SUMMIX: A METHOD FOR DETECTING AND ADJUSTING FOR POPULATION STRUCTURE IN GENETIC SUMMARY DATA

Ian Arriaga-MacKenzie

OVERVIEW

- Background
- Motivation
- Mixture model
- Simulations
- Ancestry Estimation
- Ancestry Adjusted Allele Frequencies
- Importance
- Future Work

BACKGROUND

Publicly and easily available online genomic resources

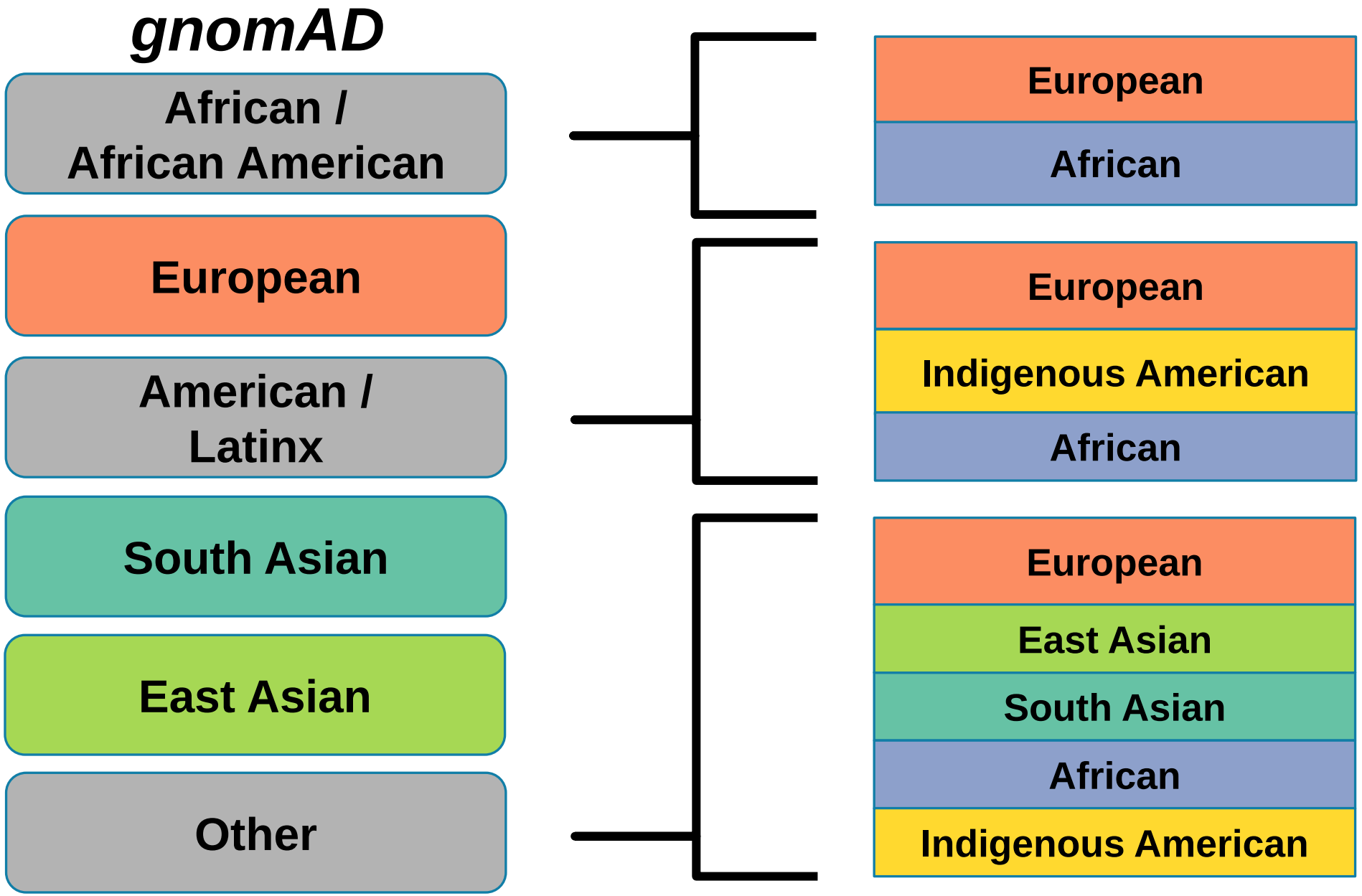
- Used to prioritize putative causal variants in rare diseases
- External pseudo controls in case-control analysis

genome Aggregation Database (gnomAD)

- ~140,000 sequenced invididuals
- V2.1 contains precise information (e.g. controls, precise European ancestry)

Reported in genotype counts/frequencies

Can mask within or between sample heterogeneity



MOTIVATING EXAMPLE

Incomplete ancestry information can result in incorrect analyses and conclusions about genetic associations

Variant *PADI3* in Central Centrifugal Cicatricial Alopecia

Liron Malki, M.Sc., Ofer Sarig, Ph.D., Maria-Teresa Romano, M.Sc., Marie-Claire Méchin, Ph.D., Alon Peled, B.Med.Sci., Mor Pavlovsky, M.D., Emily Warshauer, M.D., Liat Samuelov, M.D., Laura Uwakwe, M.D., Valeria Briskin, Ph.D., Janan Mohamad, B.Med.Sci., Andrea Gat, M.D., et al.

The NEW ENGLAND JOURNAL of MEDICINE

February 13, 2019
DOI: 10.1056/NEJM[illegible]1816614

In a post hoc analysis, the *PADI3* mutation frequency among 58 women of African ancestry who had CCCA (116 alleles) was found to differ significantly from that calculated for a control cohort of women of African ancestry (from the gnomAD V2.1 control set) according to the chi-square test (P=0.002) and Fisher's exact test (P=0.006). The difference remained significant after adjustment for relatedness of persons according to the chi-square test (P=0.03) and Fisher's exact test (P=0.04). We did not control for population stratification. However, the mutation frequency was similar across various African subpopulations (Table S5 in the Supplementary Appendix).

MOTIVATION CONT.

Individuals from understudied and admixed populations most likely to lack large public resources with precisely matched ancestry

Difficult choice:

- Use closest but still poorly matched ancestral group
  - Can bias results
- Not use publicly available data as resource
  - Reduced sample sizes and thus power

OUR METHOD

Goal:

Precisely estimate continental ancestry from allele frequency data

Adjust allele frequency data to match target population

THE MIXTURE MODEL

Observed summary allele frequency, AF_{OBS}, can be modeled as a mixture of reference ancestries

$$AF_{OBS} = \sum_{k=1}^{K} \left(\pi_k \times AF_{REF,k} \right)$$

- Estimate proportion of each reference ancestry, π_k
- AF are vectors of N SNPs

MIXTURE MODEL CONT.

Estimate π_k using least squares

$$minimize: f(\pi) = \left(AF_{OBS} - \sum_{k=1}^{K} (\pi_k \times AF_{REF,k}) \right)^2$$

Subject to the constraints: $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$

REFERENCE DATA

Four 1000 Genomes continental populations

- **African (AFR)** – Removed Americans of African Ancestry (ASW) and African Caribbean's in Barbados (ACB)
- **East Asian (EAS)**
- **European (EUR)** – removed Finnish (FIN)
- **South Asian (SAS)**

Indigenous American (IAM)

- Affy 6.0 previously harmonized with 1000Genomes
- Martin et al., 2017

Greater than 1% MAF in at least one population

OPTIMIZATION

$$minimize: f(\pi) = \left(AF_{OBS} - \sum_{k=1}^{K} (\pi_k \times AF_{REF,k}) \right)^2$$

Need to find the solution over K dimensions:

Our loss function is: continuous, convex, twice differentiable

Allows us to use Sequential Quadratic Programming (SQP)

- "gradient descent" to find global minimum
- Equivalent to applying Karush-Kuhn-Tucker conditions

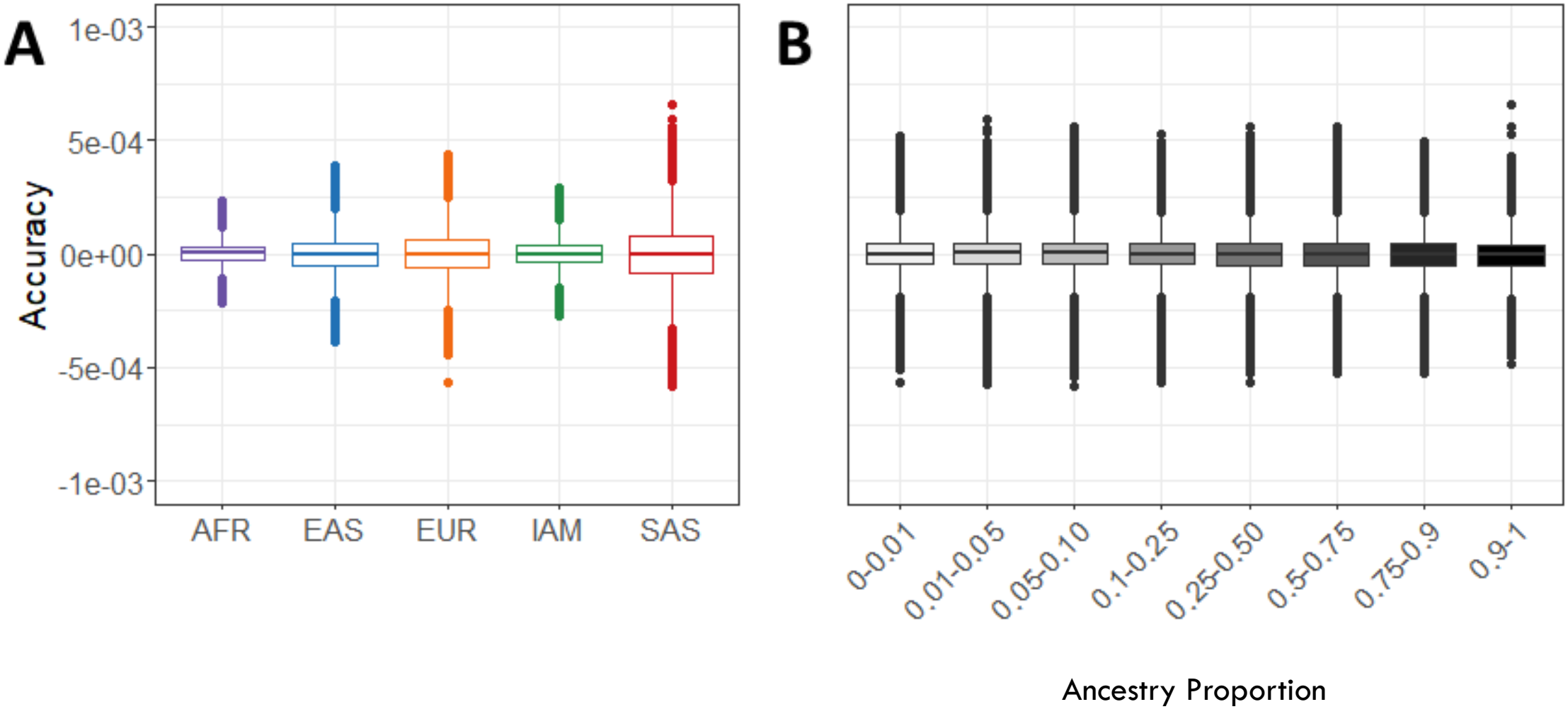
SIMULATION DESIGN

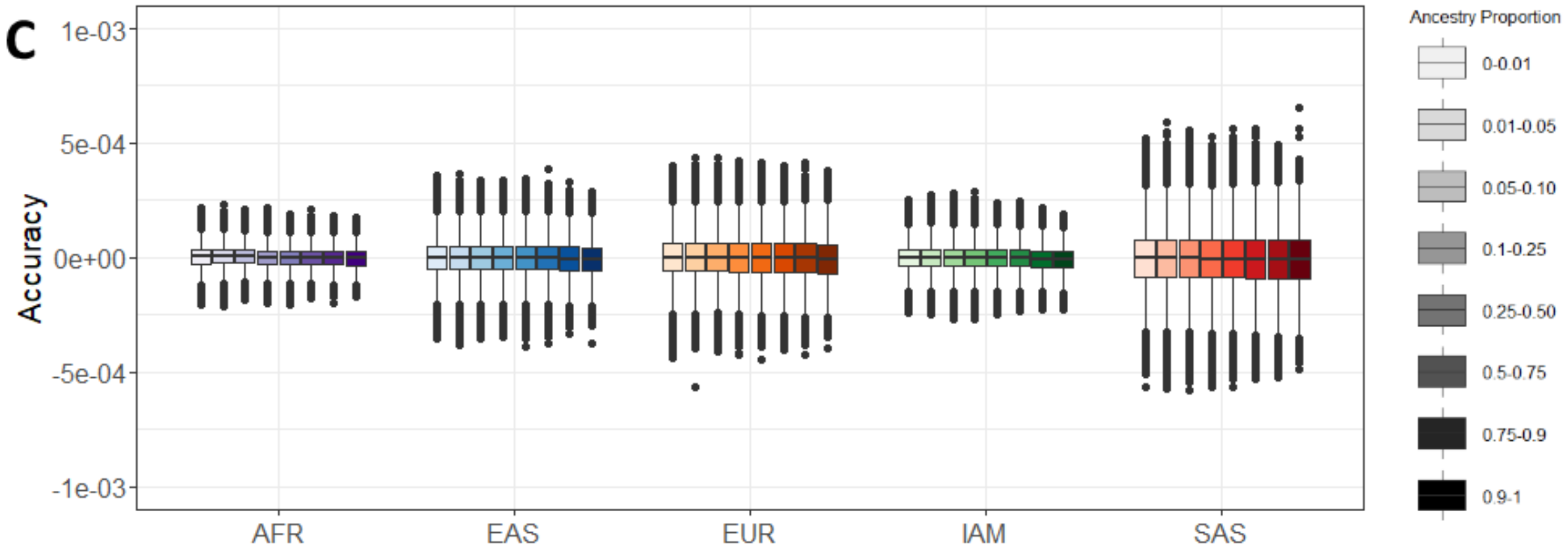
Using reference data:

- Randomly chose 100,000 variants
- Assuming HWE, simulate variants from K ancestries
- Weighted sum of ancestry to create summary data

Simulation parameters:

- N (people) = 10,000
- Ancestry proportions randomly chosen from bin
  - 0-0.015, 0.010-0.055, 0.05-0.105, 0.10-0.255, 0.25-0.505
- All combinations of ancestries
- Over 5000 simulation scenarios in total





APPLICATION TO GNOMAD V2.1

- Estimated ancestry proportions for all groups and sub-groups (non-topmed, non-neuro, controls)
- 1000 Genomes used for reference panel
- After gnomAD-reference merge
  - 9,835 exome SNPs
  - 582,550 genome SNPs

BLOCK BOOTSTRAPPING

- SQP estimates are fast (within seconds)
- Use block bootstrapping to estimate error and 95% confidence intervals
  - 3357 cM blocks across genome
  - 1000 bootstrap replicates

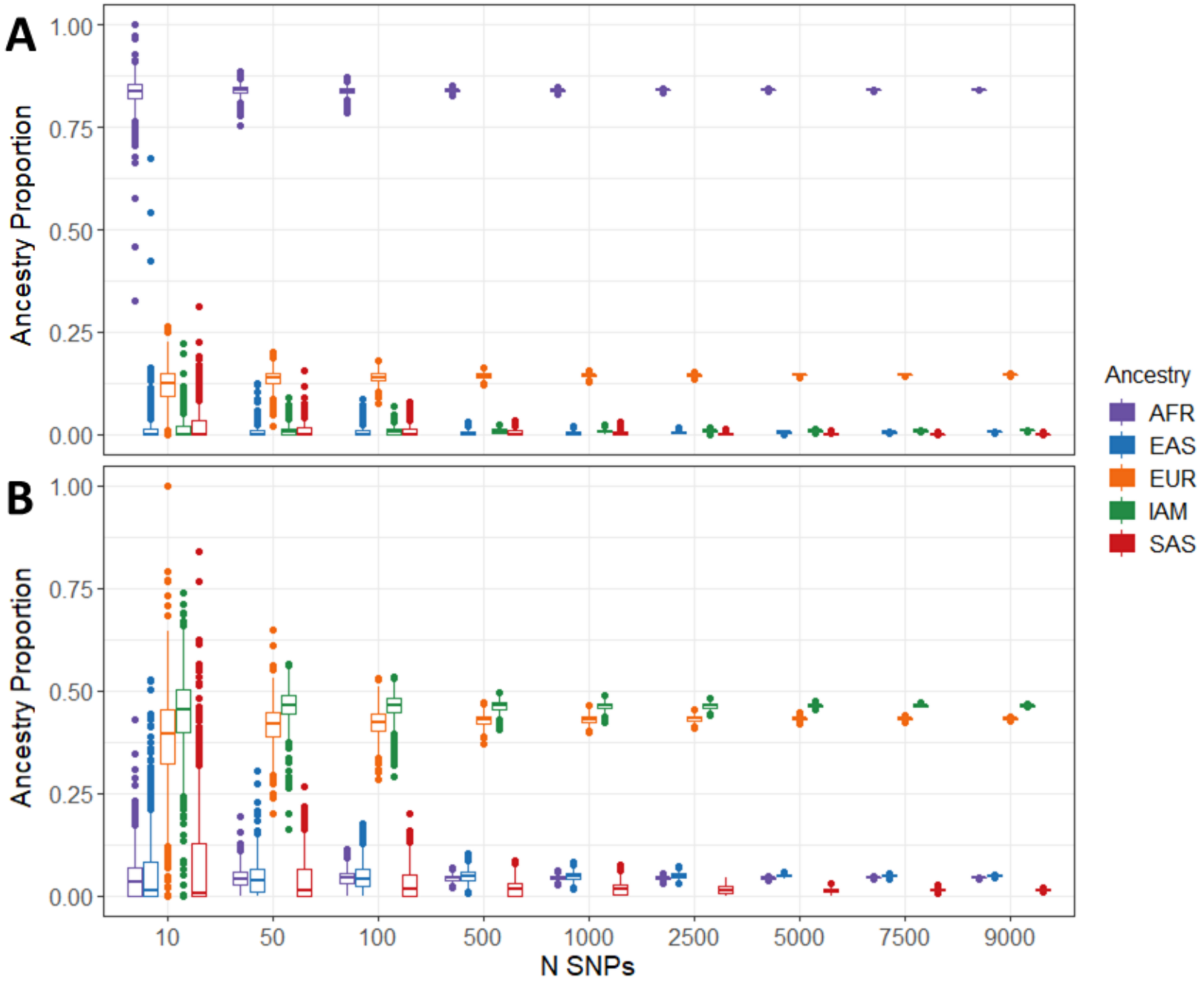
ANCESTRY ESTIMATION RESULTS

Estimated ancestry proportions for gnomAD groups (95% Block Bootstrap CI)						
	Exome					
	African/ African American (N=8,128)	American/ Latinx (N=17,296)	Other (N=3,070)	Non-Finnish European (N=56,885)	East Asian (N=9,197)	South Asian (N=15,308)
AFR	**0.840 (0.838, 0.842)**	0.043 (0.039, 0.047)	0.034 (0.032, 0.037)	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	0.002 (0.000, 0.004)
EAS	0.005 (0.001, 0.008)	0.049 (0.040, 0.059)	0.046 (0.041, 0.051)	0.000 (0.000, 0.000)	**1.000 (1.000, 1.000)**	0.000 (0.000, 0.000)
EUR	0.146 (0.142, 0.149)	**0.432 (0.423, 0.441)**	**0.780 (0.775, 0.787)**	**0.975 (0.969, 0.982)**	0.000 (0.000, 0.000)	0.150 (0.144, 0.157)
IAM	0.009 (0.006, 0.012)	**0.463 (0.455, 0.471)**	0.051 (0.047, 0.054)	0.008 (0.005, 0.012)	0.000 (0.000, 0.000)	0.002 (0.000, 0.005)
SAS	0.000 (0.000, 0.006)	0.013 (0.000, 0.027)	0.089 (0.080, 0.098)	0.017 (0.009, 0.023)	0.000 (0.000, 0.000)	**0.845 (0.838, 0.852)**

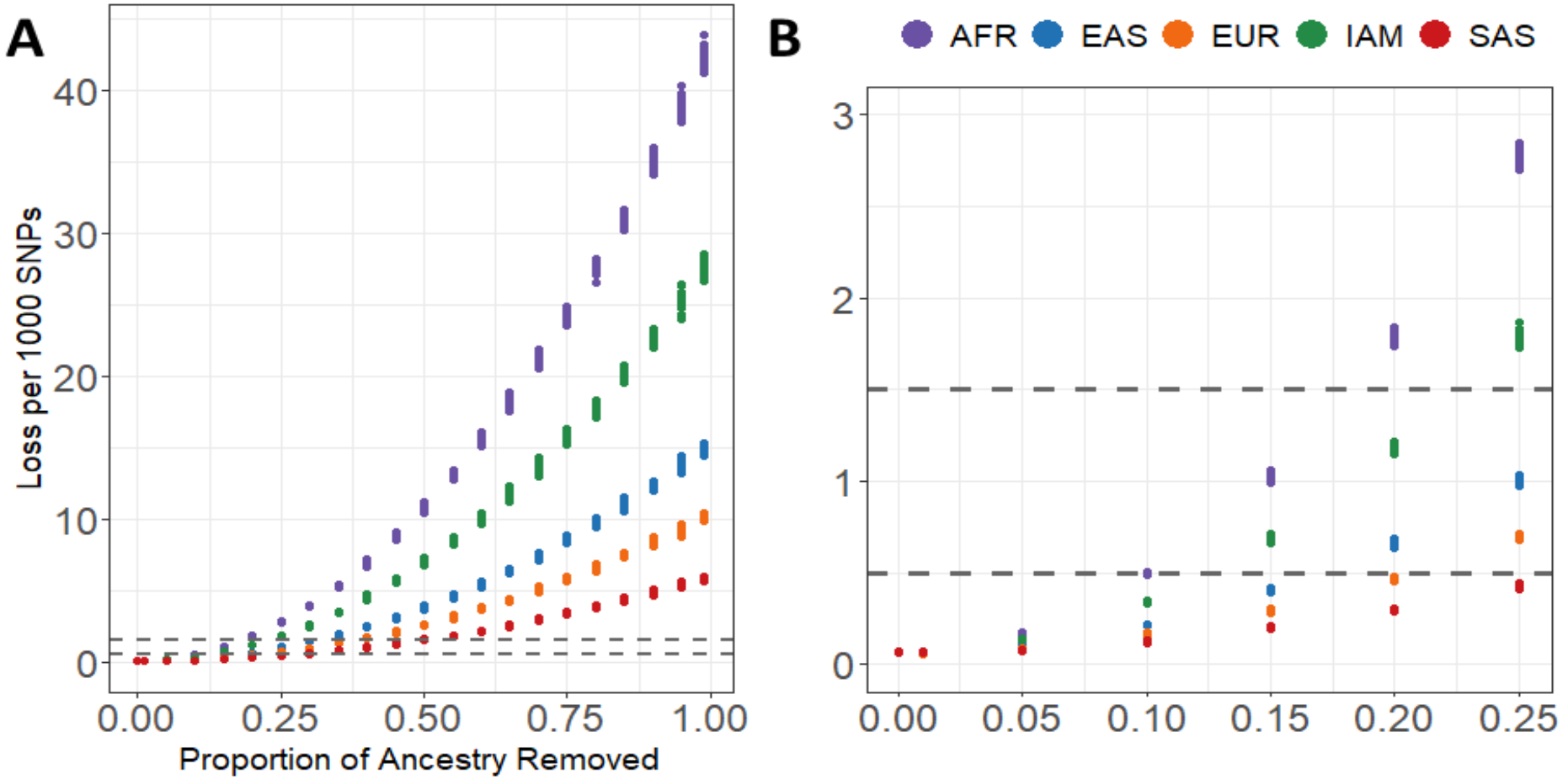
NUMBER OF SNPS/VARIANTS

- Simulations used sets of 100K randomly sampled SNPs
- gnomAD exome data is limited to ~10K SNPs

- What is the lower limit of SNPs we can reliably use?



SENSITIVITY OF REFERENCE DATA



LOSS PER 1000 SNPS

Ancestry	Exome			Genome		
	SNPs	Iterations	Loss/1000	SNPs	Iterations	Loss/1000
African/African-American	9750	24	0.234	582156	32	0.221
American/Latinx	9722	38	1.080	582155	45	0.824
Other	9749	25	0.451	582156	42	0.680
Non-Finnish European	9763	29	0.374	582156	24	0.500
East Asian	9732	30	0.346	582155	20	0.433
South Asian	9719	44	0.337	--	--	--
Finnish	9728	49	2.617	582155	85	2.618
Ashkenazi Jewish	9749	121	2.047	582156	68	2.462

ANCESTRY ADJUSTED ALLELE FREQUENCIES

Estimate ancestry-adjusted AF matching ancestry proportions for a target individual or sample

Use K-1 reference ancestries

$$x = \frac{\pi_{target,l}}{\hat{\pi}_l}\left(AF_{observed} - \sum_{k \neq l} \hat{\pi}_k AF_{ref,k}\right) + \sum_{k \neq l} \pi_{target,k}\, AF_{ref,k}$$

$$AF^*_{adjusted} = \begin{cases} 0, & x < 0 \\ x, & 0 \leq x \leq 1 \\ 1, & x > 1 \end{cases}$$

Where,

$AF^*_{adjusted}$ *is the ancestry-adjusted allele frequency*

l *- ancestry group for which the reference allele frequency data is not used*

k *- ancestry group index*

$\pi_{target,k}$ *- population k ancestry proportion for target individual or sample*

$\hat{\pi}_k$ *- estimated ancestry proportion of population k for observed publicly available summary data*

$AF_{observed}$ *- allele frequency for observed publicly available summary data (e.g. gnomAD)*

$AF_{ref,k}$ *- reference allele frequency for ancestry k*; K-1 homogenous reference ancestries are used.

ANCESTRY ADJUSTED AF

Estimated ancestry-adjusted AF for gnomAD v2.1 exomes

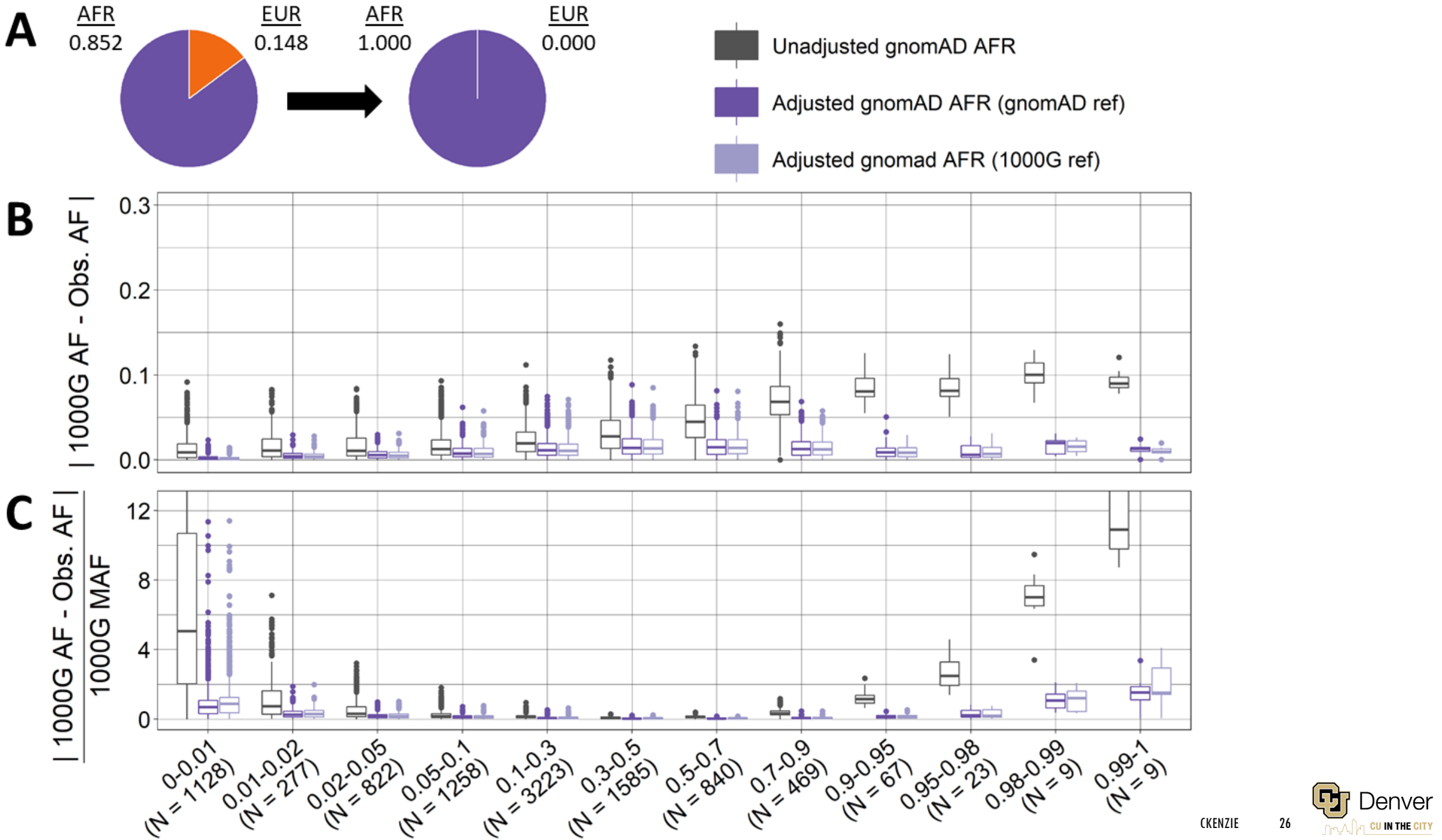
- African/African-American (leave out African reference)
- American/Latinx (leave out Indigenous American reference)

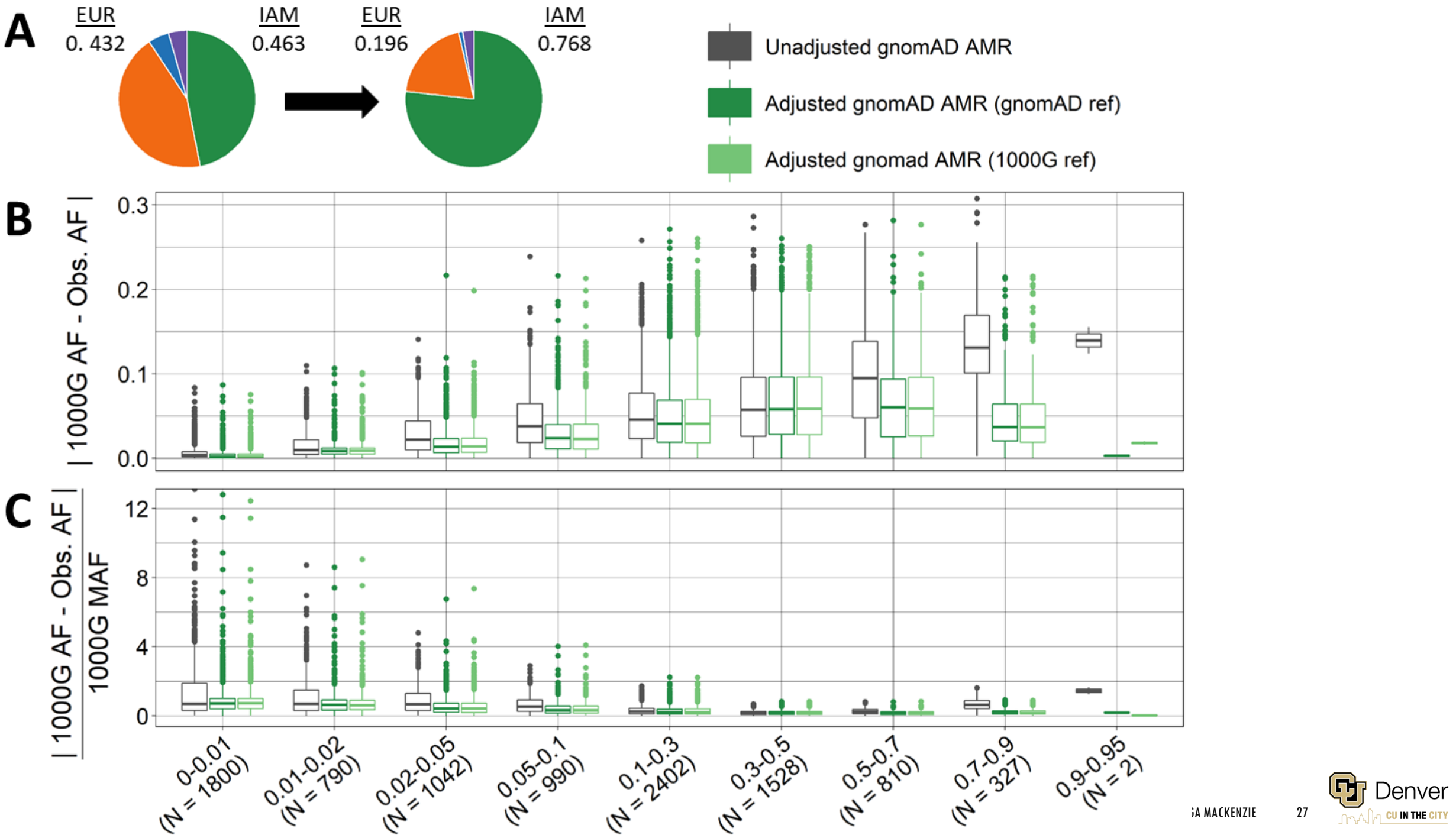
Reference AF

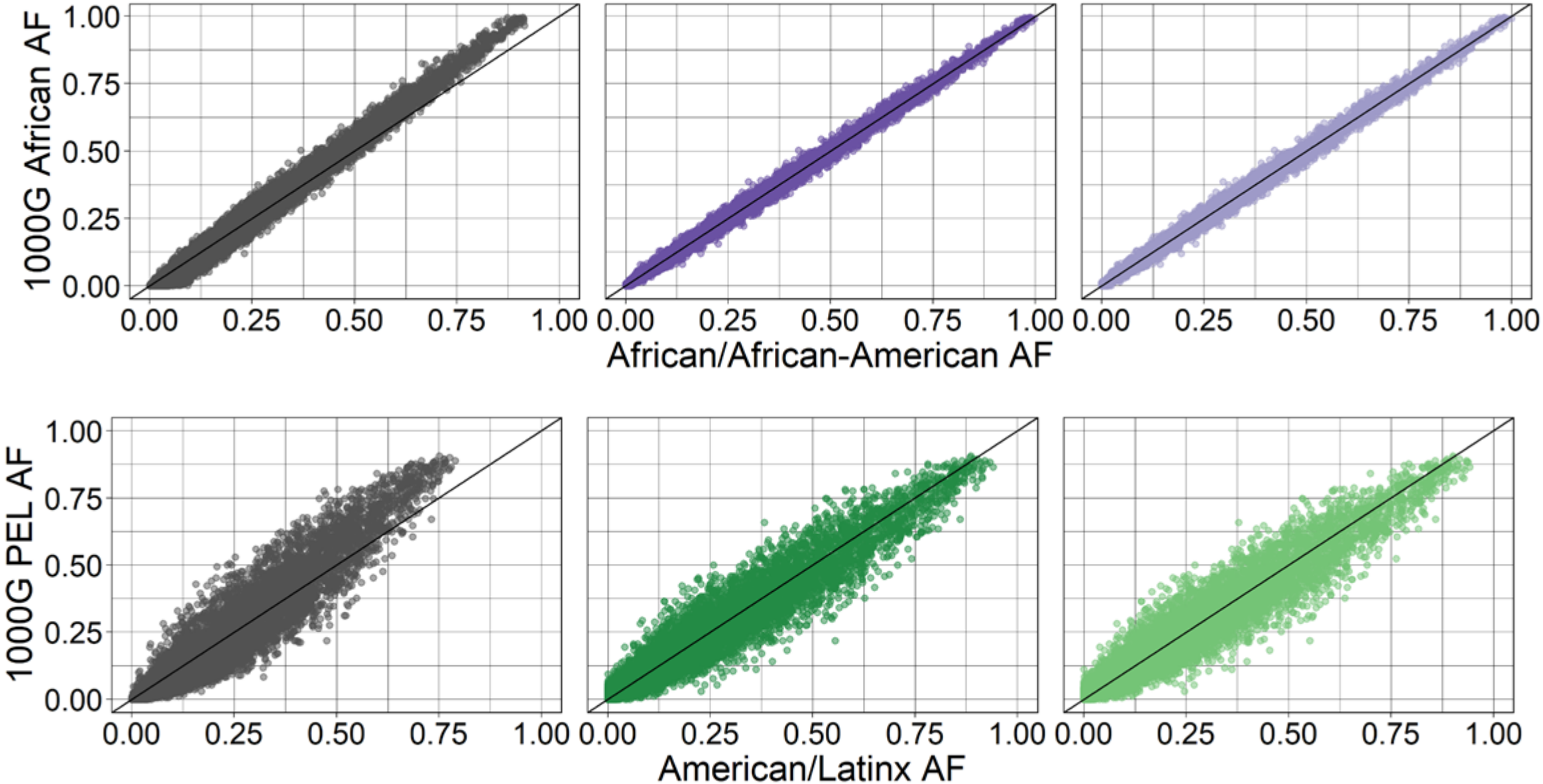
- 1000Genomes
- gnomAD
  - Use ancestry-adjusted African AF as reference for American/Latinx

Compare to 1000 Genomes to evaluate

- Homogenous African
- Peruvian (admixed)







RE-ANALYSIS OF PADI3

Minor alleles in *PADI3* variants reported from Malki et al.

	Cases	gnomAD v2.1 African	
		unadjusted	100% AFR adjusted
Minor allele	14	1677	1960
No minor allele	44	10810	10527
Chi Square p-value		0.029	0.114
Fisher's exact test p-value		0.031	0.101

IMPORTANCE

Publically available genetic resources

- Contain population structure
- May not have precise ancestry information
- May result in misunderstanding and misuse of resources

Summix estimates and adjusts for reference ancestry proportions in summary data

FUTURE WORK

More research needed

- Diverse reference panels
- Fine-scale and local ancestry estimates
- Other sources of heterogeneity (e.g. genetic predisposition to disease)

Estimation of Unknown Reference Data

- **Binomial distribution model**
- **Block Relaxation Algorithm**

THANK YOU

University of Colorado Denver

uchealth | University of Colorado Anschutz Medical Campus

COLORADO CENTER FOR PERSONALIZED MEDICINE

Greg Matesi
Sam Chen
Alex Ronco
Katie M. Marker
Jordan Hall
Haley Stoneman
Ryan Scherenberg
Mobin Khajeh-Sharafabadi
Yinfei Wu
Chris Gignoux
Megan Null
Audrey Hendricks

NHGRI R35HG011293

GSP: NHGRI U01HG009080

EUReCA! at CU Denver
Program to support undergraduate research

UROP at CU Denver
Undergraduate Research Opportunities Program