Statistical Associations

Katie Pollard

BMI 206

In this unit we will learn …

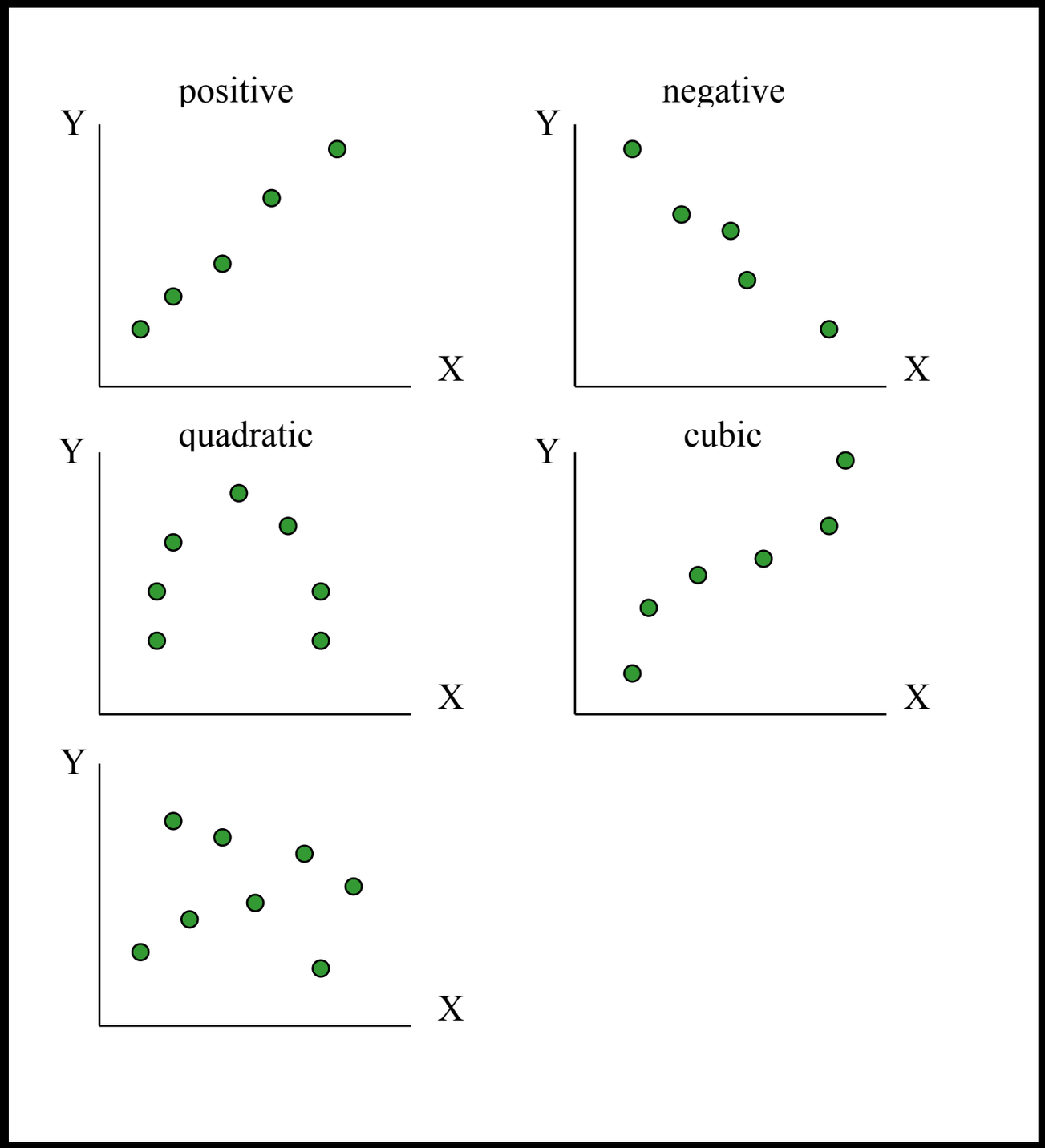
- Methods for relating variables of different kinds
- Measures of association for different kinds of data
- Mathematical definitions of independence and of statistical association
- Correlation versus enrichment

Relating Different Data Types

Covariate (independent variable)

Outcome (dependent variable)		Continuous or Both	Categorical
	Continuous	Linear Regression / ANCOVA	ANOVA
	Categorical	Generalized Linear Model Regression	Contingency Tables / Log-linear Model Regression

Relating Continuous Variables



Linear relationship

Non-linear relationship

No obvious relationship

Relating Categorical Variables

rs80265967	Disease	No disease
A	1	6721
C	2	2

Association

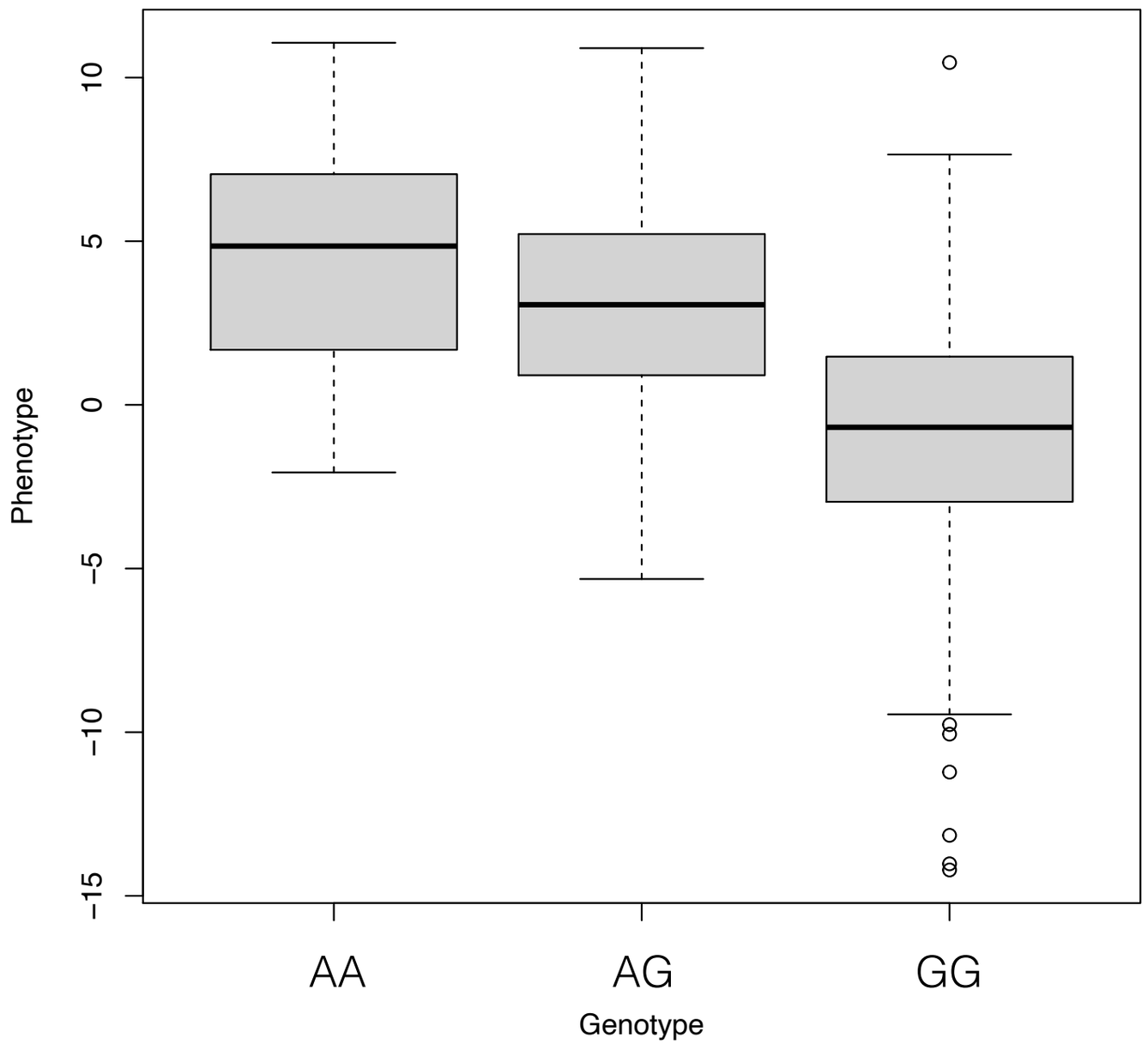
rs17880490	Disease	No disease
G	360	1981
A	2	11

No association

* joint = product of marginals

1000 Genomes Allele Frequencies, hypothetical disease

Relating Continuous and Categorical Variables



Different means

Association

Statistical association is <u>any</u> dependence between two random variables.

Dependence means that mathematically probabilistic independence is not satisfied.

Conditional Probabilities

Outcomes are independent if the conditional probability equals the marginal probability:

- Written P(A|B)=P(A). Equivalently, (B|A)=P(B).
- P(A and B) also written P(A,B) is the joint probability

Multiplicative Rule: P(A and B) = P(A|B) P(B)

- Rearranged is Bayes Rule: P(A|B) = P(A and B)/P(B)
- If A and B are independent, P(A and B) = P(A) P(B)

Enrichment

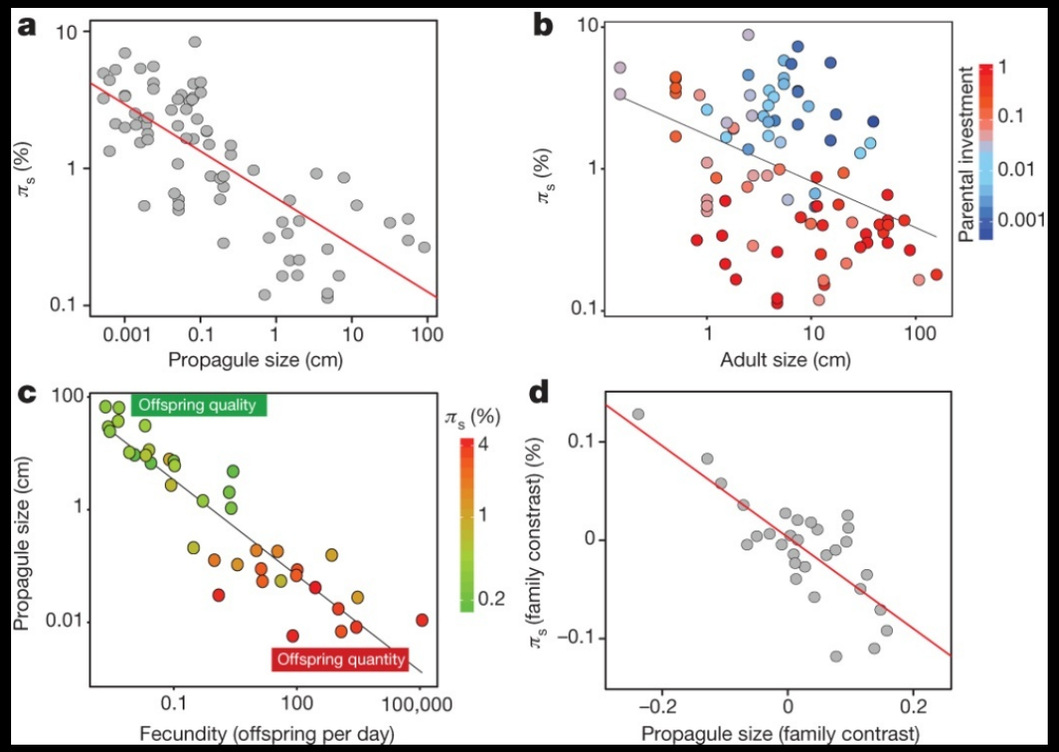
Quantifies excess overlap in sets versus expectation

- Refers to counts of observations in sets
- Not applicable to quantitative data
- Expectation is relative to a null distribution,

  e.g.,

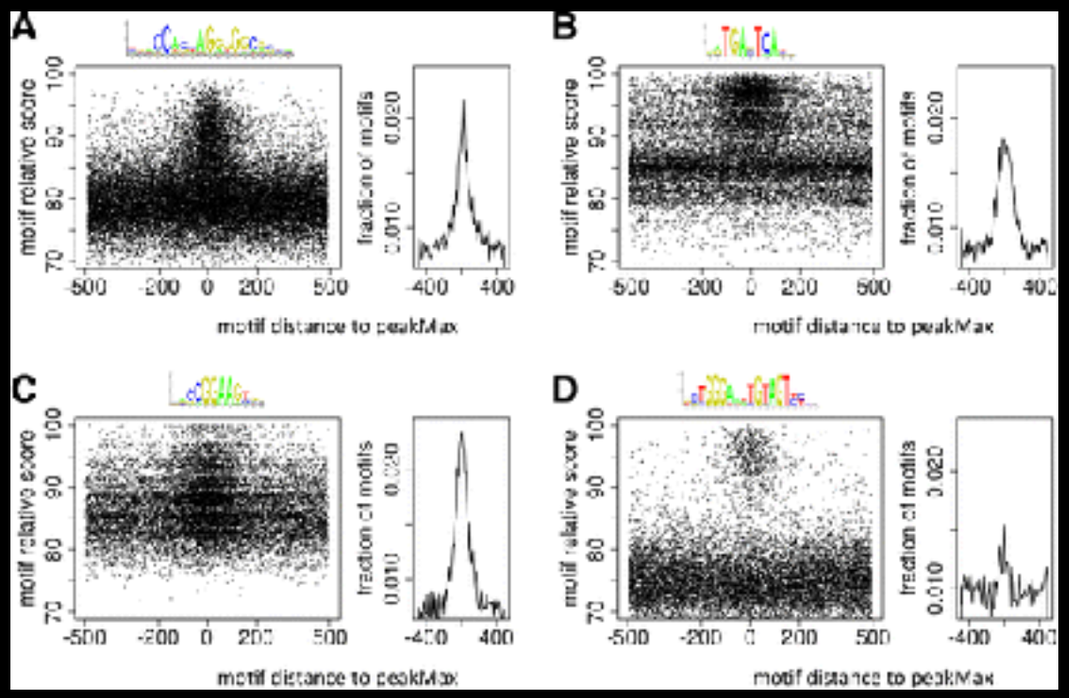
  - Independence
  - Background level of dependence

Correlation versus Enrichment in Bioinformatics



Life-history traits are correlated with population genetic diversity across animals

Romiguier et al. (2014) Nature



Zinc finger motifs are enriched in ChIP-seq peaks for non-zinc-finger transcription factors

Hunt & Wasserman (2014) Genome Biology

Points to Remember About Statistical Associations

- Association is more general than correlation, e.g.,
  - Odds ratio, relative risk and other measures of association for categorical data
  - Mutual information, dual total correlation, maximal information coefficient
- Association does NOT imply causality.
- Conditional association depends on other variables.