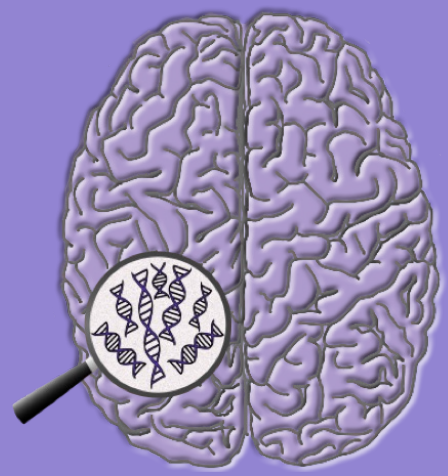
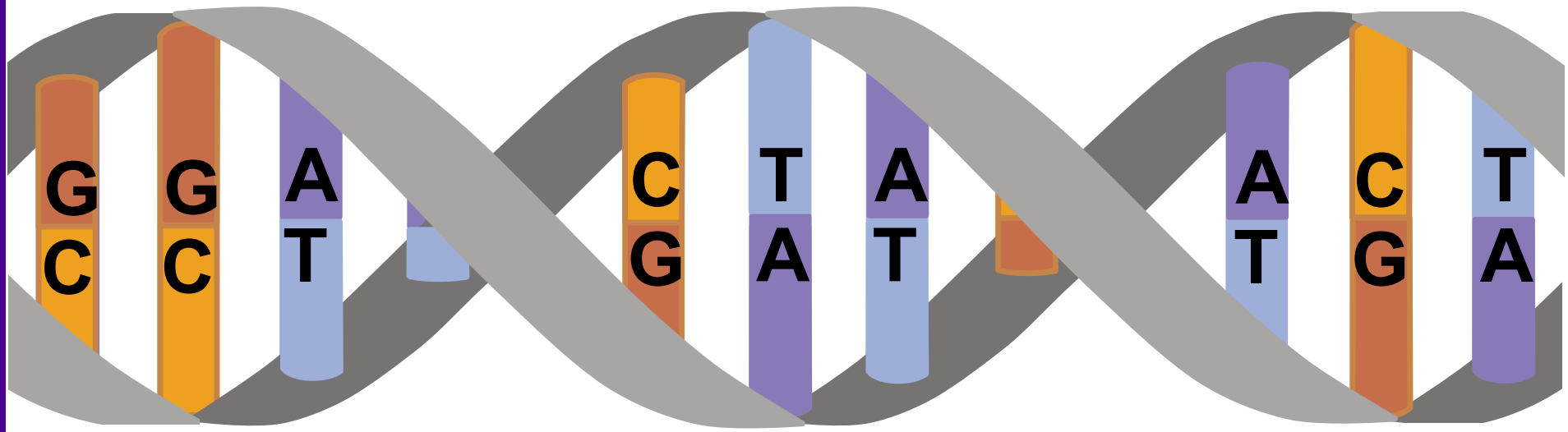
Single Nucleotide Polymorphisms (SNP) analysis

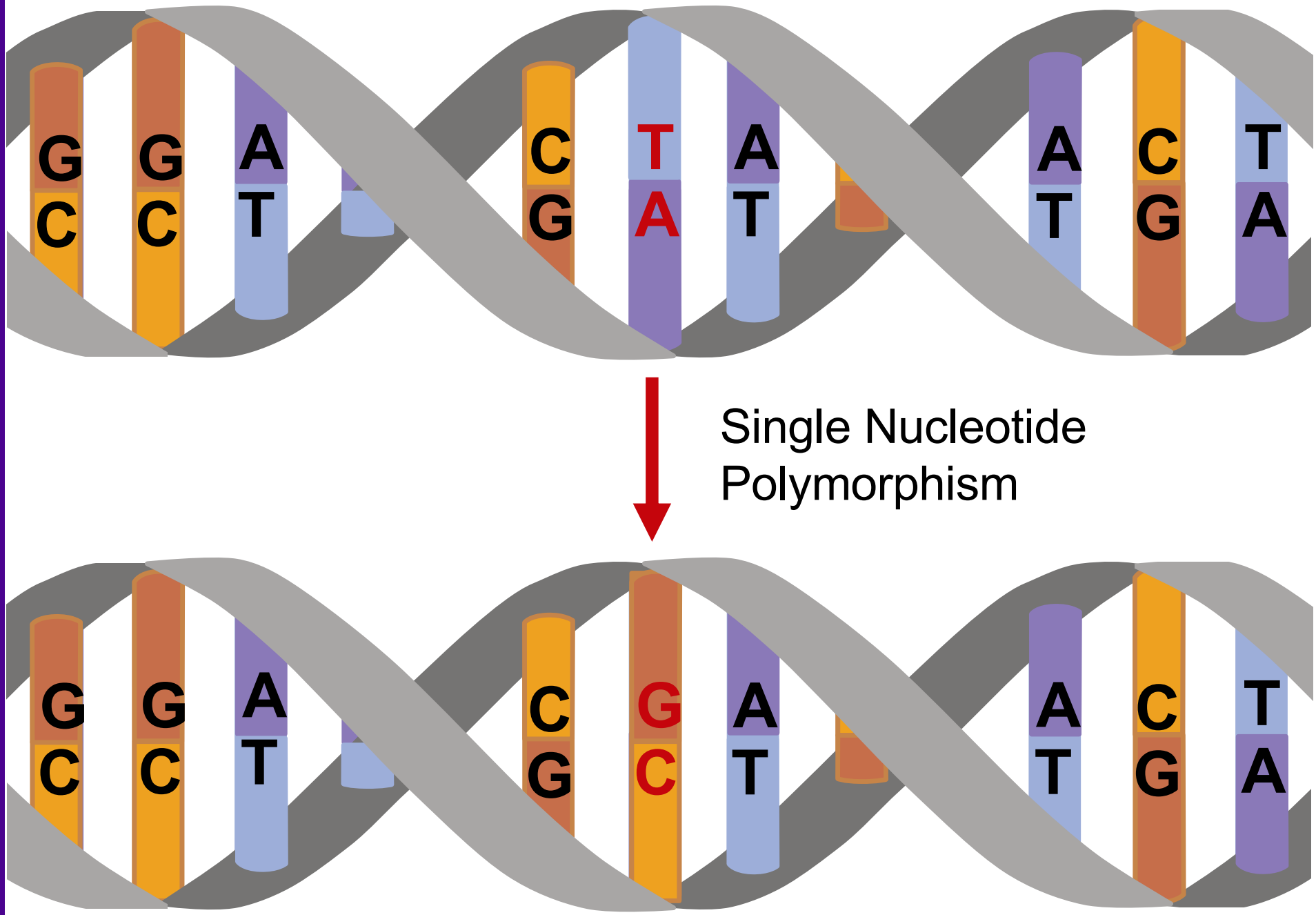
Yuetiva Robles, PhD

Basics of Genetic Analysis: SNP analysis and GWAS

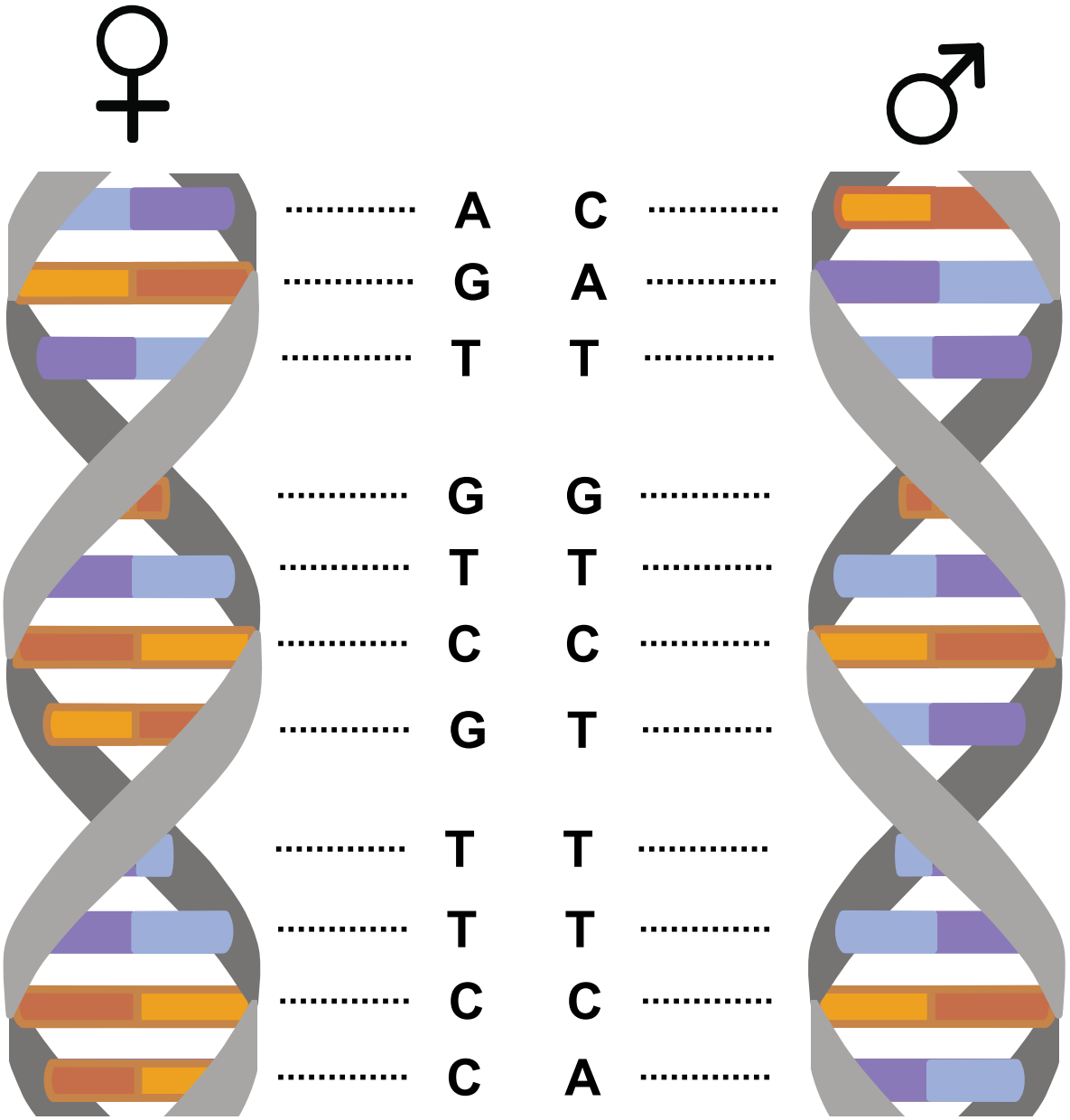
July 25, 2025




G
C
G
C
A
T
C
G
T
A
T
A
A
T
C
G
T
A

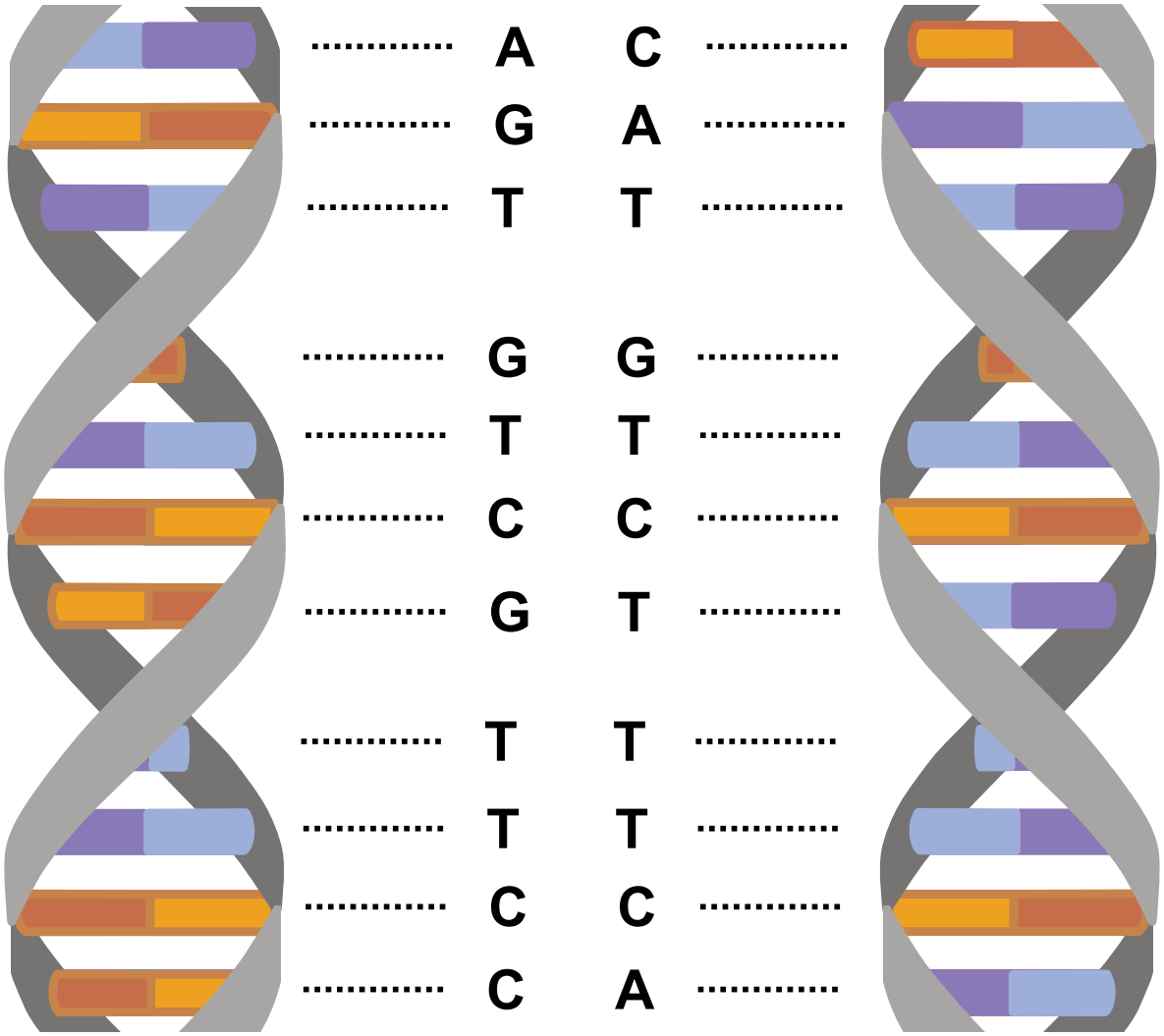

G C
G C
A T
C G
T A
A T
A T
C G
T A
Single Nucleotide Polymorphism
G C
G C
A T
C G
G C
A T
A T
C G
T A

Genotype: Allele pairs



*flipped strand to mirror alleles

Genotype: Allele pairs



		minor allele	#minor alleles
A	C	C	1
G	A	G	1
T	T	G	0
G	G	G	2
T	T	T	2
C	C	A	0
G	T	G	1
T	T	T	2
T	T	G	0
C	C	C	2
C	A	A	1

*flipped strand to mirror alleles

SNP Analysis

- Usually logistic or linear regression
- SNP usually coded as an additive 0, 1, 2 based on number of minor alleles a person has

Outcome = Intercept + SNP + covariates

SNP regression results - example

CSF $A\beta_{42}$ = Intercept + rs429358_C + age_lp + sex + Summary_Diagnosis

(aka APOE ε4)

```
Residuals:
    Min      1Q  Median      3Q     Max
-744.07 -232.79  -31.43  211.44  925.44

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             676.957    189.692   3.569 0.000404 ***
rs429358_C             -130.776     28.540  -4.582 6.24e-06 ***
age_lp                   -1.654      2.379  -0.695 0.487261
genderMale              -18.938     35.069  -0.540 0.589505
Summary_DiagnosisMCI-AD 183.143     82.101   2.231 0.026280 *
Summary_DiagnosisNormal 378.629     67.587   5.602 4.05e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 330.4 on 383 degrees of freedom
  (25 observations deleted due to missingness)
Multiple R-squared:  0.2019,    Adjusted R-squared:  0.1915
F-statistic: 19.38 on 5 and 383 DF,  p-value: < 2.2e-16
```

SNP regression results - example

CSF $A\beta_{42}$ = Intercept + rs429358_C + age_lp + sex + Summary_Diagnosis

(aka APOE ε4)

Note the allele here
results = effect of # C alleles

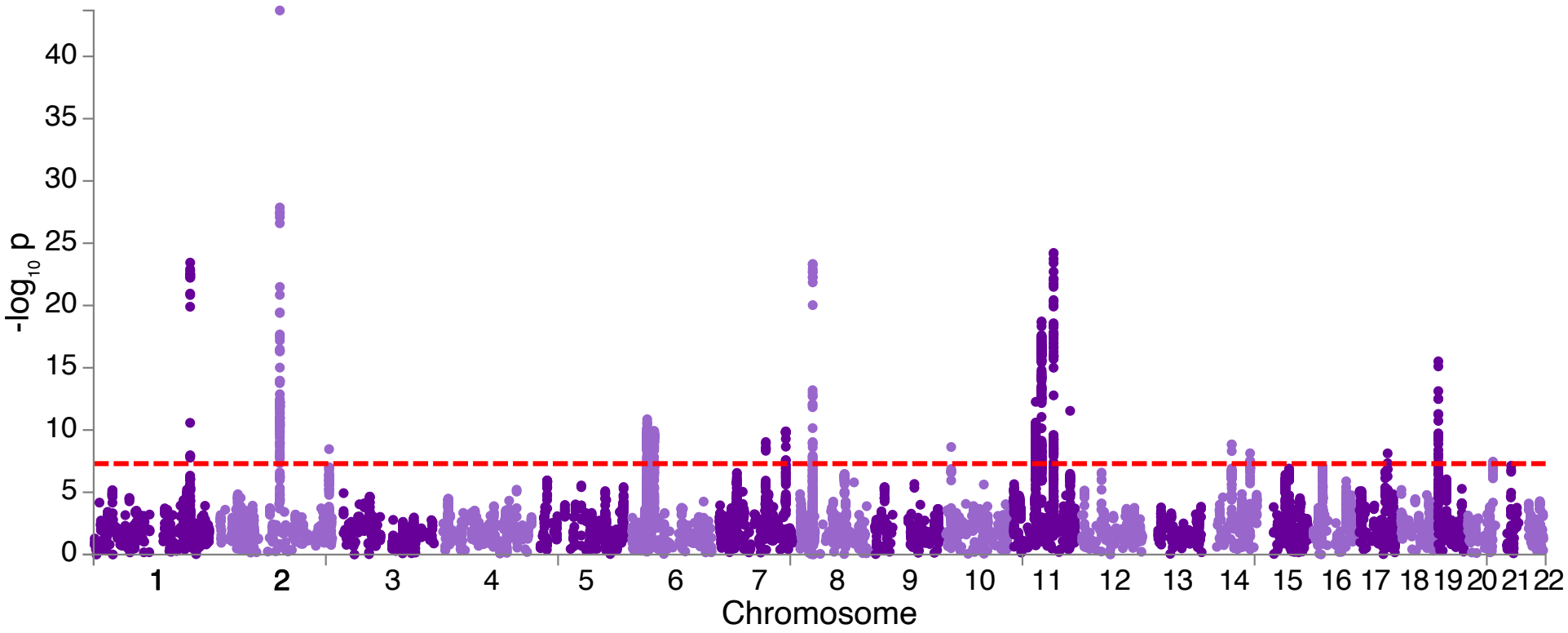
```
Residuals:
    Min      1Q  Median      3Q     Max
-744.07 -232.79  -31.43  211.44  925.44

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             676.957    189.692   3.569 0.000404 ***
rs429358_C             -130.776     28.540  -4.582 6.24e-06 ***
age_lp                   -1.654      2.379  -0.695 0.487261
genderMale              -18.938     35.069  -0.540 0.589505
Summary_DiagnosisMCI-AD 183.143     82.101   2.231 0.026280 *
Summary_DiagnosisNormal 378.629     67.587   5.602 4.05e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 330.4 on 383 degrees of freedom
  (25 observations deleted due to missingness)
Multiple R-squared:  0.2019,    Adjusted R-squared:  0.1915
F-statistic: 19.38 on 5 and 383 DF,  p-value: < 2.2e-16
```

Genome-Wide Association Studies (GWAS)

GWAS – multiple testing

- Individual SNP tests repeated hundreds of thousands to millions
- Traditional Bonferroni threshold 5×10^{-8} based on assumption of 1 million independent tests
- Alternatives available including FDR, permutation, etc

Manhattan plot created from GWAS published by Kunkle et al, 2019



- Results from meta-analysis of logistic regressions (AD case/control defined by clinical status)
- Each dot represents a SNP, x-axis genomic location, y-axis -log10 p
- Dotted line is the genome-wide significance threshold (5×10^{-8})