Data Quality Control

DNA Collection

- Blood samples
  - For unlimited supply of DNA
    - Transformed cell lines
      - Is expensive
    - Whole genome amplification
      - Allows for the creation of large amounts of DNA from initial small DNA sample
        - Perform WGA on each sample three or more times and use pooled samples
      - Can experience lower call rates and higher genotyping error rates
      - Not recommend to use WGS samples for Copy Number Variant analysis
- Buccal Swabs
    - Small amounts of DNA
    - DNA not stable
- Saliva (Origene collection kit)

Measurement of DNA Concentrations

- Nanodrop
- Picogreen

Genotype SNPs (~20-96) before Exome or Whole Genome Sequencing

- Genotype markers which can be used as DNA fingerprint
- Allows for Assessment of DNA quality
- Determines the sex of the individual
  - To aid in identification of sample swaps
- Detects cryptic duplicates
- For family data
  - Aids in determining close familial relationships
    - Non-paternity
    - Sample swaps
    - Cryptic relationships

Detecting Genotyping Errors

- Duplicate samples genotyped to detect inconsistencies
  - Can use duplicate samples that are inconsistent to adjust clusters to improve allele calls
    - Will not detect systematic errors
- Usually not performed for exome and whole genome sequencing studies

Effects of Genotyping Error

- If there is no bias in genotyping error between cases and controls
  - Same rates of genotyping errors in case and control data
- For family based association studies - Trios
  - Can increase both type I and II error
- Population based studies
  - Increases type II error only
- Cases and controls are genotyped
  - At different times
  - Different institutions
  - Or one group, case or control, is predominately genotyped at one time/same batch
- Can lead to different genotyping error rates in cases and controls
  - In this situation both type I and II error can be increased
- If genotyping cases and controls
  - Randomize cases and controls so they are spread evenly across genotyping runs.

Convenience Controls

- Can reduce the cost of a study
- Genotype data
- Type I error can be increased
  - Ascertainment from different population
  - Differential genotyping error
    - Even if performed at the same facility
- Proper QC can reduce or remove biases

Convenience Controls–Sequence Data

- Obtain BAM files and recall cases and control together
  - Can still have differential errors between cases and controls
  - Check variant frequency by variant types in cases and control
    - Synonymous variants should have the same frequencies
    - Would not expect large differences in numbers of variants between cases and controls
- For single variants can compare difference in frequencies with gnomAD but is problematic
  - Differences in frequencies can be due to differences in populations and errors
  - Can not adjust for confounders
    - e.g. sex, population substructure/admixture
- Don’t perform an aggregate test using frequency information from gnomAD

Genotype Data QC – Population Based Studies

- Remove DNA samples from individuals who are missing >3% or their genotype data
  - May choose to use an even more stringent criteria
- Low DNA quality can lead to higher genotypes error rates at markers with available genotype data
- Lower call rates in individuals may be due to DNA contamination with another DNA sample
- To avoid markers with higher genotyping error rates
  - For markers with a minor allele frequency (MAF)≥0.05
    - Remove markers missing >5% of their genotype data
  - For makers with a MAF<5%
    - Remove markers missing > 1% of their genotype data

Additional QC Family based studies

- Detect double recombination events over small distances
  - Can be an indication of genotyping error
    - Merlin
- Detect non-Mendelian errors of segregation
  - Can be due to genotyping errors*
  - If a larger number are observed
    - Could be due to incorrect specification of pedigree structure
      - e.g. non-paternity
    - Pedcheck

*Many genotyping errors will not be detected for single nucleotide variants (SNVs). The probability of detecting non-Mendelian segregation due genotyping errors decreases with increasing MAF

Data Clean – Accessing Sex

- Males with an excess of heterozygous SNPs on the X chromosome can denote
  - Males mislabeled as females
  - Males with Klinefelter syndrome
  - Note: Males will be heterozygous for markers in the pseudoautosomal regions
- Females with an excess of homozygous genotypes on the X chromosome can denote
  - Females mislabeled as males
  - Females with Turner Syndrome

Data Clean – Accessing Sex

- Males mislabeled as females and females mislabeled as males
- Can be observed due to sample mix-ups
- Samples for which the sex is incorrect
  - Should be removed from the analysis
- Probably not the person you think it is

Checking for Potential DNA Contamination

- DNA samples which have been contaminated by another DNA sample will have a larger proportion of their genotypes being heterozygous*
- If cross contamination of samples within the same study will observe "relatedness" amongst study subjects
- Can also be observed from principal components analysis (PCA)/multidimensional scaling (MDS)
  - Samples which are cross contaminated will cluster together and not cluster with other samples

*Higher levels of heterozygous markers will be observed in individuals of sub-Saharan African ancestry compared to those of European and Asian ancestry.

Checking for Duplicate and Related Individuals

- Duplicate samples are sometimes included in a study as part of quality control
  - These can easily removed before data quality control
- Cryptic duplicates (unintentional)
  - DNA sample aliquoted more than once
  - Individual ascertained more than once for a study
    - e.g. The same individual undergoes the same operation more than once and is ascertained each time
- Individuals who are related to each other may participate in the same study
  - Unknown to the investigator

Identifying Duplicate and Related Individuals

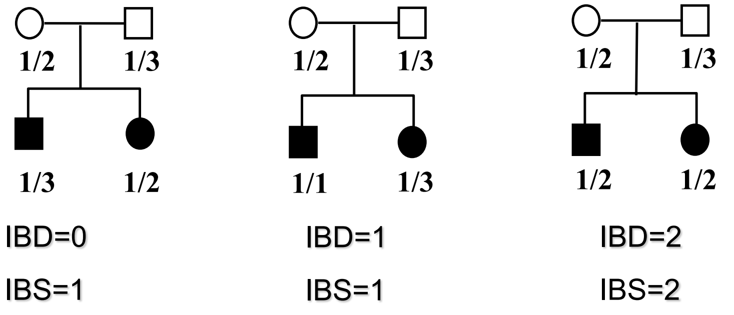
- Genotype data from one of the duplicates needs to be removed from the dataset
- Only one related individual should be retained in the data set
  - If related individuals remain in the data set mixed-models need to be used to analyze the data*
    - Case-Control
      - Generalized linear mixed models
    - Quantitative traits
      - Linear mixed models
  - If not type I error rates can be increased

*If only a few related individuals in sample, may wish to remove them or use mixed-models to control type I errors. Must use mixed models if many related individuals in dataset. Due to the construction of the generalized relationship matrix (GRM) can be problematic to apply for large samples sizes, e.g. UKbiobank.

Identifying Duplicate and Related Individuals

- Duplicate and related individuals can be detected by examining identify by state (IBS) adjusted for allele frequencies (p-hat) between all pairs of individuals within a sample
  - To estimate IBD sharing

IBD/IBS



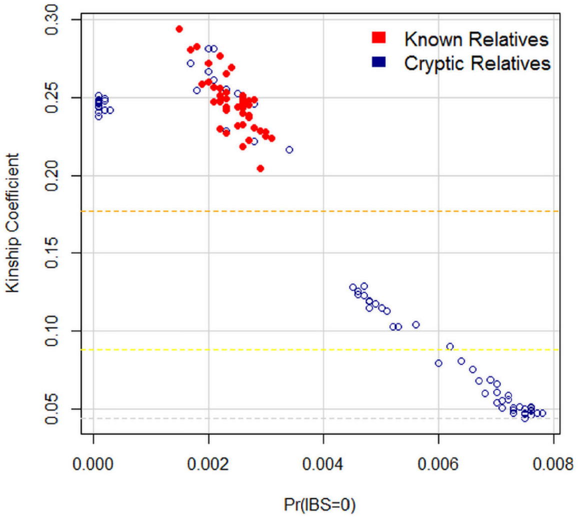
Identifying Duplicate and Related Individuals

- IBS is the number of alleles of alleles which are shared between a pair of individuals
  - Can either share 0, 1, and 2 alleles
- Duplicate individuals will have IBS measures of 1 or close to 1
    - If there is genotyping error could make IBS< 1
- IBS measures are adjusted for by allele frequencies p-hat
  - Approximates IBD sharing

Identifying Duplicate and Related Individuals

- Siblings and child-parent pairs will share 0.50 of their alleles IBD
  - For parent-child IBD=1 is ~1.0
  - For sibs IBD=1 is ~0.50
    - For more distantly related individuals the IBD measure will be lower
- Can use whole genome scan data to check IBD
  - Should "trim" markers so that they are not in LD
  - Caution should be used if using candidate genes are being studied since many markers will be in LD which can inflate the IBD measure
- PLINK or KING can be used to identify duplicate and related individuals

King Graphical Output



Observing Low Levels of IBD Sharing for Multiple Individuals

- P-hat is calculated using the "population" allele frequency
- If individuals in sample come from different populations
  - Individuals from the same population within the sample will have inflated p-hat values due to incorrect allele frequencies
    - Appear incorrectly to be related to each other
- "Relatedness" amongst many individuals can also be observed when batches are combined if they have different error rates
  - Individuals from the same batch appear to be related
- DNA contamination can cause "relatedness" between multiple individuals

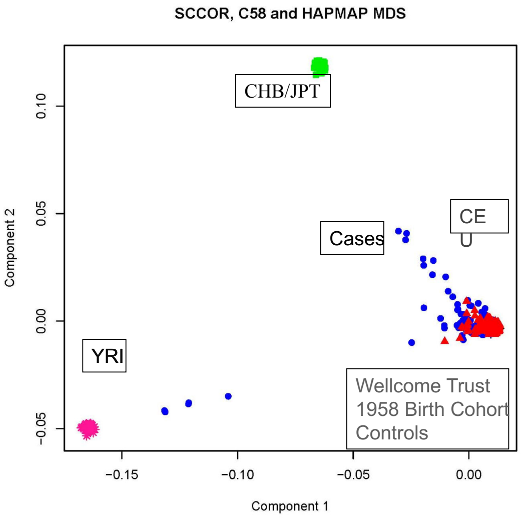
PCA/MDS

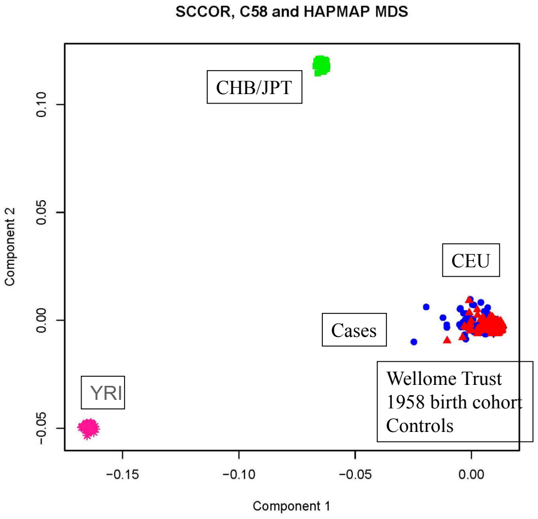
- Can be used to find outliers
- Individuals from different ethnic backgrounds
  - African American Samples included in samples of European Americans
- Use a subset of markers which have been LD pruned
  - So there is only very low levels of LD between marker loci
- Plot 1st component vs. 2nd component
  - Additional components should also be plotted to determine potential biases in the data

PCA/MDS

- Used to identify outliers
  - Individuals from different ancestries
    - e.g. African American samples included with European Americans samples
      - Can use samples from HapMap to help to determine what ethnic group outliers belong to.
        - » Should not include HapMap samples when calculating components to control for population substructure/admixture
  - Related individuals
  - Problems with genotype quality
    - Batch effects

Detecting Outliers Using PCA





Detecting Genotyping Error – Examining HWE

- Testing for deviations from Hardy-Weinberg Equilibrium (HWE) not very powerful to detect genotyping errors
- The power to detect deviations from HWE is dependent on
  - Error rates
  - Underlying Error Model
    - Random
    - Heterozygous genotypes -> homozygous genotypes
    - Homozygous genotypes ->Heterozygous genotype
  - MAFs

Detecting Genotyping Error – Examining HWE

- Controls and Cases are evaluated separately
  - Deviation found only in cases can be due to an association
- Test for deviation from HWE only in a sample of the same ancestry
  - Population substructure can introduce deviations from HWE
- Do not include related individuals when testing for deviations from HWE
  - Can cause deviation from HWE
- Quantitative Traits
  - Caution should be used removing markers which deviate from HWE may be due to an association
    - Remove markers with extreme deviations from HWE and Flag markers with less extreme deviations from HWE

Detecting Genotyping Error

- Quantitative Traits
  - Caution should be used removing markers which deviate from HWE may be due to an association
    - Remove markers with extreme deviations from HWE and Flag markers with less extreme deviations from HWE
- In the presents of genotyping error
  - Random error and equal allele frequencies
  - Power to detect deviation from HWE is α
- Pseudo SNPs lend themselves to readily be detected through testing for deviation from HWE

Testing for Deviations in HWE - α value

- For different studies a variety of α values are used
  - Need to consider that multiple testing is being performed
- WTCCC used a criterion $p<5 \text{ X } 10^{-7}$ to remove SNPs
- A large variety of criterion are used to reject the null hypothesis of HWE
  - A criterion of 1.0×10^{-4} is often used in published studies
- Deviations from HWE can be used to flag SNPs
  - e.g. Remove those SNPs with deviations from HWE with $p<5.0 \times 10^{-8}$
    - Note should be more conservative if performing quality control for imputation
  - SNPs can then be investigated in more detail later for reasons for the observation of deviation from HWE
    - If there are significant association results
- A significant result with a large deviations from HWE in cases and controls is probably due to genotyping error or a pseudoSNP

Deviation from HWE

- HWD coefficient (Weir 1996)

$$D = P_{11} - p^{2}$$

- For SNPs (2 allele system)
- The proportion observed for N genotypes G_{11}, G_{12} and G_{22}
  - Is P_{11}, P_{12} and P_{12} respectively
- p is the allele frequency which is estimated by

  $(2*G_{11} + G_{22})/2N$

Deviation from HWE

- Under HWE D=0
- Negative values of D indicate an excess of heterozygote genotypes
- Positive values of D indicate an excess of homozygote genotypes
- For a diallelic system
  - D can range from -0.25 to 0.25
- Markers not in HWE
  - May be due to population admixture
    - Excess of heterozygous genotypes (-D)
  - Copy number repeats (CNVs)
    - Either an excess of heterozygous (-D) or homozygous genotypes (+D)
- Heterozygous Advantage
  - Excess of heterozygous genotypes (-D)

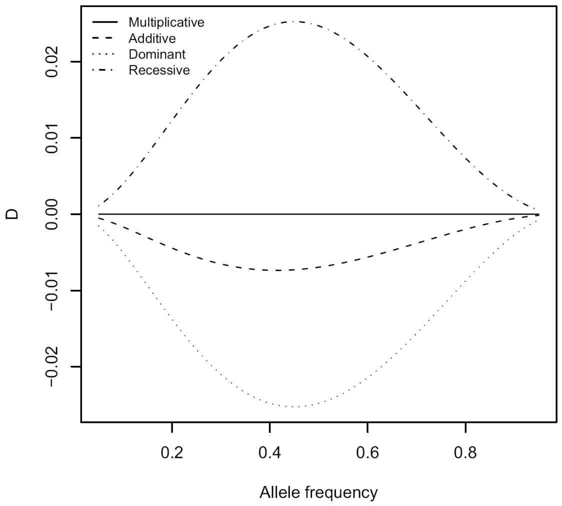
Detecting Genotyping Error

- Chance
  - Either an excess of heterozygous (-D) or homozygous genotypes (+D)
- Deviation from HWE due to genotyping error
  - Either an excess of heterozygous (-D) or homozygous genotypes (+D)
- Deviation from HWE due to pseudo SNPs
  - Excess of heterozygous (-D) genotypes
- Indication that deviation from HWE due to genotyping error
  - Higher genotype drop out rates for specific markers
- Pseudo SNPs
  - Primers map to multiple genomic regions

Deviation from HWE

- Genotype data from cases deviate from HWE when the SNP which is being tested is functional or in LD with a functional variant
  - Additive genetic model (D- excess heterozygous genotypes)
  - Dominant genetic model (D- excess heterozygous genotypes)
  - Recessive genetic model (D+ excess homozygous genotypes)
  - Multiplicative genetic model there is no deviation from HWE (D=0.0)
- Controls – unaffected individuals
  - Display an extremely small deviation from HWE
  - Deviation from HWE is higher with higher disease prevalence
- Only unascertained samples have no deviation from HWE at the functional locus

Deviation from HWE



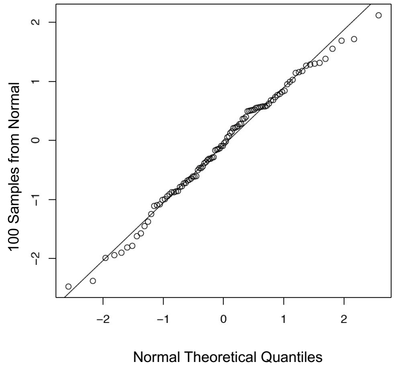
Odds Ratio 1.5

QQ plot

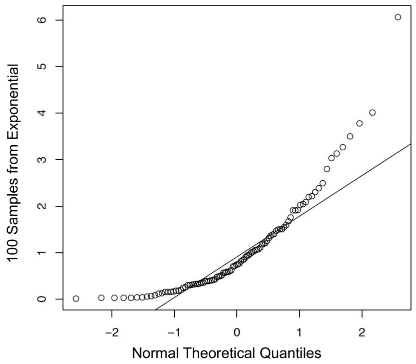
- QQ (quantile-quantile) plot is a graphical method for diagnosing differences between a random sample and a probability distribution
- Assume there are n points, ordered as x(1)<…<x(n). Then for i=1,…,n, we plot x(i) against the ith quantile (qi) of the specified distribution (e.g., normal).
- If the random sample is from the specified distribution, the QQ plot will be a straight line
- Let F() be the cumulative distribution function of a random variable, e.g., F(z)=P(x<z).
- The ordered n samples, x(1)<…<x(n), are treated as n empirical quantiles and the ith theoretical quantile, qi, corresponding to x(i) can be calculated as qi=F-1((i-0.5)/n)
- Plot qi on x axis and x(i) on y axis

QQ Plot Examples

Normal-Normal QQ plot



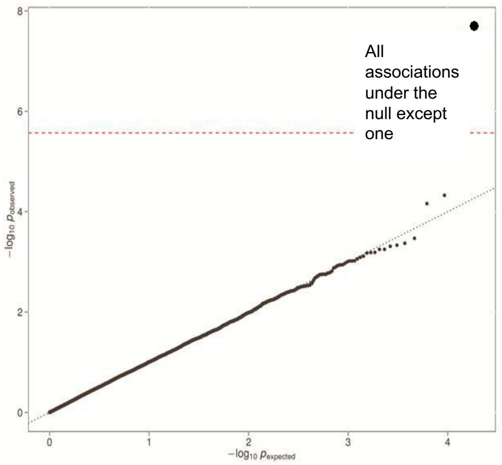
Exponential-Normal QQ plot



Genome Wide Association Diagnosis

- Thousands of markers are tested simultaneously
- The p values of neutral markers follow the uniform distribution
- If there are systematic biases, e.g., population substructure, genotyping errors, there will be a deviation from the uniform distribution
- QQ plots offers a intuitive way to visually detect biases

QQ Plot of Exome Wide P Values



All associations under the null except one

Genomic Inflation Factor to Evaluate Inflation of the Test Statistic

- Genomic Inflation Factor (GIF): ratio of the median of the test statistics to expected median and is usually represented as λ
  - No inflation of the test statistic λ=1
  - Inflation λ>1
  - Deflation λ<1
- Problematic to examine the mean of the test statistic
  - If for a number of variants the null can be rejected
    - Particularly if they have very small p-values
      - The mean test statistic will be inflated

Phenotype	Covariate	Mean Chi-Square	GIF (λ)
BP		1.23829	1.16932
BP	Age	1.24119	1.18025
BP	**Age-EV1**	1.09471	*1*
BP	Age-EV2	1.0881	1
BP	Age-EV4	1.08385	1
BP	Age-EV10	1.09582	1.00402
BPI		1.14931	1.08921
BPI	Age	1.15139	1.08113
BPI	Age-EV1	1.05079	1.01148
BPI	**Age-EV2**	1.0428	**1**
BPI	Age-EV4	1.04204	1
BPI	Age-EV10	1.05421	1.01724
BPII		1.17283	1.25664
BPII	Age	1.17583	1.26996
BPII	Age-EV1	1.09874	1.15065
BPII	Age-EV2	1.09904	1.16425
BPII	Age-EV4	1.09502	1.14609
BPII	Age-EV10	1.10046	1.1418
BPII	Sex,Age-EV1	1.05958	1.06424
BPII	**Sex,Age-EV4**	1.05817	**1.05323**
BPII	Sex,Age-EV10	1.06338	1.05581

Significance Results

- For those SNPs with strong associations
- Cluster plots should be examined
- Poor clustering – overlap between clusters
  - Could be the reason for the strong association
- Cluster caller should be adjusted when possible and data reanalyzed

Order of Data Cleaning

- Remove samples missing >10% genotype data
- Remove SNPs with missing genotype data
  - If MAF >5%
    - Remove markers with >5% missing genotypes
  - If MAF <5%
    - Remove markers with >1% missing genotypes
- Remove samples missing >3% genotype calls
- Check for sex of individuals based on X-chromosome markers
  - Remove individual whose reported sex is inconsistent with genetic data
    - Could be due to a sample mix-up
- Check for cryptic duplicates and related individuals
  - Used “trimmed data set of markers which are not in LD
    - E.g. $r^2 < 0.5$
  - Retain duplicate with best genotyping quality

Order of Data Cleaning

- Perform principal components analysis to check for outliers
  - Use trimmed data set of markers which are not in LD
    - E.g. $r^2 < 0.5$
  - Remove outliers from data
- Check for deviations from Hardy Weinberg Equilibrium
  - Separately in cases and controls
  - If more than one ethnic group
    - Separately for each ethnic group
- Examine QQ plots for potential problems with the data
  - e.g. not controlling adequately for population admixture