

# 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: 6P4UB8E](#)

### 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](#) (unfortunately, 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](#) (Bradham 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)

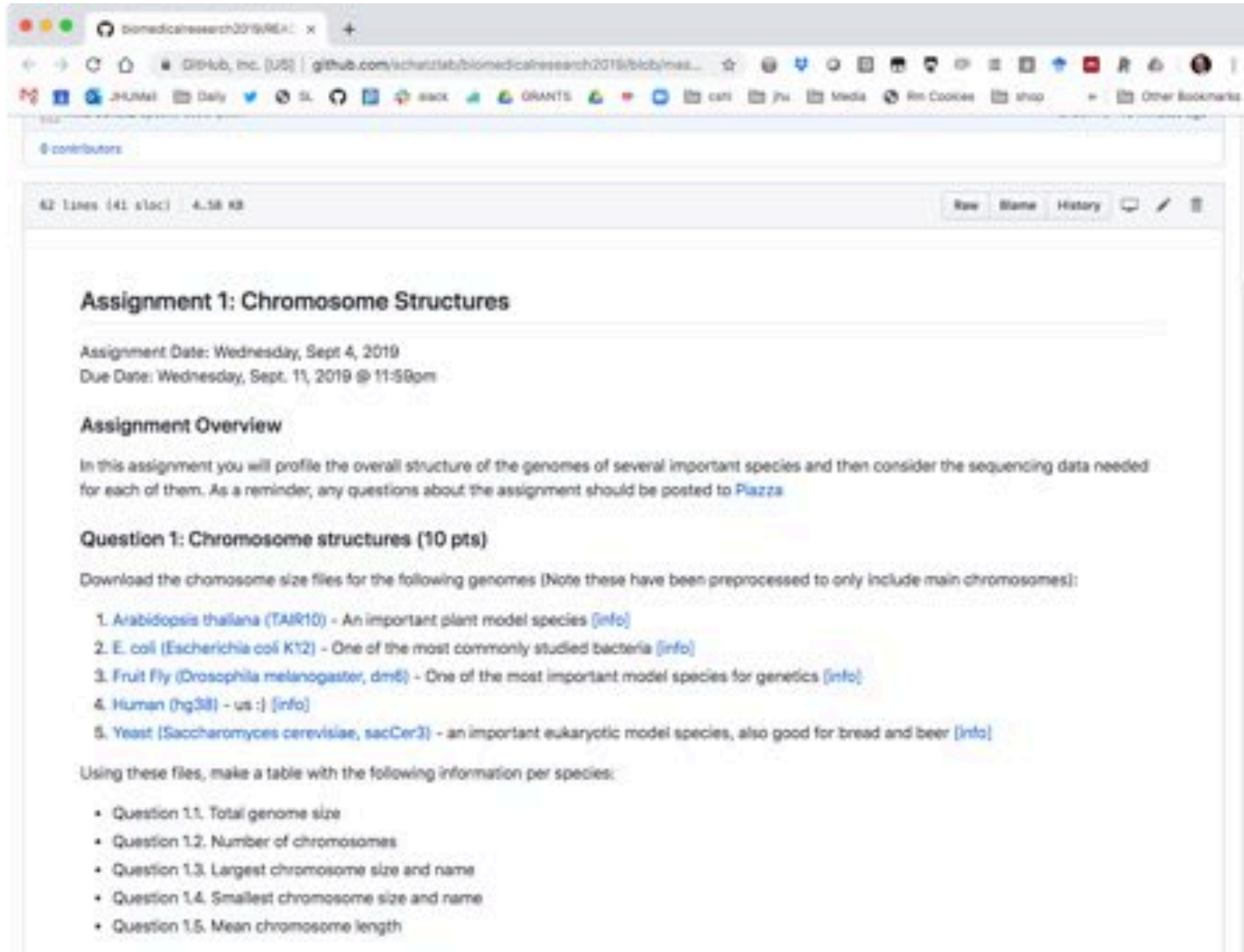
### Schedule

#	Date	Lecture	Readings & Resources	Assignment
1	Tu, 8/23	Lecture 1: Introduction	• <a href="#">Biological data sciences in genome research</a> (Schatz, 2015, Genome Research)	<a href="#">Download Syllabus</a>

<https://github.com/schatzlab/biomedicalresearch2019>

# Assignment I: Chromosome Structures

## Due Wed Sept 11 @ 11:59pm



The screenshot shows a web browser displaying a GitHub repository page. The browser's address bar shows the URL: <https://github.com/schatzlab/biomedicalresearch2019/blob/master/Assignment%201%20Chromosome%20Structures.md>. The repository name is 'biomedicalresearch2019' and the file name is 'Assignment 1: Chromosome Structures'. The file size is 4.58 KB and it has 62 lines of code. The document content includes the assignment title, dates, overview, and a list of five model organisms for which chromosome size files are provided. The organisms are: 1. Arabidopsis thaliana (TAIR10), 2. E. coli (Escherichia coli K12), 3. Fruit Fly (Drosophila melanogaster, dm6), 4. Human (hg38), and 5. Yeast (Saccharomyces cerevisiae, sacCer3). The document also lists five questions to be answered based on the data.

**Assignment 1: Chromosome Structures**

Assignment Date: Wednesday, Sept 4, 2019  
Due Date: Wednesday, Sept. 11, 2019 @ 11:58pm

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

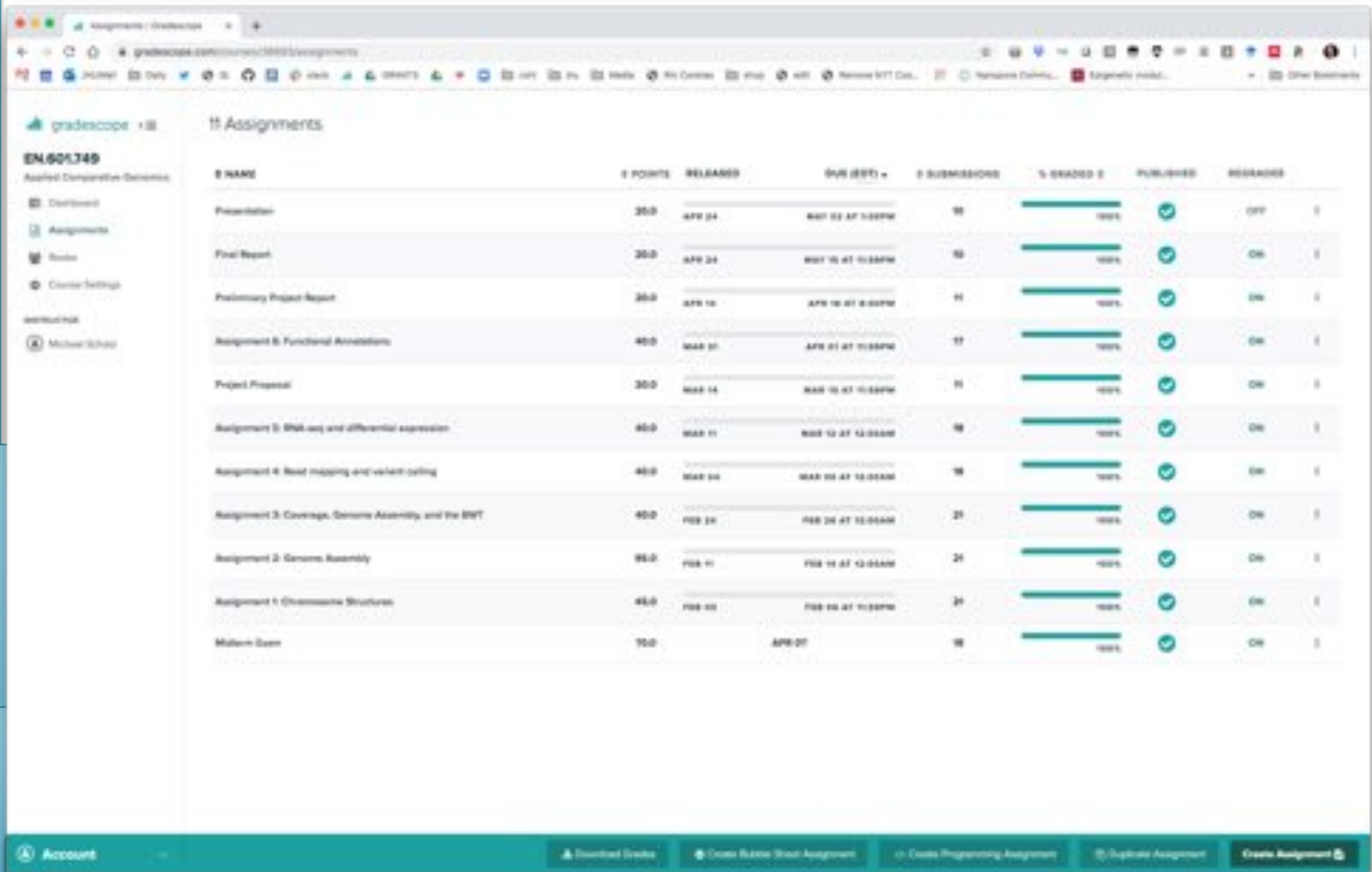
1. Arabidopsis thaliana (TAIR10) - An important plant model species [\[info\]](#)
2. E. coli (Escherichia coli K12) - One of the most commonly studied bacteria [\[info\]](#)
3. Fruit Fly (Drosophila melanogaster, dm6) - One of the most important model species for genetics [\[info\]](#)
4. Human (hg38) - 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



NAME	POINTS	RELEASED	DUE (EST)	SUBMISSIONS	% GRADED	PUBLISHED	GRADED
Presentation	30.0	APR 24	MAY 02 AT 11:00PM	10	100%	✓	OK
Final Report	30.0	APR 24	MAY 10 AT 11:00PM	10	100%	✓	OK
Preliminary Project Report	30.0	APR 10	APR 10 AT 8:00PM	11	100%	✓	OK
Assignment 6: Functional Annotations	40.0	MAR 31	APR 01 AT 11:00PM	17	100%	✓	OK
Project Proposal	30.0	MAR 14	MAR 10 AT 11:00PM	11	100%	✓	OK
Assignment 5: RNA-seq and differential expression	40.0	MAR 11	MAR 12 AT 12:00AM	18	100%	✓	OK
Assignment 4: Read mapping and variant calling	40.0	MAR 04	MAR 05 AT 10:00AM	18	100%	✓	OK
Assignment 3: Coverage, Genome Assembly, and the BWT	40.0	FEB 25	FEB 26 AT 12:00AM	25	100%	✓	OK
Assignment 2: Genome Assembly	35.0	FEB 11	FEB 11 AT 12:00AM	25	100%	✓	OK
Assignment 1: Chromosome Structures	45.0	FEB 03	FEB 03 AT 11:00PM	21	100%	✓	OK
Midterm Exam	10.0		APR 07	18	100%	✓	OK

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

# Piazza

The screenshot shows the Piazza website interface for a course. The top navigation bar includes the Piazza logo, a course identifier 'EN 601.452', and links for 'Resources', 'Syllabus', and 'Manage Class'. The user 'Michael Schen' is logged in. On the left sidebar, there are tabs for 'Global', 'Unread', 'Unresolved', and 'Following'. Below these is a 'New Post' button and a search bar. A list of pinned posts is visible, including 'Search for Teammate!', 'Introduce Piazza to your stu...', 'Get familiar with Piazza', and 'Tips & Tricks for a successful...'. The main content area displays a 'Welcome to Piazza!' message, followed by a list of tips for getting started. The bottom of the page shows a table with columns for 'Average Response Time' and 'Special Mentions', both currently showing 'N/A'.

EN 601.452 + Q & A Resources Syllabus Manage Class

Michael Schen

Global Unread Unresolved Following

New Post Search or add a post...

SEARCH FOR TEAMMATE! 8/25/19

Introduce Piazza to your stu... 10/28PM

Get familiar with Piazza 10/28PM

Tips & Tricks for a successful... 10/28PM

Welcome to Piazza! 10/28PM

Piazza is a Q&A platform designed to get you great answers from classmates and instructors fast. We've put together this list of tips you might find handy as you get started!

1. Ask questions!

The best way to get answers is to ask questions! Ask questions on Piazza rather than emailing your teaching staff so everyone can benefit from the response (and so you can get answers from classmates who are up as late as you are).

2. Edit questions and answers wiki-style.

Think of Piazza as a Q&A wiki for your class. Every question has just a single **students'** answer that students can edit collectively (and a single **instructors'** answer for instructors).

3. Add a followup to comment or ask further questions.

To comment on or ask further questions about a post, start a **followup discussion**. Mark it resolved when the issue has been addressed, and add any relevant information back into the Q&A above.

4. Go anonymous.

Shy? No problem. You can always opt to post or edit anonymously.

5. Tag your posts.

It's far more convenient to find all posts about your Homework 3 or Midterm 1 when the posts are tagged. Type a "t" before a key word to tag. Click a blue tag in a post or the question feed to filter for all posts that share that tag.

6. Format code and equations.

Adding a code snippet? Click the **pre** or **tt** button in the question editor to add pre-formatted or inline text. Mathematical equation? Click the **fx** button to access the LaTeX editor to build a nicely formatted equation.

7. View and download class details and resources.

Click the **Course Page** button in your top bar to access the class syllabus, staff contact information, office hours details, and course resources—all in one place!

Contact the Piazza Team anytime with questions or comments at [team@piazza.com](mailto:team@piazza.com). We love feedback!

Average Response Time: N/A Special Mentions: There are no special mentions at this time.

Online Now: 1 This Week: 1

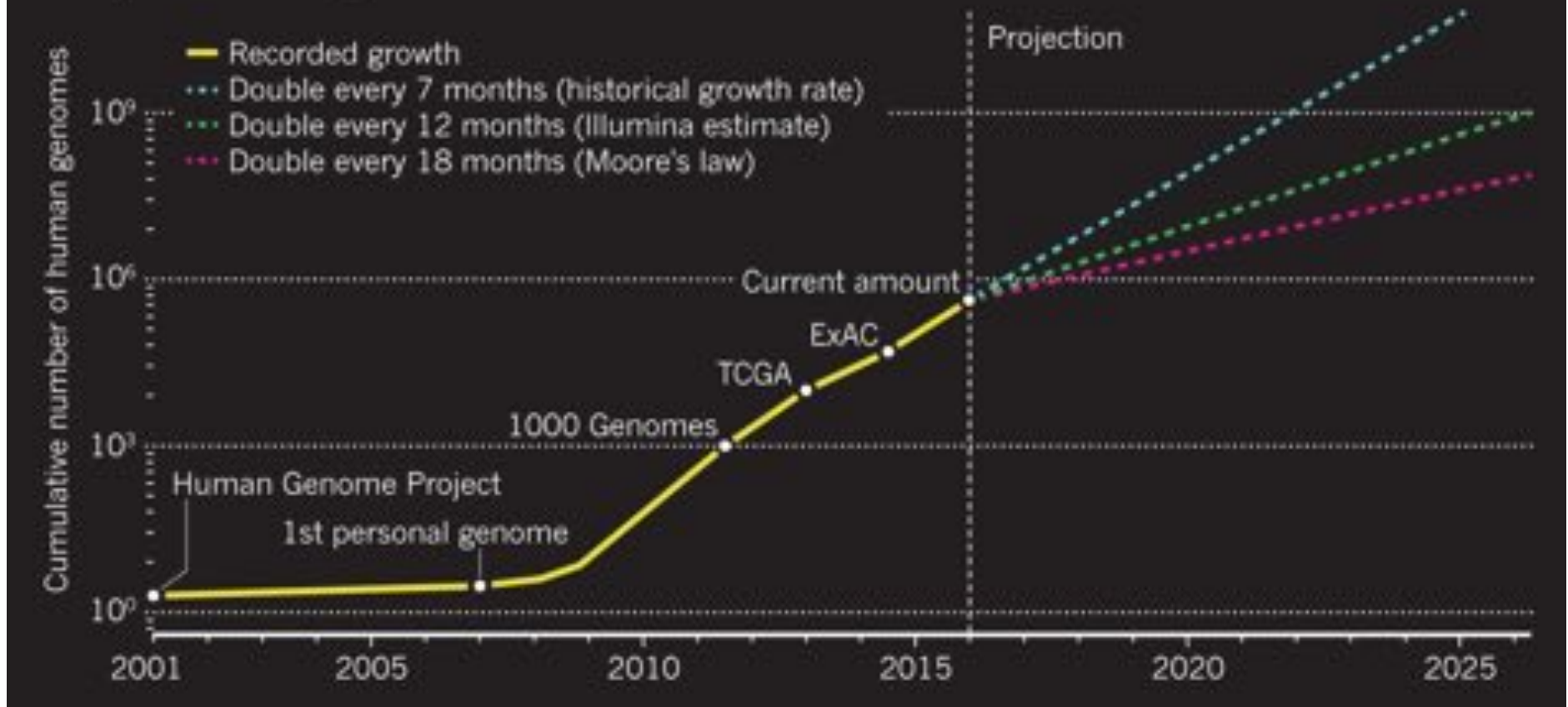
<http://piazza.com/jhu/fall2019/en601452>



# Sequencing Capacity

## DNA SEQUENCING SOARS

Human genomes are being sequenced at an ever-increasing rate. The 1000 Genomes Project has aggregated hundreds of genomes; The Cancer Genome Atlas (TCGA) has gathered several thousand; and the Exome Aggregation Consortium (ExAC) has sequenced more than 60,000 exomes. Dotted lines show three possible future growth curves.



### Big Data: Astronomical or Genomical?

Stephens, Z, et al. (2015) PLOS Biology DOI: [10.1371/journal.pbio.1002195](https://doi.org/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 FMRP
  - Related to neuron development and synaptic plasticity
  - Also strong overlap with chromatin remodelers

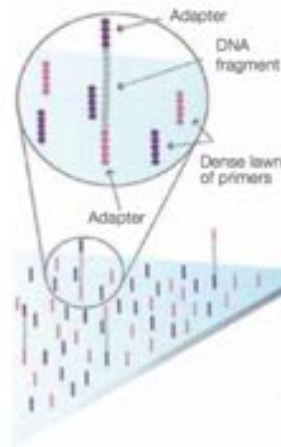


# Second Generation Sequencing

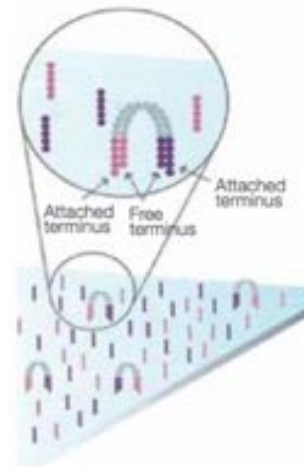


**Illumina NovaSeq 6000**  
*Sequencing by Synthesis*

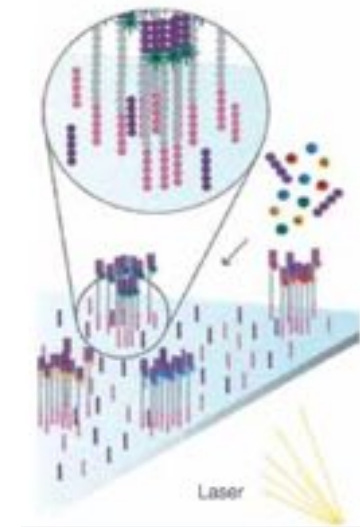
>3Tbp / day



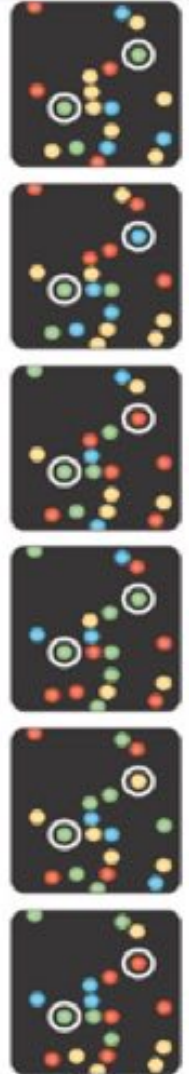
1. Attach



2. Amplify



3. Image

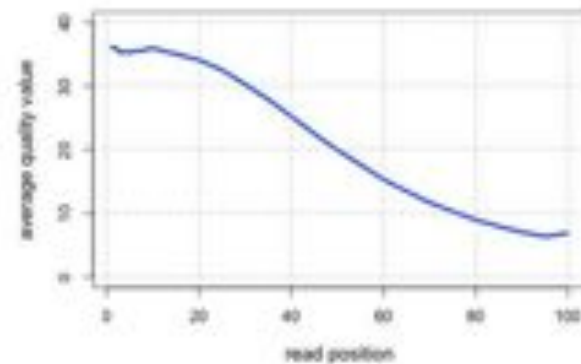


Metzker (2010) Nature Reviews Genetics 11:31-46  
<https://www.youtube.com/watch?v=fCd6B5HRaZ8>

# Illumina Quality

QV	P <sub>error</sub>
40	1/10000
30	1/1000
20	1/100
10	1/10

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



```

SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....
.....XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....
.....IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....
.....JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....
!"#$%&'()*+,-./0123456789;:<=>?@ABCDEFGHIJKLMN
OPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|          |          |          |          |
33         59        64         73         104        126

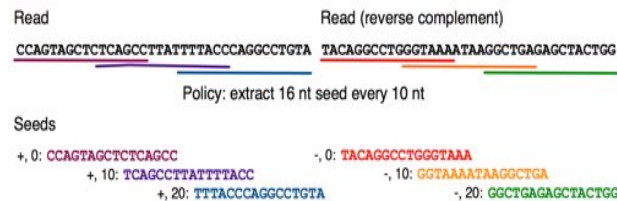
```

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)

# Assembly, Mapping & Genotyping

## Week 2/3/4

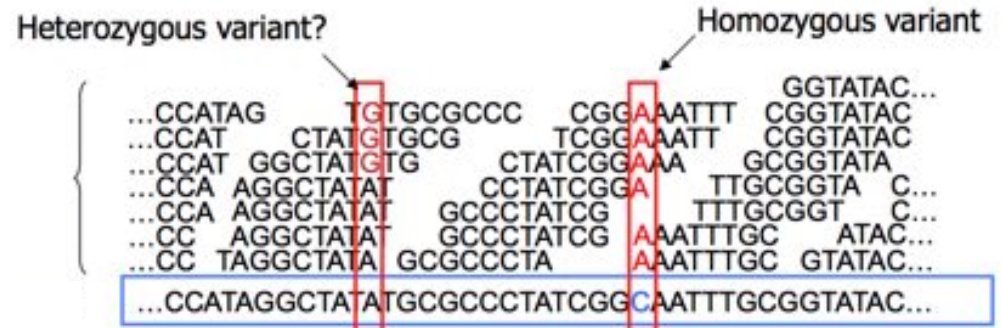
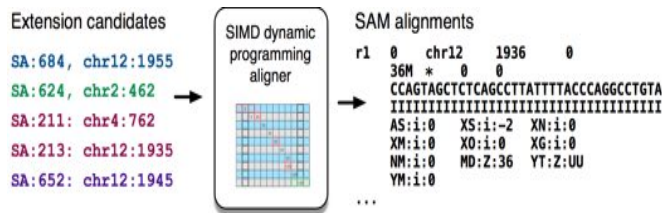
### 1. Split read into segments



### 2. Lookup each segment and prioritize



### 3. Evaluate end-to-end match



- 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

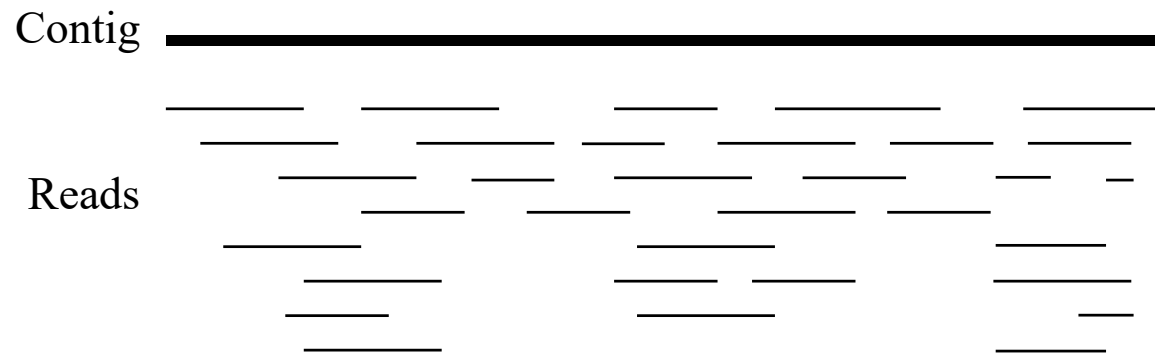
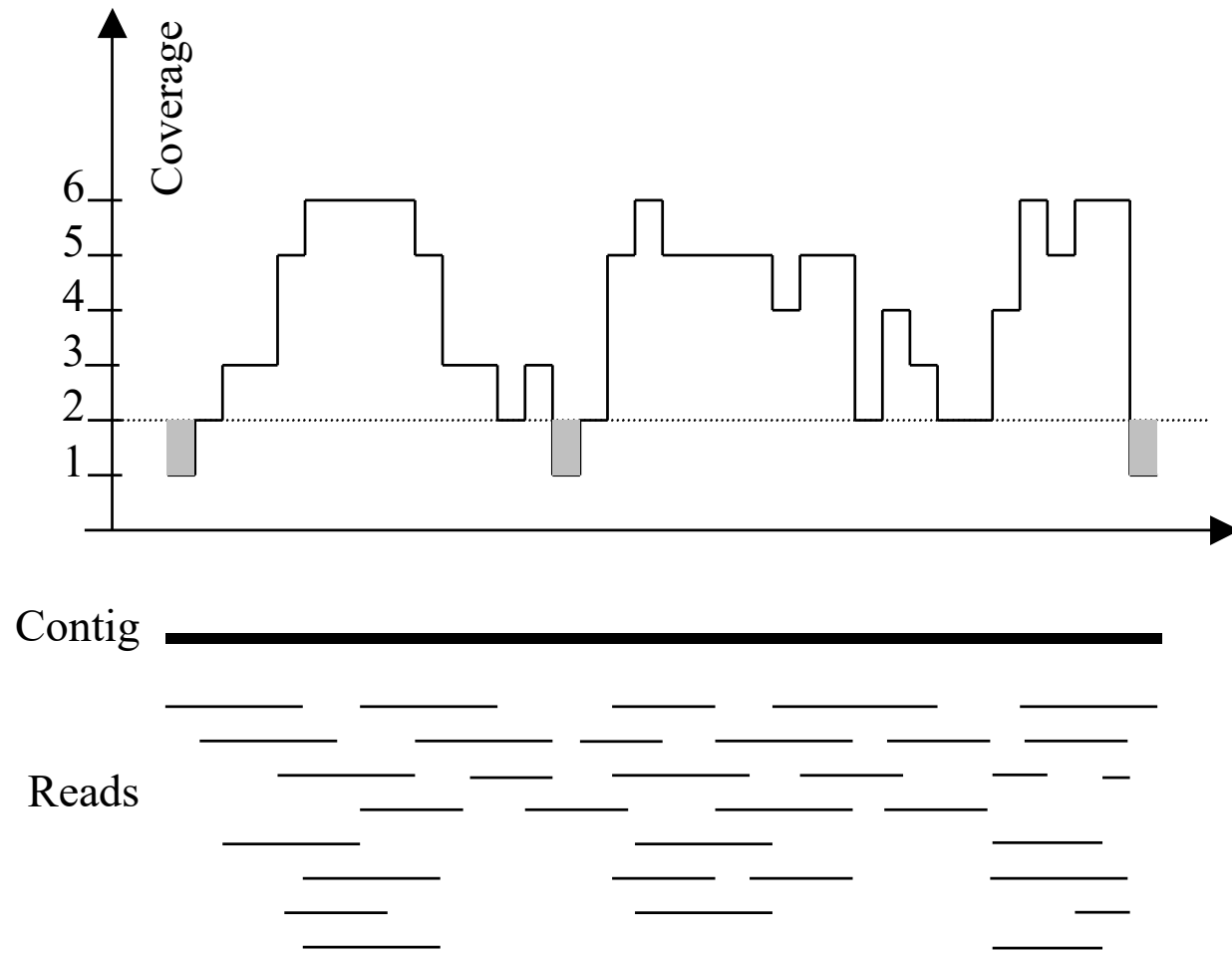
### Fast gapped-read alignment with Bowtie 2

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

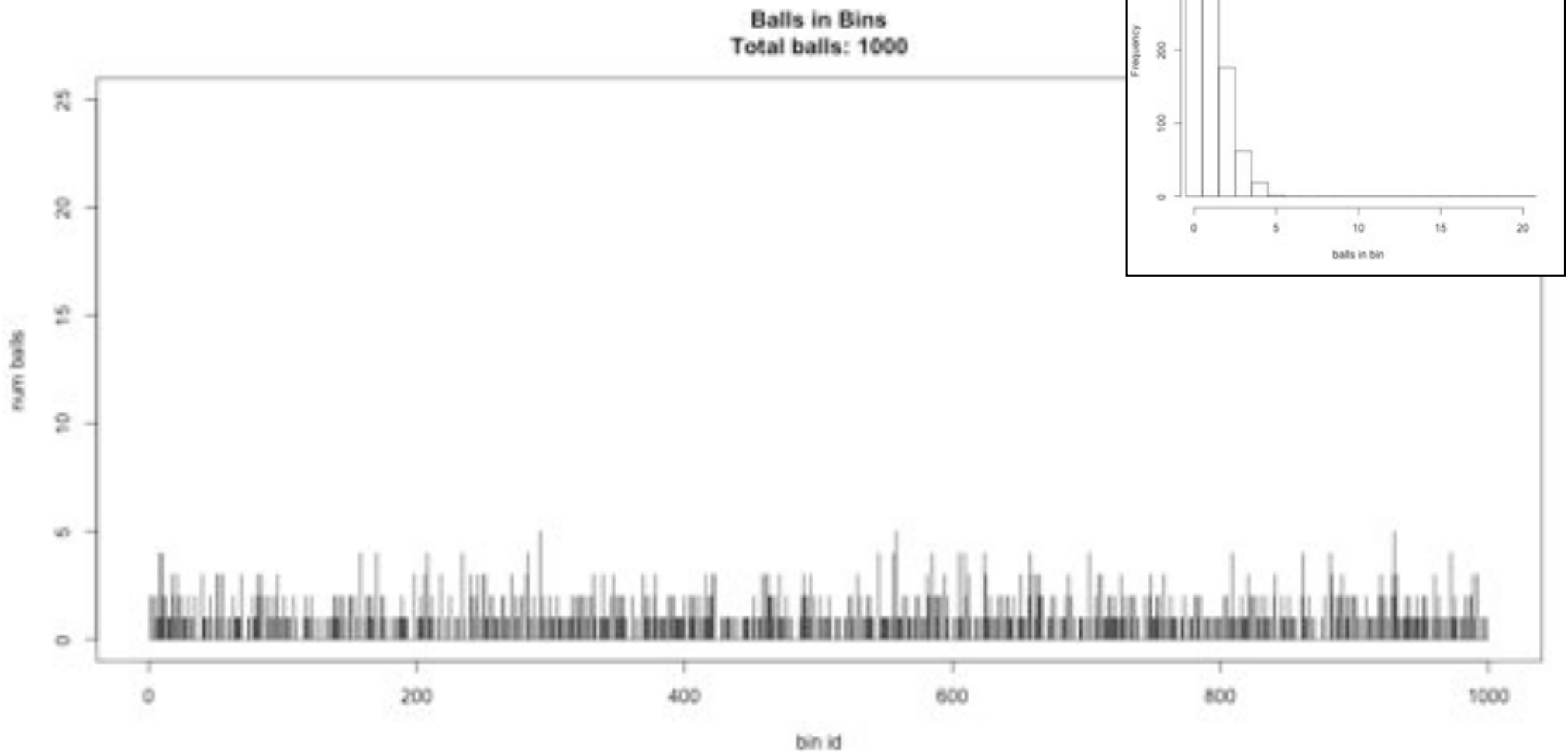


Imagine raindrops on a sidewalk

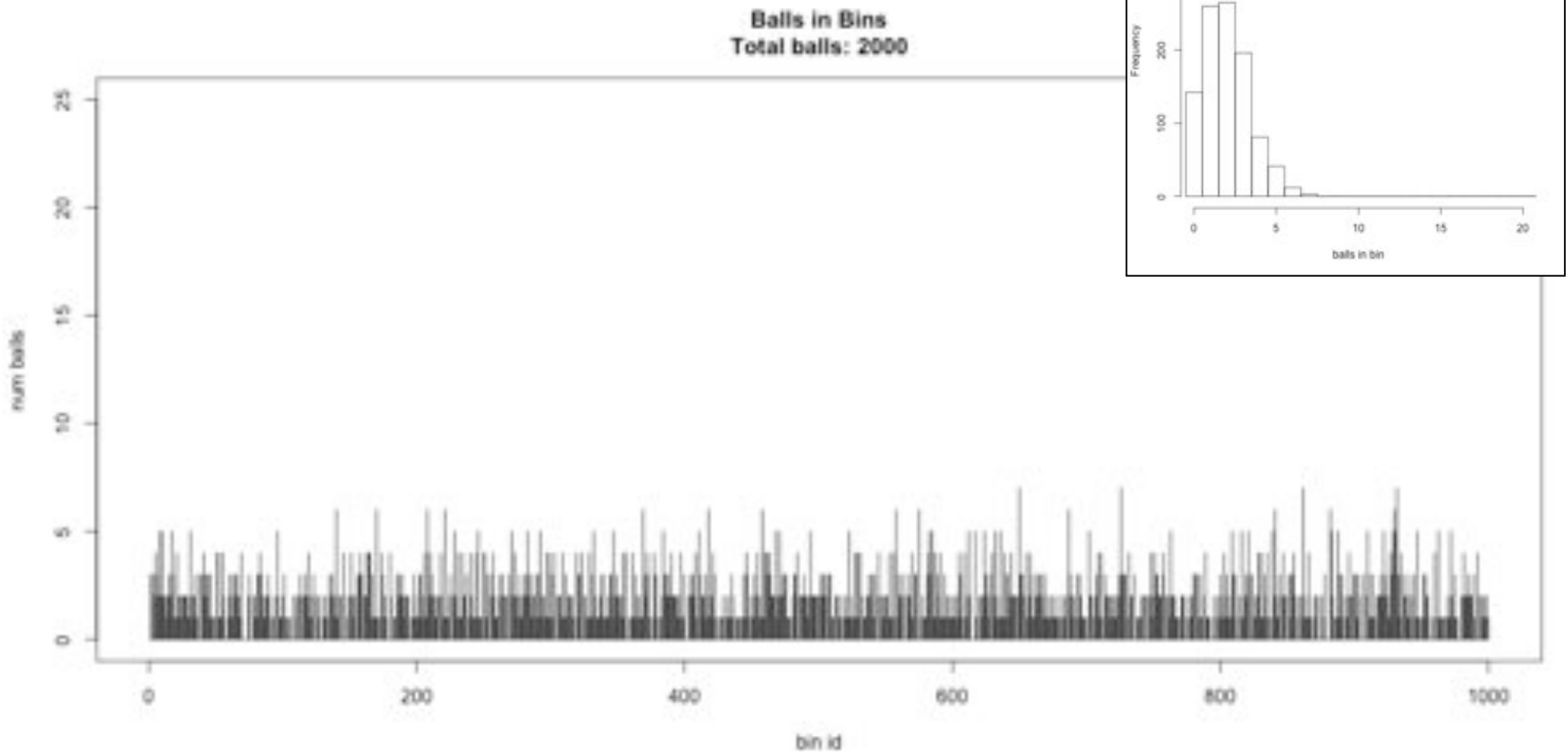
We want to cover the entire sidewalk but each drop costs \$1

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

# 1x sequencing

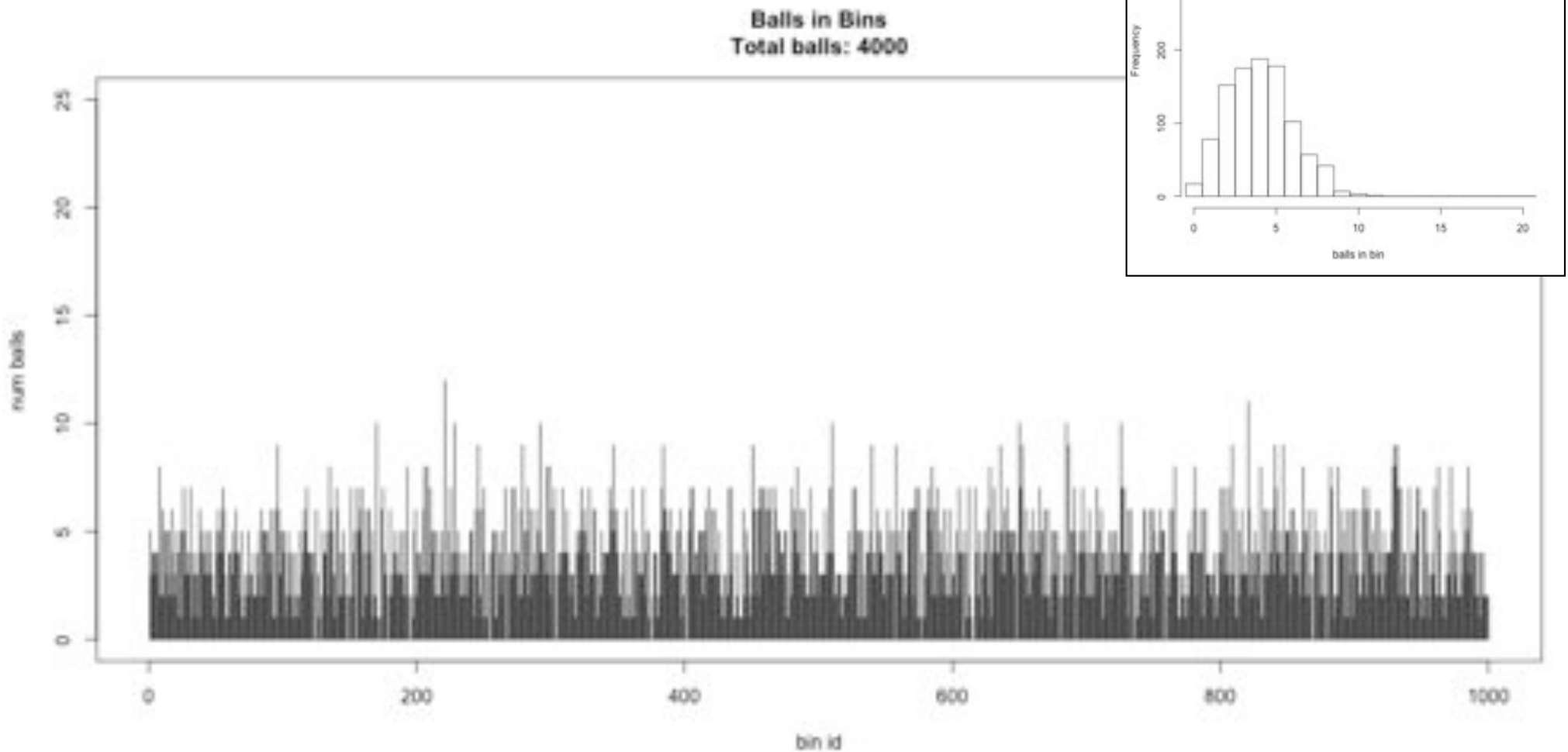


# 2x sequencing

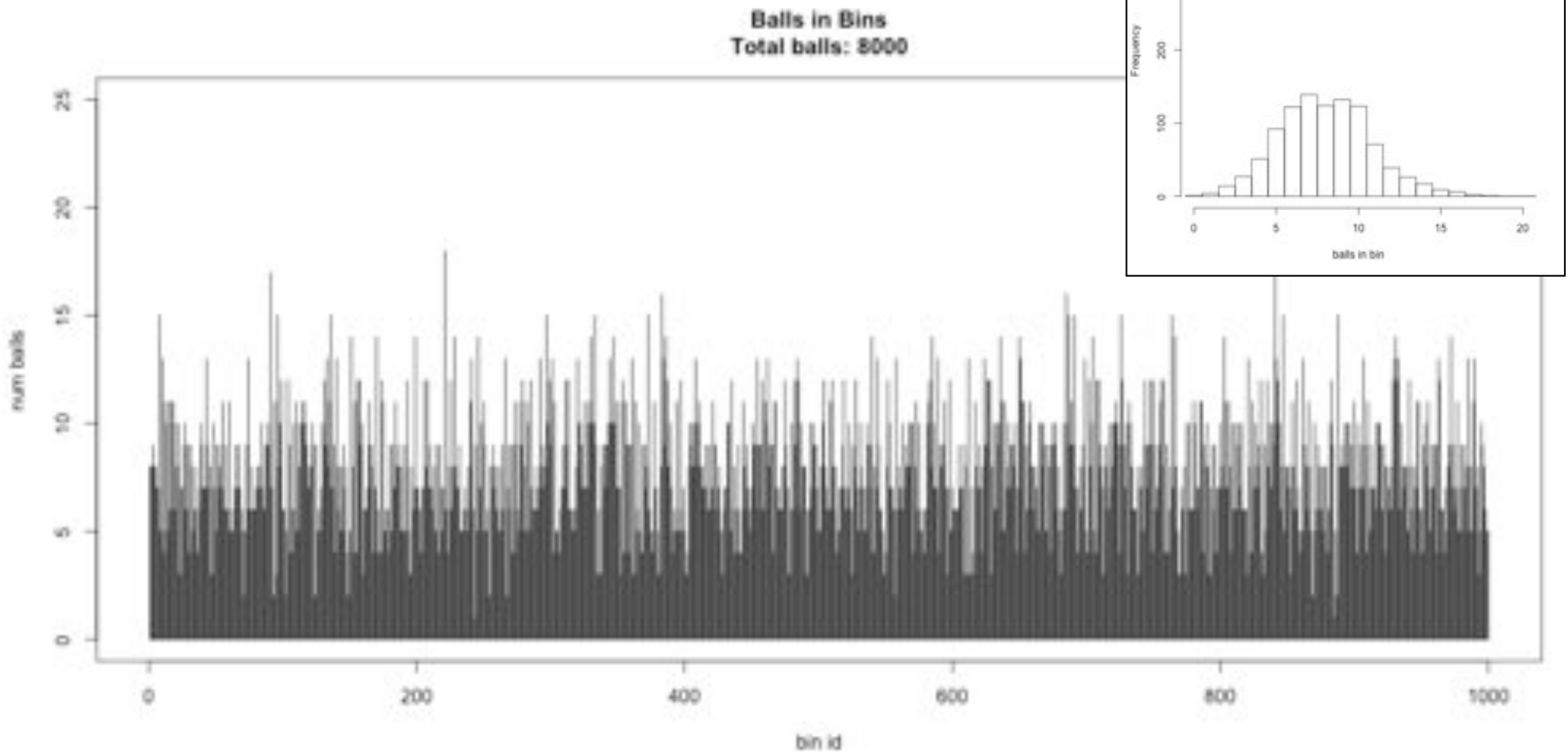




# 4x sequencing



# 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.

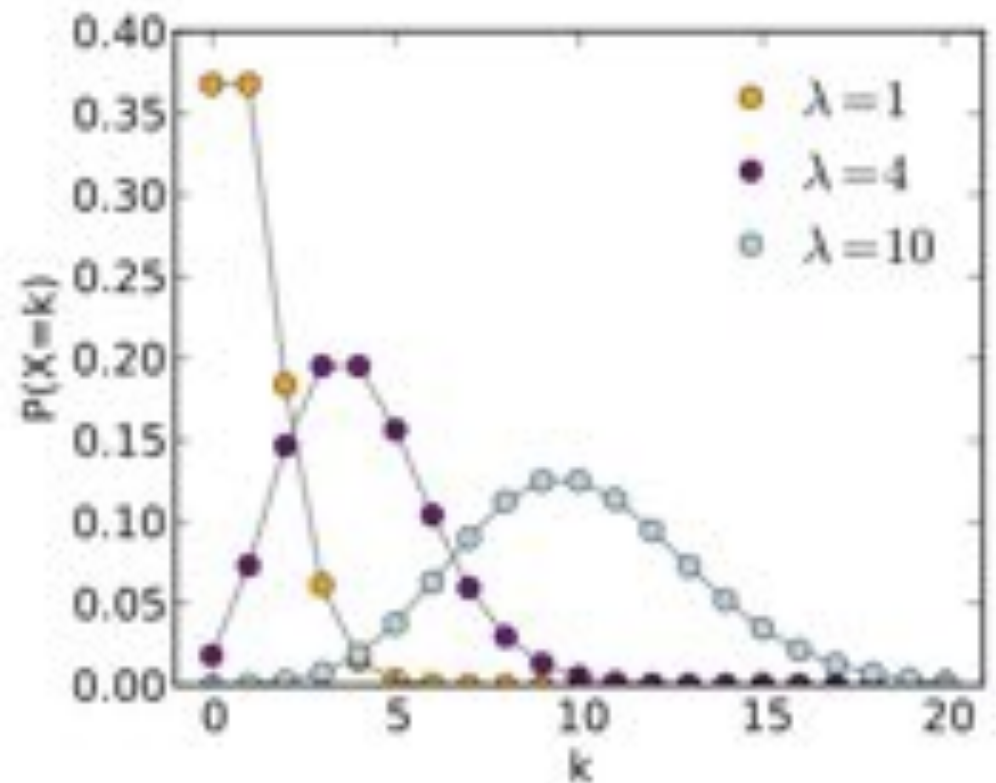
Formulation comes from the limit of the binomial equation

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

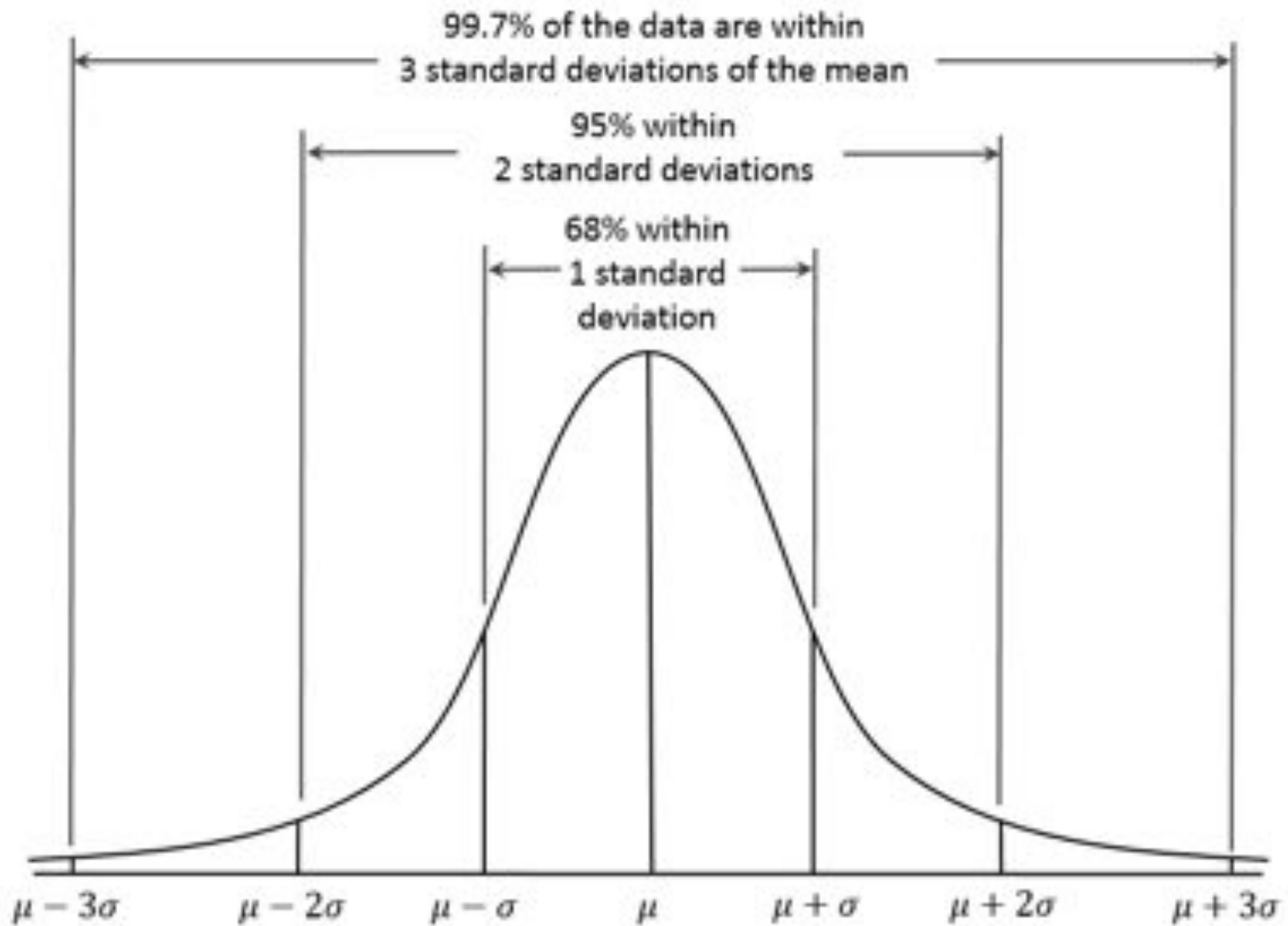
## **Key properties:**

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

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



# 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  $10\text{Mbp} \times 24x = 240\text{Mbp}$  of data  
 $240\text{Mbp} / 120\text{bp} / \text{read} = 2\text{M reads}$



# 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 I
  1. Set up Dropbox for yourself!
  2. Get comfortable on the command line