Computational
Biology @
Berkeley

Analysis of Dataset Equity and Genetic Diversity in the 1000 Genomes Project and TCGA

Cyan, Sreyas, Ashwin, Bren, Frank

Cal Undergraduate Bioengineering Symposium

12/9/2025

Table of Contents

I Research Background

II Hypothesis & Objectives

III Methodology

IV Results

V Conclusion & Discussion

Introduction

- Many genomic and biomedical datasets that are used in biology research tend to under represent certain populations

- This can lead to biased insights in downstream research

- Students and researchers alike tend to use these datasets without actually questioning how representative it is

Knowledge Gap

The 1000 Genomes Project and TCGA are both large scale genomics consortia that generate publicly available datasets

- The diversity in terms of populations and genetically is not quantified

Chose 1000 Genomes and TCGA as they represent opposite ends of the representation spectrum

Both are widely used benchmark datasets in genomics and precision medicine, so inequities or biases in their composition have real downstream impact on research, tools, and clinical translation

Questions

1. How are the various superpopulations represented within each dataset?
2. How do the various superpopulations vary genetically?
3. How do the datasets compare in regards to these metrics?

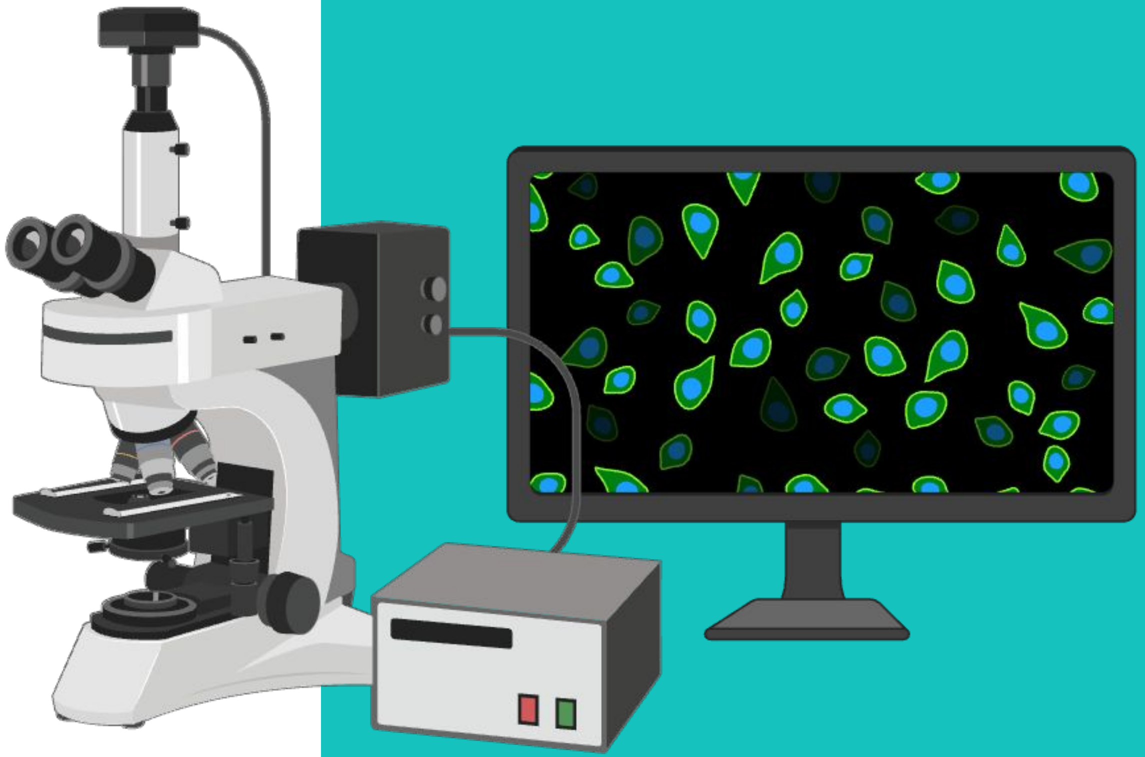
Hypothesis:

We hypothesize that both datasets will underrepresented the AFR superpopulation in their sampling

Research Methodology

We ran elementary statistics on both datasets

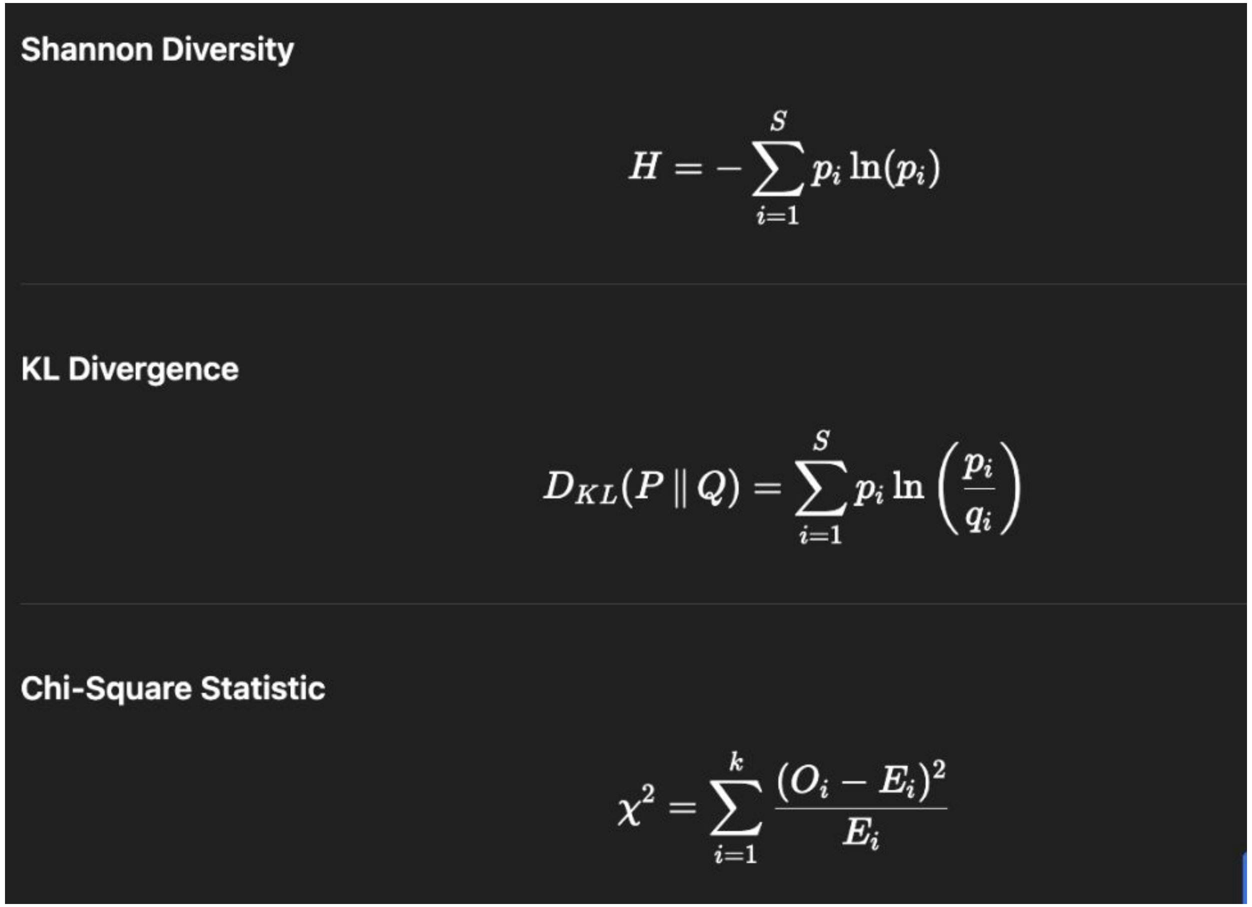
- Calculated population distribution
- Measured genomic diversity for available groups
- Compared representation gaps between datasets
- Interpreted patterns using demographic and evolutionary theory



Statistics We Used

- **Shannon Diversity Index**: Measures richness and evenness of categories in a sample, where higher values indicate greater and more balanced diversity
- **KL Divergence:** Measures how different one probability distribution is from another, where values closer to 0 indicate the distributions are more similar (better fit)
- **Chi-Square (χ^2):** Measures how far observed data deviate from expected values, where lower values indicate a better fit to the expected model
- **P-value:** Measures how likely the observed results are under the null hypothesis, where lower values (typically **< 0.05**) indicate stronger evidence against the null
- **π (nucleotide diversity):** the average number of pairwise nucleotide differences per site between two randomly chosen individuals in a population
- **H□ (expected heterozygosity):** the probability that two alleles sampled from the same population are different (i.e., the fraction of sites expected to be heterozygous)

Equations



Shannon Diversity

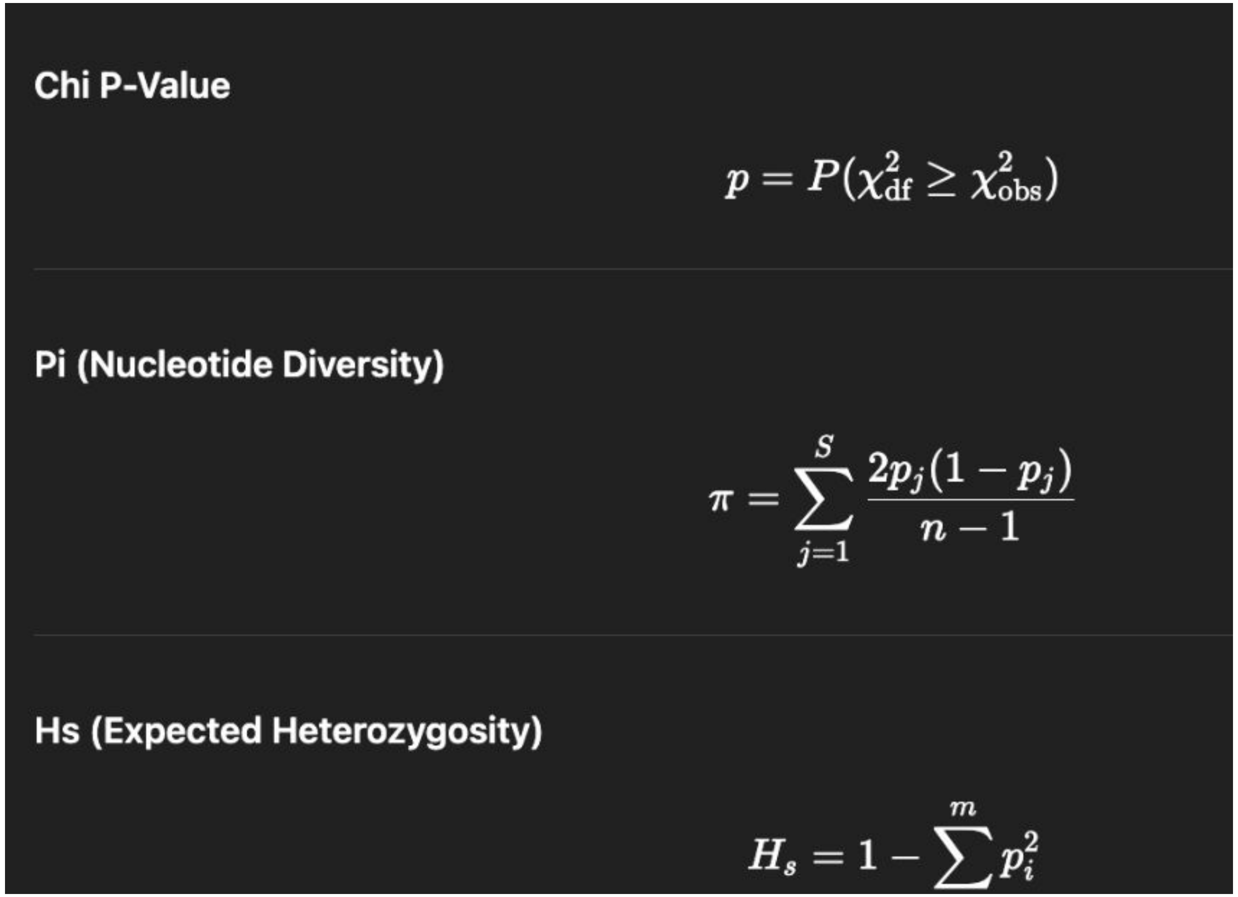
$$H = -\sum_{i=1}^{S} p_i \ln(p_i)$$

KL Divergence

$$D_{KL}(P \parallel Q) = \sum_{i=1}^{S} p_i \ln\left(\frac{p_i}{q_i}\right)$$

Chi-Square Statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$



Chi P-Value

$$p = P(\chi^2_{\mathrm{df}} \geq \chi^2_{\mathrm{obs}})$$

Pi (Nucleotide Diversity)

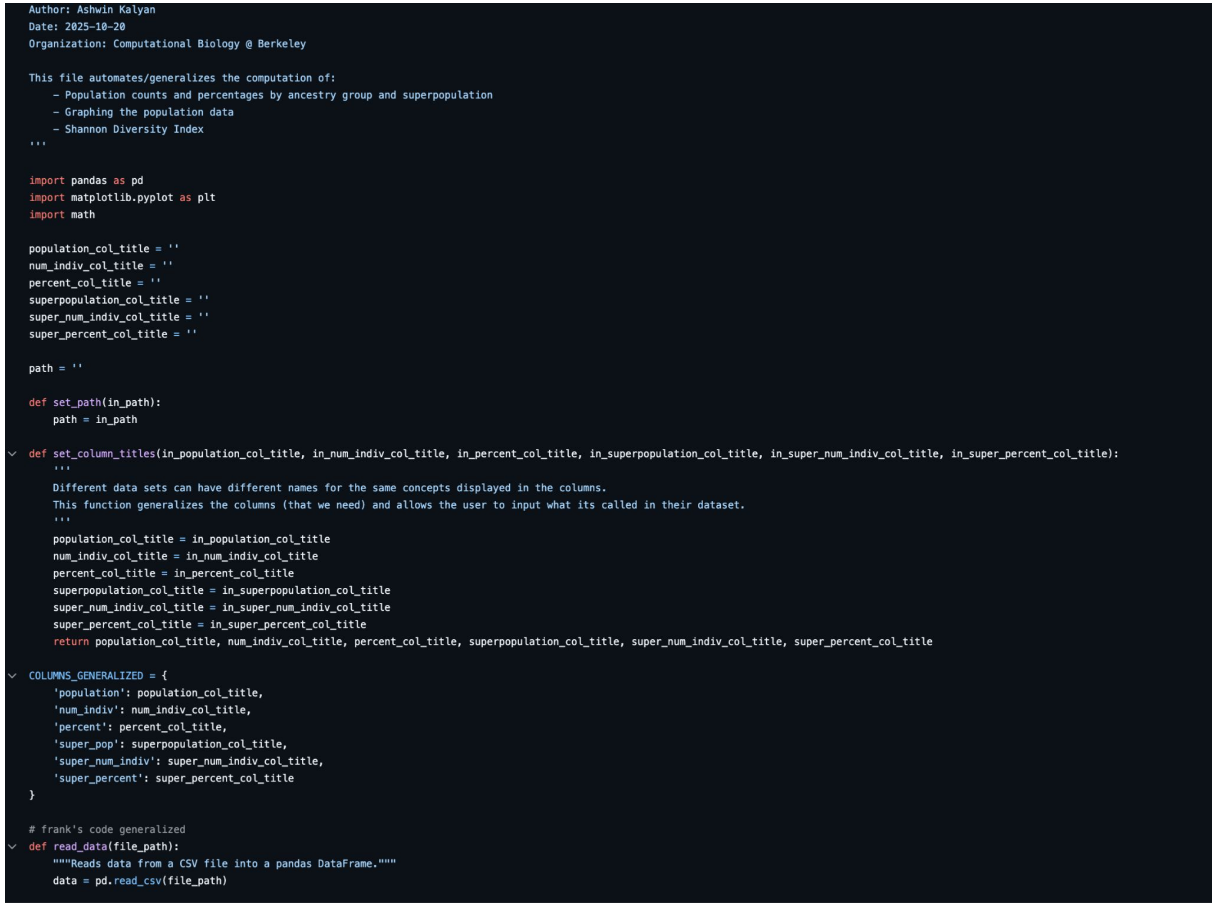
$$\pi = \sum_{j=1}^{S} \frac{2p_j(1 - p_j)}{n - 1}$$

Hs (Expected Heterozygosity)

$$H_s = 1 - \sum^{m} p_i^2$$

Comparison of Diversity Metrics in 1000 Genomes and TCGA

- Computed same statistics for both datasets
- Compared them to find any significant results
- Brainstormed ways to improve possible inequities

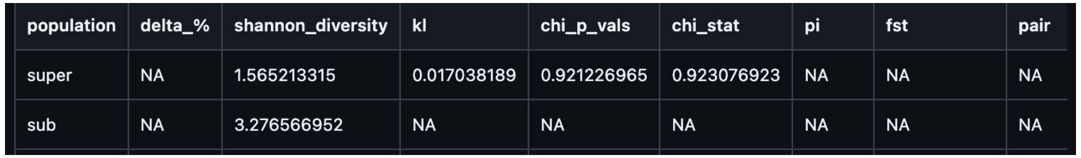


```
Author: Ashwin Kalyan
Date: 2025-10-20
Organization: Computational Biology @ Berkeley

This file automates/generalizes the computation of:
    - Population counts and percentages by ancestry group and superpopulation
    - Graphing the population data
    - Shannon Diversity Index
'''

import pandas as pd
import matplotlib.pyplot as plt
import math

population_col_title = ''
num_indiv_col_title = ''
percent_col_title = ''
superpopulation_col_title = ''
super_num_indiv_col_title = ''
super_percent_col_title = ''

path = ''

def set_path(in_path):
    path = in_path

def set_column_titles(in_population_col_title, in_num_indiv_col_title, in_percent_col_title, in_superpopulation_col_title, in_super_num_indiv_col_title, in_super_percent_col_title):
    '''
    Different data sets can have different names for the same concepts displayed in the columns.
    This function generalizes the columns (that we need) and allows the user to input what its called in their dataset.
    '''
    population_col_title = in_population_col_title
    num_indiv_col_title = in_num_indiv_col_title
    percent_col_title = in_percent_col_title
    superpopulation_col_title = in_superpopulation_col_title
    super_num_indiv_col_title = in_super_num_indiv_col_title
    super_percent_col_title = in_super_percent_col_title
    return population_col_title, num_indiv_col_title, percent_col_title, superpopulation_col_title, super_num_indiv_col_title, super_percent_col_title

COLUMNS_GENERALIZED = {
    'population': population_col_title,
    'num_indiv': num_indiv_col_title,
    'percent': percent_col_title,
    'super_pop': superpopulation_col_title,
    'super_num_indiv': super_num_indiv_col_title,
    'super_percent': super_percent_col_title
}

# frank's code generalized
def read_data(file_path):
    """Reads data from a CSV file into a pandas DataFrame."""
    data = pd.read_csv(file_path)
```

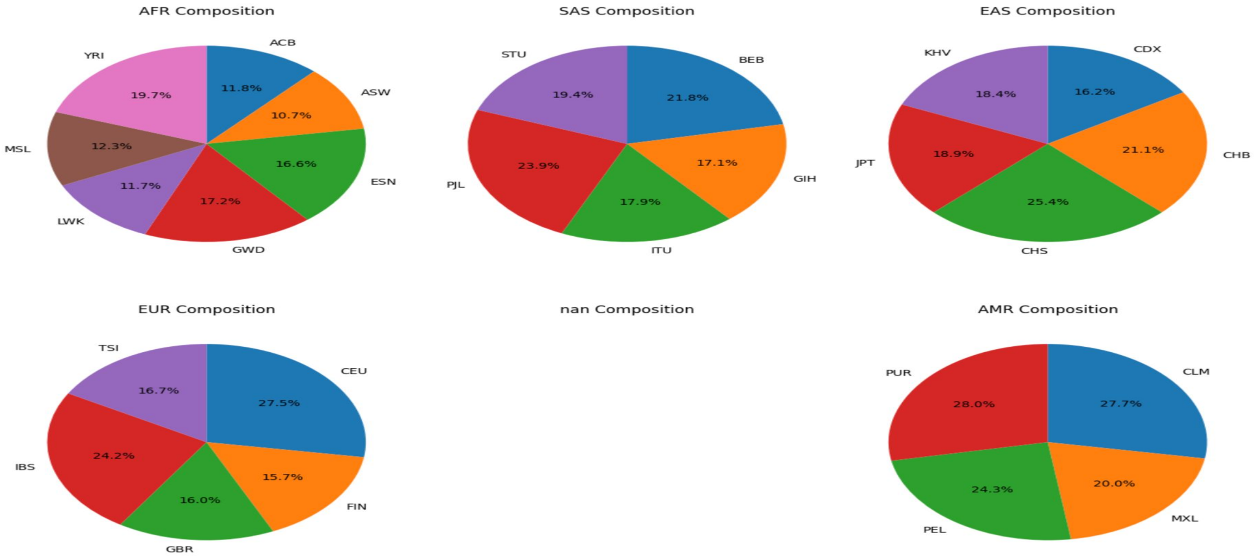
Example code

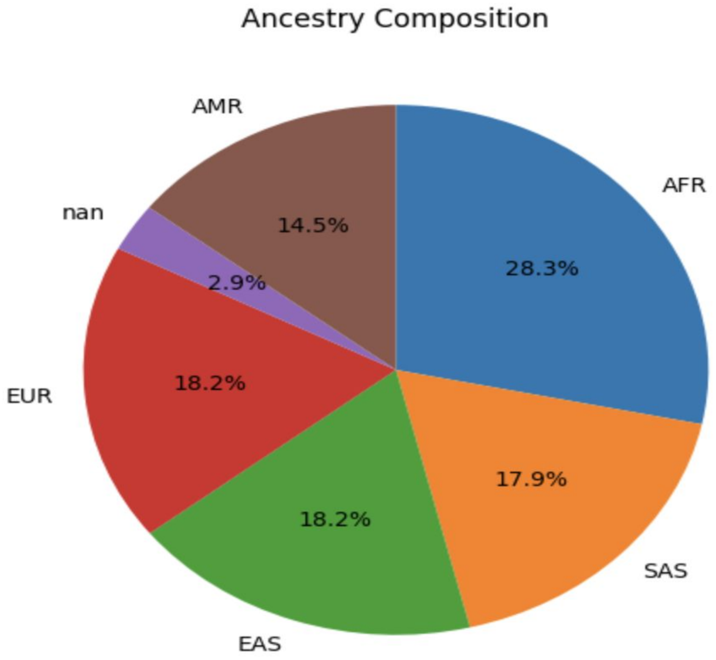
Comparison of Diversity Metrics in 1000s Genomes

- Pretty equitable
- Overall even distribution
- Signifies more equal sampling and a better representation of data to be used



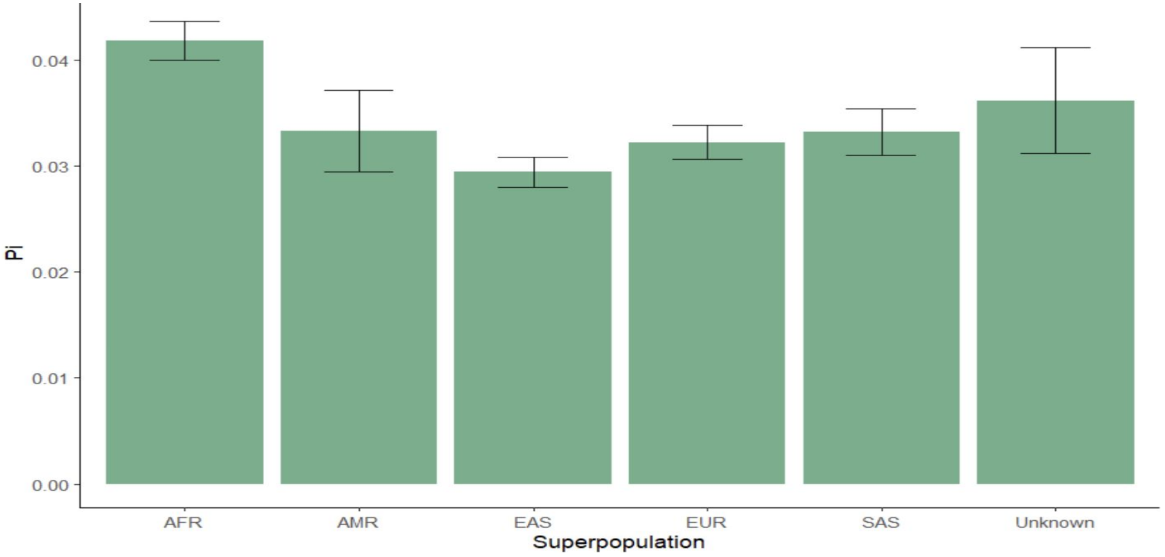
population	delta_%	shannon_diversity	kl	chi_p_vals	chi_stat	pi	fst	pair
super	NA	1.565213315	0.017038189	0.921226965	0.923076923	NA	NA	NA
sub	NA	3.276566952	NA	NA	NA	NA	NA	NA

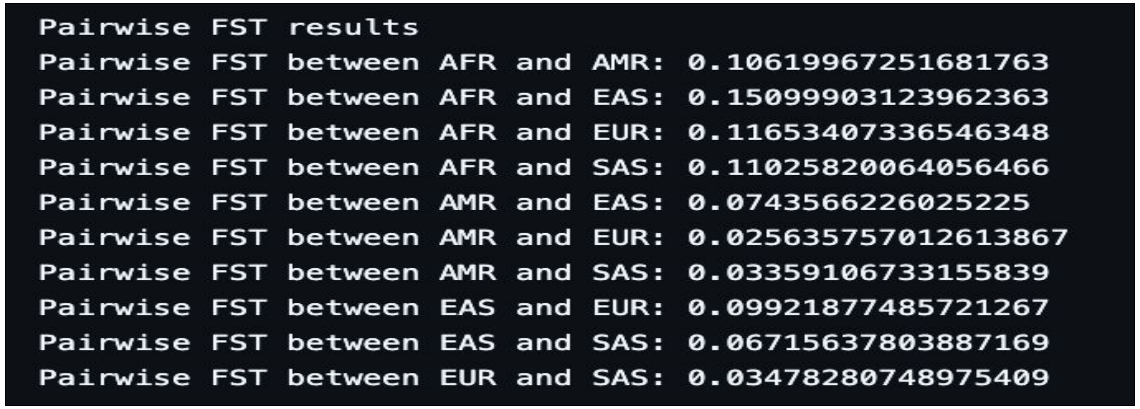




Diversity Metrics in 1000s Genomes

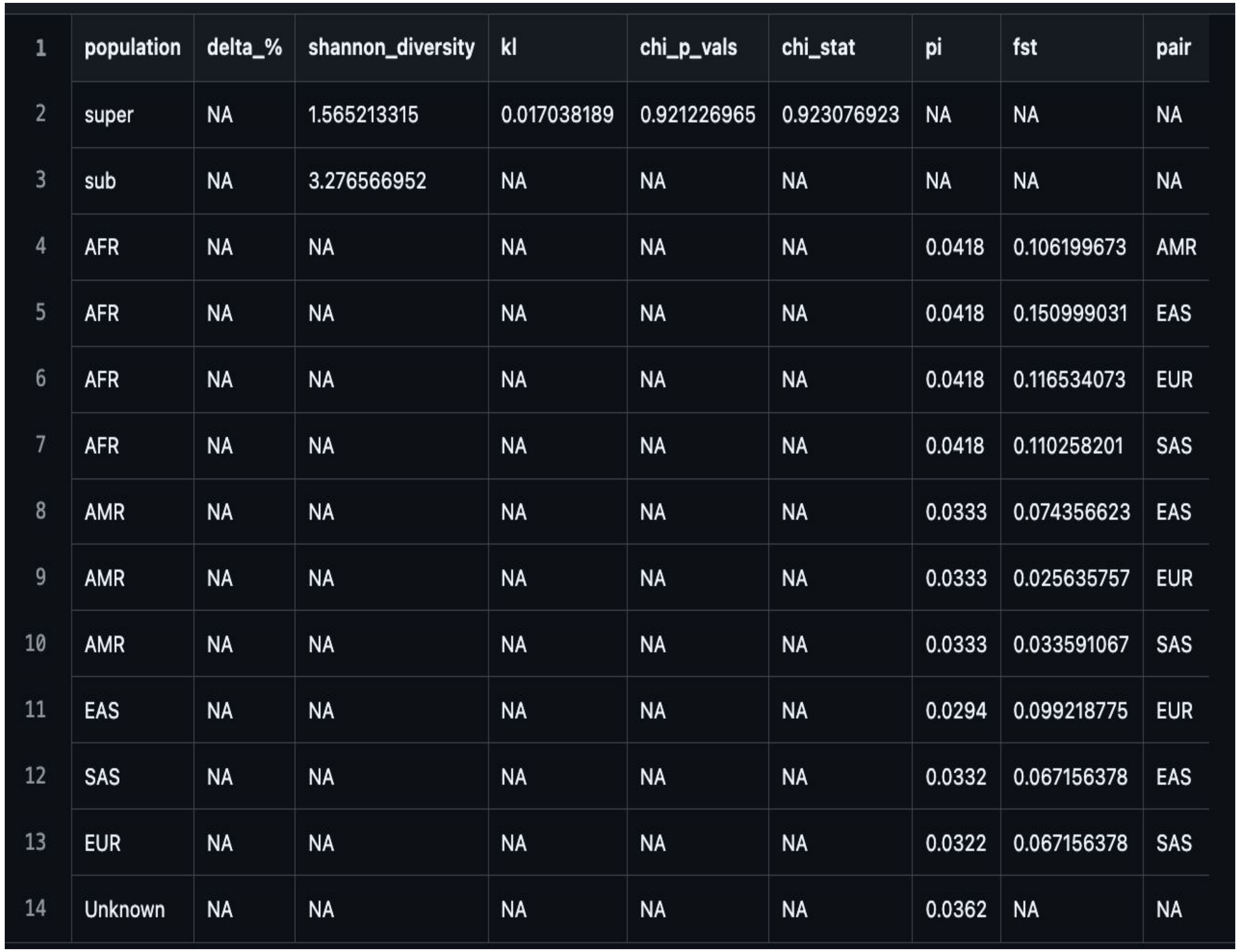
- Genomic diversity shows similar pattern
- Shows a realistic distribution, with AFR the highest
- AFR shows the highest π and H□ because of larger long-term effective population size and weaker drift, allowing more mutations to arise and persist
    - Diversity declines in non-African groups due to bottlenecks during the Out-of-Africa expansion
- SAS > EUR ≈ EAS because South Asians experienced less severe bottlenecks and more admixture, while Europeans and East Asians underwent additional founder events that reduced diversity to similar levels




```
Pairwise FST results
Pairwise FST between AFR and AMR: 0.10619967251681763
Pairwise FST between AFR and EAS: 0.1509990312396236З
Pairwise FST between AFR and EUR: 0.11653407336546348
Pairwise FST between AFR and SAS: 0.11025820064056466
Pairwise FST between AMR and EAS: 0.0743566226025225
Pairwise FST between AMR and EUR: 0.025635757012613867
Pairwise FST between AMR and SAS: 0.03359106733155839
Pairwise FST between EAS and EUR: 0.09921877485721267
Pairwise FST between EAS and SAS: 0.06715637803887169
Pairwise FST between EUR and SAS: 0.03478280748975409
```

Diversity Metrics in 1000s Genomes

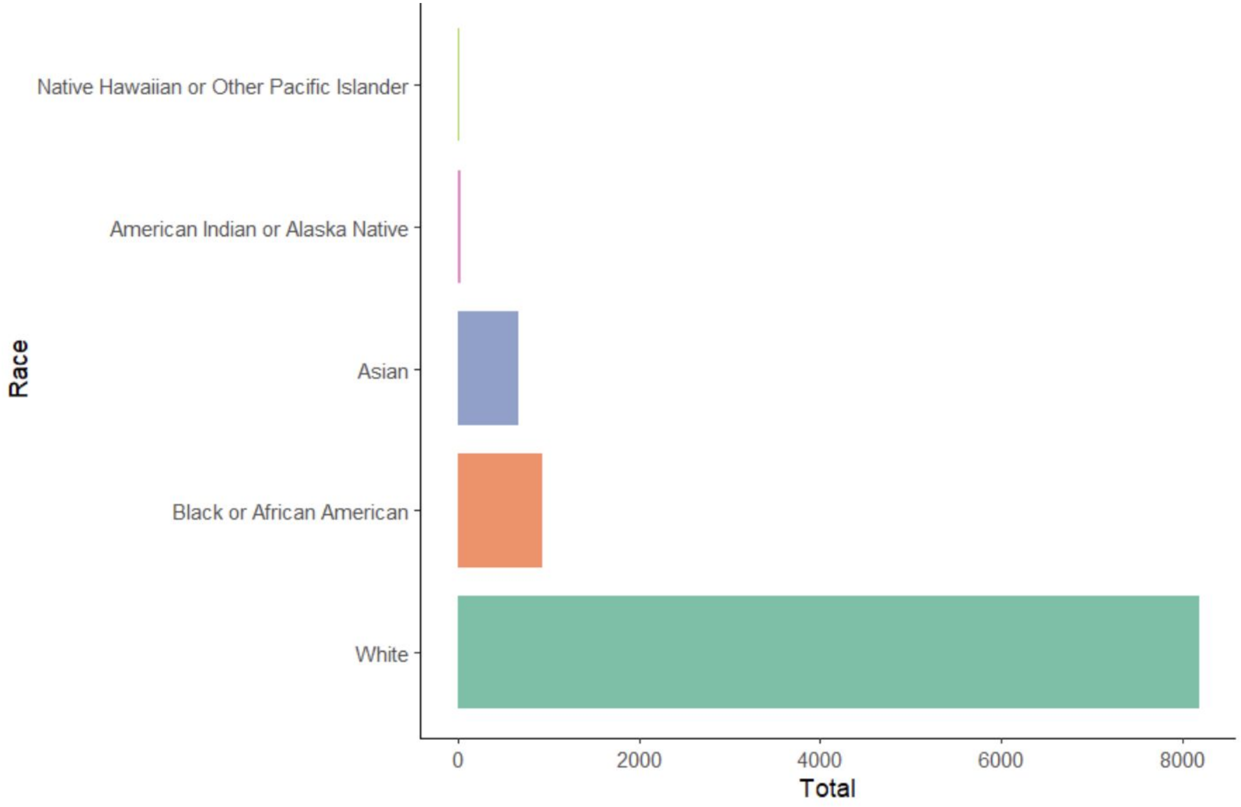
- Genomic diversity shows similar pattern
- Shows a realistic distribution, with AFR the highest
- AFR shows the highest π and H□ because of larger long-term effective population size and weaker drift, allowing more mutations to arise and persist
    - Diversity declines in non-African groups due to bottlenecks during the Out-of-Africa expansion
- SAS > EUR ≈ EAS because South Asians experienced less severe bottlenecks and more admixture, while Europeans and East Asians underwent additional founder events that reduced diversity to similar levels



population	delta_%	shannon_diversity	kl	chi_p_vals	chi_stat	pi	fst	pair
super	NA	1.565213315	0.017038189	0.921226965	0.923076923	NA	NA	NA
sub	NA	3.276566952	NA	NA	NA	NA	NA	NA
AFR	NA	NA	NA	NA	NA	0.0418	0.106199673	AMR
AFR	NA	NA	NA	NA	NA	0.0418	0.150999031	EAS
AFR	NA	NA	NA	NA	NA	0.0418	0.116534073	EUR
AFR	NA	NA	NA	NA	NA	0.0418	0.110258201	SAS
AMR	NA	NA	NA	NA	NA	0.0333	0.074356623	EAS
AMR	NA	NA	NA	NA	NA	0.0333	0.025635757	EUR
AMR	NA	NA	NA	NA	NA	0.0333	0.033591067	SAS
EAS	NA	NA	NA	NA	NA	0.0294	0.099218775	EUR
SAS	NA	NA	NA	NA	NA	0.0332	0.067156378	EAS
EUR	NA	NA	NA	NA	NA	0.0322	0.067156378	SAS
Unknown	NA	NA	NA	NA	NA	0.0362	NA	NA

Diversity Metrics in TCGA

- Much more skewed
- More inequitable
- Signifies more biased sampling and a worse representation of data, similar to how the healthcare industry looks like
- No verified and published statistics for genomic diversity, but we assume it to be low due to TCGA's well documented sampling biases



Summary of Results and Drawbacks

- Two very different sides of the spectrum, showing clear disparities in which groups are represented in major genomic datasets
- Indicates that health inequity still exists, because underrepresented populations receive fewer research insights, weaker diagnostic tools, and less effective precision-medicine advances
- Highlights the need for more inclusive data collection and equitable distribution of genomic resources

1000 Genomes Project — Drawbacks

- Underrepresentation of many global populations, which limits downstream analyses of diversity, allele frequencies, and ancestry-specific variant interpretation
- Low sequencing depth (~4–6× WGS for many samples) reduces sensitivity for detecting rare variants and structural variants

TCGA — Drawbacks

- Limited demographic diversity and strong sampling bias, with overrepresentation of White patients and certain cancer types
- Highly heterogeneous data generation protocols (different platforms, labs, and time periods), which can introduce batch effects and complicate cross cohort comparisons

Remaining Questions

- How consistent are these representation gaps across other major biomedical and genomic datasets?
- How much do missing ancestry annotations distort our estimates of diversity and equity?
- Can we quantify how these disparities translate into downstream clinical or research biases?

Next Steps

- Run the same analyses on additional datasets to test whether the disparities we found generalize
- Build a GAN-based model that can automatically detect inequities, generate insights, and provide recommendations for improving dataset representation

THANKS FOR LISTENING