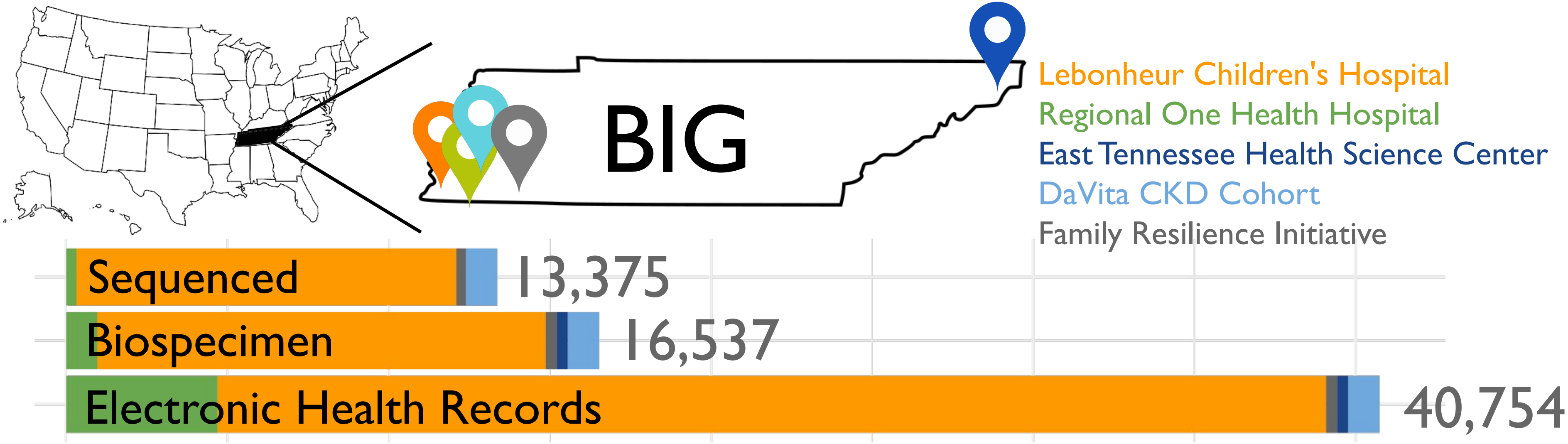
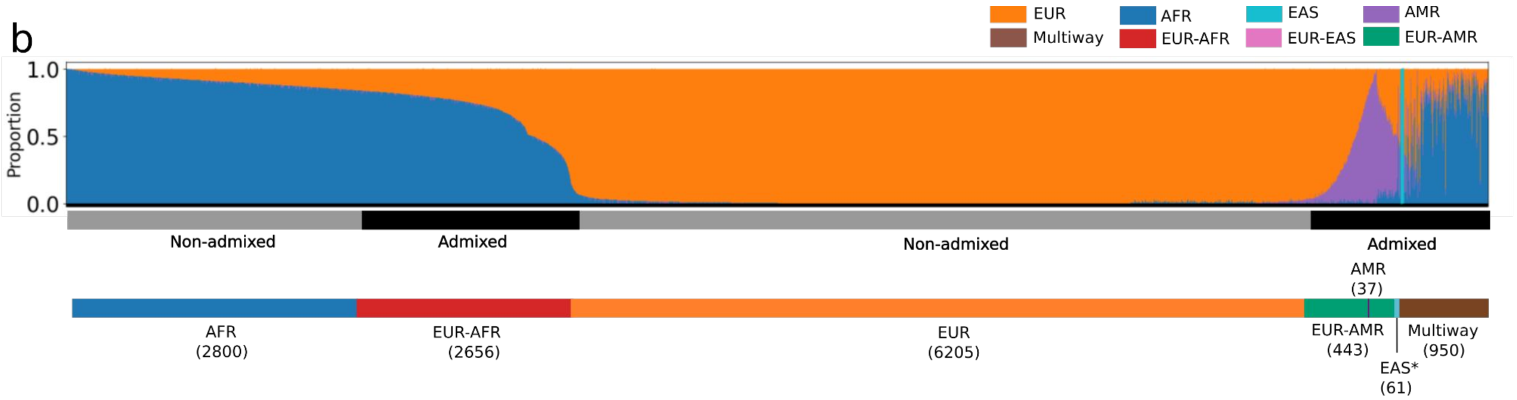
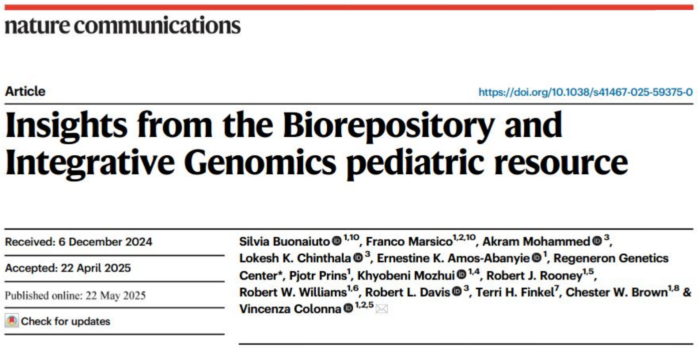
The Biorepository for Integrative Genomics initiative



Ancestry inference in BIG



- Four continental level ancestries
- Considerable proportion of admixed individuals
- High genetic diversity geographically concentrated



nature communications

Article

https://doi.org/10.1038/s41467-025-59375-0

Insights from the Biorepository and Integrative Genomics pediatric resource

Received: 6 December 2024

Accepted: 22 April 2025

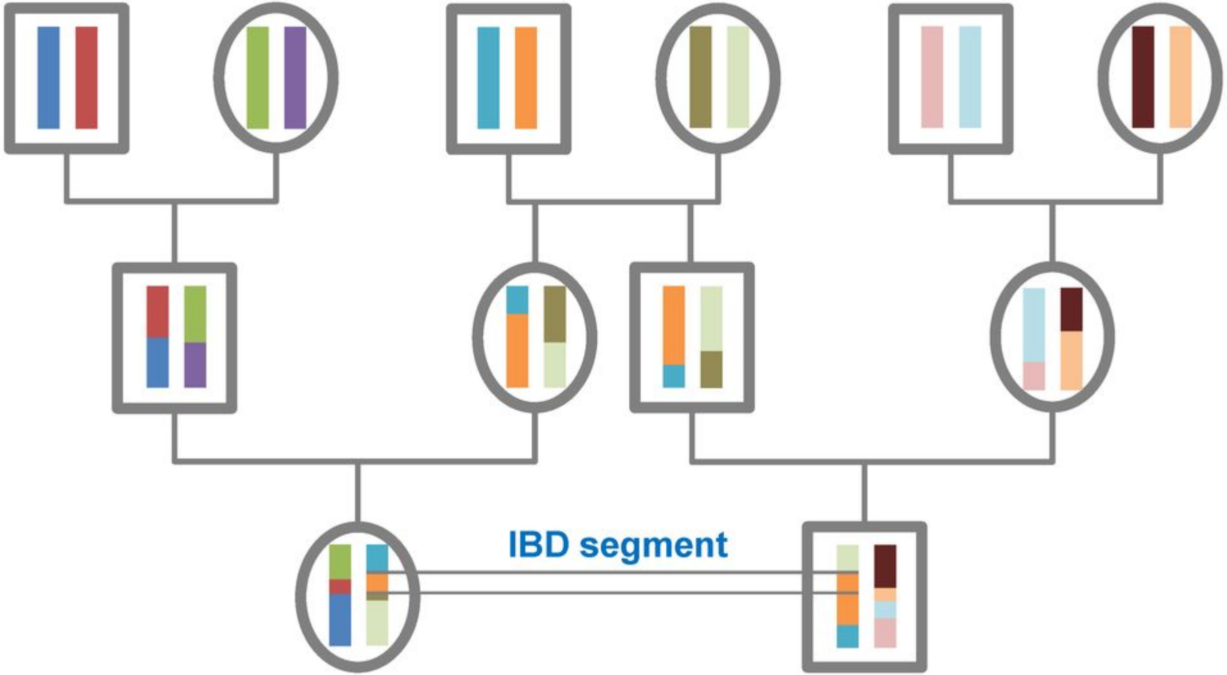
Published online: 22 May 2025

Check for updates

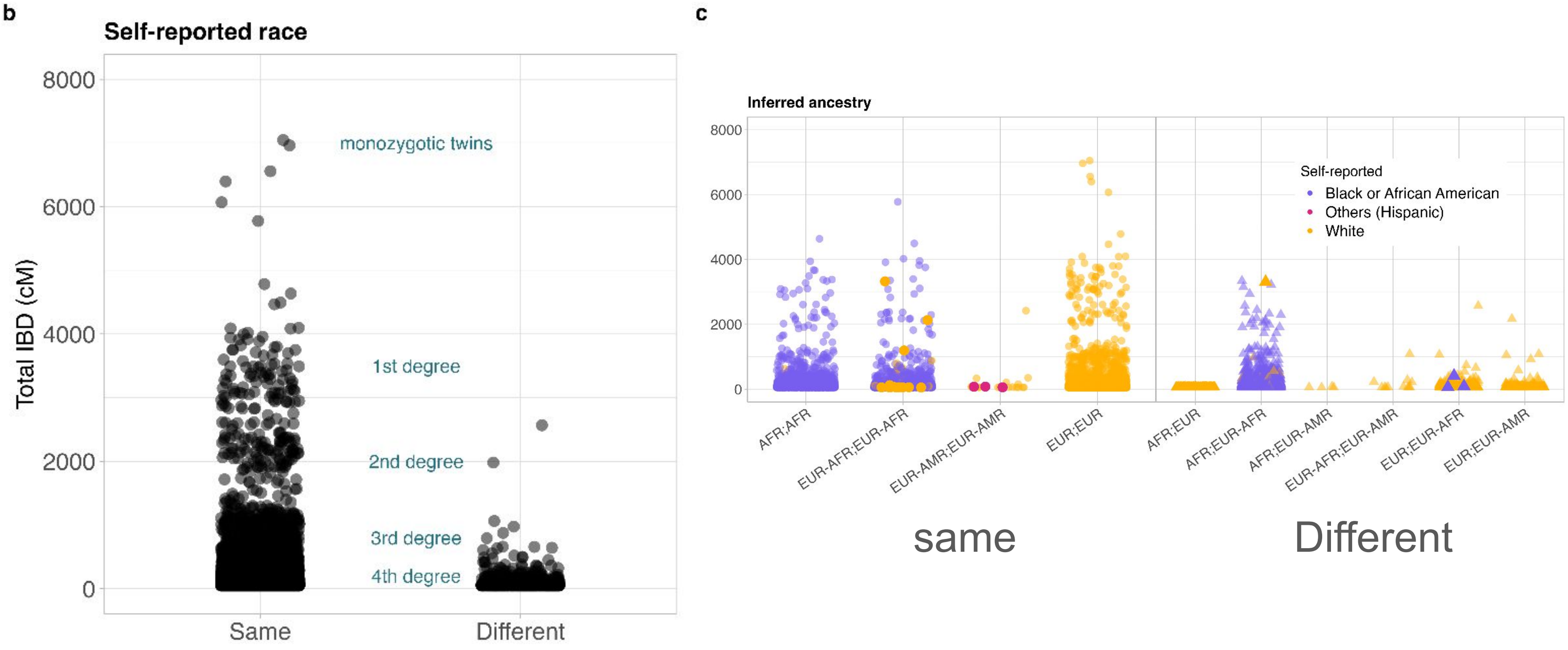
Silvia Buonaiuto[1,10], Franco Marsico[1,2,10], Akram Mohammed[3], Lokesh K. Chinthala[3], Ernestine K. Amos-Abanyie[1], Regeneron Genetics Center*, Pjotr Prins[1], Khyobeni Mozhui[1,4], Robert J. Rooney[1,5], Robert W. Williams[1,6], Robert L. Davis[3], Terri H. Finkel[7], Chester W. Brown[1,8] & Vincenza Colonna[1,2,5]

Buonaiuto* , Marsico*, Nat Comm, 2025

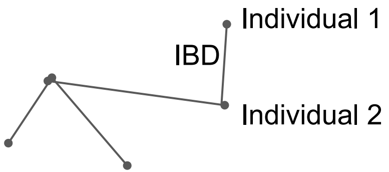
Identity by Descent (IBD)

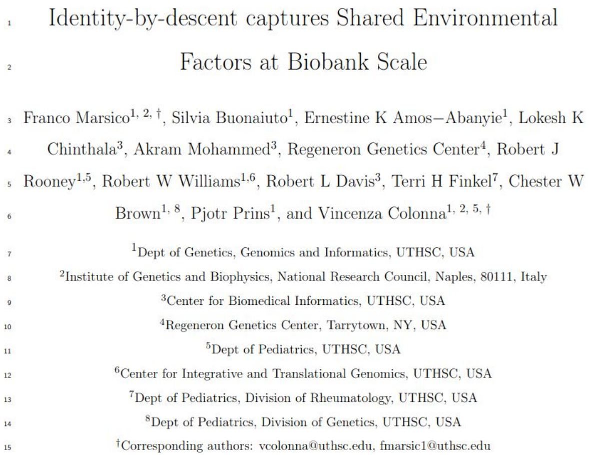


Self-reported race and genetic ancestry through the lens of relatedness



From pairwise IBD to population structure



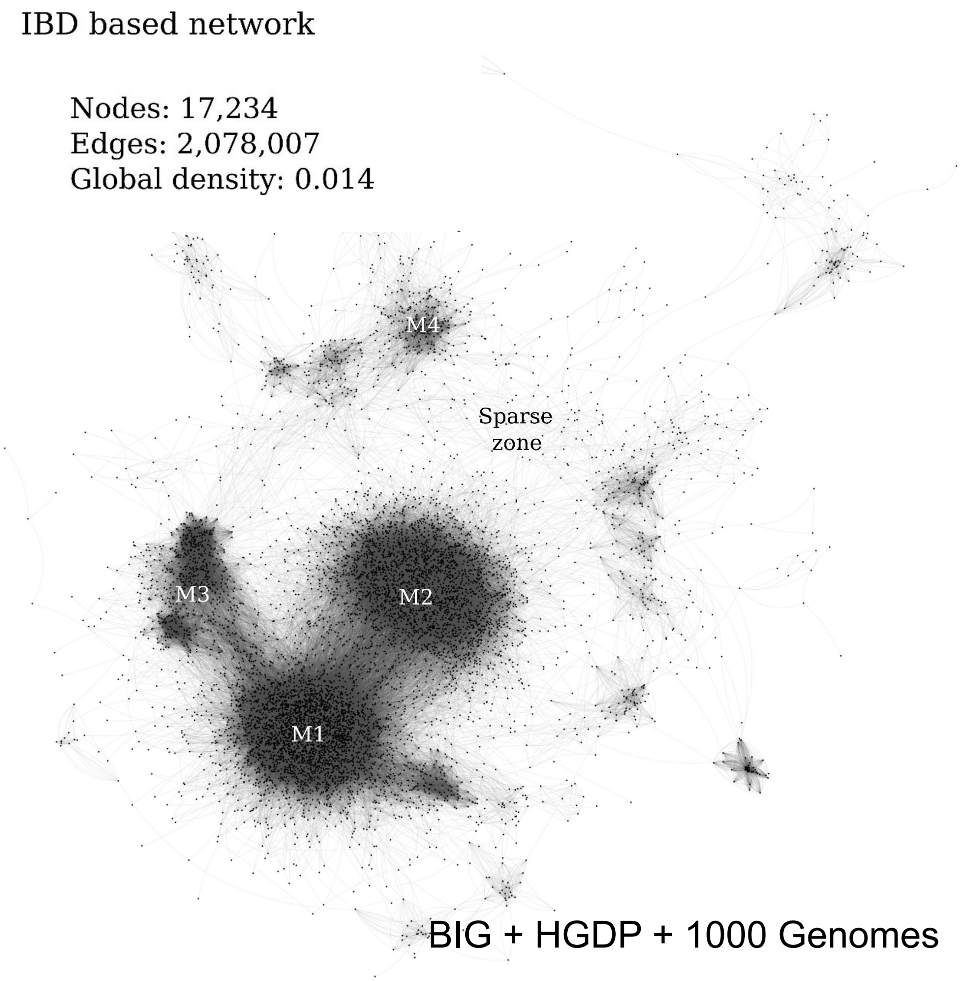


Identity-by-descent captures Shared Environmental Factors at Biobank Scale

Franco Marsico[1, 2, †], Silvia Buonaiuto[1], Ernestine K Amos–Abanyie[1], Lokesh K Chinthala[3], Akram Mohammed[3], Regeneron Genetics Center[4], Robert J Rooney[1,5], Robert W Williams[1,6], Robert L Davis[3], Terri H Finkel[7], Chester W Brown[1, 8], Pjotr Prins[1], and Vincenza Colonna[1, 2, 5, †]

[1]Dept of Genetics, Genomics and Informatics, UTHSC, USA
[2]Institute of Genetics and Biophysics, National Research Council, Naples, 80111, Italy
[3]Center for Biomedical Informatics, UTHSC, USA
[4]Regeneron Genetics Center, Tarrytown, NY, USA
[5]Dept of Pediatrics, UTHSC, USA
[6]Center for Integrative and Translational Genomics, UTHSC, USA
[7]Dept of Pediatrics, Division of Rheumatology, UTHSC, USA
[8]Dept of Pediatrics, Division of Genetics, UTHSC, USA
[†]Corresponding authors: vcolonna@uthsc.edu, fmarsic1@uthsc.edu

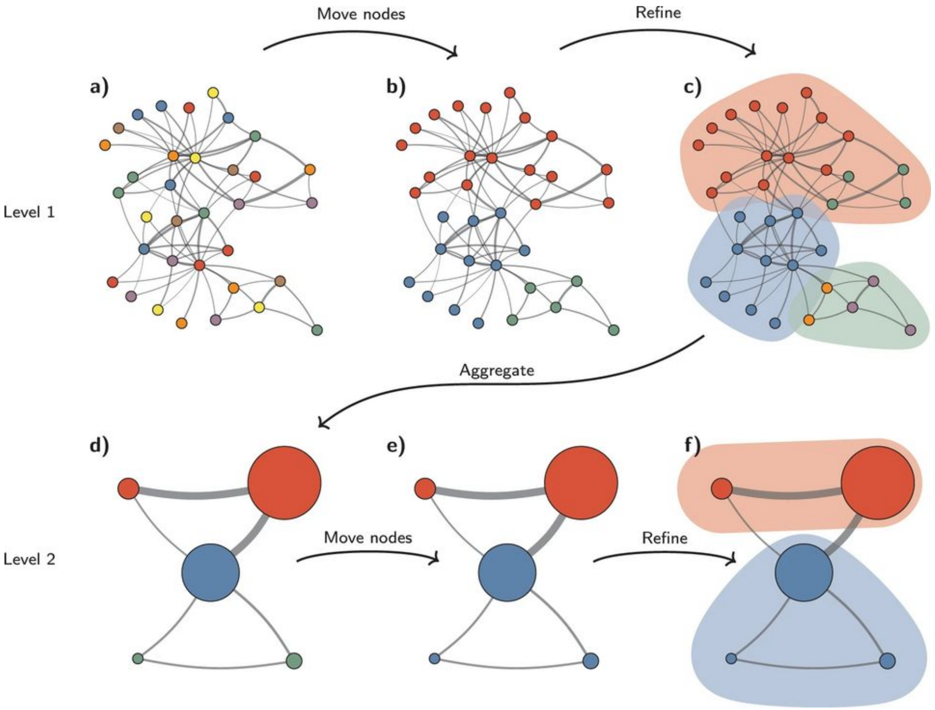
https://www.biorxiv.org/content/10.1101/2025.05.03.652048v1.full


IBD based network

Nodes: 17,234
Edges: 2,078,007
Global density: 0.014

BIG + HGDP + 1000 Genomes

Leiden algorithm for community detection

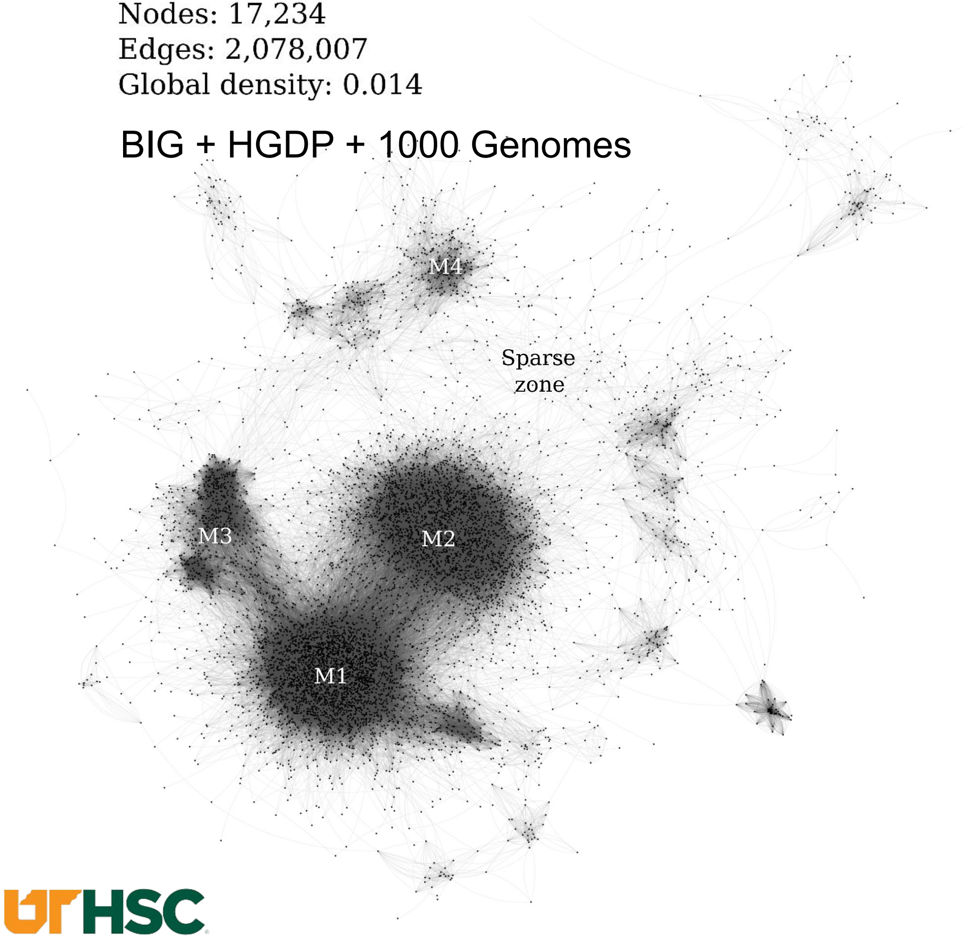


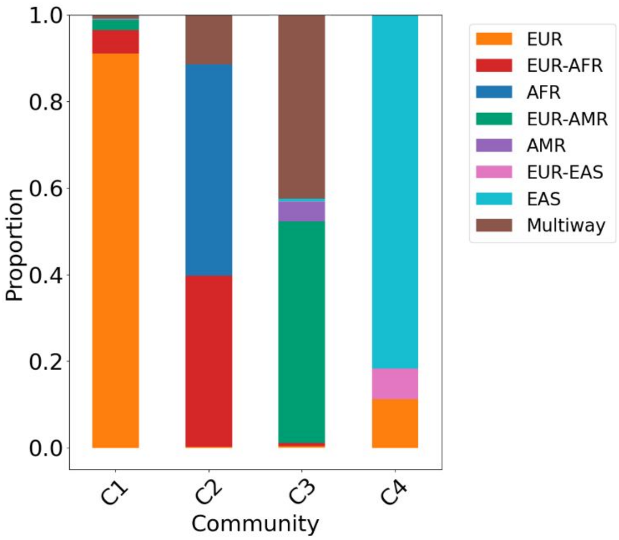
Beyond ancestry: grouping people based on how much genome they actually share

IBD based network

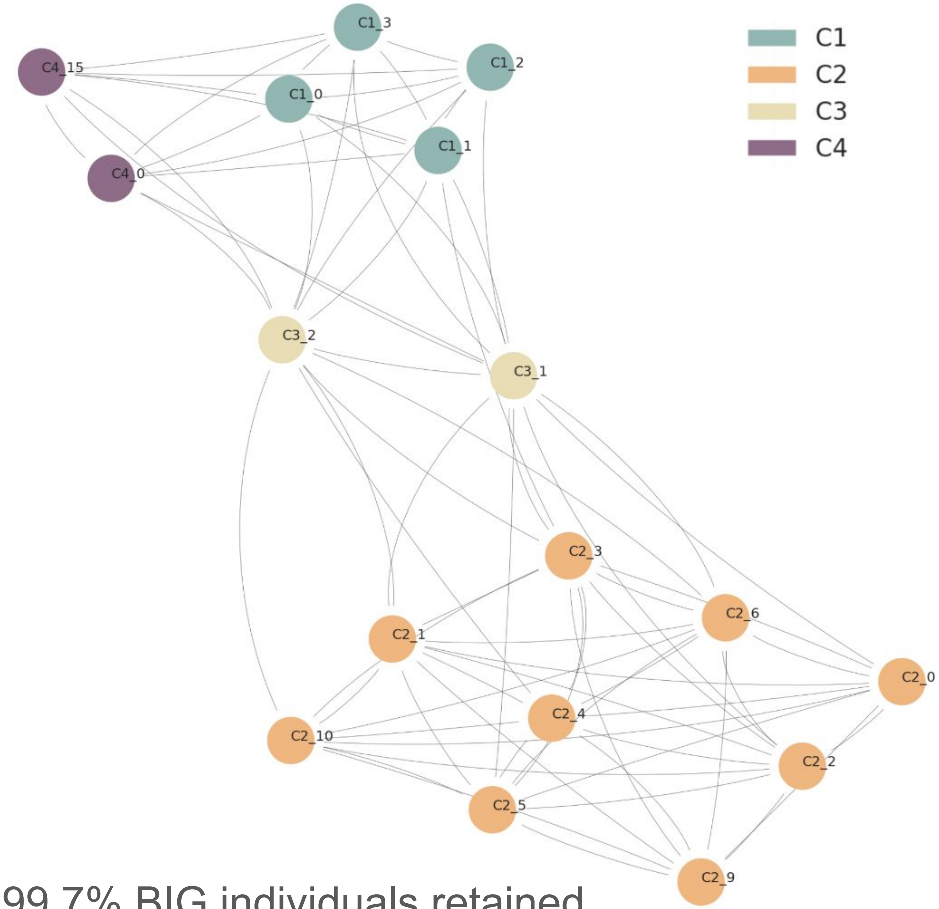
Nodes: 17,234
Edges: 2,078,007
Global density: 0.014

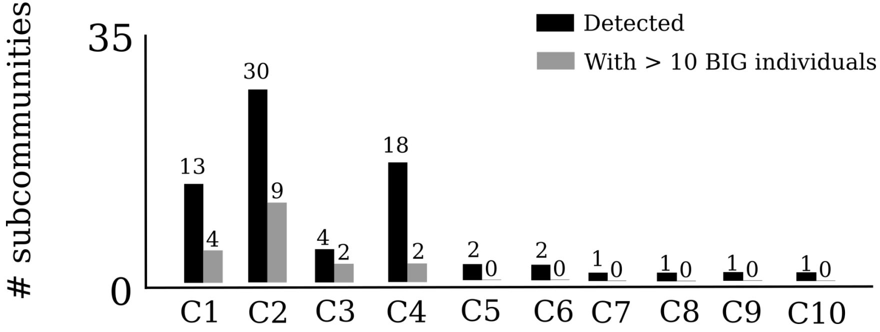
BIG + HGDP + 1000 Genomes





IBD *versus* ancestry

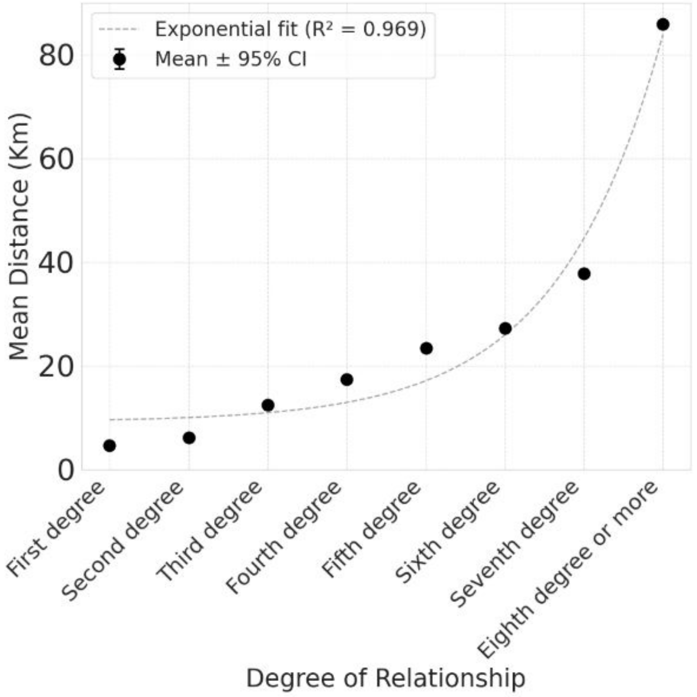


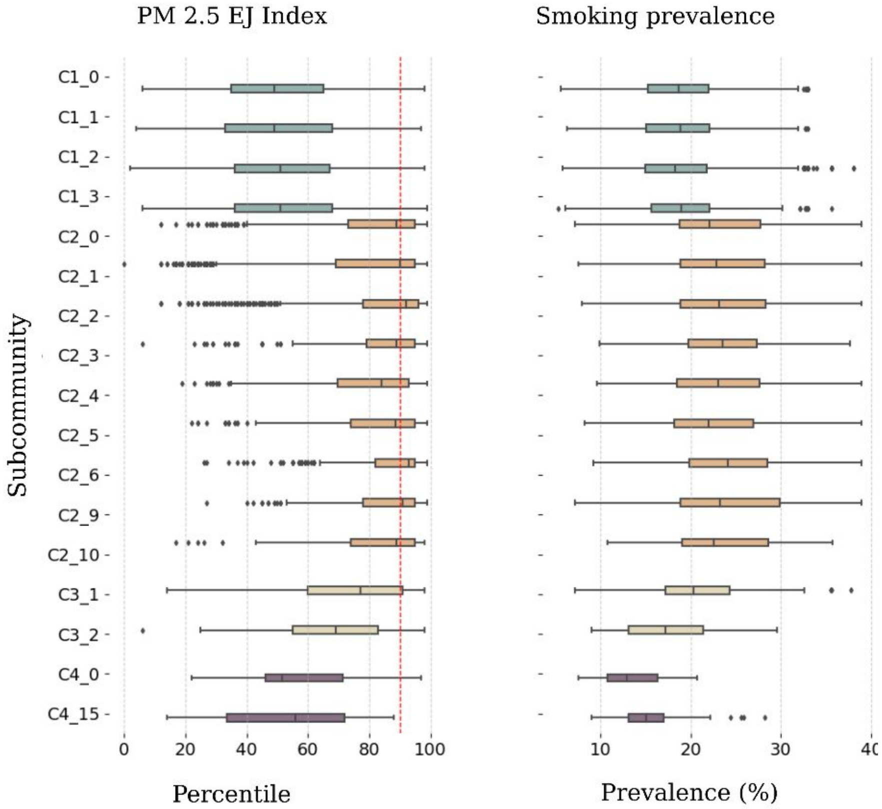


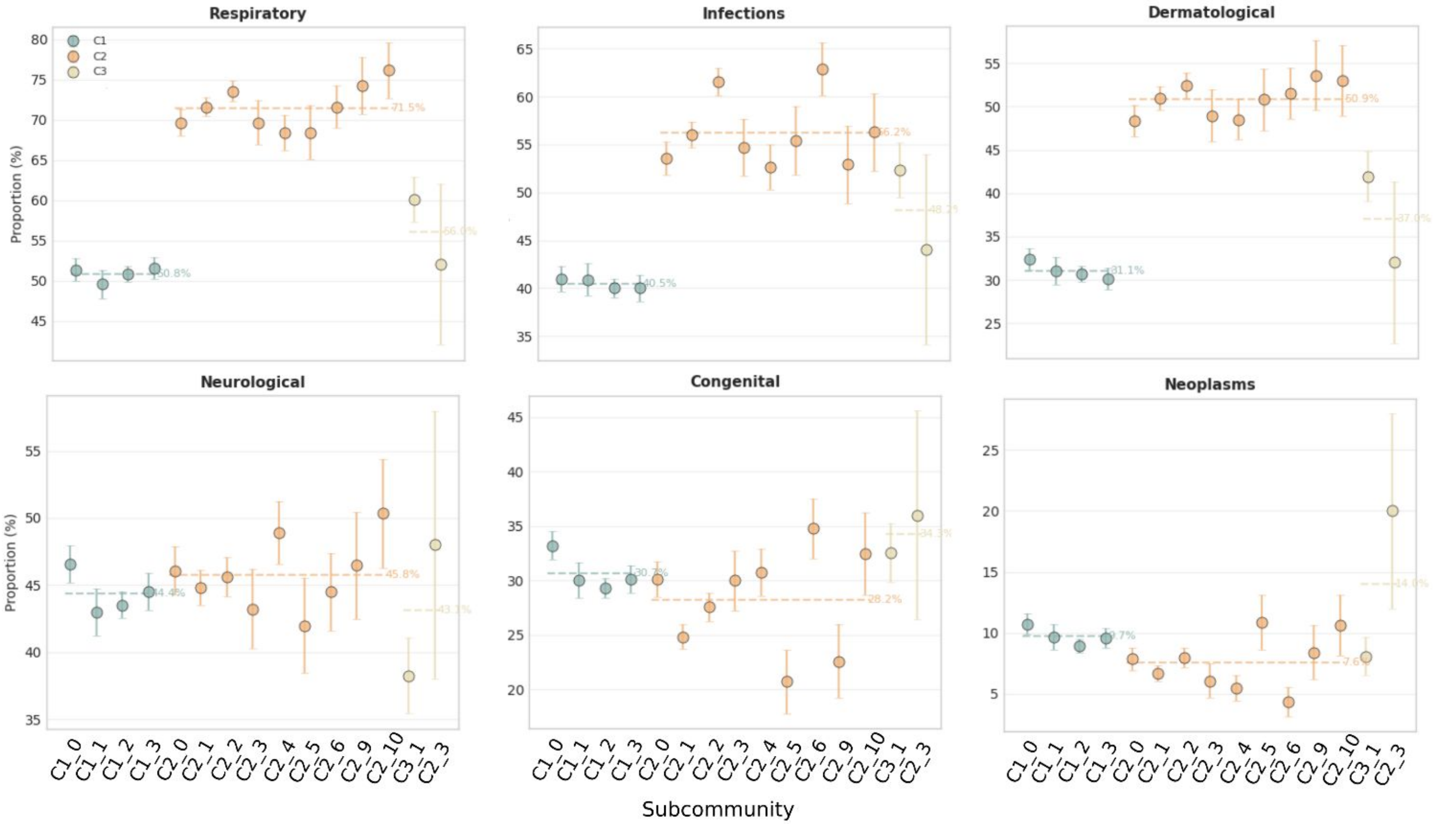
99.7% BIG individuals retained

Identity-by-descent captures shared environment

A - Genealogical and geographic distance





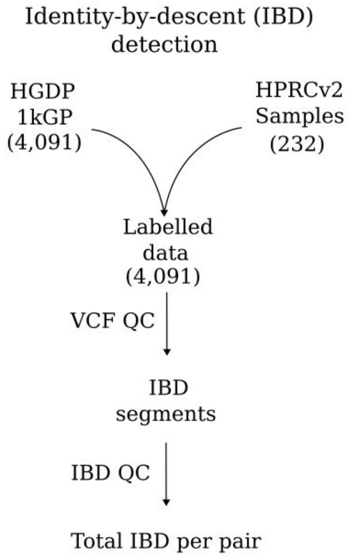

Respiratory
C1
C2
C3
Proportion (%)
80
75
70
65
60
55
50
45
71.5%
50.8%
56.0%
Infections
65
60
55
50
45
40
35
56.2%
40.5%
48.2%
Dermatological
55
50
45
40
35
30
25
50.9%
31.1%
37.0%
Neurological
Proportion (%)
55
50
45
40
35
45.8%
44.4%
43.1%
Congenital
45
40
35
30
25
20
30.7%
28.2%
34.3%
Neoplasms
25
20
15
10
5
9.7%
7.6%
14.0%
C1_0
C1_1
C1_2
C1_3
C2_0
C2_1
C2_2
C2_3
C2_4
C2_5
C2_6
C2_9
C2_10
C3_1
C2_3
Subcommunity

UTHSC

Can a similar approach be

useful on HPRCv2?

Preliminary/pilot results



Identity-by-descent (IBD) detection

HGDP
1kGP
(4,091)

HPRCv2
Samples
(232)

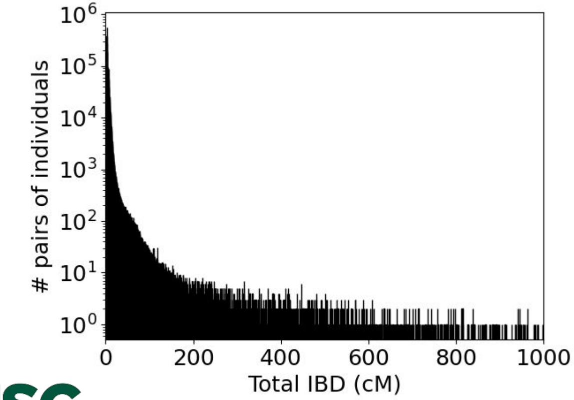
Labelled data (4,091)

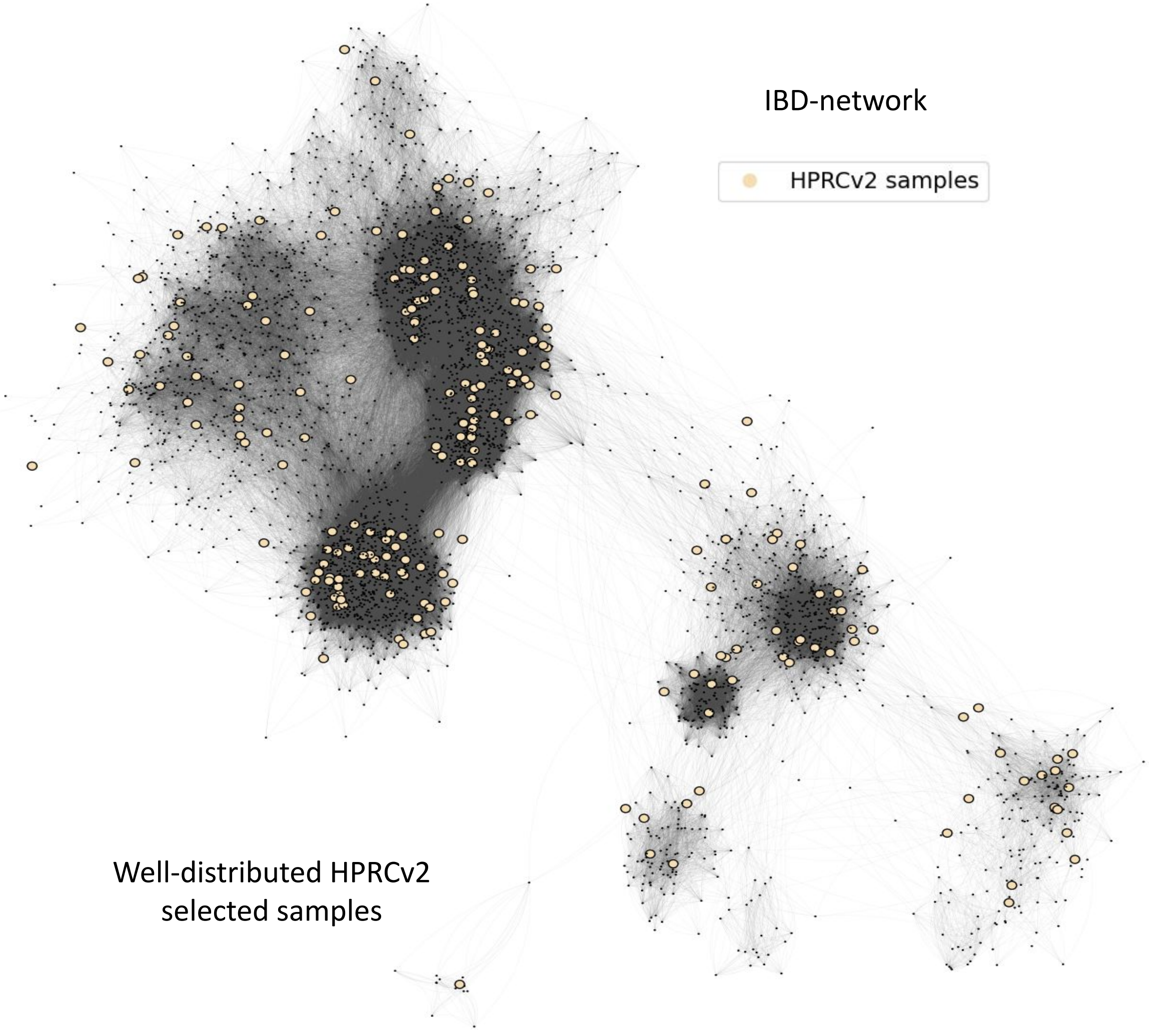
VCF QC

IBD segments

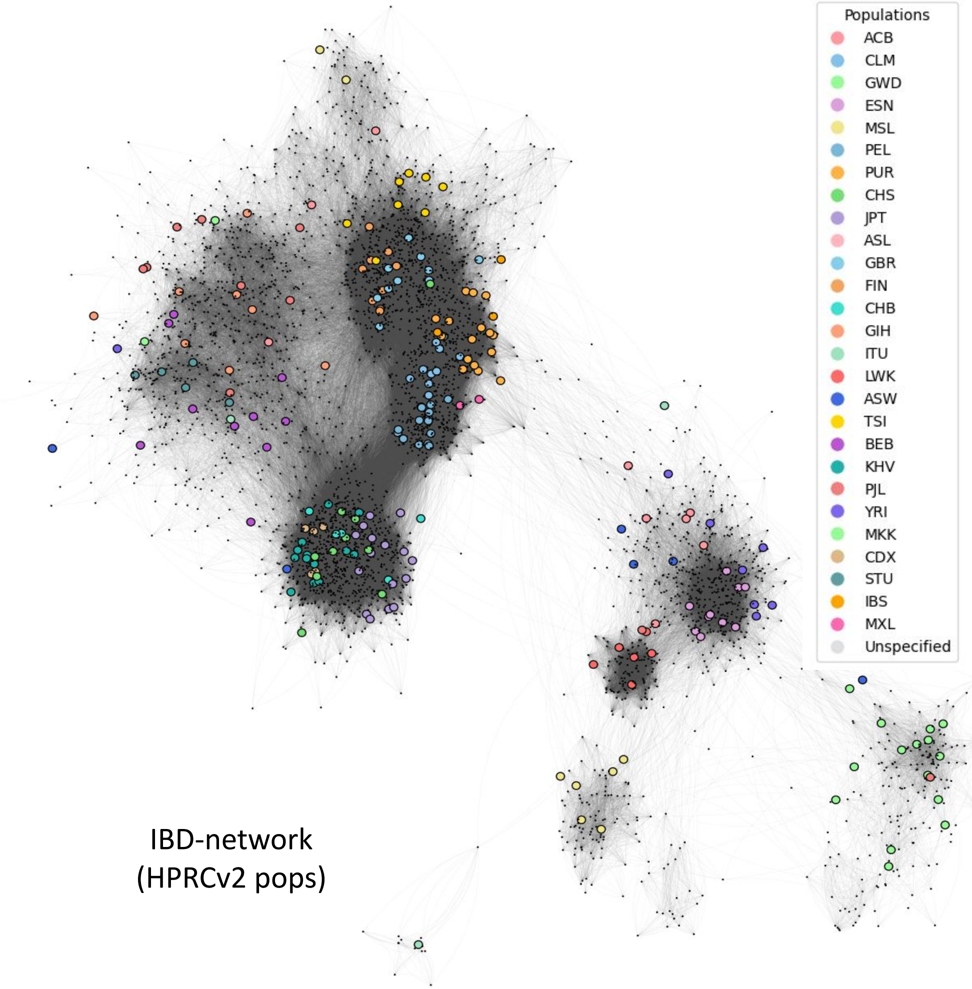
IBD QC

Total IBD per pair

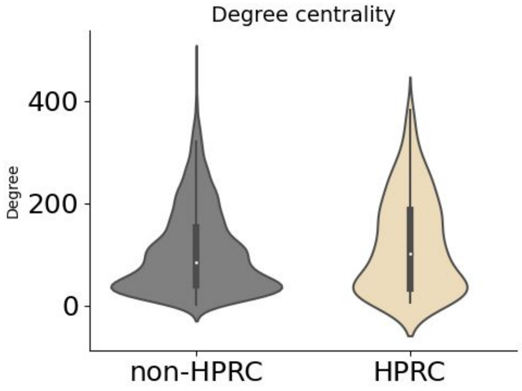


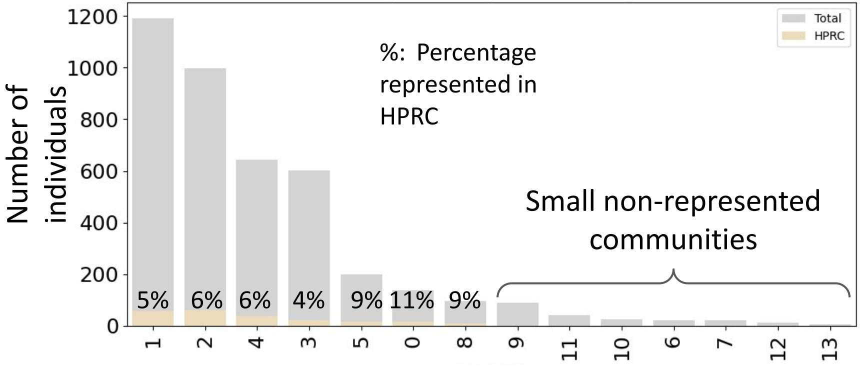


Well-distributed HPRCv2
selected samples



IBD-network (HPRCv2 pops)



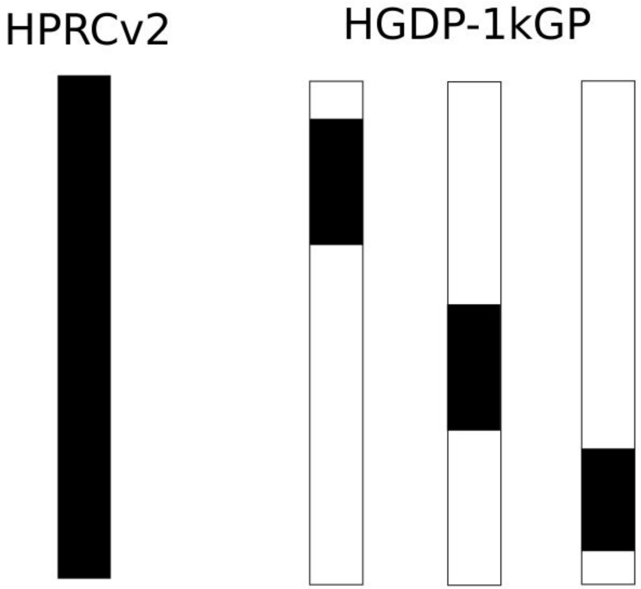


%: Percentage represented in HPRC

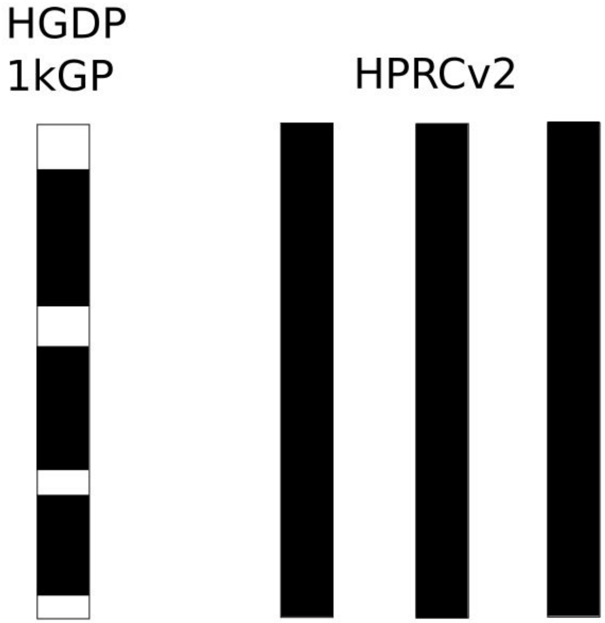
Small non-represented communities

Some small HGDP-1kGP communities (detected by Leiden) are yet unexplored in HPRCv2

Other IBD applications in sample selection



Haplotype informativeness: Compute the amount of cM explained by IBD sharing between the selected samples and public resources (HGDP-1kGP).



Proportion covered: Compute the percentage (amount of cM) already explored by IBD sharing of a potential HGDP-1kGP sample to be incorporated to HPRCv2.

Thank you