Disclosure Slide

Financial Disclosure for:

Christian Fuchsberger
Sebastian Schönherr
Lukas Forer
Cassie Spracklen
Albert Smith
Sarah Hanks

We have nothing to disclose

Section 4

Performing GWAS using imputed data

Cassie Spracklen
University of Massachusetts-Amherst
cspracklen@umass.edu

MICHIGAN IMPUTATIONSERVER

ASHG
American Society of Human Genetics

Learning objectives

Participants will be able to:

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

Imputation Quality

- How confident can we be that the imputation is accurate for a particular variant?
- “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

Available GWAS Programs

No File Reformatting (VCF from MIS)

- EPACTS
- Rvtests
- SAIGE
- SNPTEST

File Formatting Required

- PLINK
- BOLT-LMM
- BGENIE
- regenie

Each GWAS Program Has Strengths, Limitations

EPACTS/Rvtests

+ Many model options
+ Chr X analyses
+ Automatically programs for parallel processing (EPACTS only)
+ Can transform your phenotype file (e.g inverse normal; Rvtests only)
+ Produce covariance matrices for downstream 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

SAIGE

+ Similar to Rvtests, but for very large sample sizes (e.g. biobanks)
+ Designed to handle unbalanced number of cases and controls

Chr X analyses unknown

- Should not be used to examine heritability (biased variance estimates)
- Computational time can vary widely between phenotypes and sample sizes
- Can be conservative when 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

PLINK

+ Quick
+ 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/plink2/

ASHG
American Society of Human Genetics

Each GWAS Program Has Strengths, Limitations

BOLT-LMM/BGENIE/regenie

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

- Requires files to be in BGEN or PLINK format
- Linear mixed models only (quantitative traits); need to convert to log OR for binary traits (BOLT-LMM)
- Not optimal for extremely unbalanced case control ratio (especially with rare variants) (BOLT-LMM)

- BOLT-LMM: https://data.broadinstitute.org/alkesgroup/BOLT-LMM/#x1-5600011
- BGENIE: https://jmarchini.org/bgenie/ Regenie: https://rgcgithub.github.io/regenie/

Performing the GWAS

- Each program has its own code, options

- Typical input files (format varies by program)
  - 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)

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

- Straight-forward errors
  - 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

- Additional 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

- Additional 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

- Additional 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

- Additional common errors
  - Not loaded other programs (e.g. R with EPACTS, SAIGE)
    - Correct by loading other needed programs

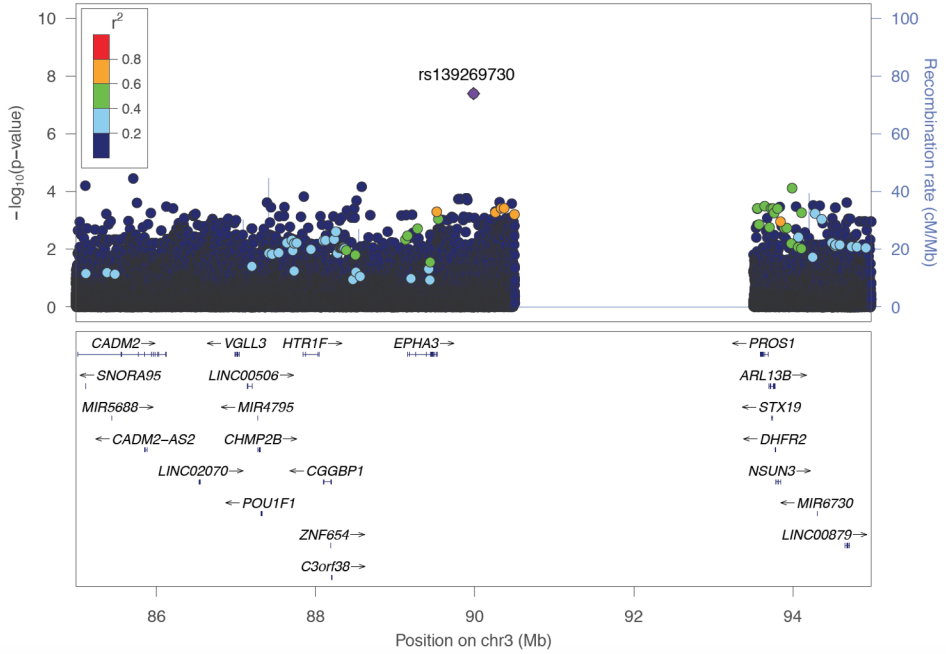
Common Errors When Running a GWAS

- Additional common errors
  - Not loaded other programs (e.g. R with EPACTS, SAIGE)
    - Correct by loading other needed programs
  - Invalid heritability estimate (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



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 situation.

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