

Report vcf file Analysis

Introduction:

The following analysis was conducted to examine differentiation between a set of populations and genomic regions under selection. The data was provided in form of an VCF file along with the utilized reference genome (house sparrow) and its gff file. The VCF file includes 16 individuals from 4 populations (8N, K, Lesina, and Naxos2), Naxos2 being an outgroup. It contains data for 2 chromosomes (chr5 and chrZ) with a total of 3,816,977 variants. The analysis included filtering of the provided data and applying a range of tools to calculate FST, Likelihood Ratio, and Tajima's D.

Methods:

Data Filtering

To identify and filter out low-quality data I analysed the data using bcftools (Li, 2011), vcftools (Danecek *et al.*, 2011), and R and decided to apply the following filters:

- Minimum quality score of 30 to filter low-quality calls
- Depth min of 7 and max of 18 to avoid low coverage calls and overly high coverage
- Min call rate of 90%, to exclude all variants with >10% missing data
- Minor allele frequency (MAF) of 0.1 to exclude rare variants
- Removal of indels and multiallelic sites to focus on biallelic SNPs

Using vcftools, I analysed how much data was left for each individual after the filtering and excluded K006 (>10% missing data). I additionally excluded Naxos2 as the outgroup is not required for any of the analyses I conducted. After these filtering steps, the dataset was reduced to 166,556 variants.

FST analysis

FST is a measure of genetic differentiation between populations, and it was used here to understand how differentiated the three populations (8N, K, and Lesina) are from each other. I performed pairwise FST calculations with vcftools between the populations (8N/K, 8N/Lesina, K/Lesina) using a sliding window approach with a window size of 20,000 base pairs.

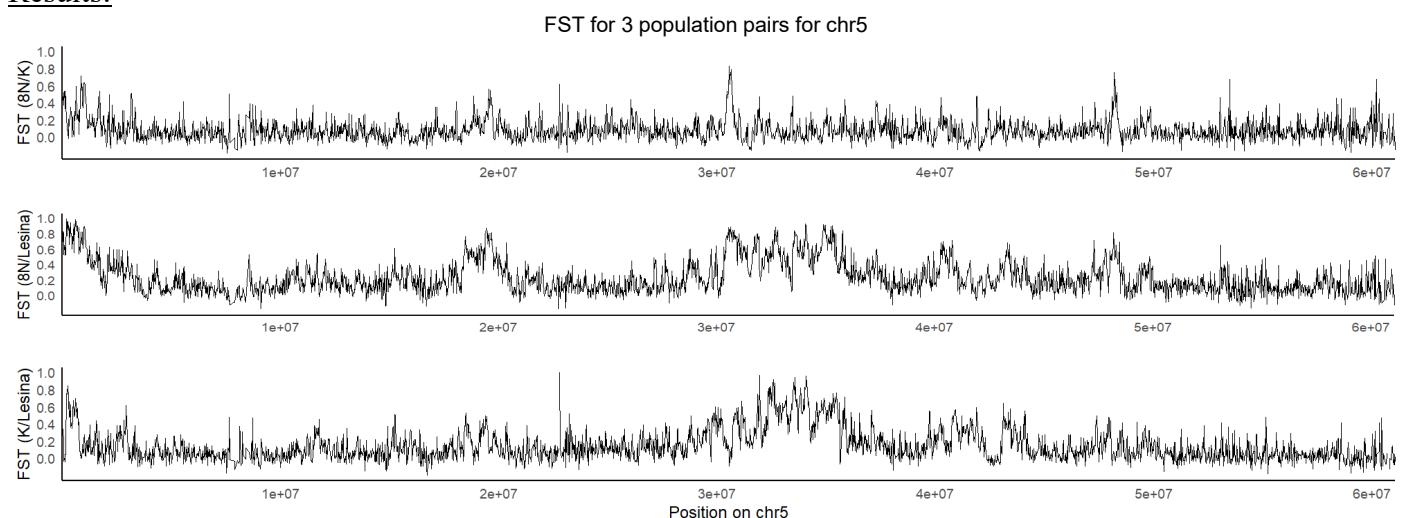
Likelihood-based detection of selection analysis

To detect regions of the genome under selection, I used SweepFinder2 (DeGiorgio *et al.*, 2016). This method estimates the probability that a region has undergone a selective sweep by analysing patterns of genetic variation. I phased the data for each chromosome (chr5 and chrZ) using Shapeit2 (Delaneau *et al.*, 2011) and then performed SweepFinder2 analysis on the phased data.

Tajimas's D analysis

Tajima's D is a statistical test used to assess whether a population is experiencing selection, genetic drift, or a recent expansion. I calculated Tajima's D using vcftools for both chromosomes with a sliding window of 20,000 base pairs, which provides insights into regions that may be under selective pressure or undergoing changes in effective population size.

Results:



FST for 3 population pairs for chrZ

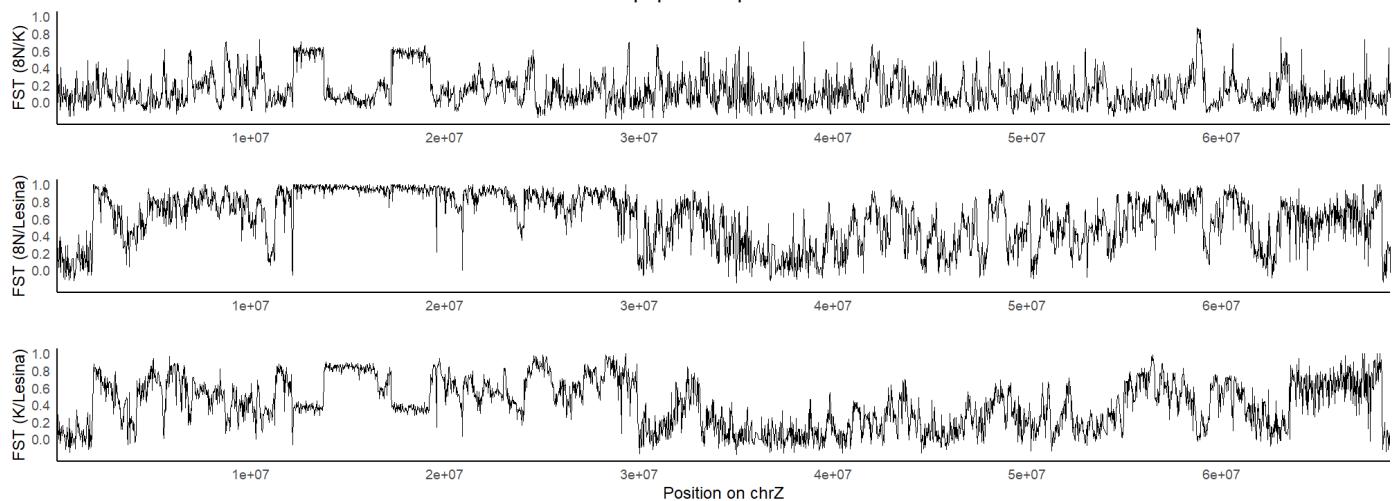
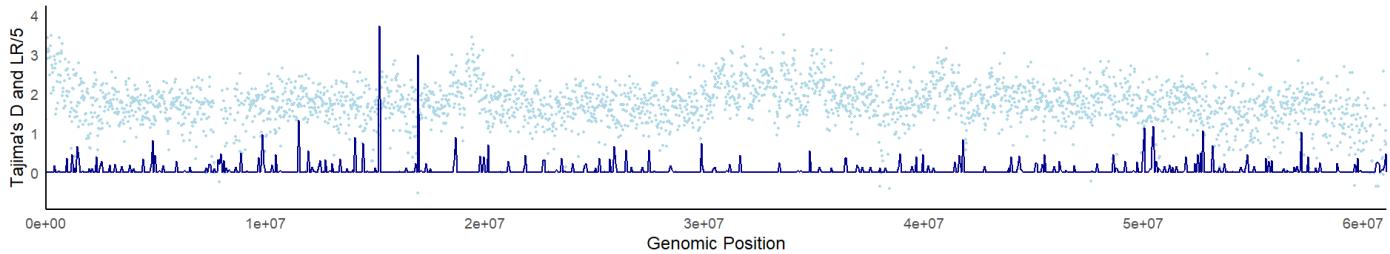


Fig1. FST values across chr5 and chrZ for population pairs (8N/K, 8N/Lesina, K/Lesina). The y-axis represents FST values, the x-axis genomic position.

The FST plots show differentiation between the three populations across both chromosomes (Fig1). In both cases, the differentiation is moderate across most regions, with some peaks indicating highly differentiated loci. ChrZ exhibits stronger peaks and elevation changes, especially in the range of 1e+07 - 2e+07.

Tajima's D and LR Across Chr5



Tajima's D and LR Across ChrZ

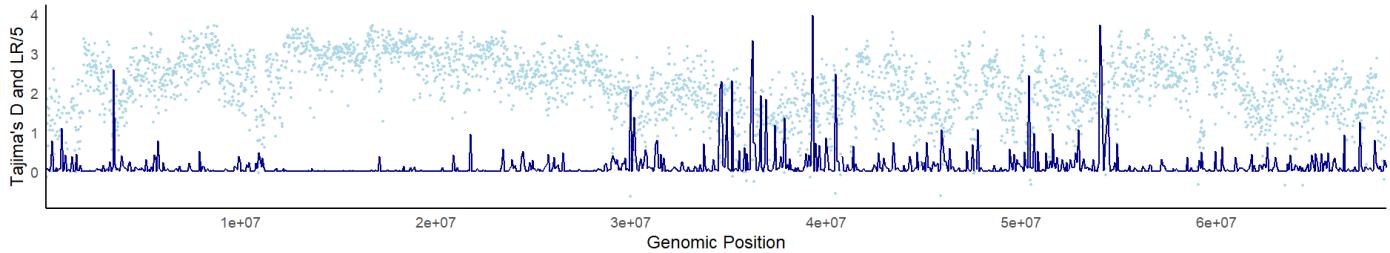


Fig2. Likelihood Ratio (LR) scores and Tajima's D values across chromosomes. The y-axis represents LR scores / 5 and Tajima's D; the x-axis represents genomic position. Peaks in LR and negative Tajima's D values suggest regions under strong selection.

The selection scans using SweepFinder2 show several sites with high likelihood ratio (LR) scores that are additionally supported by low Tajima's D values suggesting positive selection (resulting in selective sweeps of the diversity in these regions) (Fig2). Chr5 has 2 main LR peaks; chrZ exhibits overall more peaks, with the majority being in the genomic area of 3e+07 - 4e+07.

Discussion:

The results indicate population differentiation among the three groups, with varying levels across the 2 chromosomes. The higher FST values suggest loci that may be under divergent selection, restricted gene flow, or linked to local adaptation. The presence of elevated FST on chrZ could be attributed to sex-linked selection, as Z chromosomes often exhibit reduced recombination and stronger selection effects.

The selection scan highlights several regions potentially under positive selection. Genes found near by in the ref genome include IL6ST, PTGER4, and CCL21, which are involved in immune responses, potentially making them a relevant target to selection.

Danecek, P. *et al.* (2011) 'The variant call format and VCFtools', *Bioinformatics*, 27(15), pp. 2156–2158. doi:10.1093/bioinformatics/btr330.

DeGiorgio, M. *et al.* (2016) 'SweepFinder2: increased sensitivity, robustness and flexibility.', *Bioinformatics (Oxford, England)*, 32(12), pp. 1895–1897. doi:10.1093/bioinformatics/btw051.

Delaneau, O., Marchini, J. and Zagury, J.-F. (2011) 'A linear complexity phasing method for thousands of genomes.', *Nature methods*, 9(2), pp. 179–181. doi:10.1038/nmeth.1785.

Li, H. (2011) 'A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.', *Bioinformatics (Oxford, England)*, 27(21), pp. 2987–2993. doi:10.1093/bioinformatics/btr509.