess of significant environmental associations in both CA and LFMM GEA analyses (Appendix: 1.3.1 *A. Bicinctus* - Special Supplementary Analysis). For CA, this may partly reflect that population structure is not explicitly modelled in the Kendall rank correlation analysis. LFMM accounts for population structure through latent factors and genomic inflation factor correction was additionally applied, but elevated inflation persisted. Population structure analyses identified three main genetic clusters, primarily associated with the Bab-el-Mandeb Strait and Farasan Islands (S Figure 2; S Figure 5), with cross-entropy supporting K = 3 (S Figure 3). Increasing the number of latent factors to K = 4 and K = 6 only slightly reduced inflation (S Figure 7), suggesting that population structure alone did not explain the observed pattern. The first two PC axes explained 5.5% and 2.7% of genetic variation and were strongly correlated with 12 of 32 environmental variables ($R^2 > 0.6$), including average seawater velocity ($R^2 = 0.904$), average nitrate concentration ($R^2 = 0.857$), and standard deviation of oxygen concentration ($R^2 = 0.822$) (S Table 4). This overlap between genetic and environmental variation may have contributed to persistent inflation by making population structure and environmental effects difficult to distinguish. Other factors may also have contributed. The longest assembled chromosome (NC_072778.1) showed LD across its full length (S Figure 17), although this may reflect the use of a congeneric reference genome rather than biological LD. In addition, correlations between the first two PC axes and GC content suggest a possible contribution of technical or genomic factors (S Figure 5). Given these unresolved sources of inflation, the disproportionately large number of GEA associations detected in *A. bicinctus* should be interpreted with greater caution than those in the other species. This highlights the value of the cross-species approach, as repeated signals across species can provide additional support for candidate adaptive associations that may be difficult to distinguish from population structure or p-value inflation in single-species GEA analyses.

5.4 Implications and Future Directions

The repeated association between *lrp1ab* and iron concentration variability provides a clear target for experimental validation. Future studies could test whether variation in environmental iron affects *lrp1ab* expression or macrophage function in marine fish by exposing individuals to different iron concentrations. For example, iron availability has been shown to affect iron uptake and expression of the iron regulatory gene *hepcidin1* in marine medaka larvae (*Oryzias melastigma*) (Wang and Wang 2016). Similar experiments measuring *lrp1ab* expression and macrophage responses could therefore provide a direct test of the biological link suggested by the present genomic result.

Future cross-species analyses would also benefit from more rigorous orthology assignment and improved functional annotation. OrthoFinder, for example, incorporates phylogenetic information to distinguish orthologues and paralogues and could provide a more robust framework than the best-hit approach used here (Emms and Kelly 2019). More rigorous orthology assignment would help determine whether repeated associations, such as the *lrp1ab* signal, correspond to the same orthologous locus across species. Improved functional annotation, in turn, could facilitate the interpretation of candidate regions that currently lack clear functional information.

Furthermore, future studies could assess whether SNPs associated with the same environmental variable across species converge on shared biological functions or pathways rather than the same homologous genes, providing a functional framework for identifying repeated adaptive responses. For example, the GEA-association with average boat-detection frequency identified across all four species could be investigated at the functional level to determine whether different genomic regions are involved in similar biological responses to boat activity.

Lastly, increasing the number of species would provide a further test of the repeatability of the identified associations and allow patterns to be assessed across a broader range of evolutionary lineages. In particular, comparisons with more distantly related reef fishes could determine whether repeated associations are restricted to teleost lineages or extend more broadly across reef-associated fishes.

6 Conclusion

This thesis used a comparative seascape genomics approach across four reef fish species (*A. bicinctus, E. striatus, P. volitans* and *S. tubifer*) to explore potential genomic adaptation to global change and identify environmental factors associated with these adaptive signals. Species-specific potential adaptive signals were identified in all four species, although their number and distribution varied among species and environmental variables. Associations with seawater velocity, built-up area density and boat activity were detected across all four species, indicating that both climatic and anthropogenic conditions may contribute to genomic responses to global change (4.2 Genotype-Environment Analysis).

Several environmental variables were also associated with repeated adaptive signals across multiple species. The strongest and most consistent signal involved zebrafish *lrp1ab* homologs and variability in iron concentration, with the association supported across all four species and by both analytical approaches. Sea-surface temperature was associated with several repeated signals, while weaker signals were identified for phosphate concentration, land-surface density and cropland density across fewer species (4.3 Cross-Species Analysis). Together, these findings indicate that both climatic and anthropogenic environmental conditions may contribute to repeated adaptive signals in reef fish species.

Overall, the identified associations should be considered candidate signals rather than confirmed evidence of adaptation, given the variation among species and the methodological and genomic limitations discussed. The repeated association between *lrp1ab* and iron concentration variability provides a promising target for future functional and experimental studies. More broadly, these findings demonstrate the potential of comparative approaches to identify candidate mechanisms underlying repeated genomic adaptation to global change in reef fish.

7 AI Disclosure

The drafting of the manuscript was supported by the free versions of ChatGPT, DeepL Write and the pro version of Claude. Literature searches were assisted by Consensus AI and Scite. Script development and troubleshooting for Bash and R were assisted by ChatGPT's free version and Claude's pro version. The author performed all scientific decisions, study design, data analyses and interpretation of the results.

8 Code Availability

All scripts and input run tables used in this thesis are available in the following GitHub repository: [https://github.com/narosc/MT_NOS_IASoRFtGC]. The repository contains the final scripts used for genomic data processing, genotype-environment association analyses, cross-species analyses, and functional annotation, as well as the input run tables.

9 Acknowledgment and Attributions

I would like to express my sincere gratitude to my supervisor, Dr. Oliver M. Selmoni, for his exceptional guidance and continuous support throughout the completion of this Master's thesis. He introduced me to the fields of bioinformatics and genomics, in which I had no prior experience, and provided the knowledge, guidance, and encouragement that allowed me to develop both scientifically and technically. His willingness to share his expertise, answer questions, and offer advice whenever I encountered difficulties was invaluable and played an important role in the successful completion of this Master’s thesis. Beyond the scientific aspects of the project, I am particularly grateful for his encouragement during periods when I struggled to balance my work with the commitments of this Master’s thesis. His support helped me maintain a healthier work-life balance throughout the whole one and a half years of working on this project. Finally, I am grateful to him for introducing me to the emerging field of seascape genomics, which sparked a strong interest in the field and has inspired me to pursue opportunities to continue learning and working in this area.

I would also like to thank the entire Spatial Genetics of Ecosystems research group at the Remote Sensing Laboratories in the Department of Geography at the University of Zurich. Their constructive feedback, insightful discussions, and continuous encouragement created a supportive working environment that was essential throughout the challenging and rewarding stages of this project. I particularly appreciated the critical questions and discussions that encouraged me to repeatedly reassess my methods and helped me develop a broader perspective on the research questions addressed in this Master’s thesis. I am especially grateful to Prof. Dr. Meredith Christine Schuman, who leads the Spatial Genetics Lab and fosters a respectful, supportive, and collaborative environment for conducting interdisciplinary research. Her leadership has contributed greatly to creating a space in which ideas can be critically discussed while encouraging the development of innovative research across disciplines.

I would also like to thank Dr. Claudia Röösli, who offered a valuable perspective beyond the genomic aspects of this project. Her questions and discussions encouraged me to step back from the technical details and reconsider the broader research questions rather than becoming absorbed in the many specialised aspects that could be further explored. This perspective was particularly valuable in helping me maintain a clear focus on the overall objectives of the Master’s thesis.

10 Bibliography

10.1 List of Figures

Figure 1: Detailed flowchart illustrating the overall data processing and analysis workflow. Abbreviations: VCF, Variant Calling File; LEA: R package (3.2.2 Imputation of Genotypes); GT, genotype; LD, linkage disequilibrium; D. rerio FASTA, Danio rerio FASTA-format sequence file; tblastn, BLAST command-line function; SD, standard deviation; LFMM, Latent Factor Mixed Model; FDR, false discovery rate; GEA, genotype-environment association; CA, correlation analysis; D. rerio UniProt Ref, UniProt protein reference/annotation file; D. rerio ENSEMBL linker, mapping file linking ENSEMBL gene IDs to UniProt protein IDs; D. rerio GTF, Danio rerio Gene Transfer Format; SNPs, single-nucleotide polymorphisms; env., environmental. Stouffer, statistical method for combining p-values (3.5.1 Mapping SNPs to Genes); PicMin, statistical method for identifying repeated adaptive signals across species (3.5.2 PicMin); GSEA: Gene Set Enrichment Analysis. 3

Figure 2: Global distribution of sampling sites for the four species of interest, with species specific maps showing the detailed geographic distribution of sampling locations for each species. 4

Figure 3: Summary of GEA analyses across the four study species. For each environmental variable, bars represent the number of significant SNPs (q-value < 0.05) detected by LFMM or by CA. 14

Figure 4: From left to right: (left) Number of PicMin genes indicating repeated adaptation per environmental variable based on GEA method LFMM or CA results. Bars show genes with a pooled q-value < 0.25, with colour intensity indicating the number of contributing species. (right) Quality assessment of PicMin genes indicating repeated adaptation based on pooled q-values from LFMM and CA GEA method, with darker colours indi-cating lower q-values. 16

Figure 5: Potential parallel adaptive signal associated with lrp1ab (ENSDARG00000029146) and standard deviation of iron concentration [mmol/m³] across four reef fish species (A. bicinctus, E. striatus, P. volitans and S. tubifer). Results are shown for A. bicinctus and E. striatus. For each species, the left panel shows the spatial distribution of site-specific allele frequencies for the most contributing SNP identified by both methods, together with iron concentration variability; circle size indicates the number of samples per site. The top-right and bottom-right panels show environmental variation across genotypes (0, 1, 2) for LFMM and CA, respectively, including the p-values of the most strongly associated SNP within the candidate gene. ... 19

Figure 6: Potential parallel adaptive signal associated with lrp1ab (ENSDARG00000029146) and standard deviation of iron concentration [mmol/m³] across four reef fish species (A. bicinctus, E. striatus, P. volitans and S. tubifer). Results are shown for P. volitans and S. tubifer. For each species, the left panel shows the spatial distribution of site-specific allele frequencies for the most contributing SNP identified by both methods, together with iron concentration variability; circle size indicates the number of samples per site. The top-right and bottom-right panels show environmental variation across genotypes (0, 1, 2) for LFMM and CA, respectively, including the p-values of the most strongly associated SNP within the candidate gene. ... 20

10.2 List of Tables

Table 1: Summary of genomic input data for reef fish species, including, geographical region, sampling time frame, the number of samples (N samples), number of sampling locations (N sites), their bioproject in NCBI, the reference genome and species and notes on manual adjustments. 5

Table 2: List of RECIFs environmental variables, including descriptions, available time periods, and original data sources. Data were accessed from RECIFs on 15th October 2025. 7

Table 3: Summary of trimming parameters, post quality control datasets, and key results from the exploratory and preliminary analyses for each reef fish species. Corresponding figures are provided in the Supplementary Material. 11

Table 4: PicMin genes with pooled q-values < 0.25 identified from the LFMM and CA genotype-environment association (GEA) analyses. The table shows the environmental variable, Ensembl gene ID, PicMin pooled q-value, number of contributing species (n_est), and species-specific GEA p-values for each repeated adaptive signal. 18

10.3 References

Alcaraz, Javier, Laura Anton-Sanchez, and Juan Francisco Monge. 2022. “The Concordance Test, an Alternative to Kruskal-Wallis Based on the Kendall-Tau Distance: An R Package.” *The R Journal* 14 (2): 26–53.

Anaconda, Inc. 2025. *Miniconda 25.3.1*. Version 25.3.1. https://www.anaconda.com/docs/getting-started/miniconda/release/25.x.

Andrews, Simon. 2023. *FastQC*. Version 0.12.1. Github. https://github.com/s-andrews/FastQC.

BCFtools. 2025. *Bcftools*. Version 1.22. Github. https://github.com/samtools/bcftools.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 57 (1): 289–300.

Bergquist, B. A., J. Wu, and E. A. Boyle. 2007. “Variability in Oceanic Dissolved Iron Is Dominated by the Colloidal Fraction.” *Geochimica et Cosmochimica Acta* 71 (12): 2960–74.

Boeckmann, Brigitte, Amos Bairoch, Rolf Apweiler, et al. 2003. “The SWISS-PROT Protein Knowledgebase and Its Supplement TrEMBL in 2003.” *Nucleic Acids Research* 31 (1): 365–70.

Booker, Tom R., Sam Yeaman, and Michael C. Whitlock. 2023. “Using Genome Scans to Identify Genes Used Repeatedly for Adaptation.” *Evolution; International Journal of Organic Evolution* 77 (3): 801–11.

Bors, Eleanor K., Santiago Herrera, James A. Morris, and Timothy M. Shank. 2019. “Population Genomics of Rapidly Invading Lionfish in the Caribbean Reveals Signals of Range Expansion in the Absence of Spatial Population Structure.” *Ecology and Evolution* 9 (6): 3306–20.

Bradford, Yvonne M., Ceri E. Van Slyke, Leyla Ruzicka, et al. 2022. “Zebrafish Information Network, the Knowledgebase for Danio Rerio Research.” *Genetics* 220 (4): iyac016.

Bulik-Sullivan, Brendan K., Po-Ru Loh, Hilary K. Finucane, et al. 2015. “LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies.” *Nature Genetics* 47 (3): 291–95.

Buyinza, Isaac, Grace Ramena, Rebecca Lochmann, Amit Sinha, and Michele Jones. 2024. “Plasma and Tissue Transferrin and Ferritin, and Gene Expression of Ferritin, Transferrin, and Transferrin Receptors I and II in Channel Catfish Ictalurus Punctatus Fed Diets with Different Concentrations of Inorganic or Organic Iron.” *Journal of Fish Diseases* 47 (8): e13953.

Camacho, Christiam, George Coulouris, Vahram Avagyan, et al. 2009. “BLAST+: Architecture and Applications.” *BMC Bioinformatics* 10 (1): 421.

Campitelli, Elio. 2025. *Ggnewscale: Multiple Fill and Colour Scales in “Ggplot2.”* Version 0.5.2. Github. https://github.com/eliocamp/ggnewscale.

Caye, Kevin, Basile Jumentier, Johanna Lepeule, and Olivier François. 2019. “LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies.” *Molecular Biology and Evolution* 36 (4): 852–60.

Center For International Earth Science Information Network-CIESIN-Columbia University. 2018. “Gridded Population of the World, Version 4 (GPWv4): Population Density Adjusted to Match 2015 Revision UN WPP Country Totals, Revision 11 (Version 4.11).” Identifier original variable: gpw-v4-population-density-adjusted-to-2015-unwpp-country-totals-rev11. Version 4.11. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC).

Chen, Bing, Martin E. Feder, and Le Kang. 2018. “Evolution of Heat-Shock Protein Expression Underlying Adaptive Responses to Environmental Stress.” *Molecular Ecology* 27 (15): 3040–54.

Cinar, Ozan, and Wolfgang Viechtbauer. 2022. “The Poolr Package for Combining Independent and Dependent p Values.” *Journal of Statistical Software* 101 (1): 1–42.

Cinar, Ozan, and Wolfgang Viechtbauer. 2026. *Poolr: Methods for Pooling P-Values from (Dependent) Tests*. Version 1.2.0. Github. https://github.com/cran/poolr.

Conway, Tim M., and Seth G. John. 2014. “Quantification of Dissolved Iron Sources to the North Atlantic Ocean.” *Nature* 511 (7508): 212–15.

Cowen, R. K., C. B. Paris, and A. Srinivasan. 2006. “Scaling of Connectivity in Marine Populations.” *Science (New York, N.Y.)* 311 (5760): 522–27.

Cowman, Peter F., Valeriano Parravicini, Michel Kulbicki, and Sergio R. Floeter. 2017. “The Biogeography of Tropical Reef Fishes: Endemism and Provinciality through Time: Reef Fish Endemism and Provinciality.” *Biological Reviews of the Cambridge Philosophical Society* 92 (4): 2112–30.

Crozier, Lisa G., and Jeffrey A. Hutchings. 2014. “Plastic and Evolutionary Responses to Climate Change in Fish.” *Evolutionary Applications* 7 (1): 68–87.

Dai, Zhuoling, Wenbo Huang, Chong Xiao, Fengming You, and Jing Long. 2026. “Macrophages in Post-Chemoradiotherapy Hematopoietic Injury: A Double-Edged Sword.” *Cell Communication and Signaling* 24 (1): 166.

D’Aloia, Cassidy C., Steven M. Bogdanowicz, Robin K. Francis, John E. Majoris, Richard G. Harrison, and Peter M. Buston. 2015. “Patterns, Causes, and Consequences of Marine Larval Dispersal.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (45): 13940–45.

Danecek, Petr, James K. Bonfield, Jennifer Liddle, et al. 2021. “Twelve Years of SAMtools and BCFtools.” *GigaScience* 10 (2): giab008.

Dunnington, Dewey, Brent Thorne, and Diego Hernangómez. 2025. *Ggspatial*. Version 1.1.10. Github. https://github.com/paleolimbot/ggspatial.

E. U. Copernicus Land Monitoring Service Information. 2019. “Land Cover 2015-2019 (Raster 100 m), Global, Yearly – Version 3.” Identifier original variable: Land Cover100m:collection_3:epoch2019. Version 3. https://doi.org/10.2909/c6377c6e-76cc-4d03-8330-628a03693042.

E. U. Copernicus Marine Service Information. 2024. “Global Ocean Biogeochemistry Hindcast.” Product ID: GLOBAL_MULTIYEAR_BGC_001_029. Marine Data Store (MDS). https://doi.org/10.48670/moi-00019.

Eddy, Tyler D., Vicky W. Y. Lam, Gabriel Reygondeau, et al. 2021. “Global Decline in Capacity of Coral Reefs to Provide Ecosystem Services.” *One Earth (Cambridge, Mass.)* 4 (9): 1278–85.

Elvidge, Christopher-D, Tilottama Ghosh, Kimberly Baugh, et al. 2018. “Rating the Effectiveness of Fishery Closures with Visible Infrared Imaging Radiometer Suite Boat Detection Data.” *Frontiers in Marine Science* 5 (April): 308967.

Elvidge, Christopher-D, Mikhail Zhizhin, Kimberly Baugh, and Feng-Chi Hsu. 2015. “Automatic Boat Identification System for VIIRS Low Light Imaging Data.” *Remote Sensing* 7 (3): 3020–36.

Emms, David M., and Steven Kelly. 2019. “OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics.” *Genome Biology* 20 (1): 238.

Ensembl. 2017. “*Danio Rerio* (Zebrafish) Genome Assembly GRCz11. Version GCA_000002035.4.” Release: 115. EMBL-EBI. http://www.ensembl.org/Danio_rerio/Info/Index.

E.U. Copernicus Marine Service Information. 2023a. “Global Ocean Colour (Copernicus-GlobColour), Bio-Geo-Chemical, L4 (Monthly and Interpolated) from Satellite Observations (1997-Ongoing).” Product ID: OCEANCOLOUR_GLO_BGC_L4_MY_009_104. Marine Data Store (MDS). https://data.marine.copernicus.eu/product/OCEANCOLOUR_GLO_BGC_L4_MY_009_104/description.

E.U. Copernicus Marine Service Information. 2023b. “Global Ocean Physics Reanalysis.” Product ID: GLOBAL_MULTIYEAR_PHY_001_030. Marine Data Store (MDS). https://doi.org/10.48670/moi-00021.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. “MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report.” *Bioinformatics (Oxford, England)* 32 (19): 3047–48.

Faria, A., P. J. Fonseca, M. Vieira, et al. 2022. “Boat Noise Impacts Early Life Stages in the Lusitanian Toadfish: A Field Experiment.” *The Science of the Total Environment* 811 (151367): 151367.

Feng, Wei, Yizhuo Weng, Wenkai Shi, et al. 2025. “Aquatic High Iron Induces Hepatic Ferroptosis in Zebrafish (Danio Rerio) via Interleukin-22 Signaling Pathway.” *Environmental Pollution (Barking, Essex: 1987)* 367 (125627): 125627.

Fisher, Rebecca, Rebecca A. O’Leary, Samantha Low-Choy, et al. 2015. “Species Richness on Coral Reefs and the Pursuit of Convergent Global Estimates.” *Current Biology* 25 (4): 500–505.

Fitz, Kyra S., Humberto R. Montes Jr, Diane M. Thompson, and Malin L. Pinsky. 2023. “Isolation-by-Distance and Isolation-by-Oceanography in Maroon Anemonefish (Amphiprion Biaculeatus).” *Evolutionary Applications* 16 (2): 379–92.

Folashade, Daniel, Ooi Hong, Rich Calaway, and Steve Weston. 2022. *Foreach*. Version 1.5.2. Github. https://github.com/RevolutionAnalytics/foreach/tree/master.

Folashade, Daniel, Steve Weston, and Dan Tenenbaum. 2022. Doparallel. Version 1.0.17. Github. https://github.com/RevolutionAnalytics/doparallel.

Fraimout, Antoine, Céline Reisser, Émilie Farcy, et al. 2026. “Genome-Environment Associations Reveal Shared and Unique Adaptive Loci across Multiple Pollutants and Populations of the Eastern Mosquitofish Gambusia Holbrooki.” In *bioRxiv*. bioRxiv, January 7. https://doi.org/10.64898/2026.01.07.697681.

François, Olivier, Helena Martins, Kevin Caye, and Sean D. Schoville. 2016. “Controlling False Discoveries in Genome Scans for Selection.” *Molecular Ecology* 25 (2): 454–69.

Freitas, Carla, David Villegas-Ríos, Even Moland, and Esben Moland Olsen. 2021. “Sea Temperature Effects on Depth Use and Habitat Selection in a Marine Fish Community.” *The Journal of Animal Ecology* 90 (7): 1787–800.

Frichot, Eric, Kevin Caye, Clement Gain, and Olivier Francois. 2024. *LEA*. Version 3.20.0. Github. https://github.com/bcm-uga/LEA/blob/devel/vignettes/LEA.Rnw.

Frichot, Eric, and Olivier François. 2015. “LEA: An R Package for Landscape and Ecological Association Studies.” *Methods in Ecology and Evolution* 6 (8): 925–29.

Frichot, Eric, Sean D. Schoville, Guillaume Bouchard, and Olivier François. 2013. “Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models.” *Molecular Biology and Evolution* 30 (7): 1687–99.

Gain, Clément, and Olivier François. 2021. “LEA 3: Factor Models in Population Genetics and Ecological Genomics with R.” *Molecular Ecology Resources* 21 (8): 2738–48.

Gairin, Emma, Frédéric Bertucci, Natacha Roux, et al. 2024. “Coral Reef Fish Density at a Tourist Destination Responded Rapidly to COVID-19 Restrictions.” *Oecologia* 205 (3–4): 533–43.

Gertz, E. Michael, Yi-Kuo Yu, Richa Agarwala, Alejandro A. Schäffer, and Stephen F. Altschul. 2006. “Composition-Based Statistics and Translated Nucleotide Searches: Improving the TBLASTN Module of BLAST.” *BMC Biology* 4 (1): 41.

Glasauer, Stella M. K., and Stephan C. F. Neuhauss. 2014. “Whole-Genome Duplication in Teleost Fishes and Its Evolutionary Consequences.” *Molecular Genetics and Genomics* 289 (6): 1045–60.

Goreau, Thomas J. F., and Raymond L. Hayes. 2021. “Global Warming Triggers Coral Reef Bleaching Tipping Point.” *Ambio* 50 (6): 1137–40.

Goudet, Jerome, Thibaut Jombart, Zhian N. Kamvar, Eric Archer, and Olivier Hardy. 2022. *Estimation and Tests of Hierarchical F-Statistics*. Version 0.5.11. Reference Manual. Cran.r package documentation. https://cran.r-project.org/web/packages/hierfstat/index.html.

Gould, A. L., and P. V. Dunlap. 2017. “Genomic Analysis of a Cardinalfish with Larval Homing Potential Reveals Genetic Admixture in the Okinawa Islands.” *Molecular Ecology* 26 (15): 3870–82.

Gould, A. L., J. B. Henderson, and A. W. Lam. 2022. “Chromosome-Level Genome Assembly of the Bioluminescent Cardinalfish Siphamia Tubifer: An Emerging Model for Symbiosis Research.” *Genome Biology and Evolution* 14 (4): evac044.

Green, Alison L., Aileen P. Maypa, Glenn R. Almany, et al. 2015. “Larval Dispersal and Movement Patterns of Coral Reef Fishes, and Implications for Marine Reserve Network Design: Connectivity and Marine Reserves.” *Biological Reviews of the Cambridge Philosophical Society* 90 (4): 1215–47.

Gruber, Bernd. 2023. *dartR*. Version 2.9.7. Github. https://github.com/green-striped-gecko/dartR.

Guo, Yan-Lin, Wei-Dan Jiang, Pei Wu, et al. 2017. “The Decreased Growth Performance and Impaired Immune Function and Structural Integrity by Dietary Iron Deficiency or Excess Are Associated with TOR, NF-κB, p38MAPK, Nrf2 and MLCK Signaling in Head Kidney, Spleen and Skin of Grass Carp (Ctenopharyngodon Idella).” *Fish & Shellfish Immunology* 65 (June): 145–68.

Hallal, Mahmoud, Sophie Braga-Lagache, Jovana Jankovic, et al. 2021. “Inference of Kinase-Signaling Networks in Human Myeloid Cell Line Models by Phosphoproteomics Using Kinase Activity Enrichment Analysis (KAEA).” *BMC Cancer* 21 (1): 789.

Hsu, Feng-Chi, Christopher-D Elvidge, Kimberly Baugh, et al. 2019. “Cross-Matching VIIRS Boat Detections with Vessel Monitoring System Tracks in Indonesia.” *Remote Sensing* 11 (9): 995.

Hughes, Terry P., Michele L. Barnes, David R. Bellwood, et al. 2017. “Coral Reefs in the Anthropocene.” *Nature* 546 (7656): 82–90.

Hutchins, D. A., and P. W. Boyd. 2016. “Marine Phytoplankton and the Changing Ocean Iron Cycle.” *Nature Climate Change* 6 (12): 1072–79.

Jombart, Thibaut. 2025. *Adegenet*. Version 2.1.11. Github. https://github.com/thibautjombart/adegenet.

Jones, Rhiannon L., Jon R. Hawkings, Michael P. Meredith, et al. 2025. “Antarctic Glaciers Export Carbon-Stabilised Iron(II)-Rich Particles to the Surface Southern Ocean.” *Nature Communications* 16 (1): 5015.

Jumentier, Basile, Kevin Caye, and Olivier François. 2021a. *Latent Factor Mixed Models*. Reference Manual. Cran.r Package Documentation. https://cran.r-project.org/web/packages/lfmm/lfmm.pdf.

Jumentier, Basile, Kevin Caye, and Olivier François. 2021b. *Lfmm: Fitting Latent Factor Mixed Models*. Version 1.1. Github. https://github.com/bcm-uga/lfmm.

Khawar, Masooma, Zubia Masood, Habib Ul Hasan, et al. 2024. “Trace Metals and Nutrient Analysis of Marine Fish Species from the Gwadar Coast.” *Scientific Reports* 14 (1): 6548.

Kim, Sunhee, and Chang-Yong Lee. 2020. “A Consistent Approach to the Genotype Encoding Problem in a Genome-Wide Association Study of Continuous Phenotypes.” *PloS One* 15 (7): e0236139.

Kitsoulis, Christos, Vasileios Papadogiannis, Jon Kristoffersen, et al. 2023. “Near-Chromosome Level Genome Assembly of Devil Firefish, *Pterois Miles*.” In *bioRxiv*. bioRxiv, June 6. https://doi.org/10.1101/2023.01.10.523469.

Knaus, Brian. 2023. *vcfR: Tools to Work with Variant Call Format Files*. Version 1.15.0. Github. https://github.com/knausb/vcfR.

König, Daniela, and Alessandro Tagliabue. 2023. “The Fingerprint of Climate Variability on the Surface Ocean Cycling of Iron and Its Isotopes.” *Biogeosciences* 20 (19): 4197–212.

Krueger, Felix. 2019. *TrimGalore: A Wrapper around Cutadapt and FastQC to Consistently Apply Adapter and Quality Trimming to FastQ Files, with Extra Functionality for RRBS Data*. Version 0.6.10. Github. https://github.com/FelixKrueger/TrimGalore.

Lall, Santosh P., and Sadasivam J. Kaushik. 2021. “Nutrition and Metabolism of Minerals in Fish.” *Animals: An Open Access Journal From MDPI* 11 (9): 2711.

Lasky, Jesse R., Emily B. Josephs, and Geoffrey P. Morris. 2023. “Genotype-Environment Associations to Reveal the Molecular Basis of Environmental Adaptation.” *The Plant Cell* 35 (1): 125–38.

Lefcheck, Jonathan S., Simon J. Brandl, Anne A. Innes-Gold, Robert S. Steneck, Ruben E. Torres, and Douglas B. Rasher. 2019. “Response: Commentary: Tropical Fish Diversity Enhances Coral Reef Functioning across Multiple Scales.” *Frontiers in Ecology and Evolution* 7 (August). https://doi.org/10.3389/fevo.2019.00303.

Lefouili, Messaoud, and Kiwoong Nam. 2022. “The Evaluation of Bcftools Mpileup and GATK HaplotypeCaller for Variant Calling in Non-Human Species.” *Scientific Reports* 12 (1): 11331.

Lemos, M. L., and M. Balado. 2020. “Iron Uptake Mechanisms as Key Virulence Factors in Bacterial Fish Pathogens: Iron and Virulence in Fish Pathogens.” *Journal of Applied Microbiology* 129 (1): 104–15.

Luo, Lan, Zheng-Zheng Tang, Sean D. Schoville, and Jun Zhu. 2021. “A Comprehensive Analysis Comparing Linear and Generalized Linear Models in Detecting Adaptive SNPs.” *Molecular Ecology Resources* 21 (3): 733–44.

Mao, Hua, Liang Xie, and Xinchun Pi. 2017. “Low-Density Lipoprotein Receptor-Related Protein-1 Signaling in Angiogenesis.” *Frontiers in Cardiovascular Medicine* 4 (May): 34.

Massicotte, Philippe, and Andy South. 2026. *Rnaturalearth: World Map Data from Natural Earth*. Version 1.2.0. Github. https://github.com/ropensci/rnaturalearth.

Milton, Michael. 2016. *TidyMultiqc*. Version 1.0.3. Github. https://github.com/cran/TidyMultiqc.

Misumi, K., K. Lindsay, J. K. Moore, et al. 2013. “The Iron Budget in Ocean Surface Waters in the 20th and 21st Centuries: Projections by the Community Earth System Model Version 1.” *Biogeosciences Discussions* 10 (5): 8505–59.

Müller, Kirill, and Hadley Wickham. 2026. *Tibble: Simple Data Frames*. Version 3.3.1. Github. https://github.com/tidyverse/tibble/blob/main/DESCRIPTION.

MultiQC. 2025. *MultiQC: Aggregate Results from Bioinformatics Analyses across Many Samples into a Single Report*. Version 1.33. Github. https://github.com/MultiQC/MultiQC.

Munday, Philip L., Jennifer M. Donelson, and Jose A. Domingos. 2017. “Potential for Adaptation to Climate Change in a Coral Reef Fish.” *Global Change Biology* 23 (1): 307–17.

NOAA Coral Reef Watch. 2018a. “NOAA Coral Reef Watch Version 3.1 Daily Global 5km Satellite Coral Bleaching Degree Heating Week Product.” Identifier original variable: ct5km_dhw-max_v3.1. Version 3.1. https://coralreefwatch.noaa.gov/product/5km/index_5km_dhw.php.

NOAA Coral Reef Watch. 2018b. “NOAA Coral Reef Watch Version 3.1 Daily Global 5km Satellite Sea Surface Temperature (a.k.a. *CoralTemp*).” Identifier original variable: ct5km_sst-mean_v3.1. Version 3.1. https://coralreefwatch.noaa.gov/product/5km/index_5km_sst.php.

Oleksiak, Marjorie F., and Om P. Rajora, eds. 2020. *Population Genomics: Marine Organisms*. 2020th ed. Population Genomics. Springer Nature.

Otte, Kathrin A. 2024. “A Window into Local Adaptation.” *Molecular Ecology Resources* 24 (2): e13872.

Pebesma, E., and R. Bivand. 2026. *Simple Features for R*. Version 1.0.24. Github. https://github.com/r-spatial/sf.

Pedersen, Thomas Lin. 2025. *Patchwork*. Version 1.3.2. Github. https://github.com/thomasp85/patchwork.

Pinsky, Malin L., Humberto R. Montes Jr, and Stephen R. Palumbi. 2010. “Using Isolation by Distance and Effective Density to Estimate Dispersal Scales in Anemonefish: Isolation by Distance and Marine Dispersal.” *Evolution; International Journal of Organic Evolution* 64 (9): 2688–700.

Rellstab, Christian, Felix Gugerli, Andrew J. Eckert, Angela M. Hancock, and Rolf Holderegger. 2015. “A Practical Guide to Environmental Association Analysis in Landscape Genomics.” *Molecular Ecology* 24 (17): 4348–70.

Rescheneder, Philipp. 2017. *NextGenMap: NextGenMap Is a Flexible Highly Sensitive Short Read Mapping Tool*. Version 0.5.5. Github. https://github.com/Cibiv/NextGenMap.

Riginos, Cynthia, Eric D. Crandall, Libby Liggins, Pim Bongaerts, and Eric A. Treml. 2016. “Navigating the Currents of Seascape Genomics: How Spatial Analyses Can Augment Population Genomic Studies.” *Current Zoology* 62 (6): 581–601.

Ryan, William B. F., Suzanne M. Carbotte, Justin O. Coplan, et al. 2009a. “GMRT: Global Multi-Resolution Topography.” Identifier original variable: GMRTv3_9. In *Global Multi-Resolution Topography Synthesis*. https://www.gmrt.org/.

Ryan, William B. F., Suzanne M. Carbotte, Justin O. Coplan, et al. 2009b. “Global Multi-Resolution Topography Synthesis.” *Geochemistry, Geophysics, Geosystems: G(3)* 10 (3). https://doi.org/10.1029/2008gc002332.

Ryu, Taewoo, Marcela Herrera, Billy Moore, et al. 2022. “A Chromosome-Scale Genome Assembly of the False Clownfish, Amphiprion Ocellaris.” *G3 (Bethesda, Md.)* 12 (5): jkac074.

Saenz-Agudelo, Pablo, Joseph D. Dibattista, Marek J. Piatek, et al. 2015. “Seascape Genetics along Environmental Gradients in the Arabian Peninsula: Insights from ddRAD Sequencing of Anemonefishes.” *Molecular Ecology* 24 (24): 6241–55.

SAMtools. 2025. *Samtools*. Version 1.22.1. Github. https://github.com/samtools/samtools.

Sayers, Eric W., Chris O’Sullivan, and Ilene Karsch-Mizrachi. 2022. “Using GenBank and SRA.” *Methods in Molecular Biology (Clifton, N.J.)* 2443: 1–25.

Sedlazeck, Fritz J., Philipp Rescheneder, and Arndt von Haeseler. 2013. “NextGenMap: Fast and Accurate Read Mapping in Highly Polymorphic Genomes.” *Bioinformatics (Oxford, England)* 29 (21): 2790–91.

Selmoni, Oliver, Line K. Bay, Moises Exposito-Alonso, and Phillip A. Cleves. 2024. “Finding Genes and Pathways That Underlie Coral Adaptation.” *Trends in Genetics: TIG* 40 (3): 213–27.

Selmoni, Oliver, Gaël Lecellier, Véronique Berteaux-Lecellier, and Stéphane Joost. 2023. “The Reef Environment Centralized InFormation System (RECIFS): An Integrated Geo-environmental Database for Coral Reef Research and Conservation.” *Global Ecology and Biogeography: A Journal of Macroecology* 32 (5): 622–32.

Selmoni, Oliver, Estelle Rochat, Gael Lecellier, Veronique Berteaux-Lecellier, and Stéphane Joost. 2020. “Seascape Genomics as a New Tool to Empower Coral Reef Conservation Strategies: An Example on North-Western Pacific Acropora Digitifera.” *Evolutionary Applications* 13 (8): 1923–38.

Sherman, Krista D., Josephine R. Paris, Robert Andrew King, et al. 2020. “RAD-Seq Analysis and in Situ Monitoring of Nassau Grouper Reveal Fine-Scale Population Structure and Origins of Aggregating Fish.” *Frontiers in Marine Science* 7 (March): 503082.

Shoenfelt, Elizabeth M., Gisela Winckler, Frank Lamy, Robert F. Anderson, and Benjamin C. Bostick. 2018. “Highly Bioavailable Dust-Borne Iron Delivered to the Southern Ocean during Glacial Periods.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (44): 11180–85.

Simillion, Cedric. 2016. *SetRank*. Version 1.1.0. Github. https://github.com/C3c6e6/SetRank.

Simillion, Cedric, Robin Liechti, Heidi E. L. Lischer, Vassilios Ioannidis, and Rémy Bruggmann. 2017. “Avoiding the Pitfalls of Gene Set Enrichment Analysis with SetRank.” *BMC Bioinformatics* 18 (1): 151.

Smith, Steve, Louis Bernatchez, and Luciano B. Beheregaray. 2013. “RNA-Seq Analysis Reveals Extensive Transcriptional Plasticity to Temperature Stress in a Freshwater Fish Species.” *BMC Genomics* 14 (1): 375.

Steffen, Will, Wendy Broadgate, Lisa Deutsch, Owen Gaffney, and Cornelia Ludwig. 2015. “The Trajectory of the Anthropocene: The Great Acceleration.” *The Anthropocene Review* 2 (1): 81–98.

Stekhoven, Daniel. 2025. *missForest*. Version 1.6.1. Github. https://github.com/stekhoven/missForest.

Strona, Giovanni, Kevin D. Lafferty, Simone Fattorini, et al. 2021. “Global Tropical Reef Fish Richness Could Decline by around Half If Corals Are Lost.” *Proceedings. Biological Sciences* 288 (1953): 20210274.

Tagliabue, Alessandro, Andrew R. Bowie, Philip W. Boyd, Kristen N. Buck, Kenneth S. Johnson, and Mak A. Saito. 2017. “The Integral Role of Iron in Ocean Biogeochemistry.” *Nature* 543 (7643): 51–59.

Thirukanthan, Chandra Segaran, Mohamad Nor Azra, Fathurrahman Lananan, et al. 2023. “The Evolution of Coral Reef under Changing Climate: A Scientometric Review.” *Animals: An Open Access Journal From MDPI* 13 (5): 949.

Utturkar, Sagar, Asela Dassanayake, Shilpa Nagaraju, and Steven D. Brown. 2020. “Bacterial Differential Expression Analysis Methods.” *Methods in Molecular Biology (Clifton, N.J.)* 2096: 89–112.

Valieris, Renan. 2025. *Parallel-Fastq-Dump: Parallel Fastq-Dump Wrapper*. Version v.0.6.7. Github. https://github.com/rvalieris/parallel-fastq-dump.

Velotta, Jonathan P., Stephen D. McCormick, Andrew Whitehead, Catherine S. Durso, and Eric T. Schultz. 2022. “Repeated Genetic Targets of Natural Selection Underlying Adaptation of Fishes to Changing Salinity.” *Integrative and Comparative Biology* 62 (2): 357–75.

Waldvogel, Ann-Marie, Barbara Feldmeyer, Gregor Rolshausen, et al. 2020. “Evolutionary Genomics Can Improve Prediction of Species’ Responses to Climate Change.” *Evolution Letters* 4 (1): 4–18.

Wang, Jian, and Wen-Xiong Wang. 2016. “Novel Insights into Iron Regulation and Requirement in Marine Medaka Oryzias Melastigma.” *Scientific Reports* 6 (1): 26615.

Wei, Ke-Hsuan, I-Ting Lin, Kaushik Chowdhury, et al. 2023. “Comparative Single-Cell Profiling Reveals Distinct Cardiac Resident Macrophages Essential for Zebrafish Heart Regeneration.” *eLife* 12 (e84679): e84679.

Whiting, James R., Tom R. Booker, Clément Rougeux, et al. 2024. “The Genetic Architecture of Repeated Local Adaptation to Climate in Distantly Related Plants.” *Nature Ecology & Evolution* 8 (10): 1933–47.

Wickham, Hadley. 2025. *Stringr*. Version 1.6.0. Github. https://github.com/tidyverse/stringr.

Wickham, Hadley, Winston Chang, Lionel Henry, et al. 2026. *Ggplot2*. Version 4.0.2. Github. https://github.com/tidyverse/ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2024. *Dplyr*. Version 1.2.0. Github. https://github.com/tidyverse/dplyr.

Wickham, Hadley, and Lionel Henry. 2026. *Purrr: Functional Programming Tools*. Version 1.2.2. Github. https://github.com/tidyverse/purrr.

Yang, Jian, Michael N. Weedon, Shaun Purcell, et al. 2011. “Genomic Inflation Factors under Polygenic Inheritance.” *European Journal of Human Genetics* 19 (7): 807–12.

Zhai, Yue, Xin Zhang, Cheng Zhao, et al. 2023. “Rescue of Bmp15 Deficiency in Zebrafish by Mutation of Inha Reveals Mechanisms of BMP15 Regulation of Folliculogenesis.” *PLoS Genetics* 19 (9): e1010954.

Zhang, Weiwei, Yang Yang, Sijie Hua, et al. 2024. “Chromosome-Level Genome Assembly and Annotation of the Yellow Grouper, Epinephelus Awoara.” *Scientific Data* 11 (1): 151.

Zhao, Changlong, Zhiyong Yang, Yunbo Li, and Zilong Wen. 2024. “Macrophages in Tissue Repair and Regeneration: Insights from Zebrafish.” *Cell Regeneration (London, England)* 13 (1): 12.

1 Appendix

1.1 Software, R Packages, and Version Information

All command-line tools and R packages, including their versions, are listed in the tables below. Command-line analyses were run on the Science Cluster via an SSH connection, using SLURM for execution, on an Ubuntu Linux system. Most R scripts were executed through ScienceApps, some ran locally on my personal computer.

S Table 1: Command-line tools used in the analysis, including software versions and corresponding citations.

Tool	Version	Citation
BCFtools	1.22	BCFtools 2025
blast	2.17.0	Camacho et al. 2009
FastQC	v0.12.1	Andrews 2023
MultiQC.	1.33	MultiQC 2025
NextGenMap	0.5.5	Rescheneder 2017
parallel-fastq-dump fastq-dump sra-tools	0.6.7 3.2.1 3.2.1	Valieris 2025
r-base	4.5.1	*Date: June, 2025*
SAMtools	1.22.1	SAMtools 2025
Trim Galore	0.6.10	Krueger 2019
Miniconda	25.3.1	Anaconda, Inc. 2025

S Table 2: Overview of R packages used for data processing and analysis, including package versions and corresponding citations.

R package	Version	Citation	R package	Version	Citation
adegenet	2.1.11	Jombart 2025	patchwork	1.3.2	Pedersen 2025
dartR	2.9.7	Gruber 2023	PicMin	0.0.0.9000	Booker et al. 2023
doParallel	1.0.17	Folashadde et al. 2022	poolr	1.2.0	Cinar and Viechtbauer 2026
dplyr	1.2.0	Wickham et al. 2026	purrr	1.2.2	Wickham and Henry 2026
foreach	1.5.2	Folashade et al. 2022	rnaturalearth	1.2.0	Massicotte and South 2026
ggnewscale	0.5.2	Campitelli 2025	SetRank	1.1.0	Simillion 2016
ggplot2	4.0.2	Wickham et al. 2026	sf	1.0.24	Pebesma and Bivand 2026
ggspatial	1.1.10	Dunnington et al. 2025	stringr	1.6.0	Wickham 2025
hierfstat	0.5.11	Goudet et al. 2022	tibble	3.3.1	Müller and Wickham 2026

R package	Version	Citation	R package	Version	Citation
LEA	3.20.0	Frichot et al. 2024	TidyMultiqc	1.0.3	Milton 2016
lfmm	1.1	Jumentier et al. 2021b	vcfR	1.15.0	Knaus 2023
missForest	1.6.1	Stekhoven 2025			

Short note: The last species run was conducted in July 2026. Due to technical issues with the University of Zurich's Science Cluster, the analysis had to be run locally. As the cluster was unavailable, the exact package versions used for the previous species runs in March 2026 could not be verified. Therefore, the July 2026 run was performed using the most recent available package versions. The outputs of the July run were consistent with those obtained in March 2026, indicating that the differences in package versions did not affect the functionalities relevant to this analysis.

1.2 Excluded Sequencing Runs and Exclusion Criteria

S Table 3: Sequencing runs excluded from the final sample set for each species and the corresponding reasons for exclusion.

Species	Run	Reason
Epinephelus striatus	ERR3950796	Alignment failed; no output generated (computational failure, unrelated to sample quality).
Epinephelus striatus	ERR3950797	Alignment failed; no output generated (computational failure, unrelated to sample quality).
Epinephelus striatus	ERR3950798	Alignment failed; no output generated (computational failure, unrelated to sample quality).
Epinephelus striatus	ERR3950829	No mapped reads (0% mapping rate).
Epinephelus striatus	ERR3950836	Alignment failed; no output generated (computational failure, unrelated to sample quality).
Epinephelus striatus	ERR3950864	Alignment failed; no output generated (computational failure, unrelated to sample quality).
Pterois volitans	SRR8354498	No geographic coordinates (latitude/longitude) available in RunTable.
Pterois volitans	SRR8354499	No geographic coordinates (latitude/longitude) available in RunTable.
Pterois volitans	SRR8354500	Low mapping rate; abnormal GC content.
Pterois volitans	SRR8354521	Low mapping rate; abnormal GC content.
Pterois volitans	SRR8354527	No geographic coordinates (latitude/longitude) available in RunTable.
Pterois volitans	SRR8354528	No geographic coordinates (latitude/longitude) available in RunTable.
Pterois volitans	SRR8354529	No geographic coordinates (latitude/longitude) available in RunTable.
Pterois volitans	SRR8354530	No geographic coordinates (latitude/longitude) available in RunTable.
Pterois volitans	SRR8354531	No geographic coordinates (latitude/longitude) available in RunTable.
Pterois volitans	SRR8354532	No geographic coordinates (latitude/longitude) available in RunTable.
Pterois volitans	SRR8354545	No geographic coordinates (latitude/longitude) available in RunTable.
Pterois volitans	SRR8354548	No geographic coordinates (latitude/longitude) available in RunTable.

Species	Run	Reason
Pterois volitans	SRR8354549	No geographic coordinates (latitude/longitude) available in RunTable.
Pterois volitans	SRR8354550	No geographic coordinates (latitude/longitude) available in RunTable.
Pterois volitans	SRR8354551	No geographic coordinates (latitude/longitude) available in RunTable.
Pterois volitans	SRR8354553	No geographic coordinates (latitude/longitude) available in RunTable.
Pterois volitans	SRR8354554	No geographic coordinates (latitude/longitude) available in RunTable.
Pterois volitans	SRR8354567	No geographic coordinates (latitude/longitude) available in RunTable.
Pterois volitans	SRR8354570	No geographic coordinates (latitude/longitude) available in RunTable.
Pterois volitans	SRR8354579	No geographic coordinates (latitude/longitude) available in RunTable.
Pterois volitans	SRR8354580	No geographic coordinates (latitude/longitude) available in RunTable.
Siphamia tubifer	SRR5489521	Alignment failed; no output generated (computational failure, unrelated to sample quality).
Siphamia tubifer	SRR5489526	Alignment failed; no output generated (computational failure, unrelated to sample quality).
Siphamia tubifer	SRR5489527	Alignment failed; no output generated (computational failure, unrelated to sample quality).
Siphamia tubifer	SRR5489528	Alignment failed; no output generated (computational failure, unrelated to sample quality).
Siphamia tubifer	SRR5489654	Abnormal GC content.

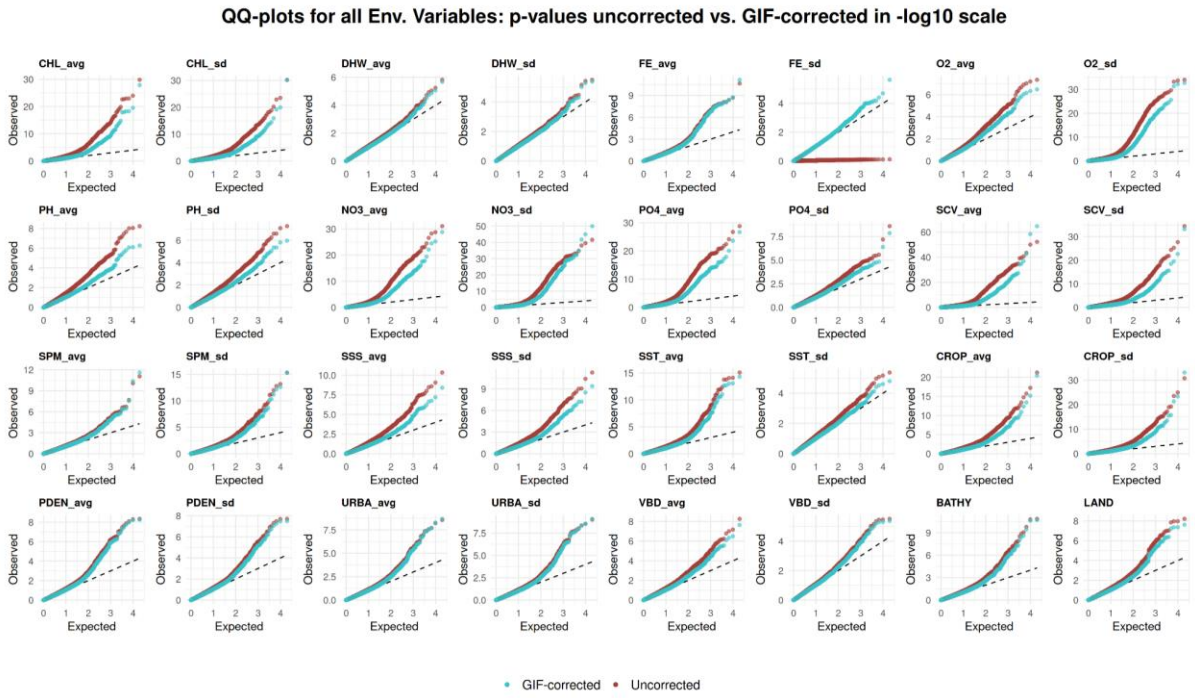
For sequencing runs where the alignment step failed, no aligned output was generated. These failures were considered computational or technical rather than related to sample quality, as the alignment process terminated without producing an output file. The failures may have resulted from SLURM resource limitations, such as insufficient runtime or memory allocation.

1.3 Supplementary Figures per Species

This part contains all supplementary figures structured per species from start to end of the analysis.

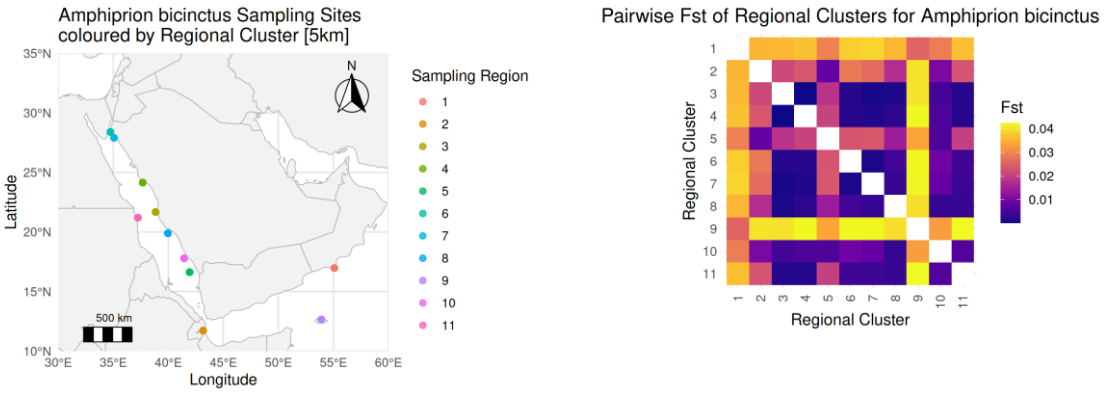
1.3.1 *A. Bicinctus* - Special Supplementary Analysis

A special section is dedicated to *A. bicinctus*, as additional analyses were performed to further investigate population structure and p-value inflation in the LFMM GEA method.



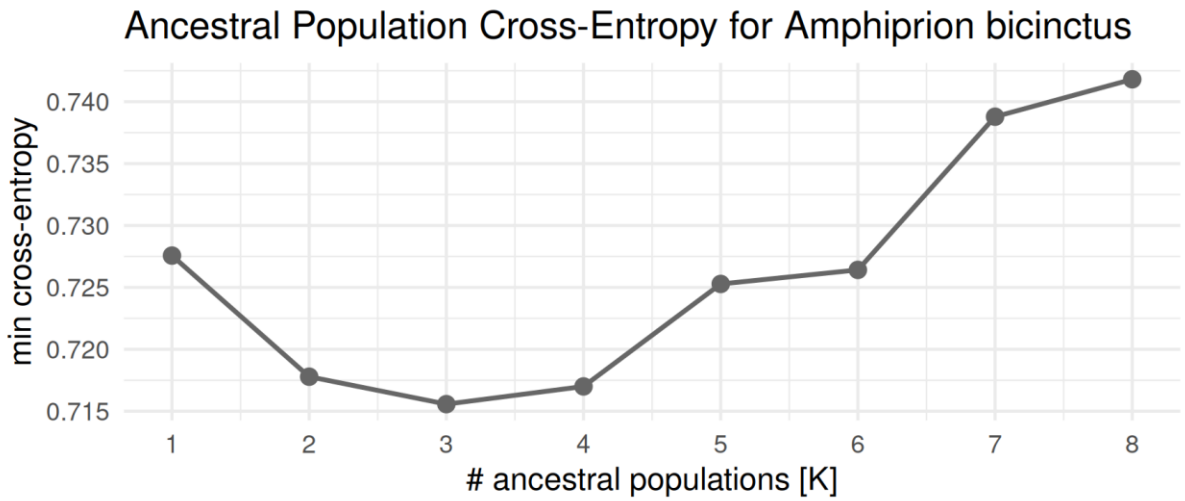
S Figure 1: QQ Plots for each environmental variable showing the distribution of expected vs. observed p-values for A. bicinctus (LFMM method). Environmental variables: CHL (Chlorophyll [mg/m³]), DHW (Degree Heating Weeks [°C·week⁻¹]), FE (Iron [mmol/m³]), O_2 (Oxygen [mmol/m³]), PH (pH), NO_3 (Nitrate [mmol/m³]), PO_4 (Phosphate [mmol/m³]), SCV (Sea-water velocity [m/s]), SPM (Suspended matter [g/m³]), SSS (Sea-surface salinity [10^{-3}]), SST (Sea-surface temperature [°C]), Bathy (Depth [m]), CROP (Cropland density [%]), LAND (Land-surface density [%]), PDEN (Population density [hab km^{-2}]), URBA (Built-up area density [%]), VBD (Boat-detection frequency [%]).

Nearly every panel shows observed p-values departing sharply upward from the expected diagonal even after GIF correction. The inflation is roughly uniform across variables rather than concentrated in a few, which points to a shared underlying cause (structure/collinearity) rather than variable-specific artefacts.



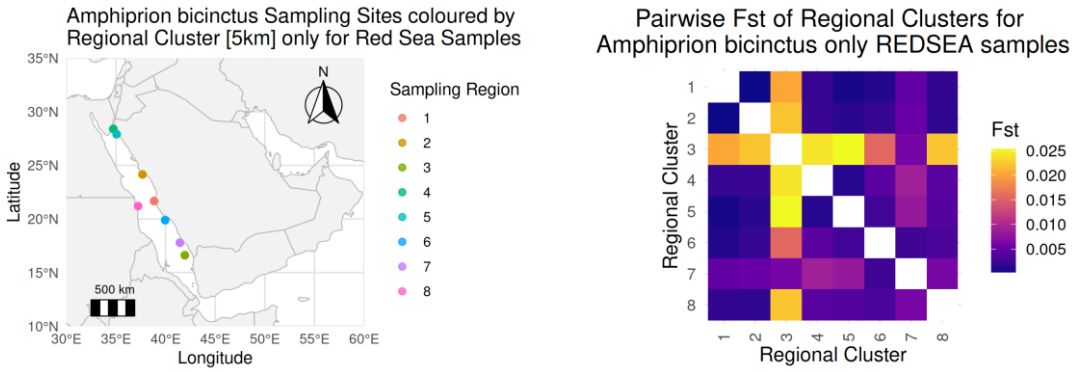
S Figure 2: Left: Map of Sample Sites for A. bicinctus coloured by Regional Cluster [5km], Right: Heatmap of pairwise F_{st} for A. bicinctus using the Regional Clusters.

Clusters 1, 2 and 9 (Gulf of Aden) and cluster 5 (Farasan Islands) stand out with elevated F_{st} against the rest, visually confirming the two-to-three population split reported in the results (4.1 Exploratory and Preliminary Analysis), with the Bab-el-Mandeb Strait as the main structuring feature.



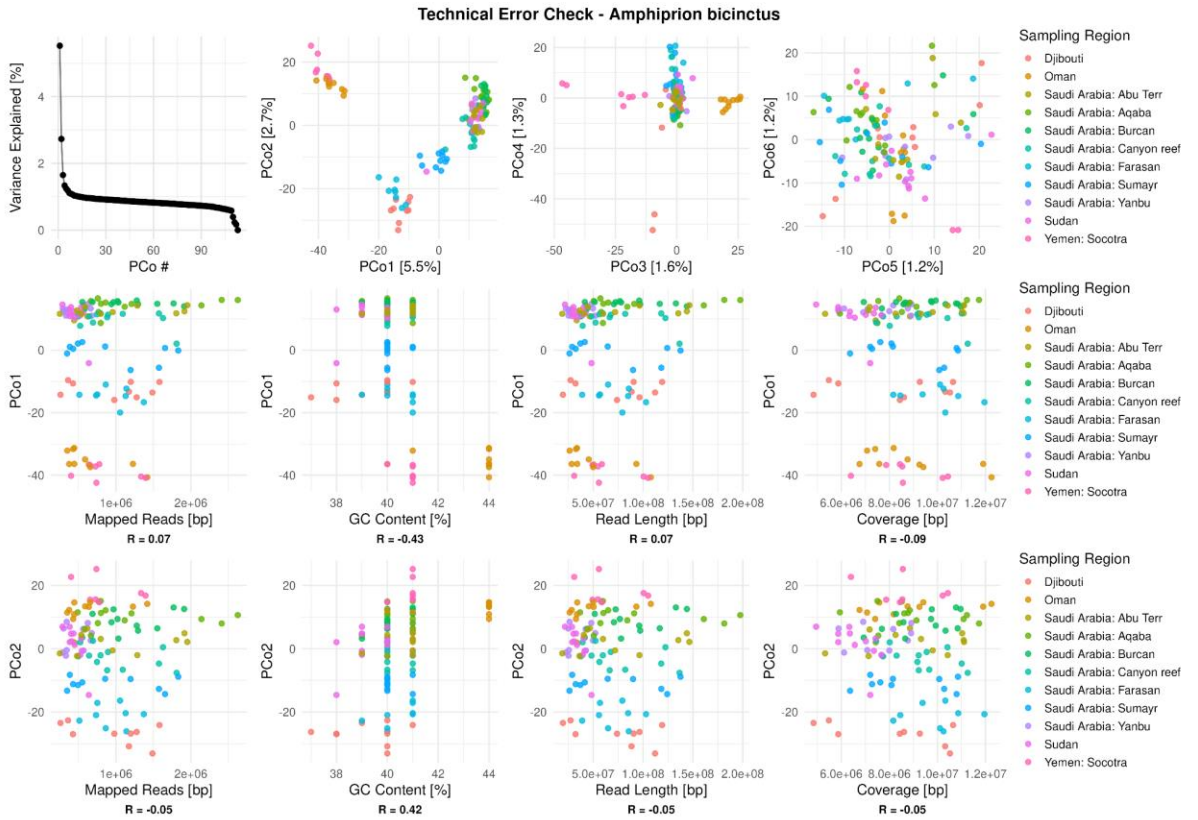
S Figure 3: This minimal cross-entropy plot indicates the number of ancestral populations that best suit A. bicinctus Samples.

Cross-entropy analysis using the snmf function from the LEA package identified K = 3 as the optimal value, although the relatively flat curve across K = 2-4 provides limited support for a clear preference for three ancestral populations.



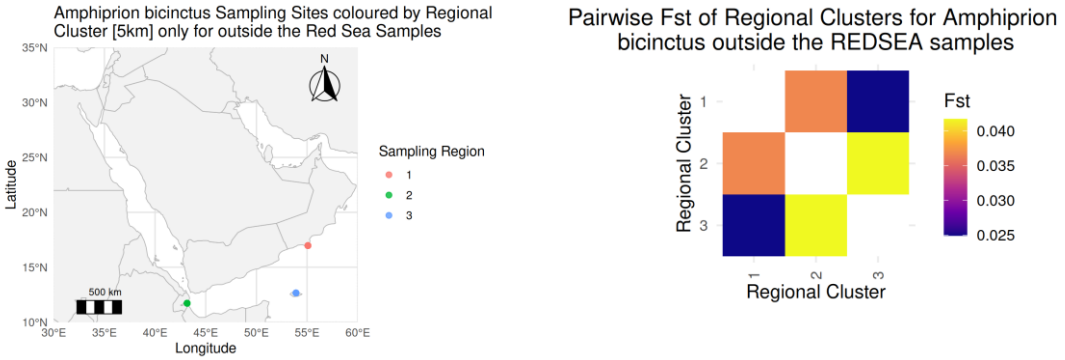
S Figure 4: Left: Map of Sample Sites for A. bicinctus coloured by Regional Cluster [5km] inside the Red Sea, Right: Heatmap of pairwise F_{st} for A. bicinctus using the Regional Clusters inside the Red Sea as Population Input.

Restricting the analysis to Red Sea samples revealed generally low genetic differentiation, with all pair-wise F_{st} values below 0.025. Differentiation was approximately tenfold lower than in the full dataset, while the Farasan Islands (cluster 3) remained the most differentiated, indicating weak but detectable north-south genetic structure within the Red Sea.



S Figure 5: PCoA of the genotype matrix for A. bicinctus, plotted against selected MultiQC metrics coloured by sampling region. Pearson's R values indicate the correlation between PC1 or PC2 and each respective MultiQC metric.

PC1 (5.5%) clearly separates the Gulf of Aden cluster from the rest, and of the QC checks only GC content shows a moderate correlation with the axes (R ≈ 0.42-0.43); read depth and mapping metrics correlate weakly, so the structure looks biological rather than a sequencing artefact.



S Figure 6: Left: Map of Sample Sites for A. bicinctus coloured by Regional Cluster [5km] outside of the Red Sea, Right: Heatmap of pairwise F_{st} for A. bicinctus using the Regional Clusters outside of the Red Sea as Population Input.

Consistent with this interpretation, the initial pairwise F_{st} heat map across all eleven regional clusters showed that Djibouti samples were genetically more similar to southern Red Sea samples than to those from Oman or Socotra. This suggests that the Bab-el-Mandeb Strait alone does not fully explain the observed genetic differentiation and that oceanographic processes may also influence gene flow.

Another possible explanation for the inflated p-values is collinearity between population structure and environmental variation. In LFMM, latent factors are used to account for background population structure while testing associations between genotypes and environmental variables. If population structure is strongly correlated with an environmental variable, LFMM may have difficulty distinguishing genetic differences associated with population structure from those associated with the environment, potentially contributing to p-value inflation. To assess this, the first two principal components from a PCoA of the genotype matrix were tested against each environmental variable using linear models. The resulting R^2 and F-test p-values are presented in the table below.

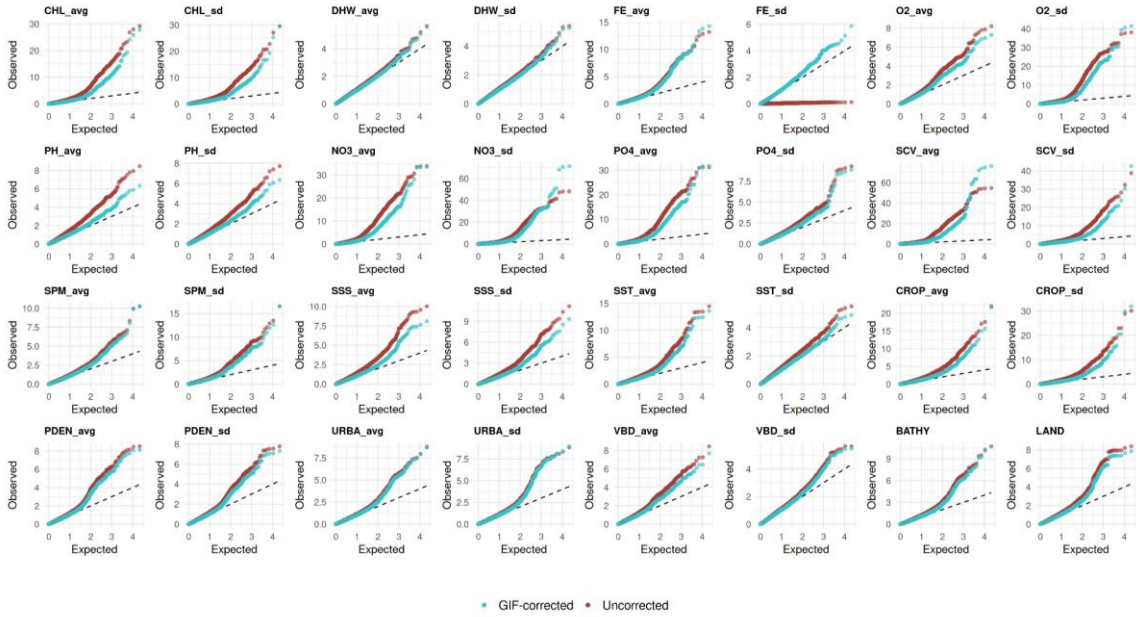
S Table 4: Linear models assessing associations between environmental variables and genetic variation (PC1 and PC2) in A. bicinctus. Environmental variables: CHL (Chlorophyll [mg/m³]), DHW (Degree Heating Weeks [°C·week⁻¹]), FE (Iron [mmol/m³]), O_2 (Oxygen [mmol/m³]), PH (pH), NO_3 (Nitrate [mmol/m³]), PO_4 (Phosphate [mmol/m³]), SCV (Sea-water velocity [m/s]), SPM (Suspended matter [g/m³]), SSS (Sea-surface salinity [10^{-3}]), SST (Sea-surface temperature [°C]), Bathy (Depth [m]), CROP (Cropland density [%]), LAND (Land-surface density [%]), PDEN (Population density [hab km^{-2}]), URBA (Built-up area density [%]), VBD (Boat-detection frequency [%]).

Environmental Variable	R^2 environmental variable ~ PC1 (5.5%) + PC2 (2.7%)	F-test p-value
s_allDB_SCV_avg	0.903962	1.08236464453023e-56
s_allDB_NO3_avg	0.856786	3.79478943883632e-47
s_allDB_NO3_sd	0.844704	3.26461028134746e-45
s_allDB_O2_sd	0.822457	5.14861518093247e-42
s_allDB_PO4_avg	0.810137	2.06156422536294e-40
s_allDB_PH_avg	0.754764	2.67369433133405e-34
s_allDB_SCV_sd	0.660002	1.7030748241232e-26
s_allDB_SSS_avg	0.659718	1.78292860521255e-26
s_allDB_SSS_sd	0.65358	4.76609368155036e-26
s_allDB_CHL_avg	0.643954	2.15187250507132e-25
s_allDB_PH_sd	0.628421	2.2533337569308e-24
s_allDB_CHL_sd	0.607696	4.45899997503829e-23
s_allDB_PO4_sd	0.591543	4.10226876152772e-22
s_allDB_SST_avg	0.544312	1.68548806521112e-19
BATHY	0.417126	1.27823981166561e-13
s_allDB_CROP_sd	0.411975	2.07383073452651e-13
s_allDB_CROP_avg	0.371968	7.74431390787001e-12
s_allDB_O2_avg	0.362861	1.70956579415095e-11
s_allDB_SPM_sd	0.335122	1.78156240216477e-10

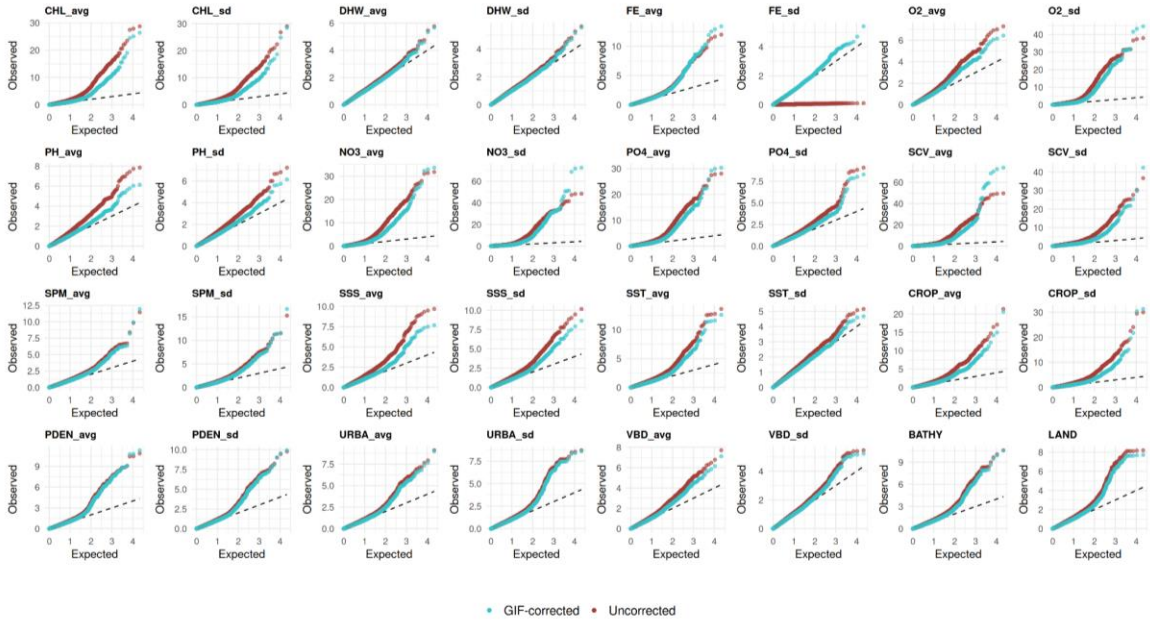
Environmental Variable	R^2 environmental variable ~ PC1 (5.5%) + PC2 (2.7%)	F-test p-value
s_allDB_DHW_avg	0.312524	1.11985874705775e-09
s_allDB_FE_avg	0.29125	5.98584177961648e-09
s_allDB_SPM_avg	0.261754	5.63733908861514e-08
s_allDB_SST_sd	0.255608	8.89361043521714e-08
s_allDB_DHW_sd	0.235377	3.88692201225712e-07
s_allDB_URBA_sd	0.140572	0.000241
s_allDB_VBD_avg	0.128462	0.00052
LAND	0.123217	0.000723
s_allDB_URBA_avg	0.094682	0.004208
s_allDB_VBD_sd	0.092419	0.004827
s_allDB_PDEN_sd	0.086633	0.006847
s_allDB_PDEN_avg	0.078686	0.011026
s_allDB_FE_sd	0.078289	0.01129

Twelve environmental variables showed associations with and $R^2 > 0.6$ with genetic variation represented by PC1 (5.5%) and PC2 (2.7%), including sea-water velocity average ($R^2 = 0.904$, $p < 0.001$), nitrate average ($R^2 = 0.857$, $p < 0.001$), nitrate standard deviation ($R^2 = 0.845$, $p < 0.001$), oxygen standard deviation ($R^2 = 0.822$, $p < 0.001$), phosphate average ($R^2 = 0.810$, $p < 0.001$), pH average ($R^2 = 0.755$, $p < 0.001$), sea-water velocity standard deviation ($R^2 = 0.660$, $p < 0.001$), sea-surface salinity average ($R^2 = 0.659$, $p < 0.001$), sea-surface salinity standard deviation ($R^2 = 0.654$, $p < 0.001$), chlorophyll average ($R^2 = 0.644$, $p < 0.001$), pH standard deviation ($R^2 = 0.628$, $p < 0.001$), and chlorophyll standard deviation ($R^2 = 0.608$, $p < 0.001$). This indicates substantial overlap between genetic and environmental variation for these variables, which may limit the ability of LFMM to distinguish environmental associations from population structure and potentially contribute to p-value inflation.

LFMM QQ-Plot output for K = 4



LFMM QQ-Plot output for K = 6



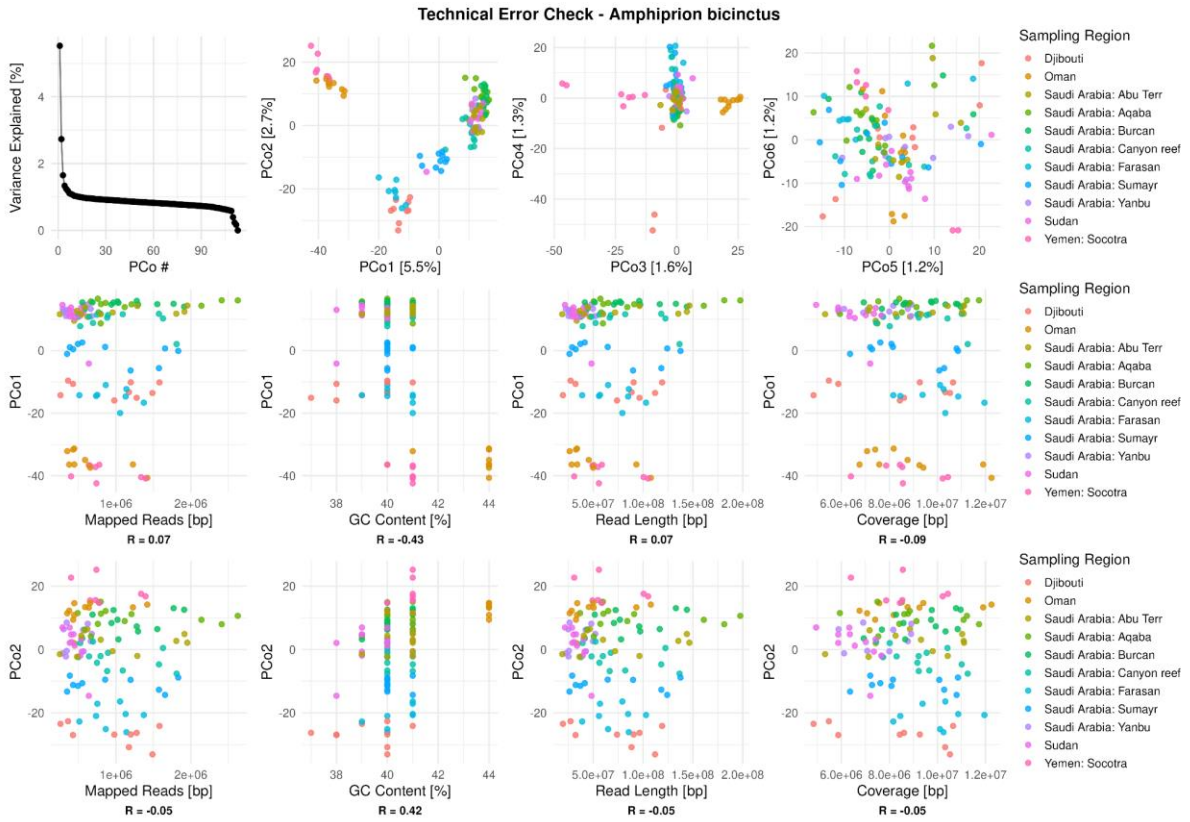
S Figure 7: QQ Plots for each environmental variable showing the distribution of expected vs. observed p-values for A. bicinctus (LFMM method), Left: LFMM run with K = 4, Right: LFMM run with K = 6.
Environmental variables: CHL (Chlorophyll [mg/m³]), DHW (Degree Heating Weeks [°C·week⁻¹]), FE (Iron [mmol/m³]), O_2 (Oxygen [mmol/m³]), PH (pH), NO_3 (Nitrate [mmol/m³]), PO_4 (Phosphate [mmol/m³]), SCV (Sea-water velocity [m/s]), SPM (Suspended matter [g/m³]), SSS (Sea-surface salinity [10^{-3}]), SST (Sea-surface temperature [°C]), Bathy (Depth [m]), CROP (Cropland density [%]), LAND (Land-surface density [%]), PDEN (Population density [hab km^{-2}]), URBA (Built-up area density [%]), VBD (Boat-detection frequency [%]).

Inflation was only marginally reduced compared with the default K = 3, indicating that adding latent factors beyond the structure-optimal K did not resolve the elevated test statistics.

Thus, the overlap between environmental and genetic variation may explain why inflated p-values persisted in LFMM.

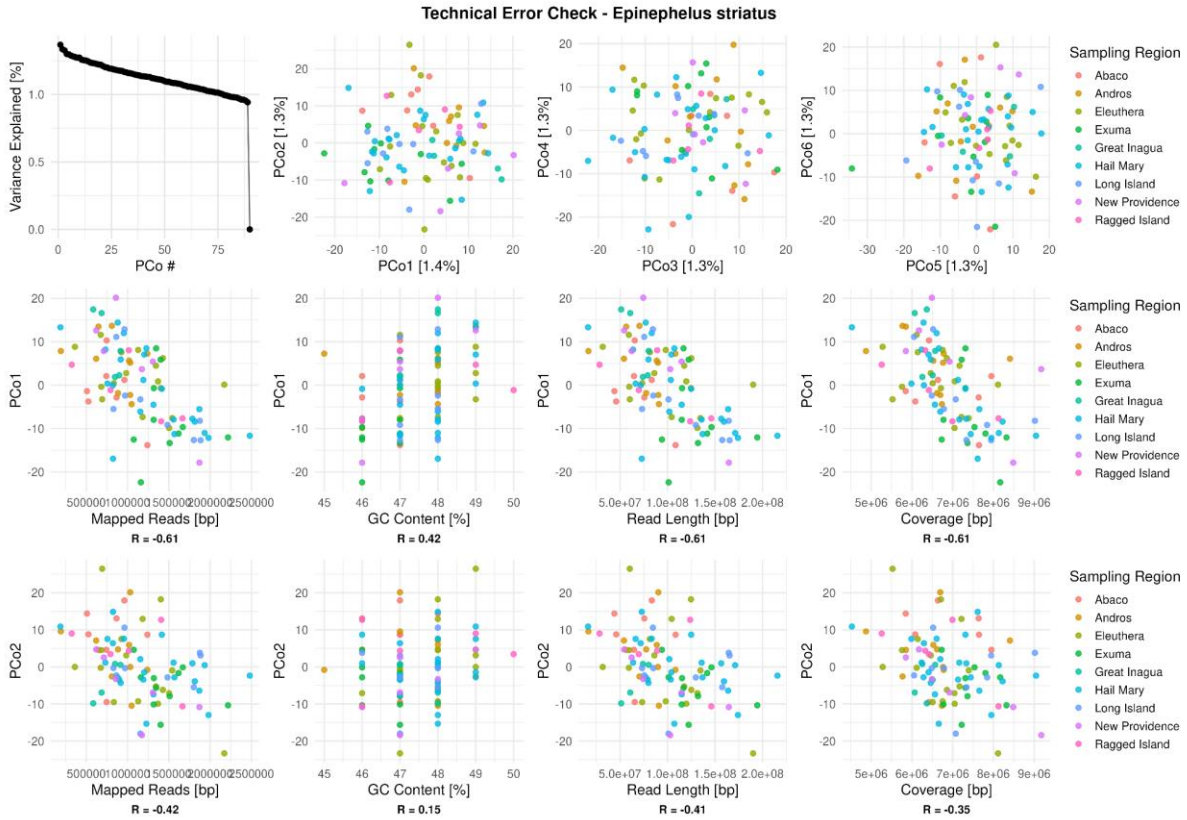
1.3.2 PCoA Analysis

A. bicinctus:



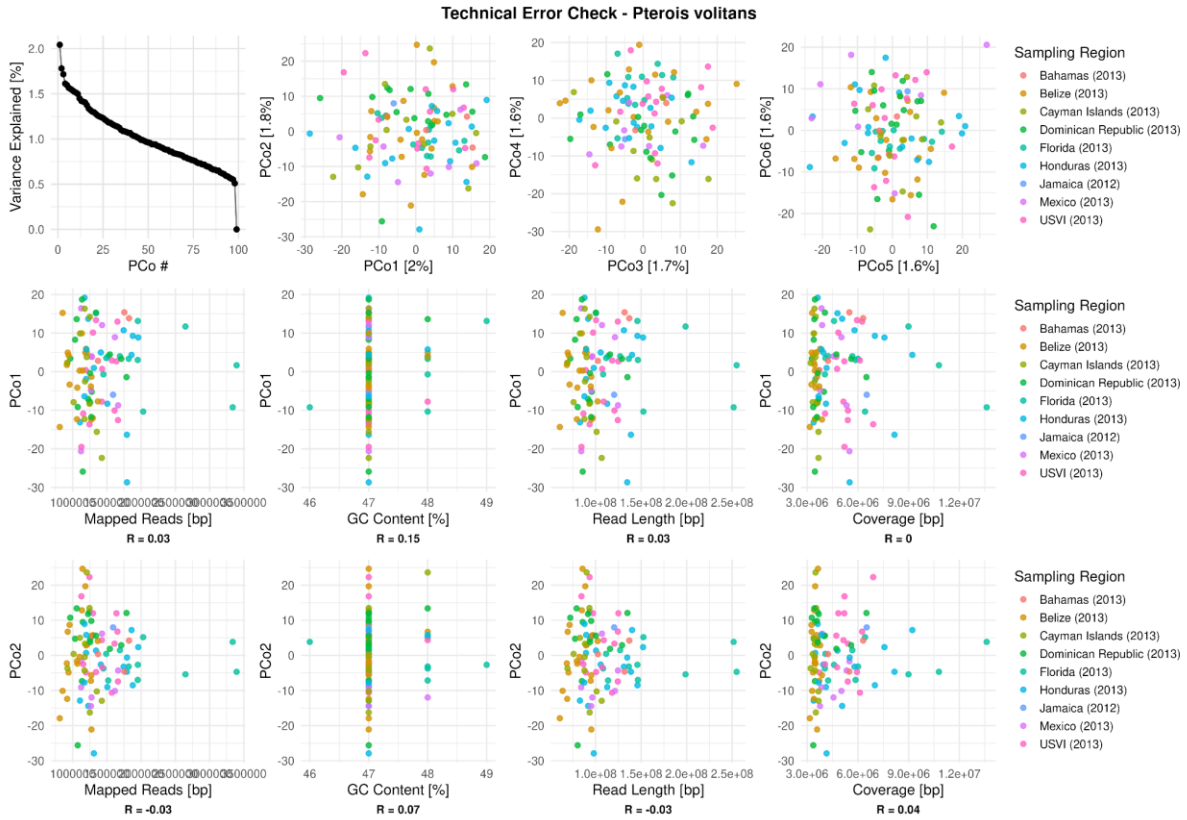
S Figure 5: PCoA of the genotype matrix for A. bicinctus, plotted against selected MultiQC metrics coloured by sampling region. Pearson's R values indicate the correlation between PC1 or PC2 and each respective MultiQC metric. Interpreted in Interpreted in *1.3.1 A. Bicinctus - Special Supplementary Analysis.*

E. striatus:



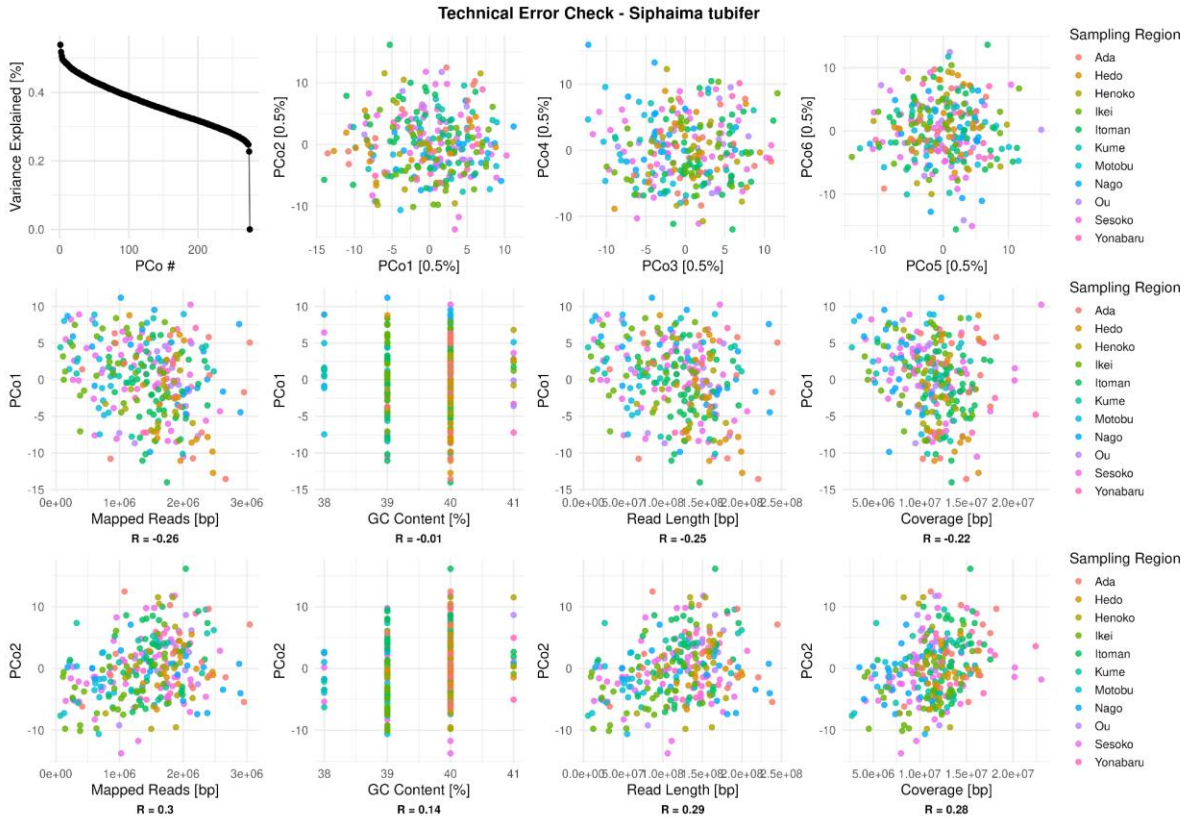
S Figure 8: PCoA of the genotype matrix for E. striatus, plotted against selected MultiQC metrics coloured by sampling region. Pearson's R values indicate the correlation between PC1 or PC2 and each respective MultiQC metric. No clear separation by sampling region was observed along any axis. Correlations with QC metrics were weak to moderate, reaching R = -0.61 for mapped reads and read length, suggesting some residual technical influence but no clear evidence of population structure.

P. volitans:



S Figure 9: PCoA of the genotype matrix for P. volitans, plotted against selected MultiQC metrics coloured by sampling region. Pearson's R values indicate the correlation between PC1 or PC2 and each respective MultiQC metric. Sampling sites overlapped strongly across all axes, with no clear regional clustering, consistent with the very low F_{st} observed for this species (S Figure 12).

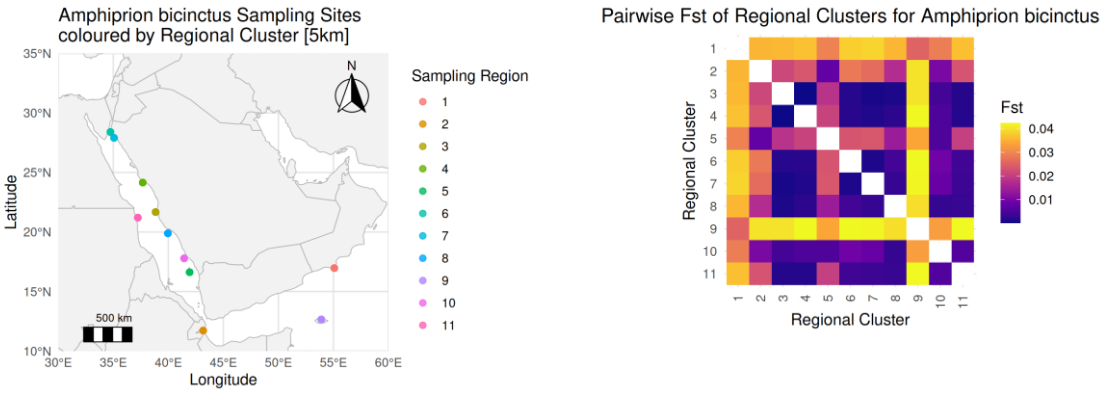
S. tubifer:



S Figure 10: PCoA of the genotype matrix for S. tubifer, plotted against selected MultiQC metrics coloured by sampling region. Pearson's R values indicate the correlation between PC1 or PC2 and each respective MultiQC metric. The variance explained curve was the flattest of the four species, with PC1 explaining only approximately 0.5% of the variance, and no clear clustering by region, indicating little detectable population structure in the largest dataset.

1.3.3 F_{st} Analysis

A. bicinctus:



S Figure 2: Left: Map of Sample Sites for A. bicinctus coloured by Regional Cluster [5km], Right: Heatmap of pairwise Fst for A. bicinctus using the Regional Clusters. Interpreted in Interpreted in 1.3.1 A. Bicinctus - Special Supplementary Analysis.

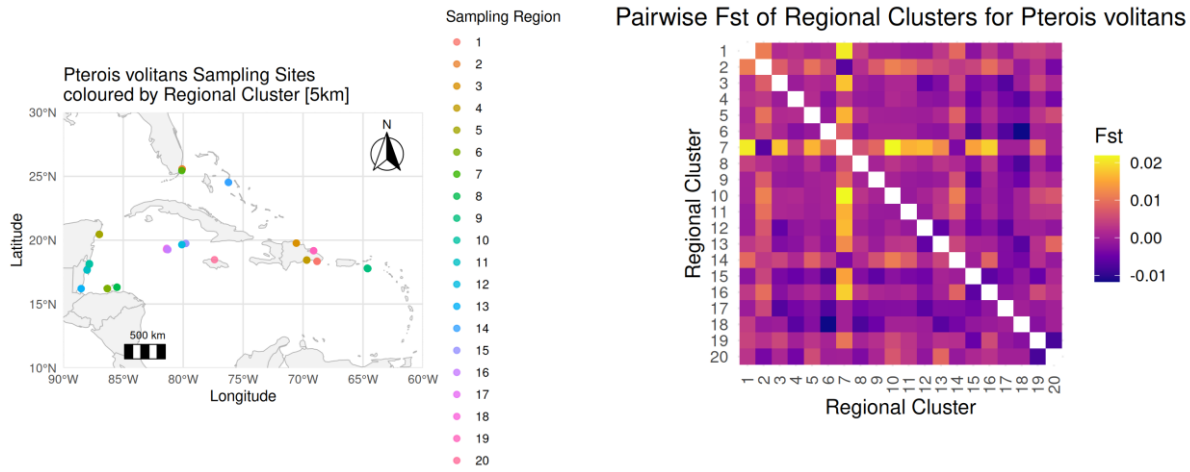
.

E. striatus:



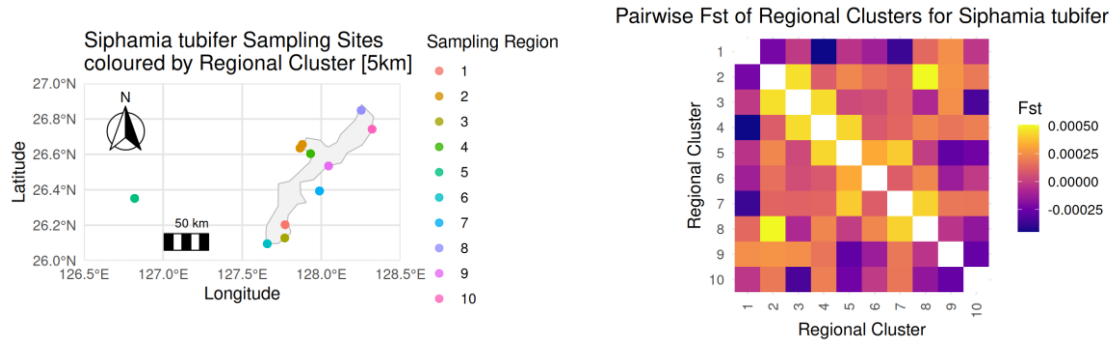
S Figure 11: Left: Map of Sample Sites for E. striatus coloured by Regional Cluster [5km], Right: Heatmap of pairwise F_{st} for E. striatus using the Regional Clusters. F_{st} values were uniformly low (approximately 0-0.0015), with no clear geographic pattern, indicating little to no genetic differentiation.

P. volitans:



S Figure 12: Left: Map of Sample Sites for P. volitans coloured by Regional Cluster [5km], Right: Heatmap of pairwise F_{st} for P. volitans using the Regional Clusters. Cluster 7 (Florida) showed higher F_{st} values relative to several other clusters, indicating localised differentiation within an otherwise weakly structured and widely distributed invasive population.

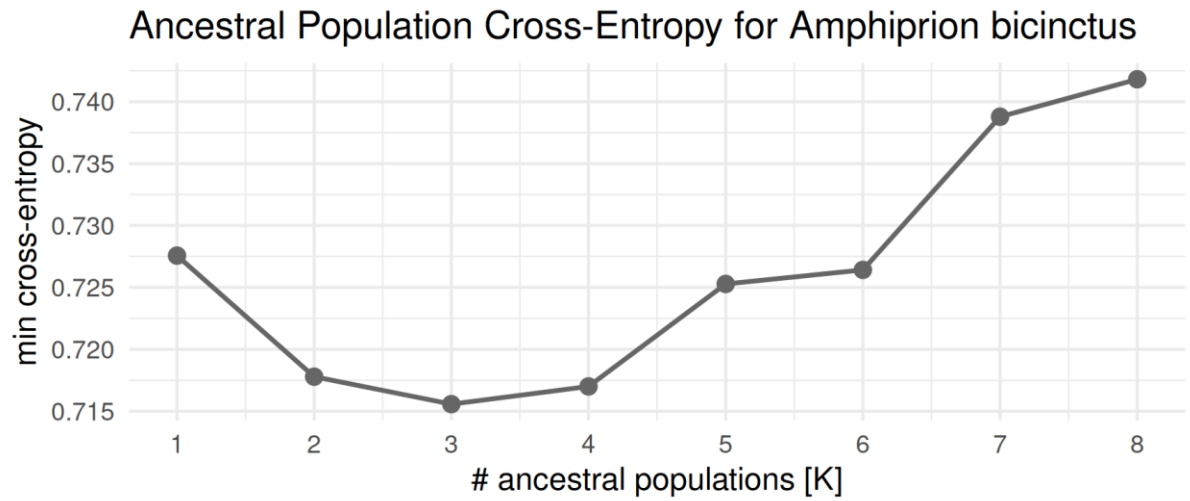
S. tubifer:



S Figure 13: Left: Map of Sample Sites for S. tubifer coloured by Regional Cluster [5km], Right: Heatmap of pairwise F_{st} for S. tubifer using the Regional Clusters. Differentiation was minimal and spatially inconsistent.

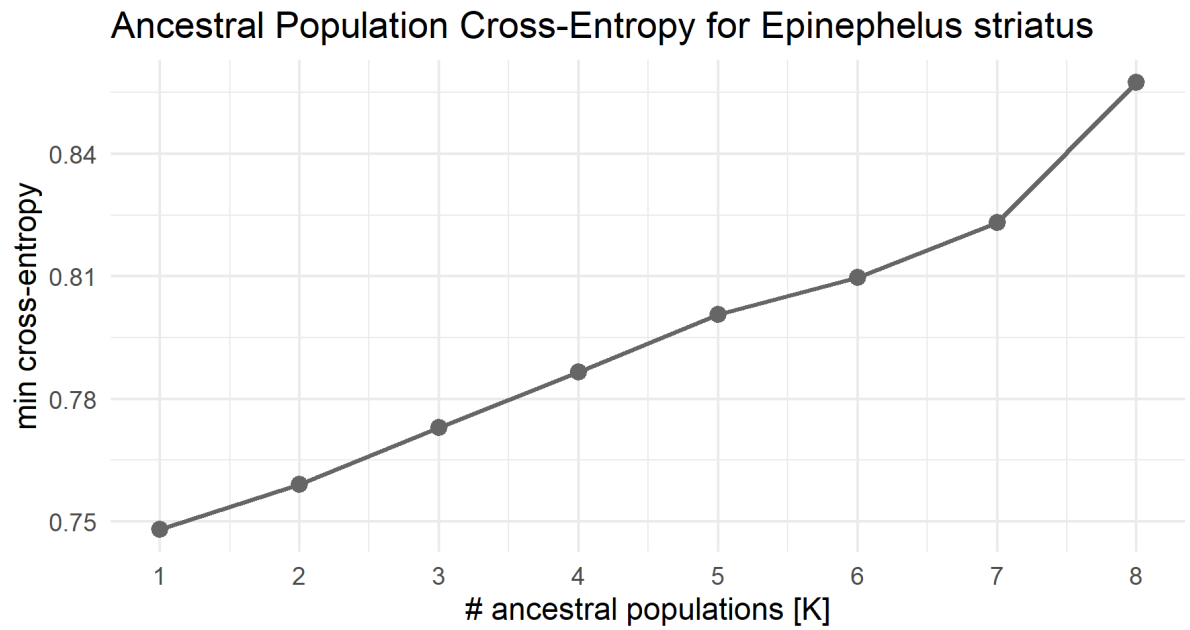
1.3.4 Ancestral Population Analysis

A. bicinctus:



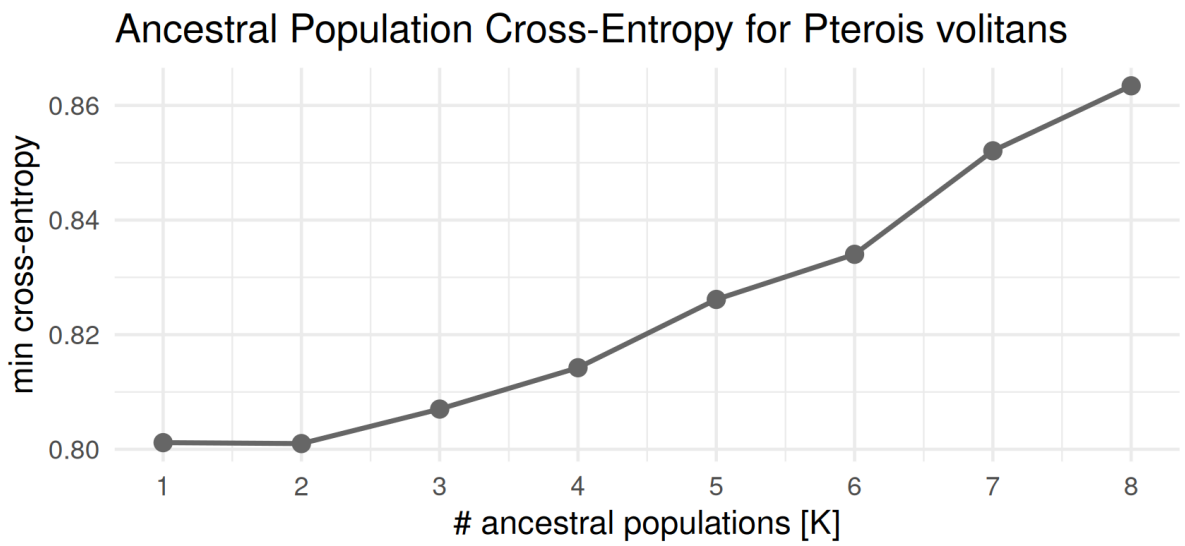
S Figure 3: This minimal cross-entropy plot indicates the number of ancestral populations that best suit A. bicinctus Samples. Interpreted in 1.3.1 A. Bicinctus - Special Supplementary Analysis.

E. striatus:



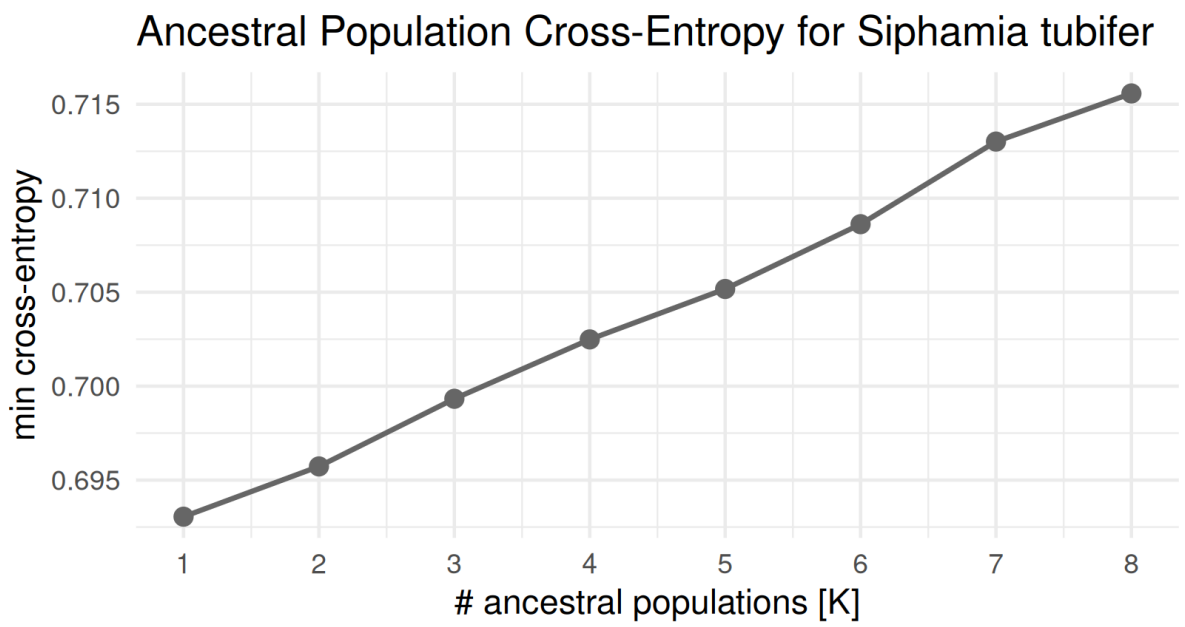
S Figure 14: This minimal cross-entropy plot indicates the number of ancestral populations that best suit E. striatus samples. The cross-entropy increased monotonically from K = 1, supporting K = 1 as the optimal value and indicating no clear population substructure.

P. volitans:



S Figure 15: This minimal cross-entropy plot indicates the number of ancestral populations that best suit P. volitans samples. A shallow decrease from K = 1 to K = 2 followed by an increase supports K = 2, although the small difference suggests weak population structure.

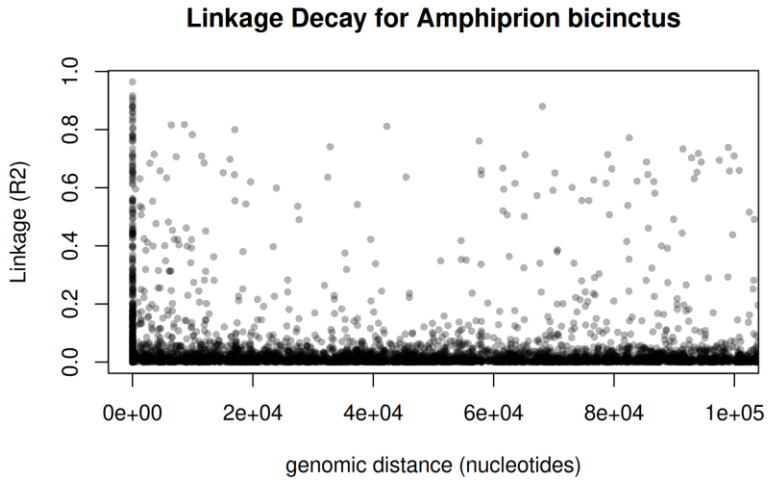
S. tubifer:



S Figure 16: This minimal cross-entropy plot indicates the number of ancestral populations that best suit S. tubifer samples. The cross-entropy increased mono-tonically from K = 1, supporting K = 1 as the optimal value and indicating no clear population substructure.

1.3.5 Linkage Disequilibrium Decay Analysis

A. bicinctus:



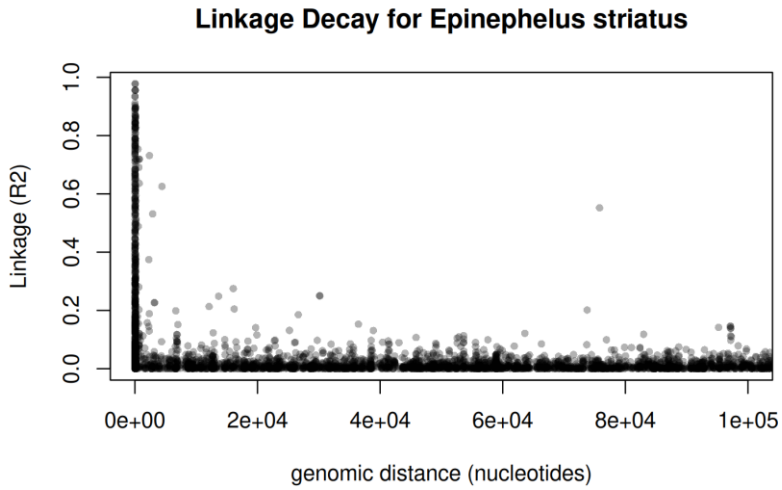
S Figure 17: LD decay figure for A. bicinctus for chromosome sequence: NC_072778.1.

A sharp decline within the first few kb was followed by a long tail of scattered higher R^2 values beyond 80 kb, consistent with the reported ≈ 9 kb decay to background. This extended LD pattern may reflect the use of a closely related species as the reference genome rather than true biological LD.

S Table 5: LD decay table for A. bicinctus for chromosome sequence: NC_072778.1.

LD decay rate	**50%**	**75%**	**90%**	**95%**
bp	2068.711	4137.423	6872.111	8940.822

E. striatus:



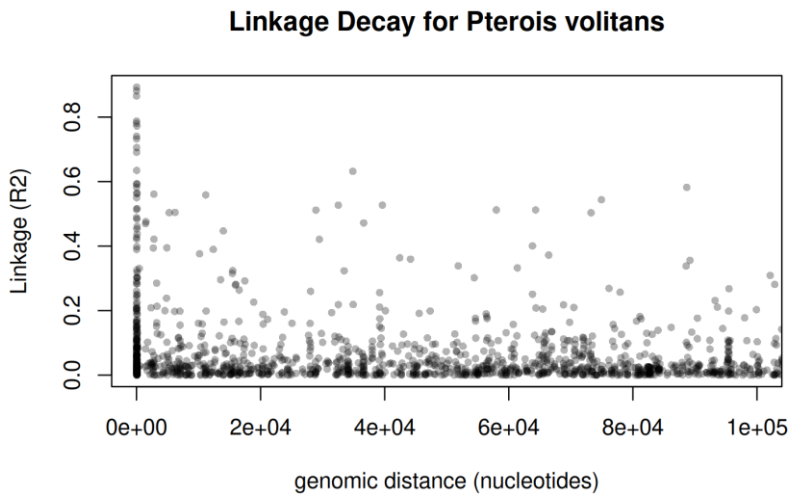
S Figure 18: LD decay figure for E. striatus for chromosome sequence: CM069296.1.

A sharp early drop but a shorter tail than *A. bicinctus*, consistent with the smaller estimated decay distance of ≈ 4 kb.

S Table 6: LD decay table for E. striatus for chromosome sequence: CM069296.1.

LD decay rate	50%	75%	90%	95%
bp	813.9134	1627.8268	2703.7618	3517.6752

P. volitans:



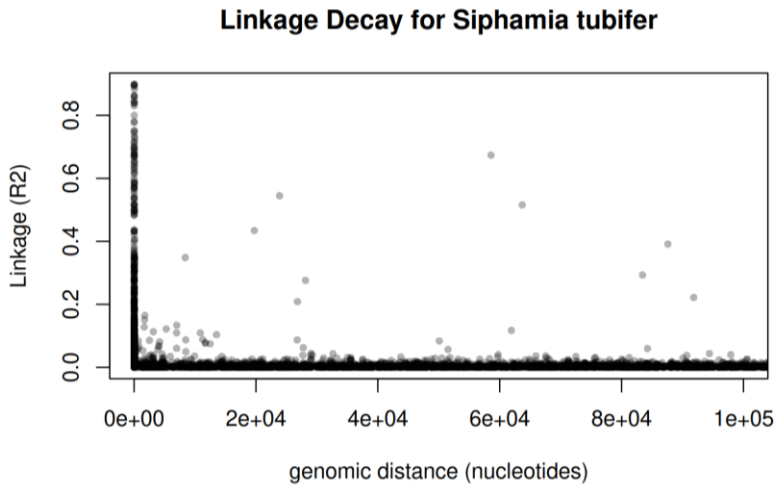
S Figure 19: LD decay figure for P. volitans for contig sequence: CAMQGK010000052.1.

The noisiest pattern of the four species, with scattered moderate to high R^2 values across the full range, consistent with the longest estimated decay distance of ≈ 11 kb. The 10 kb window therefore slightly underestimates the 95% decay distance. The use of a closely related species as the reference genome and a contig-level assembly may contribute to this pattern.

S Table 7: LD decay table for P. volitans for contig sequence: CAMQGK010000052.1.

LD decay rate	50%	75%	90%	95%
bp	2544.183	5088.366	8451.593	10995.776

S. tubifer:



S Figure 20: LD decay figure for S. tubifer for chromosome sequence: CM035794.1.

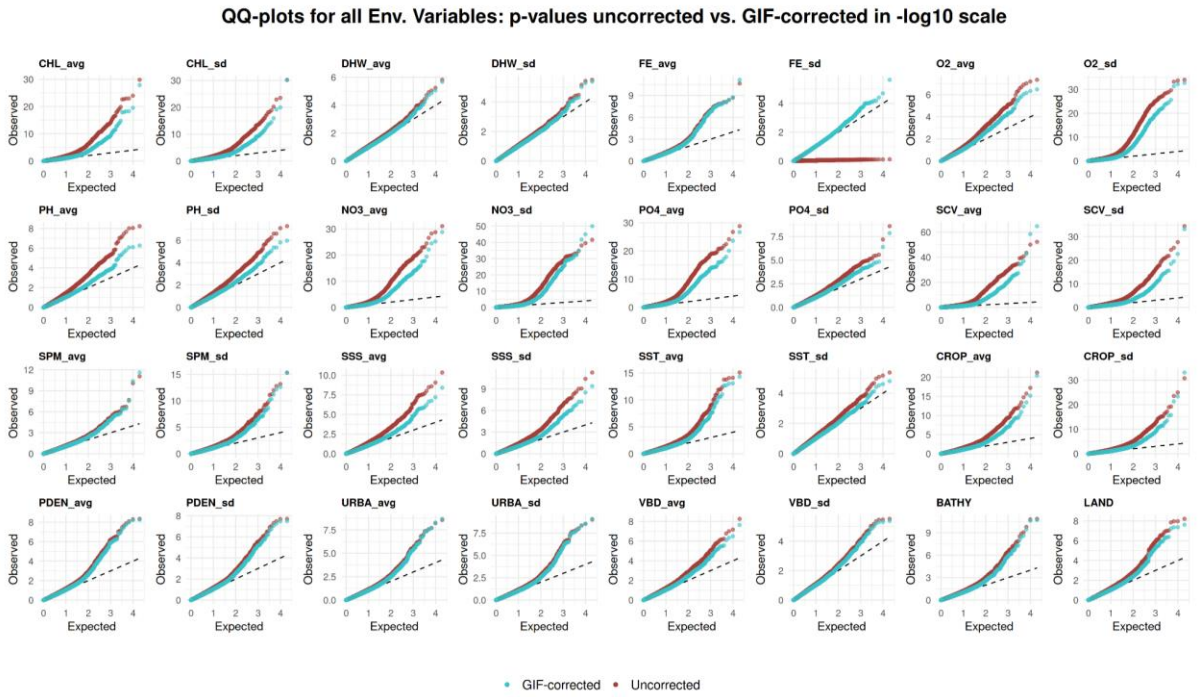
The fastest decay of the four species, with most values approaching zero beyond 1-2 kb, consistent with the short estimated decay distance of ≈ 0.5 kb. This may provide the clearest LD pattern, as this is the only species analysed using its own reference genome.

S Table 8: LD decay table for S. tubifer for chromosome sequence: CM035794.1.

LD decay rate	50%	75%	90%	95%
bp	127.2302	254.4603	422.6494	549.8796

1.3.6 Latent Factor Mixed Modell

A. bicinctus:

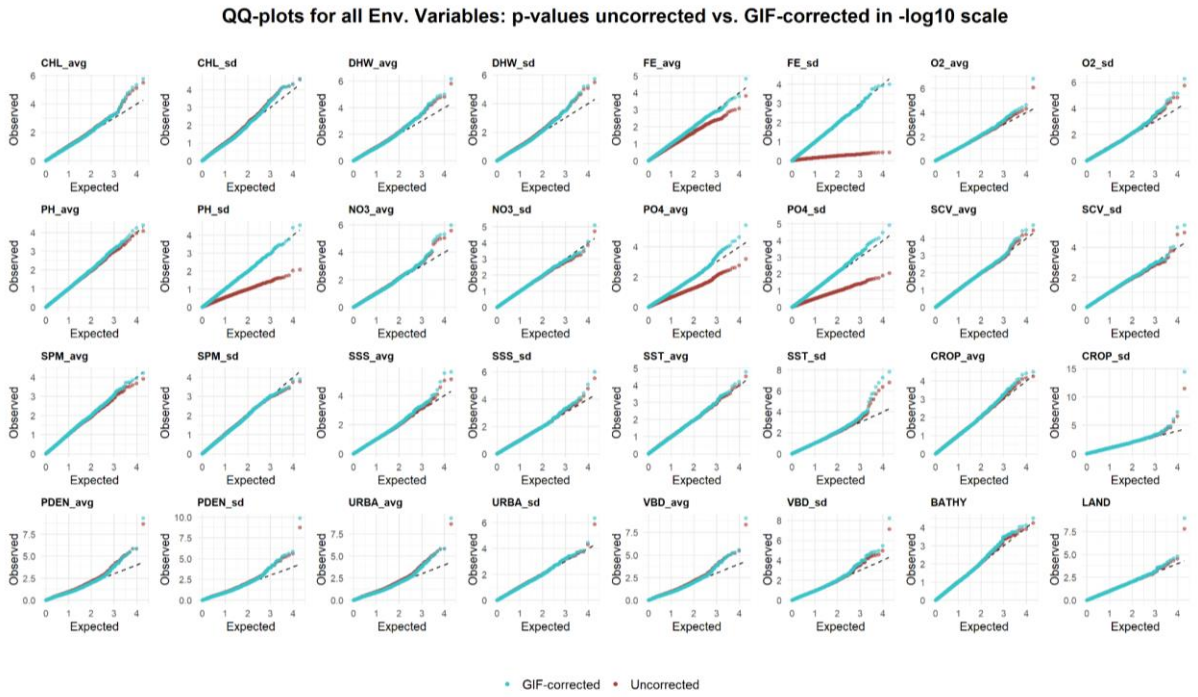


S Figure 1: QQ Plots for each environmental variable showing the distribution of expected vs. observed p-values for A. bicinctus (LFMM method).

1.3.1 A. Bicinctus - Special Supplementary Analysis.

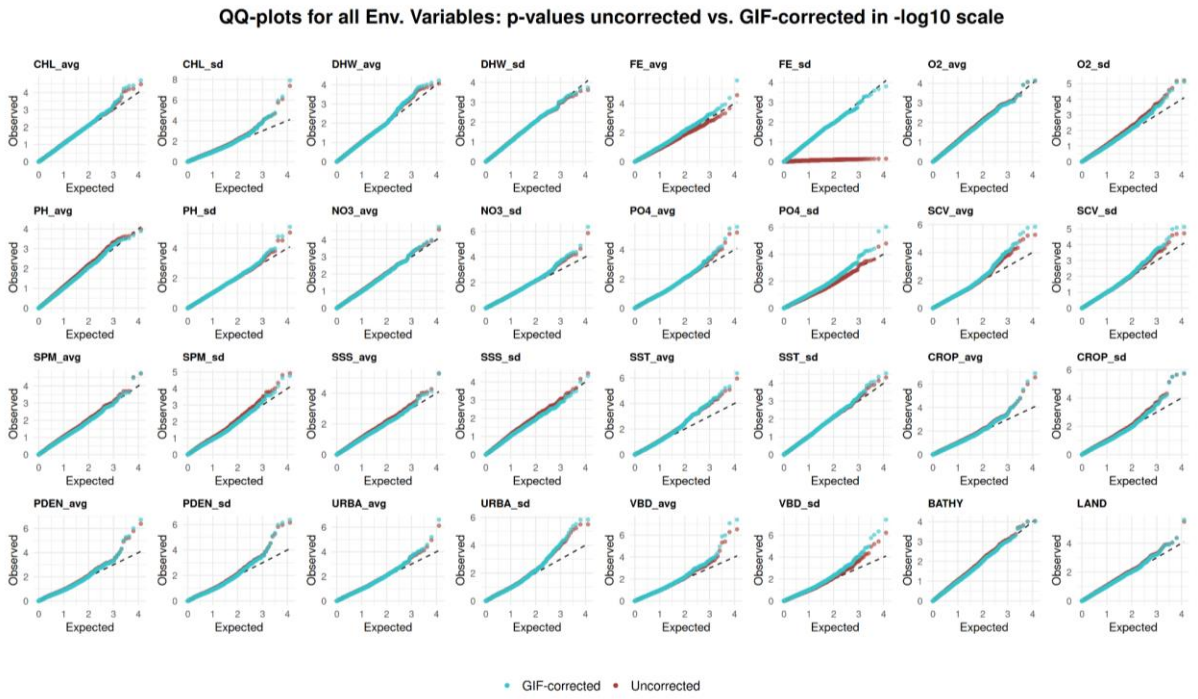
Environmental variables: CHL (Chlorophyll [mg/m³]), DHW (Degree Heating Weeks [°C·week⁻¹]), FE (Iron [mmol/m³]), O_2 (Oxygen [mmol/m³]), PH (pH), NO_3 (Nitrate [mmol/m³]), PO_4 (Phosphate [mmol/m³]), SCV (Sea-water velocity [m/s]), SPM (Suspended matter [g/m³]), SSS (Sea-surface salinity [10^{-3}]), SST (Sea-surface temperature [°C]), Bathy (Depth [m]), CROP (Cropland density [%]), LAND (Land-surface density [%]), PDEN (Population density [hab km^{-2}]), URBA (Built-up area density [%]), VBD (Boat-detection frequency [%]).

E. striatus:



S Figure 21: QQ Plots for each environmental variable showing the distribution of expected vs. observed p-values for E.striatus (LFMM method). The values mainly follow the expected diagonal across nearly all variables, indicating good calibration.
Environmental variables: CHL (Chlorophyll [mg/m³]), DHW (Degree Heating Weeks [°C·week⁻¹]), FE (Iron [mmol/m³]), O_2 (Oxygen [mmol/m³]), PH (pH), NO_3 (Nitrate [mmol/m³]), PO_4 (Phosphate [mmol/m³]), SCV (Sea-water velocity [m/s]), SPM (Suspended matter [g/m³]), SSS (Sea-surface salinity [10^{-3}]), SST (Sea-surface temperature [°C]), Bathy (Depth [m]), CROP (Cropland density [%]), LAND (Land-surface density [%]), PDEN (Population density [hab km^{-2}]), URBA (Built-up area density [%]), VBD (Boat-detection frequency [%]).

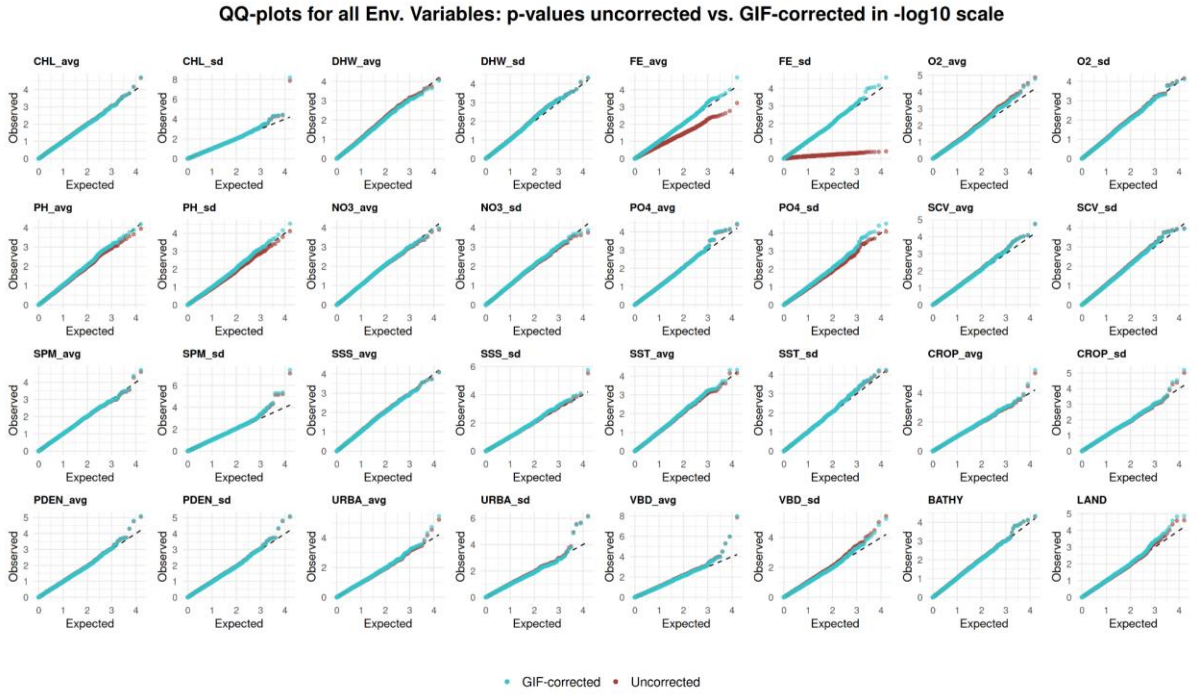
P. volitans:



S Figure 22: QQ Plots for each environmental variable showing the distribution of expected vs. observed p-values for P. volitans (LFMM method). The values mainly follow the expected diagonal across nearly all variables, indicating good calibration.
Environmental variables: CHL (Chlorophyll [mg/m³]), DHW (Degree Heating Weeks [°C·week⁻¹]), FE (Iron [mmol/m³]), O_2 (Oxygen [mmol/m³]), PH (pH), NO_3 (Nitrate [mmol/m³]), PO_4 (Phosphate [mmol/m³]), SCV (Sea-water velocity [m/s]), SPM (Suspended matter [g/m³]), SSS (Sea-surface

salinity [10⁻³]), SST (Sea-surface temperature [°C]), Bathy (Depth [m]), CROP (Cropland density [%]), LAND (Land-surface density [%]), PDEN (Population density [hab km⁻²]), URBA (Built-up area density [%]), VBD (Boat-detection frequency [%]).

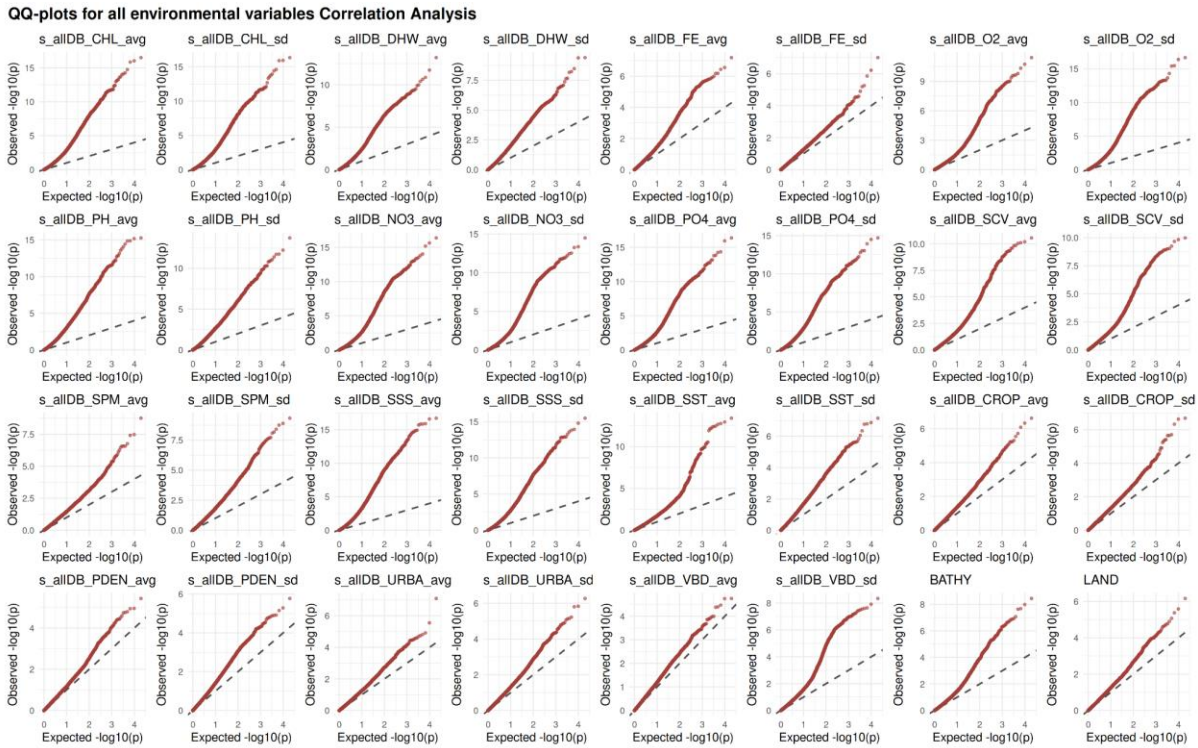
S. tubifer:



S Figure 23: QQ Plots for each environmental variable showing the distribution of expected vs. observed p-values for S. tubifer (LFMM method). The values mainly follow the expected diagonal across nearly all variables, indicating good calibration.
Environmental variables: CHL (Chlorophyll [mg/m³]), DHW (Degree Heating Weeks [°C·week⁻¹]), FE (Iron [mmol/m³]), O_2 (Oxygen [mmol/m³]), PH (pH), NO_3 (Nitrate [mmol/m³]), PO_4 (Phosphate [mmol/m³]), SCV (Sea-water velocity [m/s]), SPM (Suspended matter [g/m³]), SSS (Sea-surface salinity [10⁻³]), SST (Sea-surface temperature [°C]), Bathy (Depth [m]), CROP (Cropland density [%]), LAND (Land-surface density [%]), PDEN (Population density [hab km⁻²]), URBA (Built-up area density [%]), VBD (Boat-detection frequency [%]).

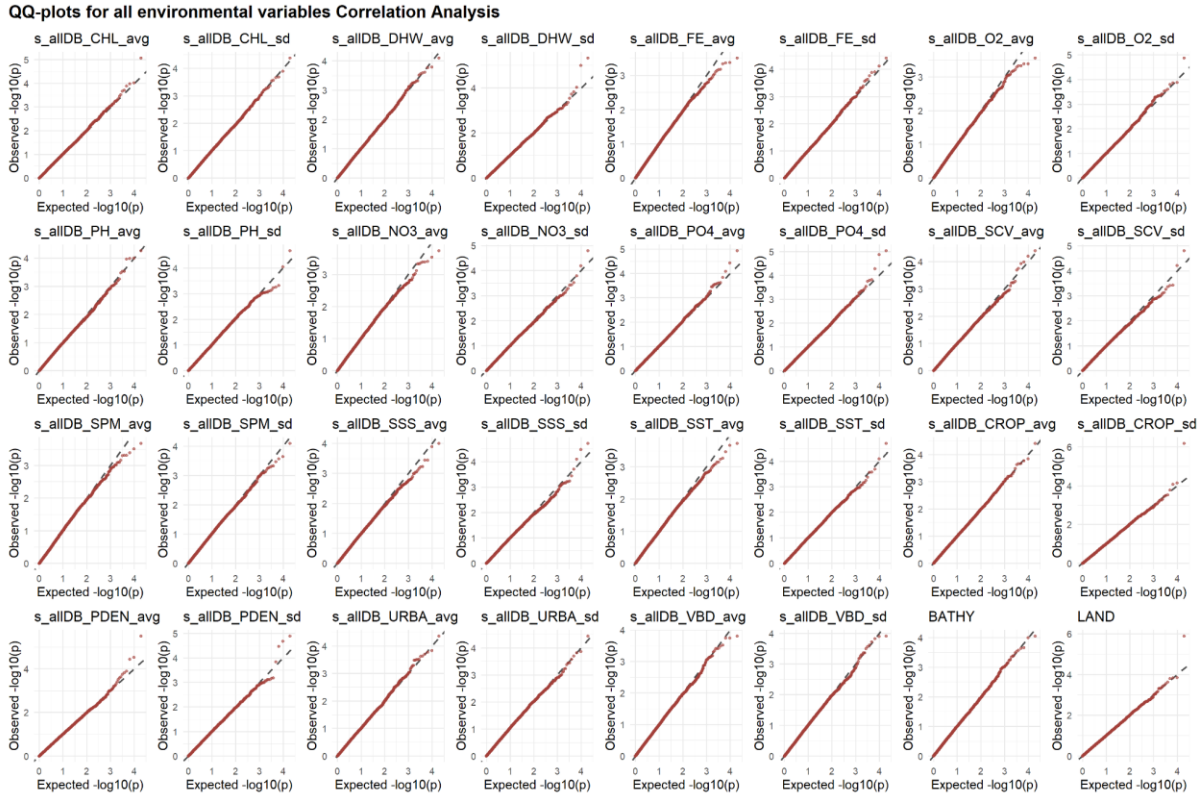
1.3.7 Correlation Analysis

A. bicinctus:



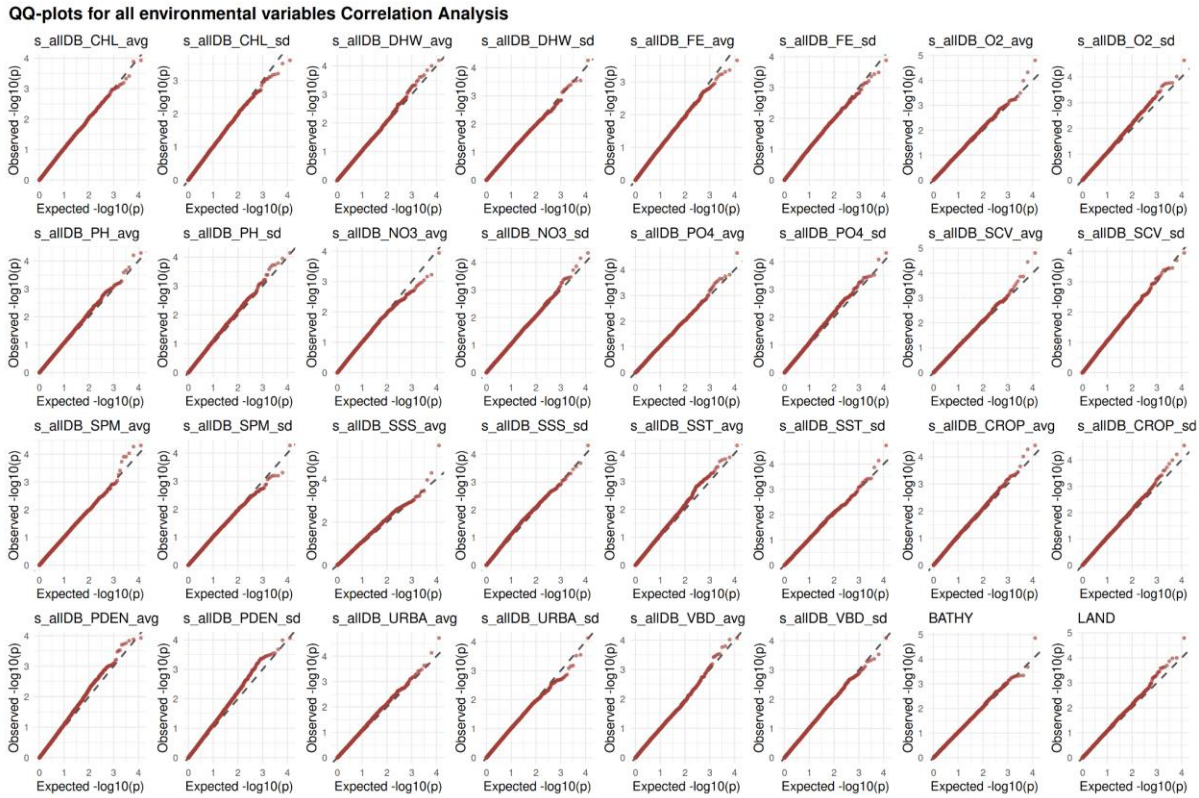
S Figure 24: QQ Plots for each environmental variable showing the distribution of expected vs. observed p-values for A. bicinctus (CA method). Observed -log10(p) values reached 10-15, compared with an expected range of approximately 3-4, indicating strong p-value inflation. This is consistent with CA not accounting for population structure.
Environmental variables: CHL (Chlorophyll [mg/m³]), DHW (Degree Heating Weeks [°C·week⁻¹]), FE (Iron [mmol/m³]), O_2 (Oxygen [mmol/m³]), PH (pH), NO_3 (Nitrate [mmol/m³]), PO_4 (Phosphate [mmol/m³]), SCV (Sea-water velocity [m/s]), SPM (Suspended matter [g/m³]), SSS (Sea-surface salinity [10^{-3}]), SST (Sea-surface temperature [°C]), Bathy (Depth [m]), CROP (Cropland density [%]), LAND (Land-surface density [%]), PDEN (Population density [hab km^{-2}]), URBA (Built-up area density [%]), VBD (Boat-detection frequency [%]).

E. striatus:



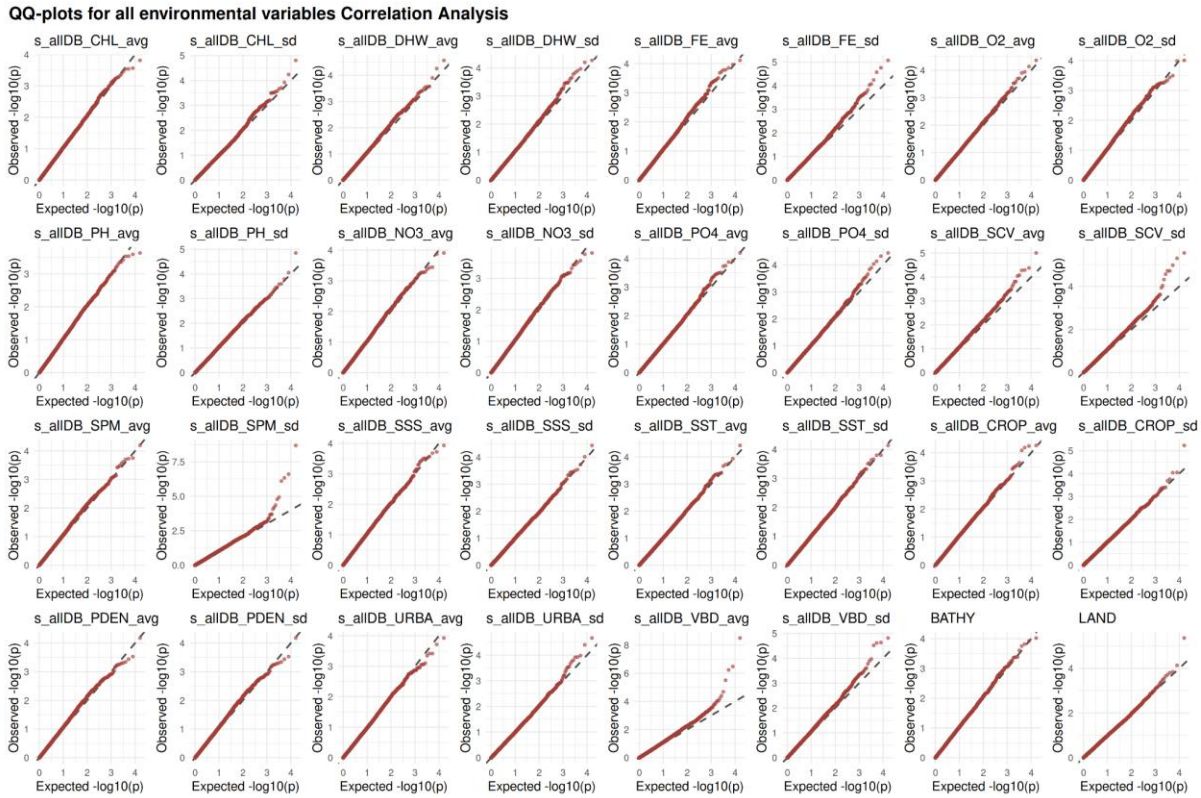
S Figure 25: QQ Plots for each environmental variable showing the distribution of expected vs. observed p-values for E. striatus (CA method). The values mainly follow the expected diagonal across nearly all variables, indicating good calibration.
Environmental variables: CHL (Chlorophyll [mg/m³]), DHW (Degree Heating Weeks [°C·week⁻¹]), FE (Iron [mmol/m³]), O_2 (Oxygen [mmol/m³]), PH (pH), NO_3 (Nitrate [mmol/m³]), PO_4 (Phosphate [mmol/m³]), SCV (Sea-water velocity [m/s]), SPM (Suspended matter [g/m³]), SSS (Sea-surface salinity [10^{-3}]), SST (Sea-surface temperature [°C]), Bathy (Depth [m]), CROP (Cropland density [%]), LAND (Land-surface density [%]), PDEN (Population density [hab km^{-2}]), URBA (Built-up area density [%]), VBD (Boat-detection frequency [%]).

P. volitans:



S Figure 26: QQ Plots for each environmental variable showing the distribution of expected vs. observed p-values for P. volitans (CA method). The values mainly follow the expected diagonal across nearly all variables, indicating good calibration.
Environmental variables: CHL (Chlorophyll [mg/m³]), DHW (Degree Heating Weeks [°C·week⁻¹]), FE (Iron [mmol/m³]), O_2 (Oxygen [mmol/m³]), PH (pH), NO_3 (Nitrate [mmol/m³]), PO_4 (Phosphate [mmol/m³]), SCV (Sea-water velocity [m/s]), SPM (Suspended matter [g/m³]), SSS (Sea-surface salinity [10^{-3}]), SST (Sea-surface temperature [°C]), Bathy (Depth [m]), CROP (Cropland density [%]), LAND (Land-surface density [%]), PDEN (Population density [hab km^{-2}]), URBA (Built-up area density [%]), VBD (Boat-detection frequency [%]).

S. tubifer:



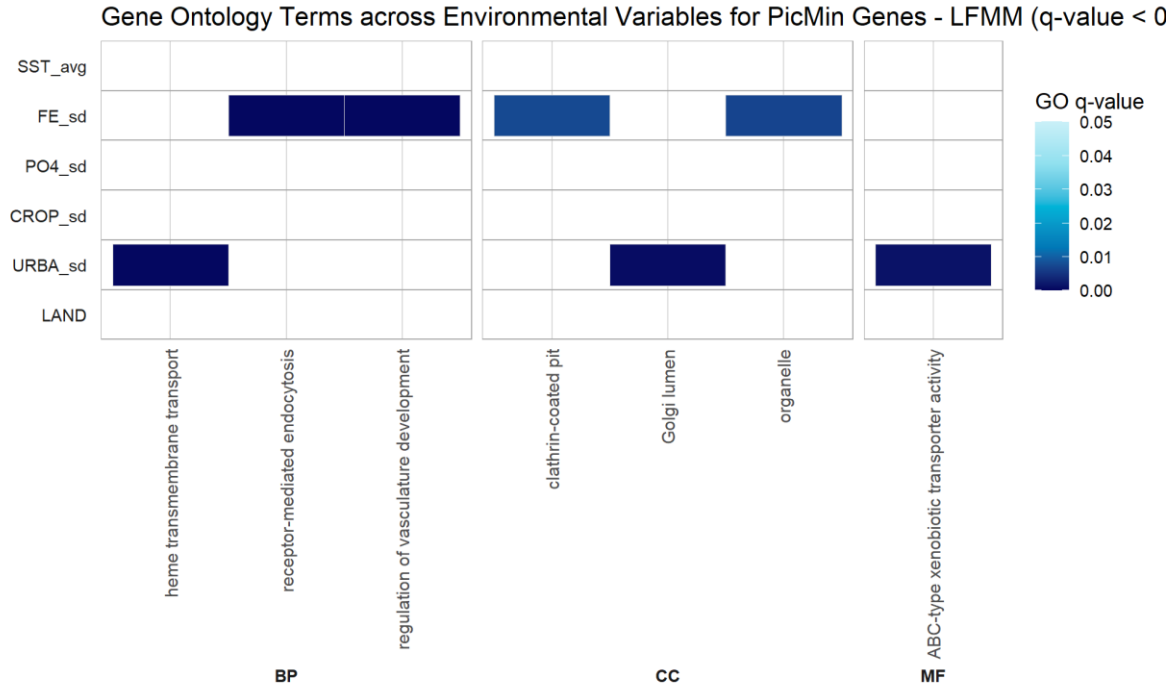
S Figure 27: QQ Plots for each environmental variable showing the distribution of expected vs. observed p-values for S. tubifer (CA method). The values mainly follow the expected diagonal across nearly all variables, indicating good calibration. Exception are SPM_sd and VBD_sd. Environmental variables: CHL (Chlorophyll [mg/m³]), DHW (Degree Heating Weeks [°C·week⁻¹]), FE (Iron [mmol/m³]), O_2 (Oxygen [mmol/m³]), PH (pH), NO_3 (Nitrate [mmol/m³]), PO_4 (Phosphate [mmol/m³]), SCV (Sea-water velocity [m/s]), SPM (Suspended matter [g/m³]), SSS (Sea-surface salinity [10^{-3}]), SST (Sea-surface temperature [°C]), Bathy (Depth [m]), CROP (Cropland density [%]), LAND (Land-surface density [%]), PDEN (Population density [hab km^{-2}]), URBA (Built-up area density [%]), VBD (Boat-detection frequency [%]).

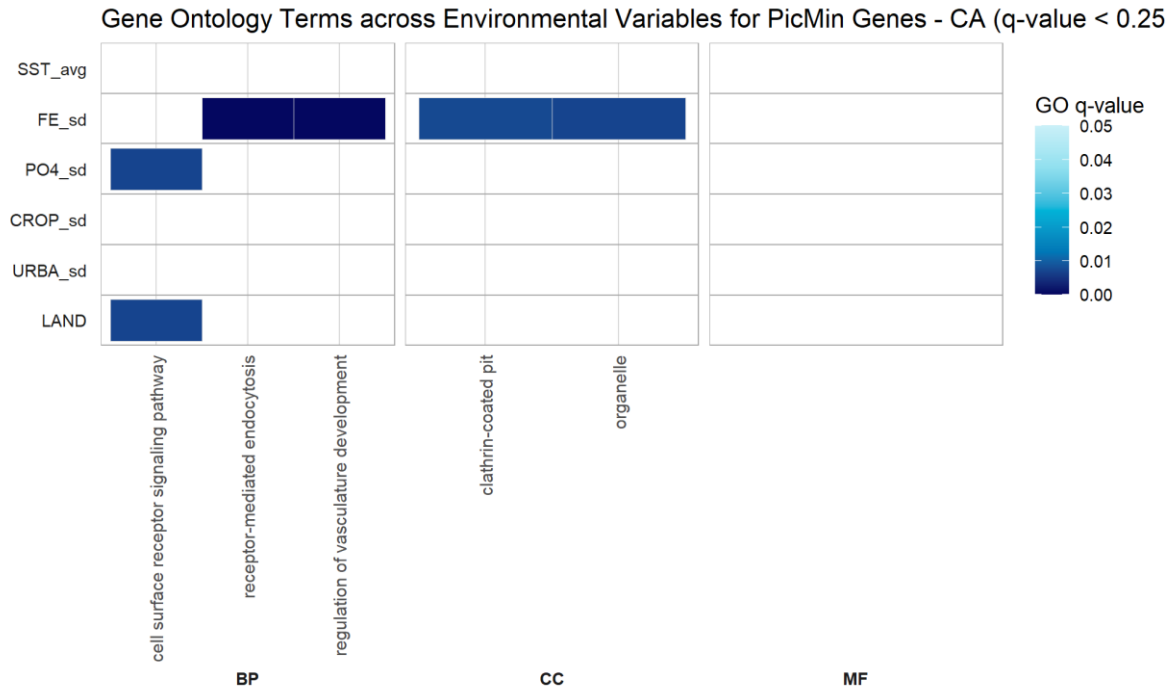
1.3.8 Overall GEA summary statistics table

S Table 9: Median and interquartile range of SNP-environment associations per environmental variable for each GEA method and species.

Grouping	GEA method / Species	Method	Median SNPs per environmental variable	IQR
method	LFMM	LFMM	69	173.8
method	CA	CA	782.5	2584
species	*Amphiprion bicinctus*	LFMM	62.5	168.2
species	*Amphiprion bicinctus*	CA	779.5	2584
species	*Epinephelus striatus*	LFMM	1	3
species	*Epinephelus striatus*	CA	0	0
species	*Pterois volitans*	LFMM	1	4
species	*Pterois volitans*	CA	0	0
species	*Siphamia tubifer*	LFMM	0	0
species	*Siphamia tubifer*	CA	0	0

1.3.9 Gene Set Enrichment Analysis





S Figure 28: Gene Ontology (GO) terms for biological process, cellular component, and molecular function associated with PicMin genes with pooled q-values < 0.25 from the LFMM (on the top) and CA (on the bottom) GEA methods. Avg = average and sd = standard deviation. Environmental variables: SST (Sea-surface temperature [°C]), FE (Iron [mmol/m³]), PO_4 (Phosphate [mmol/m³]), LAND (Land-surface density [%]), CROP (Cropland density [%]), URBA (Built-up area density [%]).

SetRank identified seven enriched GO terms in the LFMM-derived gene sets. Standard deviation of iron concentration was associated with receptor-mediated endocytosis (GO:0006898), regulation of vasculature development (GO:1901342), clathrin-coated pit (GO:0005905), and organelle (GO:0043226), all mapped to ENSDARG00000029146 (*lrp1ab*). Standard deviation of built-up area density was associated with heme transmembrane transport (GO:0035351), Golgi lumen (GO:0005796), and ABC-type xenobiotic transporter activity (GO:0008559), all mapped to ENSDARG00000061233.

For the CA-derived gene sets, SetRank identified five unique enriched GO terms. Standard deviation of iron concentration was associated with the same four terms identified in the LFMM-derived gene sets, all mapped to ENSDARG00000029146 (*lrp1ab*). Standard deviation of phosphate concentration and land-surface density were both associated with the cell surface receptor signalling pathway (GO:0007166), mapped to ENSDARG00000075962.

S Table 10: Gene Ontology terms significantly enriched in the LFMM- and CA-derived gene sets identified using SetRank, including the associated environmental variable, GO identifier, SetRank FDR, Ensembl gene ID, and ontology category.
Environmental variables: SST (Sea-surface temperature [°C]), FE (Iron [mmol/m³]), PO_4 (Phosphate [mmol/m³]), LAND (Land-surface density [%]), CROP (Cropland density [%]), URBA (Built-up area density [%]).

Environmental Variable	Gene Ontology ID	Gene Ontology Term	SetRank FDR	ENSEMBL Gene ID	Ontology
Latent Factor Mixed Model					
FE_sd	GO:0006898	receptor-mediated endocytosis	0.000265	ENSDARG00000029146	BP
FE_sd	GO:1901342	regulation of vasculature development	0.000265	ENSDARG00000029146	BP
URBA_sd	GO:0035351	heme transmembrane transport	0.000265	ENSDARG00000061233	BP
FE_sd	GO:0005905	clathrin-coated pit	0.007409	ENSDARG00000029146	CC
FE_sd	GO:0043226	organelle	0.00688	ENSDARG00000029146	CC
URBA_sd	GO:0005796	Golgi lumen	0.000794	ENSDARG00000061233	CC
URBA_sd	GO:0008559	ABC-type xenobiotic transporter activity	0.001323	ENSDARG00000061233	MF
Correlation Analysis					
FE_sd	GO:0006898	receptor-mediated endocytosis	0.000265	ENSDARG00000029146	BP
FE_sd	GO:1901342	regulation of vasculature development	0.000265	ENSDARG00000029146	BP
PO4_sd	GO:0007166	cell surface receptor signaling pathway	0.00688	ENSDARG00000075962	BP
LAND	GO:0007166	cell surface receptor signaling pathway	0.00688	ENSDARG00000075962	BP
FE_sd	GO:0005905	clathrin-coated pit	0.007409	ENSDARG00000029146	CC
FE_sd	GO:0043226	organelle	0.00688	ENSDARG00000029146	CC

1.4 List of Figures

S Figure 1: QQ Plots for each environmental variable showing the distribution of expected vs. observed p-values for A. bicinctus (LFMM method). 4
S Figure 2: Left: Map of Sample Sites for A. bicinctus coloured by Regional Cluster [5km], Right: Heatmap of pairwise F_{st} for A. bicinctus using the Regional Clusters. 4
S Figure 3: This minimal cross-entropy plot indicates the number of ancestral populations that best suit A. bicinctus Samples. 5
S Figure 4: Left: Map of Sample Sites for A. bicinctus coloured by Regional Cluster [5km] inside the Red Sea, Right: Heatmap of pairwise F_{st} for A. bicinctus using the Regional Clusters inside the Red Sea as Population Input. 5
S Figure 5: PCoA of the genotype matrix for A. bicinctus, plotted against selected MultiQC metrics coloured by sampling region. Pearson's R values indicate the correlation between PC1 or PC2 and each respective MultiQC metric. 6
S Figure 6: Left: Map of Sample Sites for A. bicinctus coloured by Regional Cluster [5km] outside of the Red Sea, Right: Heatmap of pairwise F_{st} for A. bicinctus using the Regional Clusters outside of the Red Sea as Population Input. 6
S Figure 7: QQ Plots for each environmental variable showing the distribution of expected vs. observed p-values for A. bicinctus (LFMM method), Left: LFMM run with K = 4, Right: LFMM run with K = 6. 9
S Figure 8: PCoA of the genotype matrix for E. striatus, plotted against selected MultiQC metrics coloured by sampling region. Pearson's R values indicate the correlation between PC1 or PC2 and each respective MultiQC metric. No clear separation by sampling region was observed along any axis. Correlations with QC metrics were weak to moderate, reaching R = -0.61 for mapped reads and read length, suggesting some residual technical influence but no clear evidence of population structure. 11
S Figure 9: PCoA of the genotype matrix for P. volitans, plotted against selected MultiQC metrics coloured by sampling region. Pearson's R values indicate the correlation between PC1 or PC2 and each respective MultiQC metric. Sampling sites overlapped strongly across all axes, with no clear regional clustering, consistent with the very low F_{st} observed for this species (S Figure 12). 12
S Figure 10: PCoA of the genotype matrix for S. tubifer, plotted against selected MultiQC metrics coloured by sampling region. Pearson's R values indicate the correlation between PC1 or PC2 and each respective MultiQC metric. The variance explained curve was the flattest of the four species, with PC1 explaining only approximately 0.5% of the variance, and no clear clustering by region, indicating little detectable population structure in the largest dataset. 13
S Figure 11: Left: Map of Sample Sites for E. striatus coloured by Regional Cluster [5km], Right: Heatmap of pairwise F_{st} for E. striatus using the Regional Clusters. F_{st} values were uniformly low (approximately 0-0.0015), with no clear geographic pattern, indicating little to no genetic differentiation. 14
S Figure 12: Left: Map of Sample Sites for P. volitans coloured by Regional Cluster [5km], Right: Heatmap of pairwise F_{st} for P. volitans using the Regional Clusters. Cluster 7 (Florida) showed higher F_{st} values relative to several other clusters, indicating localised differentiation within an otherwise weakly structured and widely distributed invasive population. 14
S Figure 13: Left: Map of Sample Sites for S. tubifer coloured by Regional Cluster [5km], Right: Heatmap of pairwise F_{st} for S. tubifer using the Regional Clusters. Differentiation was minimal and spatially inconsistent. 14
S Figure 14: This minimal cross-entropy plot indicates the number of ancestral populations that best suit E. striatus samples. The cross-entropy increased monotonically from K = 1, supporting K = 1 as the optimal value and indicating no clear population substructure. 15
S Figure 15: This minimal cross-entropy plot indicates the number of ancestral populations that best suit P. volitans samples. A shallow decrease from K = 1 to K = 2 followed by an increase supports K = 2, although the small difference suggests weak population structure. 16
S Figure 16: This minimal cross-entropy plot indicates the number of ancestral populations that best suit S. tubifer samples. The cross-entropy increased mono-tonically from K = 1, supporting K = 1 as the optimal value and indicating no clear population substructure. 16
S Figure 17: LD decay figure for A. bicinctus for chromosome sequence: NC_072778.1. 17
S Figure 18: LD decay figure for E. striatus for chromosome sequence: CM069296.1. 17
S Figure 19: LD decay figure for P. volitans for contig sequence: CAMQGK010000052.1. 18
S Figure 20: LD decay figure for S. tubifer for chromosome sequence: CM035794.1. 18
S Figure 21: QQ Plots for each environmental variable showing the distribution of expected vs. observed p-values for E.striatus (LFMM method). The values mainly follow the expected diagonal across nearly all variables, indicating good calibration. 20
S Figure 22: QQ Plots for each environmental variable showing the distribution of expected vs. observed p-values for P. volitans (LFMM method). The values mainly follow the expected diagonal across nearly all variables, indicating good calibration. 20
S Figure 23: QQ Plots for each environmental variable showing the distribution of expected vs. observed p-values for S. tubifer (LFMM method). The values mainly follow the expected diagonal across nearly all variables, indicating good calibration. 21
S Figure 24: QQ Plots for each environmental variable showing the distribution of expected vs. observed p-values for A. bicinctus (CA method). Observed -log10(p) values reached 10-15, compared with an expected range of approximately 3-4, indicating strong p-value inflation. This is consistent with CA not accounting for population structure. 22
S Figure 25: QQ Plots for each environmental variable showing the distribution of expected vs. observed p-values for E. striatus (CA method). The values mainly follow the expected diagonal across nearly all variables, indicating good calibration. 23

S Figure 26: QQ Plots for each environmental variable showing the distribution of expected vs. observed p-values for P. volitans (CA method). The values mainly follow the expected diagonal across nearly all variables, indicating good calibration. 24
S Figure 27: QQ Plots for each environmental variable showing the distribution of expected vs. observed p-values for S. tubifer (CA method). The values mainly follow the expected diagonal across nearly all variables, indicating good calibration. Exception are SPM_sd and VBD_sd. 25
S Figure 28: Gene Ontology (GO) terms for biological process, cellular component, and molecular function associated with PicMin genes with pooled q-values < 0.25 from the LFMM (on the top) and CA (on the bottom) GEA methods. Avg = average and sd = standard deviation. 26

1.5 List of Tables

S Table 1: Command-line tools used in the analysis, including software versions and corresponding citations. 1
S Table 2: Overview of R packages used for data processing and analysis, including package versions and corresponding citations. 1
S Table 3: Sequencing runs excluded from the final sample set for each species and the corresponding reasons for exclusion. 2
S Table 4: Linear models assessing associations between environmental variables and genetic variation (PC1 and PC2) in A. bicinctus. 7
S Table 5: LD decay table for A. bicinctus for chromosome sequence: NC_072778.1. 17
S Table 6: LD decay table for E. striatus for chromosome sequence: CM069296.1. 18
S Table 7: LD decay table for P. volitans for contig sequence: CAMQGK010000052.1. 18
S Table 8: LD decay table for S. tubifer for chromosome sequence: CM035794.1. 19
S Table 9: Median and interquartile range of SNP-environment associations per environmental variable for each GEA method and species. 25
S Table 10: Gene Ontology terms significantly enriched in the LFMM- and CA-derived gene sets identified using SetRank, including the associated environmental variable, GO identifier, SetRank FDR, Ensembl gene ID, and ontology category. . 27

Personal Declaration

I hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the thesis.

Buchs ZH, Switzerland, 28.08.2026



Naroa Schweizer