Characterization of Genomic and Phenotypic Variation in the Khoesan Speakers of Southern Africa

Nicole Da Costa (COL 2023), Dr. Michael McQuillan, Dr. Sarah Tishkoff
Department of Genetics and Biology, Perelman School of Medicine, University of Pennsylvania

Background

The African continent is home to the most genetic diversity across the planet and is where all *homo sapiens* originate from. African ethnic groups' complex history of population interactions and movements as well as changes in climate, diet, and exposure to infectious disease have led to higher levels of genetic diversity across African ethnic groups as opposed to non-African populations.[1]

The Khoesan speakers, known as the San, are hunter-gatherer populations spread across Botswana, Namibia, Angola, Zambia, Zimbabwe, Lesotho, and South Africa. They have the most basal genetic lineages, as they diverged very early from other human populations. They are descendants of the earliest diversification event among extant human populations, at least 100,000 years ago.[2] The Khoesan have unique morphological and physiological features, such as the lightest skin tone out of other African populations, a unique hair texture, and are adapted to a hunter-gatherer lifestyle in a desert climate.

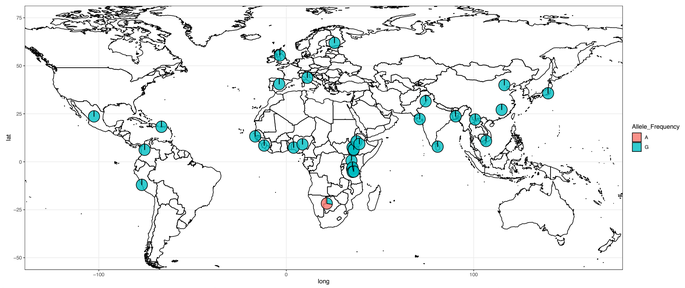

Figure 1. Map showing the allele frequencies in populations across the world of a variant in the KRT74 gene, involved with hair phenotypes. The missense variant only present in the San population

Objective

I looked for genomic signatures of natural selection to identify genetic variants or genes that may have been adaptive in the Khoesan. Through characterizing genetic and phenotypic variation in the Khoesan population and other African ethnic groups, it's possible to better understand human evolutionary history and the complex interaction of genetic and environmental factors in producing phenotypes.[1]

Image 1. The San People

Methods

I used genome-wide data from iHS (integrated haplotype score) and Di Statistic analyses, taking the top 0.01% of SNPs with the highest confidence of possibly experiencing positive selection. The iHS statistic looks to identify evidence of recent positive selection at a locus, and the Di Statistic looks for regions of the genome that are highly differentiated in a population, specifically the San in this case.

I looked at variants at the extremes of the distribution, and looked for enrichment for phenotype function, using EnrichR. I saw enrichment for genes that play a role in skin and hair phenotypes, as well as glucose metabolism and kidney function. I then identified genes that looked interesting and annotated them. I converted base pair numbers into rsid numbers with Kaviar, and then searched for SNPs identified as missense variants using Ensembl.

Results

Term	P-value	Genes
Nail disease	0.046159883	SEH1L;CEP76;RSPO4
Light vs. dark hair color	0.072014937	OCA2;HERC2;TPCN2
Abnormal hair follicle inner root sheath morphology	0.114825739	KRT71;KRT25;PKP3;FOXN1
Abnormality of nail color	0.216911549	TRPS1;ABCA12;KRT6A
Abnormal hair cortex morphology	0.220216038	KRT71;FOXN1
Thick epidermis stratum granulosum	0.220216038	KRT2;ABCA12
Melanocyte differentiation	0.279496479	MEF2C;MITF
Fair hair	0.302699447	PDE4D;UBE3A;UBR1
ADULT syndrome	0.338383064	BRF1;SATB2
Nail-patella syndrome	0.34632646	WNT7A;LDB2;RSPO4
Variegated coat color	0.395684126	OCA2;MITF
Absent eye pigmentation	0.395684126	OCA2;MITF
Skin colour saturation	0.395684126	OCA2;HERC2
Thin skin	0.421847956	KRT71;ALX4;MPV17;FOXN1;RAD23B;SLC9A1
Brittle hair	0.450570651	PKP3;FOXN1
Generalized hypopigmentation of hair	0.473667873	PDE4D;UBE3A;UBR1
Low tan response	0.487375271	OCA2;HERC2;TRPS1;TPCN2
Patchy hypopigmentation of hair	0.502494869	EDNRB;MITF

Table 1. Hair, Skin, Nail Phenotypic Results for the Di Statistics EnrichR analyses

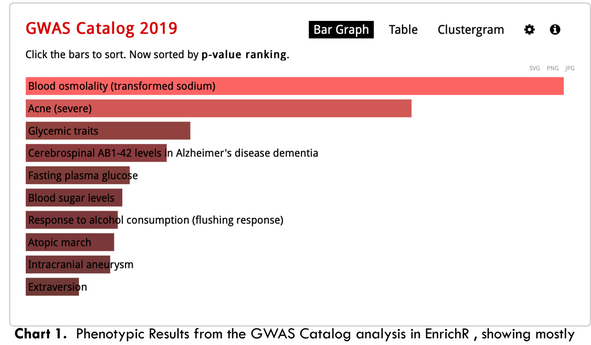

Chart 1. Phenotypic Results from the GWAS Catalog analysis in EnrichR , showing mostly kidney function and glucose metabolism

Conclusions & Future Questions

The Khoesan have specialized adaptations to their environment and have several missense variants associated with important genes worth looking further into, to see the potential pathways they may affect.

- Is the Khoesan's lighter skin color a sign of genetic adaptation or is the Khoesan's lighter skin color an ancestral trait to all human populations?
- Was it advantageous for the Khoesan to have their current phenotypes of skin color and hair texture, or are hair and skin phenotypic changes byproducts of evolution somewhere else in the body?

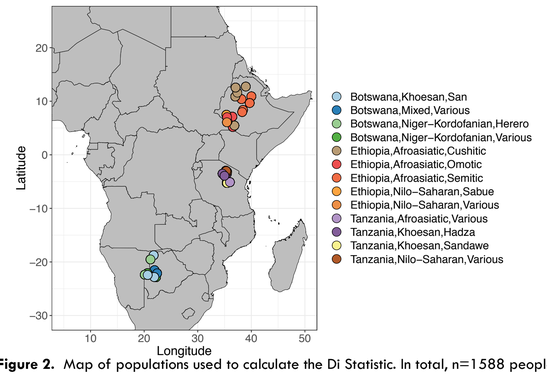

Figure 2. Map of populations used to calculate the Di Statistic. In total, n=1588 people included the in the calculation, and the analysis was performed on ~58 million SNPs

Acknowledgements

I would like to thank PURM and CURF for funding and support, as well as my mentors for the project Dr. Michael McQuillan and Dr. Sarah Tishkoff in the Tishkoff Lab for their guidance, patience, and encouragement while introducing me to the world of genetics.

Contact

Nicole Da Costa
University of Pennsylvania
Email: nicolegd@sas.upenn.edu

References

1. Campbell, M. C., & Tishkoff, S. A. (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annual review of genomics and human genetics*, 9, 403–433. https://doi.org/10.1146/annurev.genom.9.081307.164258
2. Schlebusch, C. M., Skoglund, P., Sjödin, P., Gattepaille, L. M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M. G., Soodyall, H., & Jakobsson, M. (2012). Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science (New York, N.Y.)*, 338(6105), 374–379. https://doi.org/10.1126/science.1227721

Image 1. The San people [Digital image]. (2015). Retrieved September 13, 2020, from https://learn.e-limu.org/topic/view/?t=242&c=45
Figure 1. McQuillan, M. (n.d.). [Map of Frequency Pie Charts for the Missense Variant chr12_52963690 in the KRT74 gene]. Retrieved September 13, 2020.
Figure 2. McQuillan, M. (n.d.). [Map of Populations Used to Calculate the Di Statistic]. Retrieved September 13, 2020.