Exploring genetic variation within the African population in genes associated with hypoxic-ischemic encephalopathy

Megan A. Holborn

August 2023

Click here to view the GitHub Repository

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences
Fakulteit Gesondheidswetenskappe
Lefapha la Disaense tša Maphelo

Make today matter
www.up.ac.za

ICMM
Institute for Cellular and Molecular Medicine

Background

- Hypoxic ischemic encephalopathy (HIE) is a type of brain injury resulting from a restriction in blood flow and oxygen delivery around the time of birth.
- Prior studies have revealed associations between suspected HIE and genes involved in several biological functions, including programmed cell death and inflammation.
- These studies have been performed on predominantly Asian and European populations with no studies published on African populations.
- Here, we aimed to analyse the frequency and predicted effects of genetic variants within HIE-associated genes across African population groups to provide foundational data for future studies on the genetics of HIE in African populations.
- The code generated to complete this analysis is housed in a GitHub repository.

Important Terminology

- **Gene**: A gene is a segment of DNA that contains instructions for producing a specific protein or performing a particular function within an organism.
- **Genetic variant**: A genetic variant is a change in the DNA sequence when compared to a reference or normal sequence. Genetic variants in the DNA encoding a gene can have detrimental effects on the gene's function.
- **Variant allele**: An allele refers to one of several possible versions of a variant that can exist within a population.
- **Variant allele frequency**: The frequency of a variant allele refers to the proportion of individuals in a population who carry that variant allele.
- **Deleterious variant allele**: A variant allele that is more likely to cause disease.

Data acquisition

- Variant count data on African populations was generated in-house by processing open-source genetic data on African populations retrieved from GnomAD v3.1.2.
- Variant count data on European, Asian, and Latin American populations was retrieved from the NCBI ALFA database.
- Variant effect prediction data for variants identified in African population groups was retrieved using the CADD v1.6 tool.
- Information on diseases associated with the variants identified in African populations was retrieved using the Functional Annotation of Variants - Online Resource v2.0 (FAVOR) tool.

Data preparation

The acquired data underwent preparation steps. This process entailed:

- Selecting relevant features of interest
- Removing duplicate entries and handling null values
- Merging of data from several sources if applicable
- Adding additional features
- Restructuring the data in a suitable format for further analysis

Research questions

1. Which African ethnolinguistic population groups are represented by the genetic data, and what are the proportions of samples from Central, Southern, Eastern, and Western Africa?

2. To what extent is genetic variation shared or unique within African subpopulation groups?

3. What is the distribution of rare variants within the different genes, and how do these distributions vary within African ethnolinguistic subpopulations?

4. Which variants identified in African populations have deleterious effect prediction scores and are likely to contribute to disease?

5. How do the frequencies of variants in the genes of interest compare across African subpopulation groups and between Africa and other global populations?

6. How do the frequencies of deleterious variants with known disease associations compare across African subpopulation groups and between Africa and other global populations?

Methods

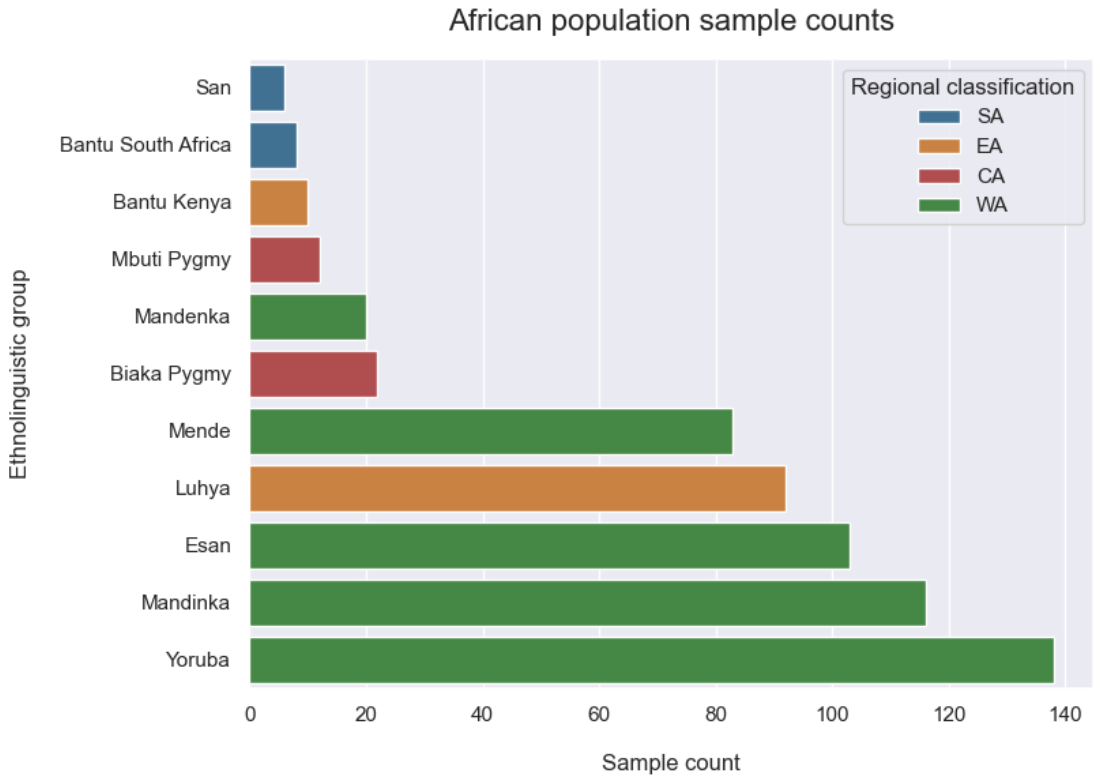
To answer each of the research questions, an analysis of the data was performed. The following notebooks detail the methods used.

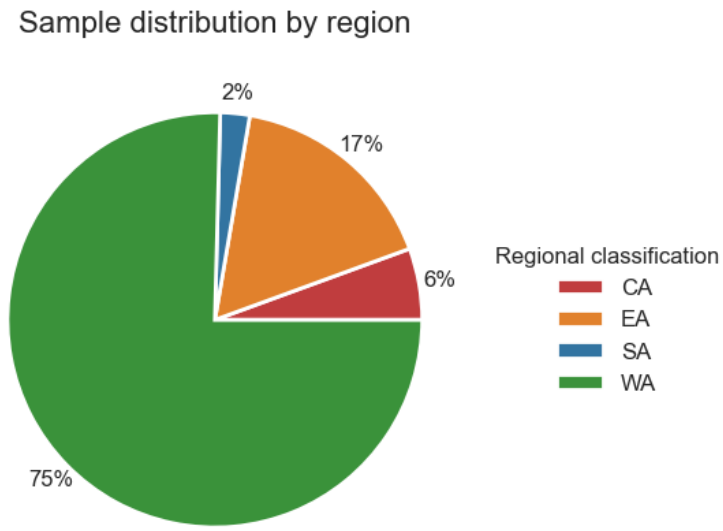
The findings are documented on the upcoming slides.

Findings

Which African ethnolinguistic population groups are represented by the genetic data, and what are the proportions of samples from Central, Southern, Eastern, and Western Africa?

- Genetic data from 610 African individuals, representative of 11 ethnolinguistic populations, was analysed.
- 75% of the individuals were representative of Western African population groups.

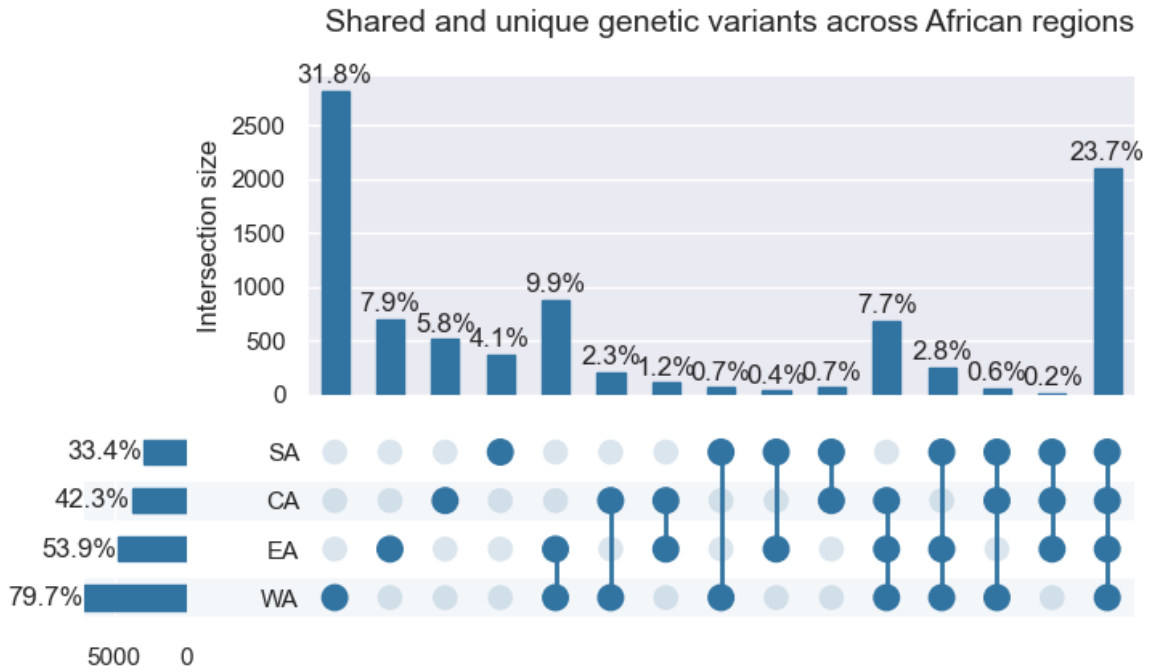




Findings

To what extent is genetic variation shared or unique within African subpopulation groups?

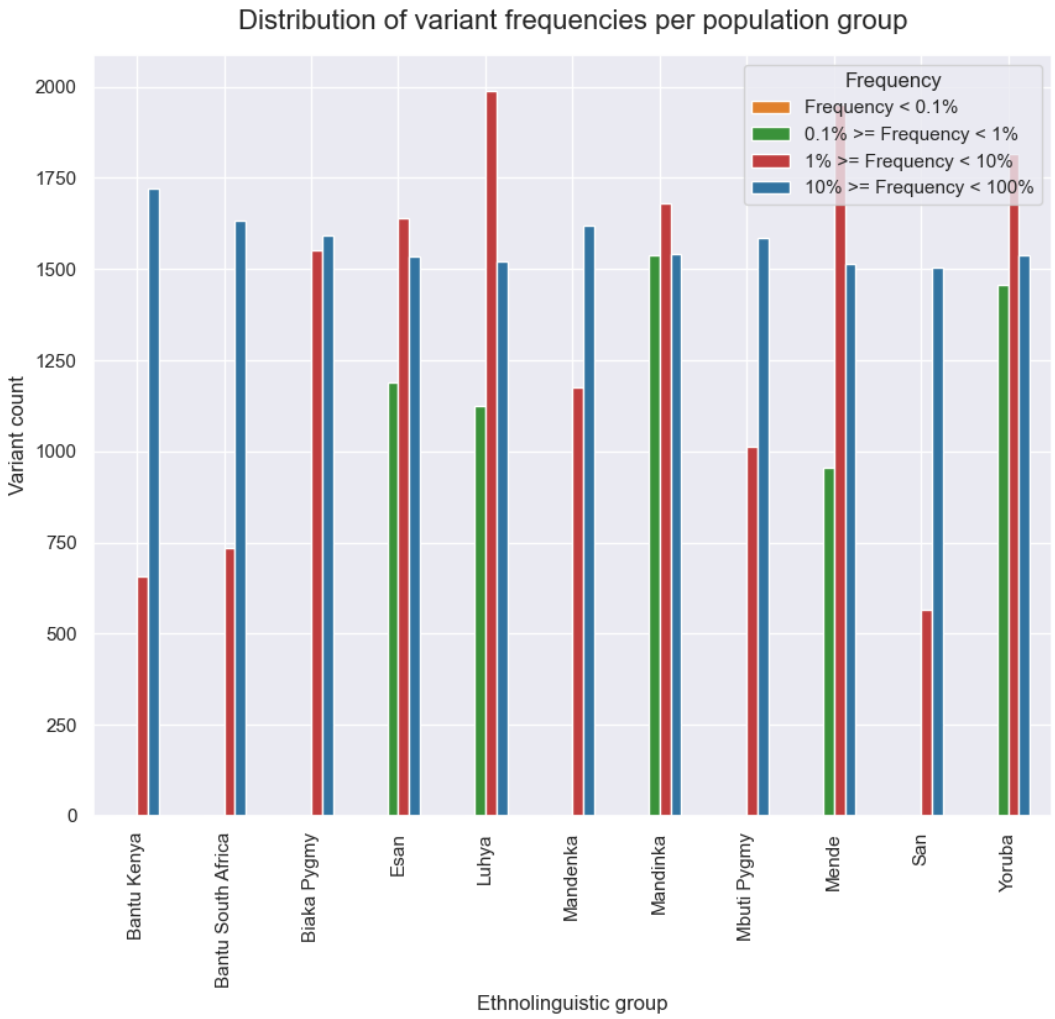
- 23.7% of genetic variants in the genes of interest were shared by populations in Central, Southern, Eastern and Western Africa.
- Western African populations contributed the most unique variants (31.8%), while Southern African populations contributed the least (4.1%).
- A strong positive correlation (> 0.87) was observed between the sample size of a region and the number of unique variants contributed.



Findings

What is the distribution of rare variants within the different genes, and how do these distributions vary within African ethnolinguistic subpopulations?

- 61.5% (n = 5457) of the variants detected in the genes of interest were rare, occurring in less than 1% of Africans overall.
- Certain populations, had a higher proportion of rare variants compared to others – notably the Esan, Mandinka and Yoruba groups.
- These populations also had the highest sample counts. Thus, rare variants might only be reliably detected within larger sample populations.

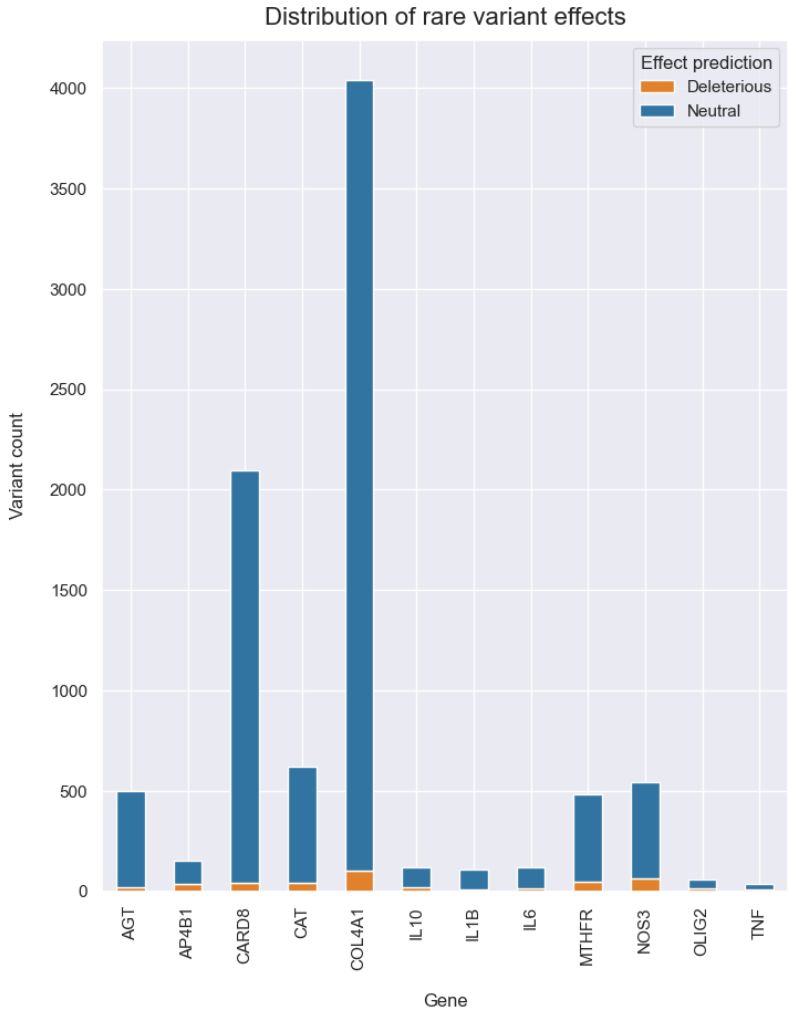


Findings

Which variants identified in African populations have deleterious effect prediction scores and are likely to contribute to disease?

- 4.7% (n = 417) of variants were predicted to be potentially harmful (deleterious).
- The variants with the highest predicted harm values are provided below.

Variant	Genomic position	Ref allele	Alt allele	Gene	Consequence
rs560166628	110150188	C	T	COL4A1	Stop-gain
rs145182838	113898727	T	C	AP4B1	Missense
rs537401710	31575514	C	T	TNF	Missense
rs183156478	206770659	C	T	IL10	Missense
chr1:206769473C-A	206769473	C	A	IL10	Missense

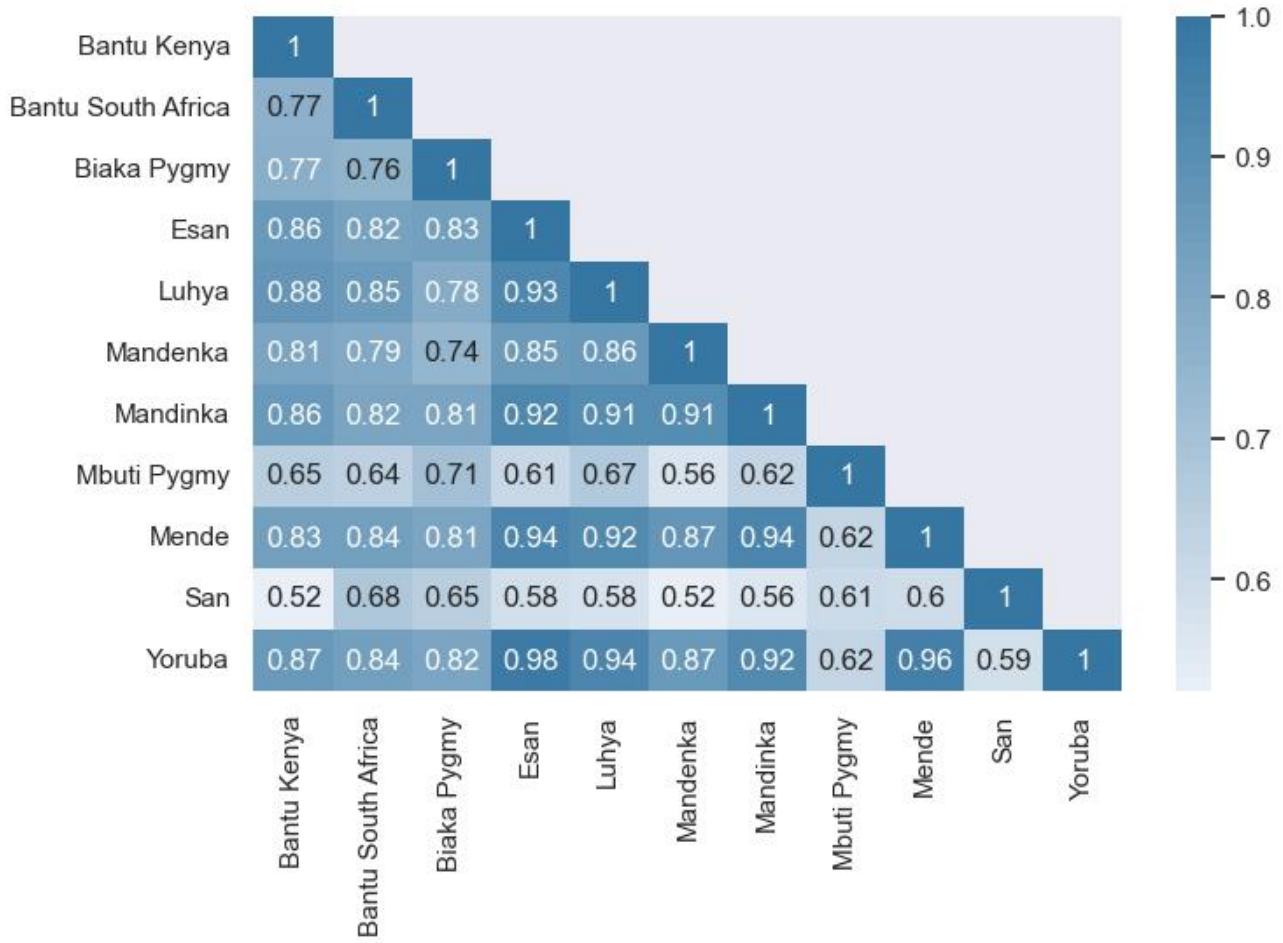


Findings

How do the frequencies of variants in the genes of interest compare across African subpopulation groups?

- Strong correlations (0.8 ≤ CC ≤ 1) were found between Esan, Luhya, Mandenka, Mandinka, Mende and Yoruba groups from Western Africa, and Bantu groups from Central and Southern Africa.

- The San group of Southern Africa and the Mbuti Pygmy group of Central Africa exhibited lower correlations with all other ethnolinguistic groups (0.5 ≤ CC ≤ 0.7).



Findings

How do the frequencies of variants in the genes of interest compare between Africa and other global populations?

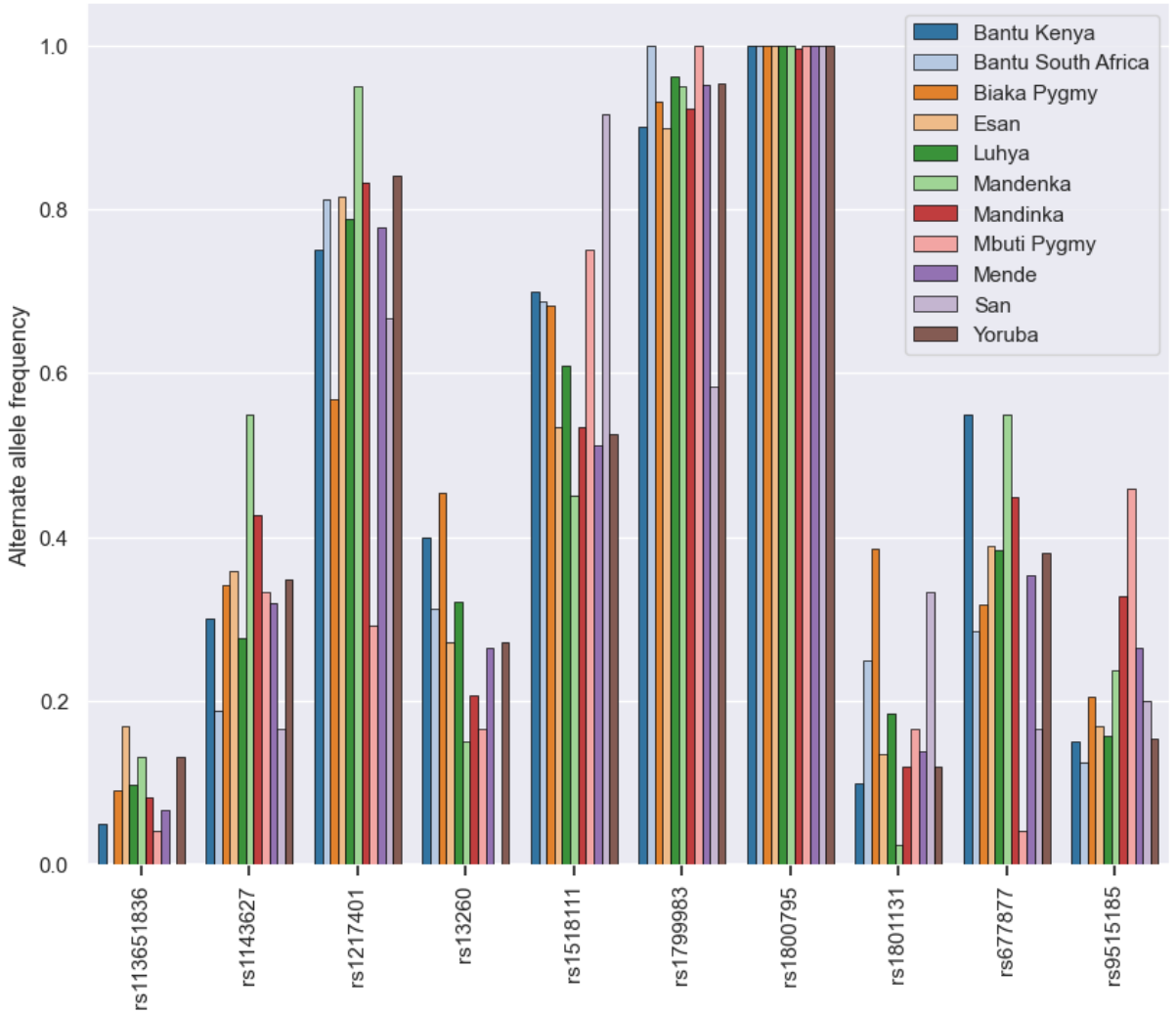
- Frequencies of variants in Africans were moderately correlated with European, East Asian, and South Asian populations ($0.5 \leq CC \leq 0.7$).
- Fisher's Tests were used to determine whether differences in frequency were statistically significant between Africans and other global populations. The results are shown below.

Comparison	**Percentage variants with significant differences ($p<0.05$)**
Africa vs Europe	76.58
Africa vs East Asia	46.26
Africa vs South Asia	37.96
Africa vs Latin America 1	36.75
Africa vs Latin America 2	64.83

Findings

How do the frequencies of deleterious variants with known disease associations compare across African subpopulation groups?

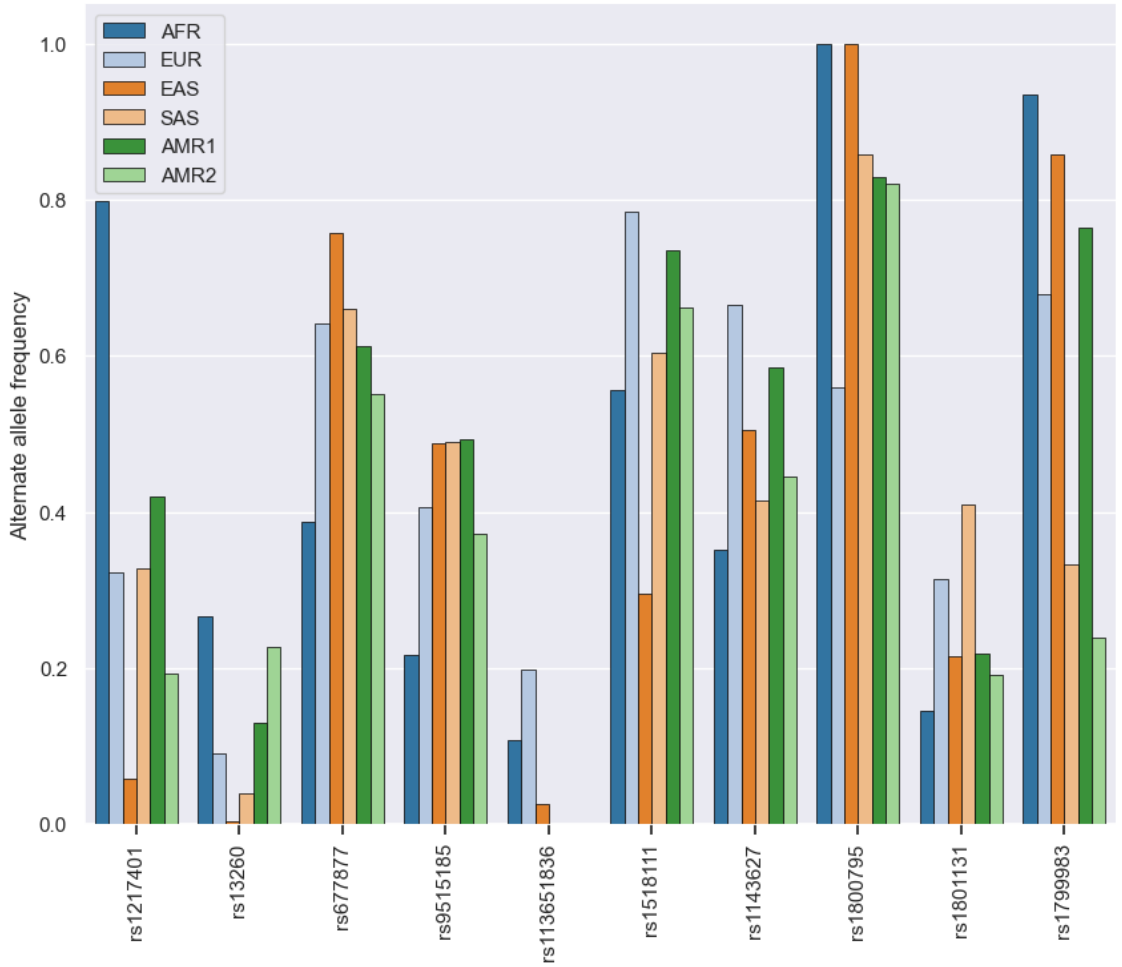
- Common variants with known disease associations showed frequency differences between African subpopulations.
- These variants are linked to conditions like spastic paraplegia, angiopathy, and homocystinuria, some of which overlap with the HIE phenotype.
- Two of the variants - *AP4B1* rs1217401 and *IL6* rs1800795 - were previously associated with HIE.



Findings

How do the frequencies of deleterious variants with known disease associations compare between Africa and other global populations?

- Common variants with known disease associations (described on the previous slide) showed frequency differences between African and European, East Asian, South Asian and Latin American populations.



Limitations

- Population Bias: Given that most samples in this analysis originate from Western Africa (80%), it is important to acknowledge that the genetic variants investigated may be skewed towards Western African populations.
- Sample Size: The sample sizes for some African ethnolinguistic subpopulations, including the San and Bantu groups, were small. Since the number of rare variants identified correlated with the sample size, it is likely that some rare variants within these populations were not captured in this analysis.

Conclusions

- A high amount of rare and population specific genetic variation was found within HIE-associated genes, emphasising the rich genetic diversity within Africa. This genetic diversity coupled with notable differences in frequency between shared genetic variants in African populations and Asians/Europeans, underscore the limitations of directly applying HIE genetic findings from Asian and European populations to Africans. This highlights the need for additional research in an African context on a genetic predisposition to HIE.

- Of the variants identified in HIE-associated genes, a notable amount were predicted to be potentially harmful or had known associations with disease. These variants may be involved in the pathogenesis of HIE or other diseases. The rich genetic diversity of African populations provides a unique opportunity to identify rare, disease-causing variants that might otherwise remain undetected within other global populations.

Acknowledgements

- Funders: The Bill and Melinda Gates Foundation and the South African Medical Research Council
- Supervisors: Prof Michael S Pepper, Prof Fourie Joubert and Dr Juanita Mellet

saMRC advancinglife

BILL & MELINDA GATES foundation

ICMM Institute for Cellular and Molecular Medicine