Hacking Race

By Anne Kim, Anvita Pandit

Basic Info

https://defcon.org/

https://www.villageb.io/

https://twitter.com/DC_BHV

Friday August 9 - Sunday August 11

What I did last year

https://docs.google.com/presentation/d/1bl3etWRFsQoORF0suejwSwxAzc85bBcjX1tA9zXN7Ko/edit?usp=sharing

https://www.youtube.com/watch?v=Jni4o947OjA

Abstract

Have you ever wondered how "race" and ancestry are "calculated" by 23andMe or Ancestry.com? Have you ever questioned why your results seemed so "wrong"? Well, in this workshop we'll be building our own admixture (ancestry) reports from publicly available data (edit: possibly sequencer data if we're sponsored?) as well as showing how we can "hack race" to make racial results different.

Workshop should be <1:30 hours

Relevant Articles

https://qz.com/765879/23andme-has-a-race-problem-when-it-comes-to-ancestry-reports-for-non-whites/

https://gizmodo.com/how-dna-testing-botched-my-familys-heritage-and-probab-1820932637

https://www.theguardian.com/lifeandstyle/2018/aug/11/question-ancestry-does-dna-testing-really-understand-race

Outline of Subjects

- Git/version control: We'll build a github repo that people can download ahead of time. This will include all the necessary files as well as dependencies. We'll probably have some kind of test/build script for people to run in order to verify that they have all the packages they need before the workshop. That way no one needs to connect to the dangerous wifi internet during the workshop.
- Basic Linear algebra: this builds the intuition for the computational biology pieces
- Computational Biology: admixture calculations

Technical Intuition - Background

We are hacking ancestry or admixture

Technical Admixture involves

- Having labeled reference populations (french, chinese, korean, indian, etc)
- Clustering and defining the genetic commonality within each population genetically
- Projecting this definition onto a new DNA sample in order to get its admixture report (ancestry)

Technical Intuition - “Vulnerability”

We are hacking ancestry or admixture

Technical Admixture involves

- Having labeled reference populations (french, chinese, korean, indian, etc)
    - No one knows exactly what the 23andMe or Ancestry.com reference populations look like, and it is suspected that it is quite shallow especially in non-european countries (All of Africa, the Middle-East, Korea, Most of South America)
    - To demonstrate how fucked up this is, we will select very small subsets of reference populations such that you can purposefully mislabel a sample of DNA. We will select these references based on the sample in order to boost goal signal and increase alternative race noise.
    - Sample reference populations: https://www.genome.gov/27528684/1000-genomes-project/
    - Sample reference populations: https://www.genome.gov/10001688/international-hapmap-project/
- Clustering and defining the genetic commonality within each population genetically
- Projecting this definition onto a new DNA sample in order to get its admixture report (ancestry)

Technical Intuition - Hack

Technical Admixture involves

- Start with a sample of DNA and a “goal” “fake race” (ex. Let’s make Anne Kim Mexican)
- Having labeled reference populations (french, chinese, korean, indian, etc)
    - Run (pca/clustering - will need to think about Math) against sample and goal_fake_race in order to identify genes expression that is most_characteristic or common (ex. Anne’s boosted Mexican signal). Select samples in the Mexican reference population that are most characteristic
    - For each reference population that is not goal_fake_race, select samples that boost the negative of the most_characteristic gene expressions
- ~~Clustering and defining the genetic commonality within each population genetically~~
- Use contrived population samples to run the admixture
- Projecting this definition onto a new DNA sample in order to get its admixture report (ancestry)
- Result should be a conclusion that is close to the goal fake race

Resources

Datasets

https://www.genome.gov/27528684/1000-genomes-project/

https://www.genome.gov/10001688/international-hapmap-project/

UK Biobank?

Code

https://github.com/snehitp/cehg16-workshop - very simple and easy to follow

https://github.com/pcgoddard/Burchardlab_Tutorials/wiki/ADMIXTURE

https://github.com/jacahill/Admixture

http://www.y-str.org/2014/04/bam-analysis-kit.html

Lit

https://genome.cshlp.org/content/19/9/1655.short

Build

- Before DEF CON
  - We probably need to think about the math a little more and double check if this actually works
  - Make a repository for this code
  - Possibly host a script on a website so people can do fun things with their 23andMe data
  - Practice present at DEF CON 617 (which is the Cambridge chapter)
- At DEF CON
  - Ask people to clone the repo before?
  - Have the code on a thumbdrive?
  - Having a website would be the easiest
  -
- Prove that race is just dumb, guys

History of race

What is race

Underdefined + Bad science

- Bring a sequencer ???

TODOs

- Actual coding
- Reaching out to Race and Bio experts once a prototype is done
- 45 minutes speaking, 45 minutes working
- Workshop
  - Public repository that people can clone - locally run analysis and tweak parameters in the code themselves
- Post-workshop collaborative discussion on race
- Stretch goals: put this on a website for public access, visual and easy-to-access demonstration on how genetic profiling is done

Racial scholars

Once prototype / initial analysis is ready, reach out to racial scholars to

Ibrahm Kendi -