Genome Sequencing

Michael Schatz

Sept 4, 2019

Lecture 2: Computational Biomedical Research

Welcome!

The goal of this course is to prepare undergraduates to understand and perform state-of-the-art biomedical research. This will be accomplished through three main components:

1. **Lectures** on cross cutting techniques for biomedical research focusing on data visualization, statistical inference, and scientific computing
2. **Research presentations** from distinguished faculty on their active research projects
3. **A major research project** with in-class research labs;
   Satisfies the CS TEAM requirement

Course Webpage:
https://github.com/schatzlab/biomedicalresearch2019

Course Discussions: http://piazza.com

Class Hours: Mon + Wed @ 3p – 3:50p Hodson 311

Office Hours: Monday @ 4-5p and by appointment
Please try Piazza first!

Course Webpage

EN.601.452 / AS.020.415 Computational Biomedical Research & Advanced Biomedical Research

Class Hours: Monday + Wednesday @ 3p - 3:50p in Hodson 311
Schatz Office Hours: Wednesday @ 4-5p in Malone 323 and by appointment
CA: M. Kirsche - Office Hours TBD

The goal of this course is to prepare undergraduates to understand and perform state-of-the-art biomedical research. This will be accomplished through three main components: (1) classroom-style lectures on cross cutting techniques for biomedical research focusing on data visualization, statistical inference, and scientific computing; (2) research presentations from distinguished faculty on their active research projects; and (3) a major research project to be performed under the mentorship of a JHU professor. Students will present their research during an in-class symposium at the end of the semester. Grading will be based on homework exercises, a written research proposal, an interim research report, an oral research presentation, and a final research report.

Course Resources:

- Syllabus and Policies
- Piazza Discussion Board
- GradeScope Entry Code: MPK8BX

Recommended Prerequisites

- Online introduction to Unix/Linux. Students are strongly recommended to complete one of the following online tutorials before class begins.
  - Code academy's Intro to Unix
  - Command line bootcamp
- Access to a Linux Machine, or Install VirtualBox (Unfortuantely, even Mac will not work correctly for some programs)

Related Courses & Readings

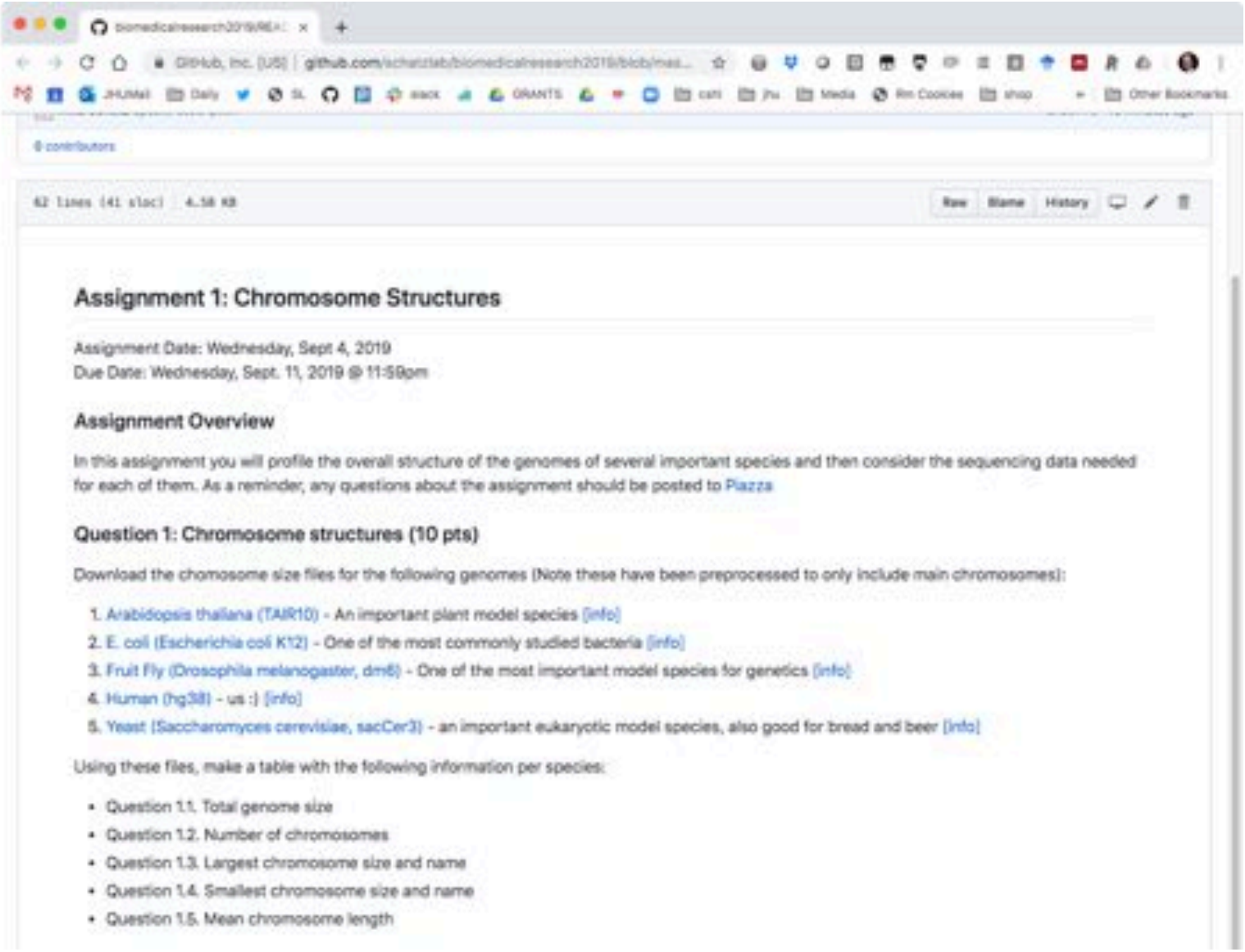
- Applied Comparative Genomics
- Bioinformatics Algorithms: An Active Learning Approach (Compeau and Pevzner)
- Unix and Perl for Biologists (Bradnam et al)
- Molecular Biology of the Gene (Watson et al)
- Molecular Biology of the Cell (Alberts)
- Biological Sequence Analysis (Durbin et al)
- Modern Statistics for Modern Biology (Holmes & Huber)

Schedule

#	Date	Lecture	Readings & Resources	Assignment
1	1/7-8/19	Lecture 1: Introduction	* Biological data sciences in genome research (Schatz, 2015, Genome Research)	Sign Up for Piazza

https://github.com/schatzlab/biomedicalresearch2019

Assignment 1: Chromosome Structures
Due Wed Sept 11 @ 11:59pm



Assignment 1: Chromosome Structures

Assignment Date: Wednesday, Sept 4, 2019
Due Date: Wednesday, Sept. 11, 2019 @ 11:59pm

Assignment Overview

In this assignment you will profile the overall structure of the genomes of several important species and then consider the sequencing data needed for each of them. As a reminder, any questions about the assignment should be posted to Piazza

Question 1: Chromosome structures (10 pts)

Download the chomosome size files for the following genomes (Note these have been preprocessed to only include main chromosomes):

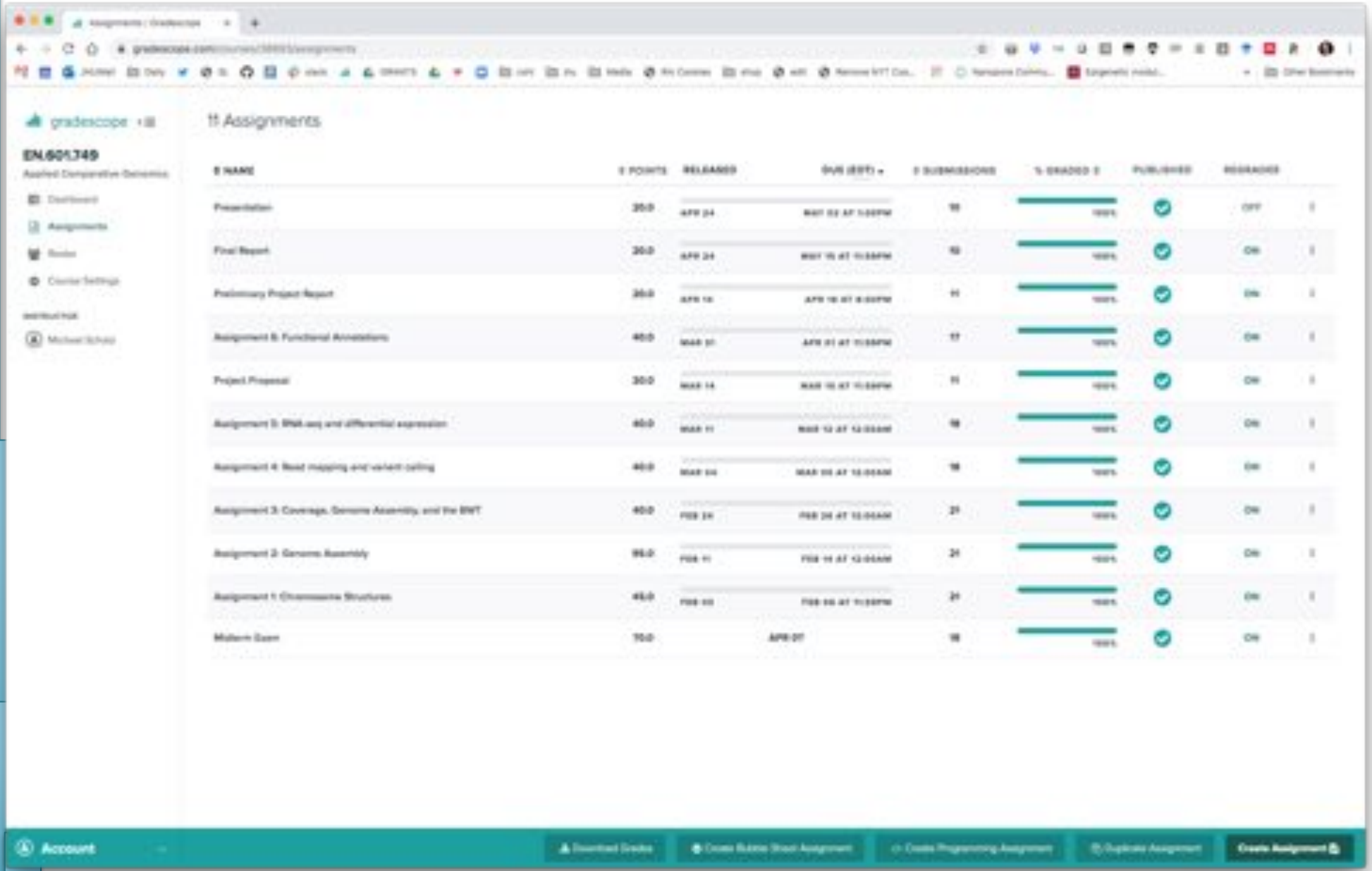
1. Arabidopsis thaliana (TAIR10) - An important plant model species [info]
2. E. coli (Escherichia coli K12) - One of the most commonly studied bacteria [info]
3. Fruit Fly (Drosophila melanogaster, dm6) - One of the most important model species for genetics [info]
4. Human (hg38) - us :) [info]
5. Yeast (Saccharomyces cerevisiae, sacCer3) - an important eukaryotic model species, also good for bread and beer [info]

Using these files, make a table with the following information per species:

- Question 1.1. Total genome size
- Question 1.2. Number of chromosomes
- Question 1.3. Largest chromosome size and name
- Question 1.4. Smallest chromosome size and name
- Question 1.5. Mean chromosome length

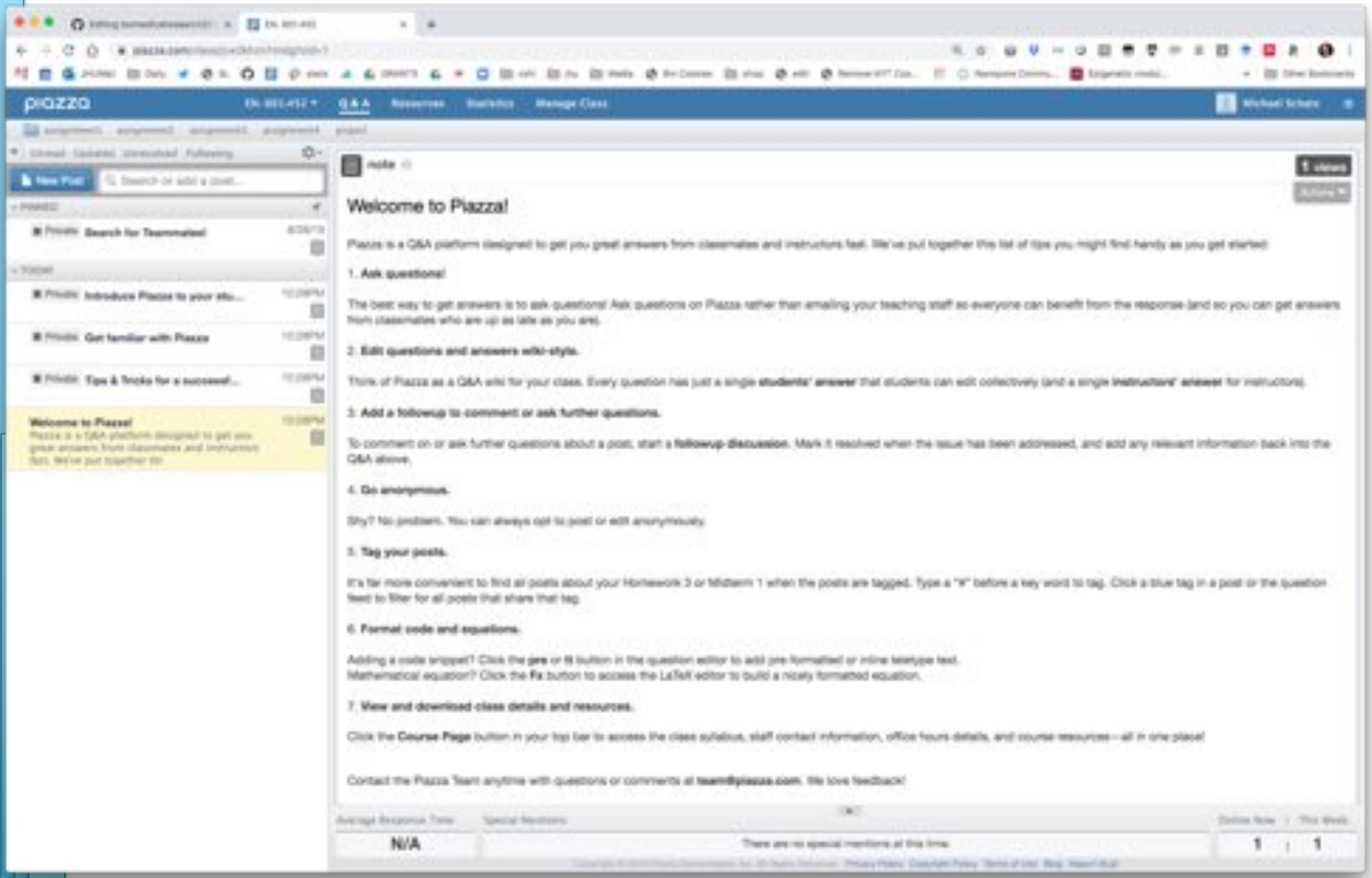
https://github.com/schatzlab/biomedicalresearch2019

GradeScope



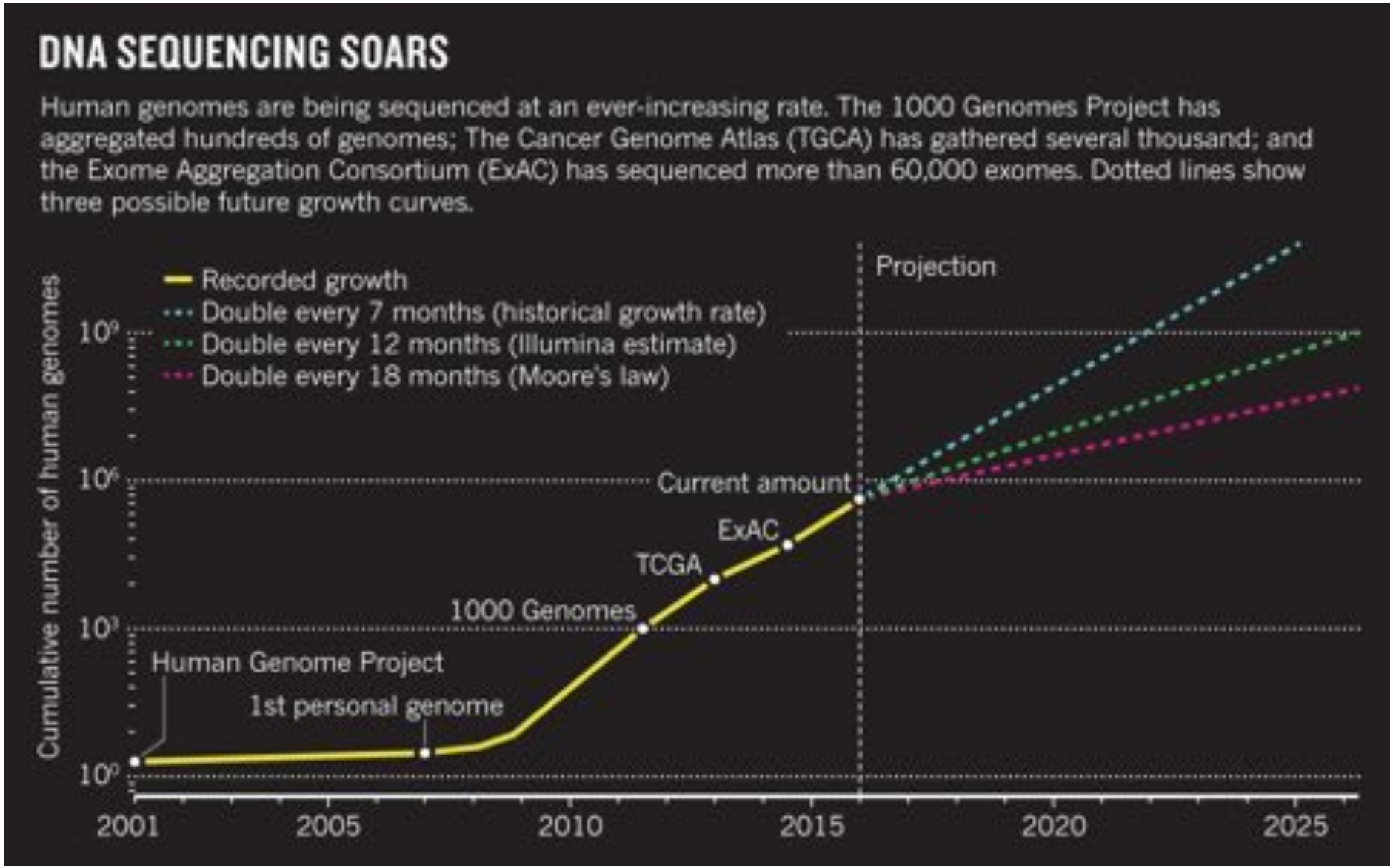
https://www.gradescope.com/courses/60230 Entry Code: MPK8BX

Piazza



http://piazza.com/jhu/fall2019/en601452

Sequencing Capacity



Big Data: Astronomical or Genomical?

Stephens, Z, et al. (2015) PLOS Biology DOI: 10.1371/journal.pbio.1002195

De novo Genetics of Autism

- In 593 family quads so far, we see significant enrichment in de novo ***likely gene killers*** in the autistic kids
  - Overall rate basically 1:1
  - 2:1 enrichment in nonsense mutations
  - 2:1 enrichment in frameshift indels
  - 4:1 enrichment in splice-site mutations
  - Most de novo originate in the paternal line in an age-dependent manner (56:18 of the mutations that we could determine)

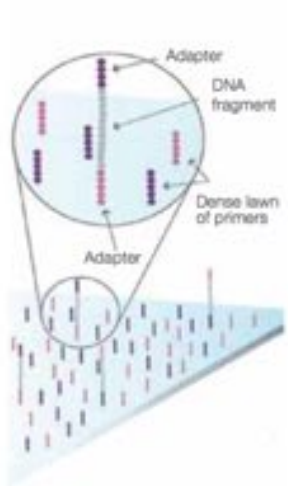
- Observe strong overlap with the 842 genes known to be associated with fragile X protein FMPR
  - Related to neuron development and synaptic plasticity
  - Also strong overlap with chromatin remodelers

Accurate de novo and transmitted indel detection in exome-capture data using microassembly.
Narzisi *et al* (2014) Nature Methods doi:10.1038/nmeth.3069

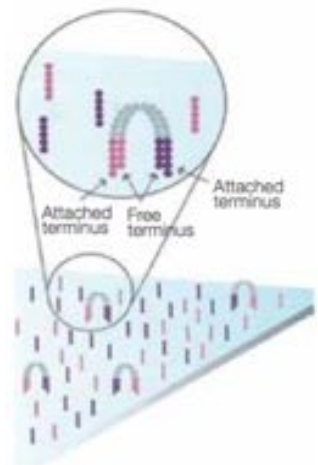
Second Generation Sequencing

Illumina NovaSeq 6000
Sequencing by Synthesis

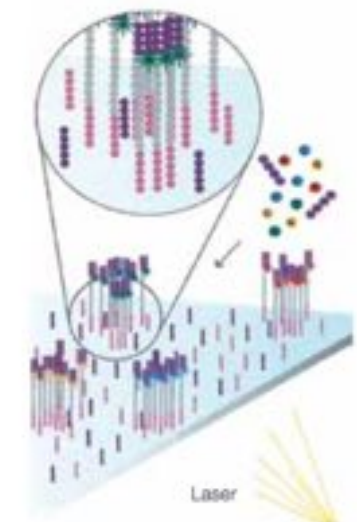
>3Tbp / day



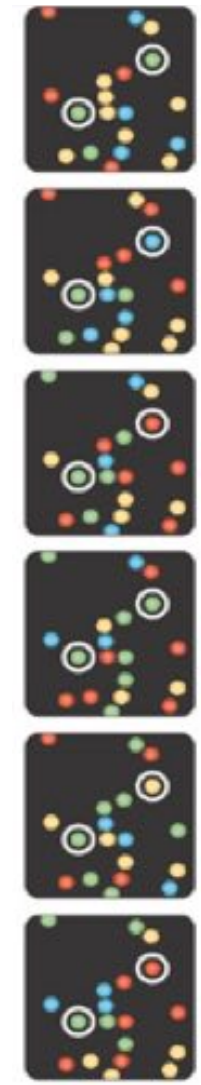
1. Attach



2. Amplify



3. Image

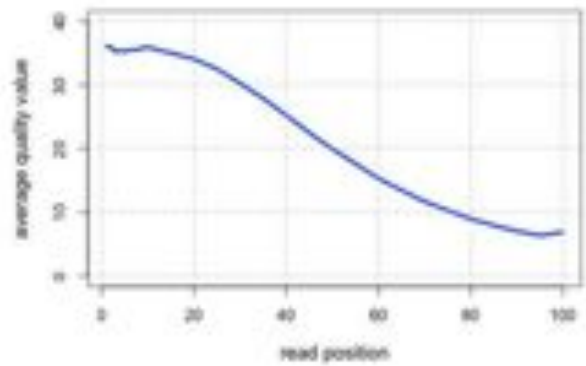


Metzker (2010) Nature Reviews Genetics 11:31-46
https://www.youtube.com/watch?v=fCd6B5HRaZ8

Illumina Quality

QV	p_{error}
40	1/10000
30	1/1000
20	1/100
10	1/10

$$Q_{\text{sanger}} = -10 \log_{10} p$$

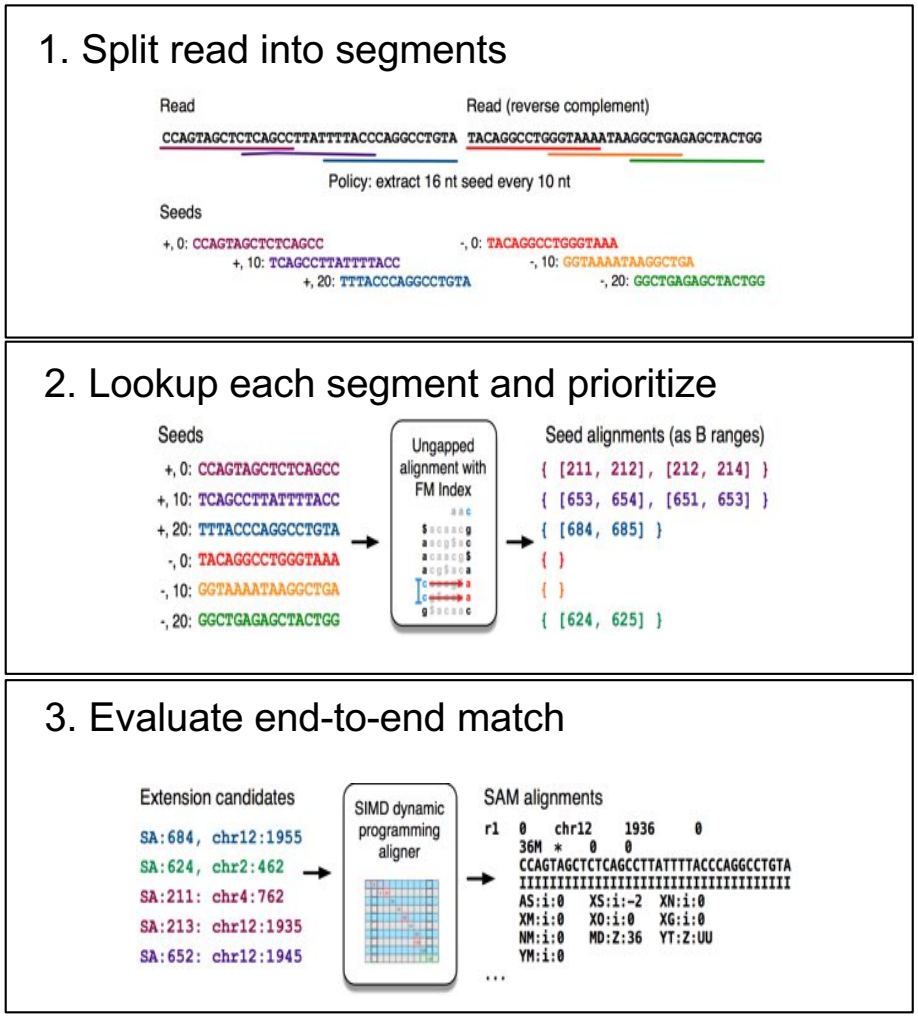


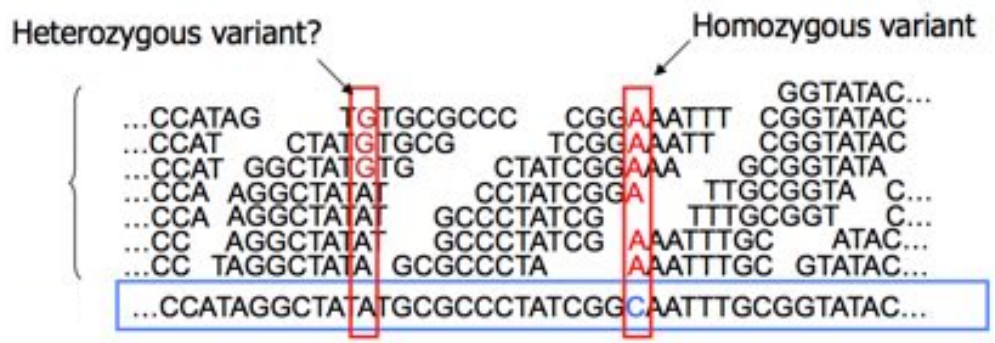

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
.........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.......................
................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.......................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                            104                   126

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
   with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
   (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

http://en.wikipedia.org/wiki/FASTQ_format

Assembly, Mapping & Genotyping

Week 2/3/4



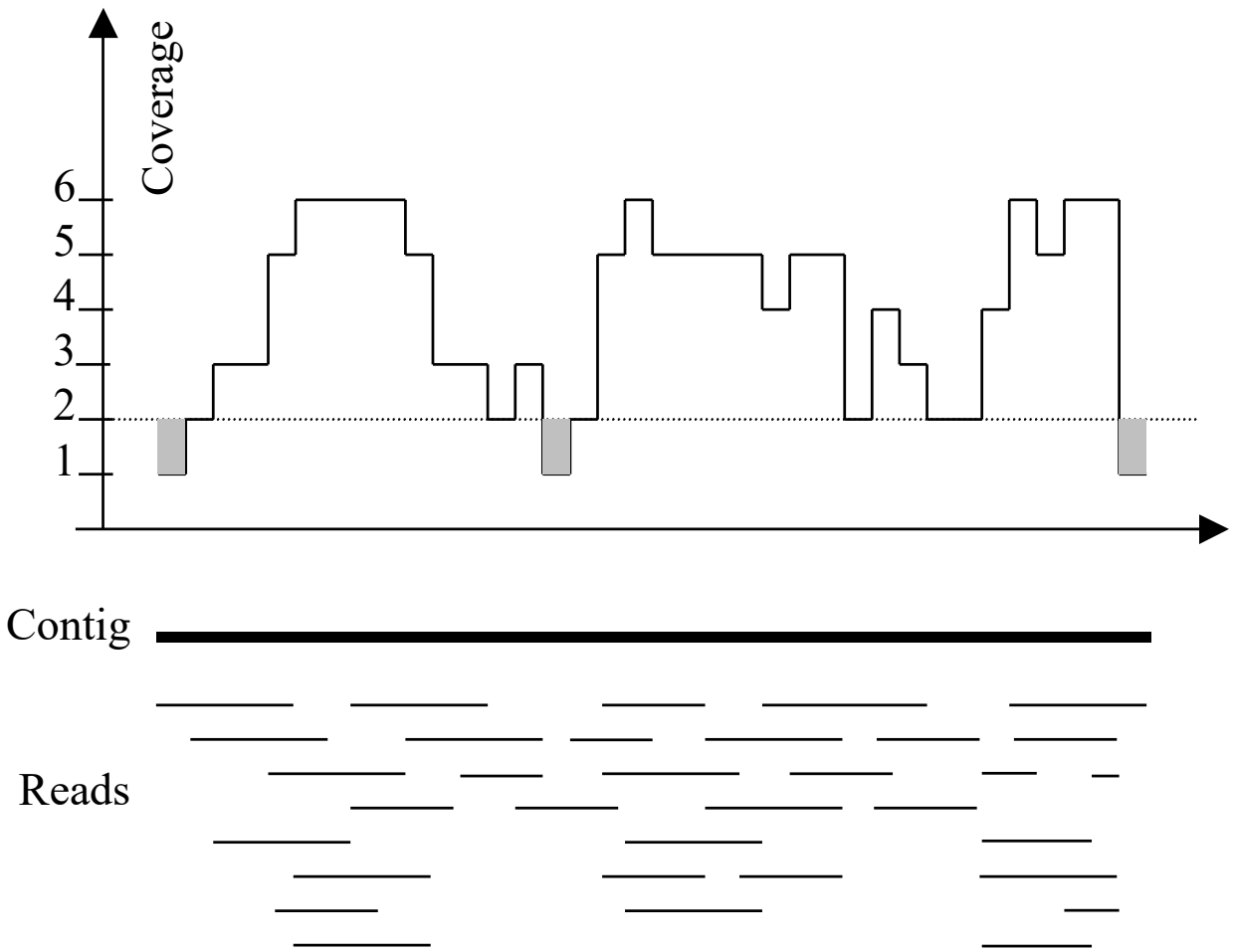


- Distinguishing SNPs from sequencing error typically a likelihood test of the coverage
  - Hardest to distinguish between errors and heterozygous SNP.
  - Coverage is the most important factor!
    - Target at least 10x, 30x more reliable

Fast gapped-read alignment with Bowtie 2
Langmead & Salzberg. (2012) *Nature Methods*. 9:357-359.

The Sequence Alignment/Map format and SAMtools
Li H et al. (2009) *Bioinformatics*. 25:16 2078-9

Typical sequencing coverage

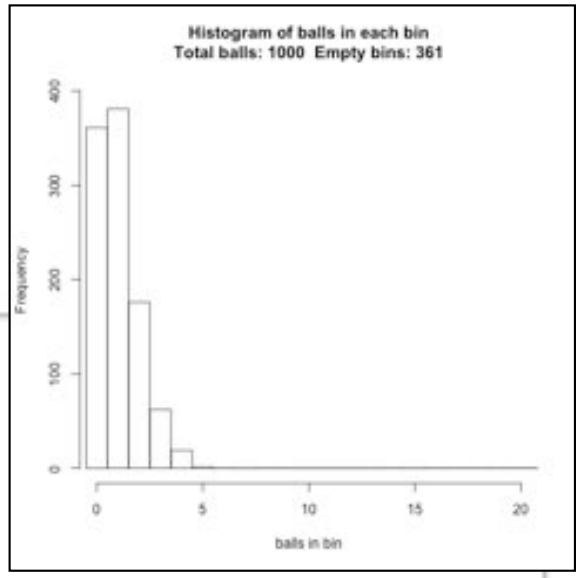


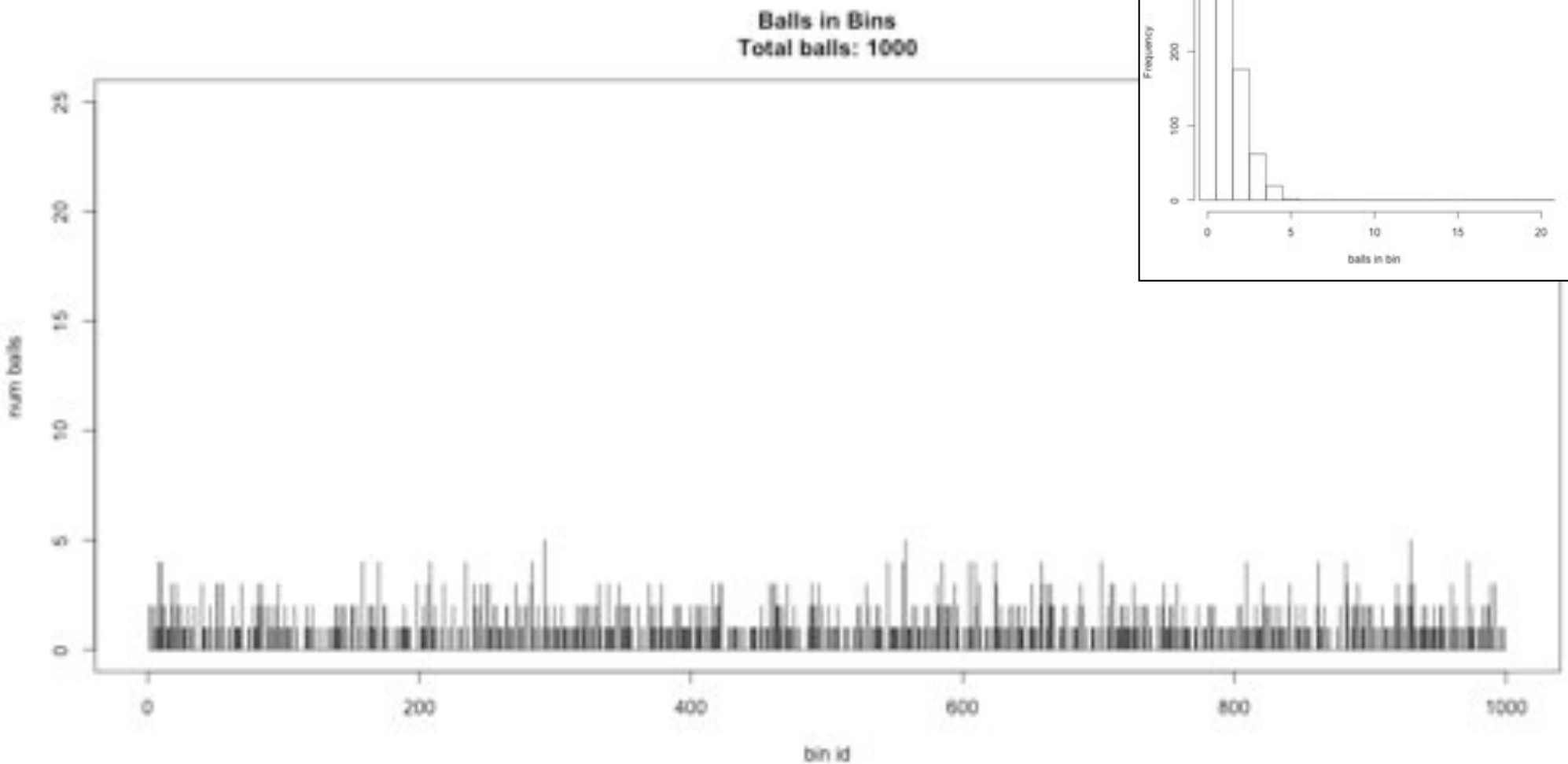
Imagine raindrops on a sidewalk

We want to cover the entire sidewalk but each drop costs $1

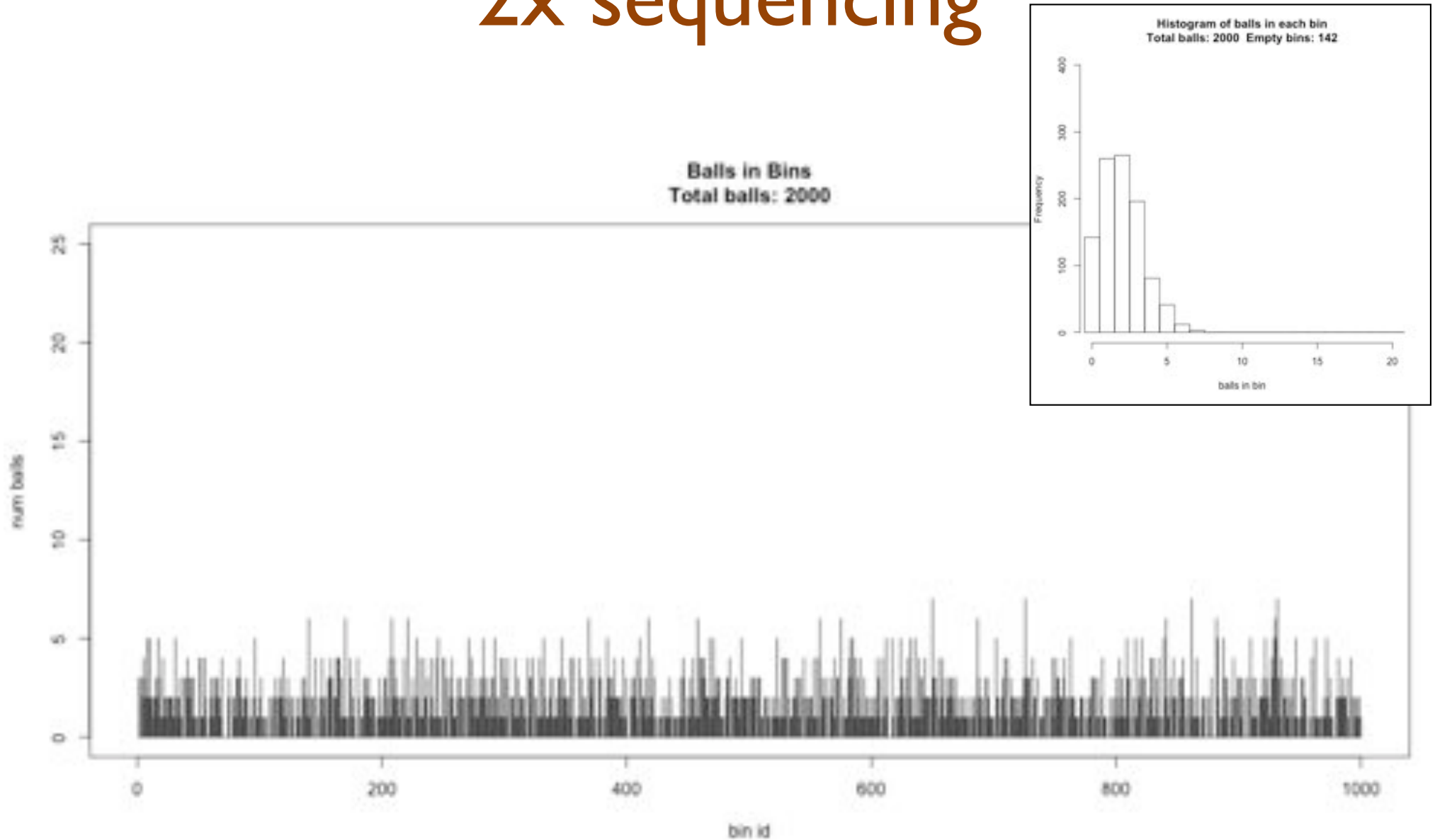
If the genome is 10 Mbp, should we sequence 100k 100bp reads?

1x sequencing

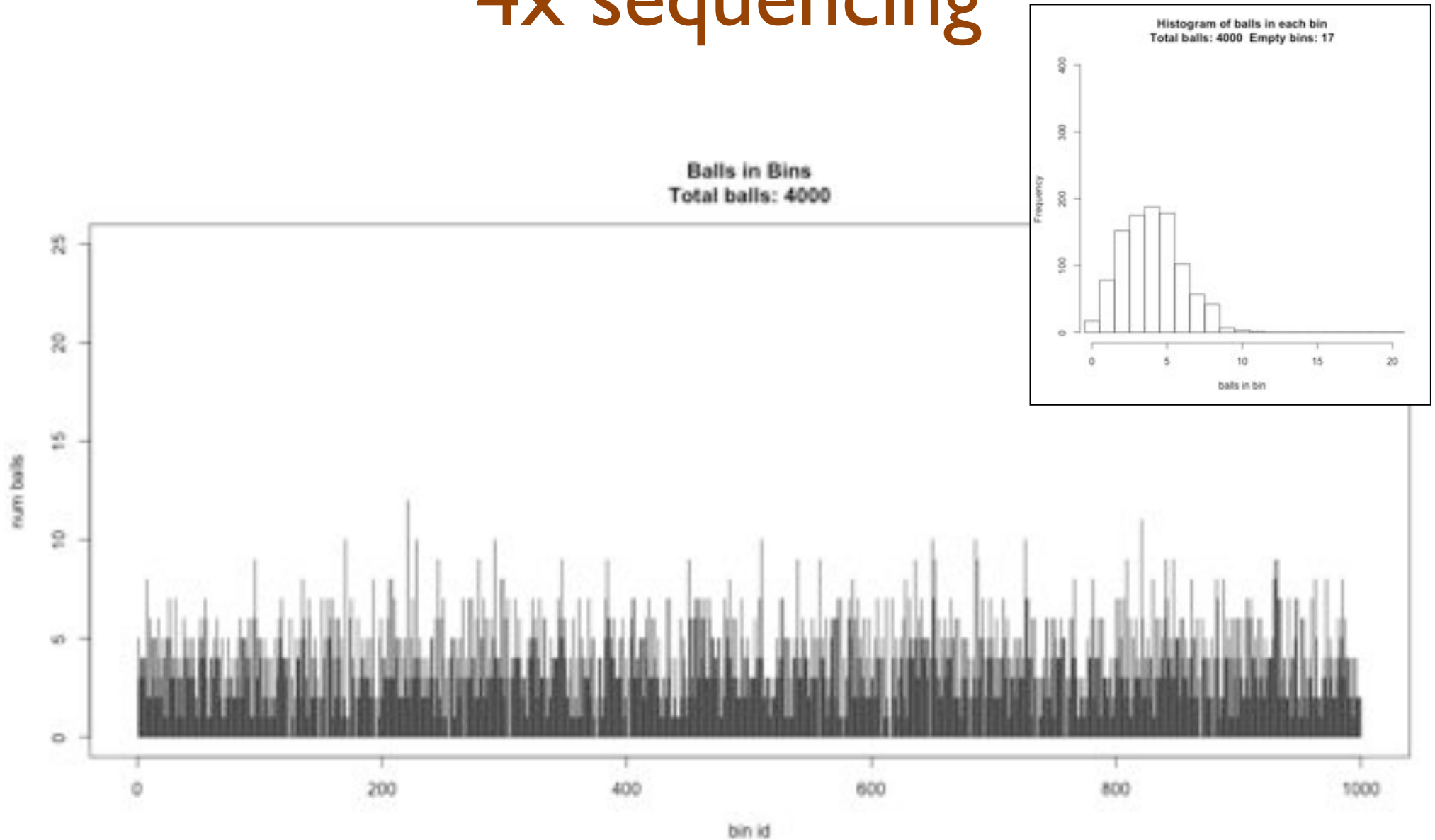




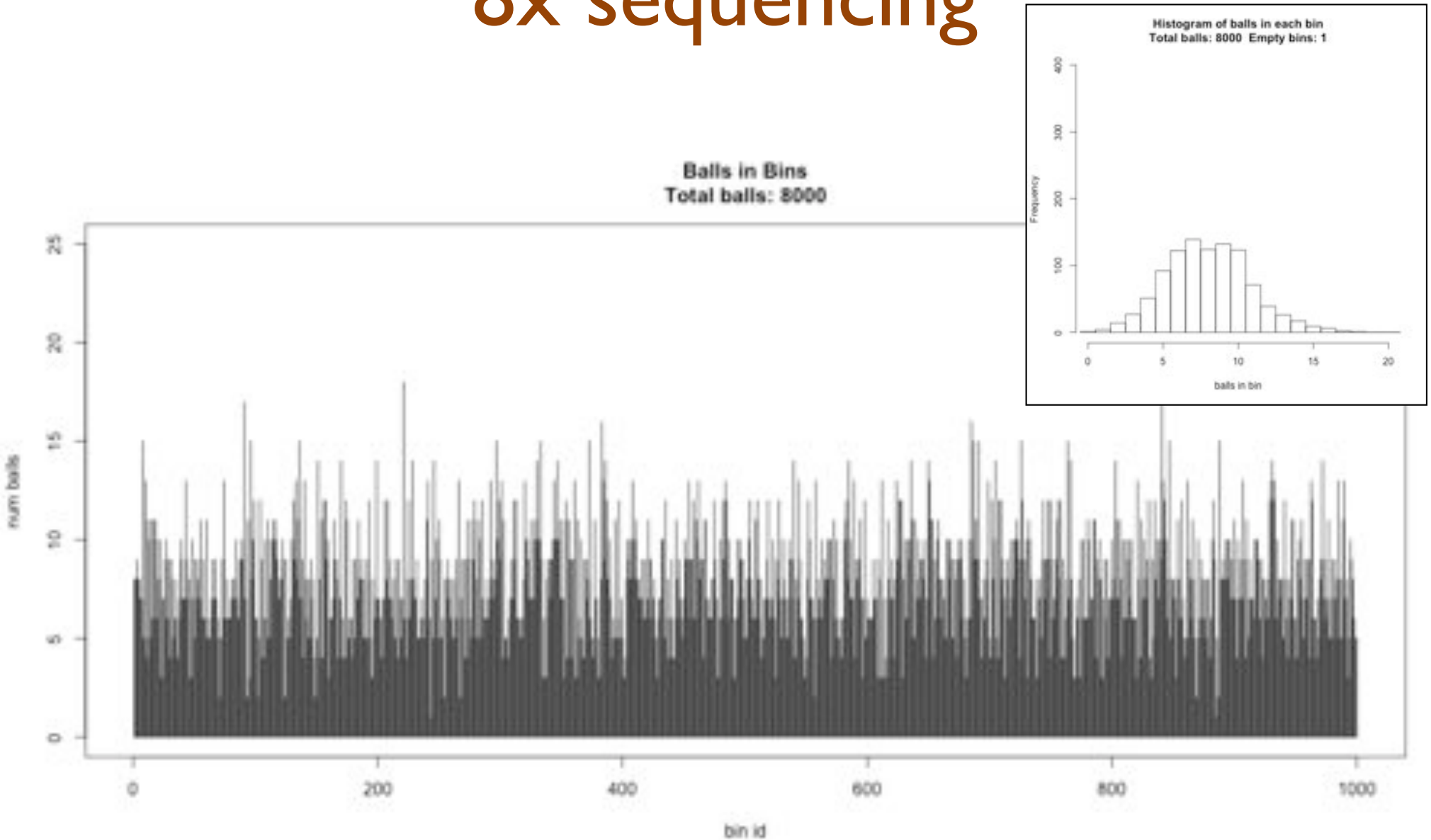
2x sequencing



4x sequencing



8x sequencing



Poisson Distribution

The probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.

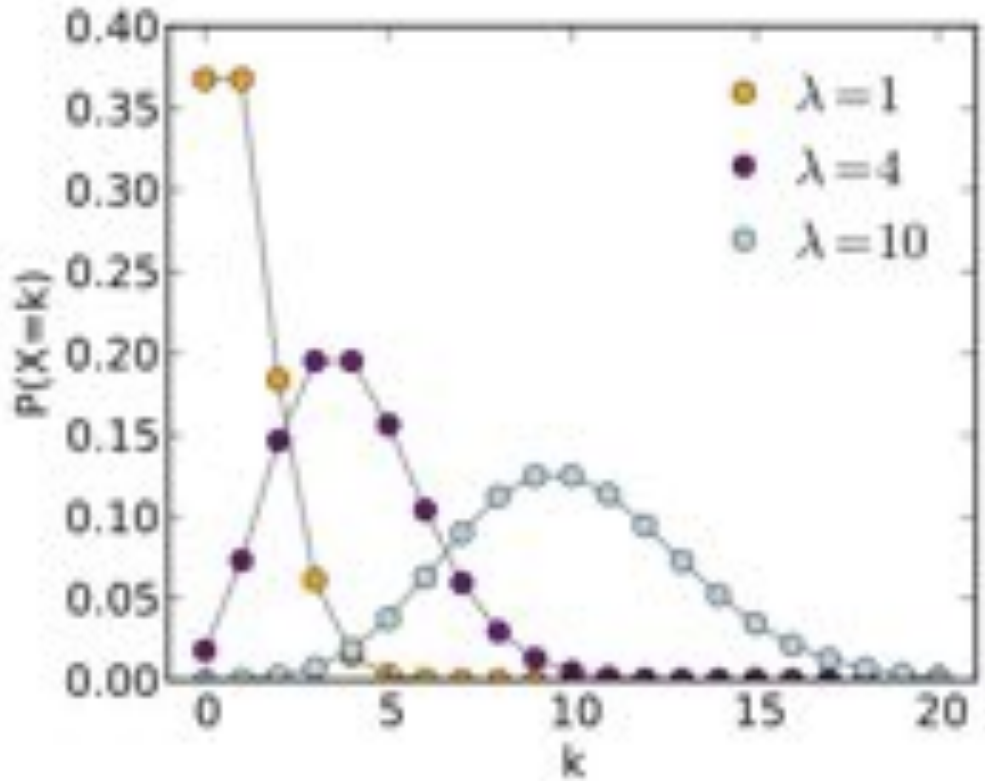
$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Formulation comes from the limit of the binomial equation

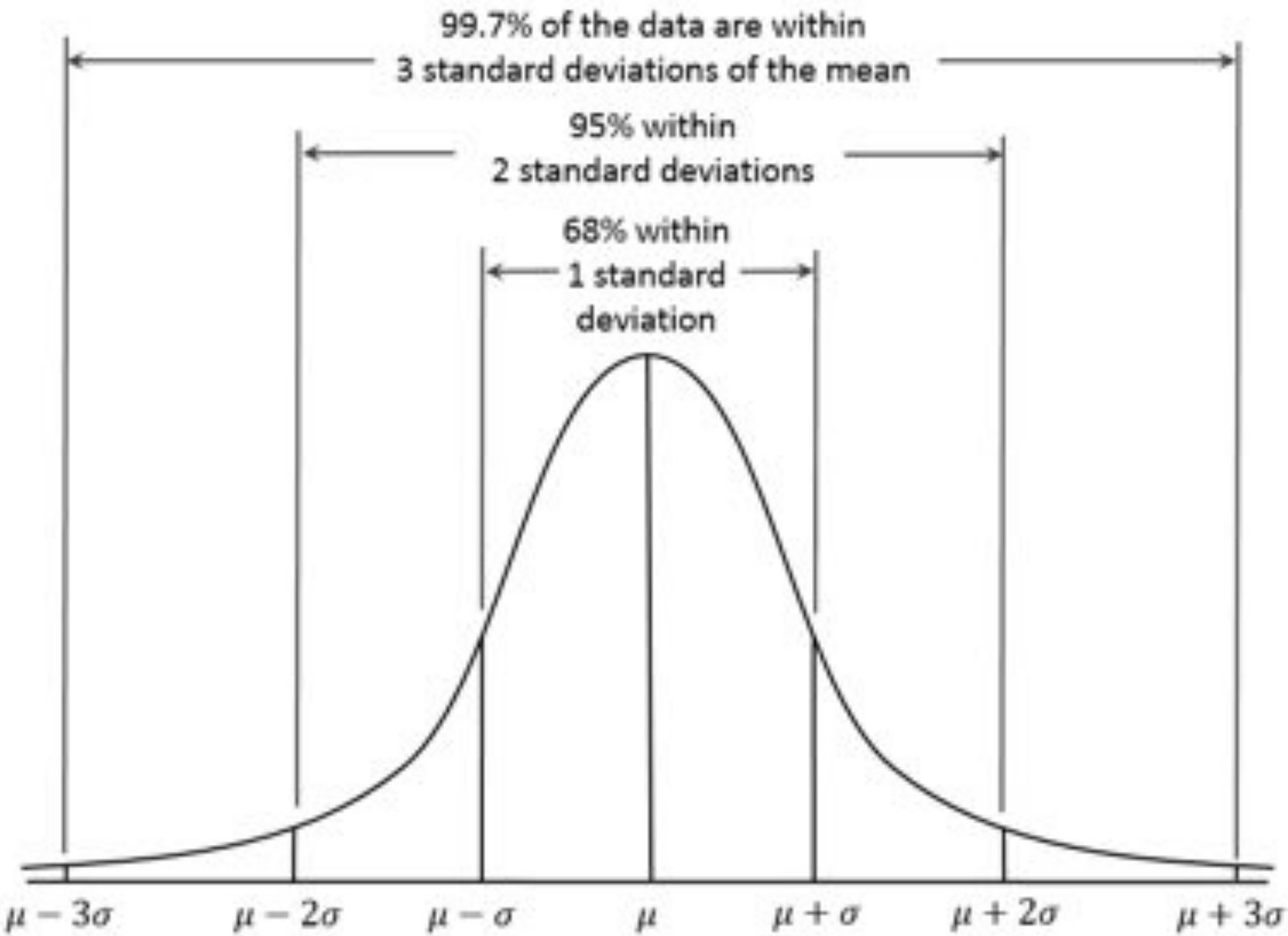
Resembles a normal distribution, but over the positive values, and with only a single parameter.

Key properties:

- ***The standard deviation is the square root of the mean.***
- ***For mean > 5, well approximated by a normal distribution***



Normal Approximation



Can estimate Poisson distribution as a normal distribution when $\lambda > 10$

Pop Quiz!

I want to sequence a 10Mbp genome to 24x coverage.
How many 120bp reads do I need?

I need 10Mbp x 24x = 240Mbp of data
240Mbp / 120bp / read = 2M reads

Next Steps

1. Reflect on the magic and power of DNA ☺
2. Check out the course webpage
3. Register on Piazza & GradeScope
4. Work on Assignment 1
    1. Set up Dropbox for yourself!
    2. Get comfortable on the command line