Supplemental Material

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

Greta Bellinzona[1], Tiago Nardi[1], Michele Castelli[1], Gherard Batisti Biffignandi[1], Karim Adjou[2], Martha Betson[3], Yannick Blanchard[4], Ioana Bujila[5], Rachel Chalmers[6], Rebecca Davidson[7], Nicoletta D'Avino[8], Tuulia Enbom[9], Jacinto Gomes[10], Gregory Karadjian[2], Christian Klotz[11], Emma Östlund[12], Judith Plutzer[13], Ruska Rimhanen-Finne[14], Guy Robinson[5], Anna Rosa Sannella[15], Jacek Sroka[16], Christen Rune Stensvold[17], Karin Troell[12], Paolo Vatta[15], Barbora Zalewska[18], Claudio Bandi[19], 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 Figures

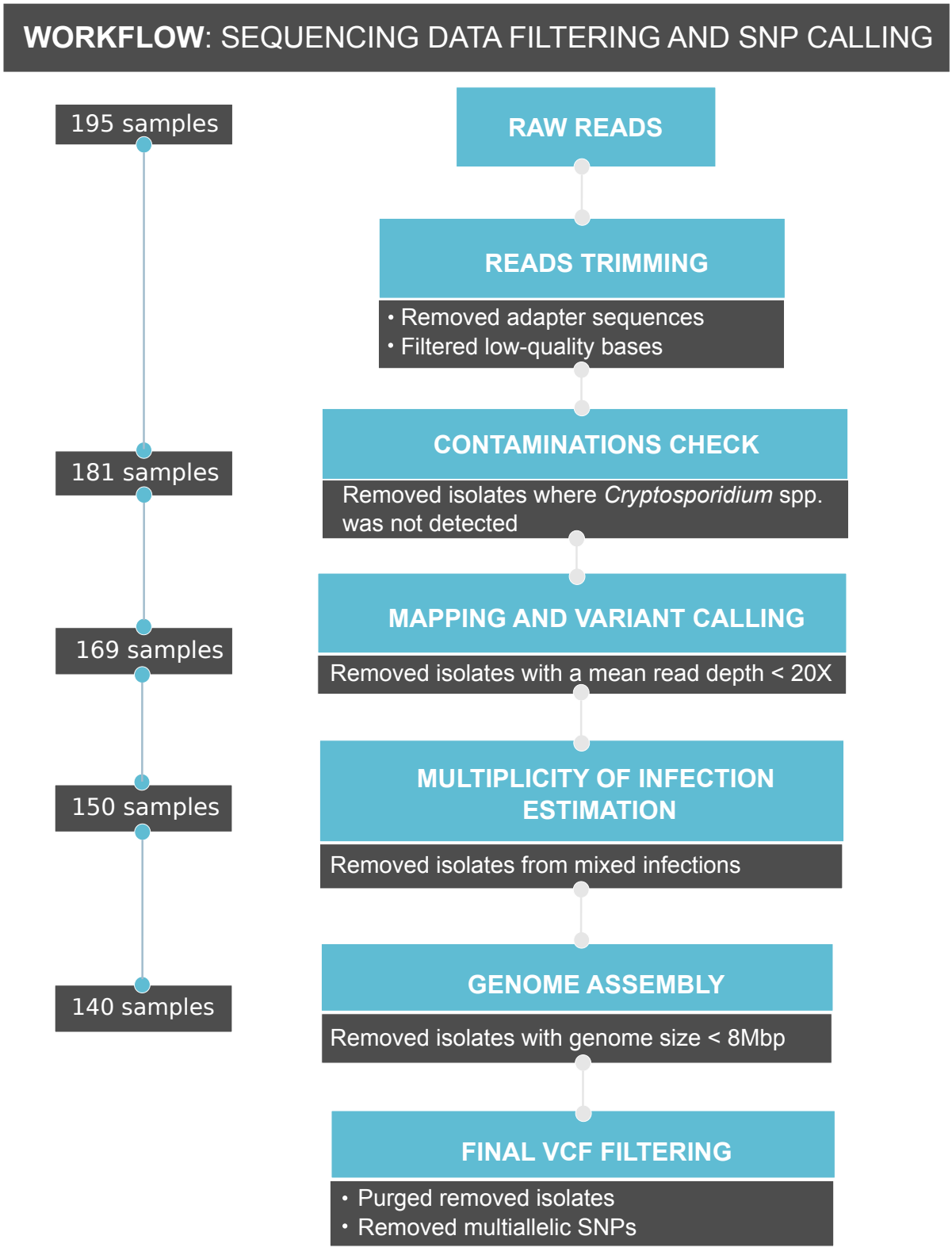


Figure S1. Workflow for data filtering and SNP calling. The figure outlines the sequential steps employed in data processing to derive the ultimate refined dataset used for the downstream analysis. The left side of the diagram enumerates the count of samples at each filtering step, underscoring the progressive 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.

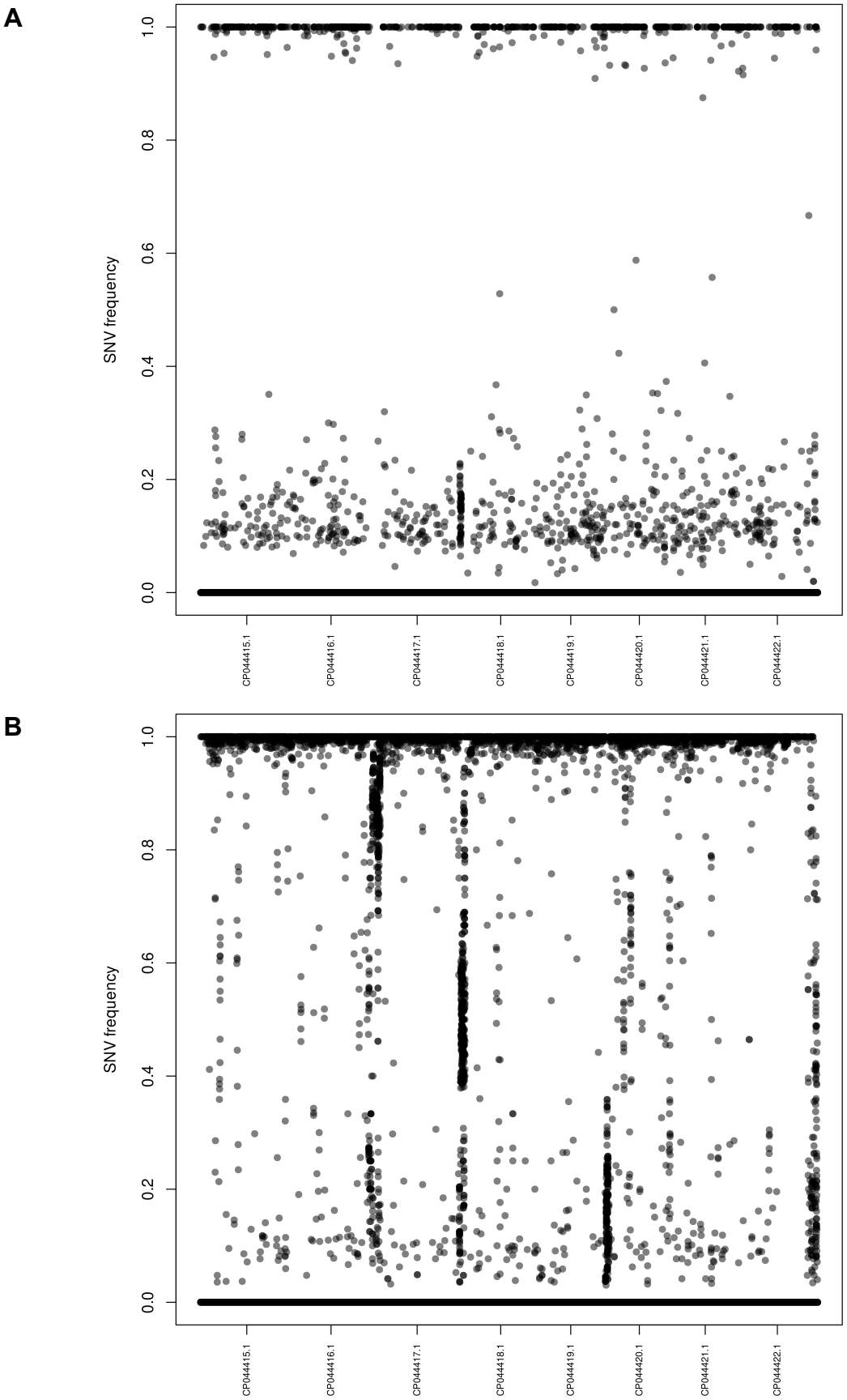


Figure S2. Estimating the presence of mixed infections. A) Isolate representing an infection with a single parasite population (FIN22), whereas B) shows an isolate representing an infection with more than one parasite population (EG44409). The latter isolate was excluded from all downstream analyses.

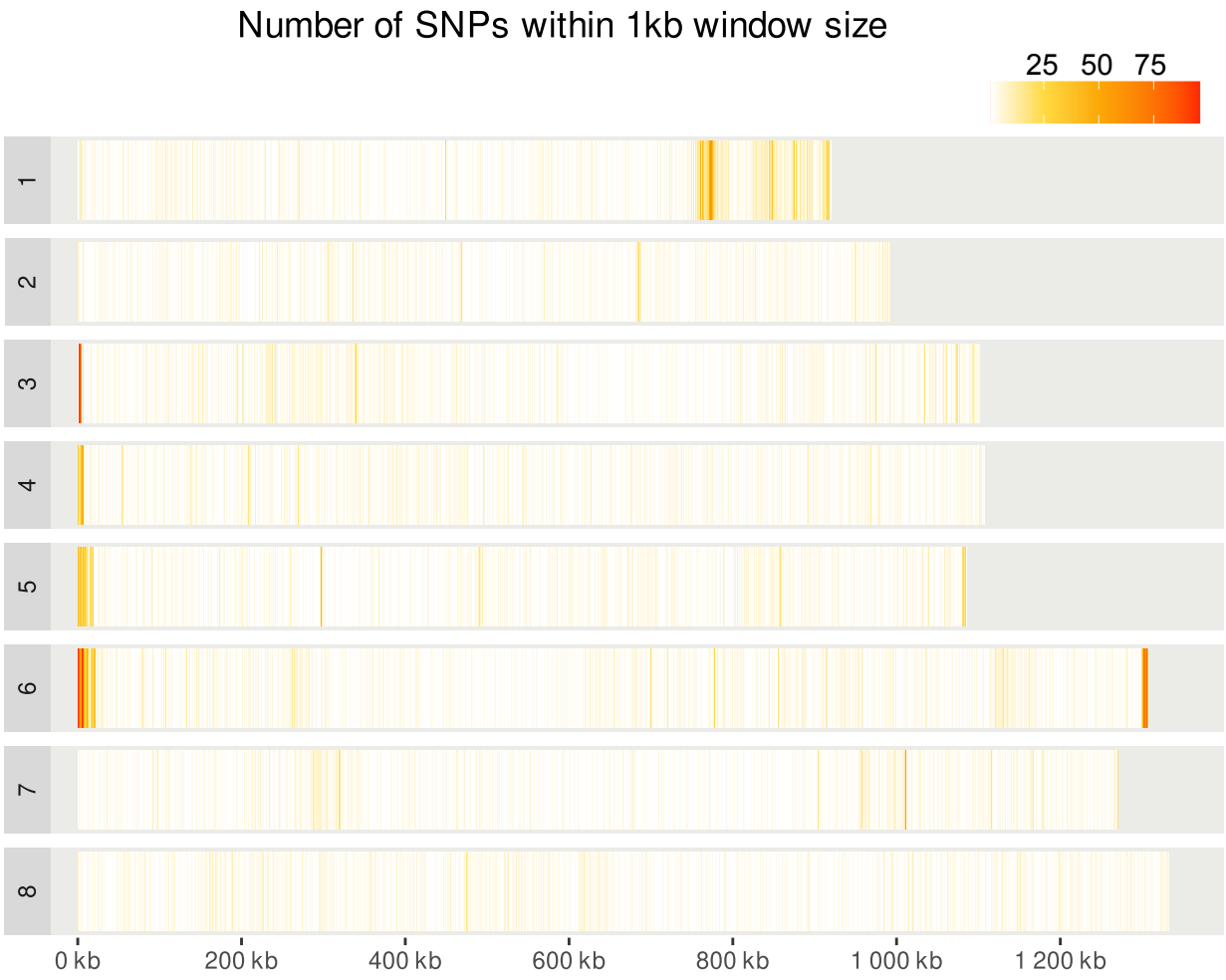


Figure S3. SNP density along the chromosomes. Number of SNPs in non-overlapping windows of 1 kb across each chromosome. Regions with higher density of SNPs are coloured in dark red.

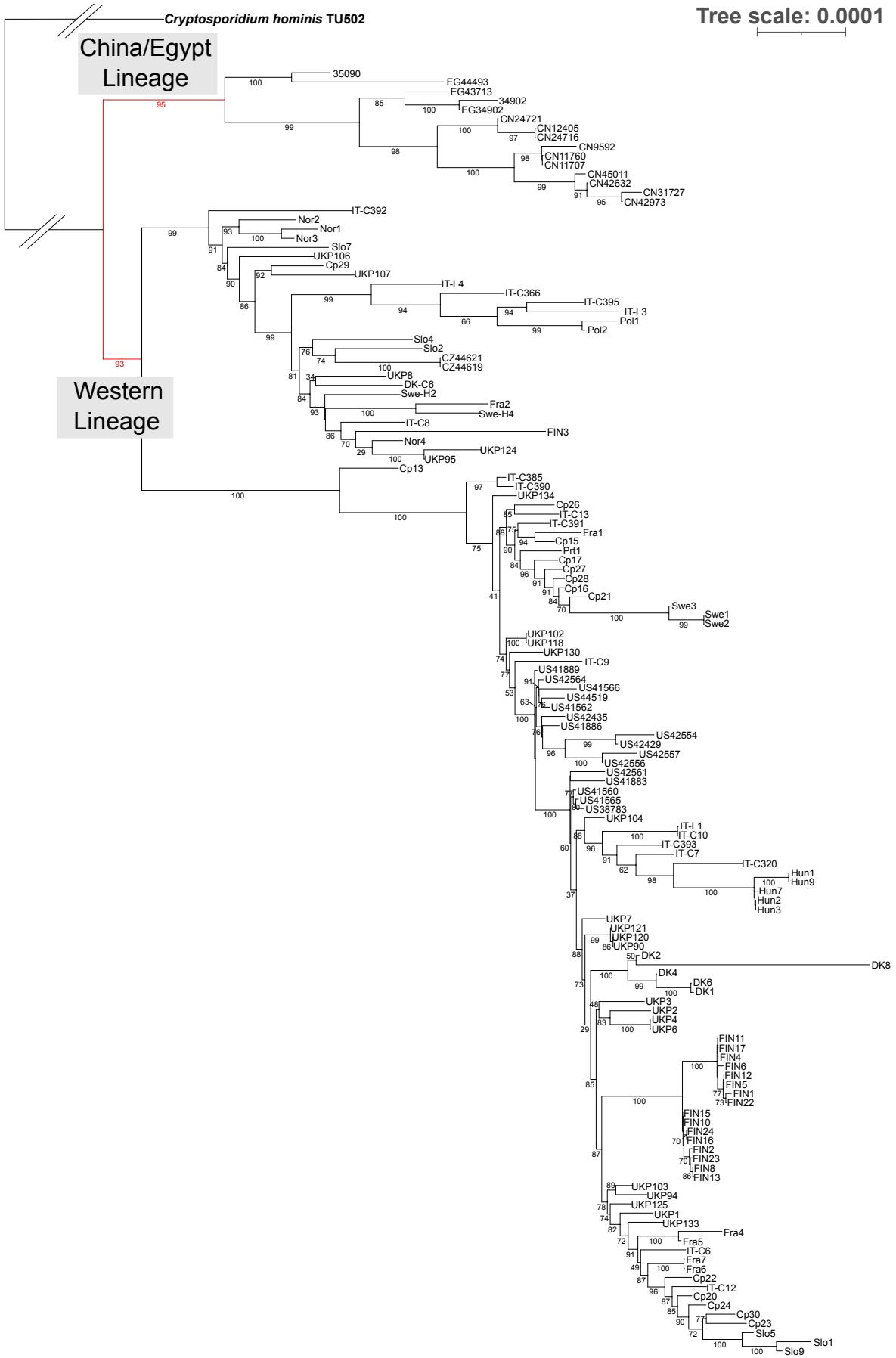


Figure S4. Maximum Likelihood (ML) phylogenetic tree inferred on a set of 179 genes. *Cryptosporidium hominis* TU502 was used as outgroup to root. The branch length leading to the outgroup has been shortened for presentation purposes.

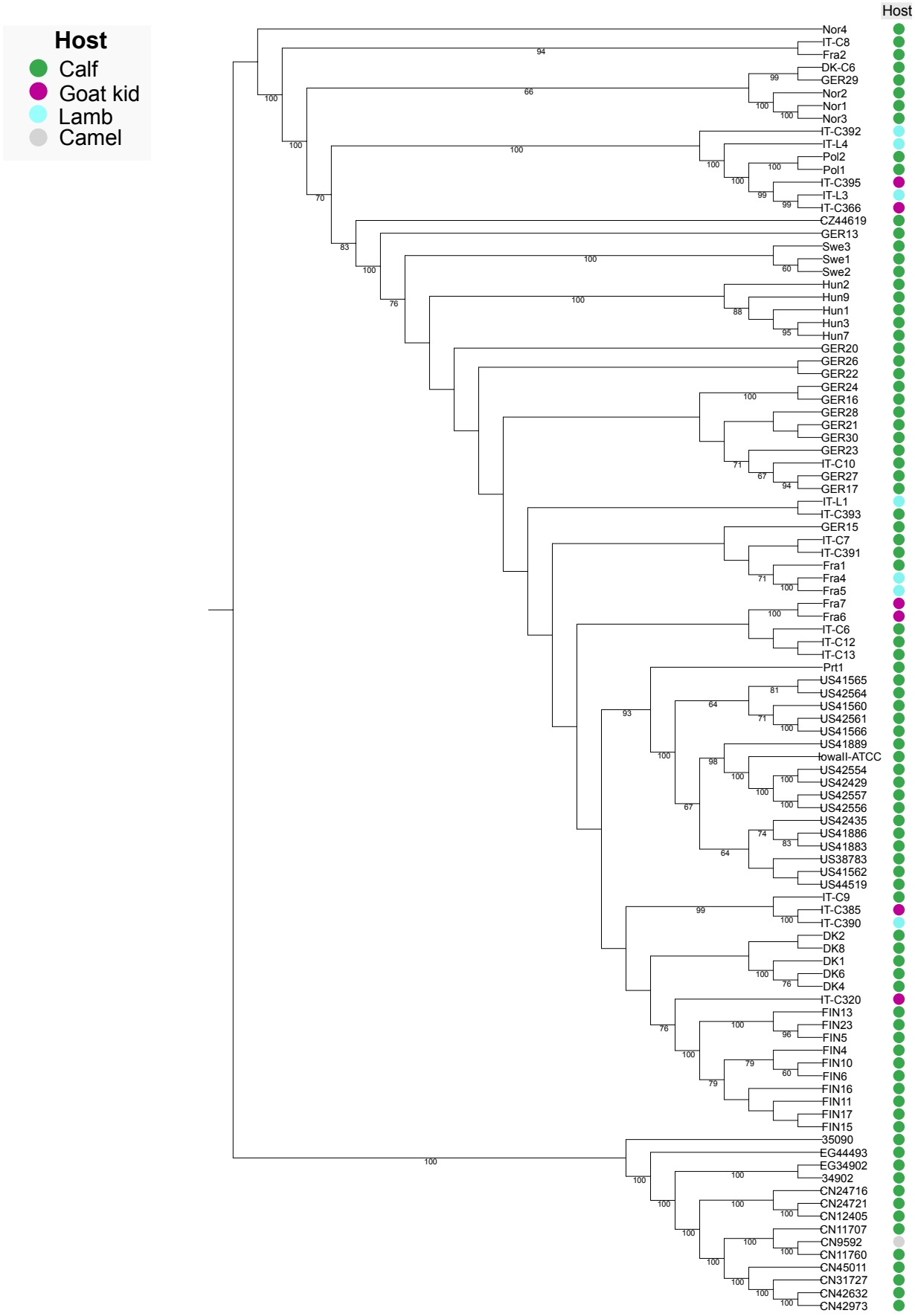


Figure S5. Maximum Likelihood tree of *C. parvum* isolates from ruminants only. Bootstrap values >60 are shown. Information about the host is mapped on the phylogeny.

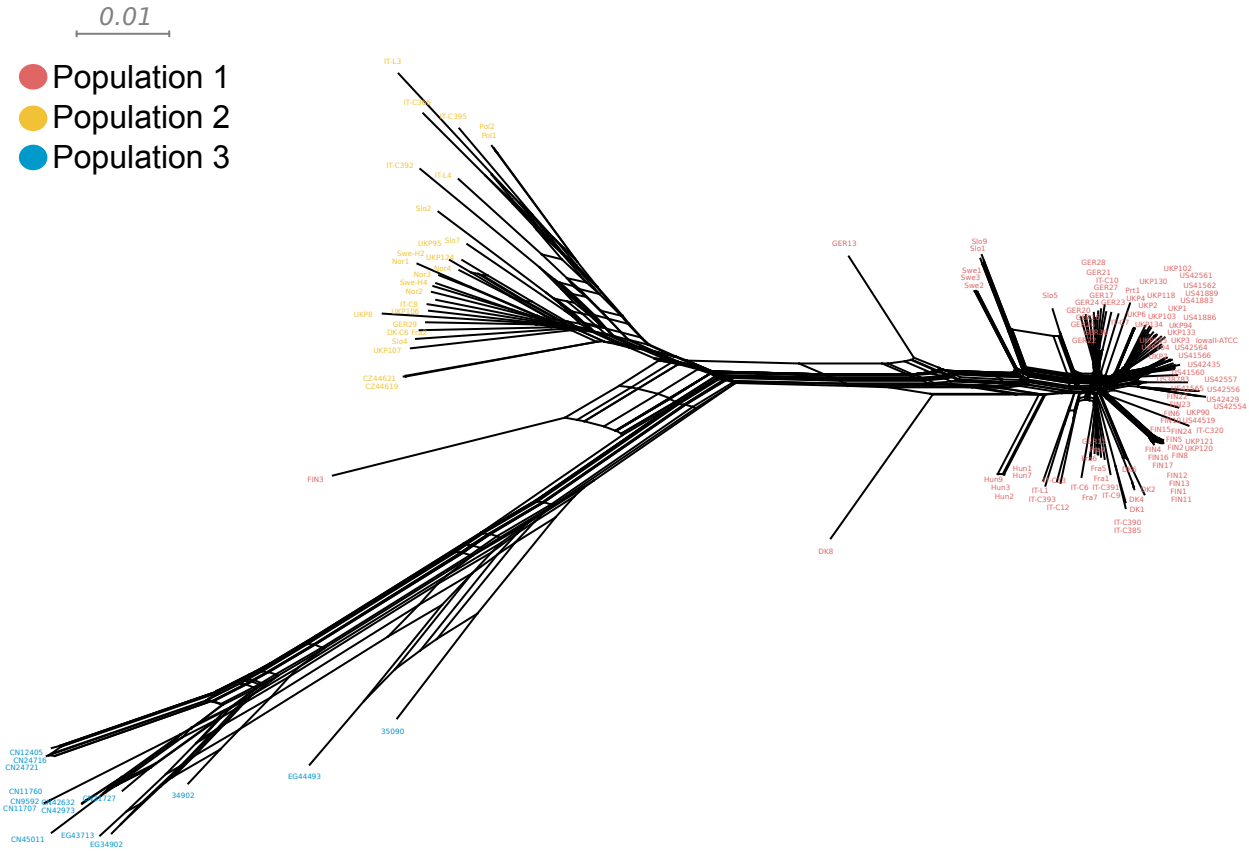
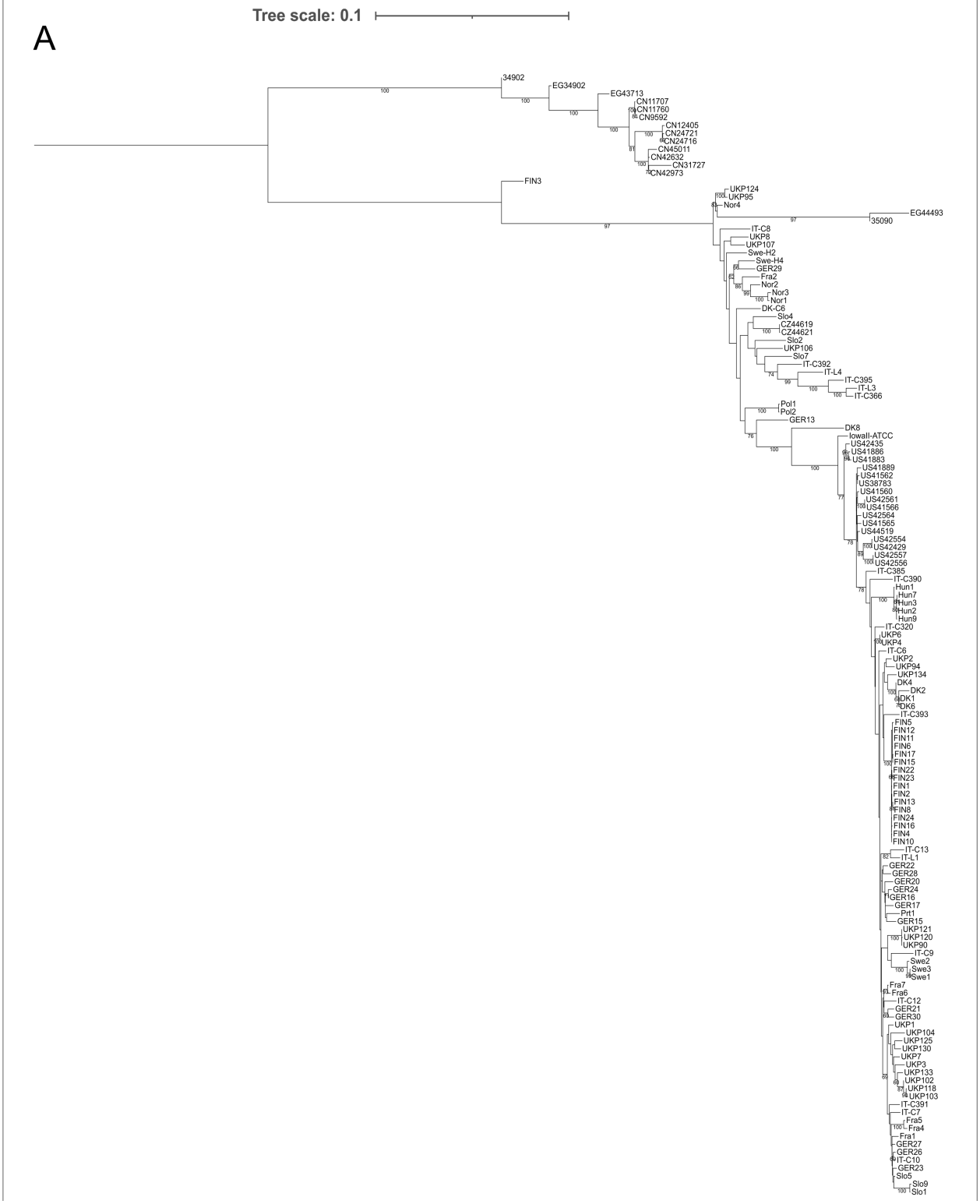


Figure S6. Phylogenetic network generated using SplitTree. The labels are color-coded according to the Population of each isolate.

Chromosome 1



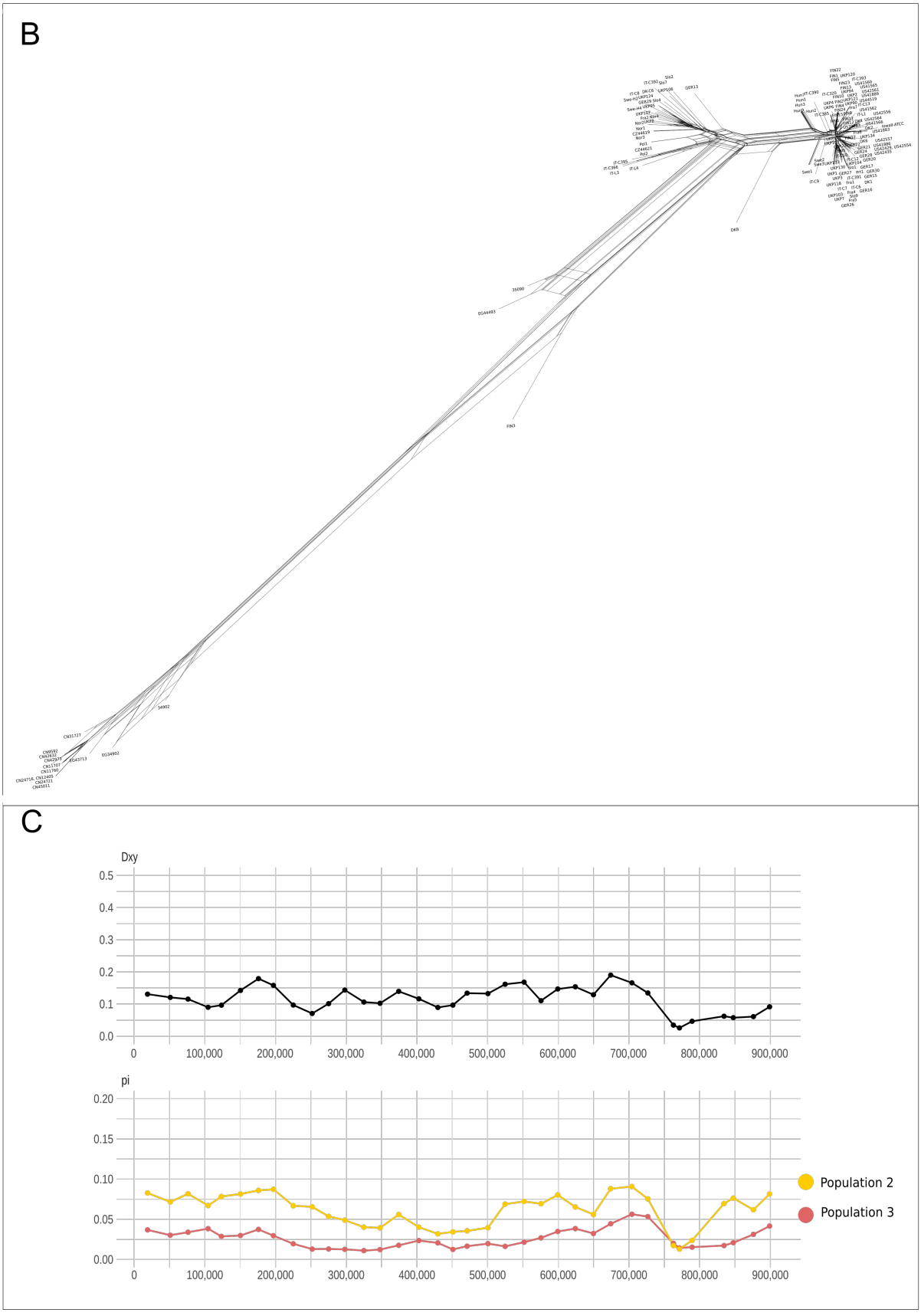


Figure S7. Analysis on SNPs located on chromosome 1: A) ML phylogenetic tree, B) Splitstree and C) Dxy and Pi.

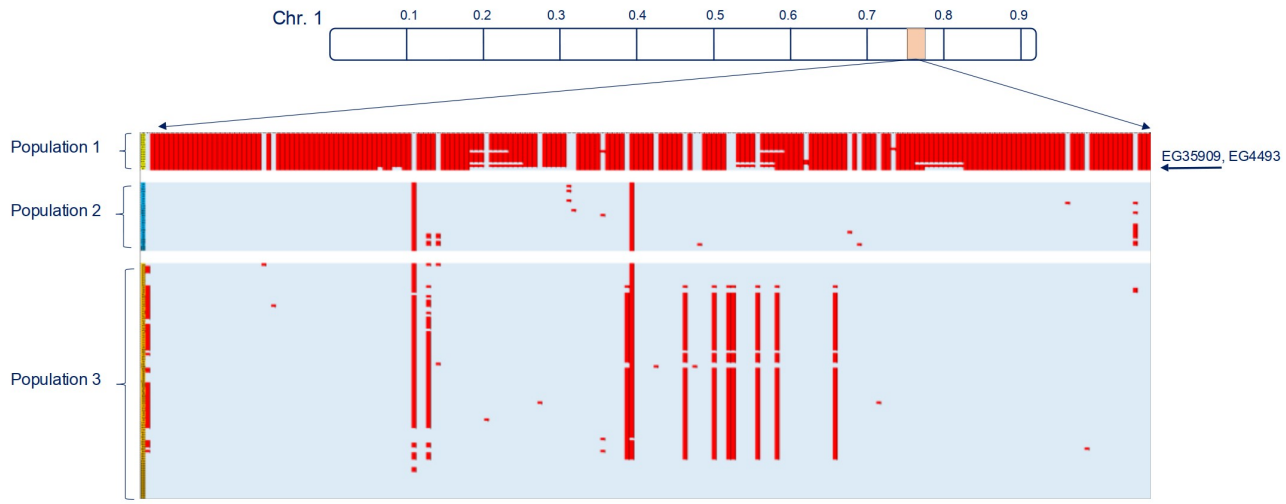


Figure S8. SNPs pattern in a 10 kb region of chromosome 1 spanning from position 755,934 to 768,672, highlighting the similarity of SNP distribution in isolates EG3509 and EG4493 (population 1) with isolates from the other two populations.

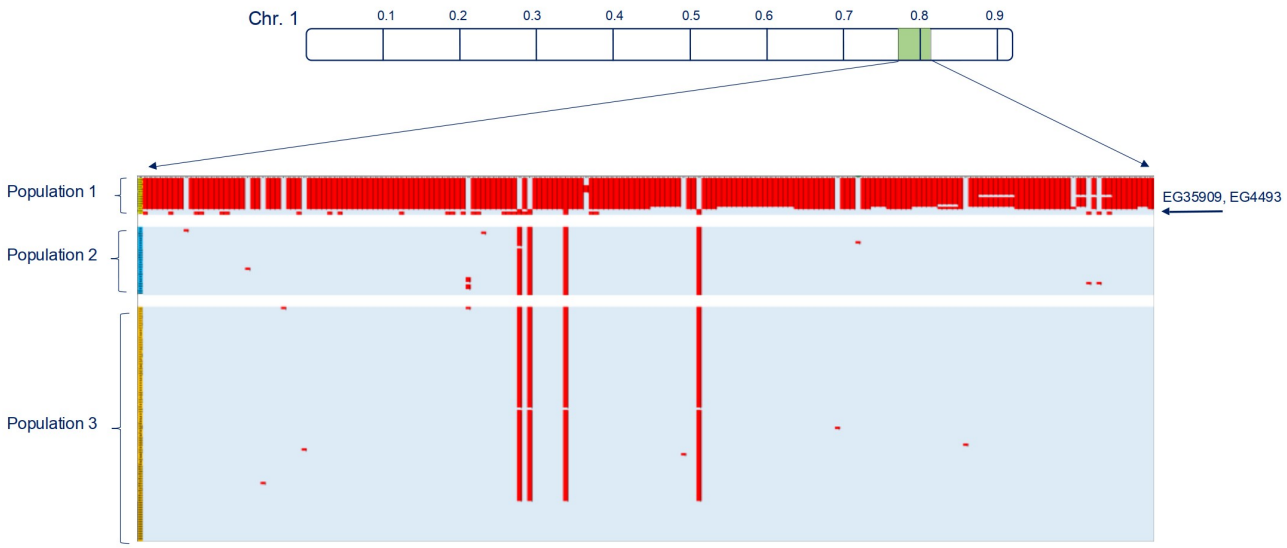


Figure S9. SNPs pattern in a 55 kb region of chromosome 1 spanning from position 768,729 to 823,729, highlighting the similarity of SNP distribution in isolates EG3509 and EG4493 (population 1) with isolates from the other two populations.

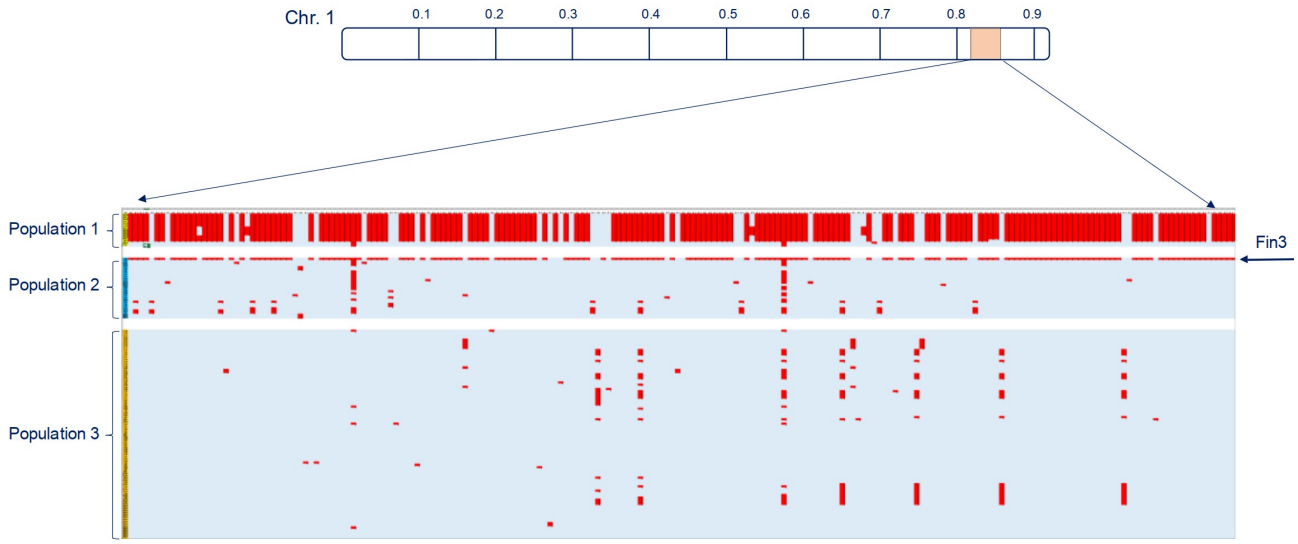
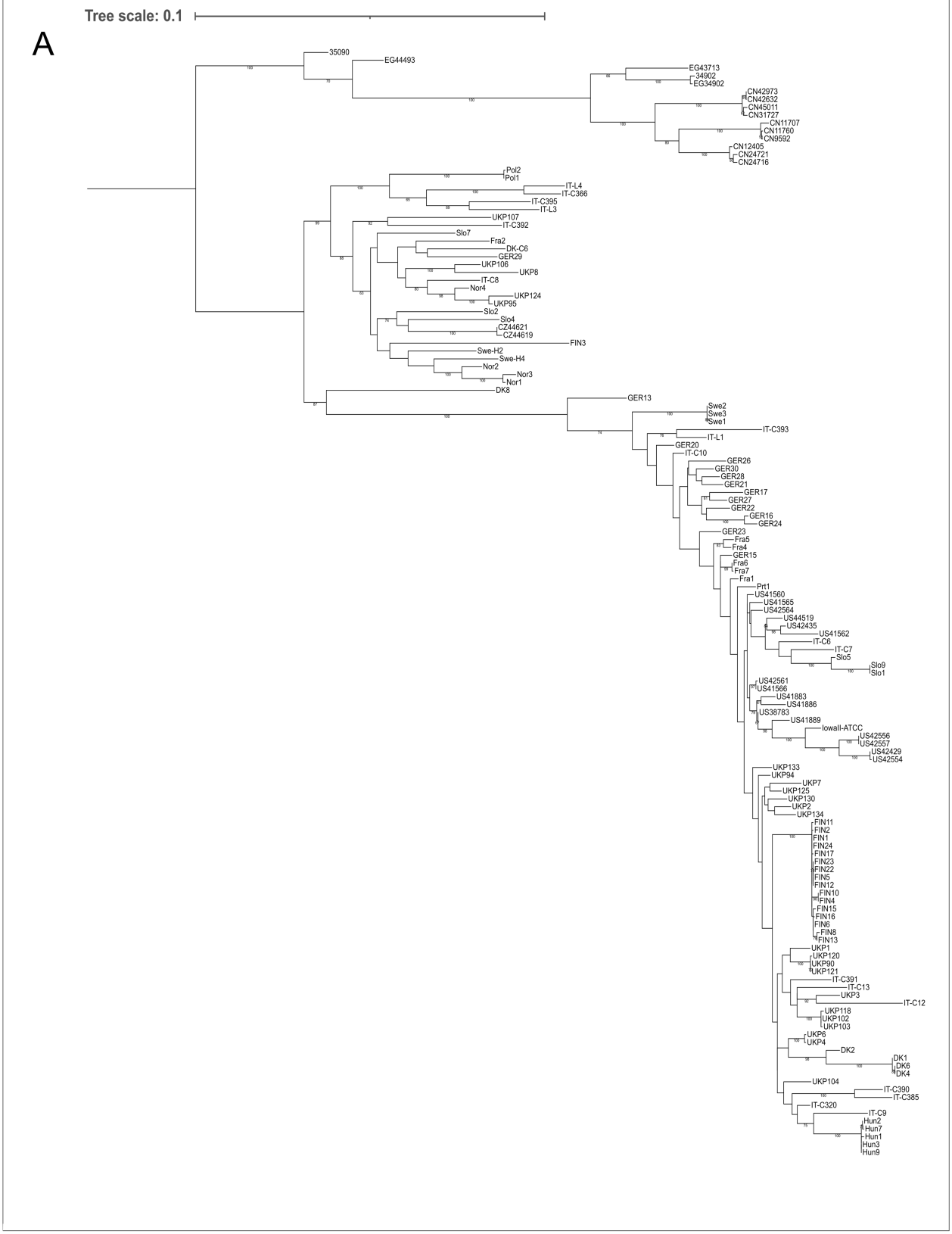


Figure S10. SNPs pattern in a 50 kb region of chromosome 1 spanning from position 824,800 to 874,170, highlighting the similarity of SNP distribution in isolate FIN3 (population 2) with isolates from population 1.

Chromosome 2



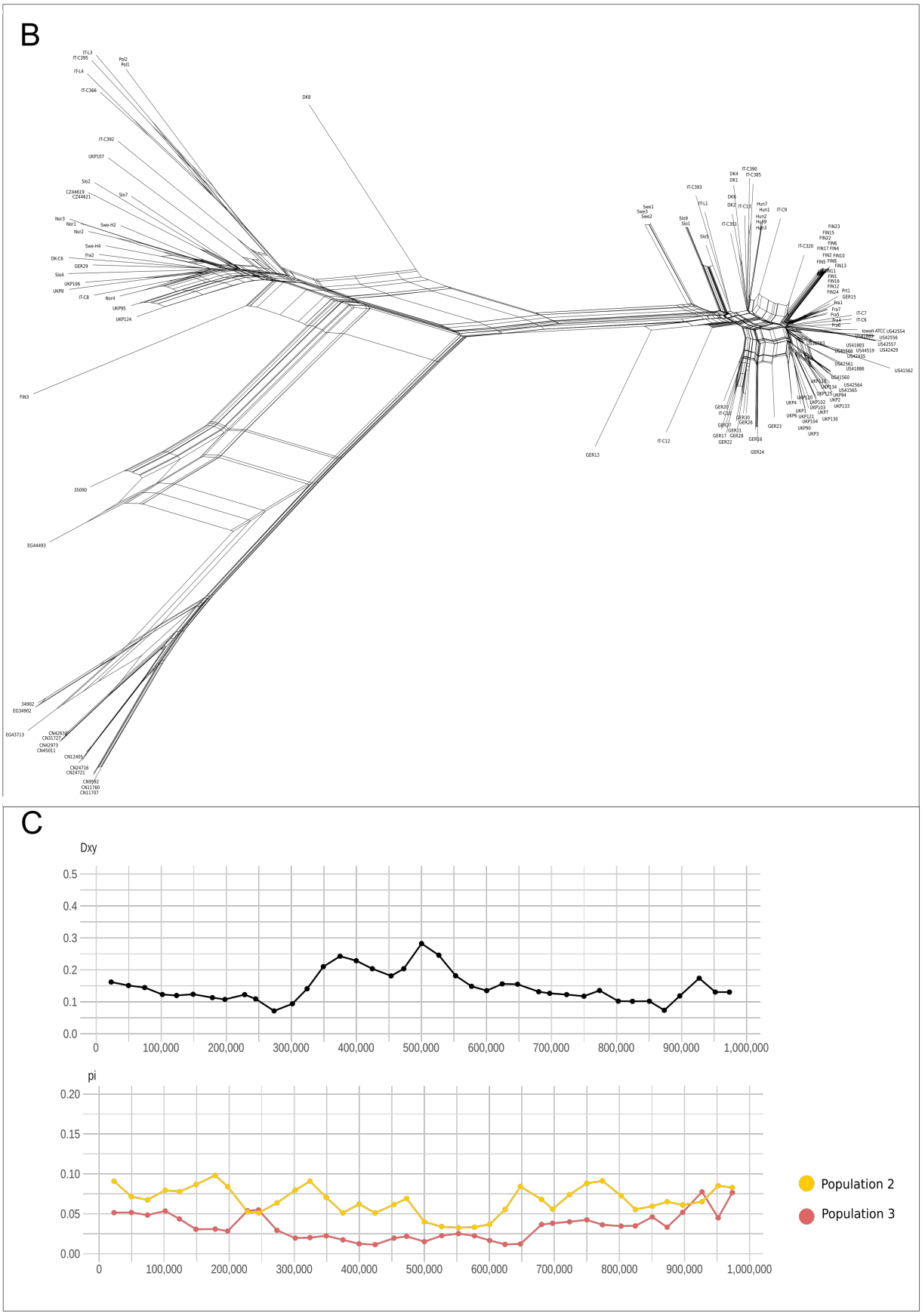


Figure S11. Analysis on SNPs located on chromosome 2: A) ML phylogenetic tree, B) Splitstree and C) Dxy and Pi.

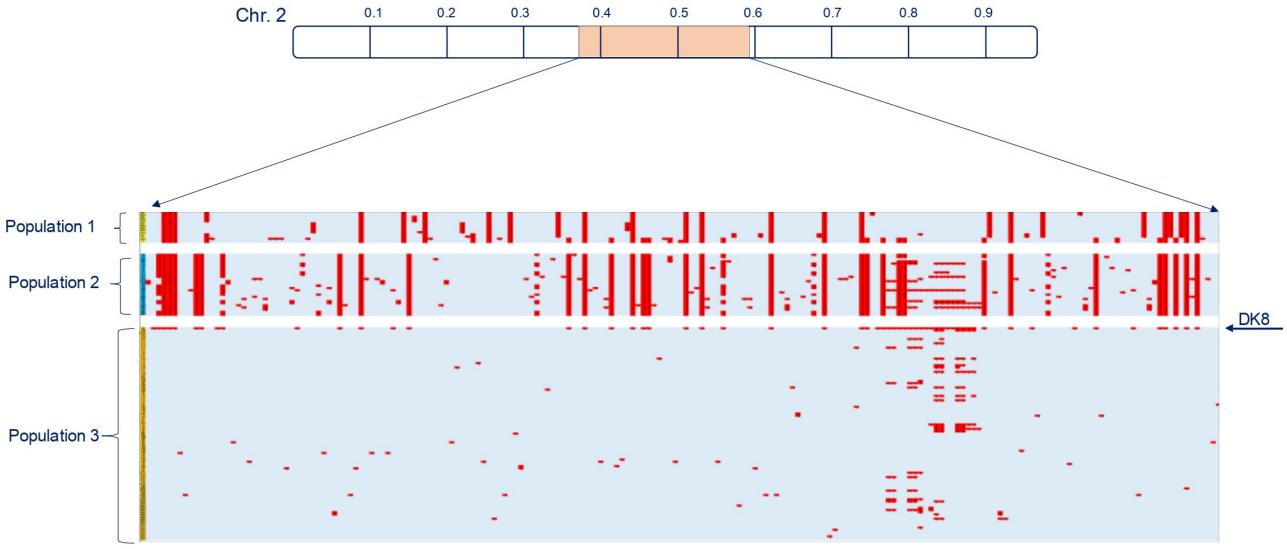
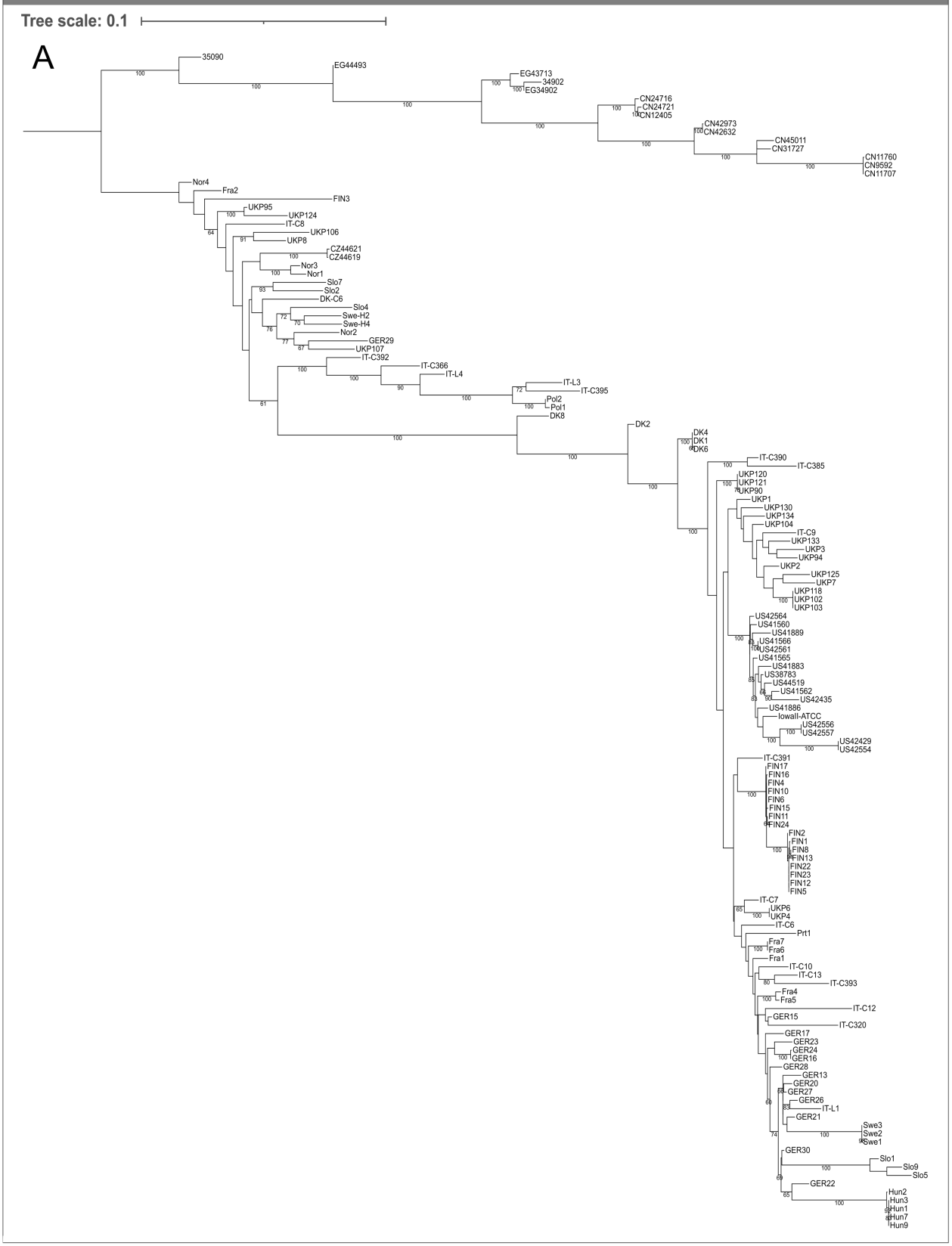


Figure S12. SNPs pattern in a 210 kb region on chromosome 2 spanning from position 384,000 to 594,000, highlighting the similarity of SNP distribution in isolate DK8 (population 3) with isolates from the other two populations.

Chromosome 3



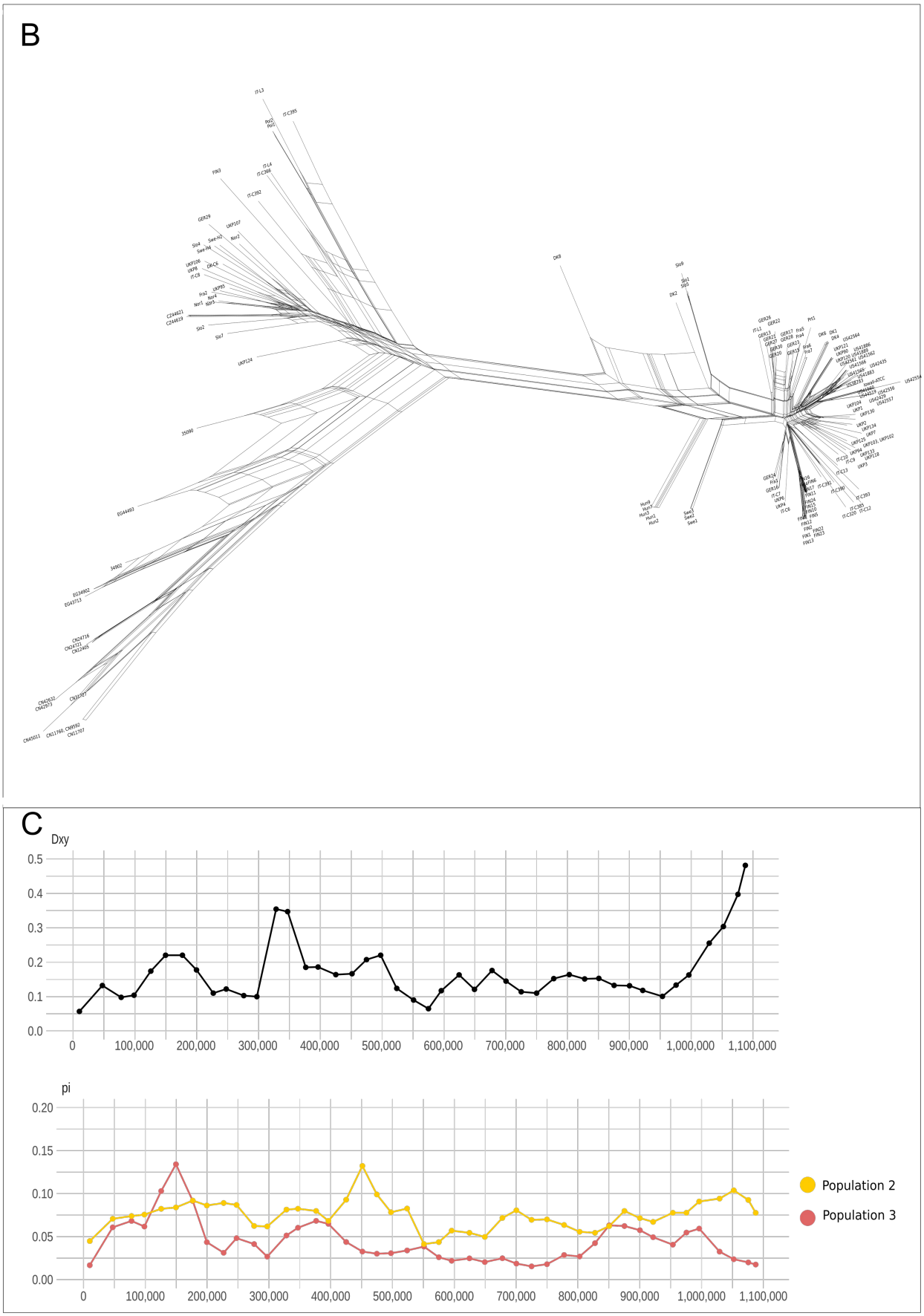
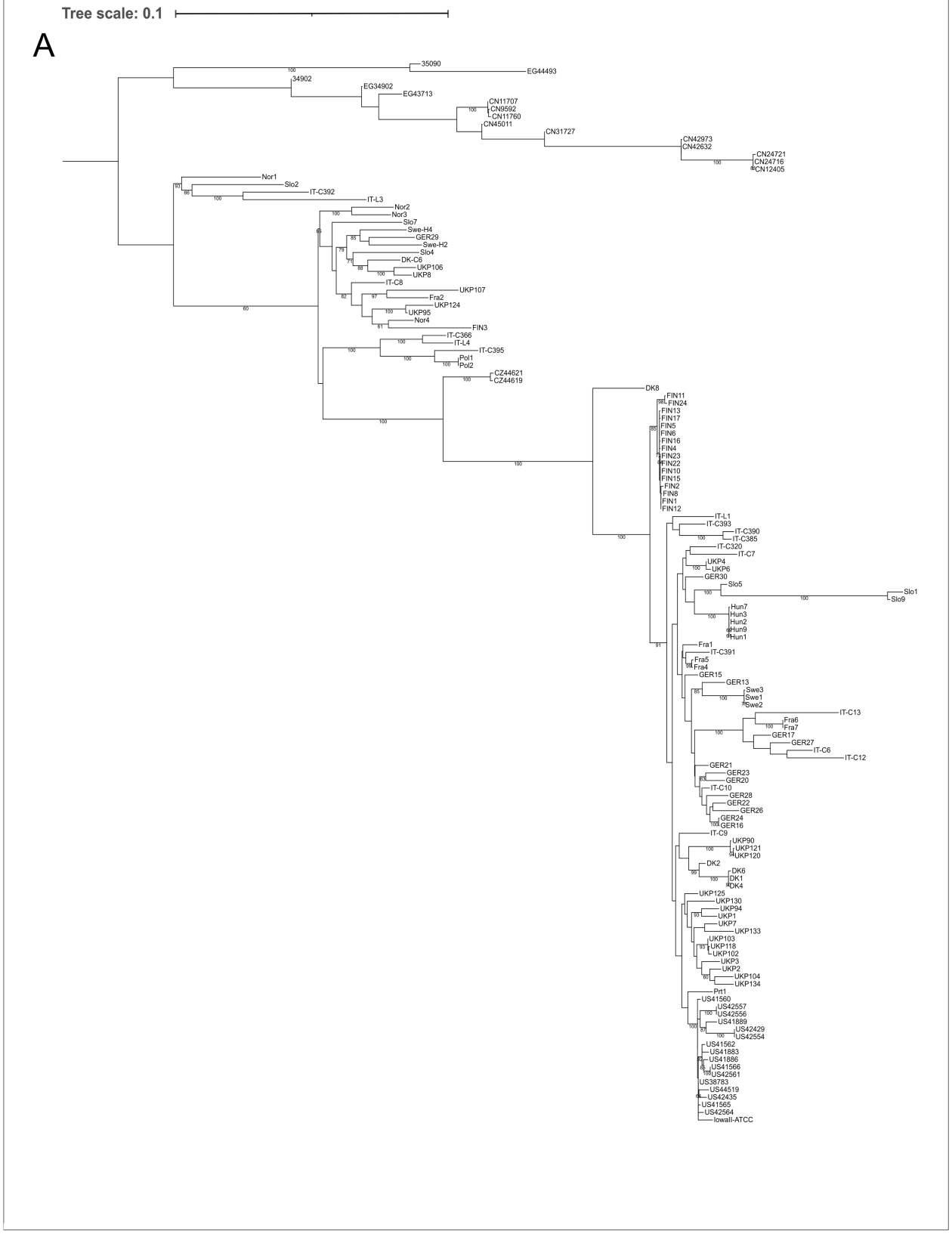


Figure S13. Analysis on SNPs located on chromosome 3: A) ML phylogenetic tree, B) Splitstree and C) Dxy and Pi.

Chromosome 4



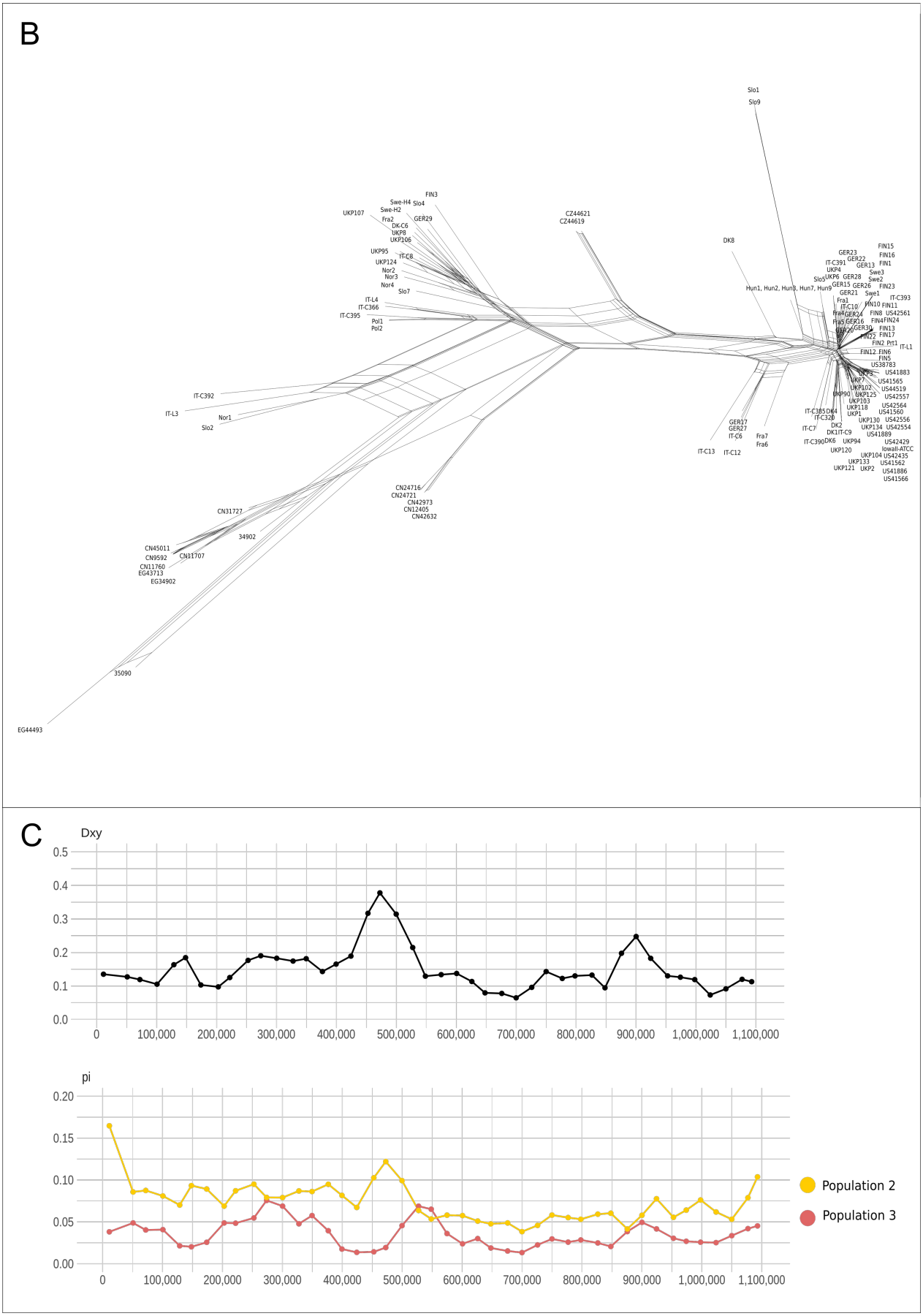


Figure S14. Analysis on SNPs located on chromosome 4: A) ML phylogenetic tree, B) Splitstree and C) Dxy and Pi.

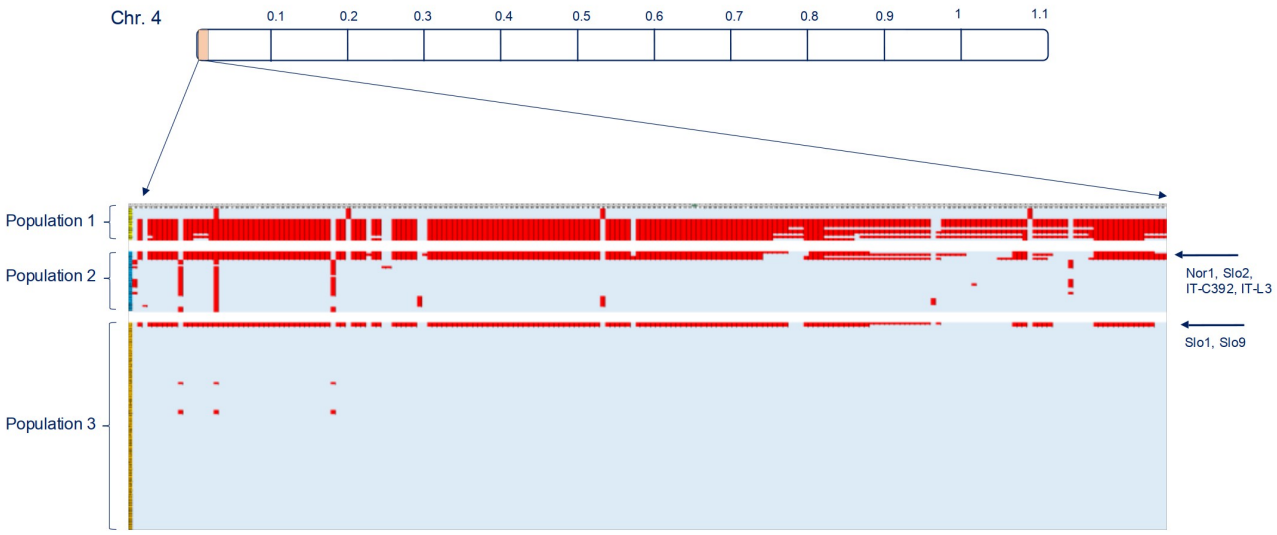
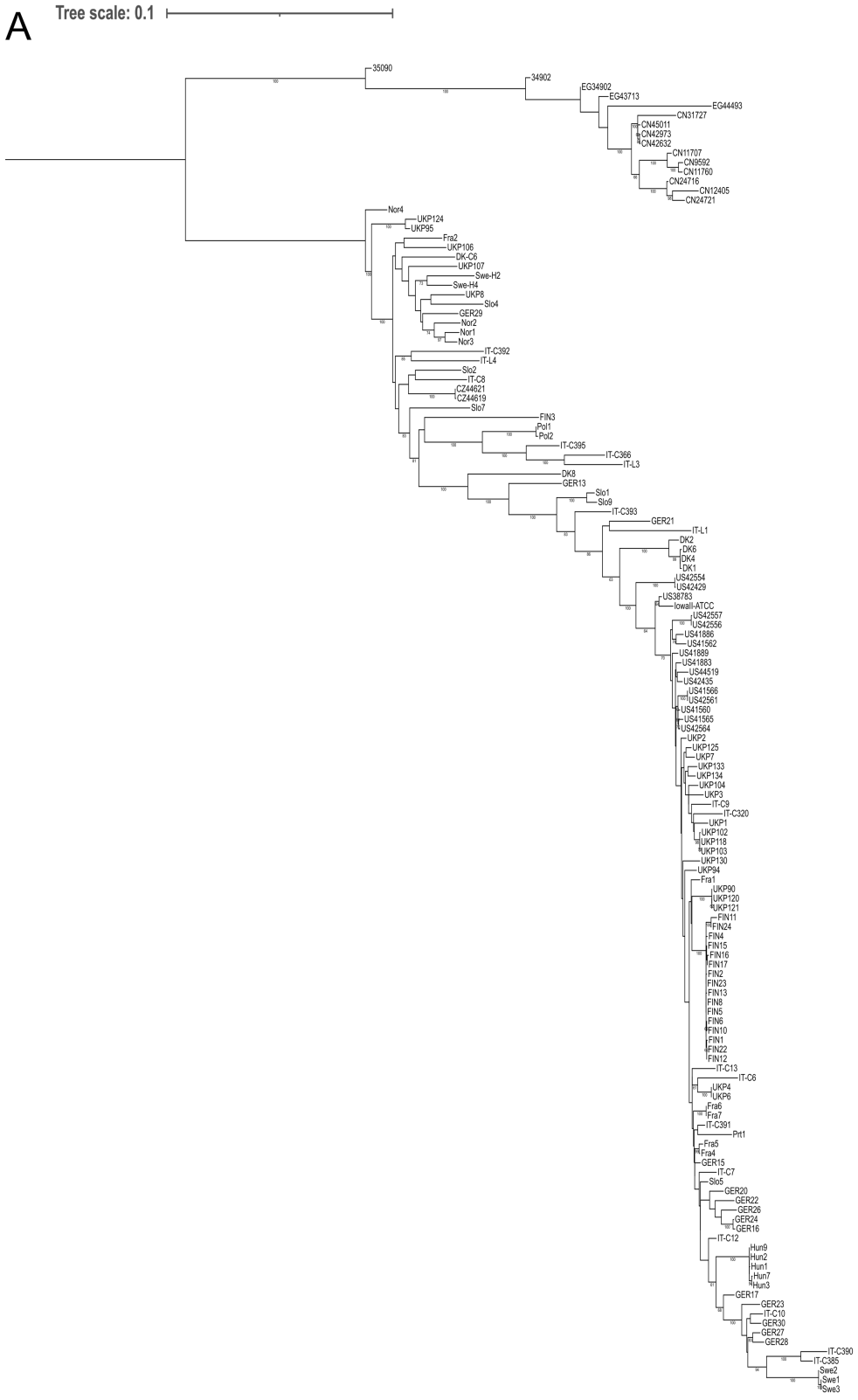


Figure S15. SNPs pattern in the first 8 kb adjacent to the 5' telomere of chromosome 4, highlighting the similarity of SNP distribution in isolates Nor1, Slo2, IT-C392 and IT-L38 (population 2), and isolates Slo2 and Slo9 (population 3), with isolates from population 1.

Chromosome 5



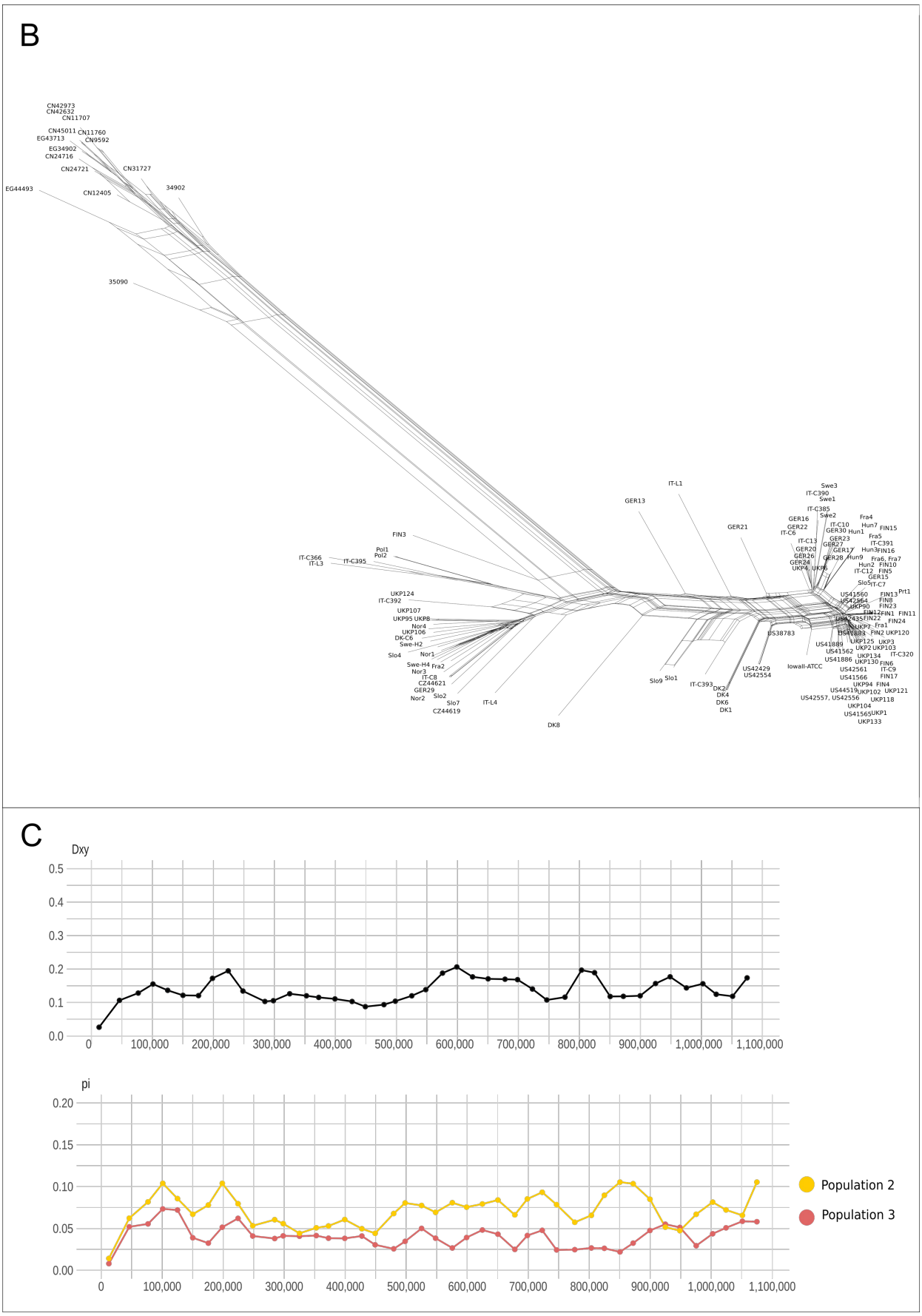
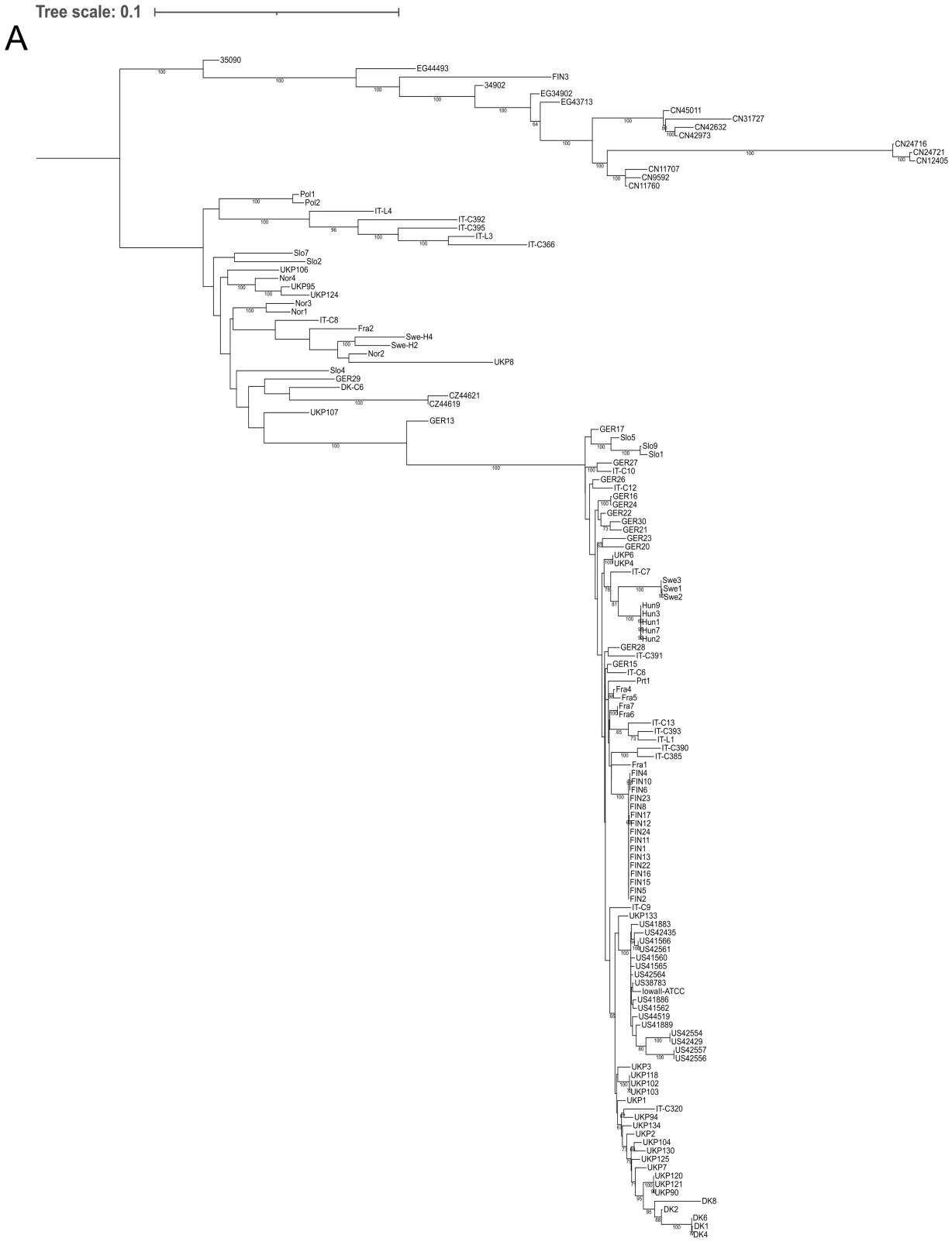


Figure S16. Analysis on SNPs located on chromosome 5: A) ML phylogenetic tree, B) Splitstree and C) Dxy and Pi.

Chromosome 6



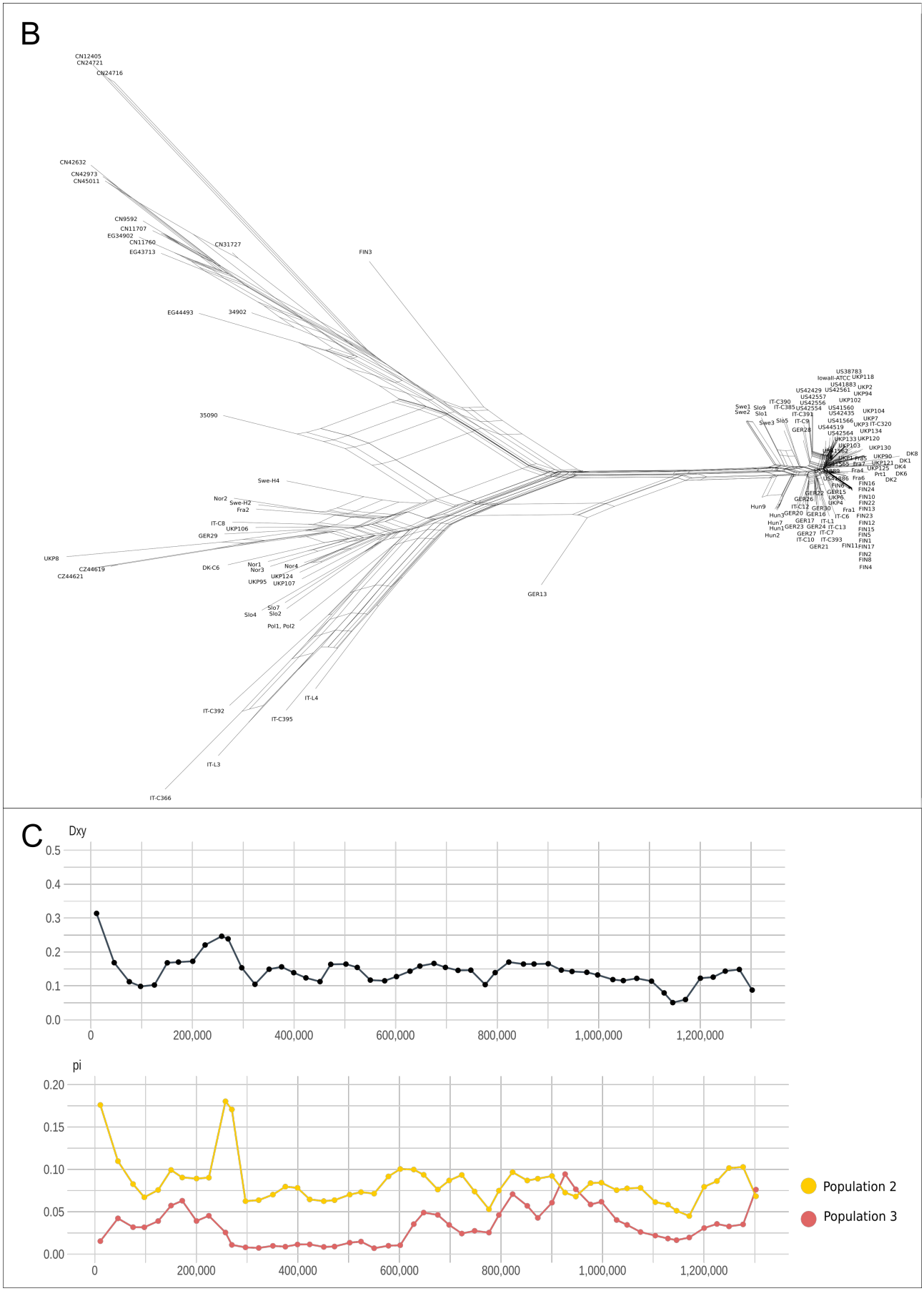


Figure S17. Analysis on SNPs located on chromosome 6: A) ML phylogenetic tree, B) Splitstree and C) Dxy and Pi.

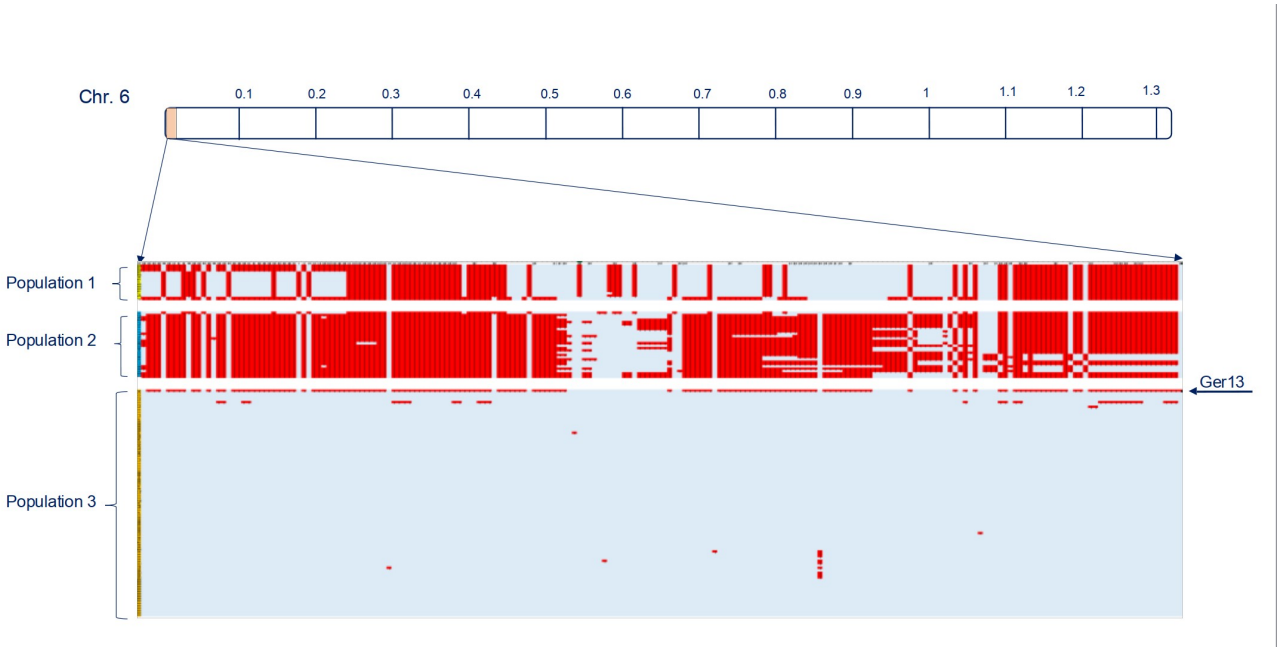


Figure S18. SNPs pattern in the first 18 kb adjacent to the 5' telomere of chromosome 6, highlighting the similarity of SNP distribution in isolate Ger13 (population 3) with isolates from the other two populations.

Chromosome 7

Tree scale: 0.01

A

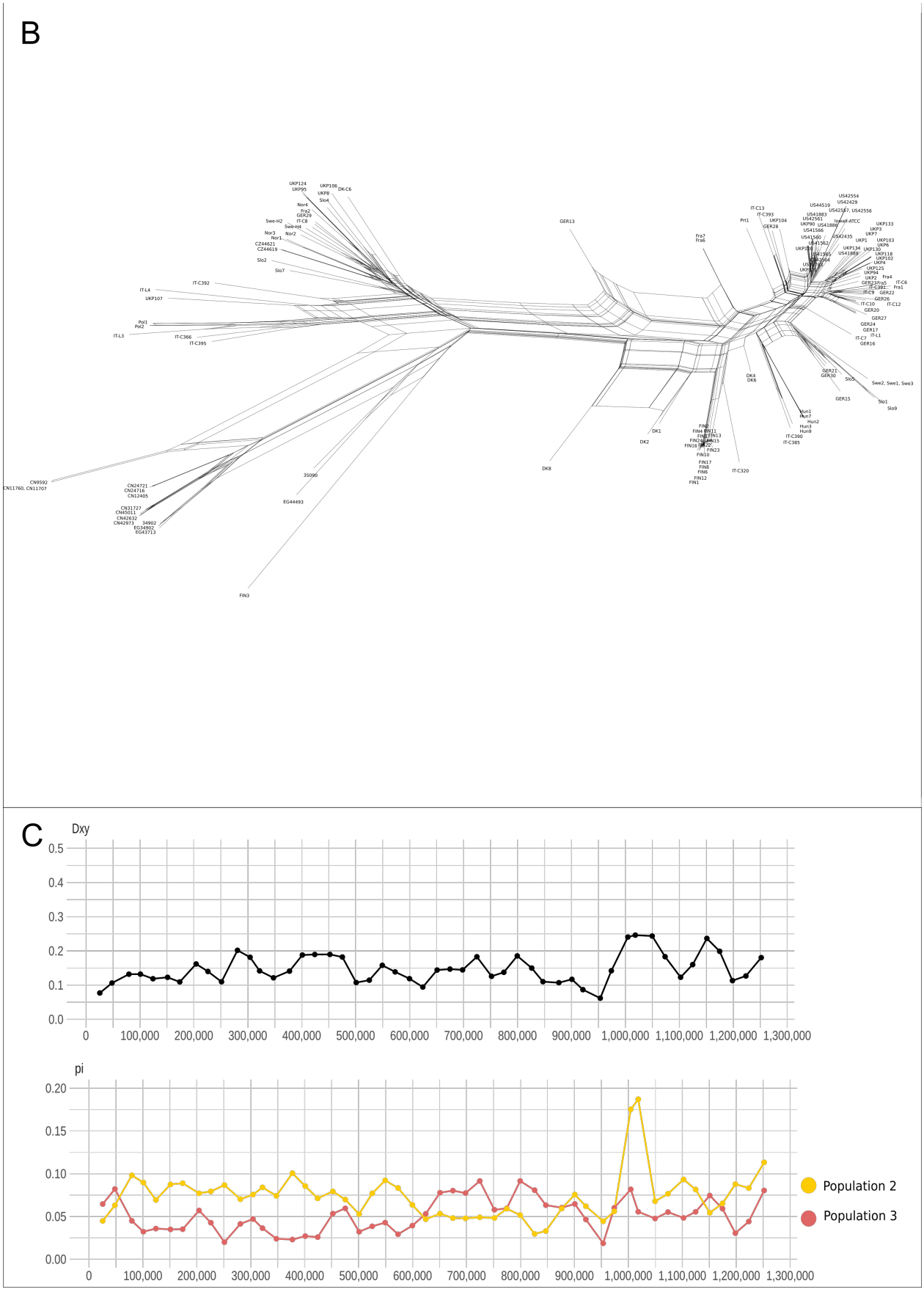
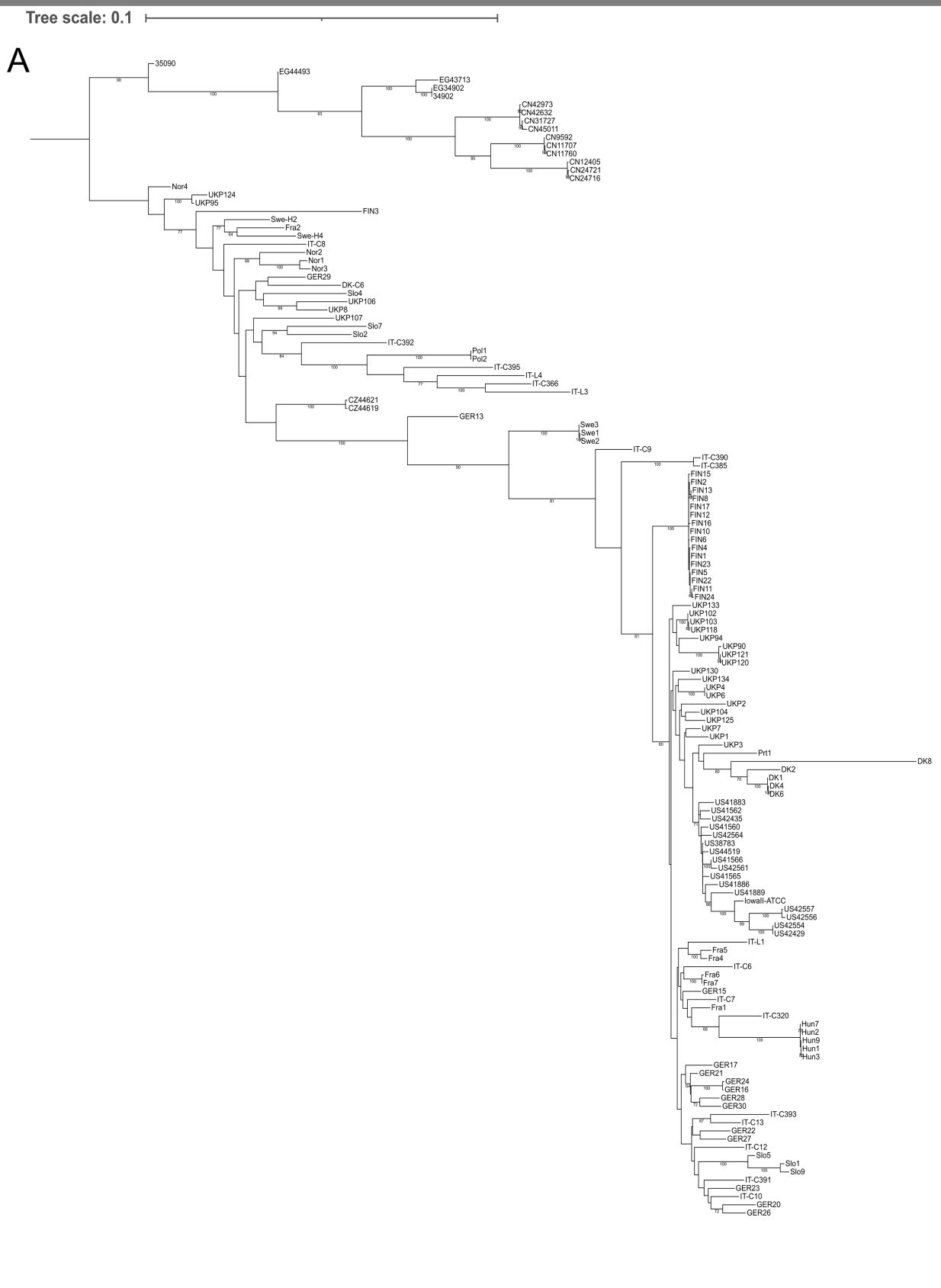


Figure S19. Analysis on SNPs located on chromosome 7: A) ML phylogenetic tree, B) Splitstree and C) Dxy and Pi.

Chromosome 8



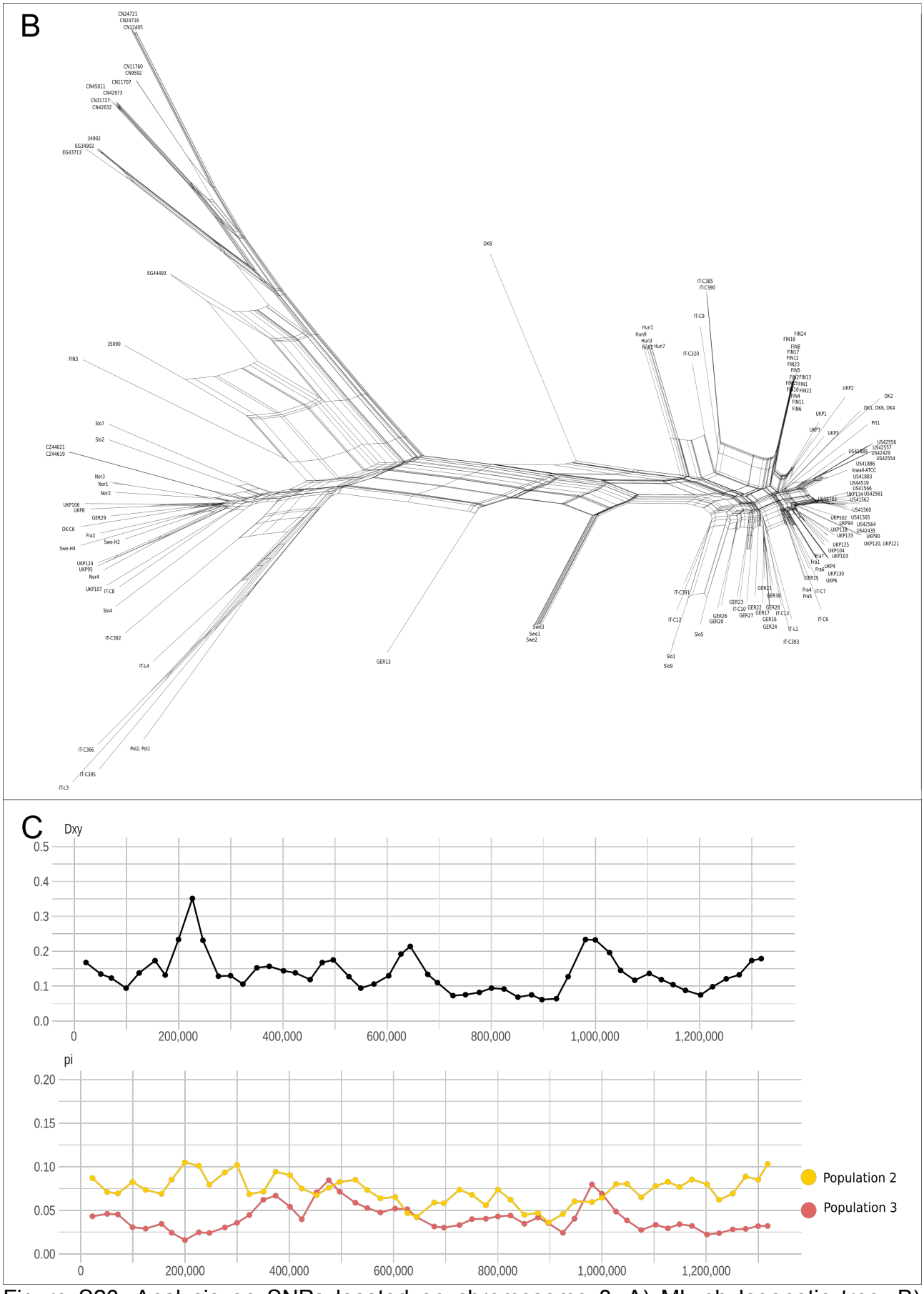


Figure S20. Analysis on SNPs located on chromosome 8: A) ML phylogenetic tree, B) Splitstree and C) Dxy and Pi.

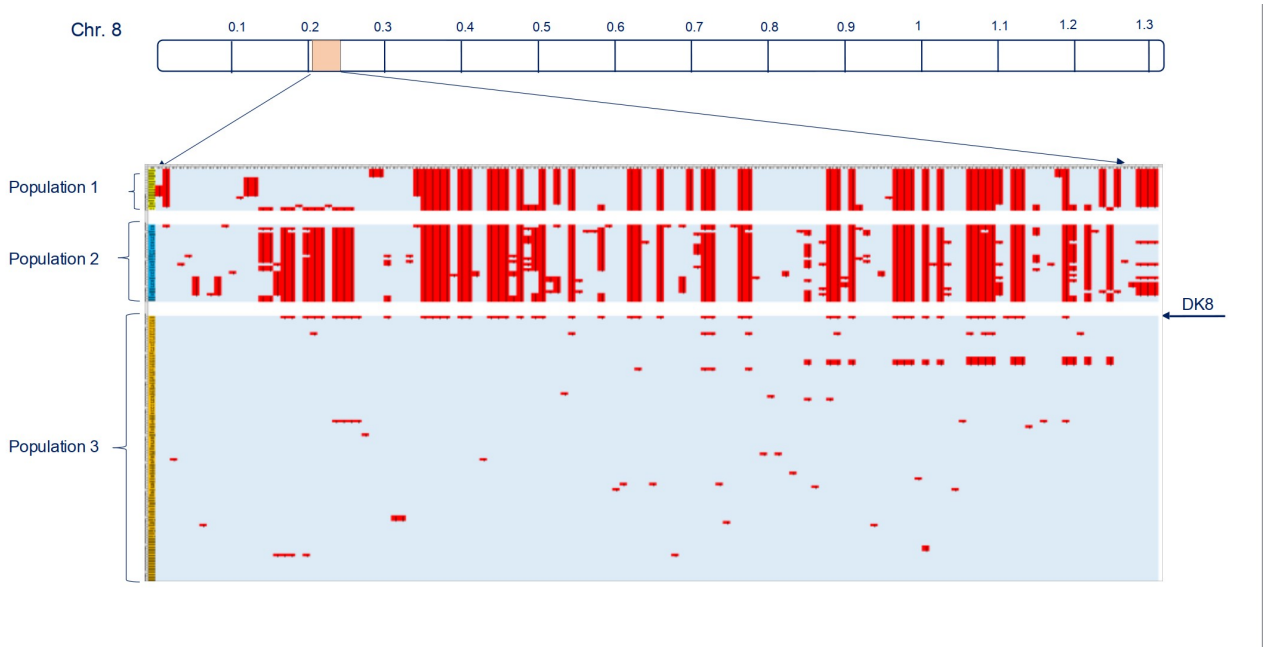


Figure S21. SNPs pattern in a 30 kb region spanning position 210,000 to 240,000 on chromosome 8, highlighting the similarity of SNP distribution in isolate DK8 (population 3) with isolates from the other two populations.S

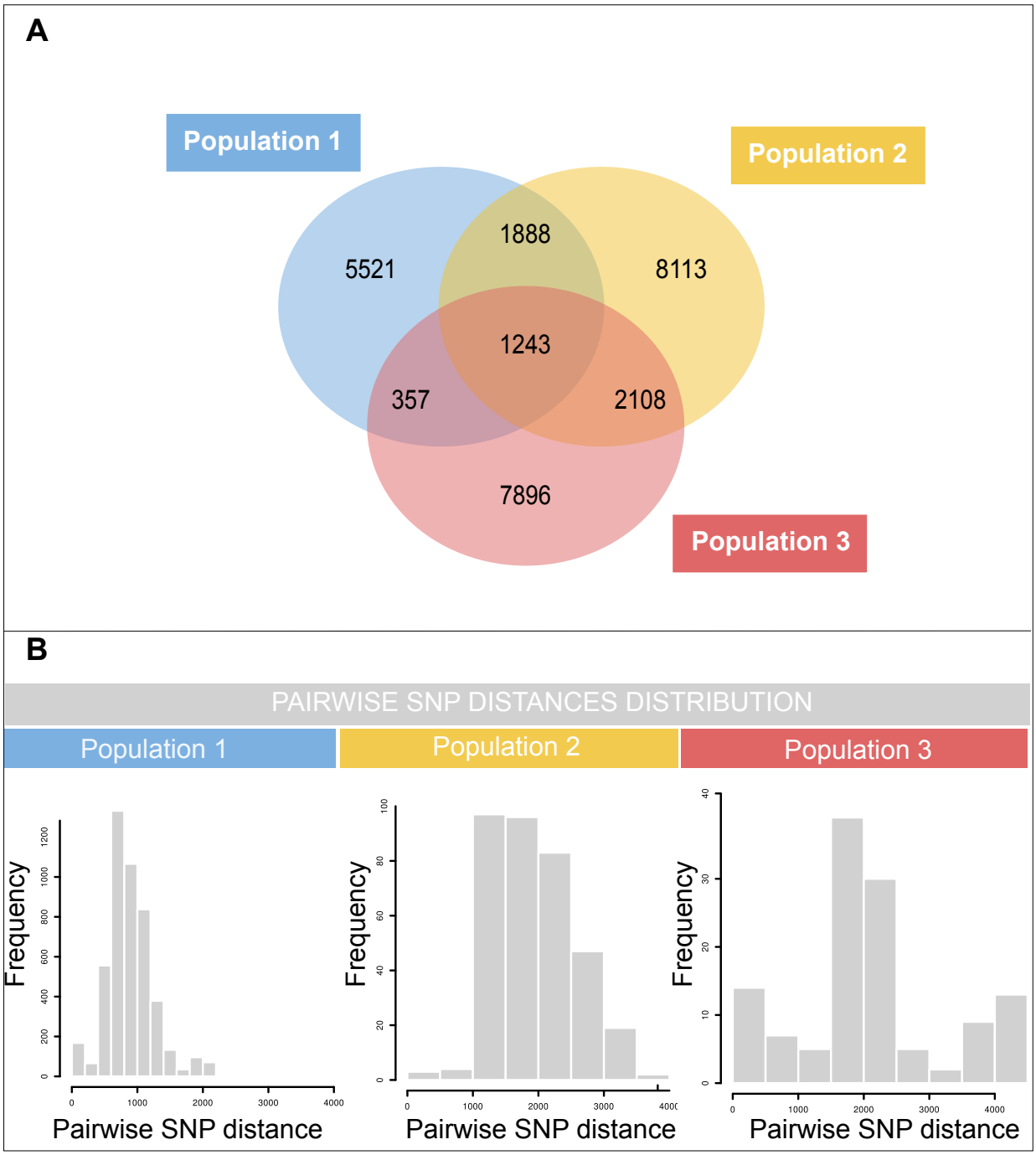


Figure S22. Distribution of shared and population-specific SNPs. A) Venn Diagram B) Histogram of the pairwise SNP distances distribution within each Population.

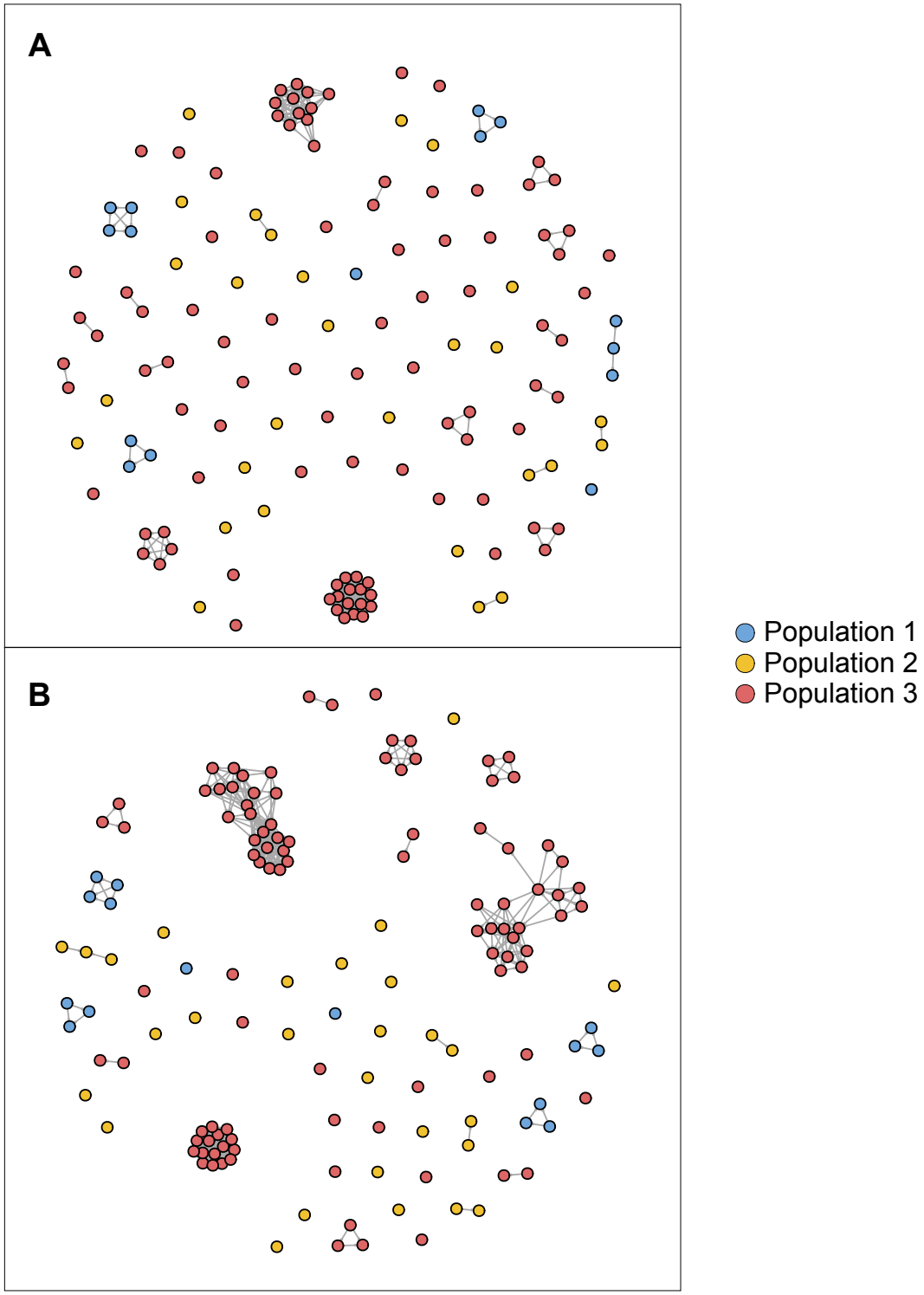


Figure S23. Relatedness network for pairs of isolates identified as having high proportions of IBD sharing. Each node identifies a unique isolate and an edge is drawn between two isolates if they share more than A) 90% of their genome IBD, and B) 80% of their genome IBD. Individuals are coloured according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

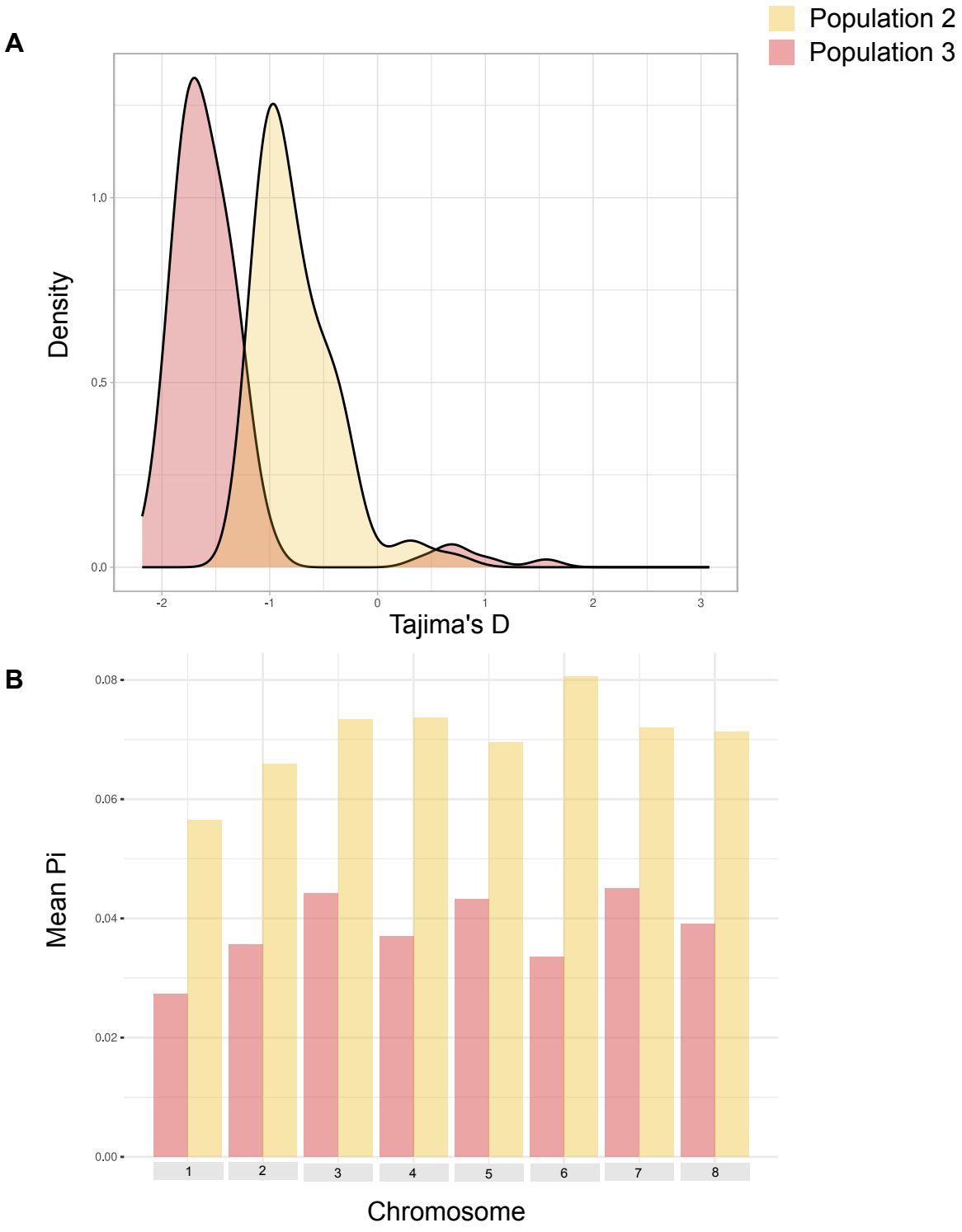


Figure S24. Distribution of A) Tajima's D values and B) nucleotide diversity (π) in population 2 and population 3. In these analyses, only one isolate per each of the 13 clusters of highly related genomes (<50 SNPs) within population 2 was used.