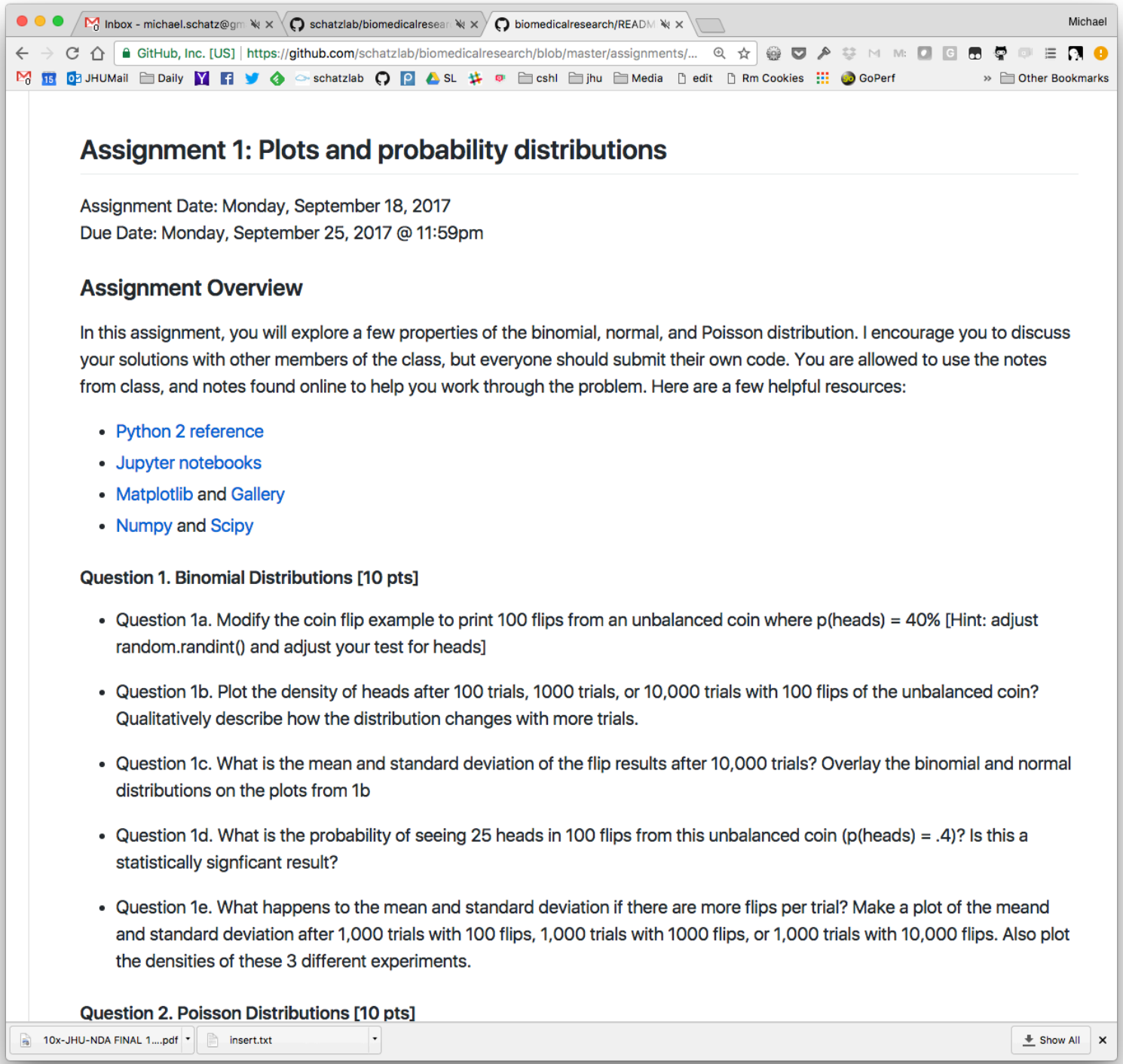
Genome Sequencing pt 2

Michael Schatz

October 11 – Lecture 11

EN.601.452 Computational Biomedical Research

AS.020.415 Advanced Biomedical Research



Assignment 1: Plots and probability distributions

Assignment Date: Monday, September 18, 2017
Due Date: Monday, September 25, 2017 @ 11:59pm

Assignment Overview

In this assignment, you will explore a few properties of the binomial, normal, and Poisson distribution. I encourage you to discuss your solutions with other members of the class, but everyone should submit their own code. You are allowed to use the notes from class, and notes found online to help you work through the problem. Here are a few helpful resources:

- Python 2 reference
- Jupyter notebooks
- Matplotlib and Gallery
- Numpy and Scipy

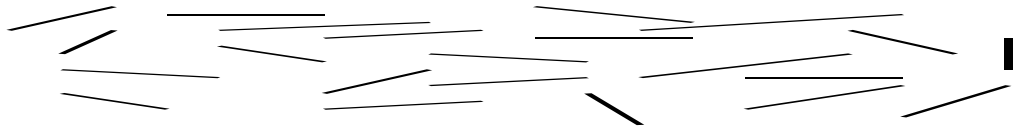
Question 1. Binomial Distributions [10 pts]

- Question 1a. Modify the coin flip example to print 100 flips from an unbalanced coin where p(heads) = 40% [Hint: adjust random.randint() and adjust your test for heads]
- Question 1b. Plot the density of heads after 100 trials, 1000 trials, or 10,000 trials with 100 flips of the unbalanced coin? Qualitatively describe how the distribution changes with more trials.
- Question 1c. What is the mean and standard deviation of the flip results after 10,000 trials? Overlay the binomial and normal distributions on the plots from 1b
- Question 1d. What is the probability of seeing 25 heads in 100 flips from this unbalanced coin (p(heads) = .4)? Is this a statistically signficant result?
- Question 1e. What happens to the mean and standard deviation if there are more flips per trial? Make a plot of the meand and standard deviation after 1,000 trials with 100 flips, 1,000 trials with 1000 flips, or 1,000 trials with 10,000 flips. Also plot the densities of these 3 different experiments.

Question 2. Poisson Distributions [10 pts]

Assembling a Genome

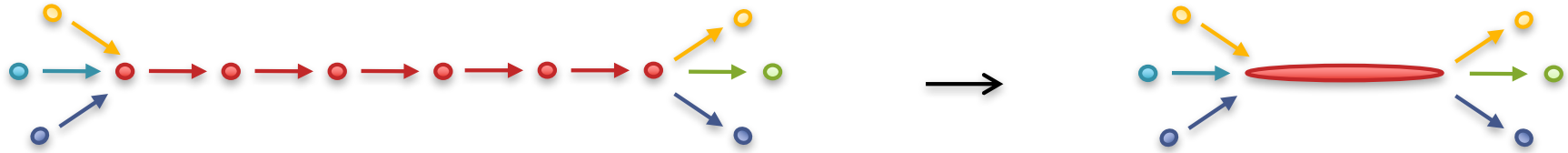
1. Shear & Sequence DNA



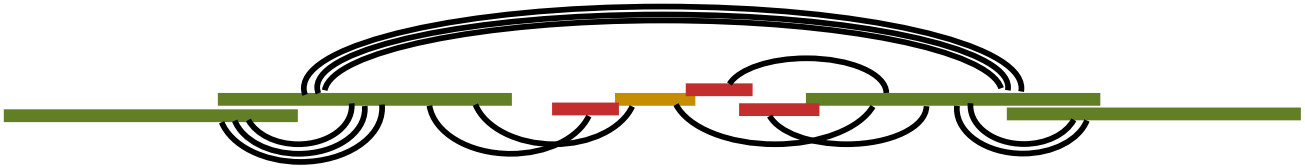
2. Construct assembly graph from reads (de Bruijn / overlap graph)

...AGCCTAGGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
CAACCTCGGACGGACCTCAGCGAA...

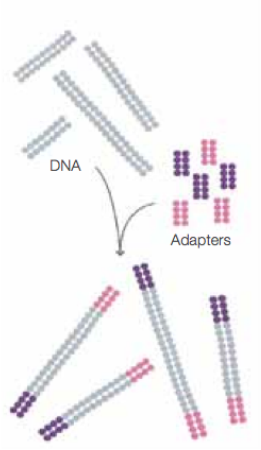
3. Simplify assembly graph



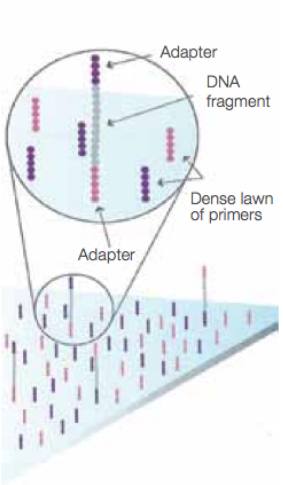
4. Detangle graph with long reads, mates, and other links



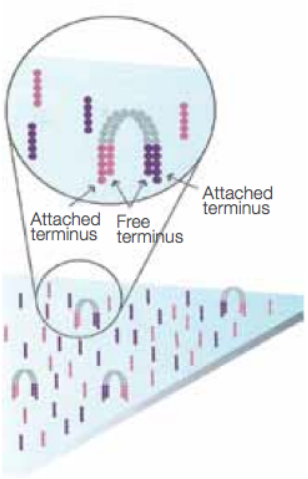
Illumina Sequencing by Synthesis



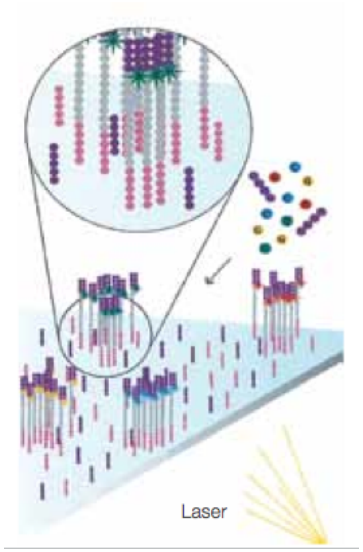
1. Prepare



2. Attach



3. Amplify



4. Image

5. Basecall

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

Illumina Roadmap

Illumina Novaseq

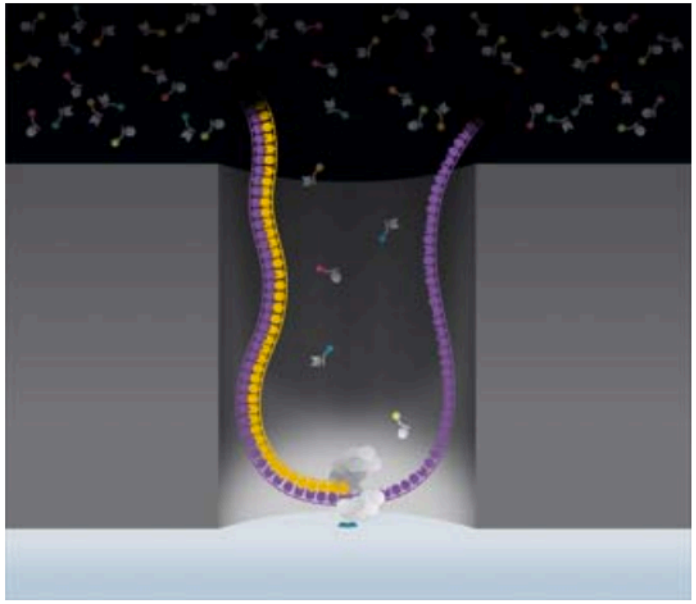
$850k instrument cost
~$1k / human @ 50x
Short reads, high throughput

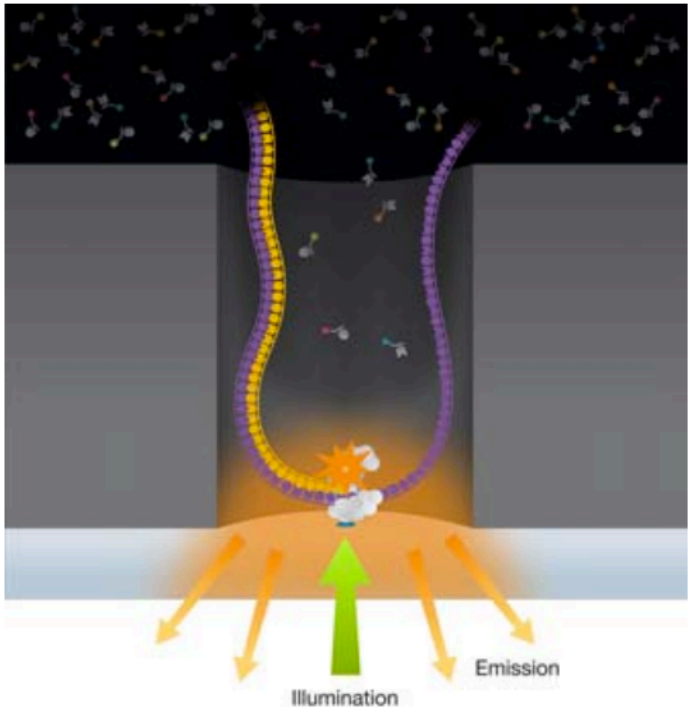
10X Chromium

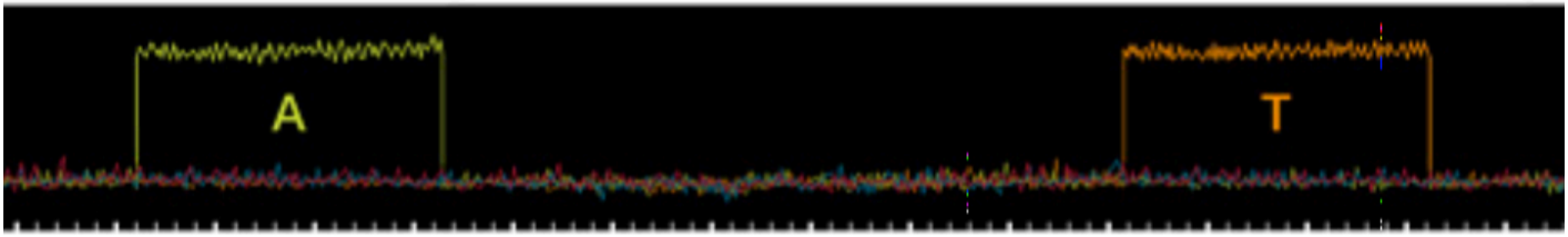
$125k instrument costs
~$2k / human
Linked reads, medium throughput

PacBio: SMRT Sequencing

Imaging of florescent phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).







http://www.youtube.com/watch?v=v8p4ph2MAvI

PacBio Roadmap

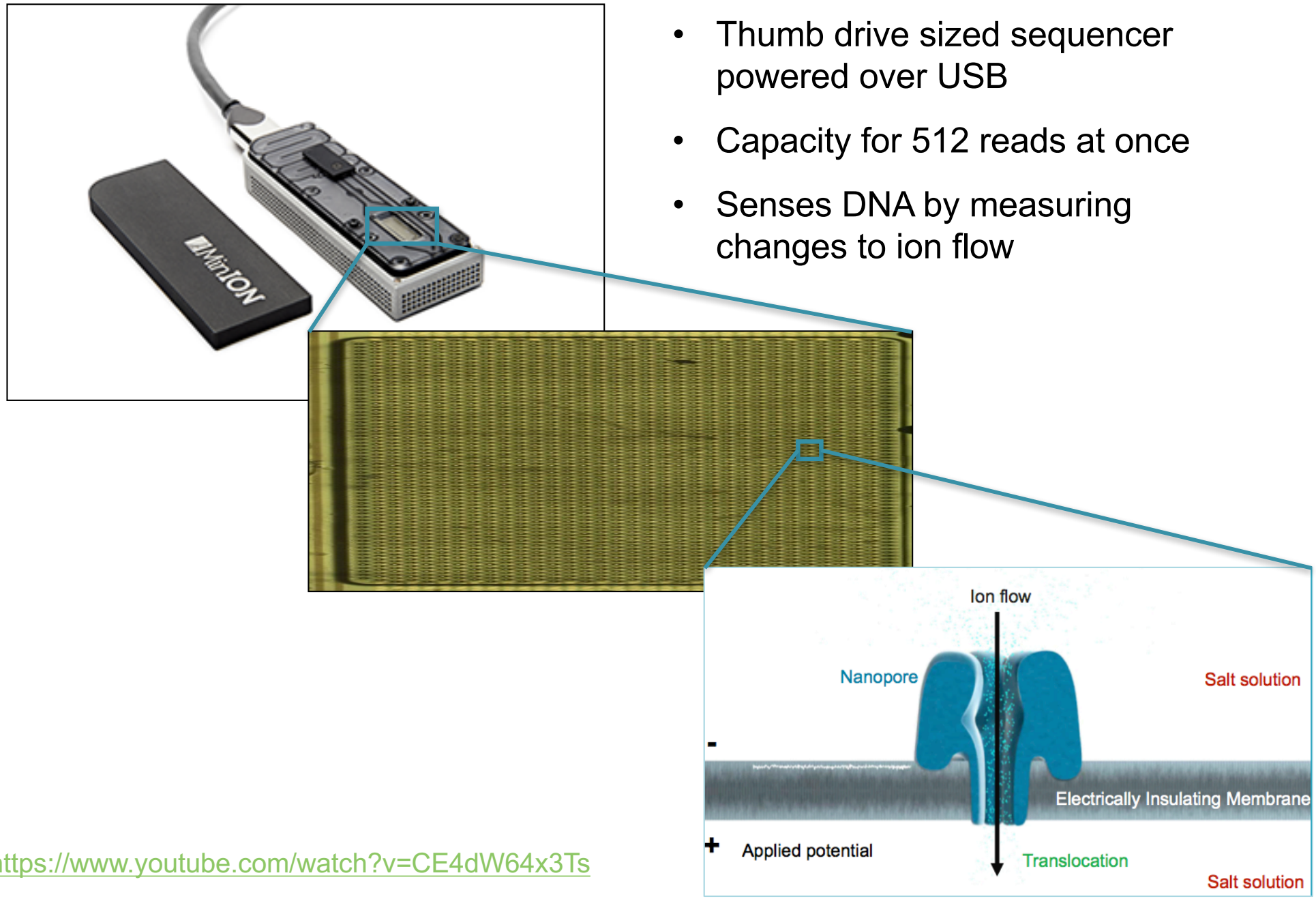
PacBio Sequel

$350k instrument cost
~$30k / human @ 50x
Long reads, Medium throughput

SMRTcell v2

1M Zero Mode Waveguides
~15kb average read length
~$1000 / SMRTcell

Oxford Nanopore MinION



- Thumb drive sized sequencer powered over USB
- Capacity for 512 reads at once
- Senses DNA by measuring changes to ion flow

https://www.youtube.com/watch?v=CE4dW64x3Ts

Oxford Nanopore

MinION

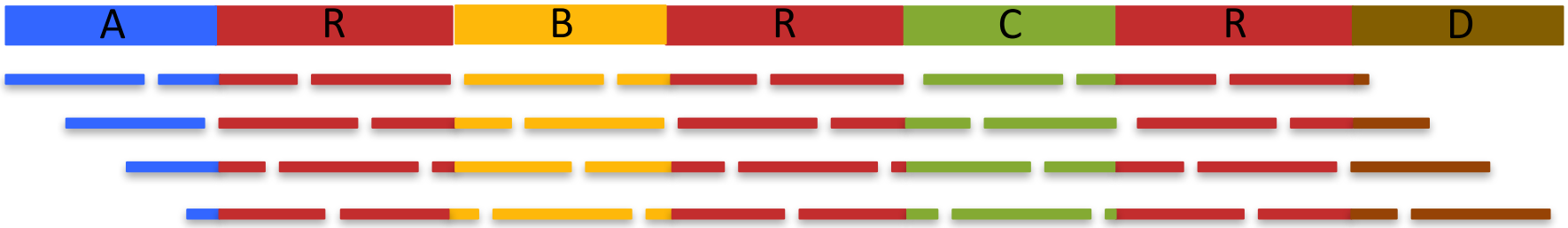
$1k / instrument
~$30k / human @ 50x
Long reads, Low throughput

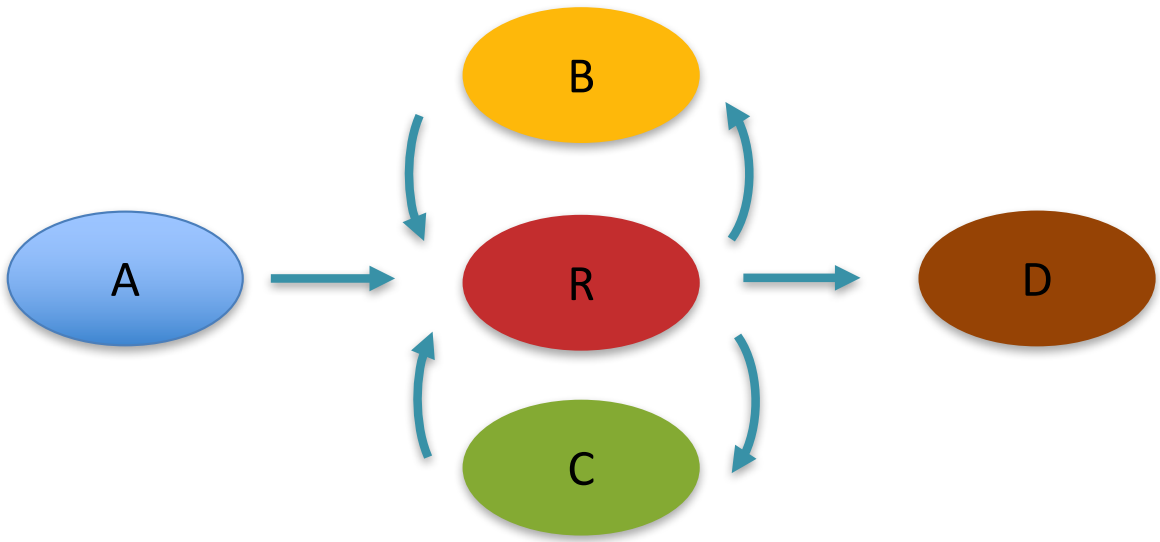
PromethION

$75k / instrument
>>100GB / day
??? / human @ 50x

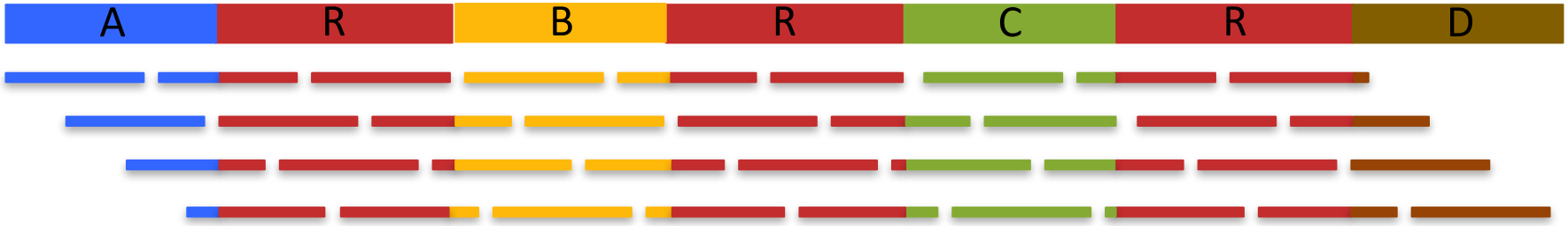
Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome
Goodwin, S, Gurtowski, J, Ethe-Sayers, S, Deshpande, P, Schatz MC* McCombie, WR* (2015) Genome Research doi: 10.1101/gr.191395.115

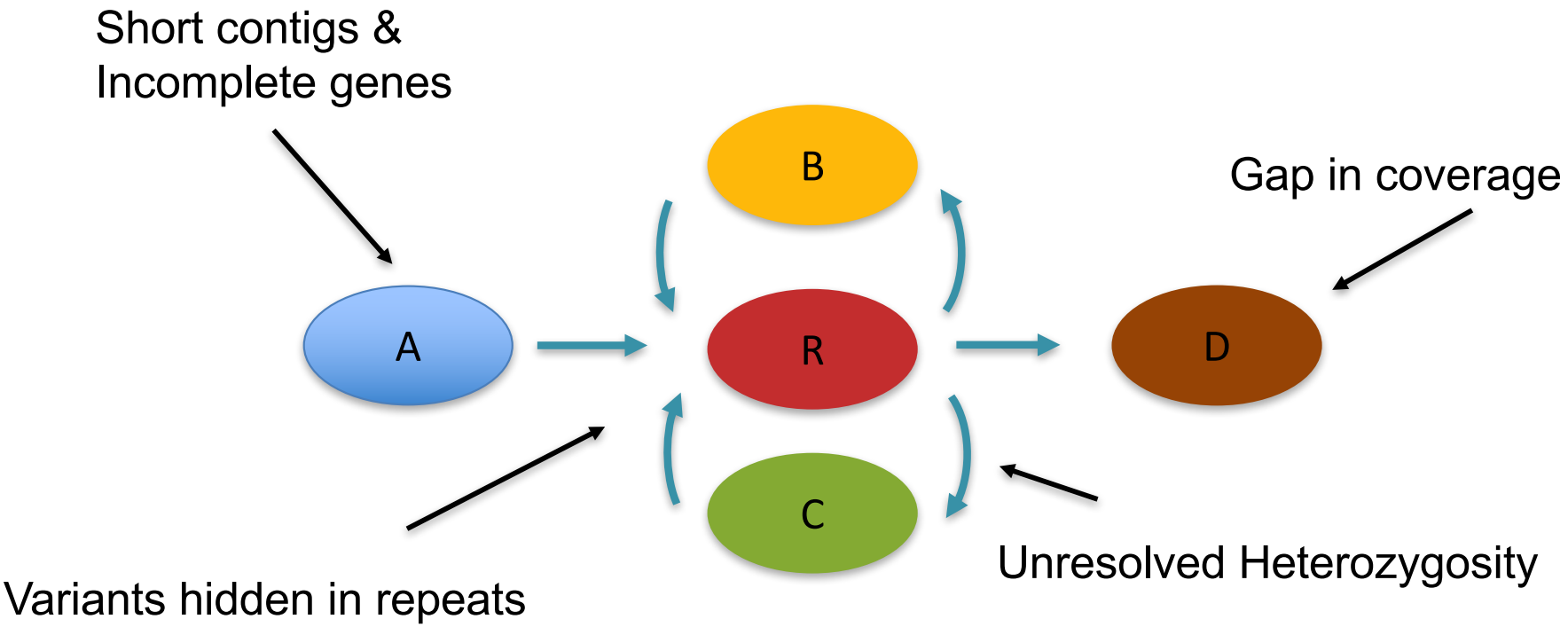
Assembly Complexity



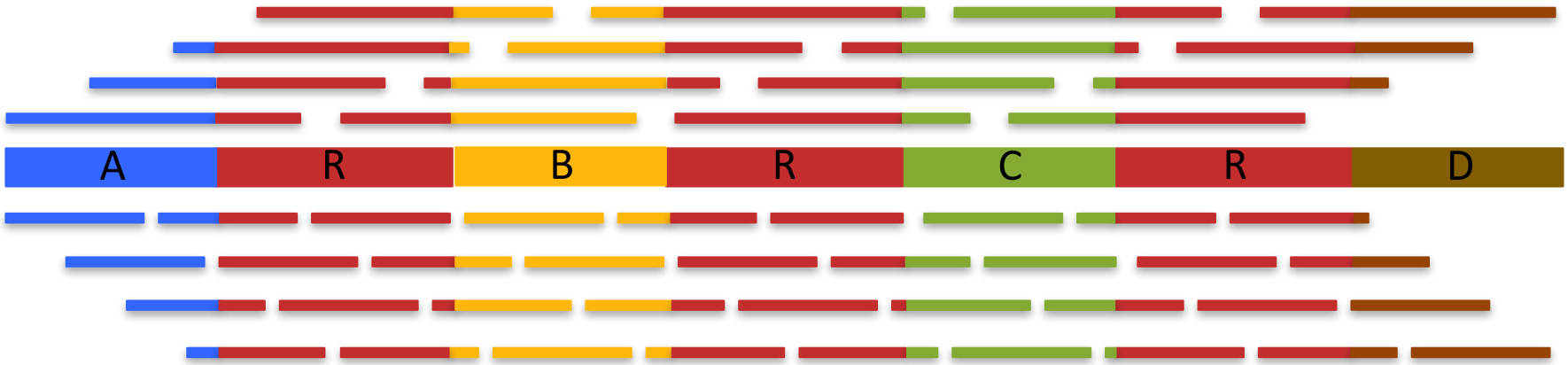


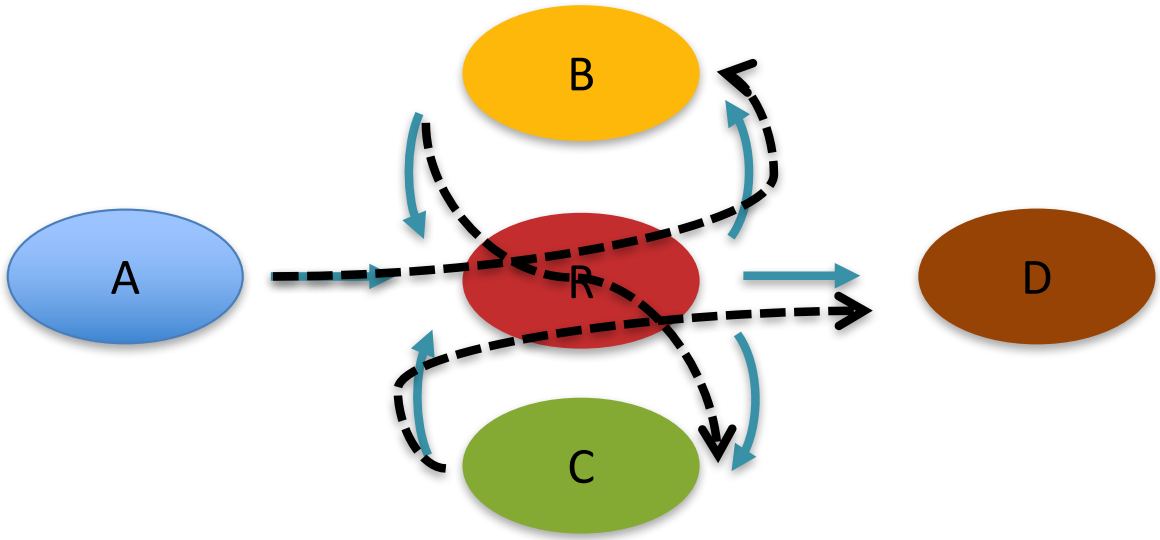
Assembly Complexity



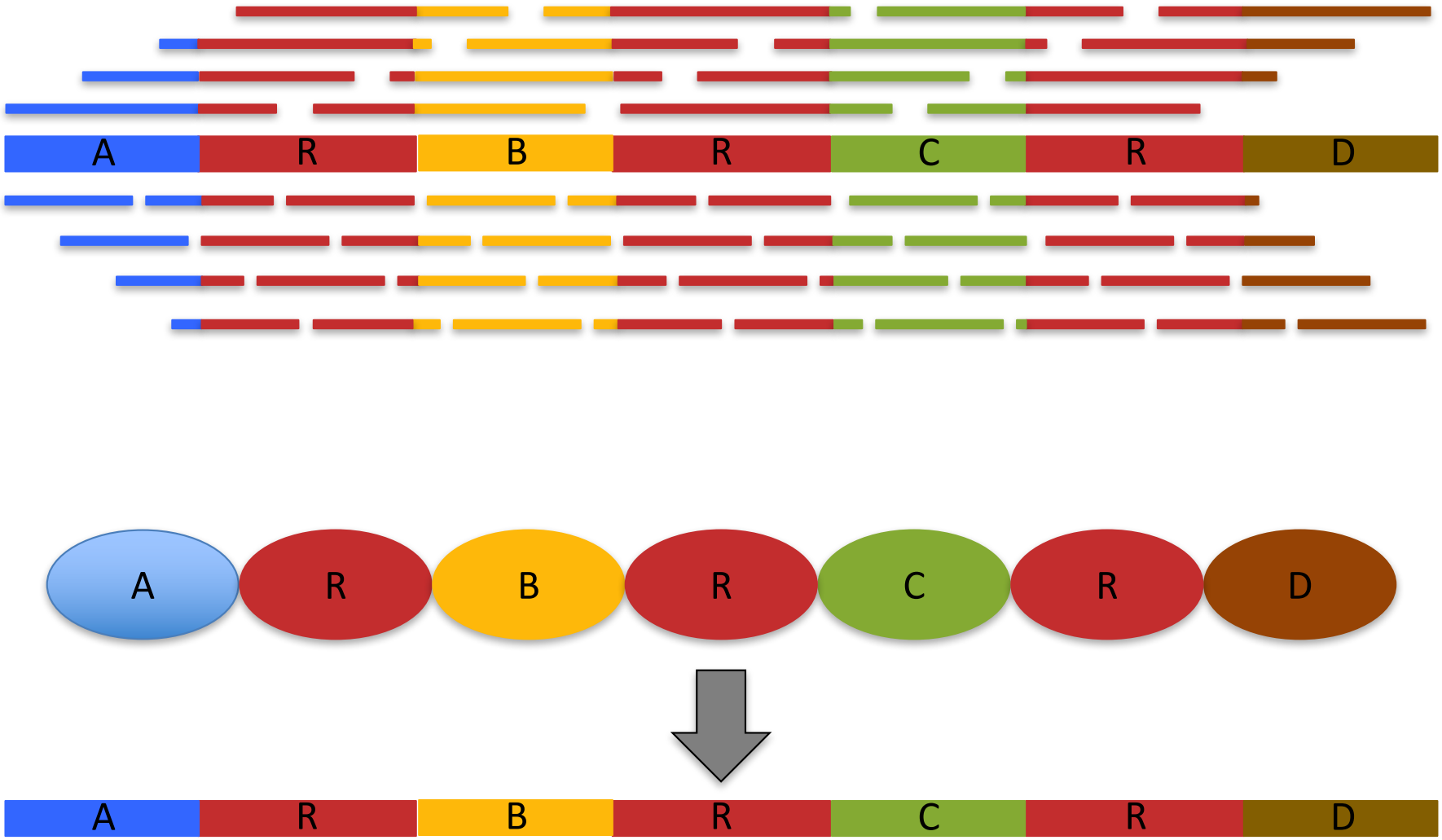


Assembly Complexity



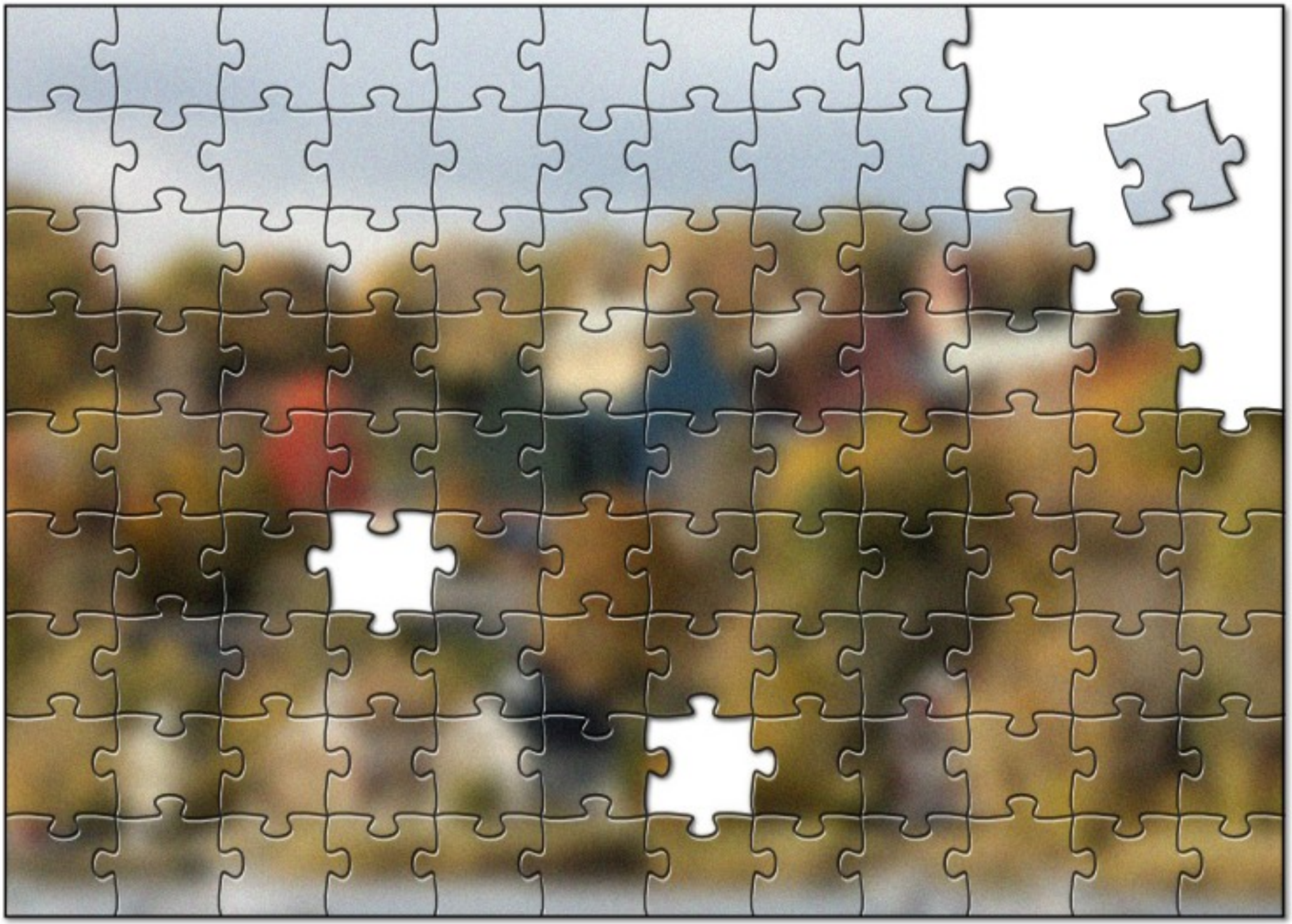


Assembly Complexity



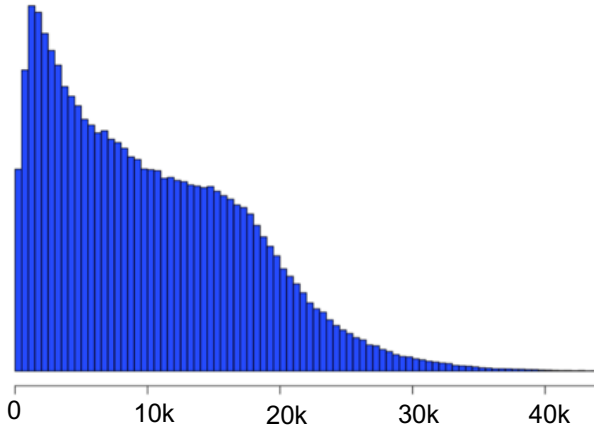
The advantages of SMRT sequencing
Roberts, RJ, Carneiro, MO, Schatz, MC (2013) *Genome Biology*. 14:405

Single Molecule Sequences



Single Molecule Sequencing

PacBio RS II

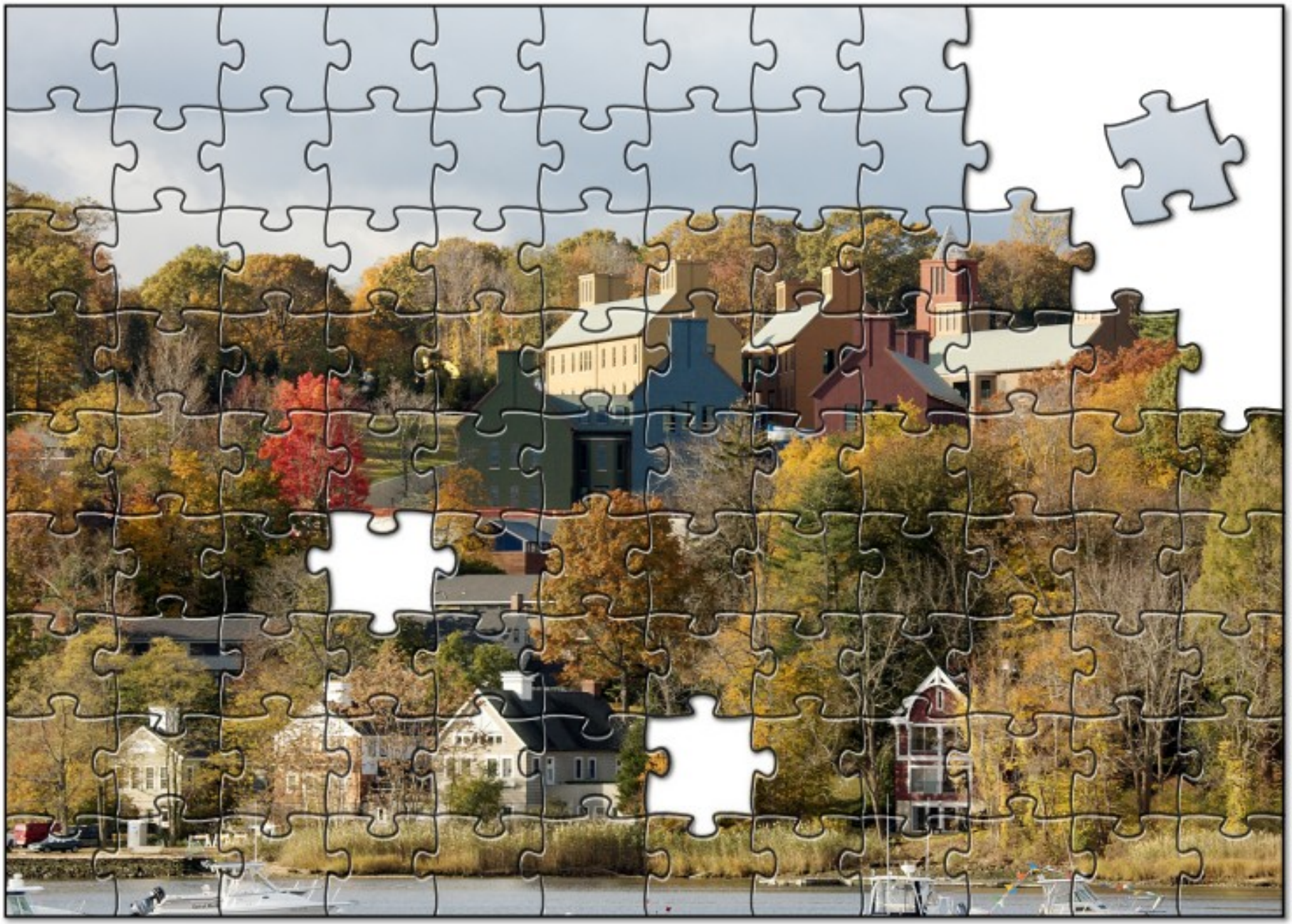


CSHL/PacBio

```
TTGTAAGCAGTTGAAAACTATGTGTGGATTTAGAATAAAGAACATGAAAG
|||||||||||||||||||||||||| ||||||| |||||||||||| |||
TTGTAAGCAGTTGAAAACTATGTGT-GATTTAG-ATAAAGAACATGGAAG

ATTATAAA-CAGTTGATCCATT-AGAAGA-AAACGCAAAAGGCGGCTAGG
| |||||| |||||||||||||| |||| | |||||| |||||| ||||||
A-TATAAATCAGTTGATCCATTAAGAA-AGAAACGC-AAAGGC-GCTAGG

CAACCTTGAATGTAATCGCACTTGAAGAACAAGATTTTATTCCGCGCCCG
| |||||| |||| ||  ||||||||||||||||||||||||||||||||
C-ACCTTG-ATGT-AT--CACTTGAAGAACAAGATTTTATTCCGCGCCCG

TAACGAATCAAGATTCTGAAAACACAT-ATAACAACCTCCAAAA-CACAA
| ||||||| |||||||||||||| || ||    |||||||||| |||||
T-ACGAATC-AGATTCTGAAAACA-ATGAT----ACCTCCAAAAGCACAA

-AGGAGGGGAAAGGGGGGGAATATCT-ATAAAAGATTACAAATTAGA-TGA
 ||||||   ||     |||||||| || |||||||||||||| || |||
GAGGAGG---AA-----GAATATCTGAT-AAAGATTACAAATT-GAGTGA

ACT-AATTCACAATA-AATAACACTTTTA-ACAGAATTGAT-GGAA-GTT
||| ||||||||| | |||||||||||| ||| ||||||| |||| |||
ACTAAATTCACAA-ATAATAACACTTTTAGACAAAATTGATGGGAAGGTT

TCGGAGAGATCCAAAACAATGGGC-ATCGCCTTTGA-GTTAC-AATCAAA
|| ||||||||| ||||||| ||| |||| |||||| ||||| |||||||
TC-GAGAGATCC-AAACAAT-GGCGATCG-CTTTGACGTTACAAATCAAA
```

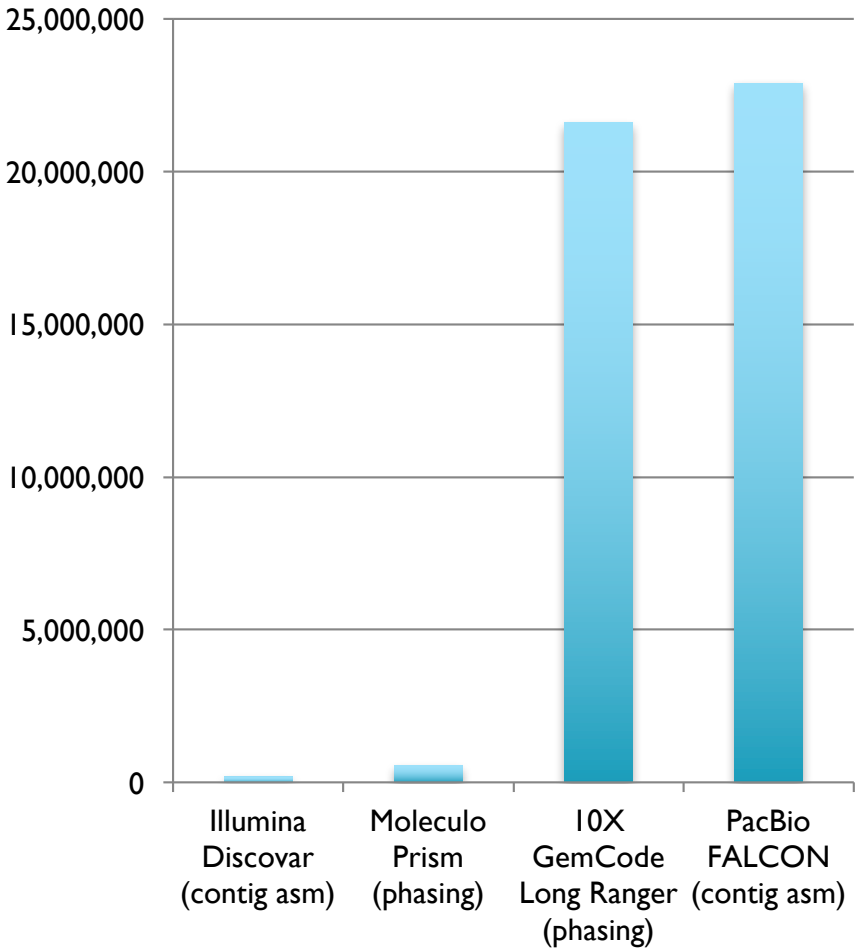
Sample of 100k reads aligned with BLASR requiring >100bp alignment
Average overall accuracy 83.7%: 11.5% insertions, 3.4% deletions, 1.4% mismatch

“Corrective Lens” for Sequencing



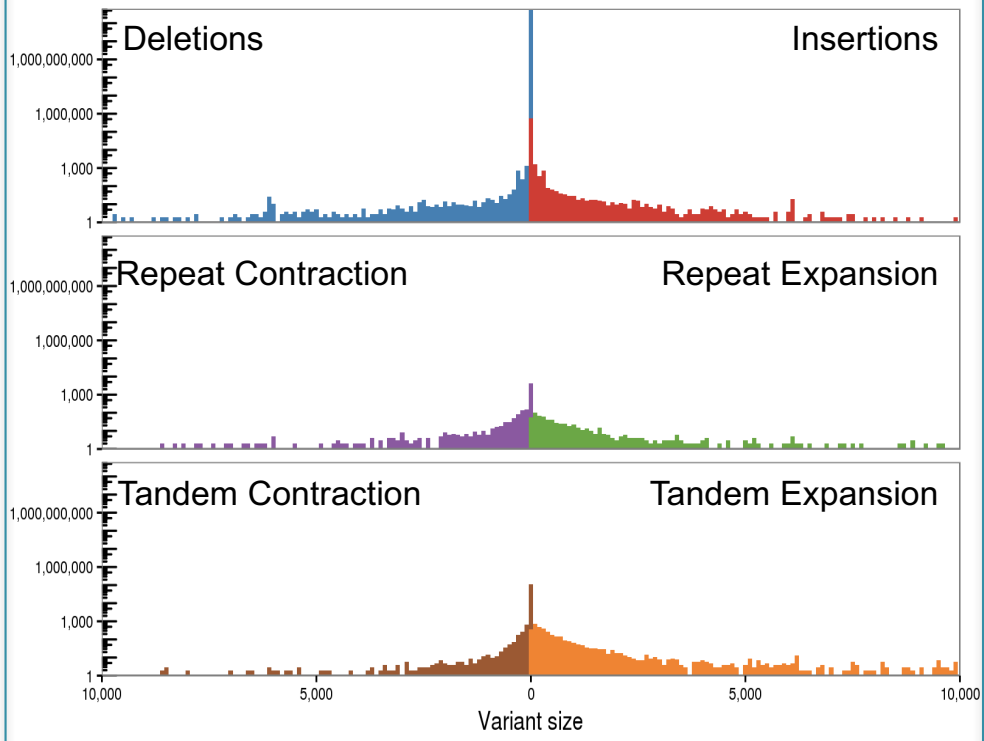
Recent Long Read Assemblies

Human Analysis N50 Sizes



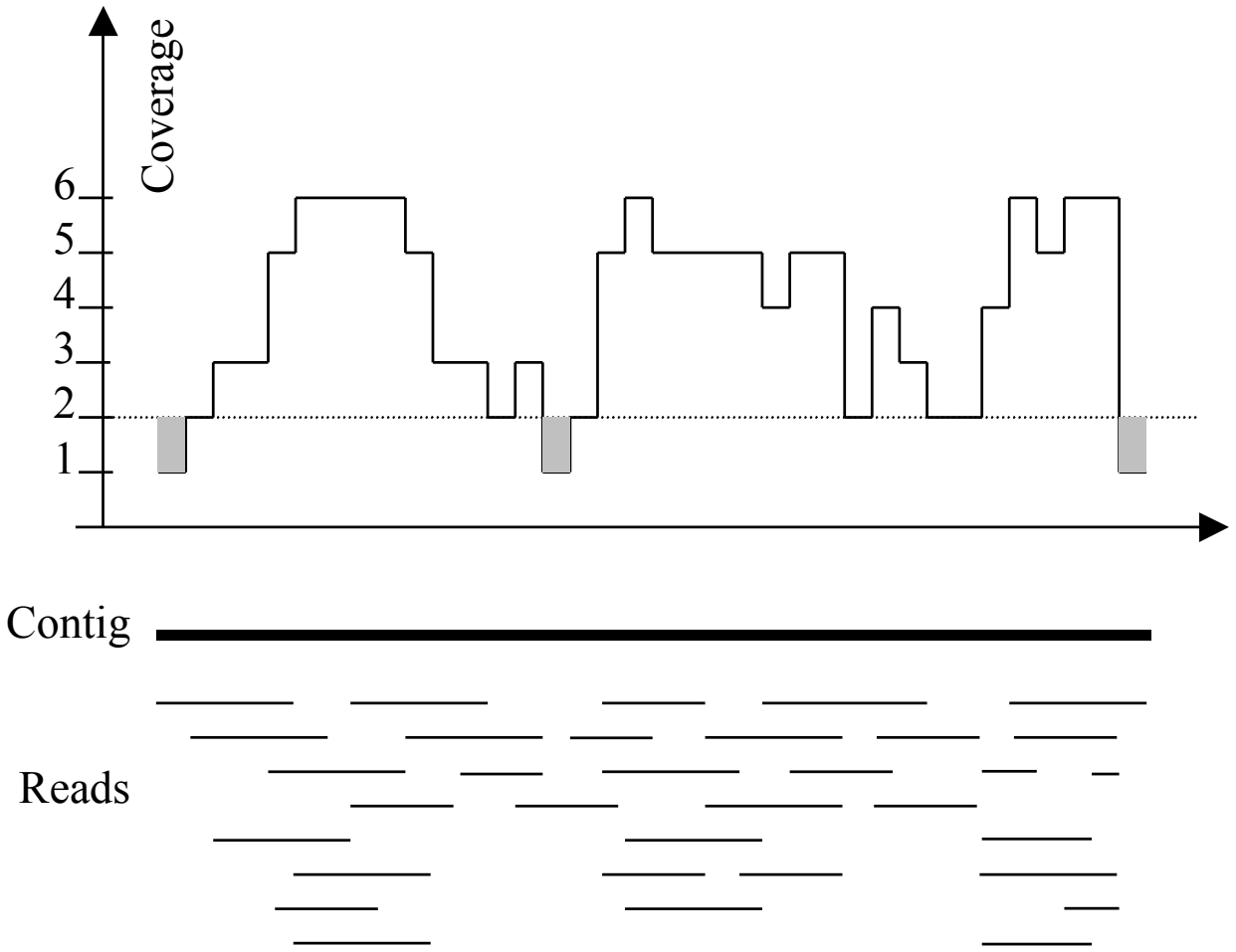
Third-generation sequencing and the future of genomics
Lee et al (2016) *bioRxiv*
doi: http://dx.doi.org/10.1101/048603

Structural Variants in CHM1



Assemblytics: a web analytics tool for the detection of variants from an assembly
Nattestad & Schatz (2016) *Bioinformatics.*
doi: 10.1093/bioinformatics/btw369

Typical sequencing coverage

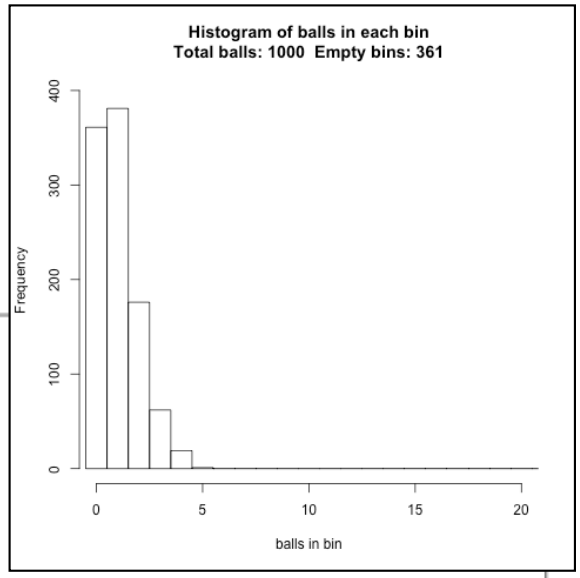


Imagine raindrops on a sidewalk

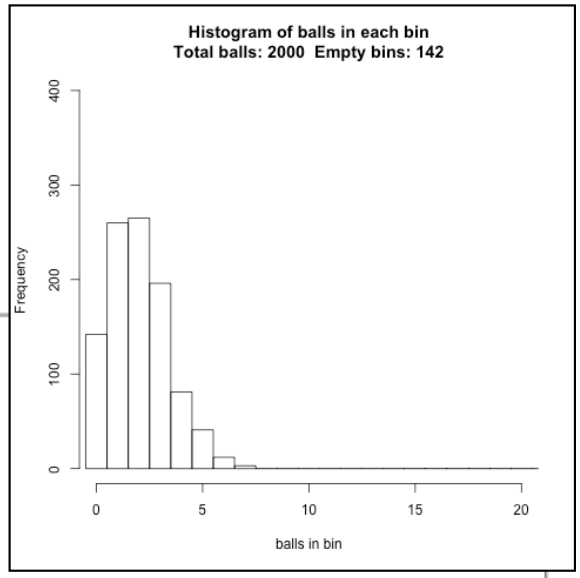
We want to cover the entire sidewalk but each drop costs $1

If the genome is 100 Mbp, should we sequence 1M 100bp reads?

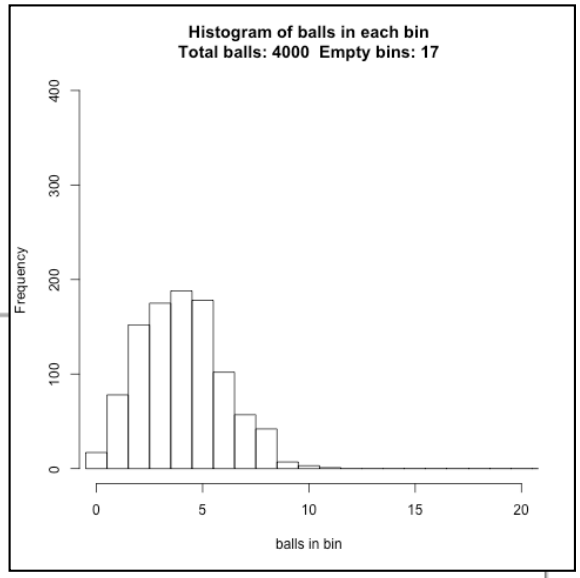
1x sequencing



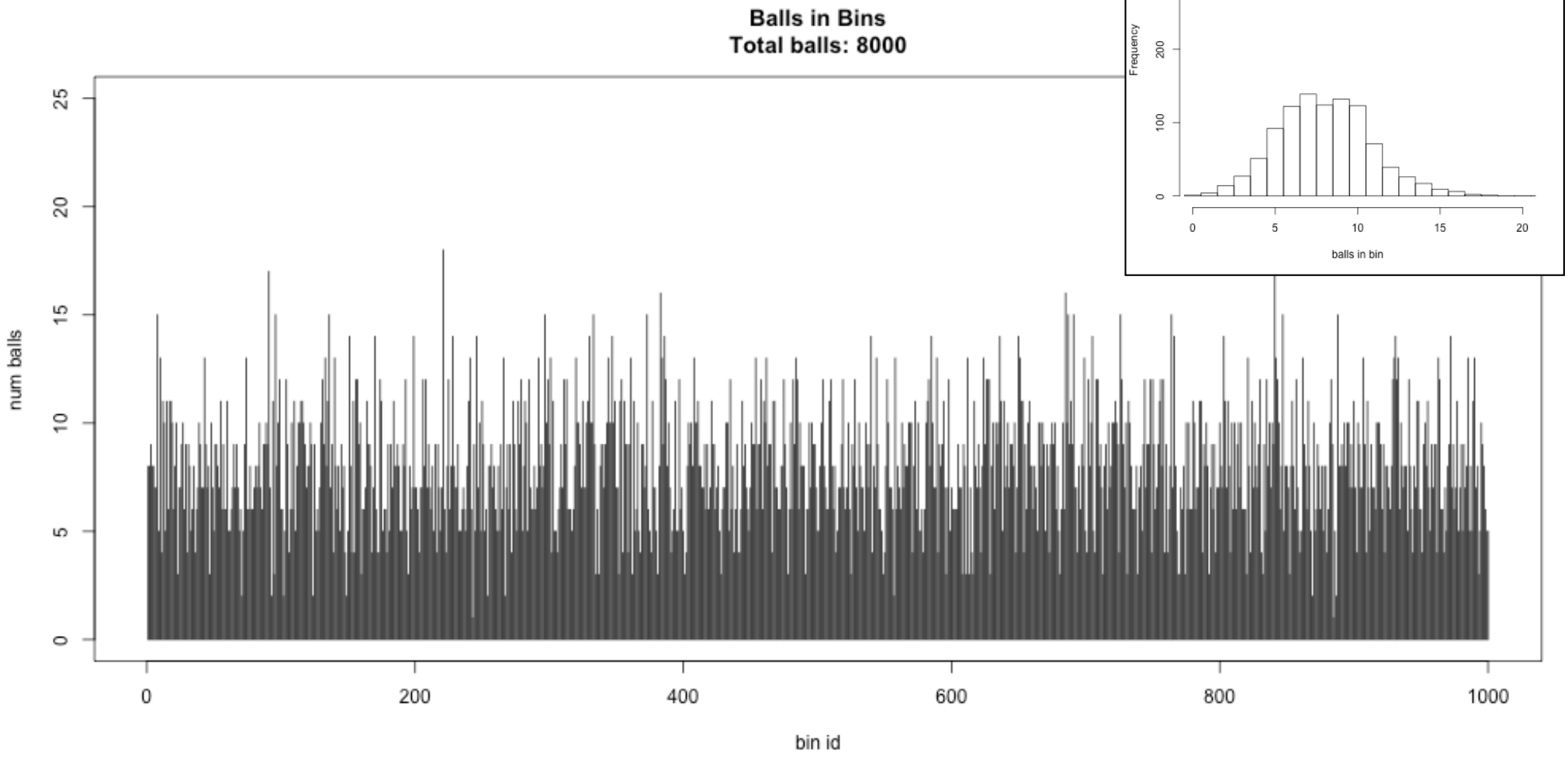
2x sequencing



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.

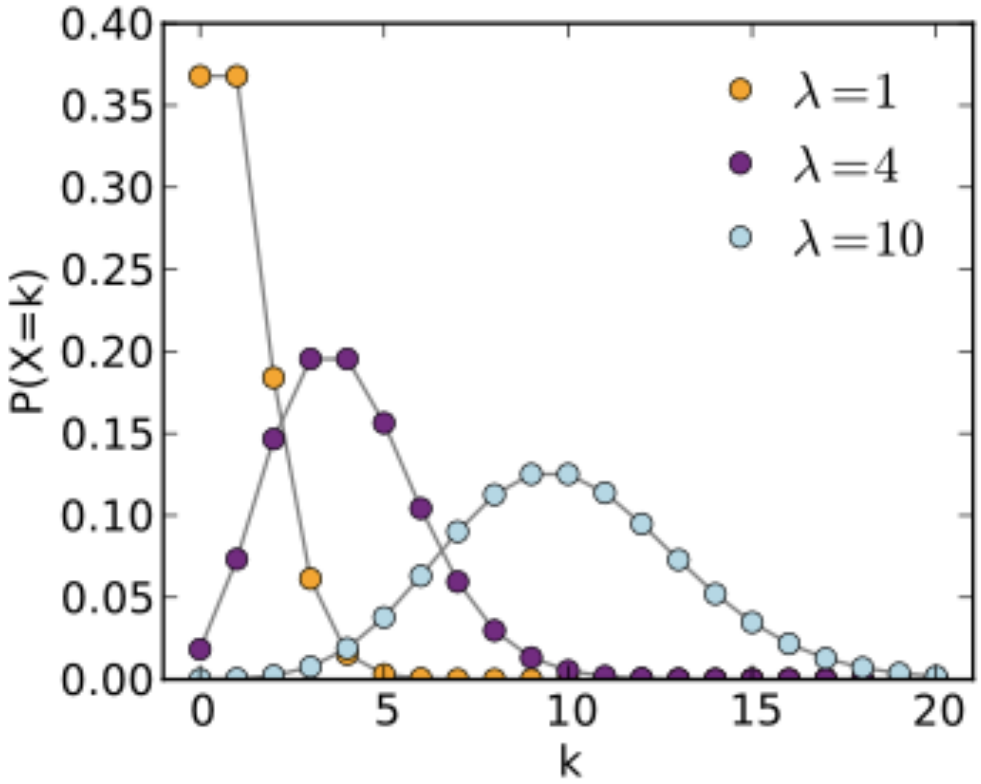
$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Formulation comes from the limit of the binomial equation

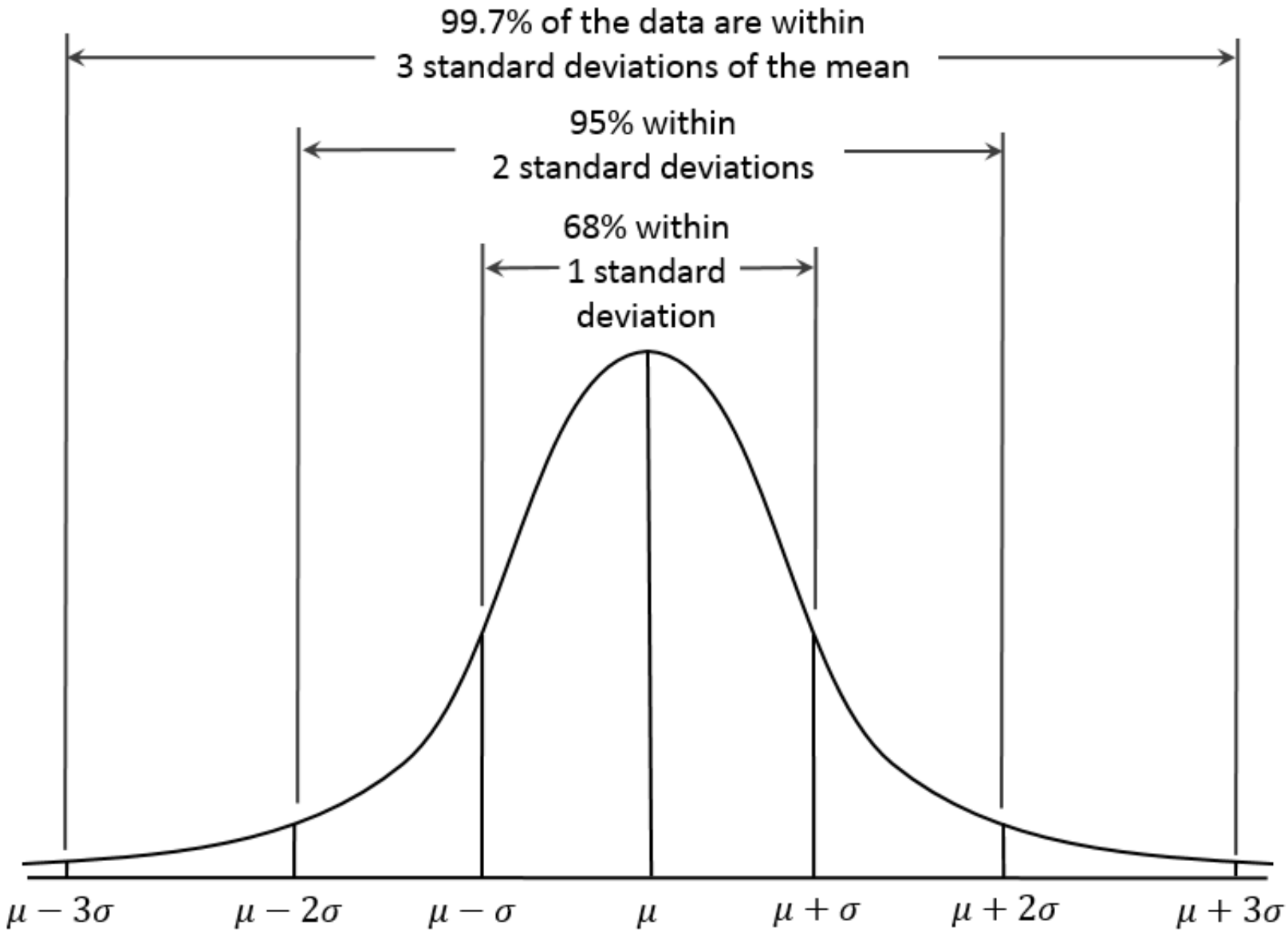
Resembles a normal distribution, but over the positive values, and with only a single parameter.

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 150bp reads do I need?

I need 10Mbp x 24x = 240Mbp of data
240Mbp / 150bp / read = 1.6M reads

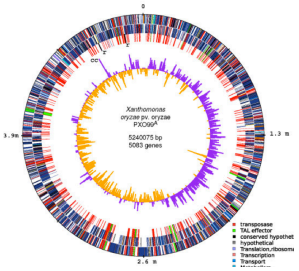
I want to sequence a 10Mbp genome so that
>97.5% of the genome has at least 24x coverage.
How many 150bp reads do I need?

Find X such that X-2*sqrt(X) = 24

36-2*sqrt(36) = 24

I need 10Mbp x 36x = 360Mbp of data
360Mbp / 150bp / read = 2.4M reads

Assembly Summary

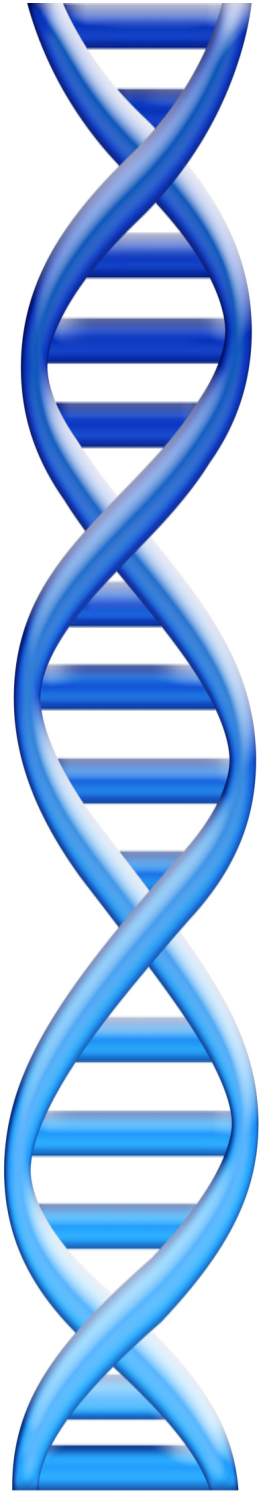


Assembly quality depends on

1. ***Coverage***: low coverage is mathematically hopeless
2. ***Repeat composition***: high repeat content is challenging
3. ***Read length***: longer reads help resolve repeats
4. ***Error rate***: errors reduce coverage, obscure true overlaps

- Assembly is a hierarchical, starting from individual reads, build high confidence contigs/unitigs, incorporate the mates to build scaffolds
  - Extensive error correction is the key to getting the best assembly possible from a given data set

- Watch out for collapsed repeats & other misassemblies
  - Globally/Locally reassemble data from scratch with better parameters & stitch the 2 assemblies together



Project Proposals

Project Proposal

Assignment Date: September 25, 2017
Due Date: Monday, October 2, 2017 @ 11:59pm

Please email a PDF of your project proposal (1/2 to 1 page) to "jhubiomedicalresearch at gmail dot com" by 11:59pm on Monday October 2, 2017.

The proposal should have the following components:

- Short title for your proposal
- Your name
- Email addresses
- Description of what you hope to do and how you will do it:
  - What is the key question you hope to address?
  - What data will you use to study it? Are all the data you need available now? When will they be available?
  - What techniques (experimental or computational) will you use to generate and analyze the data?
  - What are the desired results?
- References to relevant papers
- References/URLs to datasets that you will be studying (Note you can also use simulated data)

After submitting your proposal, we will schedule a time to discuss your proposal, especially to ensure you have access to the data that you need and you have a clear path forward for analyzing it. The sooner that you submit your proposal, the sooner we can schedule the meeting. No late days can be used for this project.

Later, you will present your project in class at the the end of the semester. You will also submit a written report (5-7 pages) of your project, formatting as scholarly article with separate sections for the Abstract, Introduction, Methods, Results, Discussion, and References. More details will be provided later in the semester.