

Supplemental Material

Comparative genomics of *Cryptosporidium parvum* reveals the emergence of an outbreak-associated population in Europe and its spread to the USA

Greta Bellinzona¹, Tiago Nardi¹, Michele Castelli¹, Gherard Batisti Biffignandi¹, Karim Adjou², Martha Betson³, Yannick Blanchard⁴, Ioana Bujila⁵, Rachel Chalmers⁶, Rebecca Davidson⁷, Nicoletta D'Avino⁸, Tuulia Enbom⁹, Jacinto Gomes¹⁰, Gregory Karadjian², Christian Klotz¹¹, Emma Östlund¹², Judith Plutzer¹³, Ruska Rimhanen-Finne¹⁴, Guy Robinson⁵, Anna Rosa Sannella¹⁵, Jacek Sroka¹⁶, Christen Rune Stensvold¹⁷, Karin Troell¹², Paolo Vatta¹⁵, Barbora Zalewska¹⁸, Claudio Bandi¹⁹, Davide Sassera^{20*}, Simone M. Cacciò^{15*}

1, Department of Biology and Biotechnology, University of Pavia, Italy

2, UMR BIPAR, Anses, Laboratoire de Santé Animale, INRAE, Ecole Nationale Vétérinaire d'Alfort, Maisons-Alfort, France

3, Department of Comparative Biomedical Sciences, School of Veterinary Medicine, University of Surrey, Guildford, UK

4, Viral Genetics and Biosecurity Unit (GVB), French Agency for Food, Environmental and Occupational Health Safety (ANSES), Ploufragan, France.

5, Department of Microbiology, Public Health Agency of Sweden, Solna, Sweden

6, *Cryptosporidium* Reference Unit, Swansea, UK

7, Norwegian Veterinary Institute, Ås, Norway

8, Istituto Zooprofilattico Sperimentale dell'Umbria e delle Marche, Perugia, Italy

9, Animal Health Diagnostic Unit, Finnish Food Authority, Kuopio, Finland

- 10, National Institute for Agricultural and Veterinary Research, Lisbon, Portugal
- 11, Department of Infectious Diseases, Unit for Mycotic and Parasitic Agents and Mycobacteria, Robert Koch Institute, Berlin, Germany
- 12, National Veterinary Institute, Uppsala, Sweden
- 13, National Institute for Public Education, Budapest, Hungary
- 14, Finnish Institute for Health and Welfare, Helsinki, Finland
- 15, Department of Infectious Diseases, Istituto Superiore di Sanità, Rome, Italy
- 16, Department of Parasitology and Invasive Diseases, National Veterinary Research Institute, Pulawy, Poland
- 17, Statens Serum Institut, Copenhagen, Denmark
- 18, Veterinary Research Institute, Department of Food and Feed Safety, Brno, Czech Republic
- 19, Department of Biosciences, University of Milan, Milan, Italy
- 20, IRCCS Fondazione Policlinico San Matteo, Pavia, Italy

*Correspondence should be addressed to:

Davide Sassera, davide.sassera@unipv.it

Simone M. Cacciò, simone.caccio@iss.it

Table of contents

Supplemental Tables

Supplemental_Table_S1: List of the *Cryptosporidium parvum* isolates included in this study, detailing information about the host, year of collection, country of origin and gp60 subtype. Refer to Supplemental_Table_S1.xlsx

Supplemental_Table_S2: Data filtering. Isolates marked with an X were removed from downstream analyses because of too high contamination, low read depth, presence of more than one population (mixed infection) and low size of the assembly, as indicated. Closely related isolates removed for computing Tajima's *D* and nucleotide diversity (π) in Supplemental Fig. S24 are marked with *. Refer to Supplemental_Table_S2.xlsx

Supplemental_Table_S3: Distribution of Single Nucleotide Polymorphisms (SNPs) in each of the eight *Cryptosporidium parvum* chromosomes. Refer to Supplemental_Table_S3.xlsx

Supplemental_Table_S4: Pairwise proportion test between chromosomes. Refer to Supplemental_Table_S4.xlsx

Supplemental_Table_S5: List of the genomic regions with $F_{st} > 0.9$. Annotation in each region has been retrieved from the reference genome IOWAII-ATCC. Refer to Supplemental_Table_S5.xlsx

Supplemental_Table_S6: List of the protein IDs in which at least one nucleotide position is considered to be significantly ($p\text{-value} < 0.5$) under selective pressure by the BUSTED algorithm. Protein IDs and gene annotations have been retrieved from the reference genome IOWAII-ATCC. Refer to Supplemental_Table_S6.xlsx

Supplemental_Table_S7: List of 16 proteins considered under selective pressure by both the F_{st} statistics and BUSTED. Refer to Supplemental_Table_S7.xlsx

Supplemental Figures

Supplemental_Figure_S1: Workflow for data filtering and SNP calling.....	6
Supplemental_Figure_S2: Estimating the presence of mixed infections.....	7
Supplemental_Figure_S3: SNP density along the chromosomes.....	8
Supplemental_Figure_S4: Maximum Likelihood (ML) phylogenetic tree inferred on a set of 179 genes.....	9
Supplemental_Figure_S5: Maximum Likelihood (ML) tree of <i>C. parvum</i> isolates from ruminants only.....	10
Supplemental_Figure_S6: Phylogenetic network generated using SplitTree.....	11
Supplemental_Figure_S7: Analysis on SNPs located on Chromosome 1.....	12
Supplemental_Figure_S8: SNPs pattern in a 10 kb region of Chromosome 1 spanning from position 755,934 to 768,672.....	14
Supplemental_Figure_S9: SNPs pattern in a 55 kb region of Chromosome 1 spanning from position 768,729 to 823,729.....	15
Supplemental_Figure_S10: SNPs pattern in a 50 kb region of Chromosome 1 spanning from position 824,800 to 874,170.....	16
Supplemental_Figure_S11: Analysis on SNPs located on Chromosome 2.....	17
Supplemental_Figure_S12: SNPs pattern in a 210 kb region on Chromosome 2 spanning from position 384,000 to 594,000.....	19
Supplemental_Figure_S13: Analysis on SNPs located on Chromosome 3.....	20
Supplemental_Figure_S14: Analysis on SNPs located on Chromosome 4.....	22
Supplemental_Figure_S15: SNPs pattern in the first 8 kb adjacent to the 5' telomere of Chromosome 4.....	24

Supplemental_Figure_S16: Analysis on SNPs located on Chromosome 5.....	25
Supplemental_Figure_S17: Analysis on SNPs located on Chromosome 6.....	27
Supplemental_Figure_S18: SNPs pattern in the first 18 kb adjacent to the 5' telomere of Chromosome 6.....	29
Supplemental_Figure_S19: Analysis on SNPs located on Chromosome 7.....	30
Supplemental_Figure_S20: Analysis on SNPs located on Chromosome 8.....	32
Supplemental_Figure_S21: SNPs pattern in a 30 kb region spanning position 210,000 to 240,000 on Chromosome 8.....	34
Supplemental_Figure_S22: Distribution of shared and population-specific SNPs.....	35
Supplemental_Figure_S23: Relatedness network for pairs of isolates identified as having high proportions of IBD sharing.....	36
Supplemental_Figure_S24: Distribution of Tajima's <i>D</i> values and nucleotide diversity (π) in population 2 and population 3.....	37

WORKFLOW: SEQUENCING DATA FILTERING AND SNP CALLING

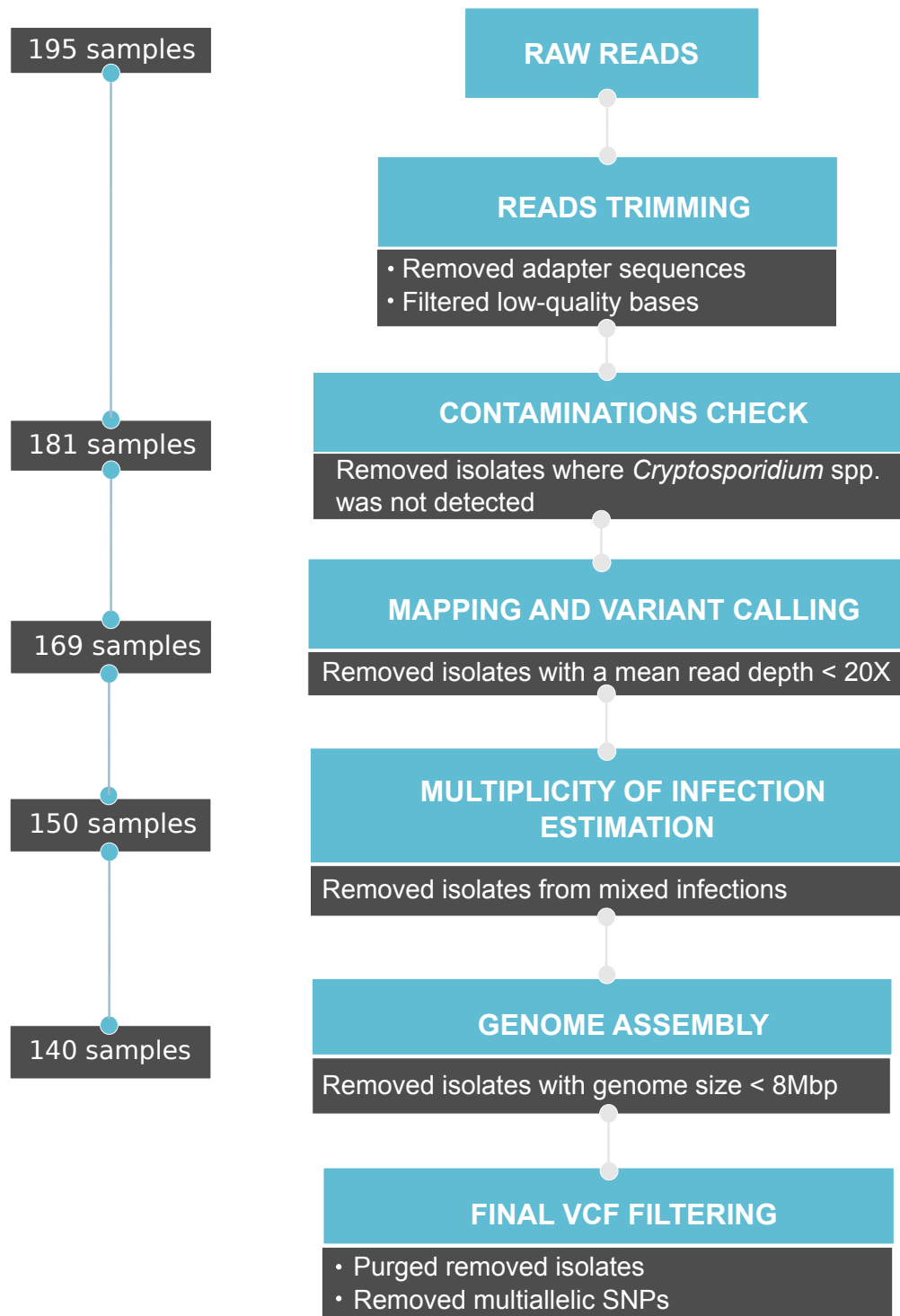


Figure S1. Workflow for data filtering and SNP calling. The figure outlines the sequential steps employed in data processing to derive the ultimate refined dataset used for the downstream analysis. The left side of the diagram enumerates the count of samples at each filtering step, underscoring the progressive refinement of the dataset (note that the reference genome is not counted). Supplementary Table 1 provides additional details about the specific samples excluded during the various filtering steps.

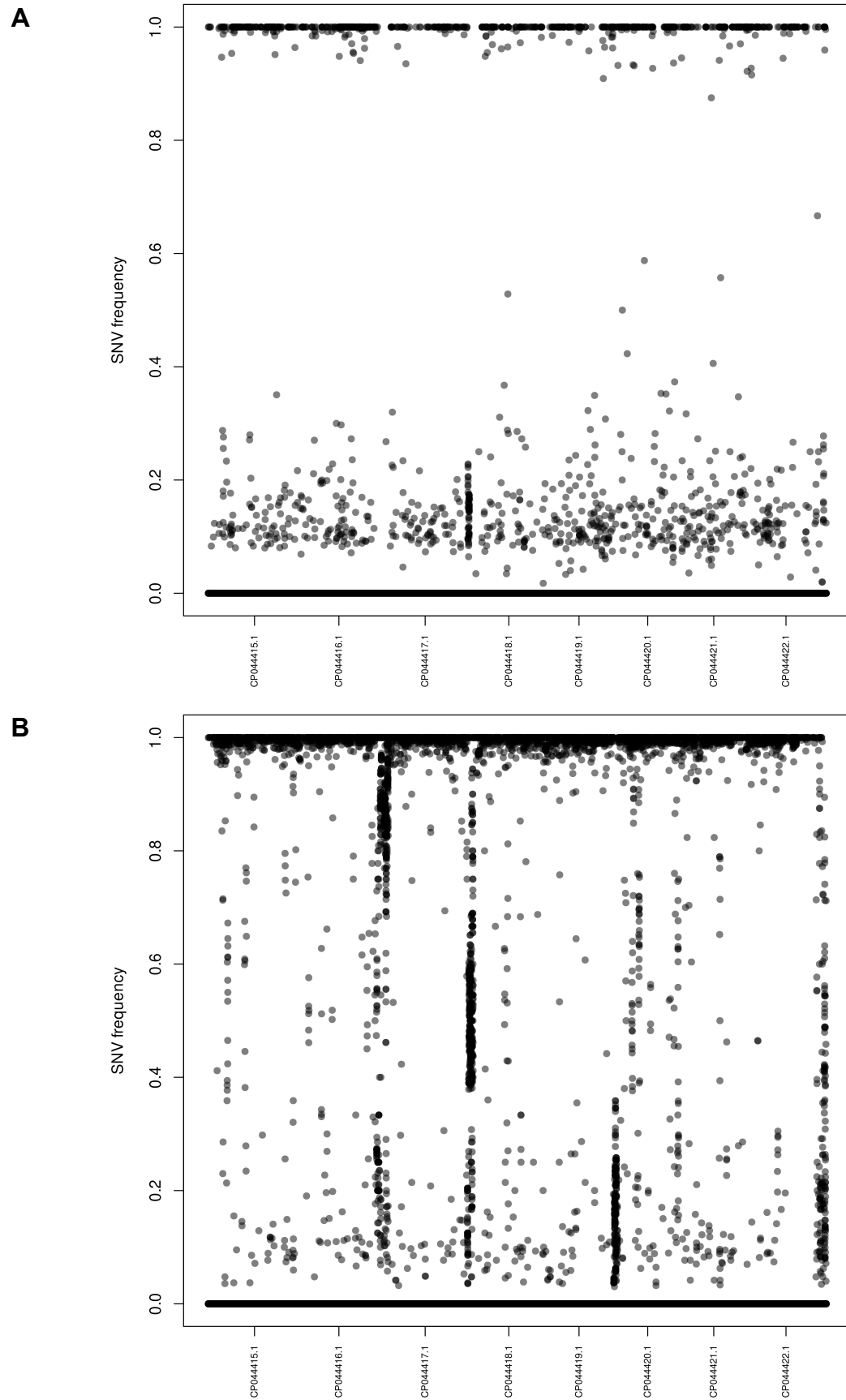


Figure S2. Estimating the presence of mixed infections. A) Isolate representing an infection with a single parasite population (FIN22), whereas B) shows an isolate representing an infection with more than one parasite population (EG44409). The latter isolate was excluded from all downstream analyses.

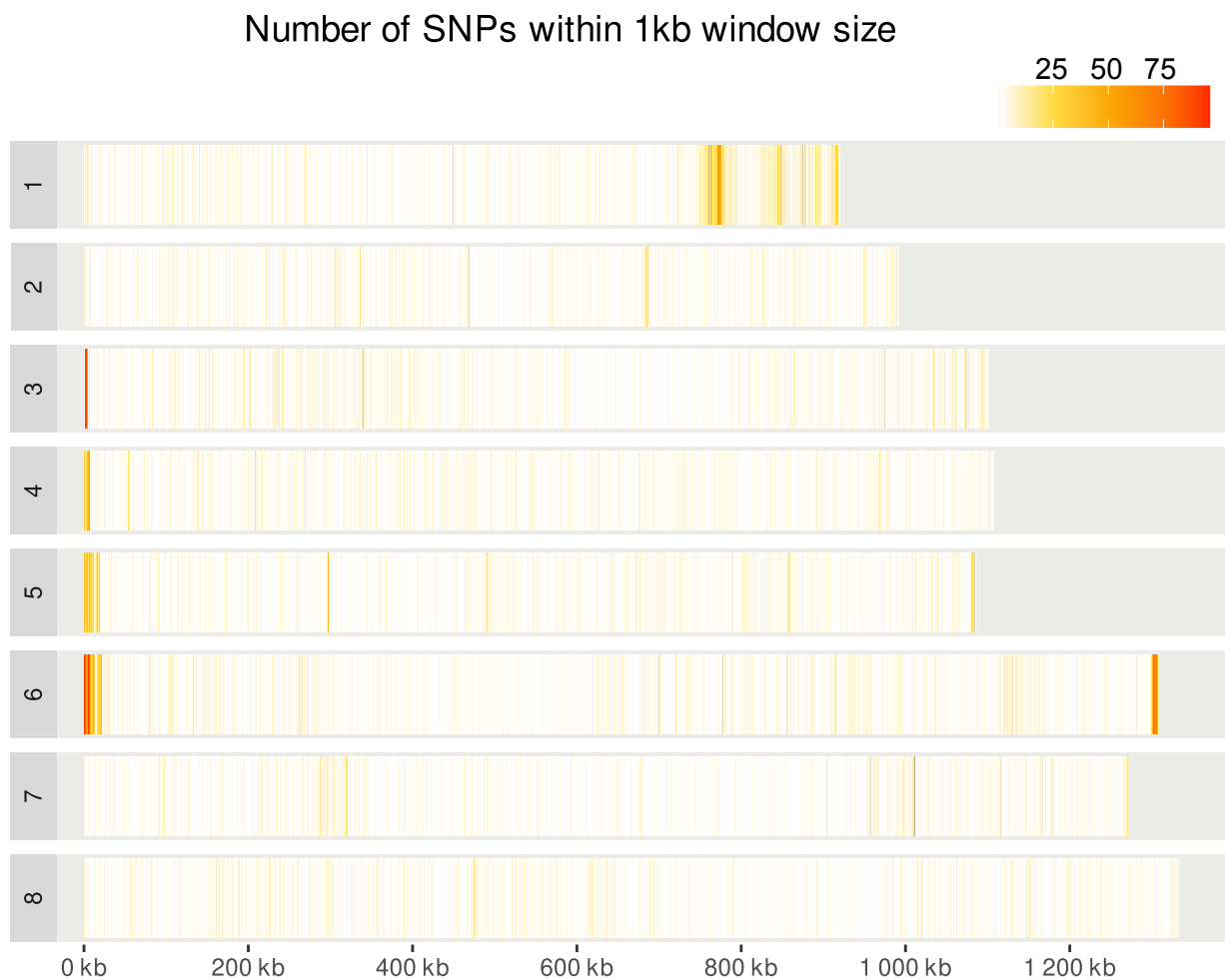


Figure S3. SNP density along the chromosomes. Number of SNPs in non-overlapping windows of 1 kb across each chromosome. Regions with higher density of SNPs are coloured in dark red.

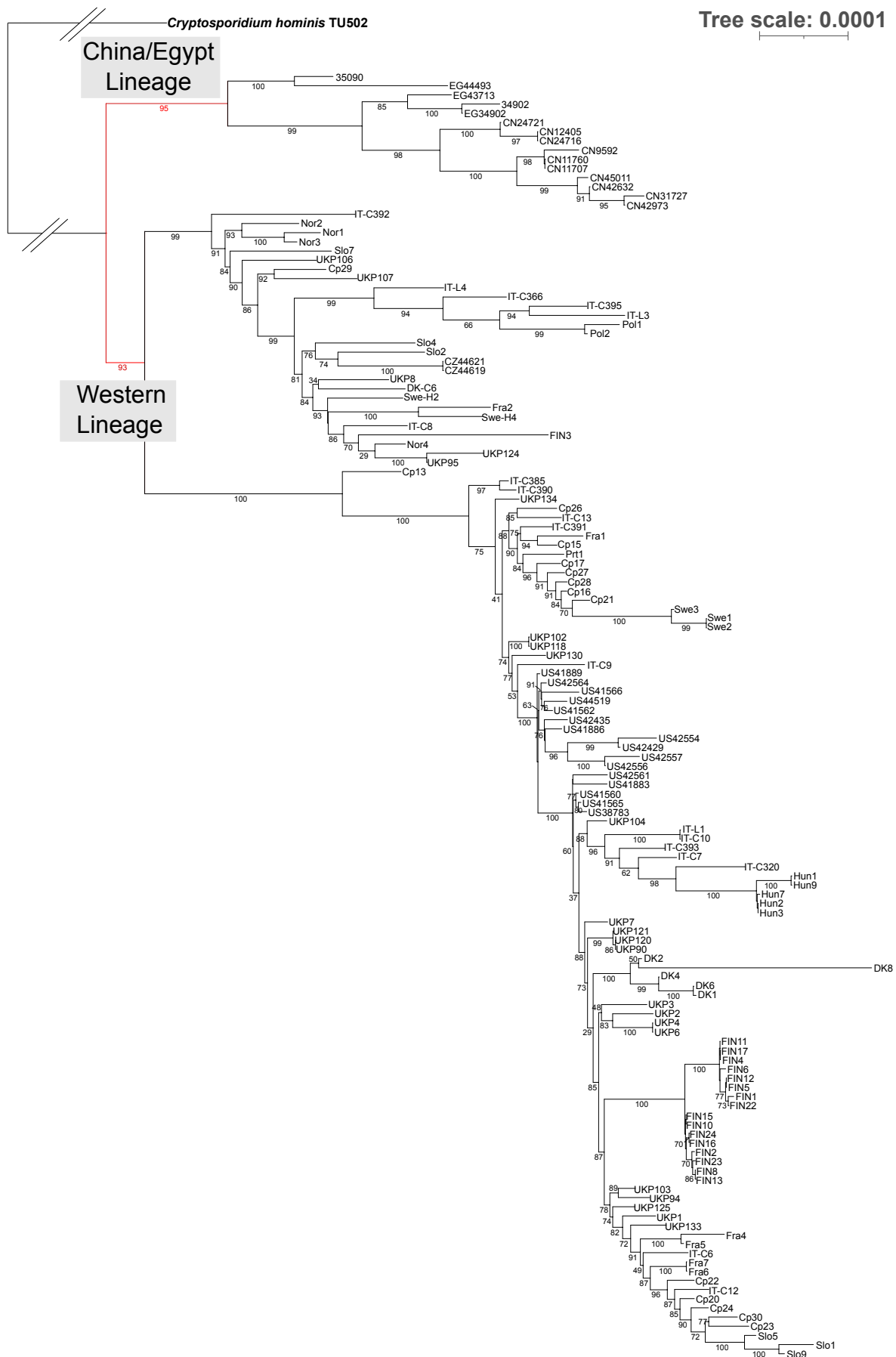


Figure S4. Maximum Likelihood (ML) phylogenetic tree inferred on a set of 179 genes. *Cryptosporidium hominis* TU502 was used as outgroup to root. The branch length leading to the outgroup has been shortened for presentation purposes.

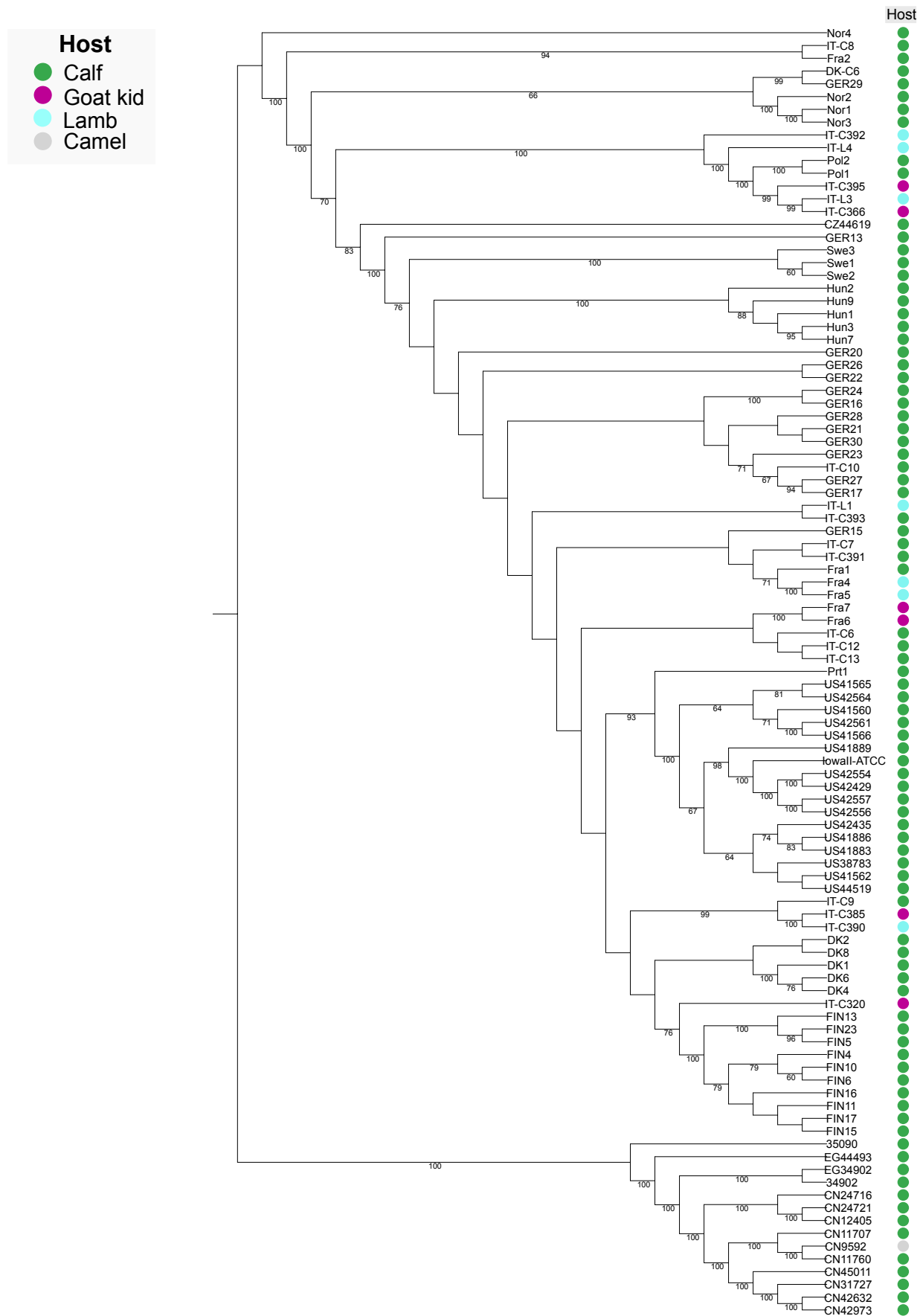
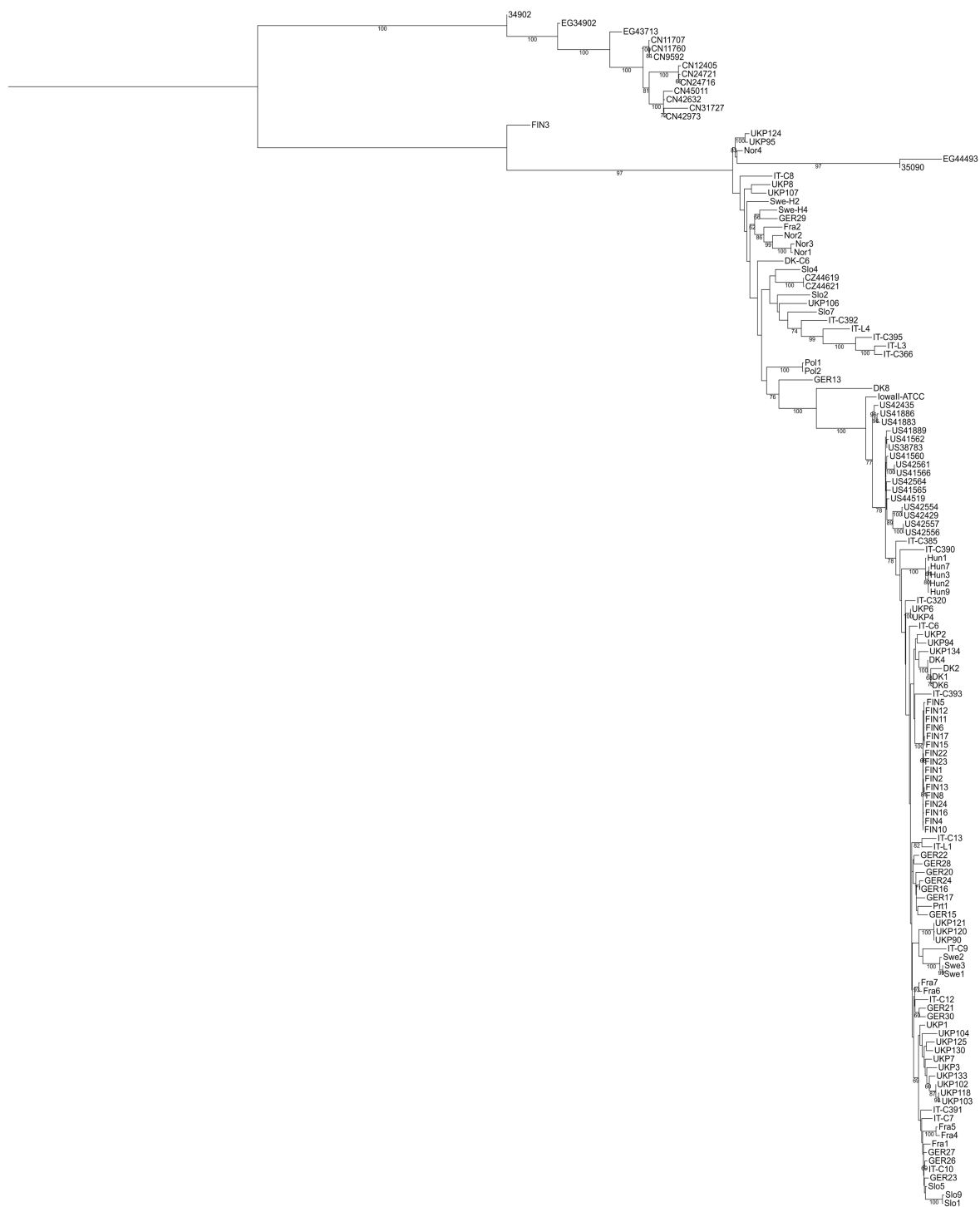


Figure S5. Maximum Likelihood tree of *C. parvum* isolates from ruminants only. Bootstrap values >60 are shown. Information about the host is mapped on the phylogeny.

Chromosome 1

A

Tree scale: 0.1



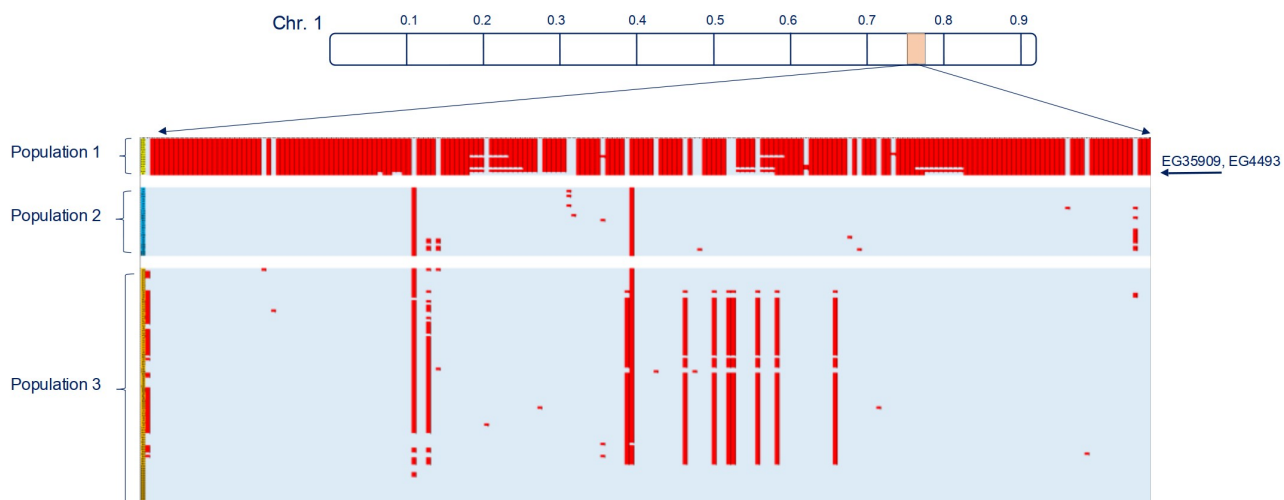


Figure S8. SNPs pattern in a 10 kb region of chromosome 1 spanning from position 755,934 to 768,672, highlighting the similarity of SNP distribution in isolates EG3509 and EG4493 (population 1) with isolates from the other two populations.

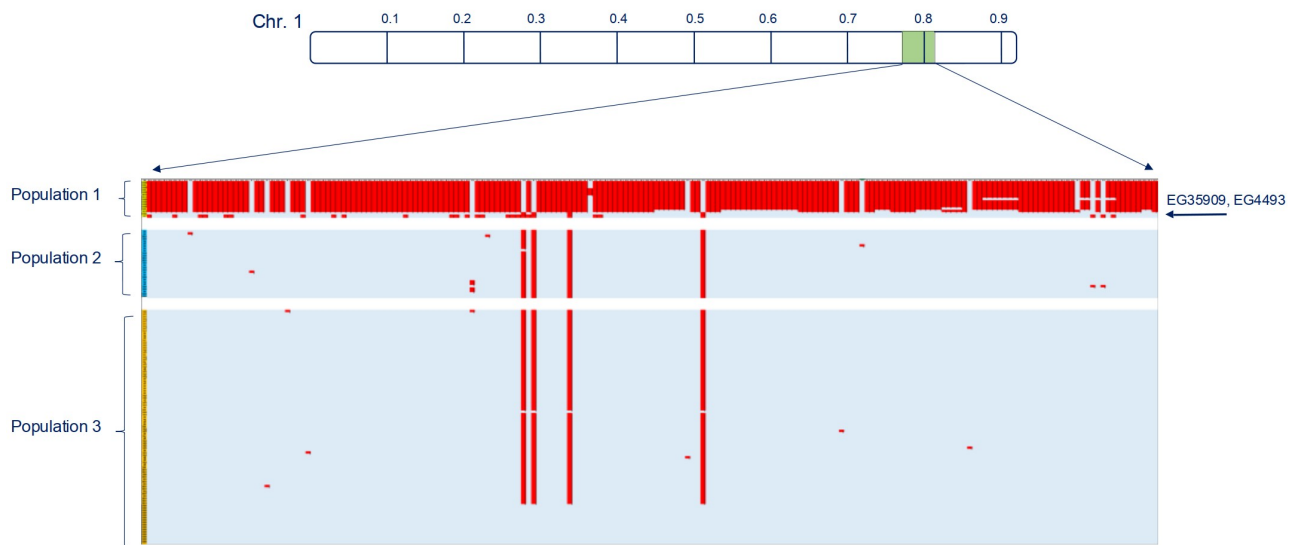


Figure S9. SNPs pattern in a 55 kb region of chromosome 1 spanning from position 768,729 to 823,729, highlighting the similarity of SNP distribution in isolates EG3509 and EG4493 (population 1) with isolates from the other two populations.

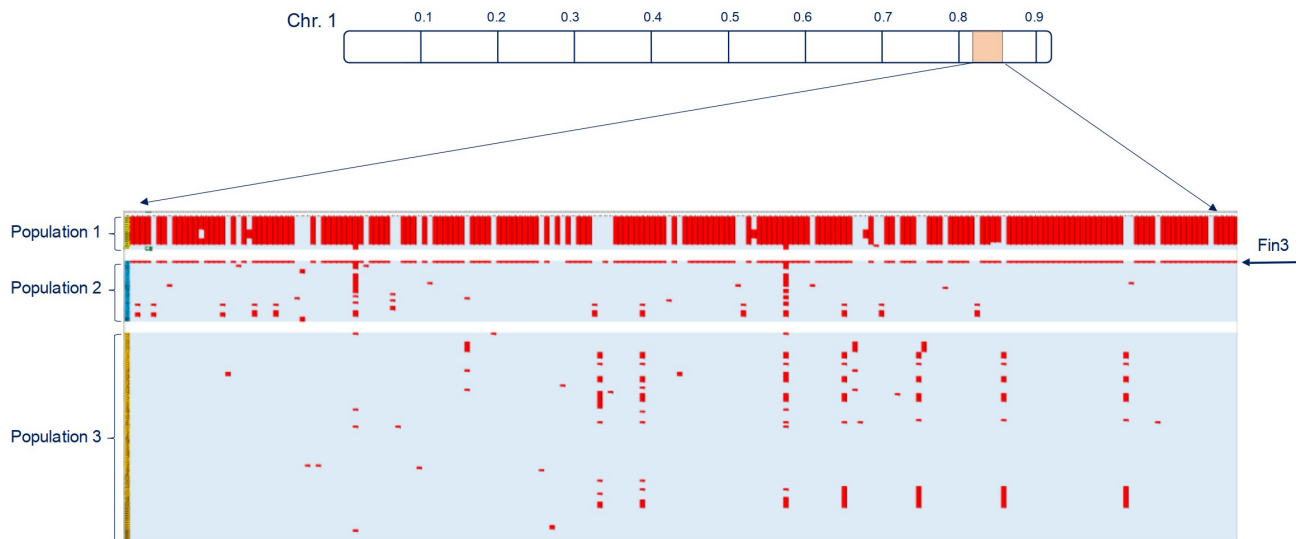
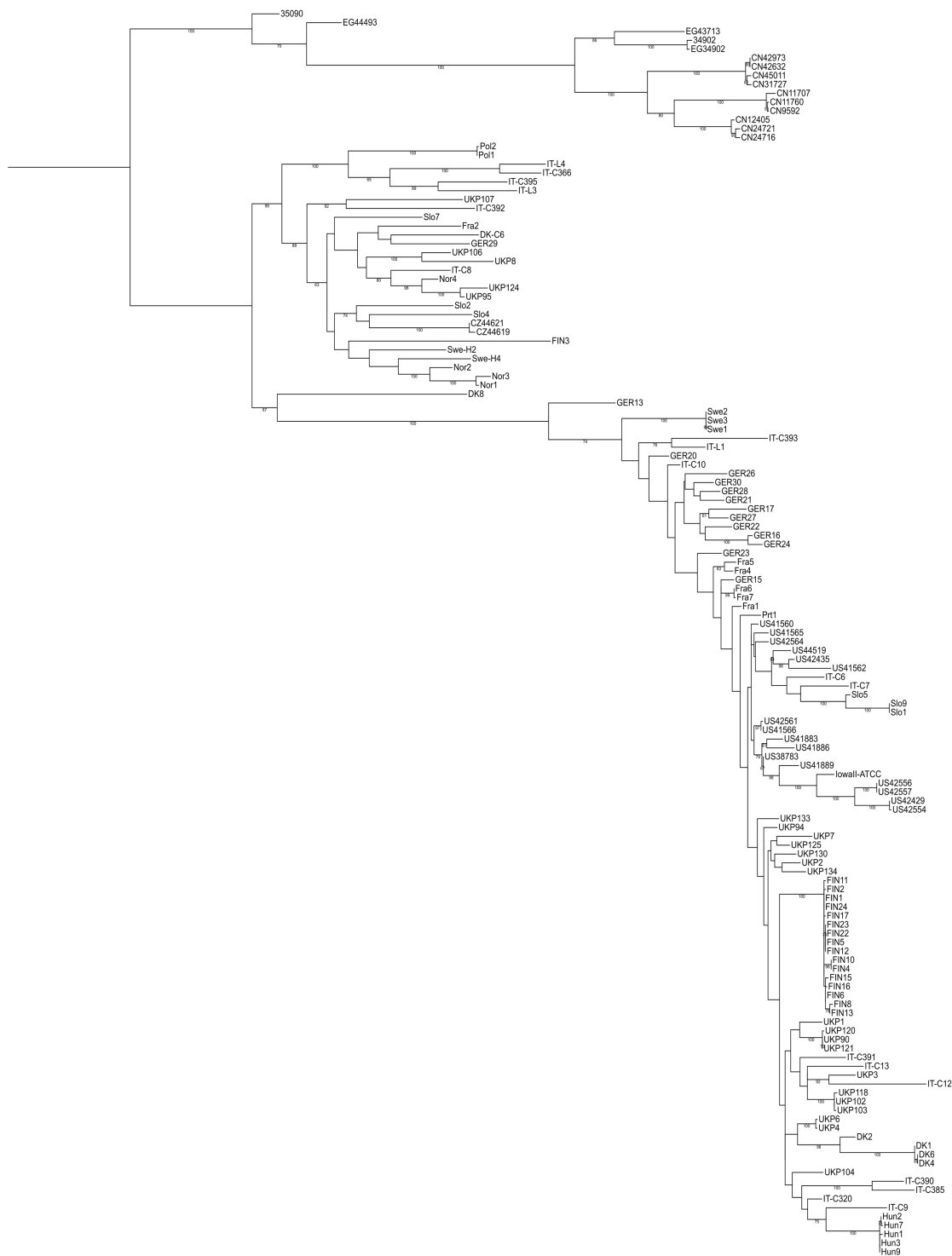


Figure S10. SNPs pattern in a 50 kb region of chromosome 1 spanning from position 824,800 to 874,170, highlighting the similarity of SNP distribution in isolate FIN3 (population 2) with isolates from population 1.

Chromosome 2

Tree scale: 0.1

A



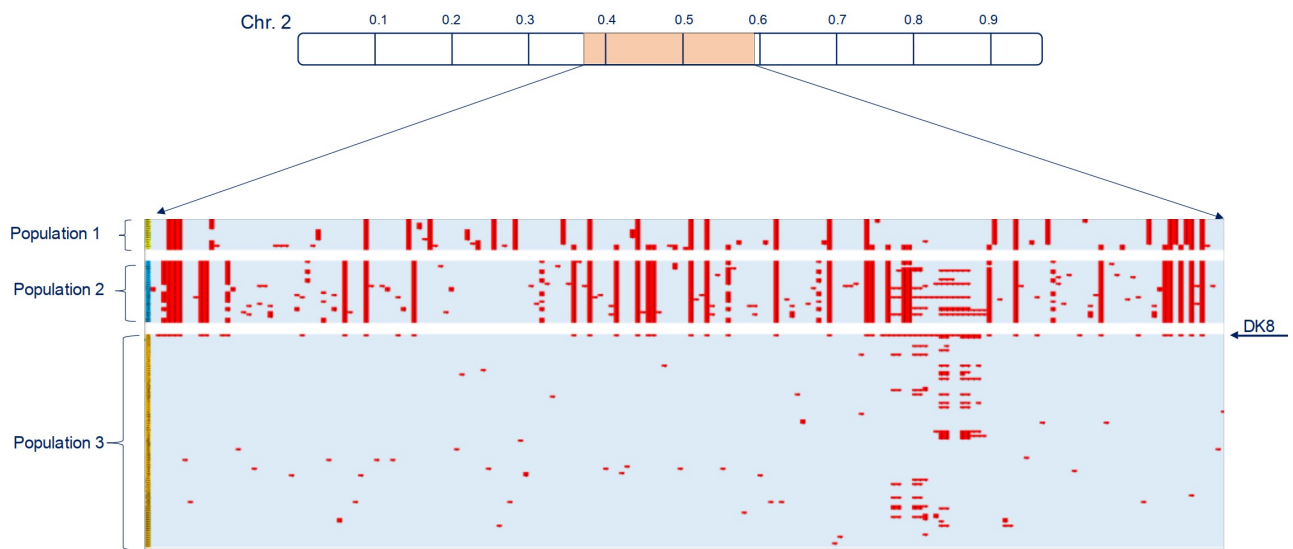
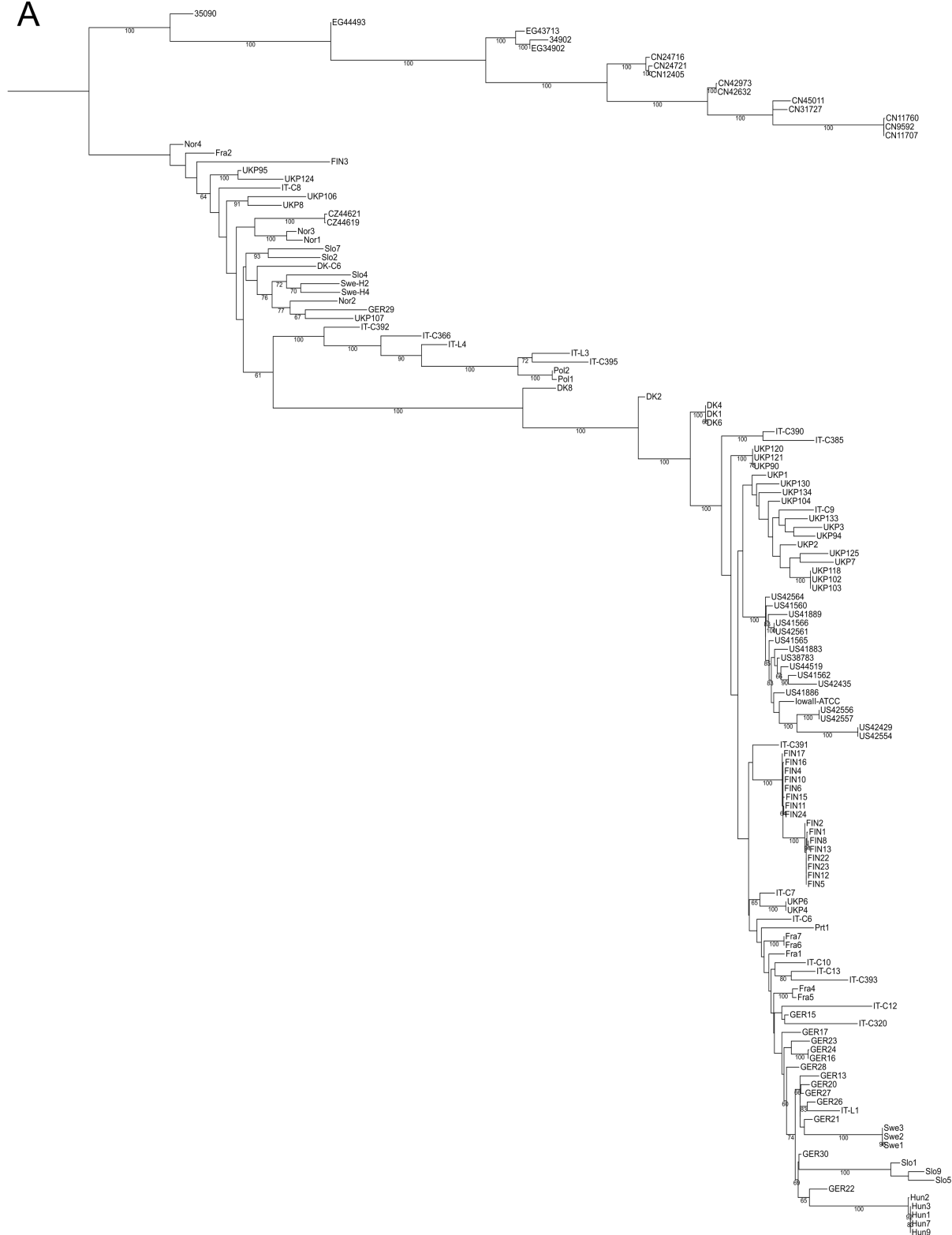


Figure S12. SNPs pattern in a 210 kb region on chromosome 2 spanning from position 384,000 to 594,000, highlighting the similarity of SNP distribution in isolate DK8 (population 3) with isolates from the other two populations.

Chromosome 3

Tree scale: 0.1

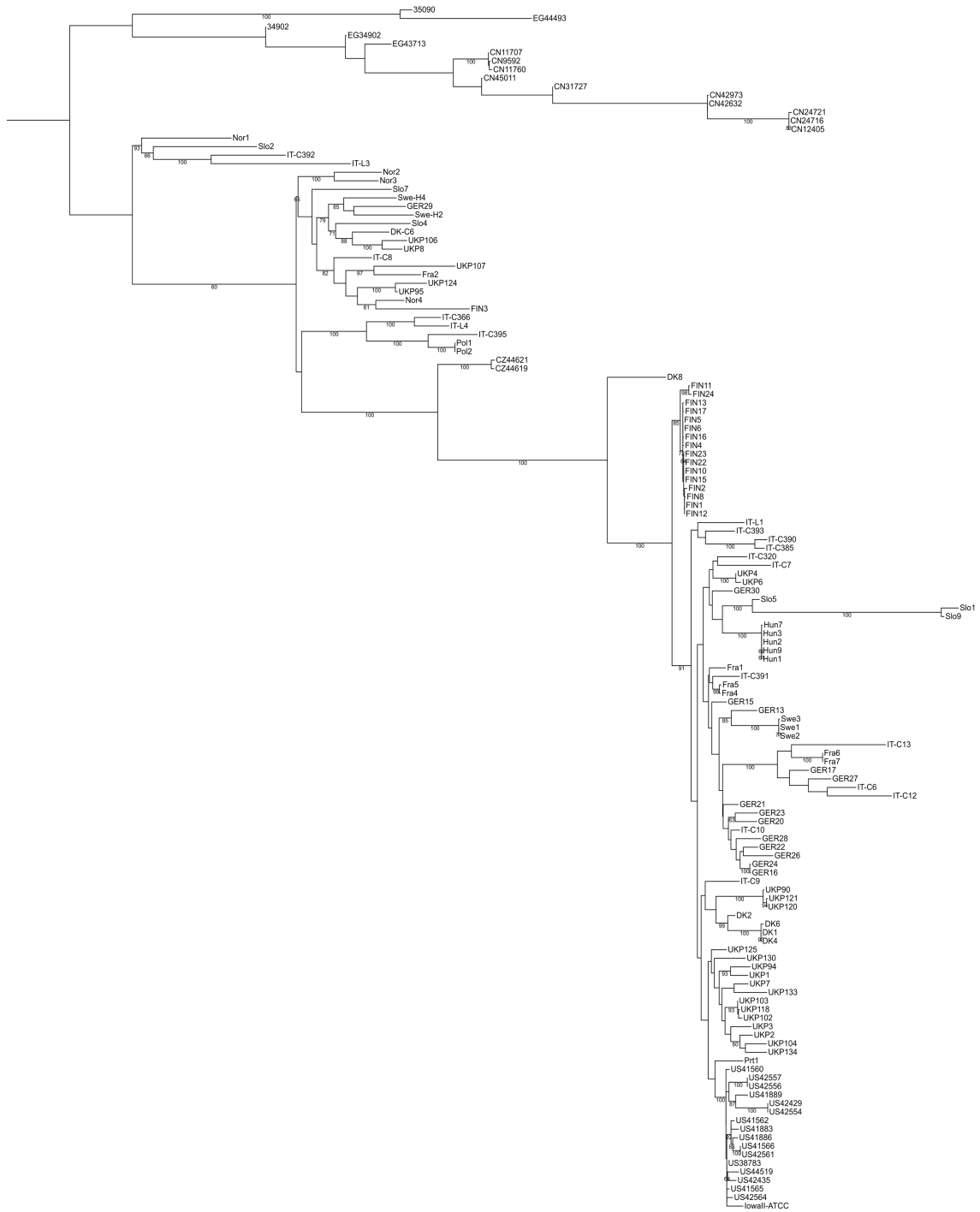
A



Chromosome 4

Tree scale: 0.1

A



B

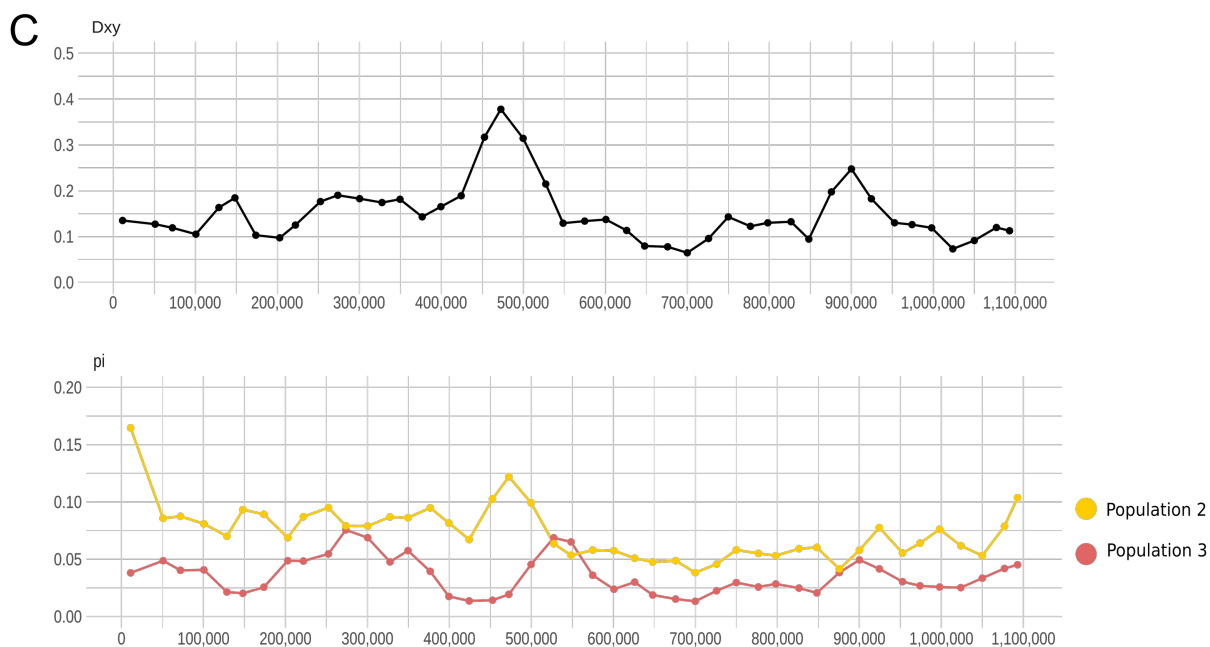
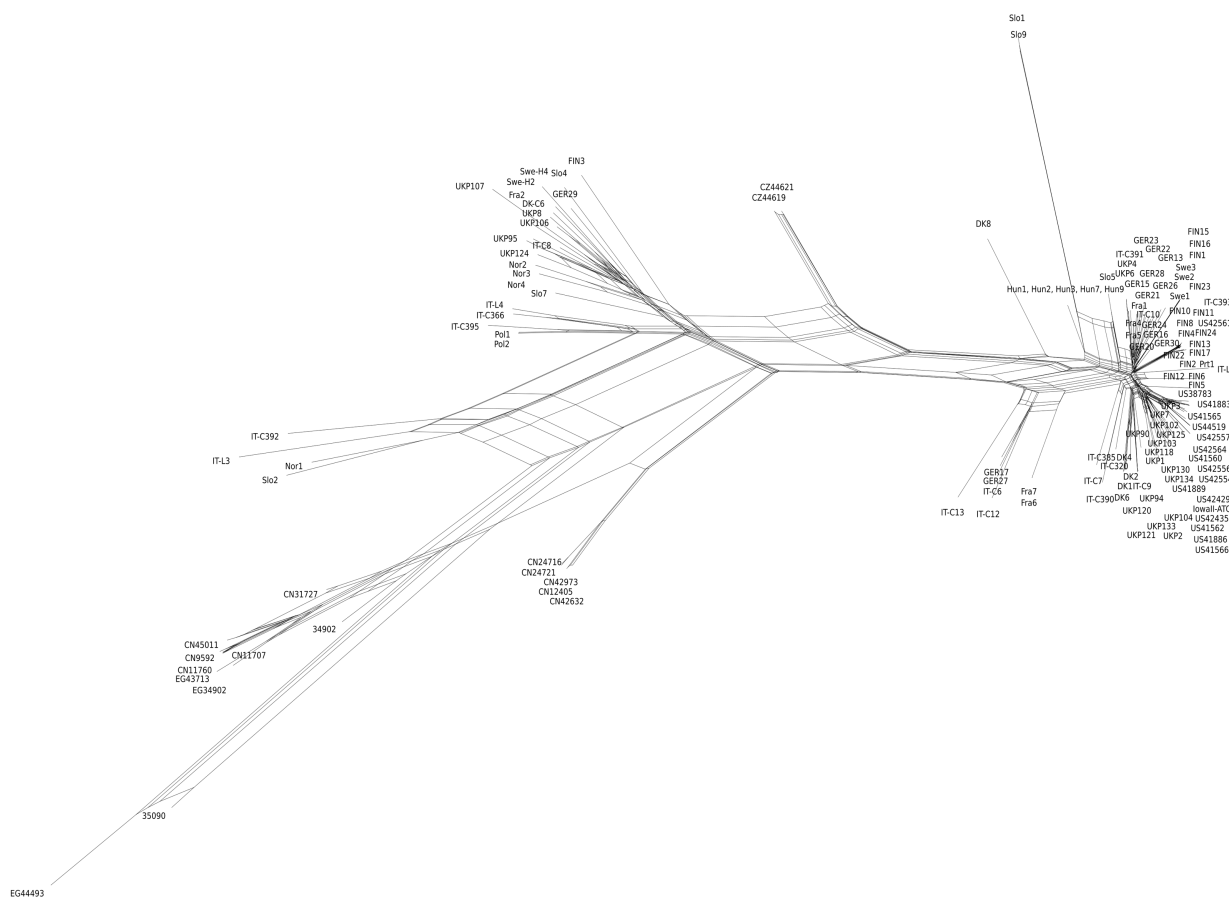


Figure S14. Analysis on SNPs located on chromosome 4: A) ML phylogenetic tree, B) Splitstree and C) Dxy and Pi.

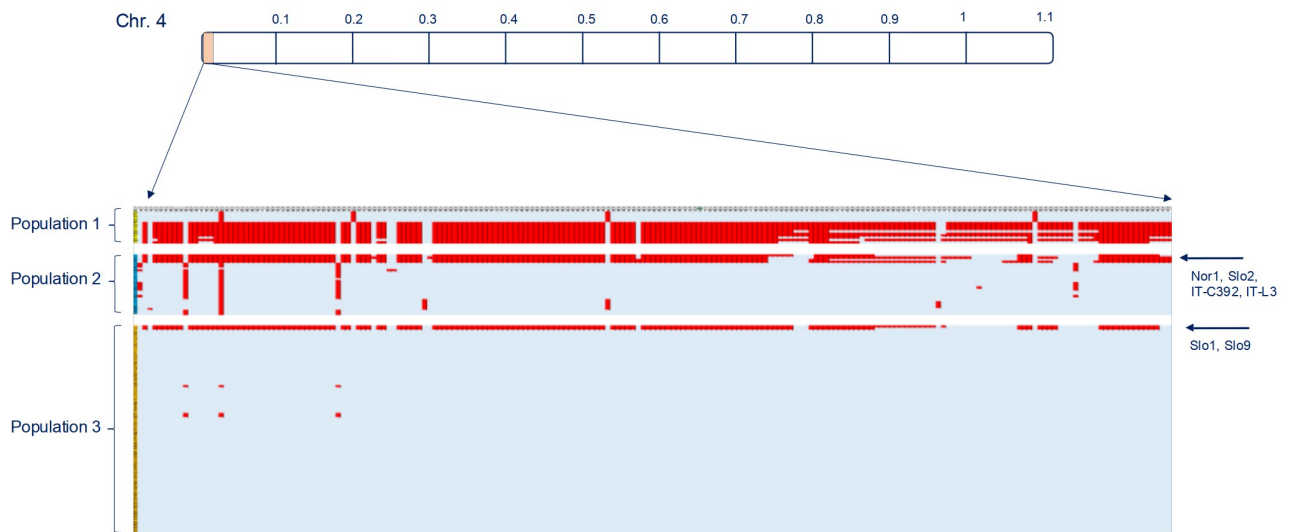
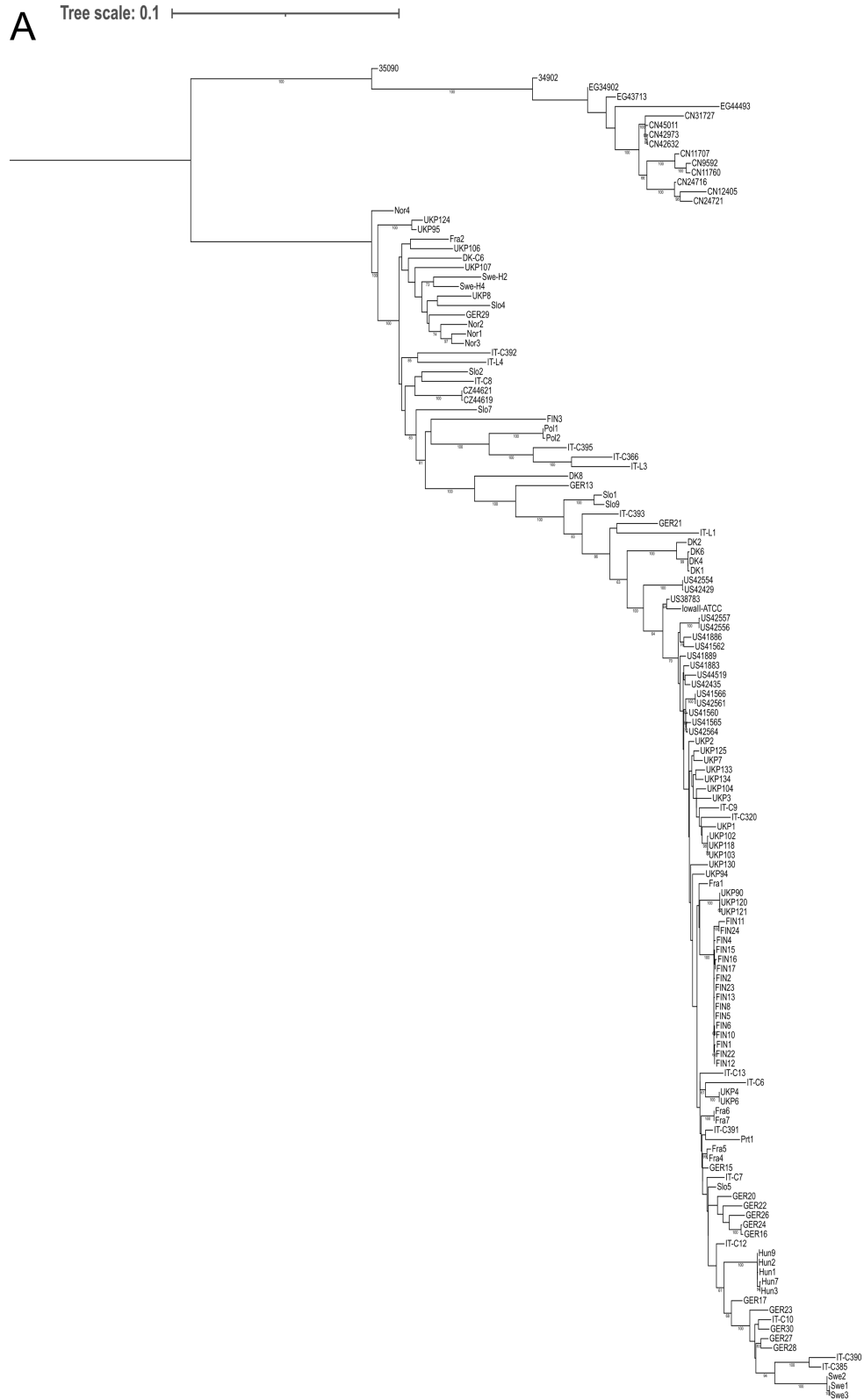


Figure S15. SNPs pattern in the first 8 kb adjacent to the 5' telomere of chromosome 4, highlighting the similarity of SNP distribution in isolates Nor1, Slo2, IT-C392 and IT-L38 (population 2), and isolates Slo2 and Slo9 (population 3), with isolates from population 1.

Chromosome 5



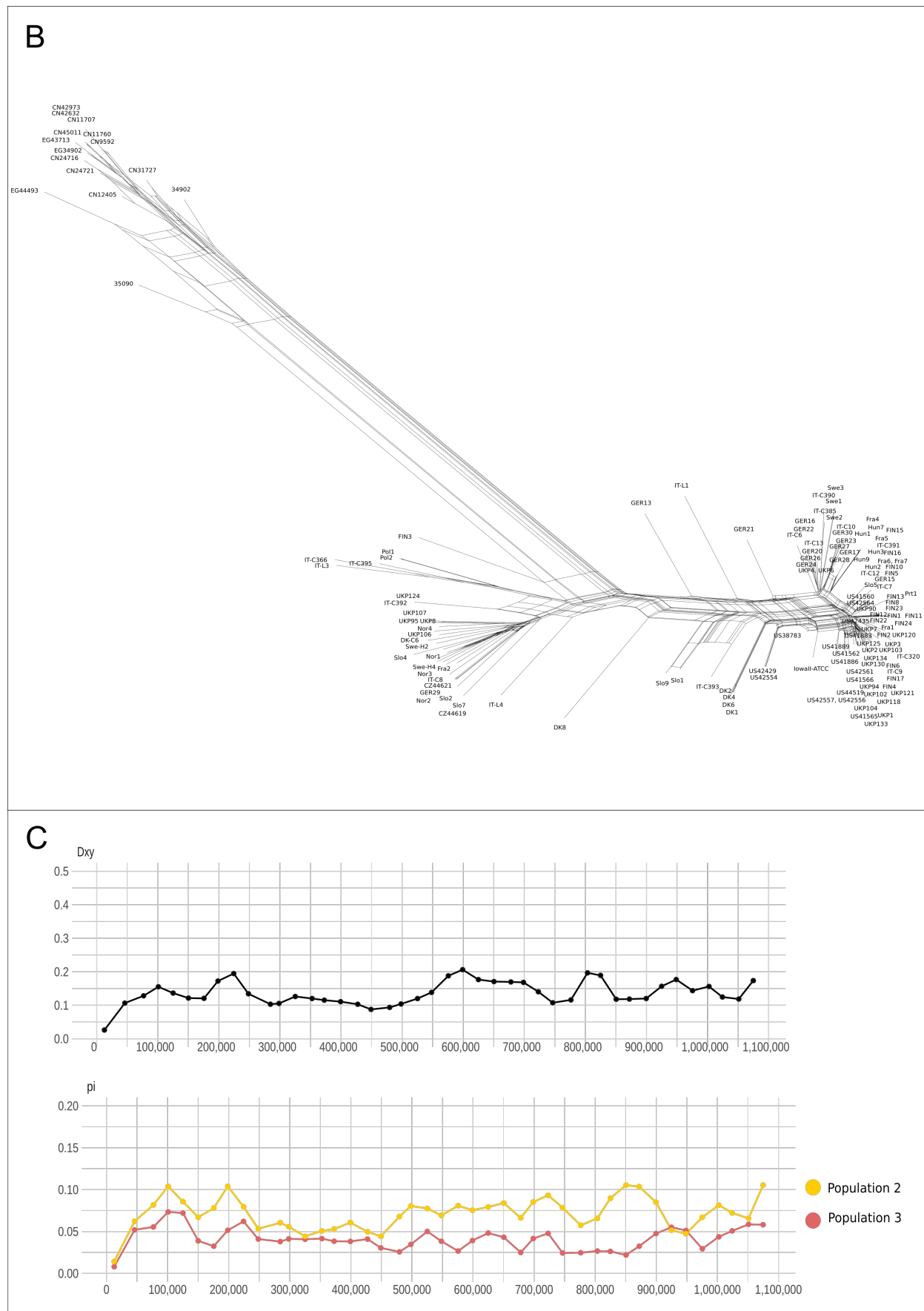
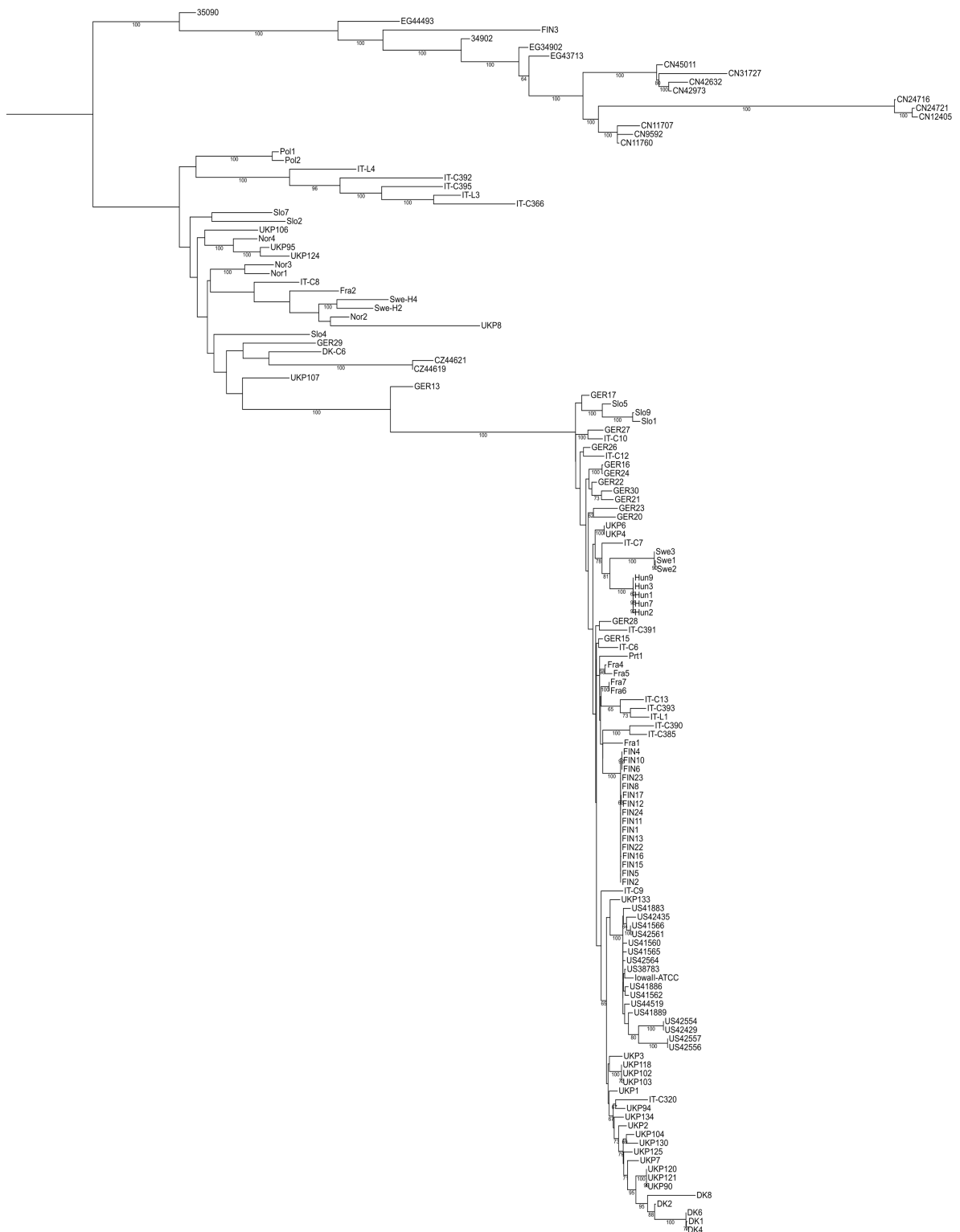


Figure S16. Analysis on SNPs located on chromosome 5: A) ML phylogenetic tree, B) Splitstree and C) Dxy and Pi.

Chromosome 6

Tree scale: 0.1

A



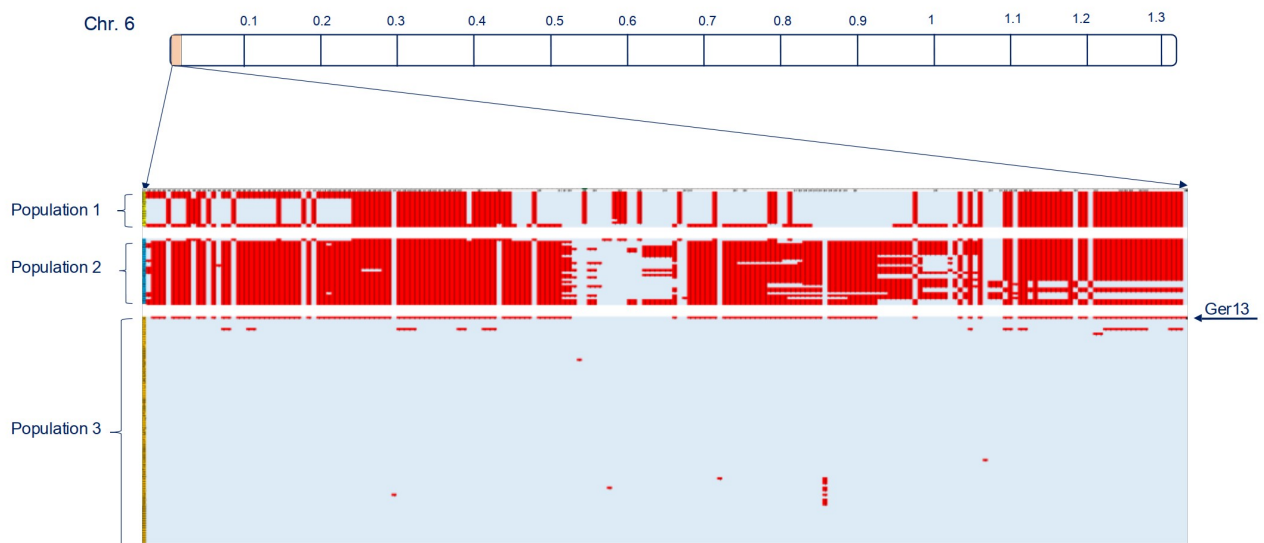
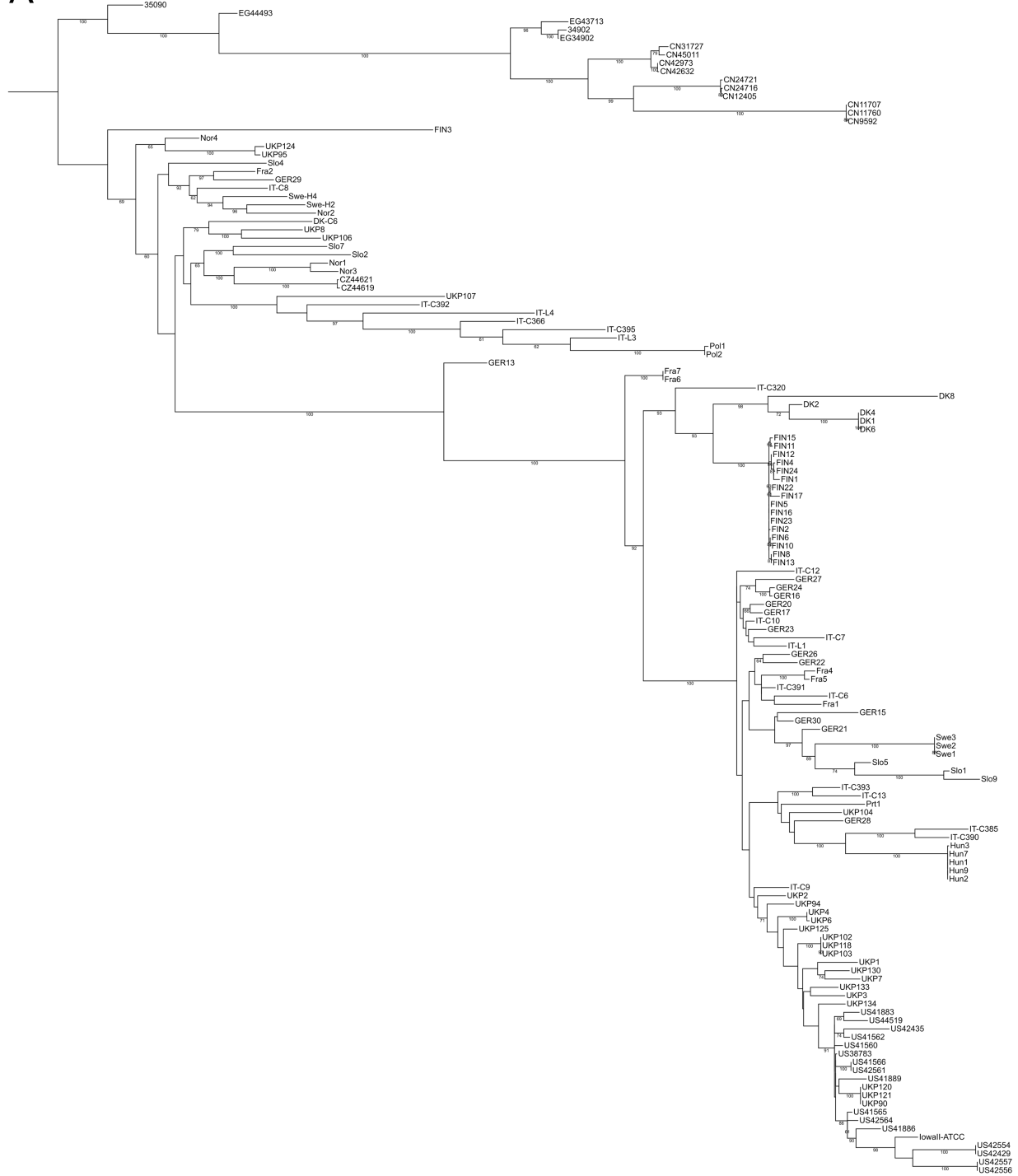


Figure S18. SNPs pattern in the first 18 kb adjacent to the 5' telomere of chromosome 6, highlighting the similarity of SNP distribution in isolate Ger13 (population 3) with isolates from the other two populations.

Chromosome 7

Tree scale: 0.01

A



Chromosome 8

Tree scale: 0.1

A



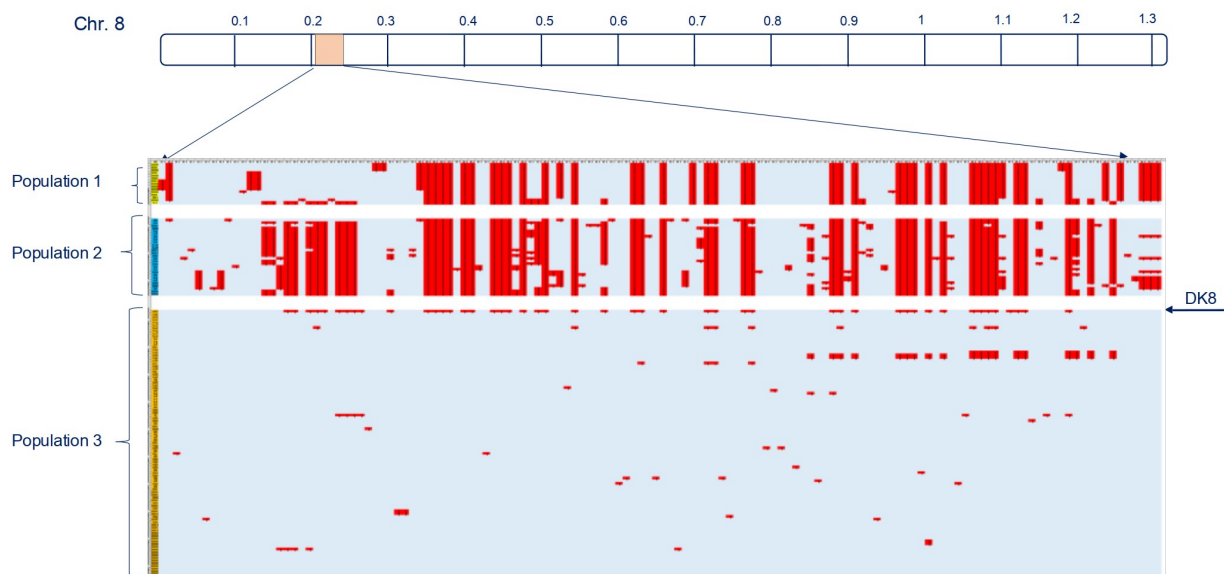


Figure S21. SNPs pattern in a 30 kb region spanning position 210,000 to 240,000 on chromosome 8, highlighting the similarity of SNP distribution in isolate DK8 (population 3) with isolates from the other two populations.S

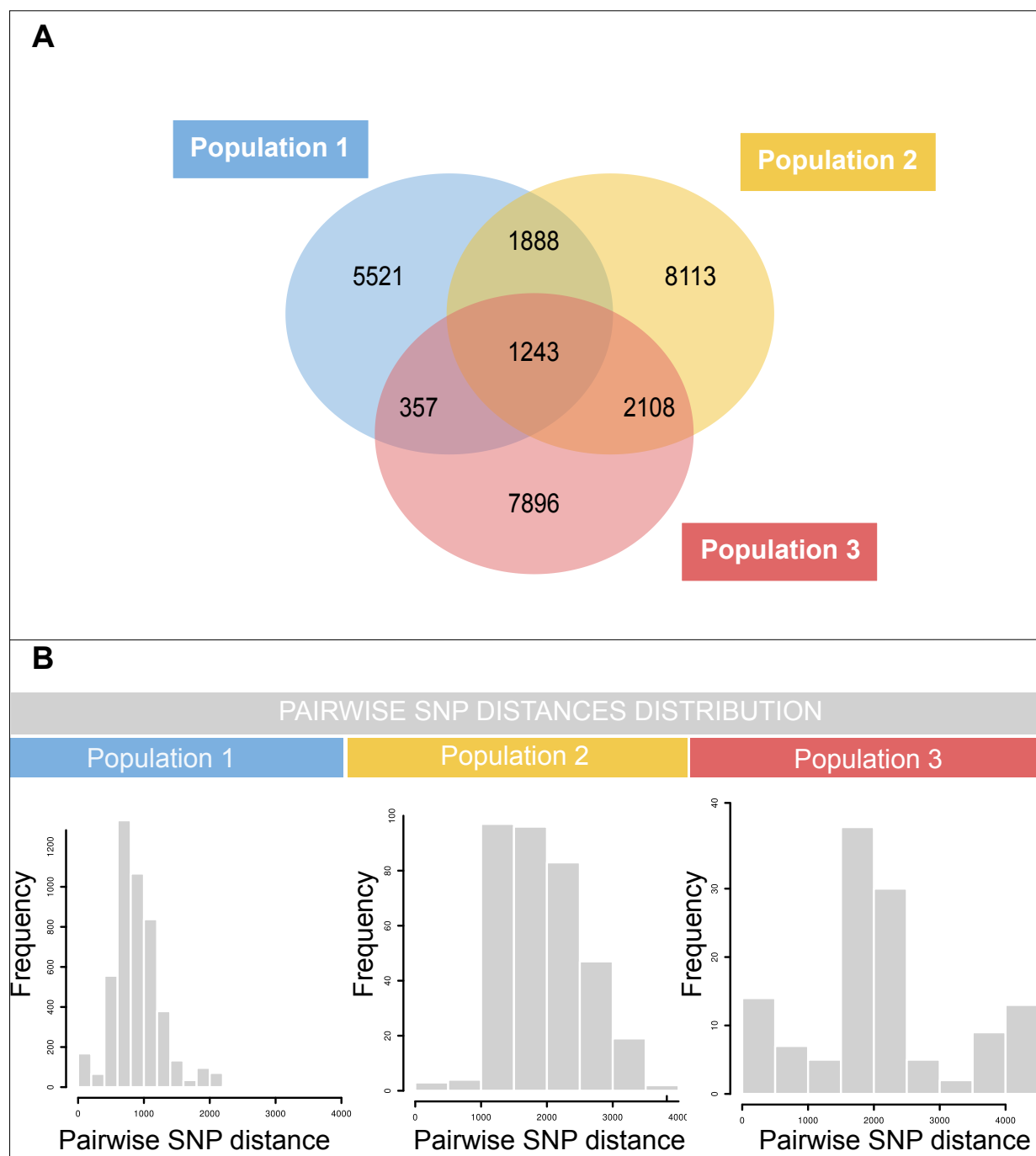


Figure S22. Distribution of shared and population-specific SNPs. A) Venn Diagram B) Histogram of the pairwise SNP distances distribution within each Population.

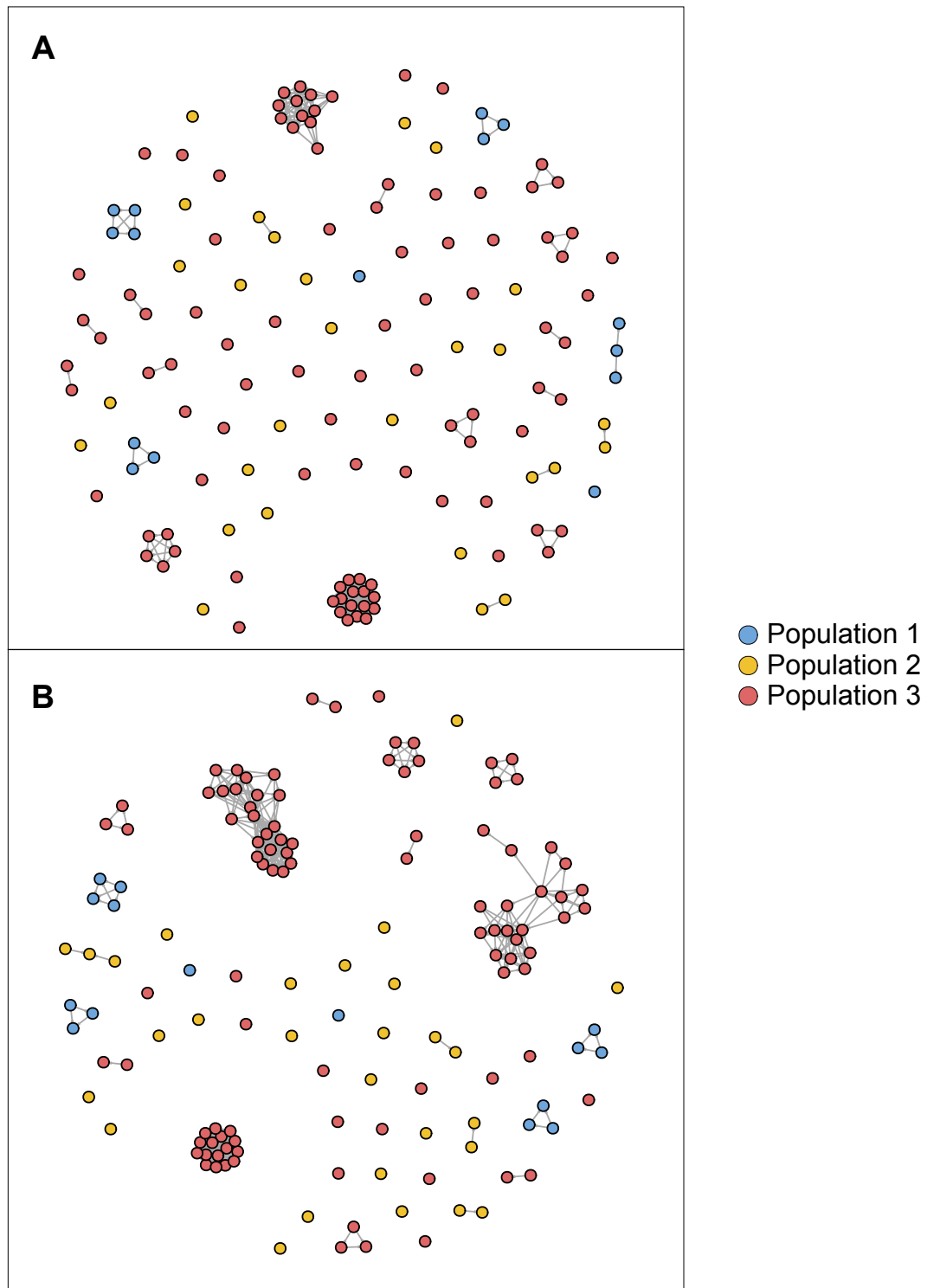


Figure S23. Relatedness network for pairs of isolates identified as having high proportions of IBD sharing. Each node identifies a unique isolate and an edge is drawn between two isolates if they share more than A) 90% of their genome IBD, and B) 80% of their genome IBD. Individuals are coloured according to the ADMIXTURE population.

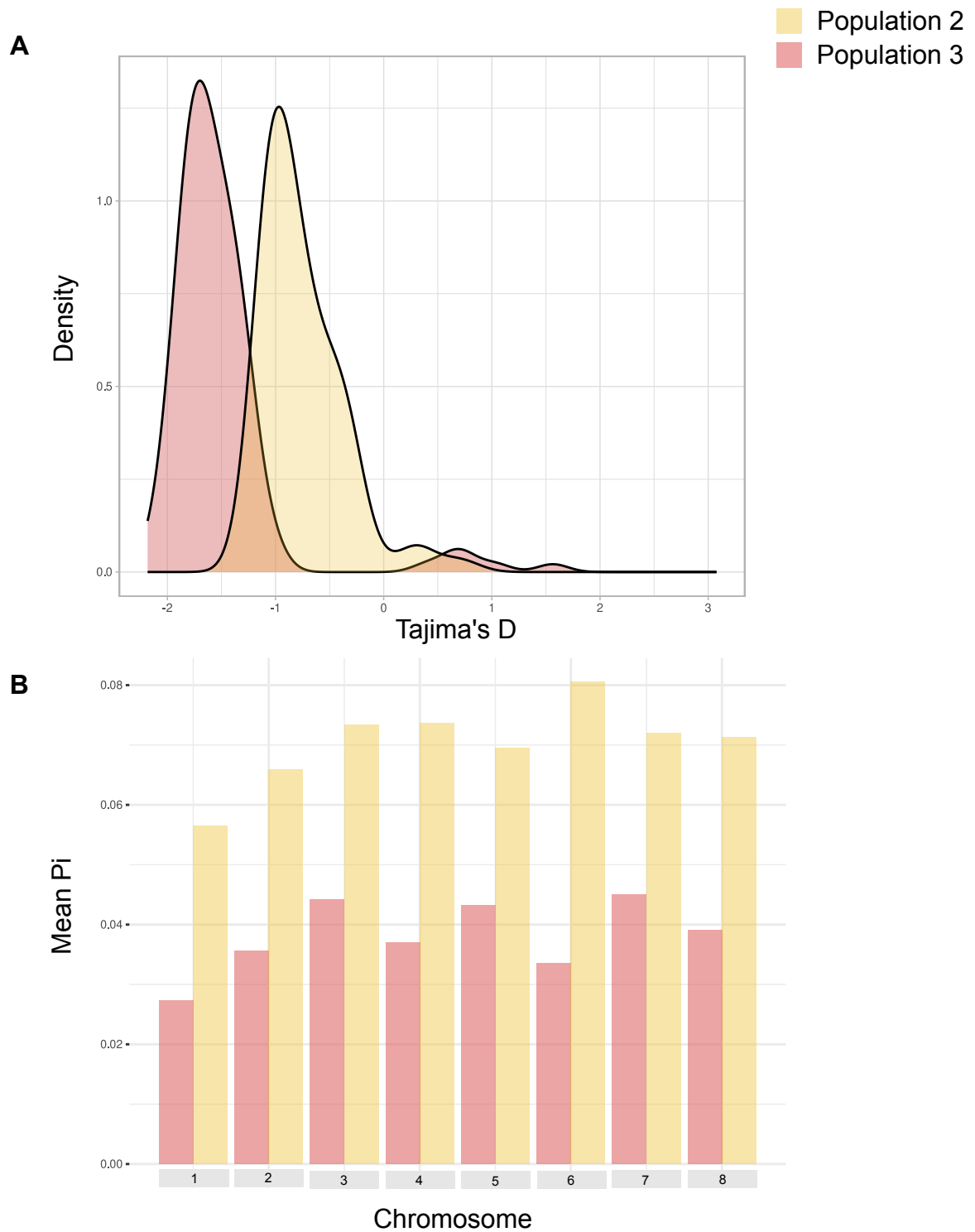


Figure S24. Distribution of A) Tajima's D values and B) nucleotide diversity (π) in population 2 and population 3. In these analyses, only one isolate per each of the 13 clusters of highly related genomes (<50 SNPs) within population 2 was used.