



Handling and QC data from a genomewide association study

Bahareh Sedaghati-Khayat

Cellular and Molecular Endocrine Research Center, Shahid Beheshti
University of Medical Sciences



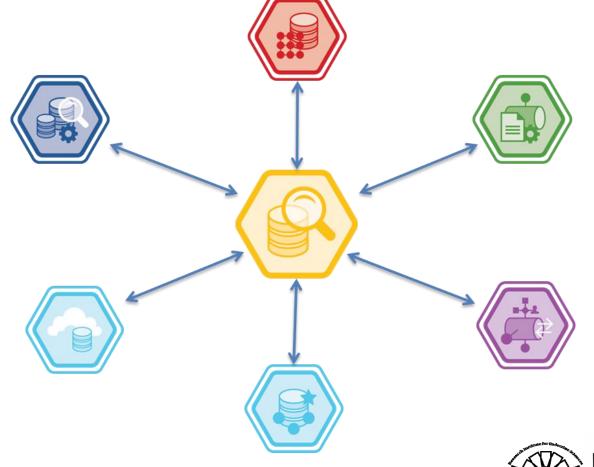
Outline

- Define the role of genetic study in cohorts
- How to establish a standard genomic bank?
- How to collect and develop family relationships in a cohort
- Share the genetic data and define the genomic map
- Create a database of phenotypes in existing cohort
- How to control data quality and integration of genomic data?
- Draw a road map for personalized medicine in the context of cohort studies



•Storing large amounts of genotype data

Quality control







Genotype data is huge

- 500,000 SNPs * 2000 cases + controls = 1,000,000,000 genotypes!
- Need compact ways to store data

 - Total file space for 300K SNPs: 4 Gigabytes
 - Largest chromosome file: .4 Gigabytes





Need to do extensive planning for genotype data before it arrives

- Chromosome datasets are too large for ordinary and commonly used analytic packages (SAS,SPSS,...)
- Need programs to select and write out genotype data in multiple formats
- Tests of procedures with large-scale trial datasets





Gather other data needed for analysis

- SNP information
 - Chromosome
 - Position
- SNP annotation
 - Gene
 - Function
- Translation of called allele to a standard allele
 - Example forward strand of given genome build





How good is the data?

- Identify and remove bad samples and SNPs
- Compute summary statistics
 - Percent successfully genotyped samples
 - Average genotyping success rate
 - Duplicate sample error rate
 - Non-Mendelian inheritance error rates (errors not consistent with normal transmission of chromosomes in family members)







Identify bad samples and remove

- Poor quality samples
 - Sample genotype success rate < 95 to 97.5%
 - Greater proportion of heterozygous genotypes than expected
- Related individuals (if independent samples)
 - Based on pair-wise comparisons of similarity of genotypes
- Sample switches
 - Wrong sex





Identify poor quality SNPs and remove

- Expected proportions of genotypes are not consistent with observed allele frequency (Hardy Weinberg Equilibrium (HWE))
 - HWE p-value $< 10^{-4}$ to 10^{-6}
 - Look for deviation from expected distribution of p-values under the null
- Genotyping success rate < 95%
- Duplicate sample or Non-Mendelian error rate is elevated
- Differential missingness in cases and controls





Initial analysis is straightforward once have everything in place

- Case/control association
 - Use test that is not affected by deviations from HWE
 - Cochran-Armitage test for trend
 - Equivalent to score test in logistic regression
- TDT or other family-based test
- Quantitative trait association





Are the results believable?

- Are stronger associations correlated with poorer quality control measures?
- Is there a strong deviation from expected distribution of p-values?
- Is there confounding from differences in the genetic origins of case and control samples (population stratification)?
 - Genomic control
 - Eigenstrat analysis





Getting more for your genotyping dollars: Imputation of SNP genotypes



- Impute/predict genotypes for:
 - Missing data within genotyped markers
 - Untyped markers
- Uses haplotype structure of existing sample such as HapMap samples to infer data for samples with sparser marker set
- Defining and Using of Reference Panel for Iranian population





Imputed data takes care to generate, analyze and understand

- Requires large scale computing resources
- Need to assess quality of imputation
 - Compare imputed gentoypes to actual genotypes
- Error rates are higher than for genotyped SNPs
- Works less well for rarer alleles
- Etc...





Summary: Ideally what needs to happen before getting the data

- Ability to store, select and write out genotype data in multiple formats for quality control and association analysis
- Identification of primary quality control and analysis programs
- Systems to store, view, merge results
- Adequate computing resources to do intensive computing
- Testing of standard and specialized processes with largescale trial datasets











Thank you for your attention



