BiP 2014: Module 4

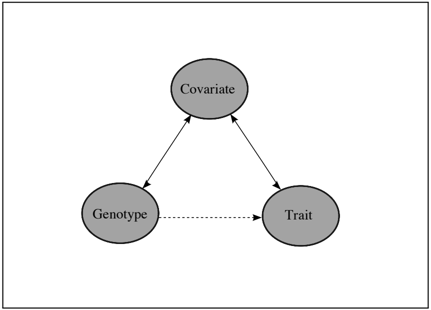
Author: Andrea S. Foulkes

Genome Wide Association Studies (GWAS)

The overarching **goal** *of genome wide association studies is to identify genes associated with complex traits.*

Three broadly-defined data components:

1. Genetic information
2. Trait (phenotype) measuring disease progression or status
3. Demographic and clinical covariates

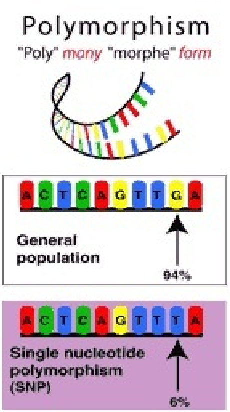


Data components and terminology

Genetic information

- Humans carry 2 *homologous chromosomes*:
  - segments of DNA, one inherited from each parent.
  - code for same trait, may carry different genetic information.
- *Nucleotide*:
  - DNA base + sugar molecule + phosphate.
  - used interchangeably with *base*.
- Gene:
  - region of DNA
  - code for proteins or involved in regulation of production of proteins from other segments of DNA

Data components and terminology

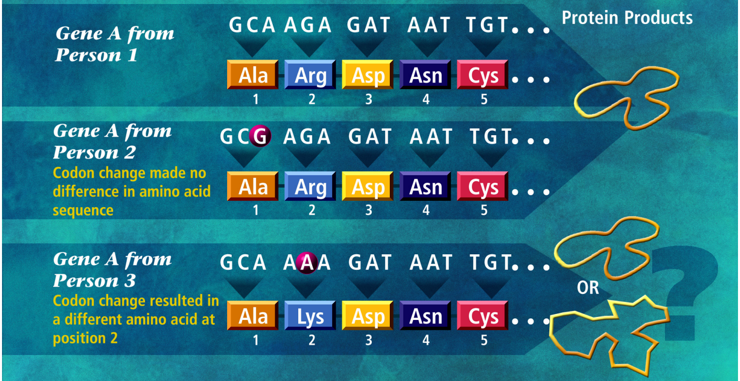


- SNP (x): basic unit of analysis, typically coded 0, 1, 2 for number of variant alleles on 2 chromosomes
- Trait (y): measure of disease progression or disease status.

Definitions:

- **polymorphism**: genetic variant occurring in greater than 1% of a population
- **single nucleotide polymorphism (SNP)**: variant at a single site (base pair position) on the genome.

DNA Sequence Variation in a Gene Can Change the Protein Produced by the Genetic Code



Protein Products

Gene A from Person 1

GCA AGA GAT AAT TGT...

Ala Arg Asp Asn Cys...

1 2 3 4 5

Gene A from Person 2

Codon change made no difference in amino acid sequence

GCG AGA GAT AAT TGT...

Ala Arg Asp Asn Cys...

1 2 3 4 5

Gene A from Person 3

Codon change resulted in a different amino acid at position 2

GCA AAA GAT AAT TGT...

Ala Lys Asp Asn Cys...

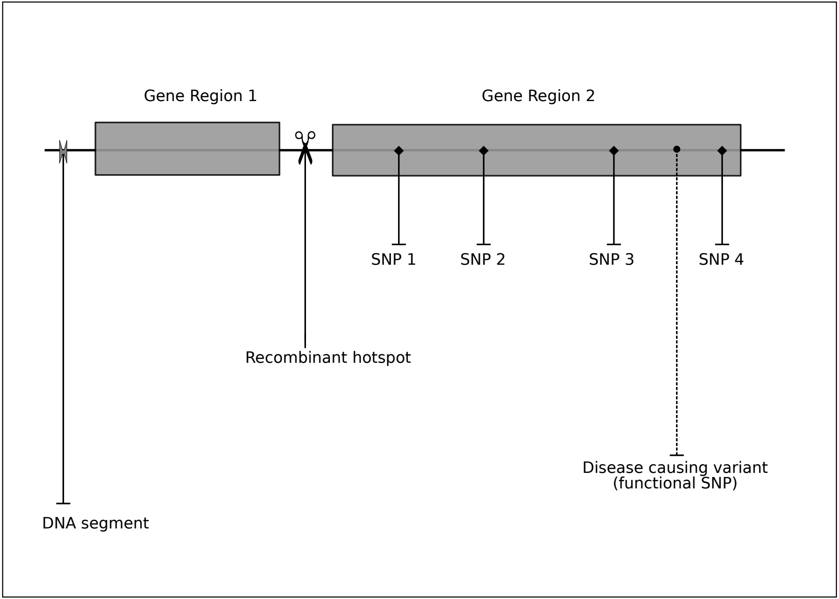
1 2 3 4 5

OR

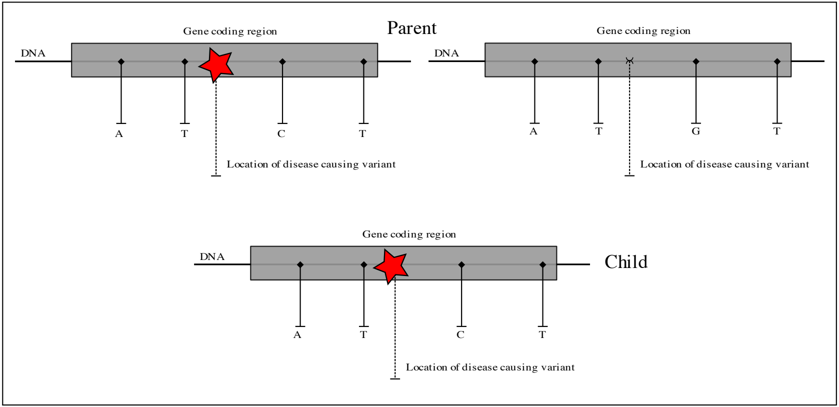
?

Y-GA 98-649

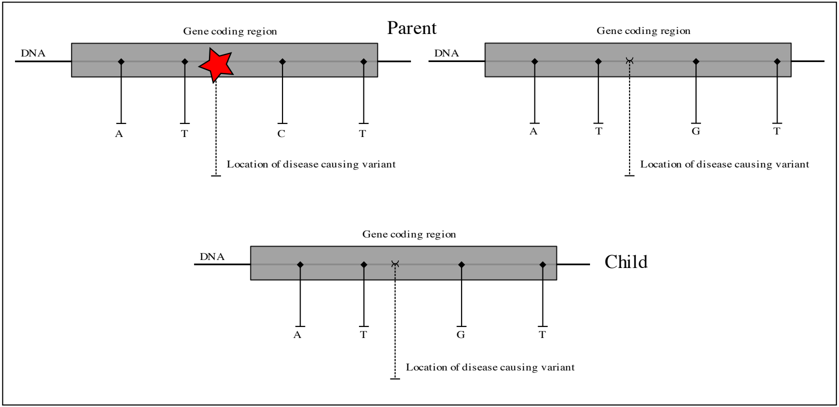
Data components and terminology



Data components and terminology


Parent
Gene coding region
DNA
A
T
C
T
Location of disease causing variant
Gene coding region
DNA
A
T
G
T
Location of disease causing variant
Gene coding region
DNA
A
T
C
T
Location of disease causing variant
Child

Data components and terminology


Parent
Gene coding region
DNA
A
T
C
T
Location of disease causing variant
Gene coding region
DNA
A
T
G
T
Location of disease causing variant
Gene coding region
DNA
Child
A
T
G
T
Location of disease causing variant

Data components and terminology

Trait

- clinical outcome or phenotype, measured *in vivo* or *in vitro*.
- **quantitative**, **binary** (diseased or not diseased), survival (censored), longitudinal/multivariate.
- e.g. total cholesterol, triglyceride levels, heart attack, CD4+ cell count, viral load, AIDS defining event, time to death, repeated measures of total cholesterol, etc.

Covariates

- environmental, clinical and demographic data.
- potential predictors, confounders, effect modifiers, effect mediators (causal pathway variables).
- also referred to as *predictors*, *confounders*, *explanatory variables*, *independent variables*.
- e.g. age, gender, race/ethnicity, BMI, smoking status, etc.

GWAS Analysis

“Typical” analysis approach:

- Separate test of association (based on multivariable linear model) for each SNP → p-value for each SNP.

$$Y = X\beta + Z\gamma + \epsilon$$

$$H_0 : \beta = 0$$

- Adjust to control Family Wise Error Rate (FWER) in context of multiple testing:

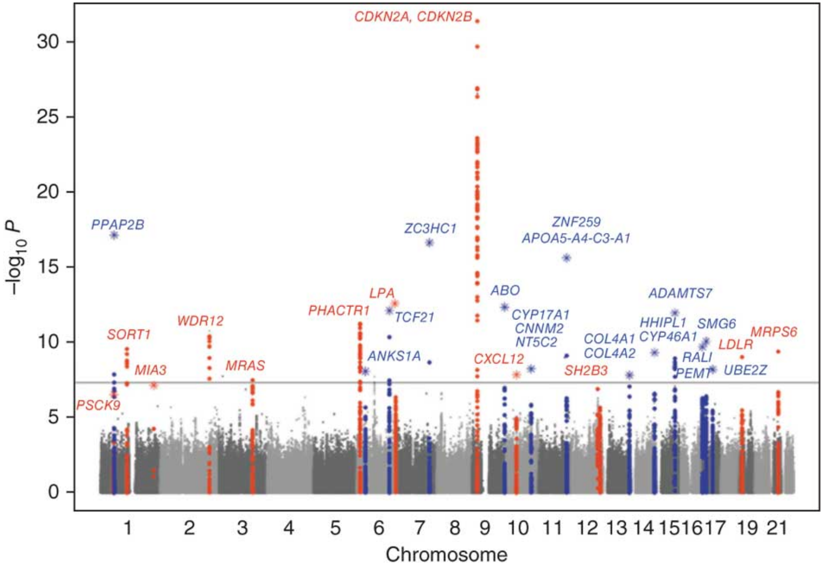
$$FWER = Pr(\text{reject at least one } H_0^k \,|\text{all } H_0^k \text{ are true})$$

- Typically control at level $\alpha = 0.05$ using Bonferonni adjustment → P-value statistically significant if less than $0.05/1,000,000 = -5 \times 10^{-8}$.

CARDIoGRAM summary level data

Coronary **AR**tery **DI**sease **G**enome-wide **R**eplication **A**nd **M**eta-anaylsis (CARDIoGRAM) data:

- Meta-analysis of 14 GWAS of coronary artery disease (CAD): 22,233 cases and 64,762 controls
- Replication study in additional $56,682$ individuals
- Available data (after pre-processing): p-values for $965,220$ SNPs in $19,216$ genes.



Schunkert et al. Nature Genetics 43, 333-338 (2011) doi:10.1038/ng.784

Lab Assignment

1. Conduct a simulation study to identify an appropriate p-value threshold for statistical significance of the minimum p-value (assuming independence of SNPs):
   - generate $965,220$ p-values from a uniform distribution
   - determine value corresponding to the minimum
   - repeat 500 times and record 5th percentile of this distn.
2. Repeat (1) for the second smallest p-value, the third smallest, etc.
3. Challenge Problems:
   - Repeat (1) while accounting for within gene correlation
     - assume inverse normally transformed p-values (p_{ij}) arise from a random effects model (i indicates gene and j indicates SNP):

$$y_{ij} = b_i + \epsilon_{ij}$$

$$p_{ij} = \Phi^{-1}(y_{ij}),\ b_i \sim N(0, 0.4),\ \epsilon_{ij} \sim N(0,1) \text{ and } b_i \perp \epsilon_{ij}.$$

   - Repeat (2) where the random gene level effects arise from a $N(0, \sigma_b^2)$ and σ_b^2 ranges from 0.2 to 1.2 in increments of 0.2.

In this lab we will use the Cardiogram summary level data. To begin, read in these data and look at the first 10 rows:

```
setwd("~/BiP/2013HPCwithR/modules/module4/source/labs/")

## Error:  cannot change working directory

cardioDat <- read.csv("cardioPvalues.csv")
print(cardioDat[1:10, ])

##          name     type    gene    pval
## 1        rs10 intronic    CDK6 0.19381
## 3  rs10000010 intronic  KCNIP4 0.03015
## 4  rs10000012 intronic   UVSSA 0.94010
## 8  rs10000023 intronic  BMPR1B 0.30638
## 13 rs10000037 intronic FAM114A1 0.30434
## 16 rs10000042 intronic  STK32B 0.35717
## 19 rs10000062 intronic  STK32B 0.24720
## 27 rs10000092 intronic  KCNIP4 0.00138
## 30 rs10000109 intronic  CCSER1 0.66652
## 32  rs1000012 intronic     AK8 0.33443
```

Exercise 1:

Conduct a simulation study to identify an appropriate p-value threshold for statistical significance of the minimum p-value (assuming independence of SNPs):

- generate $965,220$ p-values from a uniform distribution
- determine value corresponding to the minimum

```
sim1 <- runif(nrow(cardioDat))
sort(sim1)[nrow(cardioDat) * 0.05]

## [1] 0.04977
```

- repeat 500 times and record 5th percentile of this distn of minimums

```r
## locally parallelized permutation test BiP HPC workshop 2014

require(doMC)
require(foreach)
nSim <- 500  ## number of simulations
nCores <- 10
registerDoMC(nCores)
# setwd('/home/ngr67a/BiP/')
```

Run simulation loop, storing minimum each time

```r
matNullP <- foreach(i = 1:nSim, .combine = rbind) %dopar% {
    sim <- runif(nrow(cardioDat))
    c(i, min(sim))
}
```

Determine 5th percentile of the distribution of minimum p-values:

```
sort(matNullP[, 2])[nSim * 0.05]

## result.291
##   6.24e-08
```

Compare to minimum observed pvalue in Cardiogram

```
cardioDat[cardioDat[, 4] == min(cardioDat[, 4]), ]

##             name     type gene      pval
## 92922 rs10455872 intronic  LPA 3.079e-13
```