An 'Ancestral 850k' SNP panel for studies of human genetic history using outgroup-ascertainment

Pontus Skoglund
Francis Crick Institute
pontus.skoglund@crick.ac.uk

Targeted enrichment of nuclear SNPs allows for population genetic analysis of ancient skeletal remains that are not amenable to direct shotgun sequencing due to the overwhelming presence of microbial DNA[1]. However, analysis of restricted sets of SNPs that have been discovered in test populations can bias population genetic analyses[2,3], except if ascertainment is performed using an outgroup[4,5], in which case the ascertainment is symmetric with regards to the populations under study (**Figure 1**). Under outgroup-ascertainment, the goal is to discover mutations that appeared in the ancestral population of all the individuals of interest, after which its frequency would have evolved largely randomly under genetic drift, ideally on an evolutionary time scale short enough that variants do not reach fixation in any population. Outgroup ascertainment allows empirical studies of ancestry[5], divergence times[4], and conditional heterozygosity [6], in multiple species[7–12].

However, no current SNP array or ancient DNA capture reagent includes SNPs that are outgroup-ascertained with respect to all present-day human ancestry, most notably African-ancestry individuals and groups. Here we propose such a set of SNPs, by discovering variants polymorphic in Denisovan and Neanderthal genomes, which are largely outgroups to modern human ancestry[13,14], resulting in an "Ancestral 850k" panel of SNPs.

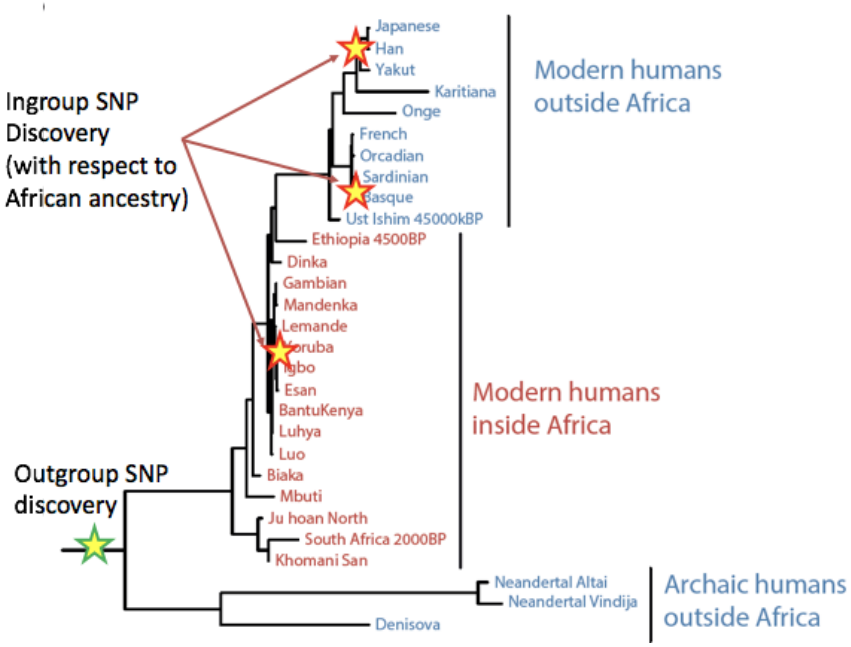

Figure 1. Illustration of outgroup SNP discovery. Adapted from ref. 15.

SNP selection

We first identified a list of 1,531,017 transversion SNPs in four archaic high-coverage genomes: Denisova[16], Altai[17], Vindija[18], and Chagyrskaya[19], using VCF files obtained from http://cdna.eva.mpg.de/neandertal/Vindija/VCF/ and http://cdna.eva.mpg.de/neandertal/Chagyrskaya/VCF/ downloaded 25 July 2018. We created consensus fasta files based on these VCFs using "bcftools consensus -I", and removed all triallelic and transition SNPs.

Separately, we processed a genotype file of the Simons Genome Diversity project[20,21] (https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/variant_set/cteam_extended.v4.maf0.1perc.bim.zip), where we also excluded all triallelic SNPs and transition SNPs, as well as requiring that the allele of the chimpanzee genome panTro2 could be determined. We identified 10,828,817 such transversion SNPs in the set of 359 SGDP individuals ('C' and 'B' groups).

We intersected genotypes of the archaic genomes with the Simons Genome and then restricted to those SNPs that show the exact same two alleles in the 359 SGDP genomes, resulting in a double-ascertained set of 852,068 SNPs autosomal and X-chromosome SNPs. For some purposes, we recommend additionally removing ~30k SNPs polymorphic only in archaic genomes and SGDP Oceanian individuals (Australians, Papuans, and Bougainville Islanders), due to the ~3% Denisovan ancestry identified in these groups[14] creating asymmetry with other non-African ancestry individuals.

Content

20,115 of the SNPs are X-chromosomal, the remainder autosomal. Removing the Chagyrskaya Neandertal genome from the set reduces the number of identified SNPs between the archaic genomes to 1,467,331. We can thus predict that if a fifth Neandertal genome was available it would only result in <30k additional nuclear SNPs for the ancestral diversity set.

Only ~30k of the suggested SNPs are singletons in the SGDP. In fact, >500k of the SNPs have MAF>5% in the SGDP, and >300k have MAF>20%, suggesting that the majority of variants are polymorphic in present-day individuals due to ancestral polymorphism rather than mutations that appeared in Neanderthal and Denisovan ancestors and introgressed into non-African ancestries.

Outgroup ascertainment reduces population genetic biases

We computed F_{ST}[22] for 1) the whole genomes in the SGDP 2) the currently widely used 1240k panel[23,24], and for the new 850k Ancestral SNP panel. Differentiation between African and non-African populations is overestimated for the 1240k, but no similar bias is observed for the Ancestral SNP panel (**Figure 2**).

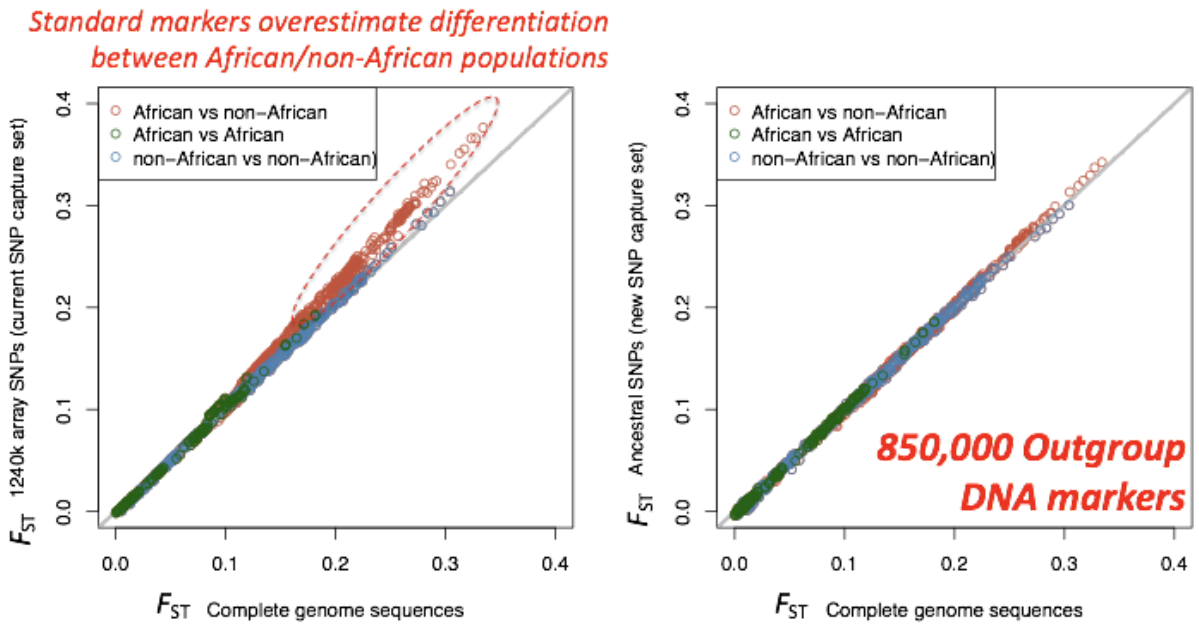

Figure 2. Comparison of SNP ascertainment panels and whole-genome sequence data. (Left) Comparison between complete genome sequences (x-axis) and the 1240k autosomal SNPs for the same individuals. (Right) Comparison between complete genome sequences (x-axis) and the 850k autosomal SNPs for the same individuals.

References

1. Fu, Q., Meyer, M., Gao, X., Stenzel, U., Burbano, H.A., Kelso, J., and Pääbo, S. (2013). DNA analysis of an early modern human from Tianyuan Cave, China. Proceedings of the National Academy of Sciences *110*, 2223–2227.
2. Albrechtsen, A., Nielsen, F.C., and Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. Mol. Biol. Evol. *27*, 2534–2547.
3. Lachance, J., and Tishkoff, S.A. (2013). SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. Bioessays *35*, 780–786.
4. Wang, Y., and Nielsen, R. (2012). Estimating population divergence time and phylogeny from single-nucleotide polymorphisms data with outgroup ascertainment bias. Mol. Ecol. *21*, 974–986.
5. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. Genetics *192*, 1065–1093.
6. Skoglund, P., Malmström, H., Omrak, A., Raghavan, M., Valdiosera, C., Günther, T., Hall, P., Tambets, K., Parik, J., Sjögren, K.-G., et al. (2014). Genomic Diversity and Admixture Differs for Stone-Age Scandinavian Foragers and Farmers. Science *344*, 747–750.
7. Skoglund, P., Thompson, J.C., Prendergast, M.E., Mittnik, A., Sirak, K., Hajdinjak, M., Salie, T., Rohland, N., Mallick, S., Peltzer, A., et al. (2017). Reconstructing Prehistoric African Population Structure. Cell *171*, 59–71.e21.
8. van der Valk, T., Pečnerová, P., Díez-Del-Molino, D., Bergström, A., Oppenheimer, J., Hartmann, S., Xenikoudakis, G., Thomas, J.A., Dehasque, M., Sağlıcan, E., et al. (2021). Million-year-old DNA sheds light on the genomic history of mammoths. Nature *591*, 265–269.
9. Palkopoulou, E., Lipson, M., Mallick, S., Nielsen, S., Rohland, N., Baleka, S., Karpinski, E., Ivancevic, A.M., To, T.-H., Kortschak, R.D., et al. (2018). A comprehensive genomic history of extinct and living elephants. Proc. Natl. Acad. Sci. U. S. A. *115*, E2566–E2574.
10. Bergström, A., Frantz, L., Schmidt, R., Ersmark, E., Lebrasseur, O., Girdland-Flink, L., Lin, A.T.,

Storå, J., Sjögren, K.-G., Anthony, D., et al. (2020). Origins and genetic legacy of prehistoric dogs. Science *370*, 557–564.
11. Fu, Q., Posth, C., Hajdinjak, M., Petr, M., Mallick, S., Fernandes, D., Furtwängler, A., Haak, W., Meyer, M., Mittnik, A., et al. (2016). The genetic history of Ice Age Europe. Nature *534*, 200–205.
12. Hajdinjak, M., Fu, Q., Hübner, A., Petr, M., Mafessoni, F., Grote, S., Skoglund, P., Narasimham, V., Rougier, H., Crevecoeur, I., et al. (2018). Reconstructing the genetic history of late Neanderthals. Nature *555*, 652–656.
13. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W.W., Fritz, M.H.Y., et al. (2010). A Draft Sequence of the Neandertal Genome. Science *328*, 710–722.
14. Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L.F., et al. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature *468*, 1053–1060.
15. Skoglund, P., and Mathieson, I. (2018). Ancient Genomics of Modern Humans: The First Decade. Annu. Rev. Genomics Hum. Genet. *19*, 381–404.
16. Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., et al. (2012). A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science *338*, 222–226.
17. Prufer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43–49.
18. Prüfer, K., de Filippo, C., Grote, S., Mafessoni, F., Korlević, P., Hajdinjak, M., Vernot, B., Skov, L., Hsieh, P., Peyrégne, S., et al. (2017). A high-coverage Neandertal genome from Vindija Cave in Croatia. Science.
19. Mafessoni, F., Grote, S., de Filippo, C., Slon, V., Kolobova, K.A., Viola, B., Markin, S.V., Chintalapati, M., Peyrégne, S., Skov, L., et al. (2020). A high-coverage Neandertal genome from Chagyrskaya Cave. Proc. Natl. Acad. Sci. U. S. A. *117*, 15132–15136.
20. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature *538*, 201–206.
21. Fan, S., Kelly, D.E., Beltrame, M.H., Hansen, M.E.B., Mallick, S., Ranciaro, A., Hirbo, J., Thompson, S., Beggs, W., Nyambo, T., et al. (2019). African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. Genome Biol. *20*, 82.
22. Hudson, R.R., Slatkin, M., and Maddison, W.P. (1992). Estimation of levels of gene flow from DNA sequence data. Genetics *132*, 583–589.
23. Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. Nature *522*, 207.
24. Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. Nature *528*, 499–503.