Population Stratification

Human Genomic Epidemiology – Asia
Virtual Course
June 13-17, 2022

Professor Qasim Ayub

qasim.ayub@monash.edu

Director, Monash University Malaysia Genomics Facility
Deputy Head of School (Research)
School of Science

Learning Outcomes

- Understand modern human population structure.
- Interpret principal components and ADMIXTURE analyses.
- Describe how population structure confounds genome wide association studies and how to control for it.

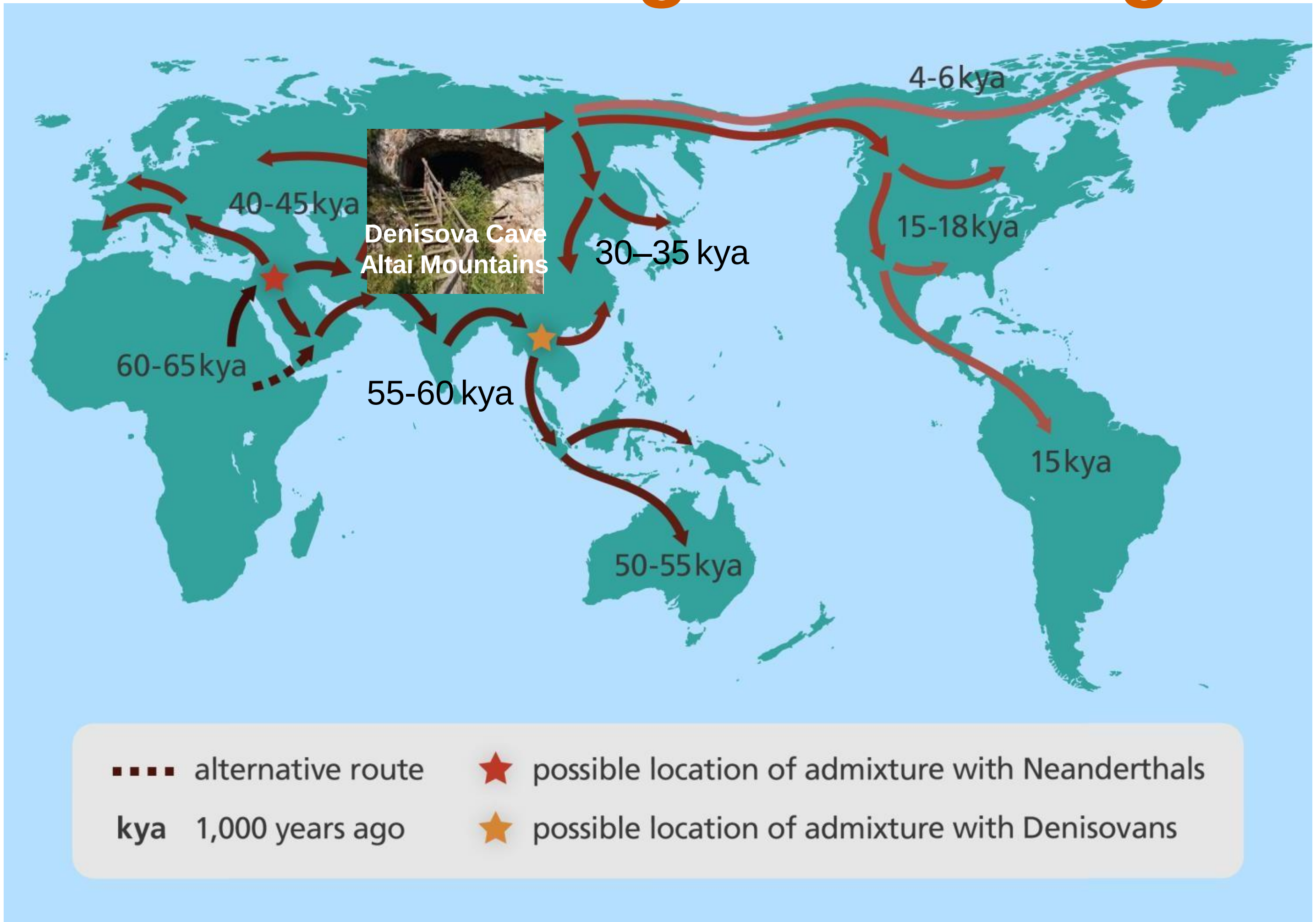
Outline

- Modern human populations and datasets.
- Principal Components and ADMIXTURE analysis.
- Population stratification and GWAS.

What is a Population?

- **Population** is a **spatial-temporal group of interbreeding individuals who share a common gene pool.**
- Population genomics aims to understand population structure and relationships.
- Population structure is defined by the organization of genetic variation and is driven by the combined effects of evolutionary processes that include recombination, mutation, genetic drift, demographic history (origins, migrations and admixtures) and evolutionary adaptations by natural selection.

Modern Human Origins and Migrations



- Modern human bony conformation was established in Africa around 330 – 200,000 years ago.
- Genetic evidence supports the fossil and archaeological evidence.

Population Variation Databases

International HapMap Project
International HapMap Project
Home | About the Project | Data | Publications | Tutorial

http://hapmap.ncbi.nlm.nih.gov/

A computer security audit revealed security flaws in the legacy HapMap site and NCBI has took it down in June 2016.

http://www.internationalgenome.org/

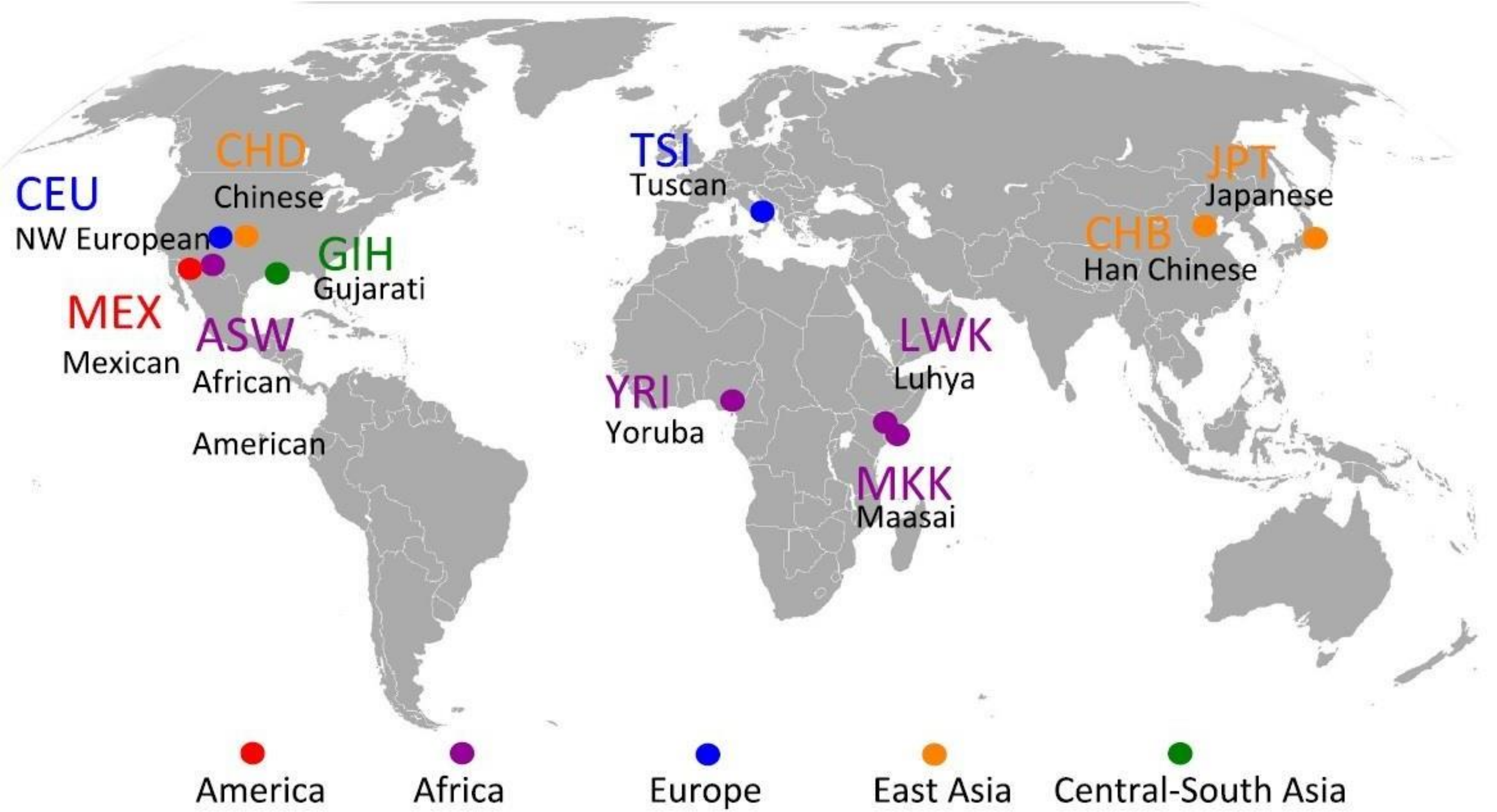
IGSR: The International Genome Sample Resource

Providing ongoing support for the 1000 Genomes Project data

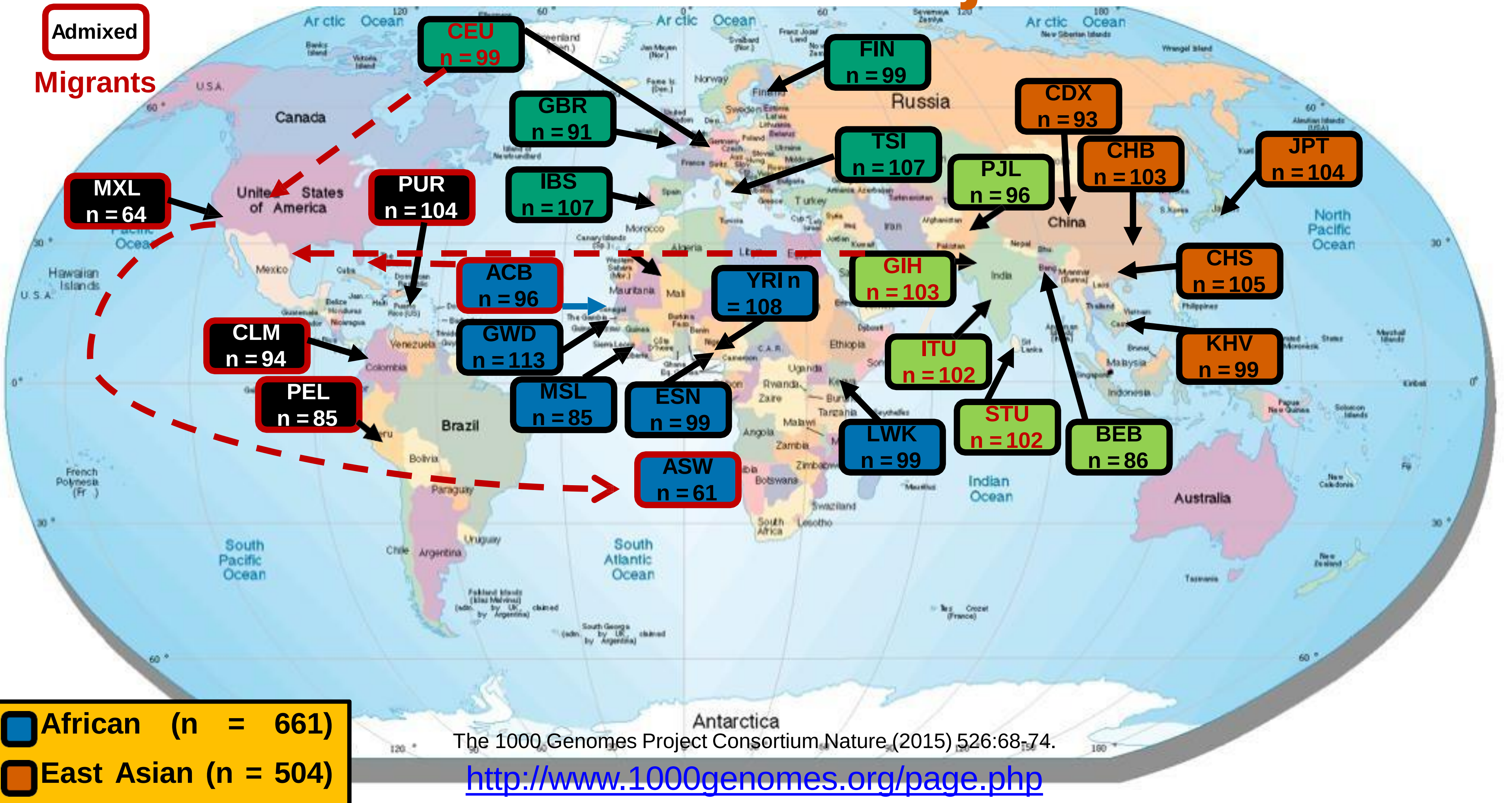
http://gnomAD.broadinstitute.org/

Genome Aggregation Database and gnomAD Browser

International HapMap Project



The 1000 Genomes Project Dataset



🟦 African (n = 661)
🟧 East Asian (n = 504)
🟩 South Asian(n = 489)
🟩 European (n = 503)
⬛ American (n =347)

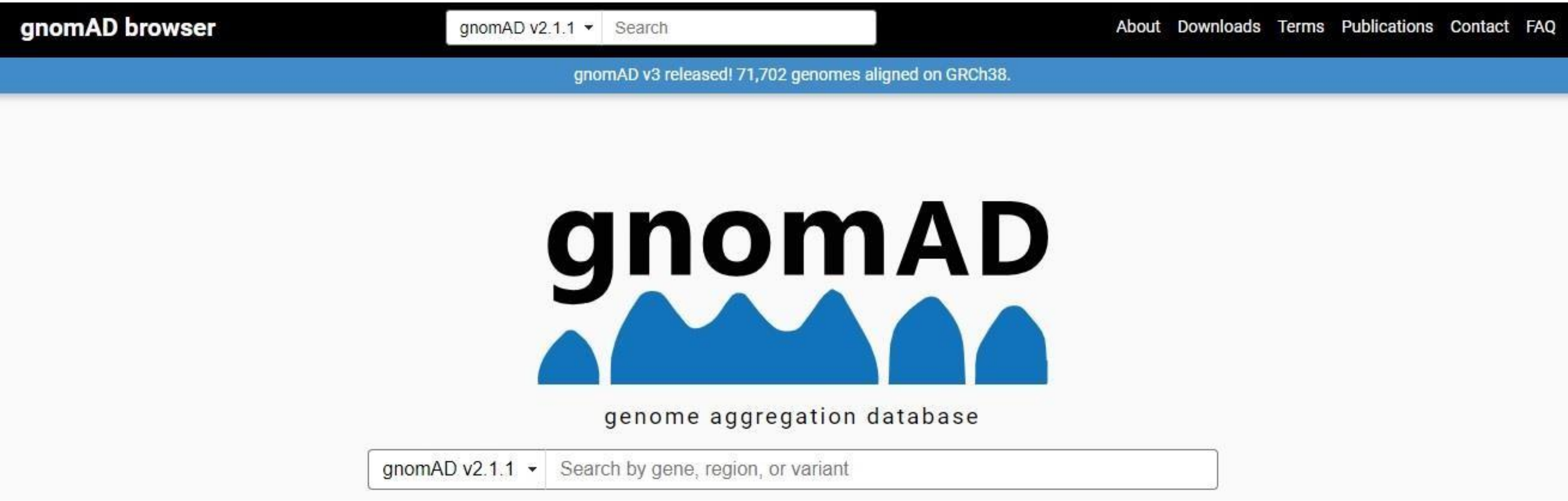
The 1000 Genomes Project Consortium Nature (2015) 526:68-74.

http://www.1000genomes.org/page.php

Samples	Populations	Mean Coverage	SNPs
2,504	26	7.4 X	84.7 M

Genome Aggregation Database

Whole exomes 125,748
Whole genomes 15,708



gnomAD browser

gnomAD v2.1.1 ▾ Search

About Downloads Terms Publications Contact FAQ

gnomAD v3 released! 71,702 genomes aligned on GRCh38.

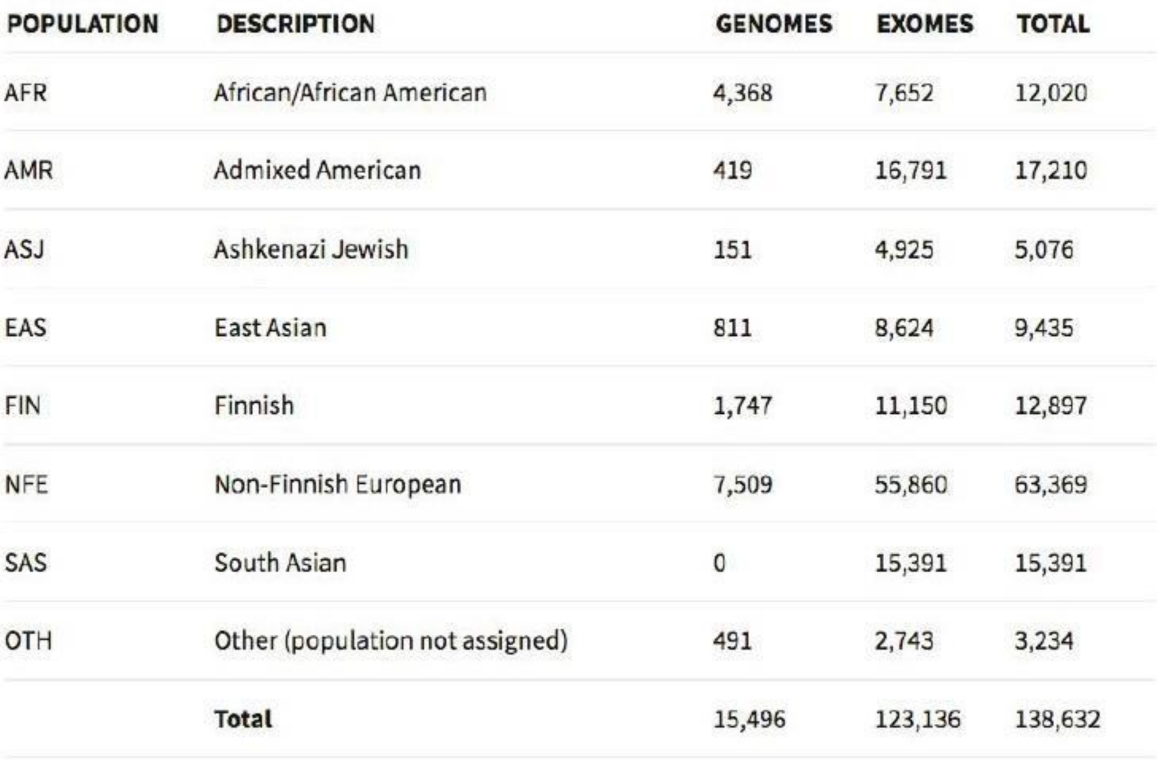
gnomAD

genome aggregation database

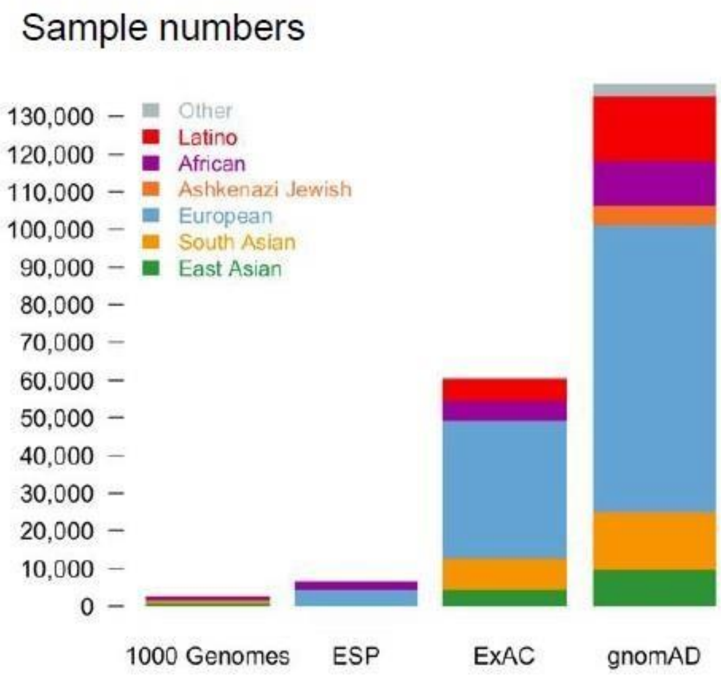
gnomAD v2.1.1 ▾ Search by gene, region, or variant

http://gnomad.broadinstitute.org

gnomAD Dataset

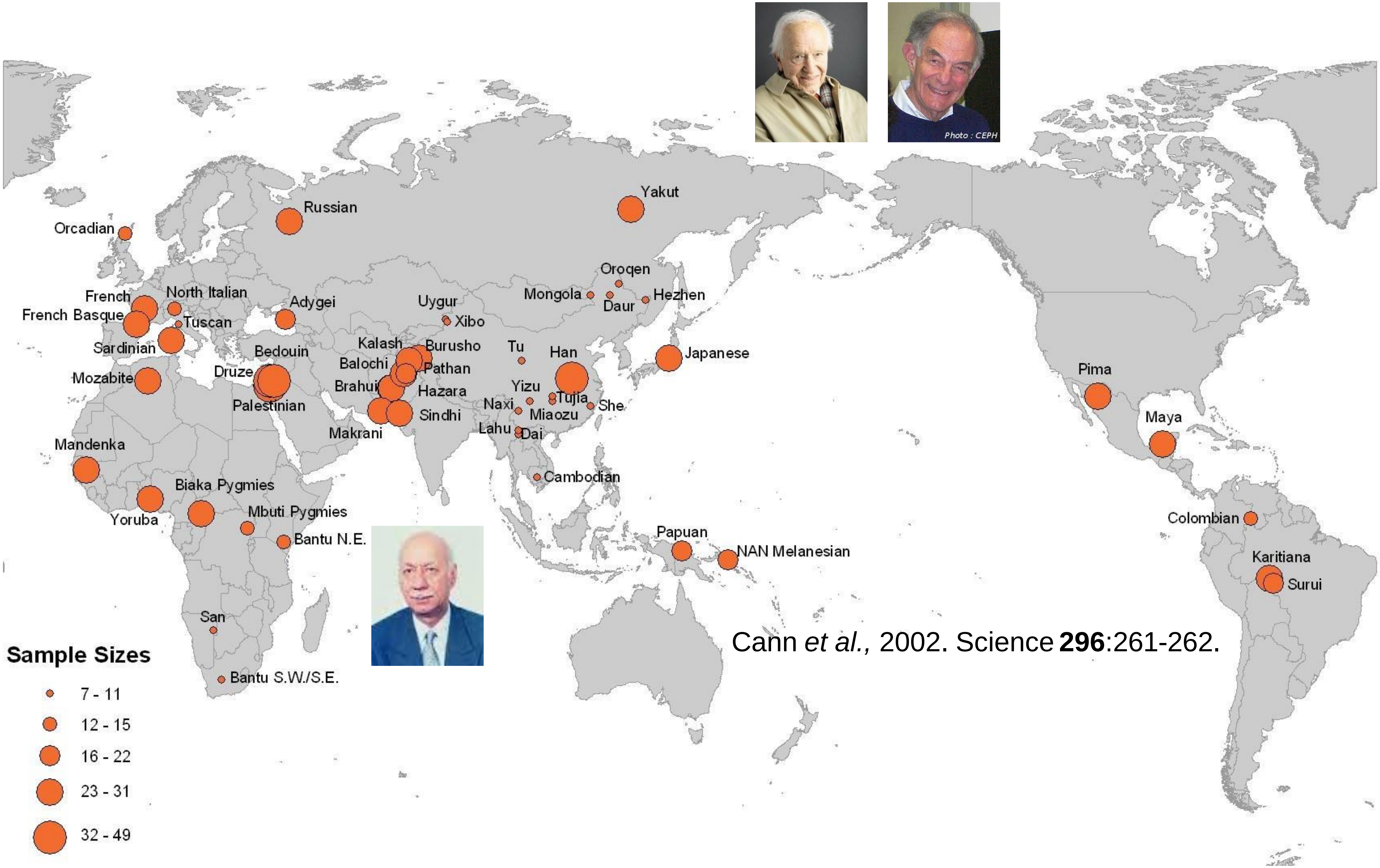


POPULATION	DESCRIPTION	GENOMES	EXOMES	TOTAL
AFR	African/African American	4,368	7,652	12,020
AMR	Admixed American	419	16,791	17,210
ASJ	Ashkenazi Jewish	151	4,925	5,076
EAS	East Asian	811	8,624	9,435
FIN	Finnish	1,747	11,150	12,897
NFE	Non-Finnish European	7,509	55,860	63,369
SAS	South Asian	0	15,391	15,391
OTH	Other (population not assigned)	491	2,743	3,234
	Total	15,496	123,136	138,632



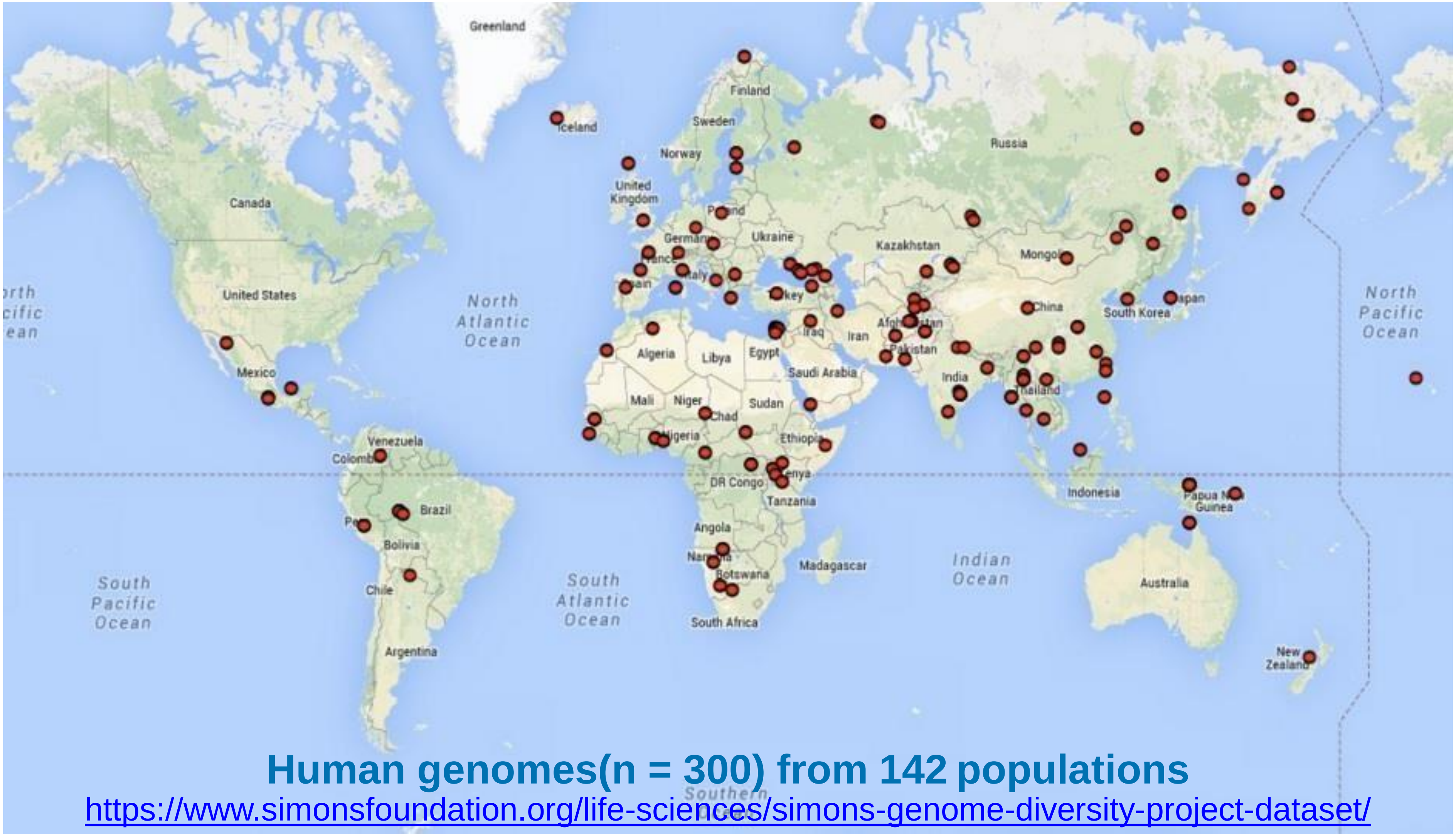
https://gnomad.broadinstitute.org/about

The HGDP-CEPH Cell Line Panel



Cann *et al.*, 2002. Science **296**:261-262.

Simons Genome Diversity Project Dataset

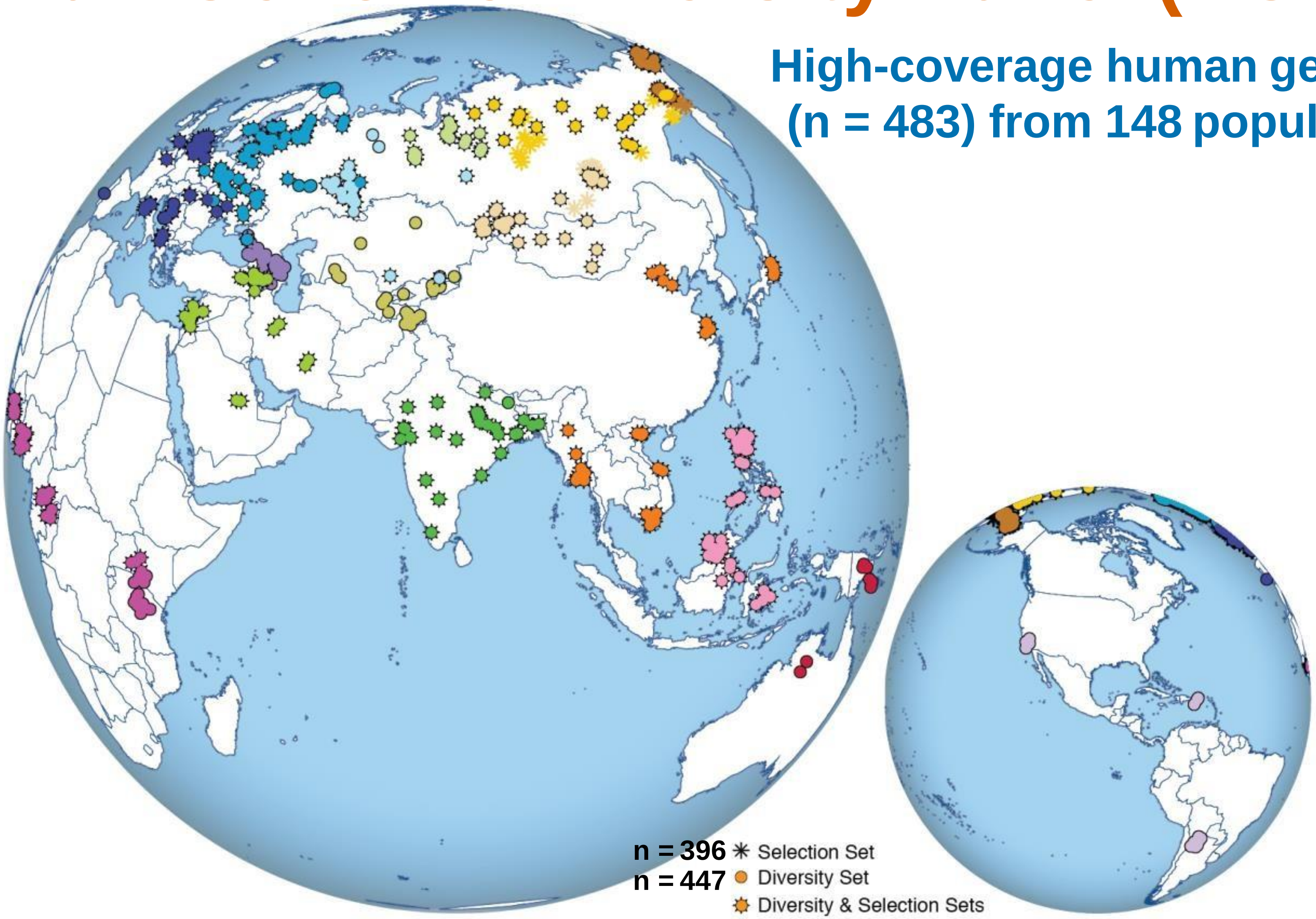


Human genomes(n = 300) from 142 populations

https://www.simonsfoundation.org/life-sciences/simons-genome-diversity-project-dataset/

Mallick *et al*., 2016. Nature 538:201-206.

Estonian Biocentre
Human Genome Diversity Panel (EGDP)



High-coverage human genomes (n = 483) from 148 populations

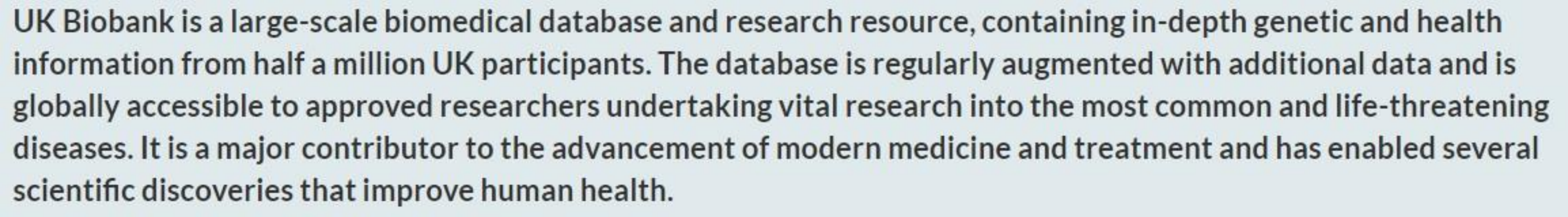
n = 396 ✳ Selection Set
n = 447 ● Diversity Set
✸ Diversity & Selection Sets

Pagani *et al*., 2016. Nature 538:238-242.

UK Biobank

https://www.ukbiobank.ac.uk/

Enabling scientific discoveries that improve human health



UK Biobank is a large-scale biomedical database and research resource, containing in-depth genetic and health information from half a million UK participants. The database is regularly augmented with additional data and is globally accessible to approved researchers undertaking vital research into the most common and life-threatening diseases. It is a major contributor to the advancement of modern medicine and treatment and has enabled several scientific discoveries that improve human health.

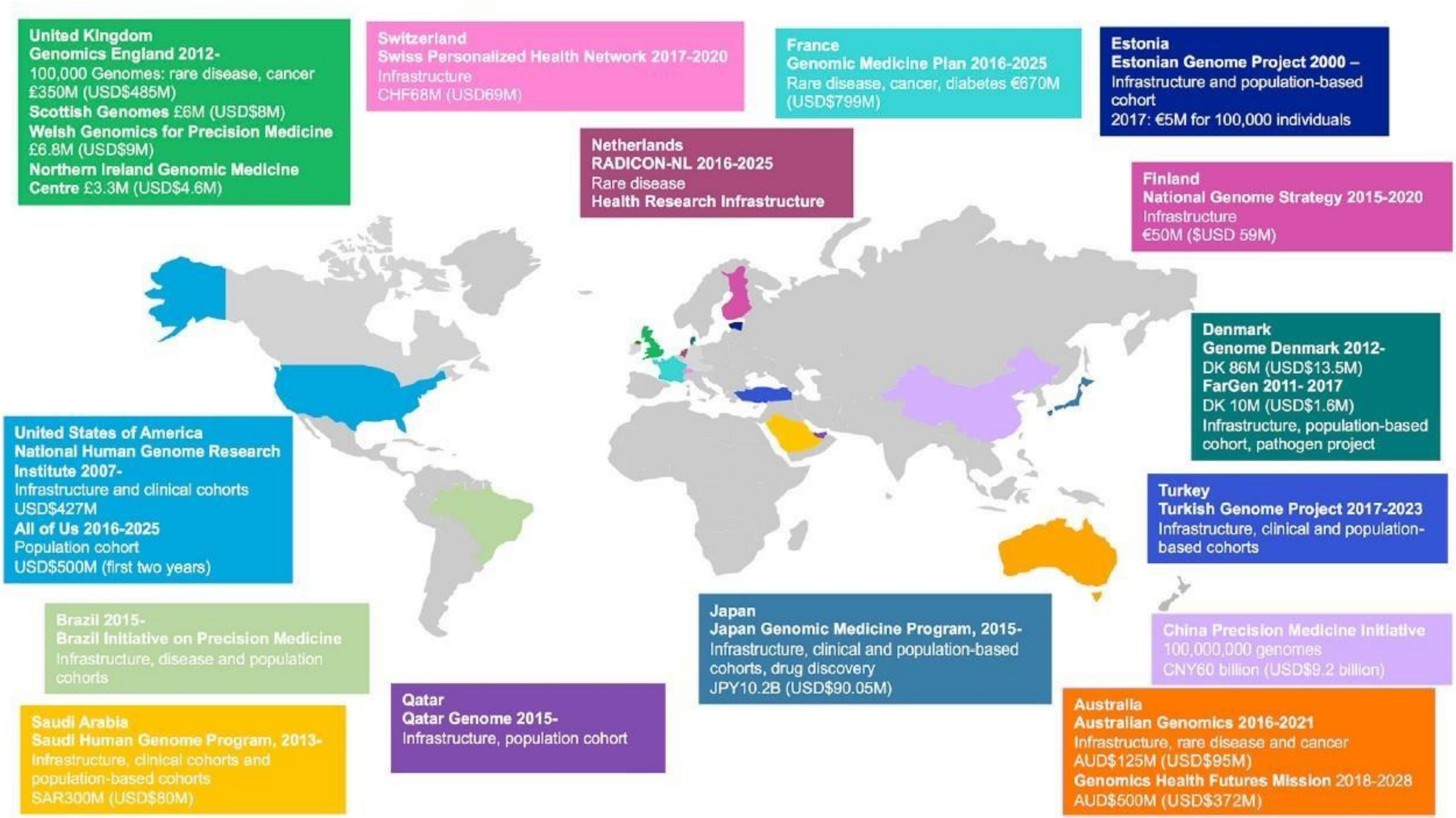
TOPMed Program

- **Trans-Omics for Precision Medicine (TOPMed) Program funded by NIH.**

- The goal of the TOPMed program is to generate scientific resources that will improve the understanding of heart, lung, blood, and sleep disorders and advance precision medicine.

https://www.nhlbiwgs.org/

Revolution in Personalized Medicine



United Kingdom
Genomics England 2012-
100,000 Genomes: rare disease, cancer
£350M (USD$485M)
Scottish Genomes £6M (USD$8M)
Welsh Genomics for Precision Medicine
£6.8M (USD$9M)
Northern Ireland Genomic Medicine Centre £3.3M (USD$4.6M)

Switzerland
Swiss Personalized Health Network 2017-2020
Infrastructure
CHF68M (USD69M)

France
Genomic Medicine Plan 2016-2025
Rare disease, cancer, diabetes €670M (USD$799M)

Estonia
Estonian Genome Project 2000 –
Infrastructure and population-based cohort
2017: €5M for 100,000 individuals

Netherlands
RADICON-NL 2016-2025
Rare disease
Health Research Infrastructure

Finland
National Genome Strategy 2015-2020
Infrastructure
€50M ($USD 59M)

Denmark
Genome Denmark 2012-
DK 86M (USD$13.5M)
FarGen 2011- 2017
DK 10M (USD$1.6M)
Infrastructure, population-based cohort, pathogen project

United States of America
National Human Genome Research Institute 2007-
Infrastructure and clinical cohorts
USD$427M
All of Us 2016-2025
Population cohort
USD$500M (first two years)

Turkey
Turkish Genome Project 2017-2023
Infrastructure, clinical and population-based cohorts

Brazil 2015-
Brazil Initiative on Precision Medicine
Infrastructure, disease and population cohorts

Japan
Japan Genomic Medicine Program, 2015-
Infrastructure, clinical and population-based cohorts, drug discovery
JPY10.2B (USD$90.05M)

China Precision Medicine Initiative
100,000,000 genomes
CNY60 billion (USD$9.2 billion)

Saudi Arabia
Saudi Human Genome Program, 2013-
Infrastructure, clinical cohorts and population-based cohorts
SAR300M (USD$80M)

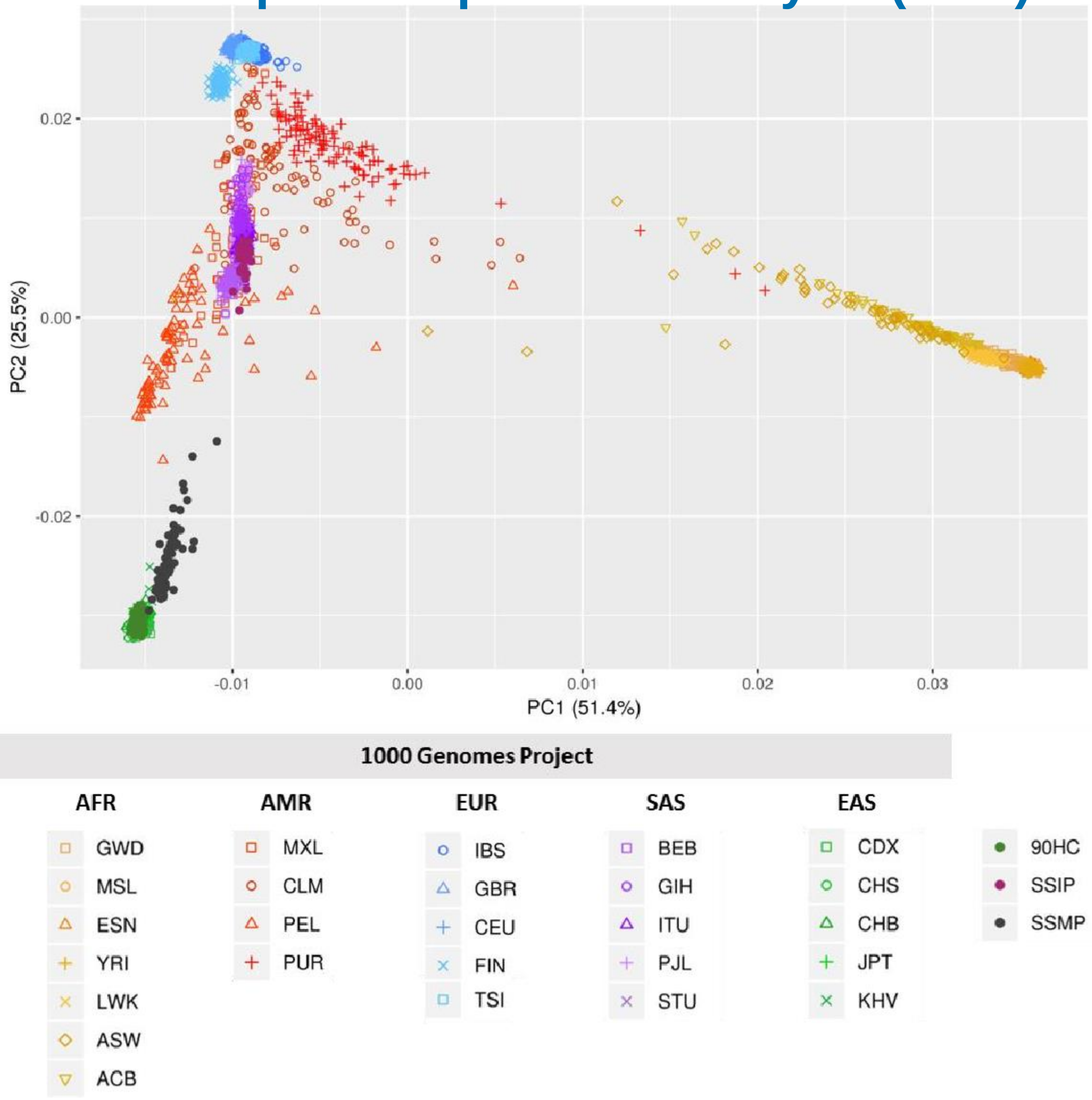
Qatar
Qatar Genome 2015-
Infrastructure, population cohort

Australia
Australian Genomics 2016-2021
Infrastructure, rare disease and cancer
AUD$125M (USD$95M)
Genomics Health Futures Mission 2018-2028
AUD$500M (USD$372M)

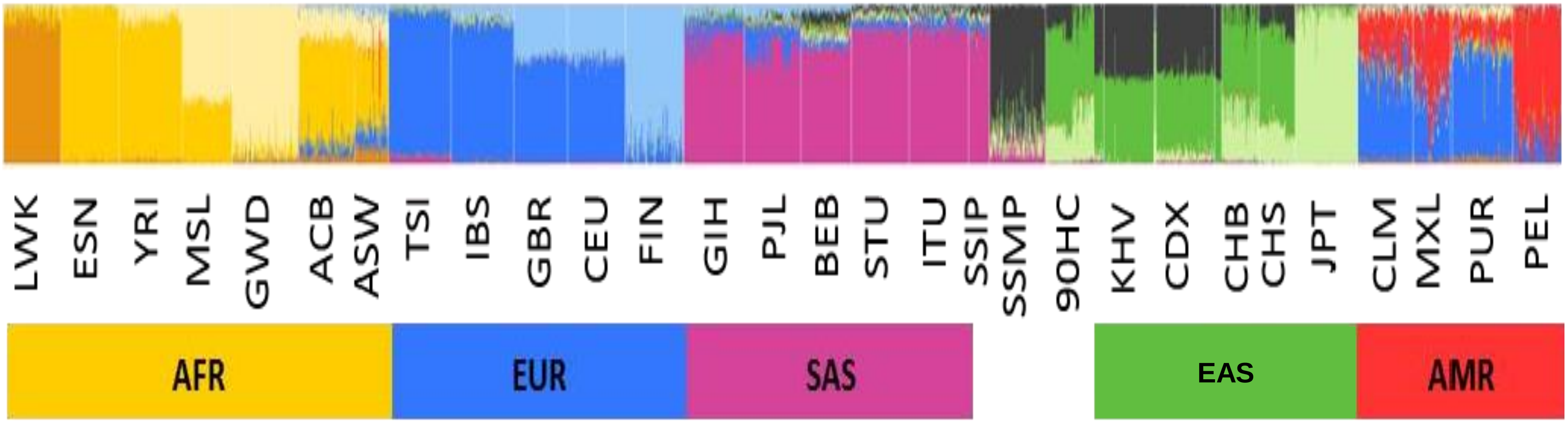
Stark *et al.* Am J Hum Genet (2019) 104:13-20.

Worldwide Population Relationships

Principal Components Analysis (PCA)



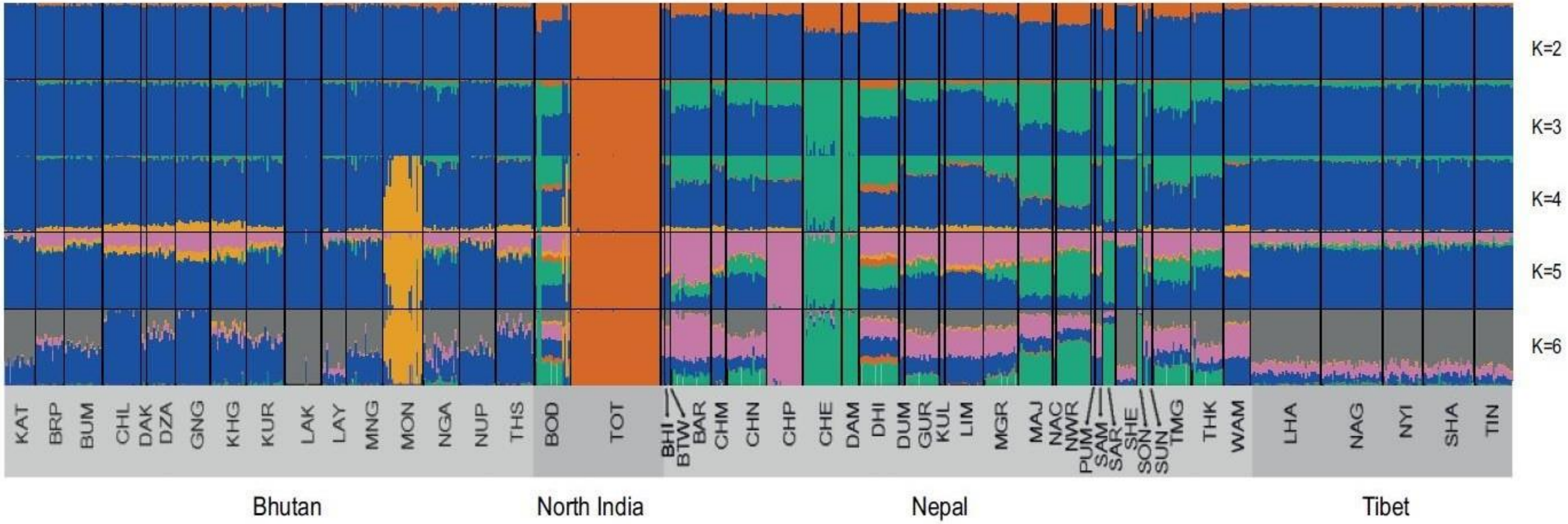
ADMIXTURE Analysis



- ➢ African populations are genetically more diverse than non-Africans.
- ➢ Genetic diversity outside of Africa tends to be a subset of the diversity within Africa.

Genetic Drift

- Strong effect of random fluctuations in allele frequencies due to population isolation and bottlenecks.



Arciero *et al.* Mol Biol Evol (2018) 35:1916-1933.

Population Stratification and GWAS

- GWAS can be confounded by population stratification—**systematic ancestry differences between cases and controls leading to a spurious association.**

- These associations may appear to be significant, but they are driven by the cohort’s relatedness rather than variants that truly affect trait or disease risk.

- Failure to control for it may lead to confounding, causing a study to fail for lack of significant results or resources to be wasted following false positive signals.

How Does it Occur?

- Whenever there are substantial variation across ethnicities in the frequency of the variant genotype being considered.

- If there is substantial variation across ethnicities in disease rates after adjustment for risk factors, other than the genotype of interest, that were collected in the study.

- The allele frequencies track with the disease rates across ethnicities, for reasons other than the effects of the allele of interest. For example, an allele with a clade or gradient of increasing frequency from North to South Asia might track with another factor, such as dietary differences or air pollution, that affects disease risk, thus, introducing bias from population stratification when studying the effect of the allele.

- Self-reported ethnic information from study participants does not reduce bias to an acceptable level.

Wacholder *et al. Cancer Epidemiology, Biomarkers and Prevention* (2002) 11:513-520.

Population Stratification

	Cases	Controls
A	101	20
a	20	101

OR = 25.5

	Cases	Controls
A	100	10
a	10	1

OR = 1

Europeans

	Cases	Controls
A	1	10
a	10	100

OR = 1

East Asians

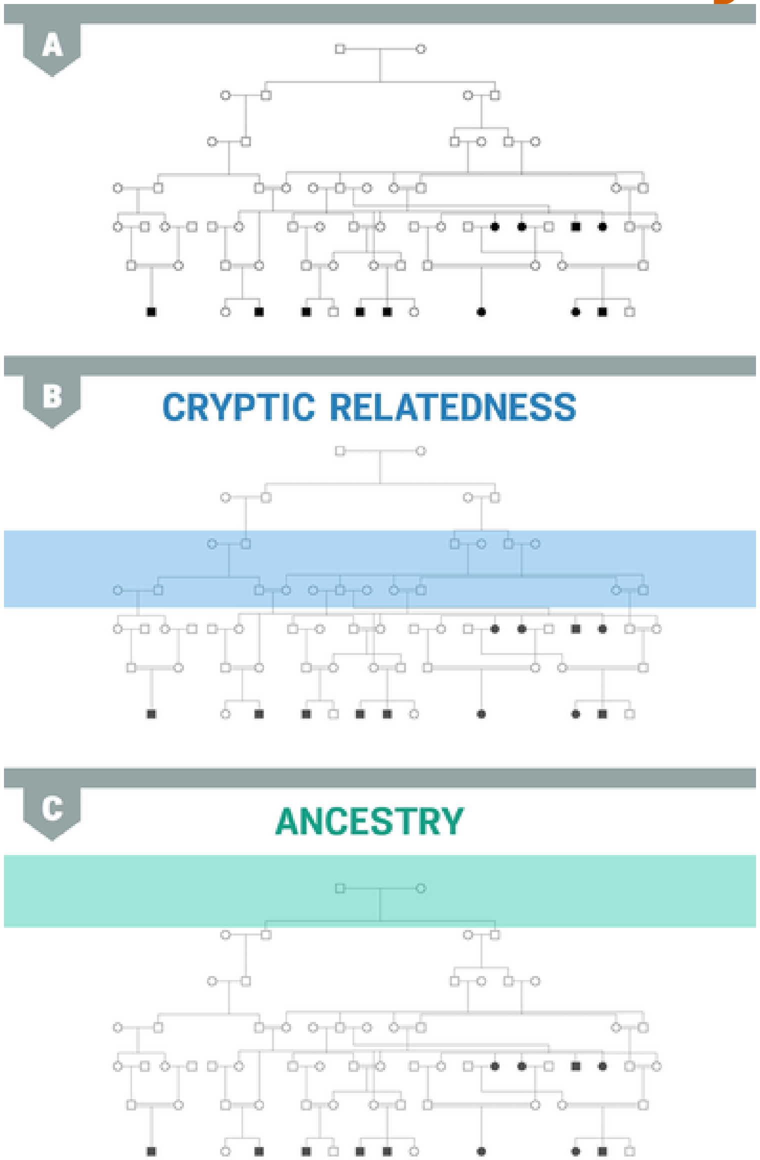
Individual Relatedness

- Ancestry differences:
  - Ancestry differences refer to different ancestry among individuals in a study.
  - If an association study contains individuals from different populations or differing degrees of admixture, the individuals will have different degrees of relatedness among them
- Cryptic relatedness:
  - Cryptic relatedness exists when some individuals are closely related, but this shared ancestry is unknown to the investigators and the study subjects.

Sul *et al.* PLoS Genetics (2018) 14:e1007309.

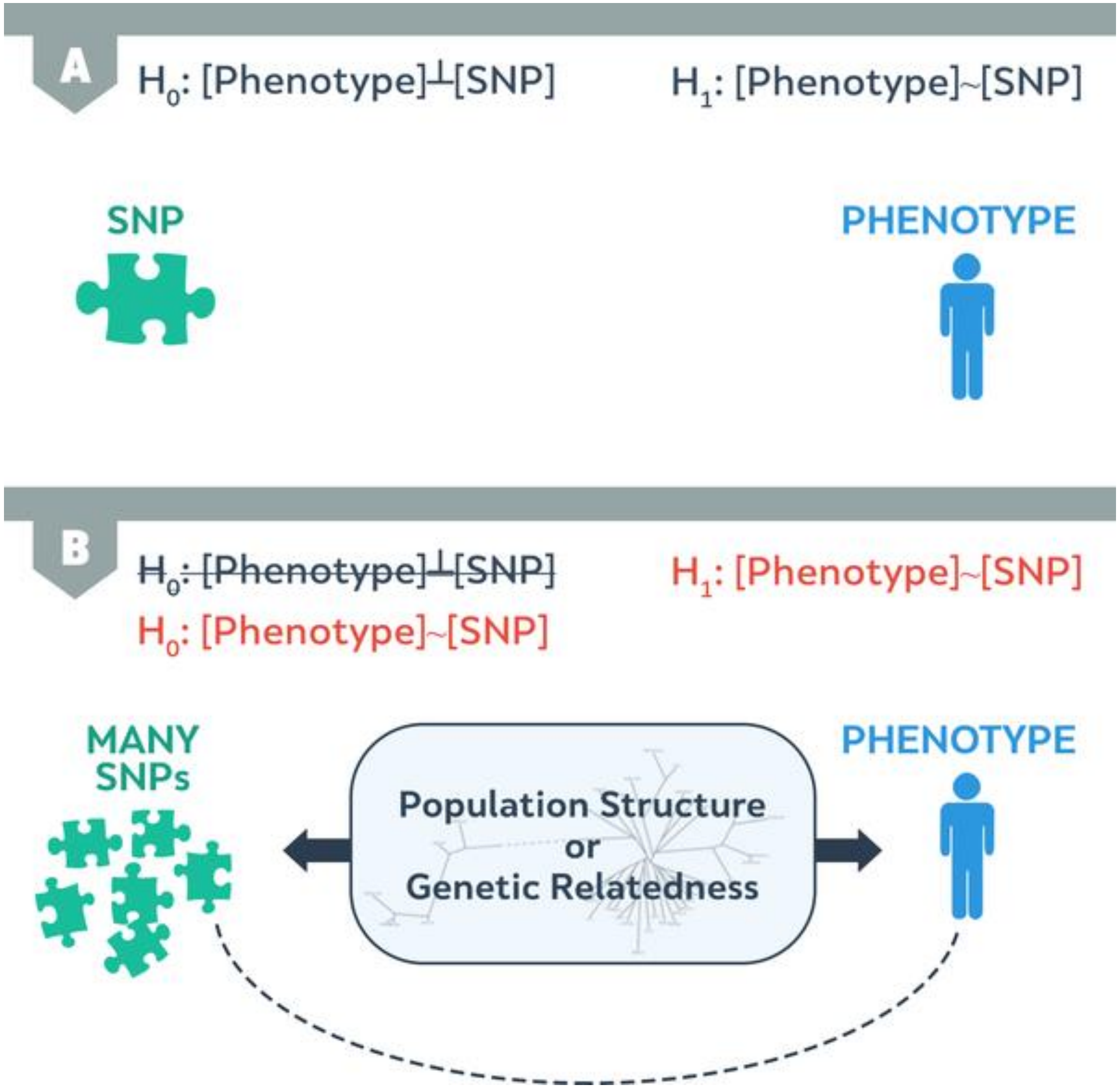
https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007309

Shared Ancestry



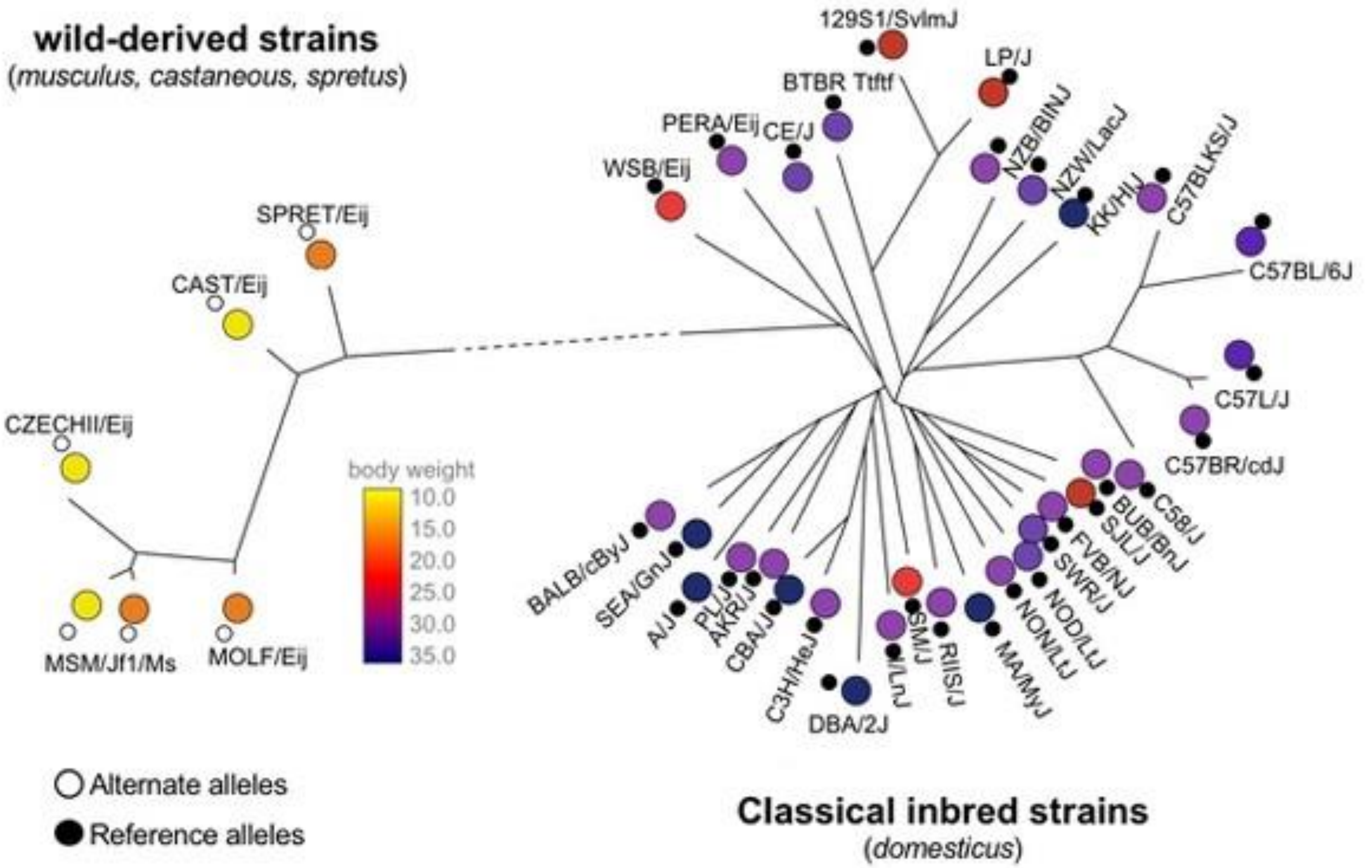
Sul *et al.* PLoS Genetics (2018) 14:e1007309.

Population Stratification and GWAS



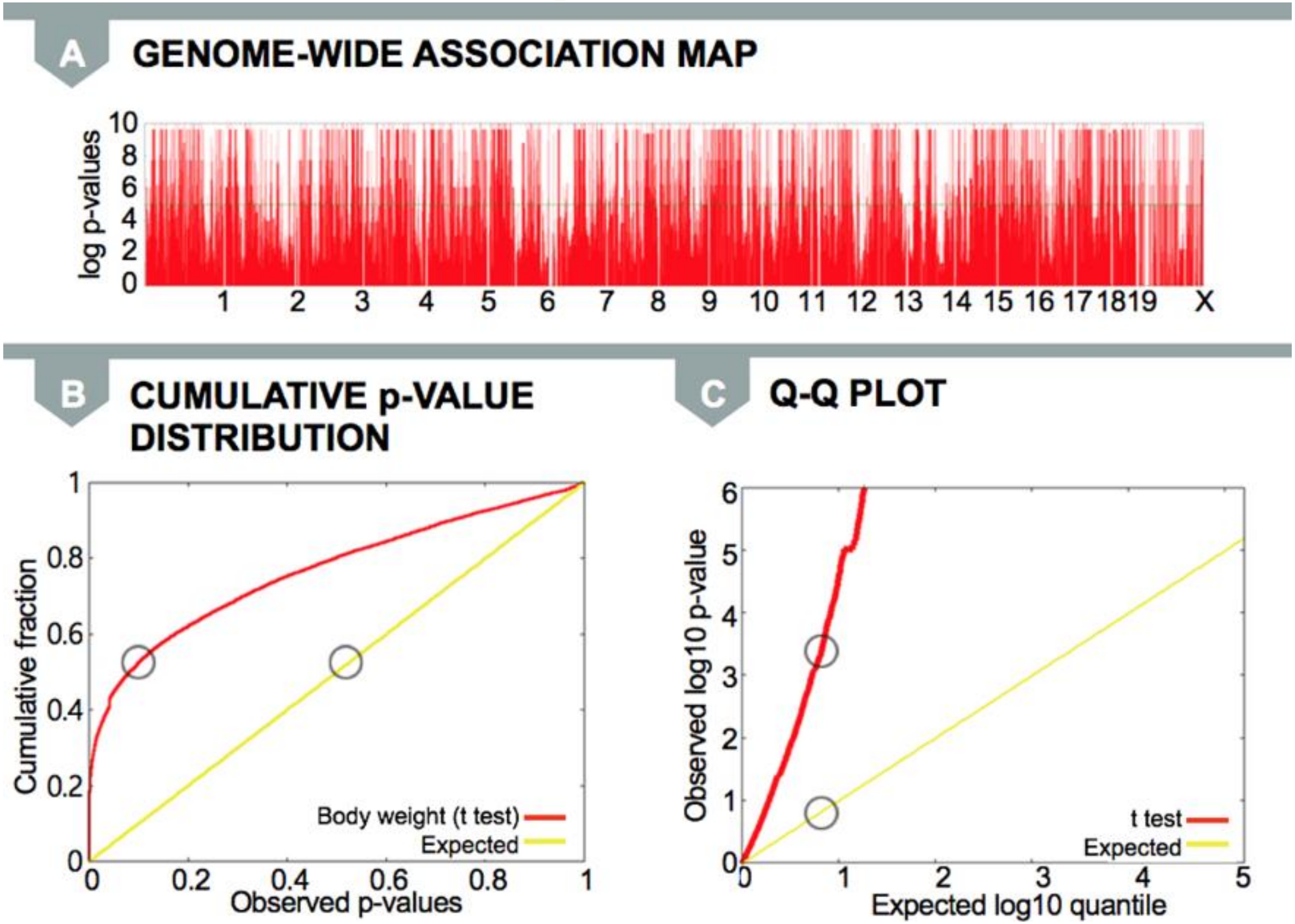
Sul *et al.* PLoS Genetics *(*2018) 14:e1007309.

Individual Relatedness



Sul *et al.* PLoS Genetics (2018) 14:e1007309.

Effect of Population Structure

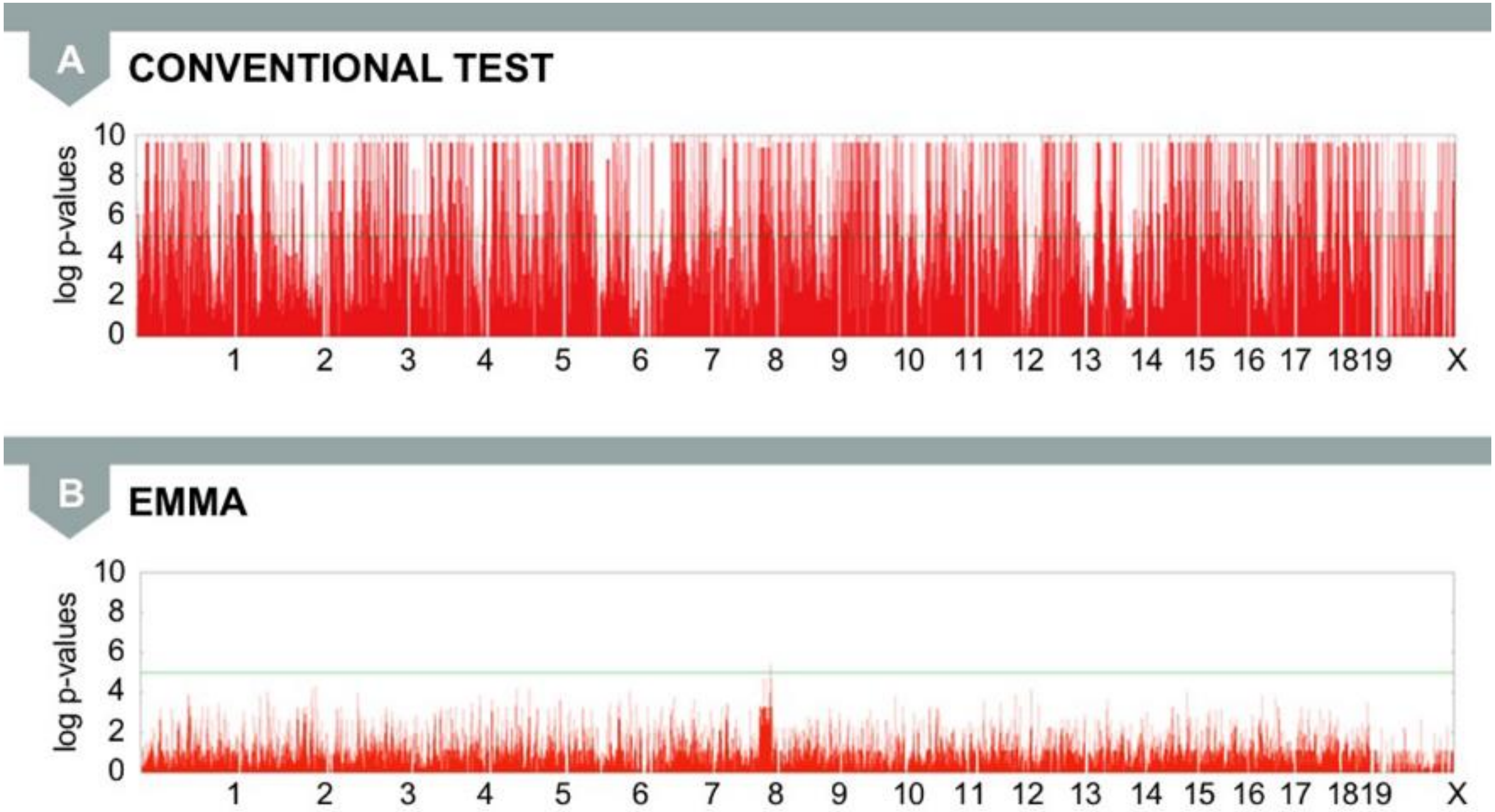


Sul *et al.* PLoS Genetics *(*2018) 14:e1007309.

Correcting for Population Stratification

- Replication in different populations:
  - Major bias in the same direction in populations with substantially different ethnic mixes is very unlikely because the conditions that allowed major bias are unlikely to be repeated.
- Genomic control markers:
  - Genomic control uses markers unrelated to disease to correct the bias.
- Correcting for population substructure:
  - To resolve the problem with population structure with the use of principal components (PC) of the genotyped dataset to model population relationships, which could be interpreted as a proxy for ancestry information, and included in the model as covariates.
  - Implemented as mixed linear model such as **E**fficient **M**ixed **M**odel **A**ssociation e**X**pedited (EMMAX) in which one SNP is fit in the model as a fixed covariate and, at the same time, a relationship matrix corrects for population structure.

Effect of Population Structure



Sul *et al.* PLoS Genetics *(*2018) 14:e1007309.

Questions?

Post Questions on the Slack Channel

qasim.ayub@monash.edu

Practical Exercise