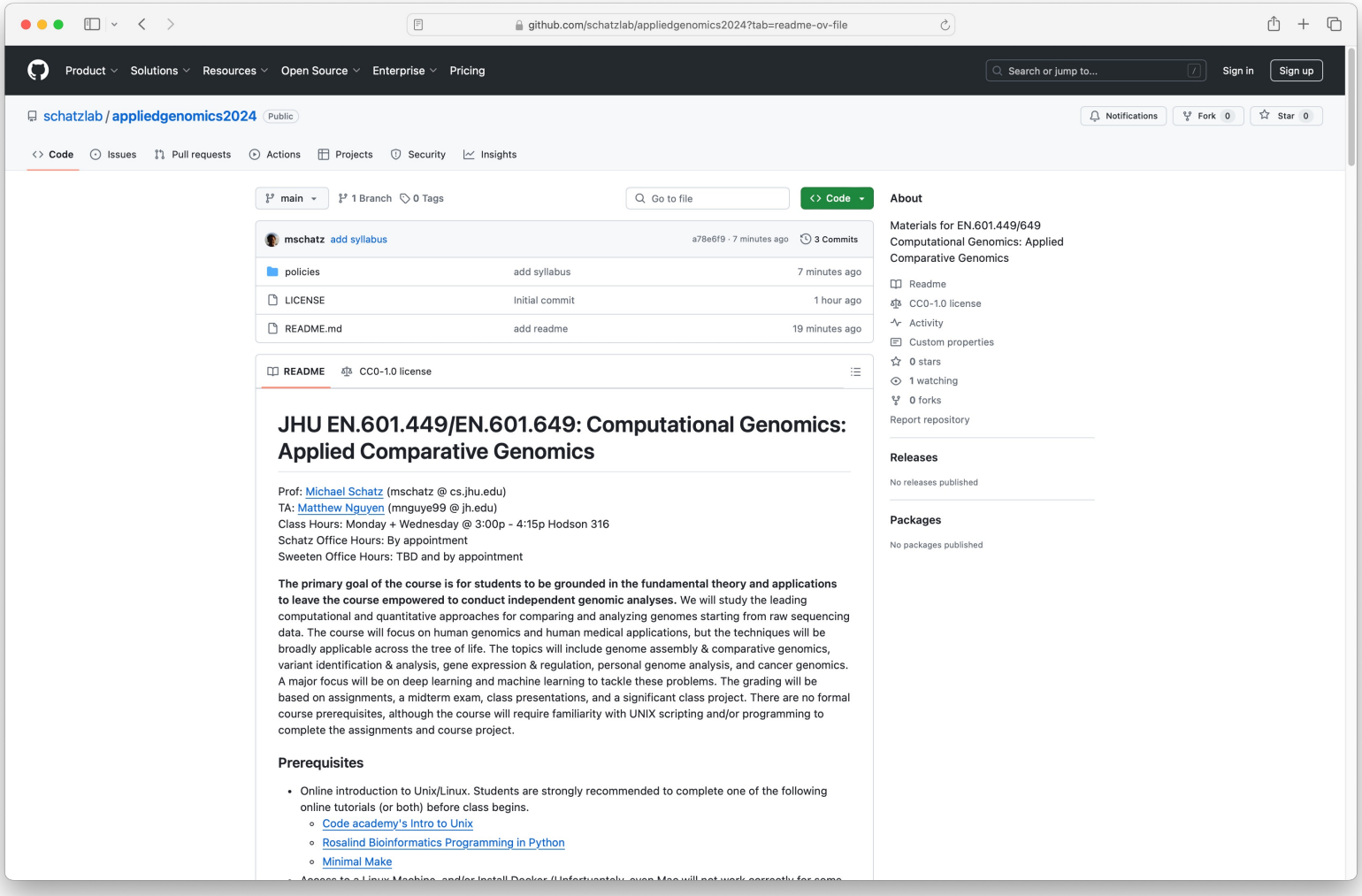
Genome Sequencing

Michael Schatz

August 28, 2024

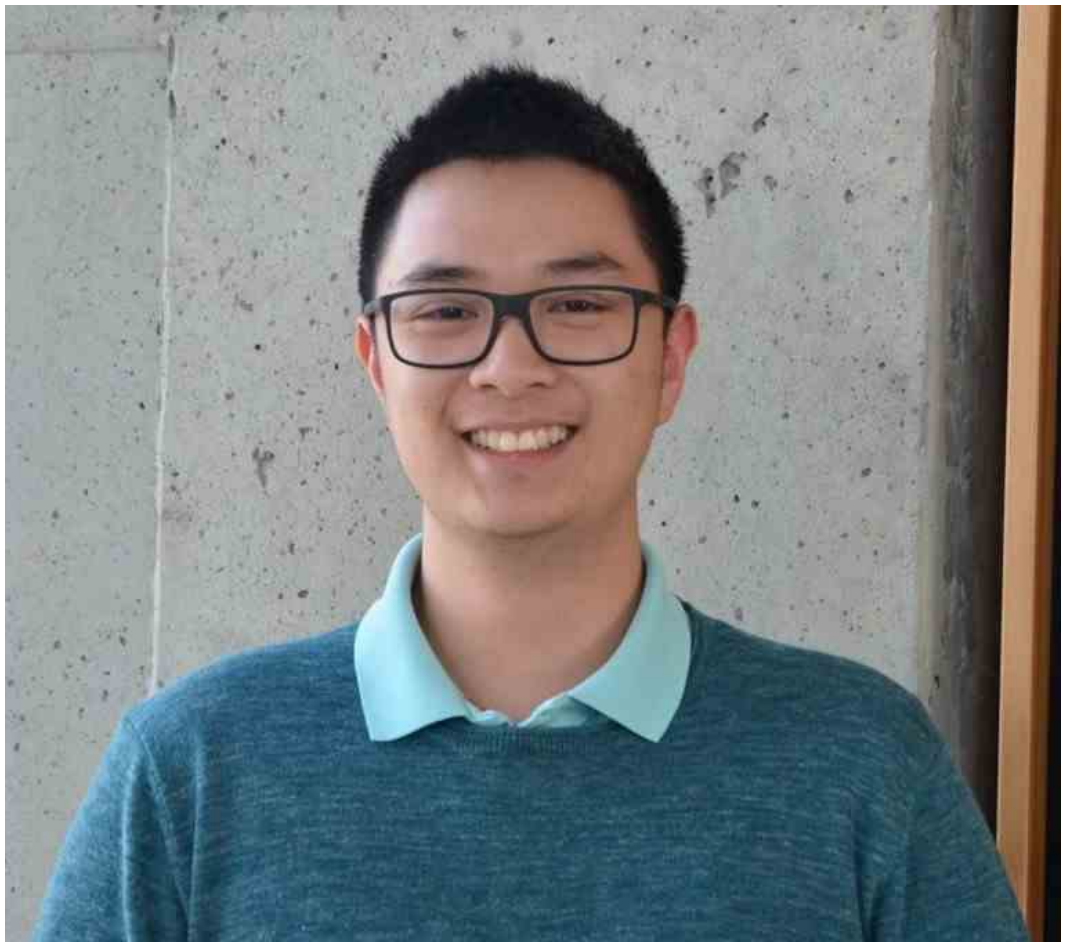
Lecture 2: Applied Comparative Genomics

Course Webpage



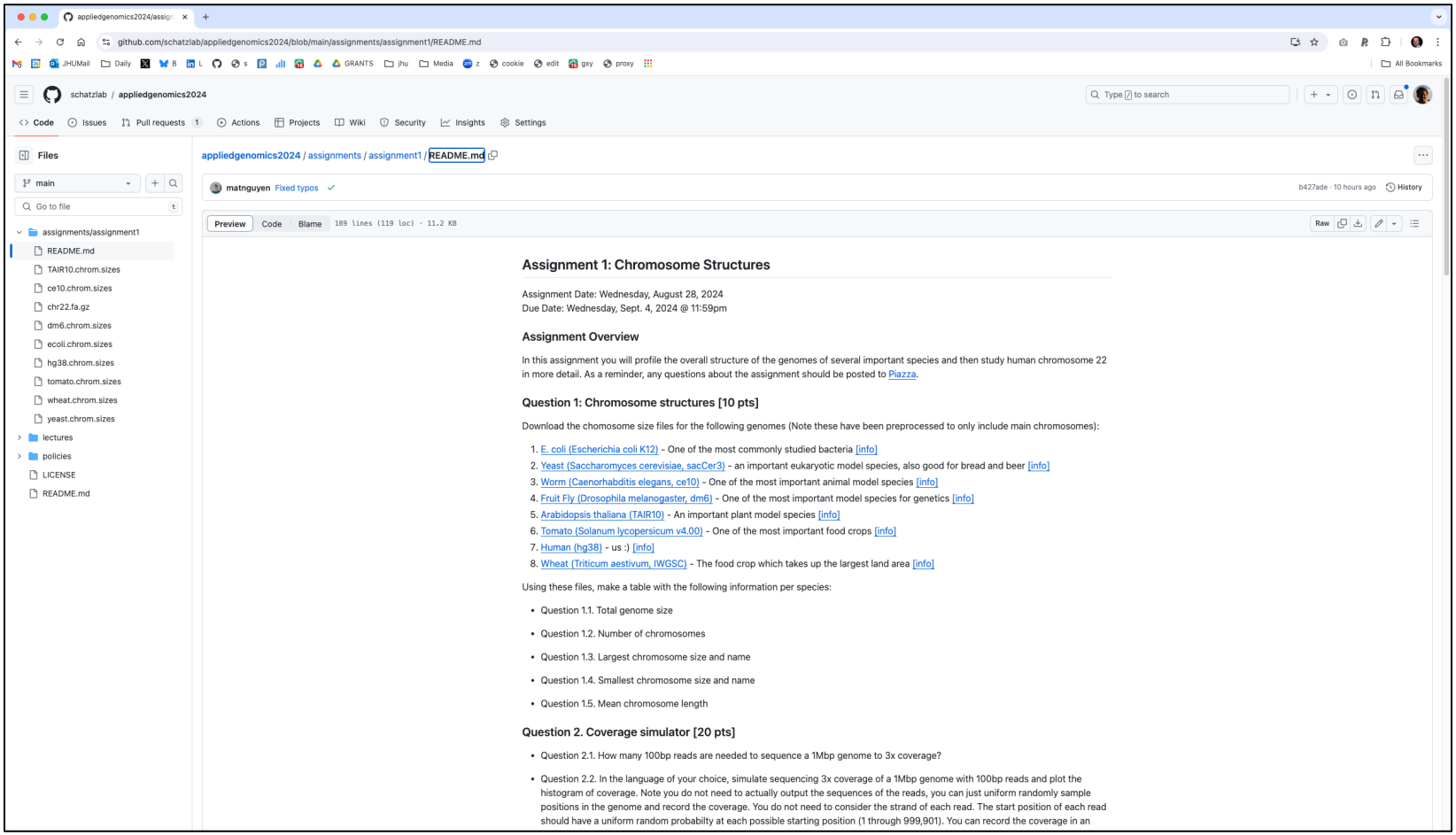
https://github.com/schatzlab/appliedgenomics2024

TA: Matthew Nguyen



Check Piazza for Poll

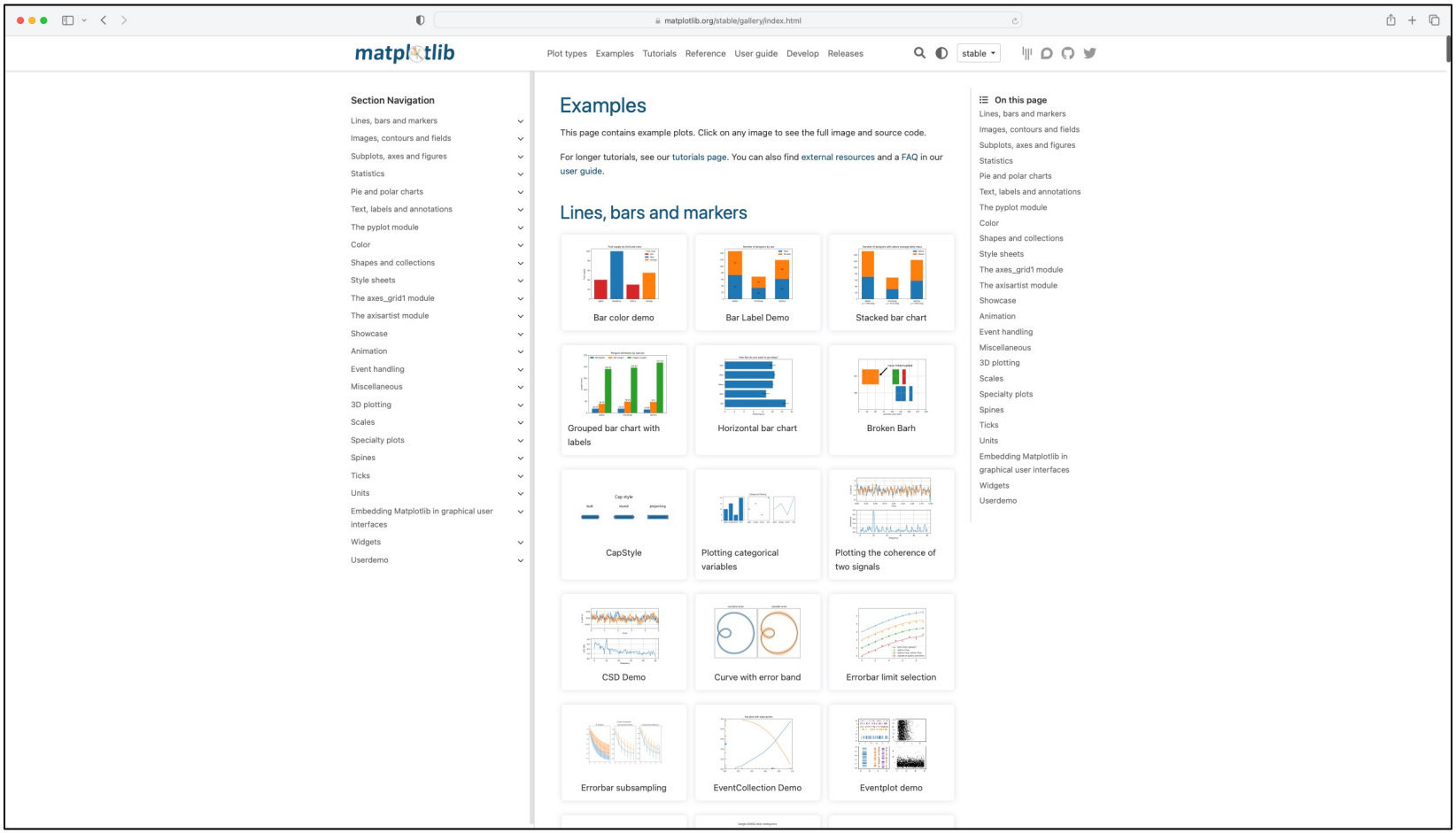
Assignment 1



https://github.com/schatzlab/appliedgenomics2024/tree/main/assignments/assignment1

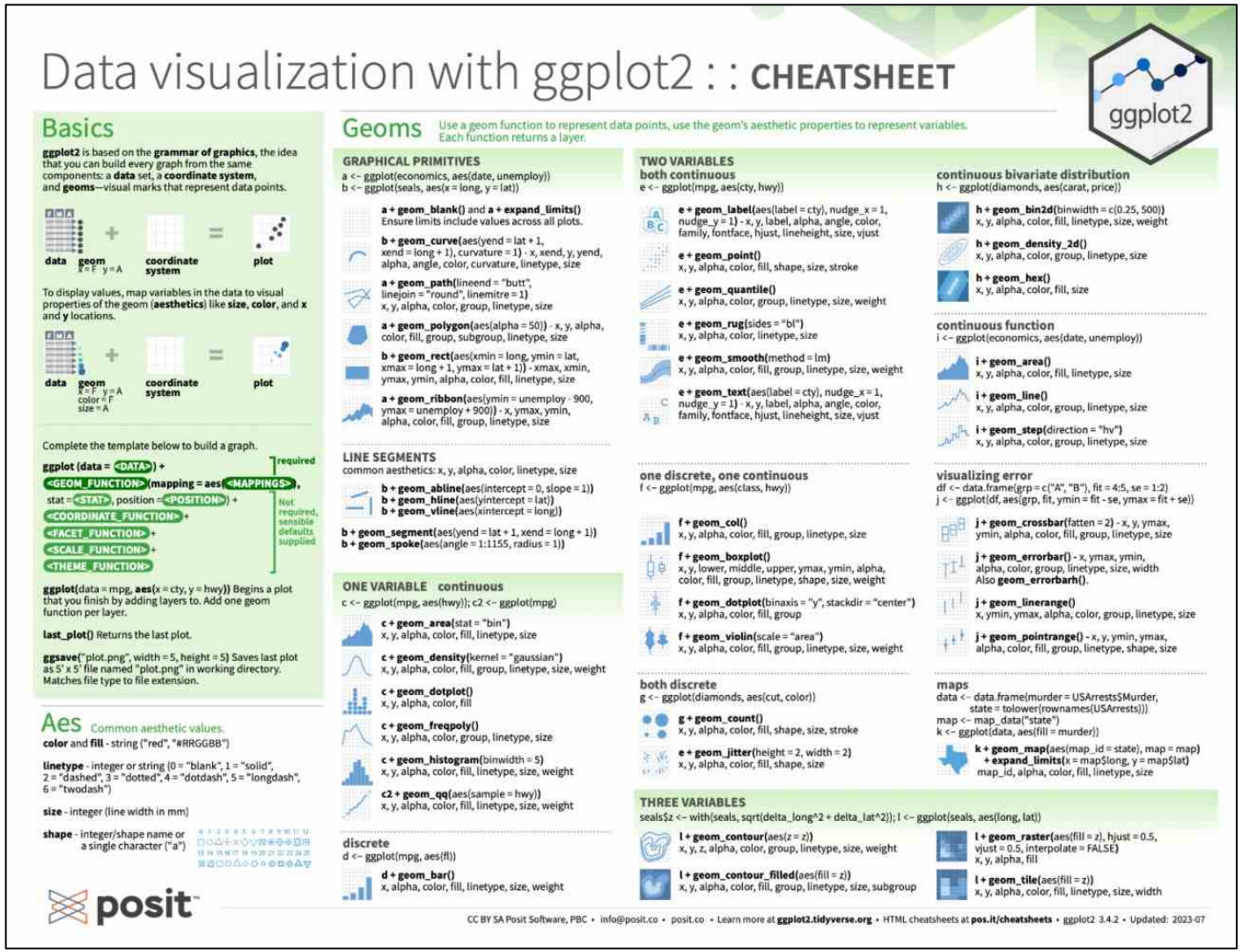
Due end of day on Sept 4 (right before midnight)

Plotting in Python



https://matplotlib.org/

Plotting in R / ggplot2



Data visualization with ggplot2 : : CHEATSHEET

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 (x = F, y = A) = coordinate system + plot

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 (x = F, y = A, color = F, size = A) = coordinate system + plot

Complete the template below to build a graph.

```
ggplot (data = <DATA>) +                                   required
<GEOM_FUNCTION>(mapping = aes(<MAPPINGS>),
 stat = <STAT>, position = <POSITION>) +                   Not required, sensible defaults supplied
<COORDINATE_FUNCTION> +
<FACET_FUNCTION> +
<SCALE_FUNCTION> +
<THEME_FUNCTION>
```

ggplot(data = mpg, **aes**(x = cty, 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.

Aes

Common aesthetic values.

color and **fill** - string ("red", "#RRGGBB")

linetype - integer or string (0 = "blank", 1 = "solid", 2 = "dashed", 3 = "dotted", 4 = "dotdash", 5 = "longdash", 6 = "twodash")

size - integer (line width in mm)

shape - integer/shape name or a single character ("a")

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, xend, y, yend, 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)) - x, ymax, ymin, alpha, color, fill, group, linetype, size

LINE SEGMENTS

common aesthetics: x, y, alpha, color, linetype, size

b + geom_abline(aes(intercept = 0, slope = 1))
b + geom_hline(aes(yintercept = lat))
b + geom_vline(aes(xintercept = long))

b + geom_segment(aes(yend = lat + 1, xend = long + 1))
b + geom_spoke(aes(angle = 1:1155, radius = 1))

ONE VARIABLE continuous

```
c <- ggplot(mpg, aes(hwy)); c2 <- ggplot(mpg)
```

c + geom_area(stat = "bin") x, y, alpha, color, fill, linetype, size

c + geom_density(kernel = "gaussian") x, y, alpha, color, fill, group, linetype, size, weight

c + geom_dotplot() x, y, alpha, color, fill

c + geom_freqpoly() x, y, alpha, color, group, linetype, size

c + geom_histogram(binwidth = 5) x, y, alpha, color, fill, linetype, size, weight

c2 + geom_qq(aes(sample = hwy)) x, y, alpha, color, fill, linetype, size, weight

discrete

```
d <- ggplot(mpg, aes(fl))
```

d + geom_bar() x, alpha, color, fill, linetype, size, weight

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

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(binaxis = "y", stackdir = "center") x, y, alpha, color, fill, group

f + geom_violin(scale = "area") x, y, alpha, color, fill, group, linetype, size, weight

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, 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

both discrete

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

g + geom_count() x, y, alpha, color, fill, shape, size, stroke

e + geom_jitter(height = 2, width = 2) x, y, alpha, color, fill, shape, size

maps

```
data <- data.frame(murder = USArrests$Murder,
                   state = tolower(rownames(USArrests)))
map <- map_data("state")
k <- ggplot(data, aes(fill = murder))
```

k + geom_map(aes(map_id = state), map = map) **+ expand_limits**(x = map$long, y = map$lat) map_id, alpha, color, fill, linetype, size

THREE VARIABLES

```
seals$z <- with(seals, sqrt(delta_long^2 + delta_lat^2)); l <- ggplot(seals, aes(long, lat))
```

l + geom_contour(aes(z = z)) x, y, z, alpha, color, group, linetype, size, weight

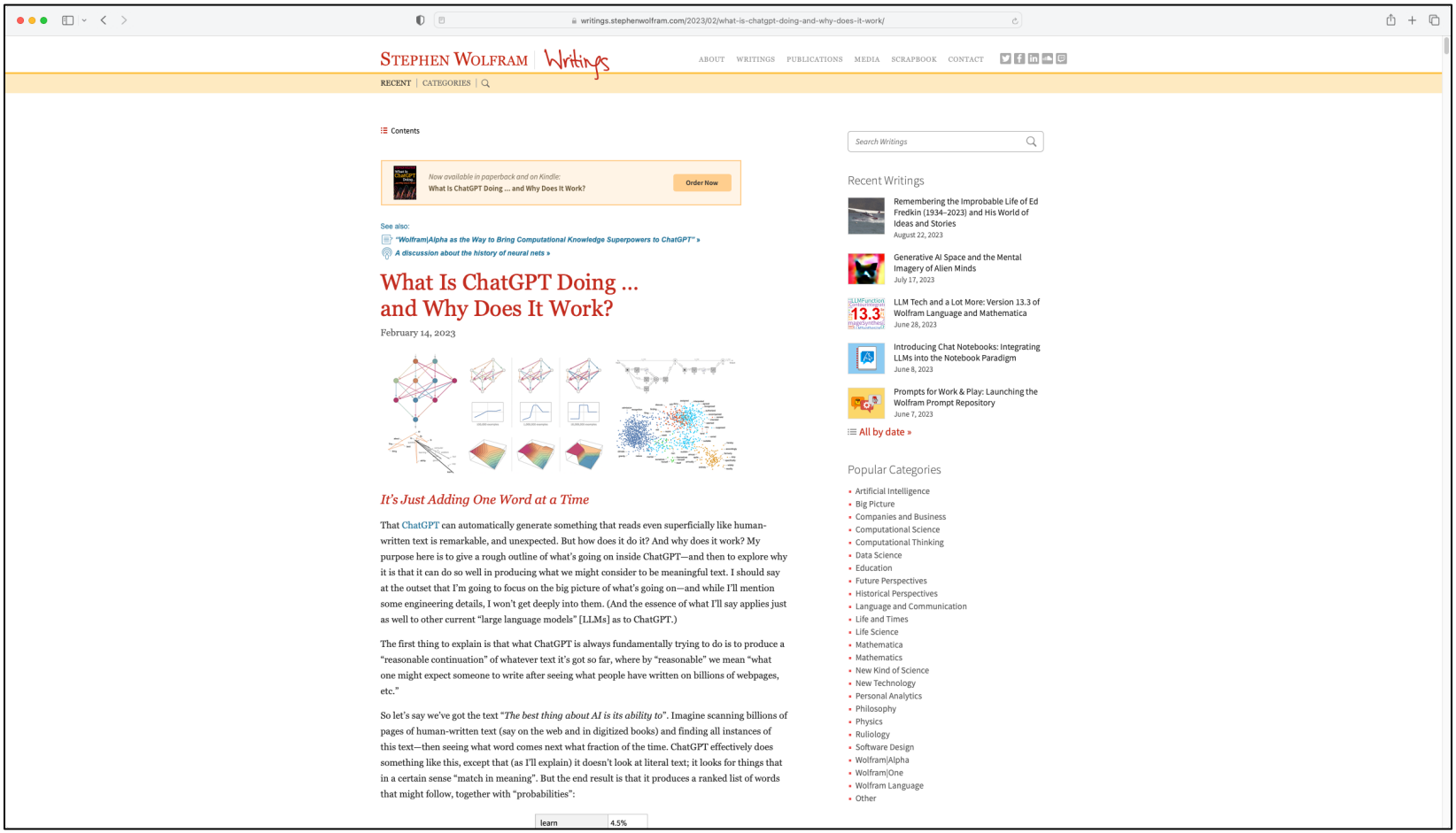
l + geom_contour_filled(aes(fill = z)) x, y, alpha, color, fill, group, linetype, size, subgroup

l + geom_raster(aes(fill = z), hjust = 0.5, vjust = 0.5, interpolate = FALSE) x, y, alpha, fill

l + geom_tile(aes(fill = z)) x, y, alpha, color, fill, linetype, size, width

https://ggplot2.tidyverse.org/

What is ChatGPT and Why Does it Work?



writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

STEPHEN WOLFRAM | Writings

ABOUT WRITINGS PUBLICATIONS MEDIA SCRAPBOOK CONTACT

RECENT | CATEGORIES |

Contents

Now available in paperback and on Kindle:
What Is ChatGPT Doing ... and Why Does It Work?
Order Now

See also:
"Wolfram|Alpha as the Way to Bring Computational Knowledge Superpowers to ChatGPT" »
A discussion about the history of neural nets »

What Is ChatGPT Doing ... and Why Does It Work?

February 14, 2023

It's Just Adding One Word at a Time

That ChatGPT can automatically generate something that reads even superficially like human-written text is remarkable, and unexpected. But how does it do it? And why does it work? My purpose here is to give a rough outline of what's going on inside ChatGPT—and then to explore why it is that it can do so well in producing what we might consider to be meaningful text. I should say at the outset that I'm going to focus on the big picture of what's going on—and while I'll mention some engineering details, I won't get deeply into them. (And the essence of what I'll say applies just as well to other current "large language models" [LLMs] as to ChatGPT.)

The first thing to explain is that what ChatGPT is always fundamentally trying to do is to produce a "reasonable continuation" of whatever text it's got so far, where by "reasonable" we mean "what one might expect someone to write after seeing what people have written on billions of webpages, etc."

So let's say we've got the text "*The best thing about AI is its ability to*". Imagine scanning billions of pages of human-written text (say on the web and in digitized books) and finding all instances of this text—then seeing what word comes next what fraction of the time. ChatGPT effectively does something like this, except that (as I'll explain) it doesn't look at literal text; it looks for things that in a certain sense "match in meaning". But the end result is that it produces a ranked list of words that might follow, together with "probabilities":

learn	4.5%

Search Writings

Recent Writings

- Remembering the Improbable Life of Ed Fredkin (1934–2023) and His World of Ideas and Stories — August 22, 2023
- Generative AI Space and the Mental Imagery of Alien Minds — July 17, 2023
- LLM Tech and a Lot More: Version 13.3 of Wolfram Language and Mathematica — June 28, 2023
- Introducing Chat Notebooks: Integrating LLMs into the Notebook Paradigm — June 8, 2023
- Prompts for Work & Play: Launching the Wolfram Prompt Repository — June 7, 2023

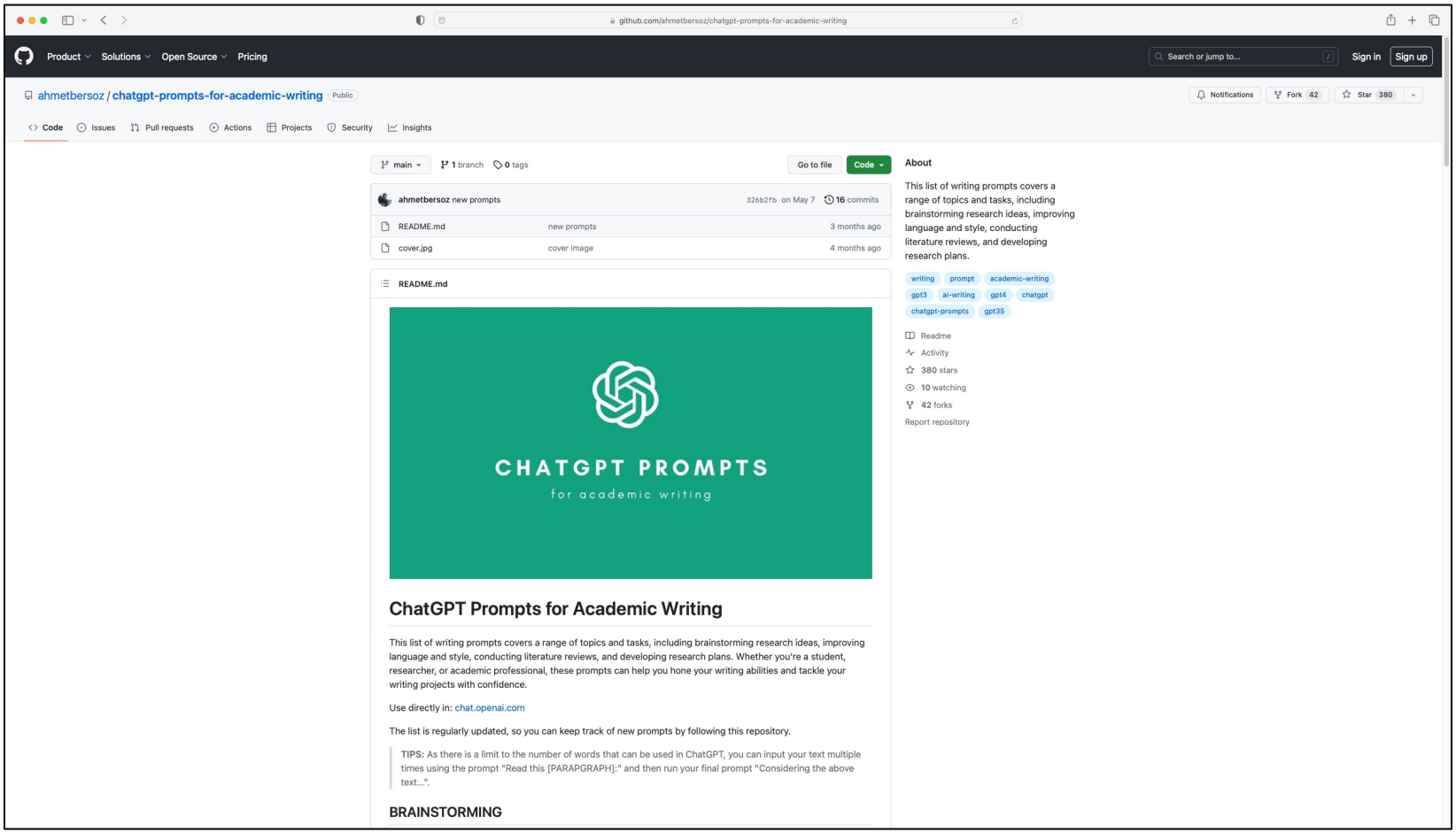
All by date »

Popular Categories

- Artificial Intelligence
- Big Picture
- Companies and Business
- Computational Science
- Computational Thinking
- Data Science
- Education
- Future Perspectives
- Historical Perspectives
- Language and Communication
- Life and Times
- Life Science
- Mathematica
- Mathematics
- New Kind of Science
- New Technology
- Personal Analytics
- Philosophy
- Physics
- Ruliology
- Software Design
- Wolfram|Alpha
- Wolfram|One
- Wolfram Language
- Other

https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

ChatGPT Prompts for Academic Writing



https://github.com/ahmetbersoz/chatgpt-prompts-for-academic-writing

Discovery of the Double Helix

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

[1] Young, F. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).
[2] Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **5**, 285 (1949).
[3] Von Arx, W. S., Woods Hole Papers in Phys. Oceanog. Meteor., **11** (3) (1956).
[4] Ekman, V. W., *Arkiv. Mat. Astron. Fysik. (Stockholm)*, **2** (11) (1905).

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

This figure is purely diagrammatic. The two ribbons symbolize the two phosphate–sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

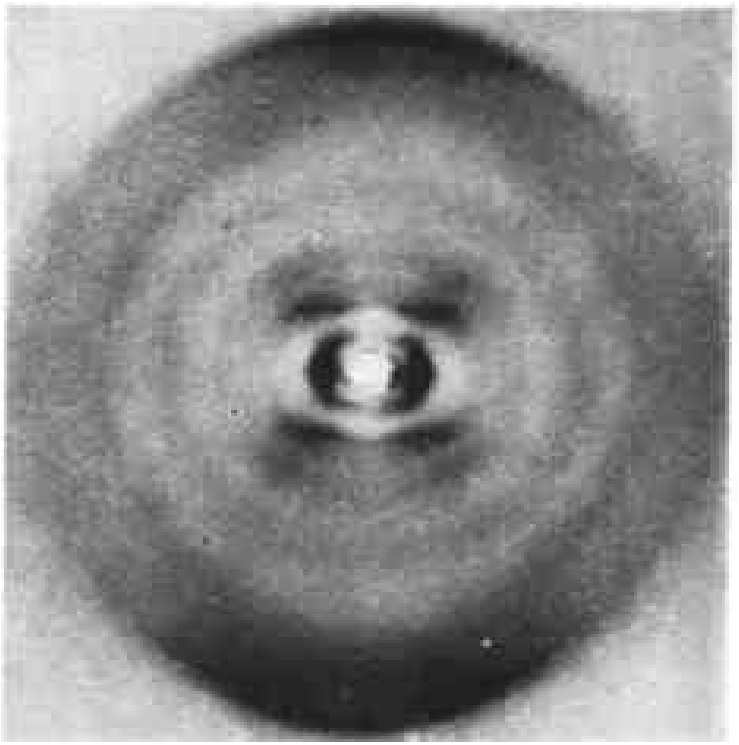
It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at



It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the con-

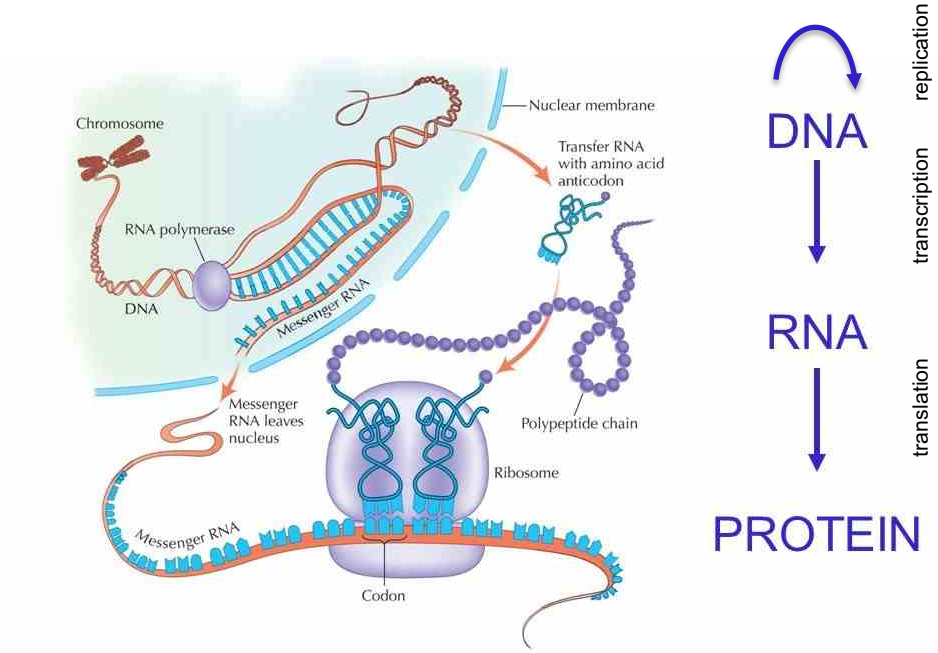
Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid

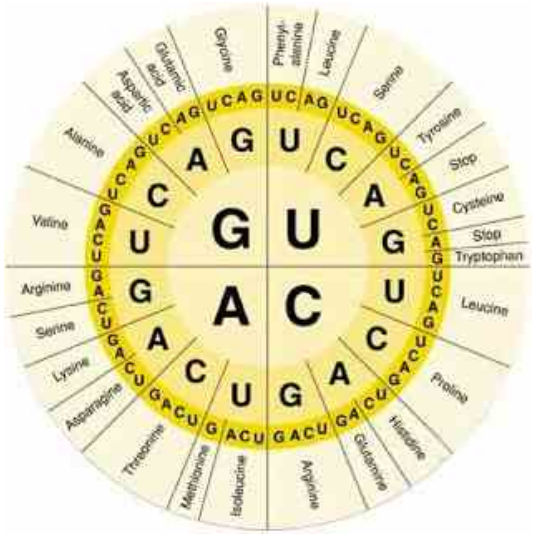
Watson JD, Crick FH (1953). *Nature* 171: 737–738.

Nobel Prize in Physiology or Medicine in 1962

Central Dogma of Molecular Biology

"Once 'information' has passed into protein it cannot get out again. In more detail, the transfer of information ***from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible***, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein"

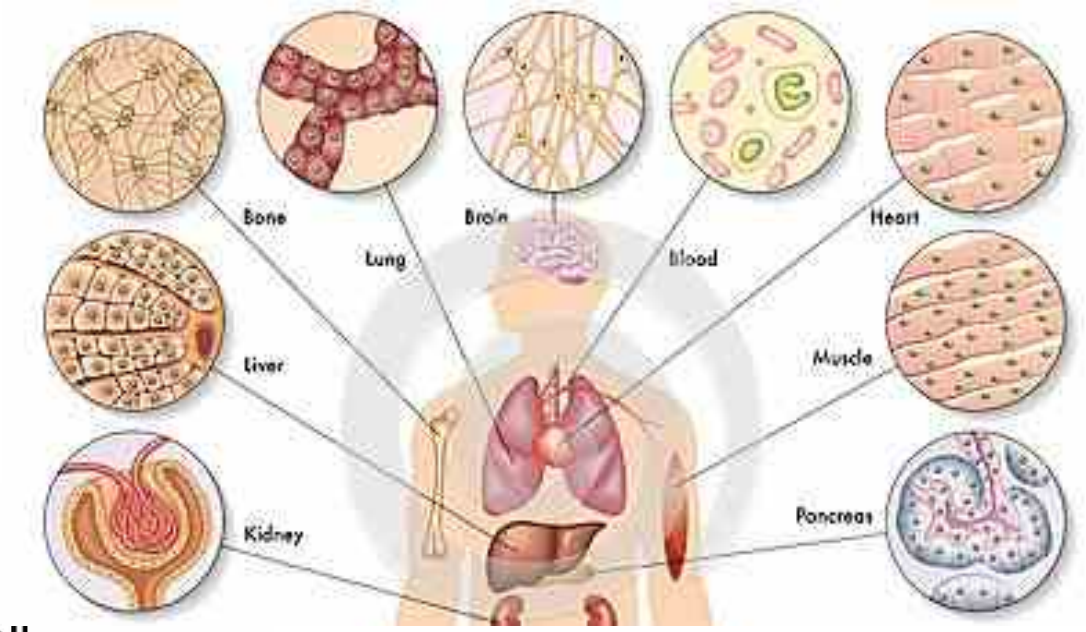


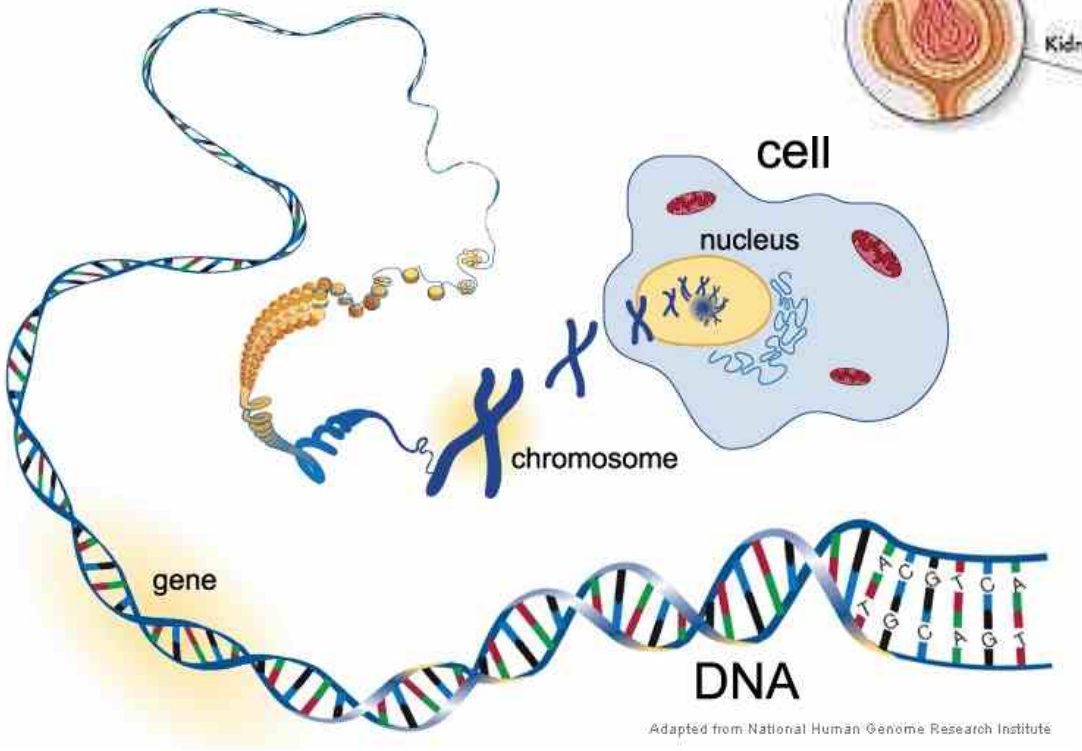


On Protein Synthesis
Crick, F.H.C. (1958). Symposia of the Society for Experimental Biology pp. 138–163.

One Genome, Many Cell Types

Each cell of your body contains a nearly exact copy of your 3 billion base pair genome.





Your body has a few hundred (thousands?) major cell types, largely defined by the gene expression patterns

Unsolved Questions in Biology

- What is your genome sequence?
- How does your genome compare to my genome?

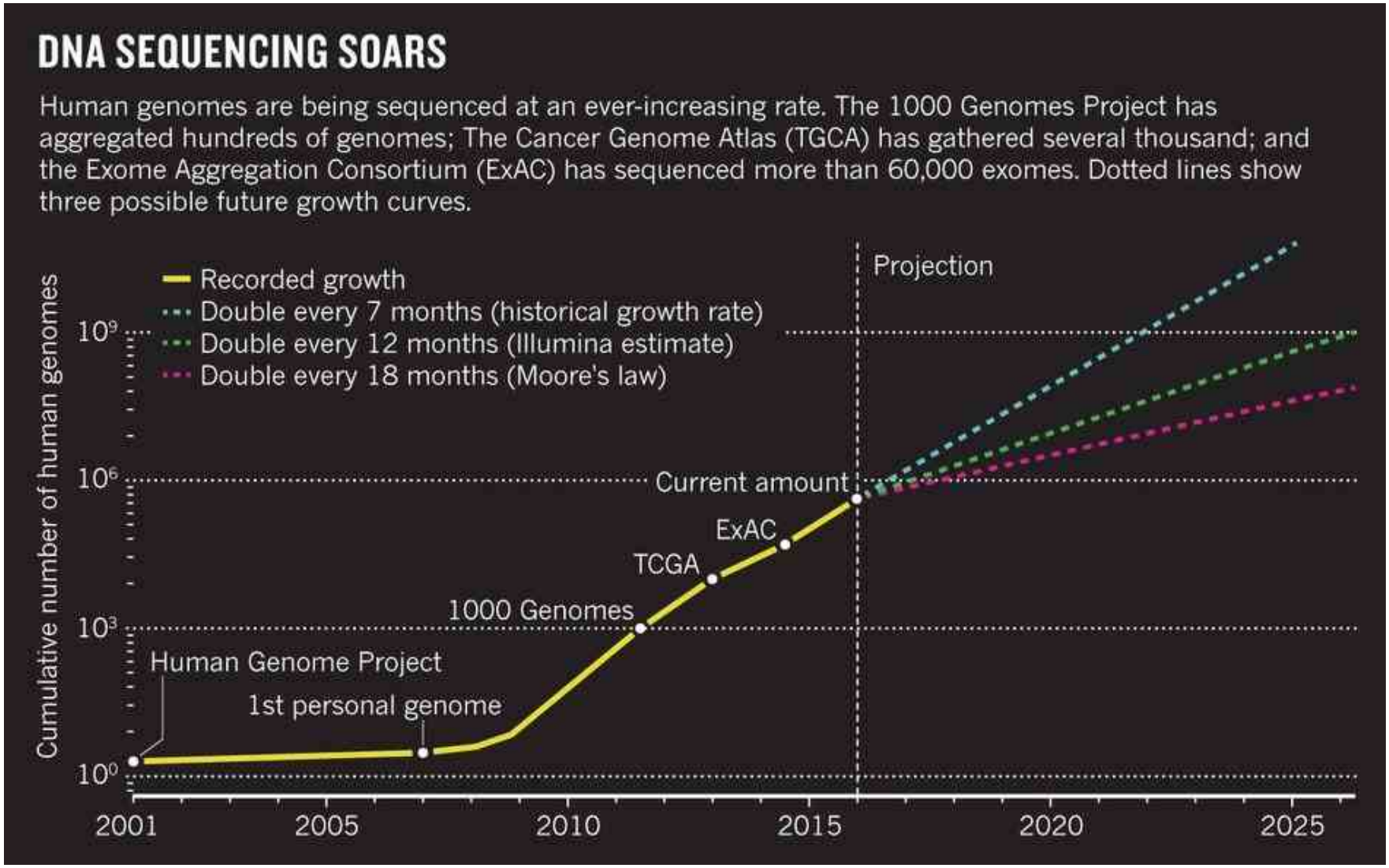
- Where are the genes and how active are they?
- How does gene activity change during development?
- How does splicing change during development?

- How does methylation change during development?
- How does chromatin change during development?
- How does is your genome folded in the cell?
- Where do proteins bind and regulate genes?

- What virus and microbes are living inside you?
- How do your mutations relate to disease?
- What drugs and treatments should we give you?

- ***Plus thousands and thousands more***

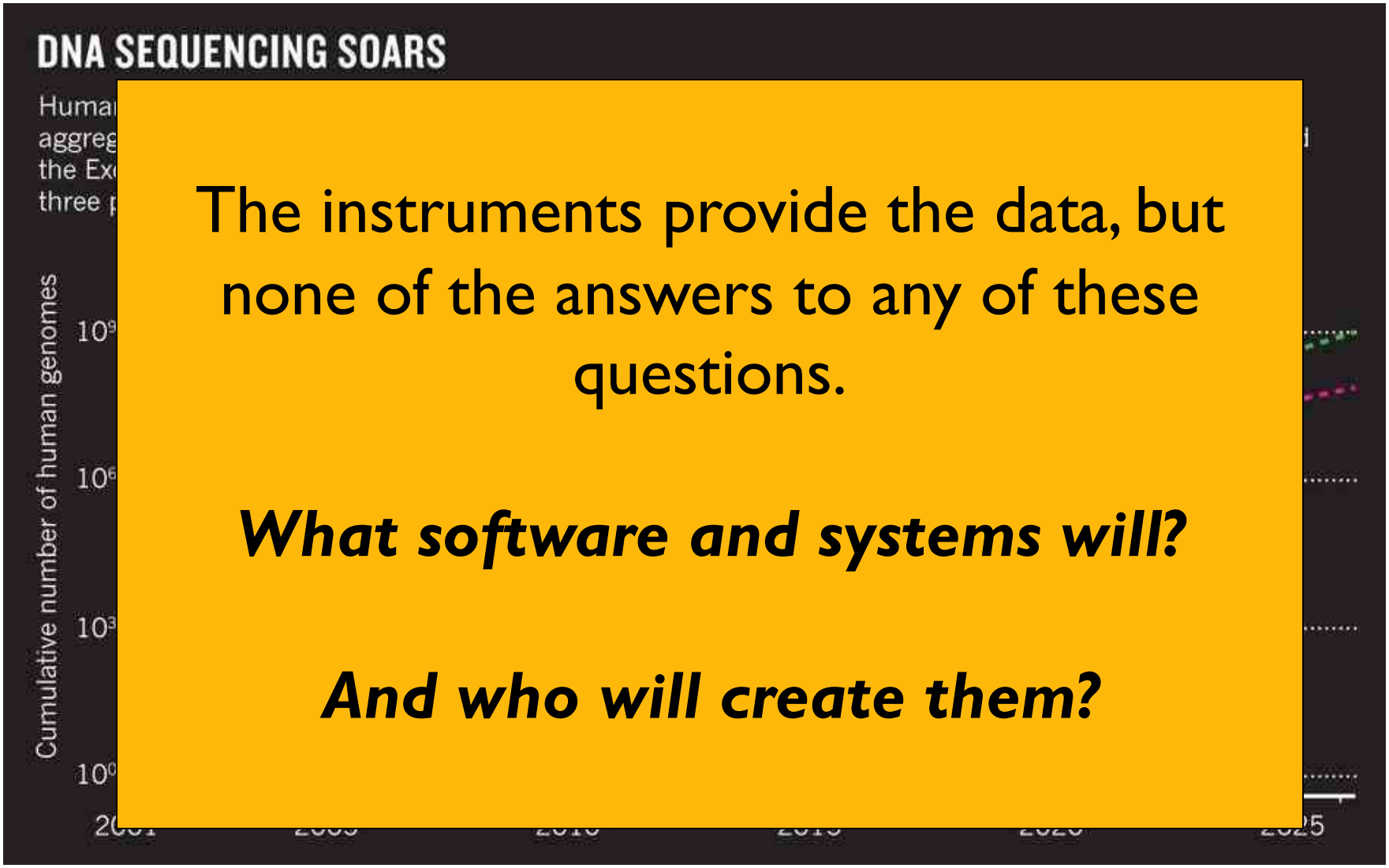
Sequencing Capacity



Big Data: Astronomical or Genomical?

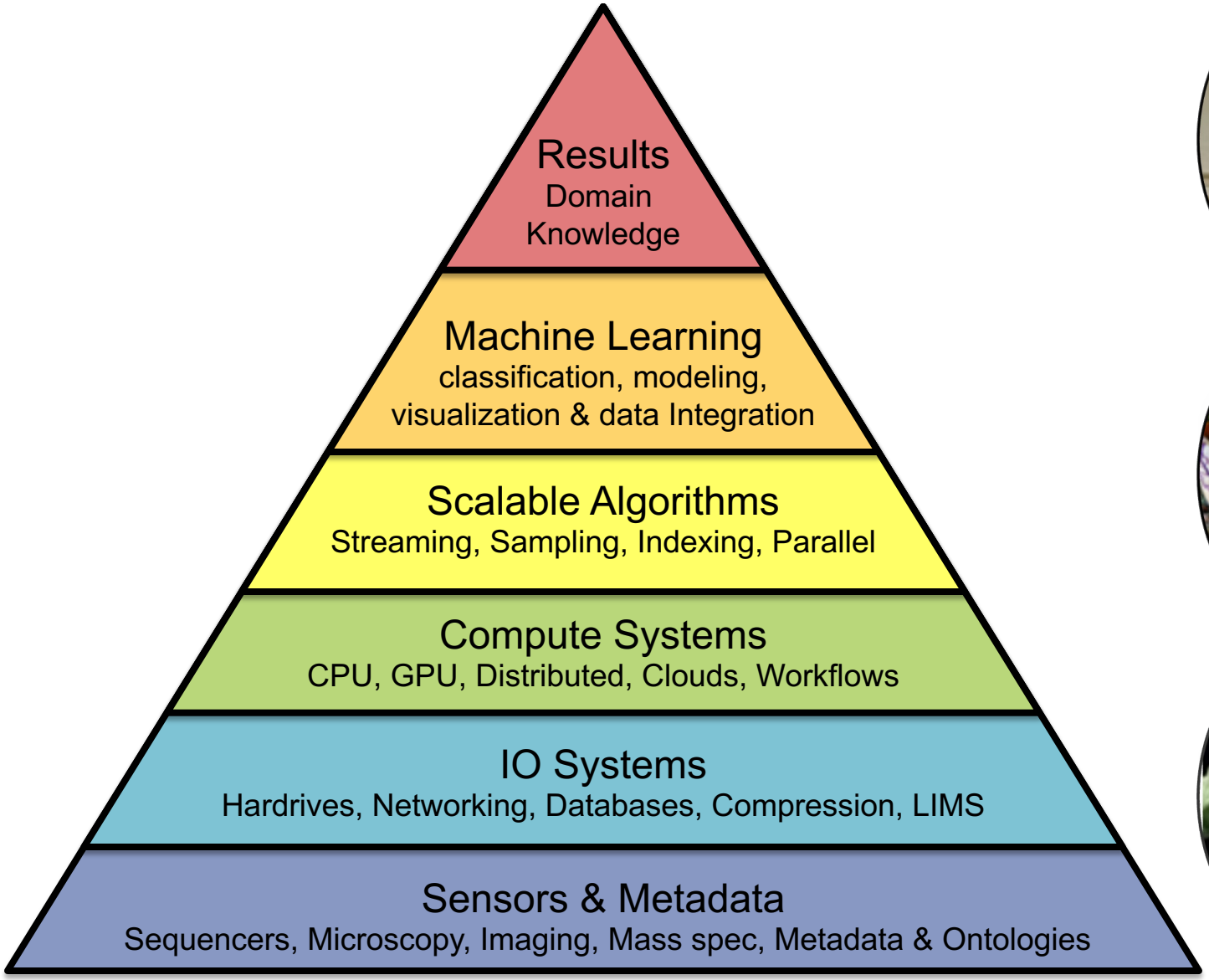
Stephens, Z, et al. (2015) PLOS Biology DOI: 10.1371/journal.pbio.1002195

Sequencing Capacity

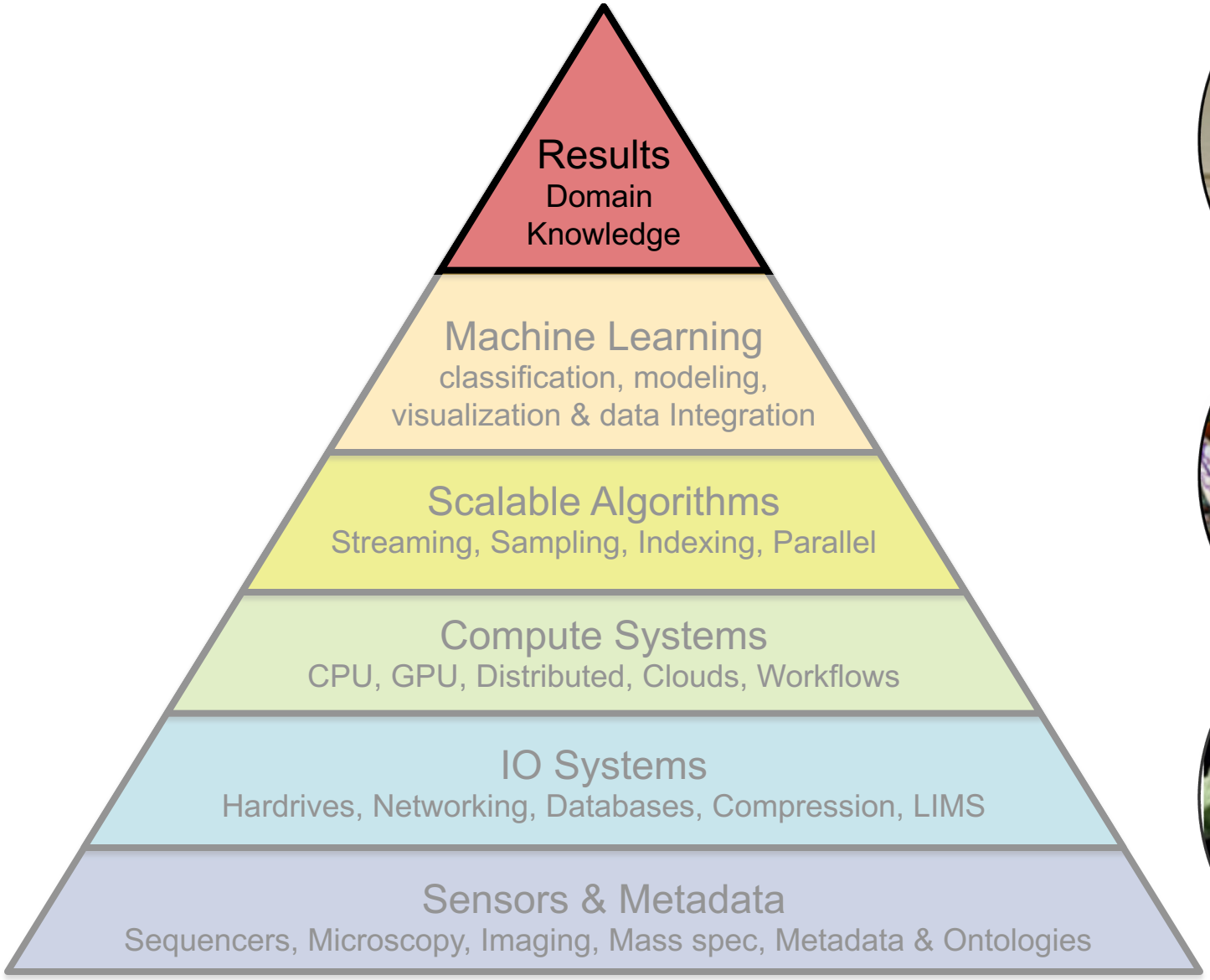


Big Data: Astronomical or Genomical?
Stephens, Z, et al. (2015) PLOS Biology DOI: 10.1371/journal.pbio.1002195

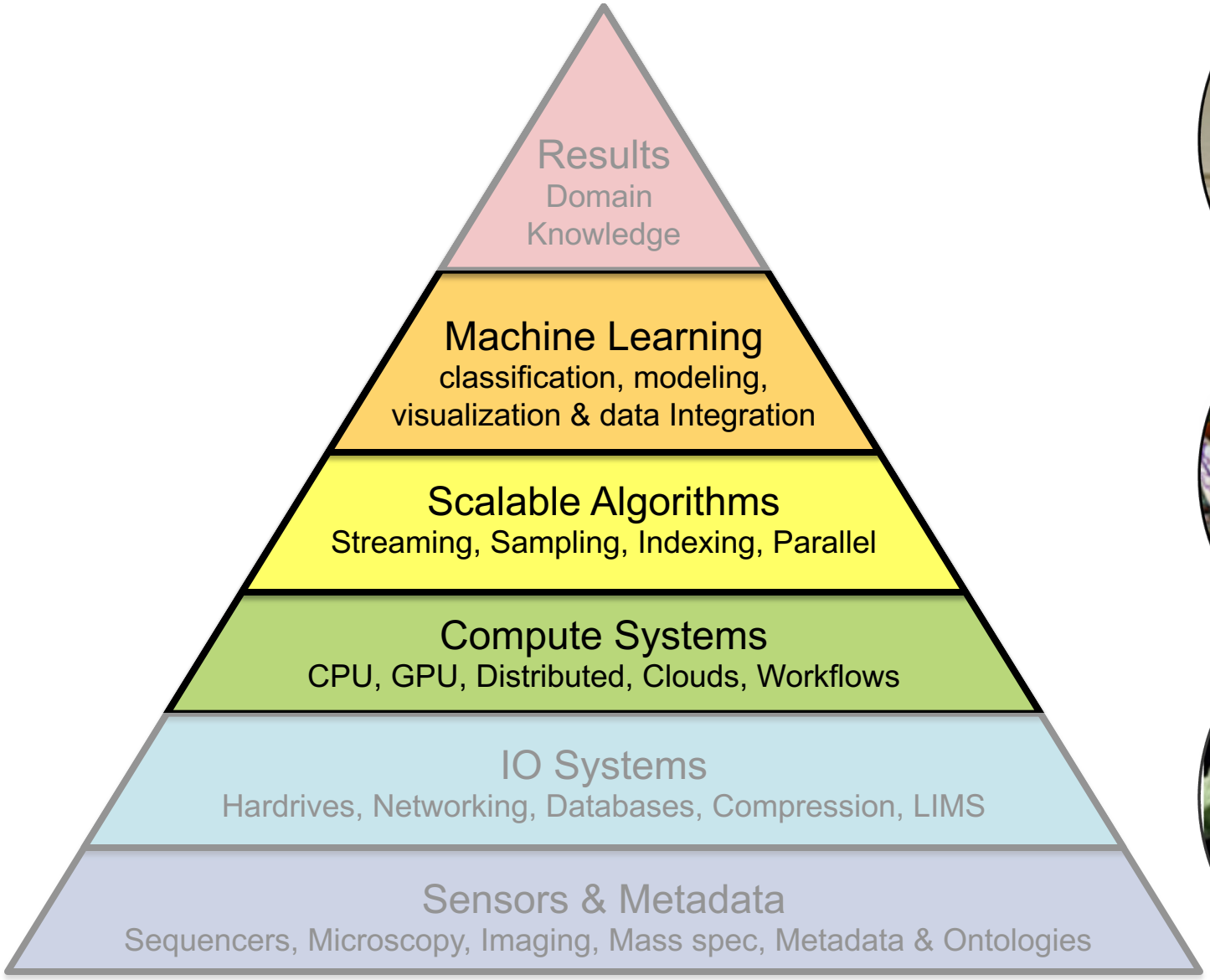
Comparative Genomics Technologies



Comparative Genomics Technologies



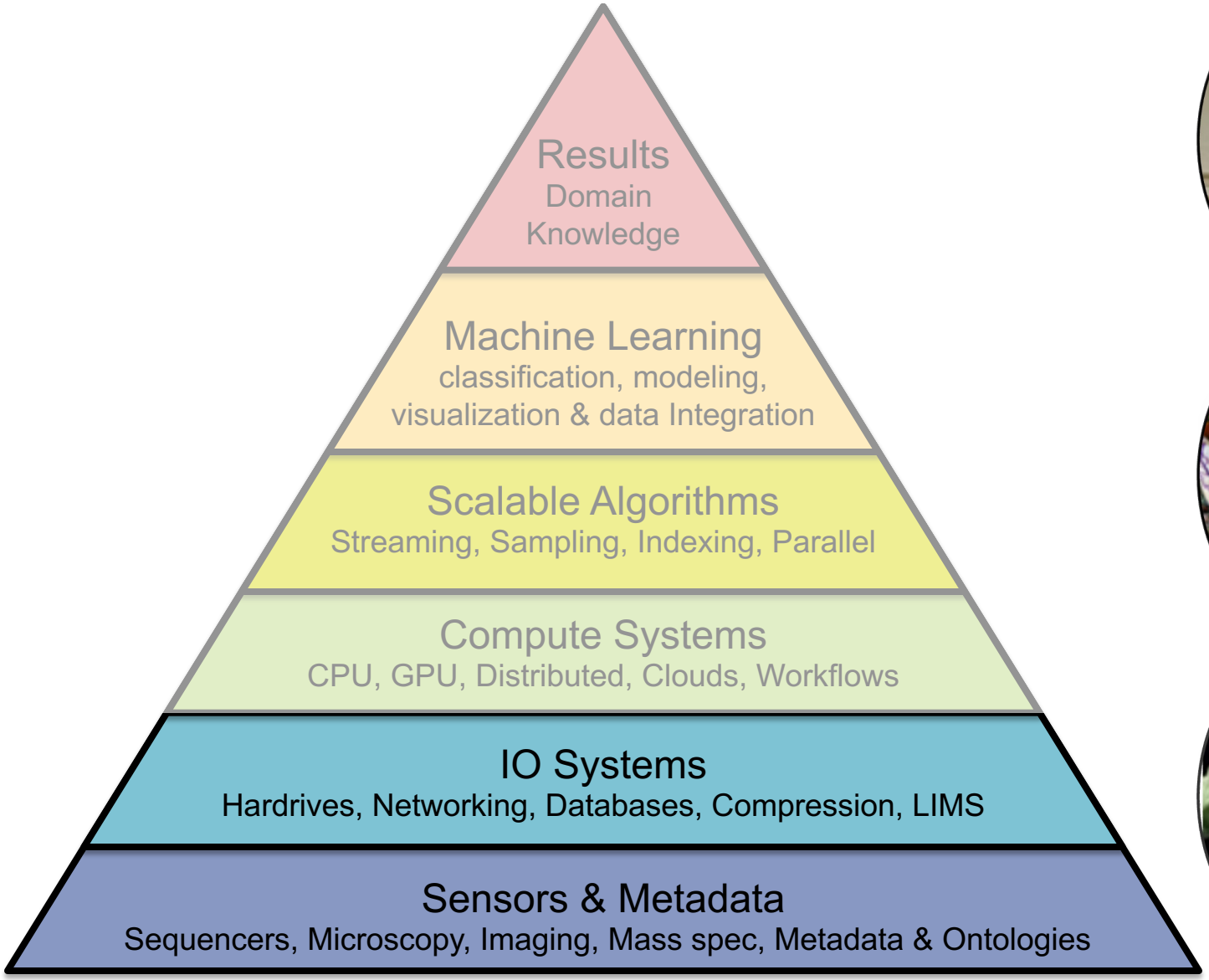
Comparative Genomics Technologies



Selected Topics

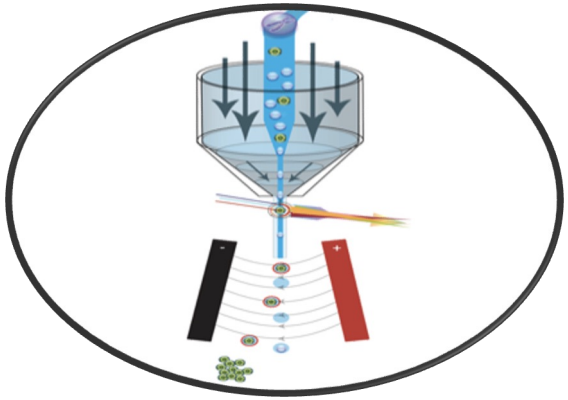
- Genome assembly, whole genome alignment
- Full text indexing: Suffix Trees, Suffix Arrays, FM-index
- Dynamic Programming: Edit Distance, sequence similarity
- Read mapping & Variant identification
- Gene Finding: HMMs, Plane-sweep algorithms
- RNA-seq: mapping, assembly, quantification
- ChIP-seq: Peak finding, motif finding
- Methylation-seq: Mapping, CpG island detection
- HiC: Domain identification, scaffolding
- Chromatin state analysis: ChromHMM
- Scalable genomics: Cloud computing, scalable data structures
- Population & single cell analysis: clustering, pseudotime
- Disease analysis, cancer genomics, Metagenomics
- Deep learning in genomics

Comparative Genomics Technologies



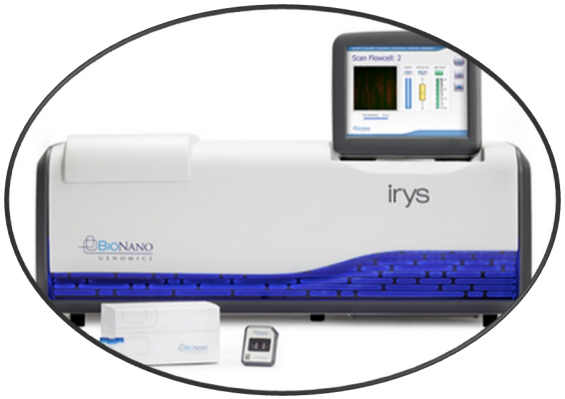
Genomics Arsenal in the year 2024

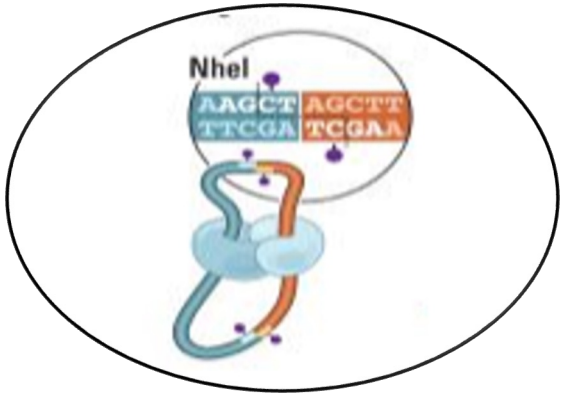
Sample Preparation

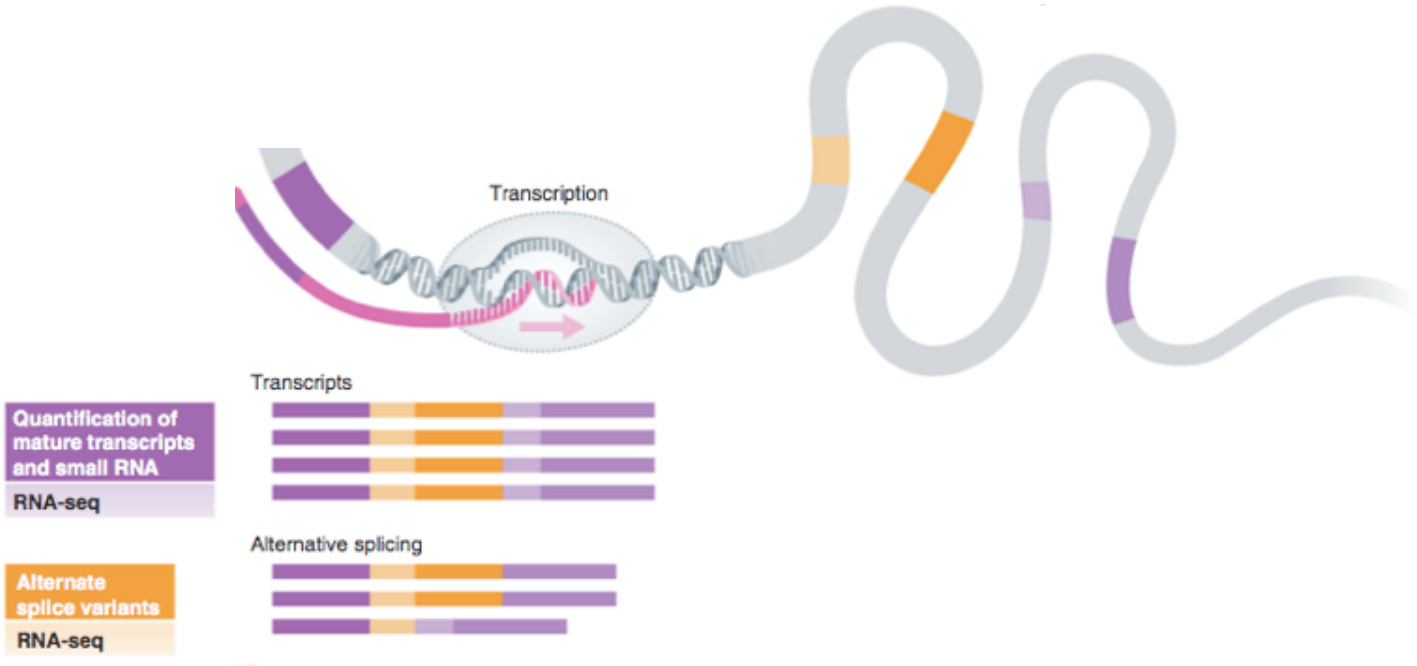


Sequencing

Chromosome Mapping

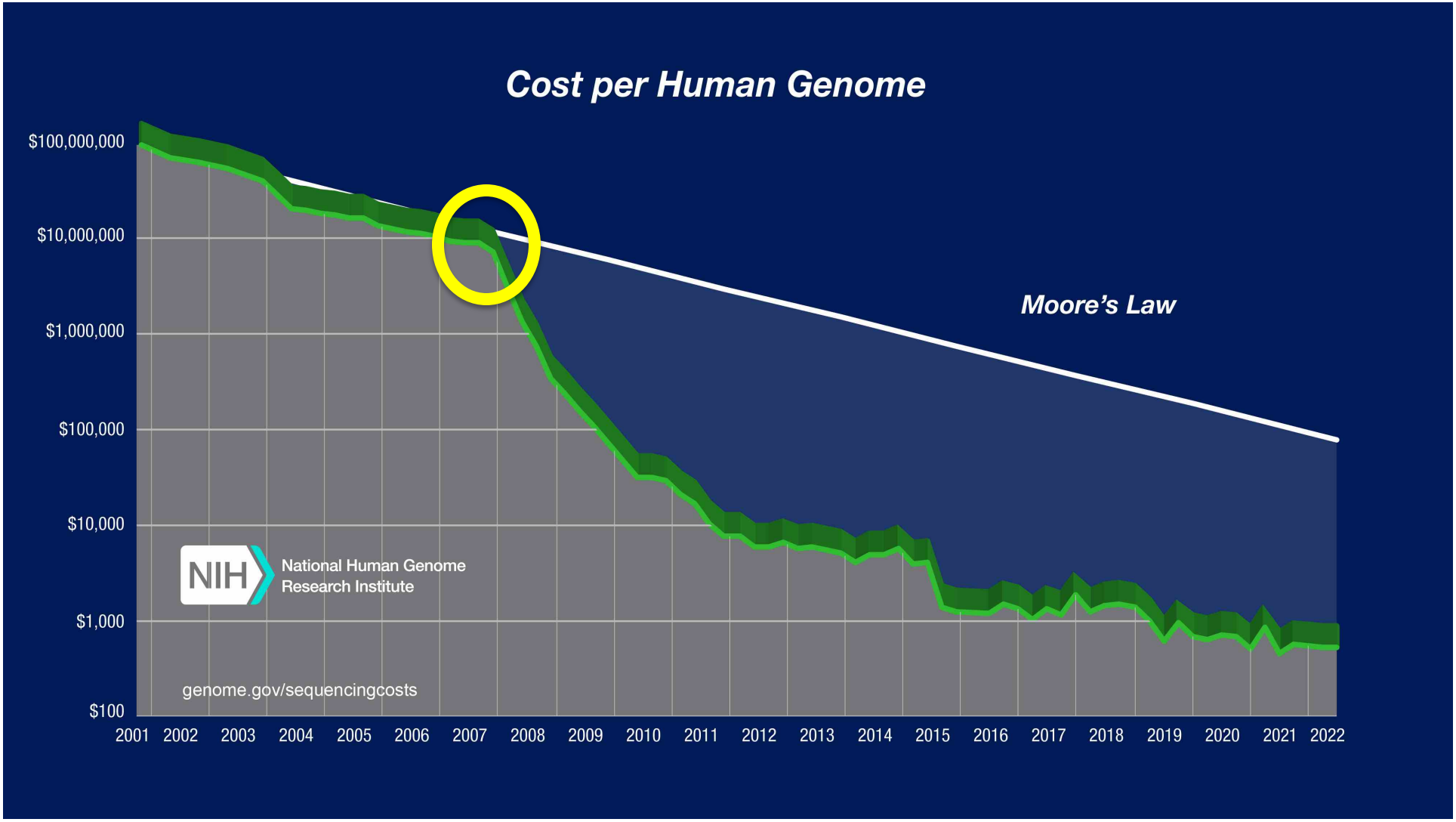






Soon et al., Molecular Systems Biology, 2013

Cost per Genome

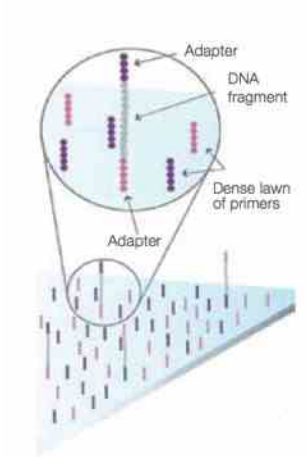


https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

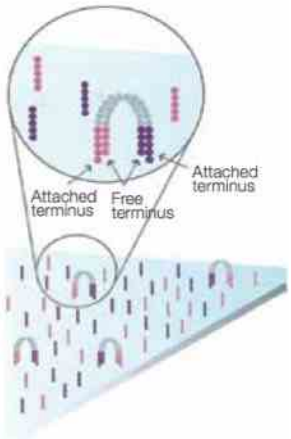
Second Generation Sequencing

Illumina NovaSeq 6000
Sequencing by Synthesis

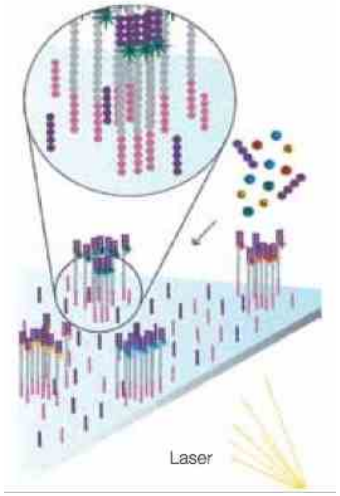
>3Tbp / day
(JHU has 4 of these!)



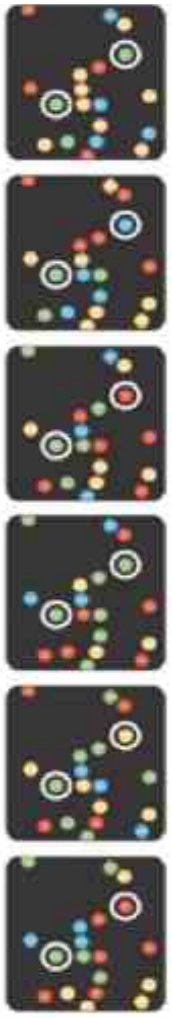
1. Attach



2. Amplify

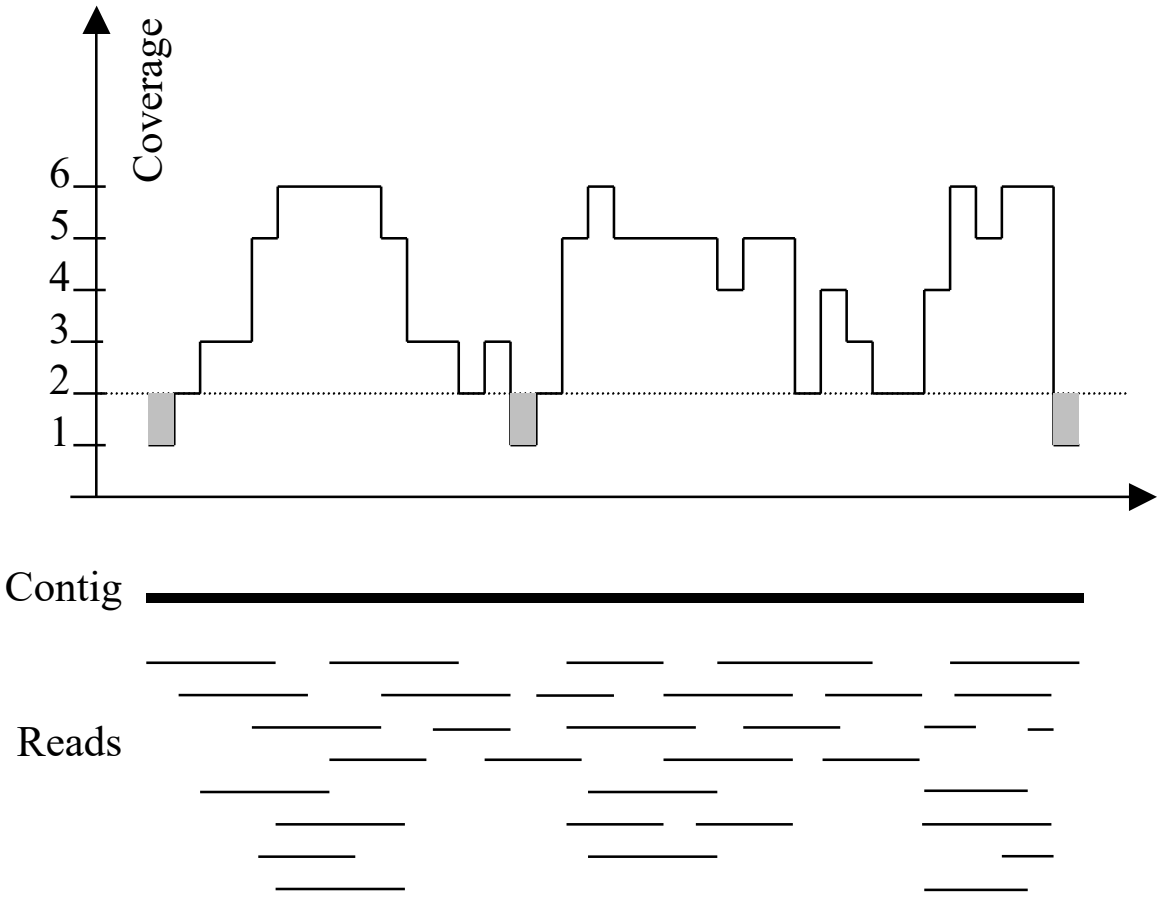


3. Image



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

Typical sequencing coverage

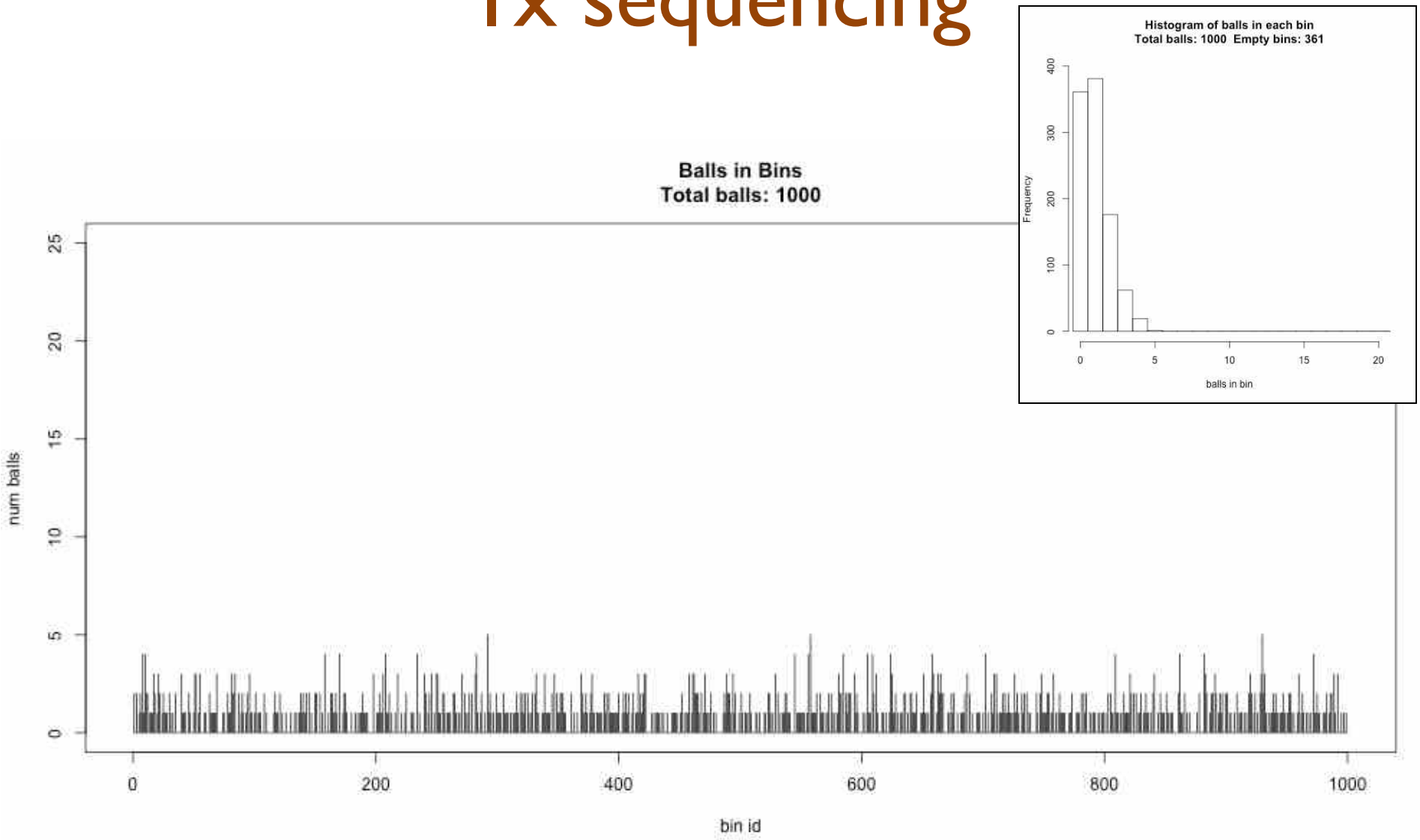


Imagine raindrops on a sidewalk

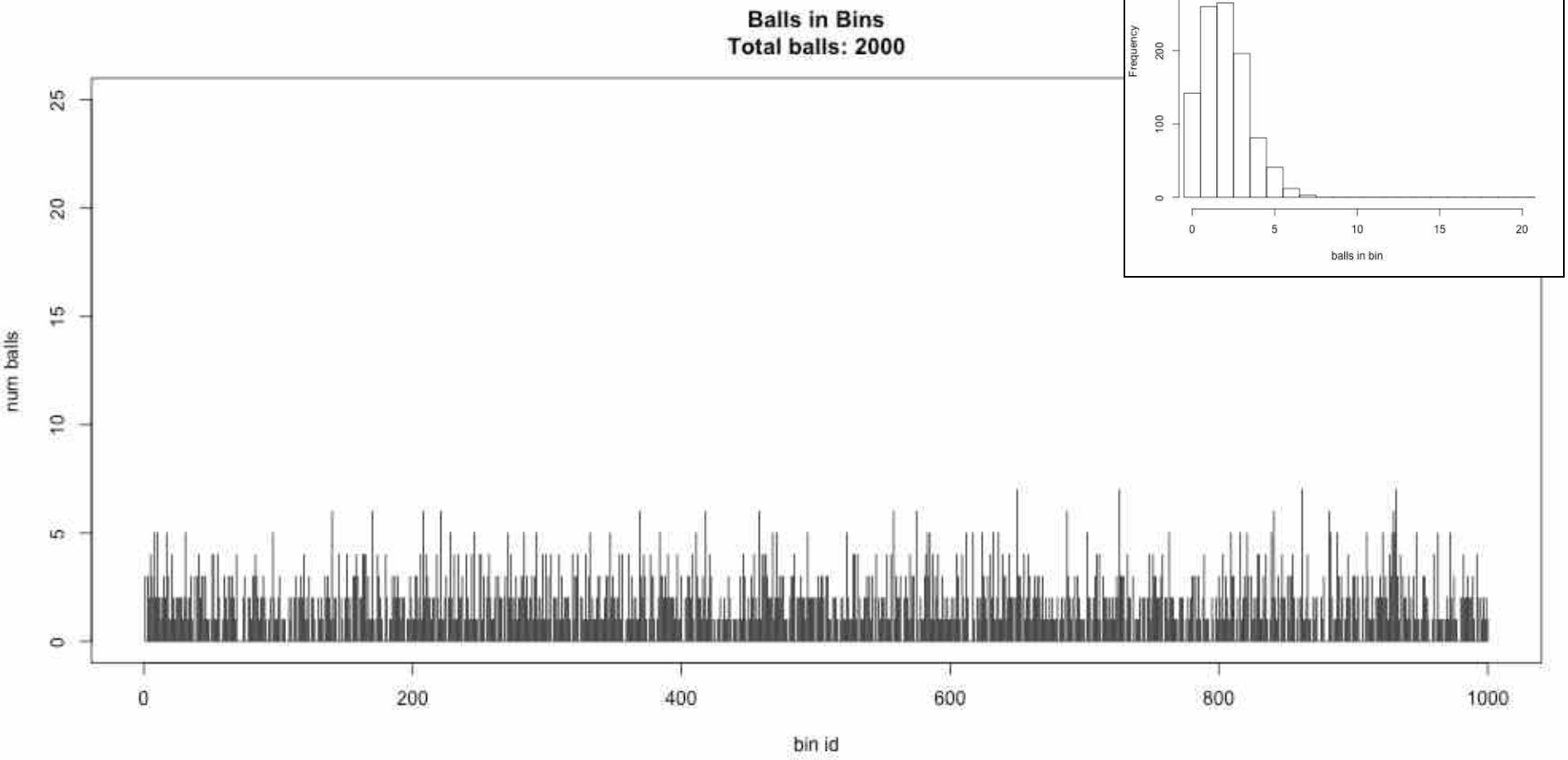
We want to cover the entire sidewalk but each drop costs $1

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

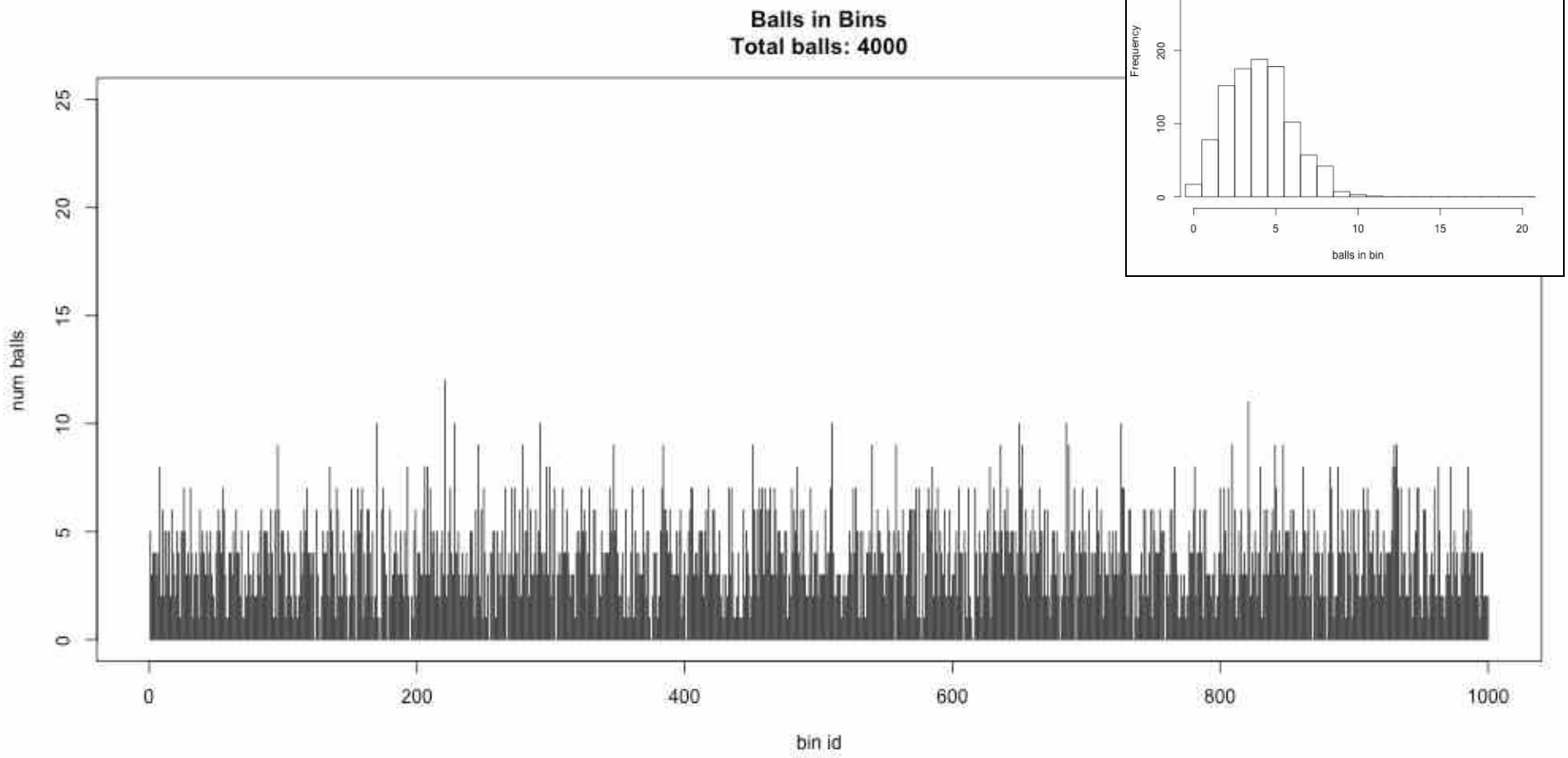
1x sequencing



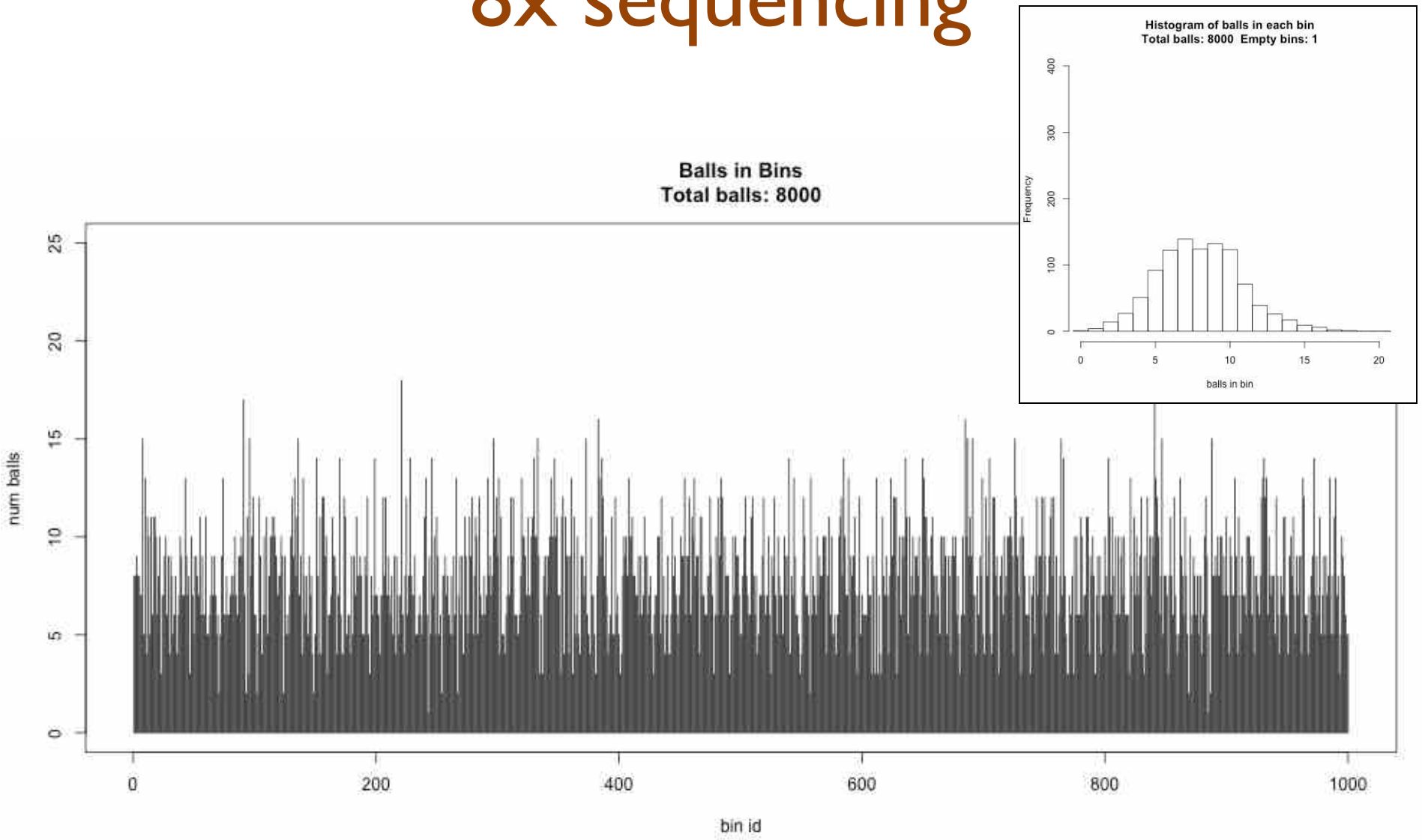
2x sequencing



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.

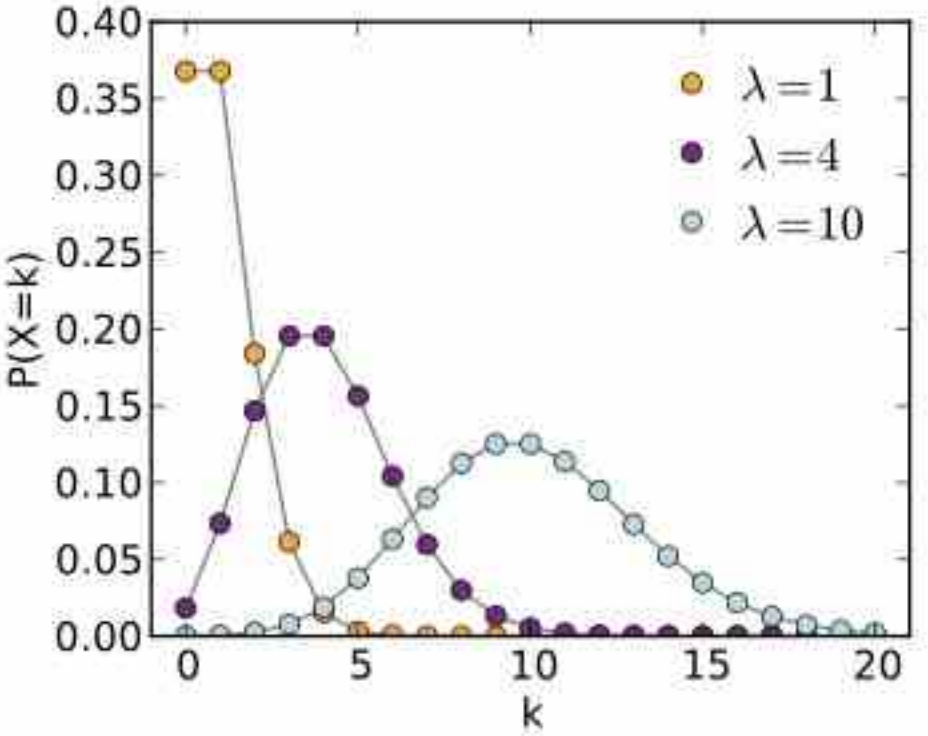
$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Formulation comes from the limit of the binomial equation

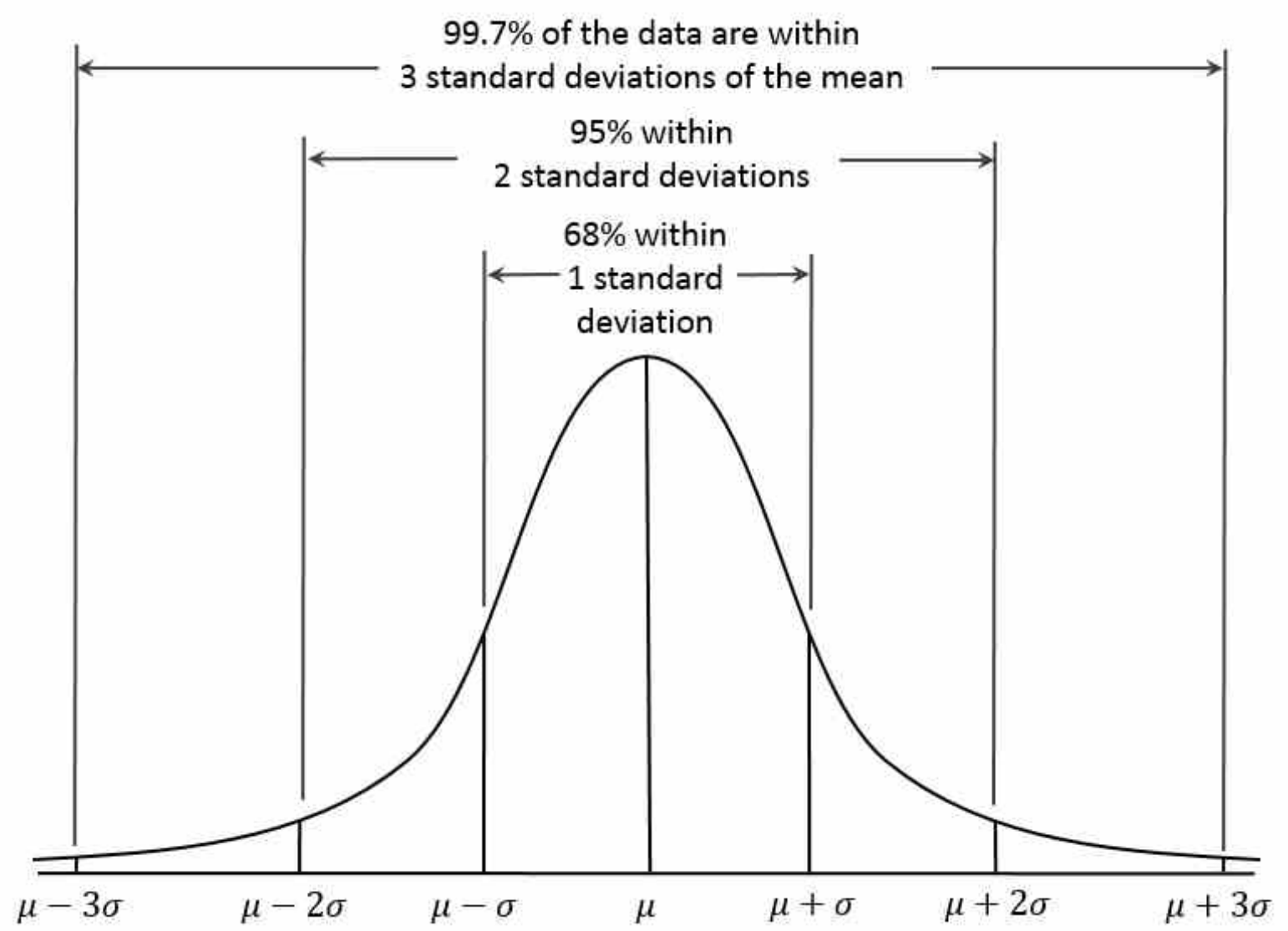
Resembles a normal distribution, but over the positive values, and with only a single parameter.

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

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 10Mbp x 36x = 360Mbp of data
360Mbp / 120bp / read = 3M reads

K-mers and K-mer counting

GATTACATACACATTGGATG

K-mers and K-mer counting

GATTACATACACATTGGATG

GAT ACA ACA ATT GAT

ATT CAT CAC TTG ATG

TTA ATA ACA TGG

TAC TAC CAT GGA

Kmers:

- Divide a string into substrings of length k
- Notice every position is covered k times
- Notice there are G – k + 1 kmers from a string of length G

K-mers and K-mer counting

GATTACATACACATTGGATG

GAT ACA ACA ATT GAT

ATT CAT CAC TTG ATG

TTA ATA ACA TGG

TAC TAC CAT GGA

Kmers:
- Divide a string into substrings of length k
- Notice every position is covered k times
- Notice there are G – k + 1 kmers from a string of length G

Computation: Very easy to compute, exact matches, represent 32mers in 64 bits
Biological: The “atomic unit” of a sequence, creates a fingerprint of a genome/read

K-mers and K-mer counting

GATTACATACACATTGGATG

GAT ACA ACA ATT GAT

ATT CAT CAC TTG ATG

TTA ATA ACA TGG

TAC TAC CAT GGA

GAT:2 CAT:2 ATG:1 TGG:1

ACA:3 CAC:1 TTA:1 TAC:2

ATT:2 TTG:1 ATA:1 GGA:1

K-mers and K-mer counting

GATTACATACACATTGGATG

GAT:2 CAT:2 ATG:1 TGG:1
ACA:3 CAC:1 TTA:1 TAC:2
ATT:2 TTG:1 ATA:1 GGA:1

1: 7 (ATG, TGG, …)
2: 4 (GAT, CAT, ATT, TAC)
3: 1 (ACA)

K-mers and K-mer counting

GATTACATACACATTGGATG

```
1:  7  (ATG, TGG, …)
2:  4  (GAT, CAT, ATT, TAC)
3:  1  (ACA)
```

How long should k be?

K-mers and K-mer counting

GATTACATACACATTGGATG

```
1:  7  (ATG, TGG, …)
2:  4  (GAT, CAT, ATT, TAC)
3:  1  (ACA)
```

How long should k be?

K=1 : Too short, every base is present
K=2 : Too short, every pair of bases will be present

Pick k so that G/(4^k) << 1
k = log_4 (G)
At least 15 for human, often a bit longer
But not too long or could loose resolution

Kmer counting applications

Coverage Statistics

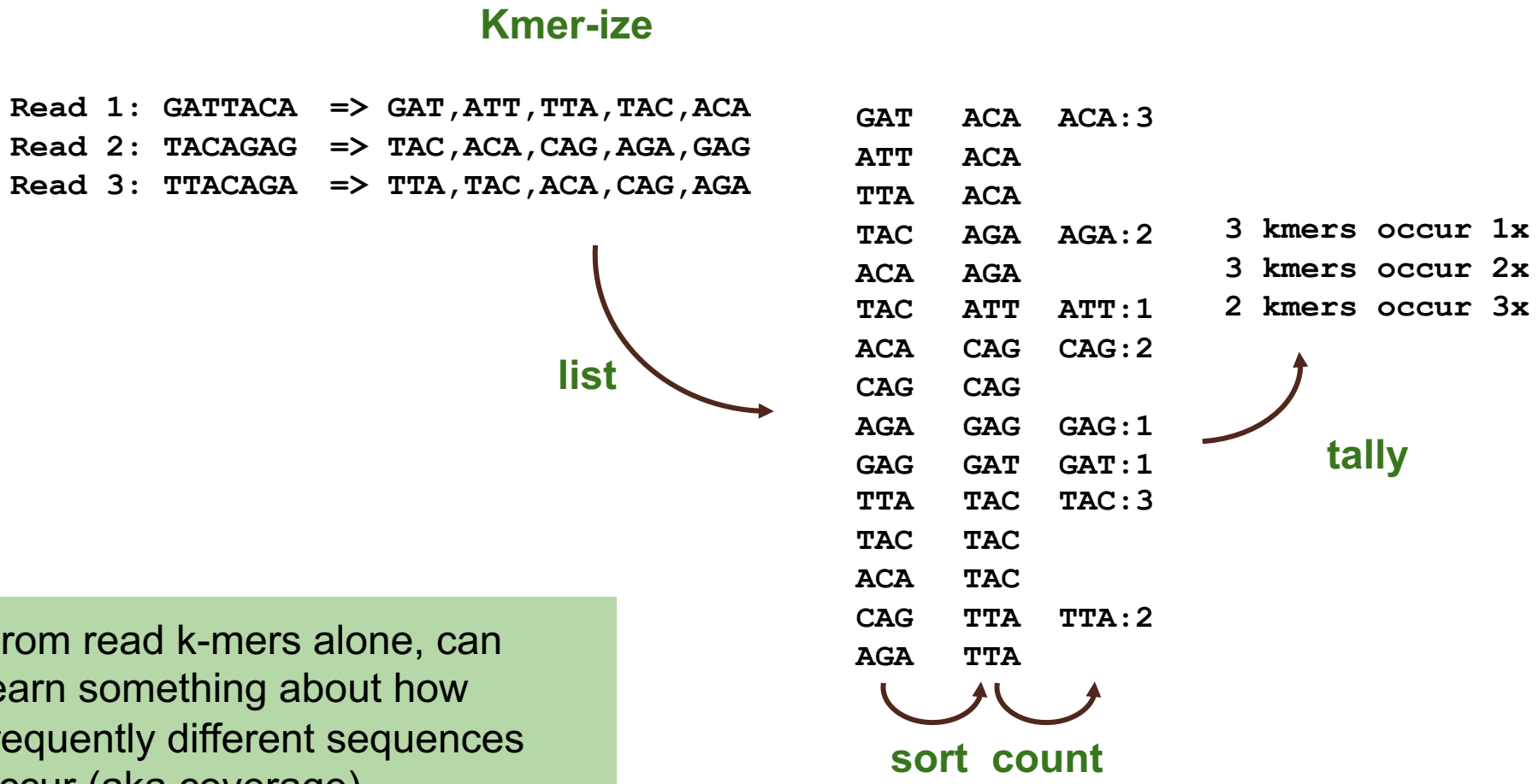
$$\text{sequencing\_coverage} = \frac{\text{total\_bases\_sequenced}}{\text{genome\_size}}$$

$$\text{genome\_size} = \frac{\text{total\_bases\_sequenced}}{\text{sequencing\_coverage}}$$

$$\text{genome\_size} = \frac{100\text{Gb}}{50\text{x}} = 2\text{Gb}$$

But how can you figure out
the coverage without a genome?

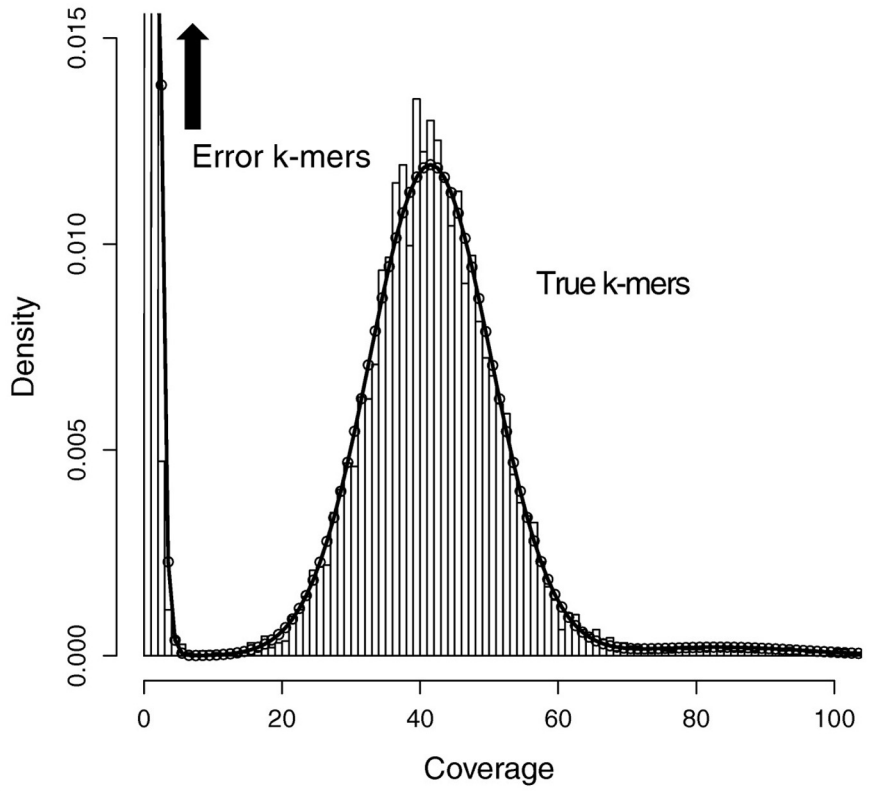
K-mer counting



From read k-mers alone, can learn something about how frequently different sequences occur (aka coverage)

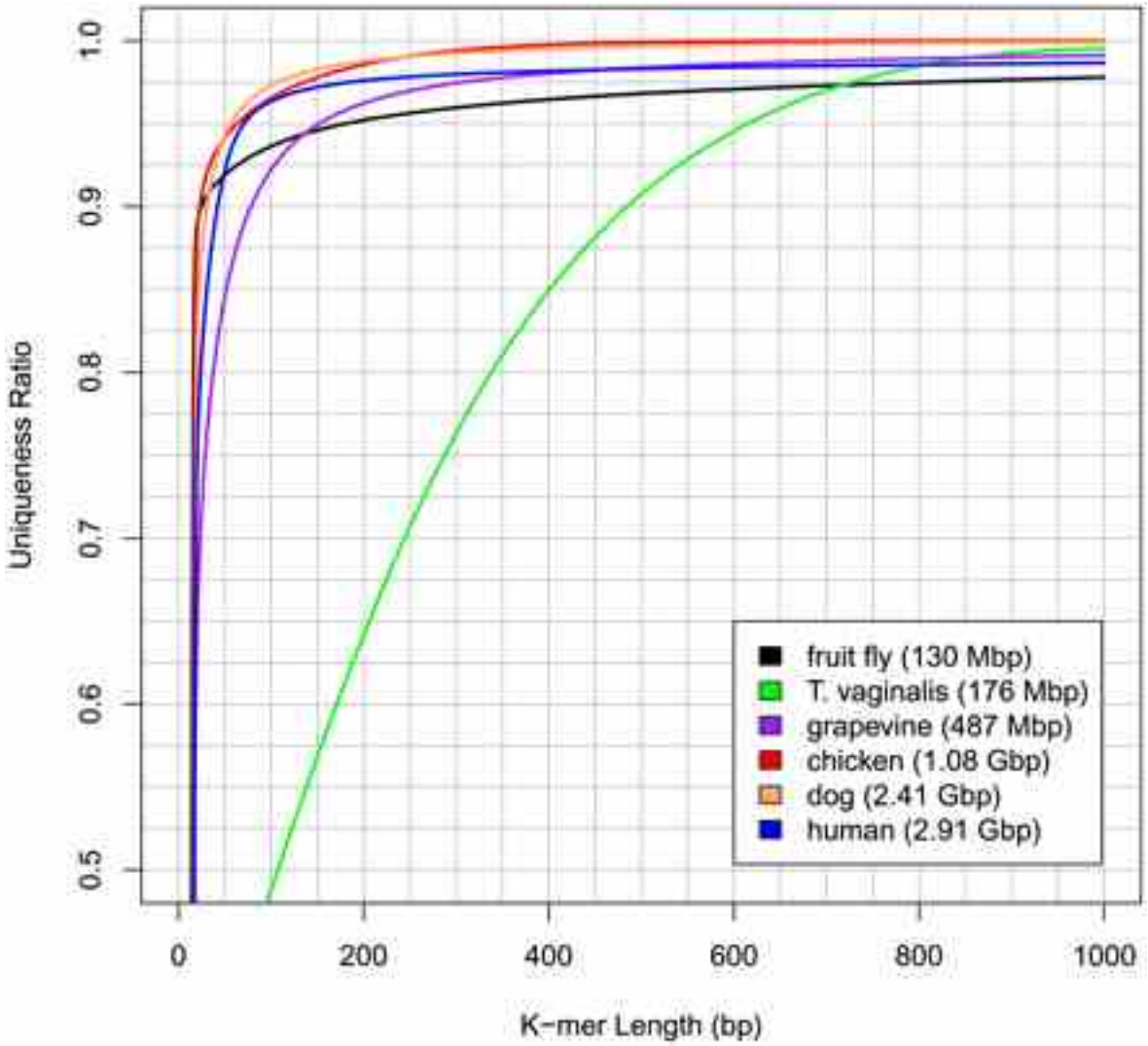
Fast to compute even over huge datasets

K-mer counting in real genomes



- The tally of k-mer counts in real genomes reveals the coverage distribution.
- Here we sequenced 120Gb of reads from a female human (haploid human genome size is 3Gb), and indeed we see a clear peak centered at 40x coverage
- There are also many kmers that only occur <5 times. These are from errors in the reads
- There are also kmers that occur many times (>>70 times). These are repeats in the genome

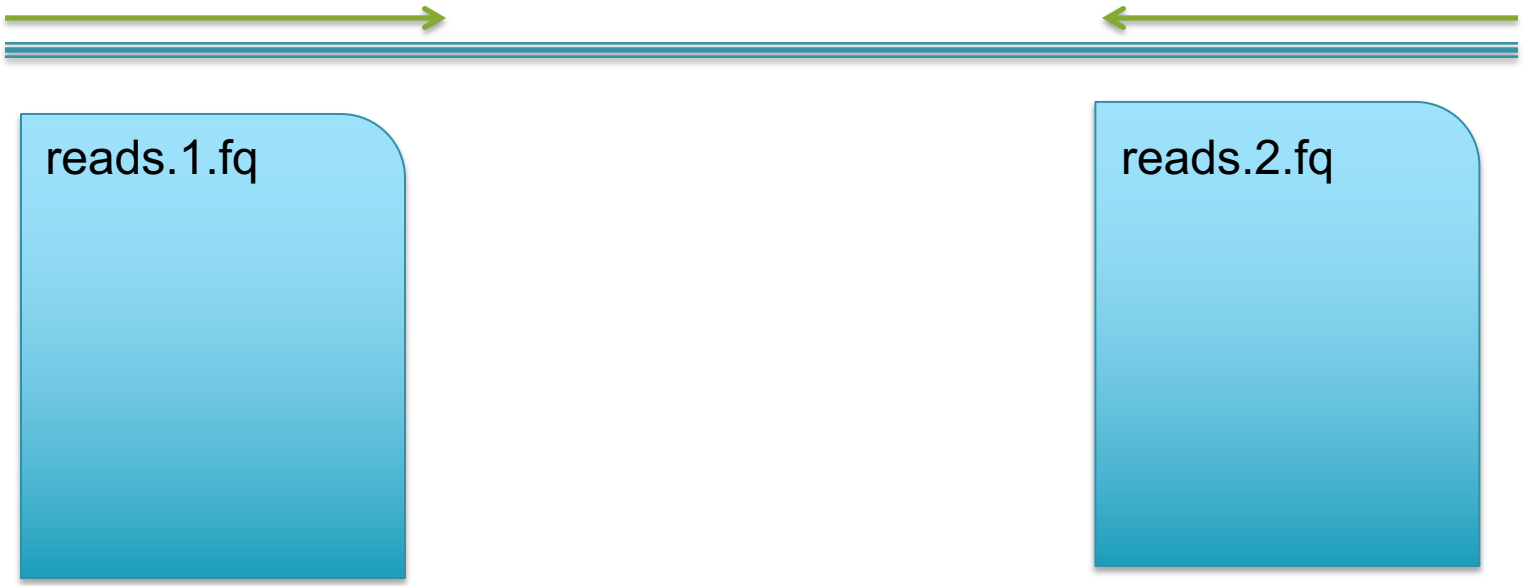
K-mer Uniqueness



Assembly of large genomes using second-generation sequencing
Schatz et al. (2010) Genome Research. doi: 10.1101/gr.101360.109

Illumina Data Characteristics

FASTQ Files




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

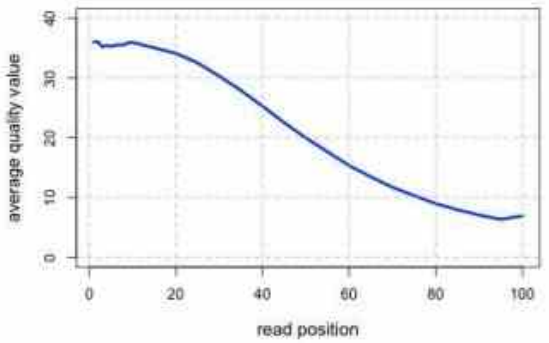
@Identifier
Sequence
+Separator
Quality Values
…

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

Illumina Quality

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

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

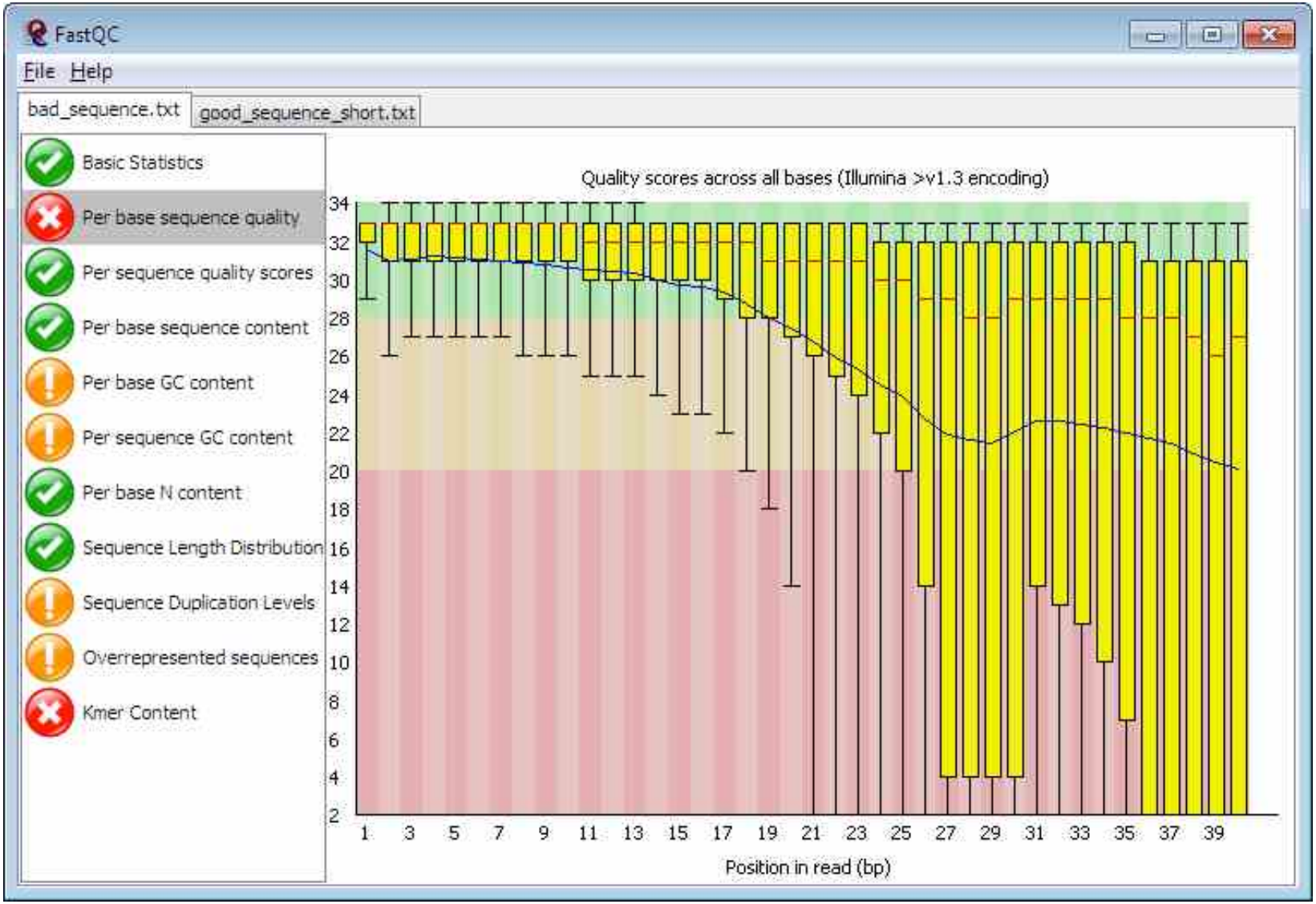



```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
  ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
  ...............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  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)
```

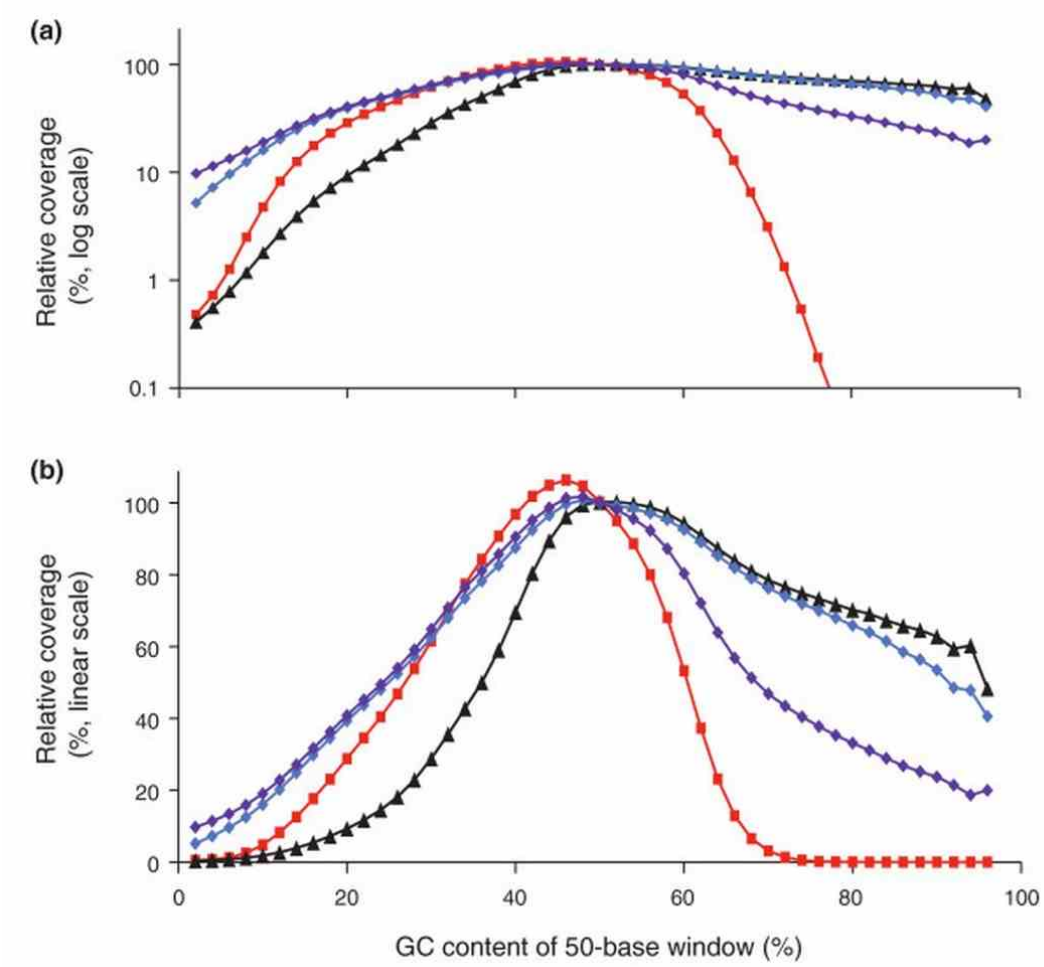
http://en.wikipedia.org/wiki/FASTQ_format

FASTQC: Is my data any good?



http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Beware of GC Biases

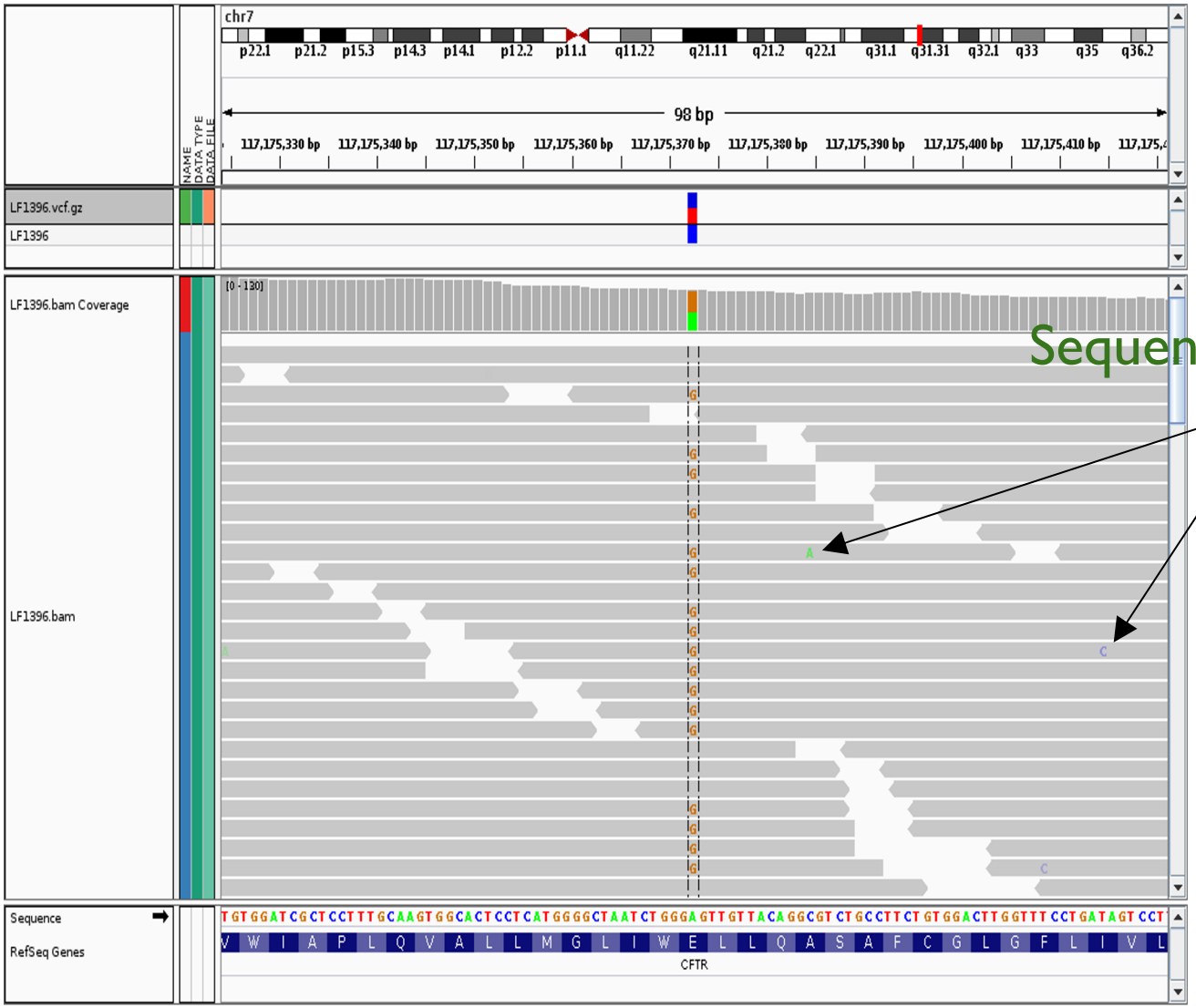


Illumina sequencing does not produce uniform coverage over the genome

- 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.

Sequencing errors fall out as noise (most of the time)



https://jchoigt.files.wordpress.com/2012/07/igv_e217g_snapshot.png

Question?

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

What can we do?

Illumina Hacking

BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 12 2013, pages 1492–1497
doi:10.1093/bioinformatics/btt178

Genome analysis

Advance Access publication May 22, 2013

Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data

Inanc Birol[1,2,3,*], Anthony Raymond[1], Shaun D. Jackman[1], Stephen Pleasance[1], Robin Coope[1], Greg A. Taylor[1], Macaire Man Saint Yuen[4], Christopher I. Keeling[4], Dana Brand[1], Benjamin P. Vandervalk[1], Heather Kirk[1], Pawan Pandoh[1], Richard A. Moore[1], Yongjun Zhao[1], Andrew J. Mungall[1], Barry Jaquish[5], Alvin Yanchuk[5], Carol Ritland[6], Brian Boyle[7], Jean Bousquet[7,8], Kermit Ritland[6], John MacKay[7,8], Jörg Bohlmann[4,6], Steven J.M. Jones[1,2,9]

[1]Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC V5Z 4S6, Canada, [2]University of British Columbia, Department of Medical Genetics, Vancouver, BC V6H 3N1, Canada, [3]School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada, [4]Michael Smith Laboratories, University of British Columbia, Vancouver, BC V6T 1Z4, Canada, [5]British Columbia Ministry of Forests, Lands and Natural Resource Operations, Victoria, BC V8W 9C2, Canada, [6]Department of Forest Sciences, University of British Columbia, Vancouver, BC V6T 1Z4, Canada, [7]Institute for Systems and Integrative Biology, Université Laval, Québec, QC G1V 0A6, Canada, [8]Department of Wood and Forest Sciences, Université Laval, Québec, QC G1V 0A6, Canada and [9]Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

Associate Editor: Michael Brudno

ABSTRACT

White spruce (*Picea glauca*) is a dominant conifer of the boreal forests of North America, and providing genomics resources for this commercially valuable tree will help improve forest management and conservation efforts. Sequencing and assembling the large and highly repetitive spruce genome though pushes the boundaries of the current technology. Here, we describe a whole-genome shotgun sequencing strategy using two Illumina sequencing platforms and an assembly approach using the ABySS software. We report a 20.8 giga base pairs draft genome in 4.9 million scaffolds, with a scaffold N50 of 20 356 bp. We demonstrate how recent improvements in the sequencing technology, especially increasing read lengths and paired end reads from longer fragments have a major impact on the assembly contiguity. We also note that scalable bioinformatics tools are instrumental in providing rapid draft assemblies.

Availability: The *Picea glauca* genome sequencing and assembly data are available through NCBI (Accession#: ALWZ010000000 PID: PRJNA83435). http://www.ncbi.nlm.nih.gov/bioproject/83435.

Contact: ibirol@bcgsc.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 20, 2013; revised on April 10, 2013; accepted on April 11, 2013

1 INTRODUCTION

The assembly of short reads to develop genomic resources for non-model species remains an active area of development (Schatz *et al.*, 2012). The feasibility of the approach and its scalability to large genomes was demonstrated b (Simpson *et al.*, 2009) using human and was later used to assemble the SOAPdenovo tool (Li *et al.*, 2010). high quality results, as demonstrate *et al.*, 2012; Ladner *et al.*, 2013; Rib been successfully applied numerous ti omes (Chan *et al.*, 2011; Chu *et al.*, 2 2011; Godel *et al.*, 2012; Swart *et al.*,

Estimated at 20 giga base pairs (G cing and assembly of the genome of the pine (*Pinaceae*) family present uni generation end, those challenges inclu whole-genome shotgun sequencing da ing reduced representation resources tude of the problem. On the bioint massive sequencing datasets is extrem ing cycles, memory usage, storage req programming implementations on co

We addressed the data representati and sequencing multiple whole-genon HiSeq 2000 and MiSeq sequencers f CA, USA). Compared with localized as building and sequencing fosmi approach of isolating ~10 kb DNA s sequencing fragments in high through CA, USA), a shotgun only sequencing sequence data effectively covering th that can be an order of magnitude les especially substantial when sequencin

In this work, we demonstrate that s at this scale remains viable and pro

*To whom correspondence should be addressed.

assemble the spruce genome, we used the ABySS algorithm (Simpson *et al.*, 2009), which captures a representation of read-to-read overlaps by a distributed de Bruijn graph and uses parallel computations to build the target genome. The modular nature of the tool allowed us to execute a large number of tests to tune the message passing interface for a successful execution, train the assembly parameters for an optimal assembly and quantify the utility of long reads for large genome assemblies. To the best of our knowledge, the ABySS algorithm is unique in its ability to enable genome assemblies of this scale using whole-genome shotgun sequencing data.

2 METHODS

2.1 Sample collection

Apical shoot tissues were collected in April 2006 from a single white spruce (*Picea glauca*, genotype PG29) tree at the Kalamalka Research Station of the British Columbia Ministry of Forests and Ranges, Vernon, British Columbia, Canada. Genomic DNA was extracted from 60 gm tissue by BioS&T (http://www.biost.com/, Montreal, QC, Canada) using an organelle exclusion method yielding 300 µg of high quality purified nuclear DNA.

2.2 Library preparation and sequencing

DNA quality was assessed by spectrophotometry and gel electrophoresis before library construction. DNA was sheared for 45 s using an E210 sonicator (Covaris) and then analysed on 8% PAGE gels. The 200–300 bp (for libraries with 250 bp insert size) or 450–550 bp (for libraries with 500 bp insert size) DNA size fractions were excised and eluted from the gel slices overnight at 4°C in 300 µl of elution buffer (5:1 [vol/vol] LoTE buffer [3 mM Tris–HCl (pH 7.5), 0.2 mM EDTA]/7.5 M ammonium acetate) and was purified using a Spin-X Filter Tube (Fisher Scientific) and ethanol precipitation. Genome libraries were prepared using a modified paired-end tag (PET) protocol supplied by Illumina Inc. This involved DNA end repair and formation of 3′ adenosine overhangs using the Klenow fragment of DNA polymerase I (3′–5′ exonuclease minus) and ligation to Illumina PE adapters (with 5′ overhangs). Adapter-ligated products were purified on QIAquick spin columns (Qiagen) and amplified using Phusion DNA polymerase (NEB) and 10 PCR cycles with the PE primer 1.0 and 2.0 (Illumina). PCR products of the desired size range were purified from adapter ligation artifacts using 8% PAGE gels. DNA quality was assessed and quantified using an Agilent DNA 1000 series II assay (Agilent) and Nanodrop 7500 spectrophotometer (Nanodrop). DNA was subsequently diluted to 8 nM. The final concentration was confirmed using a Quant-iT dsDNA HS assay kit and Qubit fluorometer (Invitrogen).

The mate pair (MPET, a.k.a. jumping) libraries were constructed using 4 µg of genomic DNA with the Illumina Nextera Mate-Pair library construction protocol and reagent (FC-132-1001). The genomic DNA sample was simultaneously fragmented and tagged with a biotin containing mate pair junction adapter, which left a short sequence gap in the tagmented DNA. The gap was filled by a strand displacement reaction using a polymerase to ensure that all fragments were flush and ready for circularization. After an AMPure Bead cleanup, size selection was done on a 0.6% agarose gel to excise 6–9 kb and 9–13 kb fractions, which were purified using a Zymoclean Large Fragment DNA Recovery Kit. The fragments were circularized by ligation, followed by a digestion to remove any linear molecules and left circularized DNA for shearing. The sheared DNA fragments that contain the biotinylated junction adapter (mate pair fragments) were purified by means of binding to streptavidin magnetic beads, and the unwanted unbiotinylated molecules were washed away. The DNA fragments were then end repaired and A-tailed following the protocol and ligated to indexed TruSeq adapters. The final library was enriched by a 10-cycle PCR and purified by AMPure bead clean-up. Library quality and size were assessed by Agilent DNA 1000 series II assay and KAPA Library Quantification protocol. The two fractions were pooled for sequencing paired end 100 bp using Illumina HiSeq2000.

The construction of the 12 kb mate pair libraries was achieved by a hybrid 454/Illumina procedure. Briefly, 50 µg of genomic DNA was fragmented for 20 cycles at speed code 12 using a Hydrosh Marlborough, MA) equipped with a large assembly modu mented DNA was loaded on a 1% agarose gel, and fragmen 18 kb were extracted. Biotinylated circularization adapters Titanium Paired-end Adaptor set (454 Life Sciences/Roch CT) were added to ends of the gel-extracted fragments. recombination of the ends was performed with Cre recom England Biolabs, Ipswich, MA), and linear molecules remai tion were removed with Plasmid Safe (Epicentre, Madison, molecules were fragmented using GS Rapid Library Nebuli Sciences/Roche, Branford, CT), and fragment end-repair fo tailing was performed with the GS Rapid Library preparation Sciences/Roche, Branford, CT). TruSeq Adaptors (Illumina CA) were ligated to the repaired/A-tailed ends. Biotinylat were enriched using Streptavidin-coupled Dynabeads (Life T Grand Island, NY) and amplified by PCR using Illu primers.

Random bacterial artificial chromosome (BAC) seq performed using DNA from the same genotype on 4 Titanium with 6 kb paired-end libraries at the PlateForm Génomiques of the Institute for Systems and Integra (Université Laval, Quebec City, QC). A single paired-end prepared on a pool of 15 BACs (equimolar concentrations) earlier in the text with the following modifications: 15 µg fragmented using a Hydroshear with a standard assembly at speed code 18, 6–10 kb fragments were extracted from GS-FLX library adaptors were ligated to the repaired/A ments. GS-FLX sequencing using the titanium chemistry w according to manufacturer's instructions (454 Life Sci Branford, CT). Sanger sequencing method was used to ob BAC sequencing data as previously described (Hamberger Keeling *et al.*, 2010).

2.3 MiSeq modification

In sequencing the spruce genome, we generated longer rea modifying the MiSeq platform. The MiSeq uses a clamshell s (Supplementary Fig. S1A) to hold reagent tubes in an arra cessed by the MiSeq's sippers. Most of the reagents are u length independent steps such as denaturation and cluster ge three reagents, the Scan, Cleavage and Incorporation mi sumed at each cycle. Although the MiSeq allows any read specified in the control software, the reagent cartridge canno during the run without stopping it. Increasing the read len requires increasing the quantity of the length-dependent cartridge. This led to the solution of combining the lengt reagents of two kits into one.

A tool was designed that opens the snap-hook latches cartridge together (Supplementary Figs S1B and S2), givi the reagent tubes, yet allowing the cartridge to be put b without damage to its components (Supplementary Fi 40 ml, the stock length-dependent reagent containers allow imum of ~650 cycles in total. To maximize the potential of t kit approach, a new reagent tray with 70 ml wells was d placed in a modified clamshell base.

Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data

Inanc Birol[1,2,3,*], Anthony Raymond[1], Shaun D Jackman[1], Stephen Pleasance[1], Robin Coope[1], Greg A Taylor[1], Macaire Man Saint Yuen[4], Christopher I Keeling[4], Dana Brand[1], Benjamin P Vandervalk[1], Heather Kirk[1], Pawan Pandoh[1], Richard A Moore[1], Yongjun Zhao[1], Andrew J Mungall[1], Barry Jaquish[5], Alvin Yanchuk[5], Carol Ritland[4,6], Brian Boyle[7], Jean Bousquet[7,8], Kermit Ritland[6], John MacKay[7,8], Jörg Bohlmann[4,6], Steven JM Jones[1,2,9]

[1] British Columbia Cancer Agency, Genome Sciences Centre, Vancouver, BC V5Z 4S6
[2] University of British Columbia, Department of Medical Genetics, Vancouver, BC V6H 3N1
[3] Simon Fraser University, School of Computing Science, Burnaby, BC V5A 1S6
[4] University of British Columbia, Michael Smith Laboratories, Vancouver, BC V6T 1Z4
[5] British Columbia Ministry of Forests, Lands and Natural Resource Operations, Victoria, BC V8W 9C2
[6] University of British Columbia, Department of Forest Sciences, Vancouver, BC V6T 1Z4
[7] Université Laval, Institute for Systems and Integrative Biology, Québec, QC G1V 0A6
[8] Université Laval, Department of Wood and Forest Sciences, Québec, QC G1V 0A6
[9] Simon Fraser University, Department of Molecular Biology and Biochemistry, Burnaby, BC V5A 1S6

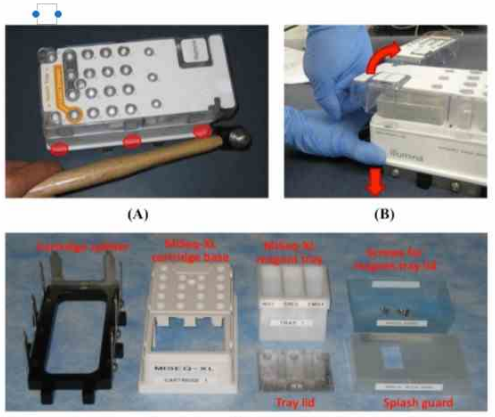


Figure S1. Modification of the MiSeq cartridge. MiSeq reagent cartridge was modified to allow for longer read lengths. **(A, B)** Opening of the clamshell style cartridge. **(C)** Contents of the modified cartridge. This was initially used to combine two PE150 kits for PE300 runs. When Illumina introduced the P250 kit, the same apparatus was used to enable PE500 runs.

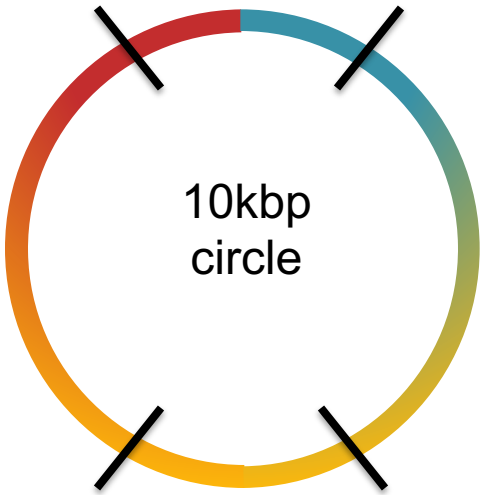
Paired-end and Mate-pairs

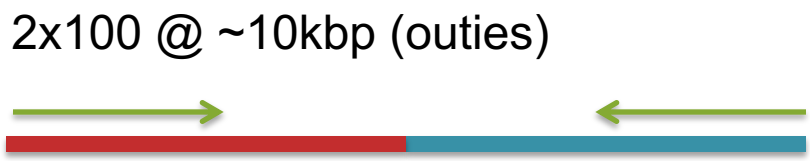
Paired-end sequencing

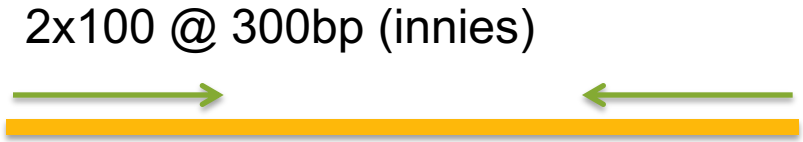
- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation

Mate-pair sequencing

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







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 NextSeq
One human genome in **<30 hours**

Illumina NovaSeq
48 Genomes / run
>8000 genomes per year
3Tb, <3 days, ~$300 / genome

Google

ilmn

All News Finance Images Videos Shopping Web More Tools

Illumina Inc

NASDAQ: ILMN

Overview Financials Compare

Market Summary > Illumina Inc

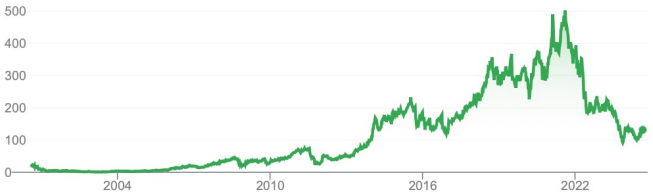
130.42 USD

+111.39 (585.34%) ↑ all time

Closed: Aug 26, 5:20 PM EDT • Disclaimer
After hours 130.42 0.00 (0.00%)

Follow

1D 5D 1M 6M YTD 1Y 5Y Max



Open	131.68	Mkt cap	20.78B	CDP score	C
High	132.72	P/E ratio	-	52-wk high	166.65
Low	130.15	Div yield	-	52-wk low	86.50

Feedback

More about Illumina Inc →

News

Explore more

Related Following

Moderna Inc	81.66 USD	↓ 0.95%
Pacific Biosciences of California Inc	1.60 USD	↑ 2.56%
Amgen Inc	326.78 USD	↓ 0.61%
Oxford Nanopore Technologies PLC	121.04 GBX	↑ 0.37%

Disclaimer

About

illumina.com

Illumina, Inc. is an American biotechnology company, headquartered in San Diego, California. Incorporated on April 1, 1998, Illumina develops, manufactures, and markets integrated systems for the analysis of genetic variation and biological function. Wikipedia

CEO: Jacob Thaysen (Sep 25, 2023–)

Founded: 1998

Google

ilmn

All News Finance Images Videos Shopping Web More Tools

illumina

Illumina Inc

NASDAQ: ILMN

Overview Financials Compare

Market Summary > Illumina Inc

130.42 USD

+111.39 (585.34%) ↑ all time

Closed: Aug 26, 5:20 PM EDT • Disclaimer
After hours 130.42 0.00 (0.00%)

+ Follow

1D 5D 1M 6M YTD 1Y 5Y Max

Open	131.68	Mkt cap	20.78B	CDP score	C
High	132.72	P/E ratio	-	52-wk high	166.65
Low	130.15	Div yield	-	52-wk low	86.50

Feedback

More about Illumina Inc →

News

www.nytimes.com/2021/07/23/science/human-genome-complete

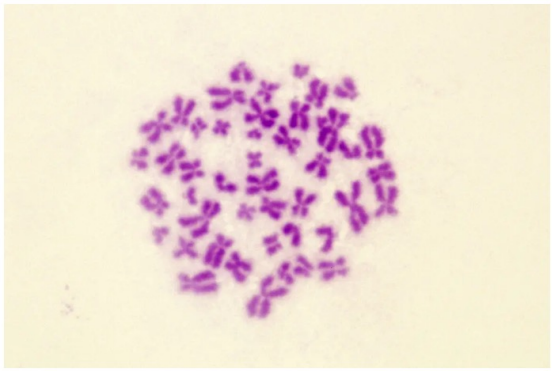
SCIENCE The New York Times GIVE THE TIMES Account

MATTER

Scientists Finish the Human Genome at Last

The complete genome uncovered more than 100 new genes that are probably functional, and many new variants that may be linked to diseases.

Share full article



A century ago, scientists knew that genes were spread across 23 pairs of chromosomes. But pinpointing any single gene and deciphering its sequence was a struggle that could have consumed a career. Michael Abbey/Science Source

By Carl Zimmer

Published July 23, 2021 Updated July 26, 2021

Sign up for Science Times Get stories that capture the wonders of nature, the cosmos and the human body. Get it sent to your inbox.

Two decades after the draft sequence of the human genome was