Variation in protein hydrophobicity across human populations

By: Asim Dave[1], Grace Brannigan[1], Matthew E. B. Hansen[2*]

[1]Center for Computational and Integrative Biology, Rutgers—Camden, NJ, 08102

[2]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104

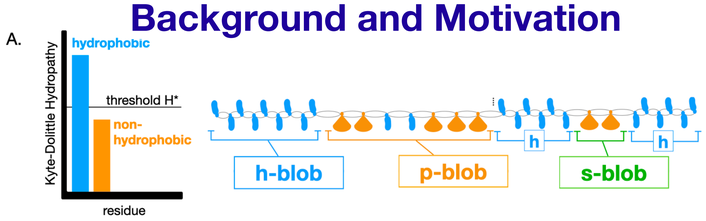
RUTGERS

Perelman SCHOOL OF MEDICINE UNIVERSITY of PENNSYLVANIA

Abstract

Genetic variation shapes diversity in human populations, influencing traits, health, and disease risk. While studies have been conducted to check the impact of genetic changes on protein sequence, we are creating a computational tool that provides population-level genetic diversity in terms of biophysical properties such as contiguous hydrophobicity. Based on haplotype data from the 1000 Genomes Project, the current pipeline translates individual coding sequences into amino acid sequences, from which hydrophobic properties are derived. Little is known about variation of hydrophobicity across human populations. By characterizing this property, it may provide insights into protein stability and interaction, giving a new perspective on human population variation. Looking ahead, the same approach can be extended to other species, studies with regard to sequence properties, including transcription factor binding affinity, and used to answer questions in evolutionary biology.

Background and Motivation



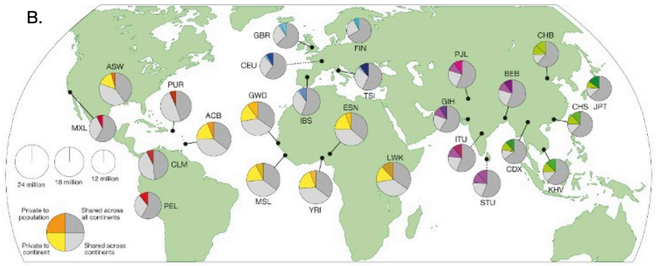


Figure 1. Global genetic variation and the measure of contiguous hydrophobicity. A. Blobulation is a method that groups amino acid residues by their contiguous hydrophobic regions (adapted from [1] Lohia, Hansen, Brannigan, PNAS 2022). **B.** The map shows the patterns of genetic polymorphism in the 1000 Genomes Project. The pie charts depict variants shared across all continents, private to a continent or population (adapted from [2] The 1000 Genomes Project Consortium, Nature 2015).

- Genetic variation in human populations is fundamental to understanding phenotype diversity and its impact on health and disease.
- The 1000 Genomes Project data genotyped 88 million variants from 2,504 individuals across 26 populations.
- The hydrophobic core in globular proteins is a key physical property that contributes to protein stability and interaction.
- By examining variation in contiguous hydrophobicity among genetic variants, we can scan for natural selection and understand evolutionary patterns that are biologically meaningful.

Research questions

Overarching questions:
How variable is contiguous hydrophobicity and is it associated with traits and disease?

Specific questions:

1. How variable is contiguous hydrophobicity across humans?
2. Are there differences in contiguous hydrophobicity between human populations?

Approach

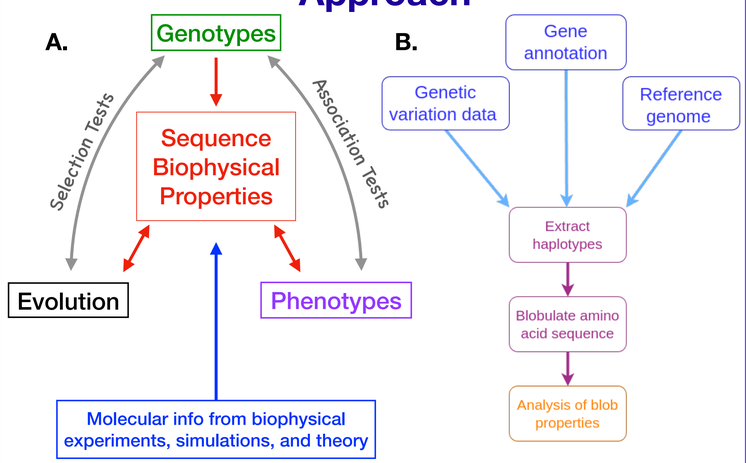


Figure 2. Overview of analysis workflow.
A. Thematic overview. This project will develop a code for extending genotype/phenotype association tests and tests of genetic selection to the biophysical properties of DNA and protein sequences with molecular effects.
B. Workflow completed to date. This panel illustrates the pipeline steps: genetic variation data are provided in VCF format, gene annotation in BED format, and haplotypes are extracted using bedtools and bcftools. Blobulation is performed using blobulator command line interface [3].

Summary

- Mean contiguous hydrophobicity varies by 0.17% across populations.
- Proteins show population-level variation in contiguous hydrophobic regions. Some proteins differ between populations, while others remain conserved.
- On average across the initial set of APOE proteins, the standard deviation in the fraction of contiguous hydrophobicity relative to the population mean for APOE is 0.8%, 1.0%, and 1.1% within the Han Chinese (CHB), British (GBR), and Yoruban (YRI) populations.

Future Work

- Apply this method to population-level variation by the fraction of variation occurring in alpha helices and beta sheets in secondary structure.
- Explore the regulation of transcription factor expression across populations by examining binding sites within non-coding sequences associated with non-functional proteins.
- Conduct a proteome-wide association study between traits and biophysical sequence properties.

Acknowledgements

- Rutgers Office of Advanced Research Computing (OARC)
- NRT, NSF DGE 2152059
- NIH 1R35GM134957

References

1. Lohia, R., Brannigan, G. (2022). Contiguously hydrophobic sequences are functionally significant throughout the human exome. *Proceedings of the National Academy of Sciences*, 119(12), e2116267119. https://doi.org/10.1073/pnas.2116267119
2. The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature, 526, 68–74. https://doi.org/10.1038/nature1539
3. Pitman, C., Santiago-McRae, E., Lohia, R., Lamb, R., Bassi, K., Riggs, L., Joseph, T. T., Hansen, M. E. B., & Brannigan, G. (2025). Revealing protein sequence organization via contiguous hydrophobicity with the blobulator toolkit. bioRxiv

Population Differences in Contiguous Protein Hydrophobicity

Table 1. Variation in protein contiguous hydrophobicity in humans. The table presents 15 proteins and variation in the fraction of peptide in a contiguous hydrophobic blobs (F_h) across populations. The mean percent relative difference for these proteins is 0.17% across populations.

Protein	Total protein size	Mean F_h	StdDev [F_h]	% Relative difference
APOE	317	37.4	0.4	1.0
BDNF	247	50.2	0.0	0.0
BDNF-203	247	50.2	0.0	0.0
CLDN18	261	67.8	0.0	0.0
GSTO2	243	54.7	0.0	0.0
HFE	348	43.1	0.0	0.0
HTT	3144	58.7	0.1	0.1
LMNA	664	32.8	0.0	0.1
MAPT	833	32.5	0.1	0.3
MAPT-205	776	30.2	0.1	0.3
MAPT-212	758	31.0	0.1	0.3
PCSK9	692	56.8	0.1	0.1
POU2F3	436	42.4	0.1	0.3
PRNP	253	36.0	0.0	0.0
TRPV2	764	55.2	0.1	0.2

Table 2. Variation in protein contiguous hydrophobicity between populations. The table presents variation in F_h and counts of residues between populations.

Protein	Population	Mean F_h[1]	StdDev[F_h]	Mean N_h[2]	StdDev[N_h]
APOE	CHB	37.5	0.3	118.8	0.9
	GBR	37.4	0.4	118.5	1.2
	YRI	37.4	0.4	118.3	1.3
MAPT	CHB	32.5	0.1	270.3	0.4
	GBR	32.5	0.1	271	0.8
	YRI	32.5	0.1	270.8	0.5
TRPV2	CHB	55.3	0.1	422.1	0.4
	GBR	55.2	0.0	422	0
	YRI	55.2	0.2	421.6	1.1
PCSK9	CHB	56.8	0.0	393.1	0.2
	GBR	56.8	0.0	393.1	0.3
	YRI	56.9	0.1	393.5	0.8

[1]F_h: N_h/N_{tot}, where N_{tot} is the protein size.
[2]N_h: The count of amino acids in h blobs.

Contiguous Hydrophobicity Distribution of Selected Proteins by Population

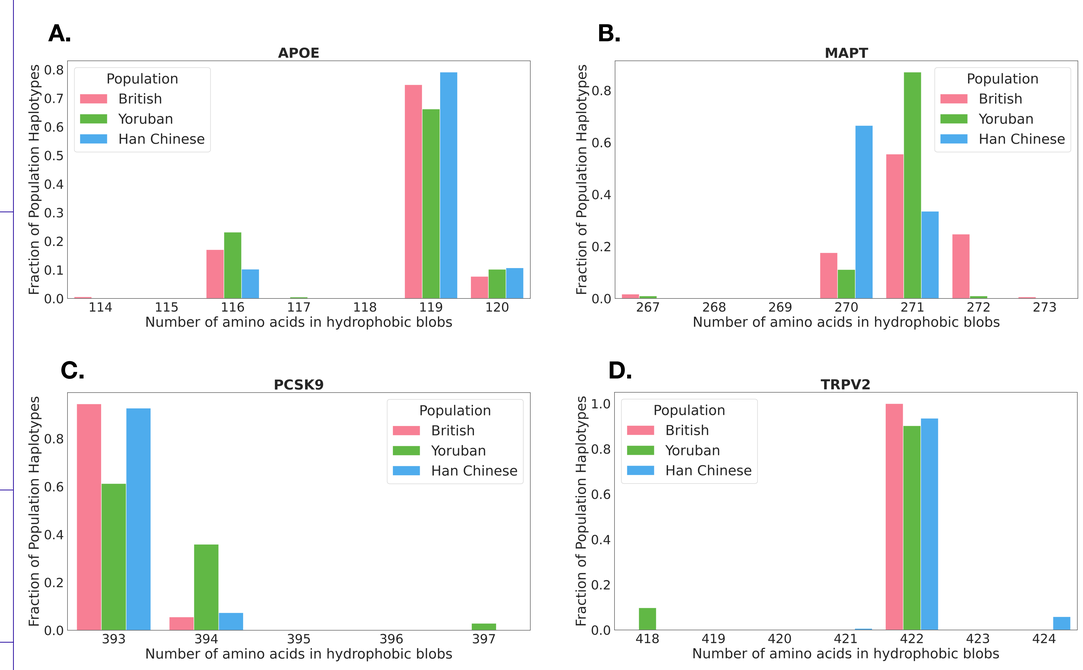


Figure 3. Population-level difference in protein contiguous hydrophobicity. The figure shows the distribution of haplotypes from individuals across Han Chinese (CHB), Yoruban (YRI), and British (GBR) populations for specific proteins. For blobulation, a hydrophobicity cut-off of 0.4 was applied, with a minimum blob length threshold set to 4 residues.