Computational Genetics Winter 2013 Lecture 5

Eleazar Eskin

University of California, Los Angeles

Midterm Review + Final Project Information

Lecture 5.

January23rd, 2013

Midterm Review

Lecture 5.

January 23rd, 2013

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

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

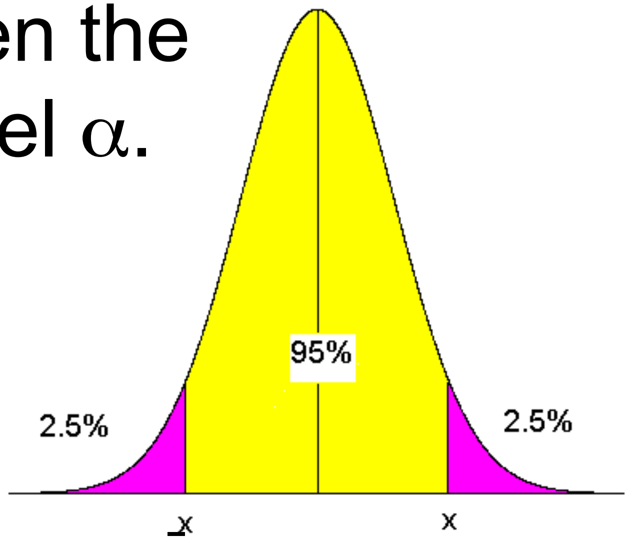
then if $p^{+}_{A} = p^{-}_{A}$

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

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

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

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

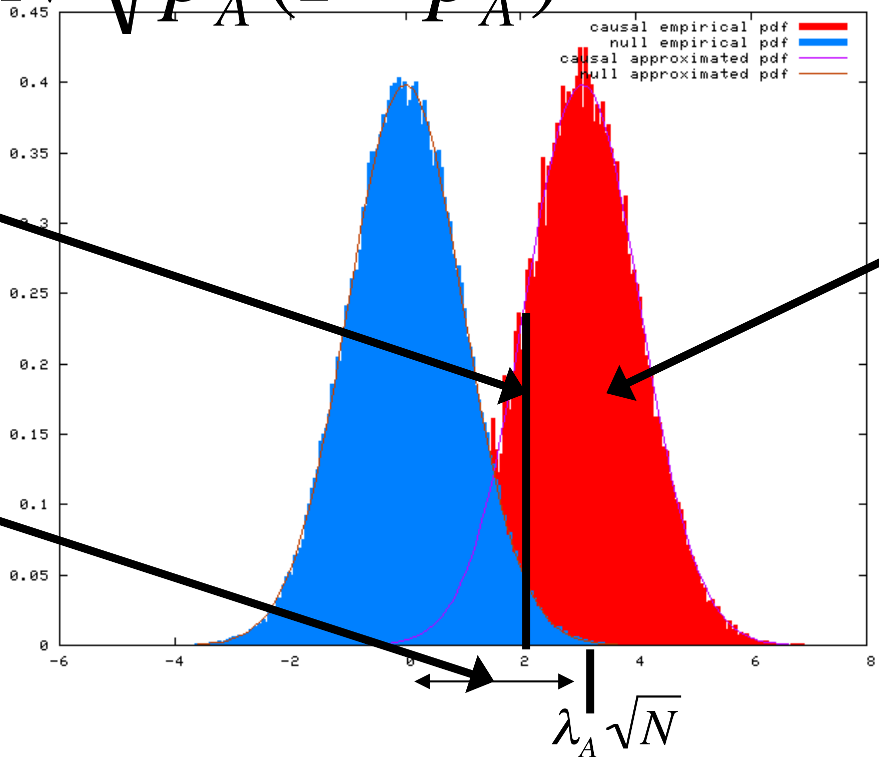
$$S_A = \frac{\hat{p}^{+}{}_{A} - \hat{p}^{-}{}_{A}}{\sqrt{2/N}\sqrt{\hat{p}_A(1-\hat{p}_A)}} \sim N\left(\lambda_A\sqrt{N},1\right)$$

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

Non-centrality parameter.



Power of association test

Association Power

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

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

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

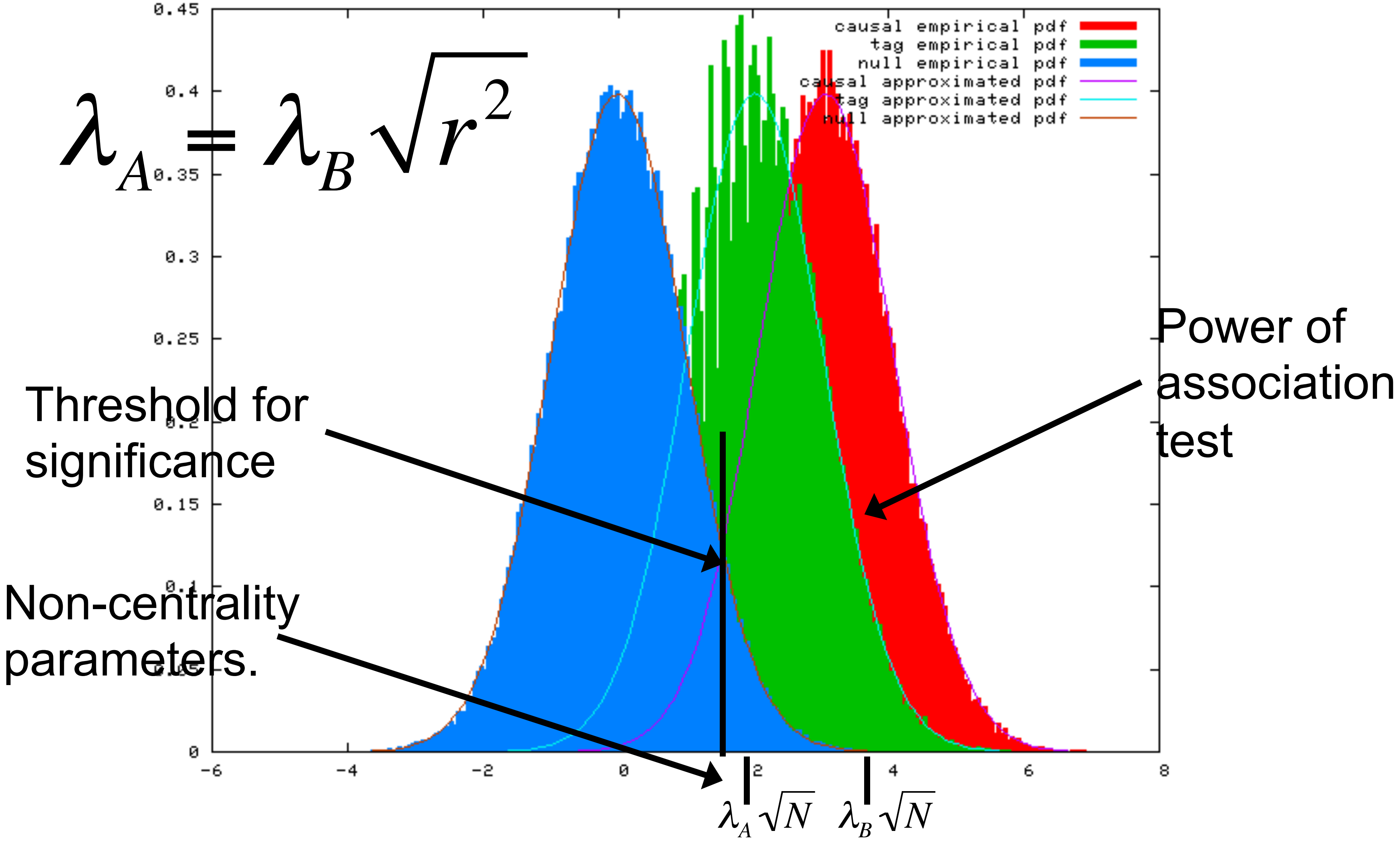
$$\lambda_A = \lambda_B \sqrt{r^2}$$

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

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

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

Visualization in terms of Power



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

Threshold for significance

Non-centrality parameters.

Power of association test

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

$$
\begin{array}{lllll}
\hat{p}_1^+ = \tfrac{120}{200} = .6 & \hat{p}_2^+ = \tfrac{80}{200} = .4 & \hat{p}_3^+ = \tfrac{60}{200} = .3 & \hat{p}_4^+ = \tfrac{100}{200} = .5 & \hat{p}_5^+ = \tfrac{120}{200} = .6 \\
\hat{p}_1^- = \tfrac{100}{200} = .5 & \hat{p}_2^- = \tfrac{75}{200} = .375 & \hat{p}_3^- = \tfrac{65}{200} = .325 & \hat{p}_4^- = \tfrac{95}{200} = .475 & \hat{p}_5^- = \tfrac{125}{200} = .625 \\
\hat{p}_1 = .55 & \hat{p}_2 = .3825 & \hat{p}_3 = .3125 & \hat{p}_4 = .4875 & \hat{p}_5 = .6125
\end{array}
$$

MultiSNP Association Example

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

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

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

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

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

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

Per-marker threshold α_s=α/5=0.01 (Bonferroni)

-Φ^{-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

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

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

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

MultiSNP Association Example

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

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

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

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

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

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

Per-marker threshold α_s=α/5=0.01 (Bonferroni)

-Φ^{-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, α_s= α/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

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

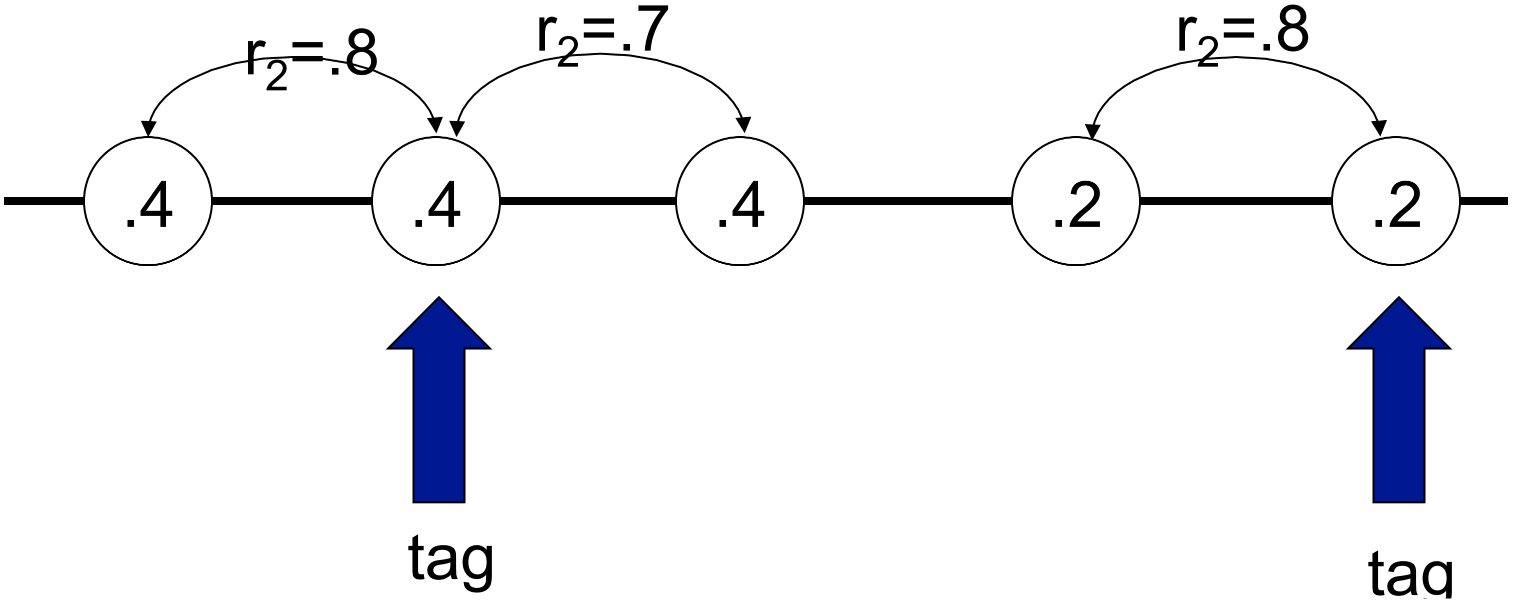
MultiSNP Power

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

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

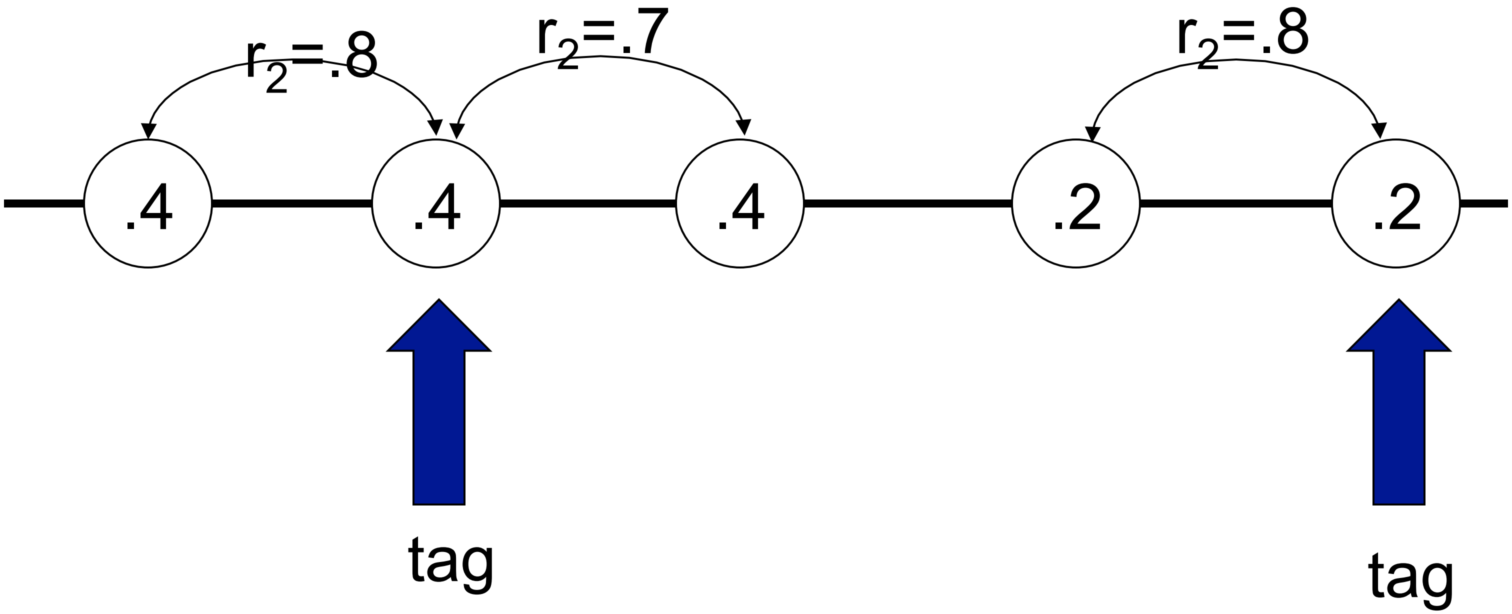
MultiSNP Power with Tags

- 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

$3.4*\sqrt{.8}=3.04$ $3.4*\sqrt{1}=3.4$ $3.4*\sqrt{.7}=2.84$ $2.9*\sqrt{.8}=2.59$ $2.9*\sqrt{1}=2.9$

MultiSNP Power with Tags

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

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

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

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

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

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

Final Projects Information

Lecture 5.

April 16th, 2012

Final Project Requirements

- Project Wiki
- Project Topic Selection (January 25th)
- Project Status Update Email (February 12th)
- Final Project Due (March 21st)
- Project Presentation Workshop – February 15th
- Project Presentation
  - **February 22nd March 1th – Easy Projects**
  - **March 8th – Medium Projects**
  - **March 6th, March 11th – Hard Projects**
  - **March 13th – Very Hard Projects**
  - **10 minutes maximum.**

Final Project Wiki

- Create a wiki page for your project at:
- **http://cs124project-2013.wikidot.com/**
- Password = cs124project
- Content:
  - □ **Paragraph Description of the project**
  - □ **Paragraph about you.**
  - □ **Goal for end of quarter (constantly revised)**
  - □ **Weekly schedule.**
- Update weekly (by Thursday midnight)
  - □ **1. Weekly progress**
  - □ **2. Next week plan**
  - □ **3. Grade for week.**
  - □ **4. Problems that came up.**
  - □ **5. Problems solved this week.**

Final Project Requirements

- Complete final project on approved topic
- Individual Project only
  - **Groups only allowed if approved in advance for very hard projects**
- Minimum difficulty level
  - **Undergraduate – Easy**
  - **Graduate - Medium**
- 10 minute powerpoint presentation in class
- Peer evaluate classmates projects
  - **MANDATORY ATTENDANCE FOR PRESENTATIONS**
- Submit all code and data on wiki
- No written component

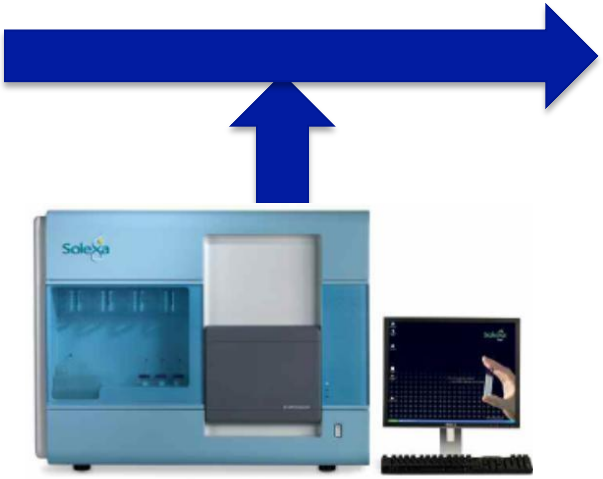
Final Project Requirements

- Must analyze data in project
- Must have experiments in project
- Must describe experiments and results in presentation
- No screenshots, no code. Prove you did experiments by showing results.
- Something new is required! Can not just use ideas I give in class.

Final Project Presentation Structure

1. Motivate the biological problem.
2. Define your computational problem with benchmarks.
3. Define baseline methods for solving the problem.
4. Define your solution to the problem and explain why it solves the problem.
5. Analyze the performance of your method.
6. Say something interesting! (Implications, Observations, etc).

How do we get someone's DNA sequence? Where are my mutations?


Sequencing Technology

Illumina / Solexa
Genetic Analyzer 1G
1000 Mb/run, 35bp reads

AGAGC**A**GTCGAC
A**G**GTATAG**T**CTA
CATGAGATC**G**AC
ATGAGATC**G**GTA
GAGC**C**GTGAGAT
C**G**ACATGATAG**C**
CAGAGC**A**GTCGA
CA**G**GTATAG**T**CT
ACATGAGATC**G**A
CATGAGATC**G**GT
AGAGC**C**GTGAGA
TC**G**ACATGATAG
CCAGAGC**A**GTCG
ACA**G**GTATAG**T**C
TACATGAGATC**G**
ACATGAGATC**G**G
TAGAGC**C**GTGAG
ATC**G**ACATGATA
G**C**CAGAGC**A**GTC
GACA**G**GTATAG**T**
CTACATGAGATC
GACATGAGATC**G**
GTAGAGC**C**GTGA
GATC**G**ACATGAT

- Next generation sequencing.
  - □ **Cheap sequencing.**
  - □ **"Short Reads"**

Short Read Sequencing Problem (A Computer Science Problem)

Full DNA Sequence

AGAGC**A**GTCGAC
A**G**GTATAG**T**CTA
CATGAGATC**G**AC
ATGAGATC**G**GTA
GAGC**C**GTGAGAT
C**G**ACATGATAG**C**
CAGAGC**A**GTCGA
CA**G**GTATAG**T**CT
ACATGAGATC**G**A
CATGAGATC**G**GT
AGAGC**C**GTGAGA
TC**G**ACATGATAG
CCAGAGC**A**GTCG
ACA**G**GTATAG**T**C
TACATGAGATC**G**
ACATGAGATC**G**G
TAGAGC**C**GTGAG
ATC**G**ACATGATA
G**C**CAGAGC**A**GTC
GACA**G**GTATAG**T**
CTACATGAGATC

→

- Short read sequencers generate random short substrings from the DNA sequence of a certain length.

ATGAGATC**G**GTAGAGC**C**GTGAGAT
GAGC**A**GTCGACA**G**GTATAG**T**CTAC
AGAGC**A**GTCGACA**G**GTATAG**T**CTA
TGAGATC**G**ACATGATAG**C**CAGAGC
TAG**C**CAGAGC**A**GTCGACA**G**GTATA
GATAG**C**CAGAGC**A**GTCGACA**G**GTA
GAGATC**G**ACATGATAG**C**CAGAGC**A**
GC**A**GTCGACA**G**GTATAG**T**CTACAT
AGC**A**GTCGACA**G**GTATAG**T**CTACA
TC**G**ACATGAGATC**G**GTAGAGC**C**GT
C**A**GTCGACA**G**GTATAG**T**CTACATG
GAGATC**G**ACATGATAG**C**CAGAGC**A**
GTAGAGC**C**GTGAGATC**G**ACATGAT

How do we recover the original sequence?

Short Reads Difficulties

```
ATGAGATCGGTAGAGCCGTGAGAT
GAGCAGTCGACAGGTATAGTCTAC
AGAGCAGTCGACAGGTATAGTCTA
TGAGATCGACATGATAGCCAGAGC
TAGCCAGAGCAGTCGACAGGTATA
GATAGCCAGAGCAGTCGACAGGTA
GAGATCGACATGATAGCCAGAGCA
GCAGTCGACAGGTATAGTCTACAT
AGCAGTCGACAGGTATAGTCTACA
TCGACATGAGATCGGTAGAGCCGT
CAGTCGACAGGTATAGTCTACATG
GAGATCGACATGATAGCCAGAGCA
GTAGAGCCGTGAGATCGACATGAT
```

- We don’t know where each read comes from!
- Can’t identify where the mutations are!
- What do we do?

Key Idea: “Re”-Sequencing

We know that my genome is very close to the Human genome.

My Genome:

TACATGAGATCGACATGAGATCGGTAGAGCCGTGAGATC

A Sequence Read:

TCGACATGAGATCGGTAGAGCCGT

The Human Genome:

```
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
         TCGACATGAGATCGGTAGAGCCGT
```

Recovered Sequence:

TACATGAGATCGACATGAGATCGGTAGAGCCGTGAGATC

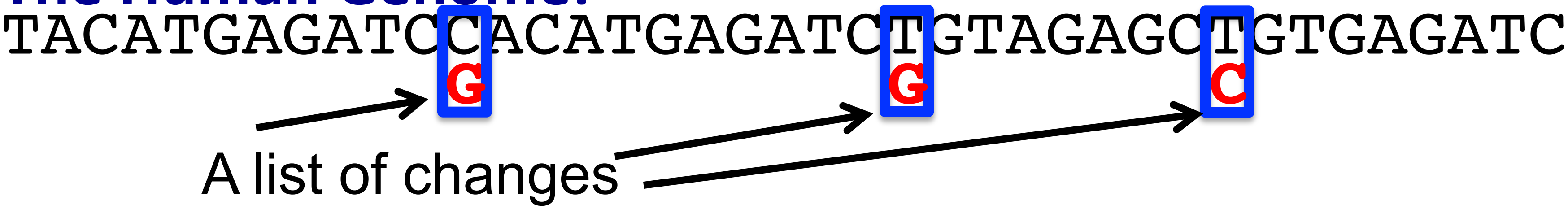
“Re”-Sequencing Output

Resequencing provides a list of changes to make from the reference to change it to the target. Similar to unix “diff”.

My Genome:

```
TACATGAGATCGACATGAGATCGGTAGAGCCGTGAGATC
```

The Human Genome:



```
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
           G          G       C
```

A list of changes

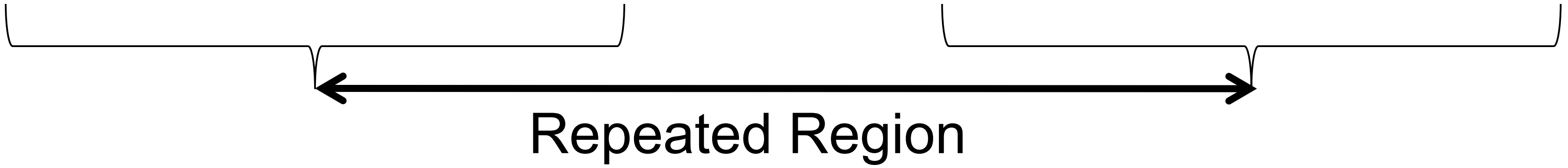
Recovered Sequence:

```
TACATGAGATCGACATGAGATCGGTAGAGCCGTGAGATC
```

“Re”-Sequencing Problems

The Human Genome:

TACATGAGATCCACATGAGATCTGTACATGAGATCCACAT



My Genome:

TACATGAGATC**G**ACATGAGATCGGTACATGAGATCCACAT

A Sequence Read:

ACATGAGATC**G**ACAT

The Human Genome:

```
TACATGAGATCCACATGAGATCTGTACATGAGATCCACAT
 ACATGAGATCGACAT          ACATGAGATCGACAT
```

Error!

Recovered Sequence:

TACATGAGATC**G**ACATGAGATCGGTACATGAGATC**G**ACAT

“Re”-Sequencing Problems

The Human Genome:

TACATGAGATCCACATGAGATCTGTACATGAGATCCACAT

My Genome:

TACATGAGGGGGGGGGGAGATCGGTACATGAGATCCACAT

A Sequence Read:

GAGGGGGGGGG

The Human Genome:

```
TACATGAGATCCACATGAGATCTGTACATGAGATCCACAT
     GAGGGGGGGGG
```

Too many mismatches to match the read to the reference. Since we don’t know where it came from, we can’t identify the difference in the target seqeunce.

Key Question: When does resequencing work?

- We must be able to map a substring from the target to its corresponding place in the reference.
- Why can this not happen?
  - **Reference has repeated sequences. In this case reads from target will map to multiple places.**
  - **Target sequence differs that resemblance to reference sequence is lost.**

Formalizing the Problem

- Target sequence – Sequence of the genome that we are analyzing and collecting reads from.
- Reference sequence – Sequence of the similar genome which we have available.
- Constraints on the reference sequence
  - **Non repetitive sequences (or non-repetitive portion)**
- Constrains on difference between the target and reference.
  - **Assume that there are a small number of structured differences.**

Simple Resequencing Formulation

- Assume that the reference sequence is of length N.
- Assume target sequence is of length N.
- Constraint on Mutations - Assume that target sequence differs from reference by less than D mutations in any window of L.
- Unique Sequence Assumption – Assume that any 2 positions in the reference sequence differ by more than D+1 mismatches.

Algorithmic “Re”-Sequencing Challenges

- Sequences are long!
  - **Human Genome is 3,000,000,000 long.**
- Sequencers generate many reads!
  - **A single run generates over 300,000,000 reads.**

- We need efficient algorithms to “map” each read to its location in the genome.

There are other challenges which we are not mentioning.

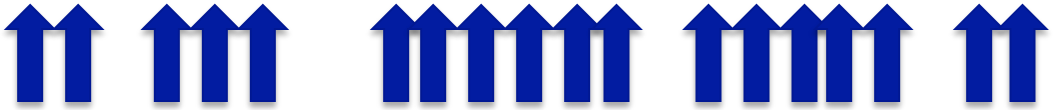
Trivial Mapping Algorithm

The Human Genome:
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

A Sequence Read:
TCGACATGAGATCGGTAGAGCCGT

- We can slide our read along the genome and count the total mismatches between the read and the genome.
- If the mismatches are below a threshold, we say that it is a match.

```
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
TCGACATGAGATCGGTAGAGCCGT
```



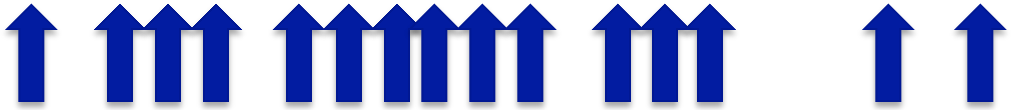
Total of 18 mismatches. Not below threshold. Not a match.

Trivial Mapping Algorithm

The Human Genome:
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

A Sequence Read:
TCGACATGAGATCGGTAGAGCCGT

```
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
 TCGACATGAGATCGGTAGAGCCGT
```



Total of 15 mismatches. Not below threshold. Not a match.

Trivial Mapping Algorithm

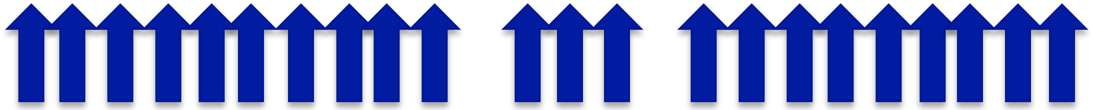
The Human Genome:

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

A Sequence Read:

TCGACATGAGATCGGTAGAGCCGT

```
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
  TCGACATGAGATCGGTAGAGCCGT
```

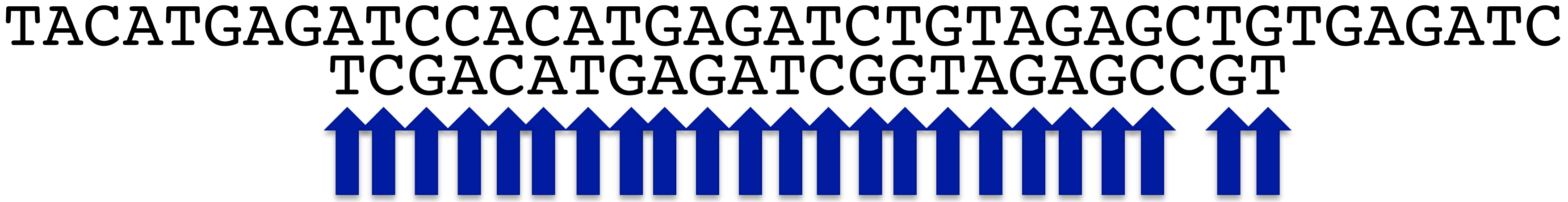


Total of 23 mismatches. Not below threshold. Not a match.

Trivial Mapping Algorithm

The Human Genome:
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

A Sequence Read:
TCGACATGAGATCGGTAGAGCCGT



```
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
        TCGACATGAGATCGGTAGAGCCGT
```

Total of 23 mismatches. Not below threshold. Not a match.

Trivial Mapping Algorithm

The Human Genome:

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

A Sequence Read:

TCGACATGAGATCGGTAGAGCCGT

```
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
         TCGACATGAGATCGGTAGAGCCGT
           ↑          ↑       ↑
```

Total of 3 mismatches. Below threshold. A match!

Complexity of Trivial Algorithm

- 3,000,000,000 length genome (N)
- 300,000,000 reads to map (M)
- Reads are of length 30 (L)
- Number of mismatches allowed is 2 (D).
- Each comparison of match vs. mismatch takes 1/1,000,000 seconds (t).

Total Time = N*M*L*t = 27,000,000,000,000 seconds or 864,164 years!

- Important: Trivial algorithm only solves problem under assumptions.

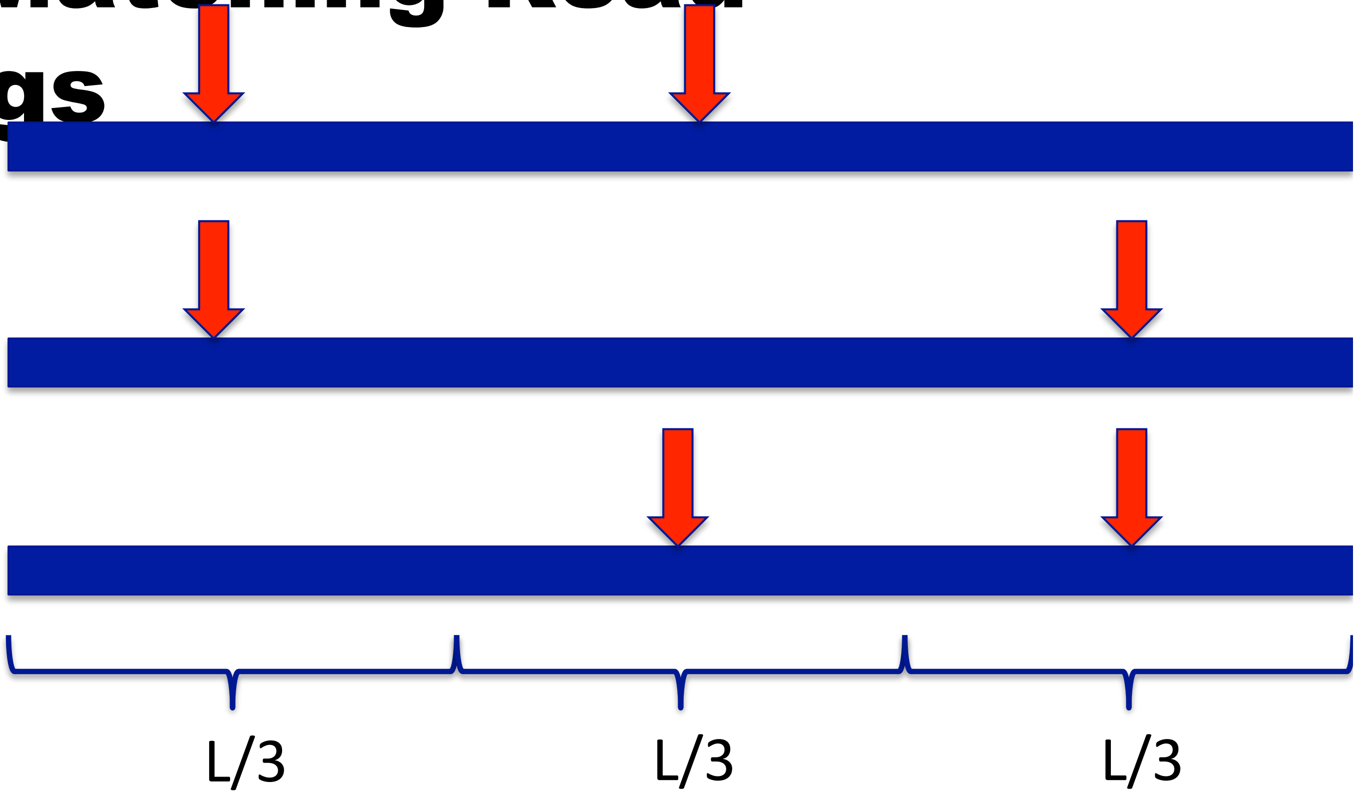
Some observations

- Most positions in the genome match very poorly.
- We are looking for only a few mismatches.

  (D is small)
- A substring of our read will match perfectly.

Perfect Matching Read Substrings

Three "worst" possible cases for placement of mutations.



- In each case, there is a perfect match of L/3.

Finding a perfect match of length L/3

- Intuition: Create an index (or phone book) for the genome.
- We can look up an entry quickly.

Sequence	Positions
AAAAAAAAAA	32453, 64543, 76335
AAAAAAAAAC	64534, 84323, 96536
AAAAAAAAAG	12352, 32534, 56346
AAAAAAAAAT	23245, 54333, 75464
AAAAAAAACA	
AAAAAAAACC	43523, 67543
…	
CAAAAAAAAA	32345, 65442
CAAAAAAAAC	34653, 67323, 76354
…	
TCGACATGAG	54234, 67344, 75423
TCGACATGAT	11213, 22323
…	
TTTTTTTTTG	64252
TTTTTTTTTT	64246, 77355, 78453

If L=30, each entry will have a key of length 10. Each entry will contain on average $N/4^{10}$ positions. (Approximately 3,000).

If L=45, each entry will have a key of length 15. Each entry will contain on average 3 positions.

Complexity of Indexing Algorithm

- We need to look up each third of the read in the index.
- For L=30, our index will contain entries of length 10. Each entry will contain on average N/($4^{L/3}$) or 3,000 positions.
- For each position, we need to compute the number of mismatches.
- Our running time is L* M*3*N/($4^{L/3}$)*T=81,000,000 seconds or 937 days.
- If L=45, then the time is 81,000 seconds or 22.5 hours.

Read Mapping Project

- Use the simulator:

  http://cs124project-2009.wikidot.com/readsimulation

- Six requirements:
  - □ **Motivate the biological problem.**
  - □ **Define your computational problem with benchmarks.**
  - □ **Define baseline methods for solving the problem.**
  - □ **Define your solution to the problem and explain why it solves the problem.**
  - □ **Analyze the performance of your method.**
  - □ **Say something interesting! (Implications, Observations, etc).**

Read Mapping Simulator Example

- Assume we were following what we were doing in class.
- Step 1: Motivate the problem:
  - □ **Sequencing technologies have been recently developed that are much cheaper than traditional sequencing technologies which will allow us to sequence everyone's genome.**
  - □ **However, these technologies generate short reads which make it difficult to recover the sequence since we do not know where the short reads come from.**
  - □ **Our idea is to use apply "resequencing" and use a "reference genome" to identify where reads come from to allow us to recover the sequence.**

Read Mapping Simulator Example

- Step 2: Define the problem and benchmarks.
  - □ **Given a set of reads which are random substrings of length L from an unknown target sequence, and a reference sequence which we assume is "very similar" to the target sequence.**
  - □ **By "very similar" we mean that the number of mismatches within a region of length L is less than D.**
  - □ **We also assume that any 2 regions of length L in the reference genome differ by at least 2D positions.**
  - □ **Our goal is to output the target sequence.**
  - □ **The benchmarks are the computational time to map a read, the accuracy of mapping a read and the memory requirements of mapping reads.**

Read Mapping Simulator Example

- Step 3: Define baseline method(s).
  - **The baseline method for mapping reads is the trivial mapping algorithm where you just slide a read along the genome and check to see if it matches in each position.**

Read Mapping Simulator Example

- Step 4: Define your solution and explain why it solves the problem:
  - **This part is the longest. You want to describe your approach, data structures, how it works, etc.**
  - **For example, if I was showing what I did in class I would describe the data structure used in the index, etc.**
  - **I would also explain why our approach will solve the problem. I.e. under the assumptions that I made, if I have enough reads, I would find all of the mismatches and recover the sequence.**
  - **I would emphasize what are my ideas, what is my contribution.**

Read Mapping Simulator Example

- Step 5: Analyze the performance:
  - **Measure how fast your algorithm is.**
  - **How fast is the baseline method.**
  - **How does the input size, read length, number of mismatches affect the speed?**
  - **What range of parameters does your method work on?**
  - **When does it outperform the baseline?**

Read Mapping Simulator Example

- Step 6: Say something interesting!
  - **When does your approach work better?**
  - **What are its limitations?**
  - **What can you do about its limitations?**
  - **What other things can you try?**
  - **What is still left to figure out?**

Presentation Requirements

- Use powerpoint. No chalkboard.
- Send me draft slides for review more than 48 hours in advance.
- Practice to a friend outside the class.
- No outline slides.
- No “question slides”
- No discussion with class.

- Only funny jokes.