Final Project in Population Genetics on Genomes

Asger Svenning | 201907234

Exploring non-African archaic segments

Description In this project you will be looking at segments of archaic genomes identified individual modern humans. You will investigate how alike non-African genomes are in terms of how archaic segments are distributed. You will be working with an extended version of the data set that you worked with in the admixture exercise. In this version you also have an addtional file with the positions of candidate archaic SNPs.

Investigate the following In your project you must address the questions below, but you are also expected to expand the project to answer your own questions. How you do this is up to you. You do not need to answer them in the order they are listed. Make a project plan with a set of analyses that will allow you to answer the questions.

A. To what extent do individuals share SNPs contributed by archaic human introgression? In other words, how correlated are the archaic contents in two individuals?
B. How does this correlation change when you compare individuals from different populations?
C. How does it change when comparing individuals from different geographical regions?
D. Does the region containing the EPAS1 gene stand out in any way? (Redo the analysis above for a 1Mb window surrounding this gene).
E. What is the total amount of admixture (Archiac genomic sequence) in each non-African individual genome?
F. What is the total amount of admixture (Archiac genomic sequence) in the region around EPAS1 in each individual?
G. Do individuals with large admixture totals have more correlated admixture patterns? Do individuals with large admixture totals in the EPAS1 region have more correlated admixture patterns in the EPAS1 region? Can you find any evidence of adaptive introgression?
H. Perform any additional analyses of your own choice.

OBS Answers to questions are marked as follows: "… answer.$^{\mathbf{(X)}}$", where **X** refers to the question which the section relates to.

Abstract

This report builds upon the work of Skov *et al.* (2018), who developed a model to identify archaic segments in individual genomes without the need for an archaic reference genome. Skov *et al.* (2018) were able to identify introgressed archaic segments, estimate admixture proportions and times, and initial divergence times between human and archaic populations. In this report I utilize the positions of the archaic segments identified by Skov et al. to explore the distribution and similarity of archaic segments among individuals from different populations and geographical regions.

Using this rich data foundation I investigate patterns of shared archaic state across diverse populations and geographical regions at the genome level, as well as assessing the significance of the region encompassing the EPAS1 gene, quantifying the level of admixture in non-African individual genomes.

To quantify the similarity of archaic ancestry among individuals, I employ the Intersection over Union (IoU) index. This index measures the proportion of shared archaic ancestry between individuals, allowing for efficient similarity calculations using bit-wise Boolean operations. However, due to non-linear computational scaling with vector length of the similarity computation (which are whole genomes in this report) I implement an approximation method downsizing the archaic segments by a factor of 1000, speeding up the computation by ~$100\,000\%$ while keeping the approximation error negligible ~$\pm 0.1\%$.

My analysis reveal distinct clusters of shared archaic ancestry, with Papua New Guinea, Asia, and West-Eurasia displaying different patterns. The distribution of shared archaic ancestry is found to be multi-modal, primarily influenced by regional differences. Moreover, I find a positive monotonic relationship between total of archaic ancestry and the degree of archaic similarity at the whole-genome scale, mostly driven by inter-regional patterns, while the relationship is more modest and or non-monotonic intra-regionally. However, in the region around the EPAS1 gene, which has been shown to have undergone adaptive introgression[3], the inter-regional relationship is replaced completely by almost identical and positive monotonic intra-regional relationships. This result shows that archaic introgression around the EPAS1 gene is not driven by global population structure providing supporting evidence for adaptive introgression. This is supported by a strong cluster of shared archaic ancestry in the EPAS1 region, consisting mostly of East Asians primarily from Tibet, Bhutan and Nepal.

1 Introduction

Modern human genomes are characterized by introgressed archaic segments from hybridization events with, at least, Neanderthals and Denisovans during the global expansion of *Homo sapiens* from Africa.[3, 4, 6] In this report I investigate the patterns of these segments, and simultaneously attempt to replicate the findings of adaptive introgression in previous works.[3, 4]

1.1 Quantifying Patterns in Archaic Ancestry

In order to allow an adequate investigation of the patterns in the archaic ancestry of modern humans, we must be able to quantify the relationship between archaic ancestry in two individuals and/or characteristics of archaic ancestry within an individual. In this report I have chosen to quantify the relationship between individuals with the similarity index **I**ntersection **o**ver **U**nion (hereafter called ”IoU”, sometimes called the Jaccard index), which is bounded in $[0, 1]$, and can thus readily be transformed into a dissimilarity index:

$$\text{IoU}(A, B) = \frac{A^a \cap B^a}{A^a \cup B^a} = \frac{\sum A_i \,\&\, B_i}{\sum A_i \mid B_i} \tag{1}$$

$$d_{\text{IoU}}(A, B) = 1 - \text{IoU}(A, B) \tag{2}$$

Where A and B are vectors encoding the archaic ancestry state for the two individuals (A and B), containing 0 on indices representing genomic positions without archaic ancestry and by symmetry containing 1 on indices representing genomic positions with archaic ancestry. The notation X^a describes the set of genomic positions of vector X with archaic ancestry, while X^m describes the set of genomic positions modern ancestry. The IoU between two individuals A & B; $\text{IoU}(A, B)$, can readily be understood as the proportion of shared archaic ancestry. That is if $\text{IoU}(A, B) = 0.25$ then person A and B share 25% of their archaic ancestry. It should be noted however that the IoU can be large even if both individuals have only a small amount of archaic ancestry, all information about positions which are of modern ancestry in both individuals is disregarded.[1]

The IoU (dis)similarity index was chosen by twofold arguments: (1) interpretability and (2) allows extremely efficient similarity calculations using bit-wise Boolean operations (Figure 1[2]). However the memory footprint of the binary vectors is still prohibitive at the whole-genome scale. I have therefore approximated the binary archaic state vectors by dividing the indices of the archaic fragments by a constant and rounding. This reduced the memory usage by several orders of magnitude, while perhaps not sacrificing much in terms of precision. It is however very important to consider the way in which the indices are rounded after division, to ensure that the calculated distances are unbiased and low-variance.

1.2 Comparison of Suggested Methods for Efficient & Unbiased Approximation of d_{IoU}

I suggest six rather straight-forward rounding methods:

1. Ceiling
2. Floor
3. Round
4. Random
5. Expand
6. Contract

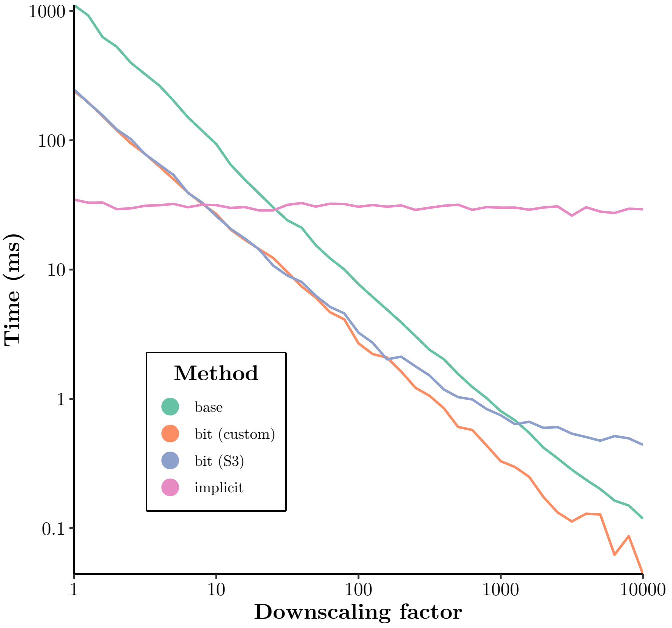

Fig. 1 Log-log benchmark comparison of computational speed using different methods for calculating d_{IoU} with increasingly downscaled archaic segments. All downscaling is done using the "contract" method and segments are already downscaled by a factor of 10 before the benchmark experiment. The "base" method simply uses base R[1] logical vectors and Boolean operations, where both "bit (S3)" and "bit (custom)" uses the `bit` package[2] where the former uses the S3-methods for Boolean and sum operations, while the latter bypasses the S3-methods and calls the `bit C++`-functions directly. The "implicit" method does not rely on Boolean operations, but instead calculates the intersection and union by directly calculating the intersection and union of all segments using only the start and end indices, which is why it does not benefit from approximating the segments.

The first three (”Ceiling”, ”Floor” and ”Round”) are perhaps the most naïve. These three methods simply round both the start and end indices the same way: either up, down or to the closest integer. The next three methods are slightly more complicated. The ”Random” method also treats the start and end index identically, however as the name suggest it is a stochastic rounding method. First the number is split into it's integer, x_Z, and decimal, X_R, parts and then the decimal part is sampled as a Bernoulli variable, and the final rounded number, X_{round} is then the sum of these such that: $X_{\text{round}} \sim X_Z + \text{B}(1, X_R)$. The last two methods treat the start and end indices differently: The ”Expand” method *expands* the rounded intervals by rounding the start index down and the end index up. The ”Contract” method *contracts* the rounded intervals by doing the opposite of the ”Expand” method; rounding the start index up and the end index down.

I evaluate these different methods by randomly selecting two individuals and calculating d_{IoU} for increasingly downscaled representations (approximations) of the archaic segments (Figure 3). This experiment clearly validates the use of the ”contract” segment approximation method for downscaling factors up to on the order of ~10 000 for obtaining unbiased estimates d_{IoU} with an error between ~$[-1\%, 1\%]$.

[1] I have chosen to use d_{IoU} for my empiric analysis instead of IoU, since many methods in genetics rely on distance or dissimilarity measures not similarities.

[2] It is necessary to downscale the Boolean vectors by factor of 10 to be able to conduct this experiment for two reasons (1) the `bit` package only supports vectors up to size $< 2^{3}1$ and (2) an R Boolean vector of size 3 billion has a size of 12GB which barely fits in my machines' RAM.

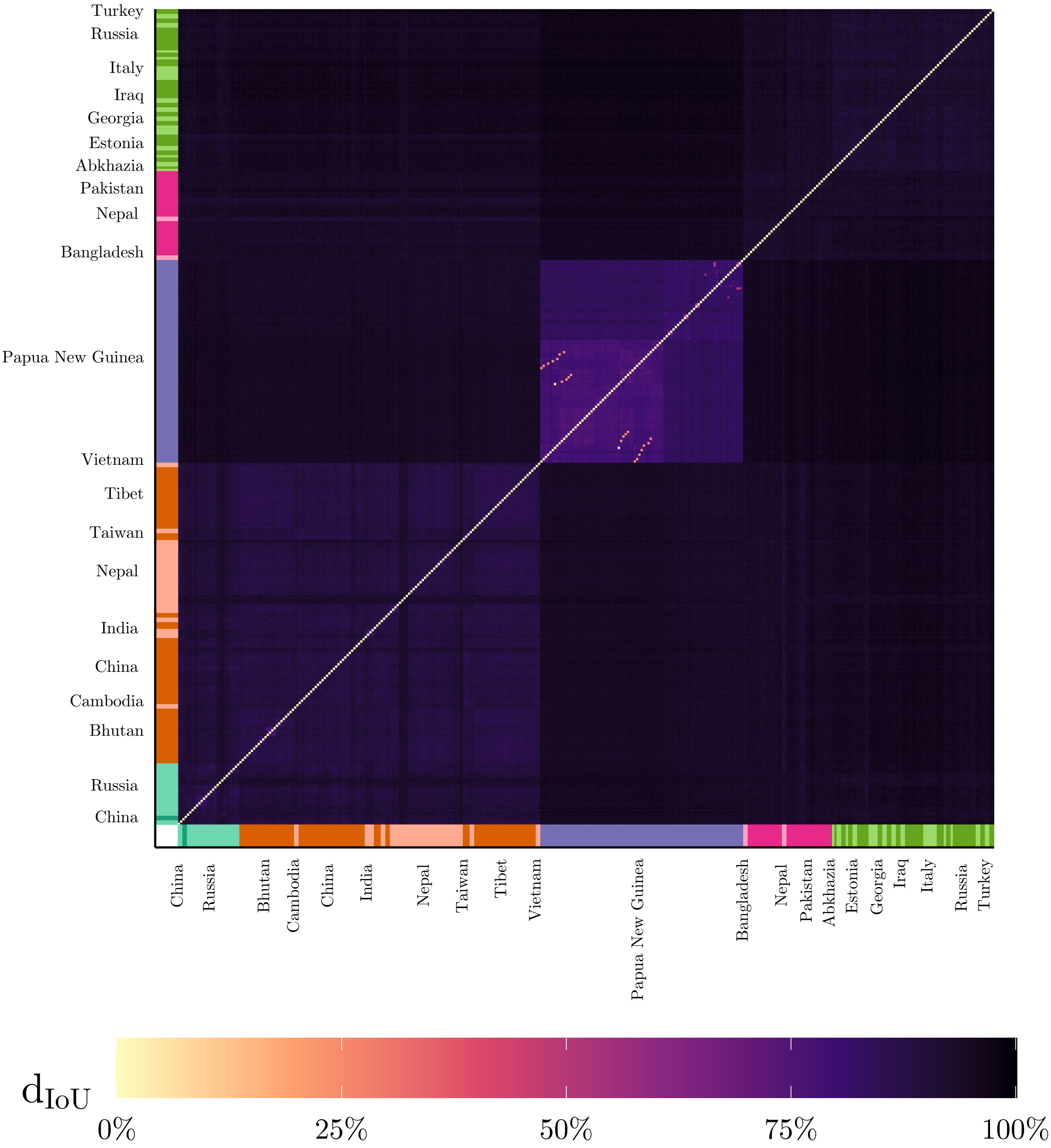

Fig. 2 Full pairwise d_{IoU} matrix between all individuals included in this report (n = 358). d_{IoU} is approximated using a downscaling factor (=1000) and "contract" rounding (Figure 3). The colored bars on the left and bottom of the heat-map indicate rows with individuals from the same region (discrete colors) and country (alternating luminosity), some country labels are suppressed to avoid overlapping text for clarity. Some countries may appear multiple times (such as Russia) since they span multiple regions.

Based on these experiments I have chosen to approximate all segments such that $A^* = \left\{\text{ceiling}\left(\frac{A_{\text{start}}}{1000}\right), \text{floor}\left(\frac{A_{\text{end}}}{1000}\right)\right\}$ (i.e. ”contract” and downscale = 1000), where A^* is the approximated implicit representation of the archaic segments, while A is the supplied implicit representation from Skov *et al.* (2018), where A_{start} and A_{end} respectively are the start and end positions of the archaic segments, unless otherwise specified. For simplicity I also ignore the archaic origin of each segment.

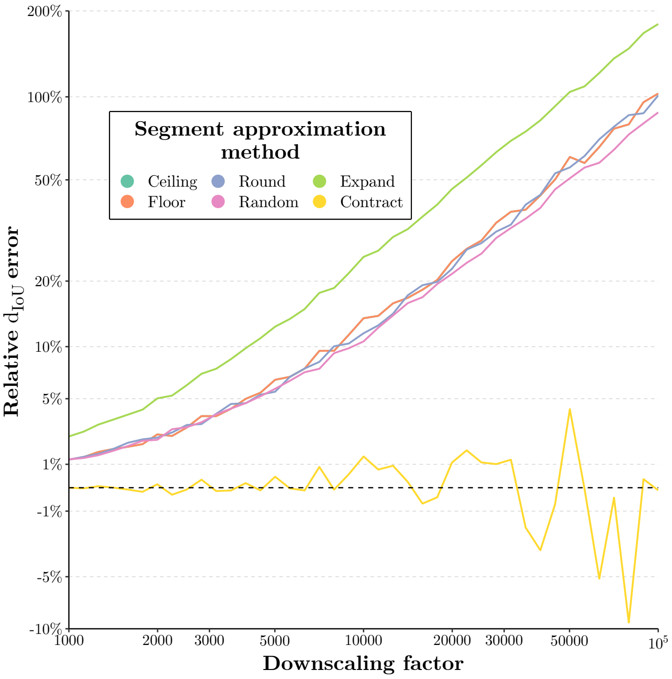

Fig. 3 Comparison of the relationship between downscaling (approximation) factor and IoU error for different segment approximation methods. All errors are based on whole-genome archaic segment d_{IoU} between two randomly sampled individuals from Finland (S-Finnish-1) and Papua New Guinea (13784_5) and the errors are calculated by comparison to d_{IoU} calculated on non-approximated segments. In this experiment I have chosen to use a minimum downscaling factor of 1000, both due to the polynomial scaling of the boolean d_{IoU} algorithms and considering that the error is negligible for downscaling factors less than 1000. Although this figure includes only two randomly sampled individuals, which is done for simplicity, I have re-run the experiment multiple times with individuals with both very high and low IoU, and the results are similar or better (lower error).

2 Results

2.1 Patterns in Shared Archaic Ancestry

To explore the patterns in the shared archaic ancestry I computed the pairwise d_{IoU} matrix for all individuals revealing at least three distinct clusters (Figure 2); Papua New Guinea, Asia & West-Eurasia from least to most clustered. However it is also clear that most individuals share only a very small proportion of their archaic ancestry ($\overline{\text{IoU}} \approx 4.1\%$), but that the IoU distribution is heavily left-skewed with a long right-tail (Figure 4).(A) However by splitting the histogram by region (not shown, but easily discernible in Figure 2), it becomes apparent that the distribution of d_{IoU} is actually multimodal around four clusters, the two largest ones separating the Papuan population of Papua New Guinea, followed by Asia and West-Eurasia. These clusters are even more distinct when analysing the hierarchical relationships in the distance matrix using a neighbour-joining tree computed on the distance matrix (Figure 5). To quantify how much of the patterns of shared archaic ancestry can be explained by region, country and population, I conducted a nested sequential anova (Table 1). Perhaps surprisingly this showed that almost all variation is explained at the region level, while country and population are only able to explain a small proportion of the remaining variation. The analysis also highlights the fact that most of the variation, ~53%, in shared archaic ancestry is found when comparing individuals from different regions, while differences in mean shared archaic ancestry between regions only accounts for half as much of the variation, ~22% (Table 1). Interestingly the pattern is reversed at the country and population level, suggesting that within regions the total and shared archaic ancestry between individuals from different countries is similar to that of individuals from the same country, and likewise for populations within countries.(B,C) Conducting this analysis while only considering segments in a 2Mb window around the EPAS1 gene reveals that the inter-regional pattern weakens substantially compared to the intra-regional pattern, while the intra-population is much stronger.(D)

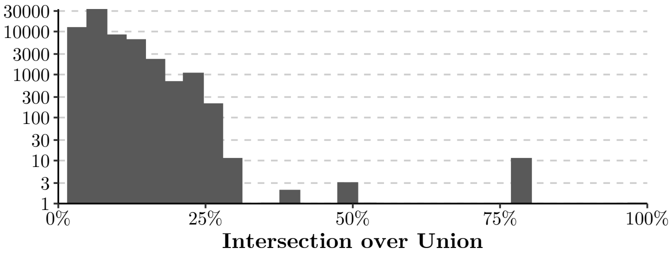

Fig. 4 Logarithmic histogram of IoU between all individuals (non-symmetric duplicates and non-identical).

	Proportion of d_{IoU} explained by	
	Between	**Within**
Region	53.11%/3.41%	22.19%/3.41%
Country	0.10%/1.15%	0.22%/2.76%
Population	0.06%/0.08%	0.11%/0.41%
Residuals	24.22%/91.93%	

Table 1 Sequential (left-right, top-bottom) proportion of sum of d_{IoU} residual squares explained by different geographical scales. The analysis is done at both the whole-genome scale (left numbers) and in a 2Mb window around the EPAS1 gene (right numbers) separately, the forward slash in each cell separates the results from each analysis. For the EPAS1 model all individuals with no archaic segments in the window are omitted since IoU is undefined between empty sets. For this analysis d_{IoU} is modelled using ordinary linear regression and sum of residuals squares are calculated using a type I anova. For simplicity I have ignored the geographical sampling bias, however this is sure to account for a major part of the patterns.

2.2 The Relationship Between Total Archaic Ancestry & Archaic Similarity

In the following part of the report I will use the abbreviation ”TAS” for ”**T**otal **A**rchaic **S**egment length” and the terms shared archaic ancestry and archaic similarity interchangeably, where they both refer to IoU. To investigate if individuals with a larger TAS are more archaically similar than individuals with a lower TAS, I binned all individuals by their TAS and calculated the median d_{IoU} within each bin. I only considered comparisons between individuals in the same TAS bin and region. This was done minimize the effect of geography and re-

Total Archaic Segment Length	$Q_{\%50}[d_{\text{IoU}}]$ (n)				
	West-Eurasia	**South-Asia**	**East-Asia**	**Central-Asia-Siberia**	**Melanesia**
[50M, 60M]	92.7% (*21*)	- (*0*)	- (*0*)	- (*0*)	- (*0*)
(60M, 70M]	91.7% (*300*)	93.4% (*1*)	- (*0*)	- (*0*)	- (*0*)
(70M, 80M]	90.7% (*595*)	92.6% (*45*)	- (*0*)	- (*0*)	- (*0*)
(80M, 90M]	89.1% (*6*)	92.1% (*276*)	88.7% (*78*)	88.3% (*66*)	- (*0*)
(90M, 100M]	- (*0*)	92.0% (*3*)	88.3% (*3003*)	88.4% (*78*)	- (*0*)
(100M, 110M]	- (*0*)	- (*0*)	86.8% (*780*)	- (*0*)	- (*0*)
(180M, 190M]	- (*0*)	- (*0*)	- (*0*)	- (*0*)	77.0% (*120*)
(190M, 200M]	- (*0*)	- (*0*)	- (*0*)	- (*0*)	76.5% (*231*)
(240M, 250M]	- (*0*)	- (*0*)	- (*0*)	- (*0*)	82.5% (*15*)
(250M, 260M]	- (*0*)	- (*0*)	- (*0*)	- (*0*)	81.7% (*66*)
(260M, 270M]	- (*0*)	- (*0*)	- (*0*)	- (*0*)	81.1% (*45*)
(270M, 280M]	- (*0*)	- (*0*)	- (*0*)	- (*0*)	80.3% (*10*)
(310M, 320M]	- (*0*)	- (*0*)	- (*0*)	- (*0*)	75.9% (*6*)
(320M, 330M]	- (*0*)	- (*0*)	- (*0*)	- (*0*)	76.6% (*6*)
(330M, 340M]	- (*0*)	- (*0*)	- (*0*)	- (*0*)	75.4% (*10*)

Table 2 Median d_{IoU} between individuals with similar total archaic segment lengths. Comparisons between individuals with dissimilar segment lengths or from different regions are omitted. The median d_{IoU} in the total population is ~93.5%. n is the number of comparisons made in the bin, i.e. the number of individuals in each bin is $^1/_2 + \sqrt{^1/_4 + 2\text{n}}$. Cells containing "- (*0*)" represent combinations between total archaic segment length bins and region containing 0 or 1 individual.

stricting the analysis to the marginal effect of TAS (i.e. excluding the interaction between individuals' TAS). As can by comparing between regions in Table 2 there is a clear tendency of individuals with higher TAS to also be more archaically similar, however this pattern is likely explained by general inter-regional patterns of genetic relatedness. The pattern is also present, although weaker, at the intra-regional scale, at least in West Eurasia, South and East Asia, while Melanesia (Papua New Guinea) has a more hump-shaped relationship between TAS and d_{IoU} which is likely explained in part by the strong (Figure 2 and Figure 5) intra-population structure in Melanesia.(G)

Redoing this analysis subsetting for only segments in 1Mb window around the EPAS1 gene reveals that the TAS in this region is substantially more uni-modal (Figure B.1 and Figure B.2), but with a fat right-tail consisting solely of individuals from the East-Asia region and Tibet specifically.(E,F) The intra-region relationship between TAS and d_{IoU} are also much stronger across all regions except Melanesia, while the inter-region relationship has mostly disappeared. Following the results from Huerta-Sánchez *et al.* (2014) I have chosen to split out individuals from Tibet as a pseudo-region, which shows that Tibetans are part of a sub-regional structure from East-Asia with both much higher archaic similarity and TAS in the region around EPAS1 (Table 3, Figure B.2 and Figure B.3 subfigure "EPAS1").(G)

Total Archaic Segment Length	$Q_{\%50}(d_{\text{IoU}})$ (n)					
	West-Eurasia	**South-Asia**	**East-Asia**	**Tibet**	**Central-Asia-Siberia**	**Melanesia**
[0, 0]	- (*666*)	- (*120*)	- (*120*)	- (*0*)	- (*15*)	- (*2080*)
[1, 50K)	100.0% (*66*)	100.0% (*28*)	100.0% (*190*)	- (*0*)	100.0% (*10*)	100.0% (*210*)
[50K, 100K)	91.9% (*45*)	100.0% (*6*)	100.0% (*91*)	- (*0*)	84.0% (*3*)	- (*0*)
[100K, 150K)	56.0% (*10*)	100.0% (*10*)	88.5% (*78*)	- (*0*)	80.5% (*15*)	- (*0*)
[150K, 200K)	82.3% (*10*)	99.1% (*1*)	100.0% (*45*)	- (*0*)	94.4% (*15*)	29.6% (*1*)
[200K, 300K)	100.0% (*1*)	83.0% (*6*)	85.0% (*78*)	30.0% (*28*)	- (*0*)	- (*0*)
[300K, 400K)	- (*0*)	- (*0*)	53.9% (*36*)	60.4% (*28*)	- (*0*)	- (*0*)
[400K, 500K)	- (*0*)	- (*0*)	52.5% (*28*)	45.3% (*21*)	- (*0*)	- (*0*)
[500K, 600K)	- (*0*)	- (*0*)	39.0% (*1*)	53.9% (*1*)	- (*0*)	- (*0*)

Table 3 Median d_{IoU} in a 1Mb window around the EPAS1 gene between individuals with similar total archaic segment lengths. Comparisons between individuals with dissimilar segment lengths or from different regions are omitted. The median d_{IoU} in the total population is ~97.5%. n is the number of comparisons made in the bin, i.e. the number of individuals in each bin is $^1/_2 + \sqrt{^1/_4 + 2\text{n}}$. Cells containing "- (*0*)" represent combinations between total archaic segment length bins and region containing 0 or 1 individual. For this analysis segments are not approximated (i.e. downscaling factor = 1), since it is computationally feasible to compute d_{IoU} between ~2Mb vectors.

Neighbour-Joining Tree of the Whole-Genome Pairwise $\mathrm{d_{IoU}}$ Matrix

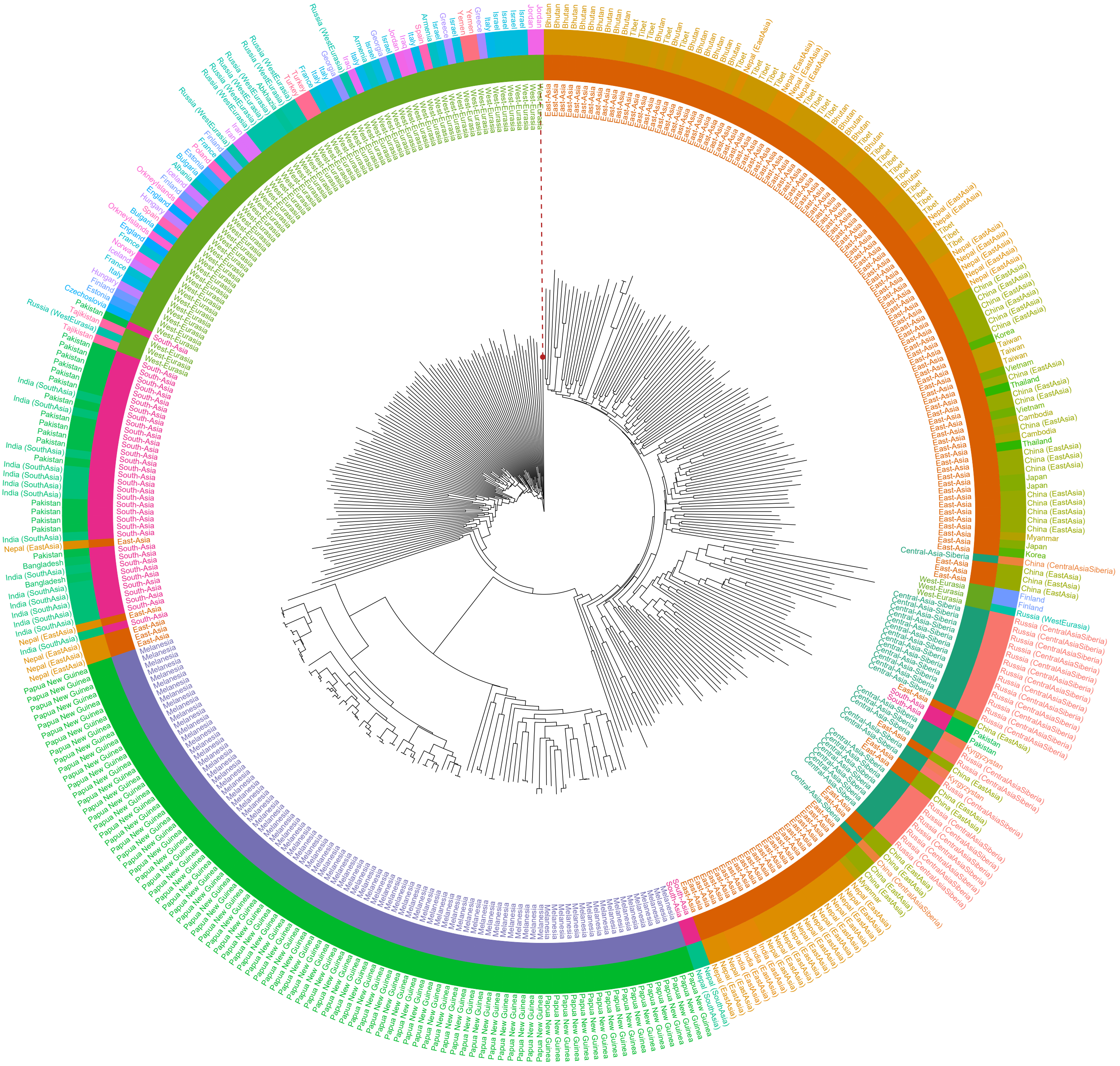

Fig. 5 Neighbour-Joining pseudo-phylogeny based on the (approximate, see Figure 2 and Figure 3) full whole genome pairwise d_{IoU} matrix between all individuals annotated by country (outside of colored ring) and region (inside of colored ring). For visualization purposes the phylogeny is rooted using the individual with the highest average divergence $\overline{d_{\mathrm{IoU}}}$ as outgroup, annotated with the red dot and dashed line, and node height is defined by distance to the root node, which is strongly either positively (West Eurasia, South Asia and Melanesia) or negatively (East Asia and Central Asia Siberia) with total archaic segment length (not shown). Interestingly the chosen outgroup is an individual from the Middle-East, which might be an indication of hybridization with non-admixed individuals i.e. Africans (This is also the case when redoing the analysis for segments which are subset by their ancestry, as well as segments with 1Mb of the EPAS1 gene). For robustness d_{IoU} is clamped to the 0.1%-quantile before computing the Neighbour-Joining tree.

2.3 Tree-Based Analysis of Archaic Similarity

As a qualitative non-parametric analysis I computed pseudo-phylogenies using neighbour-joining[9], rooted by using the most divergent individual (i.e. highest mean d_{IoU}) as an out-group. This was done both on whole-genome scale (Figure 5) and for only archaic segments in a 2Mb window around the EPAS1 gene, as well as using only archaic segments with specific archaic origins (Figure B.3). Neighbour-joining is chosen as a more robust alternative to traditional distance-based tree building algorithms such as UPGMA, considering that d_{IoU} is not a metric and avoiding the necessity of implying a ”molecular clock”[3], however certainly other more sophisticated tree-building algorithms from the Population Genetics might be more appropriate. I leave this field of inquiry open to future works. I use these trees as a qualitative non-reductive supplement to enrich the discussion of geographical patterns of introgressed archaic segments as well as the differences between introgression from Neanderthals, Denisovans and unknown archaic human species (see Section 3).**(H)**

3 Discussion

By analysing the patterns of shared archaic ancestry (d_{IoU}) I show that modern human genomes contain a large amount of shared archaic segments, with a strong structure on a macro-geographical level (Figure 5) and moderate structure on the country and population level (Table 1). This pattern is mostly driven by archaic segments of Neanderthal and unknown origin, while segments with Denisovan origin only show significant geographic structure in Asia and Melanesia, with Papua New Guinea forming an especially deep cluster with significant intra-population structure (Figure B.3). This result is in clear agreement with prior work on Denisova introgression by Browning *et al.* (2018)[4]. The similarity between the patterns of archaic ancestry between archaic segments with known and unknown origin (Table B.3 subfigure ”Known” and ”Unknown”, bottom row), suggest at least on of three things (1) introgression from unknown archaic human species/populations, (2) that much of the genetic variation in Neanderthals and Denisovans is still unknown (at least at the time of Skov *et al.* (2018)s’ work) or (3) methodological errors, misspecifications or sequencing errors. However the fact that Papua New Guinea forms an even deeper cluster based on only unknown archaic segments is easiest explained if a substantial amount of the variants from the unique second Denisova pulse in Papua New Guineas’ ancestry[4] are unknown, which could be explained if no archaic genomes from the second Denisova pulse have been found/are available.**(H)**

The strong inter-regional, but weak intra-regional, relationship between TAS and archaic similarity at the whole genome level (Table 2) is likely explained the strong isolation by distance in humans[6], suggesting that the majority of archaic introgression is non-adaptive and that most archaic segments in the modern human genome has a neutral or weak fitness effect. The fact that the strength of inter- and intra-regional relationships between TAS and d_{IoU} is reversed in the region around EPAS1 (Table 3) supports previous findings of adaptive introgression for EPAS1[3, 4] by symmetry. This is supported by the strong clustering of the Himalayan populations of Tibetans, Bhutans and Nepalese (Table B.3, subfigure ”EPAS1”) based on d_{IoU} using only segments from the region around EPAS1.**(G)**

4 Conclusions

In this report I show how using the Intersection over Union/Jaccard index (IoU) is an information rich and extremely computationally efficient similarity metric for analyses of archaic ancestry in modern humans. The use of Intersection over Union has been applied to population genetics research on population structure[7], however to the best of my knowledge it has never been applied to analyses of archaic ancestry. Using this approach I show strong geographical patterns in introgressed archaic segments in modern human genomes, and discover an equally if not stronger structure in introgressed archaic segments with unresolved archaic origin, providing further motivation for future works on uncovering unknown archaic introgression sources and dynamics. By analysing the patterns of pairwise IoU between individuals and total archaic segment length (TAS) I also find strong evidence for adaptive introgression in the EPAS1 gene, in support previous works[3, 4], based on divergent patterns of shared archaic ancestry in the EPAS1 region as compared to the whole genome level.

5 Software & Data

All analysis are conducted in the `R` programming language[1]. Data wrangling is primarily done using the `tidyverse` related packages[11] and graph handling is done using the `tidygraph` package[8], while all plots are created using the `ggplot2` package[10]. The `bit` package[2] is used for memory efficient Boolean arrays and fast Boolean operations. All data is provided by Kasper Munch through the course ”Population Genetics on Genomes” at Bioinformatics at Aarhus University. The analysed archaic segments are originally sourced from Skov *et al.* (2018)[5] with gratitude.

References

1. R Core Team. *R: A Language and Environment for Statistical Computing* Vienna, Austria, 2022.
2. Oehlschlägel, J. & Ripley, B. *bit: Classes and Methods for Fast Memory-Efficient Boolean Selections* (2020).
3. Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512, 194–197. ISSN: 14764687. (2014).
4. Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S. & Akey, J. M. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell* 173, 53–61.e9. ISSN: 10974172. (2018).
5. Skov, L. *et al.* Detecting archaic introgression using an unadmixed outgroup. *PLoS Genetics* 14, e1007641. ISSN: 15537404. (2018).
6. Ramachandran, S. *et al.* Support from the relationship of genetic and geographic in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America* 102, 15942–15947. ISSN: 00278424. (2005).
7. Prokopenko, D. *et al.* Utilizing the Jaccard index to reveal population stratification in sequencing data: A simulation study and an application to the 1000 Genomes Project. *Bioinformatics* 32, 1366–1372. ISSN: 14602059. (2016).
8. Pedersen, T. L. *tidygraph: A Tidy API for Graph Manipulation* (2023).
9. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528 (2019).
10. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* ISBN: 978-3-319-24277-4. (Springer-Verlag New York, 2016).
11. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* 4, 1686 (2019).

[3] What a ”molecular clock” means in terms of my analysis is also somewhat unclear.

Appendix

A Supplementary Materials

All code and figures are available on my GitHub; https://github.com/asgersvenning/PopulationGeneticsFinalProject.

B Supplementary Figures

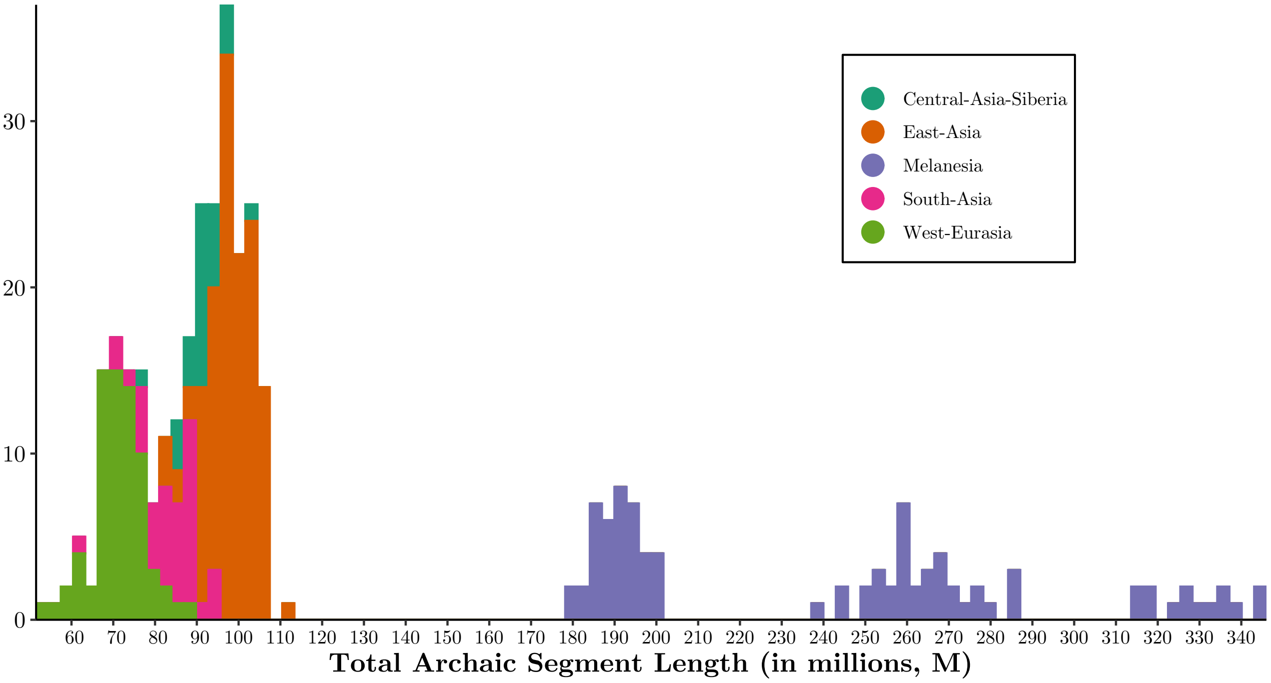

Fig. B.1 Whole genome total archaic segment length histogram, stratified by region.

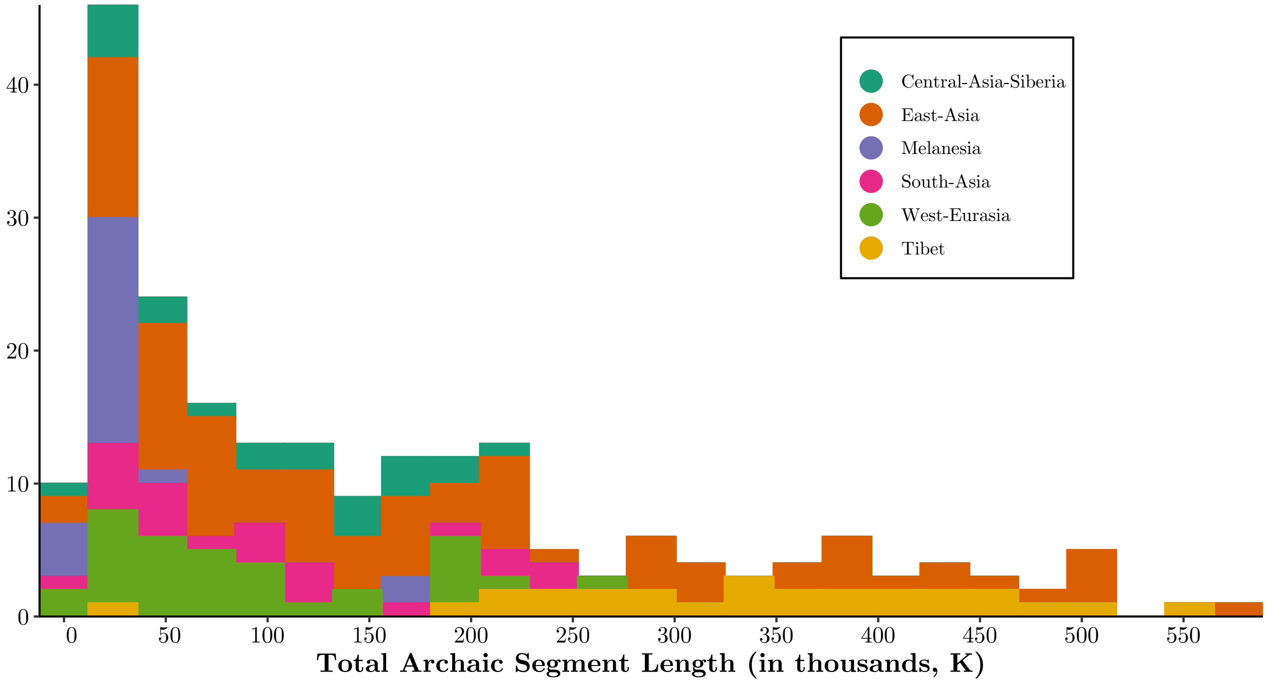

Fig. B.2 Total archaic segment length histogram for archaic segments in a 1Mb window around the EPAS1 gene, stratified by region.

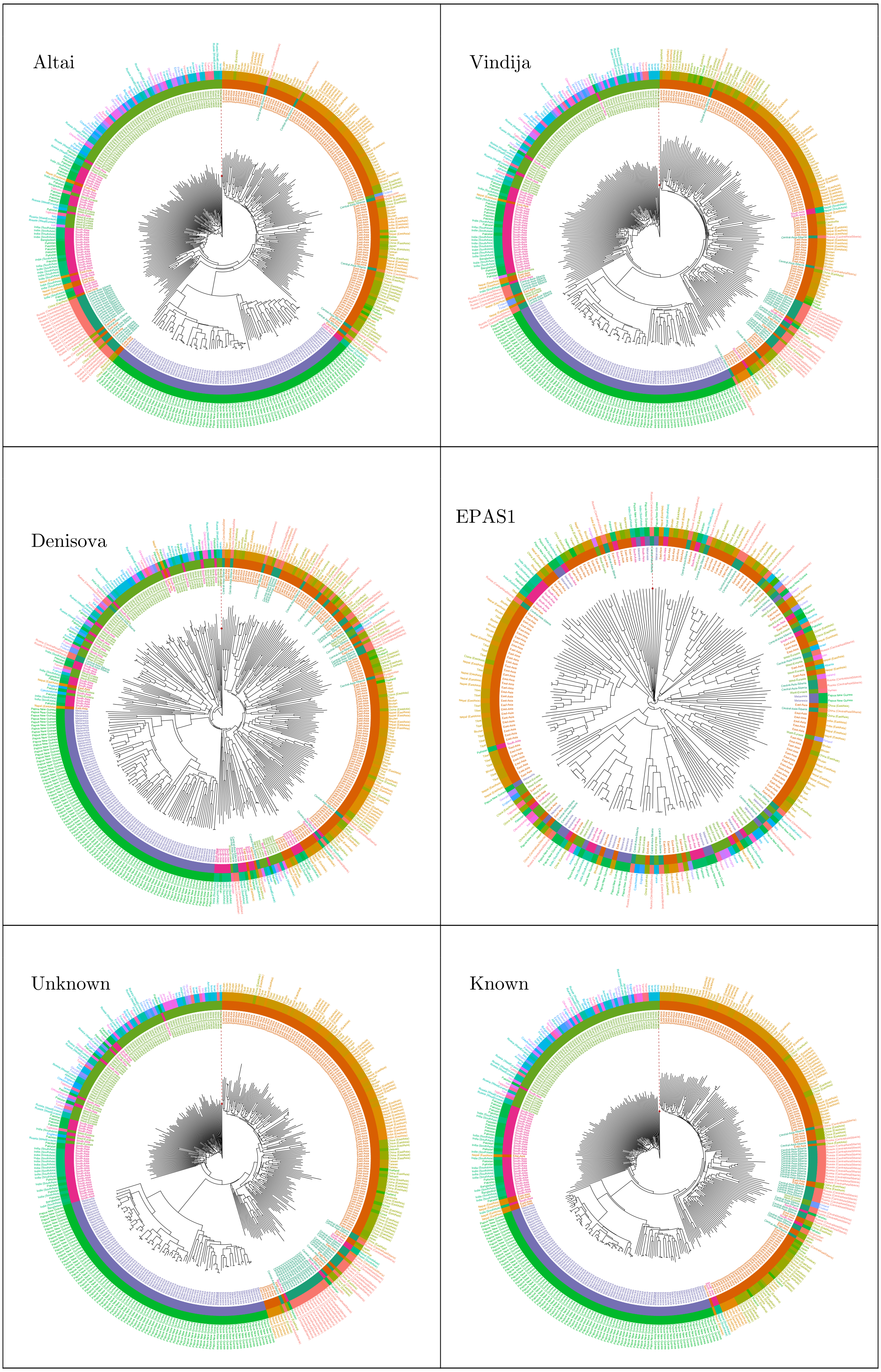

Fig. B.3 Pseudo-Phylogenies computed exactly as Figure 5, but using different subsets of archaic segments based on their ancestry or if they are within a 1Mb window of the EPAS1 gene. To subset the archaic segments I have classified them into six categories using two simple rules; (1) if more than 50% of SNPs in the segment are unknown then the segment is classified as "Unknown" otherwise "Known", (2) if the segment is "Known" then it is classified to the archaic ancestry with the largest number of SNPs in the segment; either "Altai", "Vindija" (Neanderthal) or "Denisova". For the "EPAS1" pseudo-phylogeny all segments within a 2Mb window around the EPAS1 gene are used and individuals without any segments in the window are omitted, and segments are not approximated (i.e. downscaling factor = 1).