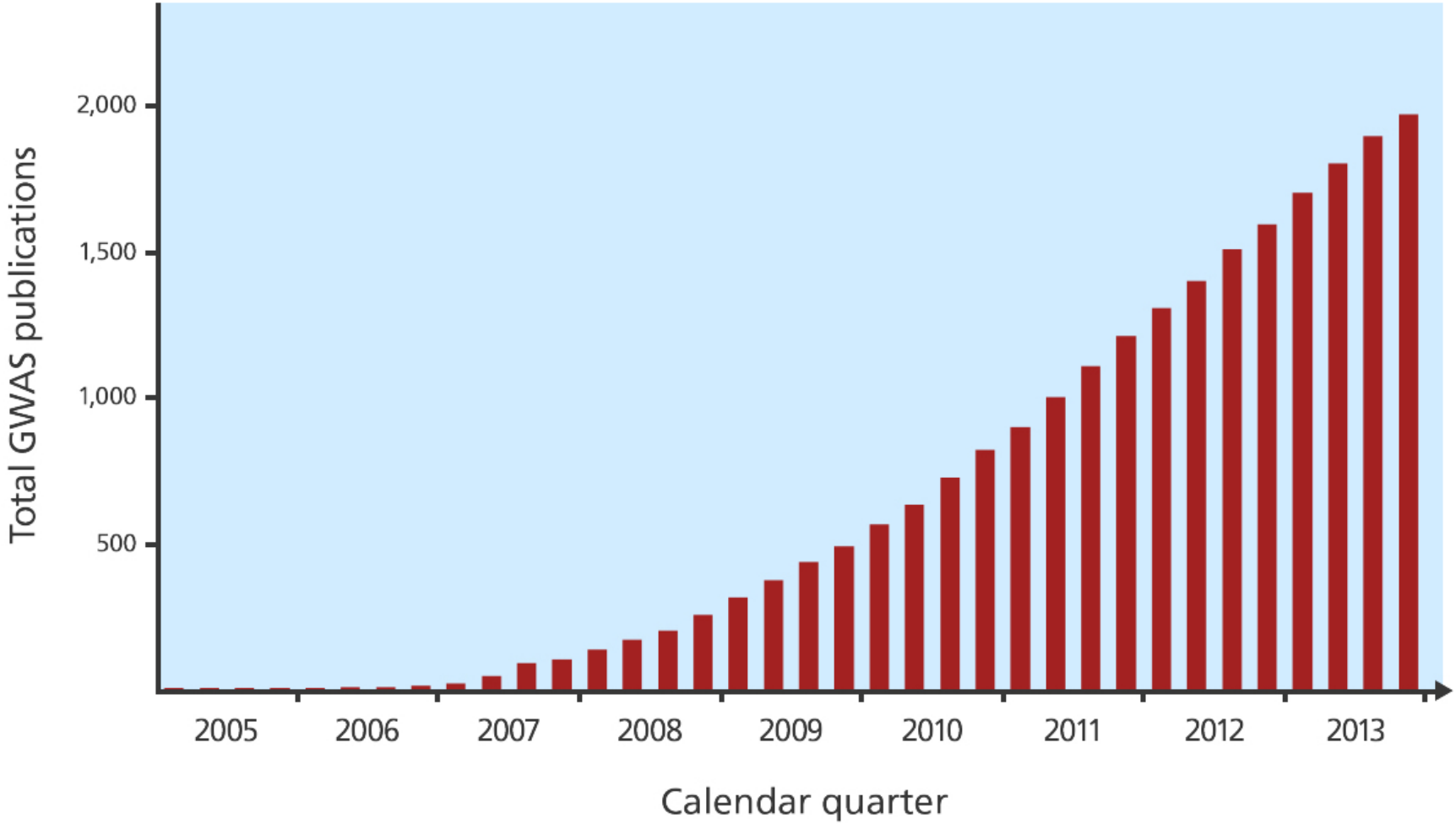
STAT021 Statistical Methods II

Lecture 24 Genome-Wide Association Studies

Lu Chen
Swarthmore College
12/11/2018

Genome-Wide Association Studies (GWAS)



Genome Research Limited

Outline

- Genome-wide association study (GWAS)
- Single-nucleotide polymorphism (SNP)
- Illustration example
  - Logistic regression model
  - inference, odds ratio and predictive accuracy
- Multiple testing and family-wise type I error
- Manhattan plot
- Final exam

Genome-Wide Association Studies (GWAS)

PLOS | COMPUTATIONAL BIOLOGY

A central goal of human genetics is to identify genetic risk factors for common, complex diseases such as schizophrenia and type II diabetes, and for rare Mendelian diseases such as cystic fibrosis and sickle cell anemia. There are many different technologies, study designs and analytical tools for identifying genetic risk factors.

Bush, W. S., Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, 8(12), e1002822.

Genome-Wide Association Studies (GWAS)

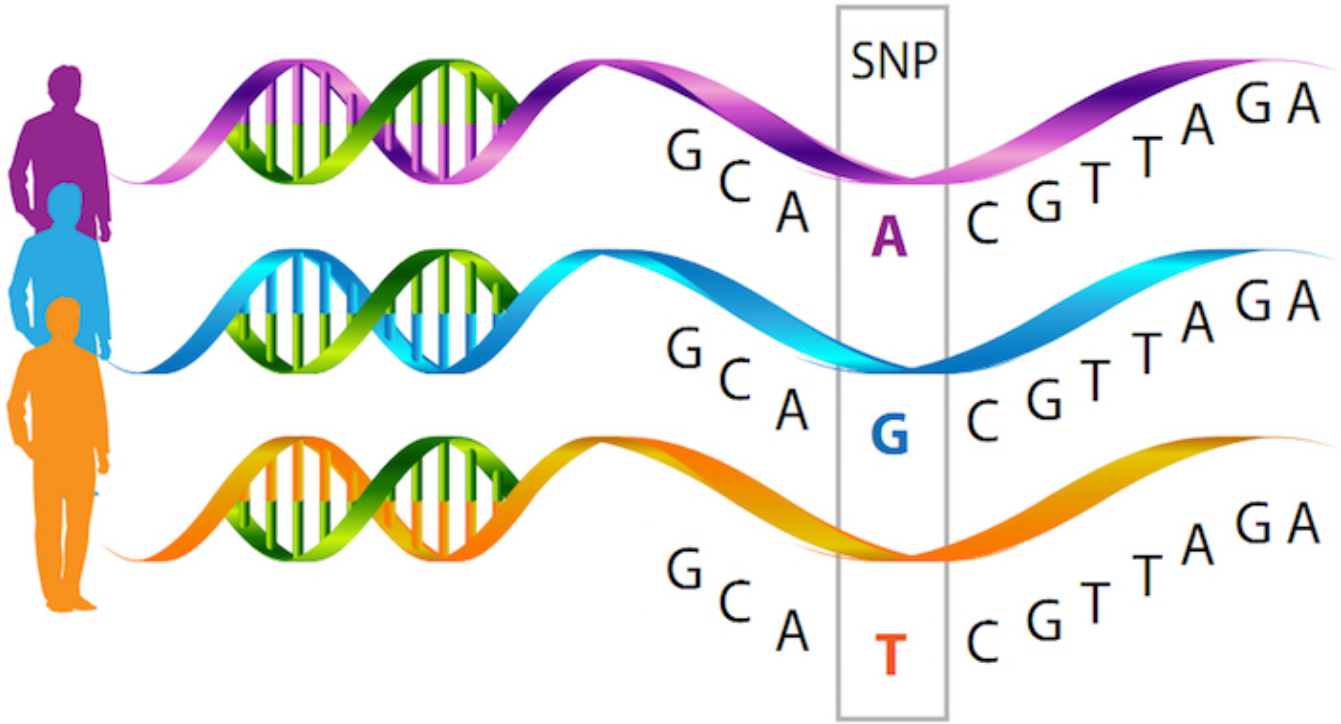
PLOS | COMPUTATIONAL BIOLOGY

We will focus here on the genome-wide association study or GWAS that measures and analyzes DNA sequence variations from across the human genome in an effort to identify genetic risk factors for diseases that are common in the population.

Bush, W. S., Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, 8(12), e1002822.

Single-nucleotide polymorphism (SNP)

GWAS typically uses **single-nucleotide polymorphism (SNP)** as the source of genetic variation. SNPs are single base-pair changes in the DNA sequence.

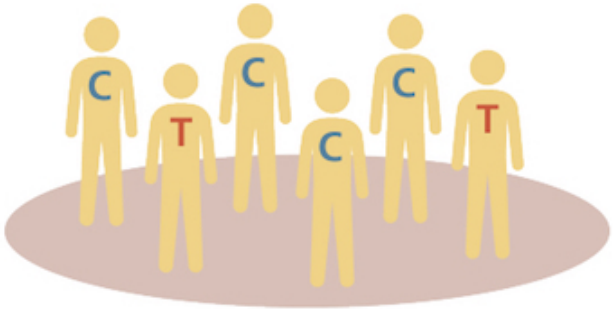


Single-nucleotide polymorphism (SNP)

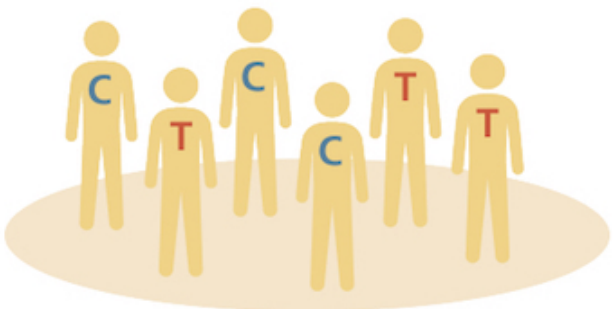
In human genome, there are

- 23×2 chromosomes
- $\sim 20,000$ genes
  1.5% of the genome
- $\sim 3,100,000,000$ base pairs
- $\sim 1,400,000$ SNPs
- The International HapMap Project focuses only on common SNPs, which are present in at least 1% of the population. Currently, the HapMap project lists 887,450 SNPs.
- A typical GWAS tests 1,000,000 SNPs.

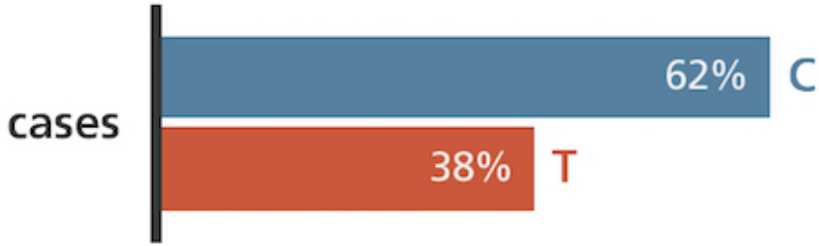
GWAS: Case-control study



cases (n=1,000)
people with heart disease



controls (n=1,000)
people without heart disease



Genome Research Limited

GWAS: Case-control study

To study the relationship between a disease and a SNP, we denote

- $D = 1$ for a subject with the disease (case) and 0 for a healthy subject (control).
- $G = 1$ for genotype C and 0 for T.
- The data

	$G = 1$ (C)	$G = 0$ (T)	Total
Cases $D = 1$	620	380	**1000**
Controls $D = 0$	490	510	**1000**

- What's the logistic regression model for this relationship?

$$\log \frac{\pi}{1-\pi} = \beta_0 + \beta_1 G, \text{ where } \pi = P(D = 1 \mid G)$$

```
head(DG)
```

```
##   Disease Genotype
## 1       1        C
## 2       0        C
## 3       0        T
## 4       0        C
## 5       1        T
## 6       0        T
```

```
dim(DG)
```

```
## [1] 2000    2
```

```
table(DG)
```

```
##          Genotype
## Disease    T   C
##       0  510 490
##       1  380 620
```

GWAS: Logistic regression model

```
summary(m1 <- glm(Disease ~ Genotype, family="binomial", data=DG))
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.29424    0.06777  -4.342 1.41e-05 ***
## GenotypeC    0.52955    0.09081   5.832 5.49e-09 ***
```

```
exp(confint(m1))
```

```
##                 2.5 %    97.5 %
## (Intercept) 0.6521113 0.8506117
## GenotypeC   1.4217714 2.0297771
```

- $b_0 = -0.29, b_1 = 0.53$

$$\log \frac{\hat{\pi}}{1-\hat{\pi}} = -0.29 + 0.53 \times G$$

- Estimated odds ratio $= e^{b_1} = e^{0.53} = 1.70$. The odds of getting the disease for subjects with genotype C is 1.70 times higher than subjects with genotype T.

GWAS: Logistic regression model

```
summary(m1 <- glm(Disease ~ Genotype, family="binomial", data=DG))
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.29424    0.06777  -4.342 1.41e-05 ***
## GenotypeC    0.52955    0.09081   5.832 5.49e-09 ***
```

```
exp(confint(m1))
```

```
##                 2.5 %    97.5 %
## (Intercept) 0.6521113 0.8506117
## GenotypeC   1.4217714 2.0297771
```

- $z = 5.83, P = 5.49 \times 10^{-9} < 0.05$. There is a significant relationship between the diesease and the genotype.
- 95% confidence interval for the odds ratio: $[1.43, 2.03]$. We are 95% confident (about the method) that the interval $[1.43, 2.03]$ will contain the true population odds ratio.

GWAS: Logistic regression model

$$\log \frac{\hat{\pi}}{1-\hat{\pi}} = -0.29 + 0.53 \times G$$

	$G = 1$ (C)	$G = 0$ (T)	Total
Cases $D = 1$	620	380	**1000**
Controls $D = 0$	490	510	**1000**

Estimated probabilities and OR

- $G = 0$ (T), $\hat{\pi}_T = \frac{e^{-0.29}}{1+e^{-0.29}} = 0.43$
- $G = 1$ (C), $\hat{\pi}_C = \frac{e^{-0.29+0.53}}{1+e^{-0.29+0.53}} = 0.56$
- OR $= e^{b_1} = e^{0.53} = 1.70$

$$\frac{Odds_C}{Odds_T} = \frac{\hat{\pi}_C/(1-\hat{\pi}_C)}{\hat{\pi}_T/(1-\hat{\pi}_T)} = 1.70$$

Empirical prabability and OR

- $\hat{p}_T = \frac{380}{380+510} = 0.43$
- $\hat{p}_C = \frac{620}{620+490} = 0.56.$
- $$\text{OR} = \frac{Odds_C}{Odds_T} = \frac{\hat{p}_C/(1-\hat{p}_C)}{\hat{p}_T/(1-\hat{p}_T)} = \frac{620 \times 510}{490 \times 380} = 1.70$$

GWAS: Logistic regression model

- Logistic regression model with a binary/categorical predictor estimate the probabilities and odds ratio exactly the same as the emprical probabilities and odds ratio.
  - The estimated/emprical odds ratio is computed as the ratio of the products of the diagonal elements of a 2 by 2 table.

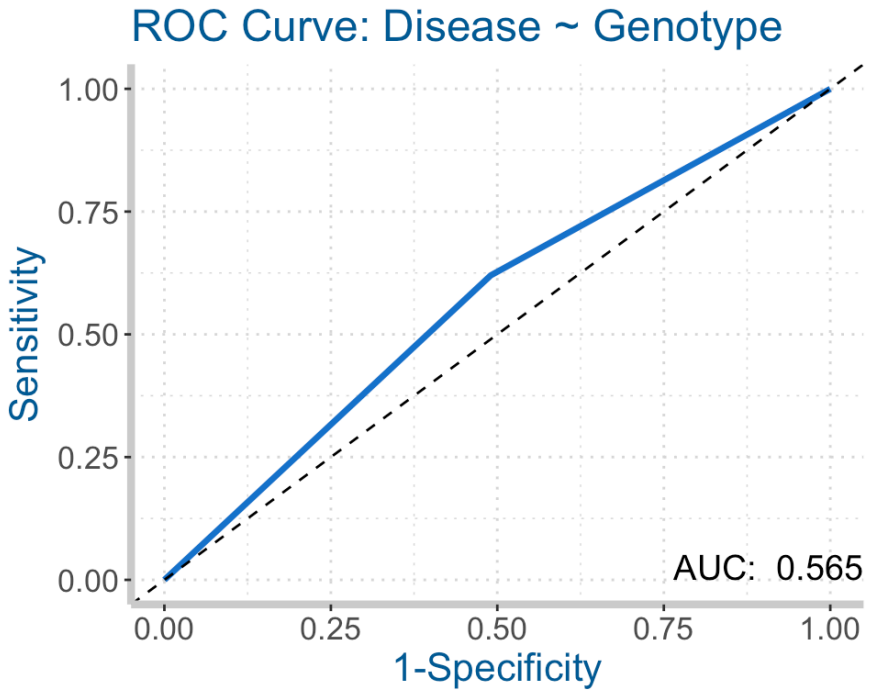
```
library(lmtest)
lrtest(m1)
```

```
## Likelihood ratio test
##
## Model 1: Disease ~ Genotype
## Model 2: Disease ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   2 -1369.1
## 2   1 -1386.3 -1 34.317  4.683e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The likelihood ratio test has $\chi^2 = 34.3$ and $P = 4.7 \times 10^{-9} < 0.05$. This model with *Genotype* is highly significant in explaining *Disease*.

GWAS: Logistic regression model

Predictive accuracy



$c = 0.5$	$\widehat{Disease} = 1$	$\widehat{Disease} = 0$
$Disease = 1$	620	380
$Disease = 0$	490	510

- Set cutoff to be any value between 0.43 and 0.56, say 0.5.
- $Sensitivity = \frac{620}{620+380} = 0.62$
- $Specificity = \frac{510}{510+490} = 0.51$
- $AUC = 0.565$

Multiple testing and family-wise type I error

- $10^6 = 1,000,000$ SNPs $\Rightarrow 10^6 = 1,000,000$ tests
- Type I error rate for each test: $\alpha = 0.05$
- Family-wise type I error rate: $1 - 0.95^{1,000,000} = 1 \Rightarrow$ inflated type I error
- In GWAS, **Bonferroni correction** is commonly used.
- Significance level for each test: $0.05/10^6 = 5 \times 10^{-8}$
- A SNP with P-value $< 5 \times 10^{-8}$ is significant.
- Usually, GWAS needs very large sample size to achieve such a small P-value. This is one of the reasons why GWAS is used for **common diseases**. A GWAS for rare disease requires an even larger sample size, which is usually not feasible.

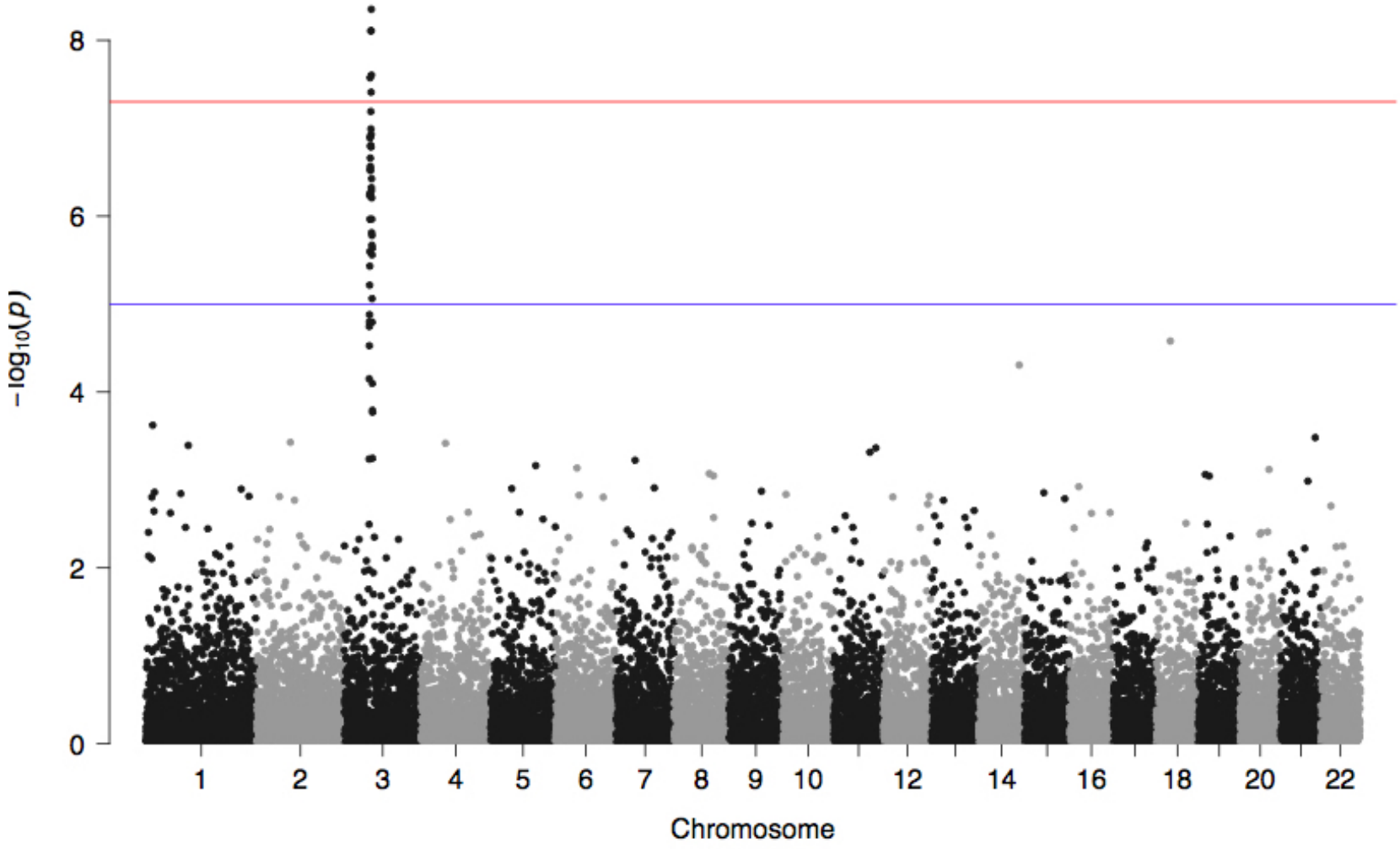
GWAS: Manhattan plot

Report the test results for all the 1,000,000 SNPs?

> A **Manhattan plot** is a type of scatterplot, usually used to display data with a **large number of data-points** - many of non-zero amplitude, and with a distribution of **higher-magnitude values**, for instance in genome-wide association studies (GWAS). - Wikipedia

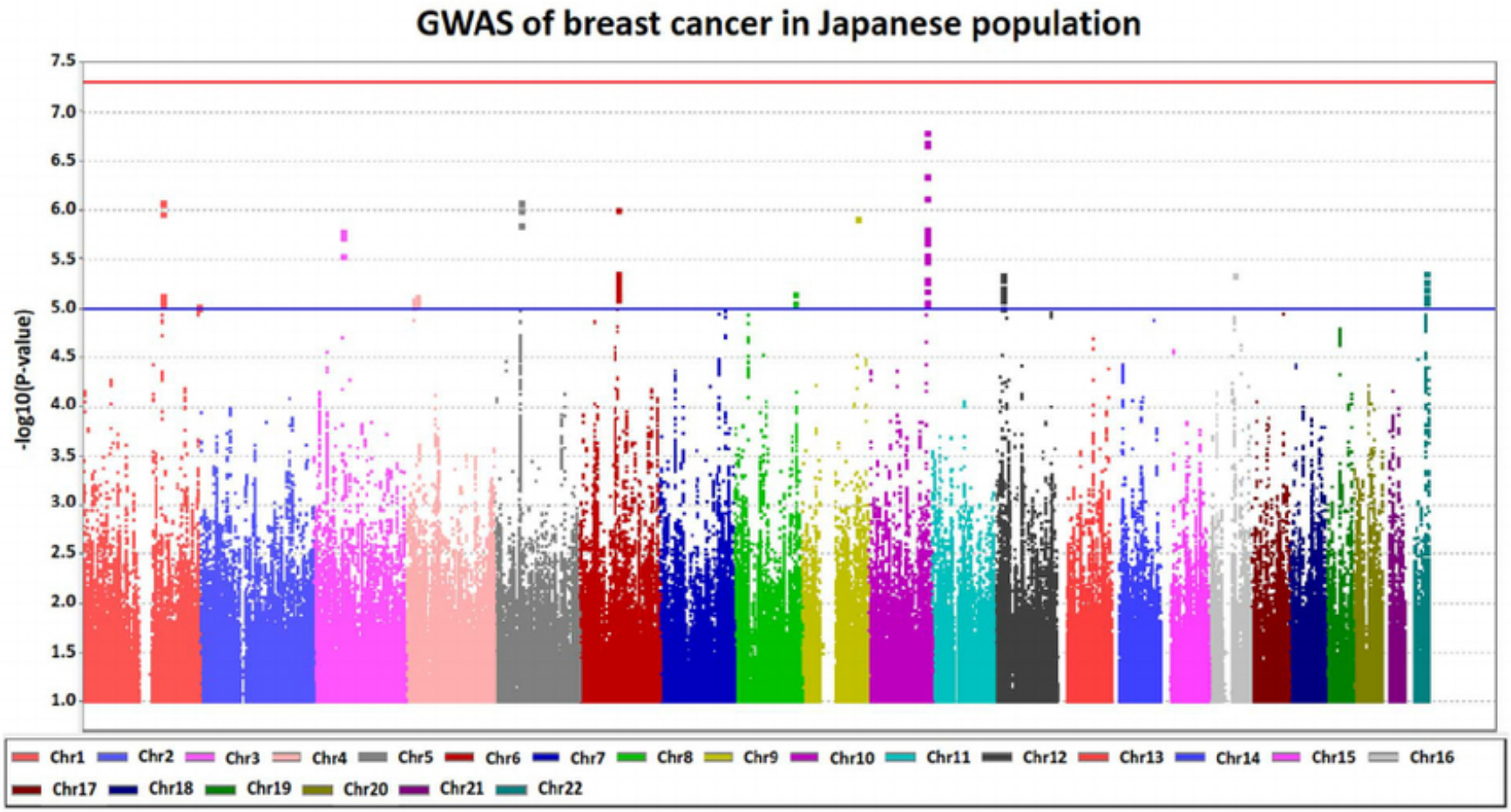
- In a Manhattan plot, P-values are transformed using $-log_{10}(P)$ and plotted against positions of the SNPs on chromosomes.
- Each point on the plot represents the P-value and location of the SNP.
- The higher the point, the smaller the P-value, the more significant of the effect.
- **Genome-wide significance level** (with Bonferroni correction):
  $P < 5 \times 10^{-8} \Rightarrow -log_{10}(5 \times 10^{-8}) = 7.30$
- **Suggestive significance level**: $P < 10^{-6}$ or $10^{-5} \Rightarrow -log_{10}(P) = 6 \text{or } 5$.

GWAS: Manhattan plot



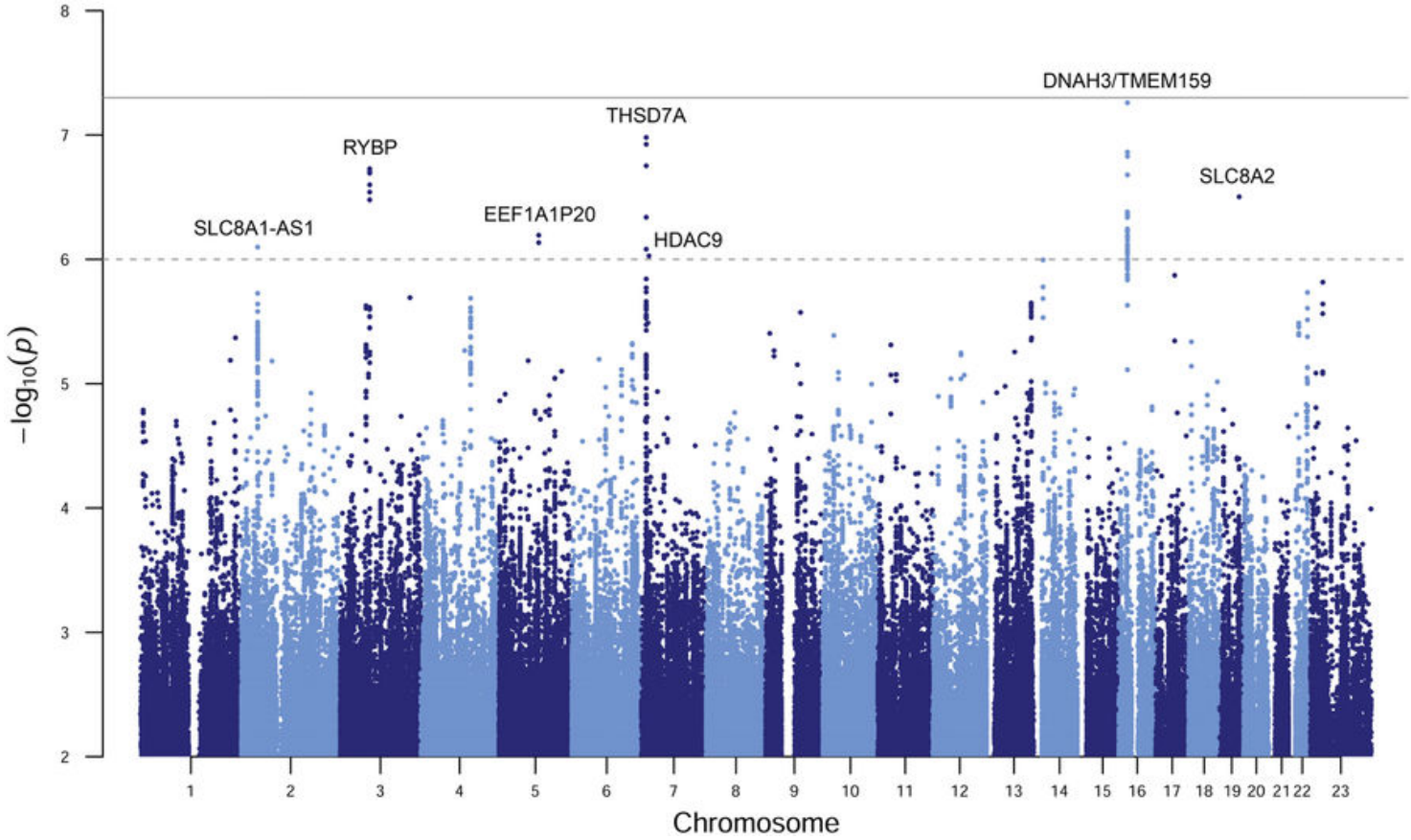
http://jojoshin.hatenablog.com

GWAS: Manhattan plot



http://www.researchgate.net

GWAS: Manhattan plot



P. Jawinski, *et al.* Human brain arousal in the resting state: a genome-wide association study. *Molecular Psychiatry* (2018). (available here)

Final Examination

- Time & Location: Tuesday 12/18 9am-12pm at SC L26
- Covers all meterials in Lecture 14~24
- Mainly short answer questions similarly as the midterm exam
- Closed-book; one two-sided letter size cheat sheet allowed
- You'll need a calculator
- Show your work and explain your reasoning
- HW 10: solutions available on Thursday 12/13
- Stat Clinics: Saturday 12/15 4-7pm, Monday 12/17 6-9pm
- Office hours: Tuesday 12/11 2:40-4:10pm, Wednesday 12/12 2:20-3:20pm, Monday 12/17 12-2pm
- Check your homework grades on Moodle and DataCamp completion
- Complete course evaluation on Moodle by Monday 12/17