Plan Overview

A Data Management Plan created using DMPonline

Title: Genomic Diversity of Migratory Fish

Creator: Stan Coppis

Principal Investigator: Stan Coppis

Affiliation: Wageningen University and Research (Netherlands)

Template: Data Management Plan | Wageningen University and Research

ORCID iD: 0009-0006-3110-8550

Project abstract:

Migratory fish species such as salmon, sturgeon, lamprey and shad depend on well-connected aquatic systems to reach feeding, spawning and nursery habitats throughout their life cycle. However, river connectivity is increasingly affected by climate change and human activities. Seasonal drying, altered hydrology and the construction of dams, sluices and weirs can fragment river systems and restrict natural migration routes. Such changes may influence not only the movement and abundance of migratory fish, but also their ability to complete their life cycle, as well as the genetic diversity, connectivity and long-term resilience of their populations. This project investigates how fragmentation and changing environmental conditions are reflected in the genomics of migratory fish populations. Whole-genome sequencing will be used to study several migratory fish species across Northwestern European river systems. Tissue samples will be collected from individuals at multiple locations within hydrological networks, including sites upstream and downstream of migration barriers. Genomic analyses will be used to assess patterns of genetic diversity, population connectivity, relatedness, inbreeding and demographic history. In addition to these processes, the project will investigate the genomic architecture underlying ecological and life-history differences within and between closely related populations, forms or ecotypes. For many migratory fish species, detailed genomic information remains limited and is rarely incorporated into species management. By combining genomic data with ecological and spatial information, this project aims to improve our understanding of migratory fish genomics and of how modified and changing river systems affect their populations. Ultimately, the project will explore how genomic information can complement existing ecological monitoring and be applied to the conservation and management of migratory fish in changing systems.

ID: 211165

Start date: 01-04-2026

End date: 30-03-2030

Last modified: 26-11-2026

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Genomic Diversity of Migratory Fish

A. Describe the research project

1. Name researcher (please, add your full name):

S.C.H.V. (Stan) Coppis

2. What is the name of your department(s)?

- Animal Sciences

3. What is the name of your chair group(s) or business unit(s)? English name and abbreviation for chair groups from this page; business units from this page (expand to Wageningen Research and keep expanding to find your specific division / group). Examples: Bioprocess Engineering (BPE) or Contract Research Organization (CRO).

Marine Animal Ecology (MAE)

4. Describe the organisational context of your research project.

DMP version (or date last modified)	29-september-2026
Supervisor / (co-)promotors	Dr. R. (Reindert) Nijland, Dr. Ir. HJ (Hendrik-Jan) Megens
Graduate School (WU only)	WIAS
Start date of project	2026-april-01
End date of project	2030-march-31
Project number	n.a.
Funding body	Marine Animal Ecology department

5. Give a short description of your research project.

Title	Genomic Diversity of Migratory Fish
Summary	Migratory fish species such as salmon, sturgeon, lamprey and shad depend on well-connected aquatic systems to reach feeding, spawning and nursery habitats throughout their life cycle. However, river connectivity is increasingly affected by climate change and human activities. Seasonal drying, altered hydrology and the construction of dams, sluices and weirs can fragment river systems and restrict natural migration routes. Such changes may influence not only the movement and abundance of migratory fish, but also their ability to complete their life cycle, as well as the genetic diversity, connectivity and long-term resilience of their populations. This project investigates how fragmentation and changing environmental conditions are reflected in the genomics of (migratory) fish populations. Whole-genome sequencing will be used to study several fish species across Northwestern European river systems. Tissue samples will be collected from individuals at multiple locations within hydrological networks, including sites upstream and downstream of migration barriers. Genomic analyses will be used to assess patterns of genetic diversity, population connectivity, relatedness, inbreeding and demographic history. In addition to these processes, the project will investigate the genomic architecture underlying ecological and life-history differences within and between closely related populations, forms or ecotypes. For many fish species, detailed genomic information remains limited and is rarely incorporated into species management. By combining genomic data with ecological and spatial information, this project aims to improve our understanding of migratory fish genomics and of how modified and changing river systems affect their populations. Ultimately, the project will explore how genomic information can complement existing ecological monitoring and be applied to the conservation and management of migratory fish in changing systems.

6. List the individuals responsible for the following data management tasks.

Data collection	Stan Coppis, Hendrik-Jan Megens, Reindert Nijland
Data quality	Stan Coppis, Hendrik-Jan Megens, Reindert Nijland
Storage and backup	Stan Coppis, Hendrik-Jan Megens, Reindert Nijland
Data archiving / publishing	Stan Coppis, Hendrik-Jan Megens, Reindert Nijland
Data stewardship / support	Diede Maas
Any other role [.....]	n.a.

7. I have requested a review of this data management plan from:

- WUR Library - Data Management Support (data@wur.nl).

8. Name of the data management support staff and / or data steward consulted during the preparation of this plan and date of consultation.

Dr. Sydney Jordan
WUR Library – Research Data Management Support
data@wur.nl
Date: 2026-09-29

B. Describe the data to be collected, software used, file formats and data size.

9. Will you use existing data for this project?

- No. Please describe below any constraints to reusing existing data.

The genomic data used in this project will be generated within the project itself. Only existing publicly available reference genome assemblies and associated annotations will be used for mapping and downstream genomic analyses.

10. Will new data be produced?

- Yes.

11. Please describe the data you expect to generate and / or use in the table below. Include reused existing data as well (as these are files that you manage and store).

File contents	Data type	Software	(Open) file format	Estimated size of each file (range)	Estimated number of files (range)
(e.g. lab analysis, gene sequence, interviews, lesion scores, etc.)	(e.g. numerical)	(e.g. Excel)	(e.g. .csv)	(e.g. 20-50 Mb)	(e.g. 50-100)
Species sample administration	Text and numerical	Excel	.xlsx	< 1 Mb	< 75
Raw sequencing reads	Genomic sequence data	Illumina software	.fastq.gz	~ 5 - 50 Gb	~ 200 - 1500
Aligned sequencing data	Genomic sequence data	e.g., SAMtools	.bam	~ 5 - 50 Gb	~ 100 - 500
Called genetic variants	Genomic and numerical	e.g., BCFtools	.vcf.gz	< 50 Gb	~ 10 - 100
Filtered/ processed genomic dataset	Genomic and numerical	e.g., PLINK, BCF/VCFtools	.vcf.gz .bed/.bim/.fam	~0.1 - 20 Gb	~ 50 - 150
Reference genomes and genomic annotations (*existing*)	Genomic and numerical	Genome browser / NCBI	.fasta / .gtf	< 5 Gb	< 50
Bioinformatics and statistical analysis scripts	Text and code	Python, R, Bash, SLURM	.py, .R, .sh, .txt	< 0.5 Gb	~ 100
Analysis results and output	Numerical and text	Python, R. and genomic software	.csv, .tsv, .txt	< 1 Gb	~ < 500
Figures and visualization	Images	Python output, Adobe, Photoshop	.png, .pdf, jpeg	< 0.1 Gb	~ < 500
Analysis and log files	Text	SLURM, bioinformatics software	.out, .err, .log, .txt	<100 Mb	~ 500 - 1000
Laboratory / sample quality check images	Image	Imaging software Biorad	.scn, .png	< 0.1 Gb	< 500
Lab journal	Text	eLabjournal	.txt, .pdf	< 2 Mb	~ 50 - 100

12. Estimate how much data storage you require in total (e.g. by using the information in the table at question 11).

- >1000 GB

Genomic (unprocessed) data is relatively large (~30 Gb) per individual at times. After sequencing, this data has to be stored and processed, whereafter it will be moved to online repositories. When sequencing multiple individuals a large amount of storage space will therefore be necessary.

C. Storage of data and data documentation / metadata during research

13. Where will the data,code and accompanying documentation / metadata be stored and backed up during the research project (see the WUR Data Storage Finder)? Include platforms you use to share data, collect data on, or send data to for processing or analysis.

- W:drive Enterprise File Storage (WUR network drive).
- W:drive Massive File Storage Disaster Recovery (WUR network drive).
- Other, please specify below the storage medium / system and describe back-up frequency, access management, and geographic location (e.g. within or outside the EU).

Sample administration files (mostly Excel, Word or PDF) will be kept in a Teams environment dedicated to the project, in MAE's /w drive, and on the High Performance Cluster.

Code, raw, and processed sequence data will be kept on the High Performance Cluster and be published in the European Nucleotide Archive or NCBI but kept private until the project, publication or report is finished.

D. Structuring your data and information

14. Give a (visual) representation of the folder structure you intend to use.

Folder structure, based on https://datascience.codata.org/articles/10.5334/dsj-2025-035:

01_Sample_ administration

```
-README.txt
  --Genetic_fish_atlas
  --Species_name
      -YYYYMMDD_Species_name_SamAdmin_SC.xlsx
  --Gel_photos
      -YYYYMMDD_Species_name_sampleIDs_SC.png
```

02_Data_analysis_admin (on HPC)

```
--Genetic_fish_atlas
--Species_name
      -README.txt
      -Species_name_sample_metadata.txt or .csv
      --1_RAW_data
         -YYYYMMDD_Species_name_RAW_SampleIDs.fastq.gz (read_only)
      --2_Processed_data
         -YYYYMMDD_Species_name_SNP_called_samplenumbers_SC.vcf.gz
      --3_Analyses
         --01_Analysis_name
               -YYYYMMDD_Species_name_analysis_name_script_SC.py
            --01_Analysis_name_Figures
            --02_Analysis_name_Logs
            --03_Analysis_name_Temp
         --02_Analysis_name
```

03_PhD_documents_and_outputs

```
--PhD_SC
    --01_documents
      --01_Personal
      --02_Administrative
         --DMP
         --Grants
         --TSP
      --03_Courses
         --YYYYMMDD_Title_name
   --02_Proposal
   --03_Presentations
      --01_Conferences
         --YYYYMMDD_Title_Conference
            -01_YYYYMMDD_title_conference_presentation_SC.pptx
            -02_YYYYMMDD_title_conference_Notes_and_text_SC.docx
            --YYYYMMDD_title_conference_administration
      --02_Miscellaneous_presentations
         --YYYYMMDD_Title_Presentation_location
            -01_YYYYMMDD_title_conference_presentation_SC.pptx
            -02_YYYYMMDD_title_conference_text_SC.docx
   --04_Publications
      --ManuscriptX_01_short_name
         --01_Drafts
         --02_Figures
            --FigureX
               -Script.txt
               -FigureX.png
         --03_Submitted
         --04_Revision
         --05_Published
   --05_Thesis
      --01_Figures
      --02_Articles
      --03_Drafts
      --04_Submitted
      --05_Revised
      --06_Defense
```

15. Describe the file naming conventions you intend to use. Please give one or multiple example(s).

Files start with a date (YYYYMMDD) and end with the writers' initials. Within the file name the associated species name is written, optionally together with the range of samplsIDs (if possible). i.e. [date]_[Species_name]_[optional: SampleIDs]_[Subject]_[version]_[Initials]_[Extension]. Subjects can, for example, be sample administration, data analysis, sequencing quote, etc.

Each separate analysis will be categorized in folders by numbers starting from 01. Each analysis folder will contain the script, and subsequent folders leading to figures, temporary files, and logs.

Versions will be indicated by v0X numbering.

16. How will you distinguish between versions of files (multiple answers possible)?

- Dates within file names are updated when files are modified.
- The designation 'vRAW' is added to file names that contain raw unaltered data (before any processing and cleaning). Any alteration of RAW data is done on a copy of the RAW data and appended with a version number which increases with each file modification (e.g. v01, v02, v03 etc.).

For files requiring low amounts of storage space, such as administration, a copy will be made of the file and kept in a designated "_old" folder. Hereafter, the date in the file name will be updated and used as the new working file.

E. Data documentation and data quality

17. Describe below what data documentation and metadata will accompany the data to help make the data findable, understandable, and reproducible.

- Electronic lab notebooks (e.g. eLabJournal).
- Elaborate documentation and notes within scripts / codes.
- The WUR readme file template (see template at https://doi.org/10.5281/zenodo.7701727).

18. Describe what data and analysis quality controls will be used?

- We will perform preliminary (pilot) experiments to validate intended experimental methods.
- We will use standard and validated protocols where appropriate.
- Statistical model assumptions are adhered to and assessed (e.g. (residual) distribution analysis, outlier analysis, (accounting for) independence, homogeneity of variance, etc.).
- We will use a statistical power analysis before and after the experiment.
- Supervisors or peers will review the data and results for any anomalies (e.g. unexpected inconsistencies, outliers, correct labeling of data and / or treatments, correct and consistent coding applied, etc.).
- We will use standardised coding and terms of data throughout all experiments so that data descriptions are equal throughout various datasets created.

Data quality controls related to sample size and preliminary experiments will follow the requirements of the applicable animal experimentation protocols and agreements with the Animal Welfare Body (IvD). Where animal experimentation is involved, sample sizes will be justified in advance, including a power analysis, with the aim of using the minimum number of animals required to obtain scientifically meaningful (genetic) results.

Data, sample and file names/codes, and other metadata will be named consistently across datasets and analyses. Bioinformatic analyses will, therefore, be performed using documented and reproducible scripts, with software versions and parameters recorded.

F. Working with sensitive data (personal data, ethics), data ownership, sharing and access

19. Who is the (rights)holder of the data (commonly known as the owner of the data)?

- WUR is the (rights)holder of the data.

20. What is the data classification for your project (for example as specified in SmartPIA) taking into account the (privacy) sensitivity of the data?

- Negligible.

21. Is this project registered in SmartPIA?

- No. Please register in SmartPIA in the case (privacy) sensitive data is collected (when applicable: via your supervisor, the project manager, see guidance).

22. Please specify the (sensitive) data and privacy protection measures. Note that any measures undertaken should be consulted with the Information Security Officer (ISO) and Privacy Officer (PO).

- Data is classified as negligible and standard WUR security measures are undertaken.
- Personal or other sensitive data will be removed when not required for verification of research.

23. Are there other ethical issues that need to be taken into account which may include approval from ethical committees?

- I work with genetic resources covered by the Convention on Biological Diversity and need to comply with the Nagoya Protocol.
- I work with animals and will seek / have approval of the ethics committees involved (Animal Welfare Body (IvD), Animal Tests Committee (DEC), Central Animal Testing Committee (CCD)).

24. Will there be any intellectual property (IP) rights or alternative applications or routes to impact (such as commercial interests) associated with the data?

- No.

G. Data archiving and publishing

25. Are there reasons to restrict access to the data or limit which data will be made publicly available?

- No.

26. Describe what data from question 11 will be archived internally (e.g. WUR network drive / Yoda@WUR) and not published, for a minimum of 10 years? Include the exact name for the storage medium chosen (see the WUR Data Storage Finder).

- Not applicable as data will be published.
- Other, please specify below.

Raw sequencing data that is not published in a peer-reviewed article but used in reports will be made publicly available within 1 year after the report's finalization.

27. What data will be published and made available for reuse via a data repository?

- Data underlying publications or reports. Please specify below which data listed in question 11.

Raw sequencing data and corresponding metadata will be published and made available for reuse.

28. When will the data be available for reuse, and for how long will the data be available?

- Data will be available for at least 10 years as soon as the article or report is published and not required for any other article publication.
- Published data becomes accessible for at least 10 years after expiring of the embargo. Please specify below the reason for an embargo.
- Publication of data not underlying an article or report will be considered at the end of the project.
- Data will be available for at least 10 years upon completion of the project.

29. Which data repository do you intend to use to make the data findable and accessible (see the WUR Repository Finder)?

- Other, please specify below.

NCBI/genbank and European Nucleotide Archive (ENA) for genomic data. GitHub for code and software. Waarnemingen.nl/Observation.org for occurrence data.

30. Which metadata standard will be used to describe the data during internal archiving and / or depositing in a data repository?

- Other, please specify below.

Simple Darwin Core (flat) i.e., notation of species identification, occurrence, geographical location, handling, sampling and DNA yield thereof.

31. Which licence/terms of use will be applied to the data?

- Open access (Creative Commons Attribution licence (CC BY); anyone can access and reuse with attribution).

H. Data management costs

33. What resources (in time and / or money) will be dedicated to data management, data archiving or publication, and ensuring that data is reusable? Indicate as well how these costs will be covered.

- The PhD candidate and supervisor will spend at least 10% of their time on research data management to approach the FAIR principles as much as possible.
- All costs for 10 year data storage and access management to that data after journal publication or report are covered by the research group / project.