NEUROHACK 2022

Challenge 1: Team B London
(Genome-Wide Association Lovers)

Modelling the Genetics of Dementia & its Risk Factors in South Asian Ancestry

TEAM MEMBERS

Isy Foote (London)

David Enoma (Nigeria)

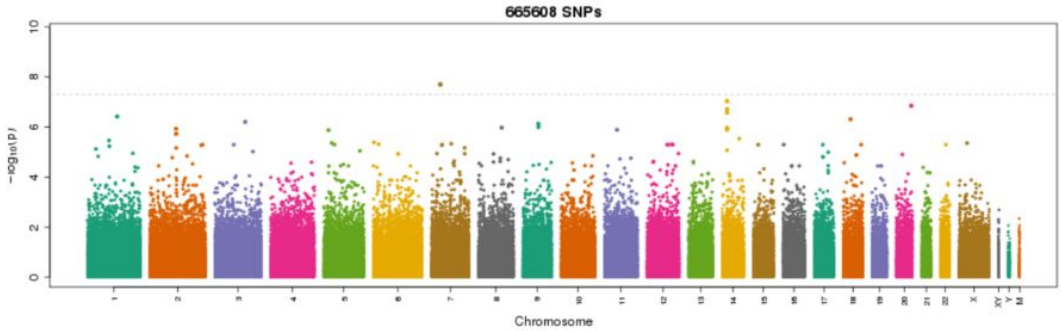
Mateus Harrington (Cardiff)

Gabriela Paulus (NYC)

Anna Furtjes (London)

First Project aim

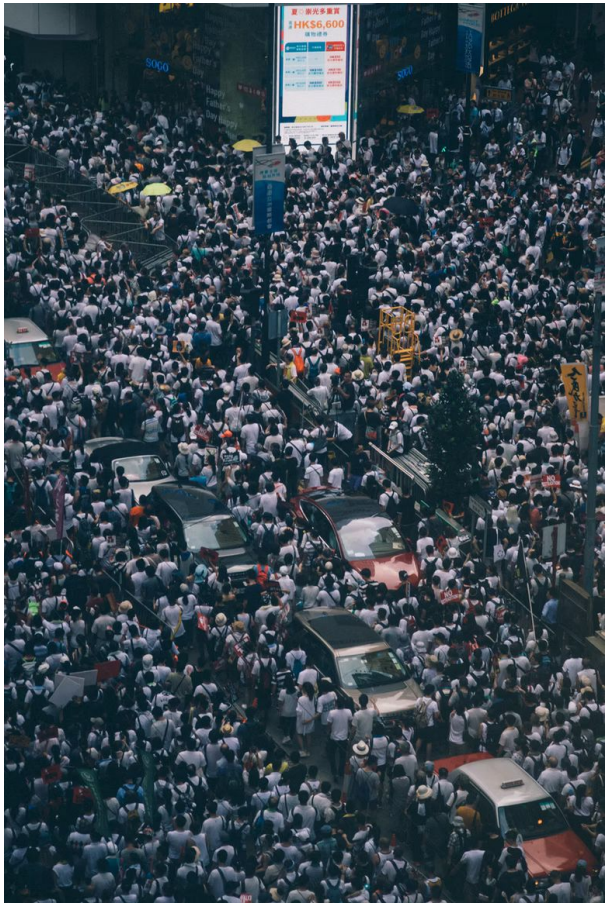
To conduct a dementia GWAS in the LASI-DAD using a continuous phenotype & compare how it performs against the binary GWAS



REF: LASI DAD NCD GWAS, QC report

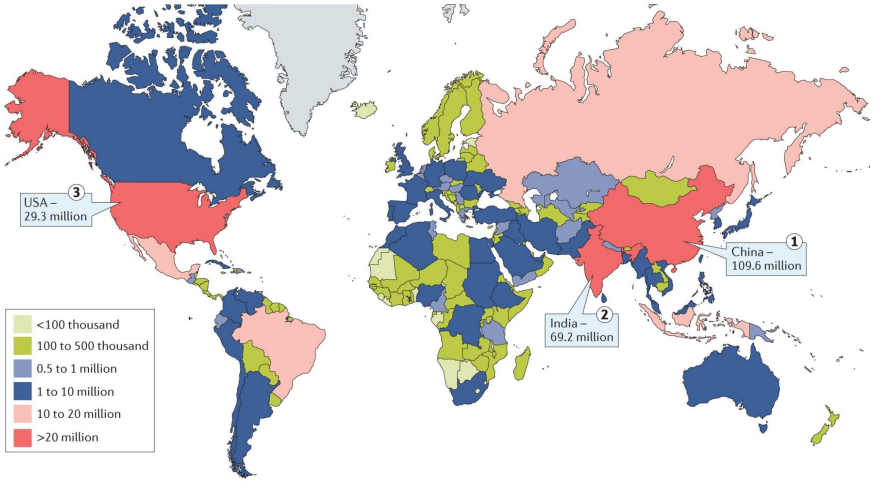
Second project aim

To use tools that consider heterogeneous ancestry & complex population stratification more reliably than traditional GWAS software.



Third project aim

To conduct GWAS of dementia risk factors that have particularly high prevalence in South Asians to explore whether there are differential shared pathways with dementia in South Asians versus European populations.



Type 2 diabetes global prevalences:
Zheng et al 2017

Fourth project aim

To restructure a seamless pipeline in NextFlow both for ease of use and to improve reproducibility & portability into cloud infrastructures.

Considerations

1. Phenotypic data cleaning
2. Covariate data cleaning
3. Genetic data cleaning
4. Regenie setup
5. Regenie pipeline
    a. Step 1
    b. Step 2
6. Pipeline integration in NextFlow
7. Future Directions

Phenotypic data cleaning

- LASI-DAD data: N = 932 unrelated individuals
- South Asian Ancestry
- Neurocognitive Disorder - none, mild, severe (0,1,2)
- Risk factors (Type 2 diabetes, high blood pressure, chronic heart disease & stroke) - all binary phenotypes
- We made 2 phenotype files (1 for continuous, 1 for binary) to be able to run separate GWASs in Regenie.

LIMITATION:

We wanted to integrate family history to stratify the sample by genetic risk as well but there wasn't enough overlapping data (<20 participants in the full data)

Covariate data cleaning

- Education (50% illiterate)
- Sex
- Age
- Genetic PCs

LIMITATION:

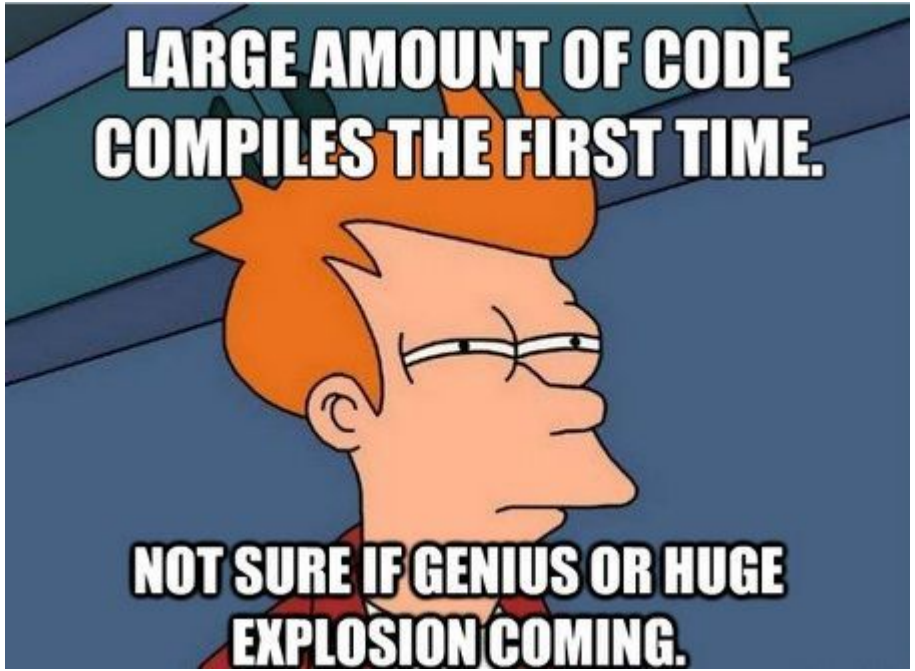
Genetic batch was unavailable in the dataset provided.

Genetic Quality control

- QC already done in LASI DAD that we want to match
  - MAF, missingness filters
  - Autosomal chromosomes
  - Sex check
  - Unrelated individuals
  - Imputed SNPs INFO > 0.4

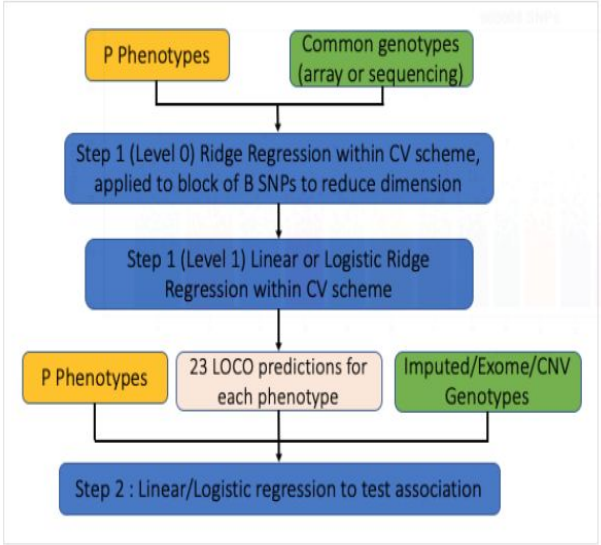
Regenie set-up

- We were unable to install Regenie via conda in the JupyterLab environment due to dependency version conflicts
- Similarly, the Docker image also wouldn’t run correctly
- Ultimately we wrote a script to compile BGEN (a dependency) and Regenie from source in the environment



Why do we use REGENIE?

- LASI-DAD is a genetically heterogeneous sample
- Relatively small sample
- Can induce false-positive genetic signal
- Account for population stratification through mixed linear model



Mbatchou et al., 2021 *Nature Genetics*

Regenie

Step 1:

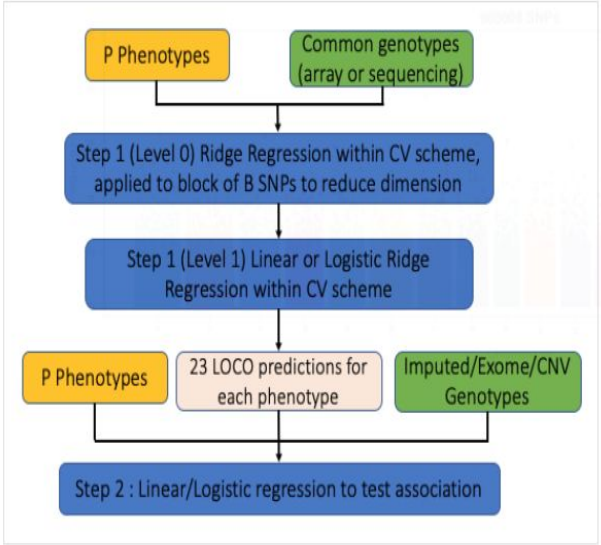
- Blocks of directly genotyped SNPs are used to fit a cross-validated whole-genome regression model

Step 2:

- Association between the phenotype and genetic variants

LIMITATION:

Regenie cannot create the mixed model in a sample of 10 (fake data) so we could not ‘check’ our pipeline.



Mbatchou et al., 2021 *Nature Genetics*

Future Directions

- Explore genetic overlap between our NCD GWAS and risk factor GWAS
  - Calculate linkage disequilibrium scores and weights in LASI-DAD
  - Model shared genetics between NCD and risk factors using Genomic SEM and genomic PCA
  - Aim: identify major dimensions of variance in NCD that could be explained by shared biological pathways with risk factors
- Explore biological function of genome-wide significant SNPs
- Test for hits across ancestries
- Compare heritability estimates and genetic correlations between ancestries (Genomic Complex Trait Analysis)

Thanks for your attention

- Phenotypic Data cleaning
- Covariate Data cleaning
- Genetic Data cleaning
- SNP list creation
- Regenie software setup
- NCD GWAS
- Risk Factor GWASs
- Expected Output
- Future directions

Hackathon DEMON

Code

Genome-Wide Association Lovers

14/01/2022

Here we present code to calculate Genome-Wide-Association Study (GWAS) Summary Statistics for Neurocognitive Disorder (NCD). Please see our Git repository for more information here.

We suggest a novel way of defining NCD as a continuous phenotype, and we employ a mixed linear model to account for pronounced population stratification in the South Asian sample, which is more diverse than European samples, for example. We decided to also run simultaneous binary GWASs for modifiable dementia risk factors that have a higher prevalence in South Asians compared to Europeans (e.g., type 2 diabetes, hypertension, chronic heart disease, stroke).