

Genome Sequencing

Michael Schatz

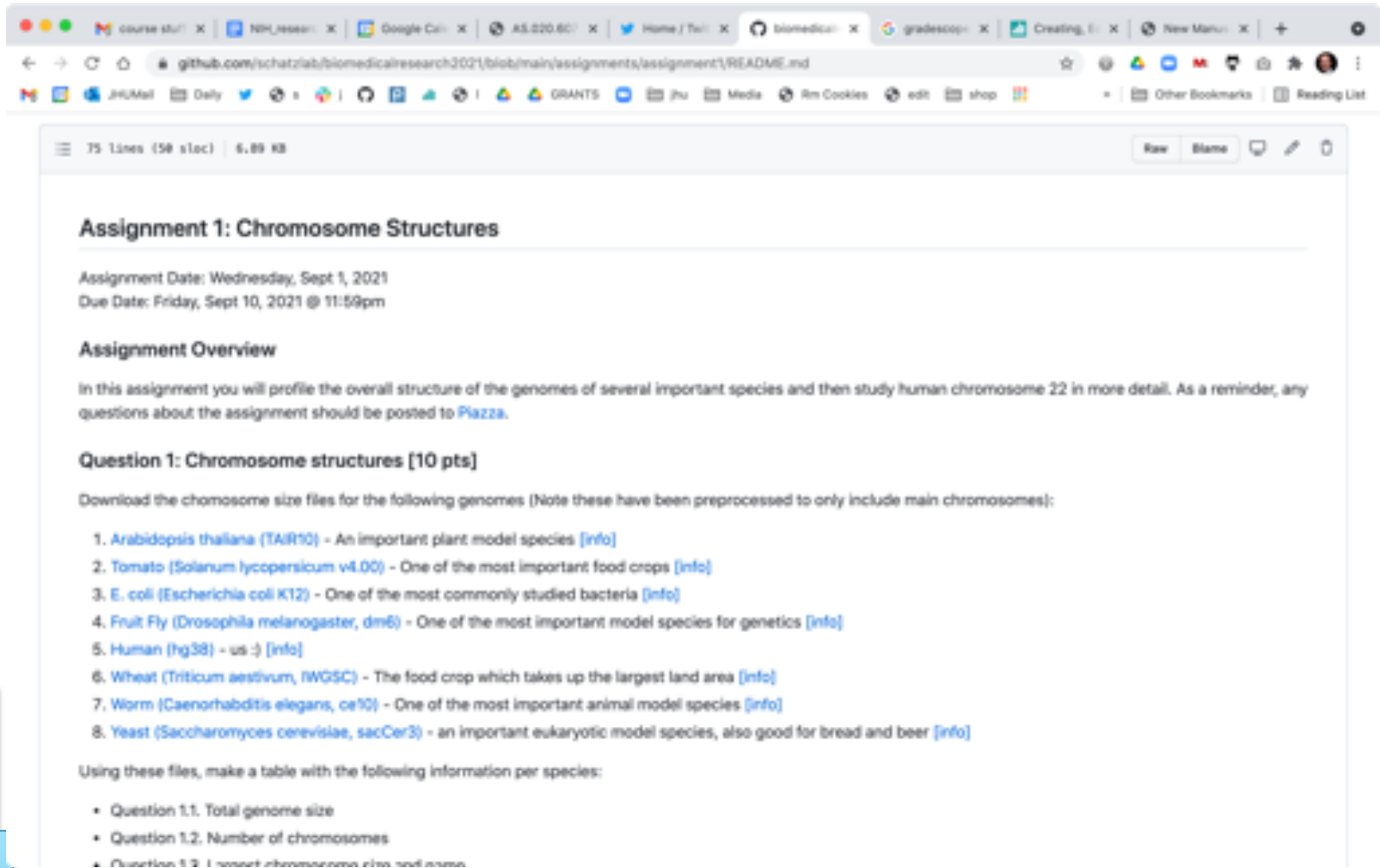
Sept 1, 2021

Lecture 2: Biomedical Research



Assignment 1: Chromosome Structures

Due Friday Sept 10 @ 11:59pm



The screenshot shows a web browser displaying a GitHub repository page. The browser's address bar shows the URL: github.com/schatzlab/biomedicalresearch2021/blob/main/assignments/assignment1/README.md. The page title is "Assignment 1: Chromosome Structures". Below the title, the assignment date is "Wednesday, Sept 1, 2021" and the due date is "Friday, Sept 10, 2021 @ 11:59pm". The "Assignment Overview" section states: "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](#)." The "Question 1: Chromosome structures [10 pts]" section instructs the user to "Download the chromosome size files for the following genomes (Note these have been preprocessed to only include main chromosomes):". A list of eight species follows, each with a number and a brief description, and a link to "info":

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

<https://github.com/schatzlab/biomedicalresearch2021>

Plotting in Python



The screenshot shows the Matplotlib website homepage. The browser's address bar displays `matplotlib.org`. The page features the Matplotlib logo with the text "Version 3.4.3" and a navigation bar with links for "Installation", "Documentation", "Examples", "Tutorials", and "Contributing". A search bar is located on the right side of the navigation bar. Below the navigation bar, the page title is "Matplotlib: Visualization with Python". A sub-header states: "Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python." Four small plots are displayed: a line plot with multiple peaks, a bell curve, a heatmap, and a 3D surface plot. Below these plots, the text reads: "Matplotlib makes easy things easy and hard things possible." Three main sections are highlighted: "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), and "Extend" (Explore tailored functionality provided by third party packages; Learn more about Matplotlib through the many external learning resources). On the right side, there is a section for "Latest stable release 3.4.3" with links to "docs" and "changelog", and "Last release for Python 2 2.2.5" with links to "docs" and "changelog". Below this is a section for "Development version" with a link to "docs". Further down is a section for "Matplotlib cheatsheets" showing a preview of a cheatsheet. At the bottom right, there is a button that says "Support Matplotlib".

matplotlib
Version 3.4.3

Installation Documentation Examples Tutorials Contributing

home | contents | Matplotlib: Python plotting

modules | index

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

github.com/instudio/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.

data + **geom** = **plot**
 $z = f(x, y)$

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

data + **geom** = **plot**
 $z = f(x, y)$
color = A
size = B

Complete the template below to build a graph.

```
ggplot(data = <DATA>) +  
  <GEOM_FUNCTION>[mapping = aes(<MAPPING>)] +  
  stat = <STAT>[position = <POSITION>] +  
  <COORDINATE_FUNCTION> +  
  <THEME_FUNCTION> +  
  <SCALE_FUNCTION>
```

ggplot(data = mpg, aes(x = cty, y = hwy)) Begins a plot that you finish by adding layers. 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(aes(x = long, y = lat))
```

- a = geom_blank()** and **a = expand_limits()**
Ensure limits include values across all plots.
- b = geom_curve()**
aes(x = long, y = lat, curvature = 10, xend = x, yend = y, alpha, angle, color, curvature, linetype, size)
- a = geom_path()**
aes(linewidth = "butt", linetype = "solid", linetype = 1)
x, y, alpha, color, group, linetype, size
- a = geom_polygon()**
aes(alpha = 500, x, y, alpha, color, fill, group, subgroup, linetype, size)
- b = geom_rect()**
aes(xmin = long, ymin = lat, xmax = long + 1, ymax = lat + 1, alpha, color, fill, linetype, size)
- a = geom_ribbon()**
aes(ymin = unemploy - 900, ymax = unemploy + 900, x, ymax, ymin, alpha, color, fill, group, linetype, size)

TWO VARIABLES both continuous

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

- a = geom_label()**
aes(x = cty, y = hwy, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust)
- a = geom_point()**
x, y, alpha, color, fill, shape, size, stroke
- a = geom_quantile()**
x, y, alpha, color, group, linetype, size, weight
- a = geom rug()**
aes(x = "cty", x, y, alpha, color, linetype, size)
- a = geom_smooth()**
method = "lm", x, y, alpha, color, fill, group, linetype, size, weight
- a = geom_text()**
aes(label = cty, nudges_x = 1, nudges_y = 10, x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust)

continuous bivariate distribution

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

- b = geom_bin2d()**
binwidth = c(0.25, 500), x, y, alpha, color, fill, linetype, size, weight
- b = geom_density_2d()**
x, y, alpha, color, group, linetype, size
- b = 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 = "hu", x, y, alpha, color, group, linetype, size

visualizing error

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

- j = geom_crossbar()**
x = g, y, ymin, ymax, alpha, color, fill, group, linetype, size
- j = geom_errorbar()**
x, y, ymin, ymax, alpha, color, group, linetype, size, width, size, geom_errorbarh()
- j = geom_linerange()**
x, y, ymin, ymax, alpha, color, group, linetype, size
- j = geom_pointrange()**
x, y, ymin, ymax, alpha, color, fill, group, linetype, shape, size

one discrete, one continuous

```
f <- ggplot(mpg, aes(class, hwy))
```

- f = geom_col()**
x, y, alpha, color, fill, group, linetype, size
- f = geom_boxplot()**
x, y, lower, middle, upper, ymax, ymin, alpha, color, fill, group, linetype, shape, size, weight
- f = geom_dotplot()**
binwidth = "y", stackdir = "center", x, y, alpha, color, fill, group
- f = geom_violin()**
scale = "area", x, y, alpha, color, fill, group, linetype, size, weight

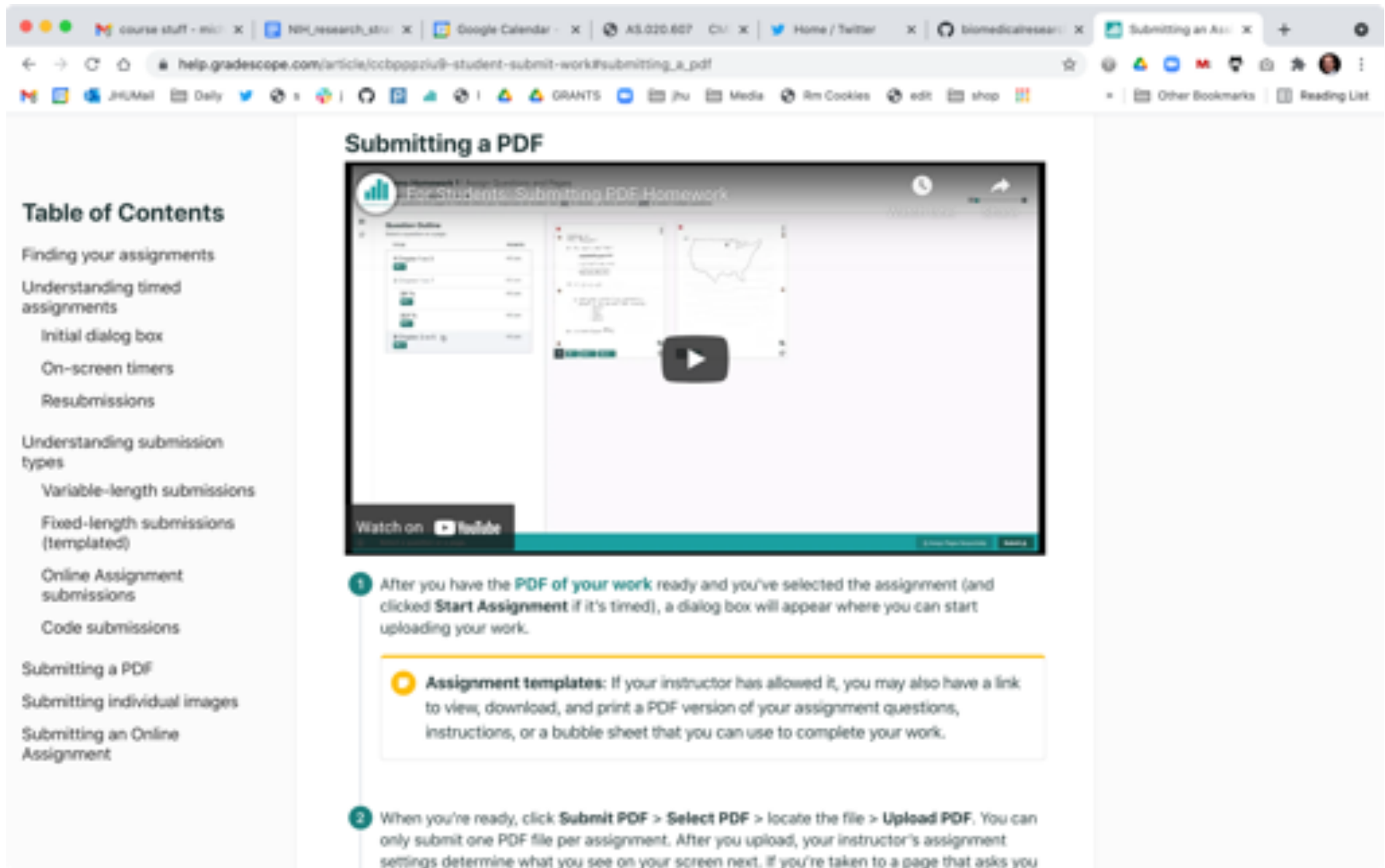
both discrete

```
g <- ggplot(diamonds, aes(cut, color))
```

- g = geom_raster()**

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

Submission with GradeScope



course stuff - mic X | NH_research_stru X | Google Calendar X | AS.020.607 CUI X | Home / Twitter X | biomedicalresearch X | Submitting an Assignment X

help.gradescope.com/article/ocbpgpziu8-student-submit-work#submitting_a_pdf

Table of Contents

- Finding your assignments
- Understanding timed assignments
 - Initial dialog box
 - On-screen timers
 - Resubmissions
- Understanding submission types
 - Variable-length submissions
 - Fixed-length submissions (templated)
 - Online Assignment submissions
 - Code submissions
- Submitting a PDF
- Submitting individual images
- Submitting an Online Assignment

Submitting a PDF

For Students: Submitting PDF Homework

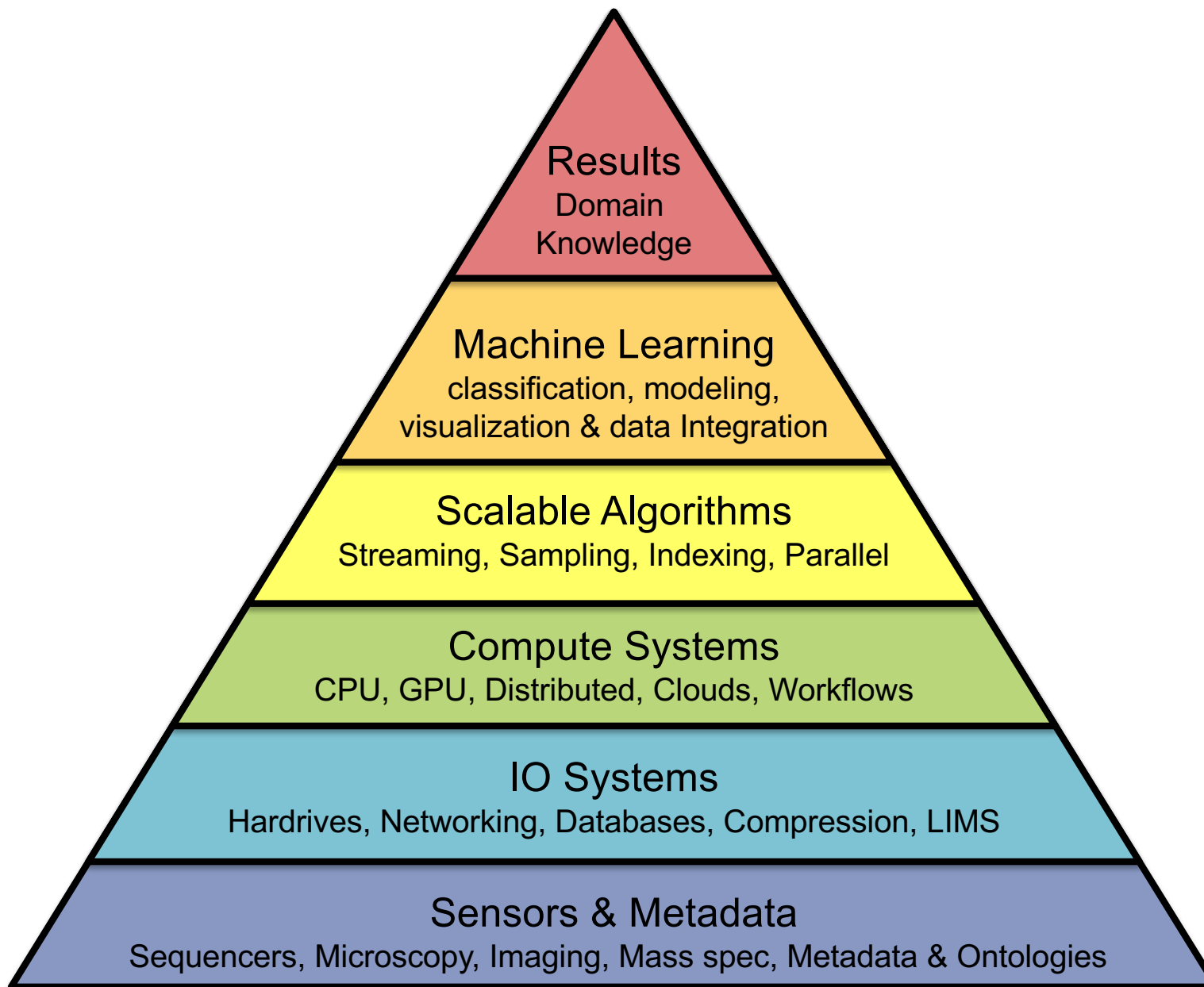
Watch on YouTube

- 1 After you have the **PDF of your work** ready and you've selected the assignment (and clicked **Start Assignment** if it's timed), a dialog box will appear where you can start uploading your work.
 - Assignment templates:** If your instructor has allowed it, you may also have a link to view, download, and print a PDF version of your assignment questions, instructions, or a bubble sheet that you can use to complete your work.
- 2 When you're ready, click **Submit PDF** > **Select PDF** > locate the file > **Upload PDF**. You can only submit one PDF file per assignment. After you upload, your instructor's assignment settings determine what you see on your screen next. If you're taken to a page that asks you

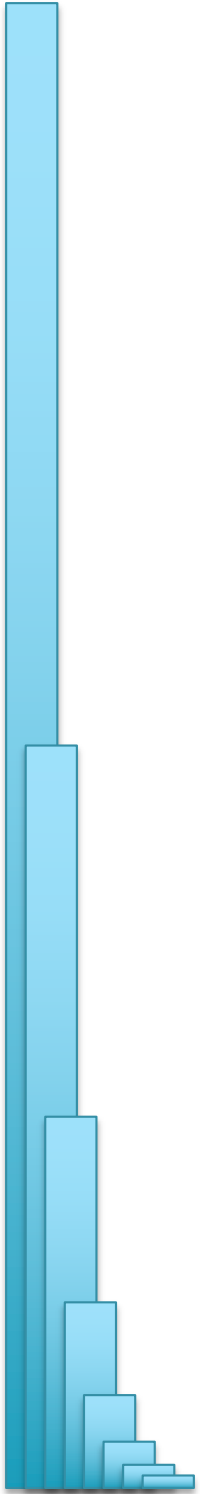
<https://www.gradescope.com/>

Entry Code:D5GDXP

Biomedical Genomics Technologies



Part I: Sequencing



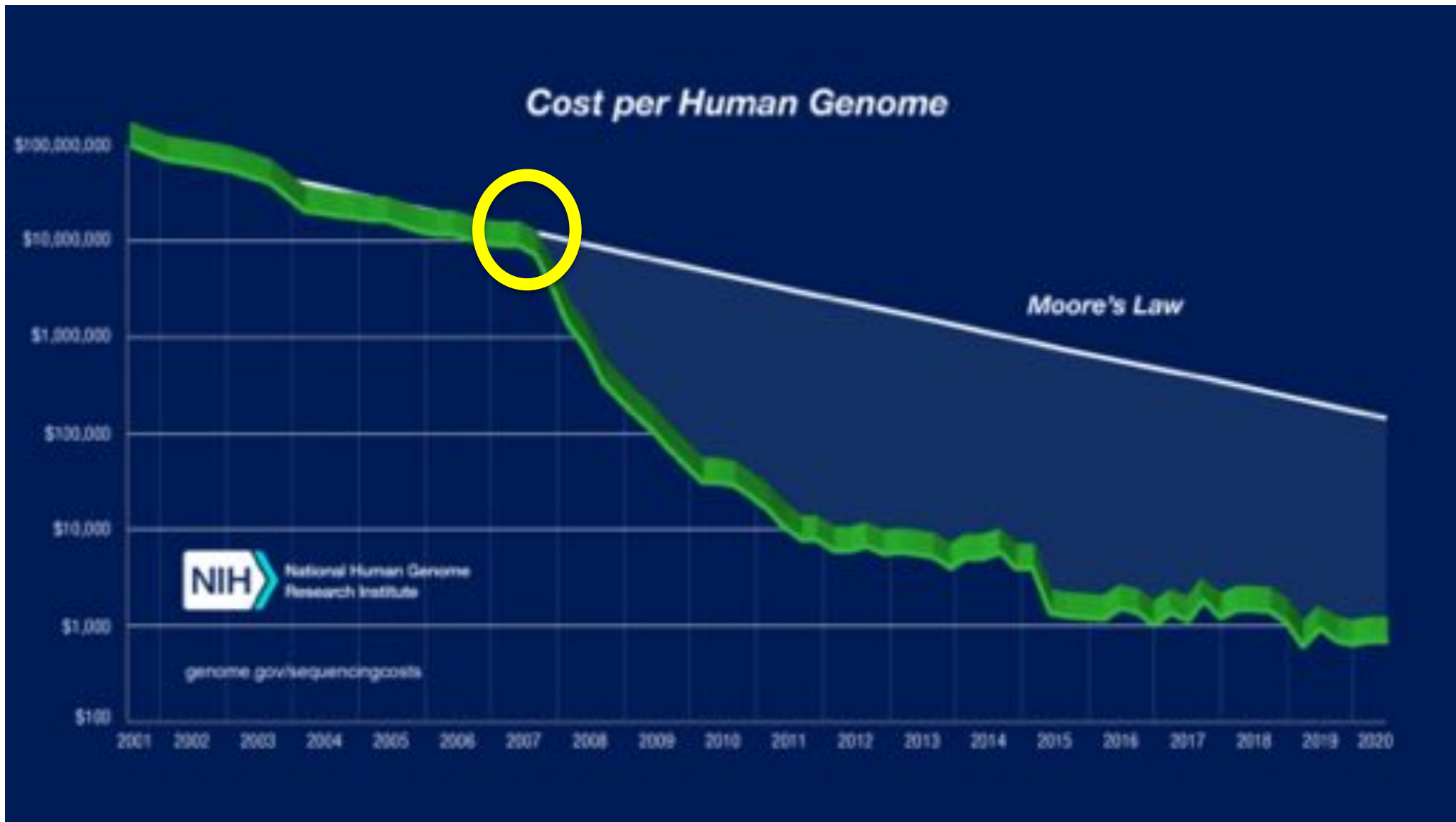
The most wondrous map...



“Without a doubt, this is the most important, most wondrous map ever produced by humankind.”

*Bill Clinton
June 26, 2000*

Cost per Genome

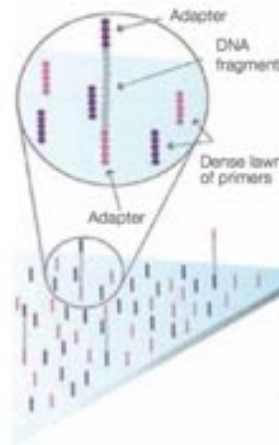


Second Generation Sequencing

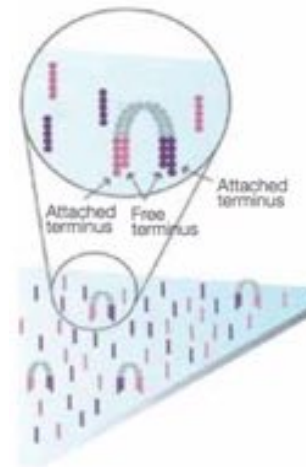


Illumina NovaSeq 6000
Sequencing by Synthesis

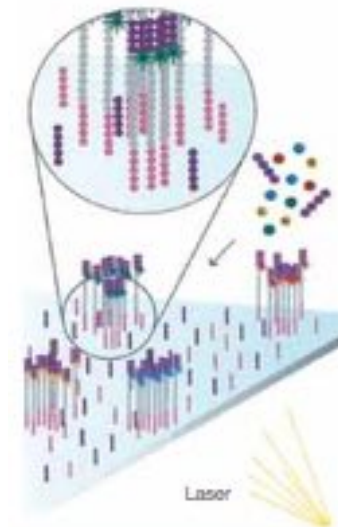
>3Tbp / day



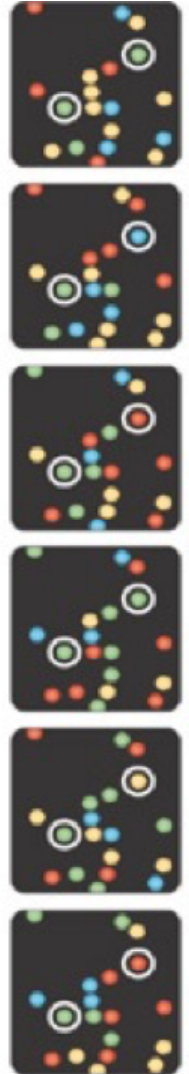
1. Attach



2. Amplify



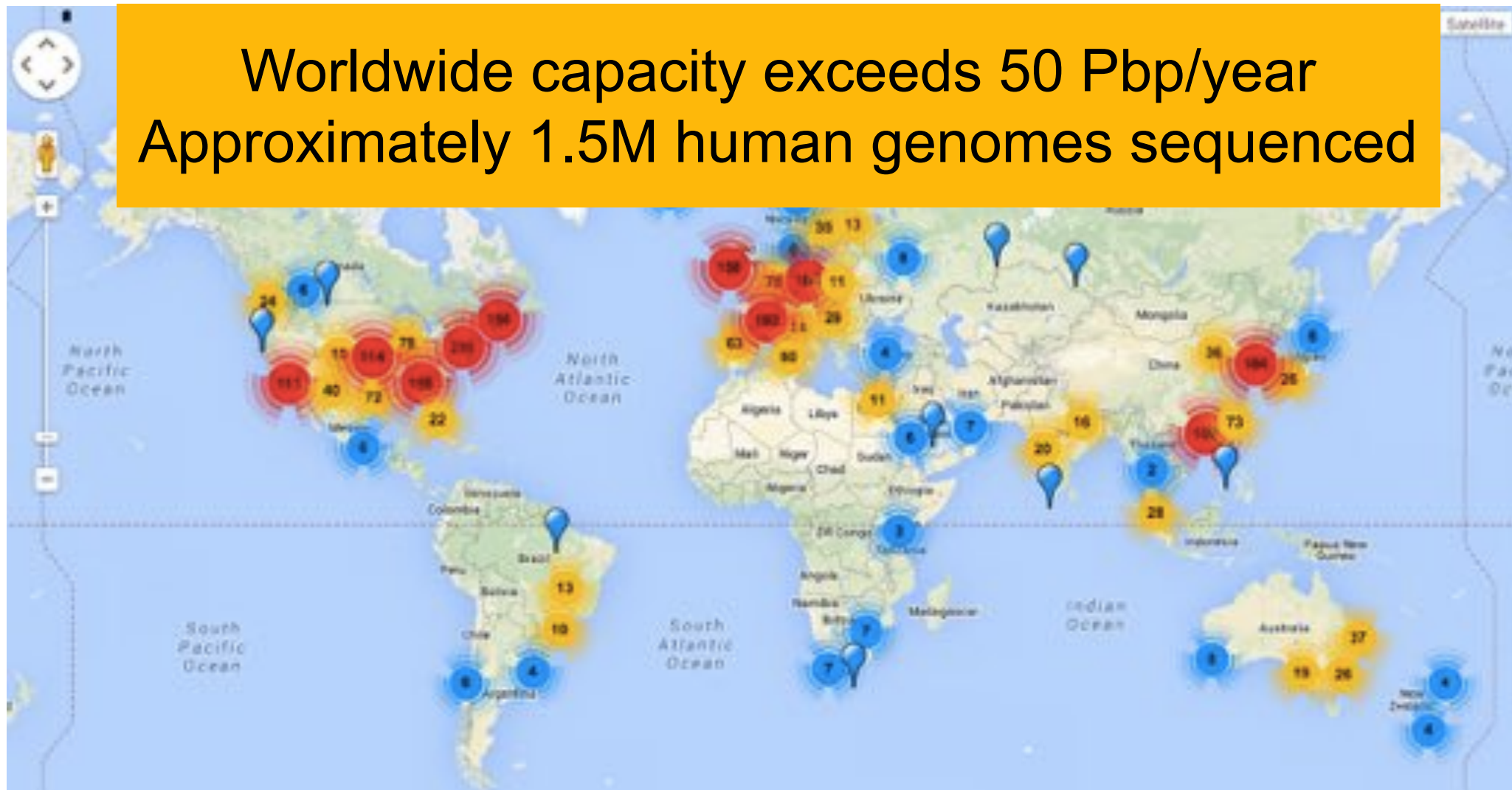
3. Image



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

Sequencing Centers

Worldwide capacity exceeds 50 Pbp/year
Approximately 1.5M human genomes sequenced



Next Generation Genomics: World Map of High-throughput Sequencers

<http://omicsmaps.com>

How much is a petabyte?

Unit	Size
Byte	1
Kilobyte	1,000
Megabyte	1,000,000
Gigabyte	1,000,000,000
Terabyte	1,000,000,000,000
Petabyte	1,000,000,000,000,000

*Technically a kilobyte is 2^{10} and a petabyte is 2^{50}

How much is a petabyte?



100 GB / Genome
4.7GB / DVD
~20 DVDs / Genome

X

10,000 Genomes

=

1PB Data
200,000 DVDs



787 feet of DVDs
~1/6 of a mile tall

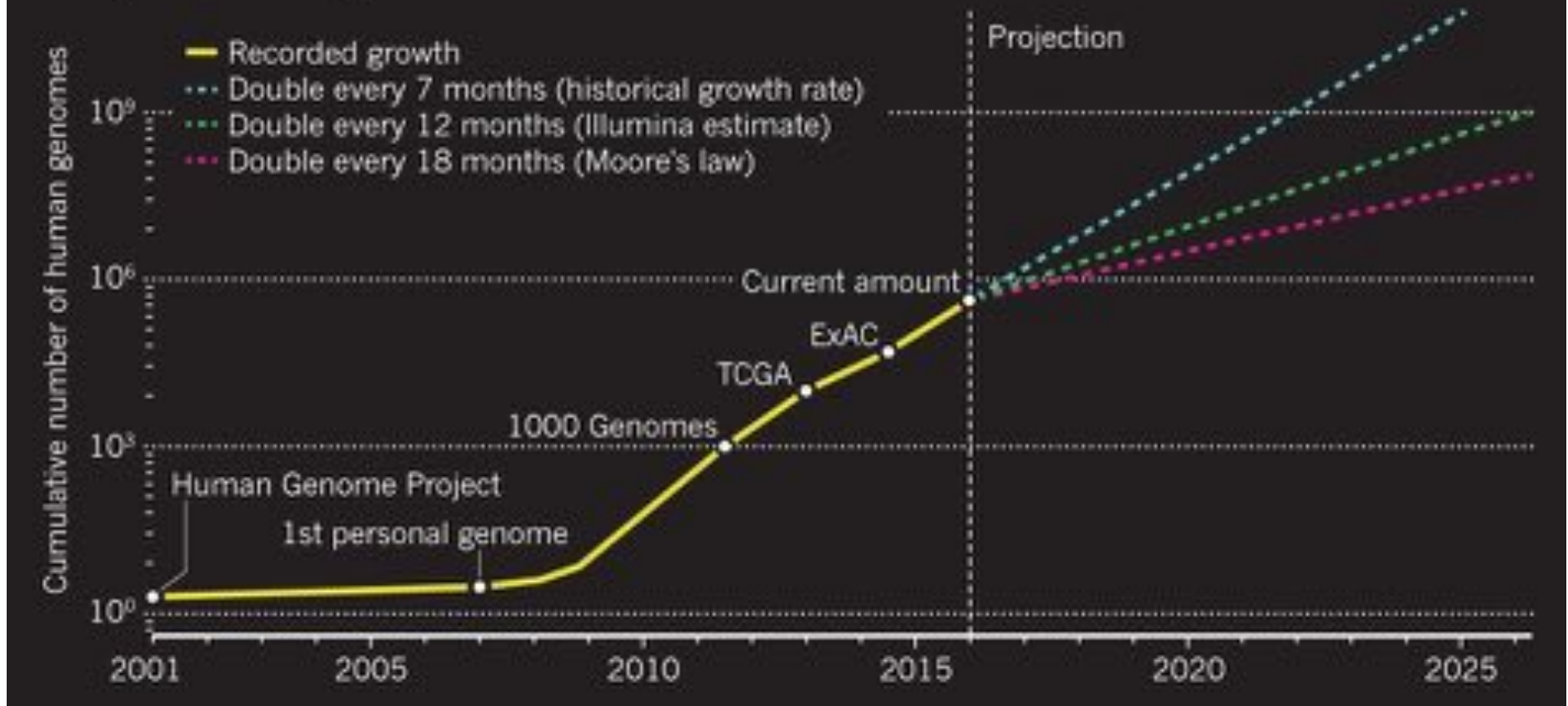


500 2 TB drives
\$50k

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)

How much is a zettabyte?

Unit	Size
Byte	1
Kilobyte	1,000
Megabyte	1,000,000
Gigabyte	1,000,000,000
Terabyte	1,000,000,000,000
Petabyte	1,000,000,000,000,000
Exabyte	1,000,000,000,000,000,000
Zettabyte	1,000,000,000,000,000,000,000

How much is a zettabyte?



100 GB / Genome
4.7GB / DVD
~20 DVDs / Genome

X

10,000,000,000 Genomes

=

1ZB Data
200,000,000,000 DVDs



150,000 miles of DVDs
~ ½ distance to moon

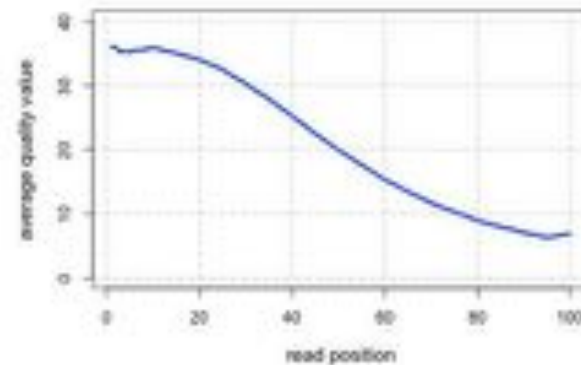


Both currently ~100Pb
And growing exponentially

Illumina Quality

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

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



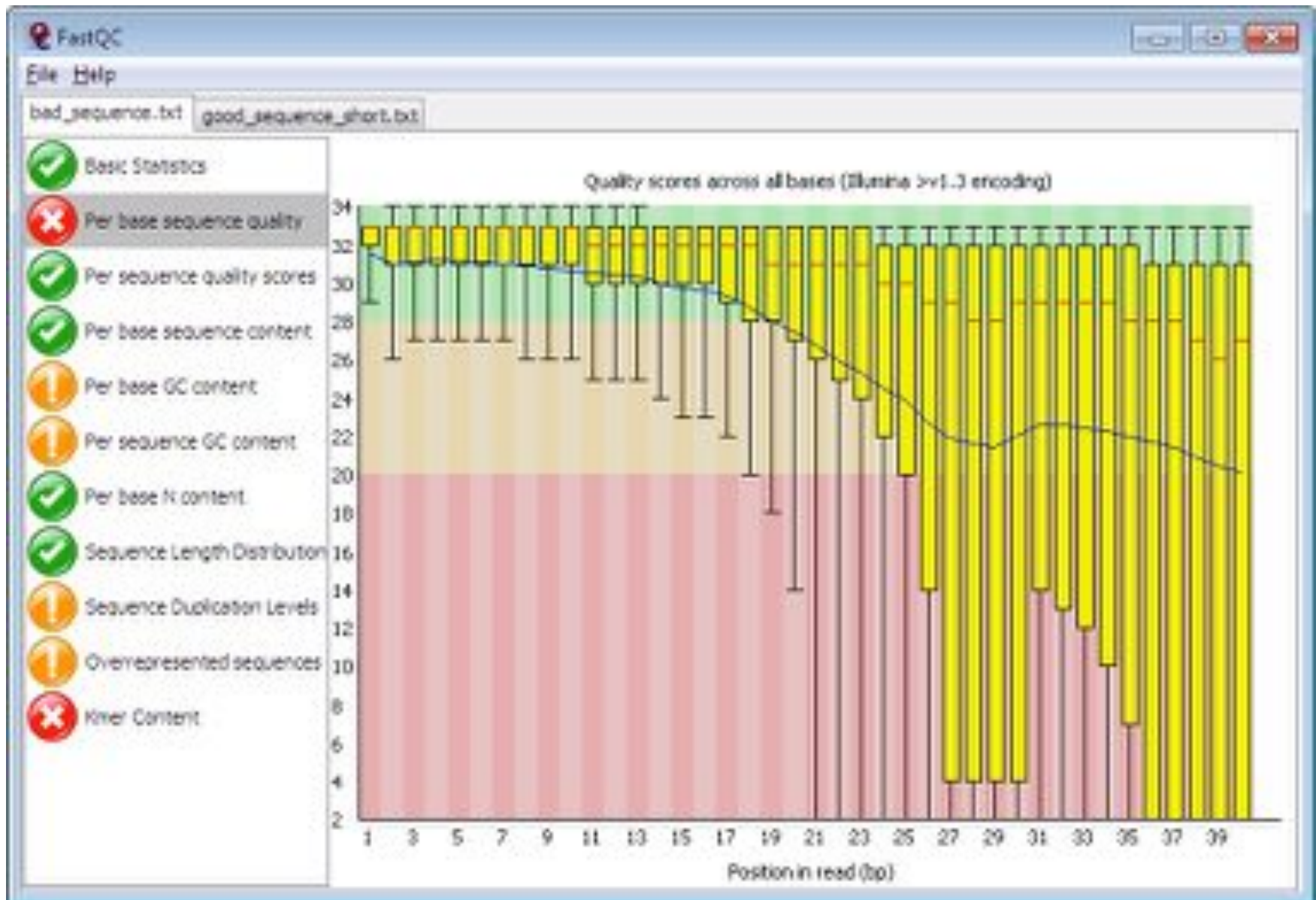
```

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

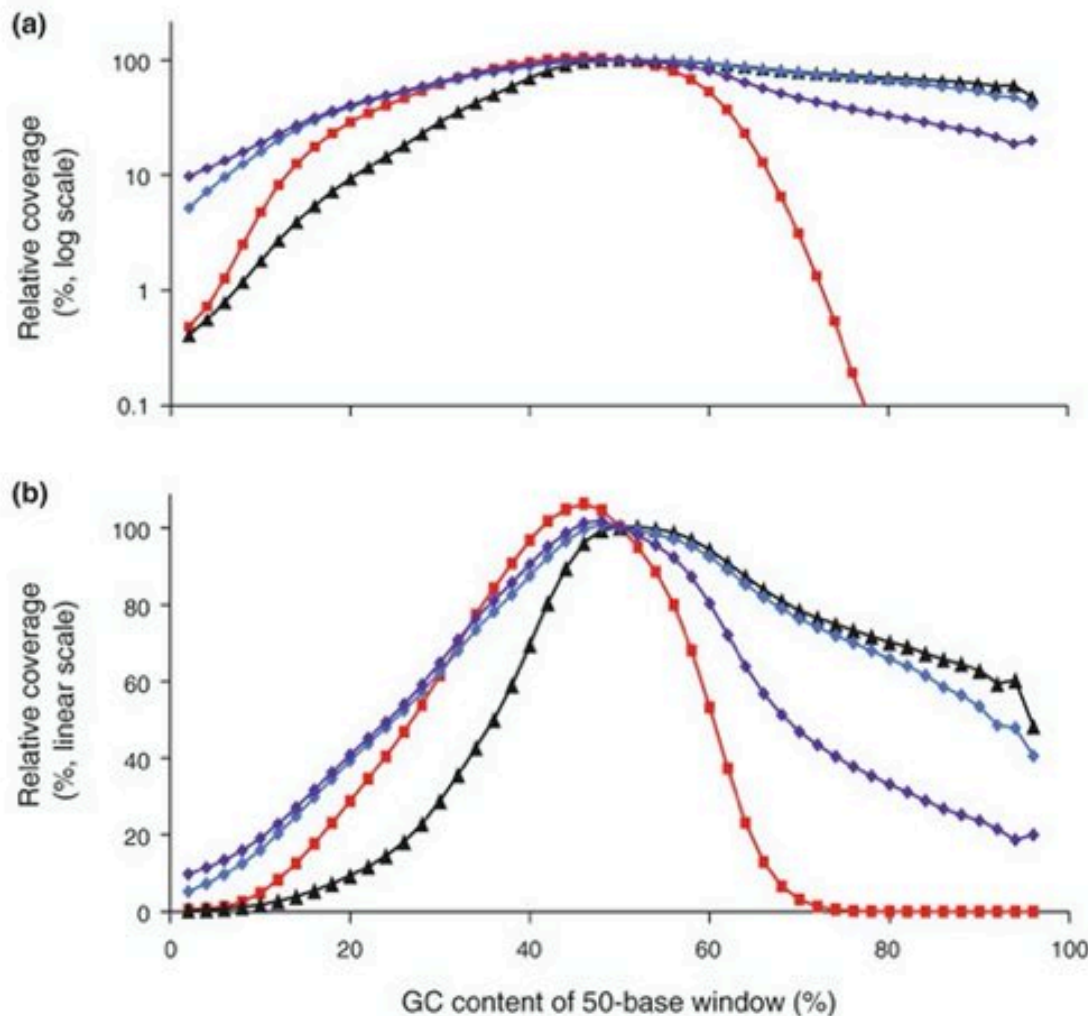
```

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)

FASTQC: Is my data any good?



Beware of GC Biases



Illumina sequencing does not produce uniform coverage over the genome

- Coverage of extremely high or extremely low GC content will have reduced coverage in Illumina sequencing
- Biases primarily introduced during PCR; lower temperatures, slower heating, and fewer rounds minimize biases
- This makes it very difficult to identify variants (SNPs, CNVs, etc) in certain regions of the genome

Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.

Aird et al. (2011) *Genome Biology*. 12:R18.

Question?

Genomes are big, Illumina reads are short.

We would love to generate
longer and longer reads with this technology

What can we do?

Paired-end and Mate-pairs

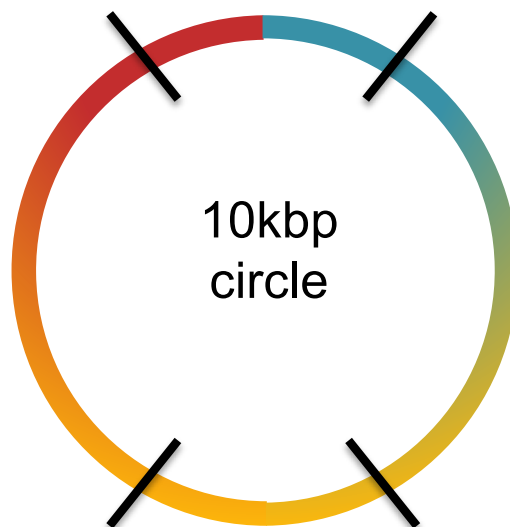
Paired-end sequencing

- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation



Mate-pair sequencing

- Circularize long molecules (1-10kbp), shear into fragments, & sequence
- Mate failures create short paired-end reads



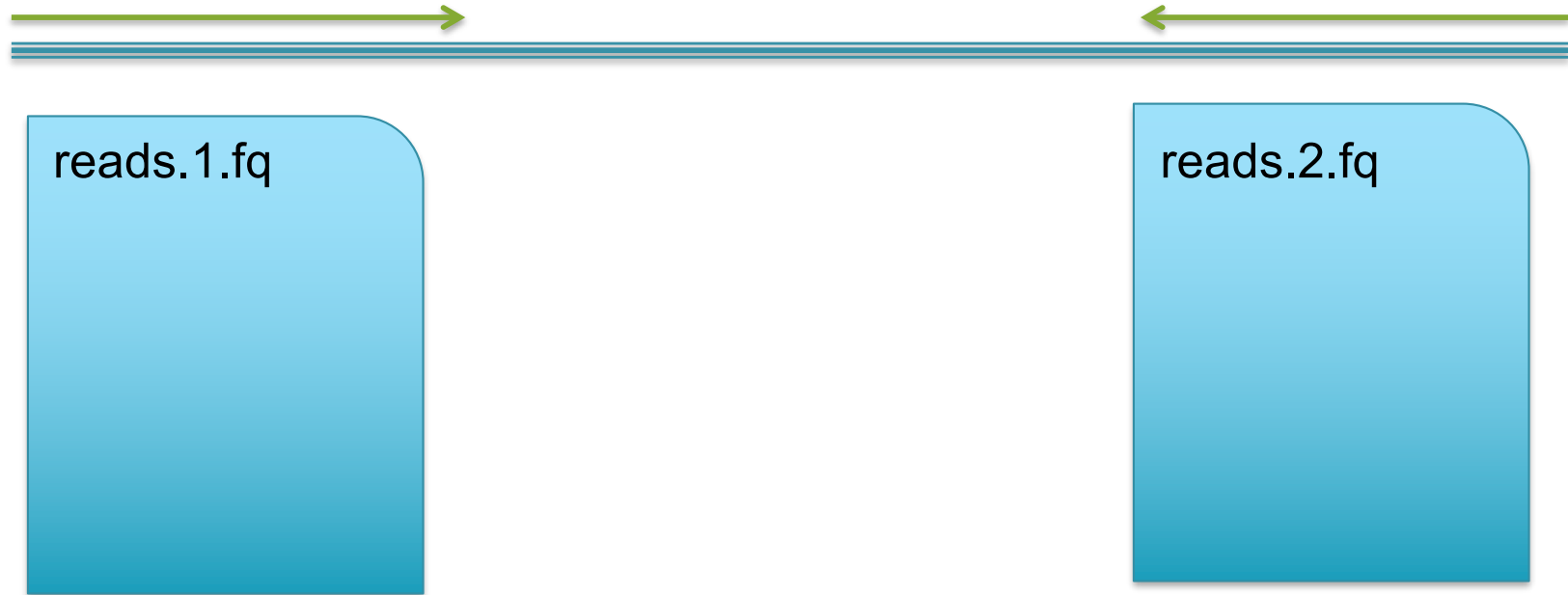
2x100 @ ~10kbp (outies)



2x100 @ 300bp (innies)



FASTQ Files



```
@SEQ_ID
GATTGTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTTT
+
!''*((( (**+))%+%++)(%%%) .1***-+*'')**55CCF>>>>>CCCCCCC65
```

@Identifier
Sequence
+Separator
Quality Values
...

Illumina Sequencing Summary

Advantages:

- Best throughput, accuracy and read length for any 2nd gen. sequencer
- Fast & robust library preparation

Disadvantages:

- Inherent limits to read length (practically, 150bp)
- Some runs are error prone
- Requires amplification, sequences a population of molecules



Illumina HiSeq

~3 billion paired 100bp reads
~600Gb, \$10K, 8 days
(or “rapid run” ~90Gb in 1-2 days)

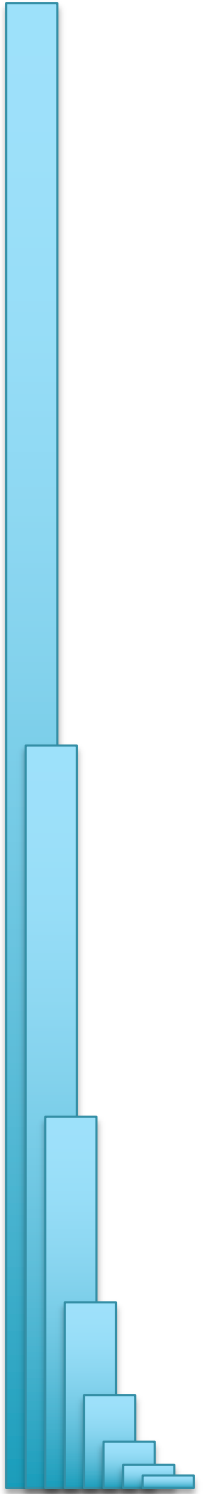
Illumina X Ten / NovaSeq

~6 billion paired 150bp reads
1.8Tb, <3 days, ~1000 / genome(\$\$)
(or “rapid run” ~90Gb in 1-2 days)

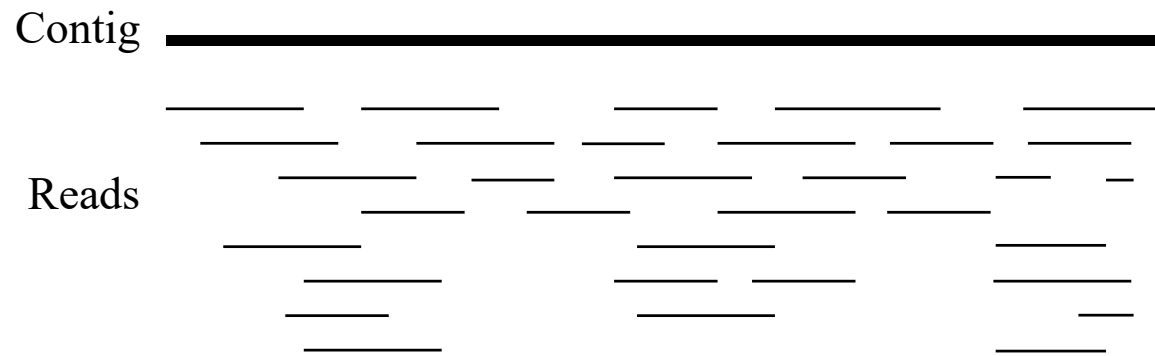
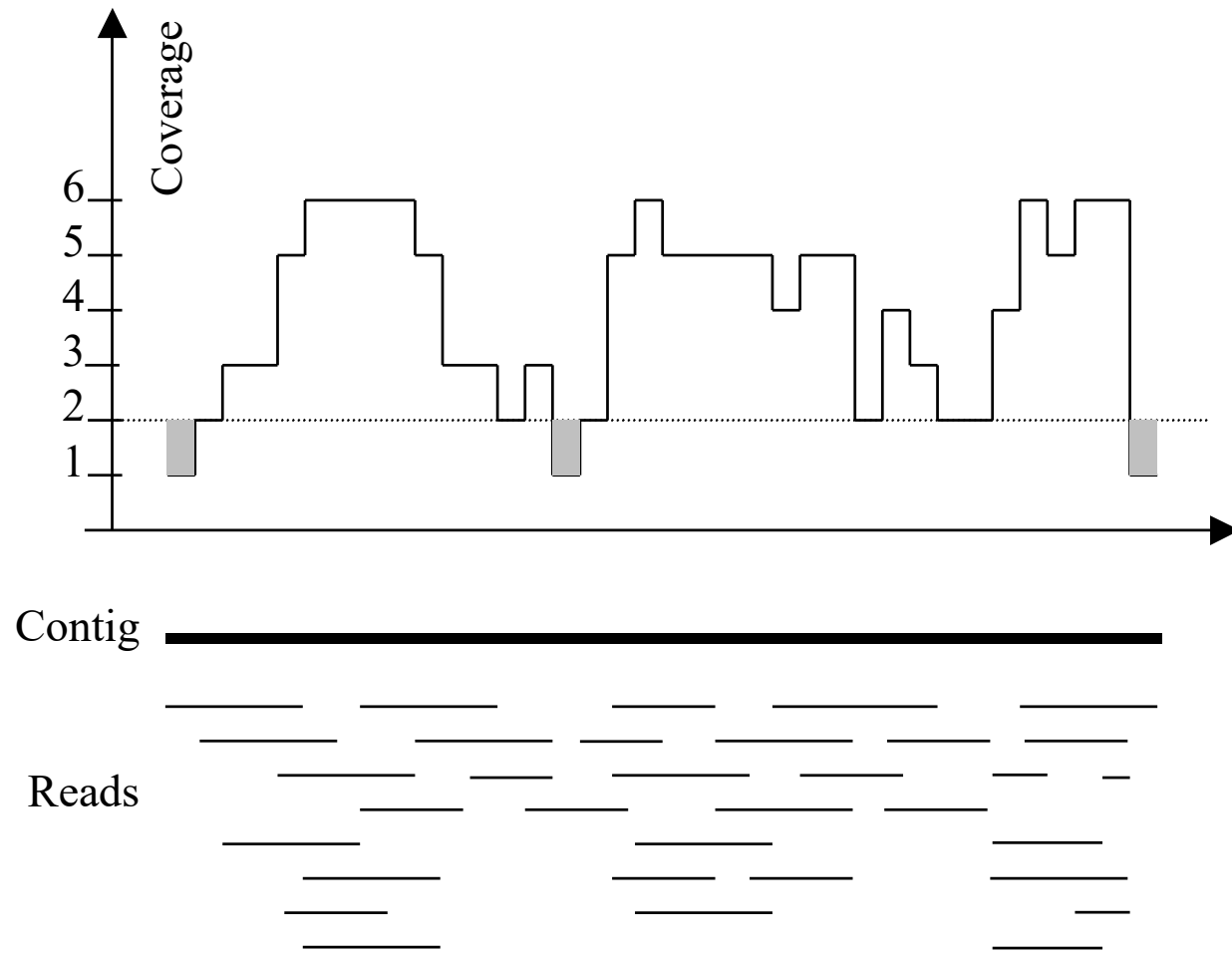
Illumina NextSeq

One human genome in **<30 hours**

Part 2: Coverage



Typical sequencing coverage

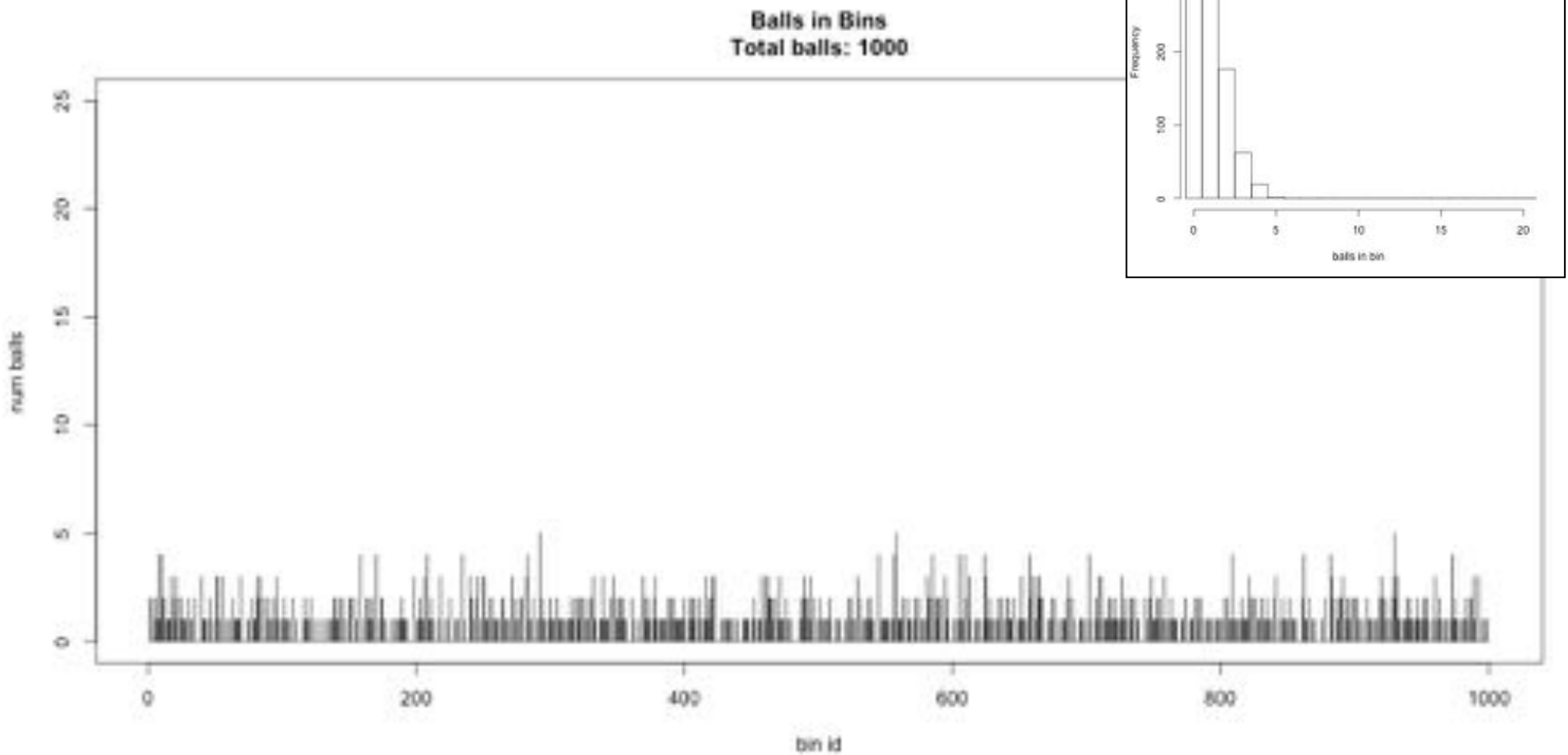


Imagine raindrops on a sidewalk

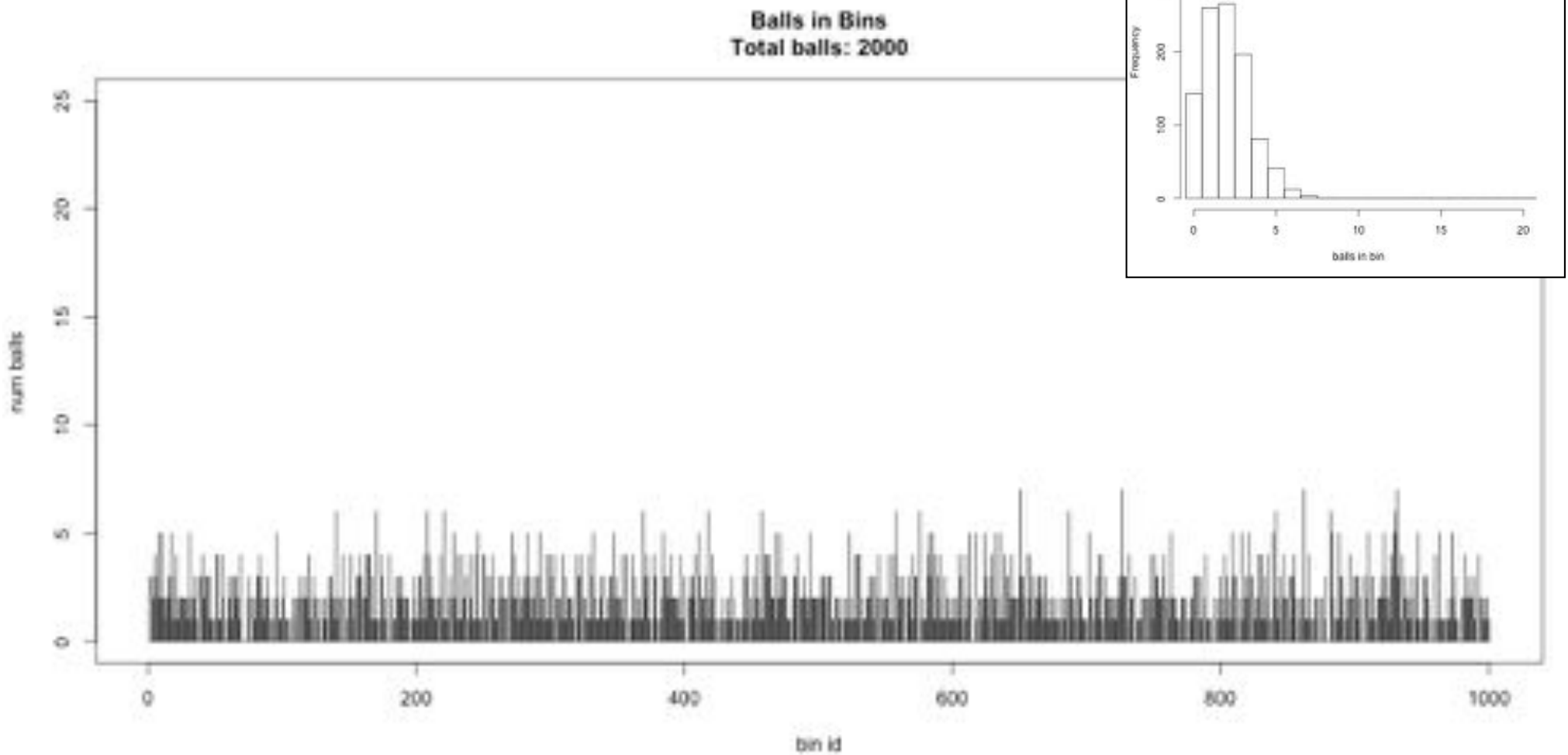
We want to cover the entire sidewalk but each drop costs \$1

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

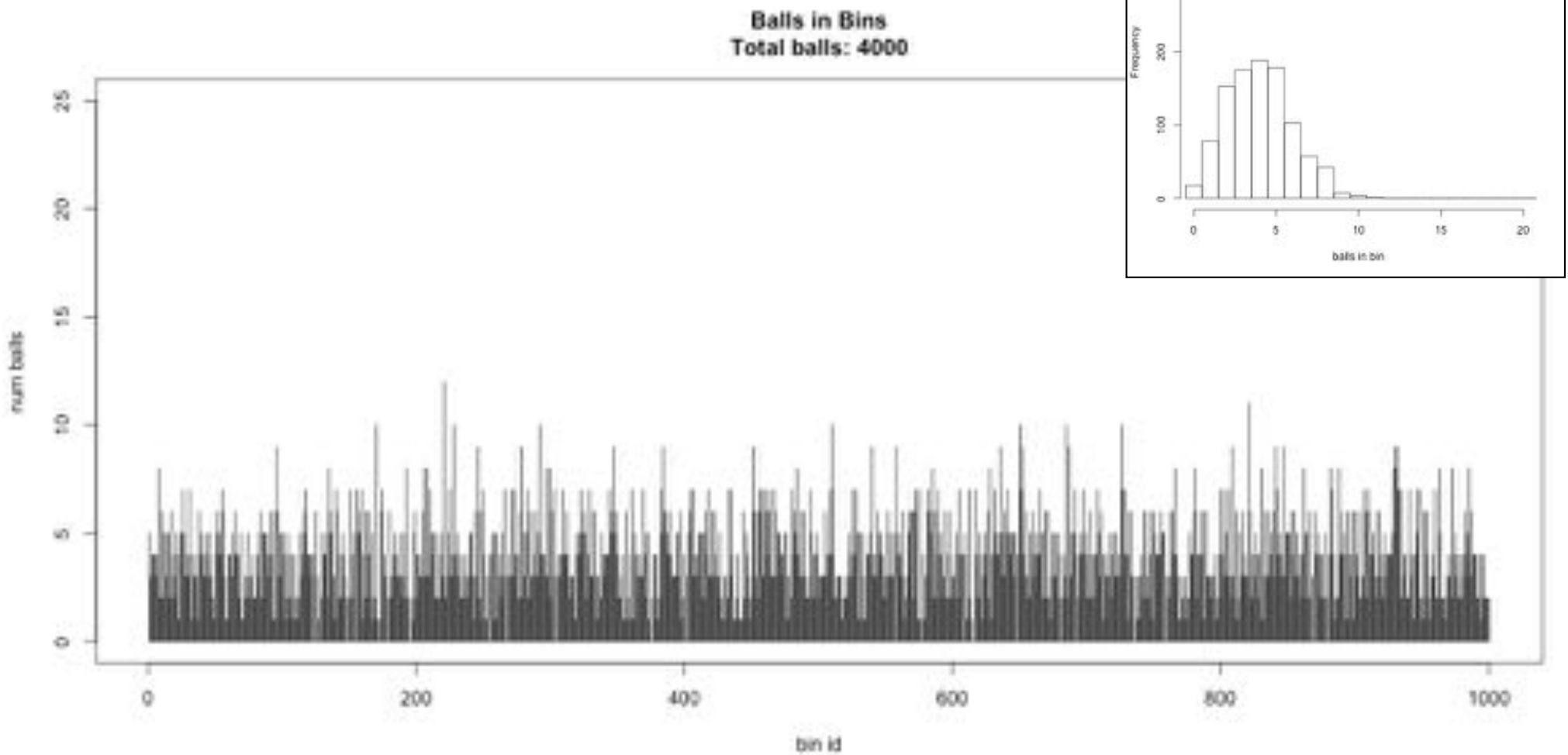
1x sequencing



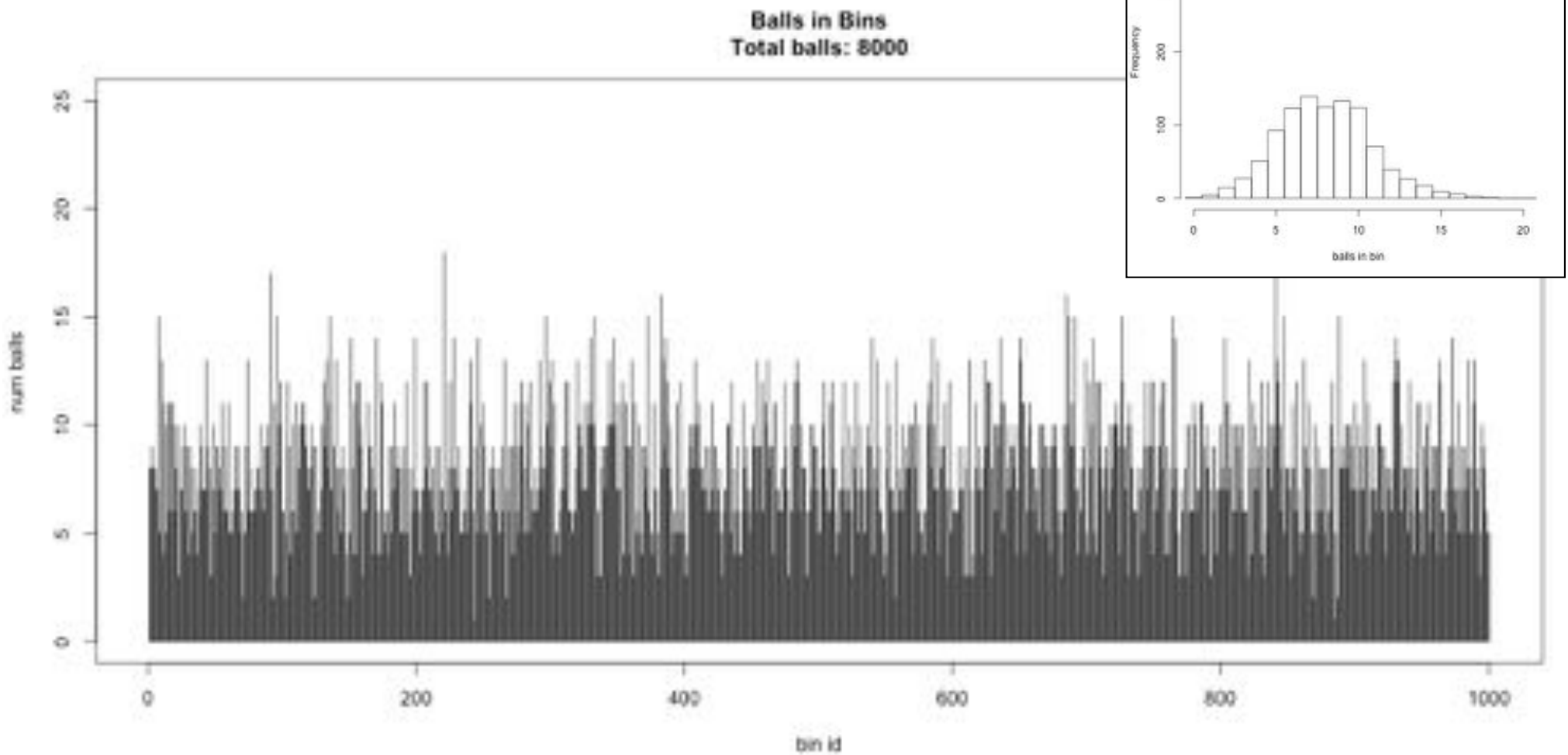
2x sequencing



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.

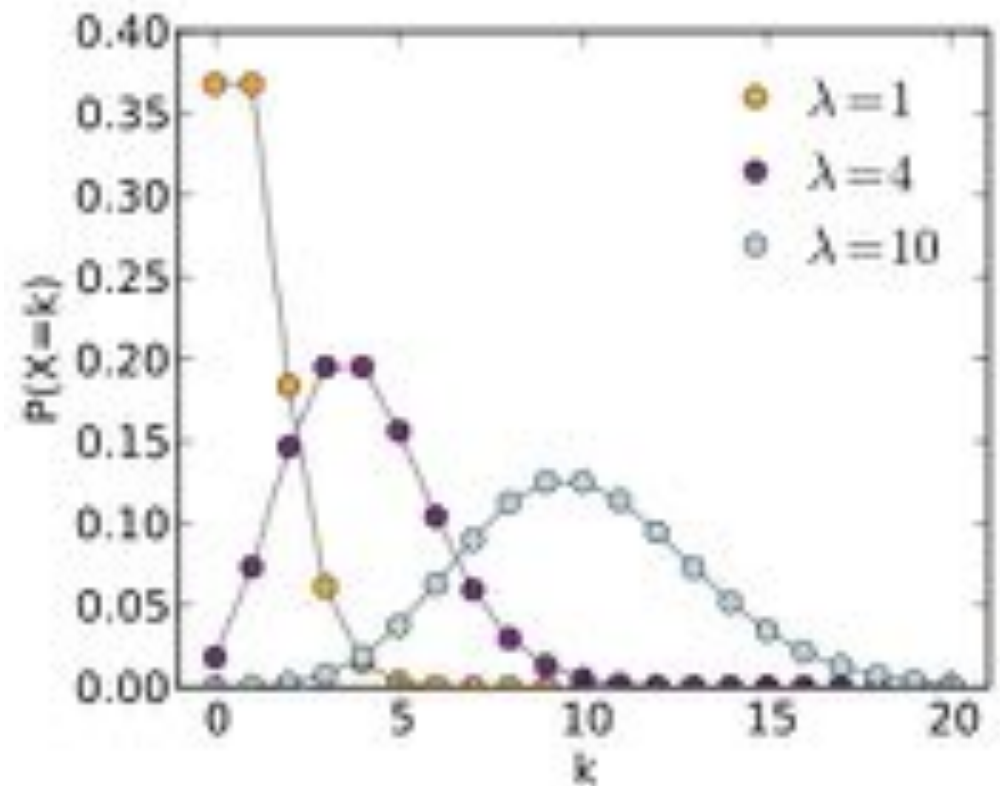
Formulation comes from the limit of the binomial equation

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

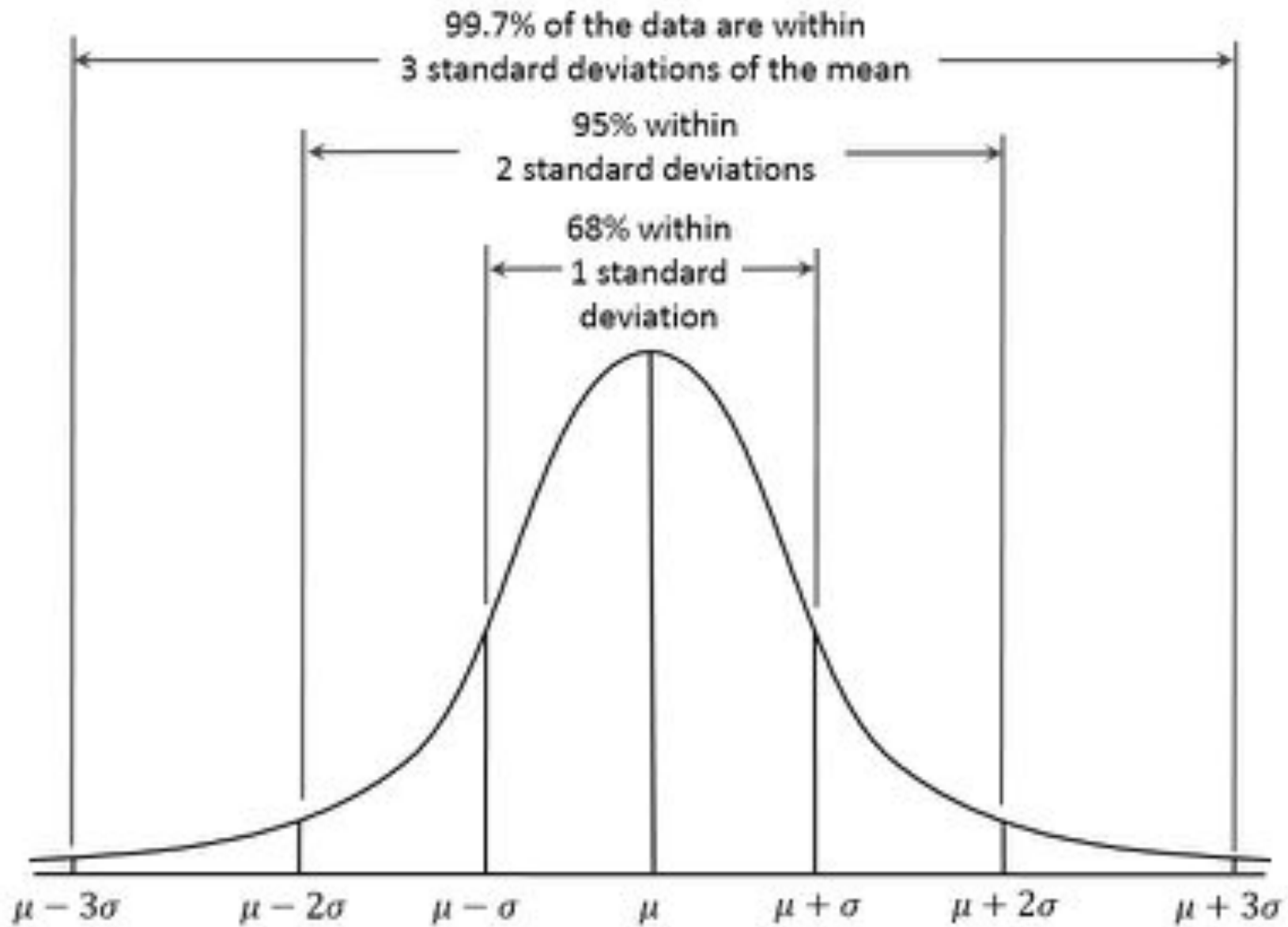
Key properties:

- ***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

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 $X - 2 \times \sqrt{X} = 24$

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

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