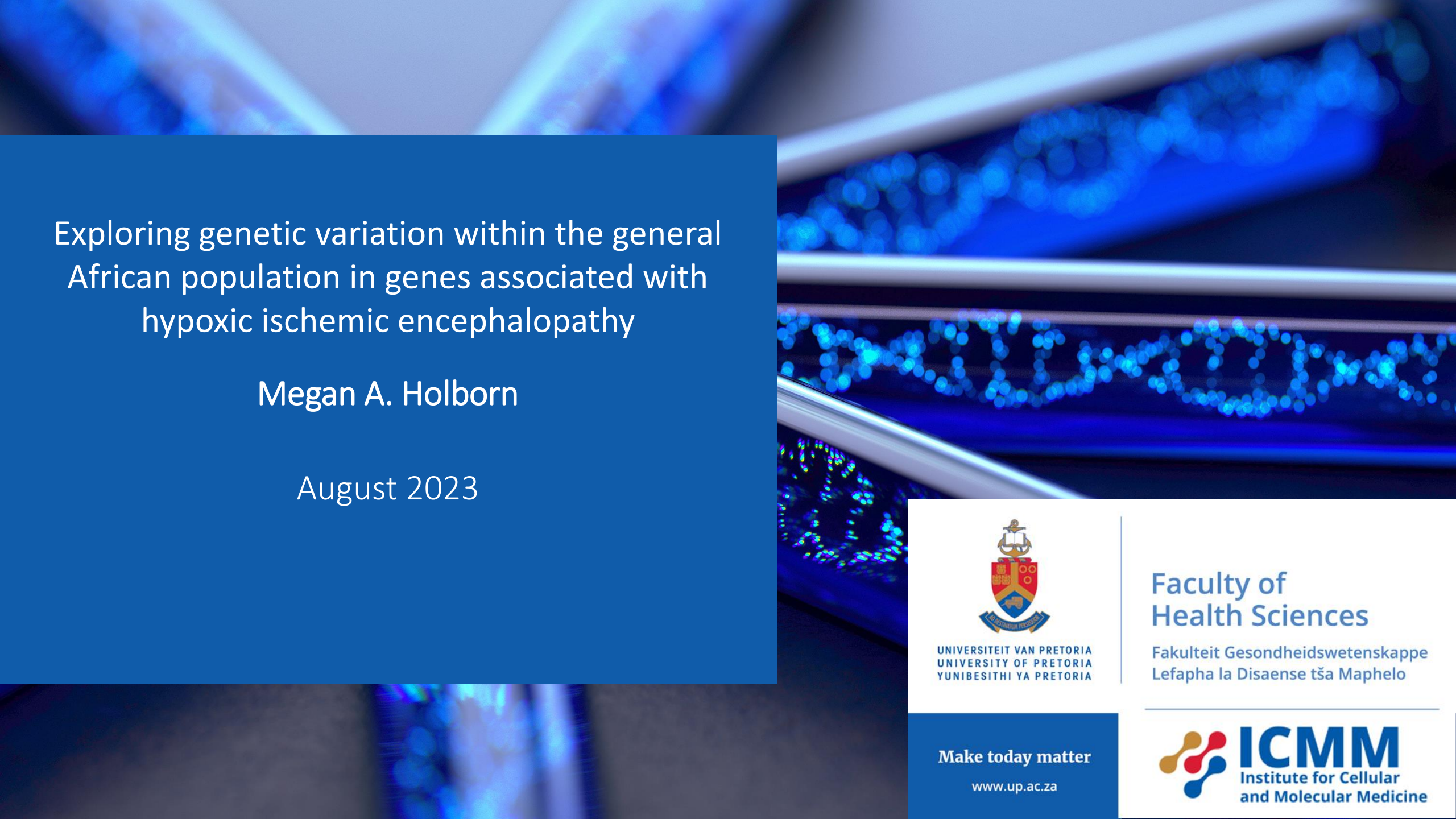
Exploring genetic variation within the general African population in genes associated with hypoxic ischemic encephalopathy

Megan A. Holborn

August 2023



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, inflammation and blood flow homeostasis.
- These studies have been performed on predominantly Asian and European populations with no studies published on African populations.
- African populations exhibit a high amount of genetic diversity, which often renders disease research findings from other global populations less applicable to Africans.
- This analysis aimed to assess the genetic variation within HIE-associated genes across African population groups to provide foundational data for a genetic association analysis involving African HIE patients and controls.
- The code generated to complete this analysis is housed in a GitHub repository.

Data acquisition

- African genomic data from the 1000 Genomes and Human Genome Diversity Project datasets was retrieved from GnomAD v3.1.2 in Variant Call Format (.vcf). A bioinformatics pipeline was utilised to process the African genomic data, resulting in population-stratified variant count information for genetic variants located in the genes of interest.

- Additional data on the impact of variants on gene functionality, processing and translation into proteins (consequences), along with predictions of the potential harm caused by variants (effect predictions) were retrieved from Ensembl and PredictSNP2, respectively.

Data preparation and cleaning

The acquired data underwent preparation and cleaning 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 African regions?

2. To what extent is genetic variation shared or unique within Central, Southern, Eastern and Western African populations?

3. What is the prevalence of rare variants within African populations, and do specific populations exhibit a higher rare variant burden?

4. Which of the variants with rare frequencies in African populations are most likely to contribute to disease, based on predicted effect on gene/protein structure and function?

5. How do frequencies of variants in the studied genes among Africans compare with those of Europeans/Asians?

6. Have any of the genetic variants within the genes of interest previously been associated with HIE? If so, are any of these variants present at significantly different frequencies in Africans compared to the population groups used in the HIE studies?

Methods

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

To answer the research question above, the following methods were utilised:

- To gain an understanding of the different African ethnolinguistic populations represented by the genetic data, the data was grouped according to ethnolinguistic classification. A bar plot was then generated to visualise the sample counts for each population group.
- Furthermore, to depict the distribution of samples across Central, Southern, Eastern, and Western African regions, the data was data was further grouped by region and used to construct a pie chart visually depicting the proportion of samples from each geographic region.

Methods

To what extent is genetic variation shared or unique within Central, Southern, Eastern and Western African populations?

To answer the research question above, the following methods were utilised:

- Genetic variant data was grouped based on geographic region (Central, Southern, Eastern and Western Africa). The variants unique to each region and shared between regions were then determined.
- To compare the shared and unique genetic variation among the different African regions, upset plots were used. Upset plots allow for visual comparison of overlapping or intersecting sets or categories.

Methods

What is the prevalence of rare variants within African populations, and do specific populations exhibit a higher rare variant burden?

To answer the research question above, the following methods were utilised:

- Genetic variant frequencies were segmented into distinct bins. The variant data was then grouped by frequency bin and ethnolinguistic population group. The distribution of variant frequencies within the different population groups was visualised using a bar plot with the grouped data as input.
- The amount and percentage of rare variants within each ethnolinguistic population was calculated, along with the overall amount and percentage of rare variants across all populations.

Methods

Which of the variants with rare frequencies in African populations are most likely to contribute to disease, based on predicted effect on gene/protein structure and function?

To answer the research question above, the following methods were utilised:

- Data pertaining to the impact of variants on gene functionality, processing and translation into proteins (consequences), along with predictions of the potential harm caused by variants (effect predictions) was selected for variants that were rare within African populations.
- The distribution of variants classified as potentially harmful (deleterious) and non-harmful (neutral) within the genes of interest was depicted using a stacked bar plot.
- Variants were arranged based on their level of predicted potential harm, with those with the most deleterious predictions by various algorithms given the highest level of potential harm.

Methods

How do frequencies of variants in the studied genes among Africans compare with those of Europeans/Asians?

To answer the research question above, the following methods were utilised:

- In-house variant frequency data for African populations, and ALFA frequency data on global populations were combined into a unified data frame for subsequent analysis.
- Scatterplots were generated to visually compare allele frequency correlations between genetic variants in African populations and those in Europeans/Asians.
- The degree of correlation between frequency values was measured by calculating the concordance correlation coefficient for respective population pairs under comparison.
- Significant differences in frequency between Africans and Europeans/Asians were measured using two-tailed Fisher's Exact Tests with correction for multiple testing using the Bonferroni method.
- Additional allele frequency comparisons between the in-house and ALFA African frequencies were also conducted to assess in-house data quality.

Methods

Have any of the genetic variants within the genes of interest been previously linked to HIE, and if so, do these variants exhibit significantly different frequencies in Africans compared to the populations studied in HIE research?

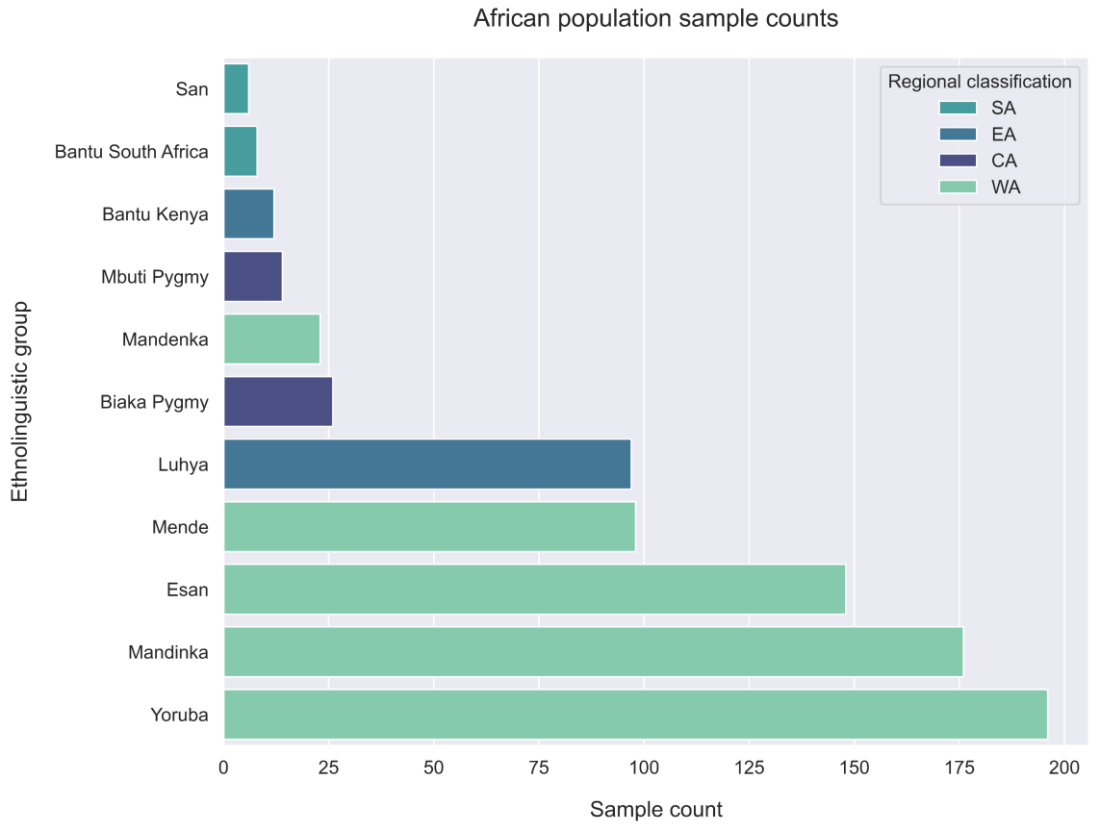
To answer the research question above, the following methods were utilised:

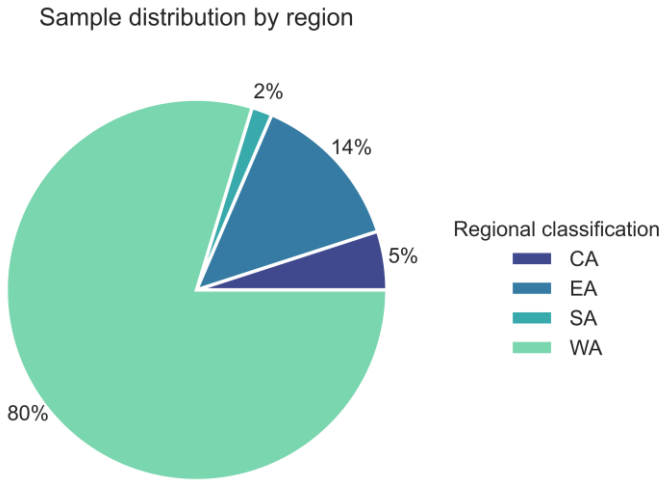
- HIE variants within both the in-house African and ALFA global variant frequency datasets were identified.
- To visually compare the variant allele frequencies for HIE variants among Africans, Europeans and Asians, a heatmap was constructed.
- Significant differences in frequency for HIE variants between Africans and Europeans/Asians were determined using a two-tailed Fisher's Exact Test with multiple testing correction using the Bonferroni method.

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 804 African individuals, representative of 11 ethnolinguistic populations, was analysed.
- 80% of the individuals were representative of Western African population groups.

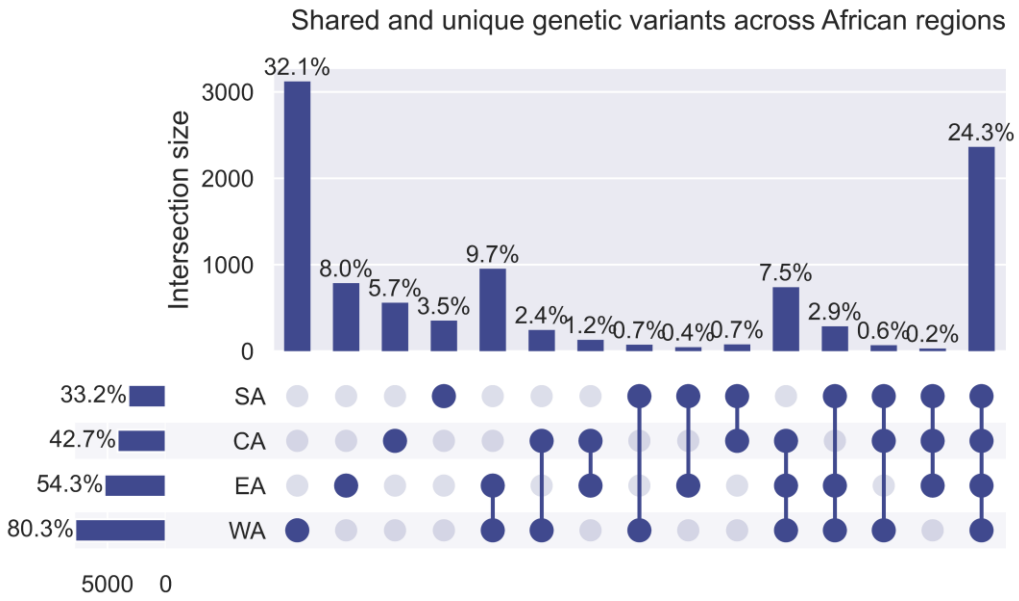




Findings

To what extent is genetic variation shared or unique within Central, Southern, Eastern and Western African populations?

- 24.3% 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 (32.1%) to the analysis, while Southern African populations contributed the least (3.5%).
- A strong positive correlation (> 0.99) was observed between the sample size of a region and the number of unique variants contributed by that region.

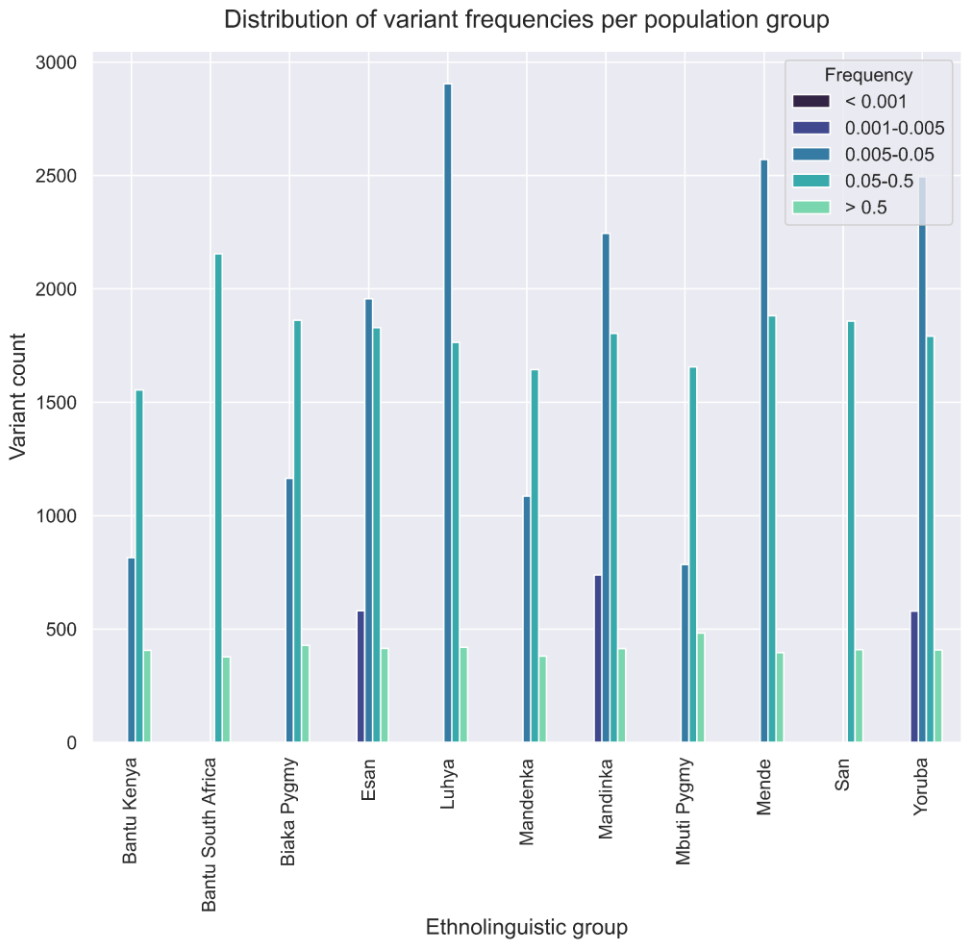


Findings

What is the prevalence of rare variants within African populations in the genes of interest, and do specific populations exhibit a higher rare variant burden?

- 20.2% (n = 1952) of the variants detected in the genes of interest within African populations were rare (frequency < 0.005) within those population groups
- All rare variants were found in the Esan, Mandinka and Yoruba population groups.
- These population groups also have the highest sample counts. This could imply that rare variants might only be reliably detected within larger sample populations.

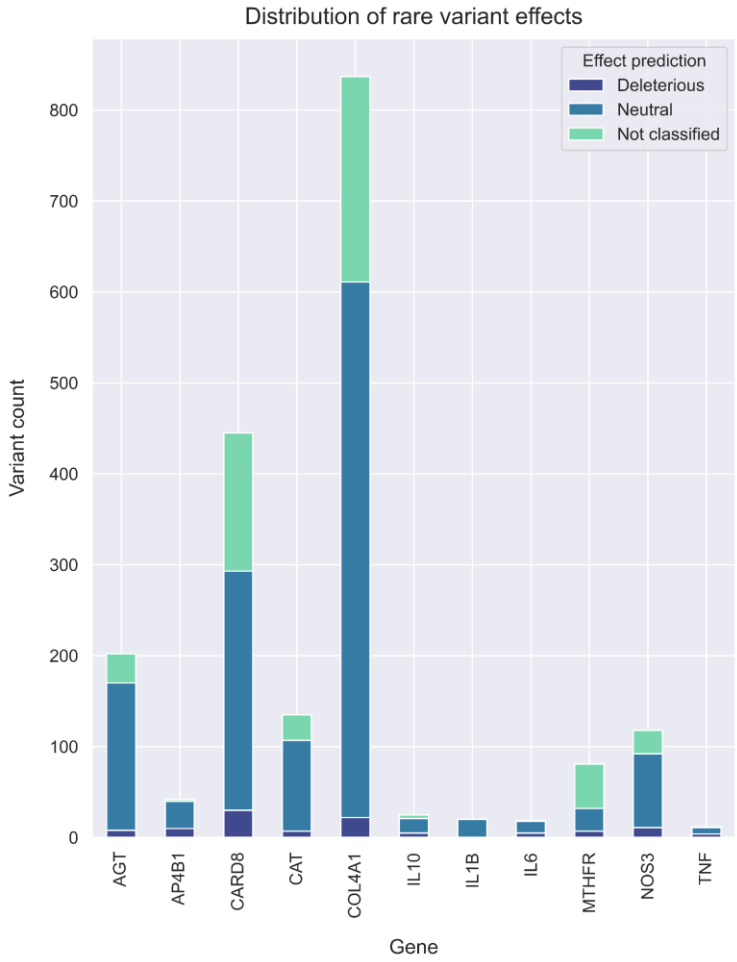
Population group	Rare Variant Count
Esan	581
Mandinka	739
Yoruba	768



Findings

Which of the variants with rare frequencies within the genes of interest in African populations are most likely to contribute to disease, based on predicted effect on gene/protein structure and function?

- 5.6% (n = 109) of rare variants were predicted to be potentially harmful (deleterious).
- Variants predicted to be deleterious were ordered according to their level of potential harm, as indicated by the consensus of deleterious predictions from multiple algorithms. A selection of variants with the highest level of potential harm is provided in the table below.



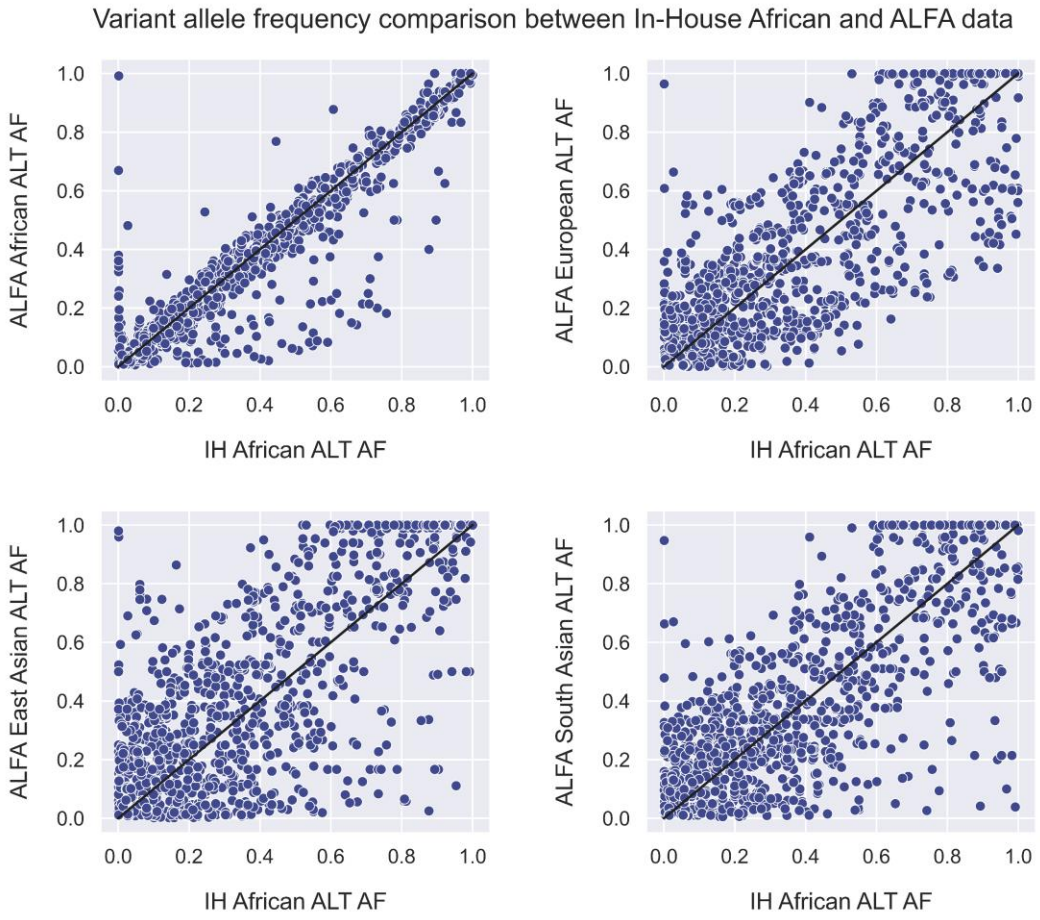
Variant	Genomic position	Reference allele	Alternate allele	Gene	Level of harm	Consequence
rs560166628	110150188	C	T	COL4A1	6	untranslated region
rs145182838	113898727	T	C	AP4B1	6	missense
rs537401710	31575514	C	T	TNF	6	upstream/downstream
rs183156478	206770659	C	T	IL10	6	intronic
chr1:206769473C-A	206769473	C	A	IL10	6	intronic

Findings

How do frequencies of variants in the studied genes among Africans compare with Europeans/Asians?

- In-house African variant frequencies matched NCBI's ALFA results closely, with only 2% of compared variants having statistically significant differences (corrected p-value < 0.05) in frequency.
- However, notable differences were observed between in-house African variant frequencies and ALFA's European, East Asian, and South Asian frequencies (46.5%, 25.3% and 20.54% significant differences, respectively).

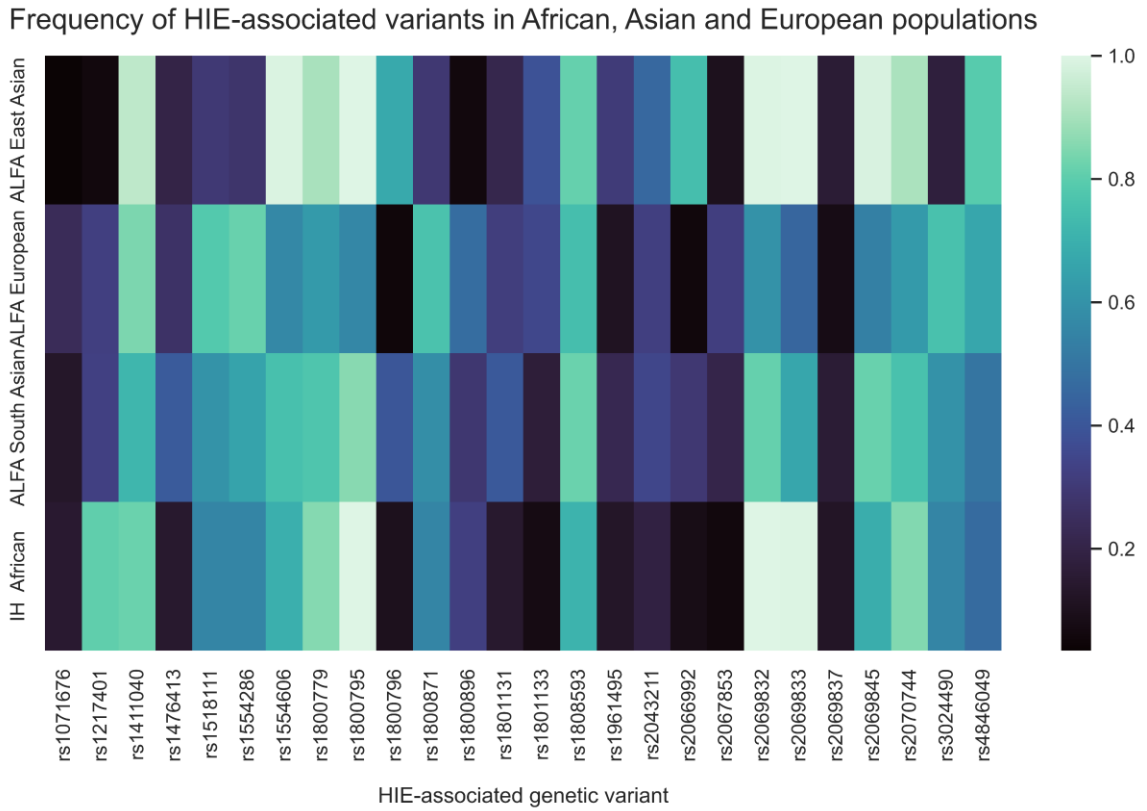
Comparison	Variants with significant differences (%) after Bonferroni correction
IH Africa vs ALFA Europe	46.523105
IH Africa vs ALFA East Asia	25.347690
IH Africa vs ALFA South Asia	20.547331
IH Africa vs ALFA Africa	2.852441



Findings

Have any of the genetic variants identified in Africans been previously linked to HIE, and if so, do these variants exhibit significantly different frequencies in Africans compared to the populations studied in HIE research?

- Of 39 genetic variants associated with HIE, 26 were identified in African populations.
- Of these 26 variants, all except rs1411040, occurred at a significantly different frequency (corrected p-value < 0.05) in Africans vs Europeans/Asians.



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.

- Data quality: While efforts were made to assess the accuracy of the African variant frequencies obtained in-house by comparing the in-house variant frequencies to that from ALFA, it is worth noting that the ALFA data may include samples representative of different African ethnolinguistic population groups from that included in-house. Consequently, while this served as the best available comparison metric, it may not be entirely representative due to these differences.

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 rare variants identified in HIE-associated genes, a notable amount were predicted to be potentially harmful. 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