Rare variant analysis
Outline

- Genotype vs phenotype
- What are common variants and what are rare variants
- Limitations of GWAS
- Rare variant analysis
- Challenges with rare variant analysis
- Analysis plan
- Conclusion
- Homework
Genotype vs phenotype

What is genotype?
- Heritable genetic identity that can be revealed by genome sequencing

What is phenotype?
- A description of your actual physical characteristics
Common variants and rare variants

- **Common variants**
  - Minor allele frequency (MAF) \( \geq 0.05 \)
  - Small effect
  - Often in a strong LD (linkage disequilibrium) with adjacent SNPs

- **Rare variants**
  - MAF \( \leq 0.05 \) (0.01)
  - May have a big effect
  - Weak LD with adjacent SNPs
  - New mutations; likely to have arisen from founder effects
Limitations of GWAS

- GWAS – an examination of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait
- GWA study focuses on common variants whose effects are usually small
- GWA studies are mostly genotype-array based
- Missing heritability → SNPs that are found with GWASs only explain a small part of disease heritability
- Not all GWAS are not replicated across studies or population
Common disease-rare variant hypothesis (CDRVH)

- CDRVH hypothesizes that if a genetic disease is common in the population, then the genetic causes may not be in the common variants
- CDRVH says that these genetic causes may be comprised of a multiplicity of risk alleles
Why rare variant analysis? Limitations

- Uses rare variants with $\text{MAF} \leq 0.05$
- Why rare variants?
  - Most human variants are rare
  - Functional variants are mostly rare
  - Rare variants are expected to have larger effects on complex traits than common variants.
- Problem is that rare variant analysis are underpowered when single SNPs are used → large sample size is needed
Different types of data for analyzing rare variants

- Based on the accumulation of minor alleles at rare variants the best and most powerful approach for assaying rare variation would be re-sequencing.
- Second approach would be using genome-wide genotyping array data (with imputation) which is cheaper way than re-sequencing.
- Another possibility is to use exome genotyping array data cost effective and that can be used for genotyping low-frequency and rare coding variants.
Rare variants and tests

- Statistical methods focus on the accumulation of minor alleles at rare variants within the same “genomic unit”. Two main classes of tests:
  - **BURDEN**: assume same direction of effect of all variants in the genomic unit.
  - **DISPERSION**: allow different directions of effect for variants in the genomic unit.
What is a genomic unit?

- Genomic unit is a unit that focuses on rare coding variants
- Genomic units can be:
  - Multiple genes in a pathway
  - All variants in a gene
  - All coding variants in a gene
  - All variants in the same exon
Burden vs Dispersion – which one to use?

- **Burden test:**
  - Number of minor alleles is higher in cases than in controls
  - Power is the highest when all variants within a gene have the same effect on the disease

- **Dispersion test:**
  - Variance in number of minor alleles is higher in cases than in controls
  - Power is the highest when variants within a gene have different effect on the disease

- **SKAT-O** combines features of burden and dispersion tests and has been demonstrated to be powerful over a range of genetic models
Analysis plan for rare variant analysis in a consortium
Exome chip analysis plan

- **Phenotype definition**
  a) Cases
  b) Controls

- **Exome chip analyses**
  - Files required
    a) Phenotype file
    b) BED file
    c) VCF with genotype data
    d) List of regions and variants to be excluded
    e) Group files for gene-wise testing
Exome chip analyses

Tools:

a) PLINK
b) PSEQ
c) CheckVCF
d) EPACTS
e) ANNOVAR
f) Perl
Preparing the exome chip data for analysis I

- Genotype calling (with Illumina GenomeStudio GENCALL and recalled using zCALL)
- Sample QC
  - Call rate (98% threshold)
  - Heterozygosity
  - Distribution of singletons per sample
  - Gender discordance
  - Outliers from the principle components analysis
Preparing the exome chip data for analysis II

- Variant QC
  - Call rate (95% threshold)
  - Hardy-Weinberg disequilibrium
  - Cluster separation score (cut-off <0.4)
  - GenTrain score (cut-off <0.6)
- Create VCF file using pseq
Cluster plot
Quality control

- Calculation of IBS (Identity By State)
  - Exclude any sample that is related to more than one other samples
  - For case-control related pairs, exclude the control
  - For case-case and control-control pairs, exclude the sample with lowest call rate

- Calculation of principle components (using PLINK)
  - To evaluate the extent of population stratification within cohort(s)
Association analyses – single variant analyses

- In association analyses can be done as a case-control single variant analyses but also as a gene-wise analyses.
- Analyses can also be adjusted for different covariates (age, gender, BMI etc).
- EPACTS software can be used (different implemented tests can be used such as Logistic Score test, Logistic Wald test).

Example:

```
${EPACTS_DIR}/epacts single \
  --vcf [input.vcf.gz] --ped [input.ped] --kin [outputprefix.kinf] \
  --sepchr (if VCF is separated by chromosome) --pheno [PCOS] \
  --cov [COV1] --cov [COV2] --test emmax \
  --out [outputprefix] --run [# of parallel jobs]
```
List on analyses that should/can be done separately

1) PCOSall ~ SNP (only autosomal variants, female and male controls)
2) PCOSwomen ~ SNP (genome-wide including X chromosome, female controls only)

And also with additional adjustments for age and BMI, when possible:

3) PCOSall_bmi_adj ~ SNP+BMI
4) PCOSwomen_bmi_adj ~ SNP+BMI
5) PCOSall_age_adj ~ SNP+age
6) PCOSwomen_age_adj ~ SNP+age
7) PCOSall_adj ~ SNP+age+BMI
8) PCOSwomen_adj ~ SNP+age+BMI
Association analyses – gene-wise analyses

- SKAT-O test implemented in EPACTS for gene-wise analyses
- Same covariates can be used as in single variant analyses
- Group files (using ANNOVAR) are required to tell EPACTS how to group variants (e.g. gene is a grouping indicator)
- Different models can be used:
  - Non-synonymous variants in a gene
  - Loss-of-function variants in a gene
  - Pathogenic variants in a gene
File naming
A file name should include:

STUDY_GENDER_TRAIT_EXOMECHIP_TEST_MODEL_DIAGNOSTIC CRITERIA_ANALYST_DATE

In these filenames:

STUDY - a short name or acronym for the study.
GENDER - ALL (when including both female and male controls), WOMEN (female controls only); for cohorts that only include women, please only upload the file with WOMEN in the gender field in the file name.
TRAIT - trait names include:
   1) PCOS (i.e. no covariates)
   2) PCOSadjAGE
   3) PCOSadjBMI
   4) PCOSadjAGE-BMI
TEST - SINGLE or GROUP.
MODEL - EMMAX, SCORE, WALD, NS1PCT, LOF, and PATH for Models 1 to 6 respectively.
DIAGNOSTIC CRITERIA - ALL (when including other criteria), Rotterdam, NIH. Please note that for phenotype group analyses (as outlined in section 2d above), this should specify the phenotype group defined as:
   Pheno1 = PCOS-ALL
   Pheno2 = PCOS-NIH
   Pheno3 = PCOS-Rotterdam
ANALYST - initials of the analyst.
DATE - date of the analysis in the form DDMMYY.
Summary

- Rare genetic variants may hold a considerable amount of the missing heritability of complex diseases/traits
- There are different tests that can be used while performing rare variant analysis
- Different kinds of data can be used: sequencing data, genome-wide or exome genotyping array data
- Single variant analyses as well as gene-wise analyses can be conducted
Homework

Please send homeworks to kreetelyll@gmail.com by 18th November 10:00
1. Try to give a short explanation why different steps are necessary in the quality control and what they show (both sample QC and variant QC)

- **Sample QC**
  - Call rate
  - Heterozygosity
  - Distribution of singletons per sample
  - Outliers from the principle component analysis

- **Variant QC**
  - Hardy-Weinberg disequilibrium
  - Cluster separation score
  - GenTrain score
2. Give arguments why you should use rare variants in one analysis and common variants in another. What are the pros and cons of these two compared to each other?

3. Why is it necessary to distinguish the terms genotype and phenotype?

4. Describe with your own words what is minor allele frequency (MAF).