

No One Left Behind: Inclusive Quality Control in Northern-Netherlands Genomic Study of Major Depressive Disorder

Julien Rietkerk^{1,2}, Madhurbain Singh^{1,3}, Bradley T. Webb^{1,4}, Hanna M. van Loo², Roseann E. Peterson^{1,5}

1. Institute for Genomics in Health, State University of New York Downstate Health Sciences University, Brooklyn, NY, USA; 2. Universitair Centrum Psychiatrie, University Medical Center Groningen, NL; 3. Virginia Institute Psychiatry Behavior Genetics, Virginia Commonwealth University, Richmond, Virginia, USA; 4. RTI International; 5. Department of Psychiatry and Behavioral Sciences, Institute for Genomics in Health, State University of New York Downstate Health Sciences University, Brooklyn, NY, USA

- ### HIGHLIGHTS
- Cohorts which exclude individuals of non-European ancestry as part of genotype Quality Control hold potential to diversify psychiatric genetic research and boost sample sizes.
 - We use updated reference panels and Mahalanobis distance to obtain ancestry assignments in the Lifelines Cohort Study

BACKGROUND

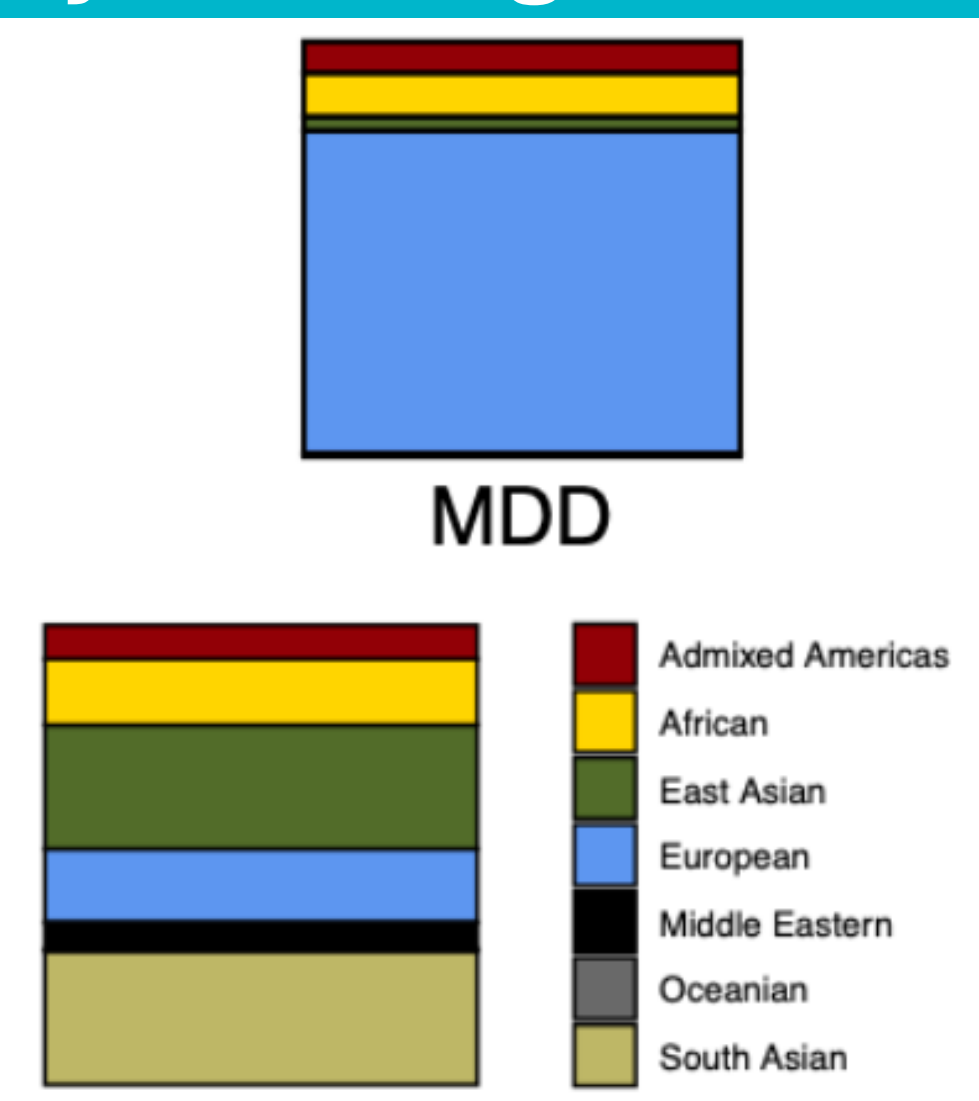
Overrepresentation of Europeans in psychiatric genetics

European-centric analyses in psychiatric genetics limits scientific discovery and increases disparities¹.

Efforts to mitigate this include: recruitment into diverse datasets²⁻⁴ and method development for diverse datasets⁵.

Some genotype Quality Control (QC) pipelines excluded diverse ancestry individuals from downstream analyses^{6,7} which present untapped diverse resources.

Future work applying our findings will focus on the study of Major Depressive Disorder (MDD).



Diversity in GWAS studies of MDD compared to world population. Adapted from Peterson et al.¹

Difference between race, ethnicity and genetic ancestry

Race

A context-dependent social construct based on physical traits, often tied to power, inequality, and health disparities¹

Ethnicity

Cultural group identity based on shared language and traditions; definitions vary and social factors can influence health risks¹

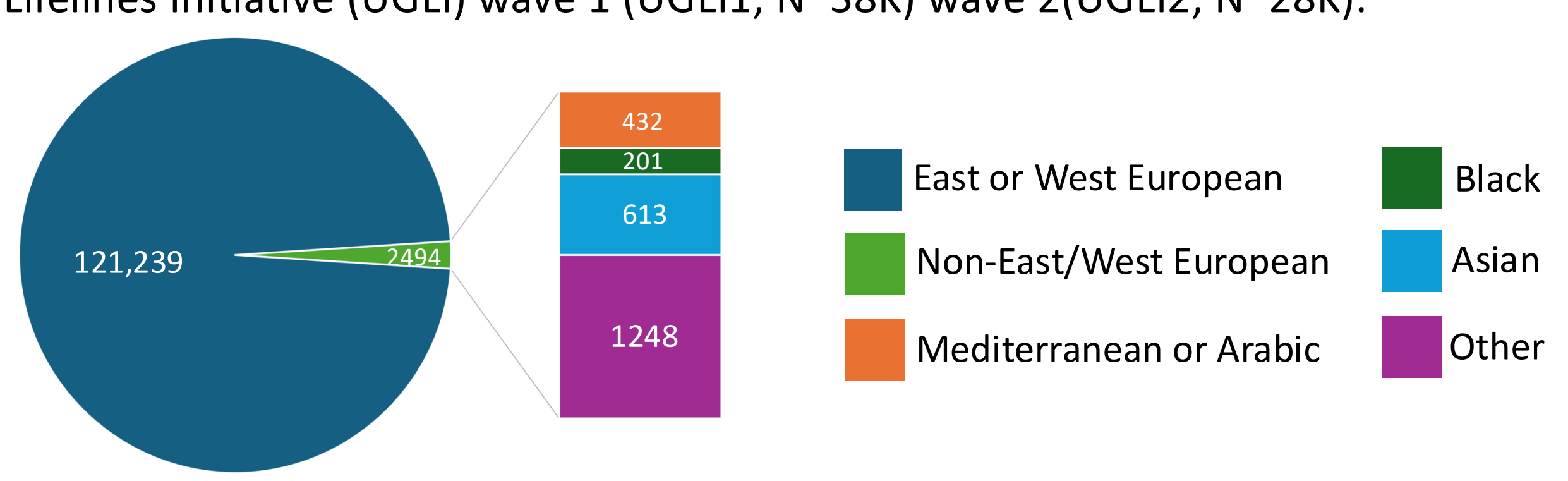
Ancestry

A genetic description of one's recent biological origins, estimated from DNA, which can be complex and context-dependent but is necessary for understanding genetic diversity

DATA

The Lifelines Cohort Study

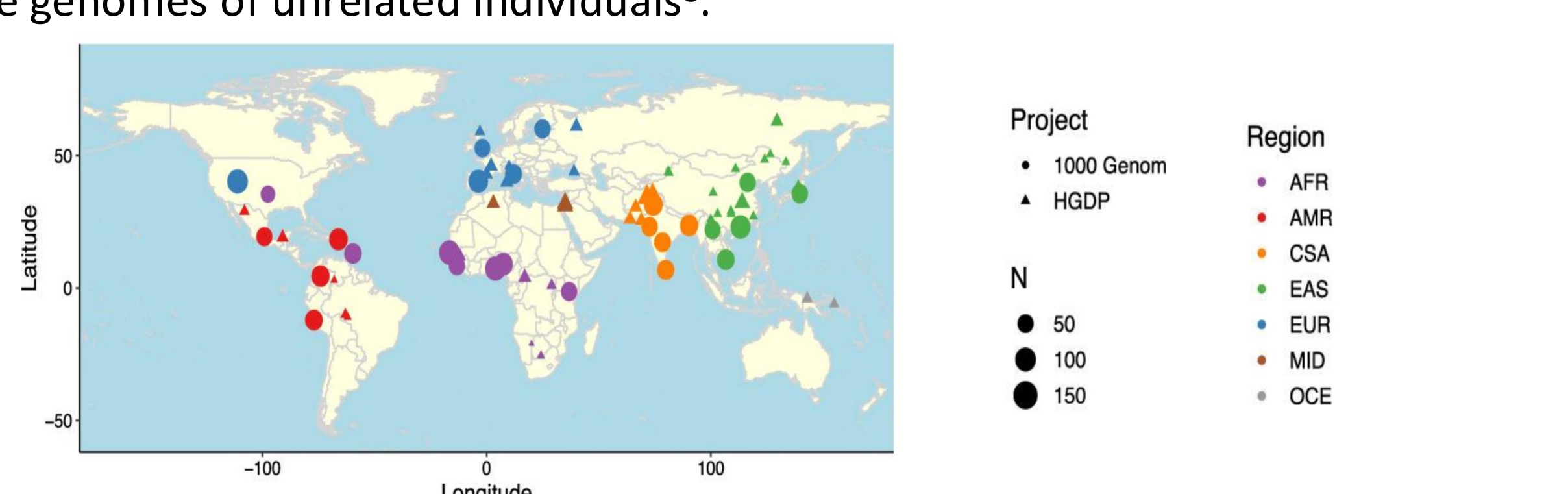
Lifelines is a multi-disciplinary prospective population-based cohort study of 167,729 persons living in the North of the Netherlands^{6,7}. We use genotyped sub-cohorts GWAS (N~15K), UMCG Genetics Lifelines Initiative (UGLI) wave 1 (UGLI1; N~38K) wave 2(UGLI2; N~28K).



Self-report race of participants in the Lifelines Cohort Study

Diverse ancestry reference panel

We use the 1000 Genomes Project and the Human Genome Diversity Project reference panel from the Genome Aggregation Database. It covers: 80 populations, 7 continental regions and 3400 whole genomes of unrelated individuals⁸.



Representation of sample size and regions within the reference panel. Figure 1A of Koenig et al.¹⁰

METHOD

Population Grouping by Mahalanobis Distance (POP-MaD)

Mahalanobis distance is calculated between a point (individual eigenvalue) and a distribution (eigenvalues of all individuals in the reference sub-population).

This considers the stretch and rotation of a distribution in PC-space, as well as applies a more flexible QC cut-off for outliers than classical PCA analyses⁹.

Population assignments are made based on the shortest distance to a reference panel using a maximum likelihood estimator^{5,10}.

Our cut-off threshold of > 3 SD of the Mahalanobis distance away from all distances assigned to that population⁵.

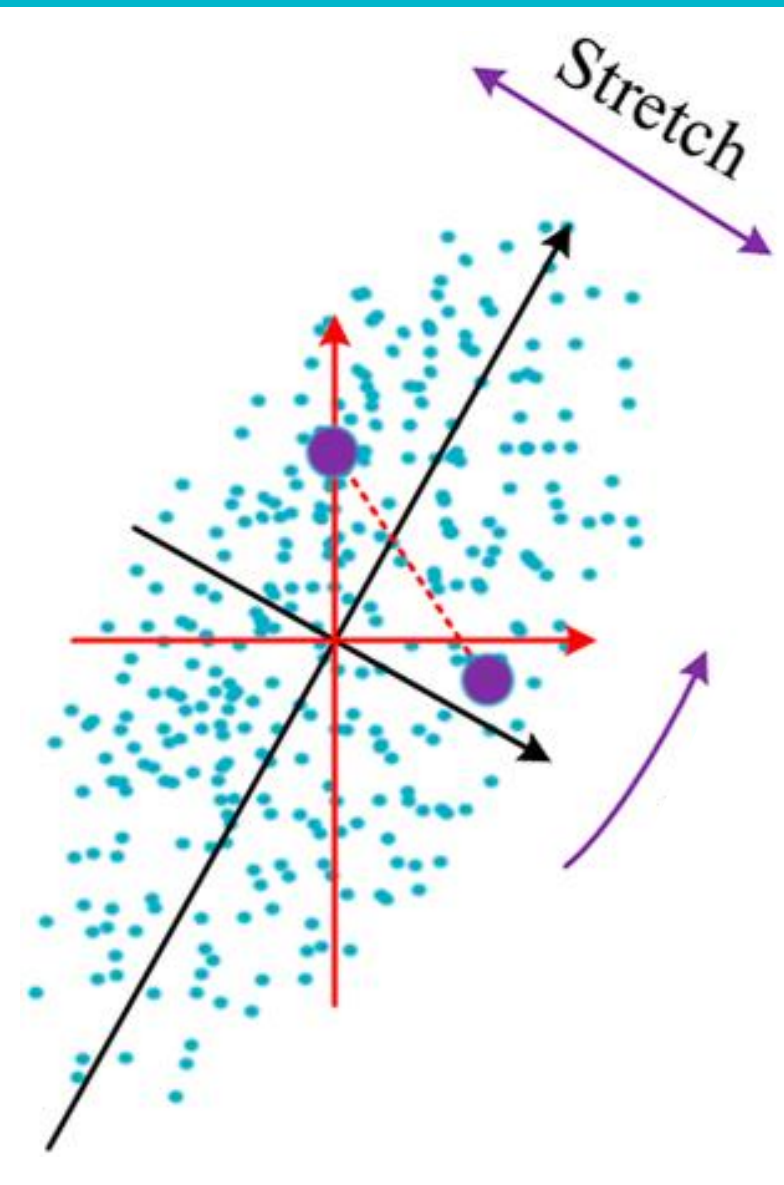
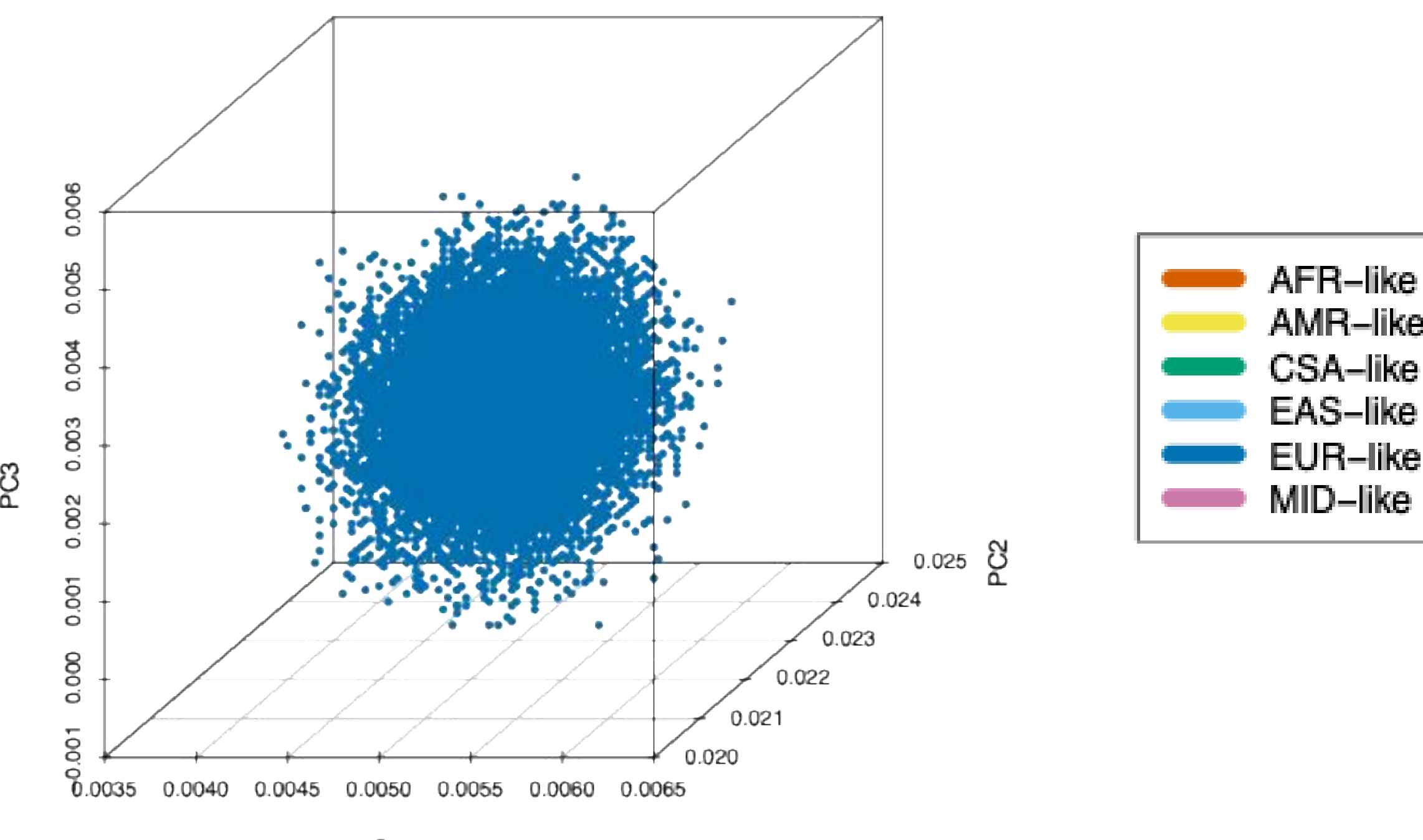
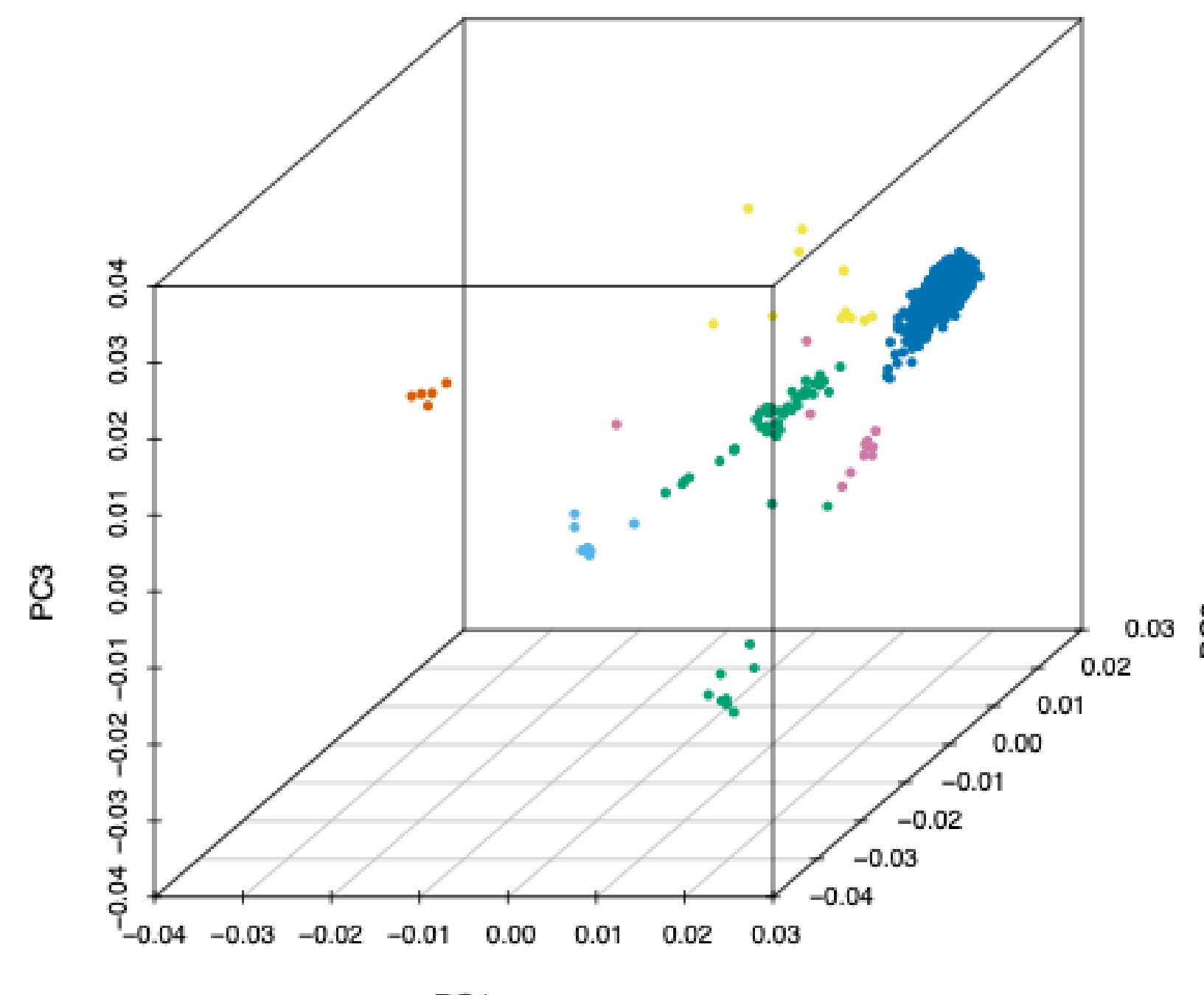


Image adapted from Kuang et al.⁹, figure 5

RESULTS



Lifelines Cohort Study sub-cohort GWAS: PCA mapping based on 1KGP HGDP reference panel and Mahalanobis population grouping



Lifelines Cohort Study sub-cohort UGLI1 & UGLI2: PCA mapping based on 1KGP HGDP reference panel and Mahalanobis population grouping

	GWAS	UGLI1 & UGLI2
AFR-like	0	5
AMR-like	0	11
CSA-like	0	56
EAS-like	0	8
EUR-like	15212	64234
MID-like	0	13
Dropped*	210	162

Ancestry assignments in three Lifelines sub-cohorts GWAS, UGLI1 and UGLI2. *Dropped individuals were EUR-like. Most likely excluded due to recent admixture