Advances in Genome Sequencing & Assembly

Michael Schatz

July 17, 2018

UCLA Computational Genomics Summer Institute

@mike_schatz

Outline

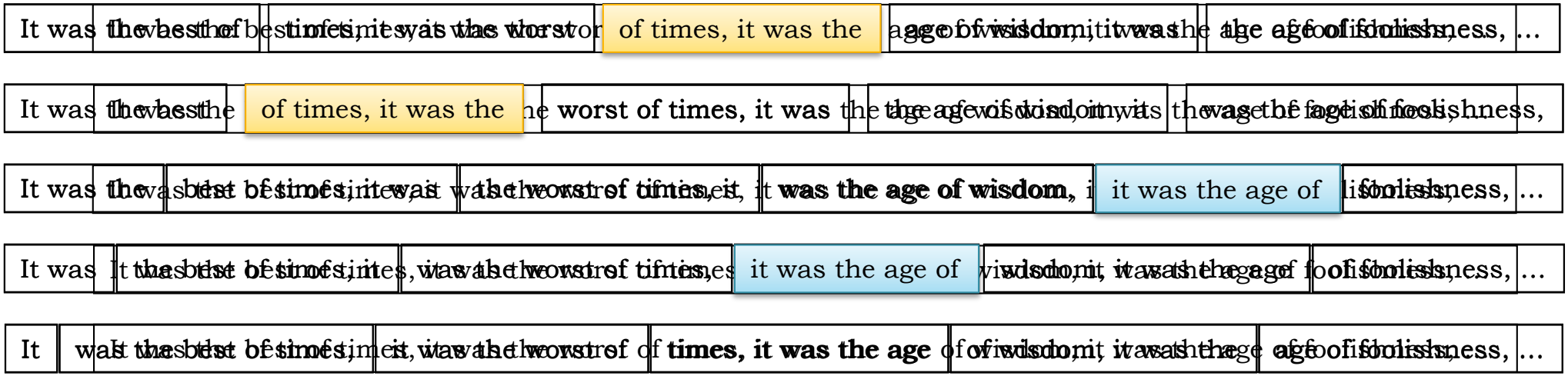
1. ***Introduction to Genome Assembly***
   - Assembly by analogy

2. ***Practical Issues***
   - Coverage, read length, errors, and repeats

3. ***Research Projects***
   - Long read sequencing of breast cancer

Shredded Book Reconstruction

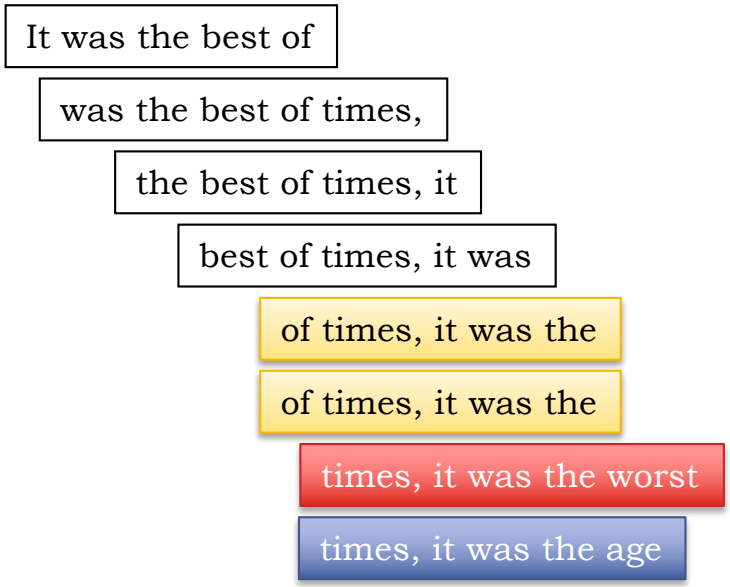
- Dickens accidentally shreds the first printing of A Tale of Two Cities
  - Text printed on 5 long spools



- How can he reconstruct the text?
  - 5 copies x 138, 656 words / 5 words per fragment = 138k fragments
  - The short fragments from every copy are mixed together
  - Some fragments are identical

Greedy Reconstruction

It was the best of
age of wisdom, it was
best of times, it was
it was the age of
it was the age of
it was the worst of
of times, it was the
of times, it was the
of wisdom, it was the
the age of wisdom, it
the best of times, it
the worst of times, it
times, it was the age
times, it was the worst
was the age of wisdom,
was the age of foolishness,
was the best of times,
was the worst of times,
wisdom, it was the age
worst of times, it was



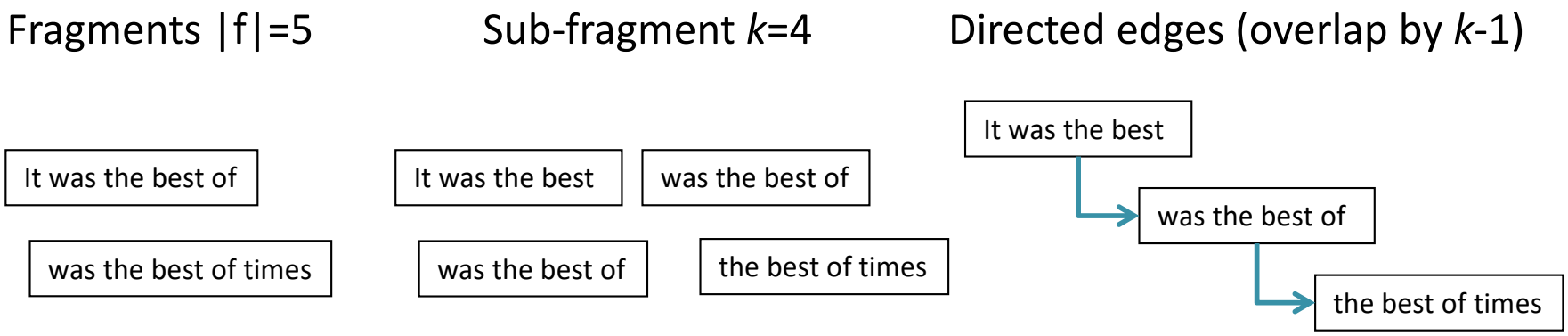
The repeated sequence make the correct reconstruction ambiguous

- It was the best of times, it was the [worst/age]

Model the assembly problem as a graph problem

de Bruijn Graph Construction

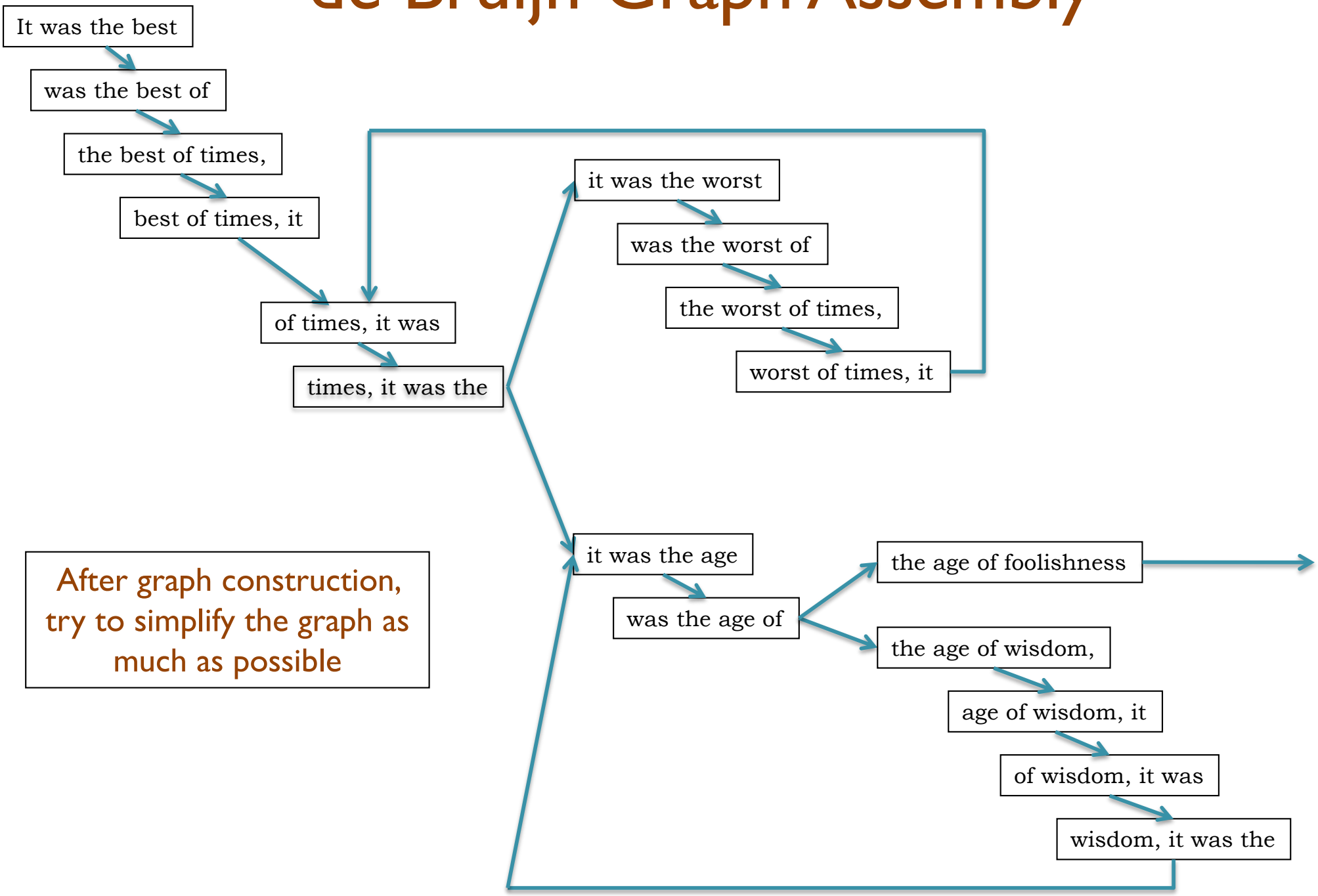
- $G_k = (V,E)$
  - V = Length-*k* sub-fragments
  - E = Directed edges between consecutive sub-fragments
    - Sub-fragments overlap by *k*-1 words



– Overlaps between fragments are implicitly computed

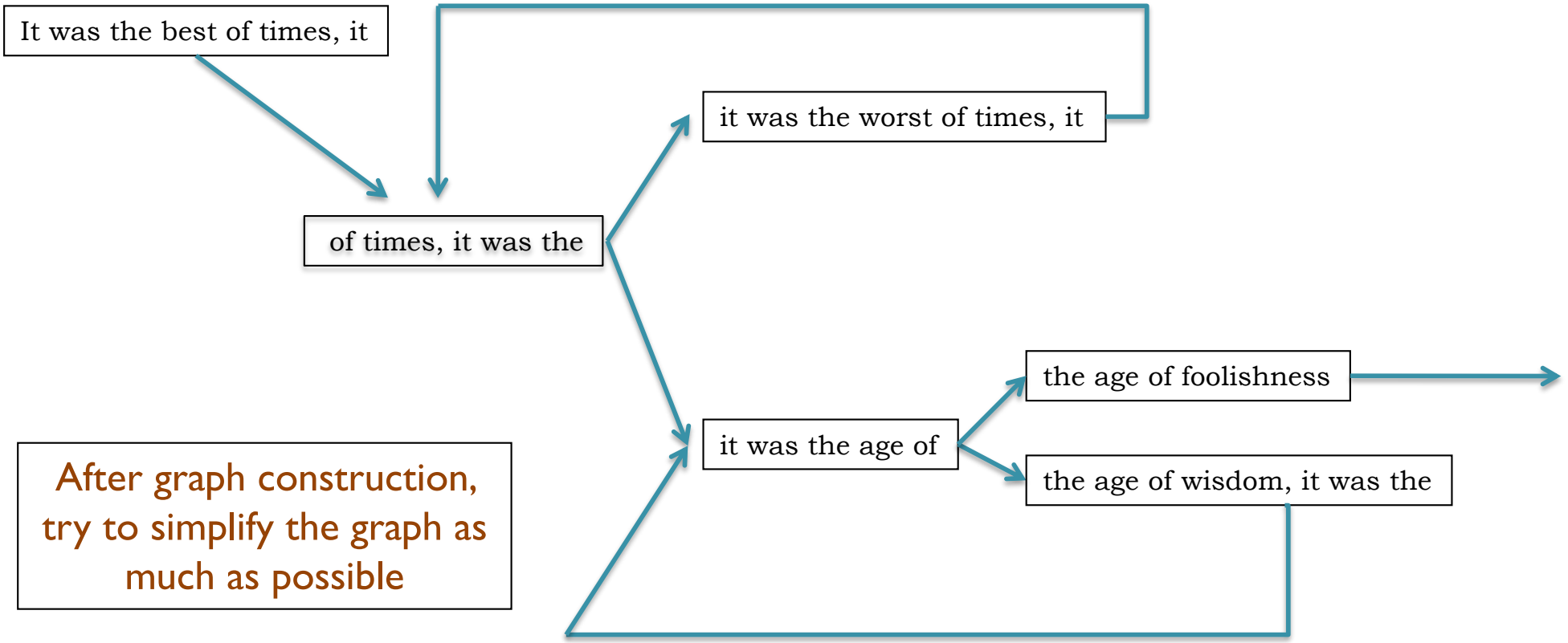
de Bruijn, 1946
Idury et al., 1995
Pevzner et al., 2001

de Bruijn Graph Assembly



After graph construction, try to simplify the graph as much as possible

Compacted de Bruijn Graph



After graph construction, try to simplify the graph as much as possible

The full tale

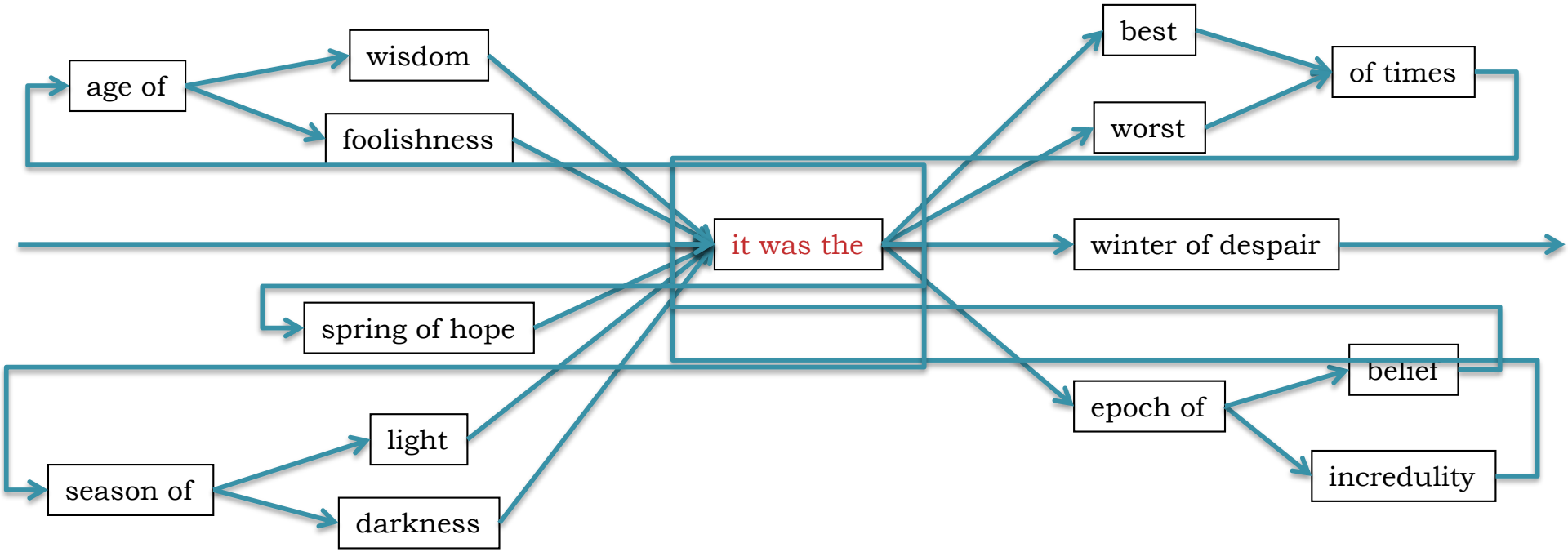
… it was the best of times it was the worst of times …

… it was the age of wisdom it was the age of foolishness …

… it was the epoch of belief it was the epoch of incredulity …

… it was the season of light it was the season of darkness …

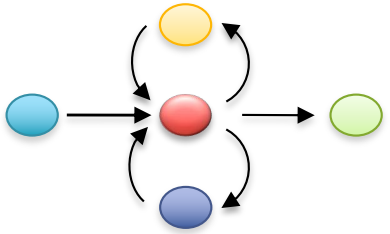
… it was the spring of hope it was the winder of despair …



Assembly Complexity

Finding possible assembly paths is easy!

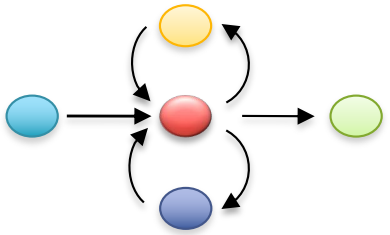
- Eulerian tour in linear time ☺



Assembly Complexity

Finding possible assembly paths is easy!

- Eulerian tour in linear time ☺



However, there is a ~~astronomical~~ ***genomical*** number of possible paths!

- Proportional to the product of the factorial of the degree of the nodes

Kingsford, Schatz, Pop (2010) *BMC Bioinformatics* 11:21

- Alternative formulations related to the shortest-common-superstring problem are NP-hard

Computability of Models for Sequence Assembly

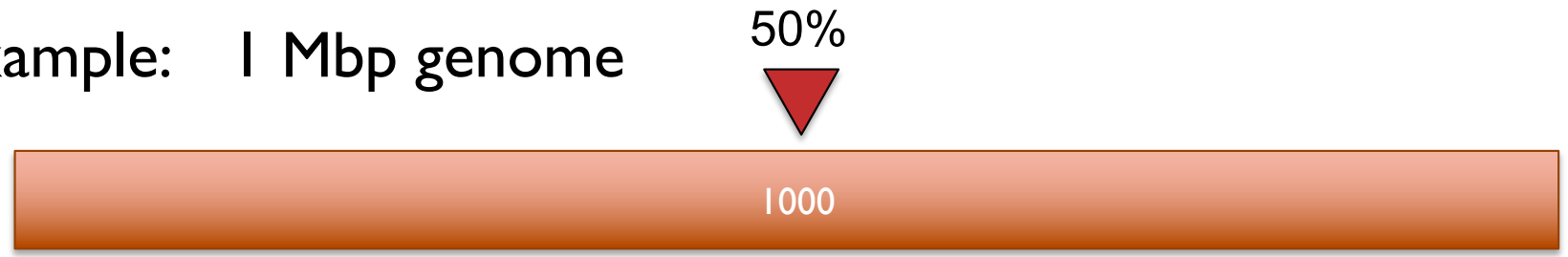
Medvedev et al (2007) *Algorithms in Bioinformatics. 978-3-540-74126-8*

Hopeless to figure out the whole genome/chromosome (with short reads): figure out the parts that you can

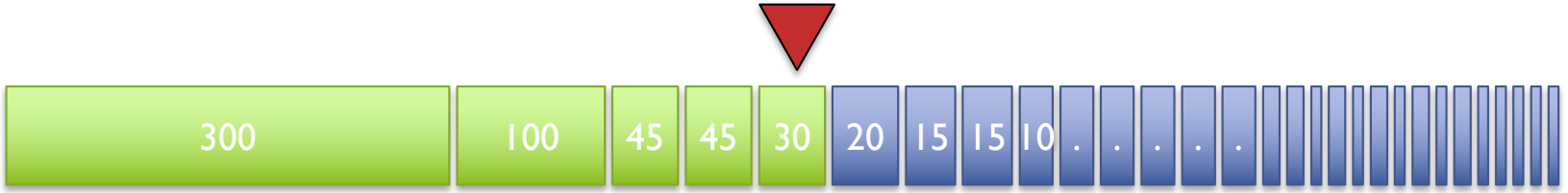
Contig N50

Def: 50% of the genome is in contigs as large as the N50 value

Example: 1 Mbp genome

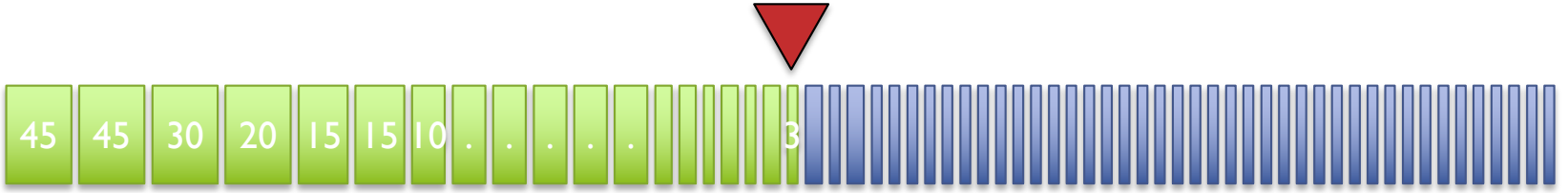


A



N50 size = 30 kbp

B



N50 size = 3 kbp

Contig N50

Def: 50% of the genome is in contigs as large as the N50 value

50%

Better N50s improves the analysis in every dimension

- Better resolution of genes and flanking regulatory regions
- Better resolution of transposons and other complex sequences
- Better resolution of chromosome organization
- Better sequence for all downstream analysis

Just be careful of N50 inflation!

- *A very very very bad assembler in 1 line of bash:*
- *cat *.reads.fa > genome.fa*

N50 size = 3 kbp

Outline

1. ***Introduction to Genome Assembly***
   - Assembly by analogy

2. ***Practical Issues***
   - Coverage, read length, errors, and repeats

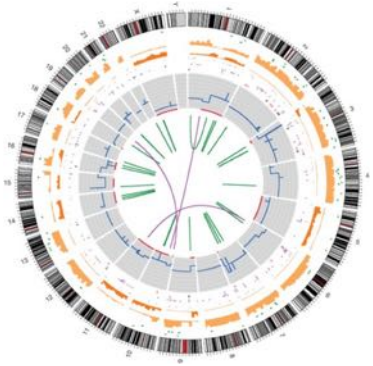
3. ***Research Projects***
   - Long read sequencing of breast cancer

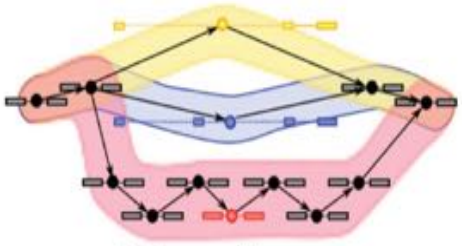
Assembly Applications

- Novel genomes

- Metagenomes

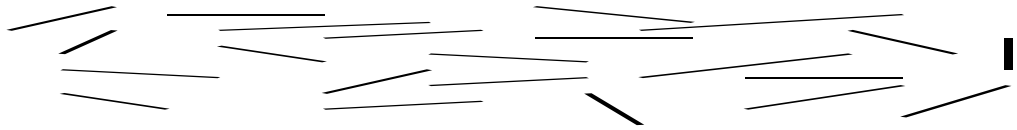
- Sequencing assays
  - Structural variations
  - Transcript assembly
  - …





Assembling a Genome

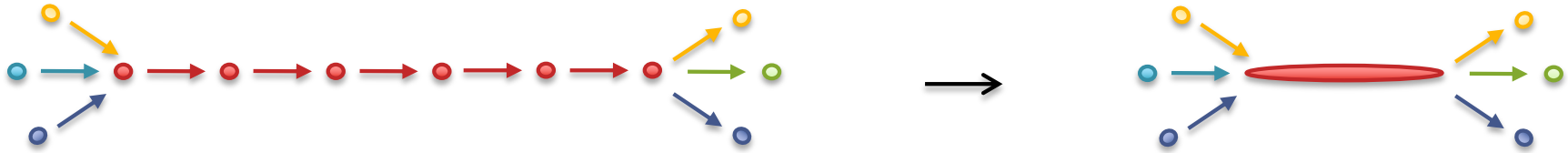
1. Shear & Sequence DNA



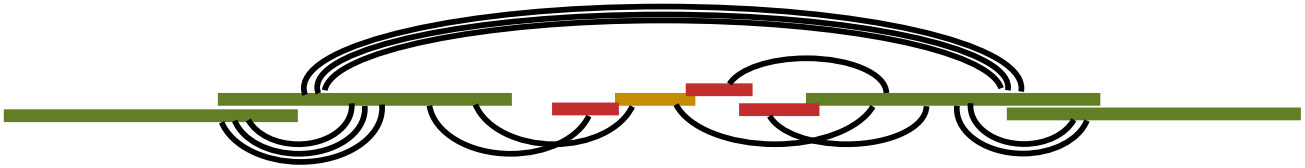
2. Construct assembly graph from reads (de Bruijn / overlap graph)

...AGCCTAGGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
CAACCTCGGACGGACCTCAGCGAA...

3. Simplify assembly graph

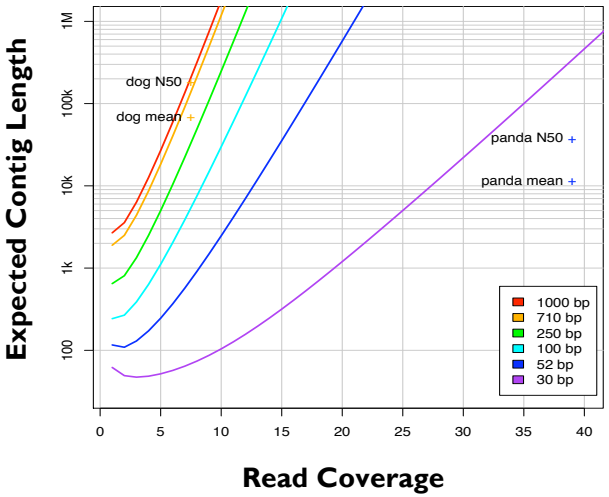


4. Detangle graph with long reads, mates, and other links



Ingredients for a good assembly

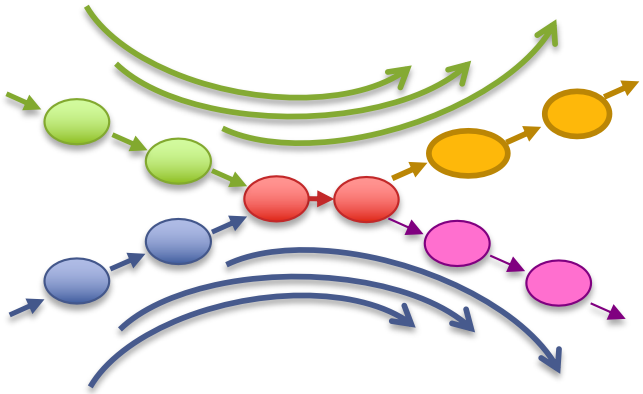
Coverage



High coverage is required

- Oversample the genome to ensure every base is sequenced with long overlaps between reads
- Biased coverage will also fragment assembly

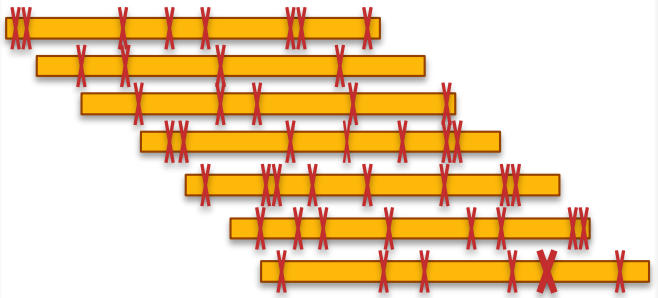
Read Length



Reads & mates must be longer than the repeats

- Short reads will have ***false overlaps*** forming hairball assembly graphs
- With long enough reads, assemble entire chromosomes into contigs

Quality

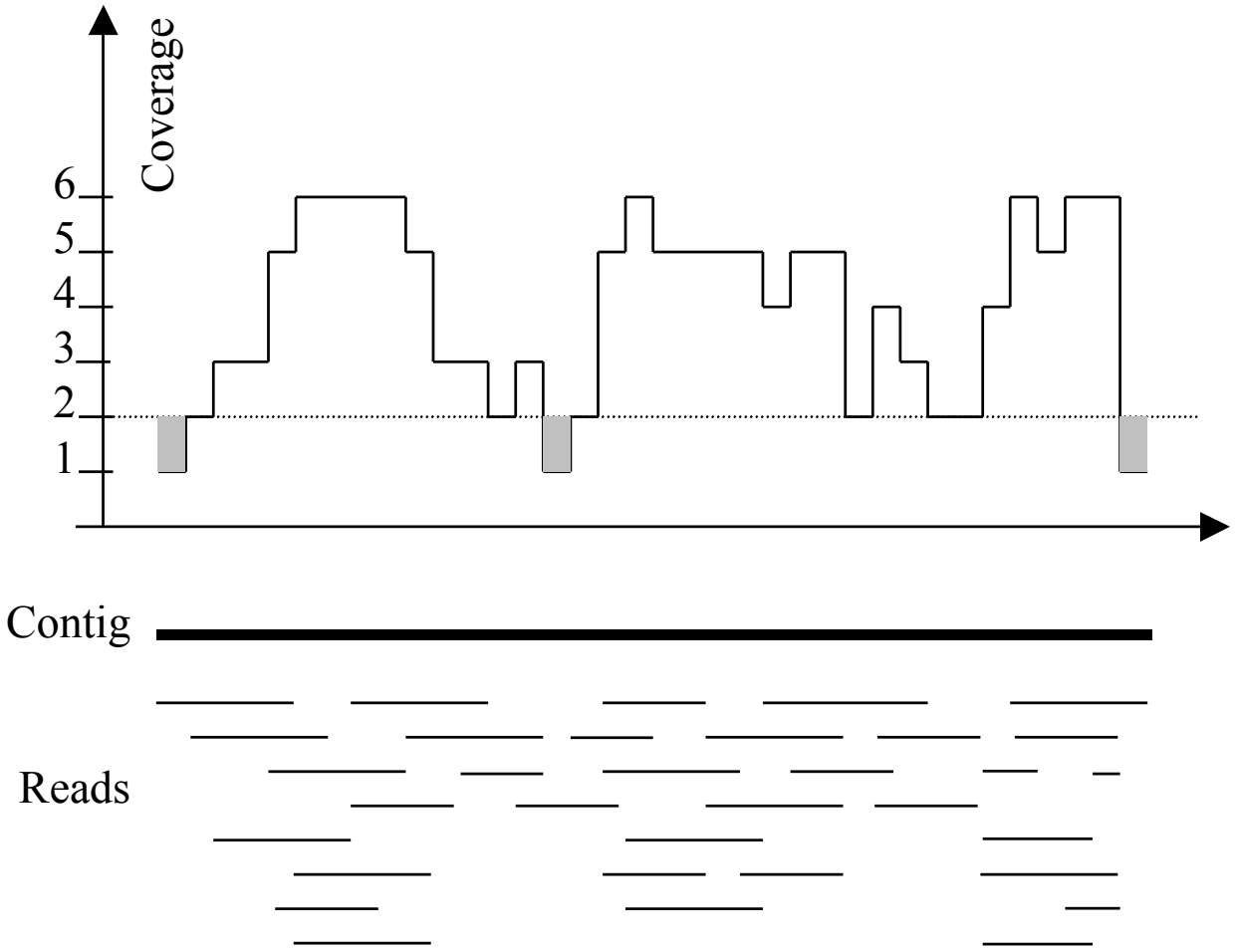


Errors obscure overlaps

- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Current challenges in *de novo* plant genome sequencing and assembly
Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology*. 12:243

Typical sequencing coverage

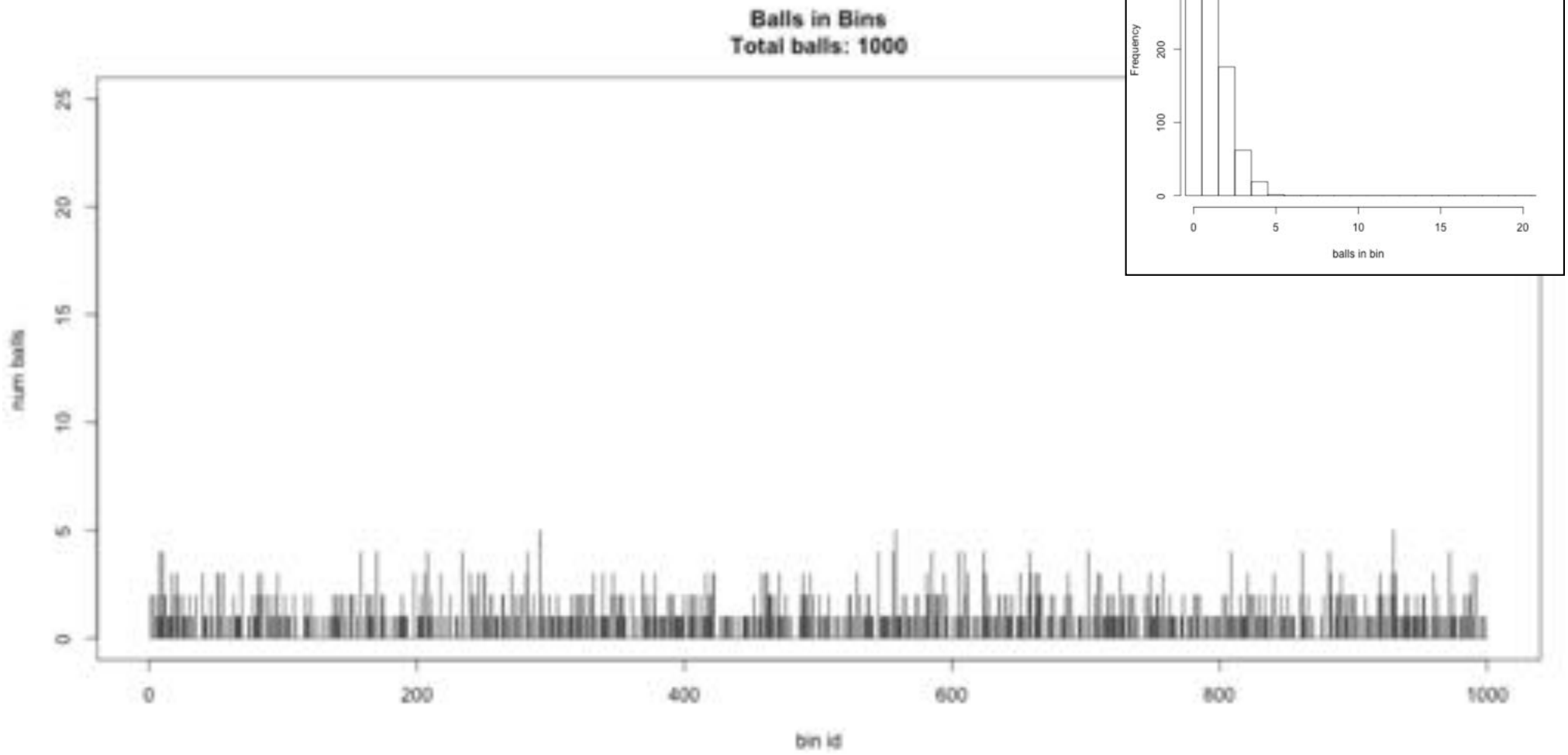


Imagine raindrops on a sidewalk

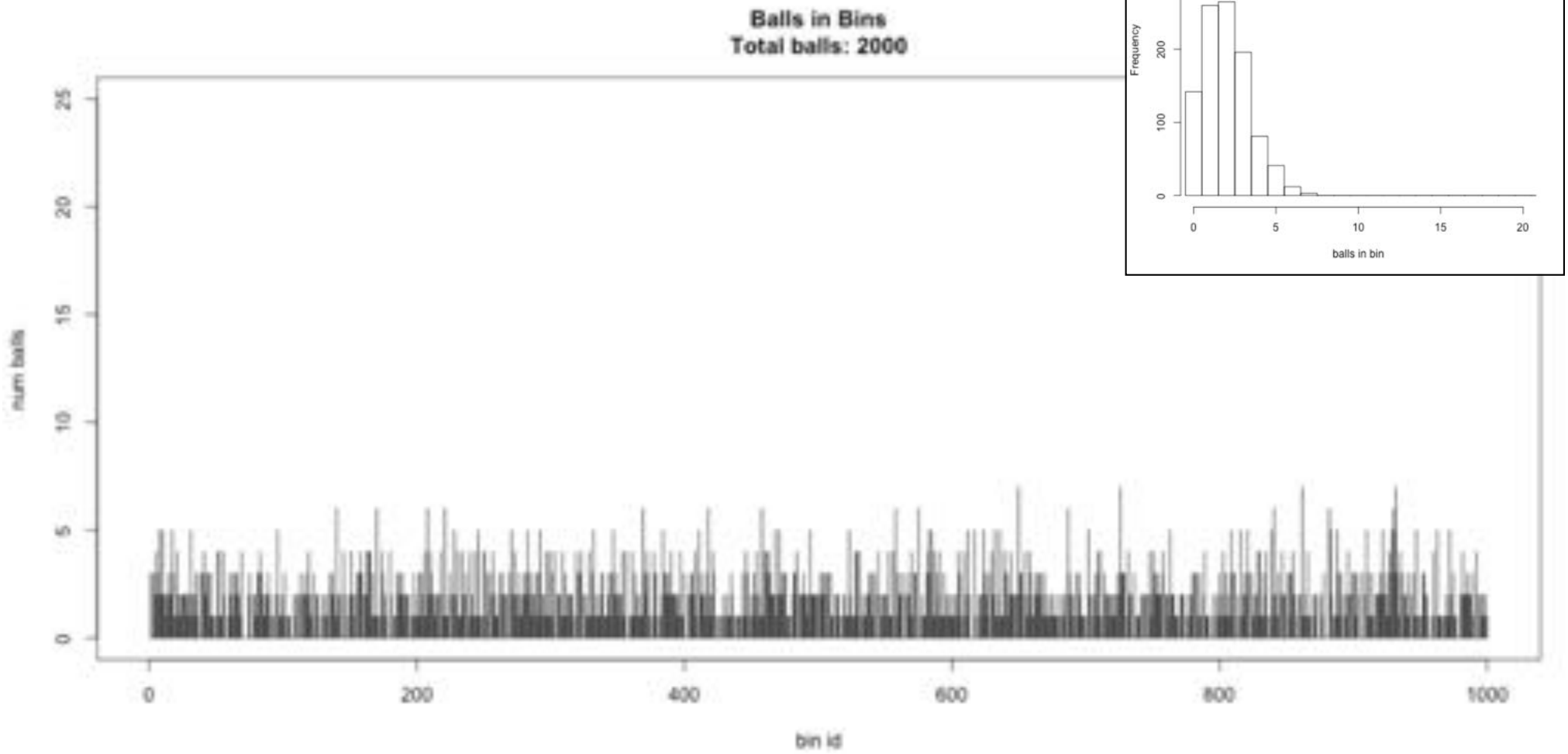
We want to cover the entire sidewalk but each drop costs $1

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

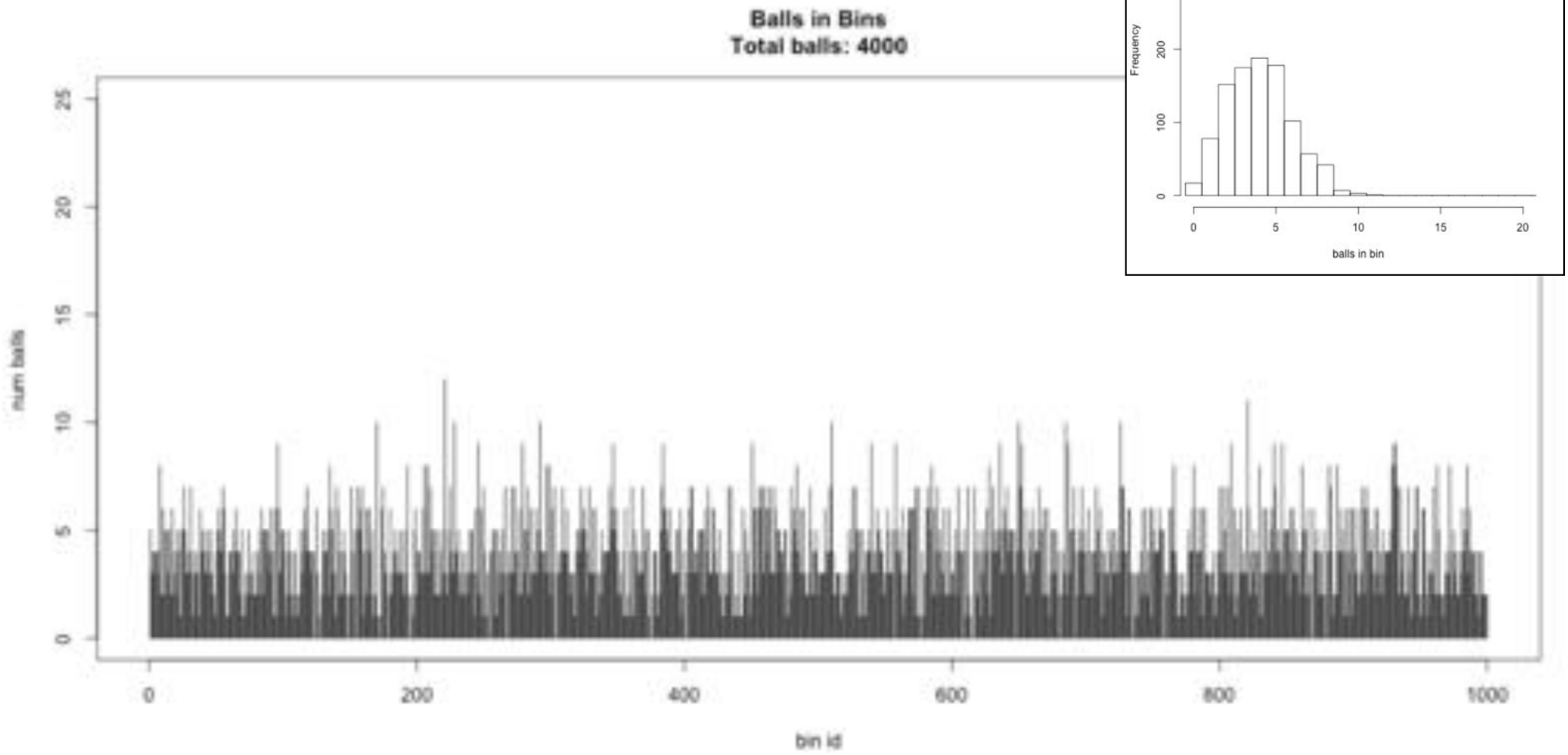
1x sequencing



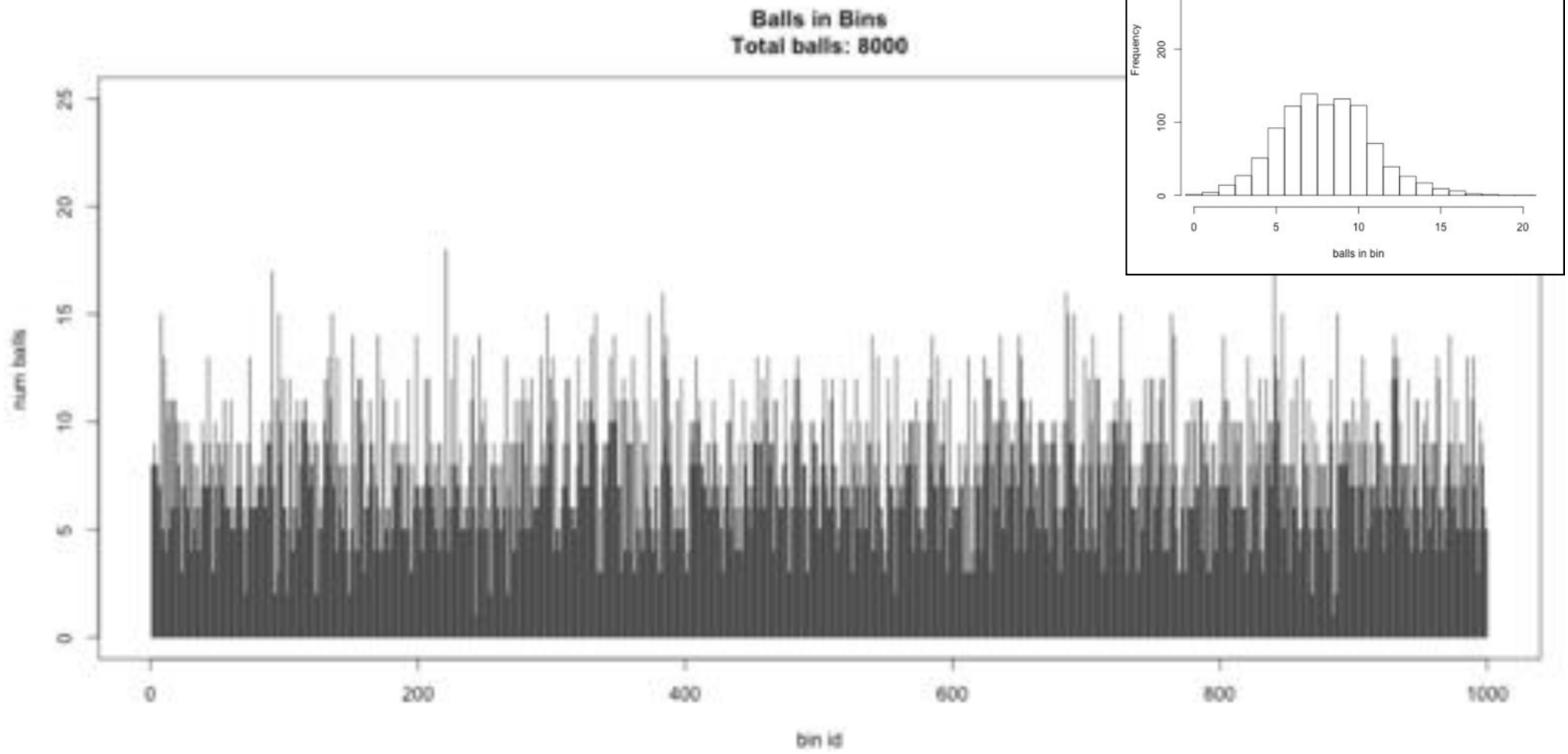
2x sequencing



4x sequencing



8x sequencing



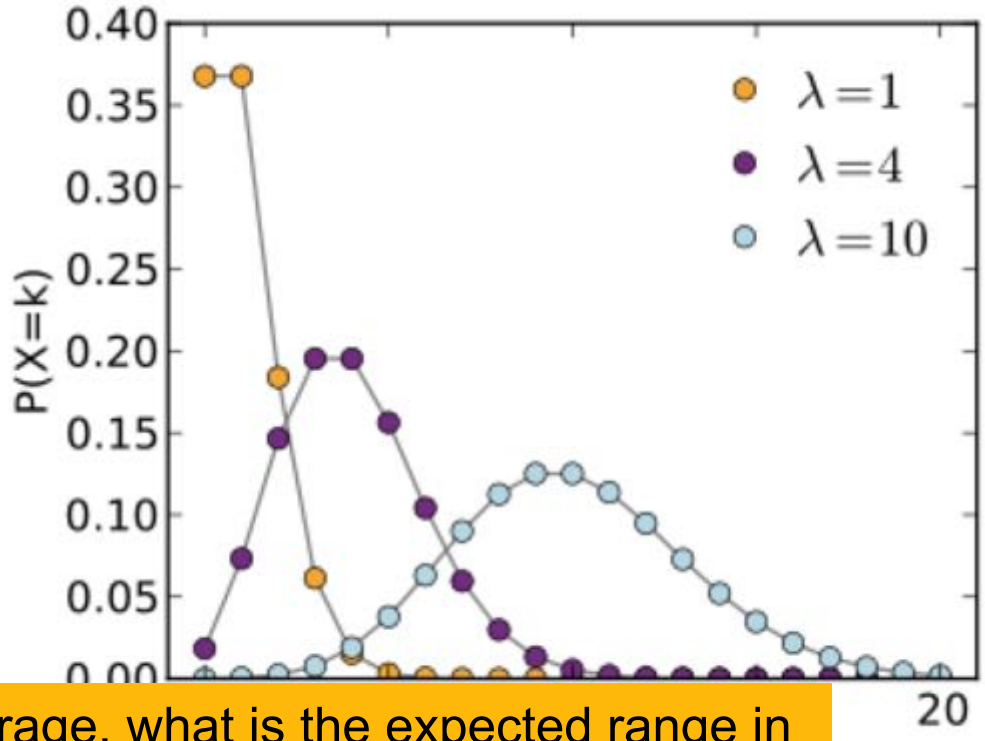
Poisson Distribution

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.

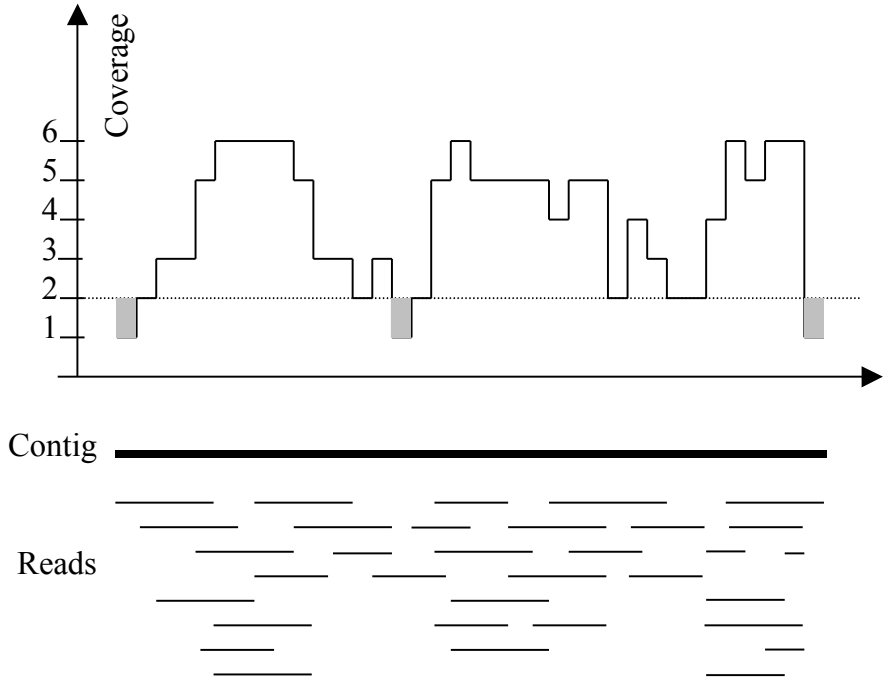


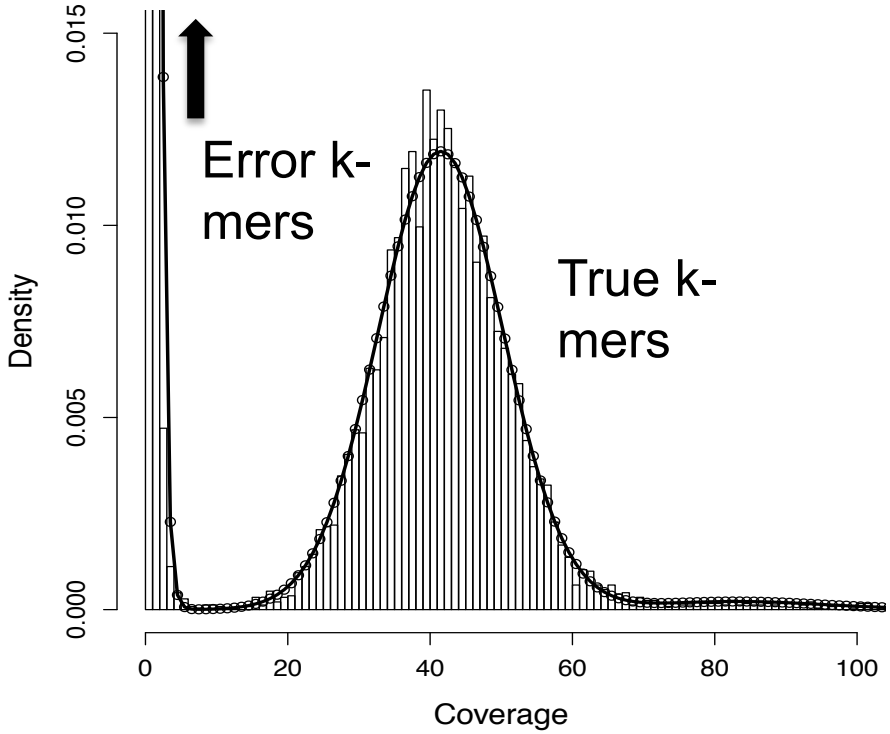
Key properties:

- ***The standard deviation is the sq***
- ***Fo***
  by a normal distribution

If I have an average of 100x coverage, what is the expected range in coverage to 2 standard deviations?

Kmer-based Coverage Analysis

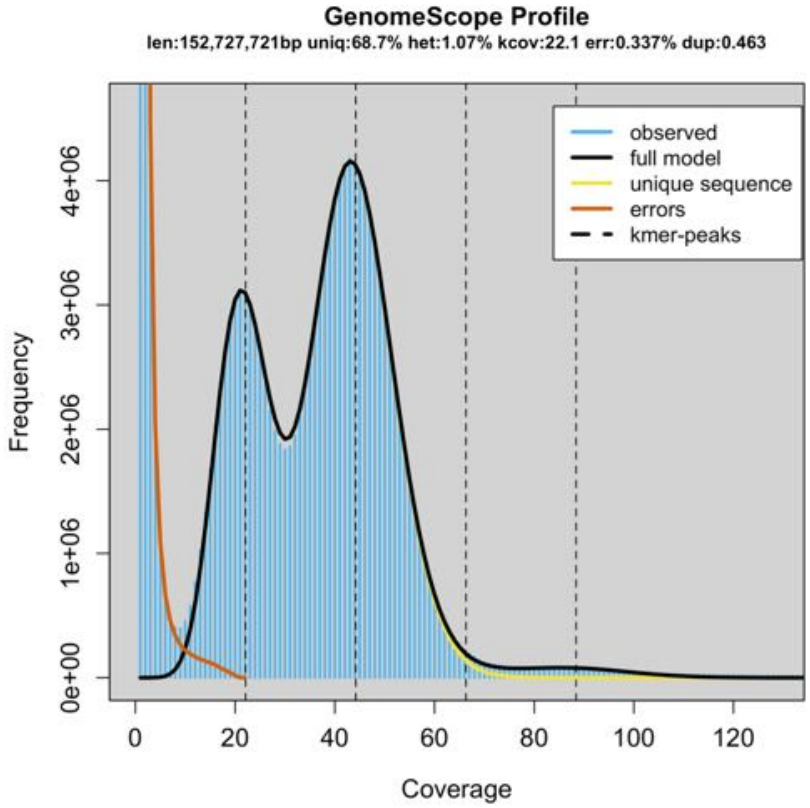


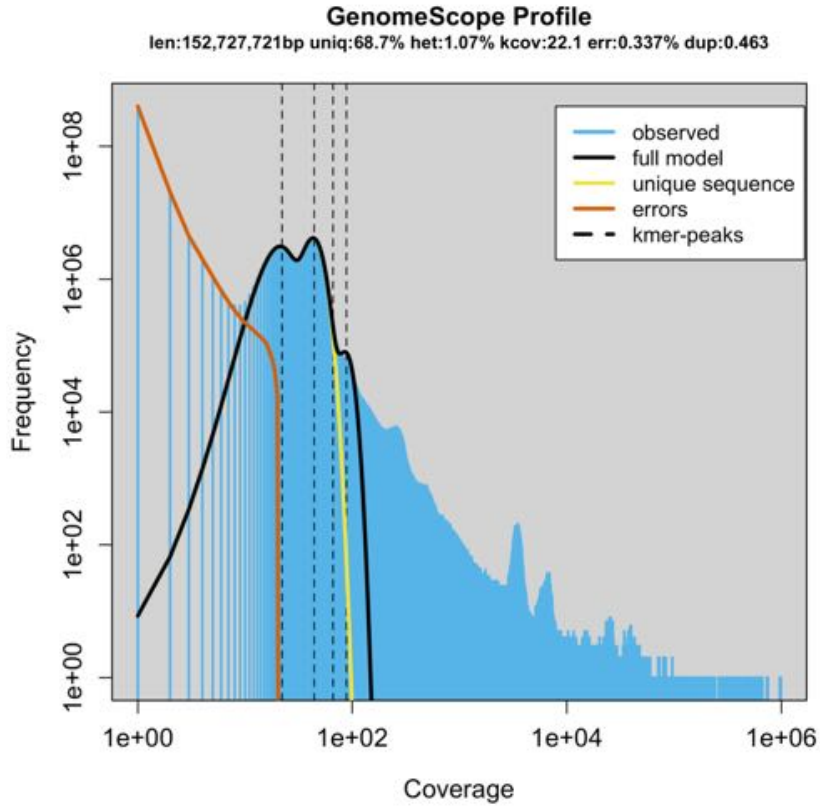


Even though the reads are not assembled or aligned (or reference available), Kmer counting is an effective technique to estimate coverage & errors

Quake: quality-aware detection and correction of sequencing reads.
Kelley, DR, Schatz, MC, Salzberg SL (2010) *Genome Biology*. 11:R116

GenomeScope: Fast reference-free genome profiling from short reads
http://qb.cshl.edu/genomescope/





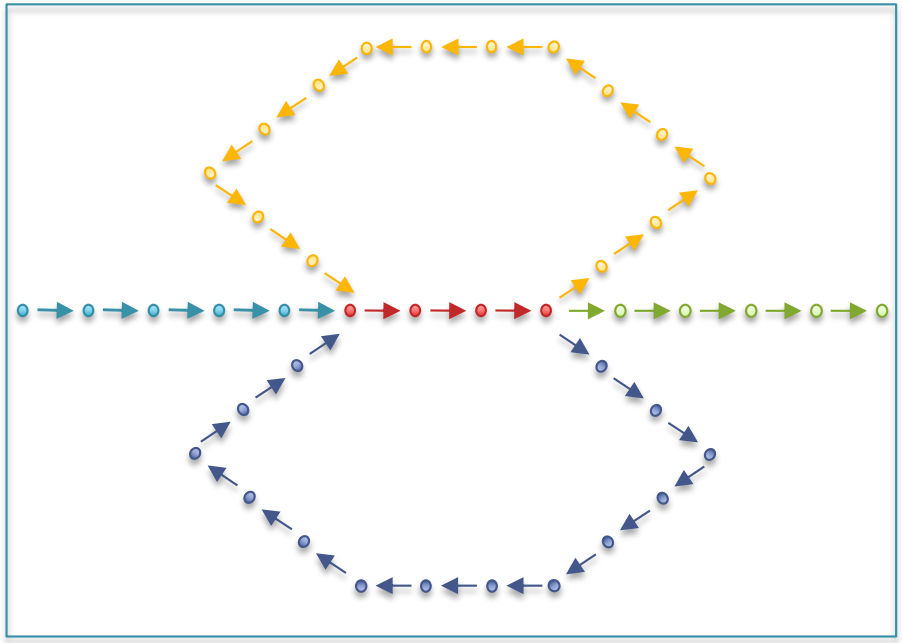
Automatically estimate several genome properties from unassembled reads

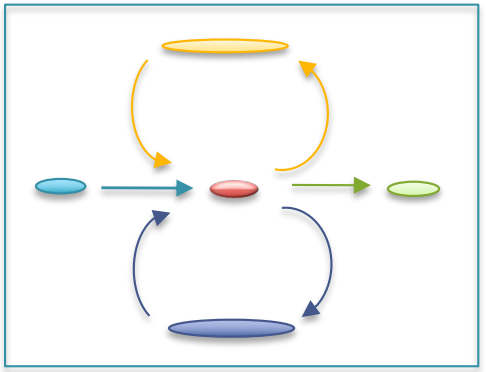
- Genome size
- Repetitiveness
- Rate of heterozygosity
- Effective Coverage
- Sequencing Error Rate
- Rate of PCR Duplicates

Vurture et al. (2017) Bioinformatics. doi: https://doi.org/10.1093/bioinformatics/btx153

Unitigging / Unipathing

- After simplification and correction, compress graph down to its non-branching initial contigs
  - Aka "unitigs", "unipaths"





Why do unitigs / unipaths end?

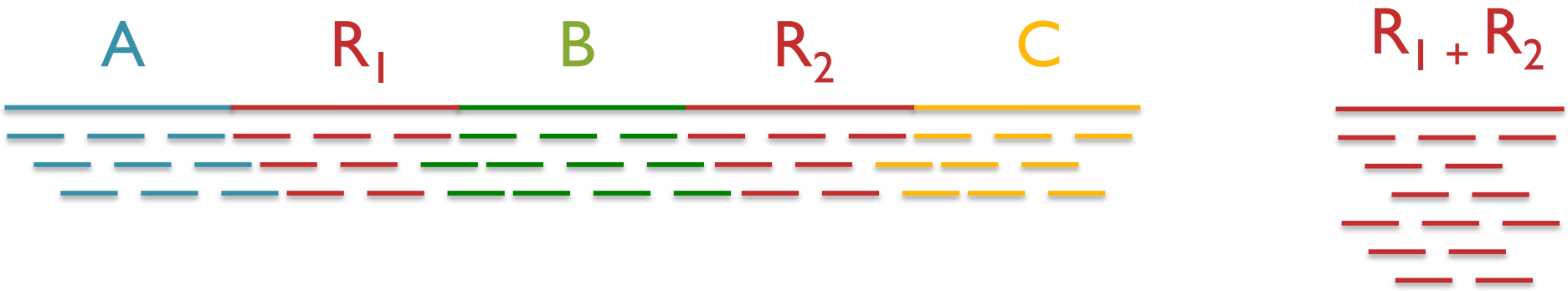
(1) lack of coverage, (2) errors, (3) heterozygosity and (4) repeats

Repetitive regions

Repeat Type	Definition / Example	Prevalence
Low-complexity DNA / Microsatellites	$(b_1 b_2 \ldots b_k)^N$ where $1 \leq k \leq 6$ CACACACACACACACACACA	2%
SINEs (Short Interspersed Nuclear Elements)	*Alu* sequence (~280 bp) Mariner elements (~80 bp)	13%
LINEs (Long Interspersed Nuclear Elements)	~500 – 5,000 bp	21%
LTR (long terminal repeat) retrotransposons	Ty1-copia, Ty3-gypsy, Pao-BEL (~100 – 5,000 bp)	8%
Other DNA transposons		3%
Gene families & segmental duplications		4%

- Over 50% of mammalian genomes are repetitive
  - Large plant genomes tend to be even worse
  - Wheat: 16 Gbp; Pine: 24 Gbp

Repeats and Coverage Statistics



- If *n* reads are a uniform random sample of the genome of length *G*, we expect $k=n\Delta/G$ reads to start in a region of length Δ.
  - If we see many more reads than k (if the arrival rate is > A) , it is likely to be a collapsed repeat

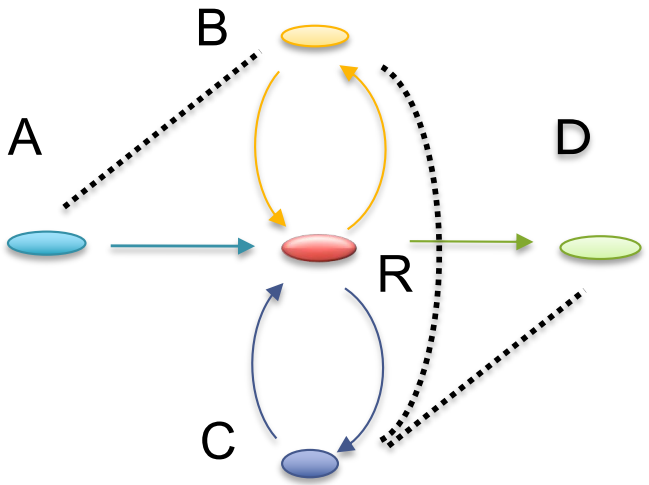
$$\Pr(X-copy)=\binom{n}{k}\left(\frac{X\Delta}{G}\right)^{k}\left(\frac{G-X\Delta}{G}\right)^{n-k}$$

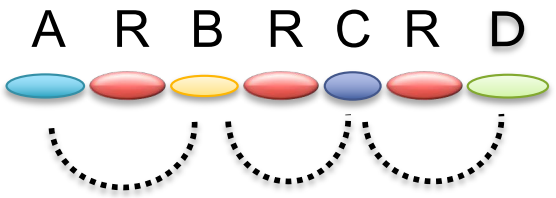
$$A(\Delta,k)=\ln\left(\frac{\Pr(1-copy)}{\Pr(2-copy)}\right)=\ln\left(\frac{\frac{(\Delta n/G)^{k}}{k!}e^{\frac{-\Delta n}{G}}}{\frac{(2\Delta n/G)^{k}}{k!}e^{\frac{-2\Delta n}{G}}}\right)=\frac{n\Delta}{G}-k\ln 2$$

The fragment assembly string graph
Myers, EW (2005) Bioinformatics. 21(suppl 2): ii79-85.

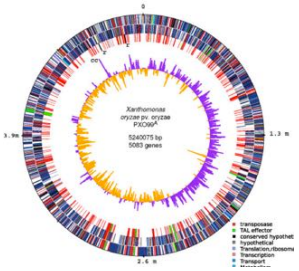
Scaffolding

- Initial contigs (*aka* unipaths, unitigs) terminate at
  - *Coverage gaps*: especially extreme GC
  - *Conflicts*: errors, repeat boundaries
- Use mate-pairs/linked-reads/HiC/ optical maps to resolve correct order through assembly graph
  - Place sequence to satisfy the mate constraints
  - Mates through repeat nodes are tangled
- Final scaffold may have internal gaps called sequencing gaps
  - We know the order, orientation, and spacing, but just not the bases. Fill with Ns instead





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

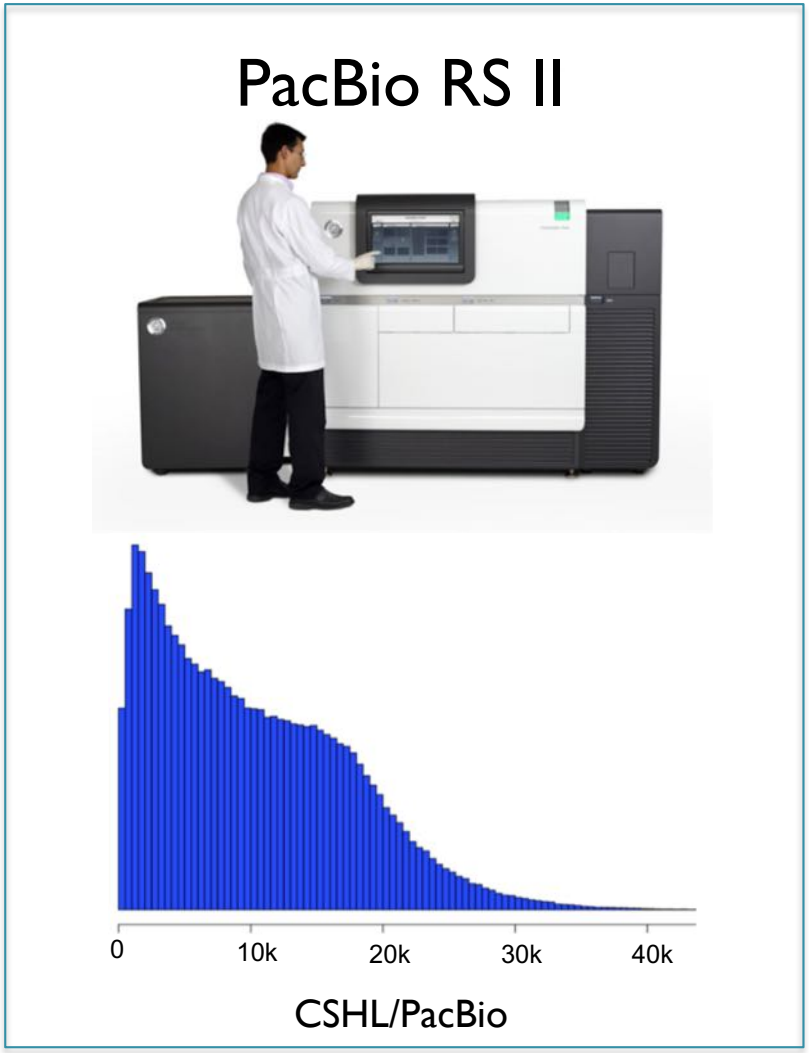
Outline

1. ***Introduction to Genome Assembly***
   - Assembly by analogy

2. ***Practical Issues***
   - Coverage, read length, errors, and repeats

3. ***Research Projects***
   - Long read sequencing of breast cancer

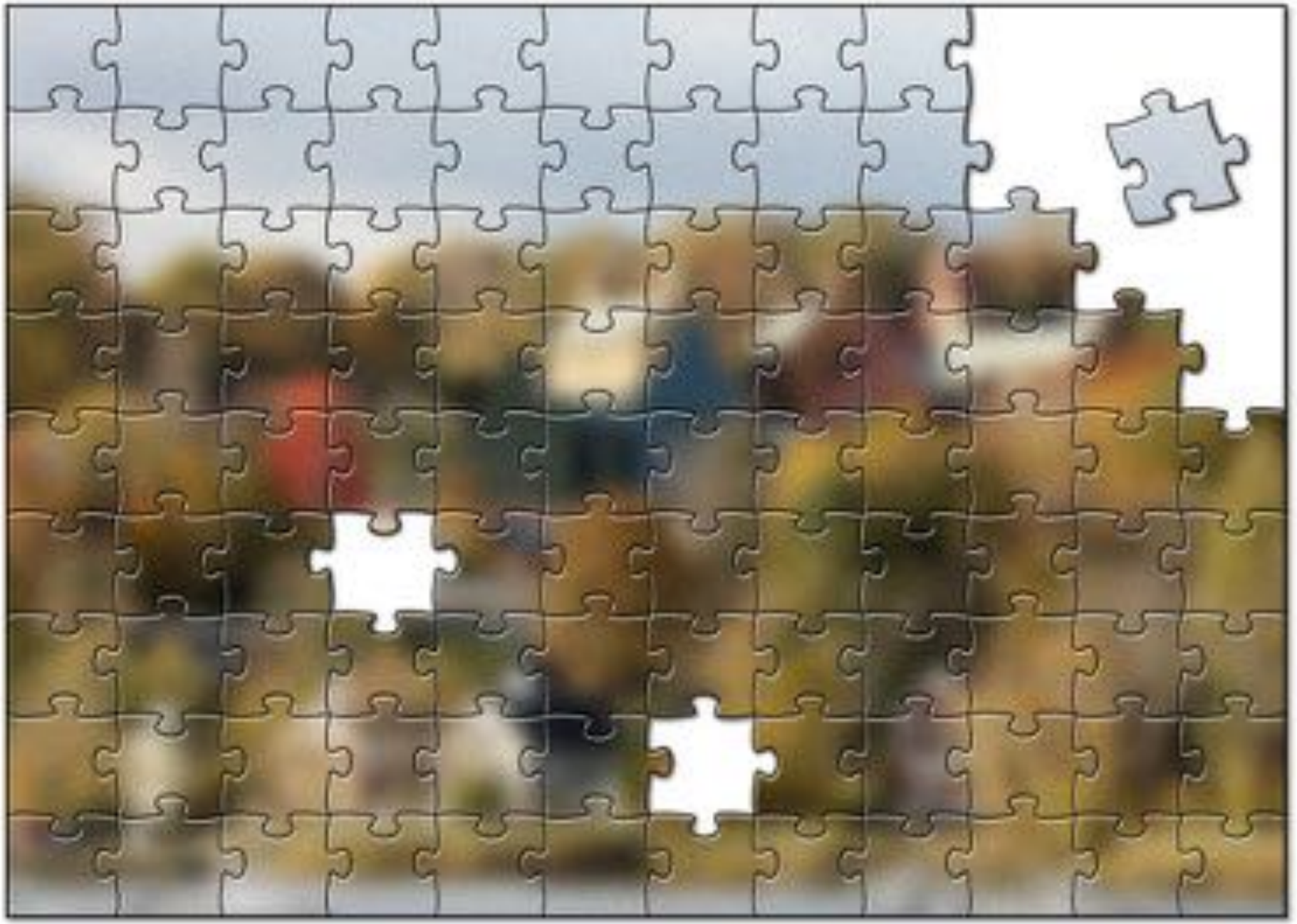
Single Molecule Sequencing



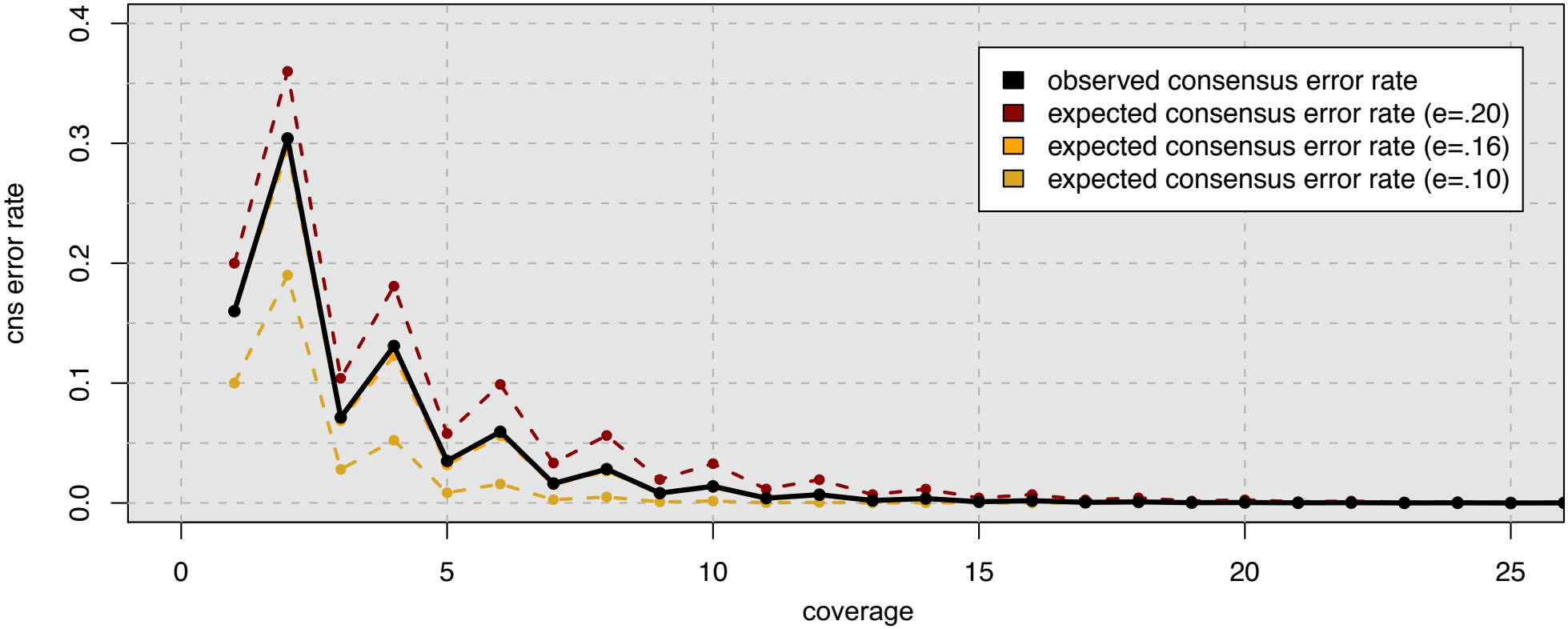

```
TTGTAAGCAGTTGAAAACTATGTGTGGATTTAGAATAAAGAACATGAAAG
|||||||||||||||||||||||||| ||||||| |||||||||||| |||
TTGTAAGCAGTTGAAAACTATGTGT-GATTTAG-ATAAAGAACATGGAAG

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

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

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

-AGGAGGGGAAAGGGGGGAATATCT-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

Single Molecule Sequences



Consensus Accuracy and Coverage



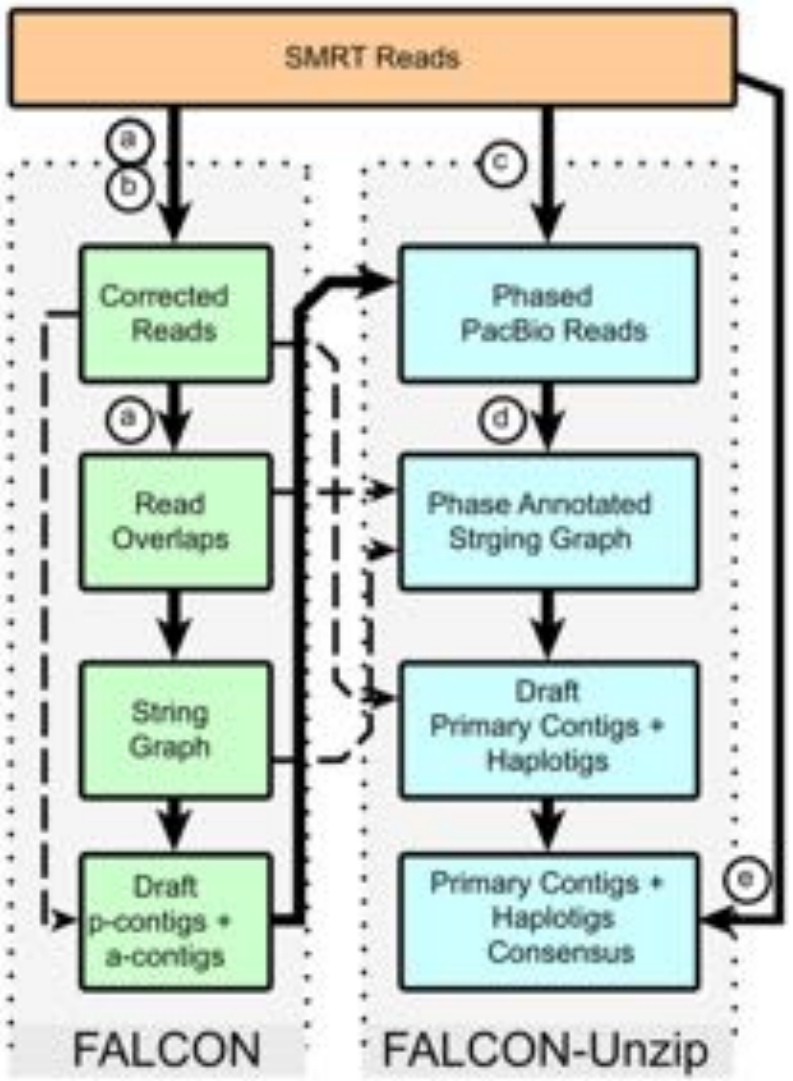
Coverage can overcome random errors

- Dashed: error model from binomial sampling
- Solid: observed accuracy

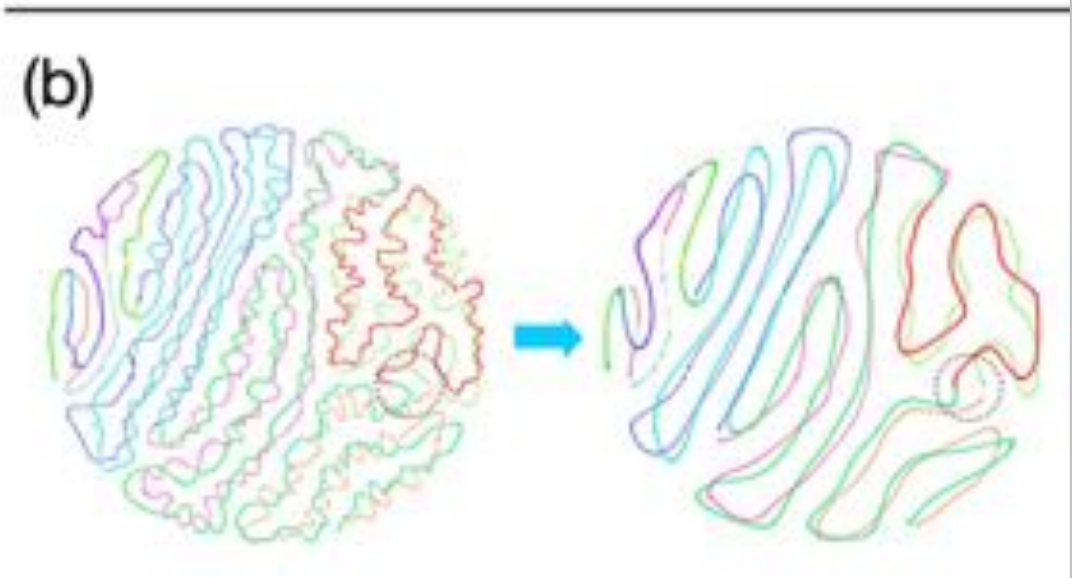
$$CNS\ Error = \sum_{i=\lceil c/2 \rceil}^{c} \binom{c}{i} (e)^{i} (1-e)^{n-i}$$

Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Koren et al (2012) *Nature Biotechnology. doi:10.1038/nbt.2280*

FALCON-unzip Accuracy

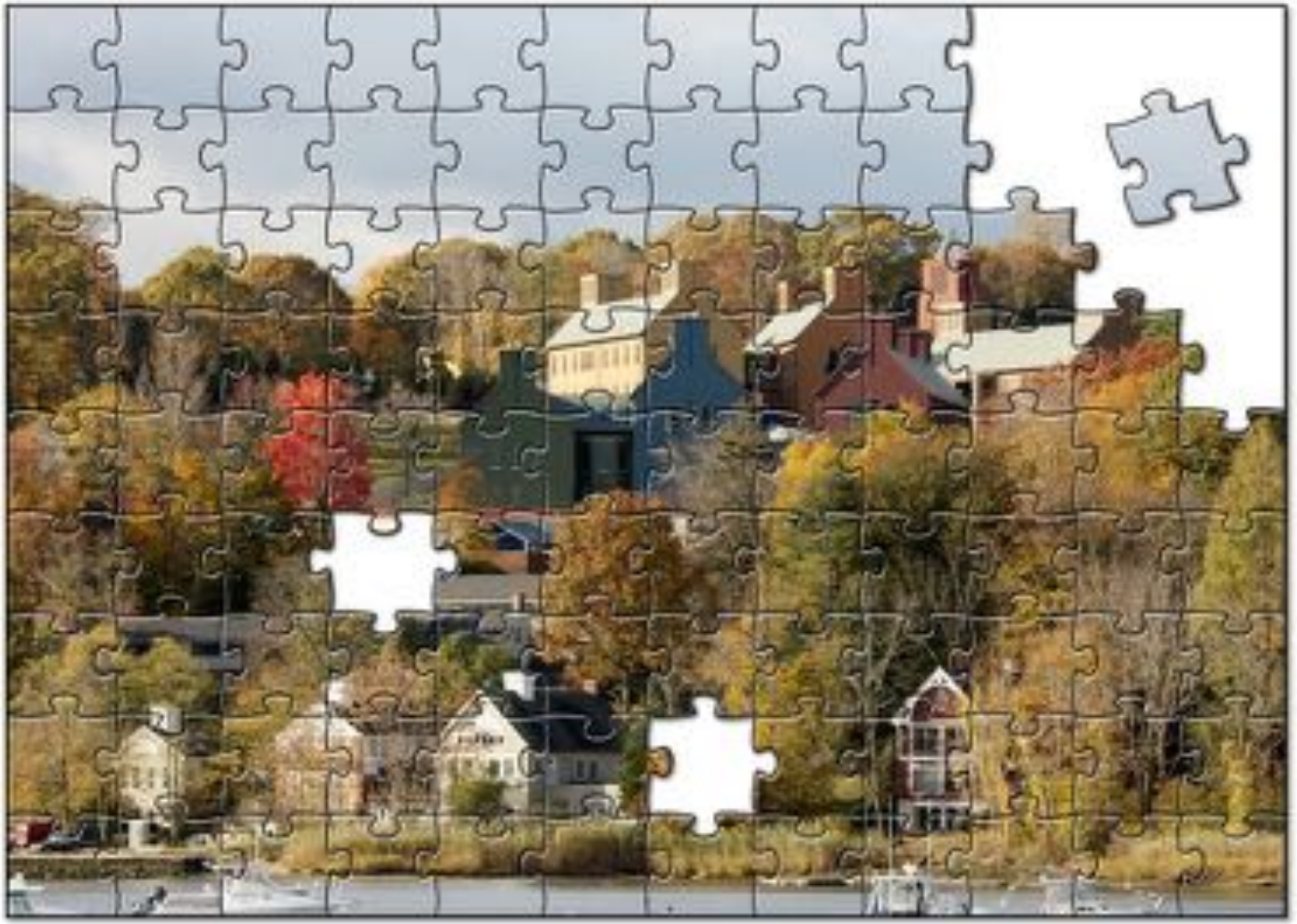


"*The overall base-to-base concordance rate is about 99.99%* (QV40 in Phred scale) in the F1 FALCON-Unzip assembly. The insertion and deletion (indel) concordances to the parental lines were lower (about QV40) than the SNP concordance rate (about QV50), with most residual errors concentrated in long homopolymer sequences"

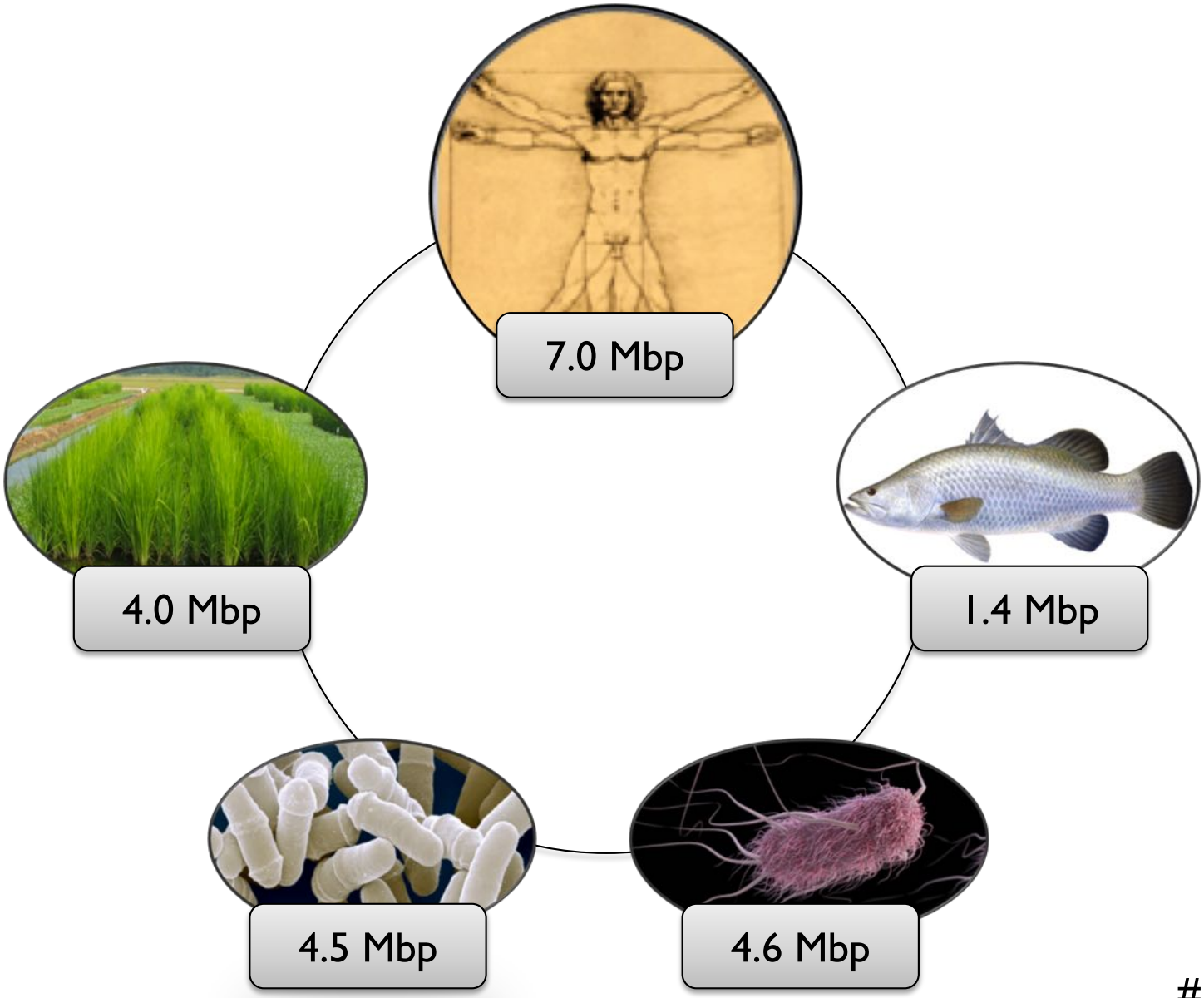


Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing
Chin et al (2016) *Nature Methods. doi:10.1038/nmeth.4035.*

“Corrective Lens” for Sequencing

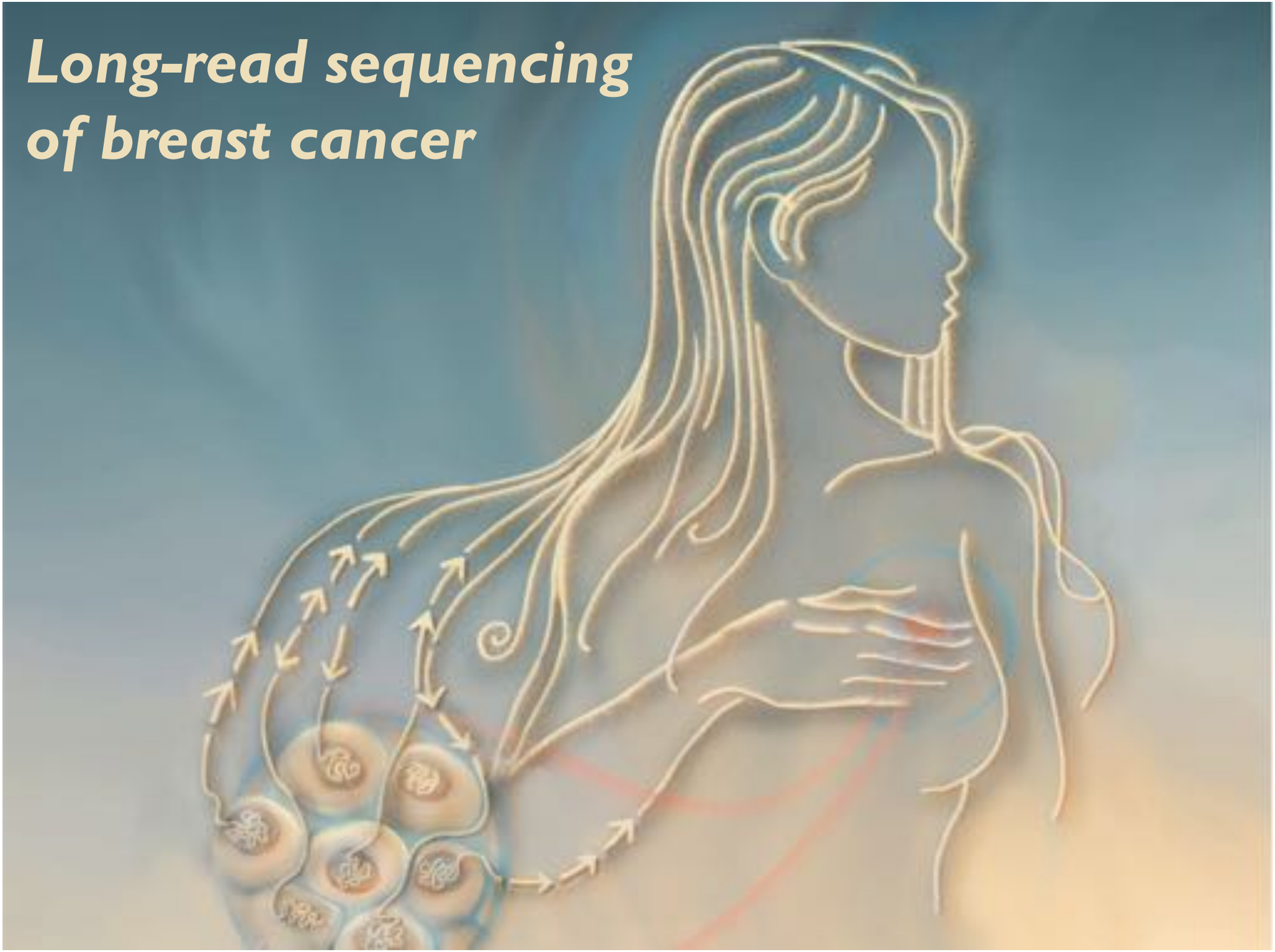


(A few) Recent Long Read Assemblies



#1mbctgclub

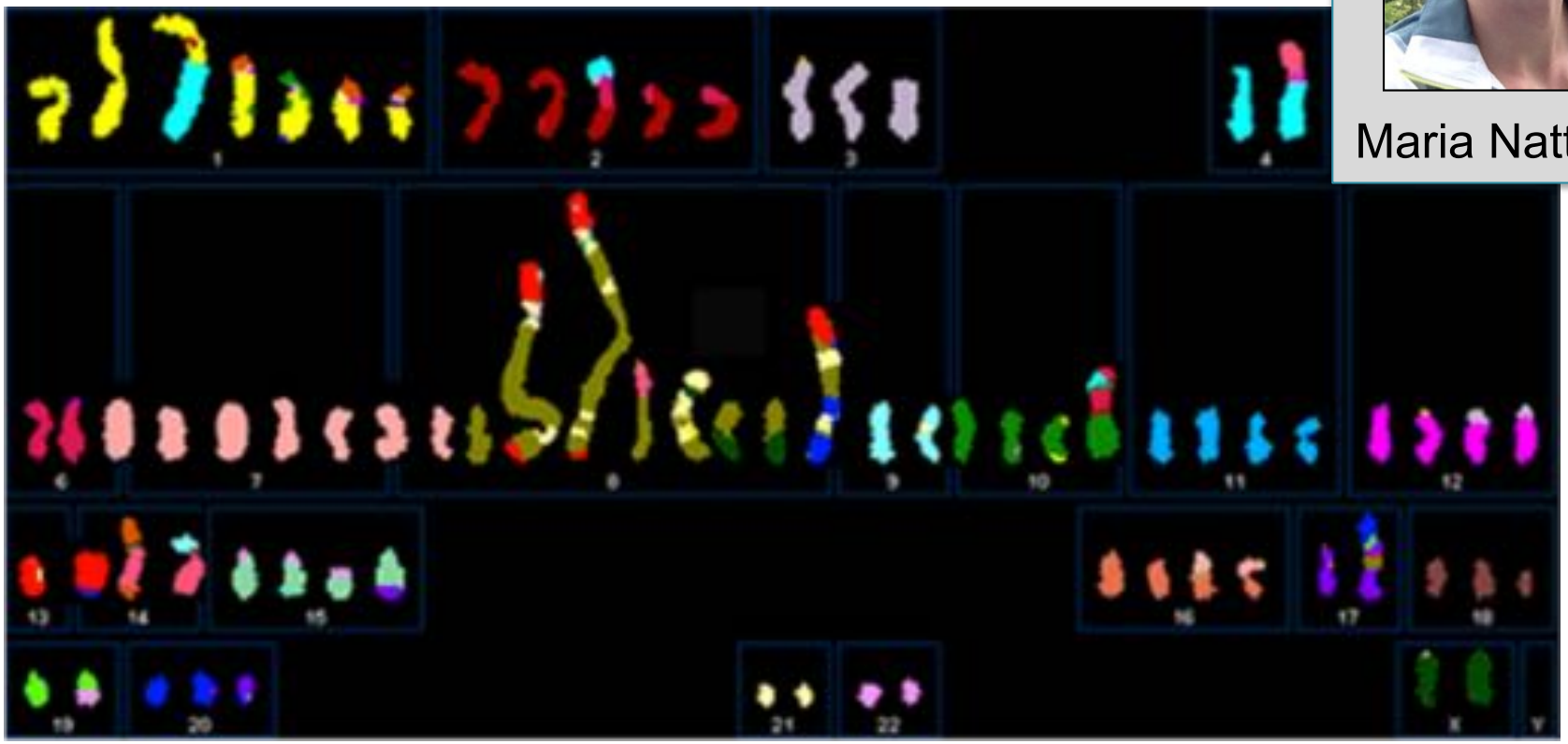
Long-read sequencing of breast cancer



SK-BR-3


Maria Nattestad

Most commonly used Her2-amplified breast cancer



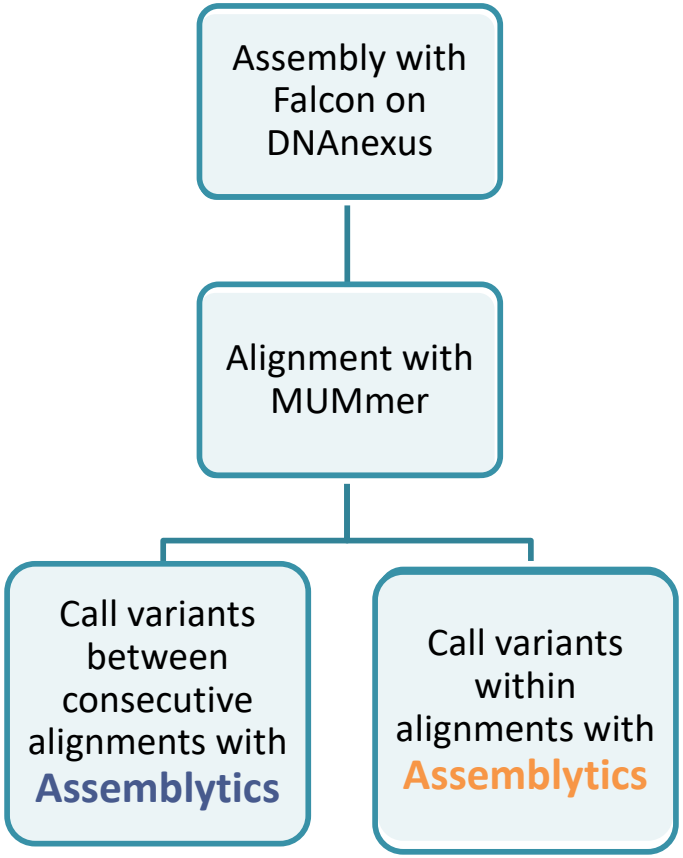
(Davidson et al, 2000)

Can we resolve the complex structural variations, especially around Her2?

Recent collaboration between JHU, CSHL and OICR to *de novo* assemble and analyze the complete cell line genome with PacBio long reads

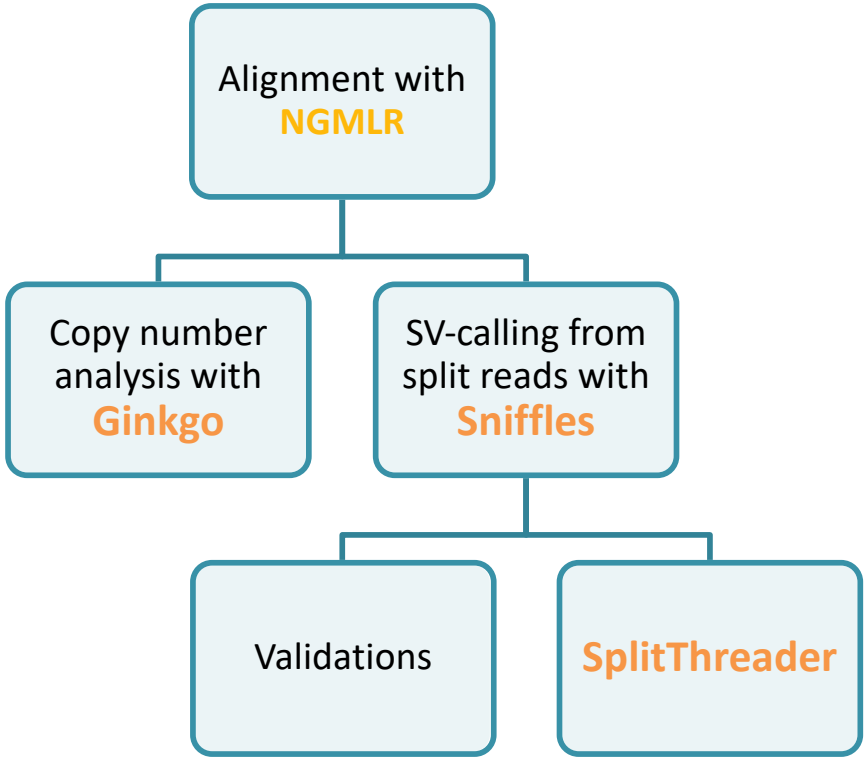
Structural Variation Analysis

Assembly-based



~ 11,000 structural variants
50 bp to 10 kbp

Split-Read based

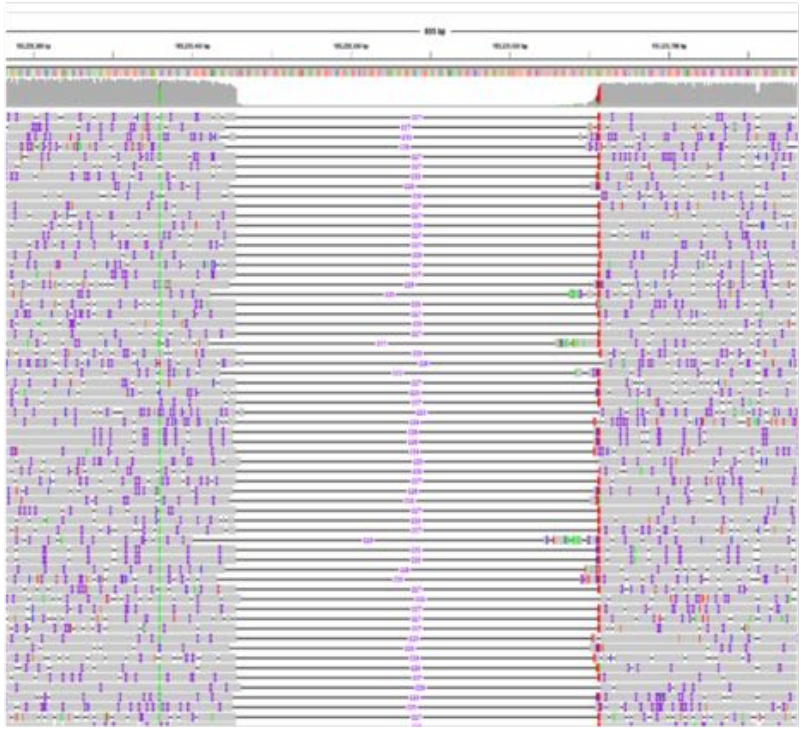


~ 20,000 structural variants
Including many inter-chromosomal rearrangements

NGMLR + Sniffles

BWA-MEM:

NGMLR:



NGMLR: Convex scoring model to accommodate many small gaps from sequencing errors along with less frequent but larger SVs

Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, Rescheneder et al (2018) *Nature Methods. doi:10.1038/s41592-018-0001-7*

Highlights

- Finding 10s of thousands of additional variants
- PCR validation confirms high accuracy of long reads
- Detect many novel gene fusions
- Identify early vs late mutations in the cancer

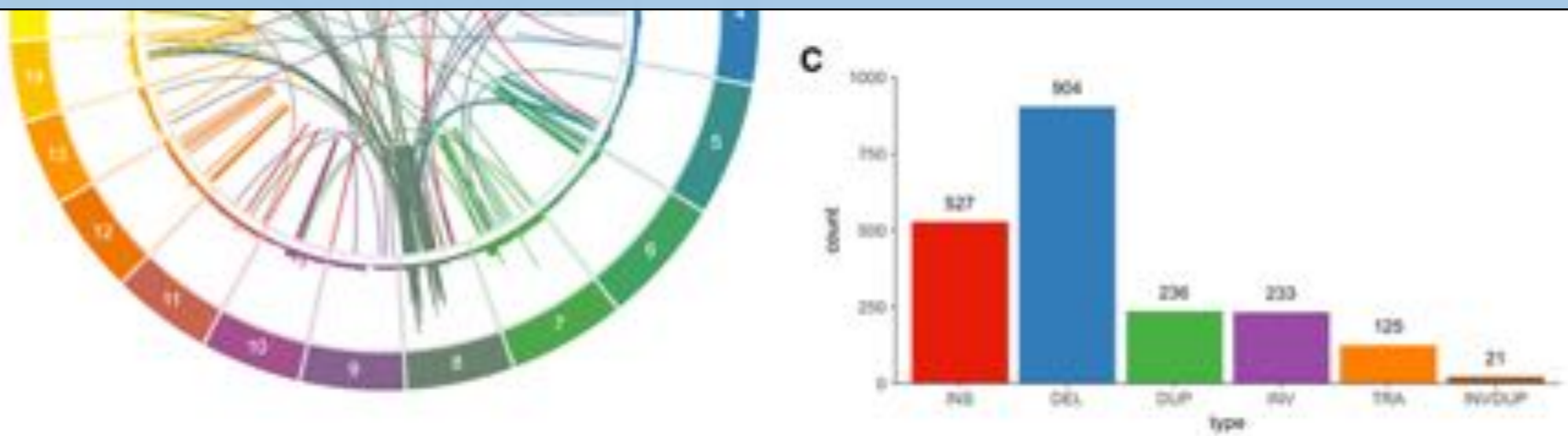


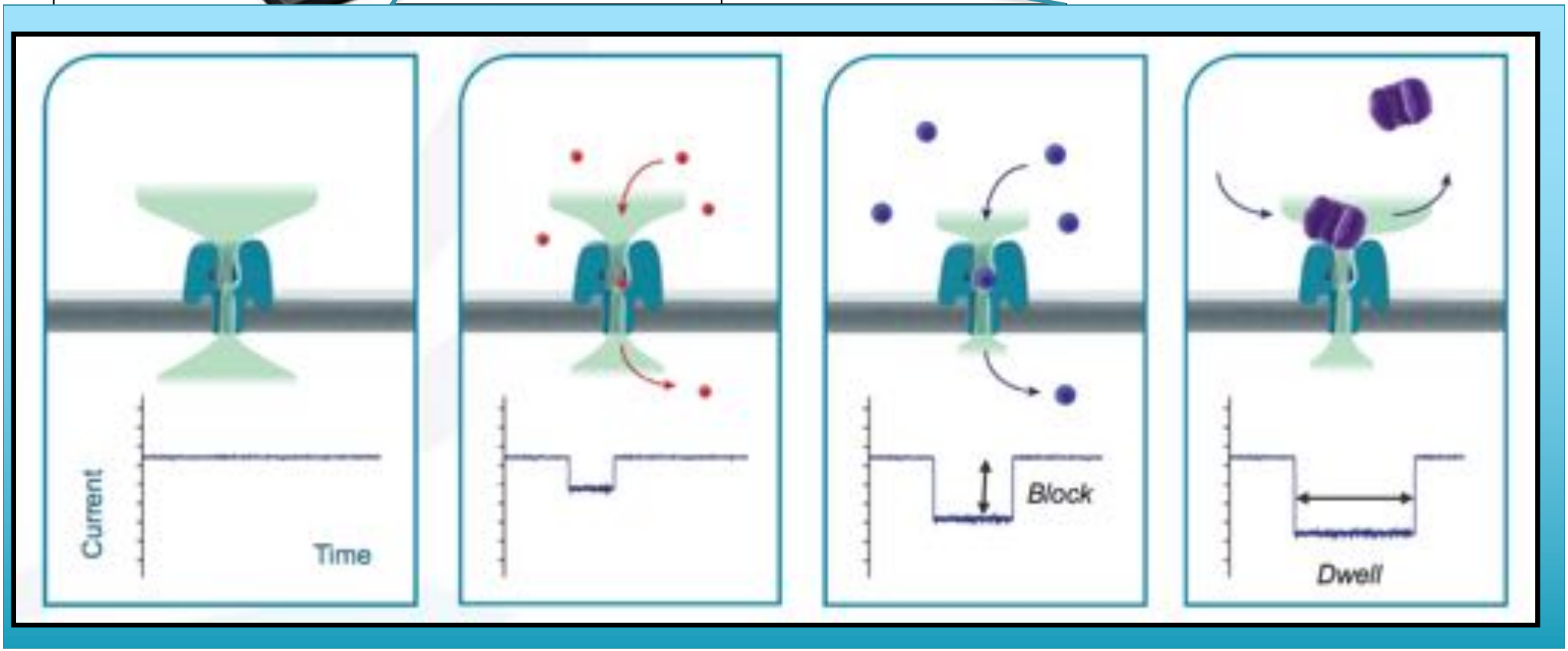
Figure 1. Variants found in SK-BR-3 with PacBio long-read sequencing. (*A*) Circos (Krzywinski et al. 2009) plot showing long-range (larger than 10 kbp or inter-chromosomal) variants found by Sniffles from split-read alignments, with read coverage shown in the *outer* track. (*B*) Variant size histogram of deletions and insertions from size 50 bp up to 1 kbp found by long-read (Sniffles) and short-read (SURVIVOR 2-caller consensus) variant calling, showing similar size distributions for insertions and deletions from long reads but not for short reads, where insertions are greatly underrepresented. (C) Sniffles variant counts by type for variants above 1 kbp in size, including translocations and inverted duplications.

Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line

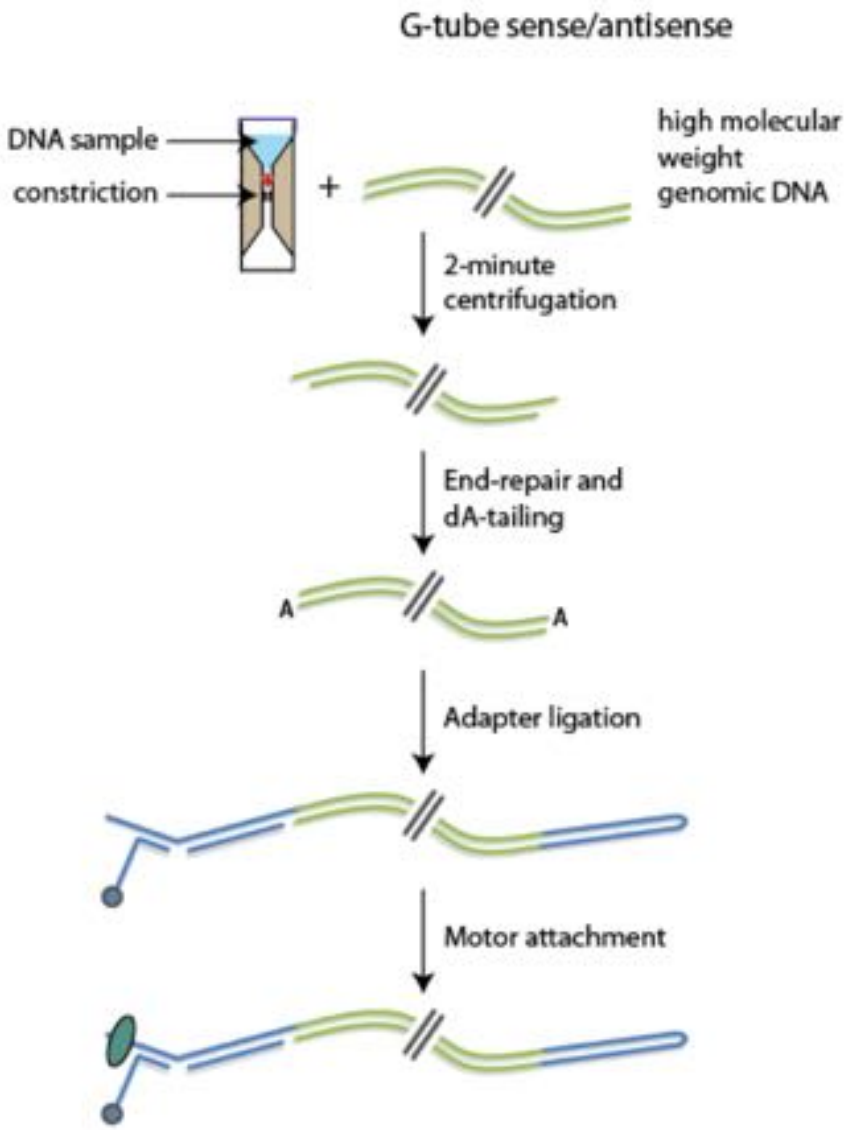
Nattestad et al. (2018) *Genome Research.* doi: 10.1101/gr.231100.117

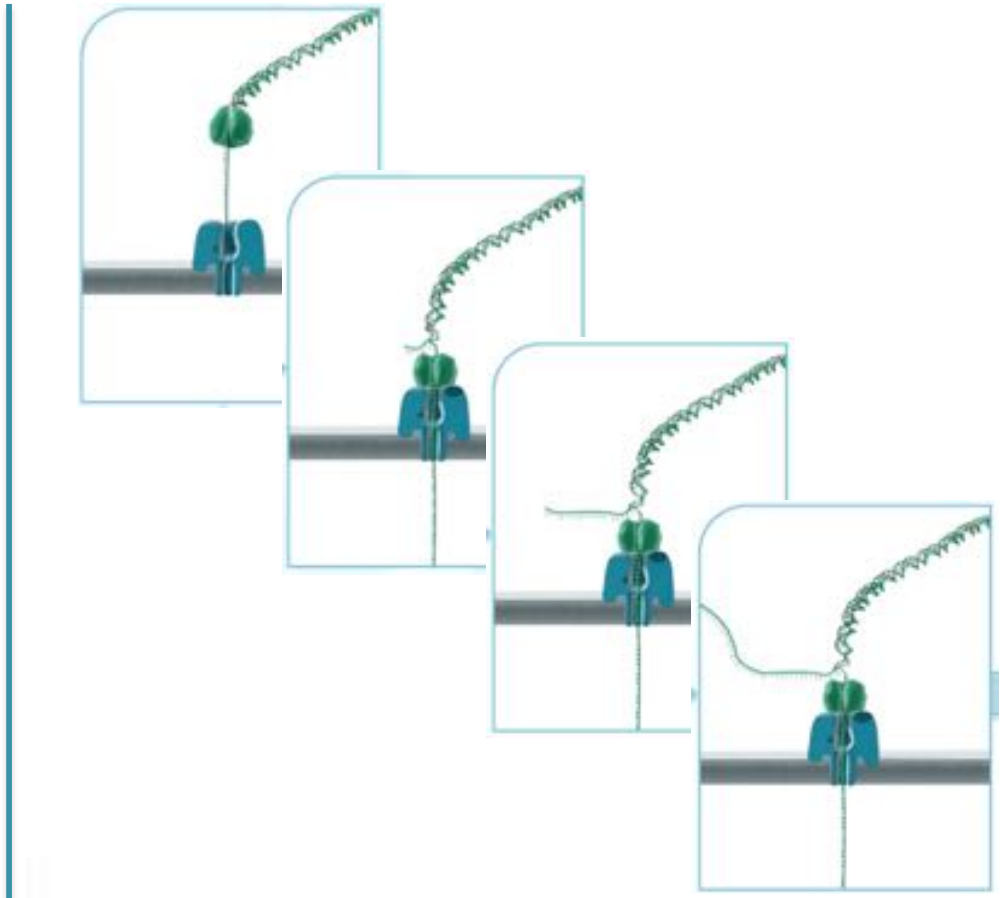
Oxford Nanopore Sequencing

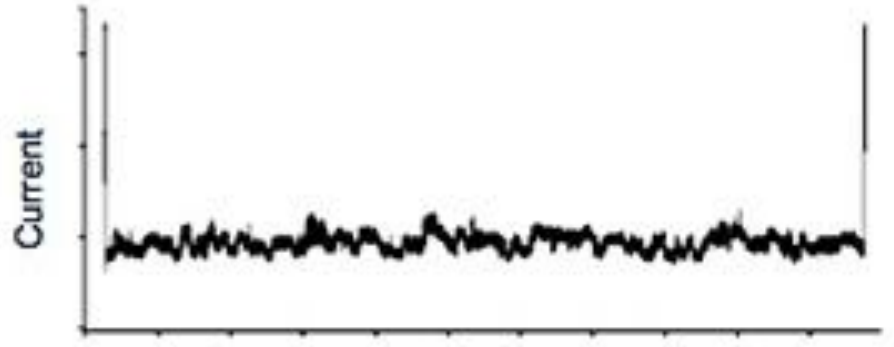
- Thumb drive sized sequencer powered over USB
- Capacity for 512 reads at once
- Senses DNA by measuring changes to ion flow



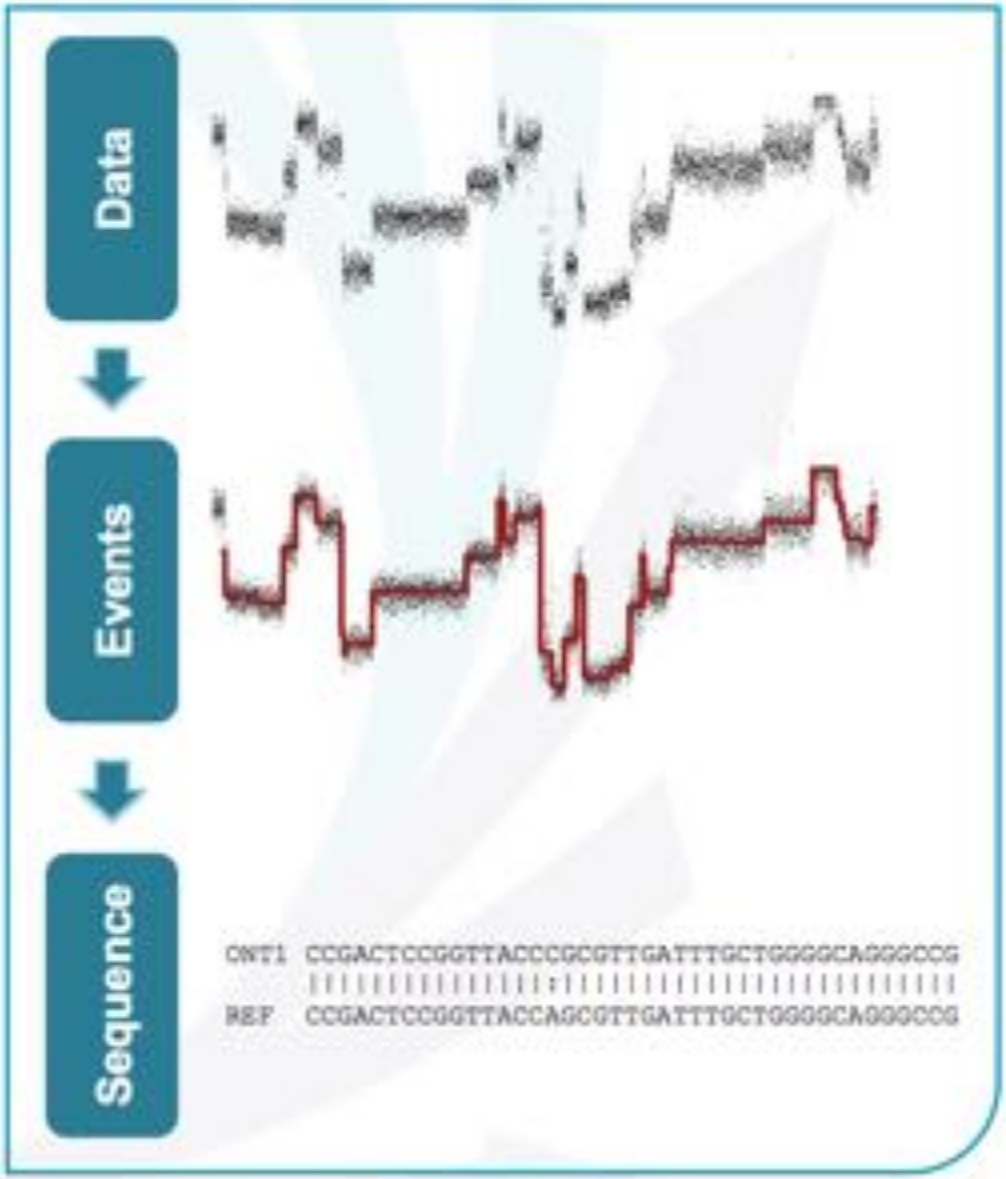
Nanopore Sequencing







Nanopore Basecalling



- Hidden Markov model
- Only four options per transition
- Pore type = distinct kmer length

- Form probabilistic path through measured states currents and transitions
  - e.g. Viterbi algorithm

Originally HMM based base calling, quickly shifting to RNN approaches

Oxford Nanopore Sequencing

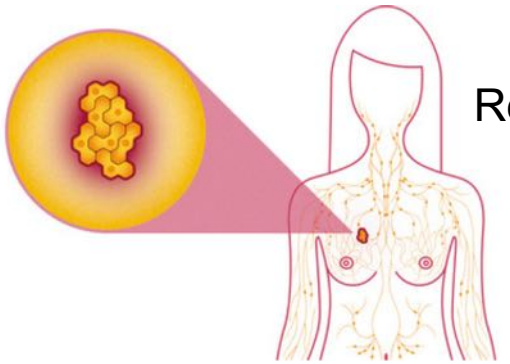
MinION

$1k / instrument
~$15k / human @ 50x
Long reads, Low throughput

PromethION

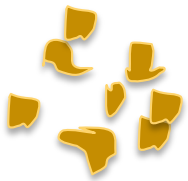
$75k / instrument
~$4k / human @ 50x
Long reads, High throughput

Taking Nanopore Sequencing into the Clinic

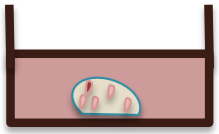


Resected primary tumor

Mechanical/ enzymatic



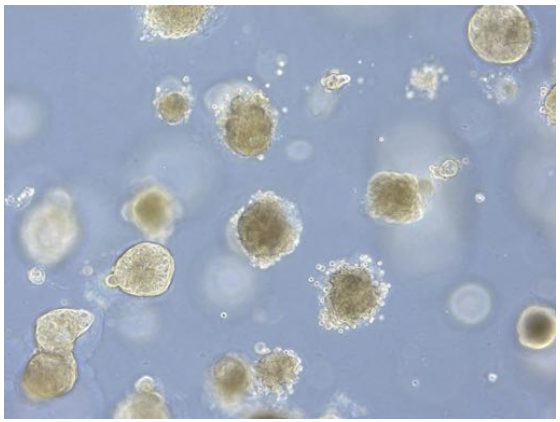
Purification



Plating on Matrigel

Add growth factors

Tumor organoids in culture



- ✓ Stable Growth in 3D
- ✓ Recapitulate tumor pathology & treatment response
- ✓ Maintenance of tissue/tumor heterogeneity
- ✓ "2017 Method of the Year" - Nature Methods

CSH

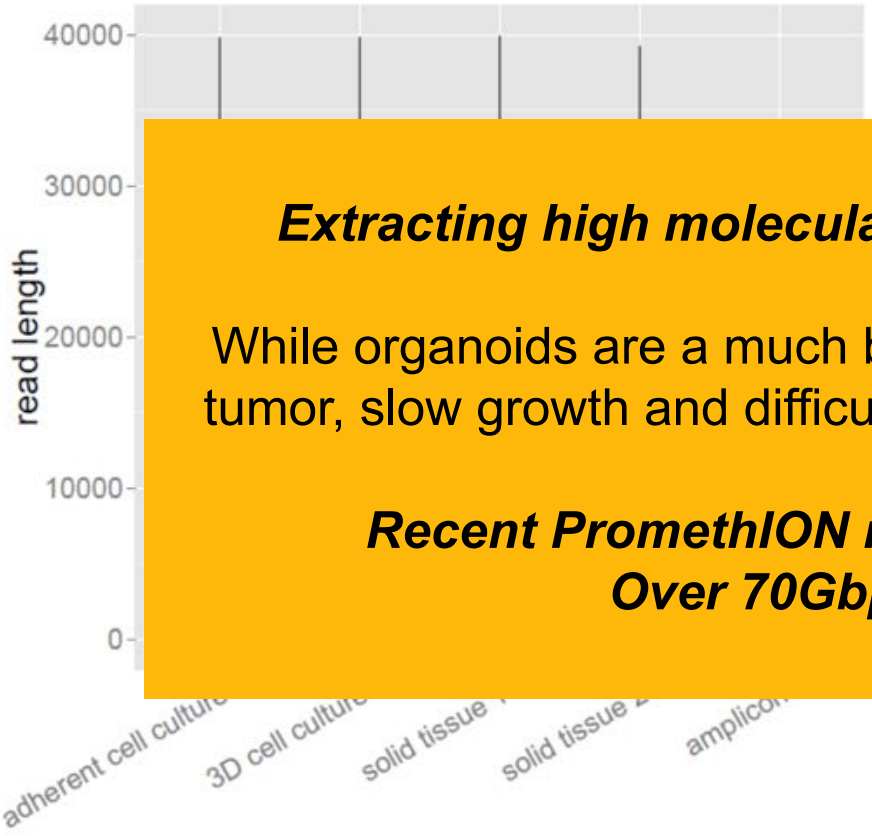
David Spector

Northwell Health

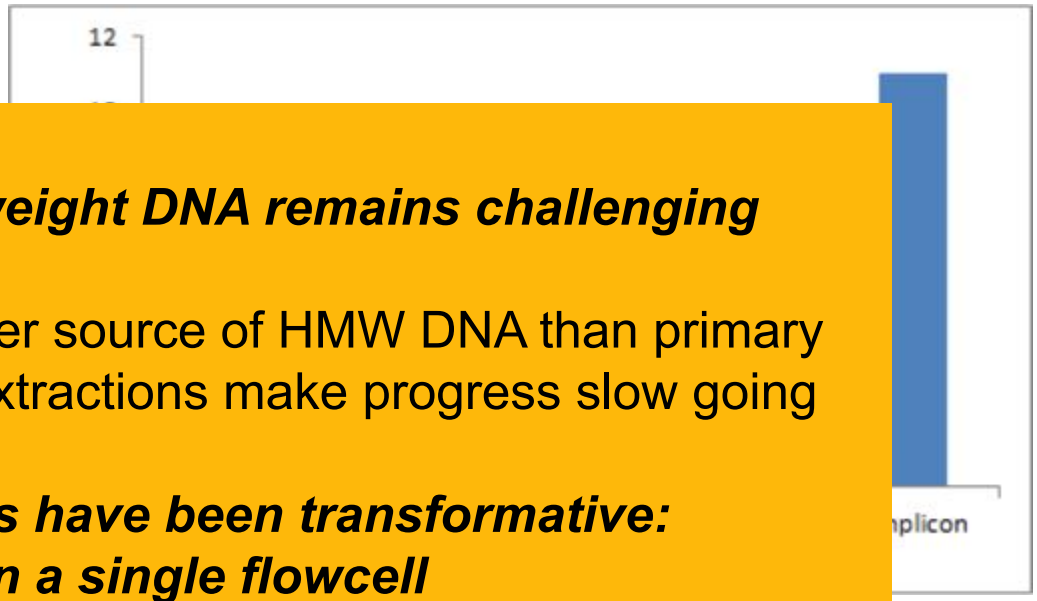
Karen Kostroff

Oxford Nanopore Sequencing Results

Tissue source impacts
read length



Tissue source impacts
yield per flow cell



Extracting high molecular weight DNA remains challenging

While organoids are a much better source of HMW DNA than primary tumor, slow growth and difficult extractions make progress slow going

Recent PromethION runs have been transformative:
Over 70Gbp on a single flowcell

Sequencing by W. Richard McCombie @ CSHL
Winston Timp @ JHU

Preliminary Structural Variations Analysis

	Total	Deletions	Duplications	Insertions	Inversions	Translocations
All SVs in normal	9816	5225	578	3727	130	156
All SVs in tumor	13737	7020	988	5292	202	235
SVs only in tumor (Also exclude NA12878)	3662	1805	420	1250	98	89

Preliminary Structural Variations Analysis

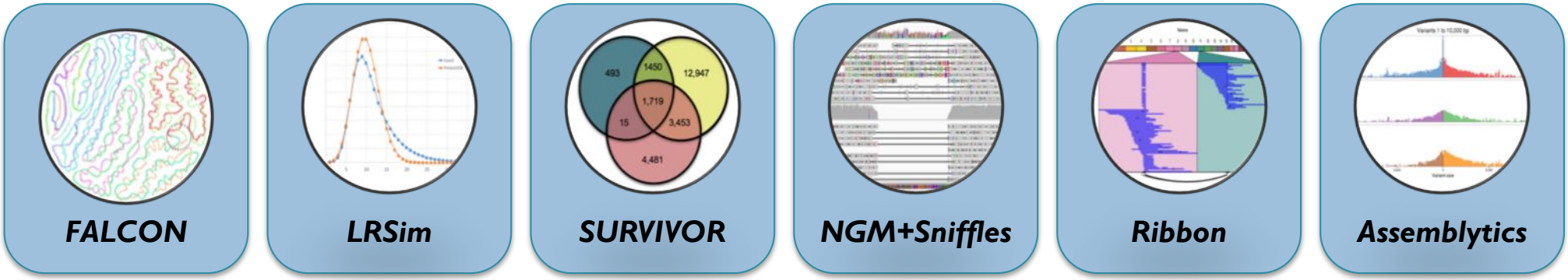
	Total	Deletions	Duplications	Insertions	Inversions	Translocations
All SVs in normal	9816	5225	578	3727	130	156
All SVs in tumor	13737	7020	988	5292	202	235
SVs only in tumor (Also exclude NA12878)	3662	1805	420	1250	98	89

In pursuit of perfect genome sequencing

New sequencing technologies combined with new algorithms are revealing a universe of new genomic variants to study

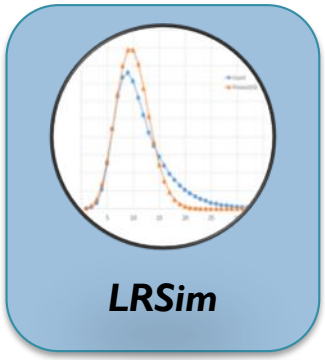
- Tens of thousands of SVs per person, many megabases of variation

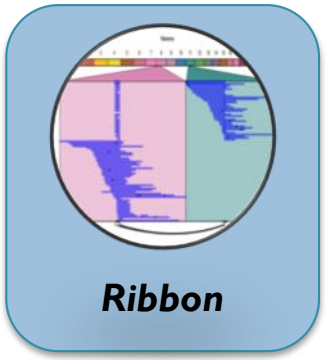
- Identification of novel cancer drivers
- Identification of novel genetic risk factors
- Identification of novel isoforms and fusion genes
- Identification of novel tumor virus and transposable element insertions
- Identification of novel genomic and transcriptomic epigenetic modifications
- Enhanced study of tumor progression, allele-specific factors
- …

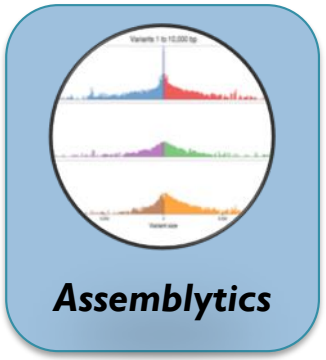


Computational Research Landscape

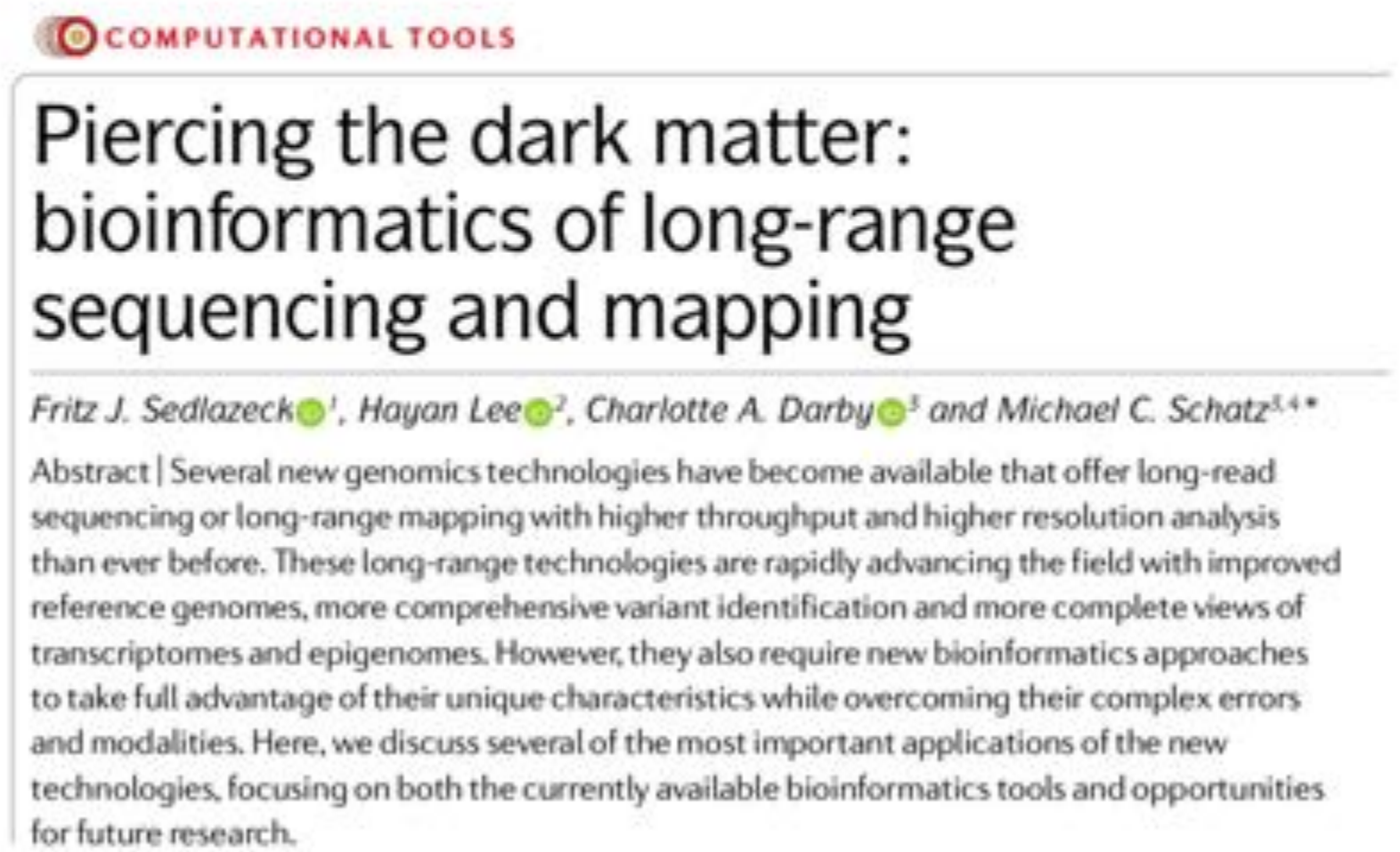
- ***Avoid***
  - New Illumina/PacBio base callers
  - Entirely new genome assembler from scratch
- ***Good***
  - Alignment/Assembly/Analysis methods robust to errors, polyploidy, aneuploidy
  - Use insights from long-reads to improve analysis of short-reads
- ***Best***
  - Synthesis of large numbers of samples ("pan-genome assembly") and/or multiple data types ("multi-omics")
  - Prioritization and interpretation of variations







REVIEWS



COMPUTATIONAL TOOLS

Piercing the dark matter: bioinformatics of long-range sequencing and mapping

*Fritz J. Sedlazeck[1], Hayan Lee[2], Charlotte A. Darby[3] and Michael C. Schatz[3,4]**

Abstract | Several new genomics technologies have become available that offer long-read sequencing or long-range mapping with higher throughput and higher resolution analysis than ever before. These long-range technologies are rapidly advancing the field with improved reference genomes, more comprehensive variant identification and more complete views of transcriptomes and epigenomes. However, they also require new bioinformatics approaches to take full advantage of their unique characteristics while overcoming their complex errors and modalities. Here, we discuss several of the most important applications of the new technologies, focusing on both the currently available bioinformatics tools and opportunities for future research.

Piercing the dark matter: bioinformatics of long- range sequencing and mapping
Sedlazeck et al (2018) *Nature Reviews Genetics. doi:10.1038/s41576-018-0003-4*

Acknowledgements

Schatz Lab
Mike Alonge
Amelia Bateman
Charlotte Darby
Han Fang
Michael Kirsche
Sam Kovaka
Laurent Luo
Srividya Ramakrishnan
T. Rhyker Ranallo-Benavide
Your Name Here

Baylor Medicine
Fritz Sedlazeck

University of Vienna
Arndt von Haeseler
Philipp Rescheneder

DNAnexus
Maria Nattestad

CSHL
Gingeras Lab
Jackson Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Tuveson Lab
Ware Lab
Wigler Lab

SBU
Skiena Lab
Patro Lab

GRC
Roderic Guido
Alessandra Breschi
Anna Vlasova

Yale
Gerstein Lab

JHU
Battle Lab
Langmead Lab
Leek Lab
Salzberg Lab
Taylor Lab
Timp Lab
Wheelan Lab

Cornell
Susan McCouch
Lyza Maron
Mark Wright

OICR
John McPherson
Karen Ng
Timothy Beck
Yogi Sundaravadanam

PacBio
Greg Concepcion

NSF

National Human Genome Research Institute

NIH NATIONAL CANCER INSTITUTE

ALFRED P. SLOAN FOUNDATION

Biological Data Science

Single Cell | Personalized Medicine | Imaging | Machine Learning | Algorithmics | Tools, Infrastructure, & Visualization

November 7-10, 2018


@cncurtis


@jtleek


@BeEngelhardt


@mike_schatz

@StevenSalzberg1
Keynote Speaker

@DrAnneCarpenter
Master Lecturer


Brenda Andrews


@hcorrada


Anna Goldenberg


@mgymrek


@michaelhoffman


@SherlockpHolmes


@jlee8usa


Catalina Vallejos


@aphillippy


Giovanni Parmigiani


Laura van 't Veer


@arjunrajlab

CSH Cold Spring Harbor Laboratory
MEETINGS & COURSES PROGRAM

#biodata18