# NBIS Support project

Admixture analysis

Verena Kutschera

2025-09-08

# Background

Admixture analysis of 57 individuals sampled in 1980, as part of a larger monitoring project.

Table 1: Clusters and sample sizes after removal of low-quality samples, total number of sequenced individuals shown in brackets.

| Cluster | number of samples |
|---|---|
| Island | 7 |
| Southern | 29 (30) |
| Northern | 15 |
| Hybridzone | 5 |

# Code availability

The ADMIXTURE Snakemake pipeline and all files to produce this Quarto report are available on https://github.com/verku/admixture.

# Methods

## Data processing and analysis

### Reference genome assembly preparations

A Hi-C scaffolded de-novo assembly of the studied species was used as reference genome for this analysis. Sex chromosome-linked scaffolds were excluded from all analyses.

Repetitive regions were de novo identified and masked using RepeatModeler (Smit and Hubley 2008) and RepeatMasker (Smit, Hubley, and Green 2013) and CpG sites were identified in each of the reference genomes using GenErode v0.6.0 (Kutschera et al. 2022).

### Whole-genome sequencing data processing

Whole-genome sequencing data from 57 samples were processed and analysed using GenErode (Kutschera et al. 2022) (https://github.com/NBISweden/GenErode) with default settings if not otherwise described. GenErode version 0.6.0 was used for mapping and BAM file processing, and version 0.7.0 (available on git branch 'dev' at the time of writing this report) for data processing and analysis of processed BAM files. Further downstream analyses were performed in a newly developed Snakemake pipeline for admixture analyses and in this Quarto document.

Briefly, sequencing reads were adapter- and quality-trimmed with fastp v0.22.0 (Chen et al. 2018) and then mapped to each of the reference genomes with BWA mem v0.7.17 (Li and Durbin 2009). PCR duplicates were identified and marked using Picard MarkDuplicates v2.26.6 (https://gatk.broadinstitute.org/hc/en-us/articles/360037225972-MarkDuplicates-Picard-) and reads around indels were realigned with GATK IndelRealigner v3.7 (McKenna et al. 2010). The genome-wide mean depth and minimum and maximum depth thresholds for downstream analyses were calculated in GenErode with default settings.

Basic FastQ and bam file statistics were obtained with FastQC v0.12.1 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), samtools v1.9 (Li et al. 2009), QualiMap v2.2.2 (Okonechnikov, Conesa, and García-Alcalde 2015) and summarized using MultiQC v1.9 (Ewels et al. 2016).

Variants were called in each sample using bcftools v1.20 (Danecek et al. 2021) mpileup and call. CpG sites as identified from the reference genomes were removed and default GenErode filters were applied to the VCF files. Finally, VCF files from all samples were merged and sites that are not biallelic as well as sites with any missing genotypes were removed along with sex-linked scaffolds.

### Population structure

A Snakemake pipeline was written to run ADMIXTURE v1.3.0 (Alexander, Novembre, and Lange 2009) to estimate individual-based ancestry and identify genetic clusters (https://github.com/verku/admixture/). This maximum likelihood approach uses a cross-validation procedure to determine the best number of possible genetic groups present in the dataset, under the assumption that individuals are unrelated. ADMIXTURE was run for $K$=1-4.

## Results

### Data processing

I mapped 57 re-sequenced genomes to the reference genome, resulting in mean genome-wide depths of coverage ranging from 5X to 30X (mean across all samples: 17X). One sample was excluded from further analysis due to low sequencing depth. After merging the remaining 56 VCF files, SNPs were filtered to keep only biallelic SNPs, and to exclude sex chromosome-linked scaffolds and sites with missing data, resulting in a final set of 483,494 SNPs.

### Population structure

Admixture was run for $K$=1-4. The highest support was obtained for $K$=2 (Figure 1). For $K$=2, a Northern cluster is distinguishable from a Southern cluster and the island samples, and individuals from the hybrid zone appear admixed (Figure 2). For $K$=3, the Northern cluster appears substructured. The

hybrid zone cluster becomes fully apparent and some Southern cluster samples appear admixed for $K$=4 (Figure 2).

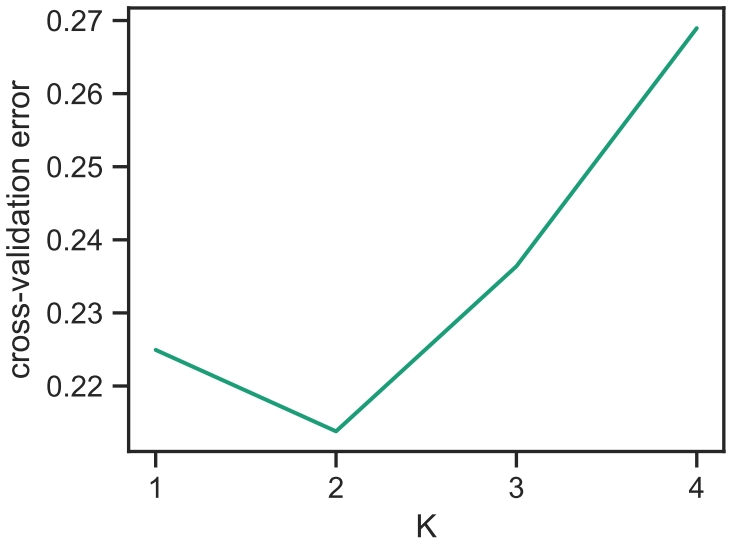
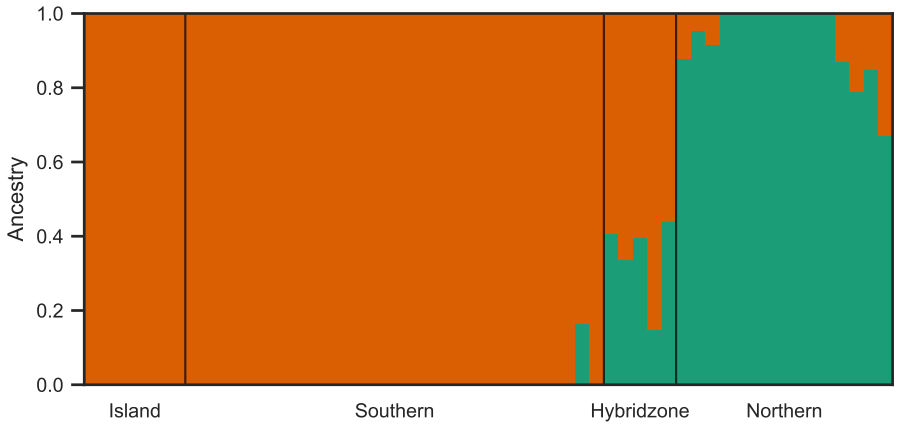
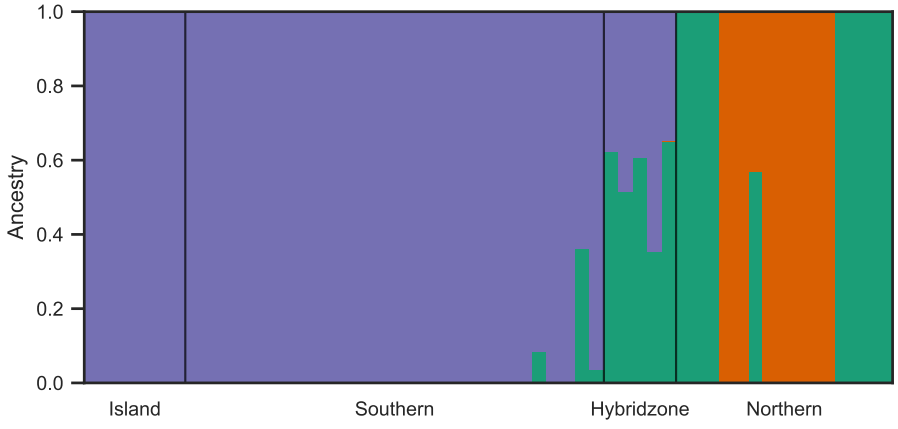


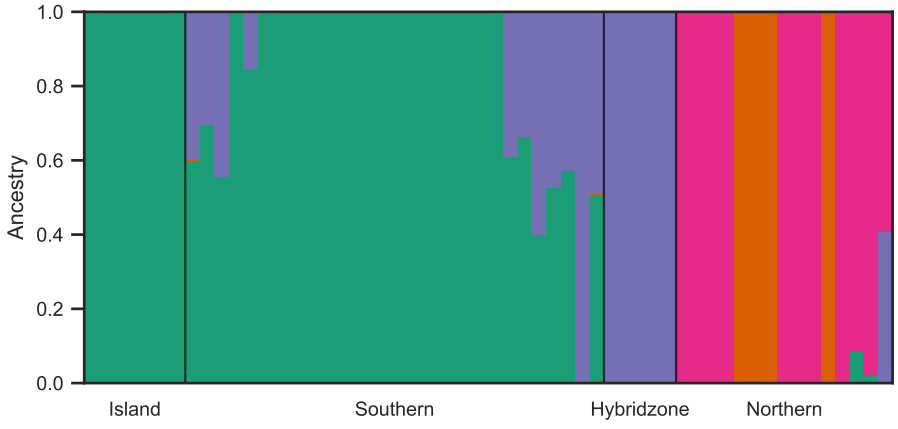

Figure 1: Cross-validation plot to determine the best-fitting number of K from an ADMIXTURE analysis of 56 genomes.




(a) K=2




(b) K=3




(c) K=4

Figure 2: Admixture plot for 56 genomes sampled from four clusters.

# References


Alexander, David H., John Novembre, and Kenneth Lange. 2009. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." *Genome Research* 19: 1655–64. https://doi.org/10.1101/GR.094052.109.

Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. "Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor." *Bioinformatics* 34: i884–90. https://doi.org/10.1093/bioinformatics/bty560.

Danecek, Petr, James K. Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O. Pollard, Andrew Whitwham, Thomas Keane, Shane A. McCarthy, and Robert M. Davies. 2021. "Twelve Years of SAMtools and BCFtools." *GigaScience* 10. https://doi.org/10.1093/gigascience/giab008.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. "MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report." *Bioinformatics* 32: 3047–48. https://doi.org/10.1093/bioinformatics/btw354.

Kutschera, Verena E., Marcin Kierczak, Tom van der Valk, Johanna von Seth, Nicolas Dussex, Edana Lord, Marianne Dehasque, et al. 2022. "GenErode: A Bioinformatics Pipeline to Investigate Genome Erosion in Endangered and Extinct Species." *BMC Bioinformatics* 23: 1–17. https://doi.org/10.1186/S12859-022-04757-0.

Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." *Bioinformatics* 25: 1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25: 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing Next-Generation DNA Sequencing Data." *Genome Res* 20: 1297–1303. https://doi.org/10.1101/gr.107524.110.

Okonechnikov, Konstantin, Ana Conesa, and Fernando García-Alcalde. 2015. "Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data." *Bioinformatics* 32 (2): 292–94. https://doi.org/10.1093/bioinformatics/btv566.

Smit, Arian F. A., and Robert Hubley. 2008. "RepeatModeler Open-1.0." http://www.repeatmasker.org.

Smit, Arian F. A., Robert Hubley, and Phil Green. 2013. "RepeatMasker Open-4.0." http://www.repeatmasker.org.