GWAS Analyses

```
$ echo "Data Sciences Institute"
```

What You Will Learn Today

- **Quality Control (QC) essentials for GWAS**: why QC matters and common filters.
- **Hands-on GWAS walkthrough (1 hour)**: end-to-end tutorial from input to association output.
- **From discovery to validation**: replication criteria and why signals may not replicate.

Quality-Control (QC)

- QC is mandatory in genetic studies.
- Sample quality depends on DNA source (e.g., blood vs. buccal), handling, and storage.
- Error profiles **vary by platform**.
- **Random** genotyping error → lower power, no systematic bias.
- **Non-random** error → can inflate **false positives**.
    - For example, case/control imbalances (different batches, labs, technologies).
- Avoid convenience controls that don’t match cases.

Genotype Calling

- Then vs. now: Early studies relied on manual inspection of electrophoresis gels; modern studies use fully automated pipelines.
- How calls are made: Algorithms use allele-specific intensity signals for each SNP and assign the most likely genotype.
- Platform specificity: Different genotyping platforms employ their own calling algorithms, tuned to their chemistry and probe design.

Genotype Calling

- Historically: manual calls from gels; now automated from allele intensity data.
- Algorithms infer genotypes for each SNP using two-allele intensities.
- Different platforms use different callers.

Typical Quality Filters

- GWAS include hundreds of thousands of SNPs and thousands of samples; even a 1% error rate yields a substantial number of SNPs and samples with errors.
- Remove SNPs and samples with excessive **missing or failed calls**.
- Extensive QC is performed prior to association testing to prevent non-random errors from inflating false positives.

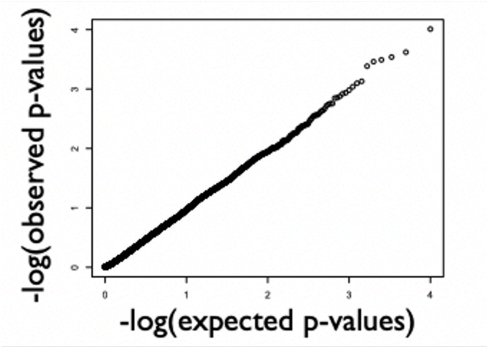
Filtering Steps

- **MAF & missingness**: Drop SNPs with MAF < 5% (study-dependent) and remove SNPs or samples with >~5% missing calls.
- Hardy-Weinberg equilibrium: **Exclude SNPs showing HWE departures** (e.g., $P_{\mathrm{HWE}} < 0.001$), which often indicate genotyping error.
- Mendelian consistency: In pedigree data, remove SNPs with excess **Mendelian errors**.
- **Relatedness checks**: Identify cryptic relatives or duplicates; unmodeled relatedness inflates test statistics and false positives.
- After QC, evaluate the dataset with a quantile-quantile (Q-Q) plot to confirm that the test statistic distribution matches expectation (no residual inflation).

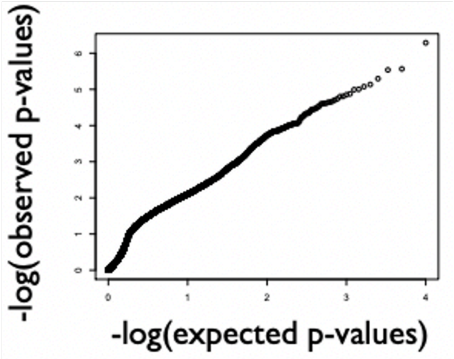
Q-Q plots

- Q-Q plots can diagnose systematic inflation or deflation of test statistics (e.g., uncorrected population stratification, batch effects).
- If most variants are truly null $\rightarrow$ points lie on the 45° line.

Good Genotyping Quality



Poor Genotyping Quality



Source: publicly available from Google Images and modified by Fan Wang

- Important caveat: Many traits are polygenic; widespread weak effects can cause mild upward deviation even with good QC.

GWAS tutorial

Replication vs. Non-replication

- After a new association is reported, it should be validated in an independent cohort.
- Genotype the same SNP in an independent sample and test the association again.
- Replication criteria (very strict):
  - The SNP attains statistical significance ($\alpha = 0.05$).
  - Same genetic model is supported (additive, dominant, recessive, etc.).
  - Effect direction matches the discovery (risk $\leftrightarrow$ protective).
  - The replication study has adequate power (typically $80 - 90\%$).
- Note: Failure to replicate does not automatically invalidate the original finding.

Why a Result Might Not Replicate

- **False positive** in the discovery study.
- **Insufficient power** in the replication sample.
- The tested SNP is a **tag** (in LD with the causal variant); **LD differences** across populations can break the signal.
- **Context dependence**: effect sizes can vary with covariates (e.g., age), study phenotype definition, or other design differences.

What's Next

- **What GWAS teaches us**: polygenicity, effect sizes, and cross-ancestry considerations.
- **Post-GWAS fine-mapping basics**: Regular approaches and Bayesian approaches.