BiP 2014: Module 4

Author: Andrea S. Foulkes

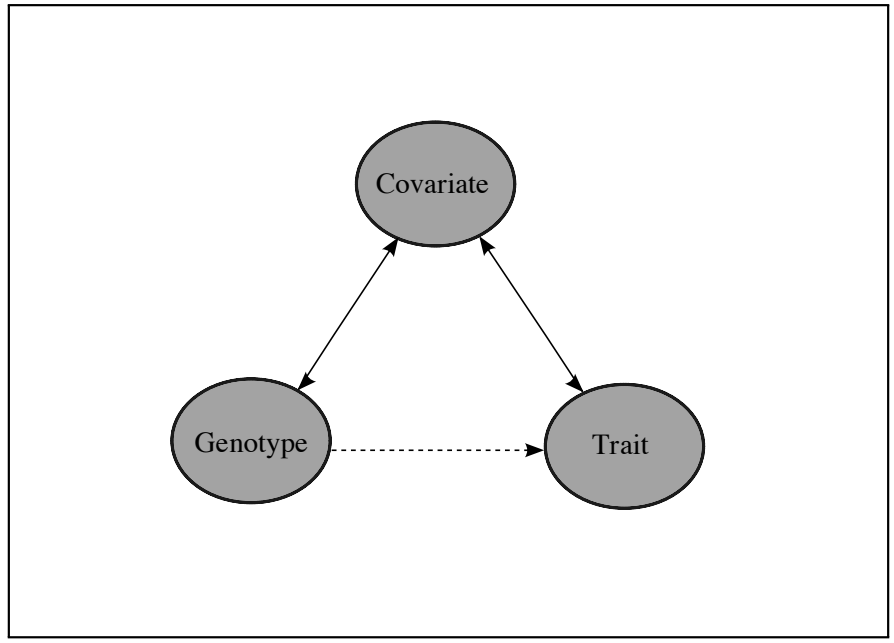
This material is part of the **statsTeachR** *project*

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

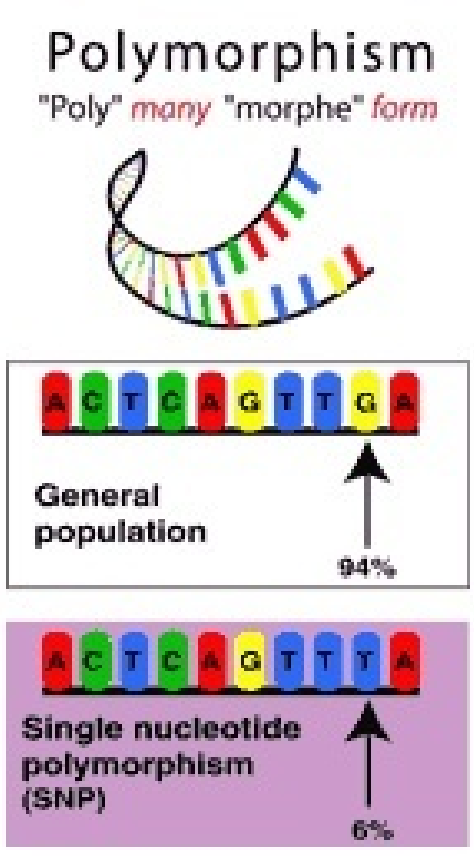


Data components and terminology

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

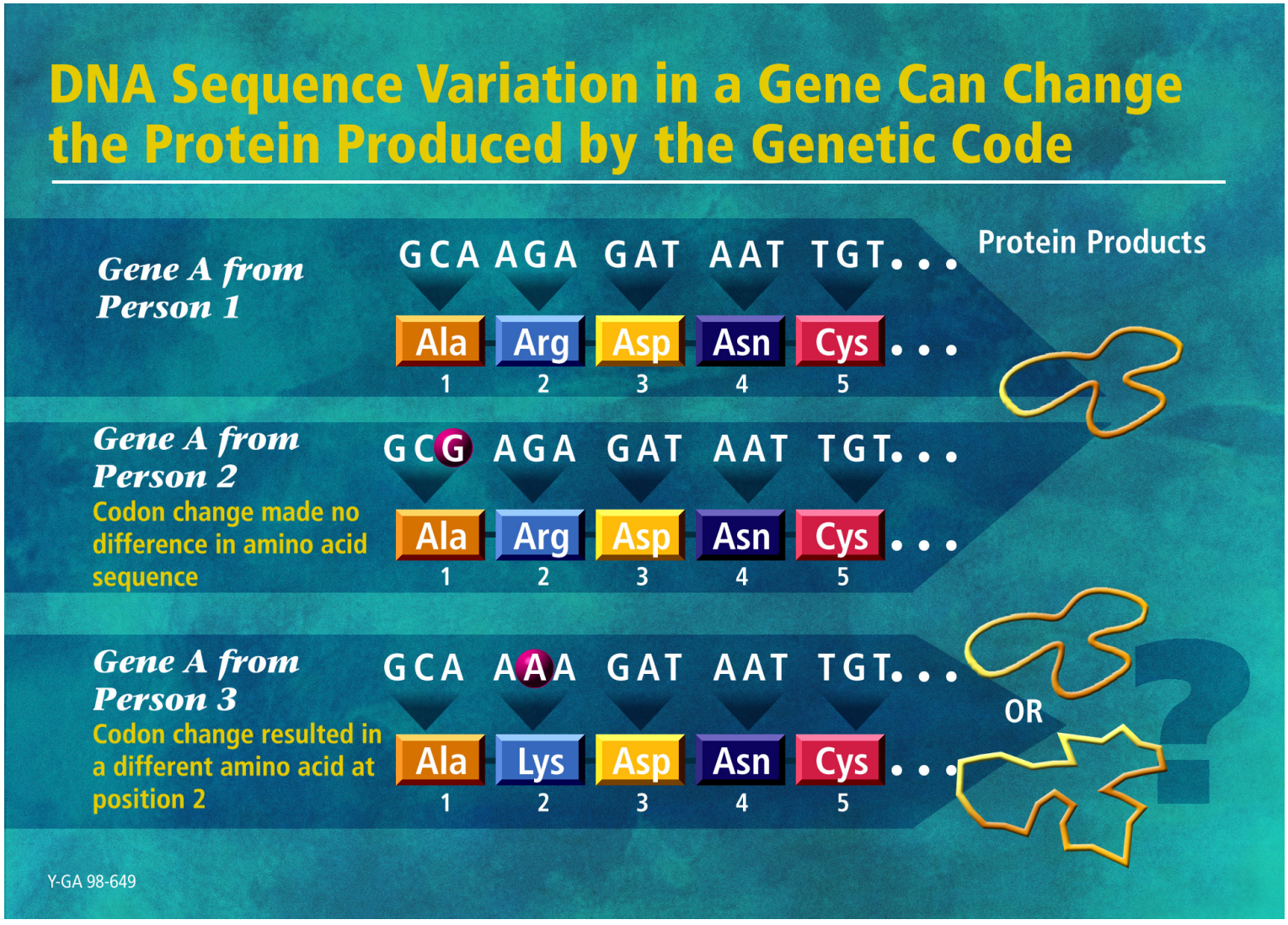


- SNP (x): basic unit of analysis, typically coded 0, 1, 2 for number of variant alleles on 2 chromosomes
- Trait (y): measure of disease progression or disease status.

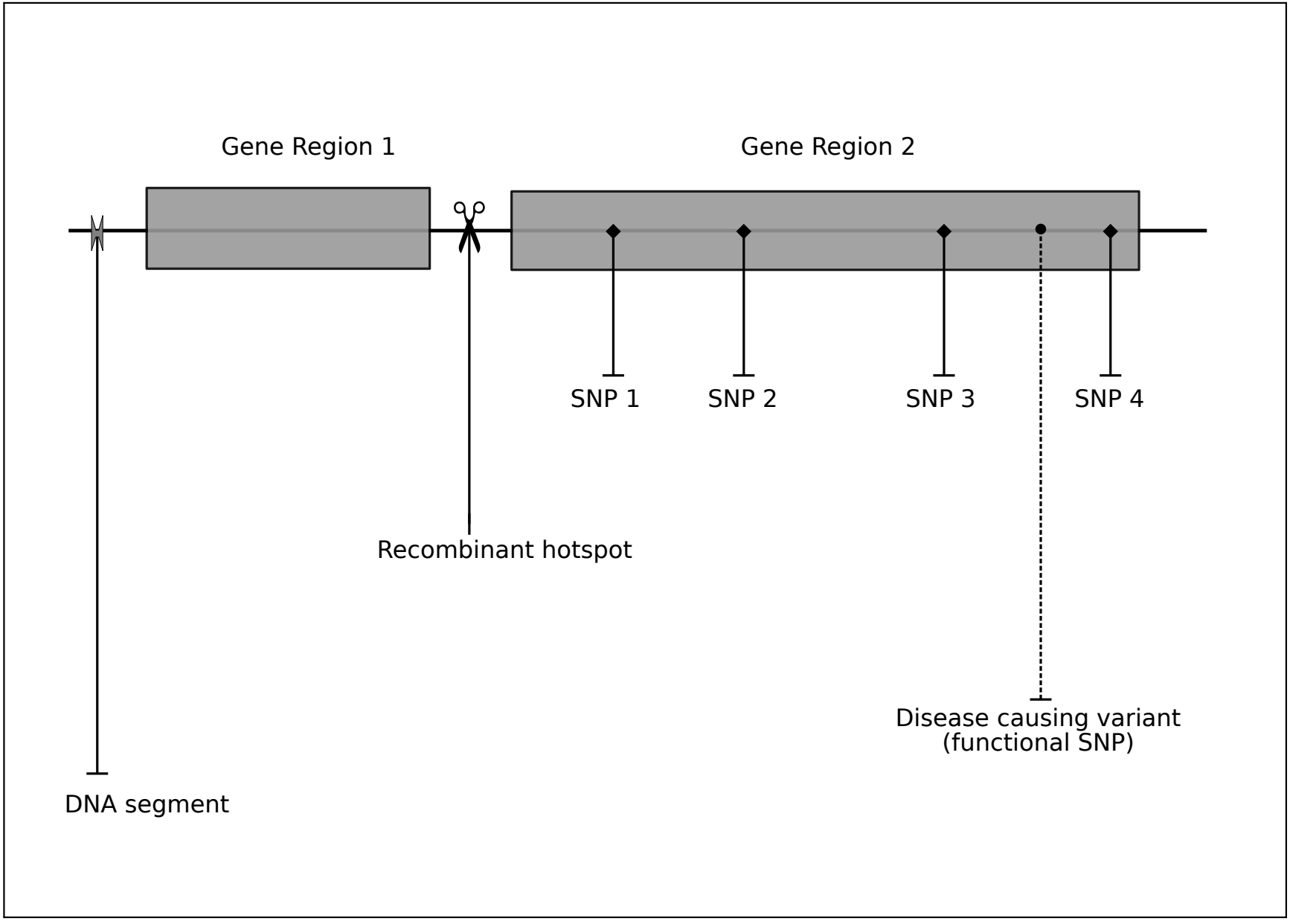
Definitions:

- **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.

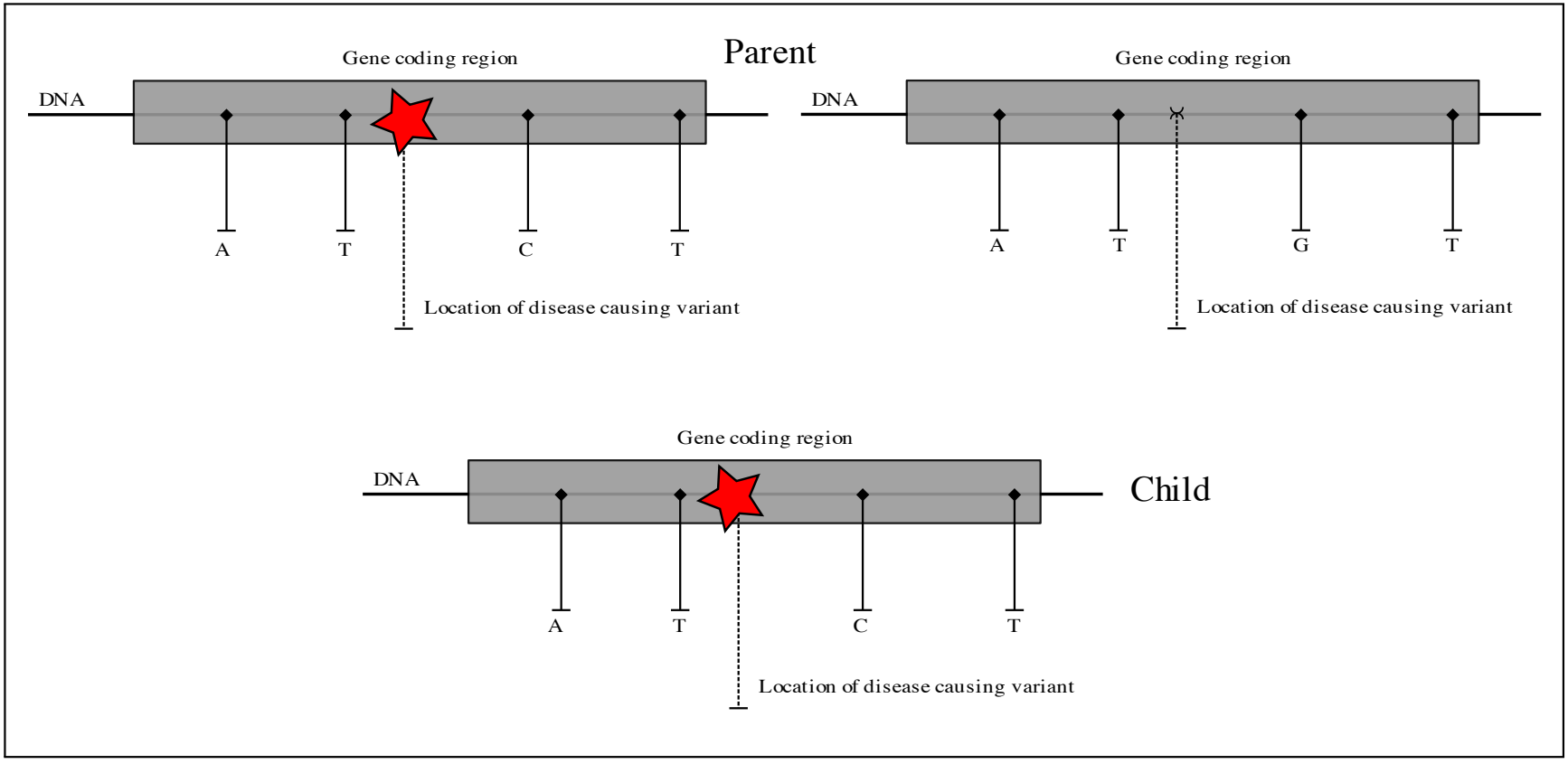
DNA Sequence Variation in a Gene Can Change the Protein Produced by the Genetic Code



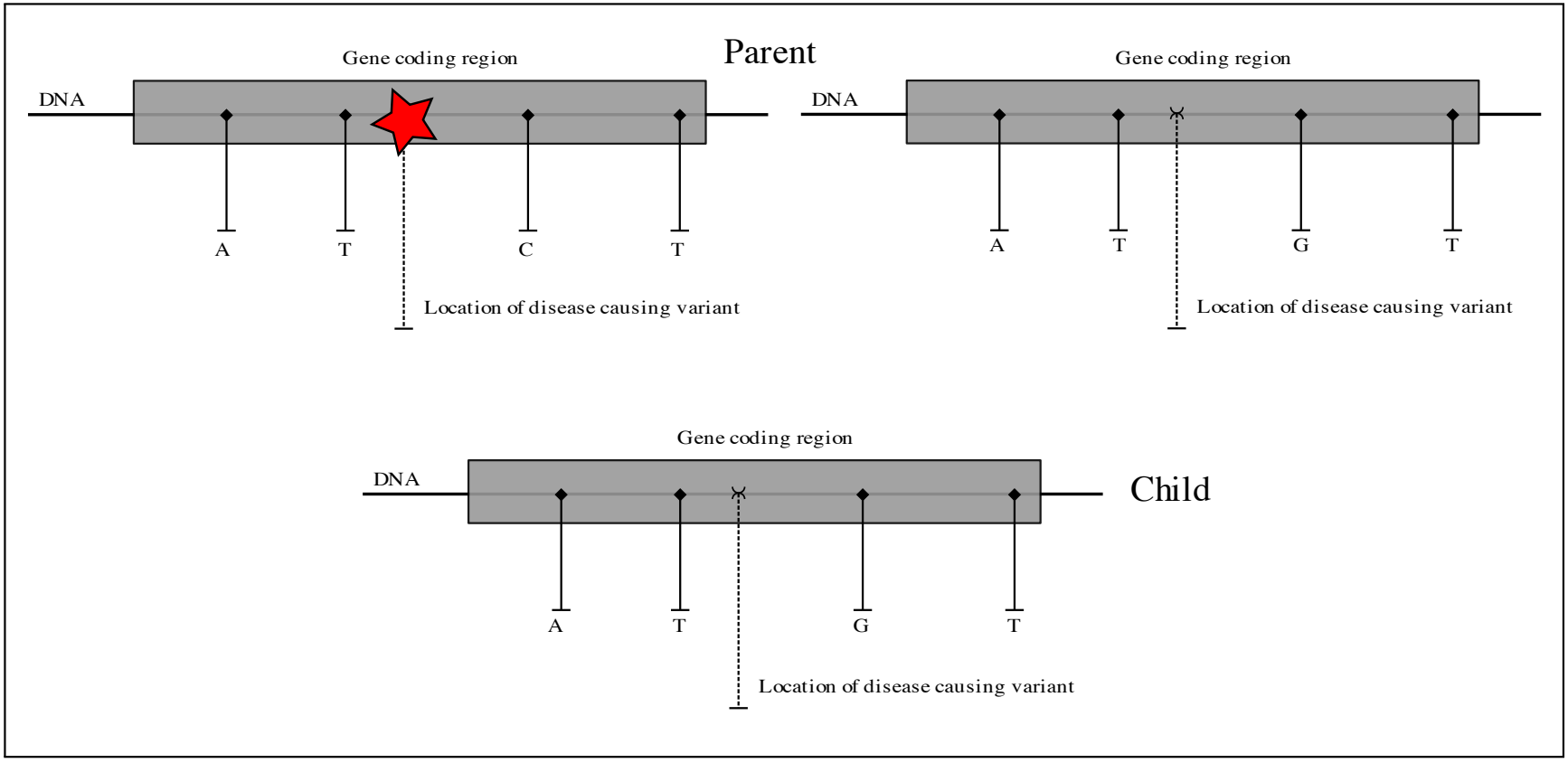
Data components and terminology



Data components and terminology



Data components and terminology



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 $\rightarrow$ 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:

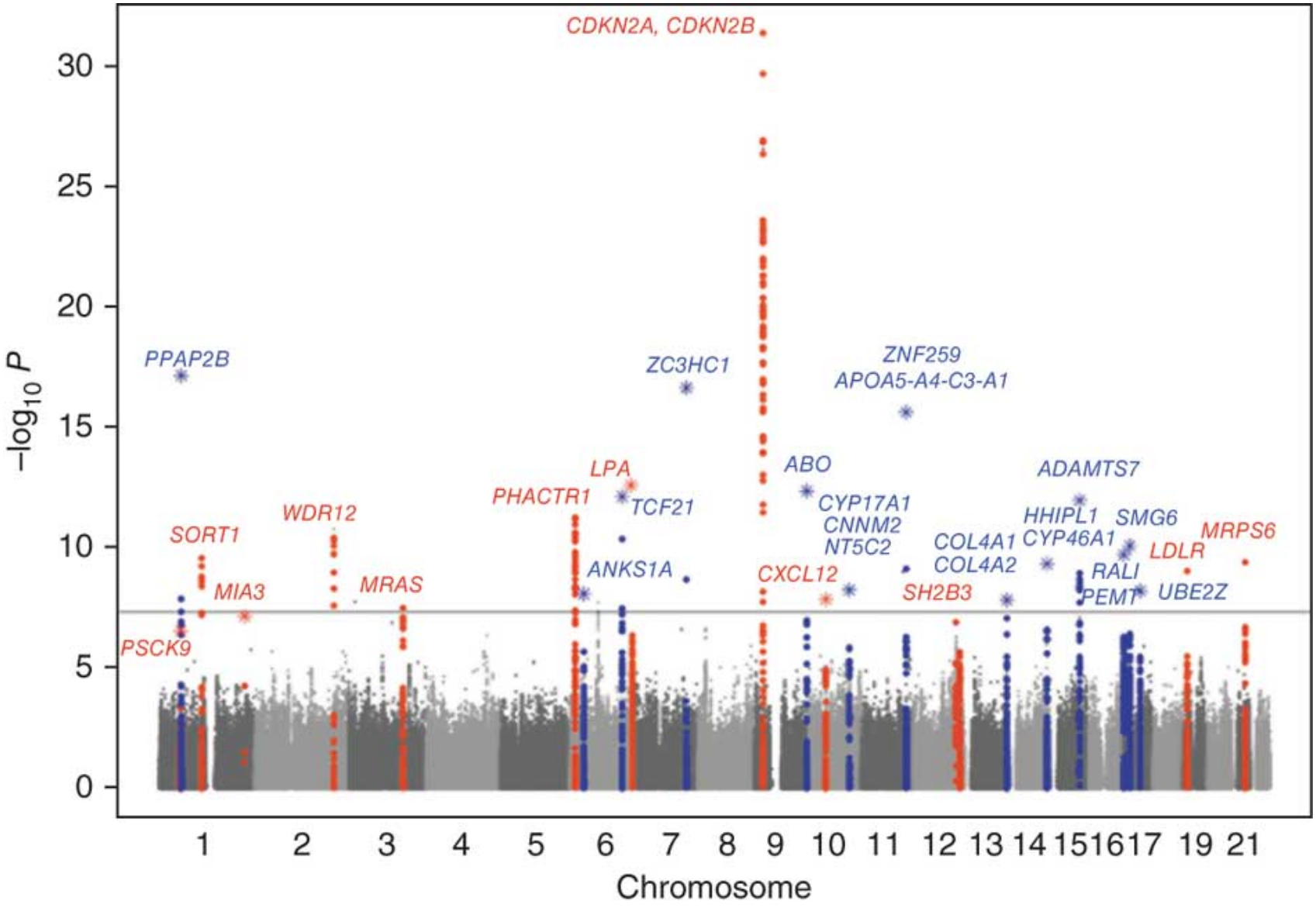
$$FWER = Pr(\text{reject at least one } H_0^k \text{ |all } H_0^k \text{ are true})$$

- Typically control at level $\alpha = 0.05$ using Bonferonni adjustment $\rightarrow$ 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 (assuming independence of SNPs):
   - generate $965,220$ p-values from a uniform distribution
   - determine value corresponding to the 5th percentile
   - repeat 500 times and record 5th percentile of this distn.
2. 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):

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

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

3. 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.