

Genomic Technologies

Michael Schatz

August 31, 2022

Lecture 2: Applied Comparative Genomics



Course Webpage

The screenshot shows a GitHub repository page for 'appliedgenomics2022'. The repository is public and contains a single commit from 'mschatz' titled 'update schedule'. The commit adds a syllabus, an initial commit, and updates the README. The README file describes the course as 'JHU EN.601.449/EN.601.649: Computational Genomics: Applied Comparative Genomics'. It lists the professor as Michael Schatz, TA as Bohan Ni, and class hours as Monday + Wednesday @ 1:30p - 2:45p Gilman 17. It also mentions office hours by appointment. The README states that the primary goal is for students to be grounded in theory and leave the course empowered to conduct independent genomic analyses. The course will focus on human genomics and medical applications. Topics include genome assembly & comparative genomics, variant identification & analysis, gene expression & regulation, personal genome analysis, and cancer genomics. The grading will be based on assignments, a midterm exam, class presentations, and a significant class project. There are no formal course prerequisites, although familiarity with UNIX scripting and/or programming is required.

About

Materials for EN.601.449/EN.601.649
Computational Genomics: Applied
Comparative Genomics

Readme
CC0-1.0 license
1 star
1 watching
0 forks

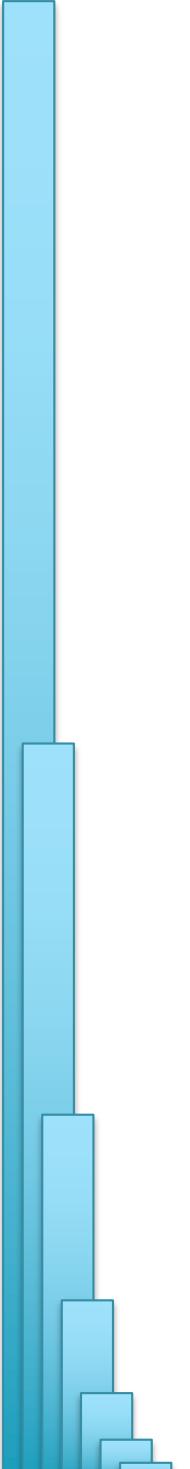
Releases

No releases published

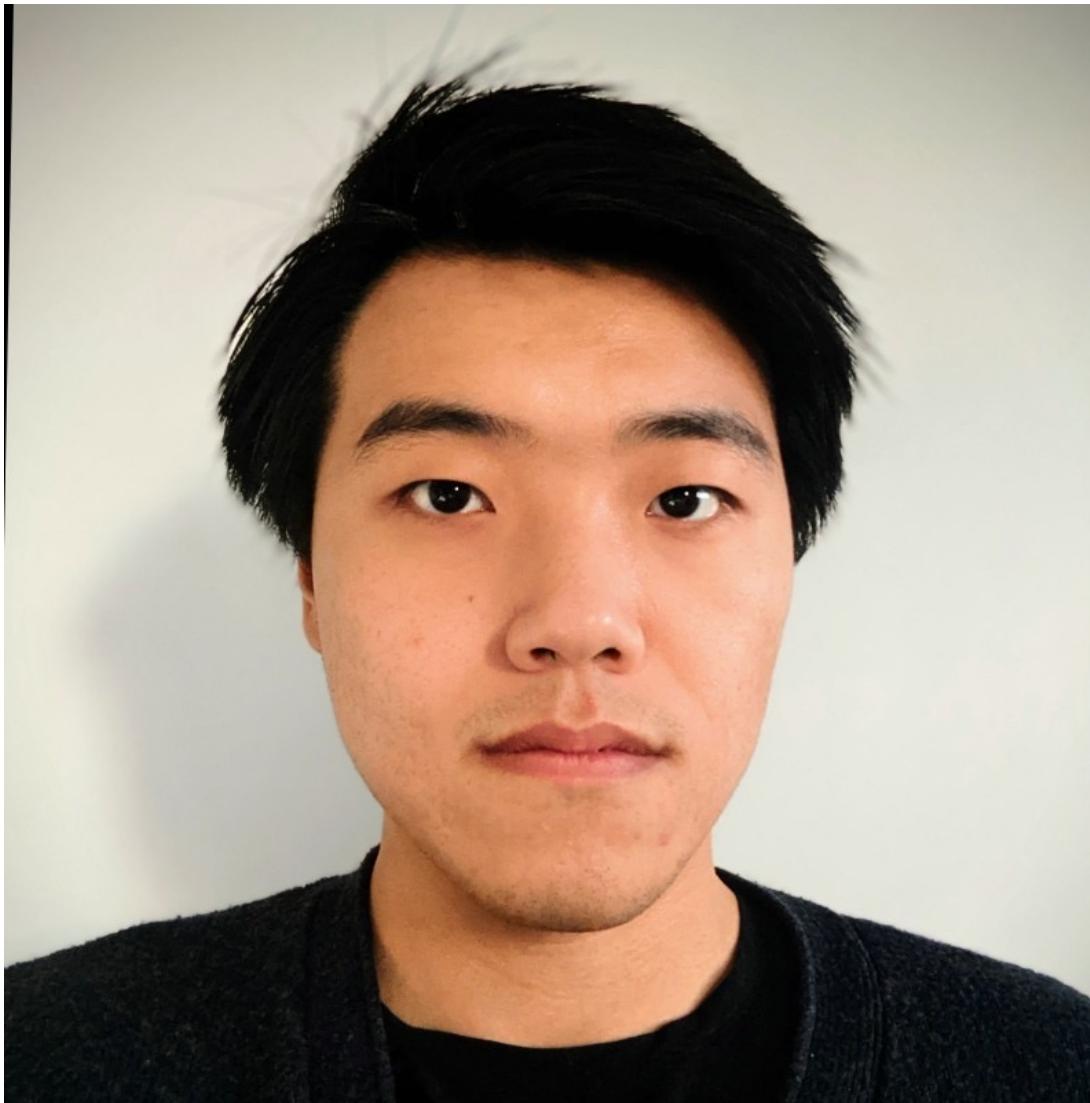
Packages

No packages published

<https://github.com/schatzlab/appliedgenomics2022>



TA: Bohan Ni



Check Piazza for Office Hours Poll

TA: Bohan Ni

The screenshot shows a Piazza poll titled "TA Office Hour Poll". The poll asks: "As the TA for the course, I will be holding office hours once a week on Zoom. In order to pick a time for them that works for most people, I have made a poll so you all can decide when would work the best. Please select the times below when you would be available to attend office hours. I'll close the poll after 48 hours and announce my office hours at that time." Below the question, it says: "And for assignment one which is due next Wednesday, I will host an office hour on Tuesday September 6th at 12pm-1pm. But please start early and post your questions on piazza to get questions answered at the earliest time." The poll has the following options:

- Thursday 10-11am
- Thursday 9-10am
- Friday 1-2pm
- Friday 2-3pm
- Thursday 11am-12pm
- Friday 11am-12pm

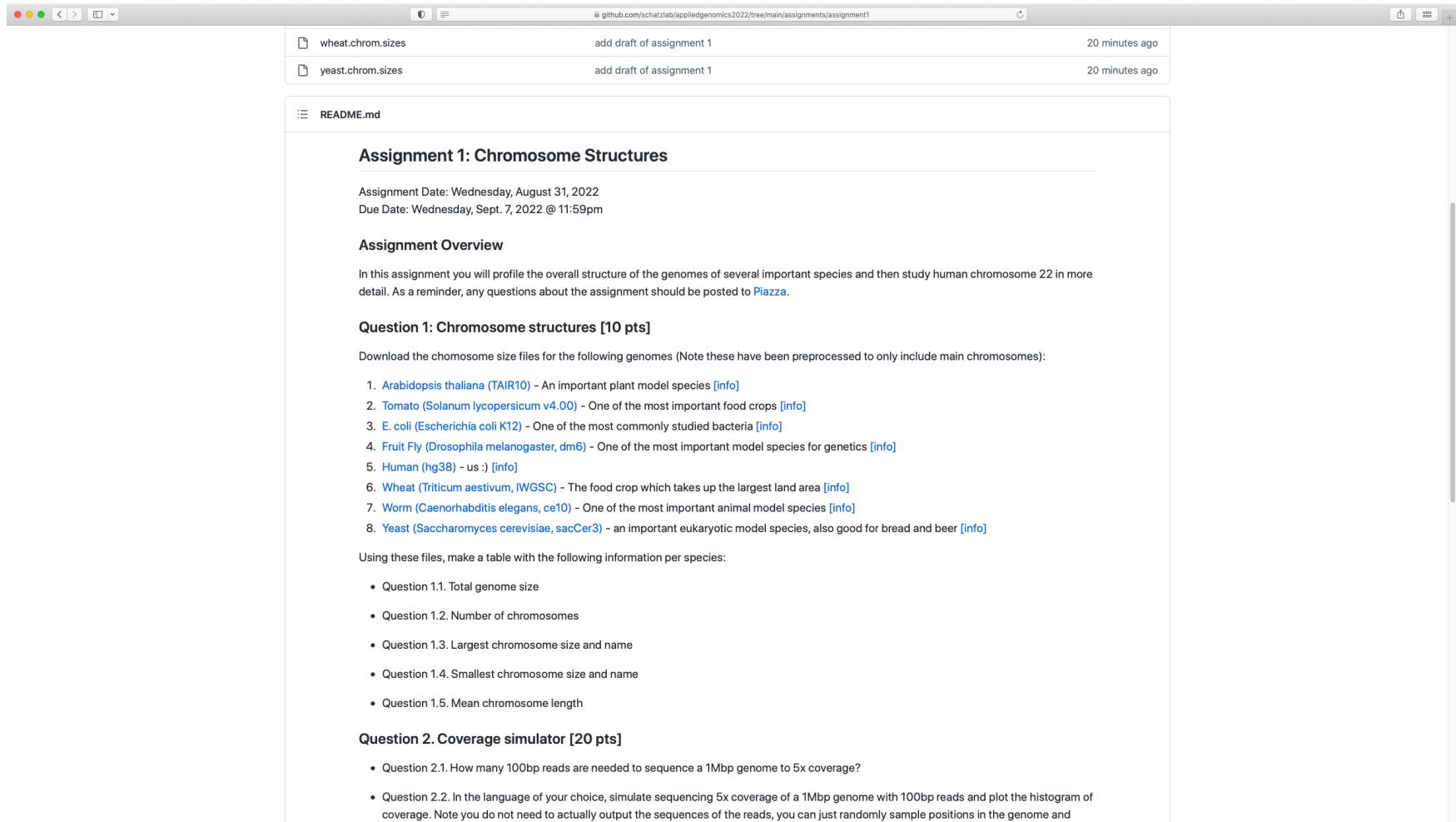
At the bottom of the poll page, there is a chart titled "TA Office Hour Poll closes in 1 day(s)" showing the distribution of votes:

Time Range	Votes	Percentage
Thursday 10-11am	1	100%
Thursday 9-10am	0	0%
Friday 1-2pm	1	100%
Friday 2-3pm	0	0%
Thursday 11am-12pm	0	0%
Friday 11am-12pm	0	0%

Below the chart, there is a section for "followup discussions, for lingering questions and comments". It includes fields for "Start a new followup discussion", "Compose a new followup discussion", "Average Response Time: N/A", "Special Mentions:", and "There are no special mentions at this time." At the bottom right, it says "Online Now | This Week: 5 | 5".

Check Piazza for Office Hours Poll

Assignment 1



The screenshot shows a GitHub repository page for 'assignment1'. At the top, there are two files: 'wheat.chrom.sizes' and 'yeast.chrom.sizes', both added 20 minutes ago. Below them is the 'README.md' file.

Assignment 1: Chromosome Structures

Assignment Date: Wednesday, August 31, 2022
Due Date: Wednesday, Sept. 7, 2022 @ 11:59pm

Assignment Overview

In this assignment you will profile the overall structure of the genomes of several important species and then study human chromosome 22 in more detail. 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. [Tomato \(Solanum lycopersicum v4.00\)](#) - One of the most important food crops [\[info\]](#)
3. [E. coli \(Escherichia coli K12\)](#) - One of the most commonly studied bacteria [\[info\]](#)
4. [Fruit Fly \(Drosophila melanogaster, dm6\)](#) - One of the most important model species for genetics [\[info\]](#)
5. [Human \(hg38\) - us :\)](#) [\[info\]](#)
6. [Wheat \(Triticum aestivum, IWGSC\)](#) - The food crop which takes up the largest land area [\[info\]](#)
7. [Worm \(Caenorhabditis elegans, ce10\)](#) - One of the most important animal model species [\[info\]](#)
8. [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

Question 2. Coverage simulator [20 pts]

- Question 2.1. How many 100bp reads are needed to sequence a 1Mbp genome to 5x coverage?
- Question 2.2. In the language of your choice, simulate sequencing 5x coverage of a 1Mbp genome with 100bp reads and plot the histogram of coverage. Note you do not need to actually output the sequences of the reads, you can just randomly sample positions in the genome and

<https://github.com/schatzlab/appliedgenomics2022/tree/main/assignments/assignment1>
Due end of day on Sept 7 (right before midnight)

Plotting in Python



The screenshot shows the official website for Matplotlib at matplotlib.org. The page features a large "matplotlib" logo with a circular icon containing a multi-colored sunburst or radar chart. Below the logo, it says "Version 3.4.3". A navigation bar at the top includes links for Installation, Documentation, Examples, Tutorials, and Contributing. On the right side of the header is a search bar and a "Fork me on GitHub" button. The main content area has a dark blue header with "Matplotlib: Visualization with Python". Below this, a sub-header states: "Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python." To the right, there's a sidebar with sections for the latest stable release (3.4.3), the last release for Python 2 (2.2.5), and the development version. It also includes a "Matplotlib cheatsheets" section with a thumbnail image of a full-page reference sheet. At the bottom right of the main content area is a "Support Matplotlib" button.

Matplotlib: Visualization with Python

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.



Matplotlib makes easy things easy and hard things possible.

Create

- Develop publication quality plots with just a few lines of code
- Use interactive figures that can zoom, pan, update...

Customize

- Take full control of line styles, font properties, axes properties...
- Export and embed to a number of file formats and interactive environments

Extend

- Explore tailored functionality provided by third party packages
- Learn more about Matplotlib through the many external learning resources

Documentation

To get started, read the [User's Guide](#).

Trying to learn how to do a particular kind of plot? Check out the [examples gallery](#) or the [list of plotting commands](#).

Latest stable release
3.4.3: [docs](#) | [changelog](#)

Last release for Python 2
2.2.5: [docs](#) | [changelog](#)

Development version
[docs](#)

Matplotlib cheatsheets


Support Matplotlib

<https://matplotlib.org/>

Plotting in R / ggplot2

Course NIH_re Google AS.02 Home biome grade Creative New Matplotlib Data cheat

github.com/rstudio/cheatsheets/blob/master/data-visualization-2.1.pdf

Data visualization with ggplot2 :: CHEAT SHEET



Basics

ggplot2 is based on the **grammar of graphics**, the idea that you can build every graph from the same components: a **data set**, a **coordinate system**, and **geoms**—visual marks that represent data points.

To display values, map variables in the data to visual properties of the geom (**aesthetics**) like **size**, **color**, and **x** and **y** locations.

Complete the template below to build a graph.

```
ggplot (data = <DATA>) +  
  <GEOM_FUNCTION> (mapping = aes(<MAPPINGS>),  
  stat = <STAT>, position = <POSITION>) +  
  <COORDINATE_FUNCTION> +  
  <FACET_FUNCTION> +  
  <SCALE_FUNCTION> +  
  <THEME_FUNCTION>
```

required

Not required, sensible defaults supplied

ggplot(data = mpg, aes(x = cyl, y = hwy)) Begins a plot that you finish by adding layers to. Add one geom function per layer.

last_plot() Returns the last plot.

ggsave("plot.png", width = 5, height = 5) Saves last plot as 5' x 5' file named "plot.png" in working directory. Matches file type to file extension.

Geoms

Use a geom function to represent data points, use the geom's aesthetic properties to represent variables. Each function returns a layer.

GRAPHICAL PRIMITIVES

```
a <- ggplot(economics, aes(date, unemploy))  
b <- ggplot(seals, aes(x = long, y = lat))
```

a + geom_blank() and **a + expand_limits()**
Ensure limits include values across all plots.

b + geom_curve(aes(yend = lat + 1, xend = long + 1), curvature = 1 - x, yend, y, alpha, angle, color, curvature, linetype, size)

a + geom_path(lineend = "butt", linejoin = "round", linemitre = 1)
x, y, alpha, color, group, linetype, size

a + geom_polygon(aes(alpha = 50)) - x, y, alpha, color, fill, group, subgroup, linetype, size

b + geom_rect(aes(xmin = long, ymin = lat, xmax = long + 1, ymax = lat + 1))- xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size

a + geom_ribbon(aes(ymin = unemploy - 900, ymax = unemploy + 900))- y, xmax, ymin, alpha, color, fill, group, linetype, size

TWO VARIABLES both continuous

```
e <- ggplot(mpg, aes(cty, hwy))
```

e + geom_label(aes(label = cty, nudge_x = 1, nudge_y = 1))- x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

e + geom_point()
x, y, alpha, color, fill, shape, size, stroke

e + geom_quantile()
x, y, alpha, color, group, linetype, size, weight

e + geom_rug(sides = "bl")
x, y, alpha, color, linetype, size

e + geom_smooth(method = lm)
x, y, alpha, color, fill, group, linetype, size, weight

e + geom_text(aes(label = cty, nudge_x = 1, nudge_y = 1))- x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

continuous bivariate distribution

```
h <- ggplot(diamonds, aes(carat, price))
```

h + geom_bin2d(binwidth = c(0.25, 500))
x, y, alpha, color, fill, linetype, size, weight

h + geom_density_2d()
x, y, alpha, color, group, linetype, size

h + geom_hex()
x, y, alpha, color, fill, size

continuous function

```
i <- ggplot(economics, aes(date, unemploy))
```

i + geom_area()
x, y, alpha, color, fill, linetype, size

i + geom_line()
x, y, alpha, color, group, linetype, size

i + geom_step(direction = "hv")
x, y, alpha, color, group, linetype, size

visualizing error

```
df <- data.frame(grp = c("A", "B"), fit = 4.5, se = 1.2)  
j <- ggplot(df, aes(grp, fit, ymin = fit - se, ymax = fit + se))
```

j + geom_crossbar(fatten = 2)- x, y, ymax, ymin, alpha, color, fill, group, linetype, size

j + geom_errorbar()- x, y, ymax, ymin, alpha, color, group, linetype, size, width
Also **geom_errorbarh()**.

j + geom_linerange()
x, ymin, ymax, alpha, color, group, linetype, size

j + geom_pointrange()- x, y, ymin, ymax, alpha, color, fill, group, linetype, shape, size

maps

```
data <- data.frame(murder = USArrests$Murder, state = tolower(rownames(USArrests)))
```

<https://ggplot2.tidyverse.org/>

Unsolved Questions in Biology

- What is your genome sequence?

The instruments provide the data, but none of the answers to any of these questions.

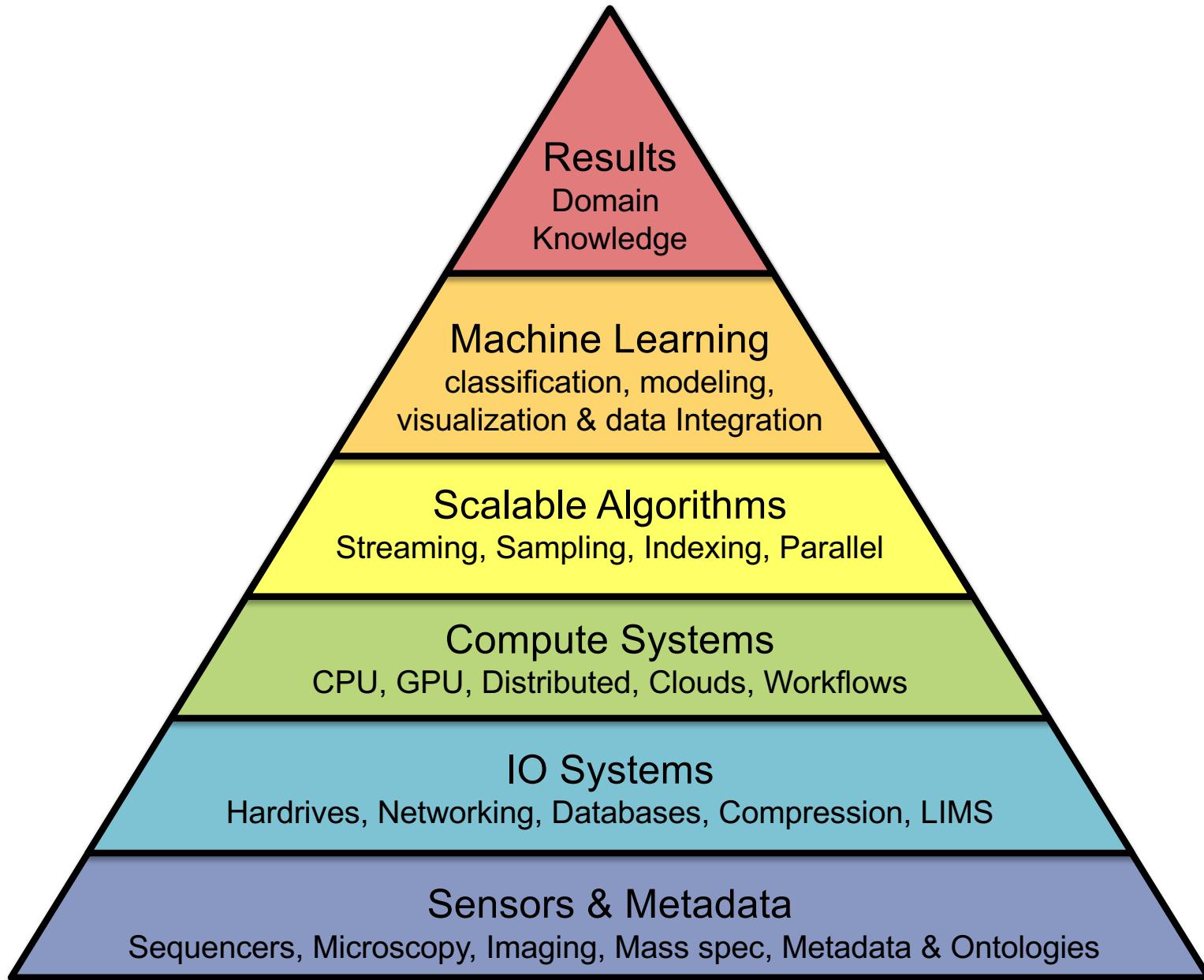
What software and systems will?

And who will create them?

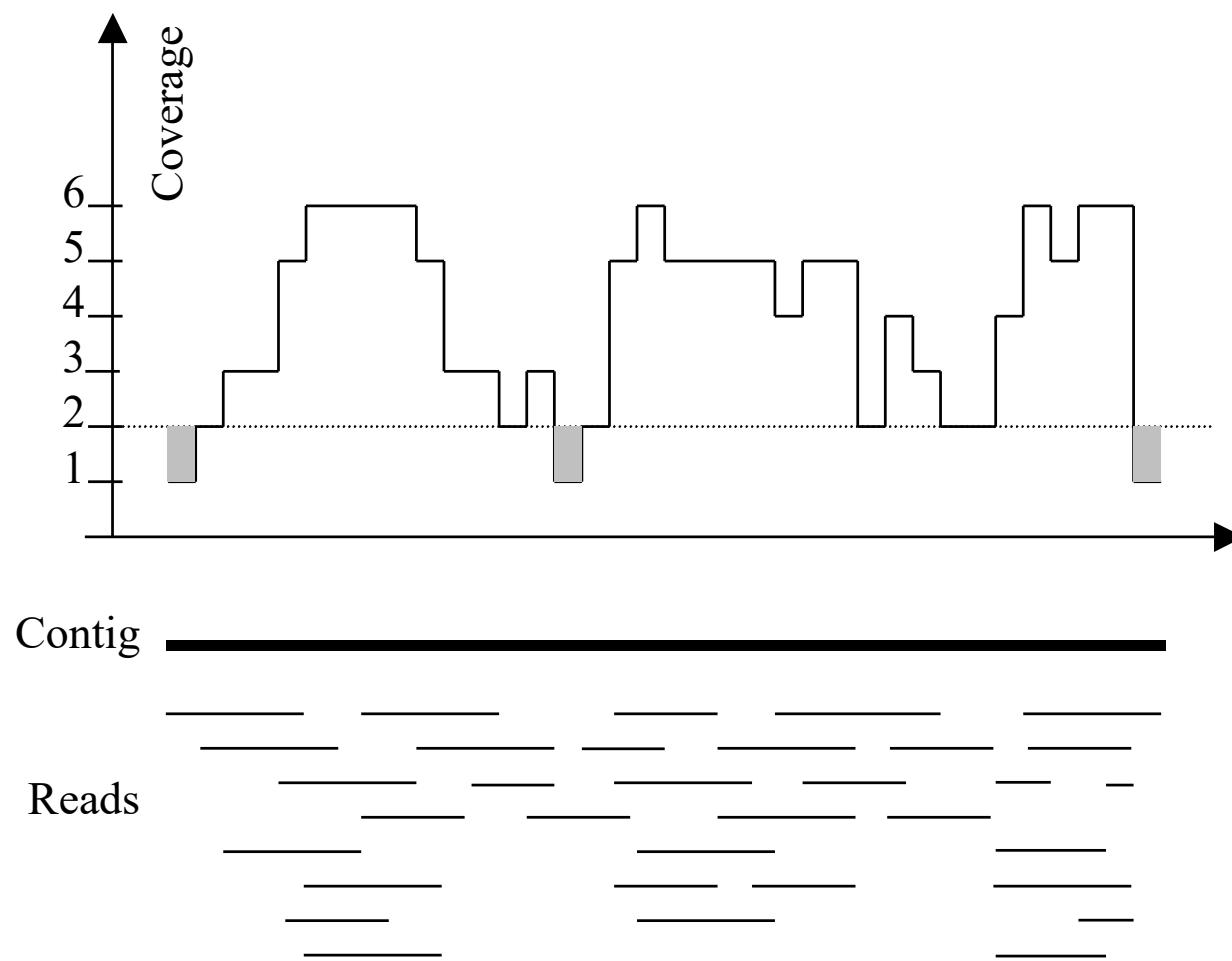
- **Plus thousands and thousands more**



Comparative Genomics Technologies



Typical sequencing coverage

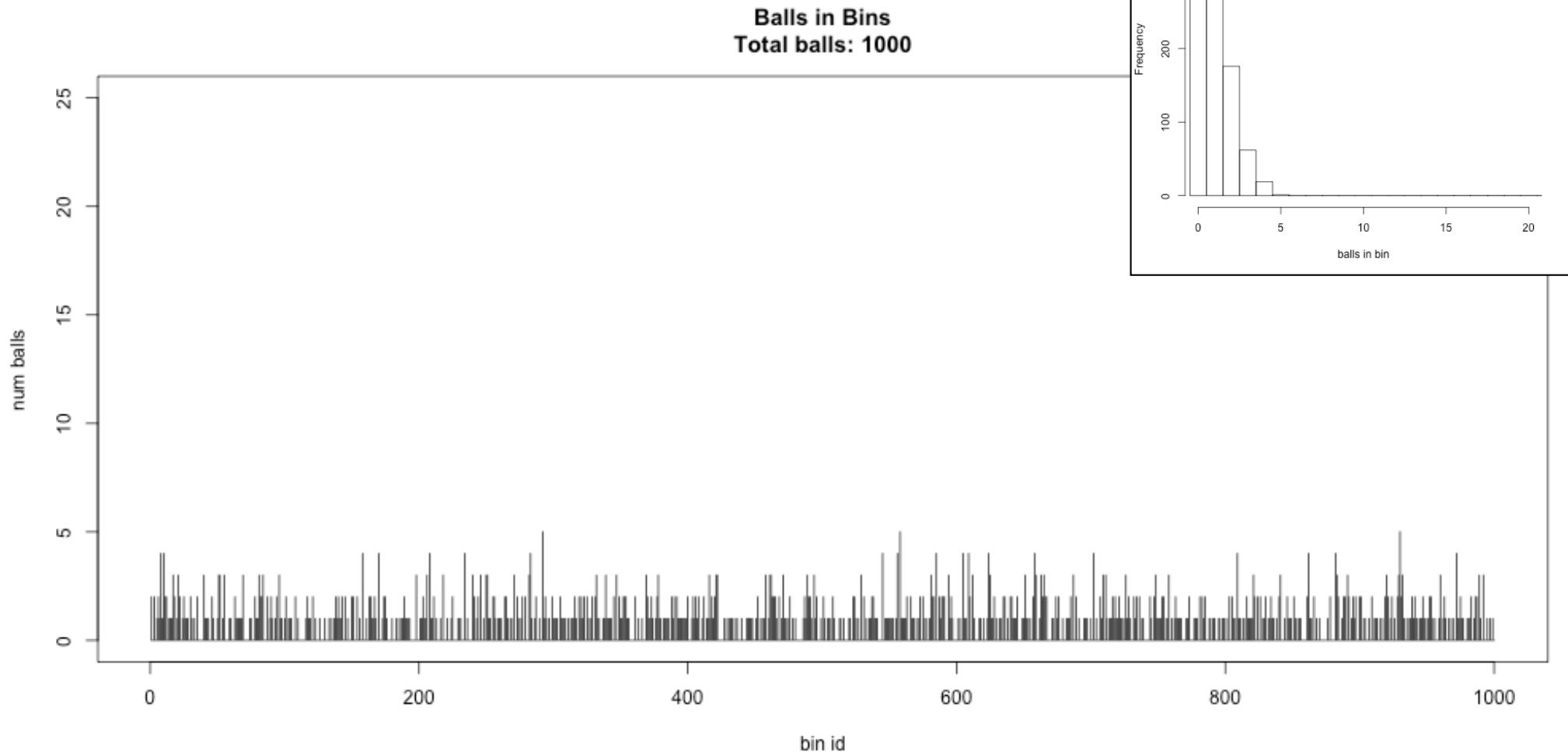


Imagine raindrops on a sidewalk

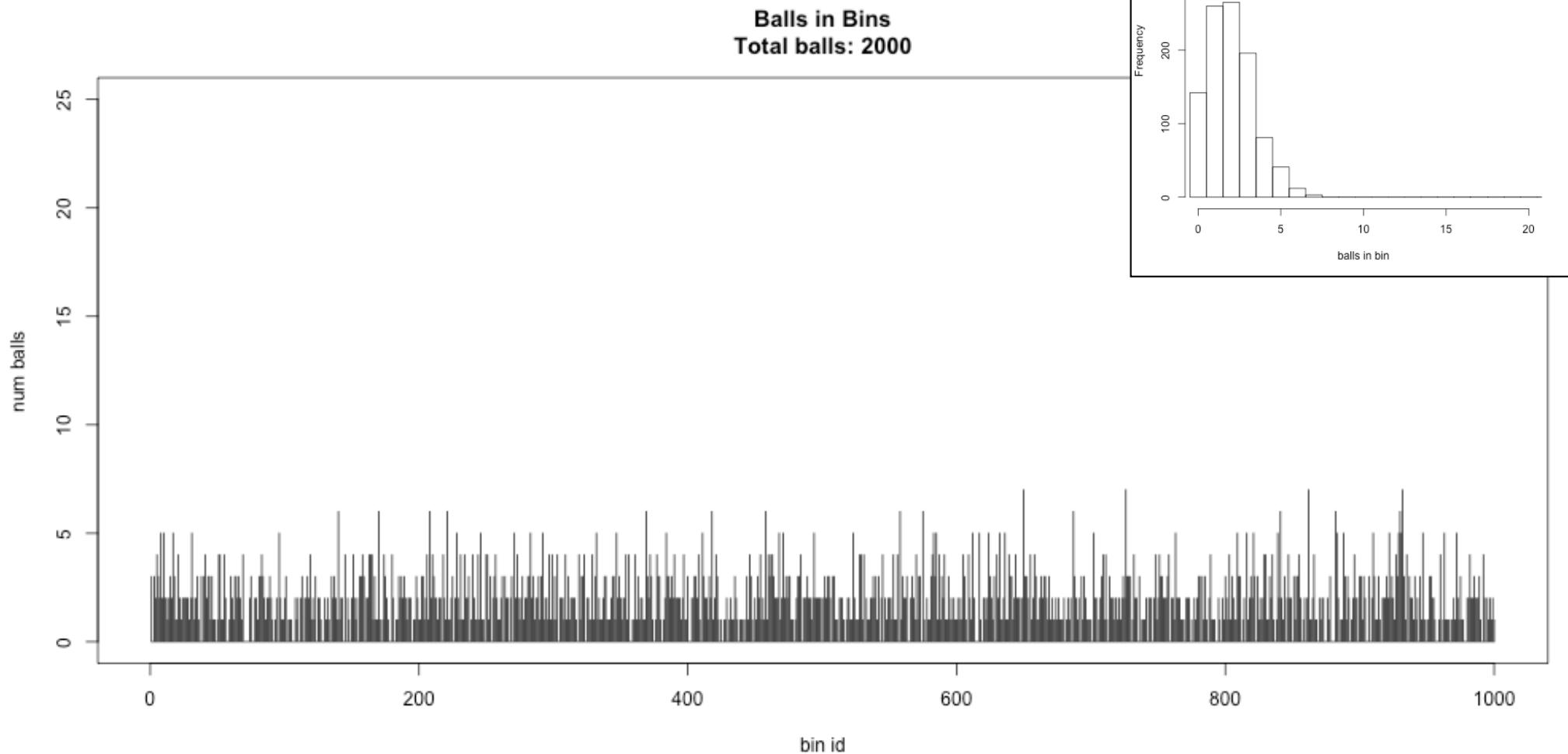
We want to cover the entire sidewalk but each drop costs \$1

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

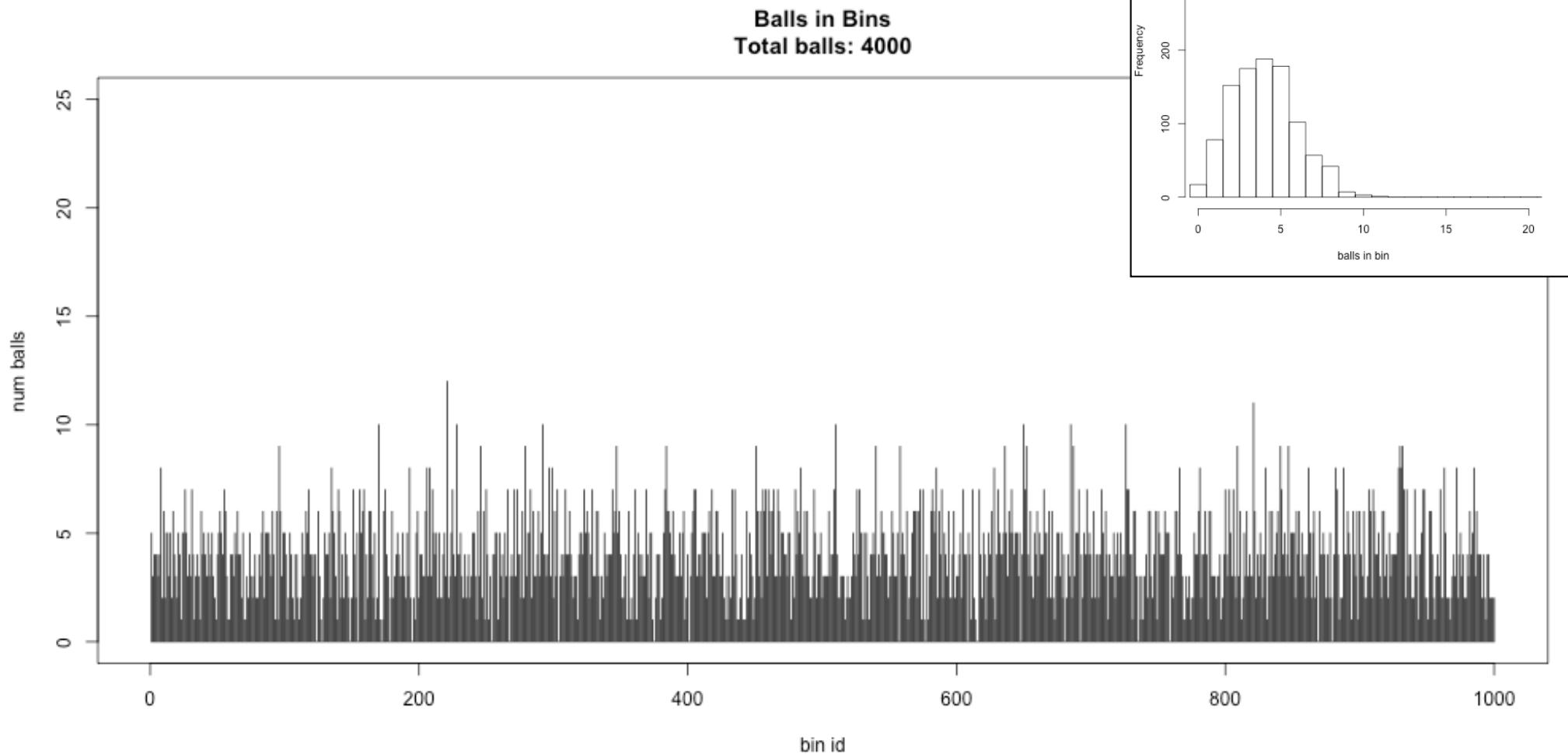
Ix sequencing



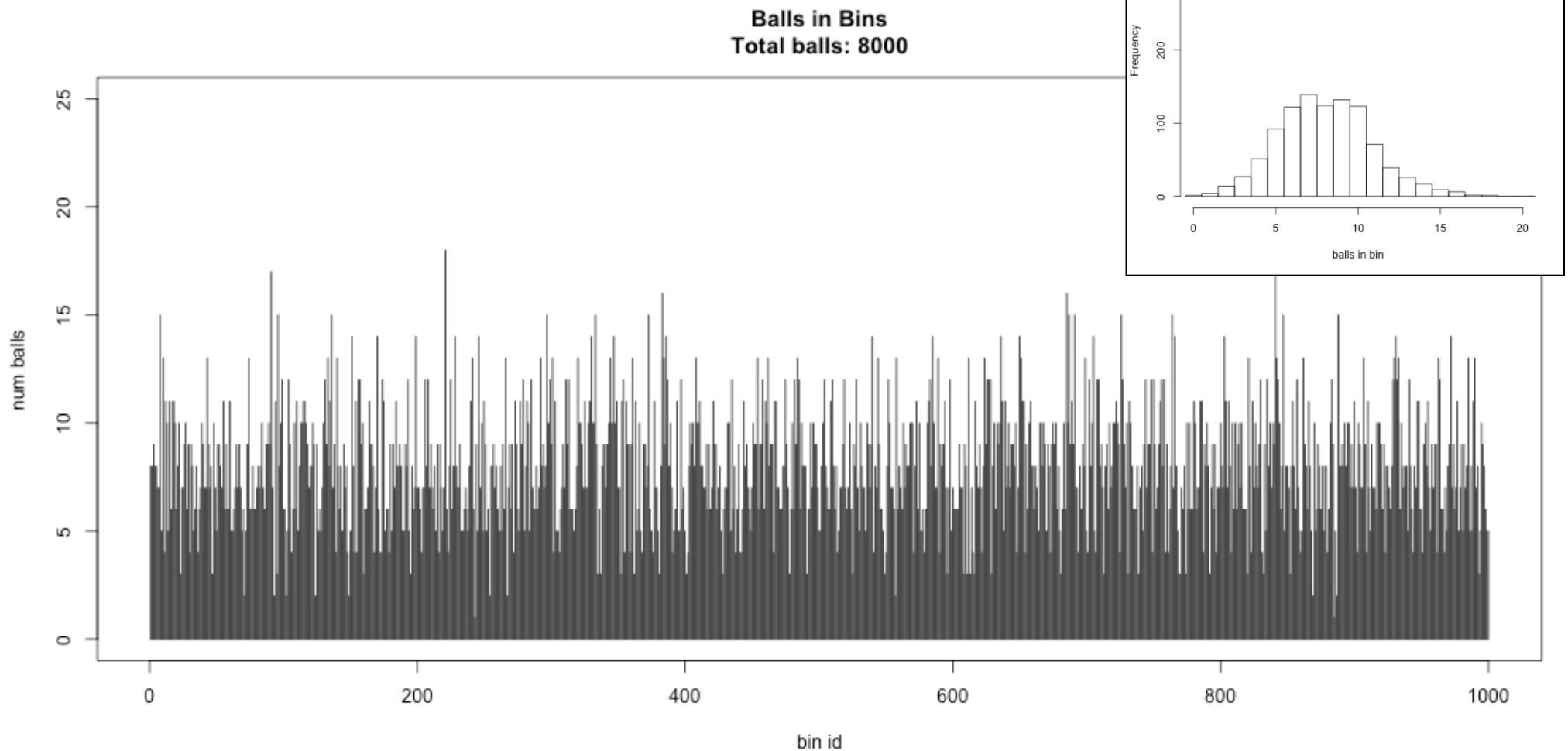
2x sequencing



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.

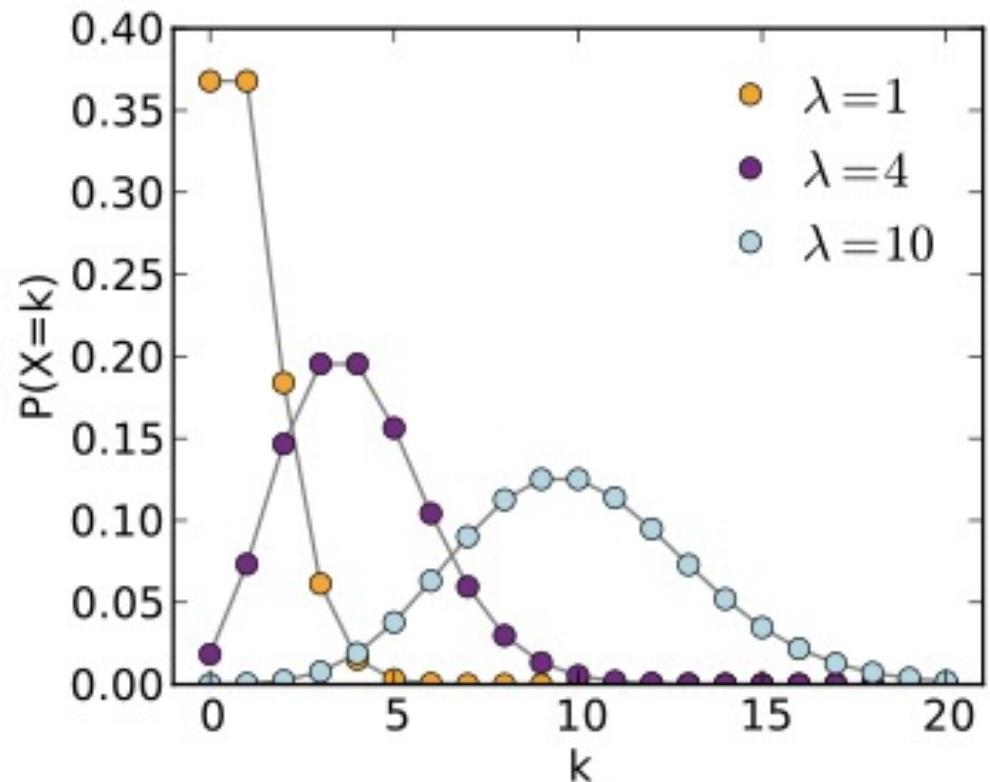
Formulation comes from the limit of the binomial equation

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

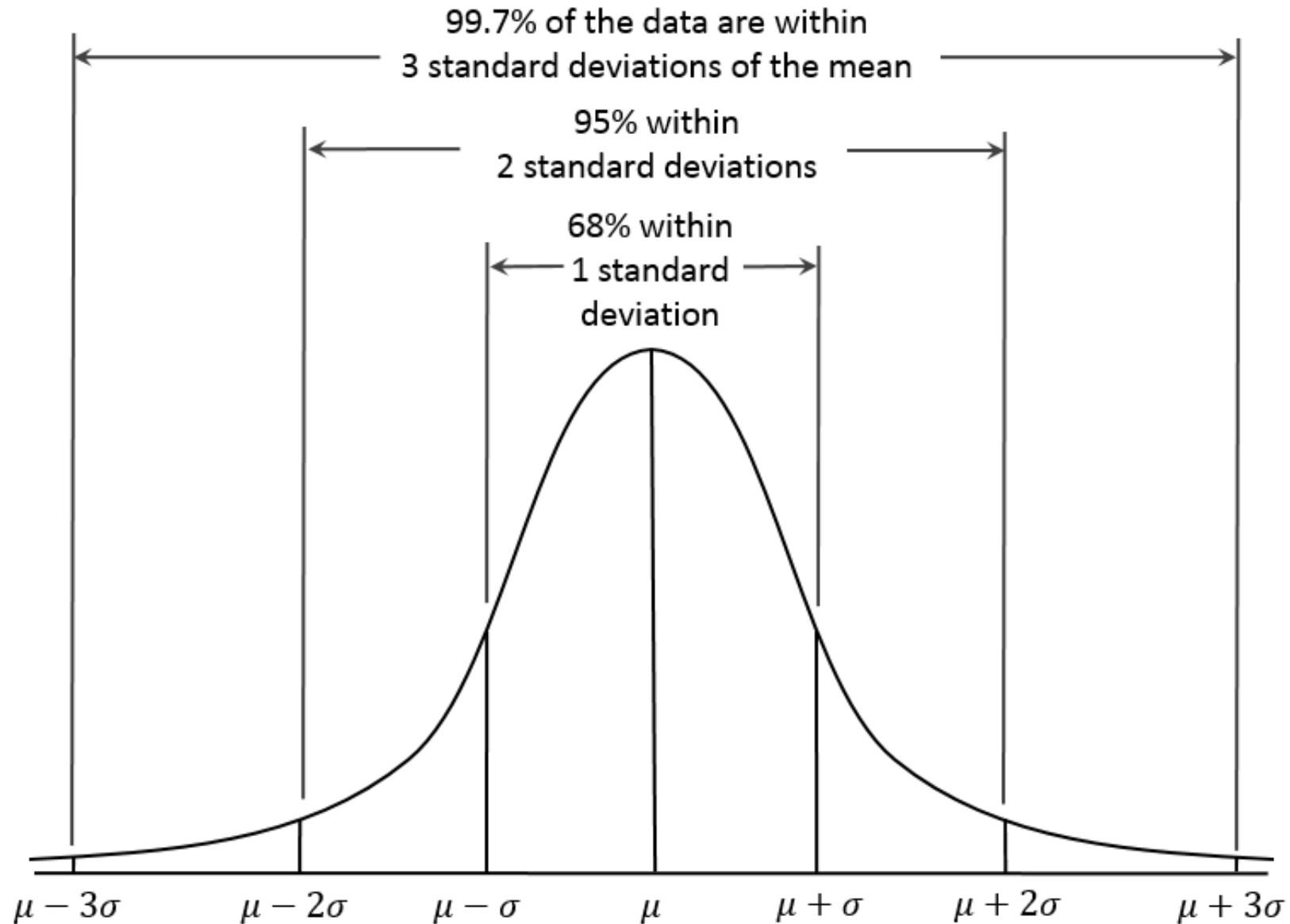
Key properties:

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

I want to sequence a 10Mbp genome so that
>97.5% of the genome has at least 24x coverage.
How many 120bp reads do I need?

Find X such that $X - 2\sqrt{X} = 24$

$$36 - 2\sqrt{36} = 24$$

I need $10\text{Mbp} \times 36\text{x} = 360\text{Mbp}$ of data
 $360\text{Mbp} / 120\text{bp} / \text{read} = 3\text{M reads}$

Genomics Arsenal in the year 2022

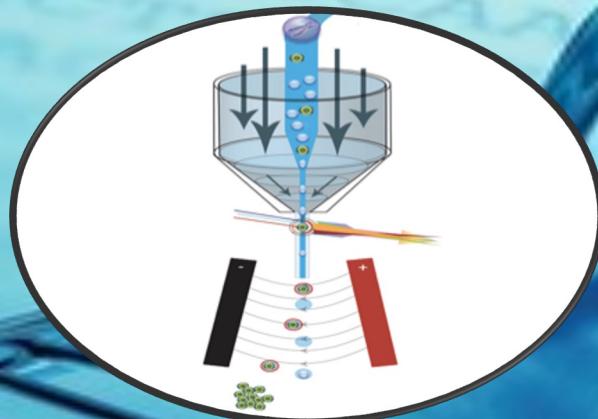
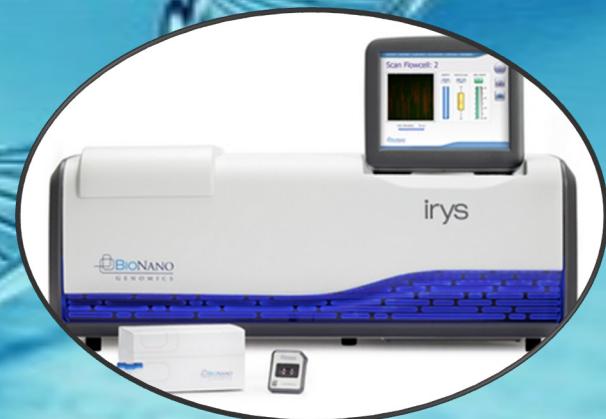
Sample Preparation



Sequencing



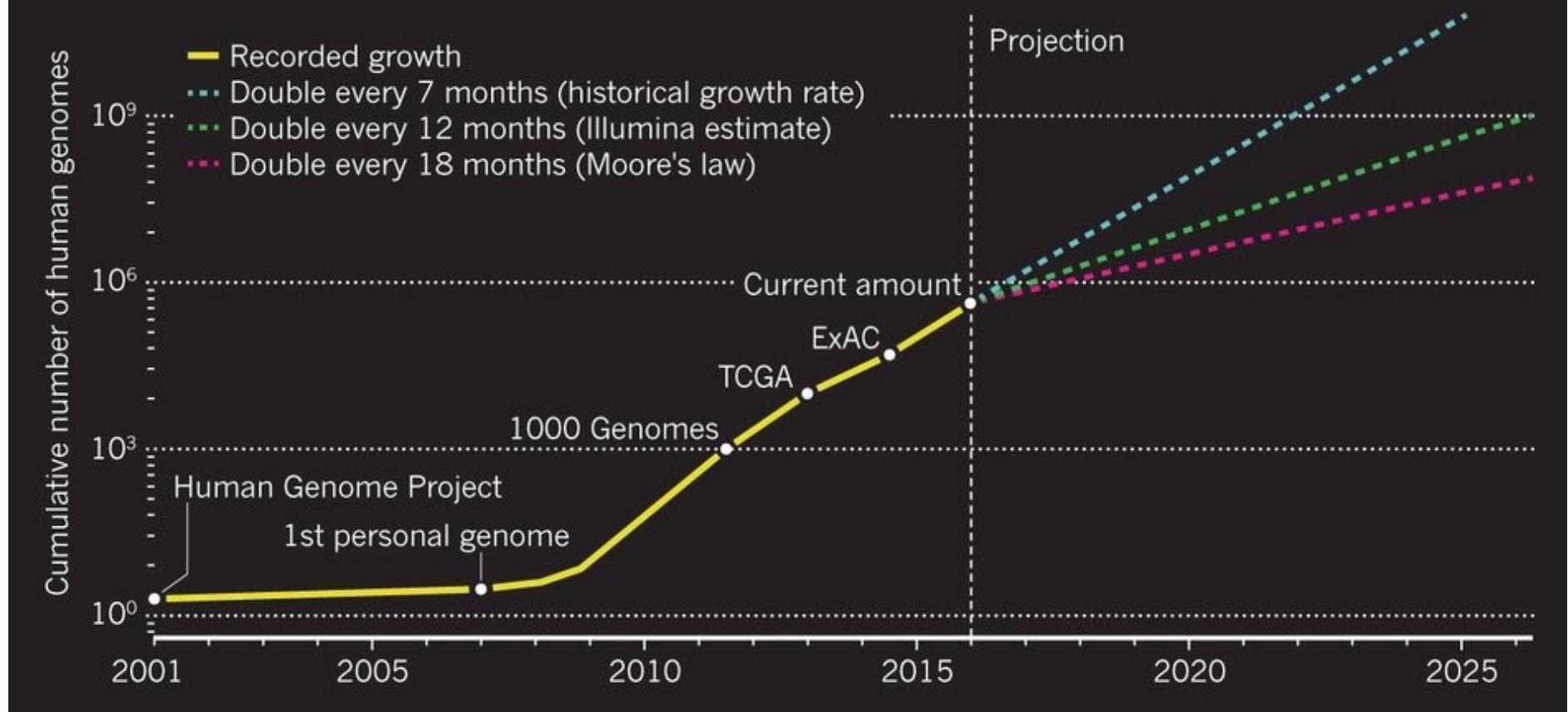
Chromosome Mapping



Sequencing Capacity

DNA SEQUENCING SOARS

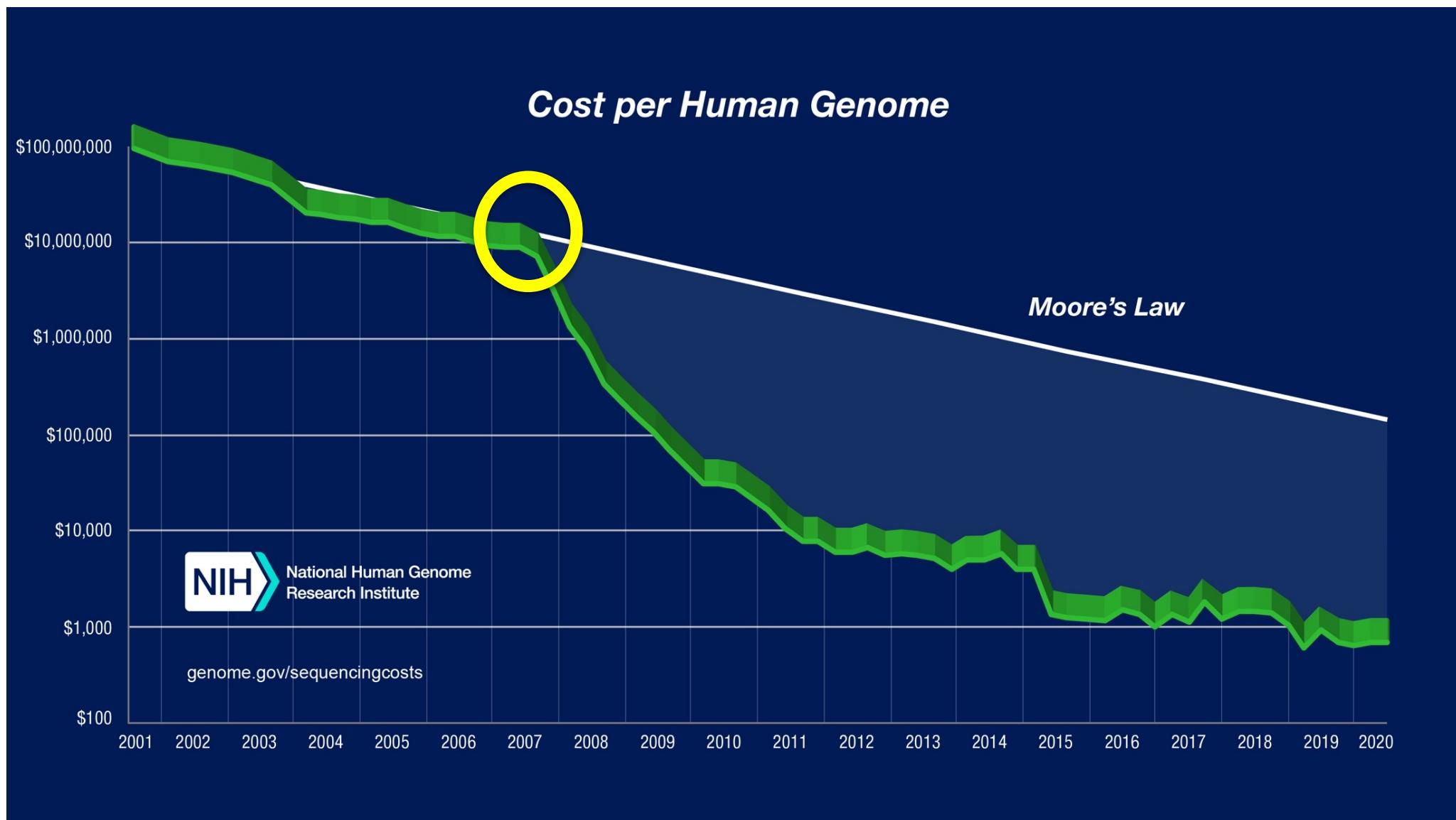
Human genomes are being sequenced at an ever-increasing rate. The 1000 Genomes Project has aggregated hundreds of genomes; The Cancer Genome Atlas (TCGA) has gathered several thousand; and the Exome Aggregation Consortium (ExAC) has sequenced more than 60,000 exomes. Dotted lines show three possible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

Cost per Genome

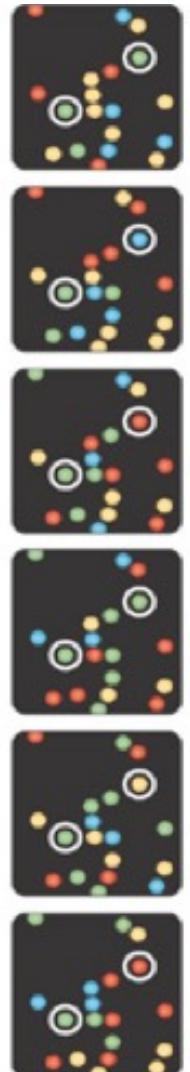
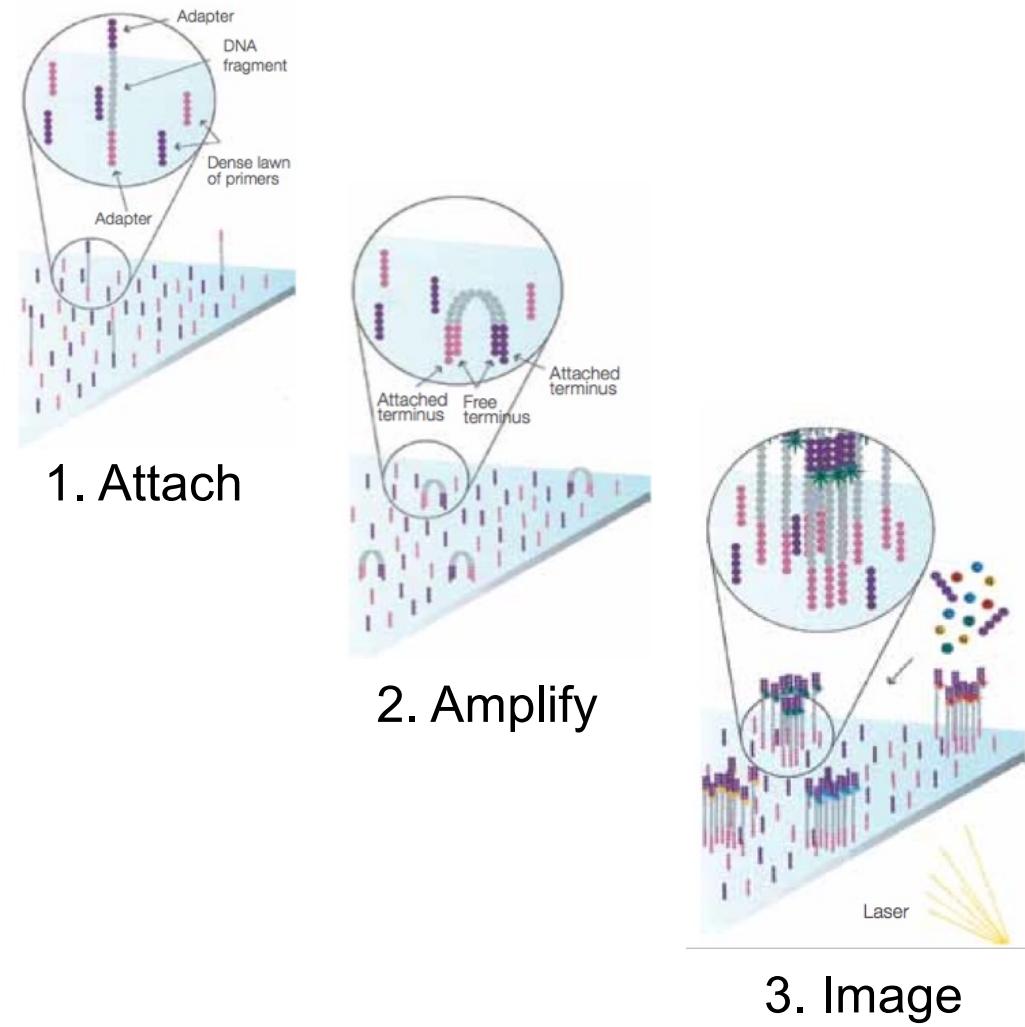


Second Generation Sequencing

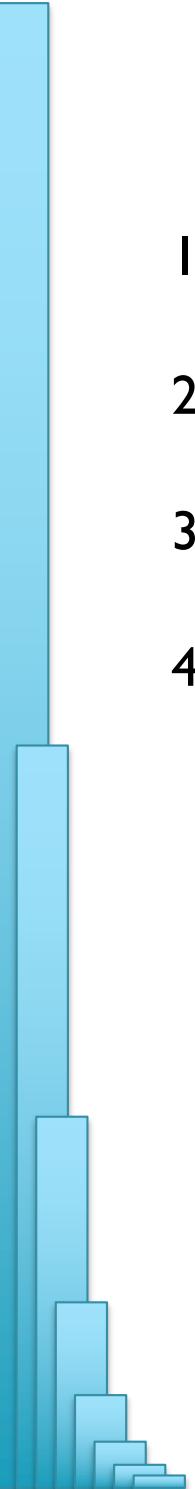


Illumina NovaSeq 6000
Sequencing by Synthesis

>3Tbp / day



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



Next Steps

1. Reflect on the magic and power of DNA 😊
2. Check out the course webpage
3. Register on Piazza
4. Work on Assignment I
 1. Set up Linux, set up Virtual Machine, set up Ubuntu
 2. Set up Dropbox for yourself!
 3. Get comfortable on the command line