Computational Genetics
Spring 2014
Lecture 3

Eleazar Eskin

University of California, Los Angeles

Indirect Association + Multiple Hypothesis Testing

Lecture 3.

April 7th, 2014

Home Work & Midterm

- HW 0A due on Monday (4/7/14)
- HW 0B due on Tuesday (4/15/14)

- HapMap Paper Question due Wednesday (4/9/14)
- HapMap Paper Responses due Friday (4/11/14)

- Project Selection due (4/11/14)

- Midterm Review (4/21/14)
- Midterm (4/23/14)

Indirect Association + Multiple Hypothesis Testing

Lecture 3.

April 7th, 2014

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

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

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

$$= \Phi(\Phi^{-1}(0.025) + 2.84)$$

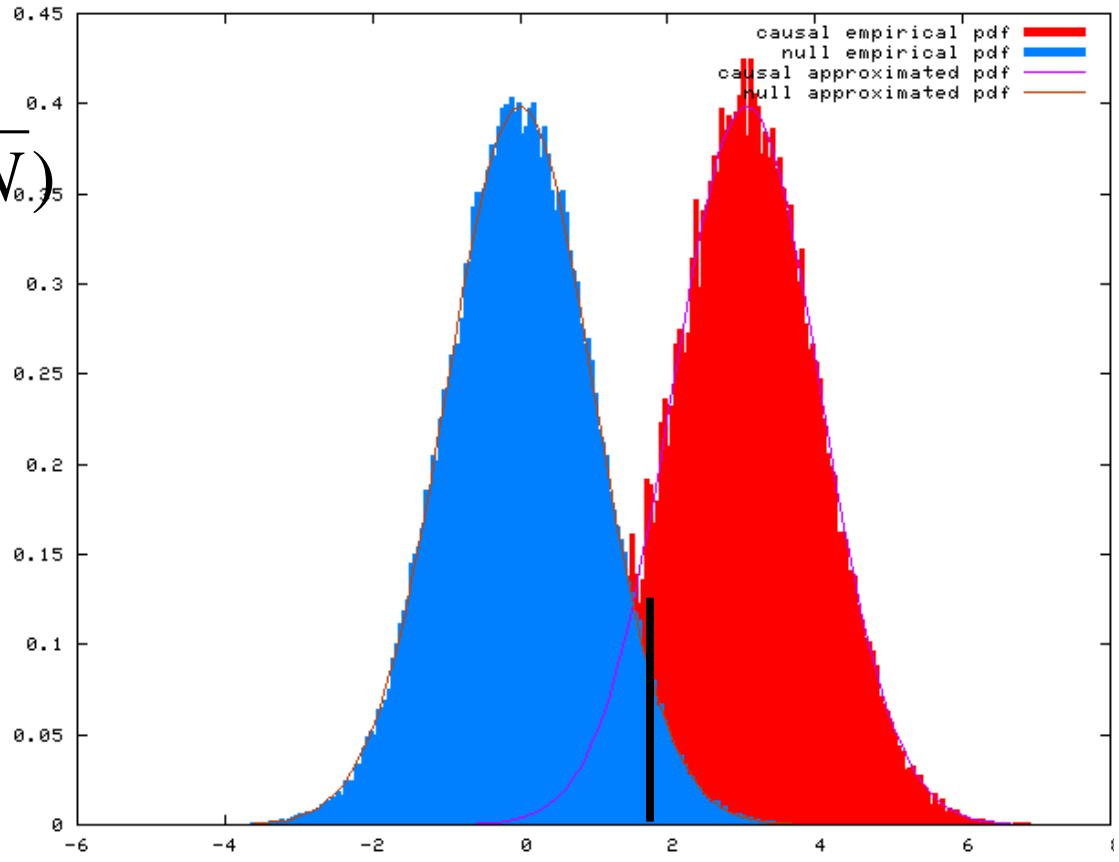
$$+1 - \Phi(-\Phi^{-1}(\alpha/2) + 2.84)$$

$$= \Phi(-1.95 + 2.84)$$

$$+1 - \Phi(1.95 + 2.84)$$

$$= \Phi(.89) + 1 - \Phi(4.79)$$

$$= .81 + 1 - .9999992 = .81$$



Relative Risk Examples

- Let the relative risk =1.5
- Let F=0.001 (small value)
- Let p_A=.2

$$p_A^+ = \frac{\gamma p_A}{(\gamma - 1)p_A + 1} = \frac{1.5 * .2}{(1.5 - 1).2 + 1} = .273 \quad p_A^- = p_A = .2$$

- We use p^+_A and p^-_A to estimate power.

Relative Risk Examples

- Let the relative risk =2.0
- Let F=0.001 (small value)
- Let p_A=.2

$$p_A^+ = \frac{\gamma p_A}{(\gamma - 1)p_A + 1} = \frac{2 * .2}{(2-1).2 + 1} = .333 \quad p_A^- = p_A = .2$$

- We use p^+_A and p^-_A to estimate power.

Relative Risk Examples

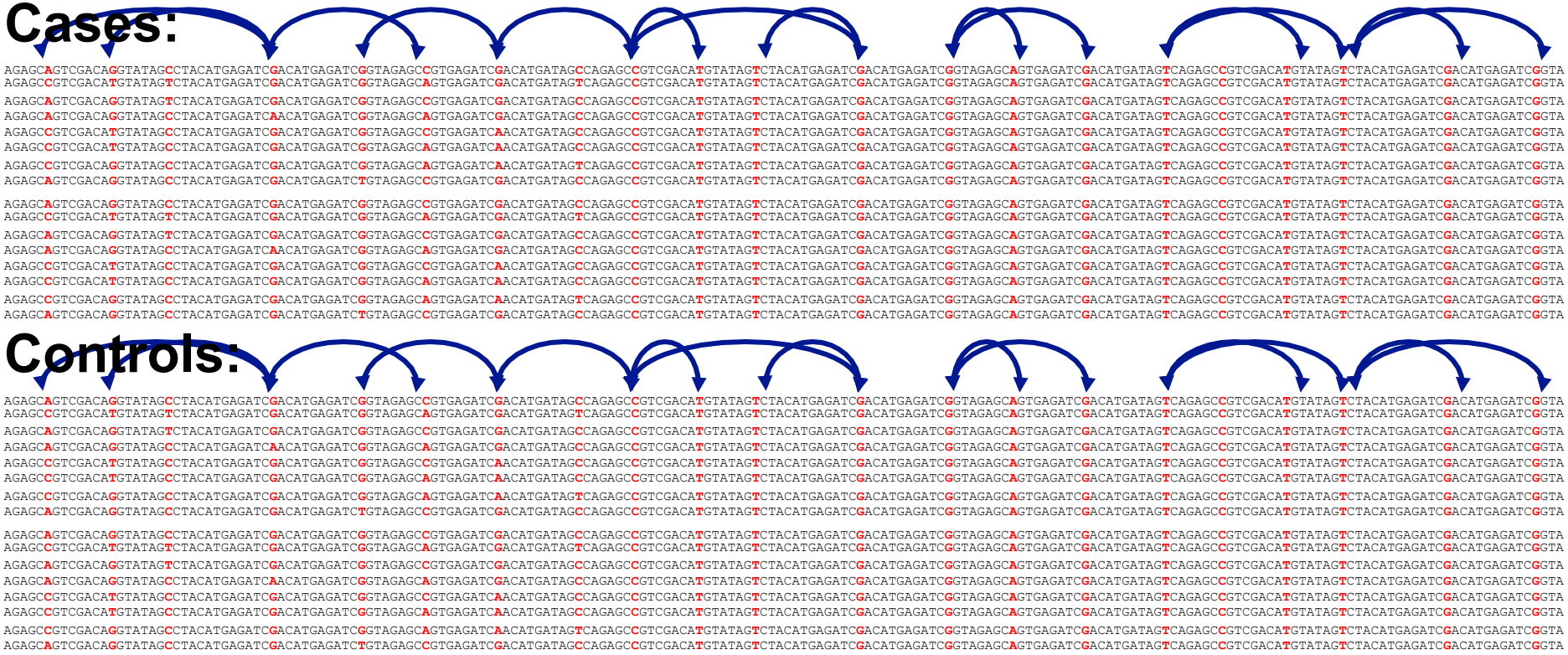
- Let the relative risk =1.5
- Let F=0.001 (small value)
- Let p_A=.4

$$p_A^+ = \frac{\gamma p_A}{(\gamma - 1)p_A + 1} = \frac{1.5 * .4}{(1.5 - 1).4 + 1} = .5 \quad p_A^- = p_A = .4$$

- We use p^+_A and p^-_A to estimate power.

Power for Correlated SNPs

Cases:



Controls:

- Key Assumption: Each SNP is equally likely to be causal.
- What is power of a study with multiple correlated SNPs?
- How does correlation affect power?

Indirect Association

- Now lets assume that we have 2 markers, A and B. Let us assume that marker B is the causal mutation, but we are observing marker A.
- If we observed marker B directly our statistic would be

$$\lambda_B = \frac{(p_B^+ - p_B^-)}{\sqrt{2p_B(1-p_B)}} \qquad S_B \sim N\left(\lambda_B\sqrt{N},1\right)$$

Indirect Association

- However, we are observing A where our statistic is

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

- What is the relation between S_A and S_B?

Fundamental Assumptions of HapMap Analysis

- In case and control samples, the correlation patterns are the same.
  - **Vague Definition.**
  - **Slippery Slope!**
- We use the following precise assumption:

  If marker B is causal with different allele frequencies, then the conditional distributions:

  $p^{+}_{A|B}=p^{-}_{A|B}=p_{A|B}$.

  Note that $p^{+}_{B|A}\neq p^{-}_{B|A}$!!!!

Indirect Association

- We want to relate

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

- to

$$\lambda_B = \frac{(p_B^+ - p_B^-)}{\sqrt{2p_B(1-p_B)}} \qquad S_B \sim N\left(\lambda_B\sqrt{N},1\right)$$

Indirect Association

- Since conditional probability distributions are equal in case and control samples

$$p_A^+ = p_{AB}^+ + p_{Ab}^+$$

$$p_A^+ = p_B^+ p_{A|B} + (1 - p_B^+) p_{A|b}$$

$$p_A^- = p_B^- p_{A|B} + (1 - p_B^-) p_{A|b}$$

$$p_A^+ - p_A^- = p_{A|B}(p_B^+ - p_B^-) - p_{A|b}(p_B^+ - p_B^-)$$

$$p_A^+ - p_A^- = (p_B^+ - p_B^-)(p_{A|B} - p_{A|b})$$

Indirect Association

- Then

$$\lambda_A = \frac{(p_A^+ - p_A^-)}{\sqrt{2p_A(1-p_A)}} = \frac{(p_B^+ - p_B^-)(p_{A|B} - p_{A|b})}{\sqrt{2p_A(1-p_A)}}$$

$$= \frac{(p_B^+ - p_B^-)(p_{A|B} - p_{A|b})}{\sqrt{2p_A(1-p_A)}} \frac{\sqrt{2p_B(1-p_B)}}{\sqrt{2p_B(1-p_B)}}$$

$$= \frac{(p_B^+ - p_B^-)}{\sqrt{2p_B(1-p_B)}} \frac{(p_{A|B} - p_{A|b})\sqrt{2p_B(1-p_B)}}{\sqrt{2p_A(1-p_A)}}$$

$$= \lambda_B \frac{(p_{A|B} - p_{A|b})\sqrt{2p_B(1-p_B)}}{\sqrt{2p_A(1-p_A)}}$$

Indirect Association

$$\lambda_A = \lambda_B \frac{(p_{A|B} - p_{A|b})\sqrt{2p_B(1-p_B)}}{\sqrt{2p_A(1-p_A)}}$$

- Note that

$$\lambda_A = \lambda_B \sqrt{r^2}$$

$$= \lambda_B \frac{\left(\frac{p_{AB}}{p_B} - \frac{p_{Ab}}{1-p_B}\right)\sqrt{p_B(1-p_B)}}{\sqrt{p_A(1-p_A)}}$$

$$= \lambda_B \frac{\left(\frac{p_{AB} - p_{AB}p_B - p_{Ab}p_B}{p_B(1-p_B)}\right)\sqrt{p_B(1-p_B)}}{\sqrt{p_A(1-p_A)}}$$

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

Indirect Association

- How many individuals, N_A, do we need to collect at marker A to achieve the same power as if we collected N_B individualsat marker B.

$$S_A \sim N\left(\lambda_A \sqrt{N_A}, 1\right)$$

$$S_B \sim N\left(\lambda_B \sqrt{N_B}, 1\right)$$

$$\lambda_A = \lambda_B \sqrt{r^2}$$

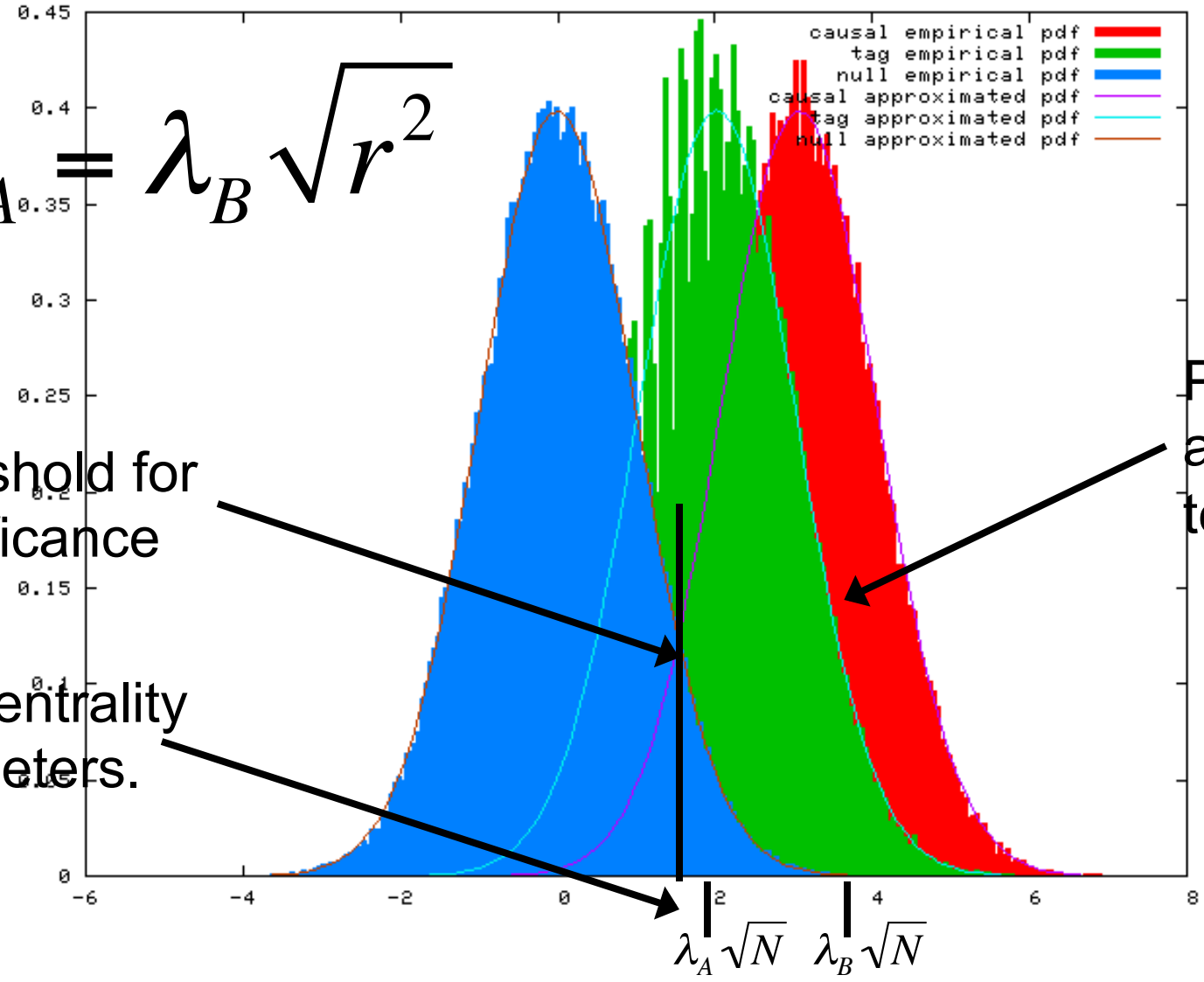
$$\lambda_A \sqrt{N_A} = \lambda_B \sqrt{N_B}$$

$$\lambda_B \sqrt{r^2} \sqrt{N_A} = \lambda_B \sqrt{N_B}$$

$$N_A = \frac{N_B}{r^2}$$

Visualization in terms of Power

$$\lambda_A = \lambda_B \sqrt{r^2}$$



Threshold for significance

Non-centrality parameters.

Power of association test

Indirect Association Power Example

- Significance Threshold α=0.05
- CollectSNP B with alleles {B,b}.
- Assume true case/control probabilities are 0.6/0.5 at SNP A.
- Assume causal SNP A and r^2_{AB}=.8
- 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

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

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

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

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

$$S_B \sim N(\lambda_B \sqrt{N}, 1) = N(\sqrt{r_{AB}^2}\,\lambda_A \sqrt{N}, 1)$$

$$S_B \sim N(\sqrt{.8} * 4.02, 1) = N(3.56, 1)$$

Power Example

power

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

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

$$= \Phi(\Phi^{-1}(0.025) + 3.56)$$

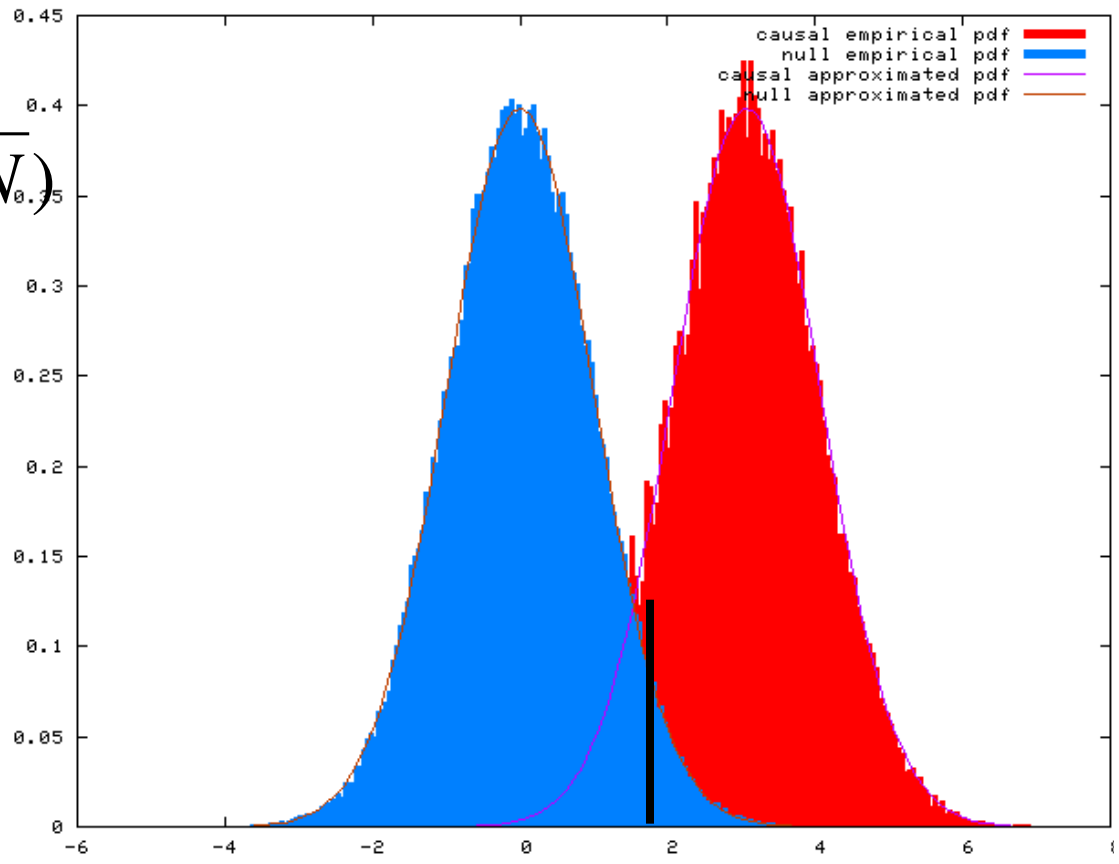
$$+1 - \Phi(-\Phi^{-1}(\alpha/2) + 3.56)$$

$$= \Phi(-1.95 + 3.56)$$

$$+1 - \Phi(1.95 + 3.56)$$

$$= \Phi(1.61) + 1 - \Phi(5.17)$$

$$= .95 + 1 - 1 = .95$$



Association with Multiple SNPs

- Let $\hat{p}_1^+, \hat{p}_2^+, \ldots, \hat{p}_M^+$ and $\hat{p}_1^-, \hat{p}_2^-, \ldots, \hat{p}_M^-$ be the allele frequencies in the cases and controls.
- Association statistic S_i is

$$S_i = \frac{\hat{P}^+_i - \hat{P}^-_i}{\sqrt{2/N}\sqrt{\hat{p}_i(1-\hat{p}_i)}} \sim N\left(\frac{p_i^+ - p_i^-}{\sqrt{2/N}\sqrt{p_i(1-p_i)}}, 1\right)$$

- We compute each S_i and take the most significant S_{max}.
- If $\Phi(-|S_{max}|) < \alpha_s/2$, association at S_{max} is significant.
- Question: How do we set α_s, the per-SNP threshold.
- Question: What is the power of such a study?

Multiple SNP Association

- Three Main Questions

  1. How do we analyze associations with multiple SNPs?
     - **Multiple hypothesis testing**

  2. How do we compute the power of associations with multiple SNPs?
     - **Assume each SNP has equal probability of being causal.**
     - **Compute the "Average" Power.**
     - **Take into account indirect associations.**

  3. How do we design multiple SNP studies?
     - **Goal: Maximize association study power.**
     - **Tag SNP selection.**

Multiple Hypothesis Testing

- Motivating Example:
  - **Chance to flip 10 heads in a row is 1/1024.**
  - **Chance to flip 10 heads in a row in 1,000 attempts is 62%.**
  - **Chance to flip 15 heads in a row in 1,000 attempts is 3%.**
  - **Chance to flip 16 heads in a row in 1,000 attempts is 1.5%.**
  - **Chance to flip 17 heads in a row in 1,000 attempts is 0.07%.**
- If threshold for association at each SNP is α_s, the chance of a significant association at any SNP is greater than α_s.
- Key question: How do we set α_s at each marker so that overall false positive rate is α.
- In other words: How significant must the association at a SNP be so that it is significant when considering all SNPs?

Multiple SNP Association

- Let $\hat{p}_1^+, \hat{p}_2^+, \ldots, \hat{p}_N^+$ and $\hat{p}_1^-, \hat{p}_2^-, \ldots, \hat{p}_N^-$ be the allele frequencies in the cases and controls.
- Association statistic S_i is

$$S_i = \frac{\hat{P}^+_i - \hat{P}^-_i}{\sqrt{2/N}\sqrt{\hat{p}_i(1-\hat{p}_i)}} \sim N\left(\frac{p_i^+ - p_i^-}{\sqrt{2/N}\sqrt{p_i(1-p_i)}}, 1\right)$$

- We compute each S_i and take the most significant S_{max}.
- Question: For correlated markers, what is the p-value for the most significant S_{max}?

Multiple SNP Association

- Given M SNPs.
- Let A_i be the event that SNP i is significant under null hypothesis.

  $P(A_i)=\alpha_s$
- Let $\neg A_i$ be the event that SNP i is not significant under null hypothesis.

  $P(\neg A_i)=1-\alpha_s$
- Probability that any SNP is significant under null hypothesis

  $\alpha = P(A_1 \text{ or } A_2 \text{ or } \ldots \text{ or } A_n)$

  $\alpha = 1-P(\neg A_1 \text{ and } \neg A_2 \text{ and } \ldots \text{ and } \neg A_n)$
- Assumption: All SNPs are independent.

  $P(\neg A_1 \text{ and } \neg A_2)=P(\neg A_1)P(\neg A_2)$

  $\alpha = 1-\prod P(\neg A_i)$

Sidak and Bonferroni Correction

- Assuming SNPs are independent

  $\alpha = 1-\prod P(\neg A_i)=1-\prod(1-\alpha_s)=1-(1-\alpha_s)^M$

- Sidak Correction

  $$\alpha_s = 1 - \sqrt[M]{1-\alpha}$$

- Bonferroni Correction

$$\alpha = 1-(1-\alpha_s)^M = 1-1+M\alpha_s - \frac{M(M-1)}{2}\alpha_s^2 + \ldots \approx M\alpha_s$$

$$\alpha_s \approx \frac{\alpha}{M} \leq 1 - \sqrt[M]{1-\alpha}$$

MultiSNP Association Example

- Collect data at 5 SNPs
- Significance Threshold α=0.05
- Sample: 100 Cases and 100 Controls
- Total of 200 Case Chromosomes and 200 Control Chromosomes

$\hat{p}_1^+ = \frac{120}{200} = .6$ $\hat{p}_2^+ = \frac{80}{200} = .4$ $\hat{p}_3^+ = \frac{60}{200} = .3$ $\hat{p}_4^+ = \frac{100}{200} = .5$ $\hat{p}_5^+ = \frac{120}{200} = .6$

$\hat{p}_1^- = \frac{100}{200} = .5$ $\hat{p}_2^- = \frac{75}{200} = .375$ $\hat{p}_3^- = \frac{65}{200} = .325$ $\hat{p}_4^- = \frac{95}{200} = .475$ $\hat{p}_5^- = \frac{125}{200} = .625$

$\hat{p}_1 = .55$ $\hat{p}_2 = .3825$ $\hat{p}_3 = .3125$ $\hat{p}_4 = .4875$ $\hat{p}_5 = .6125$

MultiSNP Association Example

$$\hat{p}_1^+ = \frac{120}{200} = .6 \quad \hat{p}_2^+ = \frac{80}{200} = .4 \quad \hat{p}_3^+ = \frac{60}{200} = .3 \quad \hat{p}_4^+ = \frac{100}{200} = .5 \quad \hat{p}_5^+ = \frac{120}{200} = .6$$

$$\hat{p}_1^- = \frac{100}{200} = .5 \quad \hat{p}_2^- = \frac{75}{200} = .375 \quad \hat{p}_3^- = \frac{65}{200} = .325 \quad \hat{p}_4^- = \frac{95}{200} = .475 \quad \hat{p}_5^- = \frac{125}{200} = .625$$

$$\hat{p}_1 = .55 \quad \hat{p}_2 = .3825 \quad \hat{p}_3 = .3125 \quad \hat{p}_4 = .4875 \quad \hat{p}_5 = .6125$$

$$S_1 = \frac{.6 - .5}{\sqrt{2/200}\sqrt{.55(1-.55)}} = 2.01 \qquad S_2 = \frac{.4 - .375}{\sqrt{2/200}\sqrt{.3825(1-.3825)}} = .514 \qquad S_3 = \frac{.3 - .325}{\sqrt{2/200}\sqrt{.3125(1-.3125)}} = -.54$$

$$S_4 = \frac{.5 - .475}{\sqrt{2/200}\sqrt{.4875(1-.4875)}} = .500 \qquad S_5 = \frac{.6 - .625}{\sqrt{2/200}\sqrt{.6125(1-.6125)}} = -0.513$$

$S_1=S_{max}=2.01$ (Is this significant?)

Per-marker threshold $\alpha_s=\alpha/5=0.01$ (Bonferroni)

$-\Phi^{-1}(0.01/2)=2.57$

Association is not significant

MultiSNP Association Example

- Collect data at 5 SNPs
- Significance Threshold α=0.05
- Sample: 1000 Cases and 1000 Controls
- Total of 2000 Case Chromosomes and 2000 Control Chromosomes

$\hat{p}_1^+ = \frac{1200}{2000} = .6$ $\hat{p}_2^+ = \frac{800}{2000} = .4$ $\hat{p}_3^+ = \frac{600}{2000} = .3$ $\hat{p}_4^+ = \frac{1000}{2000} = .5$ $\hat{p}_5^+ = \frac{1200}{2000} = .6$

$\hat{p}_1^- = \frac{1000}{2000} = .5$ $\hat{p}_2^- = \frac{750}{2000} = .375$ $\hat{p}_3^- = \frac{650}{2000} = .325$ $\hat{p}_4^- = \frac{950}{2000} = .475$ $\hat{p}_5^- = \frac{1250}{2000} = .625$

$\hat{p}_1 = .55$ $\hat{p}_2 = .3825$ $\hat{p}_3 = .3125$ $\hat{p}_4 = .4875$ $\hat{p}_5 = .6125$

MultiSNP Association Example

$$\hat{p}_1^+ = \frac{1200}{2000} = .6 \quad \hat{p}_2^+ = \frac{800}{2000} = .4 \quad \hat{p}_3^+ = \frac{600}{2000} = .3 \quad \hat{p}_4^+ = \frac{1000}{2000} = .5 \quad \hat{p}_5^+ = \frac{1200}{2000} = .6$$

$$\hat{p}_1^- = \frac{1000}{2000} = .5 \quad \hat{p}_2^- = \frac{750}{2000} = .375 \quad \hat{p}_3^- = \frac{650}{2000} = .325 \quad \hat{p}_4^- = \frac{950}{2000} = .475 \quad \hat{p}_5^- = \frac{1250}{2000} = .625$$

$$\hat{p}_1 = .55 \quad \hat{p}_2 = .3825 \quad \hat{p}_3 = .3125 \quad \hat{p}_4 = .4875 \quad \hat{p}_5 = .6125$$

$$S_1 = \frac{.6 - .5}{\sqrt{2/2000}\sqrt{.55(1-.55)}} = 6.36 \quad S_2 = \frac{.4 - .375}{\sqrt{2/2000}\sqrt{.3825(1-.3825)}} = 1.63 \quad S_3 = \frac{.3 - .325}{\sqrt{2/2000}\sqrt{.3125(1-.3125)}} = -1.71$$

$$S_4 = \frac{.5 - .475}{\sqrt{2/2000}\sqrt{.4875(1-.4875)}} = 1.58 \quad S_5 = \frac{.6 - .625}{\sqrt{2/2000}\sqrt{.6125(1-.6125)}} = -1.62$$

$S_1=S_{max}=6.36$ (Is this significant?)

Per-marker threshold $\alpha_s=\alpha/5=0.01$ (Bonferroni)

$-\Phi^{-1}(0.01/2)=2.57$

Association is significant

Correlated Markers

- If markers are correlated, Sidak (and Bonferroni) correction is conservative.

$$\alpha < 1 - (1 - \alpha_s)^M$$

- Extreme example: assume markers are perfectly correlated.

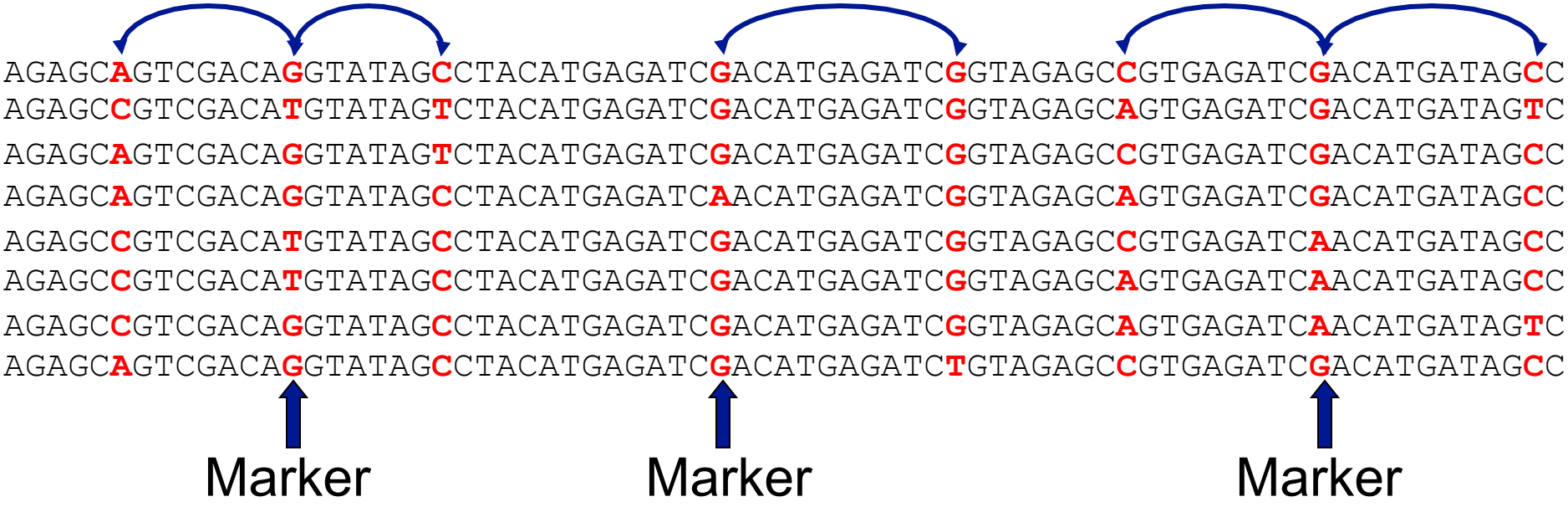
  $P(A_1 \text{ or } A_2 \text{ or } \ldots \text{ or } A_M) = P(A_i)$

$$\alpha = 1 - (1 - \alpha_s) = \alpha_s$$

- In general, it is difficult to analytically compute α_s.
- We can empirically estimate α_s using a technique called permutation.

"Best Tag" Assumption

- Partition of SNPs by Collected Markers



- Best Tag Assumption: If SNP is causal, association is detected only if corresponding marker detects association

Indirect Association

- How many individuals, N_A, do we need to collect at marker A to achieve the same power as if we collected N_B markers at marker B.

$$S_A \sim N\left(\lambda_A \sqrt{N_A}, 1\right)$$

$$S_B \sim N\left(\lambda_B \sqrt{N_B}, 1\right)$$

$$\lambda_A = \lambda_B \sqrt{r^2}$$

$$\lambda_A \sqrt{N_A} = \lambda_B \sqrt{N_B}$$

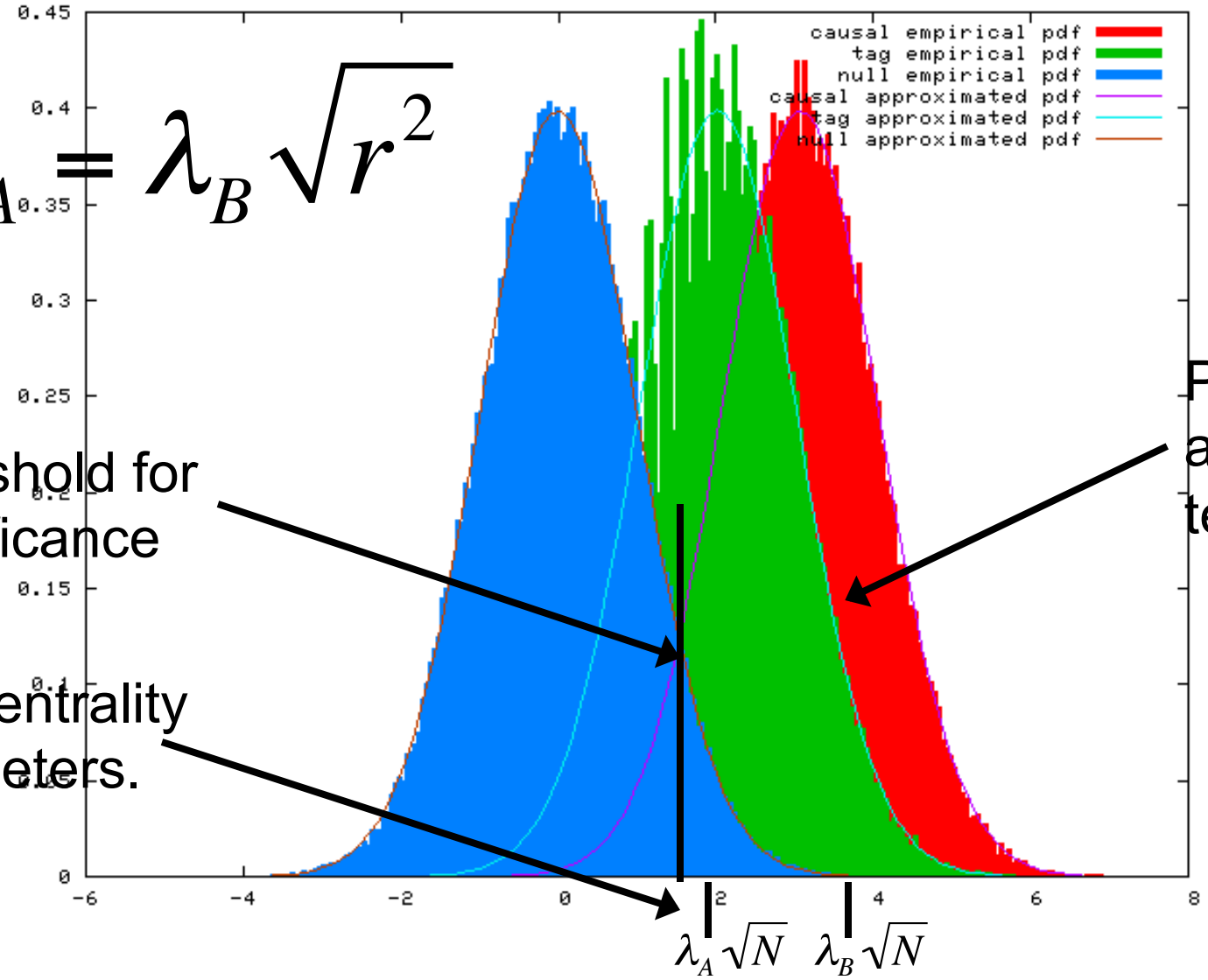
$$\lambda_B \sqrt{r^2} \sqrt{N_A} = \lambda_B \sqrt{N_B}$$

$$N_A = \frac{N_B}{r^2}$$

Visualization in terms of Power

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

$$\lambda_A = \lambda_B\sqrt{r^2}$$



Threshold for significance

Power of association test

Non-centrality parameters.

MultiSNP Association Example

- Collect data at 5 SNPs
- Significance Threshold α=0.05
- Sample: 100 Cases and 100 Controls
- Total of 200 Case Chromosomes and 200 Control Chromosomes

$\hat{p}_1^+ = \frac{120}{200} = .6$ $\hat{p}_2^+ = \frac{80}{200} = .4$ $\hat{p}_3^+ = \frac{60}{200} = .3$ $\hat{p}_4^+ = \frac{100}{200} = .5$ $\hat{p}_5^+ = \frac{120}{200} = .6$

$\hat{p}_1^- = \frac{100}{200} = .5$ $\hat{p}_2^- = \frac{75}{200} = .375$ $\hat{p}_3^- = \frac{65}{200} = .325$ $\hat{p}_4^- = \frac{95}{200} = .475$ $\hat{p}_5^- = \frac{125}{200} = .625$

$\hat{p}_1 = .55$ $\hat{p}_2 = .3825$ $\hat{p}_3 = .3125$ $\hat{p}_4 = .4875$ $\hat{p}_5 = .6125$

MultiSNP Association Example

$$\hat{p}_1^+ = \frac{120}{200} = .6 \quad \hat{p}_2^+ = \frac{80}{200} = .4 \quad \hat{p}_3^+ = \frac{60}{200} = .3 \quad \hat{p}_4^+ = \frac{100}{200} = .5 \quad \hat{p}_5^+ = \frac{120}{200} = .6$$

$$\hat{p}_1^- = \frac{100}{200} = .5 \quad \hat{p}_2^- = \frac{75}{200} = .375 \quad \hat{p}_3^- = \frac{65}{200} = .325 \quad \hat{p}_4^- = \frac{95}{200} = .475 \quad \hat{p}_5^- = \frac{125}{200} = .625$$

$$\hat{p}_1 = .55 \quad \hat{p}_2 = .3825 \quad \hat{p}_3 = .3125 \quad \hat{p}_4 = .4875 \quad \hat{p}_5 = .6125$$

$$S_1 = \frac{.6 - .5}{\sqrt{2/200}\sqrt{.55(1-.55)}} = 2.01 \qquad S_2 = \frac{.4 - .375}{\sqrt{2/200}\sqrt{.3825(1-.3825)}} = .514 \qquad S_3 = \frac{.3 - .325}{\sqrt{2/200}\sqrt{.3125(1-.3125)}} = -.54$$

$$S_4 = \frac{.5 - .475}{\sqrt{2/200}\sqrt{.4875(1-.4875)}} = .500 \qquad S_5 = \frac{.6 - .625}{\sqrt{2/200}\sqrt{.6125(1-.6125)}} = -0.513$$

$S_1=S_{max}=2.01$ (Is this significant?)

Per-marker threshold $\alpha_s=\alpha/5=0.01$ (Bonferroni)

$-\Phi^{-1}(0.01/2)=2.57$

Association is not significant

MultiSNP Association Example

- Collect data at 5 SNPs
- Significance Threshold α=0.05
- Sample: 1000 Cases and 1000 Controls
- Total of 2000 Case Chromosomes and 2000 Control Chromosomes

$\hat{p}_1^+ = \frac{1200}{2000} = .6$ $\hat{p}_2^+ = \frac{800}{2000} = .4$ $\hat{p}_3^+ = \frac{600}{2000} = .3$ $\hat{p}_4^+ = \frac{1000}{2000} = .5$ $\hat{p}_5^+ = \frac{1200}{2000} = .6$

$\hat{p}_1^- = \frac{1000}{2000} = .5$ $\hat{p}_2^- = \frac{750}{2000} = .375$ $\hat{p}_3^- = \frac{650}{2000} = .325$ $\hat{p}_4^- = \frac{950}{2000} = .475$ $\hat{p}_5^- = \frac{1250}{2000} = .625$

$\hat{p}_1 = .55$ $\hat{p}_2 = .3825$ $\hat{p}_3 = .3125$ $\hat{p}_4 = .4875$ $\hat{p}_5 = .6125$

MultiSNP Association Example

$$\hat{p}_1^+ = \frac{1200}{2000} = .6 \quad \hat{p}_2^+ = \frac{800}{2000} = .4 \quad \hat{p}_3^+ = \frac{600}{2000} = .3 \quad \hat{p}_4^+ = \frac{1000}{2000} = .5 \quad \hat{p}_5^+ = \frac{1200}{2000} = .6$$

$$\hat{p}_1^- = \frac{1000}{2000} = .5 \quad \hat{p}_2^- = \frac{750}{2000} = .375 \quad \hat{p}_3^- = \frac{650}{2000} = .325 \quad \hat{p}_4^- = \frac{950}{2000} = .475 \quad \hat{p}_5^- = \frac{1250}{2000} = .625$$

$$\hat{p}_1 = .55 \quad \hat{p}_2 = .3825 \quad \hat{p}_3 = .3125 \quad \hat{p}_4 = .4875 \quad \hat{p}_5 = .6125$$

$$S_1 = \frac{.6 - .5}{\sqrt{2/2000}\sqrt{.55(1-.55)}} = 6.36 \quad S_2 = \frac{.4 - .375}{\sqrt{2/2000}\sqrt{.3825(1-.3825)}} = 1.63 \quad S_3 = \frac{.3 - .325}{\sqrt{2/2000}\sqrt{.3125(1-.3125)}} = -1.71$$

$$S_4 = \frac{.5 - .475}{\sqrt{2/2000}\sqrt{.4875(1-.4875)}} = 1.58 \quad S_5 = \frac{.6 - .625}{\sqrt{2/2000}\sqrt{.6125(1-.6125)}} = -1.62$$

$S_1=S_{max}=6.36$ (Is this significant?)

Per-marker threshold $\alpha_s=\alpha/5=0.01$ (Bonferroni)

$-\Phi^{-1}(0.01/2)=2.57$

Association is significant

MultiSNP Power

- Assume that we have 5 independent SNPs, 3 have minor allele frequency of .4 and 2 have a minor allele frequency of .2. Assume that the relative risk of one of them is 2.0 (we do not know which one). Assume that we are collecting 100 case and 100 control individuals. With α=0.05, what is the power of this association study?

MultiSNP Power

- If a SNP with minor allele frequency of .4 is causal, then

$$p_A^+ = \frac{\gamma p}{(\gamma - 1)p + 1} = \frac{2 * .4}{(2-1).4 + 1} = .57 \quad p_A^- = p = .4 \quad p_A = \frac{p_A^+ + p_A^-}{2} = .485$$

- If a SNP with minor allele frequence of .2 is causal, then

$$p_A^+ = \frac{\gamma p}{(\gamma - 1)p + 1} = \frac{2 * .2}{(2-1).2 + 1} = .33 \quad p_A^- = p = .2 \quad p_A = \frac{p_A^+ + p_A^-}{2} = .266$$

MultiSNP Power

- If a SNP with minor allele frequency of .4 is causal, then

$$\lambda_{p=.4}\sqrt{N} = \frac{p_A^+ - p_A^-}{\sqrt{2/N}\sqrt{p_A(1-p_A)}} = \frac{.57 - .4}{\sqrt{2/200}\sqrt{.485(1-.485)}} = 3.4$$

- If a SNP with minor allele frequence of .2 is causal, then

$$\lambda_{p=.2}\sqrt{N} = \frac{p_A^+ - p_A^-}{\sqrt{2/N}\sqrt{p_A(1-p_A)}} = \frac{.33 - .2}{\sqrt{2/200}\sqrt{.266(1-.266)}} = 2.9$$

MultiSNP Power

- If α=0.05, then the per-marker threshold using the Bonferroni correction, $\alpha_s = \alpha/5 = 0.01$.
- The power at a SNP with minor allele frequency 0.4 is

$$\begin{aligned}\text{power} &= \Phi(\Phi^{-1}(\alpha_s/2) + \lambda\sqrt{N}) + 1 - (-\Phi^{-1}(\alpha_s/2) + \lambda\sqrt{N}) \\ &= \Phi(\Phi^{-1}(0.005) + 3.4) + 1 - (-\Phi^{-1}(0.005) + 3.4) \\ &= .795\end{aligned}$$

- At a SNP with minor allele frequency 0.2

$$\begin{aligned}\text{power} &= \Phi(\Phi^{-1}(\alpha_s/2) + \lambda\sqrt{N}) + 1 - (-\Phi^{-1}(\alpha_s/2) + \lambda\sqrt{N}) \\ &= \Phi(\Phi^{-1}(0.005) + 2.9) + 1 - (-\Phi^{-1}(0.005) + 2.9) \\ &= .627\end{aligned}$$

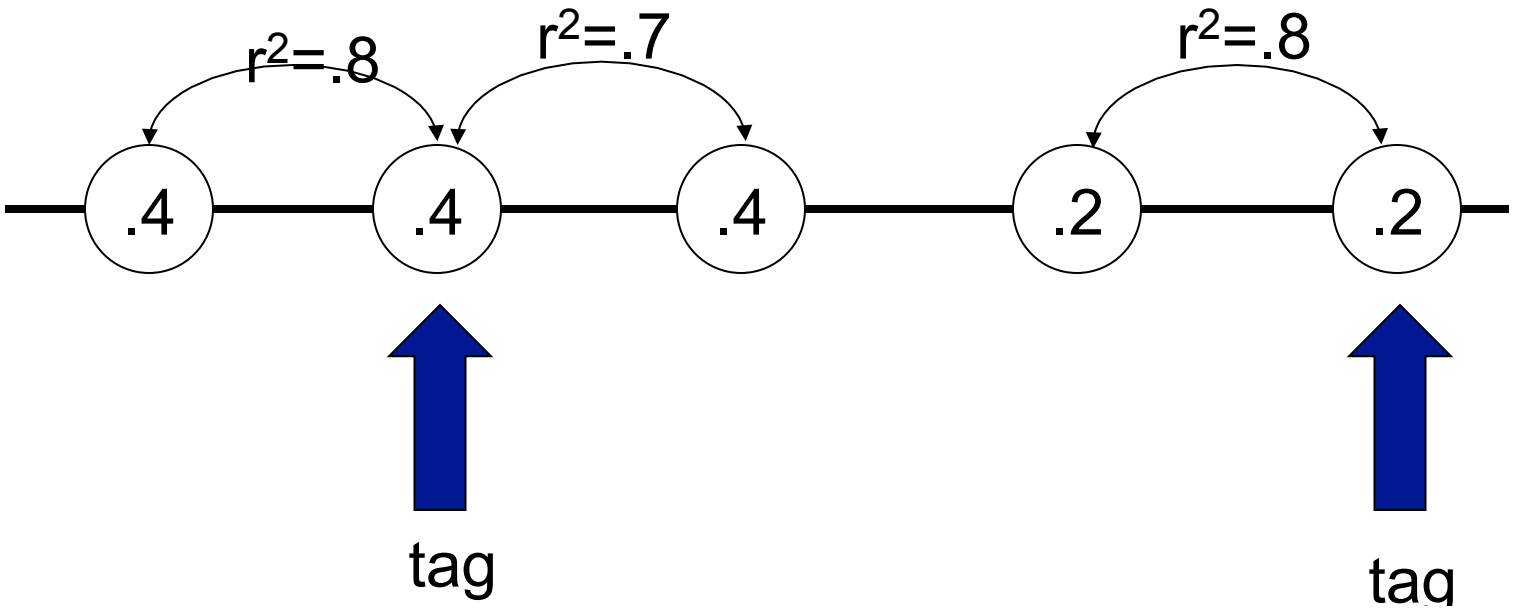
MultiSNP Power

- Since there are 3 SNPs with minor allele frequence 0.4 and 2 SNPs with minor allele frequency 0.2, the total power is

$$\text{total power} = \frac{3 * .795 + 2 * .627}{5} = .728$$

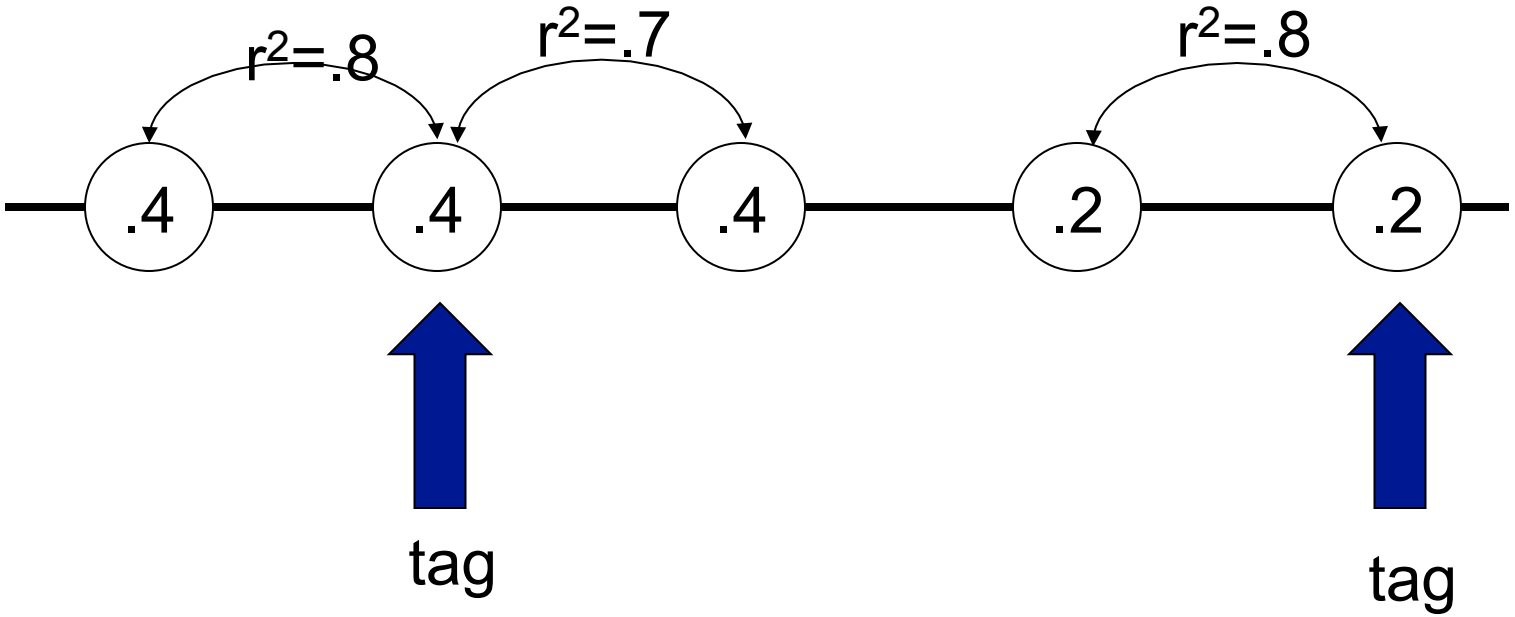
MultiSNP Power with Tags

- Assume you have 5 SNPs, 2 of them are tags. Assume that the relative risk of one of them is 2.0 (we do not know which one). Assume that we are collecting 100 case and 100 control individuals. With α=0.05, what is the power of this association study?



MultiSNP Power with Tags

- Since there are 2 tags, $\alpha_s=\alpha/2=0.05/2=0.025$



Non-centrality parameters

3.4*√.8=3.04 3.4*√1=3.4 3.4*√.7=2.84 2.9*√.8=2.59 2.9*√1=2.9

MultiSNP Power with Tags

$$\text{power at SNP 1} = \Phi(\Phi^{-1}(0.0125) + 3.04) + 1 - (-\Phi^{-1}(0.0125) + 3.04) = .787$$

$$\text{power at SNP 2} = \Phi(\Phi^{-1}(0.0125) + 3.4) + 1 - (-\Phi^{-1}(0.0125) + 3.4) = .877$$

$$\text{power at SNP 3} = \Phi(\Phi^{-1}(0.0125) + 2.84) + 1 - (-\Phi^{-1}(0.0125) + 2.84) = .725$$

$$\text{power at SNP 4} = \Phi(\Phi^{-1}(0.0125) + 2.59) + 1 - (-\Phi^{-1}(0.0125) + 2.59) = .636$$

$$\text{power at SNP 5} = \Phi(\Phi^{-1}(0.0125) + 2.9) + 1 - (-\Phi^{-1}(0.0125) + 2.9) = .745$$

$$\text{total power} = .754$$

What we lose using tags

- Among perfectly correlated SNPs, genotyping only one representative SNP will give the same result as genotyping every SNP.
- Among (not perfectly) correlated SNPs, genotyping only one SNP will reduce the cost, but will decrease the performance of the experiment:
  - **e.g. A certain SNP might show significance using every SNP, but not show significance using tags.**
- We want to make a balance between cost and performance.

Tagging tradeoff

- Hypothetical Situation: Each SNP costs $0.01 to collect. We are interested in 1,000,000 SNPs.
- We can either collect each SNP in 1,000 individuals for a cost of $10,000,000.
- There exists a set of 200,000 tags with minumum r^2 of 0.8 to the 1,000,000 SNPs.
- If we collect 1,000/0.8=1,250 individuals on these 200,000 tags, we will pay $2,500,000 for at least the same power.

Choosing tags

- Find minimum tag SNP set such that SNP is either a tag SNP or has at least r^2 with a tag SNP.
- Pairwise r^2 is directly related to sample size and power of association studies.

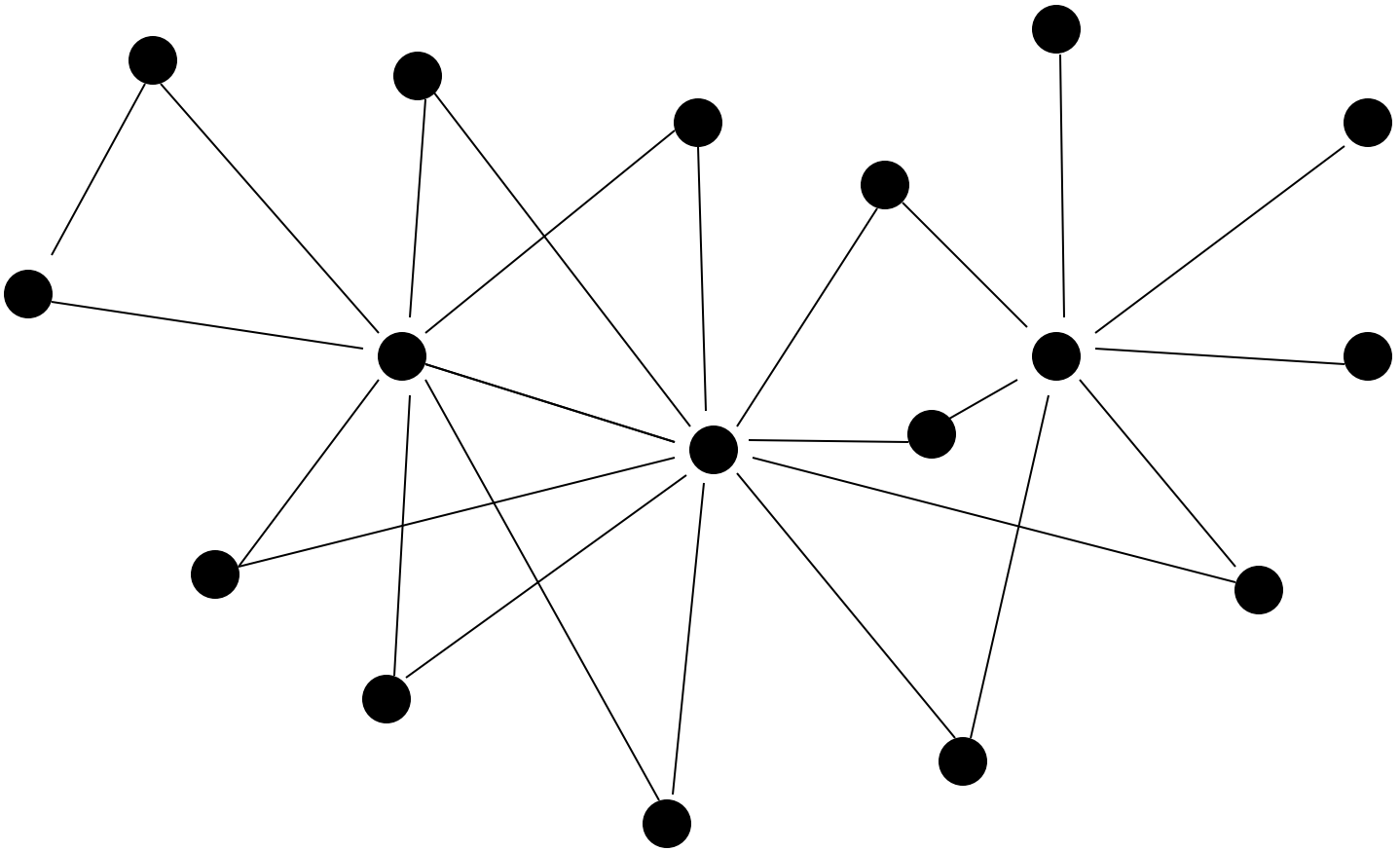
  Pritchard and Przeworski 2001.
- Greedy approach: keep selecting untagged SNP that has at least r^2 correlation with the most remaining untagged SNPs.

  Carlson et al. 2004.

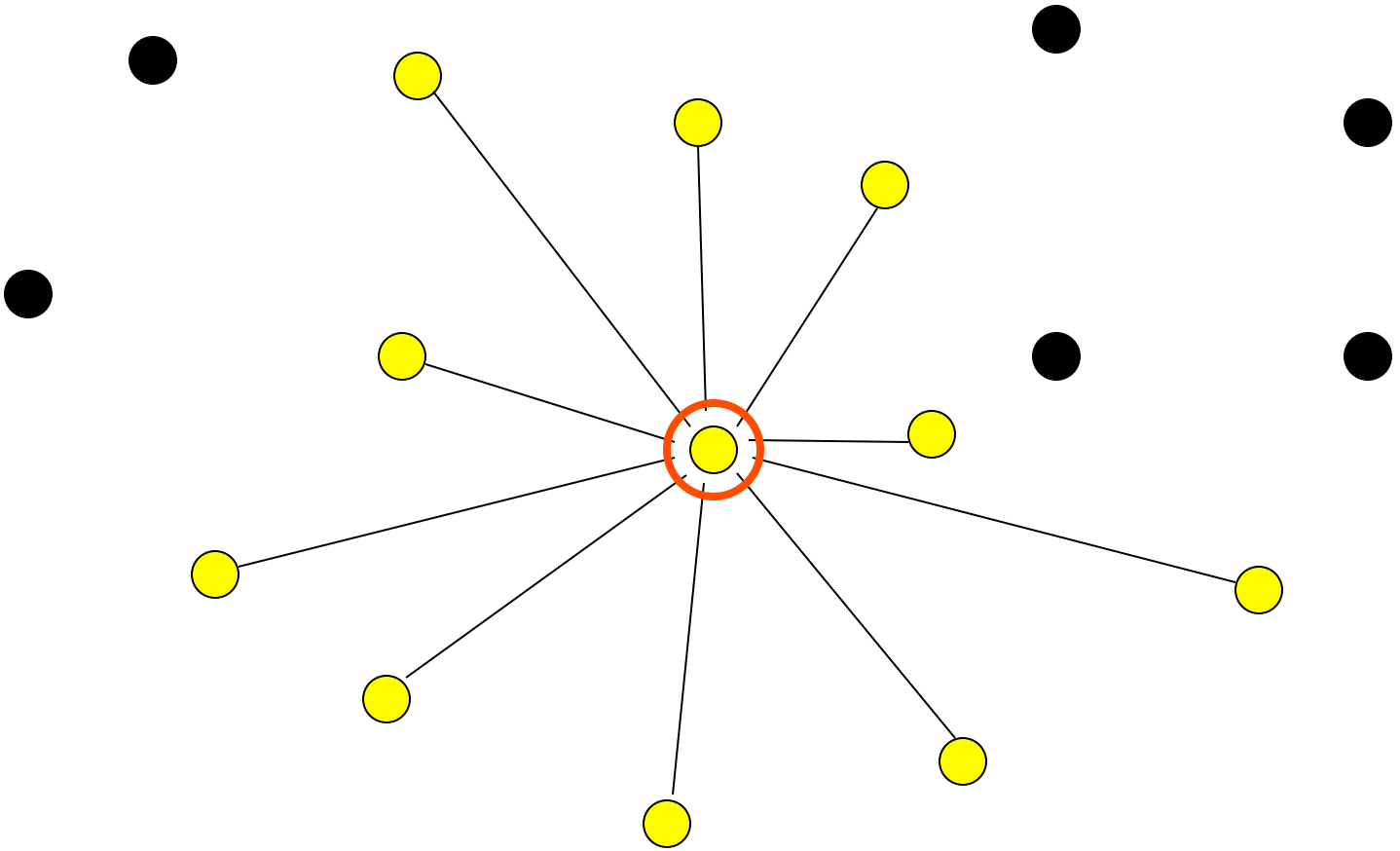
Choosing Tag SNPs
A Computer Science Problem

- NP-Complete Problem
- Connection to Vertex Cover
- Connection to Satisfiability
- Hundreds (or at least a hundred) of methods
- Defined benchmarks
- Some algorithms are better than others

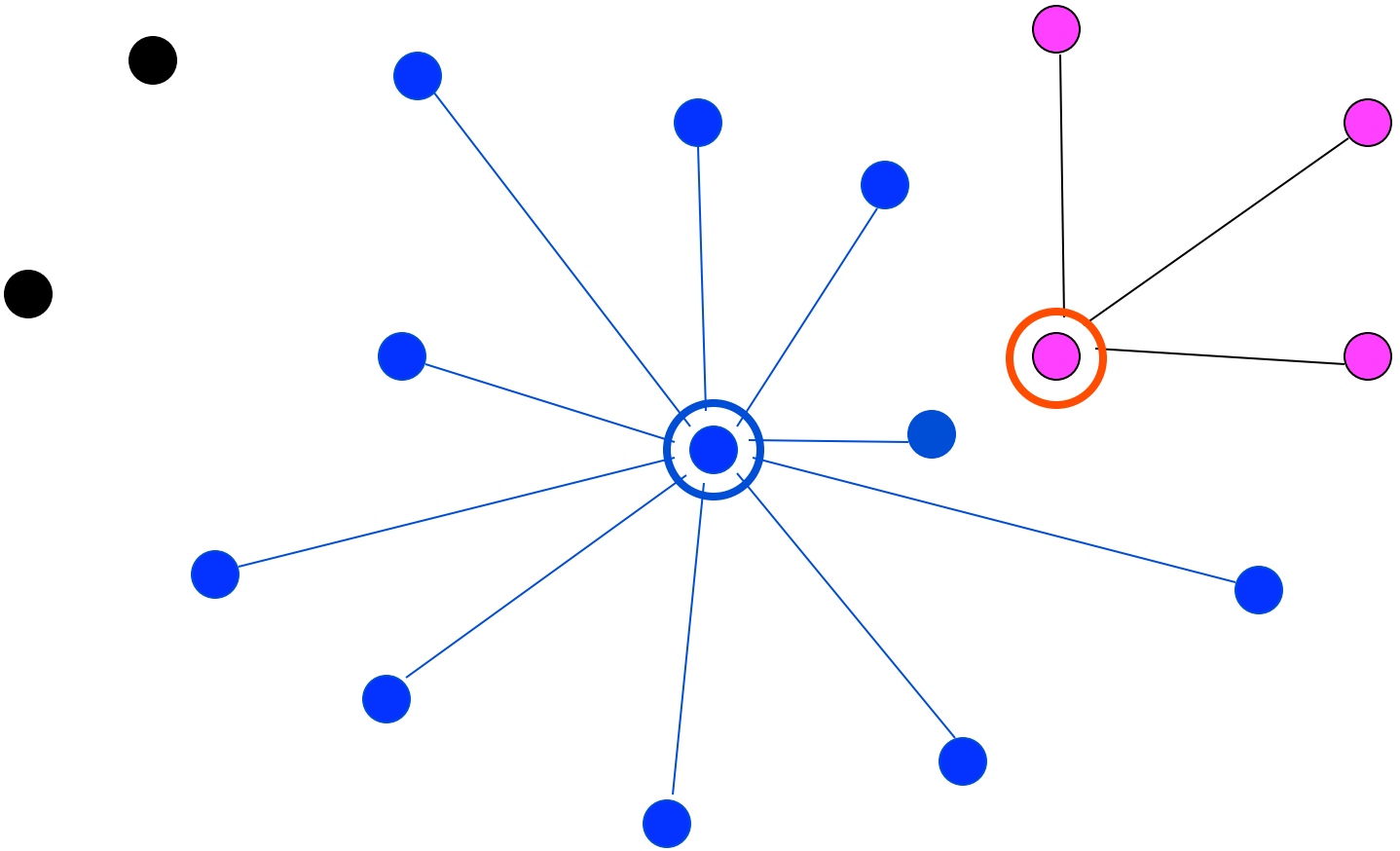
Toy Example



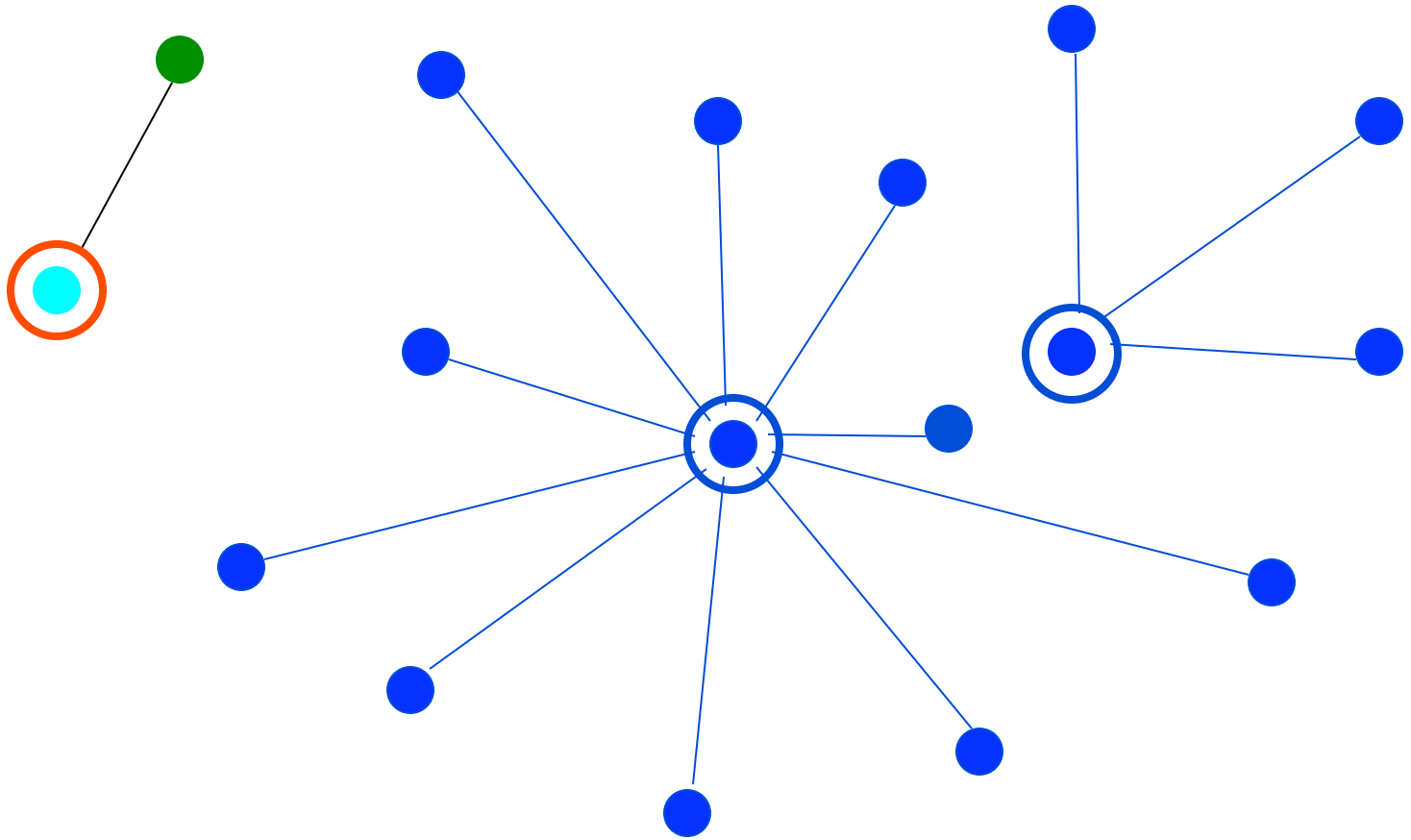
Greedy May Not Be Optimal



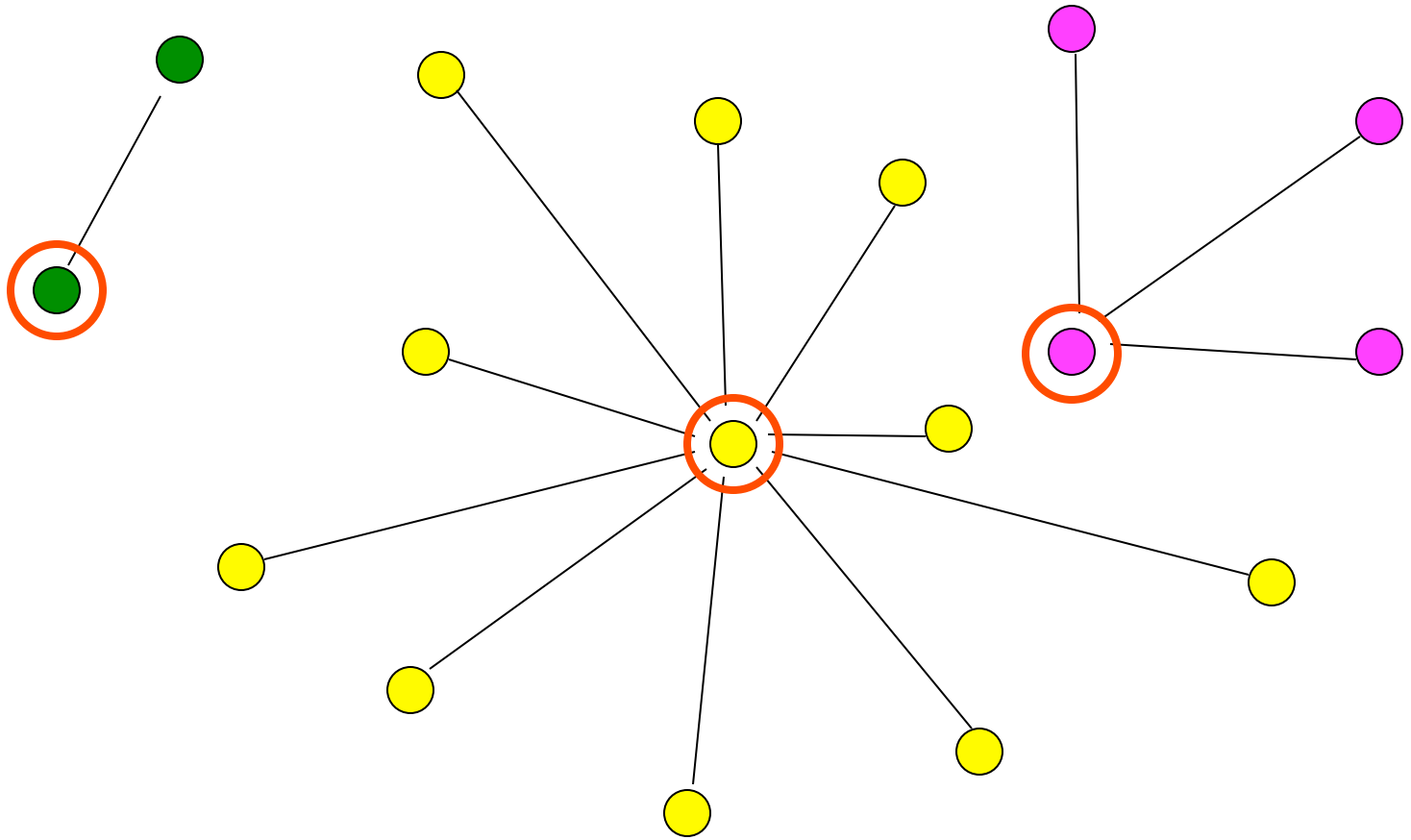
Greedy May Not Be Optimal



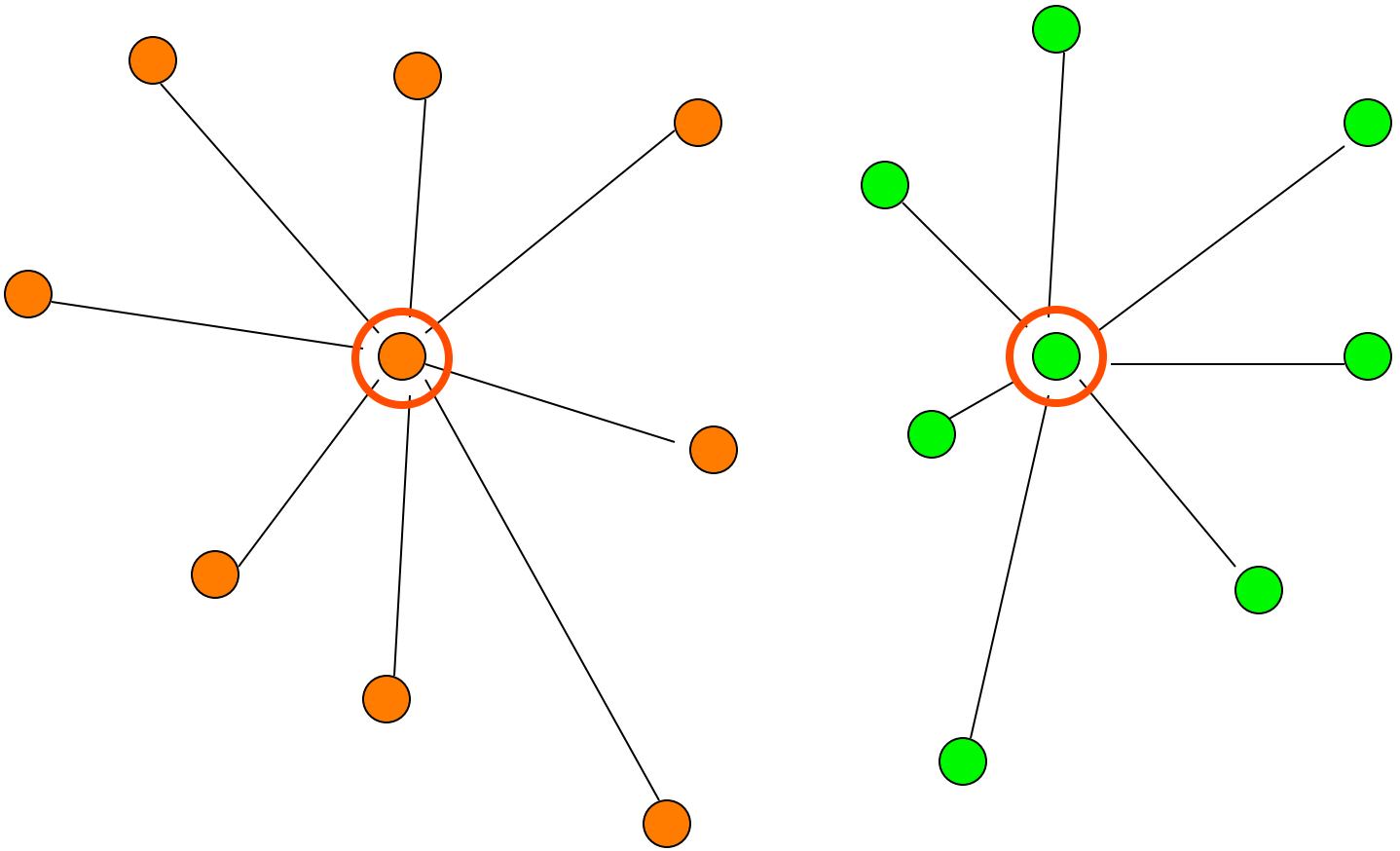
Greedy May Not Be Optimal



Greedy May Not Be Optimal



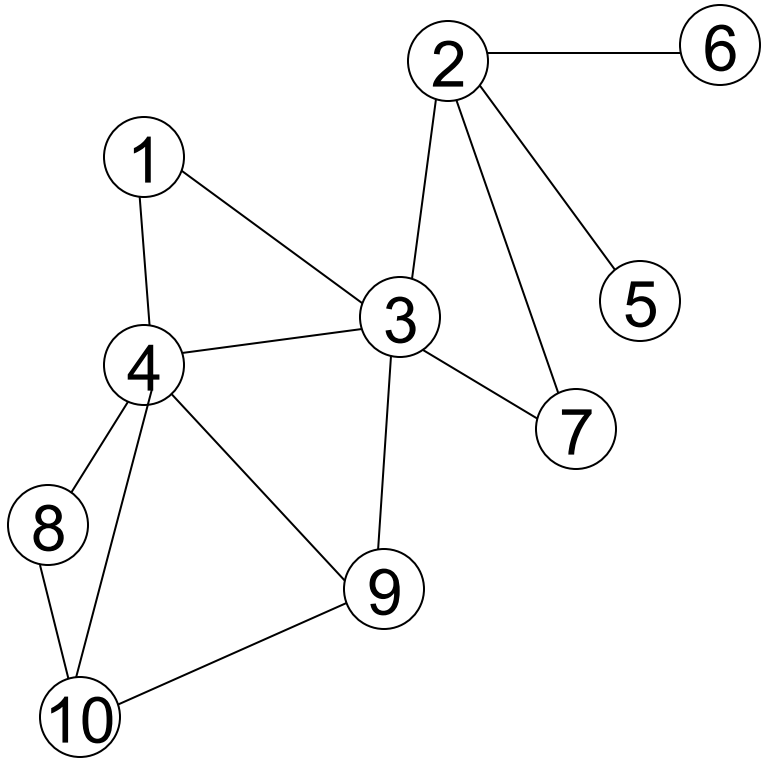
Greedy May Not Be Optimal



Greedy Tag SNP Selection

Nodes are SNPs
Edges denote r^2>.8

Out Degree Counts
1: 2
2: 4
3: 5 (highest)
4: 5 (highest)
5: 1
6: 1
7: 2
8: 2
9: 3
10: 3

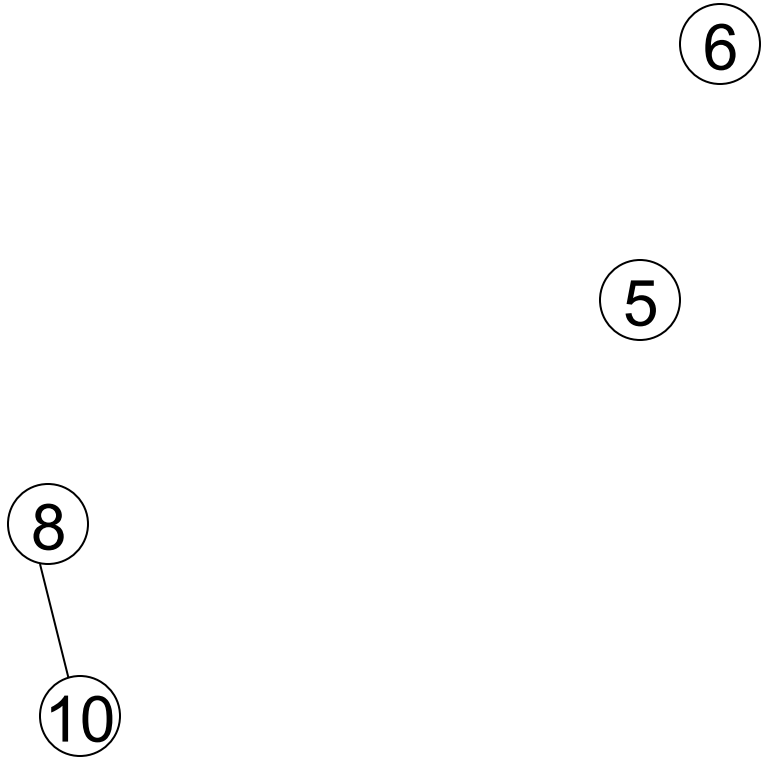


Tags 3

Greedy Tag SNP Selection

Nodes are SNPs
Edges denote r^2>.8

Out Degree Counts
5: 0
6: 0
8: 1 (highest)
10: 1



Tags 3,8

Greedy Tag SNP Selection

Nodes are SNPs
Edges denote r^2>.8

Out Degree Counts
5: 0 (highest)
6: 0

Tags 3,5,8

Greedy Tag SNP Selection

Nodes are SNPs
Edges denote r^2>.8

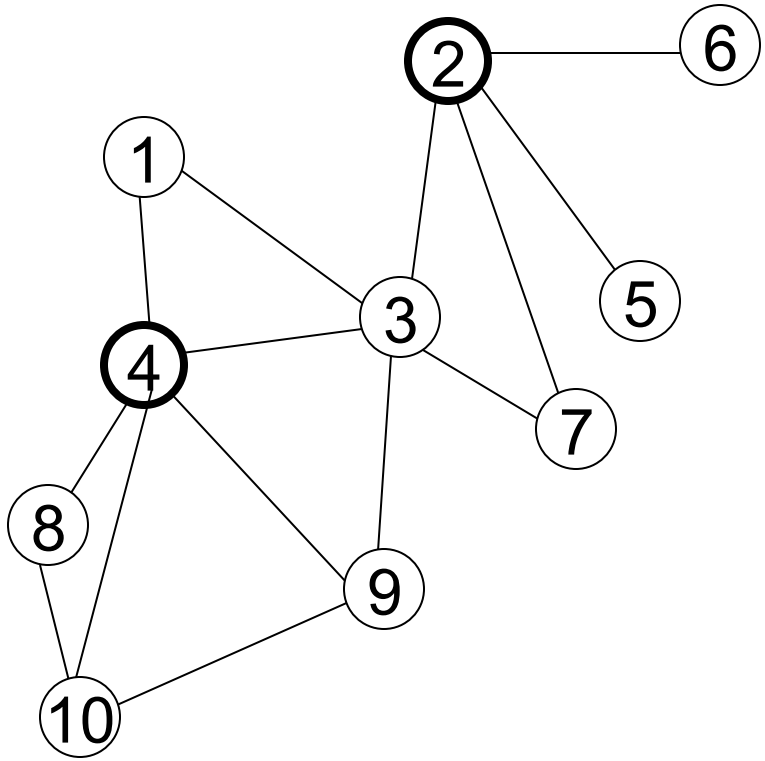
Out Degree Counts
6: 0 (highest)

Tags 3,5,6,8

Optimal Tag SNP Selection

Nodes are SNPs
Edges denote r^2>.8

Out Degree Counts
1: 2
2: 4
3: 5
4: 5
5: 1
6: 1
7: 2
8: 2
9: 3
10: 3



Tags 2 and 4 cover all other SNPs.

Different Objective Function: Power-based tagging

- Assume we know the number of individuals, the penetrance, and the relative risk.
- We can compute the non-centrality parameters between any 2 pairs of SNPs.
- We can then use the non-centrality parameters for choosing pairwise SNPs.

- We will read paper about this next week…

Break!