Efficiently integrating ancient DNA into modern Y-chromosome trees

Rui Martiniano[1,2], Lara M. Cassidy[3], Daniel G. Bradley[3], Richard Durbin[1,2]

1 - Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH, UK.
2 - Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge, CB10 1SA, UK.
3 - Smurfit Institute of Genetics, School of Genetics and Microbiology, Trinity College Dublin, Dublin, Dublin 2, Ireland.

UNIVERSITY OF CAMBRIDGE
wellcome trust sanger institute
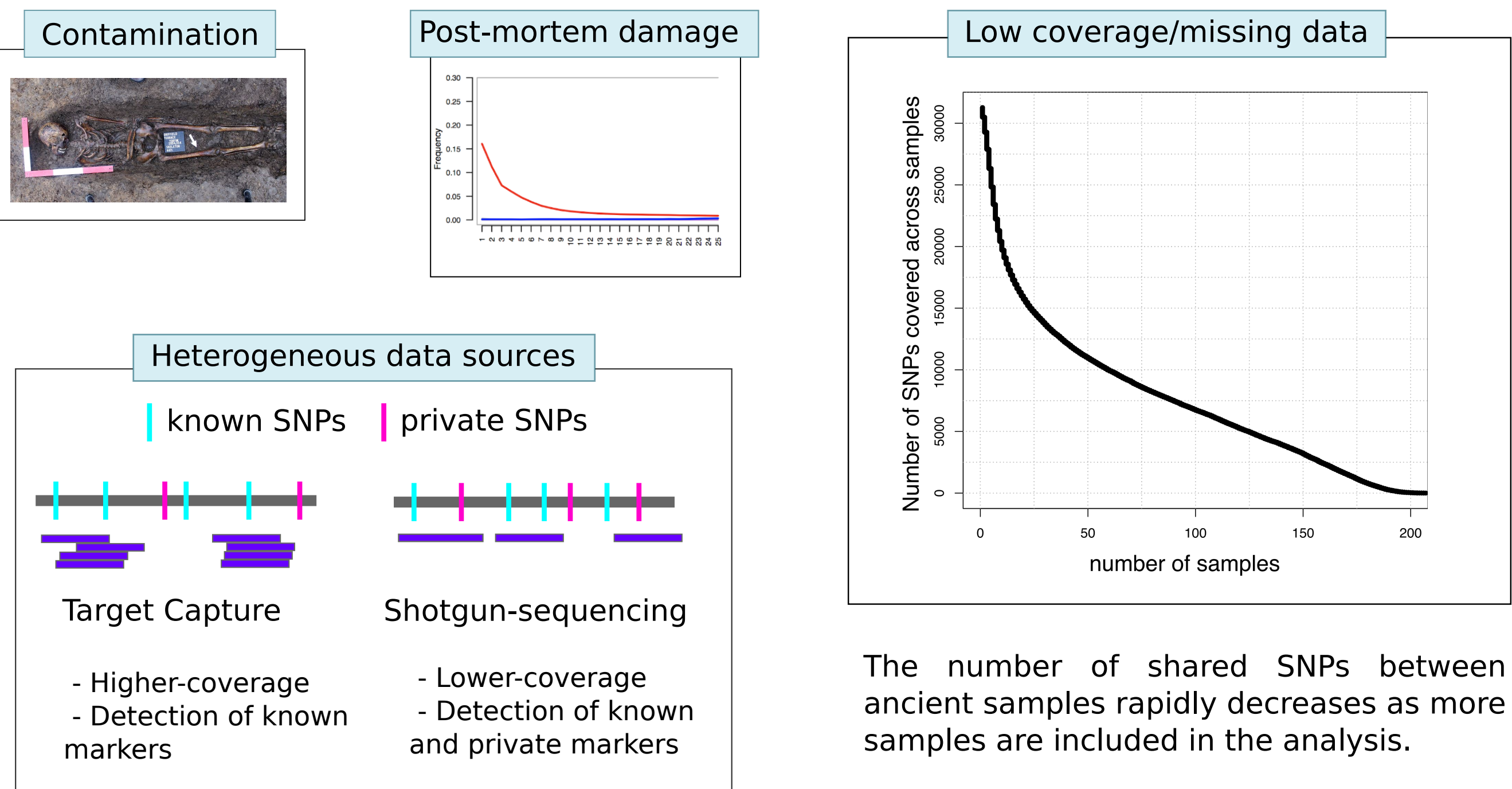Trinity College Dublin
The University of Dublin

During the last decade, a huge wealth of human ancient Y chromosome data has been generated as part of whole-genome shotgun and target capture sequencing studies. However, given the highly degraded nature of ancient DNA (aDNA) data, post-mortem deamination and often low genomic coverage, combining ancient and modern samples for phylogenetic analyses remains challenging. Most analyses use limited markers and/or extensive manual curation.

Current standard methods for the analysis of Y chromosome data focus on known, gold-standard markers, but these contain only a subset of the total Y chromosomal variation. Examining all polymorphic markers is particularly useful when dealing with low coverage aDNA data because it substantially increases the number of overlapping sites between present-day and ancient individuals and it may also help uncover relationships inaccessible via standard known variation.

We provide an automated workflow for jointly analysing ancient and present-day sequence data using all uniquely mappable regions of the Y chromosome. From a given high-coverage dataset, a maximum likelihood phylogeny is estimated and variants are assigned to each branch of the tree. Next, for each low coverage aDNA sample, we count the number of ancestral and derived alleles at each branch and use this information to map ancient lineages to their most likely place in the phylogeny. We apply this method to a large dataset of novel and publicly available data from ancient Eurasians and characterize patterns of Y chromosomal diversity across time as well as the impact of past migrations on the landscape of present-day paternal lineage distribution.

Background

Joint analysis of ancient and present-day genetic data is greatly complicated by characteristics inherent to ancient DNA.

Contamination

Post-mortem damage

Low coverage/missing data

Heterogeneous data sources

known SNPs   private SNPs

Target Capture
- Higher-coverage
- Detection of known markers

Shotgun-sequencing
- Lower-coverage
- Detection of known and private markers

The number of shared SNPs between ancient samples rapidly decreases as more samples are included in the analysis.
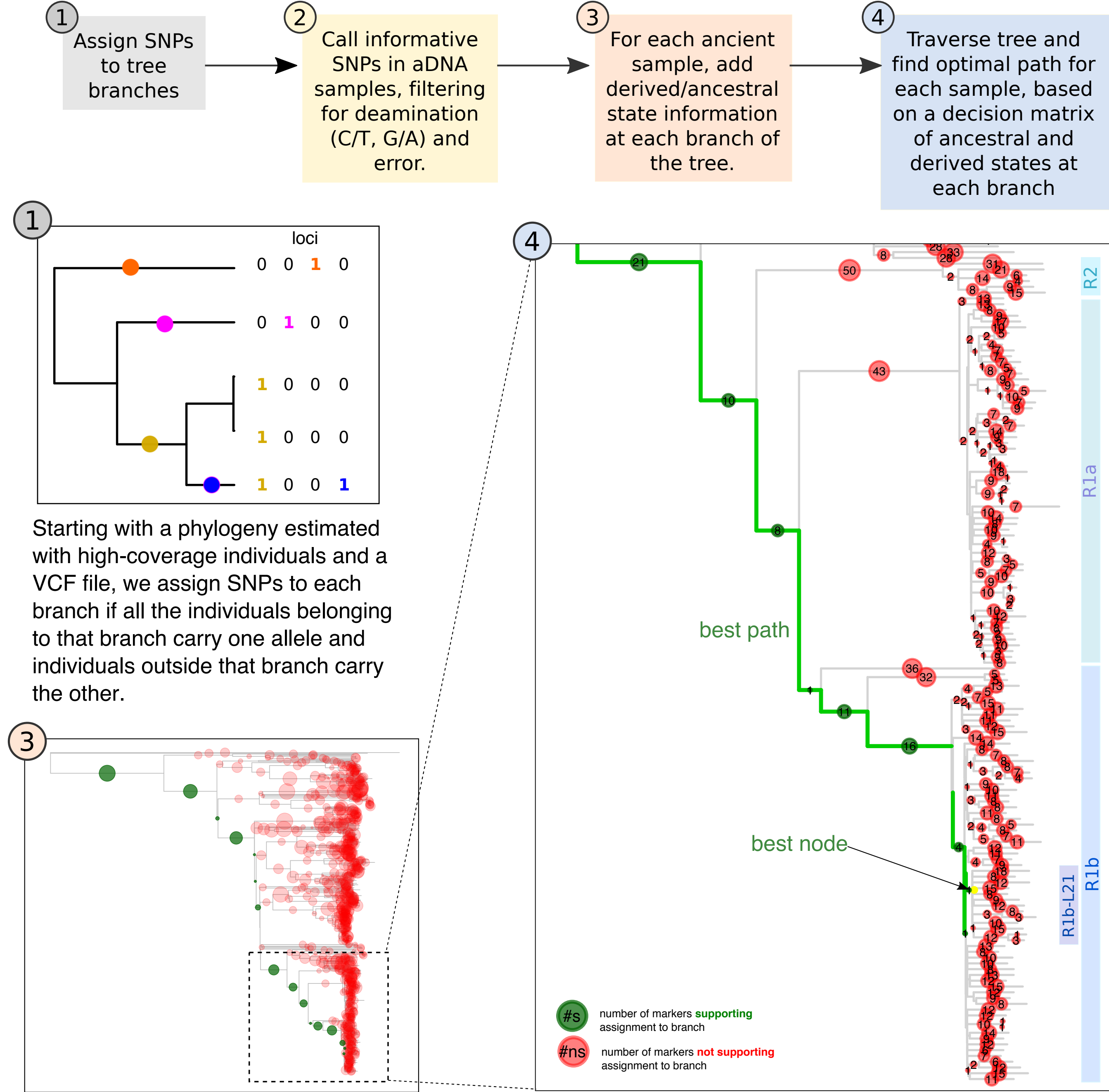
These aspects, especiially low genomic coverage, make estimating accurate Y-chromosome phylogenies with ancient and modern data using standard methods very challenging. Furthermore, ancient DNA Y-chromosome analysis restricts itself to a set of known, curated SNPs used for the identification of Y-chromosome lineages, however, many more variants exist and will continuously be generated in present-day datasets, which can be used to inform aDNA paternal affinities.

Aims

1) Integrate ancient DNA data from multiple sources (target capture and shotgun sequencing) into present-day phylogenies, which is challenging with current available methods.

2) Use all available Y-chromosomal variants in a given high-coverage dataset rather than a subset of known SNPs, and in this way provide a greater chance of detecting informative SNPs in low coverage aDNA samples;

3) Provide an interactive visualization of Y-chromosome phylogenies which will enable exploring ancient and present day lineage variability across time and geography.

Workflow

① Assign SNPs to tree branches
② Call informative SNPs in aDNA samples, filtering for deamination (C/T, G/A) and error.
③ For each ancient sample, add derived/ancestral state information at each branch of the tree.
④ Traverse tree and find optimal path for each sample, based on a decision matrix of ancestral and derived states at each branch

① loci

Starting with a phylogeny estimated with high-coverage individuals and a VCF file, we assign SNPs to each branch if all the individuals belonging to that branch carry one allele and individuals outside that branch carry the other.

③

④

best path

best node

#s   number of markers supporting assignment to branch
#ns   number of markers not supporting assignment to branch

All possible paths in the tree are traversed using a modified decision matrix based on the number of ancestral and derived markers (first implemented in Poznik, 2016) which can easily be adjusted according to the needs of the user. The path containing the highest number of supporting markers and lowest number of not supporting markers is chosen.

Interactive visualization of modern and ancient Y-chromosome lineages

Ancient and presesent-day lineages and their geographical distribution can be visualized interactively with microreact, selecting specific lineages, time-periods and populations.

Copper and Bronze Age

Bell Beakers Britain
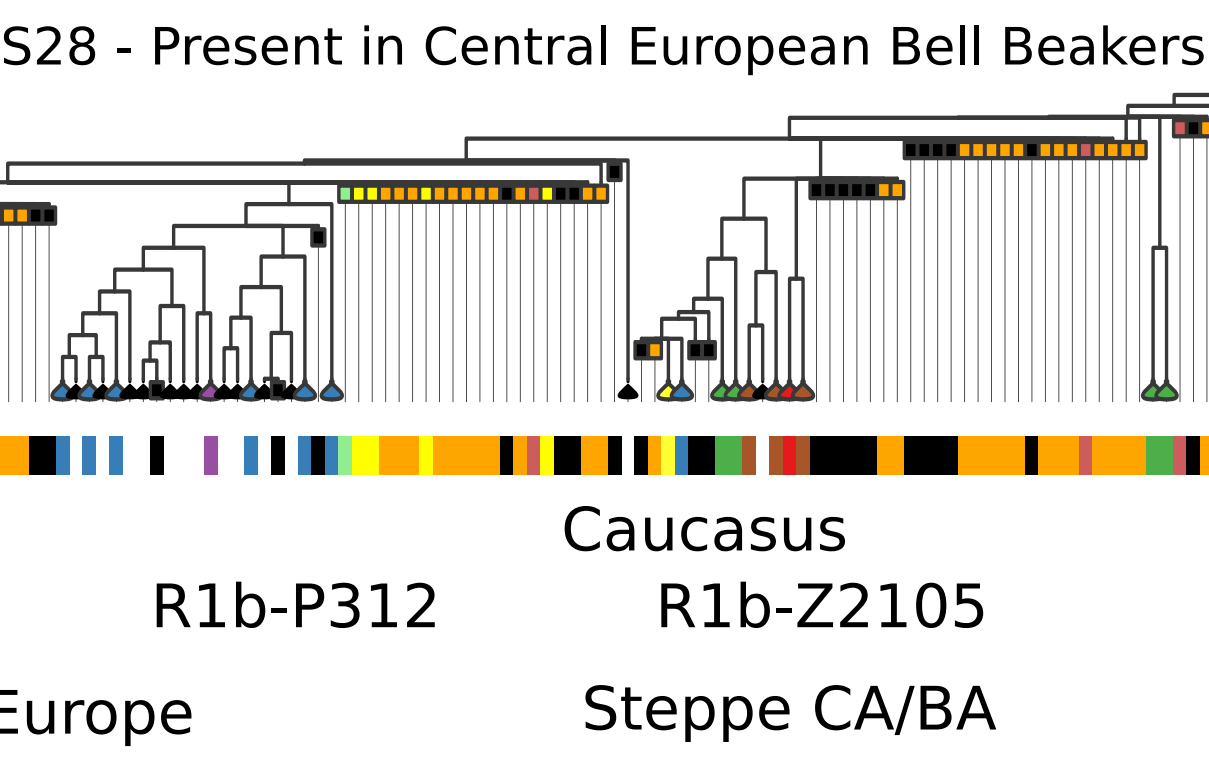CA/BA Britain

Bell Beakers Central Europe
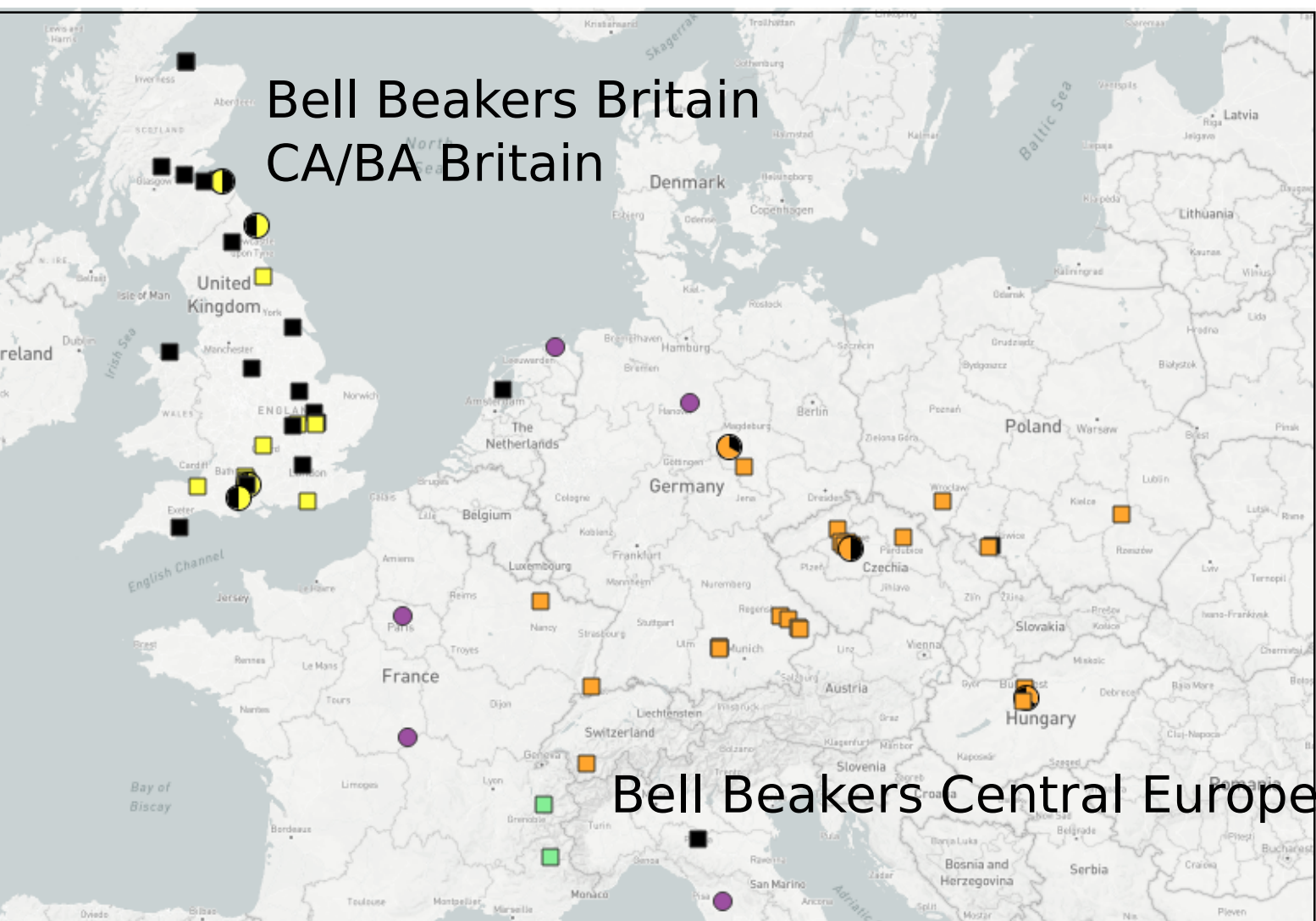
.Yamnaya (Z2105) and Beaker (P313) lineages belong to different Ychr clades
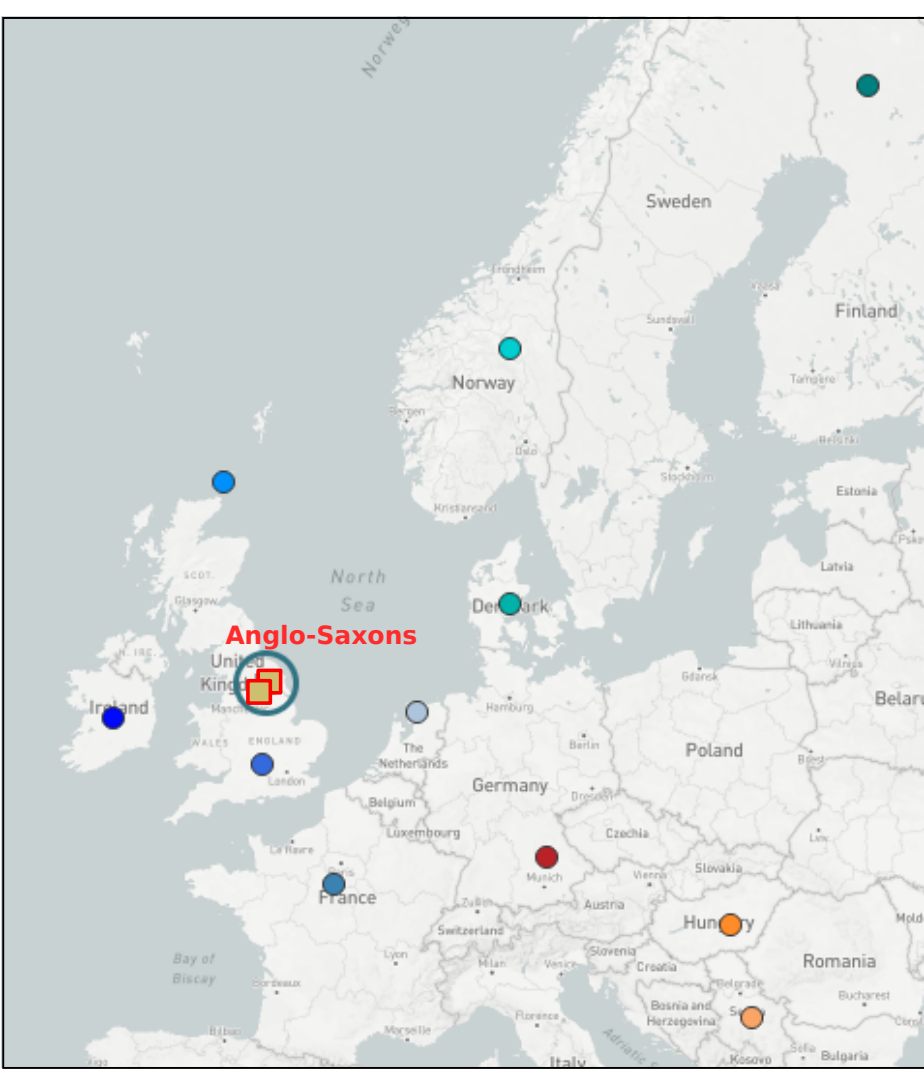
Z2105 - Yamnaya, steppe related populations

P312 - Bell Beaker onwards, Central and Western Europe. Most frequent lineage in Central/Western Europe

Descendants of P312 are highly differentiated during the Copper Age:

L21 Present in British Bell Beakers, Copper Age, Bronze Age (almost completely restricted to Britain), 1 Roman British belong to this lineage, as well as 2 Irish Bronze Age samples, 1 Bronze Age Britain. Important role in the making of Y-chromosome diversity in the British Isles.

S28 - Present in Central European Bell Beakers

WEur   WEur   Caucasus
R1b-L21   R1b-S28   R1b-P312   R1b-Z2105
Bell Beakers Britain   Bell Beakers Central Europe   Steppe CA/BA
CA/BA Britain

Anglo-Saxon period

Anglo-Saxons

Two Anglo-Saxon samples from England were placed in branches leading to I1a2b and I1a2a.

These lineages are widespread in Norway, Sweden, Germany, Frisia.

(tree estimated in Hallast et al., 2015)

I1a2b
I1a2a
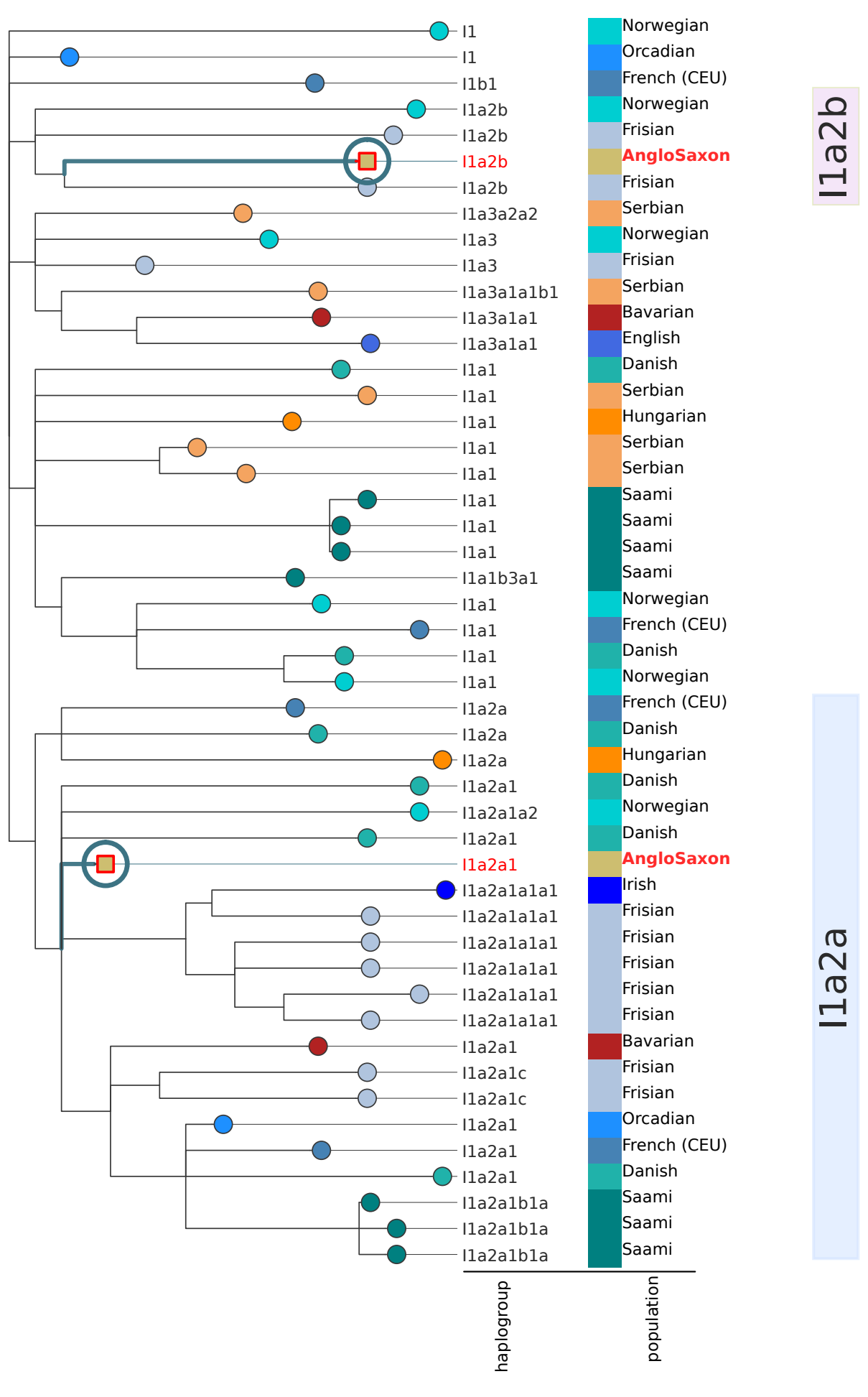
References
Poznik 2013
Poznik 2016
Kivisild 2017

Contact:
Rui Martiniano
rm890@cam.ac.uk

Datasets:
Karmin et al., 2015, Hallast et al., 2016
Mathieson et al., 2017, Mathieson et al., 2018, Olalde et al., 2018, Martiniano et al., 2016