P17.015.D Impact of control selection strategies on GWAS results: a study of Schizophrenia and Bipolar Disorder in the UK Biobank

Jingzhan Lu[1][2], Johan Thygesen[2], Andrew McQuillin[3], Harry Green[1]

Affiliations: [1]Department of Clinical and Biomedical Sciences, University of Exeter, Exeter, UK. [2]Institute of Health Informatics, University College London, London, UK. [3]Molecular Psychiatry Laboratory, Division of Psychiatry, University College London, London, UK.

Background

As GWAS studies move from array-based genotyping to whole exome (WES) and whole genome (WGS) sequencing, there is a significant increase in cost. For disease phenotypes, case-control selection may be a useful, but it is a currently underexplored technique for minimising resource intensity while maintaining statistical power. This is particularly relevant to studies of disease in a general population like UK Biobank (N=500,000), where there is often a surplus of available healthy controls. This investigation aims to explore different strategies for control selection in GWAS studies.

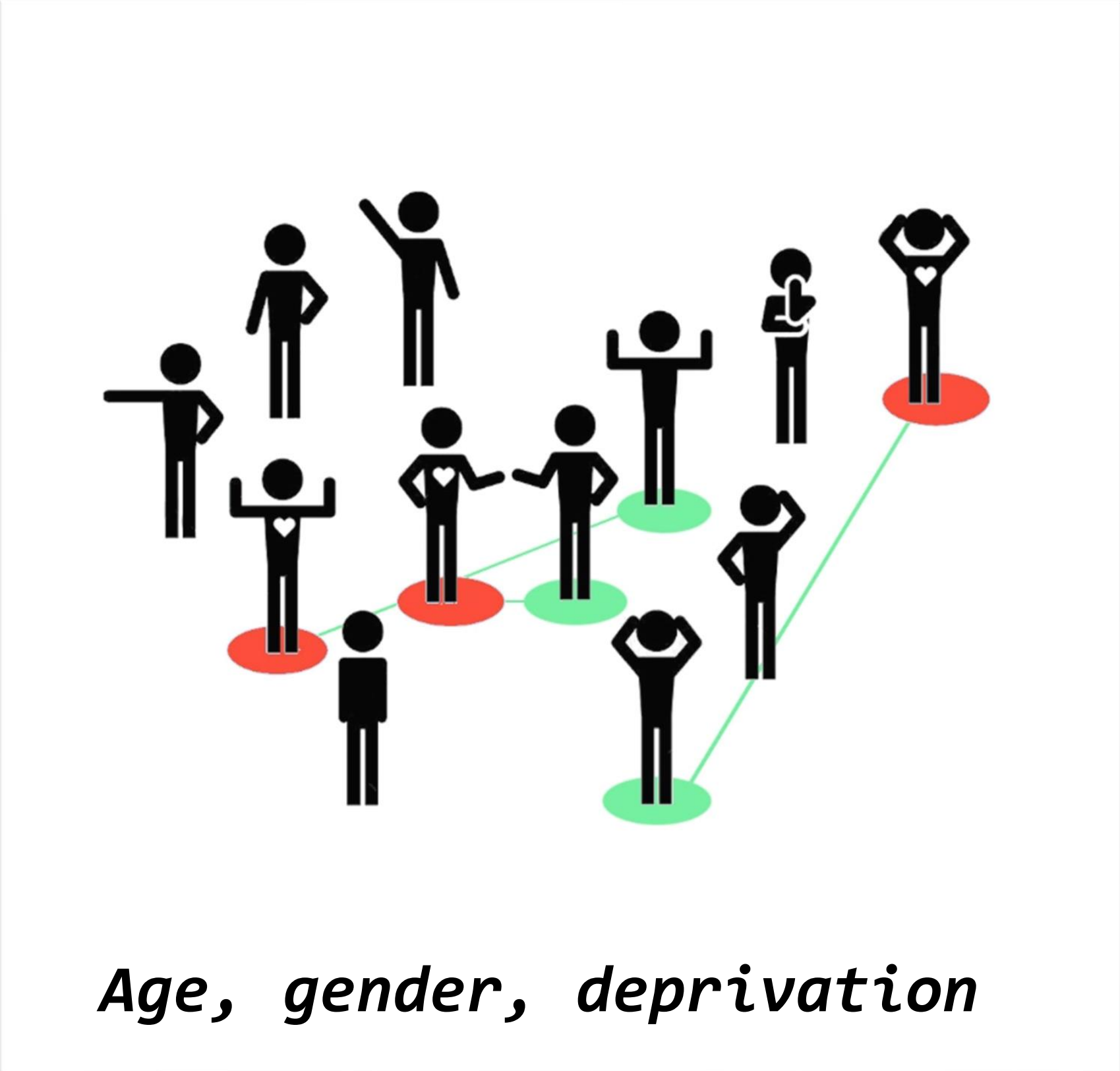
Methods

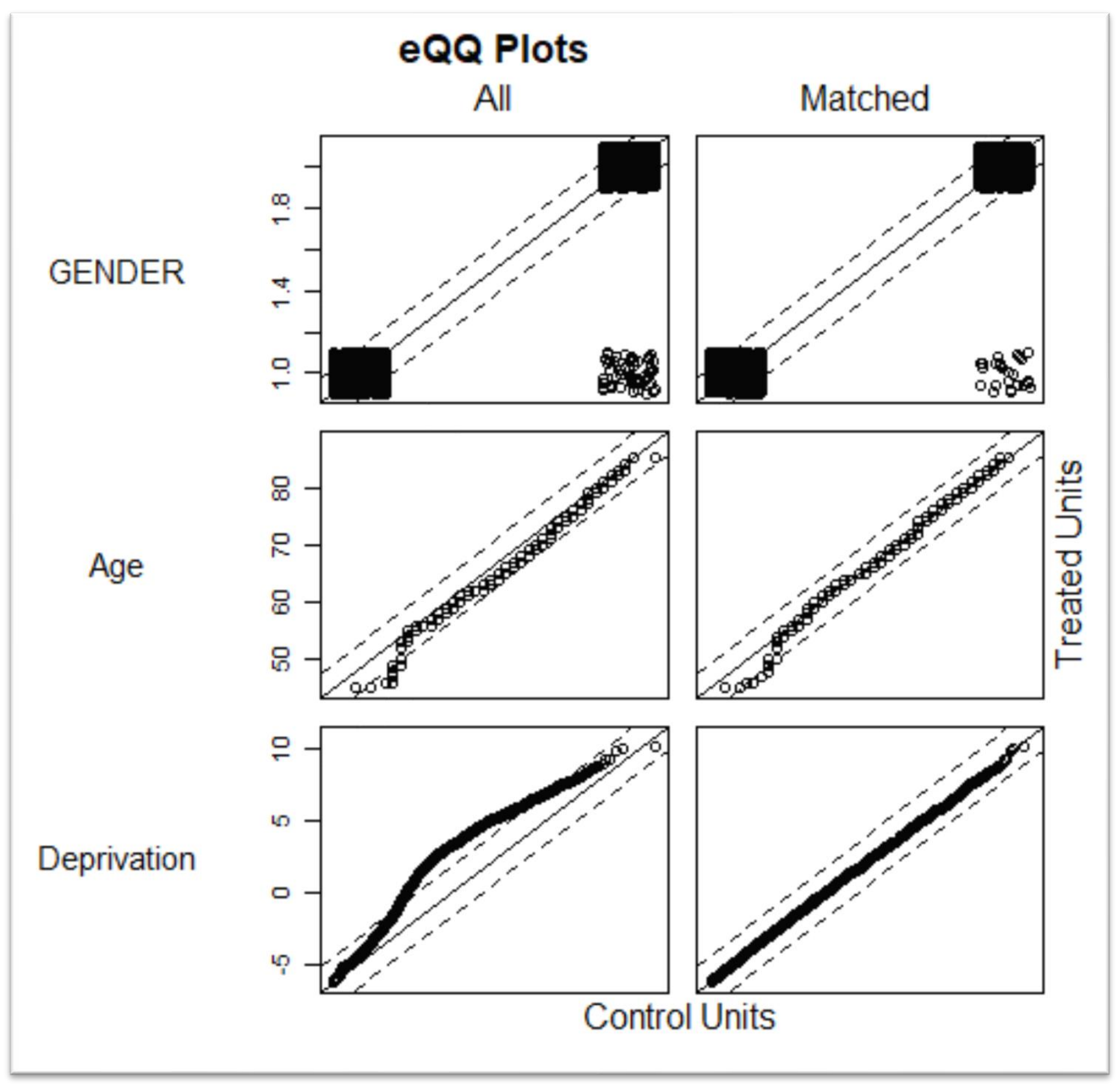
The study included 1,988 Schizophrenia (SCZ) and Bipolar Disorder (BD) cases of European ancestry from the UK Biobank (UKB), with a 1:4 case-to-control ratio. Individual GWAS were conducted under each strategy using REGENIE. We assessed the impact of three control selection strategies on GWAS combining SCZ and BD with covariates: age, gender and deprivation. These $S_{1\text{-}3}$ were:

S1: Random selection without covariates
S2: Case/control matched analysis with covariates
S3: (Another) Random selection with covariates

Results were compared with the Psychiatric Genomics Consortium (PGC 2022), the largest GWAS meta-analysis, which was considered the gold standard.

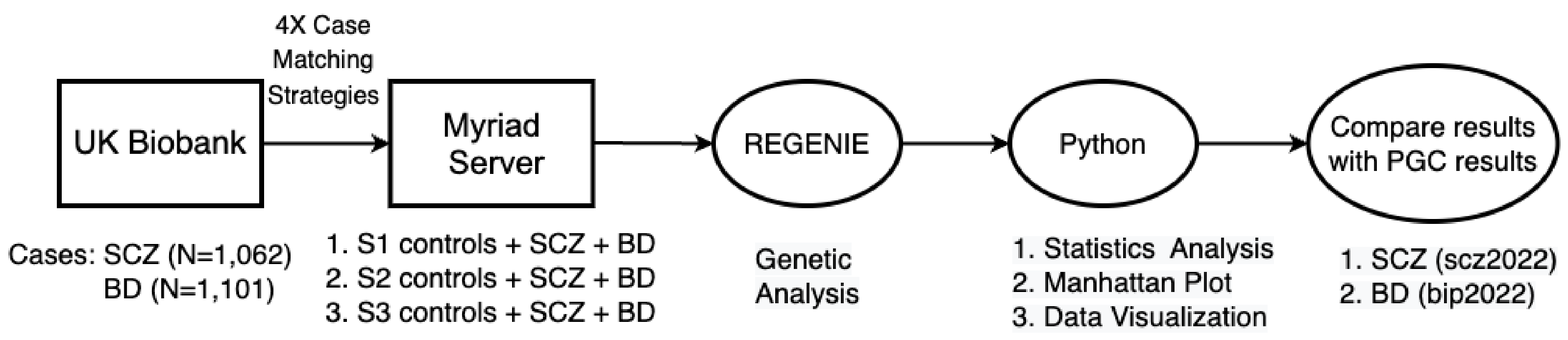
Matching


Age, gender, deprivation



A total of 1,988 SCZ and BD cases combined in the GWAS. The $S_{1\text{-}3}$ selected 4X controls from the 405,949 available cohort of UKB.

Flowchart


UK Biobank → 4X Case Matching Strategies → Myriad Server → REGENIE → Python → Compare results with PGC results

Cases: SCZ (N=1,062)
BD (N=1,101)

1. S1 controls + SCZ + BD
2. S2 controls + SCZ + BD
3. S3 controls + SCZ + BD

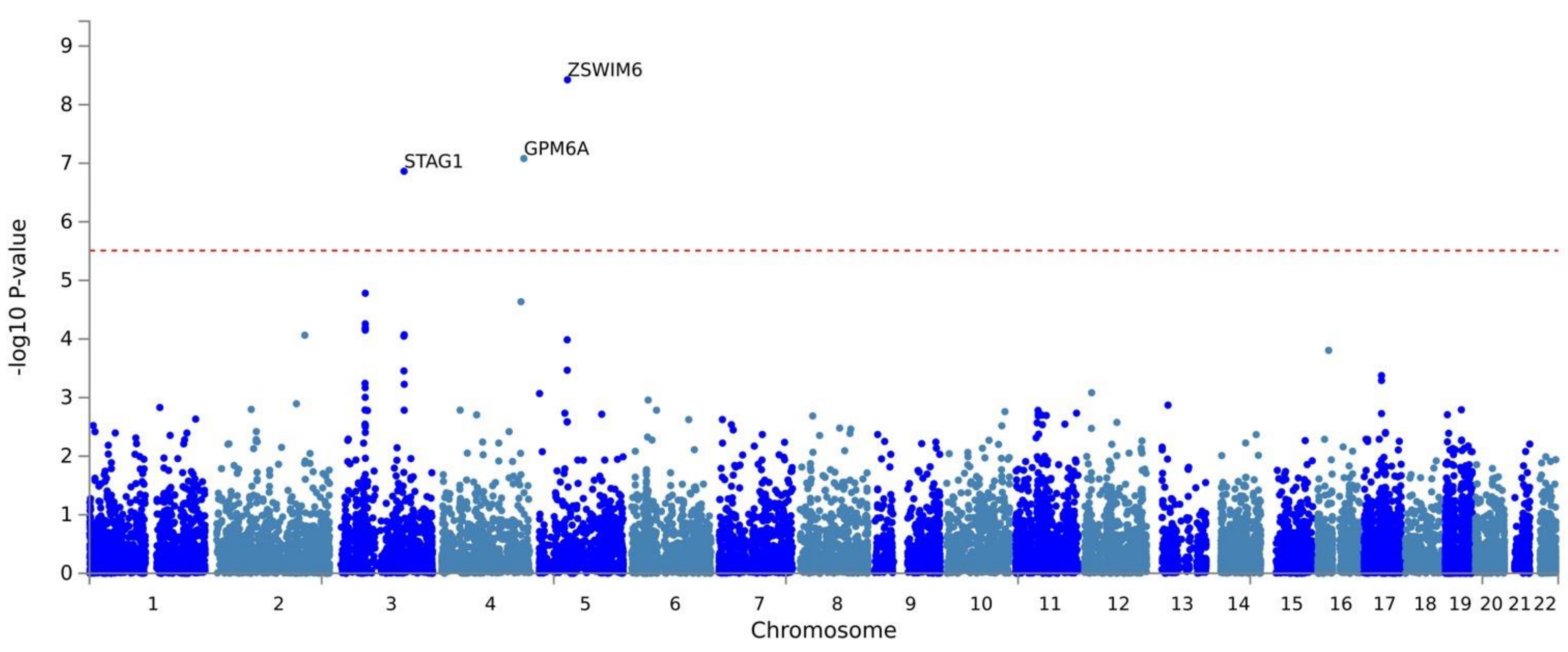
Genetic Analysis

1. Statistics Analysis
2. Manhattan Plot
3. Data Visualization

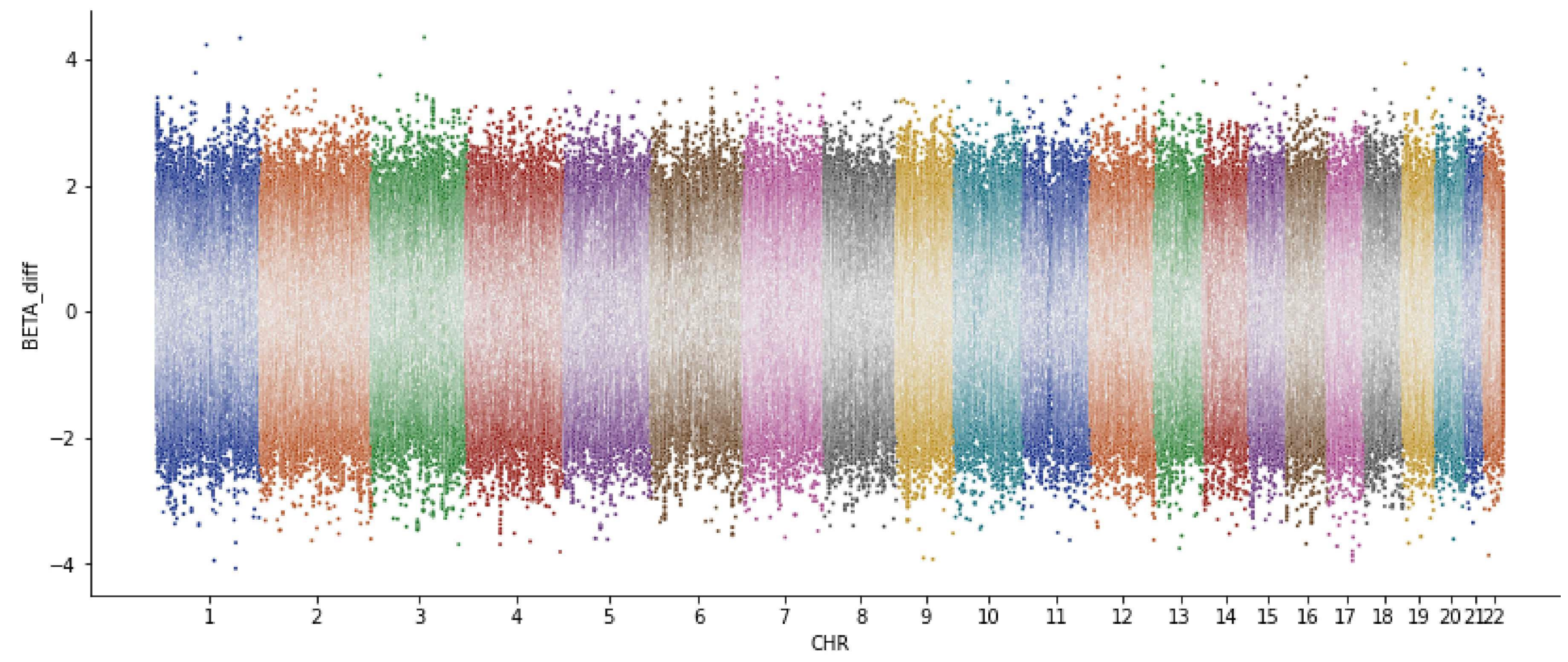
1. SCZ (scz2022)
2. BD (bip2022)

Manhattan-Comparison Plot Development

Example of Manhattan plot (one GWAS result)



S2 vs S3 Manhattan-Comparison plot (two GWAS results)



Results & Discussion

Equation: $\text{p\_diff} = \frac{|(S_{123})_p - PGC_p|}{SNP_{\text{num}}}$, num is the overlapping SNPs

PGC	Compare Items	Strategy 1	Strategy 2	Strategy 3
PGC SCZ				
	p_diff	-0.03	0.0002	-0.021
p mean = 0.426	*p_abs_diff*	0.382	**0.384**	0.375
β mean = -0.002	*β_diff*	0.002	**-0.001**	-0.010
	β_abs_diff	0.189	0.419	0.426
PGC BD				
	p_diff	-0.037	-0.003	-0.027
p mean = 0.441	*p_abs_diff*	0.383	**0.387**	0.380
β mean = -0.0016	*β_diff*	**-0.001**	**-0.004**	-0.014
	β_abs_diff	0.190	0.430	0.434

Overall, S2 (matching) showed the best performance for GWAS, matching the PGC results the closest. This suggests that the matching process makes the controls and cases more similar at the level of covariates.

Future Plans

This preliminary work shows case-control matching performs better than adjusting the model for the same covariates. Future work will compare case-control matching with using all controls, test in other disease phenotypes, and measure the impact of matching when WGS techniques are used.

@JingzhanLu

This research was made possible through access to the data and findings generated by the UK Biobank Project; https://www.ukbiobank.ac.uk/

Interested in this project and want to hear more / collaborate or contribute please write to: jl1426@exeter.ac.uk