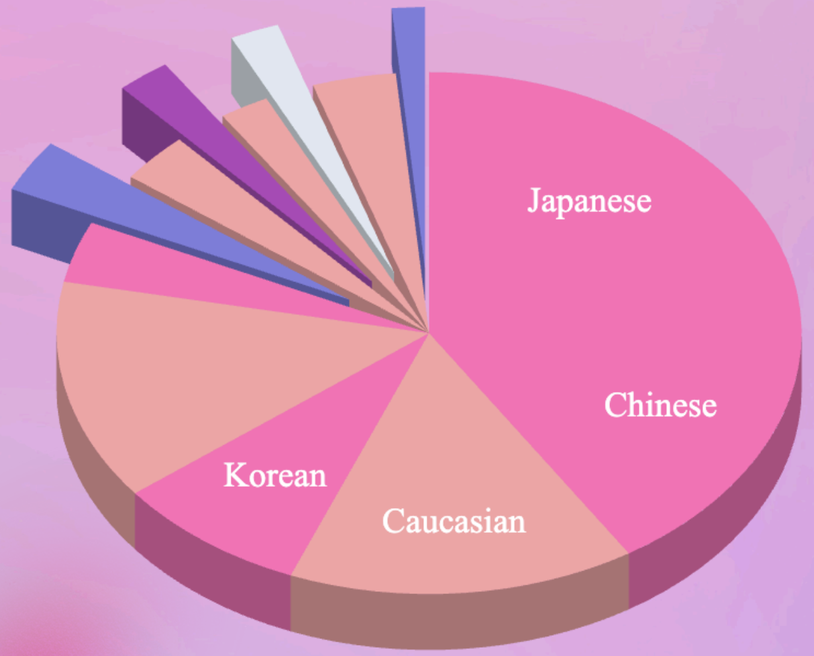
Underrepresented Populations

in the Genetic Association Database (GAD)

Jessica Meng
Khoury College of Computer Science '28

Northeastern University



Japanese Chinese Caucasian Korean Italian European
German Indian **Spanish** Finnish **Turkish** French
African American Polish Russian **Mexican**

```sql
USE gad;
-- Which populations are underrepresented in the Gene Association Database?
SELECT population, COUNT(population) AS num_records
FROM gad
WHERE population IS NOT NULL AND population <> ''
GROUP BY population
-- Finding most popular populations
HAVING num_records > 100
-- Finding the populations represented the most
ORDER BY num_records DESC;
-- However, since the data about population is not standardized
-- and the majority is null, this is a difficult question.
-- That is why I have decided to focus on this topic: I believe it highlights
-- information that the GAD can improve upon, and I hope to raise awareness of it.
```

To start, I want to emphasize that the information in the GAD is incomplete. Most data entries do not have a population listed. On top of that, entries that do have populations listed are not standardized, making data usage difficult.

With that being said, the most underrepresented populations from the data I gathered are African, Middle Eastern, and Hispanic/Latino populations. While disappointing, this is the result I expected. Research represented in the GAD primarily focuses on Caucasian and Asian populations, which limits the generalization of the findings to the global community. Also, expanding research in underrepresented populations is crucial for effective medical treatment across all diverse groups.

Becker et al. (2004). The Genetic Association Database. Nature Genetics 36(5):431-2

Northeastern University