Algorithms in Bioinformatics
Spring 2023
Lecture 1

Eleazar Eskin and Jason Ernst
University of California, Los Angeles

Course Logistics + Introduction + Read Mapping

Lecture 1.

April 4th, 2023

Lecture Outline

- Course logistics
- Introduction
- Read Mapping

Lecture Outline

- Course logistics
- Introduction
- Read Mapping

Algorithms in Bioinformatics

- Cross-listings:
  - Computer Science CM122 (undergraduate)
  - Computer Science CM222 (graduate)
  - Bioinformatics M222 (graduate)
- Course website: https://bruinlearn.ucla.edu/courses/160773
- Lectures: Tue and Thur 2-3:50pm WSYOUNG CS76
- Discussion sections:
  - 1A – Fri 12-1:50pm BROAD 2100A
  - 1B – Fri 12-1:50pm KAPLAN A65
  - 1C – Fri 2-3:50pm DODD 175
  - 1D – Fri 2-3:50pm KAPLAN 169
- Final exam: June 13th 3-6pm

Course Requisites

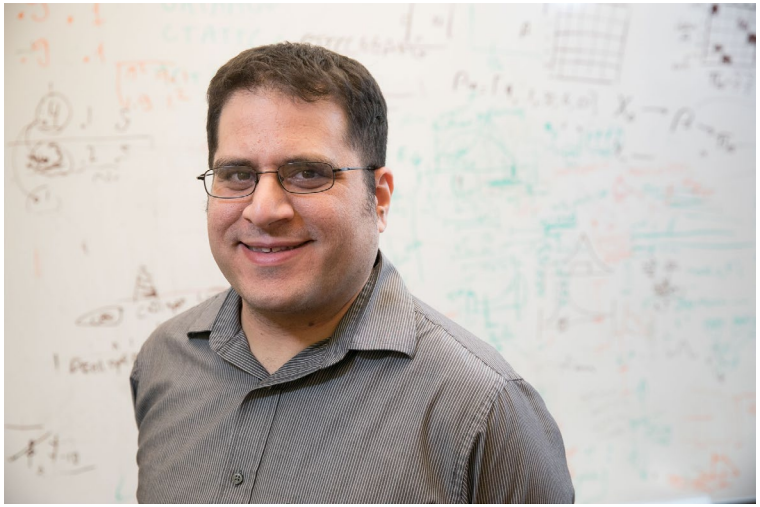
- Students taking the course should be familiar with programming.
- Required unless approval of instructor: Computer Science 32 or Program in Computing 10C with grade of C- or better
- Recommended: Civil and Environmental Engineering 110 or Electrical and Computer Engineering 131A or Mathematics 170A or Mathematics 170E or Statistics 100A.
- Note: CM121 is NOT a pre-requisite for CM122.

Waitlist and PTEs

- Waitlist/not-enrolled
  - We will issue PTEs to let you enroll. Email TA/Instructor or find us after class.

Course Instructors

Prof. Eleazar Eskin
Professor and Chair of Computational Medicine
Professor of Computer Science
Professor of Human Genetics
Email: eeskin@cs.ucla.edu
Office Hours: Tuesdays 4-5pm
Location: EVI 286

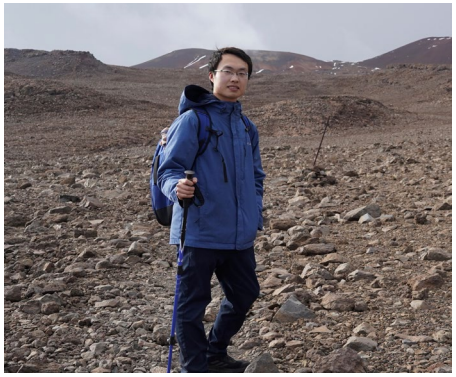


Prof. Jason Ernst
Professor of Biological Chemistry
Professor of Computer Science
Professor of Computational Medicine
Email: jason.ernst@ucla.edu
Office Hours: Tuesdays 4-5pm
Location: EVI 286 (Note: not my regular office)

Course Teaching Assistants (TAs)

Shuwen Qiu
Computer Science PhD student
janetqiu@cs.ucla.edu
Discussion 1A: Fri. 12-1:50pm
BROAD 2100A
Office Hours: Wed. 4-6pm
Boelter 3256S-E

Yihe Deng
Computer Science PhD student
yihedeng@ucla.edu
Discussion 1B: Fri. 12-1:50pm
KAPLAN A65
Office Hours: Wed. 10am-12pm
Boelter 3256S-E



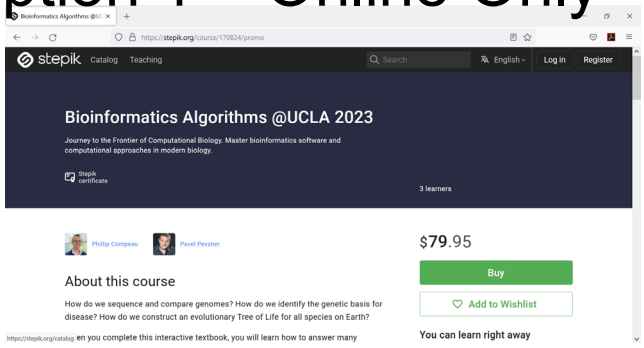
Xuheng Li
Computer Science PhD student
xuhengli99@ucla.edu
Discussion 1C: Fri. 2-3:50pm
DODD 175
Office Hours: Thur. 9-11am
Boelter 3256S-E

Runjia “Luke” Li
Bioinformatics PhD student
luke0321lrj@gmail.com
Discussion 1D: Fri. 2-3:50pm
KAPLAN 169
Office Hours: Mon. 2-4pm
Boelter 3256S-B

Course Textbook

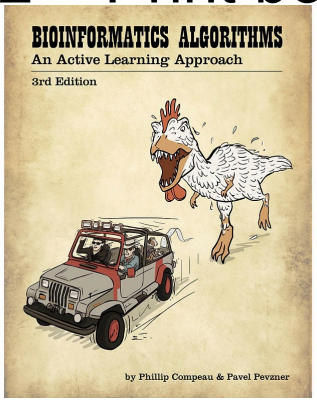
- Bioinformatics Algorithms 3rd edition by Compeau and Pevzner
- Required to be part of our course instance of interactive online book for assignments
- If you buy the print version (new or used) and provide a receipt through our form you will receive free access to the online book. Receipts should be uploaded by end of day Wed. April 5th. Late receipts will still accommodated but may cause delays. Also can get free access if you previously bought the online book not through our course instance.

Option 1 – Online Only



https://stepik.org/a/170824
(can receive refund within 30 days)

Option 2 – Print book + Online



https://www.bioinformaticsalgorithms.org/
Receipt Form:
https://forms.gle/iA7JLc7LK2a3Qiuv8

Policy on Electronic Devices During Lecture

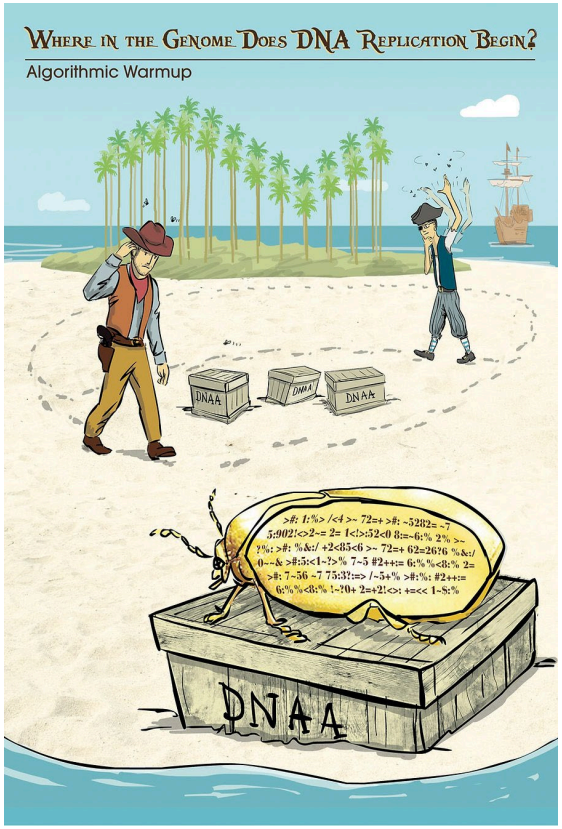
- Electronic devices including laptops and cell phones should **not** be used during lectures and discussion (unless using the interactive online textbook) without permission from the instructor

Course Objectives

- Formulate biological problems as algorithmic problems
- Have an advanced background in applied data structures and algorithms.
- Gain experience in implementing software that can scale to large datasets.

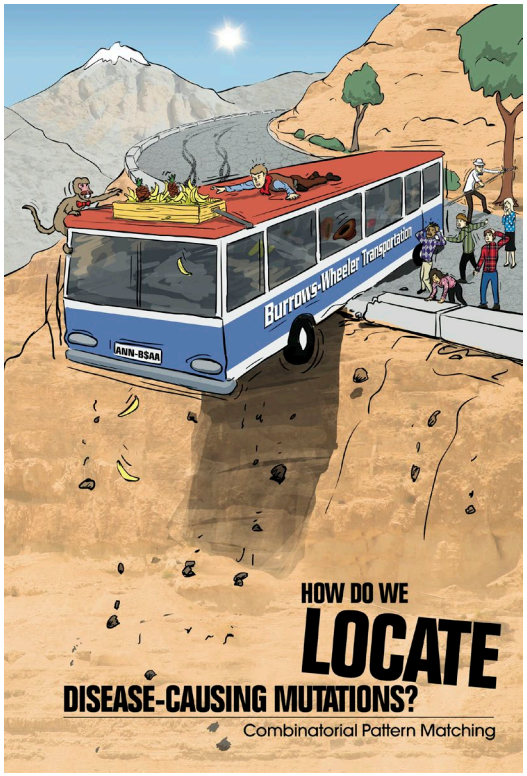
Course Overview (subject to change)

- Algorithmic Warm-up (Chapter 1)
- Discussion: Apr 7th



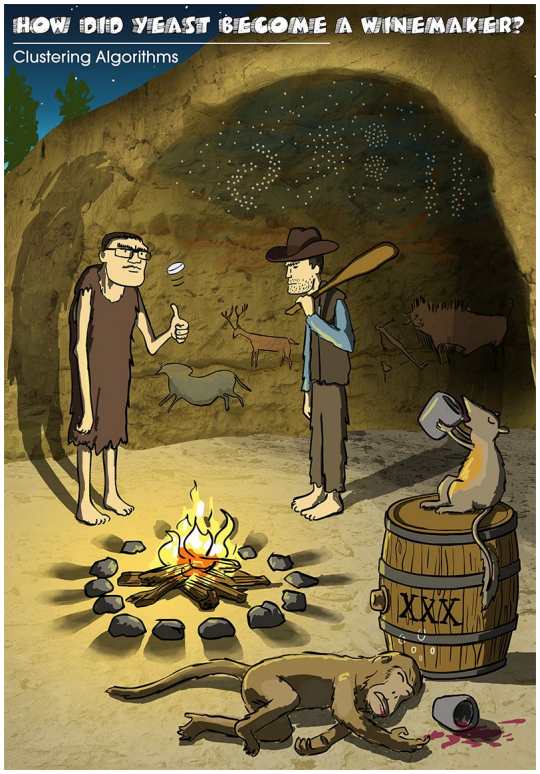
Course Overview (subject to change)

- Read Mapping (Chapter 9)
- Instructor: Prof. Eskin
- Lecture Dates: April 4th, 6th, 11th
- Discussion: Apr 7th and Apr 14th



Course Overview (subject to change)

- Clustering Algorithms (Chapter 8)
- Instructor: Prof. Ernst
- Lecture Dates: April 13th and 18th
- Discussion: April 14th and 21st



Course Overview (subject to change)

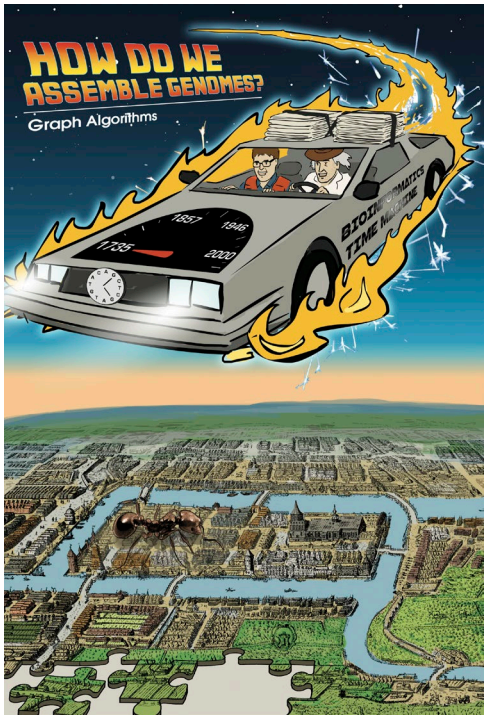
- Sequence Motifs and Prediction (Chapter 2)
- Note: will include connections to supervised Machine Learning/Deep Learning approaches not in book
- Instructor: Prof. Ernst
- Lecture Dates: April 20th, 25th, 27th
- Discussion Dates: April 21st, 28th, May 5th

Course Overview (subject to change)

- April 27th – Midterm review (Prof. Ernst)
- May 2nd – Midterm
- May 4th – Guest Speaker – Current Research Topic

Course Overview (subject to change)

- Genome Assembly (Chapter 3)
- Instructor: Prof. Eskin
- Lecture Dates: May 9th, 11th, 16th
- Discussion Dates: May 12th and 19th



Course Overview (subject to change)

- K-mer methods (Not in textbook)
- Instructor: Prof. Eskin
- Lecture Dates: May 16th, 18th, 23rd
- Discussion Dates: May 19th and 26th

Course Overview (subject to change)

- Sequence Alignment (Chapter 5)
- Instructor: Prof. Eskin
- Lecture Dates: May 25th and 30th
- Discussion Dates: May 26th and June 2nd

Course Overview (subject to change)

- Hidden Markov Models (Chapter 10)
- Instructor: Prof. Ernst
- Lecture Dates: June 1st and 6th
- Discussion Dates: June 2nd and 9th

Course Overview (subject to change)

- June 8th – Final review (Prof. Ernst/Eskin)
- June 13th – Final 3-6pm

Course Requirements

- Requirements
  - 7 Online Integrated Book Homework Assignments
  - 4 Programming Projects
  - 5 Paper and 1 Guest Speaker Question/Responses
  - Midterm Exam – Tue May 2nd
  - Final Exam – Tue June 13th 3:00pm-6:00pm
  - Graduate students: complete a more difficult programming project and extra exam problems
- Grading Basis
  - Homeworks 20%
  - Projects 25%
  - Midterm Exam 25%
  - Final Exam 25%
  - Paper/Guest Speaker Question and Responses 5%

Homework Assignments

- Integrated with reading on interactive textbook website and involve programming
- Unlimited number of submissions to solve problem
- Only a subset of problems are assigned
- Recommend to do as you do reading, well in advance of deadline. Can start at anytime.

Homework Assignments

- Homeworks are due at 12pm on due date
- Homework schedule (subject to change)
  - HW1: Chapter 1 (Warm-up) – Thur Apr 13th
  - HW2: Chapter 9 (Read-mapping) – Tue Apr 18th
  - HW3: Chapter 8 (Clustering) – Fri Apr 21st
  - HW4: Chapter 2 (Motifs) – Fri Apr 28th
  - HW5: Chapter 3 (Assembly) – Fri May 19th
  - HW6: Chapter 5 (Alignment) – Fri June 2nd
  - HW7: Chapter 10 (HMMs) – Fri June 9th

HW1 - Chapter 1

- Problems from Algorithmic Warm-up Chapter on interactive site
- This chapter is available for free if you want to start working on it before being part of course instance

https://www.bioinformaticsalgorithms.org/bioinformatics-chapter-1

- Subset of assigned problems indicate on course instance of interactive site and also listed on bruinlearn website
- Still need to submit solutions as part of course instance of online book
- Due Thur. Apr 13th at 12pm, but should try to finish this week. Due date is later in case people do not have access yet to online textbook. Two HWs due following week.

Programming Projects

- Projects will be submitted to a webserver
- Must submit project output and code
- Projects should be able to run from standard standard command line interface as described in the programming assignment
- Graduate students will have more difficult projects
- 3 of 4 projects will be broken into sub-parts with an earlier deadline for the first sub-part

Programming Projects

- Projects are due at 12pm on due date
- Project schedule (subject to change)
  - Project 1a: Read Mapping – Thur Apr 20th
  - Project 1b: Read Mapping – Tue Apr 25th
  - Project 2a: Motifs/Sequence Prediction – Thur May 4th
  - Project 2b: Motifs /Sequence Prediction – Tue May 9th
  - Project 3: Genome Assembly – Thur May 23rd
  - Project 4a: Kmer-methods – Tue May 30th
  - Project 4b: Kmer-methods – Tue June 6th

Homework and Project Policies

- Barring any extreme circumstances, extensions will be considered only if requested in advance of the deadline.
- Students may discuss homework problems and projects with others but must write any code to solve the problems themselves.
- Standard built in software libraries can be used, but not any libraries implementing the task of the homework or project.

Paper and Guest Speaker Question and Responses

- Required to post at least one question and respond to at least two other questions per assigned paper or guest speaker on class discussion forum
- Separate deadlines for posting question and responding to questions

Paper and Guest Speaker Question and Responses

- Posts due at 12pm on due dates
- Paper/Guest Speaker schedule (subject to change; Q-question; R- response)
  - Paper 1: Read Mapping – Q: Thur Apr 13th; R: Tue Apr 18th
  - Paper 2: Motifs – Q: Tue Apr 27th; R: Tue May 2nd
  - Paper 3: Sequence Prediction – Q: Thur May 4th; R: Tue May 9th
  - Guest Speaker 1 – Q: Fri May 5thth; R: Thur May 11th
  - Paper 4: Genome Assembly – Q: Thur May 18th; R: Tue May 23rd
  - Paper 5: k-mer Methods – Q: Thur May 25th; Tue May 30th

Paper 1

Software

Open Access

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome

Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg

Address: Center for Bioinformatics and Computational Biology, Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA.

Correspondence: Ben Langmead. Email: langmead@cs.umd.edu

Published: 4 March 2009

Genome ***Biology*** 2009, **10:**R25 (doi:10.1186/gb-2009-10-3-r25)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/3/R25

Received: 21 October 2008
Revised: 19 December 2008
Accepted: 4 March 2009

https://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-3-r25

Also available on coursewebsite
Question due Thur Apr. 13th
Responses due Tue Apr. 18th

Lecture Outline

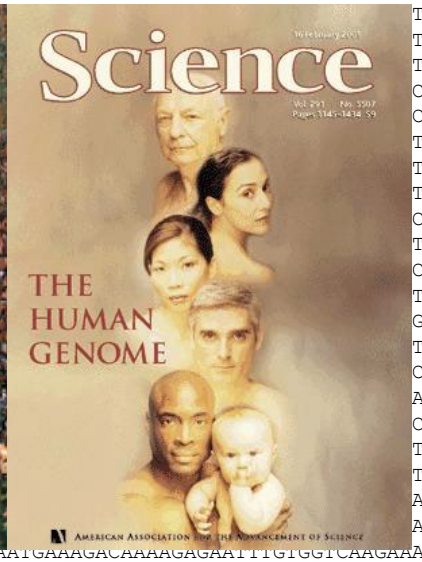
- Course logistics
- Introduction
- Read Mapping

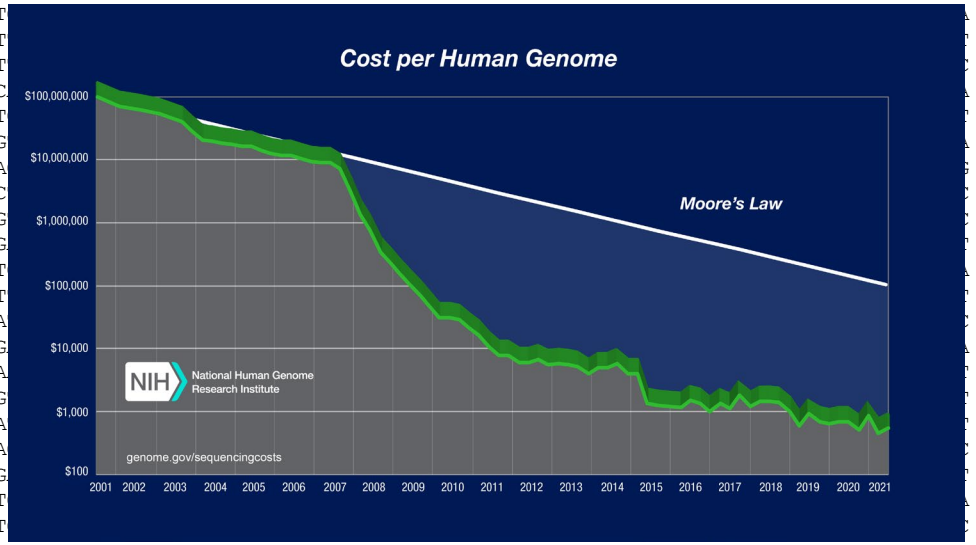
The Genomics Revolution

The Genomics Revolution

2001 – First draft of human reference genome

The Genomics Revolution





2001 – First draft of human reference genome

Rapid decline in cost of sequencing in the following two decades

Genomics Data is Pervasive

Human genetic variation data linked to phenotype variation now in large scale biobanks

Cases: (Individuals with trait)

AGAGCAGTCGACAGGTATAGCCTACATGAGATCGACATGAGATCGGTAGAGCCGTGAGATCGACATGATAGCC
AGAGCCGTCGACATGTATAGTCTACATGAGATCGACATGAGATCGGTAGAGCAGTGAGATCGACATGATAGTC
AGAGCAGTCGACAGGTATAGTCTACATGAGATCGACATGAGATCGGTAGAGCCGTGAGATCGACATGATAGCC
AGAGCAGTCGACAGGTATAGCCTACATGAGATCAACATGAGATCGGTAGAGCAGTGAGATCGACATGATAGCC
AGAGCCGTCGACATGTATAGCCTACATGAGATCGACATGAGATCGGTAGAGCCGTGAGATCAACATGATAGCC
AGAGCCGTCGACATGTATAGCCTACATGAGATCGACATGAGATCGGTAGAGCAGTGAGATCAACATGATAGCC
AGAGCCGTCGACAGGTATAGCCTACATGAGATCGACATGAGATCGGTAGAGCAGTGAGATCAACATGATAGTC
AGAGCAGTCGACAGGTATAGCCTACATGAGATCGACATGAGATCTGTAGAGCCGTGAGATCGACATGATAGCC

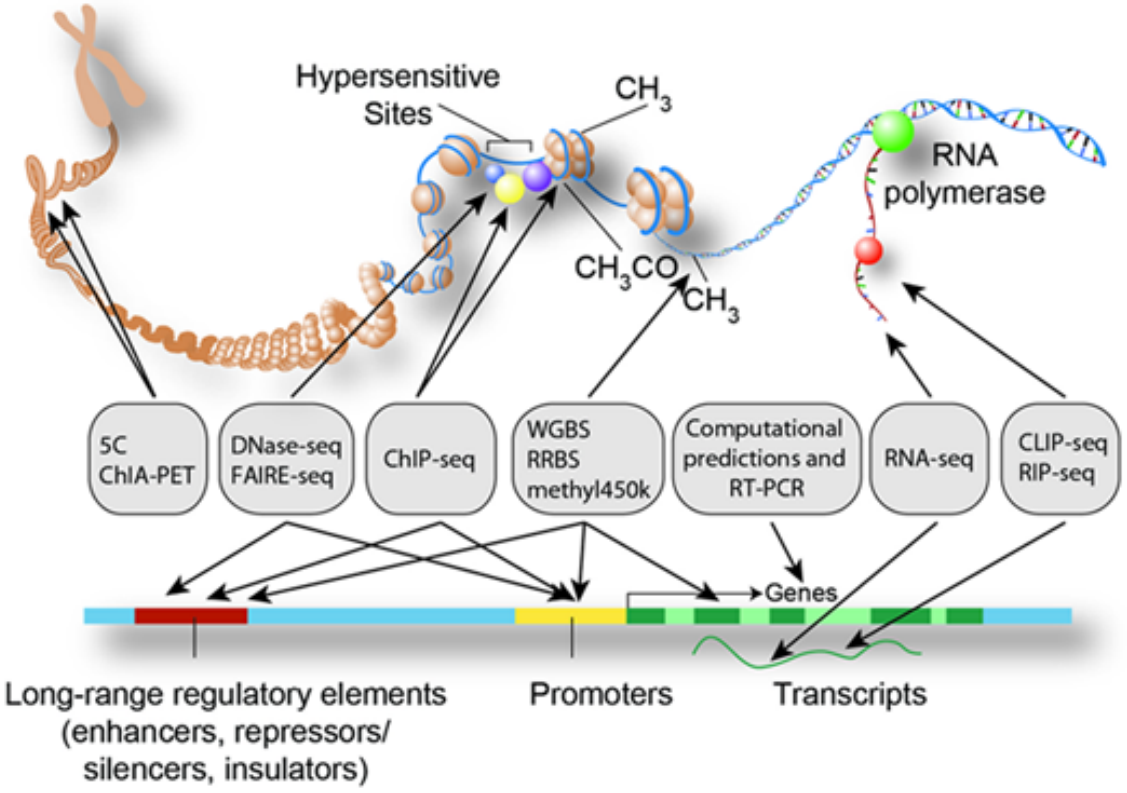
Associated SNP

Controls: (Individuals without trait)

AGAGCAGTCGACATGTATAGTCTACATGAGATCGACATGAGATCGGTAGAGCAGTGAGATCAACATGATAGCC
AGAGCAGTCGACATGTATAGTCTACATGAGATCAACATGAGATCTGTAGAGCCGTGAGATCGACATGATAGCC
AGAGCAGTCGACATGTATAGCCTACATGAGATCGACATGAGATCTGTAGAGCCGTGAGATCAACATGATAGCC
AGAGCCGTCGACAGGTATAGCCTACATGAGATCGACATGAGATCTGTAGAGCCGTGAGATCGACATGATAGTC
AGAGCCGTCGACAGGTATAGTCTACATGAGATCGACATGAGATCTGTAGAGCCGTGAGATCAACATGATAGCC
AGAGCAGTCGACAGGTATAGTCTACATGAGATCGACATGAGATCTGTAGAGCAGTGAGATCGACATGATAGCC
AGAGCCGTCGACAGGTATAGCCTACATGAGATCGACATGAGATCTGTAGAGCCGTGAGATCGACATGATAGCC
AGAGCCGTCGACAGGTATAGTCTACATGAGATCAACATGAGATCTGTAGAGCAGTGAGATCGACATGATAGTC

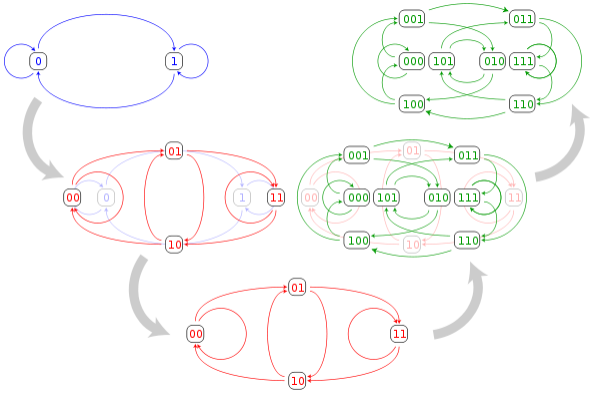
Genomics Data is Pervasive

A multitude of functional genomic assays to understand what is encoded in the genome and the function of genes

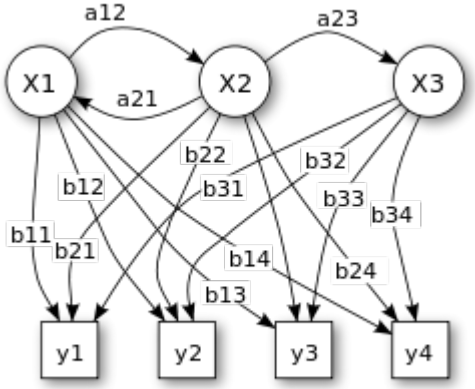


Source: ENCODE

Algorithms and computational methods are critical to genomics



De bruijn graphs – genome assembly



Hidden Markov Models – genome annotation

Many others covered in this course!

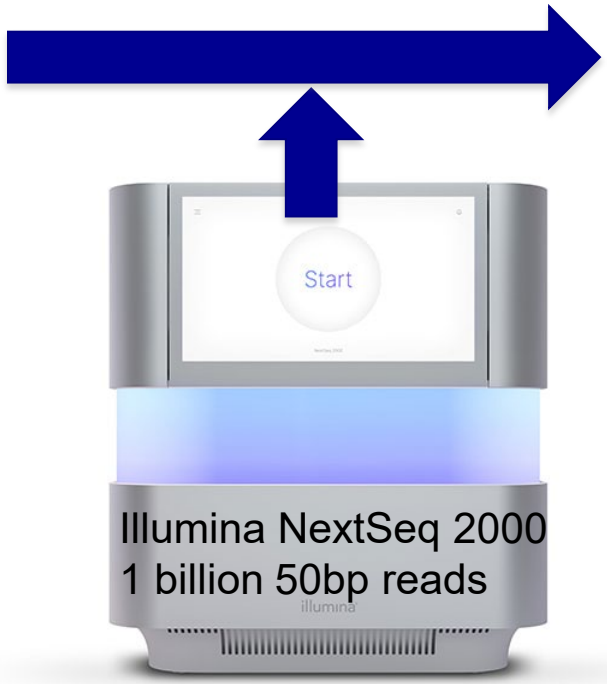
Image sources wikipedia

Lecture Outline

- Course logistics
- Introduction
- Read Mapping

Active Research Problem: Short Read Re-sequencing

Where are my mutations?



AGAGCAGTCGACAGGTA
TAGTCTACATGAGATCG
ACATGAGATCGGTAGAG
CCGTGAGATCGACATGA
TAGCCAGAGCAGTCGAC
AGGTATAGTCTACATGA
GATCGACATGAGATCGG
TAGAGCCGTGAGATCGA
CATGATAGCCAGAGCAG
TCGACAGGTATAGTCTA
CATGAGATCGACATGAG
ATCGGTAGAGCCGTGAG
ATCGACATGATAGCCAG
AGCAGTCGACAGGTATA
GTCTACATGAGATCGAC
ATGAGATCGGTAGAGCC
GTGAGATCGACATGATA
GCCAGAGCAGTCGACAG
GTATAGTCTACATGAGA
TCGACATGAGATCGGTA
GAGCCGTGAGATCGACA
TGATAGCCAGAGCAGTC
GACAGGTATAGTCTACA
TGAGATCGACATGAGAT
CGGTAGAGCCGTGAGAT
CGACATGATAGCCAGAG
CAGTCGACAGGTATAGT
CTACATGAGATCGACAT
GAGATCGGTAGAGCCGT
GAGATCGACATGATAGC

- Next generation sequencing.
  - **Cheap sequencing.**
  - **“Short Reads”**

Short Read Sequencing Problem (A Computer Science Problem)

Full DNA Sequence

AGAGCAGTCGACAGGTA
TAGTCTACATGAGATCG
ACATGAGATCGGTAGAG
CCGTGAGATCGACATGA
TAGCCAGAGCAGTCGAC
AGGTATAGTCTACATGA
GATCGACATGAGATCGG
TAGAGCCGTGAGATCGA
CATGATAGCCAGAGCAG
TCGACAGGTATAGTCTA
CATGAGATCGACATGAG
ATCGGTAGAGCCGTGAG
ATCGACATGATAGCCAG
AGCAGTCGACAGGTATA
GTCTACATGAGATCGAC
ATGAGATCGGTAGAGCC
GTGAGATCGACATGATA
GCCAGAGCAGTCGACAG
GTATAGTCTACATGAGA
TCGACATGAGATCGGTA
GAGCCGTGAGATCGACA
TGATAGCCAGAGCAGTC
GACAGGTATAGTCTACA
TGAGATCGACATGAGAT
CGGTAGAGCCGTGAGAT
CGACATGATAGCCAGAG
CAGTCGACAGGTATAGT
CTACATGAGATCGACAT
GAGATCGGTAGAGCCGT
GAGATCGACATGATAGC

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

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

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

A Sequence Read:
TCGACATGAGATCGGTAGAGCCGT

The Human Genome:
TACATGAGATC C ACATGAGATC T GTAGAGC T GTGAGATC
TC G ACATGAGATC G GTAGAGC C GT

Recovered Sequence:
TACATGAGATC**G**ACATGAGATC**G**GTAGAGC**C**GTGAGATC

"Re"-Sequencing Challenges (Why do we need Computer Science?)

- 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.
  - **A trivial mapping algorithm will take thousands of years to compute for a 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 854,164 years!

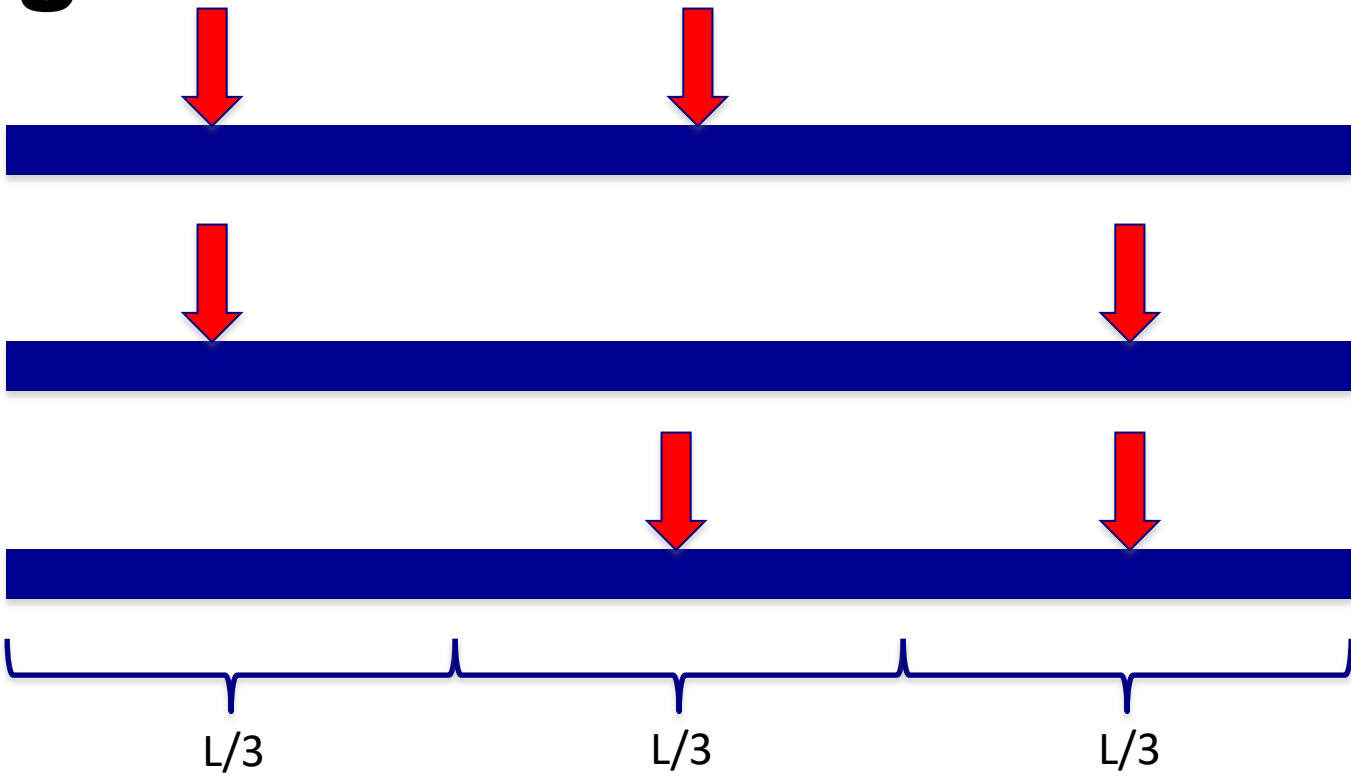
Some observations

- 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.

More problems: 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:

Recovered Sequence:

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

Sequencing Errors: Solution

- Collect redundant data.

My Genome:

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

Sequence Reads:

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

The Human Genome:

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

Recovered Sequence:

TACATGAGATC**G**ACATGAGATC**G**GTAGAGC**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 3.
- We will make a prediction with an error if two out of our three reads have an error in the same place.
- This is approximately $3e^2$.

"Re"-Sequencing Problems

The Human Genome:

TACATGAGATCCACATGAGATCTGTACATGAGATCCACAT

My Genome:

TACATGAGATCGACATGAGATCGGTACATGAGATCCACAT

A Sequence Read:

ACATGAGATCGACAT

The Human Genome:

TACATGAGATCGACATGAGATCTGTACATGAGATCCACAT
ACATGAGATCGACAT ACATGAGATCGACAT

Error!

Recovered Sequence:

TACATGAGATCGACATGAGATCGGTACATGAGATCGACAT

"Re"-Sequencing Problems

The Human Genome:

TACATGAGATCCACATGAGATCTGTACATGAGATCCACAT

My Genome:

TACATGAG**GGGGGGGG**GAGATCGGTACATGAGATCCACAT

A Sequence Read:

GAG**GGGGGGGG**

The Human Genome:

TACATGAGATCCACATGAGATCTGTACATGAGATCCACAT

GAG**GGGGGGG**G

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.

"Re"-Sequencing: Insertions

My Genome:

TACATGAGATCCACAT**A**GAGATCTGTAGAGCTGTGAGATC

A Sequence Read:

CCACATAGAGATCTGTAGAGCTGT

The Human Genome:

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

CCACATAGAGATCTGTAGAGCTGT

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

CCACATAGAGATCTGTAGAGCTGT

How do we deal with this case?

"Re"-Sequencing: Insertions

My Genome:

TACATGAGATCCACATAGAGATCTGTAGAGCTGTGAGATC

A Sequence Read:

CCACATAGAGATCTGTAGAGCTGT

The Human Genome:

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
CCACATAGAGATCTGTAGAGCTGT

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
CCACATAGAGATCTGTAGAGCTGT

Solution: Add Insertion to the Human Genome

TACATGAGATCCACAT-GAGATCTGTAGAGCTGTGAGATC
CCACATAGAGATCTGTAGAGCTGT

Many other challenges

- Coverage of sequence reads is not uniform
  - **Some places we have many reads, while some we have fewer. How do we design an approach so we can always recover the sequence.**
- Large memory requirements
  - **We need to fit our index into RAM. Often tens of Gigabytes or greater.**

Sequencing Coverage

Lecture 1.

April 4th , 2023

(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

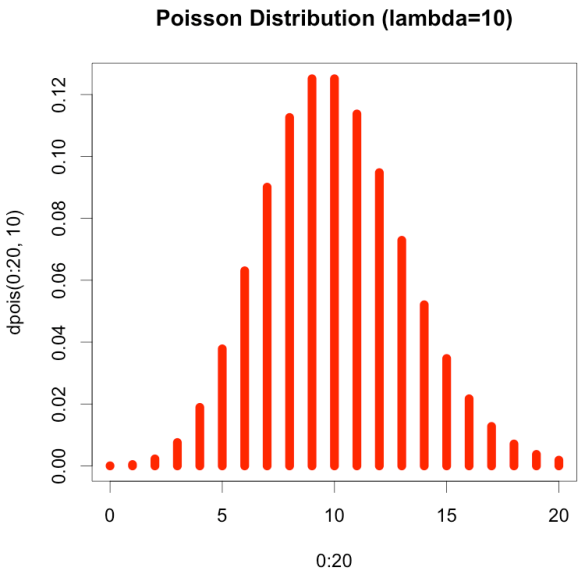
- If length of the genome is N the probability of the event that a single read of length L position spanning a single position in the genome approximately L/N (also very small).
- If the sum of the length of all K reads of length L is M=K*L, the total number of read positions that start at a single genome 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

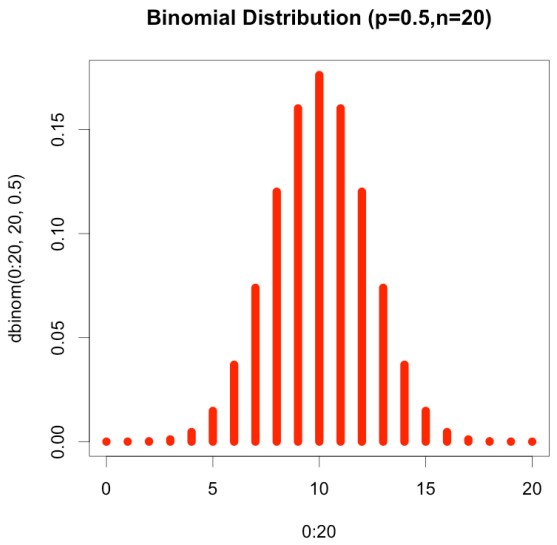
- Since humans have 2 chromosomes each read comes from one chromosome at random. If a position in the reference is covered by Y reads, the probability that X of the reads come from the first chromosome follows the binomial distribution with parameter .5.
  - **dbinom(X,Y,0.5)**
- At least X coverage 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□ACATGAGATC□GTAGAG□T□GTGAGATC

TC□ACATGAGATC□GTAGA□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?

Insertions and Deletions

Lecture 1.

April 4th, 2023

"Re"-Sequencing: Insertions

My Genome:

TACATGAGATCCACAT**A**GAGATCTGTAGAGCTGTGAGATC

A Sequence Read:

CCACATAGAGATCTGTAGAGCTGT

The Human Genome:

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

CCACATAGAGATCTGTAGAGCTGT

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

CCACATAGAGATCTGTAGAGCTGT

How do we deal with this case?

"Re"-Sequencing: Insertions

My Genome:

TACATGAGATCCACATAGAGATCTGTAGAGCTGTGAGATC

A Sequence Read:

CCACATAGAGATCTGTAGAGCTGT

The Human Genome:

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
CCACATAGAGATCTGTAGAGCTGT

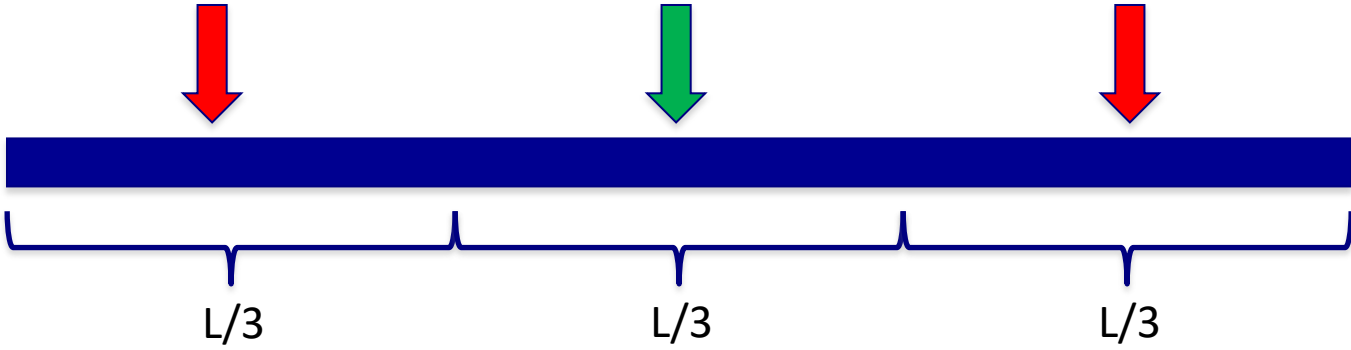
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
CCACATAGAGATCTGTAGAGCTGT

Solution: Add Insertion to the Human Genome

TACATGAGATCCACAT-GAGATCTGTAGAGCTGTGAGATC
CCACATAGAGATCTGTAGAGCTGT

Indel in Middles of Read

If indel is in the middle of read.



- Both outside regions of size L/3 will match perfectly.
- Because of coverage, indel will be in middle at least for one read.
- Important: Middle distance will be L/3+1 or L/3-1

"Re"-Sequencing: Insertions

My Genome:

TACATGAGATCCACATAGAGATCTGTAGAGCTGTGAGATC

A Sequence Read:

CCACATAGAGATCTGTAGAGCTGT

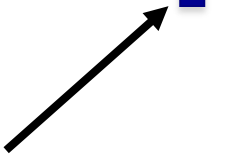
The Human Genome:

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

CCACATAGAGATCTGTAGAGCTGT

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

CCACATAGAGATCTGTAGAGCTGT



Insertion Point

Indel Algorithms

- Trivial Algorithm
  - **Try all inerstion points for a read**
  - **If read matches (with insertion) below number of mismatches, then we desclare a match and identify and indel**
- More Efficient Algorithm
  - **Look for perfect match in first part of read**
  - **Try insertion point at point of first mismatch**
  - **More complicated but faster**
- More accurate Algorithm
  - **Perform alignment between read and reference**
- Extremely Accurate Algorithm
  - **Align all reads with indel together.**
  - **Multiple Sequence Alignment!**