Day 2, PM Session

Genome-wide association studies (GWAS)

Richard Flamio Jr., Ph.D.

Madison Zimmerman

Kristina M. Ramstad, Ph.D.

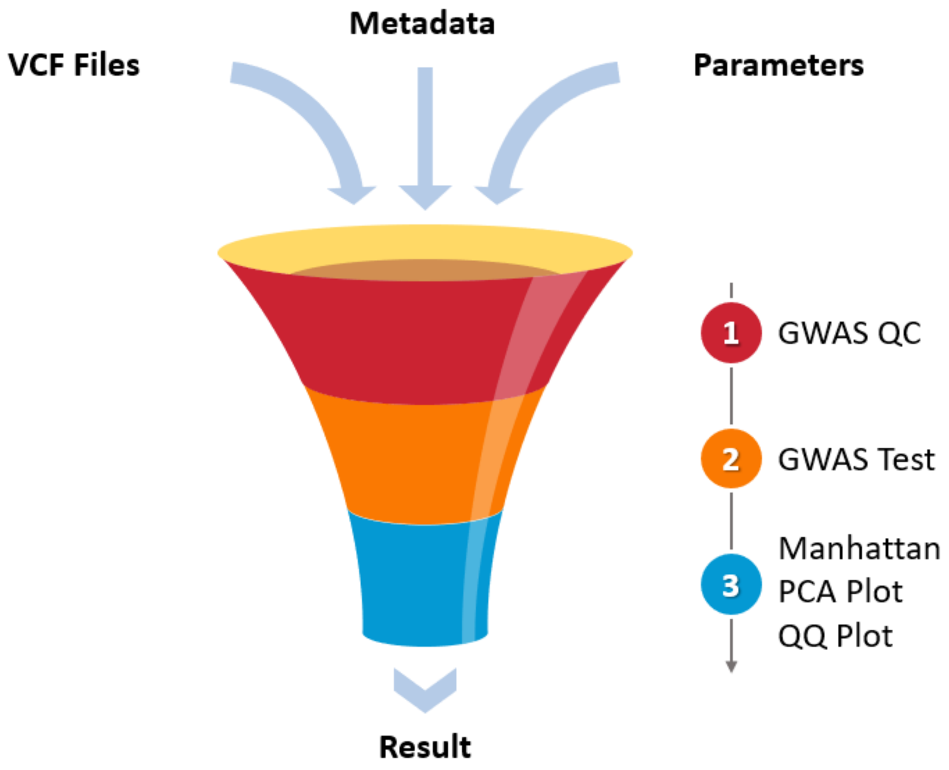
University of South Carolina Aiken

What is a GWAS?

- Tests the association between large and dense single nucleotide polymorphism (SNP) datasets and phenotypes of interest

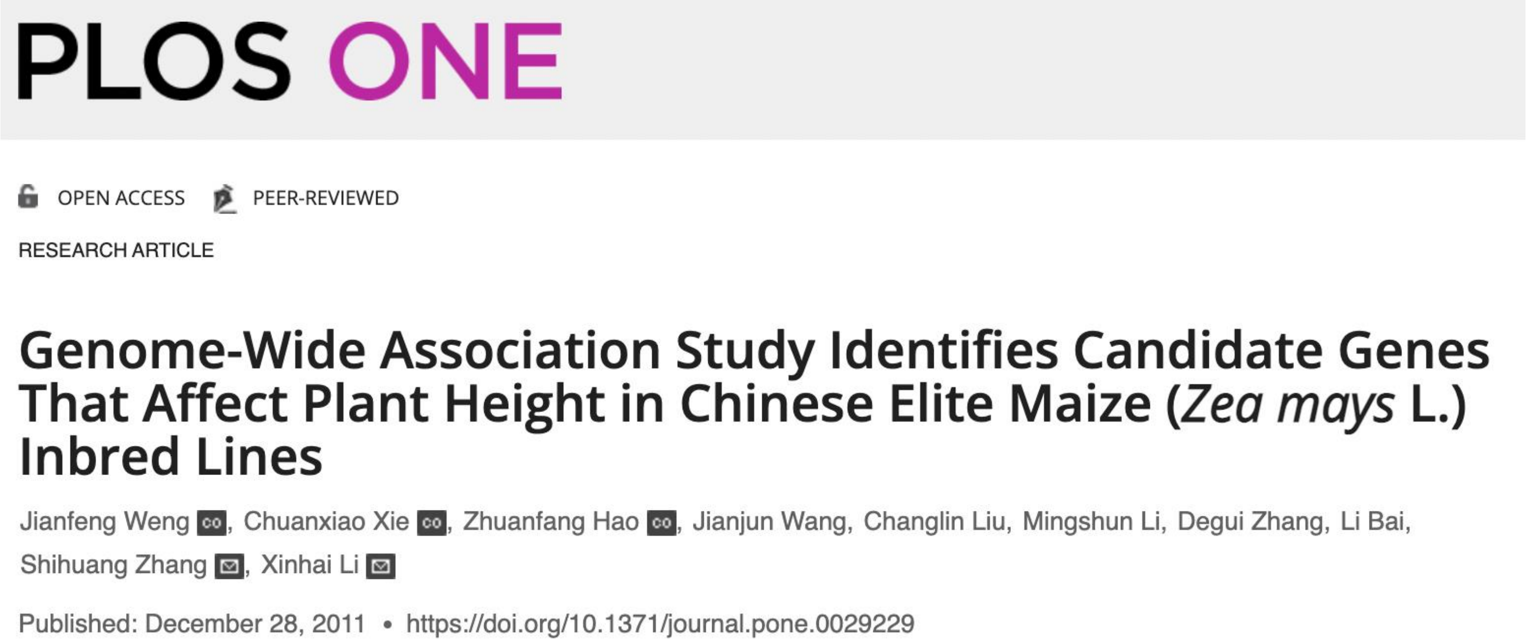
- Allele substitution effect = the effect the presence of a copy of an allele has on a phenotype

- Phenotypes can be continuous, binary, or categorical



https://shivom.gitbook.io/documentation/pipelines/gwas-pipeline

Continuous (numbers)



PLOS ONE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

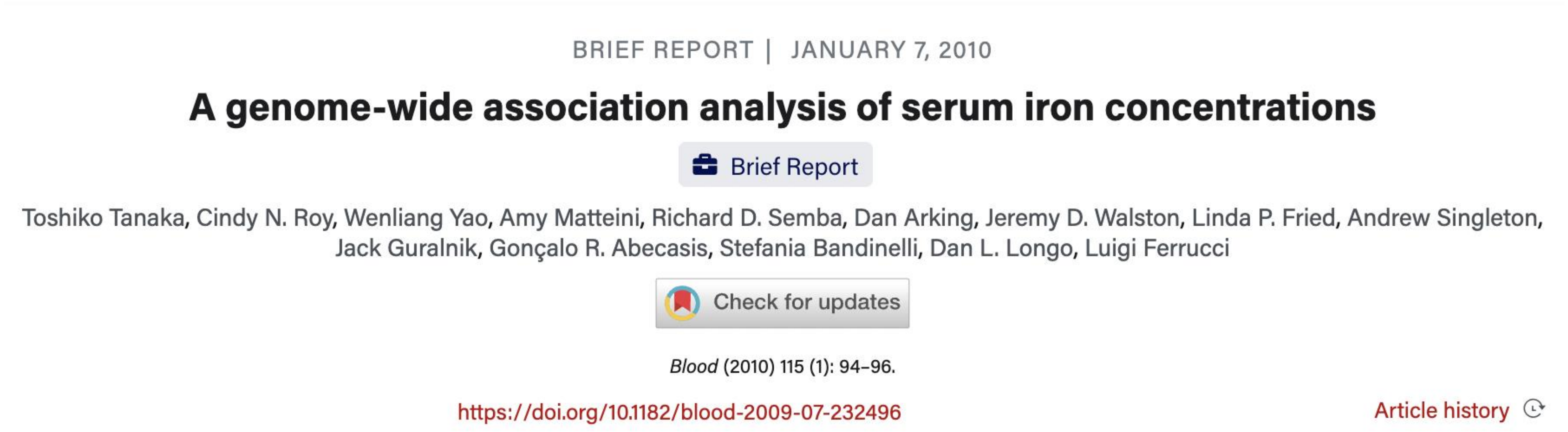
Genome-Wide Association Study Identifies Candidate Genes That Affect Plant Height in Chinese Elite Maize (*Zea mays* L.) Inbred Lines

Jianfeng Weng, Chuanxiao Xie, Zhuanfang Hao, Jianjun Wang, Changlin Liu, Mingshun Li, Degui Zhang, Li Bai, Shihuang Zhang, Xinhai Li

Published: December 28, 2011 • https://doi.org/10.1371/journal.pone.0029229

Main takeaway:
GWAS identified SNPs in genes that affected auxin and other hormones

Continuous



BRIEF REPORT | JANUARY 7, 2010

A genome-wide association analysis of serum iron concentrations

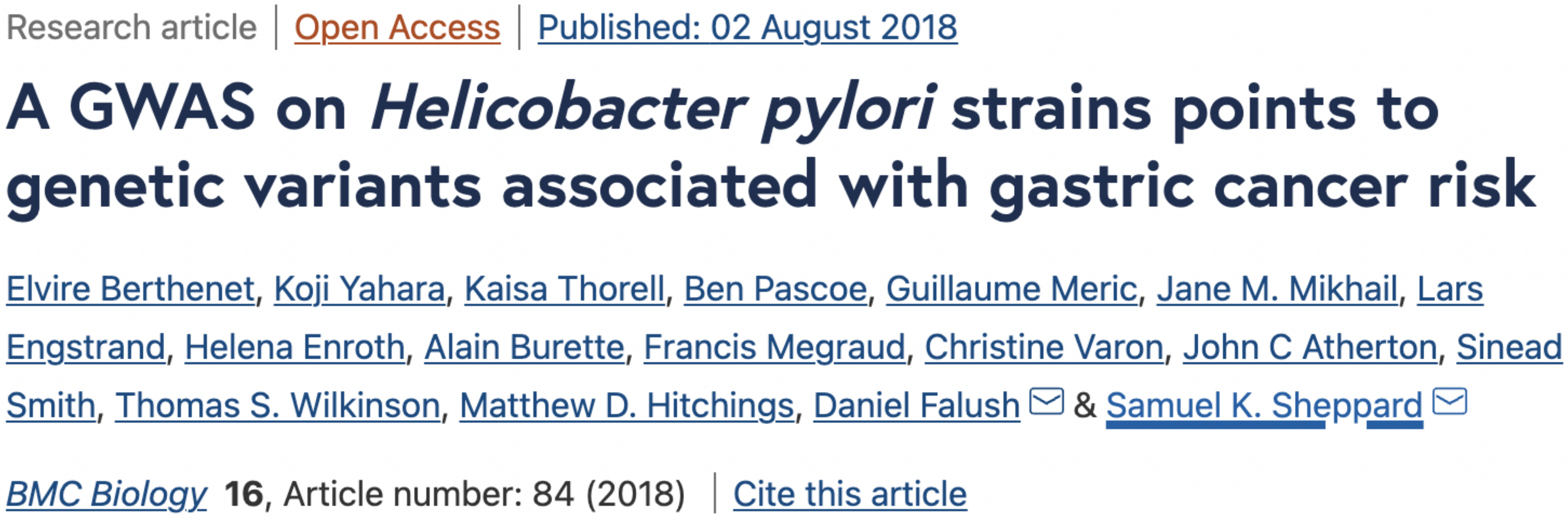
Toshiko Tanaka, Cindy N. Roy, Wenliang Yao, Amy Matteini, Richard D. Semba, Dan Arking, Jeremy D. Walston, Linda P. Fried, Andrew Singleton, Jack Guralnik, Gonçalo R. Abecasis, Stefania Bandinelli, Dan L. Longo, Luigi Ferrucci

Blood (2010) 115 (1): 94–96.

https://doi.org/10.1182/blood-2009-07-232496

Main takeaway: The researchers found a SNP in the TMPRSS6 gene which codes for an enzyme that promotes iron absorption.

Binary (presence/absence)



Research article | Open Access | Published: 02 August 2018

A GWAS on *Helicobacter pylori* strains points to genetic variants associated with gastric cancer risk

Elvire Berthenet, Koji Yahara, Kaisa Thorell, Ben Pascoe, Guillaume Meric, Jane M. Mikhail, Lars Engstrand, Helena Enroth, Alain Burette, Francis Megraud, Christine Varon, John C Atherton, Sinead Smith, Thomas S. Wilkinson, Matthew D. Hitchings, Daniel Falush ✉ & Samuel K. Sheppard ✉

BMC Biology **16**, Article number: 84 (2018) | Cite this article

Main takeaway: The presence of this bacteria is associated with the presence of certain genes (e.g. babA).

Binary

Red versus pink salmon in non-glaciated versus glaciated water bodies

Migratory versus resident wood storks as alternate life history strategies

Categorical

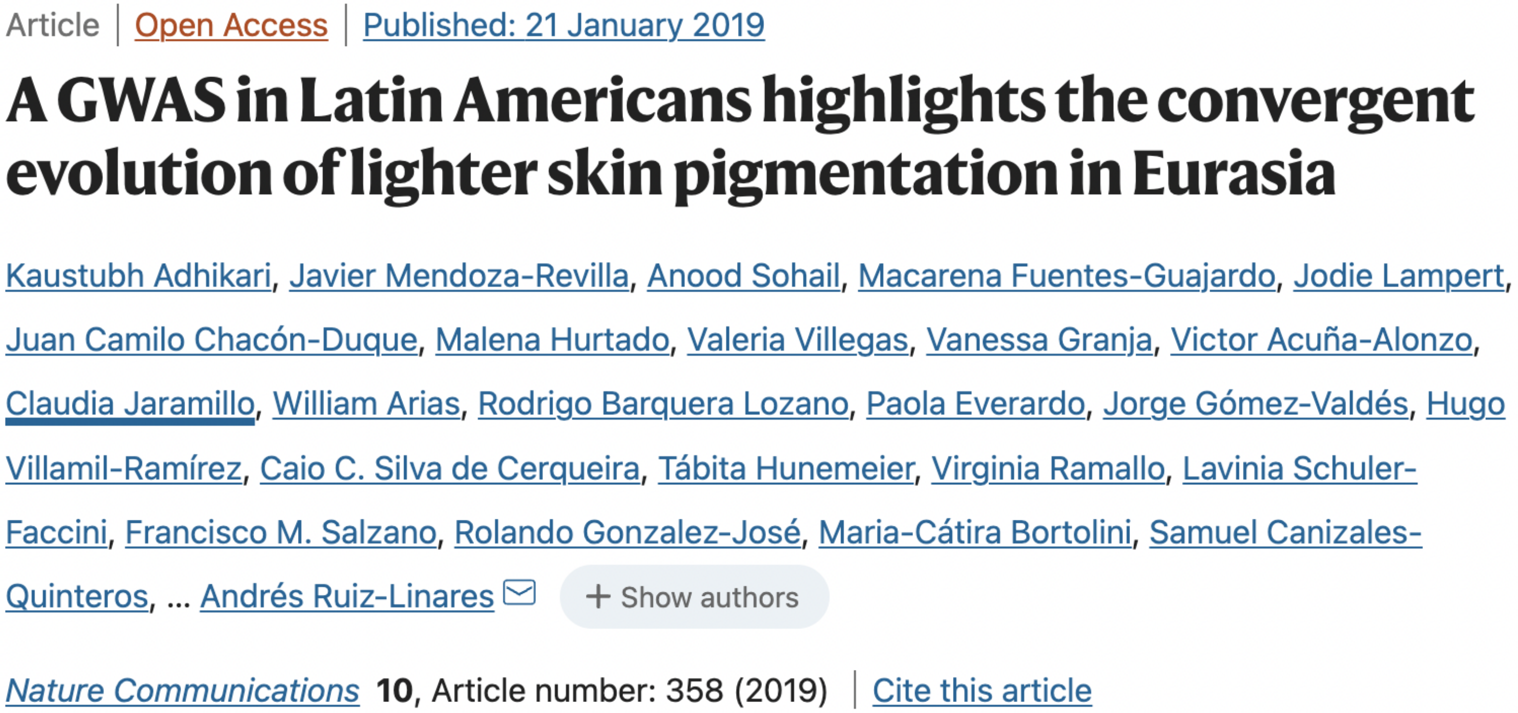
- Unordered (breed, shape, color, blood type, etc.)

- Ordered (pain scale, litter size, birth order, etc.)

https://money.com/what-dog-breed-is-best-for-me/

https://womenties.blog/tag/scales/

Unordered



Article | Open Access | Published: 21 January 2019

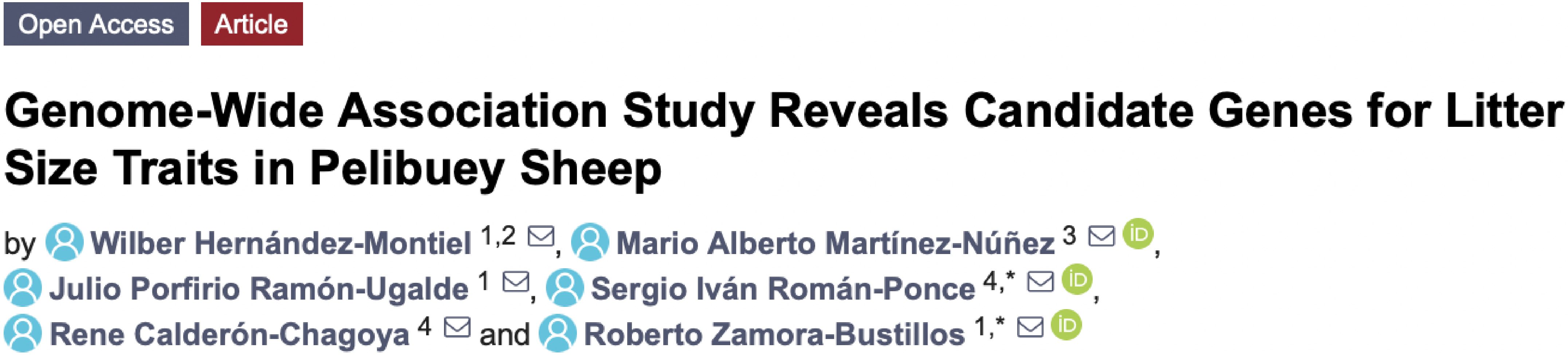
A GWAS in Latin Americans highlights the convergent evolution of lighter skin pigmentation in Eurasia

Kaustubh Adhikari, Javier Mendoza-Revilla, Anood Sohail, Macarena Fuentes-Guajardo, Jodie Lampert, Juan Camilo Chacón-Duque, Malena Hurtado, Valeria Villegas, Vanessa Granja, Victor Acuña-Alonzo, Claudia Jaramillo, William Arias, Rodrigo Barquera Lozano, Paola Everardo, Jorge Gómez-Valdés, Hugo Villamil-Ramírez, Caio C. Silva de Cerqueira, Tábita Hunemeier, Virginia Ramallo, Lavinia Schuler-Faccini, Francisco M. Salzano, Rolando Gonzalez-José, Maria-Cátira Bortolini, Samuel Canizales-Quinteros, … Andrés Ruiz-Linares ✉

Nature Communications **10**, Article number: 358 (2019) | Cite this article

Main takeaway: The study identified novel genes associated with skin and eye pigmentation.

Ordered



Open Access Article

Genome-Wide Association Study Reveals Candidate Genes for Litter Size Traits in Pelibuey Sheep

by Wilber Hernández-Montiel [1,2], Mario Alberto Martínez-Núñez [3], Julio Porfirio Ramón-Ugalde [1], Sergio Iván Román-Ponce [4,*], Rene Calderón-Chagoya [4] and Roberto Zamora-Bustillos [1,*]

Main takeaway: Many SNPs affected this trait, including SNPs within genes affecting maternal body weight.

What type of data is the following?

1. Blood type
2. Fish tail shape
3. Hair color
4. Water turbidity
5. Birth order
6. Happiness level
7. Dog breed
8. Pain scale
9. Litter size
10. Cancer presence

GWAS Pre-processing

- Phenotype data
  - Is this a representative sample?
  - Are there any outliers that need to be removed?
  - Detecting covariables

- Genotype data
  - Minor allele frequency
  - Genotyping error
  - Hardy-Weinberg Equilibrium
  - % missing data in individuals and markers

Continuous assumptions

- Normally distributed residuals
  - Check for this by using histograms and QQ-plots
- Homogeneity of variance
  - Check for this by using a scatterplot

Covariables

- Examples: sex, age, population, etc.
- Detect and correct covariables using ANOVA or linear mixed models
- Can add covariables into regression models
- For population stratification
    - Genomic control
        - Look to see if lambda is > 1
            - Lambda = 1 --> no population stratification
            - Lambda > 1 --> population stratification
    - Principal Component Analysis (PCA)
    - Kinship matrix

Imputation

- The process of replacing missing data with substituted values
- Many GWAS programs do not allow missing genotypes

- Methods:
  - Mean substitution - use the mean of the SNP
  - K-Nearest Neighbor Imputation (KNNI) - uses a distance function
  - Pedigree and kinship
  - Linkage disequilibrium and haplotypes

Regressions depending on type of data

- Continuous --> use linear regression
- Binary --> use logistic regression
- Unordered categorical --> use multinomial logistic regression
- Ordered categorical --> use ordered logistic regression

Linear Regression

$$Y = B0 + B1X$$

- Tests if B1 is significantly different than 0
- If this is the case --> SNP is associated with phenotype

Binary Logistic Regression

- Two classes (cases/controls, e.g. presence/absence)

- If sample size is low --> can use Fisher's exact test

Odds Ratio

- Degree of association between two variables
  - Phenotype and SNP
- Can range from 0 to beyond
- Odds Ratio = 1 --> no association
- Odds Ratio > 1 --> positive association between alternate allele and phenotype
- Odds Ratio < 1 --> negative association between alternate allele and phenotype

Multinomial Logistic Regression

- More than two classes and the categories are unordered

Ordered logistic regression

- Based on cumulative probabilities
- Different intercept for each category

What type of regression would you perform on the following data?

1. Hair length
2. Shoe size
3. Dental cavity presence
4. Blood type
5. Vision presence/absence
6. T-cell count
7. Fly eye diameter
8. Number of branches
9. Pain scale
10. Dog hair curliness

False positives versus false negatives

- False positives are when something is significant when it should not be
  - Example: you find an incorrect positive association between a SNP and a phenotype

- False negatives are when something is not significant when it should be
  - Example: you do not find an association between a SNP and a phenotype but that SNP does affect the phenotype in reality

Multiple Comparison Testing

- Danger of many false positives
- Example: 5000 tests and alpha = 0.05 --> how many false positives?

- Increase sample size
- Reduce number of comparisons
  - Linkage disequilibrium
  - Genomic segments of interest
- Decrease significance threshold
  - Bonferroni correction
  - False Discovery Rate (FDR; Benjamini and Hochberg)

Bonferroni Correction

- The new significance threshold is modeled by:
  - Original significance threshold/number of tests or markers
  - Example: original alpha = 0.01 and 3000 tests

- Very conservative
  - Few false positives
  - Many false negatives

What is new significance threshold for a dataset of 30,000 variants if you use a Bonferroni correction on an initial alpha = 0.05?

False Discovery Rate (FDR)

- Mathematically intensive (rely on software to calculate this)
- What it does...
  - Sets a threshold for what percent of the data is allowed to be false positive
- Less conservative than Bonferroni
  - Fewer false negatives

Additive allele effect model

Dominance genotype effect model

Additive effect model

- Allelelic test: does the frequency of allele A among cases equal the frequency of allele A among controls
- Armitage Trend test: the probability of the phenotype increases/decreases as the SNP increases

Visualization

- Manhattan Plots
- QQ-Plots

Quantile-Quantile Plots (QQ-plots)

- Detect systematic problems

Manhattan Plot

- X-axis = Chromosome
- Y-axis = negative log of p-value
- Usually two significance thresholds (example: $p < 0.05$ and $p < 0.01$)

Questions?

20 minute break

Personal dataset exercise:

1. What type of data are each of your variables?
2. What is your trait of interest?
3. What are some covariables you need to account for in your own study?
4. What type of regression would you perform on your data?
5. If there are markers in your dataset, how many?
    1. What would be an adjusted p-value if you used Bonferroni correction with an alpha value of 0.01?

Personal Dataset Presentations

References

- grunwaldlab.github
- https://bookdown.org/kdonovan125/ibis_data_analysis_r4/documenting-your-results-with-r-markdown.html