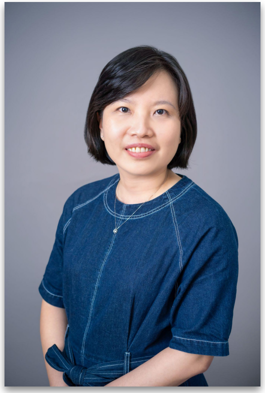
Section 3

Performing GWAS using imputed data



Xueling Sim
National University of Singapore
ephsx@nus.edu.sg

MICHIGAN IMPUTATIONSERVER

ASHG American Society of Human Genetics

Learning objectives

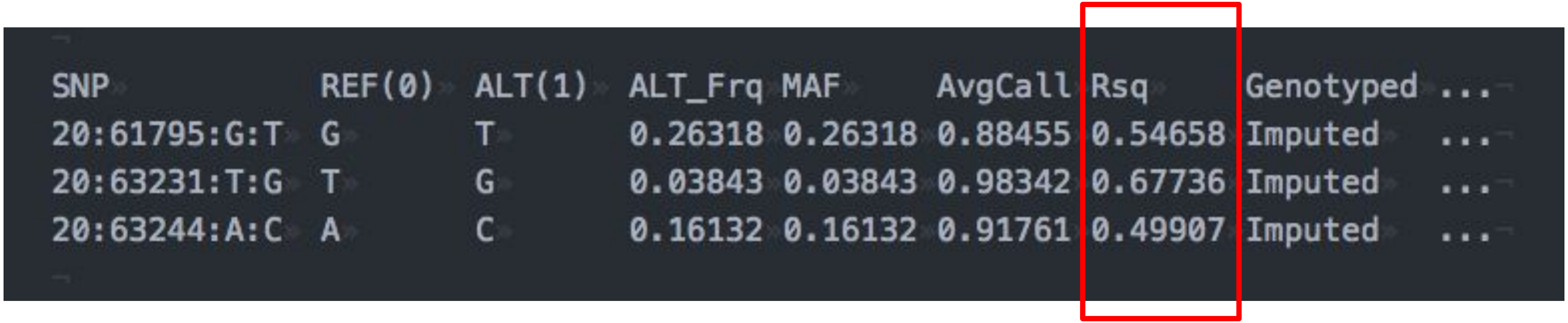
Participants will learn to:

- Identify and understand the use of variant imputation quality information following imputation in the MIS
- Distinguish between some of the available options for GWAS
- Troubleshoot common GWAS errors

Have you ever performed a GWAS?

Imputation Quality

- For each variant, how confident can we be that the imputation dosages are sufficiently “accurate” for association analyses?
- Measure of confidence in imputed dosages: “Rsq” column [range 0-1]



```
SNP            REF(0)  ALT(1)  ALT_Frq MAF     AvgCall Rsq     Genotyped ...
20:61795:G:T   G       T       0.26318 0.26318 0.88455 0.54658 Imputed   ...
20:63231:T:G   T       G       0.03843 0.03843 0.98342 0.67736 Imputed   ...
20:63244:A:C   A       C       0.16132 0.16132 0.91761 0.49907 Imputed   ...
```

From a chr20.info.gz file

Imputation Quality

- Minimal Rsq value for common variants
    - ≥ 0.30
- Minimal Rsq value for low frequency/rare variants
    - ≥0.50
- Before performing GWAS, remove variants that do not meet these thresholds
    - Suggested program: VCFtools
    - Saves computational time when performing GWAS

Which GWAS program(s) have you used?

Available GWAS Programs

No File Reformatting (VCF from MIS)

- EPACTS
- Rvtests
- SNPTEST
- SAIGE

File Formatting Required

- BOLT-LMM
- BGENIE
- regenie
- PLINK

Each GWAS Program Has Strengths, Limitations

EPACTS/Rvtests

- + Many model options - single variant, gene-based
- + Chr X analyses
- + Phenotypic transformation (e.g inverse normal; Rvtests only)
- + Linear mixed model for sample relatedness (quantitative traits only)
- + Generate covariance matrices for downstream analyses (e.g conditional analyses; Rvtests only)

- - Memory intensive
- - Sample size ~≤20,000 (better ≤10,000)

EPACTS: https://genome.sph.umich.edu/wiki/EPACTS
Rvtests: https://genome.sph.umich.edu/wiki/Rvtests

Each GWAS Program Has Strengths, Limitations

SNPTEST

+ Frequentist and bayesian methods supported

+ Chr X analyses

- Limited to unrelated individuals

- Computational intensive

SNPTEST: https://www.well.ox.ac.uk/~gav/snptest/#introduction

Each GWAS Program Has Strengths, Limitations

SAIGE

+ Similar to Rvtests, but for very large sample sizes (e.g. biobanks)
+ Able to account for sample relatedness for binary traits
+ Designed to handle unbalanced number of cases and controls
+ Chr X analyses

- Should not be used to examine heritability (biased variance estimates)
- Computational time can vary widely between phenotypes and sample sizes
- Can be conservative for extremely unbalanced case and control ratio
- Odds ratios estimated to conserve computational time

SAIGE: https://github.com/weizhouUMICH/SAIGE

Each GWAS Program Has Strengths, Limitations

BOLT-LMM/BGENIE/regenie

+ Great for very large sample sizes (e.g. biobanks)
+ Chr X analyses
+ Computationally efficient (regenie)

- Requires files to be in BGEN or PLINK format
- Nextflow pipeline for regenie using VCF: https://github.com/genepi/nf-gwas
- Not optimal for extremely unbalanced case control ratio (especially with rare variants)

BOLT-LMM: https://data.broadinstitute.org/alkesgroup/BOLT-LMM/#x1-5600011

BGENIE: https://jmarchini.org/bgenie/

Regenie: https://github.com/rgcgithub/regenie

Each GWAS Program Has Strengths, Limitations

PLINK

+ Quick
+ Multiple versions; often as intermediary tool to the other programs
+ Can run on the command line (unix not required)
+ Chr X analyses

- Requires files to be in PLINK format (.bed/.bim/.fam)
- Limited model options

PLINK: https://www.cog-genomics.org/plink/2.0/

Summary of common GWAS analysis tools

	EPACTS	Rvtests	SNPTEST	SAIGE	BLOT-LMM	Bgenie	regenie
Input VCF	Y	Y	Y				
Sample relatedness (Quantitative outcome)	Y	Y		Y	Y	Y	Y
Sample relatedness (Binary outcome)				Y		Y	Y
Case control imbalance				Y			Y
Large sample size (>20,000)				Y	Y	Y	Y

Performing the GWAS

- Each program has its own input, output formats, and options
- Typical input files
  - Genotype file (.vcf; .bgen; .bed/.bim/.fam)
  - Phenotype/covariate file (.txt; .ped)
  - Some programs use separate phenotype and covariate files
  - Kinship/relationship matrix (EPACTS, SAIGE)

Which program(s) would be the best?

- A researcher new to genetic analyses and unfamiliar to the UNIX environment wants to perform a GWAS on total cholesterol using a cohort of 5,500 unrelated individuals

Which program(s) would be the best?

- A researcher new to genetic analyses and unfamiliar to the UNIX environment wants to perform a GWAS on total cholesterol using a cohort of 5,500 unrelated individuals
    - EPACTS/Rvtests

Which program(s) would be the best?

- A researcher new to genetic analyses and unfamiliar to the UNIX environment wants to perform a GWAS on total cholesterol using a cohort of 5,500 unrelated individuals
    - EPACTS/Rvtests
- Researchers want to perform a GWAS using a cohort of 10,000 individuals with household based recruitment (i.e. includes related individuals)

Which program(s) would be the best?

- A researcher new to genetic analyses and unfamiliar to the UNIX environment wants to perform a GWAS on total cholesterol using a cohort of 5,500 unrelated individuals
  - EPACTS/Rvtests
- Researchers want to perform a GWAS using a cohort of 10,000 individuals with household based recruitment (i.e. includes related individuals)
  - EPACTS/Rvtests/BLOT-LMM or SAIGE

Which program(s) would be the best?

- A researcher new to genetic analyses and unfamiliar to the UNIX environment wants to perform a GWAS on total cholesterol using a cohort of 5,500 unrelated individuals
  - EPACTS/Rvtests
- Researchers want to perform a GWAS using a cohort of 10,000 individuals with household based recruitment (i.e. includes related individuals)
  - EPACTS/Rvtests/BLOT-LMM or SAIGE
- Researchers want to perform a GWAS using data from BioBank Japan (>200,000 individuals)

Which program(s) would be the best?

- A researcher new to genetic analyses and unfamiliar to the UNIX environment wants to perform a GWAS on total cholesterol using a cohort of 5,500 unrelated individuals
    - EPACTS/Rvtests
- Researchers want to perform a GWAS using a cohort of 10,000 individuals with household based recruitment (i.e. includes related individuals)
    - EPACTS/Rvtests/BLOT-LMM or SAIGE
- Researchers want to perform a GWAS using data from BioBank Japan (>200,000 individuals)
    - BLOT-LMM or SAIGE or begenie or regenie

Common Errors When Running a GWAS

- Wording of error messages vary by program, but the same issues will cause errors throughout all of the program
- [Unix] Errors independent of GWAS program
  - File permissions
    - Correct by changing file permissions
  - Directory/file not found
    - Correct by making sure all of the file locations and names are accurate
  - Not enough memory/time
    - Correct by restarting job with adequate memory/time allocation

Common Errors When Running a GWAS

- Common errors
    - IDs don’t match
        - Correct by ensuring that the ID in the phenotype, genotype, covariance, kinship matrix are consistent format in all files

Common Errors When Running a GWAS

- Common errors
  - IDs don’t match
    - Correct by ensuring that the ID in the phenotype, genotype, covariance, kinship matrix are consistent format in all files
  - File format(s) incorrect
    - Correct by making sure the format of all files are as the program is expecting (e.g. columns, delimiters, headers, file extension

Common Errors When Running a GWAS

- Common errors
  - IDs don’t match
    - Correct by ensuring that the ID in the phenotype, genotype, covariance, kinship matrix are consistent format in all files
  - File format(s) incorrect
    - Correct by making sure the format of all files are as the program is expecting (e.g. columns, delimiters, headers, file extension
  - Improperly specified options/command
    - Correct by checking all needed options are specified, correct order, no typos

Common Errors When Running a GWAS

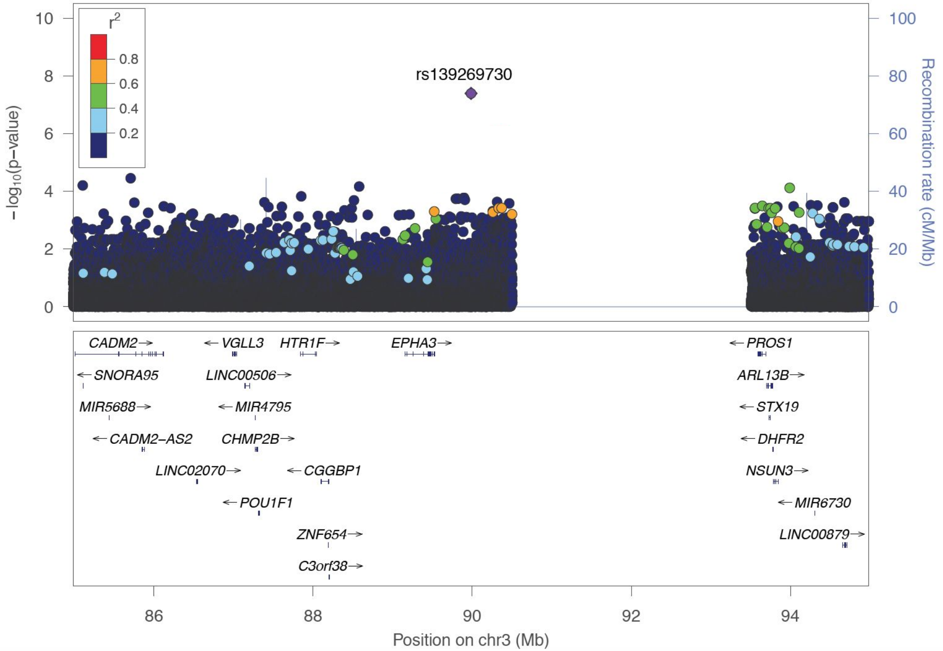
- Common errors
    - IDs don’t match
        - Correct by ensuring that the ID in the phenotype, genotype, covariance, kinship matrix are consistent format in all files
    - File format(s) incorrect
        - Correct by making sure the format of all files are as the program is expecting (e.g. columns, delimiters, headers, file extension
    - Improperly specified options/command
        - Correct by checking all needed options are specified, correct order, no typos
    - Peripheral programs not available (e.g. R with EPACTS, SAIGE)
        - Correct by installing other peripheral programs

Common Errors When Running a GWAS

- Common errors
    - IDs don’t match
        - Correct by ensuring that the ID in the phenotype, genotype, covariance, kinship matrix are consistent format in all files
    - File format(s) incorrect
        - Correct by making sure the format of all files are as the program is expecting (e.g. columns, delimiters, headers, file extension
    - Improperly specified options/command
        - Correct by checking all needed options are specified, correct order, no typos
    - Peripheral programs not available (e.g. R with EPACTS, SAIGE)
        - Correct by installing other peripheral programs
    - Invalid estimate (e.g. heritability in BOLT-LMM)
        - Sample too related and/or sample size too small
        - Correct by using a different program

Interpreting GWAS Results

- GWAS results must be carefully reviewed for:
  - Imputation quality!
  - Genomic inflation
  - False positives



- Replication datasets
- PheWas

American Society of Human Genetics

Summary

- Variants must be filtered post-imputation to remove those with poor imputation quality
- There are many GWAS programs available, each with their own strengths and limitations - so be sure to pick one that fits your analyses needs
- As these GWAS programs are widely used or adopted by consortia, there are tutorials and help-pages available

More info and FAQ can be found here:
https://imputationserver.readthedocs.io