Computational Genetics Spring 2014 Lecture 2

Eleazar Eskin

University of California, Los Angeles

Stats Review
+ Association Statistics
+ Association Power

Lecture 2.

April 2nd, 2014

Announcements

- 2 Questions due on web-forum tonight
  - **1 on Eric Lander video**
  - **1 on NOVA video**
- 4 Responses due by Friday night.
- HW0A and HW0B posted on website.
  - **http://genetics.cs.ucla.edu/cs124/**
  - **The due date for HW0B is January 15th.**
  - **Please note that all assignments must be submitted by a hard copy to TA's office by 4 pm on the day they are due. The TA's office is located in Math Sciences Building 2915.**
- HW1 Due 4/15 (Tuesday)

Stats Review
+ Association Statistics
+ Association Power

Lecture 2.

April 2nd, 2014

Common Sense Statistics 101

- Goal: Answer a question by collecting data.
- Example:

  In the general population, 25% stay up until 2am every day.

  Question: Do more Computer Science students stay up late?
- Approach: We collect data from class.

Collecting class data

- Total number of students: 80
- Number that stay up late 40.
- Is this significant?

Common Sense Statistics Method

- Four steps:

1. Assume there is no effect.
2. Compute quantity of interest.
3. Transform quantity to distribution.
4. Use distribution to measure significance.

Step 1: Assume there is no effect

- Class Data:
- Total number of students: 80
- Number that stay up late 40.
- Is this significant?

- Assumption: True frequency in class is .25.

Step 2: Compute Quantity of Interest

- Class Data:
- Total number of students: 80
- Number that stay up late 40.
- Is this significant?

- Assumption: True frequency in class is .25.
- Frequency in class is .5.

Step 3: Transform data to distribution

- Class Data:
- Total number of students: 80
- Number that stay up late 40.
- Is this significant?

- Assumption: True frequency in class is .25.
- Frequency in class is .5.
- Statistic:

$$Z = \frac{.5 - .25}{\sqrt{.25(1 - .25)/80}} = 5.16$$

Step 4: Use distribution to measure significance.

- Class Data:
- Total number of students: 80
- Number that stay up late 40.
- Is this significant?

If statistic is >2, then it is significant.

- Assumption: True frequency in class is .25.
- Frequency in class is .5.
- Statistic:

$$Z = \frac{.5 - .25}{\sqrt{.25(1 - .25)/80}} = 5.16$$

Step 4: Use distribution to measure significance.

- Class Data:
- Total number of students: 800
- Number that stay up late 250.
- Is this significant?

If statistic is >2, then it is significant.

- Assumption: True frequency in class is .25.
- Frequency in class is .3125.
- Statistic:

$$Z = \frac{.3125 - .25}{\sqrt{.25(1 - .25)/800}} = 4.08$$

Basic Statistics and Probability

- Event A occurs with probability p_A.
- Event A does not occur with probability 1- p_A.
- Event B occurs with probability p_B.
- Event B does not occur with probability $1-p_B$.
- Events A and B both occur with probability p_{AB}.
- The probability of event A occurring given that event B occurred is $p_{A|B}=p_{AB}/p_B$. (conditional probability)

Basic Statistics and Probability

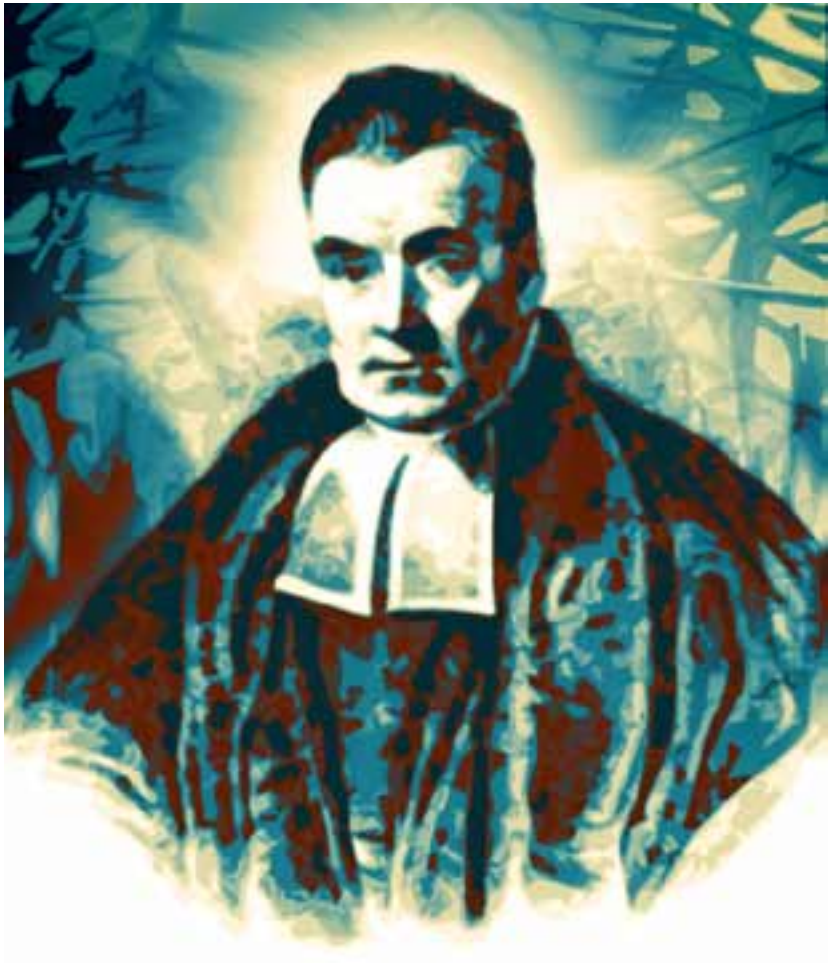
- Bayes Rule

  $p_{A|B}=p_{AB}/p_B$

  $p_{B|A}=p_{AB}/p_A$

  $p_B\, p_{A|B}=p_{AB} = p_{B|A}\, p_A$

  $p_{A|B}= p_{B|A}\, p_A/ p_B$



Basic Statistics and Probability

- Normal Distribution

  Defined by Mean μ and Variance σ^2.
- If X is a random variable with mean μ and variance σ^2, we denote this as $X \sim \mathbf{N}(\mu, \sigma^2)$.

- Central Limit Theorem (roughly):

  "The average of a large number of observations approximately follows the normal distribution."

Basic Statistics and Probability

- Normal Probability Density Function

$$\frac{1}{\sigma\sqrt{2\pi}}\,\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Normal Cumulative Density Function (CDF)

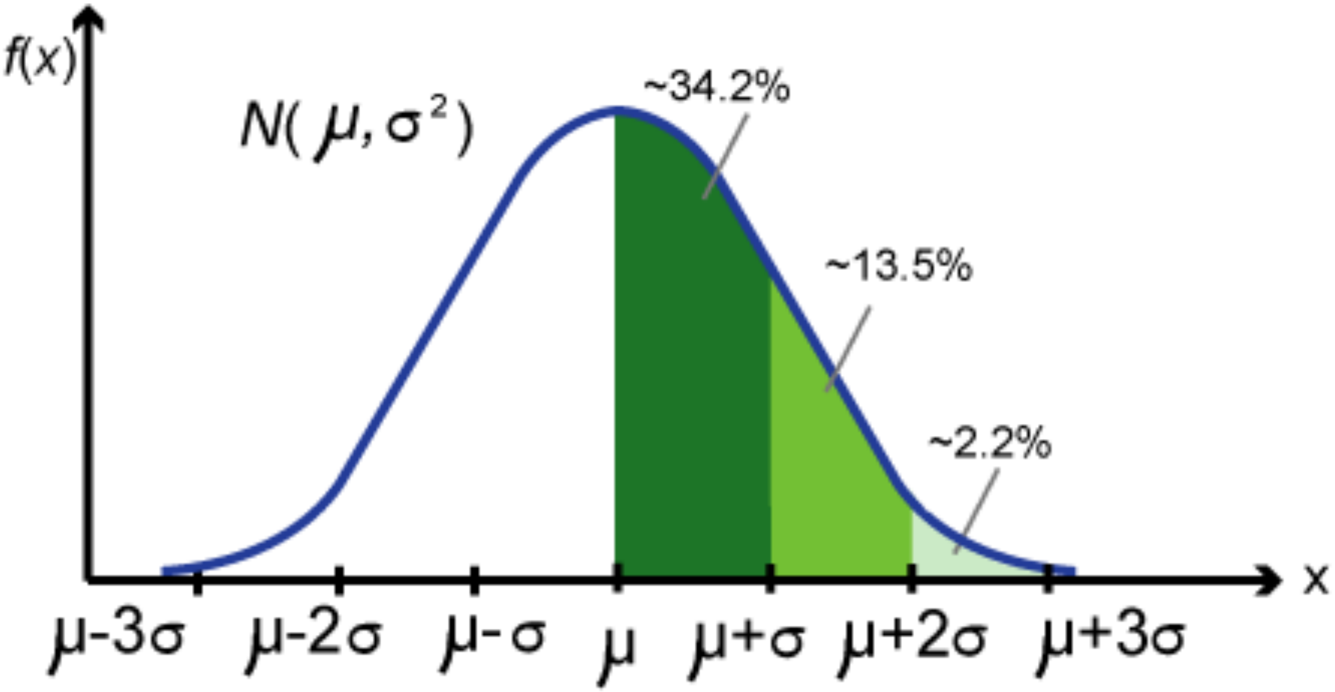
$$F(x) = P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(t-\mu)^2/2\sigma^2}\,dt$$

- Standard Normal CDF ***N***(0,1)

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2}\,dt$$

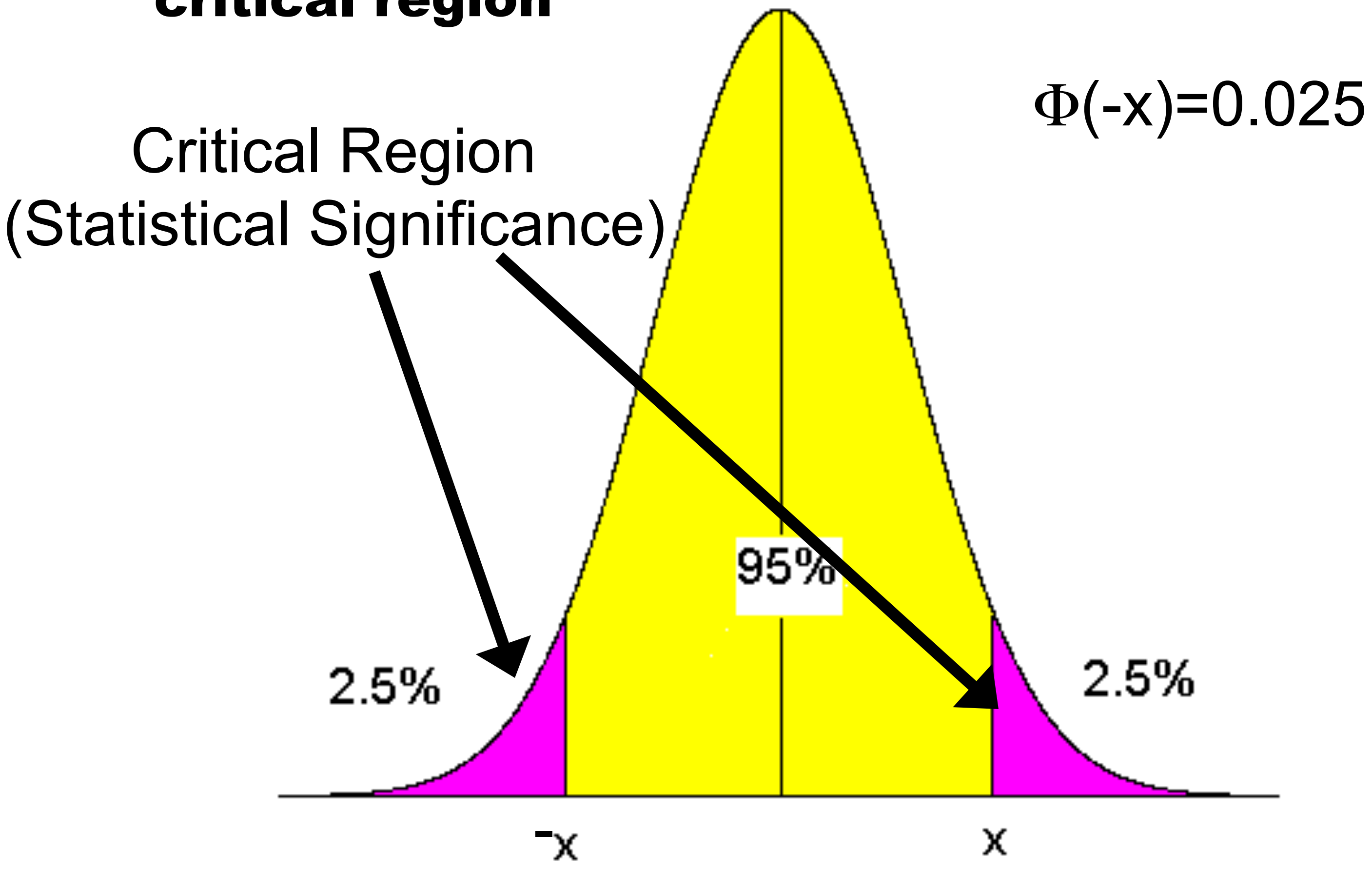
Basic Statistics and Probability

- Normal Distribution



Basic Statistics and Probability

- Normal Distribution - Hypothesis Testing
  - □ **Observe where the statistic falls on the x-axis and declare the statistic significant if it falls in the critical region**



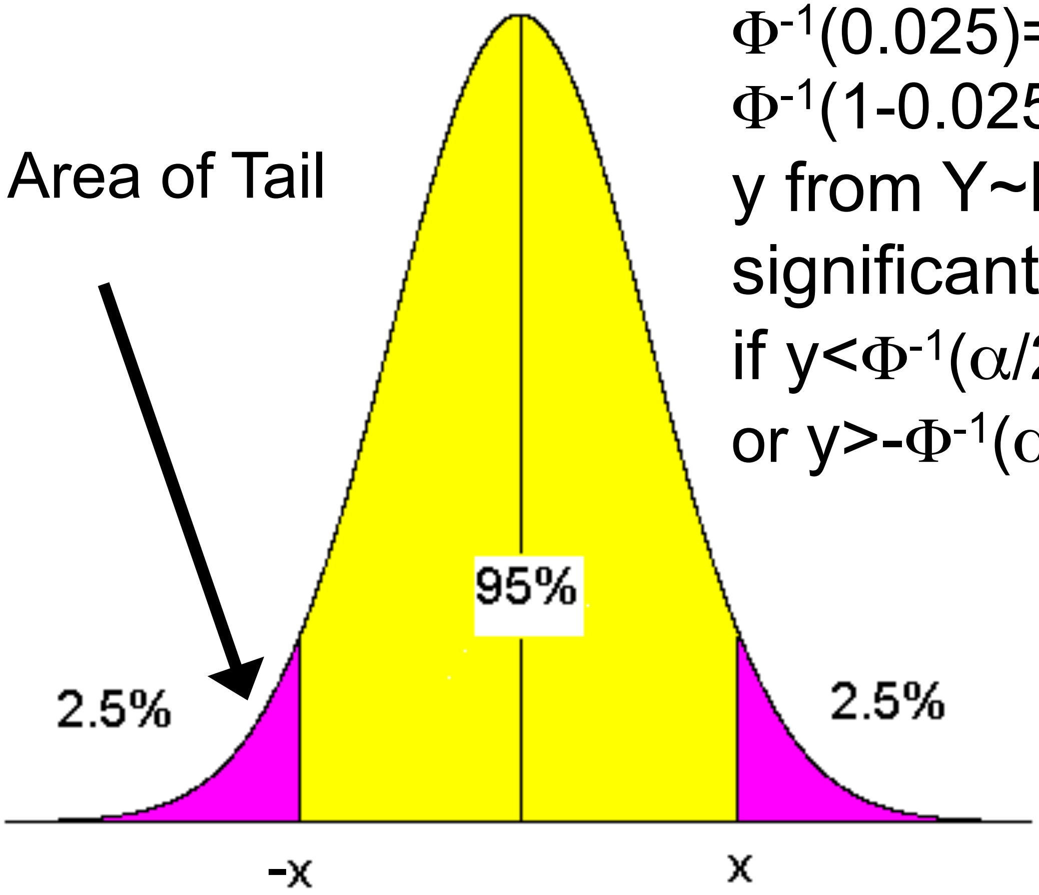
Critical Region
(Statistical Significance)

$\Phi(-x)=0.025$

Basic Statistics and Probability

- Inverse Normal Distribution

Function of Area of Tail



$\Phi^{-1}(0.025)=-x$

$\Phi^{-1}(1-0.025)=x$

y from Y~**N**(0,1) is significant at level α if $y<\Phi^{-1}(\alpha/2)$ or $y>-\Phi^{-1}(\alpha/2)$

Basic Statistics and Probability

- If X and Y are independent normally distributed random variables such that

  $X \sim \mathbf{N}(\mu_X, \sigma^2_X)$

  $Y \sim \mathbf{N}(\mu_Y, \sigma^2_Y)$

  then $X+Y \sim \mathbf{N}(\mu_X+\mu_Y, \sigma^2_X+ \sigma^2_Y)$ and

  $X-Y \sim \mathbf{N}(\mu_X-\mu_Y, \sigma^2_X+ \sigma^2_Y)$

  $aX \sim \mathbf{N}(a\mu_X, a^2\, \sigma^2_X)$
- Thus $\frac{X-\mu}{\sigma}$ $\sim \mathbf{N}(0,1)$. (called a Z-score).
- The χ^2_1 (with one degree of freedom) is approximately the square of the normal distribution.

  χ^2_1 test is approximately the square of the normal statistic test.

Basic Statistics and Probability

- If event A occurs with probability p_A in a trial, then out of N trials, the number of times A occurs is approximately normally distributed with mean Np_A and variance $Np_A(1-p_A)$.
- i.e. $N_A \sim \mathbf{N}(Np_A, Np_A(1-p_A))$.
- The observed frequency of event A in N trials is $\hat{p}_A = N_A/N \sim \mathbf{N}(p_A, p_A(1-p_A)/N)$.

Basic Statistics and Probability

- Covariance between events A and B:

  cov(A,B)=p_{AB}-$p_A p_B$

- Correlation between events A and B:

  ρ_{AB}=cor(A,B)= $\frac{p_{AB}-p_A p_B}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}$ = r

$$r^2 = \frac{(p_{AB}-p_A p_B)^2}{p_A(1-p_A)p_B(1-p_B)}$$

Basic Statistics and Probability

- P-value = the probability of observing the effect or a stronger effect under the “null” hypothesis. (i.e. the effect is due to random chance).
- P-value threshold = the maximum p-value that we will declare a significant association. Usually denoted α (i.e. α=0.05).
- Null hypothesis is independent of the strength of the effect.
- Alternative Hypothesis depends on assumed strength of effect.
- Power = Probability of significance under the alternative hypothesis.
  - **Power depends on assumed strength of effect.**

Basic Statistics and Probability

- Type I Error = False Positive. Probability of declaring an effect significant under the null hypothesis.
- Type II Error = False Negative. Probability of not declaring an effect significant under the alternative hypothesis.

Association Statistics

- Assume we are given N/2 cases and N/2 control individuals.
- Since each individual has 2 chromosomes, we have a total of N case chromosomes and N control chromosomes.

- At SNP A, let $\hat{p}^{+}_{A}$ and $\hat{p}^{-}_{A}$ be the observed case and control frequencies respectively.
- We know that:
  $\hat{p}^{+}_{A} \sim \mathbf{N}(p^{+}_{A}, p^{+}_{A}(1-p^{+}_{A})/N).$
  $\hat{p}^{-}_{A} \sim \mathbf{N}(p^{-}_{A}, p^{-}_{A}(1-p^{-}_{A})/N).$

Association Statistics

$\hat{p}^{+}_{A} \sim \mathbf{N}(p^{+}_{A}, p^{+}_{A}(1-p^{+}_{A})/N).$

$\hat{p}^{-}_{A} \sim \mathbf{N}(p^{-}_{A}, p^{-}_{A}(1-p^{-}_{A})/N).$

$\hat{p}^{+}_{A}- \hat{p}^{-}_{A} \sim \mathbf{N}(p^{+}_{A}- p^{-}_{A},(p^{+}_{A}(1-p^{+}_{A})+p^{-}_{A}(1-p^{-}_{A}))/N)$

We approximate

$p^{+}_{A}(1-p^{+}_{A})+p^{-}_{A}(1-p^{-}_{A}) \approx 2\, \hat{p}_{A}(1-\hat{p}_{A})$

$\hat{p}_{A}=(\hat{p}^{+}_{A}+\hat{p}^{-}_{A})/2$

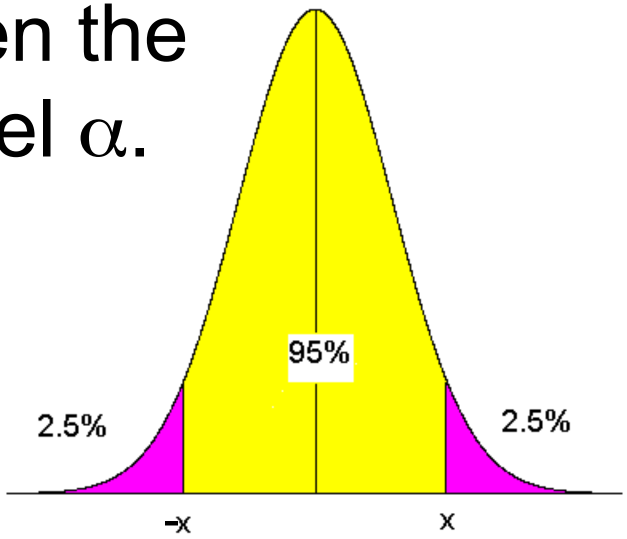
then if $p^{+}_{A} = p^{-}_{A}$

$$S_A = \frac{\hat{p}^{+}{}_{A} - \hat{p}^{-}{}_{A}}{\sqrt{2/N}\sqrt{\hat{p}_A(1-\hat{p}_A)}} \sim N(0,1)$$

Association Statistic

$$S_A = \frac{\hat{p}^+{}_A - \hat{p}^-{}_A}{\sqrt{2/N}\sqrt{\hat{p}_A(1-\hat{p}_A)}} \sim N(0,1)$$

- Under the null hypothesis $p^+_A - p^-_A = 0$
- We compute the statistic S_A.
- If $S_A < \Phi^{-1}(\alpha/2)$ or $S_A > -\Phi^{-1}(\alpha/2)$ then the association is significant at level α.



Association Study Example

- Significance Threshold α=0.05
- Sample: 100 Cases and 100 Controls
- Genotype SNP A with alleles {A,a}.
- Total of 200 Case Chromosomes and 200 Control Chromosomes
- We observe 120 A’s in the Cases and 110 A’s in the Controls.

$$\hat{p}_A^+ = \frac{120}{200} = .6 \quad \hat{p}_A^- = \frac{110}{200} = .55 \quad \hat{p}_A = \frac{\hat{p}_A^+ + \hat{p}_A^-}{2} = .575$$

Association Study Example

$$\hat{p}_A^+ = \frac{120}{200} = .6 \quad \hat{p}_A^- = \frac{110}{200} = .55 \quad \hat{p}_A = \frac{\hat{p}_A^+ + \hat{p}_A^-}{2} = .575$$

$$S_A = \frac{\hat{p}^+{}_A - \hat{p}^-{}_A}{\sqrt{2/N}\sqrt{\hat{p}_A(1-\hat{p}_A)}}$$

$$= \frac{.6 - .55}{\sqrt{2/200}\sqrt{.575(1-.575)}} = 1.01$$

Association Study Example

$$\Phi^{-1}(\alpha/2) = \Phi^{-1}(0.025) = \text{qnorm}(0.025) = -1.95$$

$$\Phi^{-1}(1-0.025) = \Phi^{-1}(0.975) = \text{qnorm}(0.975) = 1.95$$

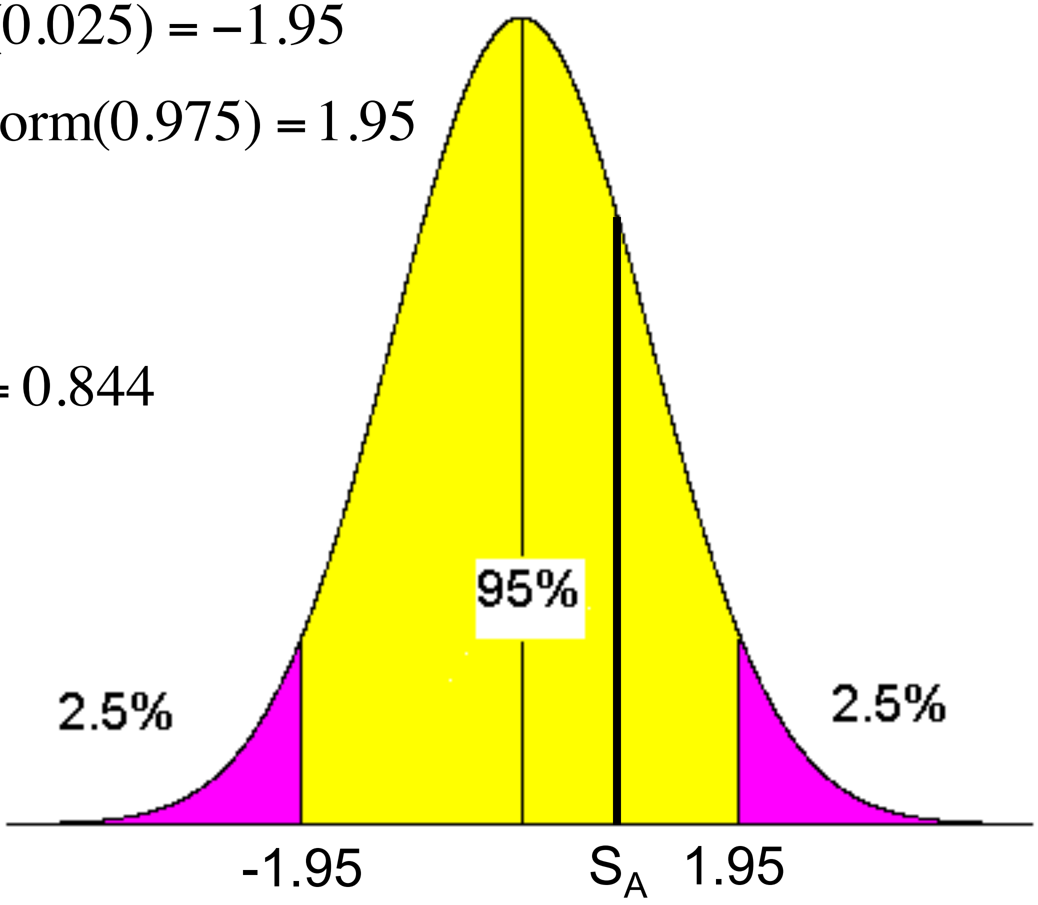
$$\Phi(-1.95) = 0.025$$

$$\Phi(1.95) = 0.975$$

$$\Phi(S_A) = \Phi(1.01) = \text{pnorm}(1.01) = 0.844$$

$$\Phi(-S_A) = \text{pnorm}(-1.01) = 0.156$$

$$\text{p} - \text{value} = 0.312$$



Association Study Example

- Significance Threshold α=0.05
- Sample: 100 Cases and 100 Controls
- Genotype SNP A with alleles {A,a}.
- Total of 200 Case Chromosomes and 200 Control Chromosomes
- We observe 130 A's in the Cases and 110 A's in the Controls.

$$\hat{p}_A^+ = \tfrac{130}{200} = .65 \quad \hat{p}_A^- = \tfrac{110}{200} = .55 \quad \hat{p}_A = \frac{\hat{p}_A^+ + \hat{p}_A^-}{2} = .6$$

Association Study Example

$$\hat{p}_A^+ = \frac{130}{200} = .65 \quad \hat{p}_A^- = \frac{110}{200} = .55 \quad \hat{p}_A = \frac{\hat{p}_A^+ + \hat{p}_A^-}{2} = .6$$

$$S_A = \frac{\hat{p}^+{}_A - \hat{p}^-{}_A}{\sqrt{2/N}\sqrt{\hat{p}_A(1-\hat{p}_A)}}$$

$$= \frac{.65 - .55}{\sqrt{2/200}\sqrt{.6(1-.6)}} = 2.04$$

Association Study Example

$$\Phi^{-1}(\alpha/2) = \Phi^{-1}(0.025) = \text{qnorm}(0.025) = -1.95$$

$$\Phi^{-1}(1-0.025) = \Phi^{-1}(0.975) = \text{qnorm}(0.975) = 1.95$$

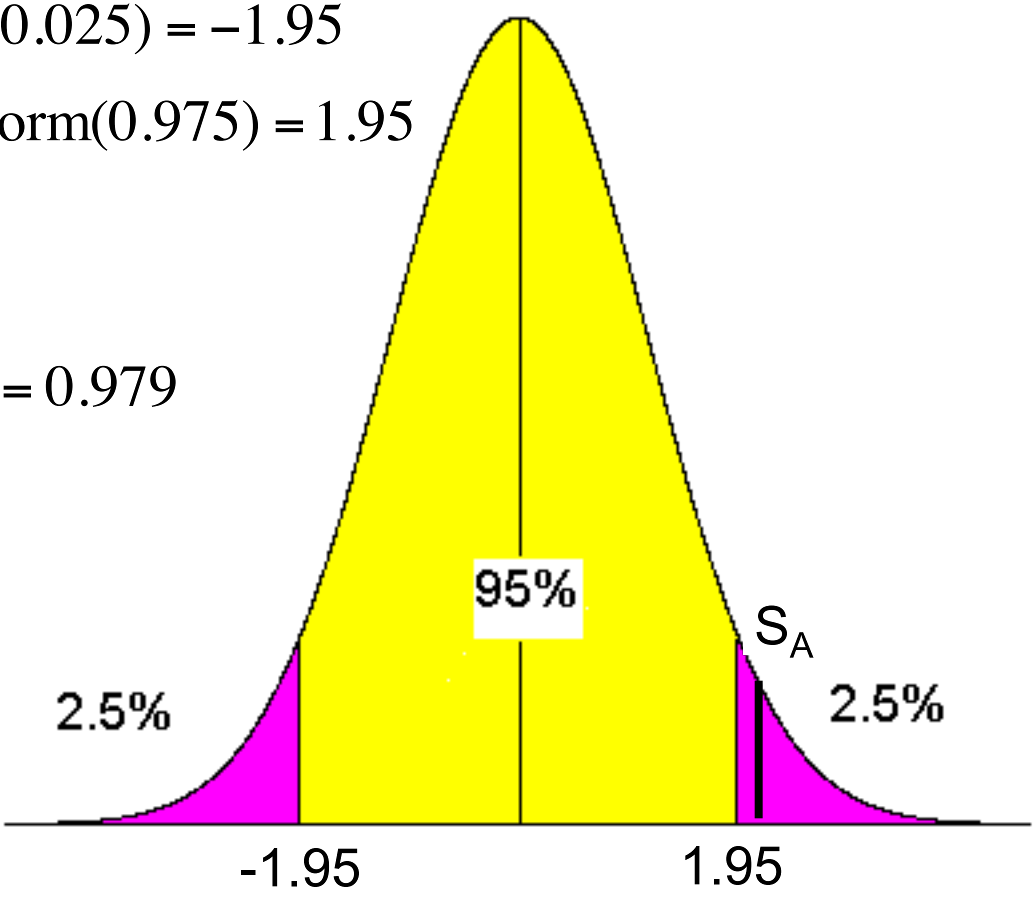
$$\Phi(-1.95) = 0.025$$

$$\Phi(1.95) = 0.975$$

$$\Phi(S_A) = \Phi(2.04) = \text{pnorm}(2.04) = 0.979$$

$$\Phi(-S_A) = \text{pnorm}(-2.04) = 0.021$$

$$\text{p} - \text{value} = 0.042$$



Association Study Example

- Significance Threshold α=0.05
- Sample: 1000 Cases and 1000 Controls
- Genotype SNP A with alleles {A,a}.
- Total of 2000 Case Chromosomes and 2000 Control Chromosomes
- We observe 1200 A's in the Cases and 1100 A's in the Controls.

$$\hat{p}_A^+ = \tfrac{1200}{2000} = .6 \quad \hat{p}_A^- = \tfrac{1100}{2000} = .55 \quad \hat{p}_A = \tfrac{\hat{p}_A^+ + \hat{p}_A^-}{2} = .575$$

Association Study Example

$$\hat{p}_A^+ = \frac{1200}{2000} = .6 \quad \hat{p}_A^- = \frac{1100}{2000} = .55 \quad \hat{p}_A = \frac{\hat{p}_A^+ + \hat{p}_A^-}{2} = .575$$

$$S_A = \frac{\hat{p}^+{}_A - \hat{p}^-{}_A}{\sqrt{2/N}\sqrt{\hat{p}_A(1-\hat{p}_A)}}$$

$$= \frac{.6 - .55}{\sqrt{2/2000}\sqrt{.575(1-.575)}} = 3.19$$

Association Study Example

$$\Phi^{-1}(\alpha/2) = \Phi^{-1}(0.025) = \text{qnorm}(0.025) = -1.95$$

$$\Phi^{-1}(1-0.025) = \Phi^{-1}(0.975) = \text{qnorm}(0.975) = 1.95$$

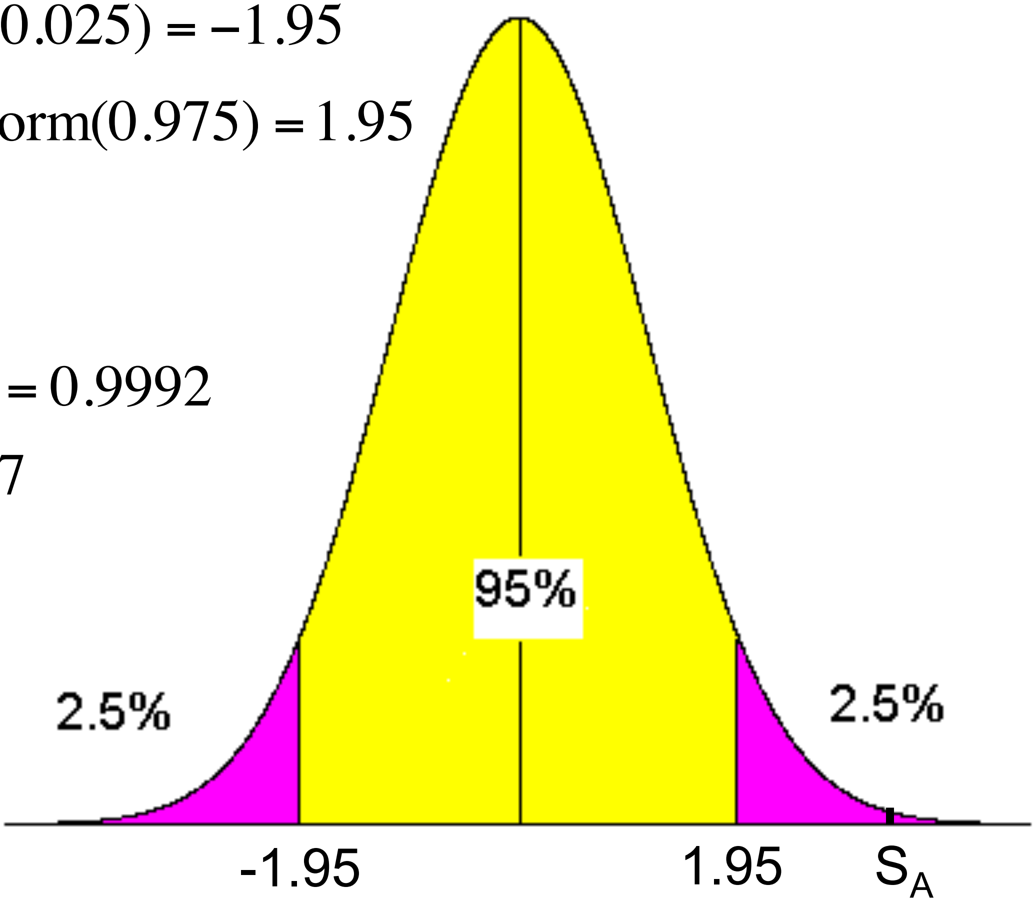
$$\Phi(-1.95) = 0.025$$

$$\Phi(1.95) = 0.975$$

$$\Phi(S_A) = \Phi(3.14) = \text{pnorm}(3.14) = 0.9992$$

$$\Phi(-S_A) = \text{pnorm}(-3.14) = 0.0007$$

$$\text{p}-\text{value} = 0.0014$$



Association Power

- Lets assume that SNP A is causal and $p^+_A \neq p^-_A$
- Given the true p^+_A and p^-_A, if we collect N individuals, and compute the statistic S_A, the probability that S_A has a significance level of α is the power.
- Power is the chance of detecting an association of a certain strength with a certain number of individuals.
- We can set the number of individuals to achieve a certain power.

Association Statistic Assuming True Association

- Lets assume that $p^{+}_{A} \neq p^{-}_{A}$ then

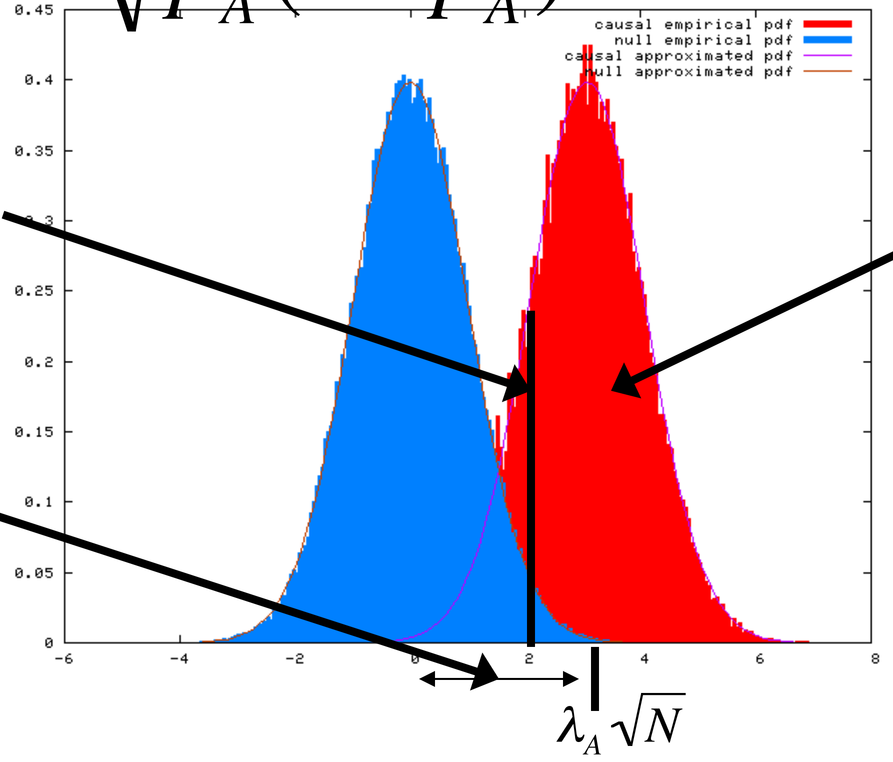
$$S_A = \frac{\hat{p}^{+}{}_{A} - \hat{p}^{-}{}_{A}}{\sqrt{2/N}\sqrt{\hat{p}_A(1-\hat{p}_A)}} \sim N\left(\frac{p_A^{+} - p_A^{-}}{\sqrt{2/N}\sqrt{p_A(1-p_A)}},1\right)$$

$$S_A = \frac{\hat{p}^{+}{}_{A} - \hat{p}^{-}{}_{A}}{\sqrt{2/N}\sqrt{\hat{p}_A(1-\hat{p}_A)}} \sim N\left(\frac{(p_A^{+} - p_A^{-})\sqrt{N}}{\sqrt{2p_A(1-p_A)}},1\right)$$

$$S_A = \frac{\hat{p}^{+}{}_{A} - \hat{p}^{-}{}_{A}}{\sqrt{2/N}\sqrt{\hat{p}_A(1-\hat{p}_A)}} \sim N\left(\lambda_A\sqrt{N},1\right)$$

Association Power

$$S_A = \frac{\hat{p}^{+}{}_A - \hat{p}^{-}{}_A}{\sqrt{2/N}\sqrt{\hat{p}_A(1-\hat{p}_A)}} \sim N\left(\lambda_A\sqrt{N},1\right)$$



Threshold for significance

Non-centrality parameter.

Power of association test

Power Example

- Significance Threshold α=0.05
- Genotype SNP A with alleles {A,a}.
- Assume true case/control probabilities are 0.6/0.5.
- How many individuals do we need to collect to observe a significant association?

Power Example

- Significance Threshold α=0.05
- Genotype SNP A with alleles {A,a}.
- Assume true case/control probabilities are 0.6/0.5.
- If we collect 100 case and 100 control individuals…

$$p_A^+ = .6 \quad p_A^- = .5 \quad p_A = \frac{p_A^+ + p_A^-}{2} = .55 \quad N = 200$$

Power Example

$$p_A^+ = .6 \quad p_A^- = .5 \quad p_A = \tfrac{p_A^+ + p_A^-}{2} = .55 \quad N = 200$$

$$\lambda_A \sqrt{N} = \frac{p_A^+ - p_A^-}{\sqrt{2/200}\sqrt{p_A(1-p_A)}}$$

$$= \frac{.6-.5}{\sqrt{2/200}\sqrt{.55(1-.55)}} = 2.01$$

$$S_A \sim N(\lambda_A \sqrt{N},1) = N(2.01,1)$$

Power Example

power

$$= \Phi(\Phi^{-1}(\alpha/2) + \lambda_A\sqrt{N})$$

$$+1 - \Phi(-\Phi^{-1}(\alpha/2) + \lambda_A\sqrt{N})$$

$$= \Phi(\Phi^{-1}(0.025) + 2.01)$$

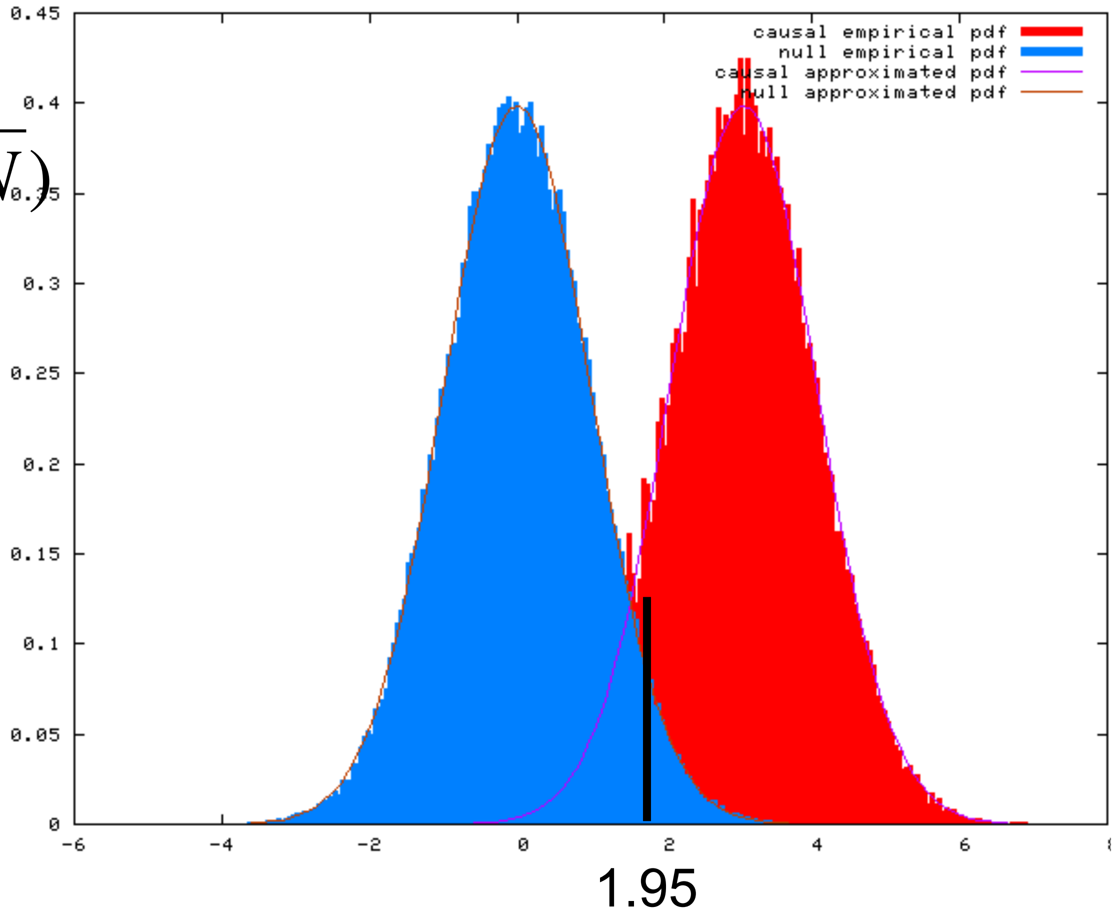
$$+1 - \Phi(-\Phi^{-1}(\alpha/2) + 2.01)$$

$$= \Phi(-1.95 + 2.01)$$

$$+1 - \Phi(1.95 + 2.01)$$

$$= \Phi(.06) + 1 - \Phi(3.96)$$

$$= .52 + 1 - .9999625 = .52$$



Power Example

- Significance Threshold α=0.05
- Genotype SNP A with alleles {A,a}.
- Assume true case/control probabilities are 0.6/0.5.
- If we collect 200 case and 200 control individuals…

$$p_A^+ = .6 \quad p_A^- = .5 \quad p_A = \frac{p_A^+ + p_A^-}{2} = .55 \quad N = 400$$

Power Example

$$p_A^+ = .6 \quad p_A^- = .5 \quad p_A = \frac{p_A^+ + p_A^-}{2} = .55 \quad N = 400$$

$$\lambda_A \sqrt{N} = \frac{p_A^+ - p_A^-}{\sqrt{2/400}\sqrt{p_A(1-p_A)}}$$

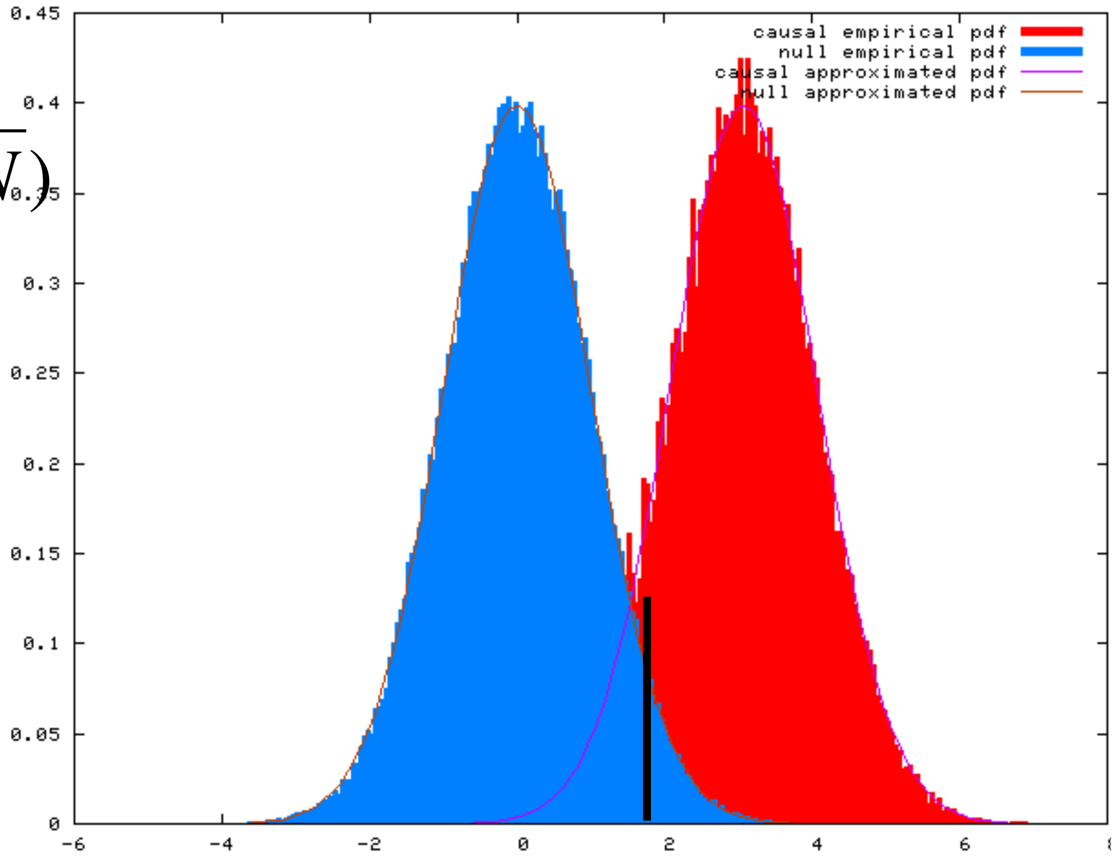
$$= \frac{.6 - .5}{\sqrt{2/400}\sqrt{.55(1-.55)}} = 2.84$$

$$S_A \sim N(\lambda_A \sqrt{N},1) = N(2.84,1)$$

Power Example

power

$$
\begin{aligned}
&= \Phi(\Phi^{-1}(\alpha/2) + \lambda_A \sqrt{N}) \\
&\quad +1 - \Phi(-\Phi^{-1}(\alpha/2) + \lambda_A \sqrt{N}) \\
&= \Phi(\Phi^{-1}(0.025) + 2.84) \\
&\quad +1 - \Phi(-\Phi^{-1}(\alpha/2) + 2.84) \\
&= \Phi(-1.95 + 2.84) \\
&\quad +1 - \Phi(1.95 + 2.84) \\
&= \Phi(.89) + 1 - \Phi(4.79) \\
&= .81 + 1 - .9999992 = .81
\end{aligned}
$$



1.95

Power Example

- Significance Threshold α=0.05
- Genotype SNP A with alleles {A,a}.
- Assume true case/control probabilities are 0.6/0.5.
- If we collect 400 case and 400 control individuals…

$$p_A^+ = .6 \quad p_A^- = .5 \quad p_A = \frac{p_A^+ + p_A^-}{2} = .55 \quad N = 800$$

Power Example

$$p_A^+ = .6 \quad p_A^- = .5 \quad p_A = \frac{p_A^+ + p_A^-}{2} = .55 \quad N = 800$$

$$\lambda_A \sqrt{N} = \frac{p_A^+ - p_A^-}{\sqrt{2/800}\sqrt{p_A(1-p_A)}}$$

$$= \frac{.6 - .5}{\sqrt{2/800}\sqrt{.55(1-.55)}} = 4.02$$

$$S_A \sim N(\lambda_A \sqrt{N}, 1) = N(4.02, 1)$$

Power Example

power

$$= \Phi(\Phi^{-1}(\alpha/2) + \lambda_A \sqrt{N})$$

$$+1 - \Phi(-\Phi^{-1}(\alpha/2) + \lambda_A \sqrt{N})$$

$$= \Phi(\Phi^{-1}(0.025) + 4.02)$$

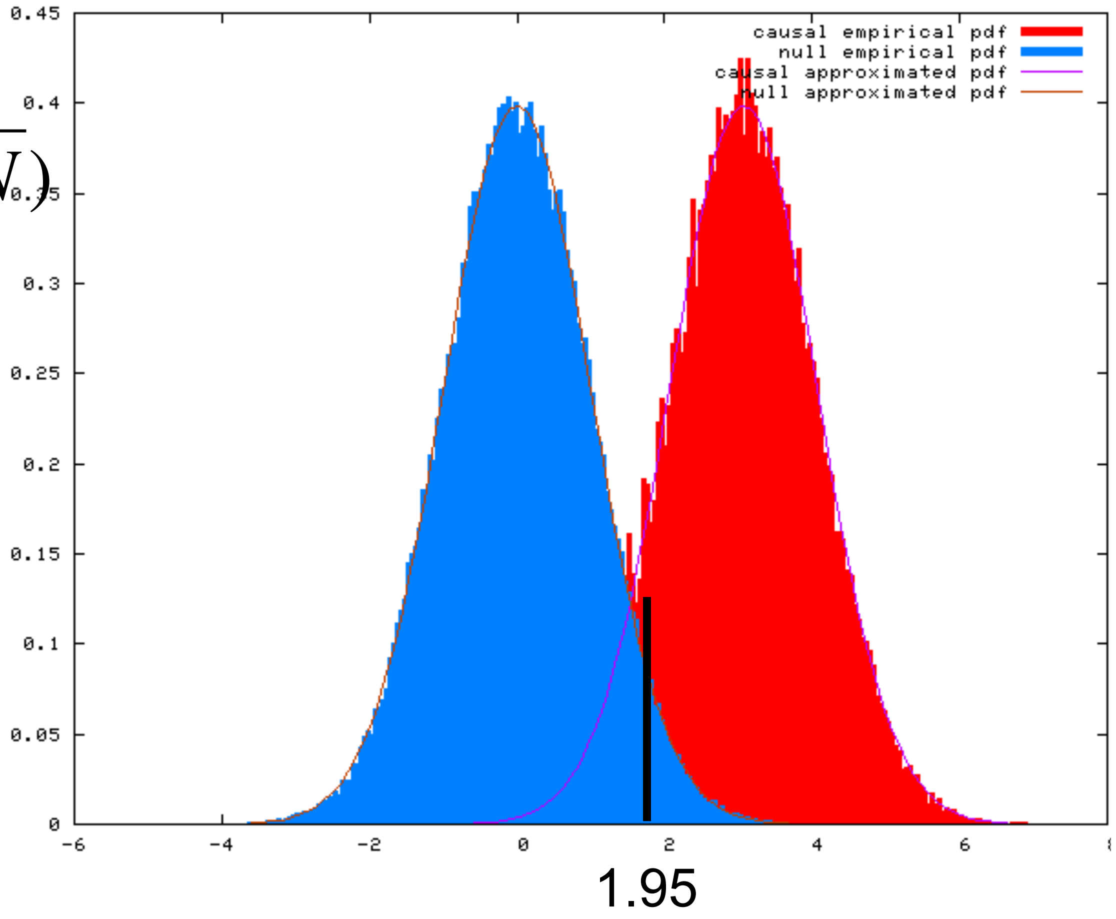
$$+1 - \Phi(-\Phi^{-1}(\alpha/2) + 4.02)$$

$$= \Phi(-1.95 + 4.02)$$

$$+1 - \Phi(1.95 + 4.02)$$

$$= \Phi(2.07) + 1 - \Phi(5.97)$$

$$= .98 + 1 - 1 = .98$$



Association Strength

- A causal SNP has a certain strength of effect on the disease.
- This effect can be parameterized by:

  γ = relative risk
- Definitions:

  p_A = allele frequency of SNP A.

  F = disease prevalence

  +/- = disease state.
- Derivation of case and control frequencies:

  $P(A)=p_A \qquad p^+_A=P(A|+) \qquad p^-_A=P(A|-) \qquad F=P(+)$

  $P(A|+)=P(+|A)P(A)/P(+)$

  $P(+|A)= \gamma P(+|\neg A)$

  $P(+)=F=p_A P(+|A)+(1-p_A)P(+|\neg A)$

  $P(+)=F= p_A P(+|A)+(1-p_A)P(+|A)/\gamma$

  $P(+)=F=P(+|A)(p_A+(1-p_A)/\gamma)= P(+|A)(p_A(\gamma-1)+1)/\gamma$

  $P(+|A)= \gamma F/(p_A(\gamma-1)+1)$

  $P(A|+)=P(+|A)P(A)/P(+)= P(+|A)p_A/F= \gamma p_A/(p_A(\gamma-1)+1)$

Association Strength

$P(-|A)=1-P(+|A)=1-\gamma F/(p_A(\gamma-1)+1)$
$P(A|-)=P(-|A)P(A)/P(-)$
If F is small, then $1-F \approx 1$ and $P(-|A) \approx 1$
then, $P(A|-) \approx P(A) = p_A$

$$\begin{aligned} p^+_A-p^-_A &\approx \gamma p_A/(p_A(\gamma-1)+1)-p_A \\ &\approx (p_A(\gamma-1)-p_A{}^2(\gamma-1))/(p_A(\gamma-1)+1) \\ &\approx (\gamma-1)p_A(1-p_A)/(p_A(\gamma-1)+1) \end{aligned}$$

$$\lambda_A = \frac{(p_A^+ - p_A^-)}{\sqrt{2p_A(1-p_A)}} \quad S_A \sim N\left(\lambda_A\sqrt{N},1\right)$$

Examples

- Relative Risk Affect
- Minor Allele Frequency Affect
- Number of Individuals Affect

$p^+_A - p^-_A \approx (\gamma-1)p_A(1-p_A)/(p_A(\gamma-1)+1)$

$$\lambda_A = \frac{(p_A^+ - p_A^-)}{\sqrt{2p_A(1-p_A)}} \qquad S_A \sim N\left(\lambda_A\sqrt{N},1\right)$$

Unequal Case and Control Size

- Let $N^+/2$ and $N^-/2$ be the size of the case and control populations respectively.
- $P^+_A - P^-_A \sim \mathbf{N}(p^+_A - p^-_A, (p^+_A(1-p^+_A)/N^+ + p^-_A(1-p^-_A)/N^-))$

  and $p^+_A(1-p^+_A)/N^+ + p^-_A(1-p^-_A)/N^-$

  $\approx p_A(1-p_A)\ (N^+ + N^-)/N^+N^-$

$$\lambda_A = \frac{(p_A^+ - p_A^-)}{\sqrt{2p_A(1-p_A)}} S_A \sim N\left(\lambda_A\sqrt{\frac{2(N^+N^-)}{N^+ + N^-}}, 1\right)$$

Infinite Control Population

- Let us assume that the size of the control population N$^-$ is unlimited. What is the increase in power?

$$S_A \sim N\left(\lambda_A \sqrt{\frac{2(N^+N^-)}{N^+ + N^-}},1\right)$$

Break!

Introduction to the HapMap

Lecture 2b.

April 2nd, 2014

- Successor to the Human Genome Project
- International consortium that aims in genotyping the genome of 270 individuals from four different populations.
- Launched in 2002. First phase was finished in October (Nature, 2005).
- Collected genotypes for 3.9 million SNPs.
- Location and correlation structure of many common SNPs.

Background

- Human Genome Project.
  - **Complete genome sequenced. 99.9% identical.**

- Goal: a “map” common human variation

- Officially launched October 27-29, 2002.

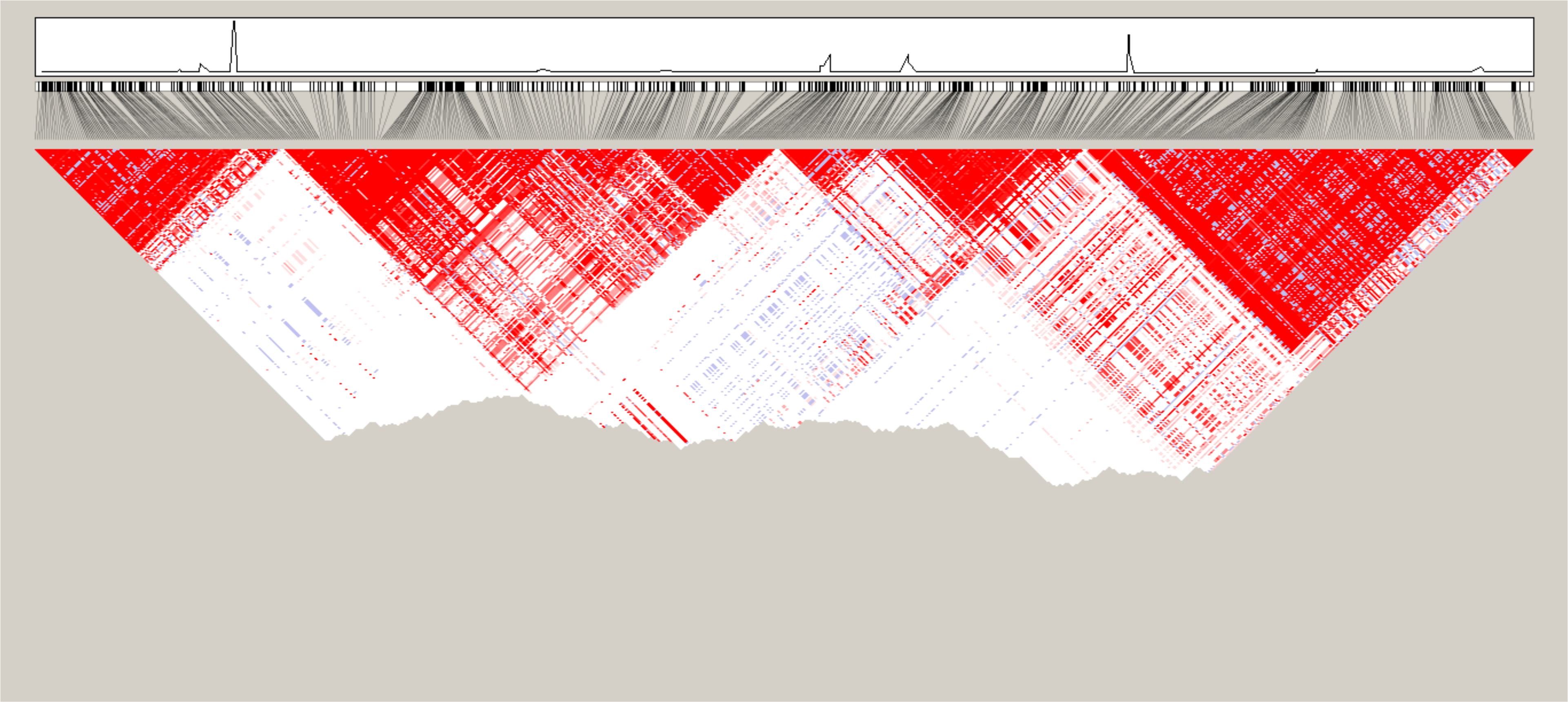
HapMap samples

- 90 Yoruba individuals (30 parent-parent-offspring trios) from Ibadan, Nigeria (YRI)
- 90 individuals (30 trios) of European descent from Utah (CEU)
- 45 Han Chinese individuals from Beijing (CHB)
- 45 Japanese individuals from Tokyo (JPT)
- Enough samples to identify common genetic variation.

Encode project

- Ten “typical” 500 kb regions.
  - □ **across a range of chromosomes**
  - □ **with different recombination rates and gene density.**
- Strategy:
  - **fully resequenced 16 CEPH individuals, 16 Nigerians, 8 Chinese and 8 Japanese from HapMap samples;**
  - **genotype SNPs in complete HapMap sample, initially SNPs in dbSNP, then additional SNPs discovered via resequencing.**
- Currently, 1 SNP every 279 bp.

Correlation structure (LD)



Phase II and Phase III

- More SNPs:
  - **Published in 2007**
- More populations.
  - **Indians, mexicans, middle easterns, etc.**
  - **Published in 2010**

- Thousand Genomes Project
  - **Sequence 1000 people**
  - **Obtain rare variants**

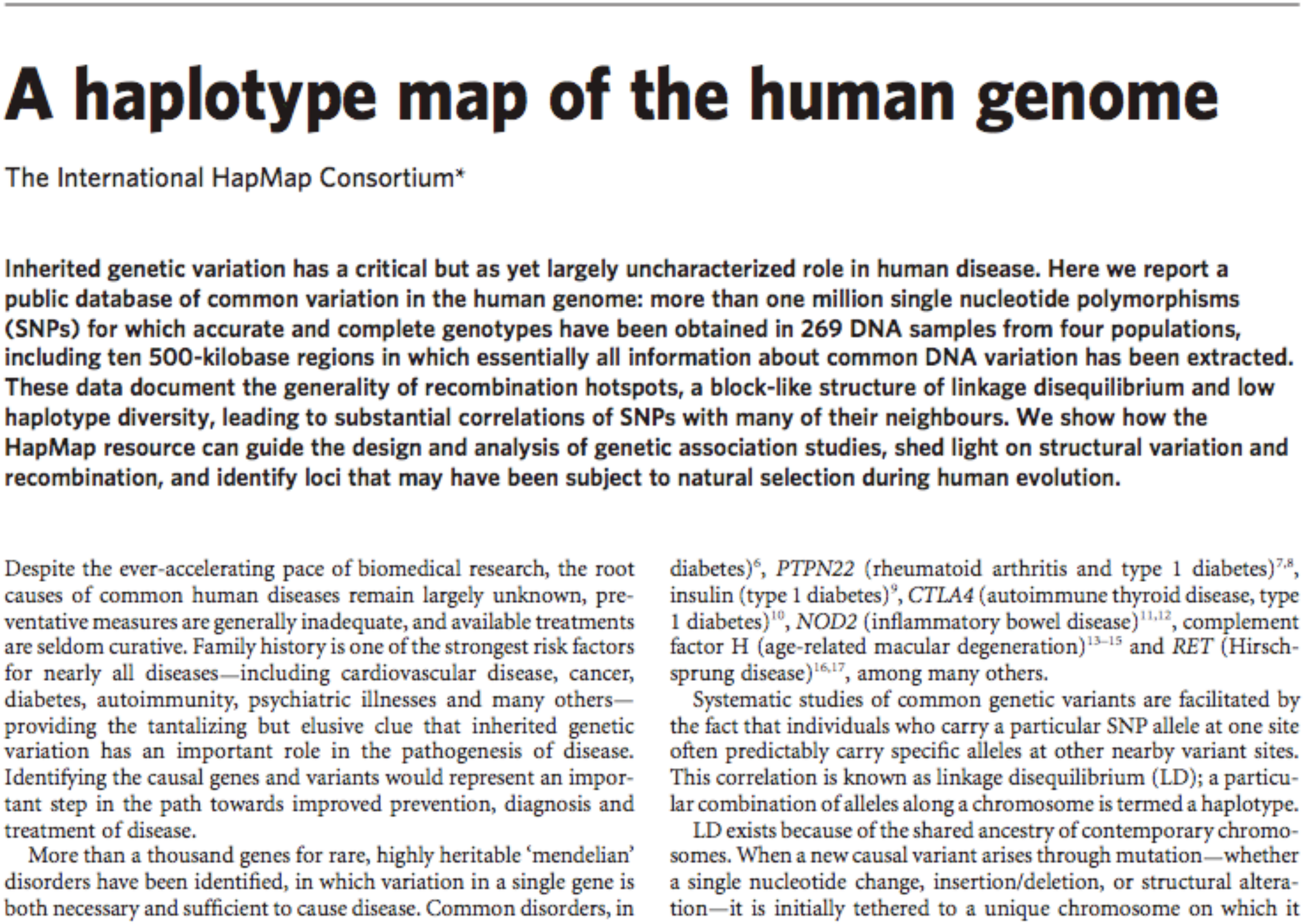
www.hapmap.org

- All data from HapMap project publicly available
- HapMart tool for easily obtaining only data you need.

References

- The International HapMap Consortium. **The International HapMap Project.** *Nature* 426, 789-796. 2003.
- The International HapMap Consortium. **A haplotype map of the human genome.** *Nature* 437, 1299-1320. 2005.
- The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851-861. 2007.
- The International HapMap Consortium. Integrating common and rare genetic variation in diverse human populations.Nature 467, 52-58. 2010
- Thorisson, G.A., Smith, A.V., Krishnan, L., and Stein, L.D. **The International HapMap Project Web site.** *Genome Research*,15:1591-1593. 2005.

The HapMap paper



A haplotype map of the human genome

The International HapMap Consortium*

Inherited genetic variation has a critical but as yet largely uncharacterized role in human disease. Here we report a public database of common variation in the human genome: more than one million single nucleotide polymorphisms (SNPs) for which accurate and complete genotypes have been obtained in 269 DNA samples from four populations, including ten 500-kilobase regions in which essentially all information about common DNA variation has been extracted. These data document the generality of recombination hotspots, a block-like structure of linkage disequilibrium and low haplotype diversity, leading to substantial correlations of SNPs with many of their neighbours. We show how the HapMap resource can guide the design and analysis of genetic association studies, shed light on structural variation and recombination, and identify loci that may have been subject to natural selection during human evolution.

Despite the ever-accelerating pace of biomedical research, the root causes of common human diseases remain largely unknown, preventative measures are generally inadequate, and available treatments are seldom curative. Family history is one of the strongest risk factors for nearly all diseases—including cardiovascular disease, cancer, diabetes, autoimmunity, psychiatric illnesses and many others—providing the tantalizing but elusive clue that inherited genetic variation has an important role in the pathogenesis of disease. Identifying the causal genes and variants would represent an important step in the path towards improved prevention, diagnosis and treatment of disease.

More than a thousand genes for rare, highly heritable 'mendelian' disorders have been identified, in which variation in a single gene is both necessary and sufficient to cause disease. Common disorders, in diabetes)[6], *PTPN22* (rheumatoid arthritis and type 1 diabetes)[7,8], insulin (type 1 diabetes)[9], *CTLA4* (autoimmune thyroid disease, type 1 diabetes)[10], *NOD2* (inflammatory bowel disease)[11,12], complement factor H (age-related macular degeneration)[13–15] and *RET* (Hirschsprung disease)[16,17], among many others.

Systematic studies of common genetic variants are facilitated by the fact that individuals who carry a particular SNP allele at one site often predictably carry specific alleles at other nearby variant sites. This correlation is known as linkage disequilibrium (LD); a particular combination of alleles along a chromosome is termed a haplotype.

LD exists because of the shared ancestry of contemporary chromosomes. When a new causal variant arises through mutation—whether a single nucleotide change, insertion/deletion, or structural alteration—it is initially tethered to a unique chromosome on which it

HapMap Paper Introduction

- Mendelian vs. Complex Diseases
- Linkage vs. Association
- Candidate Genes vs Whole Genome Association
- Correlation Structure of Genome (why?)

HapMap Data

- Phase I data - evenly spaced map of variation
- dbSNP - database of variation contains most common variation

HapMap Analysis (LD)

- Haplotypes shared among populations.
- Limited diversity in areas of low recombination
- Variation in recombination rates.
- Block like structure of human LD.

HapMap - Practical Implication

- Correlation structure can support association studies.
- Selection of tag SNPs (we will cover this in class later).
- Tags work on mutliple populations.

HapMap - Other opportunities

- Identifying structural variants
- Identifying regions undergoing selection.
- Fine scale recombination rates.
- Selective sweeps.