Computational Genetics
Spring 2014
Lecture 7

Eleazar Eskin

University of California, Los Angeles

Home Work & Midterm

- HW1 Due 4/17/13
- HW2 Due 4/22/13
- Power Tagging Paper Question due Monday (4/21/14)
- Power Tagging Paper Responses due Wednesday (4/23/14)
- Project Selection was due (4/11/14)

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

Midterm Review

Lecture 7.

April 21st , 2014

Midterm

- 60% Applied Problems
  - **Perform Associations (Examples from last class)**
  - **Compute Power for Association**
- 40% Association Derivation Questions

Midterm Questions

1. Given N/2 case individuals and N/2 control individuals. $\hat{p}^{+}_{A}$ and $\hat{p}^{-}_{A}$ are the observed frequencies. If the true frequencies are p^{+}_{A} and p^{-}_{A}, show that the difference of the observed frequencies is normally distributed with mean μ and variance σ^2.
2. Derive a statistic that is a multiple of the allele frequency difference which has variance 1. What is the mean of this statistic?

Midterm Questions

3. Now assume that we are performing an association at SNP A and while the causal mutation is at SNP B. Assume the correlation coefficient between SNPs A and B is r^2. Show power of detecting the association at SNP A by genotyping N/r^2 individuals is equal to the power of detecting the association if we genotyped SNP B with N individuals.

Midterm Questions

- (Grad student only question) Now assume that there are N^+ case and N^- control individuals in the association study. Derive a new statistic that follows the standard normal distribution. What is the power of a study compared to a study with N individuals?

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

$p^+_A(1-p^+_A)+p^-_A(1-p^-_A) \approx 2\,\hat{p}_A(1-\hat{p}_A) \quad \hat{p}_A=(\hat{p}^+_A+\hat{p}^-_A)/2$

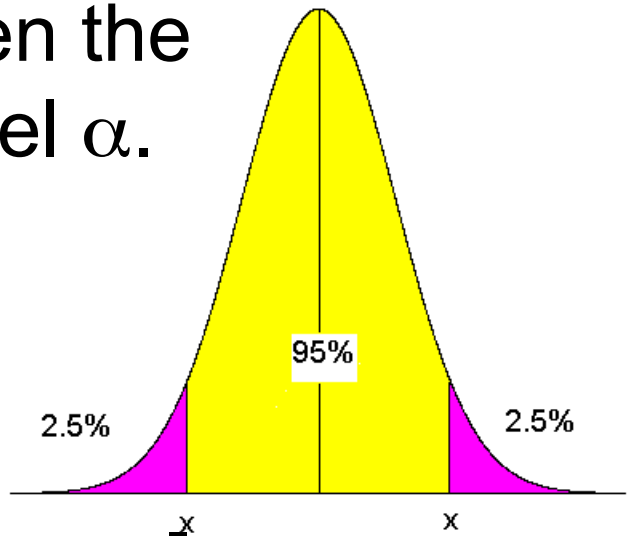
then if $p^+_A = p^-_A$

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

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 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

- Lets assume that $p^+_A \neq p^-_A$ then

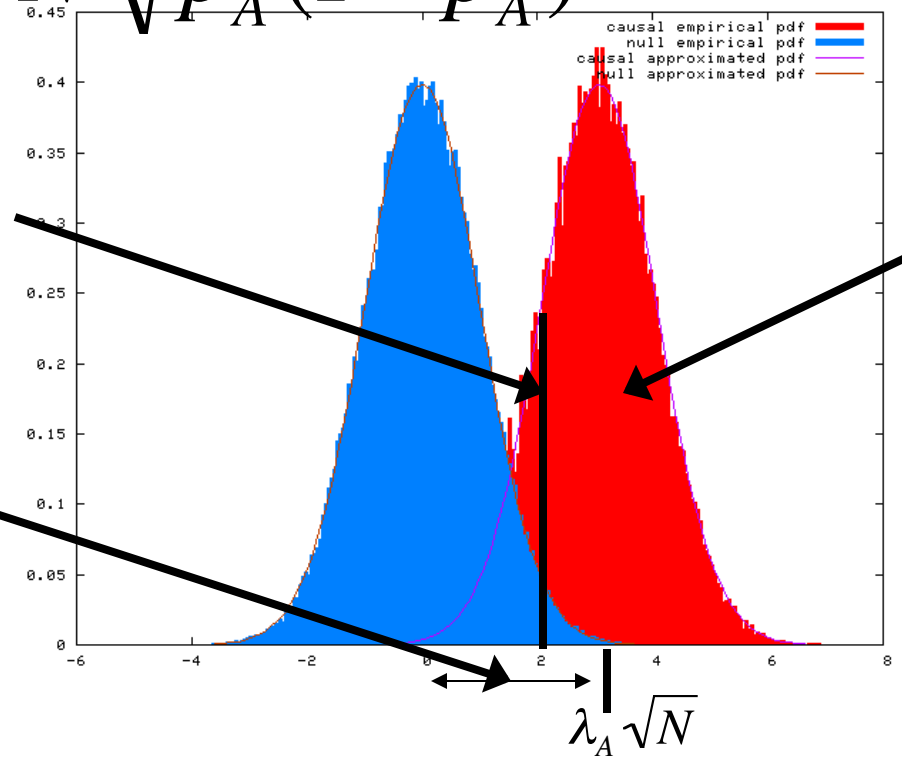
$$S_A = \frac{\hat{p}^+_A - \hat{p}^-_A}{\sqrt{2/N}\sqrt{\hat{p}_A(1-\hat{p}_A)}} \sim N\left(\frac{p^+_A - p^-_A}{\sqrt{2/N}\sqrt{p_A(1-p_A)}},1\right)$$

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

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

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

Power of association test

Non-centrality parameter.

Association Power

- Statistical Power of an association with N individuals, non-centrality parameter $\lambda\sqrt{N}$ and significance threshold α is P($\alpha, \lambda\sqrt{N}$,N)=

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

- Note that if λ=0, power is always α.

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?

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 markers at marker B.

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

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

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

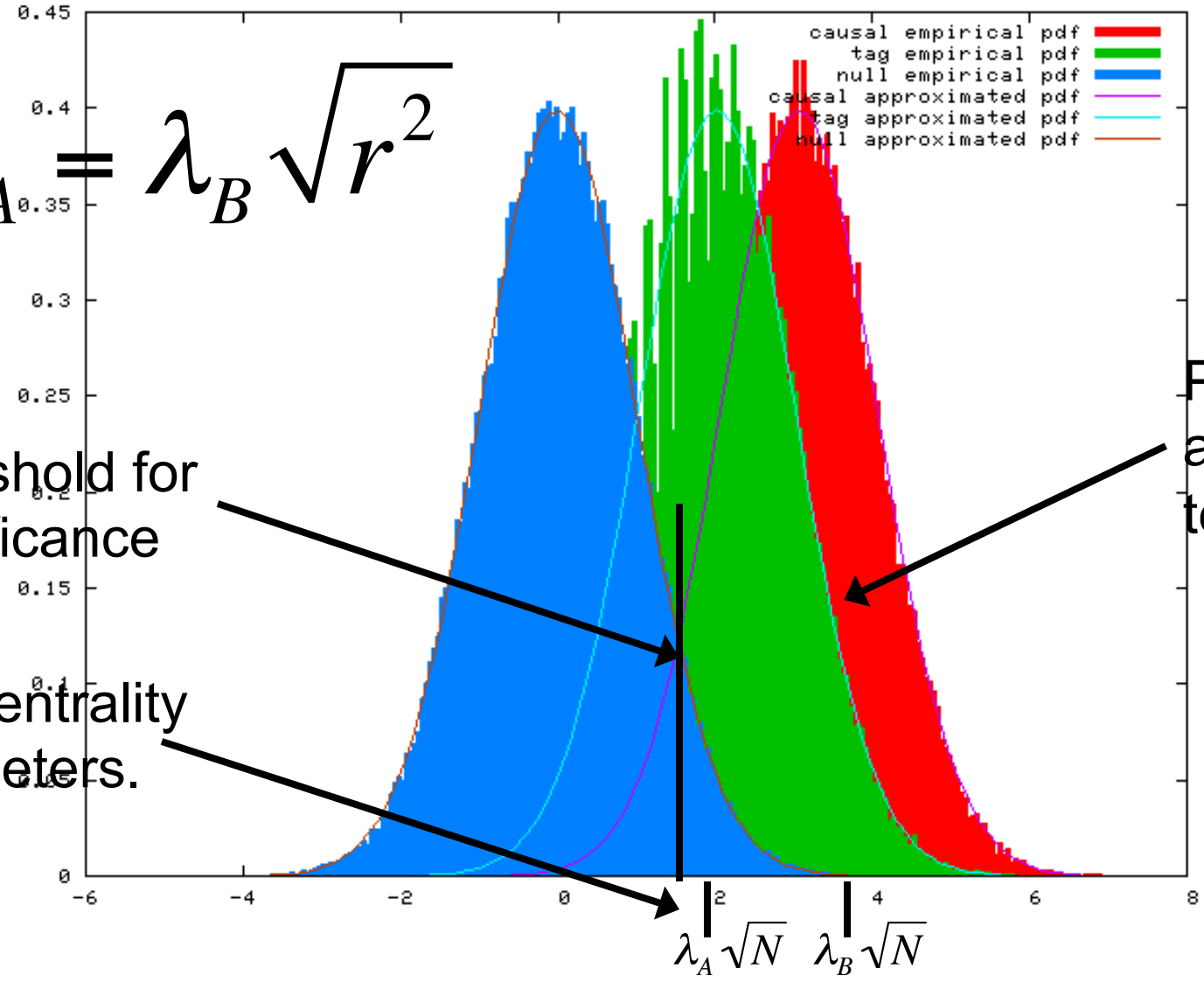
$$\lambda_A \sqrt{N_A} = \lambda_B \sqrt{N_B}$$

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

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

Visualization in terms of Power

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



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

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

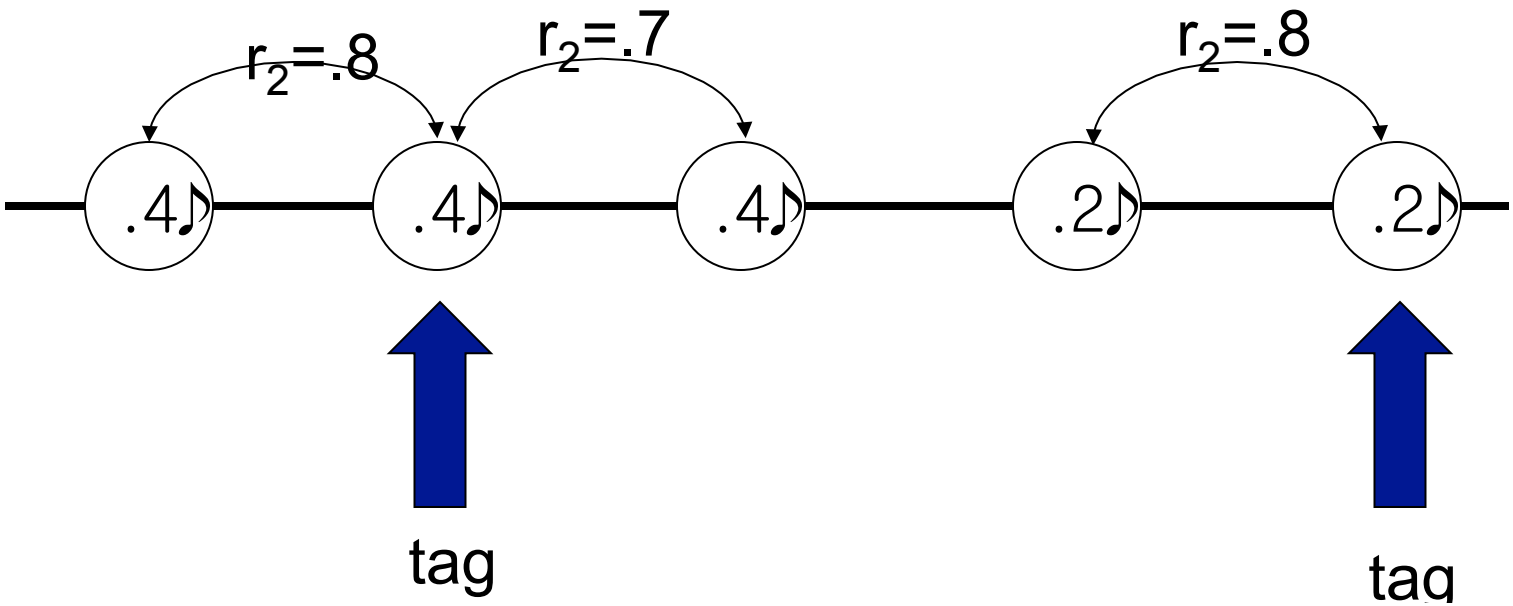
MultiSNP Power

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

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

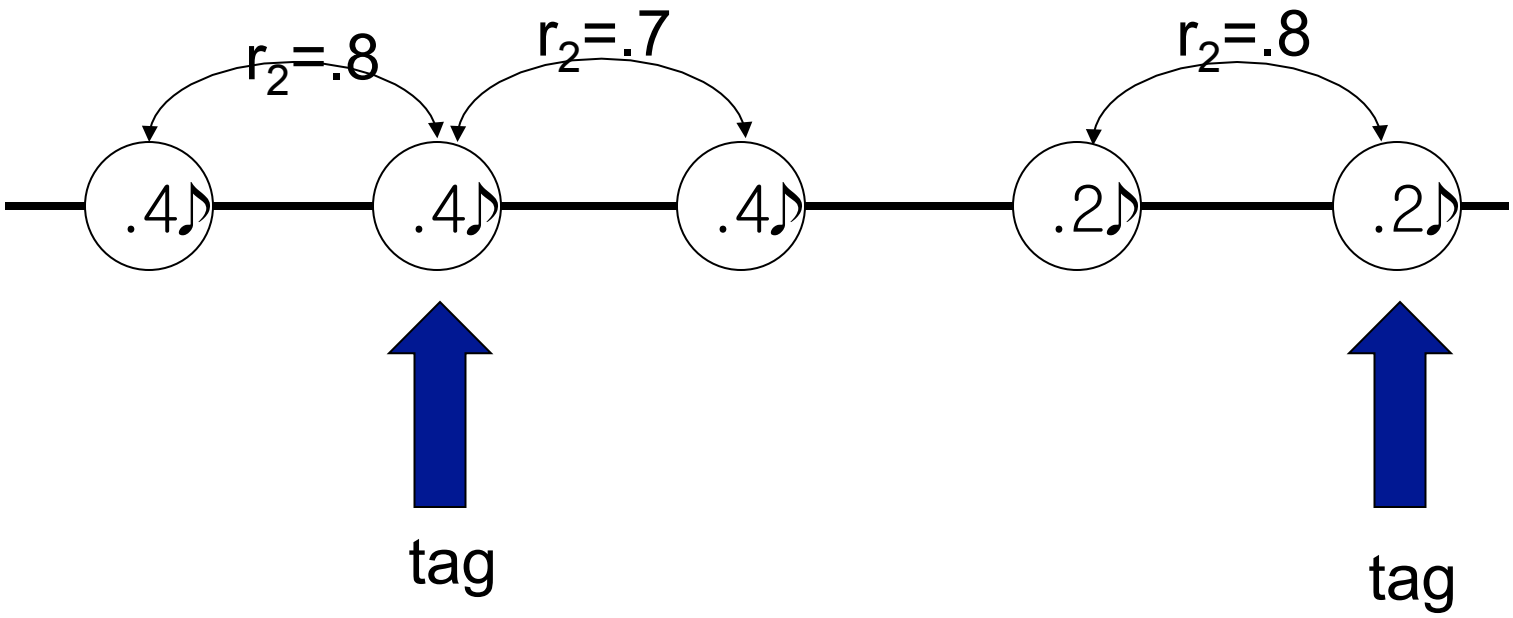
MultiSNP Power with Tags

- 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

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

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

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

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

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

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

Sequencing Coverage

Lecture 7.

April 21st, 2014

(Slides from Jae-Hoon Sul)♪

Sequence Mapping Coverage

- If a genome is length N (human is 3,000,000,000), and the total length of all sequence reads collected is M, the coverage ratio is defined at M/N.
- Often written with an “x”. For example, 10x or 20x coverage.

Sequencing Coverage Statistics

- If length of the genome is N the probability of the event that a single read position starts at a single position in the genome is 1/N (very small).
- If the number of reads is K, the total number of read positions that start at a single genome position is the number of times that an event with probability 1/N happens out of K trials.
- Poisson distribution.

Sequencing Coverage Statistics

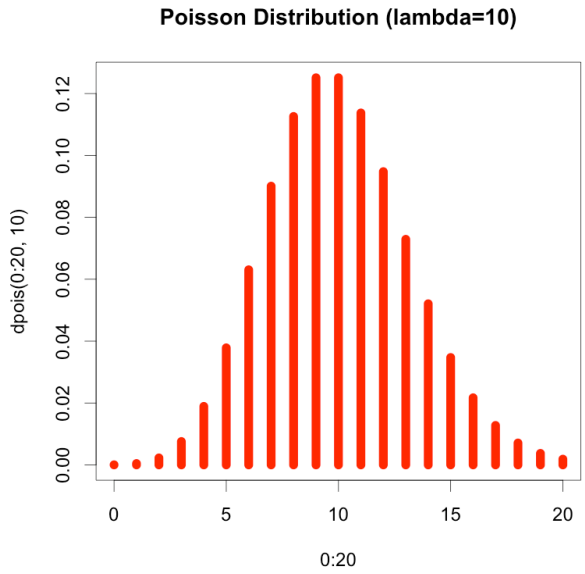
- If length of the genome is N the probability of the event that a single read of length L position spanning a single position in the genome approximately L/N (also very small).
- If the sum of the length of all M reads of length L is M=K*L, the total number of read positions that start at a single genome position is the number of times that an event with probability 1/N happens out of M trials.
- Approximately Poisson distribution.

Poisson Distribution

- Discrete probability distribution to compute probability of (rare) events given known mean
- Only one parameter: λ, mean of distribution
- Probability Mass Function

$$\Pr(N_t = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

- Mean = λ
- Variance = λ



Poisson Distribution to Sequencing Coverage

- λ=M/N.
- Probability that exactly X reads span a certain position.
  - **dpois(X, λ)**
- Probability that X or fewer reads span a certain position.
  - **ppois(X, λ)**
- At least Y% of the genome is covered with this many reads
  - **qpois(Y, λ)**

Poisson and Sequencing Coverage

- Probability that X or fewer reads span a certain position.
  - **ppois(X,λ) =** $\sum_{i=0}^{X} \text{dpois}(i,\lambda)$

Coverage examples

- For human genome (L=3,000,000,000) sequenced at 30x coverage, what is the probability that a specific location has exactly 30 coverage?
- **λ=30 dpois(30,λ)=dpois(30,30)=0.072**
- What is the probability that a specific location has at least 30 coverage?
- **1-ppois(29,λ)=1-ppois(29,30)=0.524**
- What is the probability that a specific location has at least 10 coverage?
- **1-ppois(9,30)=0.9999929**

Coverage examples

- For human genome (L=3,000,000,000) sequenced at 30x coverage, what is the probability that a specific location has exactly one read spanning it?
- **λ=30 dpois(1,λ)=2.9x10^{-12}**
- What is the probability that a specific location has at least 6 coverage?
- **λ=30 1-ppois(5,λ)=.99999**
- How many positions in the genome have less then 6 coverage ?
- **3,000,000,000*ppois(5,λ)=67.7**

Diploid Coverage

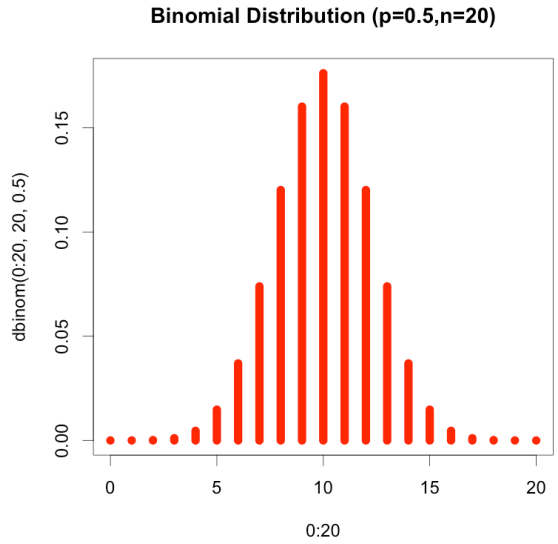
- Since humans have 2 chromosomes each read comes from one chromosome at random. If a position in the reference is covered by Y reads, the probability that X of the reads come from the first chromosome follows the binomial distribution with parameter .5.
  - **dbinom(X,Y,0.5)**
- At least X coverage for each chromosome out of Y reads $\sum_{i=X}^{Y-X} \text{dbinom}(i,Y,0.5)$

Binomial Distribution

- Discrete probability distribution to compute probability of having X successes in Y trials
- Example: What's the probability of having k heads in n tosses with fair coin (p = 0.5)?
- Probability Mass Function

$$\binom{n}{k} p^k (1-p)^{n-k}$$

- Mean = n*p
- Variance = n*p*(1–p)



Diploid Coverage Examples

- If a position is covered by 10 reads, what is the probability that exactly 3 reads come from the first chromosome?
- **dbinom(3,10,.5)=.117**
- If a position is covered by 10 reads, what is the probability that at least 4 reads come from the first chromosome?
- **1-pbinom(3,10,.5)=.828**
- If a position is covered by 10 reads, what is the probability that at least 4 reads come from each chromosome?
- **dbinom(4,10,.5)+dbinom(5,10,.5)+dbinom(6,10,.5)=.656**

Minimum Diploid Coverage

- If we want the sequence coverage is λ=M/N, the portion of the genome that has at least X coverage of each chromosome is

$$\sum_{i=2X}^{\infty} \text{dpois}(i,\lambda) \sum_{j=X}^{i-X} \text{dbinom}(j,i,0.5)$$

Diploid Coverage Examples

- If genome is covered with coverage 30, what is the probability that a position will have at least 10 reads from each chromosome?

$$\sum_{i=20}^{\infty} \text{dpois}(i,30) \sum_{j=10}^{i-10} \text{dbinom}(j,i,0.5)$$

SNP Calling

- Inferring single base differences from sequencing.
- Several challenges:
  - **Sequencing errors**
  - **Alignment “mapping” problems**
  - **Statistical Uncertainty**

SNP Calling Standard Approaches

- Consensus Algorithm
  - **Map reads to genome**
  - **Place read in best mapping position (randomly break ties)**
  - **SNP call is based on majority vote.**
- Probabilistic Algorithm
  - **Map reads to genome**
  - **Place read in best mapping position (randomly break ties)**
  - **Compute “posterior probablility”**
- Mapping uncertainty methods
  - **Map reads to genome**
  - **Record mapping uncertainty**
  - **Compute “posterior probability” incorporating mapping uncertainty**

Sequencing Errors

- Each sequence read can have some random errors.

My Genome:

TACATGAGATC**G**ACATGAGATC**G**GTAGAGC**C**GTGAGATC

A Sequence Read:

TCGACATGAGATCGGTAGA**A**CCGT

The Human Genome:

TACATGAGATC[ ]CACATGAGATC[ ]TGTAGAG[ ]CTGTGAGATC

TC[ ]GACATGAGATC[ ]GGTAGA[ ]CCGT

Recovered Sequence:

TACATGAGATC**G**ACATGAGATC**G**GTAGA**A**C**C**GTGAGATC

Consensus Algorithm

- Take majority vote.

My Genome:

TACATGAGATC**G**ACATGAGATC**G**GTAGAGC**C**GTGAGATC

Sequence Reads:

TCGACATGAGATCGGTAGA**A**CCGT
GACA**A**GAGATCGGTAGAGCCGTGA
TGAGATCGG**T**AGAGCCGTGAGATC

The Human Genome:

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
TC**G**ACATGAGATC**G**GTAGA**A**C**C**GT
GACA**A**GAGATC**G**GTAGAGC**C**GTGA
TGAGATC**G**G**T**AGAGC**C**GTGAGATC

Recovered Sequence:

TACATGAGATC**G**ACATGAGATC**G**GTAGAAC**C**GTGAGATC

How much coverage do we need?

- If error rate is *e*, and we are going to predict the consensus sequence, what is the error rate if the coverage is X.
- We will make a prediction with an error more than X/2-1 out of the X reads have an error in the same place.
- **1-pbinom(X/2,X,e)**

How much coverage do we need?

- If error rate is *e*, and we are going to predict the consensus sequence, what is the error rate if the coverage is 3.
- We will make a prediction with an error if two out of our three reads have an error in the same place.

$$\text{pbinom}(2,3,e) = e^3 + \binom{3}{2}(1-e)e^2$$

- This is approximately $3e^2$.

Diploid Sequencing

- Humans have 2 chromosomes.
- Each chromosome may have a different SNP.
- Some reads come from 1 chromosome, some come from other chromsome.
- Why does consensus method not work?
- How do we address this problem?

Break!