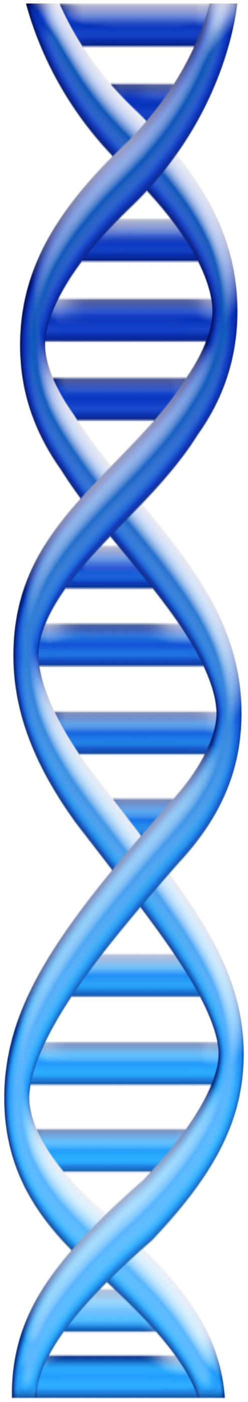
Genome Sequencing & Assembly

Michael Schatz

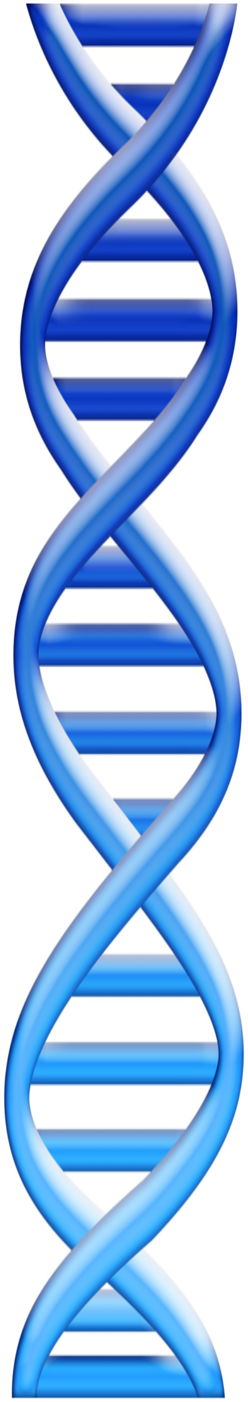
March 30, 2015

CSHL Genome Access



Outline

1. Assembly theory
    1. Assembly by analogy
    2. De Bruijn and Overlap graph
    3. Coverage, read length, errors, and repeats
2. Whole Genome Alignment
    1. Aligning & visualizing with MUMmer
3. Genome assemblers
    1. ALLPATHS-LG: recommended for Illumina-only projects
    2. Celera Assembler: recommended for PacBio/ONT projects

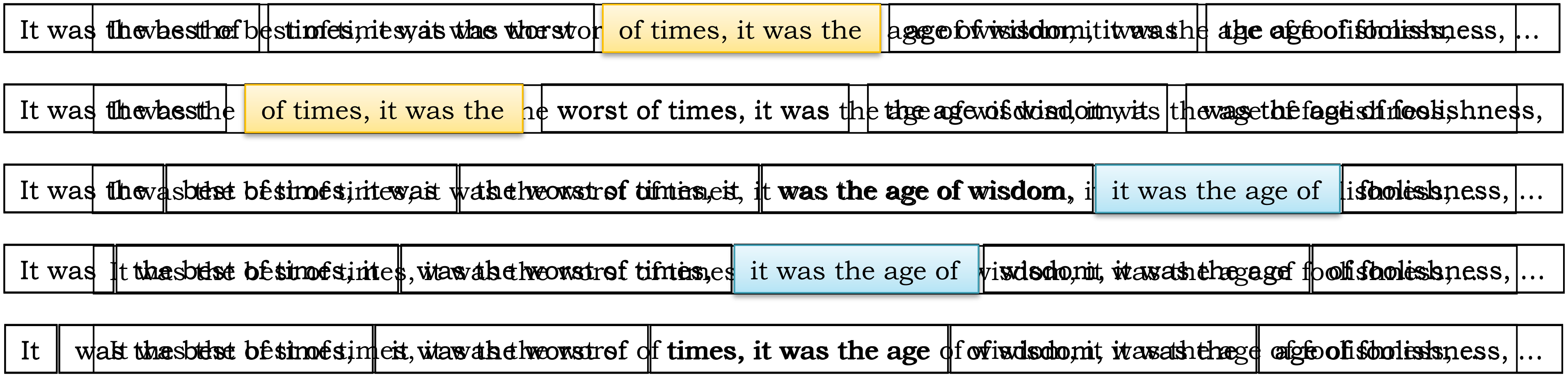


Outline

1. Assembly theory
   1. Assembly by analogy
   2. De Bruijn and Overlap graph
   3. Coverage, read length, errors, and repeats
2. Whole Genome Alignment
   1. Aligning & visualizing with MUMmer
3. Genome assemblers
   1. ALLPATHS-LG: recommended for Illumina-only projects
   2. Celera Assembler: recommended for PacBio projects

Shredded Book Reconstruction

- Dickens accidentally shreds the first printing of <u>A Tale of Two Cities</u>
  - 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,

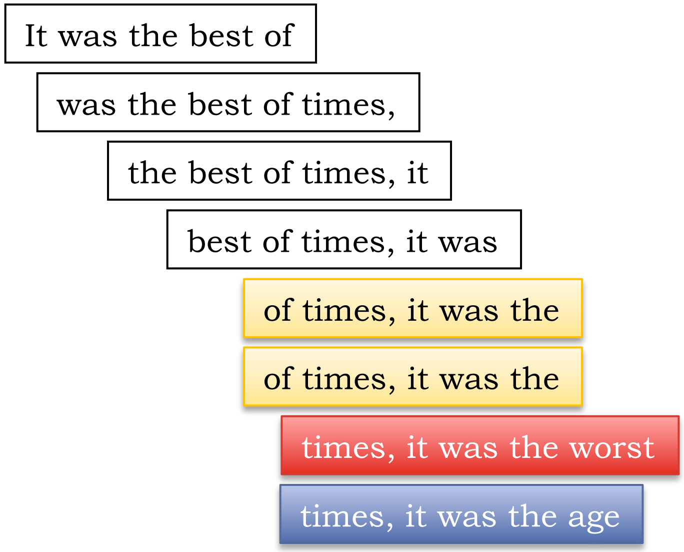
was the age of foolishness,

was the best of times,

was the worst of times,

wisdom, it was the age

worst of times, it was


It was the best of

was the best of times,

the best of times, it

best of times, it was

of times, it was the

of times, it was the

times, it was the worst

times, it was the age

The repeated sequence make the correct reconstruction ambiguous

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

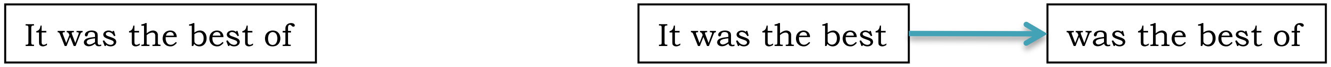
Model the assembly problem as a graph problem

de Bruijn Graph Construction

- $D_k = (V,E)$
  - V = All length-k subfragments ($k < l$)
  - E = Directed edges between consecutive subfragments
    - Nodes overlap by k-1 words

Original Fragment

Directed Edge



It was the best of

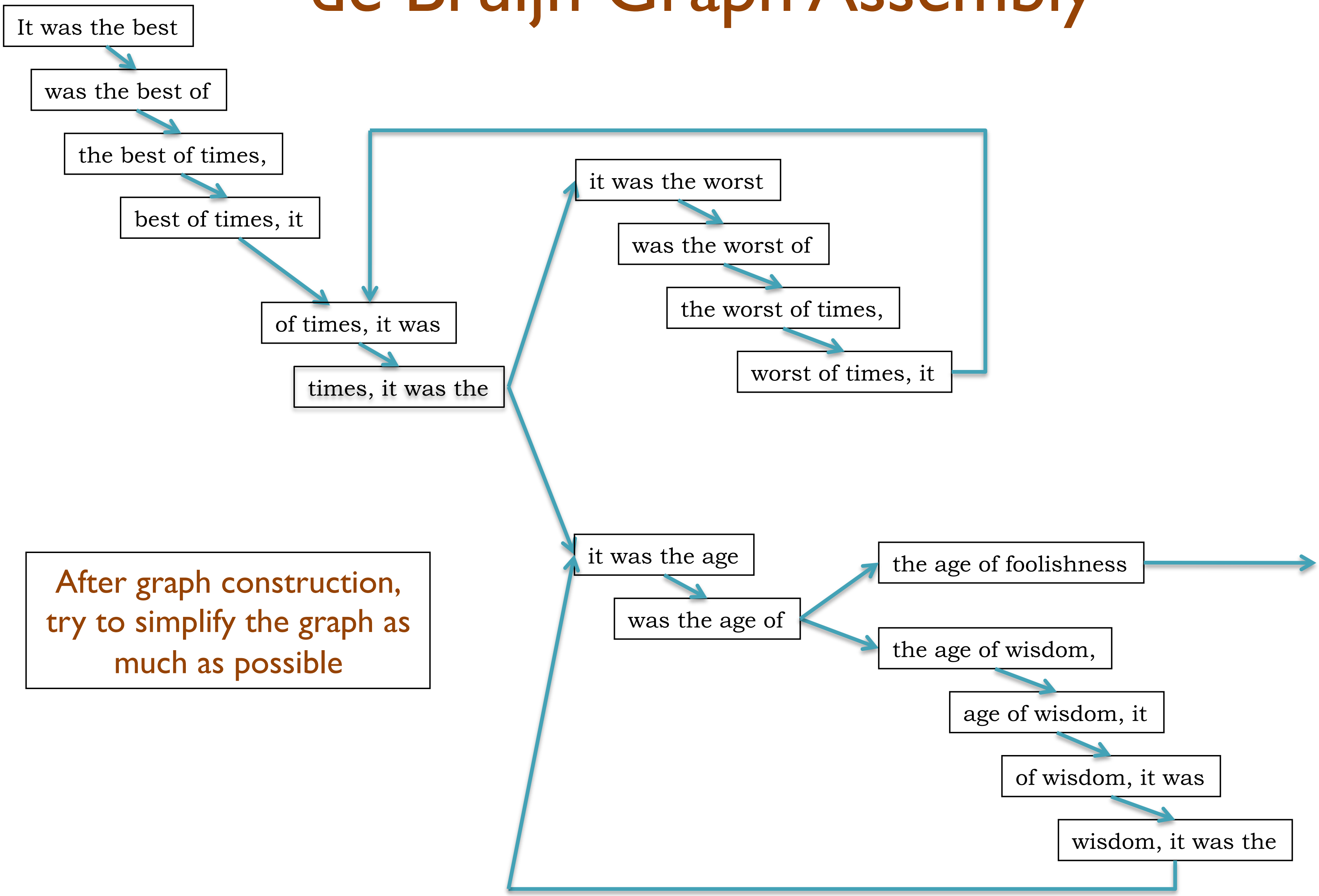
It was the best

was the best of

- Locally constructed graph reveals the global sequence structure
  - Overlaps between sequences implicitly computed

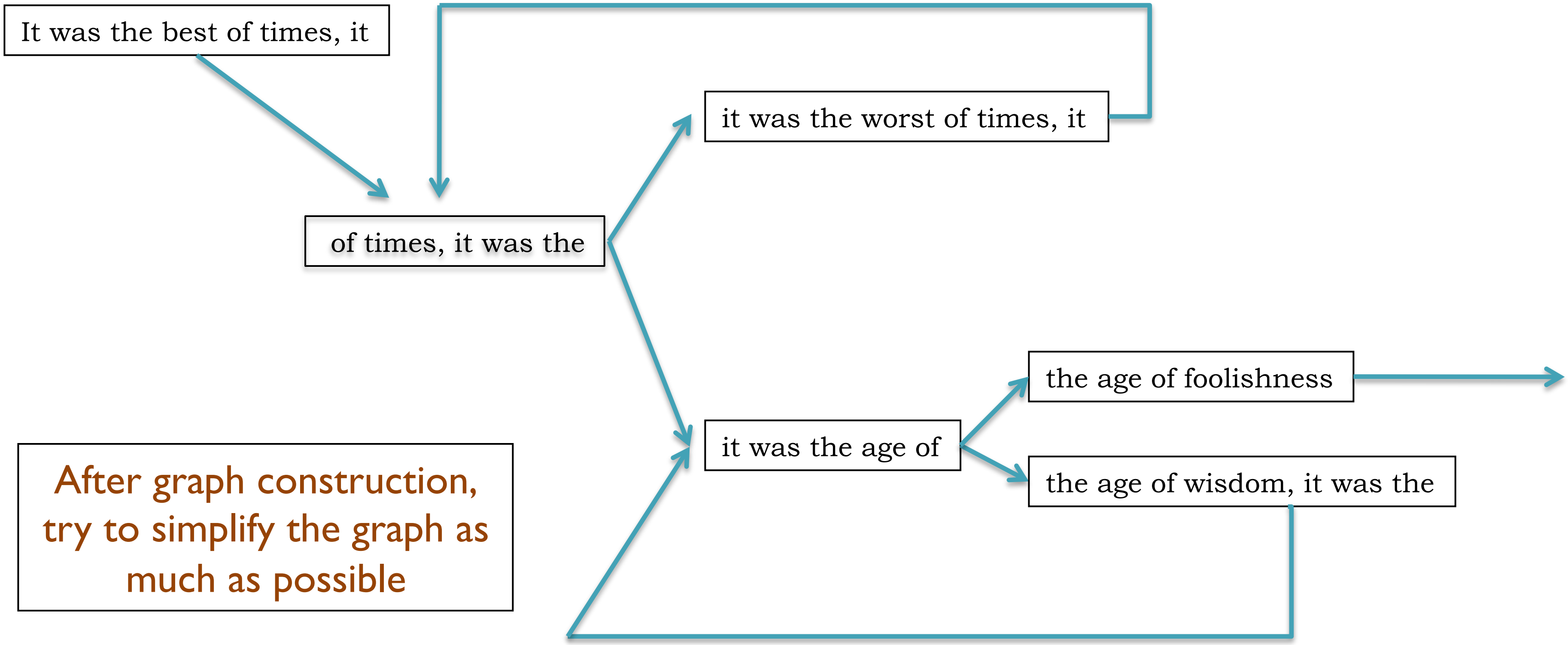
de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001

de Bruijn Graph Assembly



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

de Bruijn Graph Assembly



It was the best of times, it

of times, it was the

it was the worst of times, it

it was the age of

the age of foolishness

the age of wisdom, it was the

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

The full tale

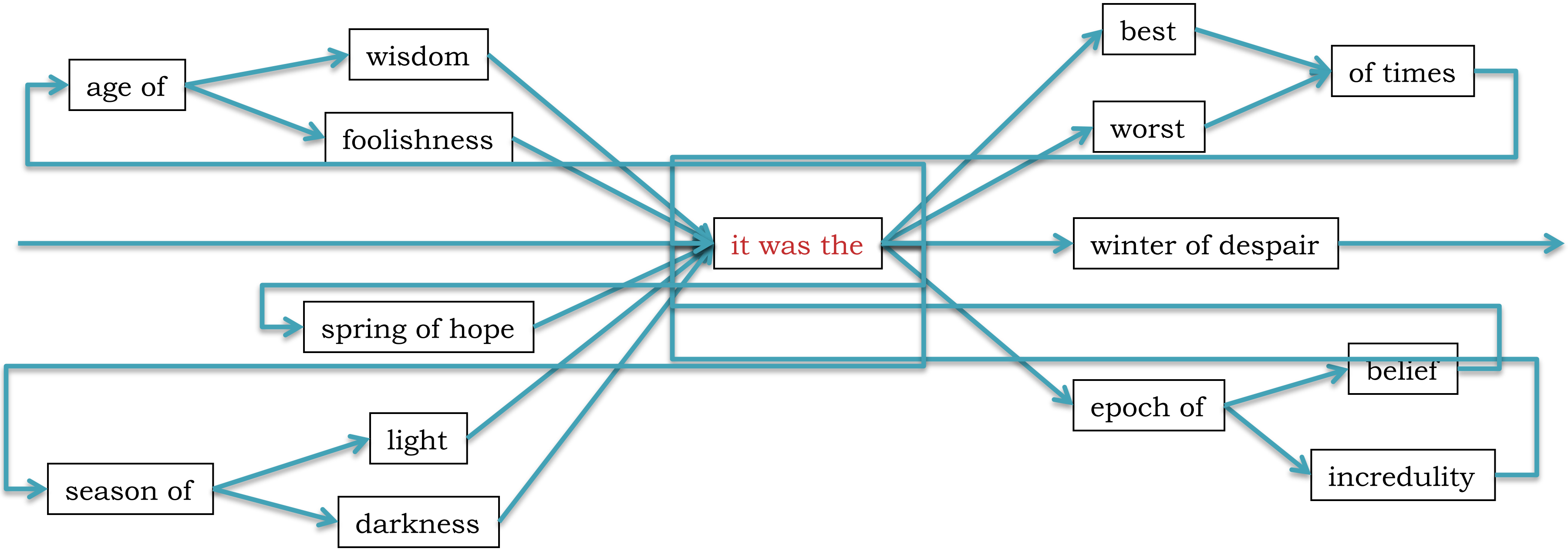
… it was the best of times it was the worst of times …

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

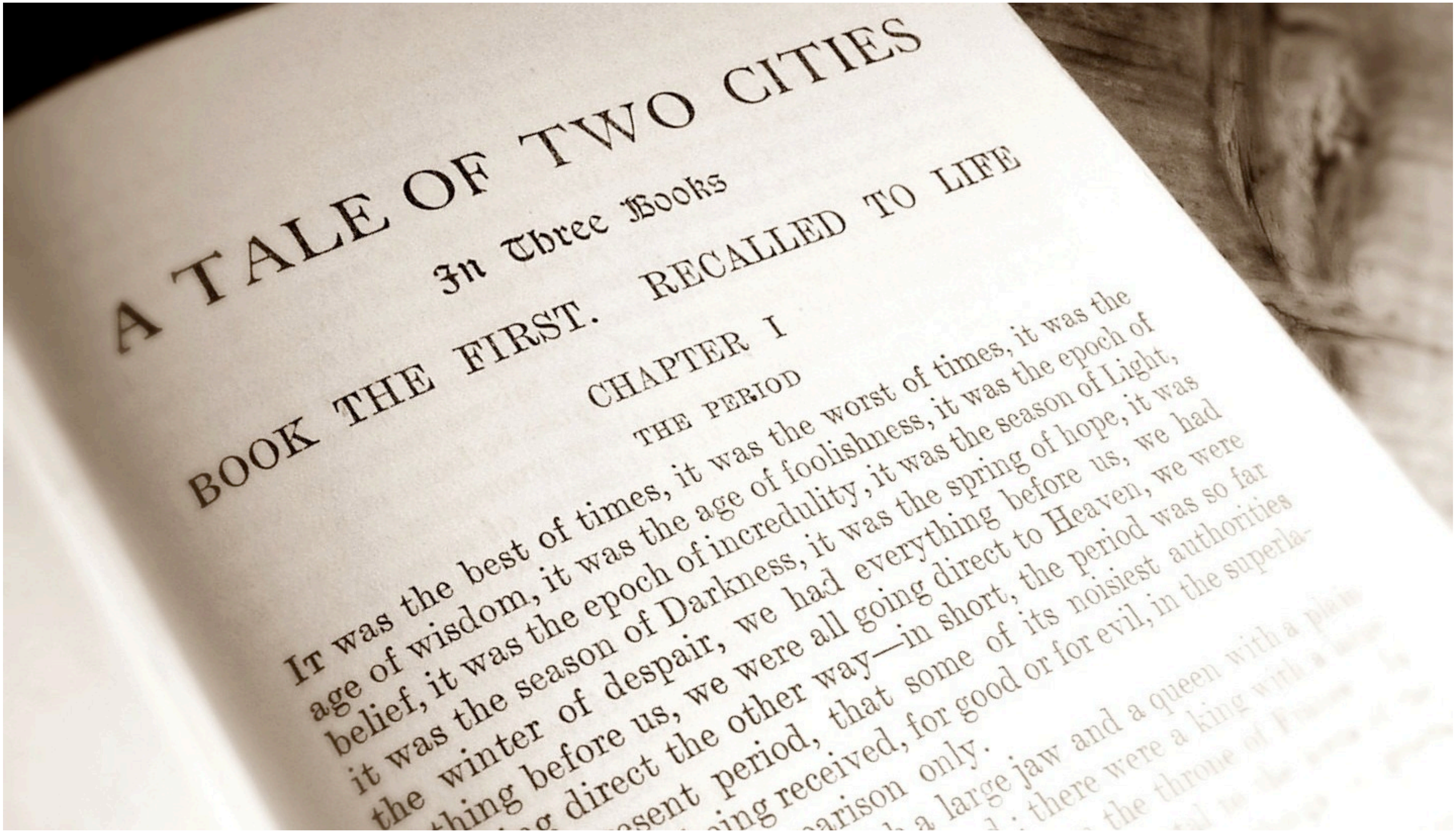
… it was the epoch of belief it was the epoch of incredulity …

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

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



The full tale



A TALE OF TWO CITIES

Jn Three Books

BOOK THE FIRST. RECALLED TO LIFE

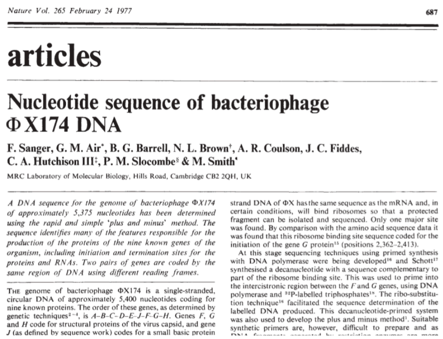
CHAPTER I

THE PERIOD

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 winter of despair, we had everything before us, we had
thing before us, we were all going direct to Heaven, we were
ng direct the other way—in short, the period was so far
esent period, that some of its noisiest authorities
ing received, for good or for evil, in the superla-
arison only.

h a large jaw and a queen with a plain
d; there were a king with a large
the throne of France

Milestones in Genome Assembly



Nature Vol. 265 February 24 1977 687

articles

Nucleotide sequence of bacteriophage ΦX174 DNA

F. Sanger, G. M. Air[*], B. G. Barrell, N. L. Brown[†], A. R. Coulson, J. C. Fiddes, C. A. Hutchison III[‡], P. M. Slocombe[§] & M. Smith[*]

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

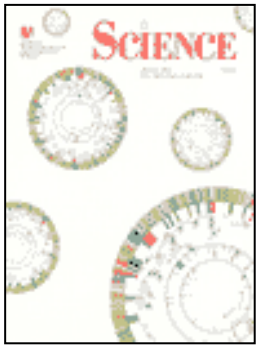
A DNA sequence for the genome of bacteriophage ΦX174 of approximately 5,375 nucleotides has been determined using the rapid and simple 'plus and minus' method. The sequence identifies many of the features responsible for the production of the proteins of the nine known genes of the organism, including initiation and termination sites for the proteins and RNAs. Two pairs of genes are coded by the same region of DNA using different reading frames.

THE genome of bacteriophage ΦX174 is a single-stranded, circular DNA of approximately 5,400 nucleotides coding for nine known proteins. The order of these genes, as determined by genetic techniques[2-4], is *A–B–C–D–E–J–F–G–H*. Genes *F*, *G* and *H* code for structural proteins of the virus capsid, and gene *J* (as defined by sequence work) codes for a small basic protein

strand DNA of ΦX has the same sequence as the mRNA and, in certain conditions, will bind ribosomes so that a protected fragment can be isolated and sequenced. Only one major site was found. By comparison with the amino acid sequence data it was found that this ribosome binding site sequence coded for the initiation of the gene *G* protein[15] (positions 2,362–2,413).

At this stage sequencing techniques using primed synthesis with DNA polymerase were being developed[16] and Schott[17] synthesised a decanucleotide with a sequence complementary to part of the ribosome binding site. This was used to prime into the intercistronic region between the *F* and *G* genes, using DNA polymerase and ^{32}P-labelled triphosphates[18]. The ribo-substitution technique[16] facilitated the sequence determination of the labelled DNA produced. This decanucleotide-primed system was also used to develop the plus and minus method[1]. Suitable synthetic primers are, however, difficult to prepare and as

1977. Sanger *et al.*
1st Complete Organism
5375 bp



1995. Fleischmann *et al.*
1st Free Living Organism
TIGR Assembler. 1.8Mbp

1998. C.elegans SC
1st Multicellular Organism
BAC-by-BAC Phrap. 97Mbp

2000. Myers *et al.*
1st Large WGS Assembly.
Celera Assembler. 116 Mbp

2001. Venter *et al.*, IHGSC
Human Genome
Celera Assembler/GigaAssembler. 2.9 Gbp

2010. Li *et al.*
1st Large SGS Assembly.
SOAPdenovo 2.2 Gbp

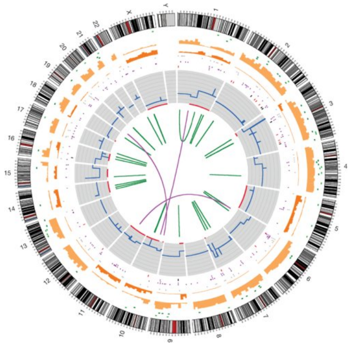
Like Dickens, we must computationally reconstruct a genome from short fragments

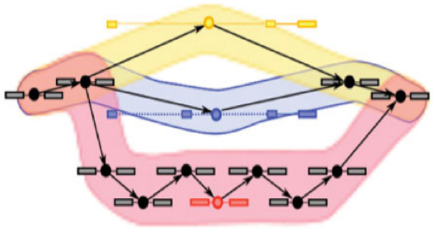
Assembly Applications

- Novel genomes

- Metagenomes

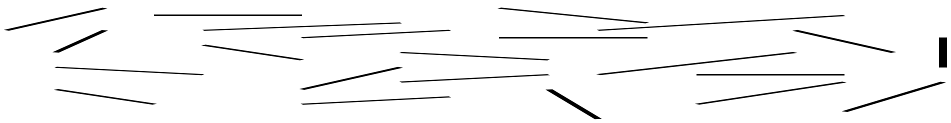
- Sequencing assays
  - Structural variations
  - Transcript assembly
  - …





Assembling a Genome

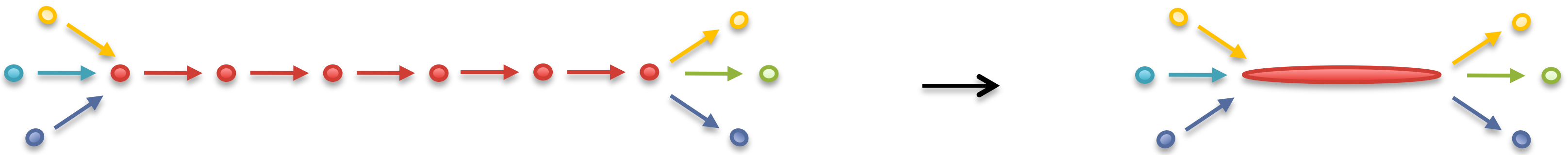
1. Shear & Sequence DNA



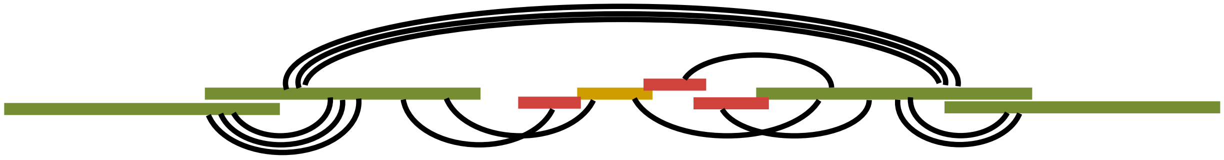
2. Construct assembly graph from overlapping reads

```
...AGCCTAGGGATGCGCGACACGT
          GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
                                          CAACCTCGGACGGACCTCAGCGAA...
```

3. Simplify assembly graph



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



Why are genomes hard to assemble?

1. ***Biological***:
   - (Very) High ploidy, heterozygosity, repeat content

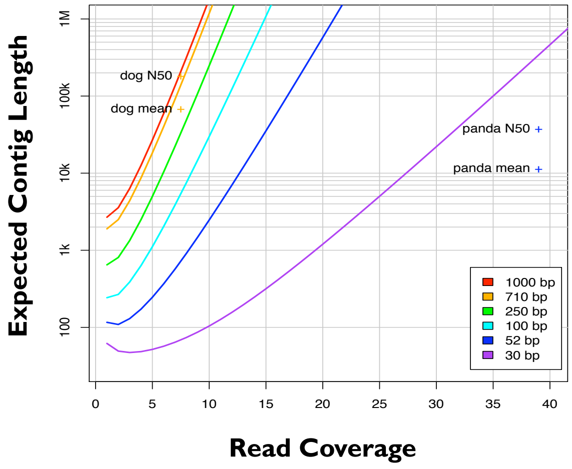
2. ***Sequencing***:
   - (Very) large genomes, imperfect sequencing

3. ***Computational***:
   - (Very) Large genomes, complex structure

4. ***Accuracy***:
   - (Very) Hard to assess correctness

Ingredients for a good assembly

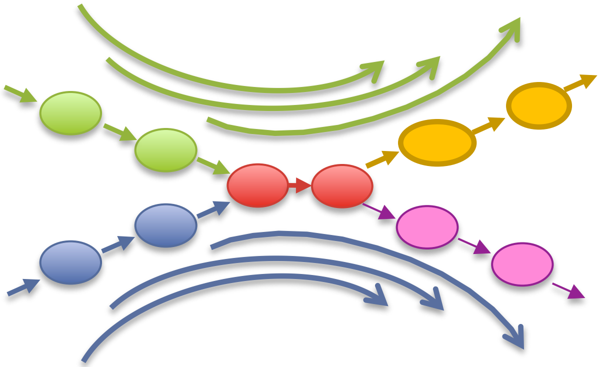
Coverage



High coverage is required

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

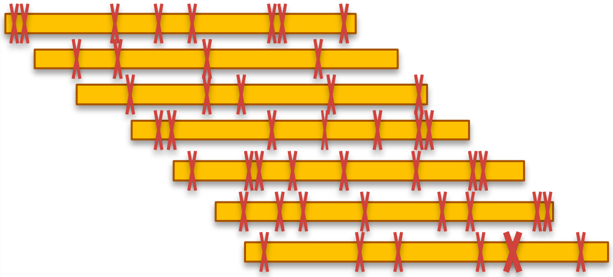
Read Length



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

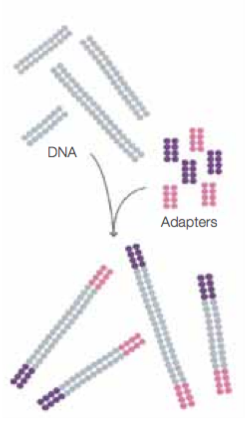


Errors obscure overlaps

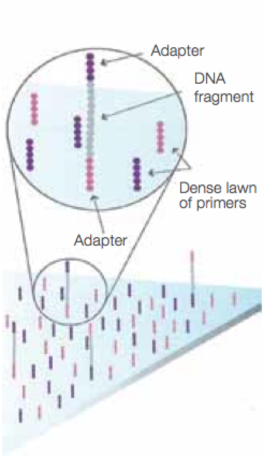
- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

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

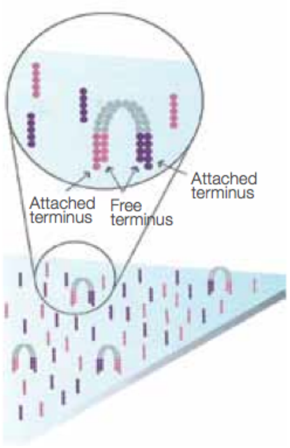
Illumina Sequencing by Synthesis



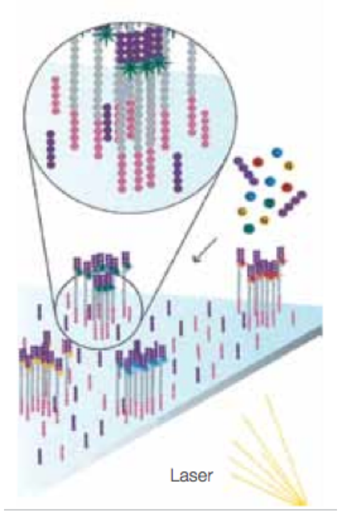
1. Prepare



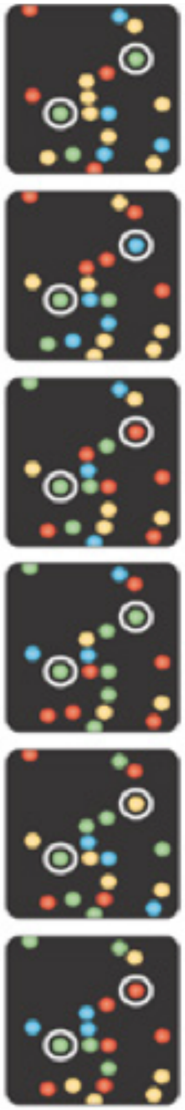
2. Attach



3. Amplify



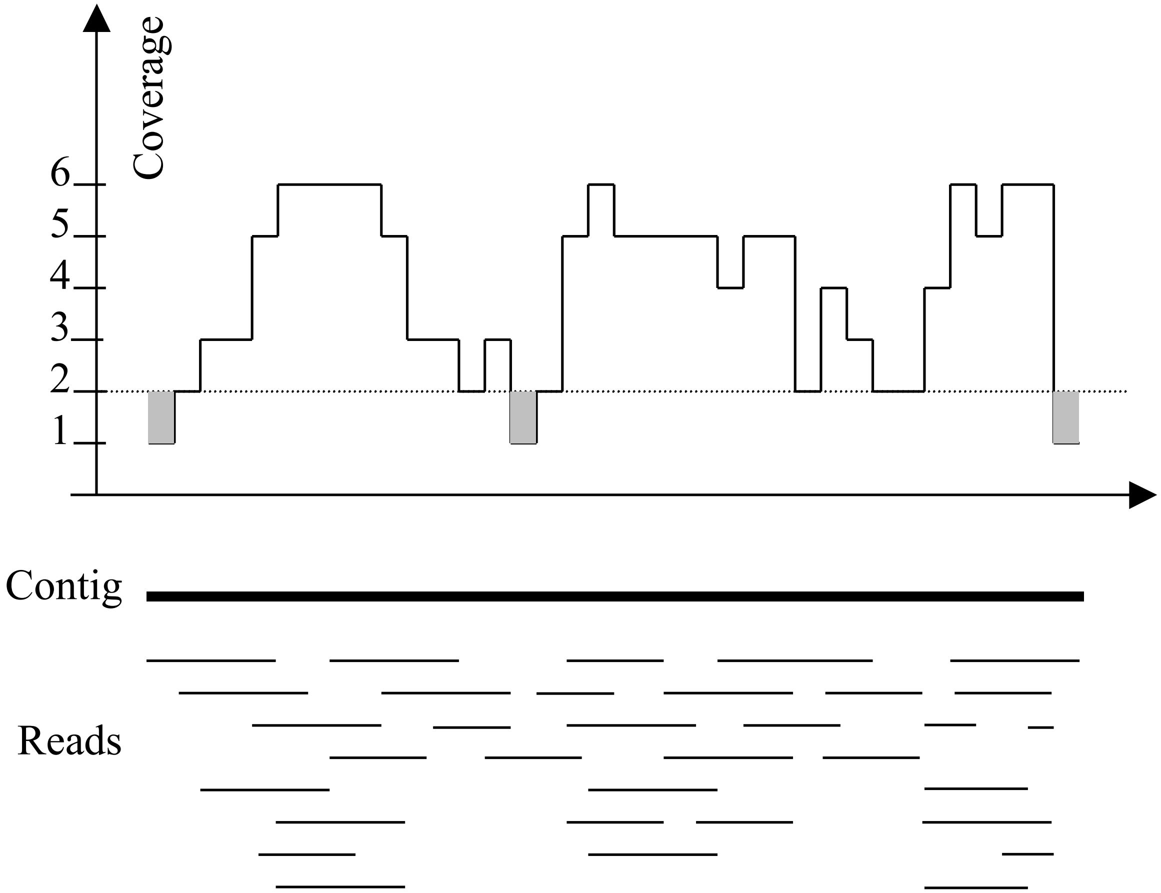
4. Image



5. Basecall

Metzker (2010) Nature Reviews Genetics 11:31-46
http://www.youtube.com/watch?v=l99aKKHcxC4

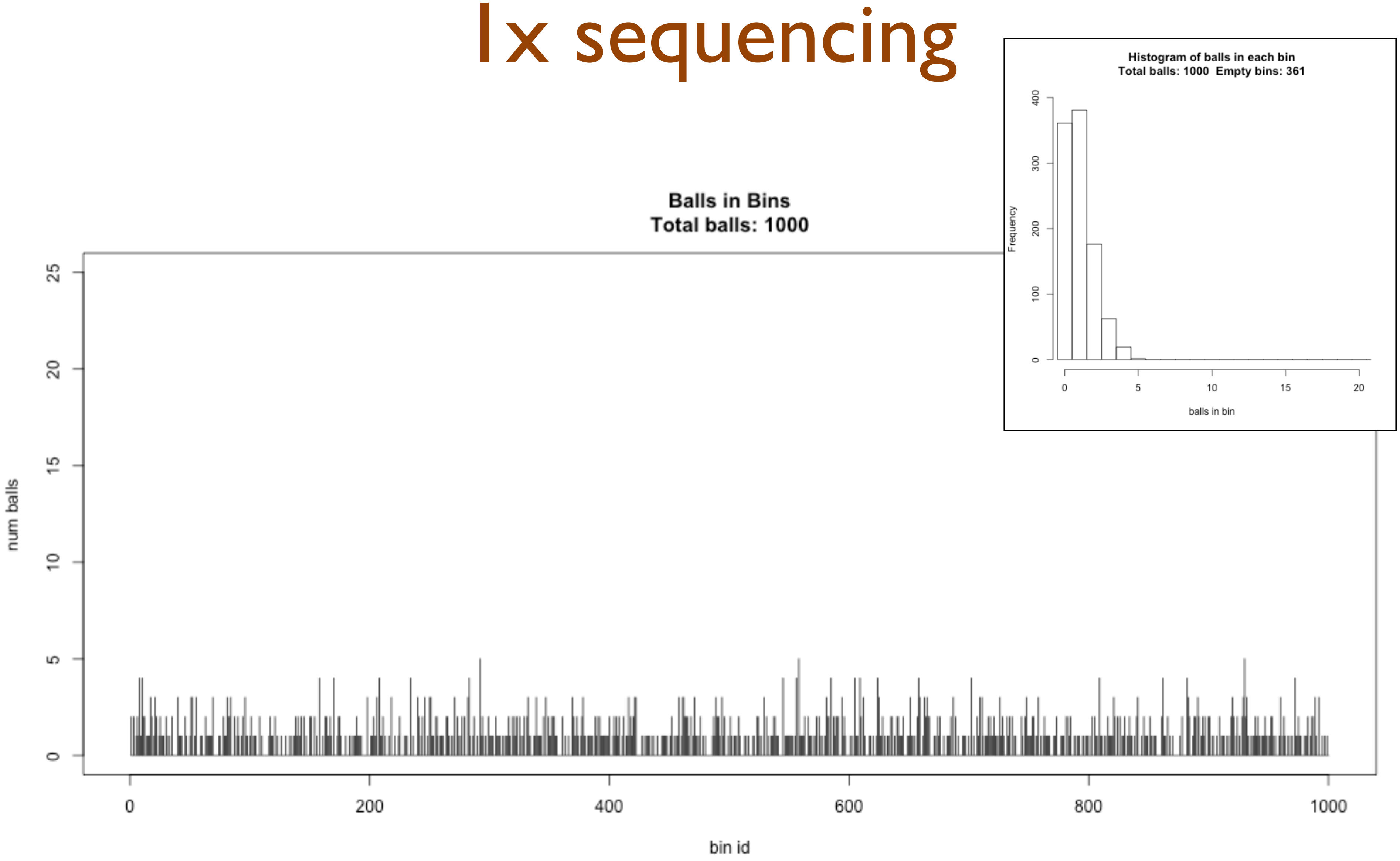
Typical sequencing coverage



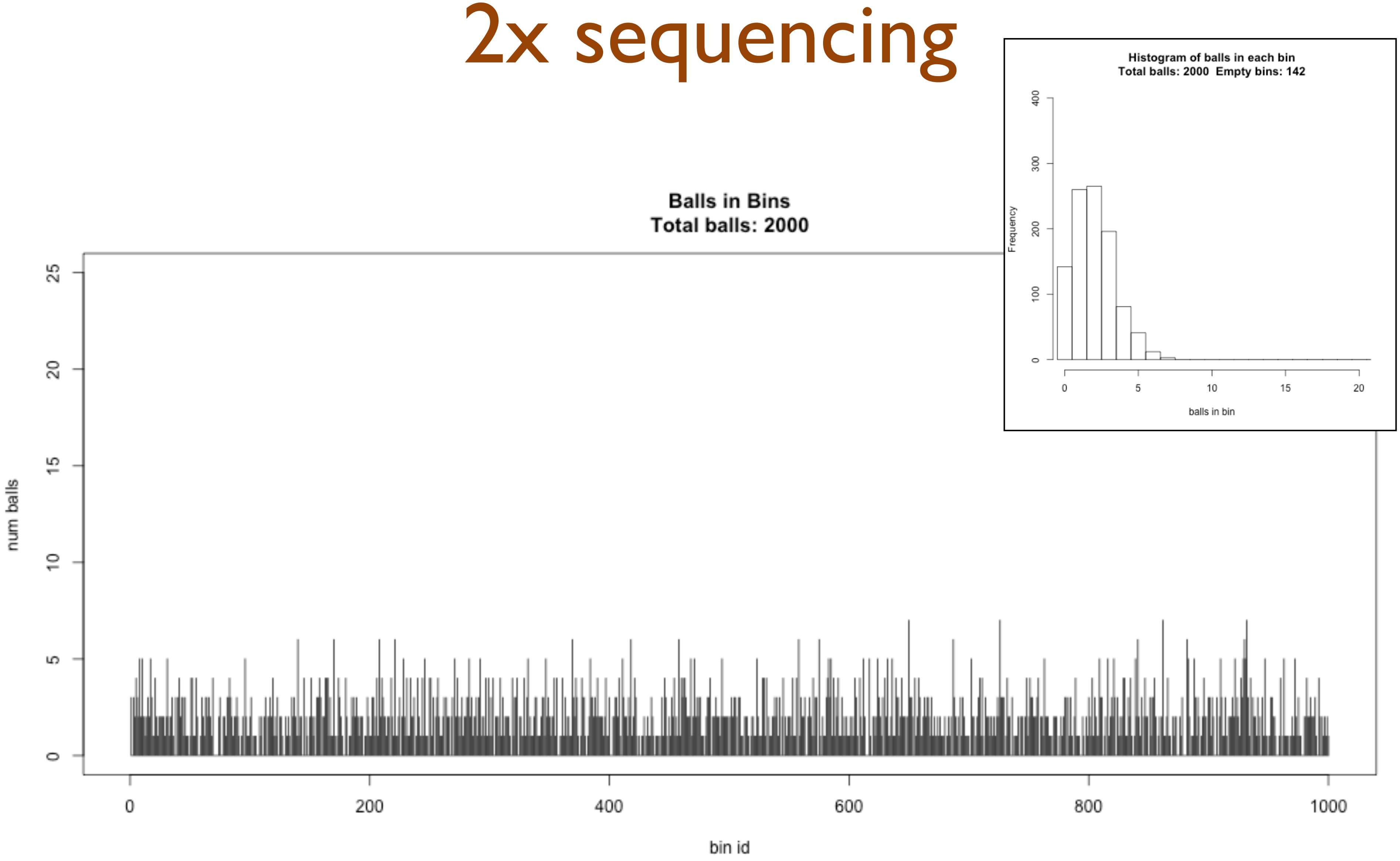
Imagine raindrops on a sidewalk

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

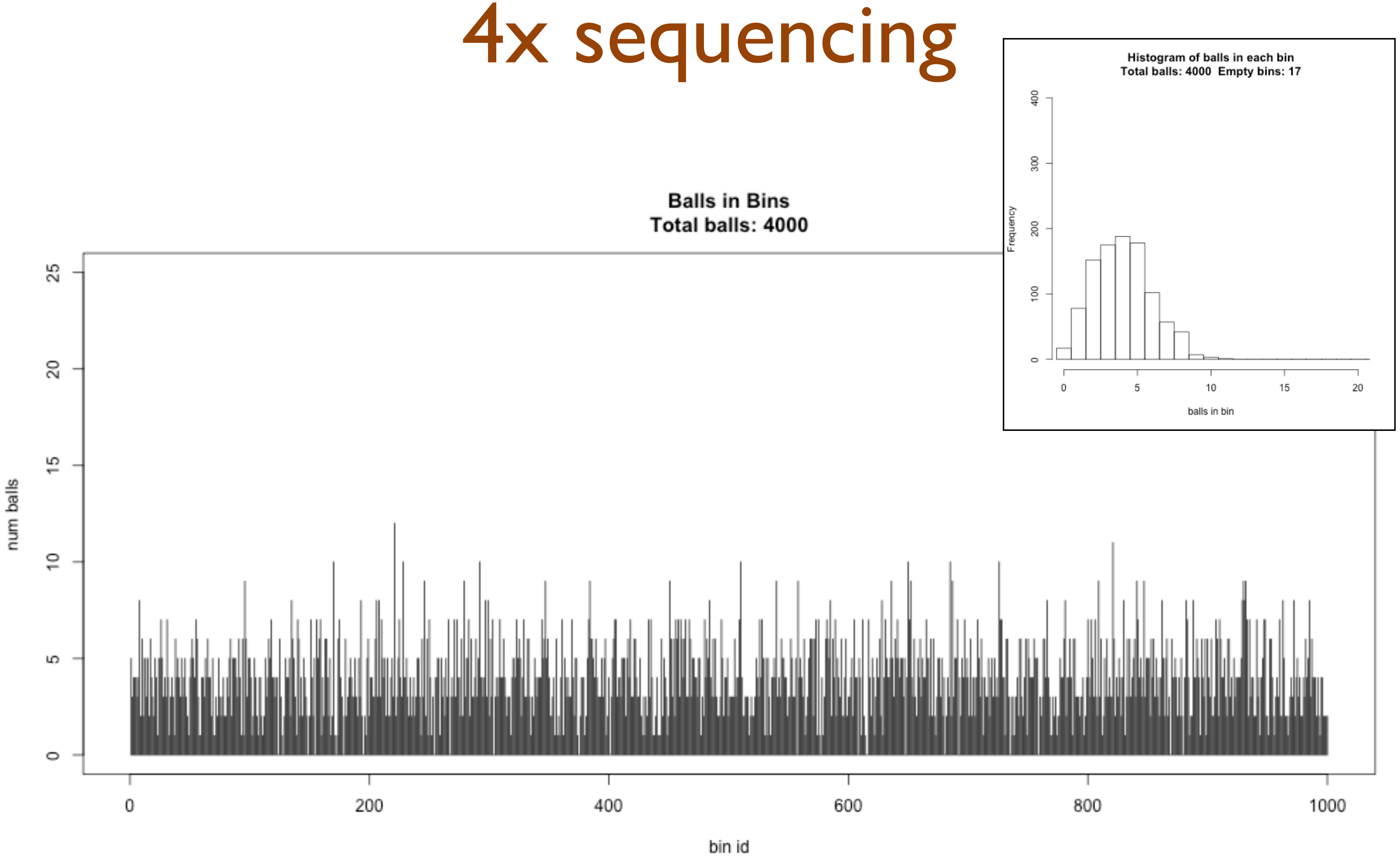
1x sequencing



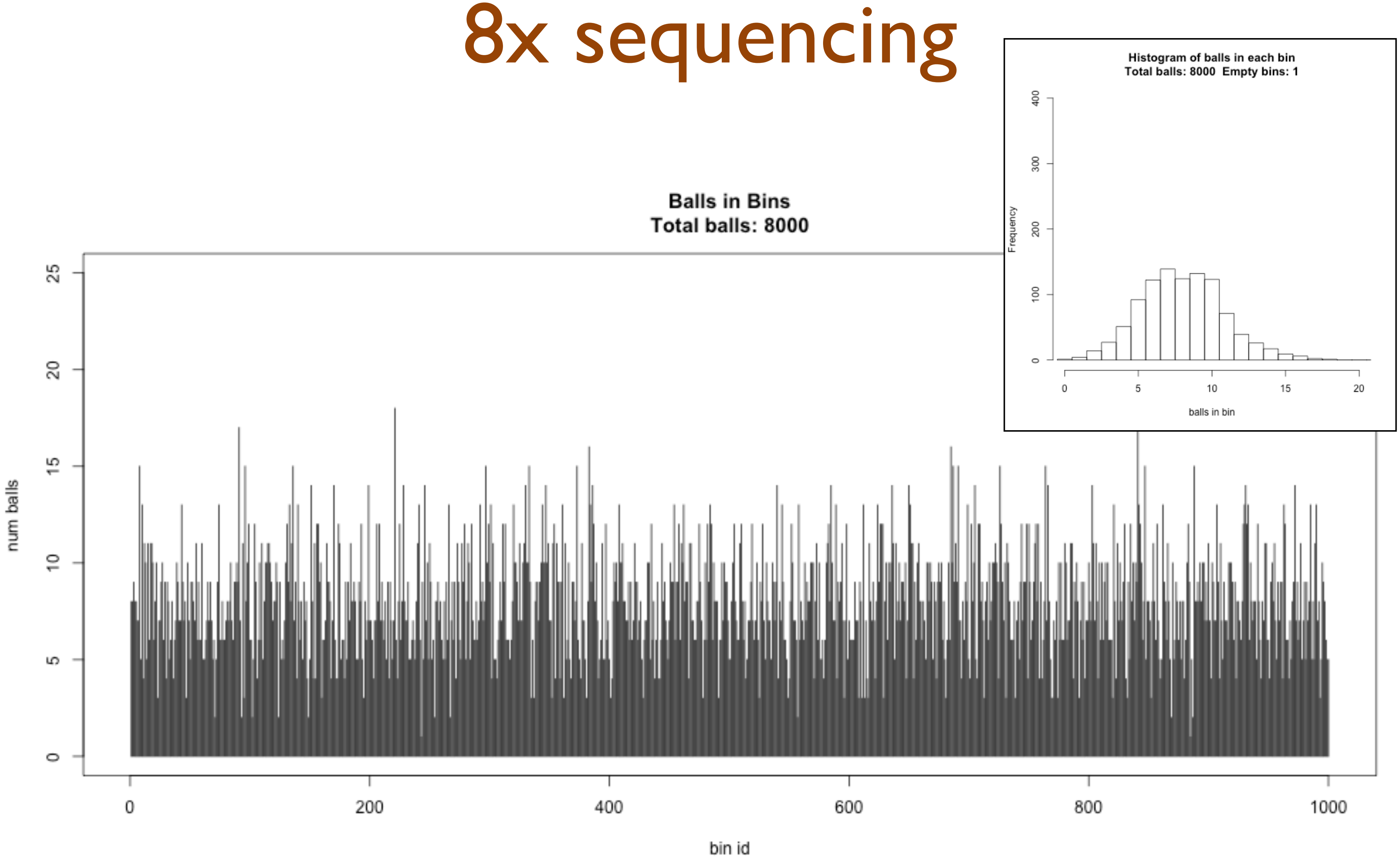
2x sequencing



4x sequencing



8x sequencing



Poisson Distribution

The probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.

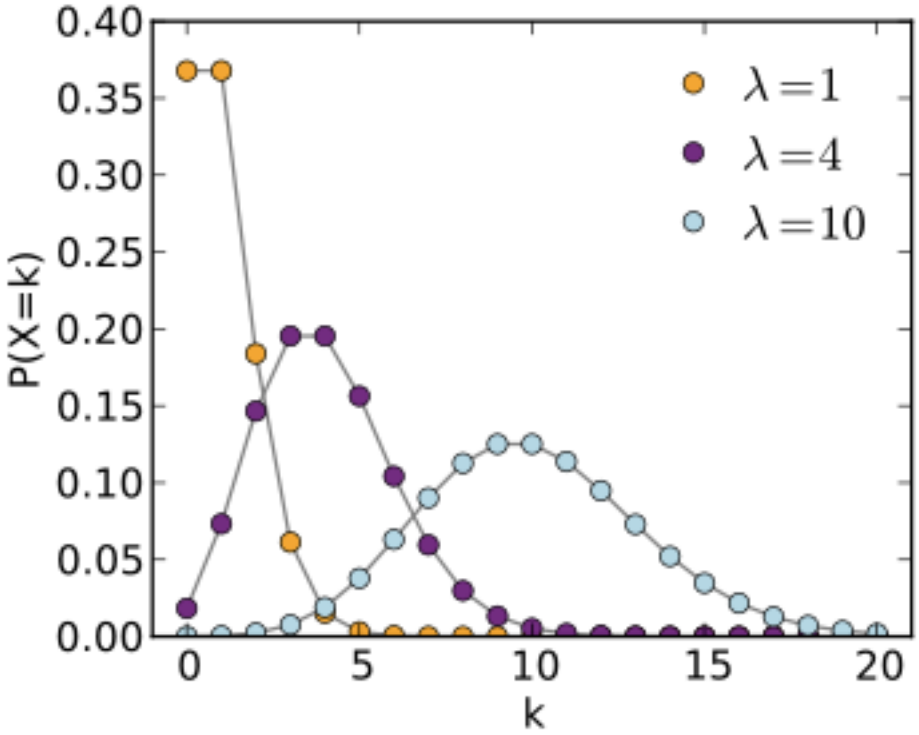
Formulation comes from the limit of the binomial equation

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

Key property:

- ***The standard deviation is the square root of the mean.***

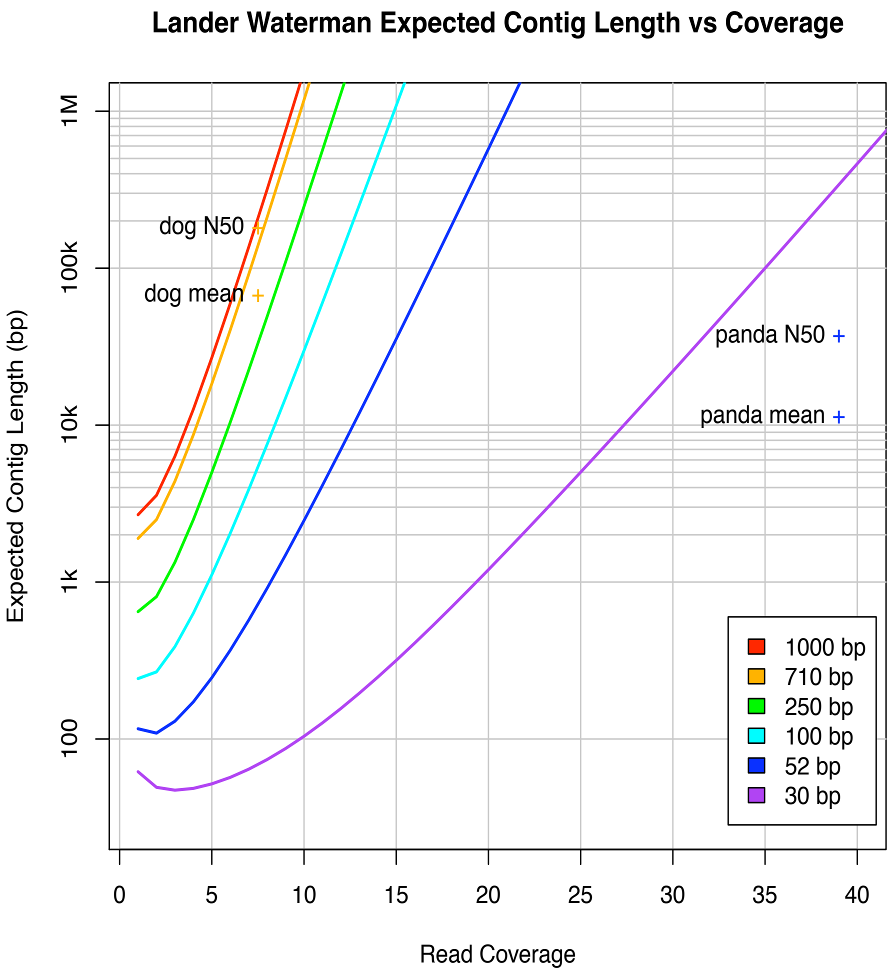
$$P(k) = \frac{\lambda^k}{k!}e^{-\lambda}$$



Coverage and Read Length

Idealized Lander-Waterman model

- Reads start at perfectly random positions
- Contig length is a function of coverage and read length
  - Short reads require much higher coverage to reach same expected contig length
- Need even high coverage for higher ploidy, sequencing errors, sequencing biases
  - Recommend 100x coverage

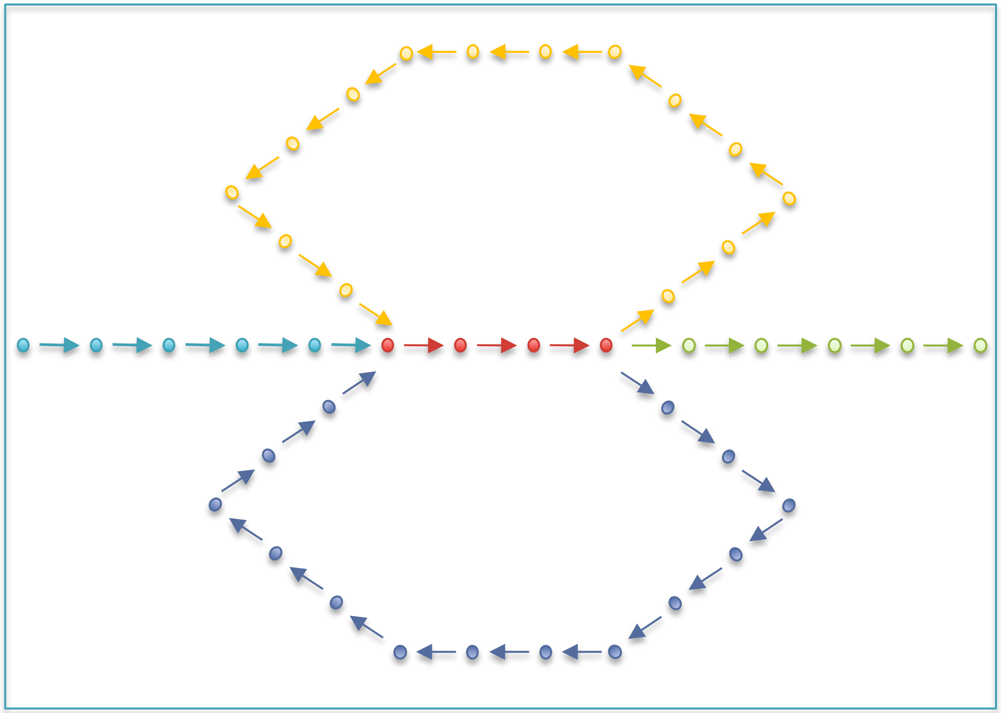


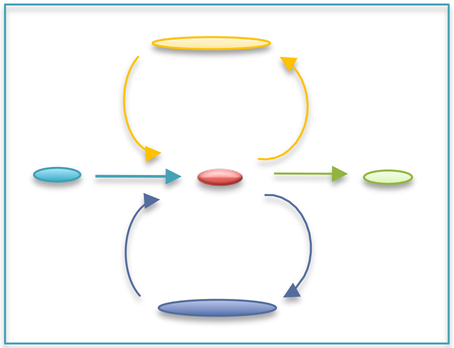
Assembly of Large Genomes using Second Generation Sequencing
Schatz MC, Delcher AL, Salzberg SL (2010) *Genome Research.* 20:1165-1173.

Errors

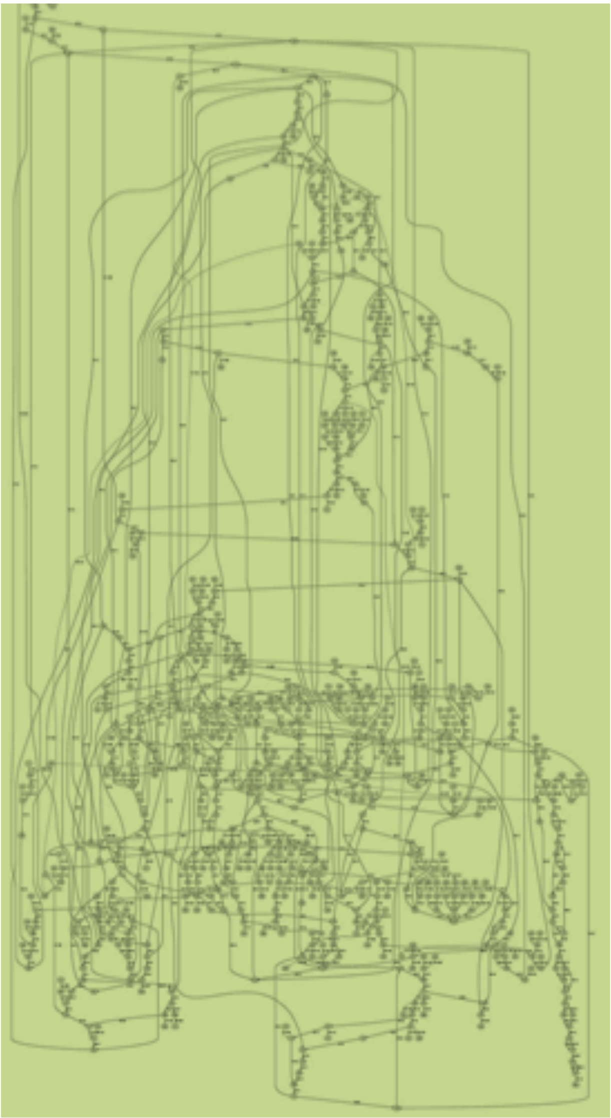
Unitigging / Unipathing

- After simplification and correction, compress graph down to its non-branching initial contigs
  - Aka "unitigs", "unipaths"
  - Unitigs end because of (1) lack of coverage, (2) errors, and (3) repeats

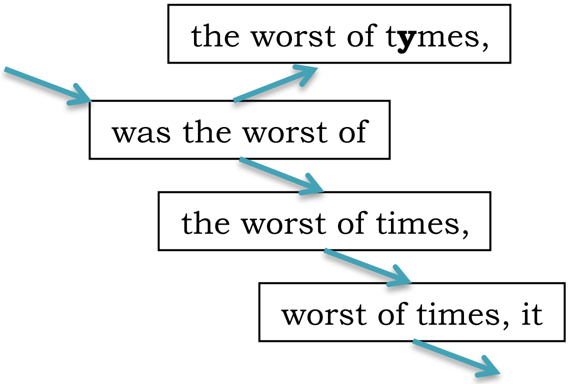
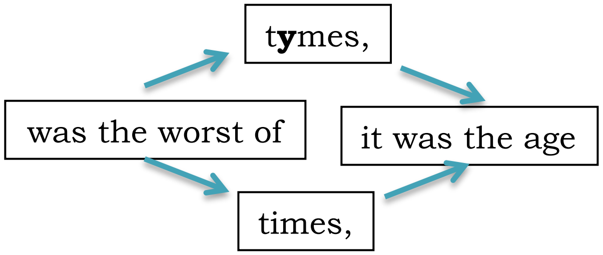




Errors in the graph



(Chaisson, 2009)

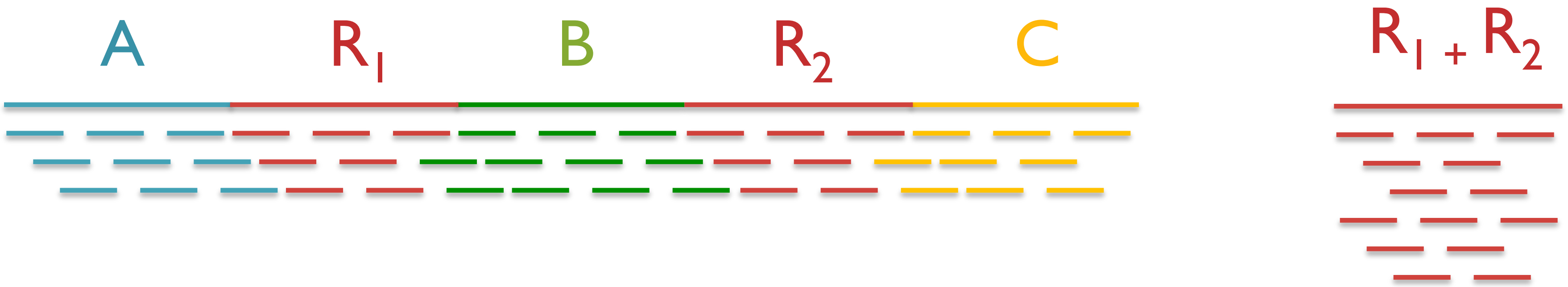
Clip Tips	Pop Bubbles
was the worst of times, was the worst of t**y**mes, the worst of times, it	was the worst of times, was the worst of t**y**mes, times, it was the age t**y**mes, it was the age
 the worst of t**y**mes, was the worst of the worst of times, worst of times, it	 t**y**mes, was the worst of it was the age times,

Repetitive regions

Repeat Type	Definition / Example	Prevalence
Low-complexity DNA / Microsatellites	$(b_1b_2\ldots b_k)^N$ 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%

- – 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 Δ* /*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

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

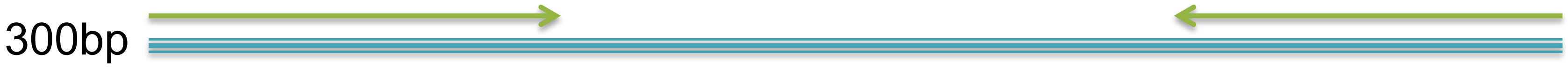
$$A(\Delta,k)=\ln\left(\frac{\Pr(1-copy)}{\Pr(2-copy)}\right)=\ln\left(\frac{\frac{(\Delta n/G)^{k}}{k!}e^{\frac{-\Delta n}{G}}}{\frac{(2\Delta n/G)^{k}}{k!}e^{\frac{-2\Delta n}{G}}}\right)=\frac{n\Delta}{G}-k\ln 2$$

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

Paired-end and Mate-pairs

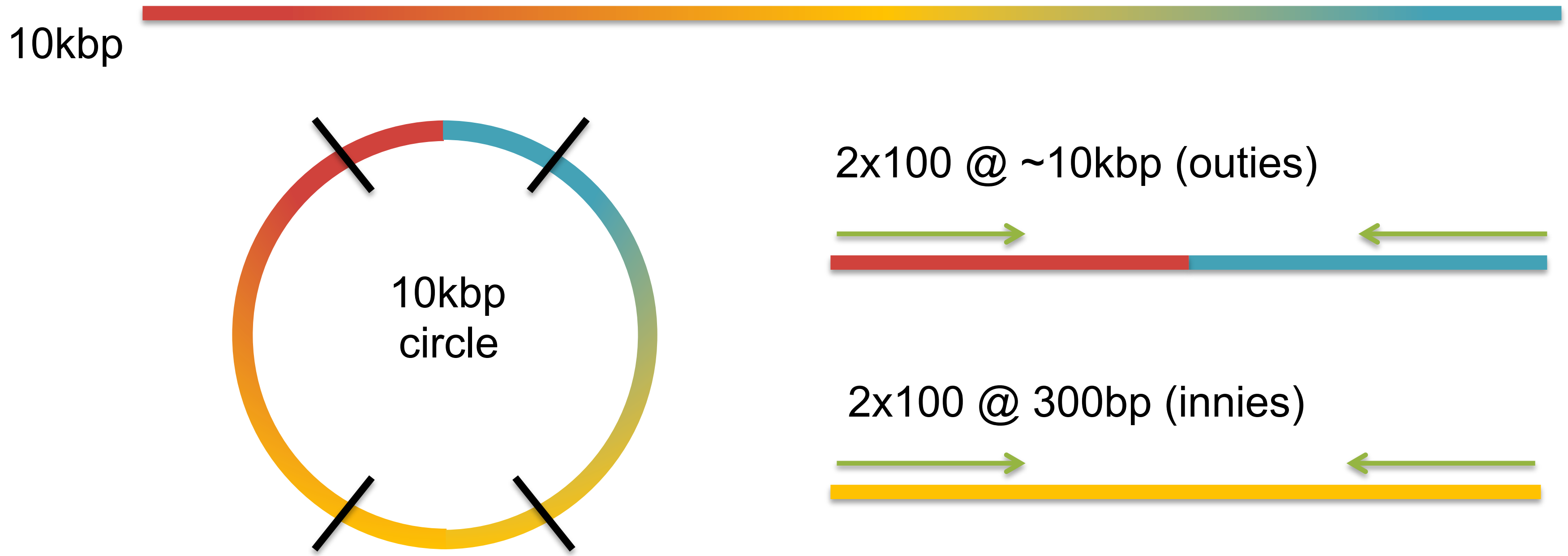
Paired-end sequencing

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



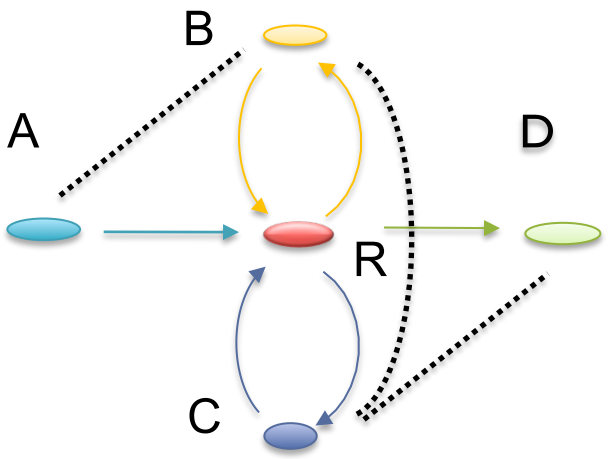
Mate-pair sequencing

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

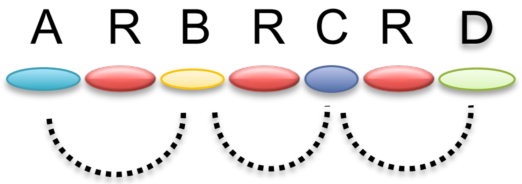


Scaffolding

- Initial contigs (*aka* unipaths, unitigs) terminate at
  - *Coverage gaps*: especially extreme GC
  - *Conflicts*: errors, repeat boundaries



- Use mate-pairs to resolve correct order through assembly graph
  - Place sequence to satisfy the mate constraints
  - Mates through repeat nodes are tangled

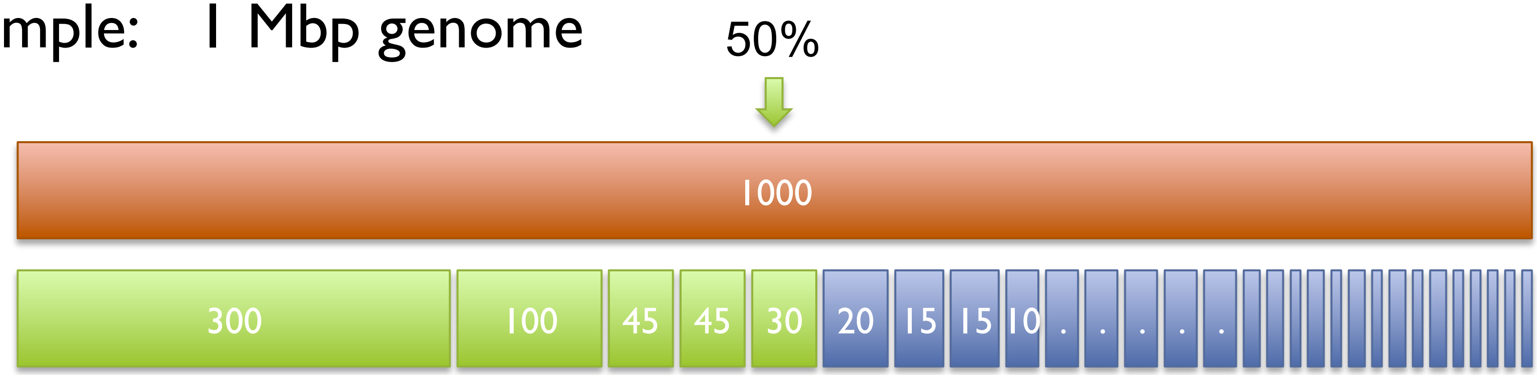


- Final scaffold may have internal gaps called sequencing gaps
  - We know the order, orientation, and spacing, but just not the bases. Fill with Ns instead

N50 size

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

Example: 1 Mbp genome

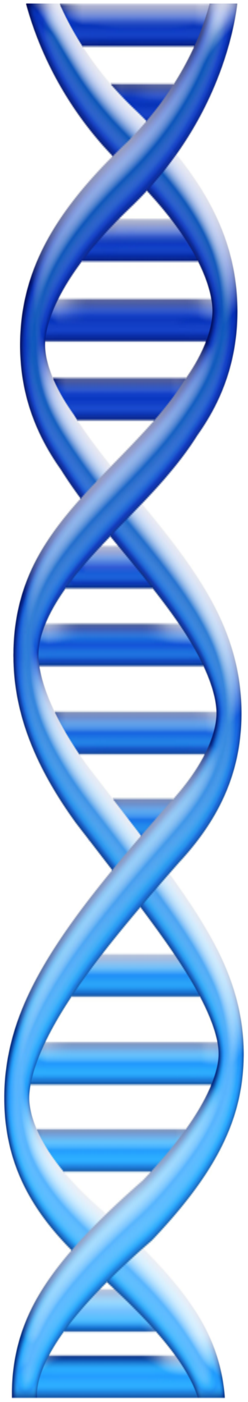


N50 size = 30 kbp

(300k+100k+45k+45k+30k = 520k >= 500kbp)

Note:

N50 values are only meaningful to compare when base genome size is the same in all cases



Outline

1. Assembly theory
   1. Assembly by analogy
   2. De Bruijn and Overlap graph
   3. Coverage, read length, errors, and repeats

2. Whole Genome Alignment
   1. Aligning & visualizing with MUMmer

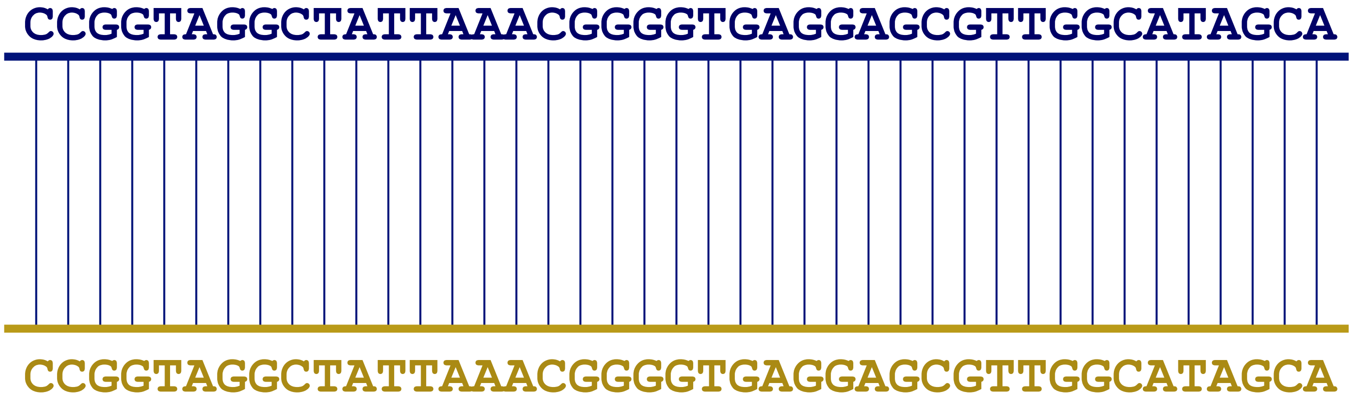
3. Genome assemblers
   1. ALLPATHS-LG: recommended for Illumina-only projects
   2. Celera Assembler: recommended for PacBio projects

Whole Genome Alignment with MUMmer

Slides Courtesy of Adam M. Phillippy
University of Maryland

Goal of WGA

- For two genomes, *A* and *B*, find a mapping from each position in *A* to its corresponding position in *B*

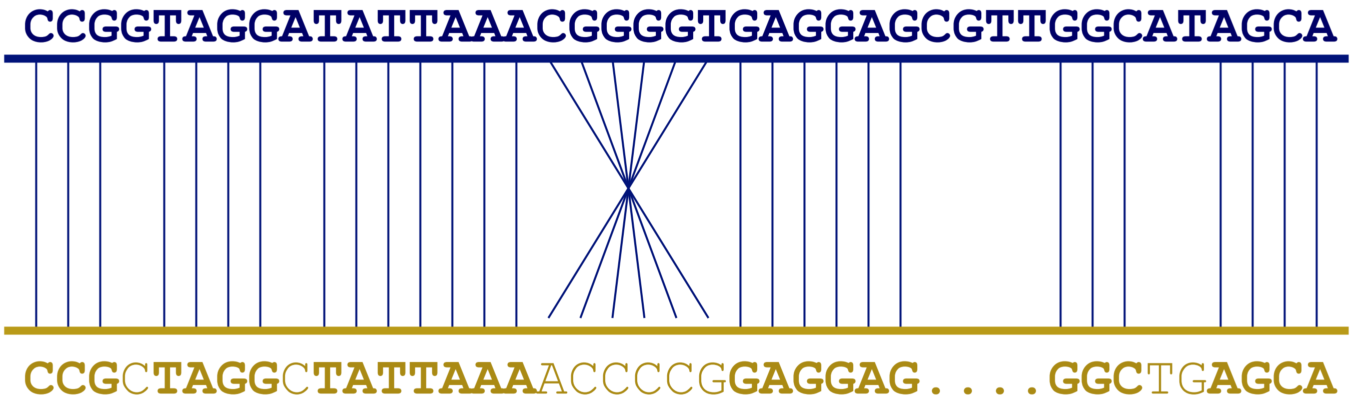


CCGGTAGGCTATTAAACGGGGTGAGGAGCGTTGGCATAGCA

CCGGTAGGCTATTAAACGGGGTGAGGAGCGTTGGCATAGCA

Not so fast...

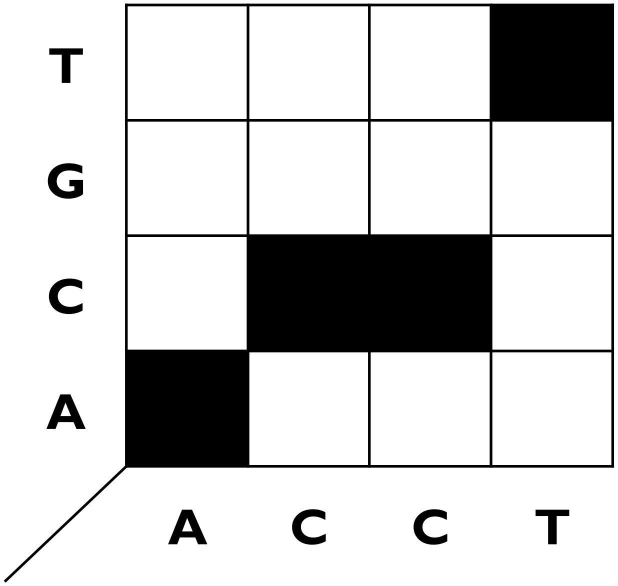
- Genome *A* may have insertions, deletions, translocations, inversions, duplications or SNPs with respect to *B* (sometimes all of the above)


CCGGTAGGATATTAAACGGGTGAGGAGCGTTGGCATAGCA
CCGCTAGGCTATTAAAACCCCGGAGGAG....GGCTGAGCA

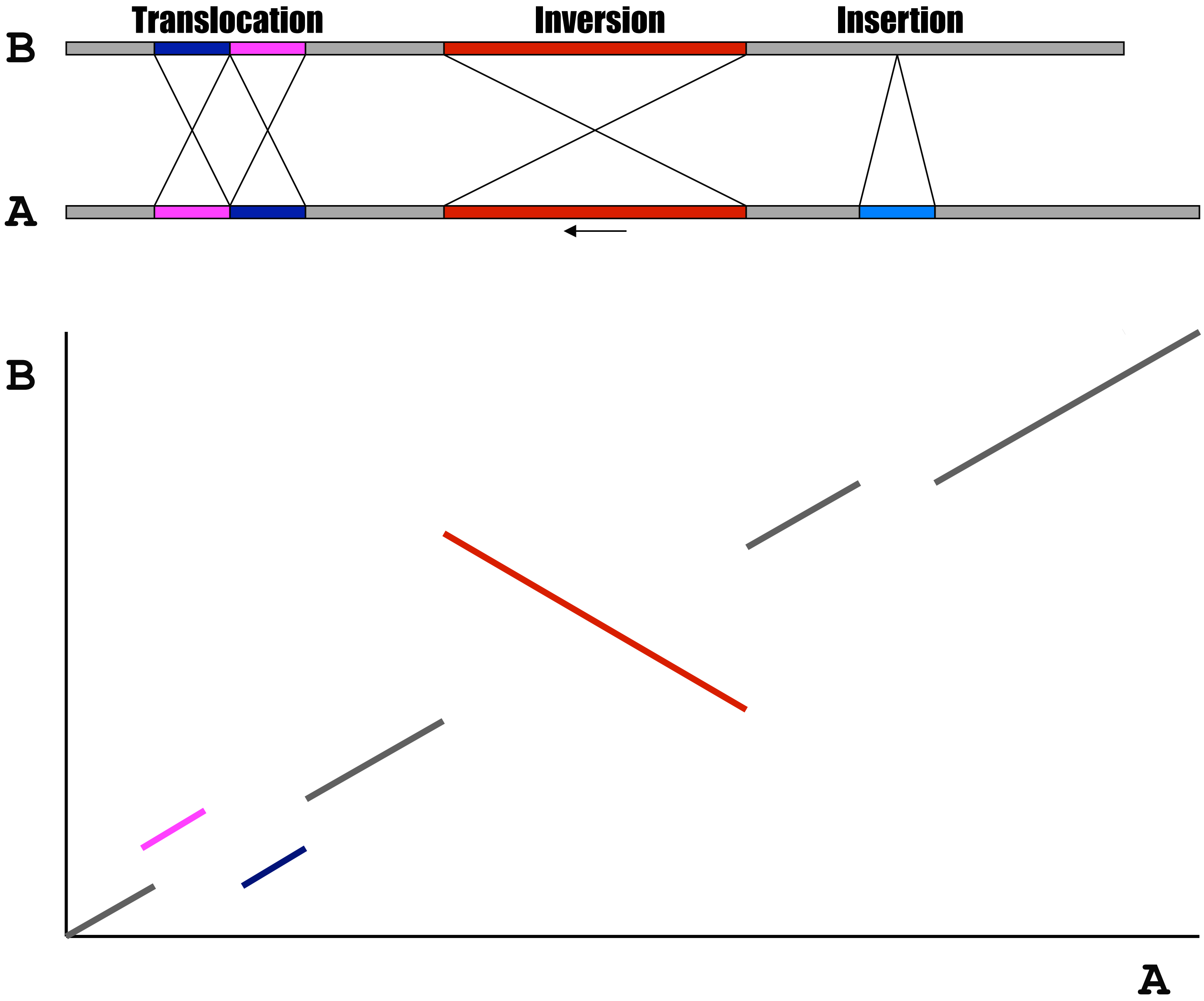
WGA visualization

- How can we visualize *whole* genome alignments?

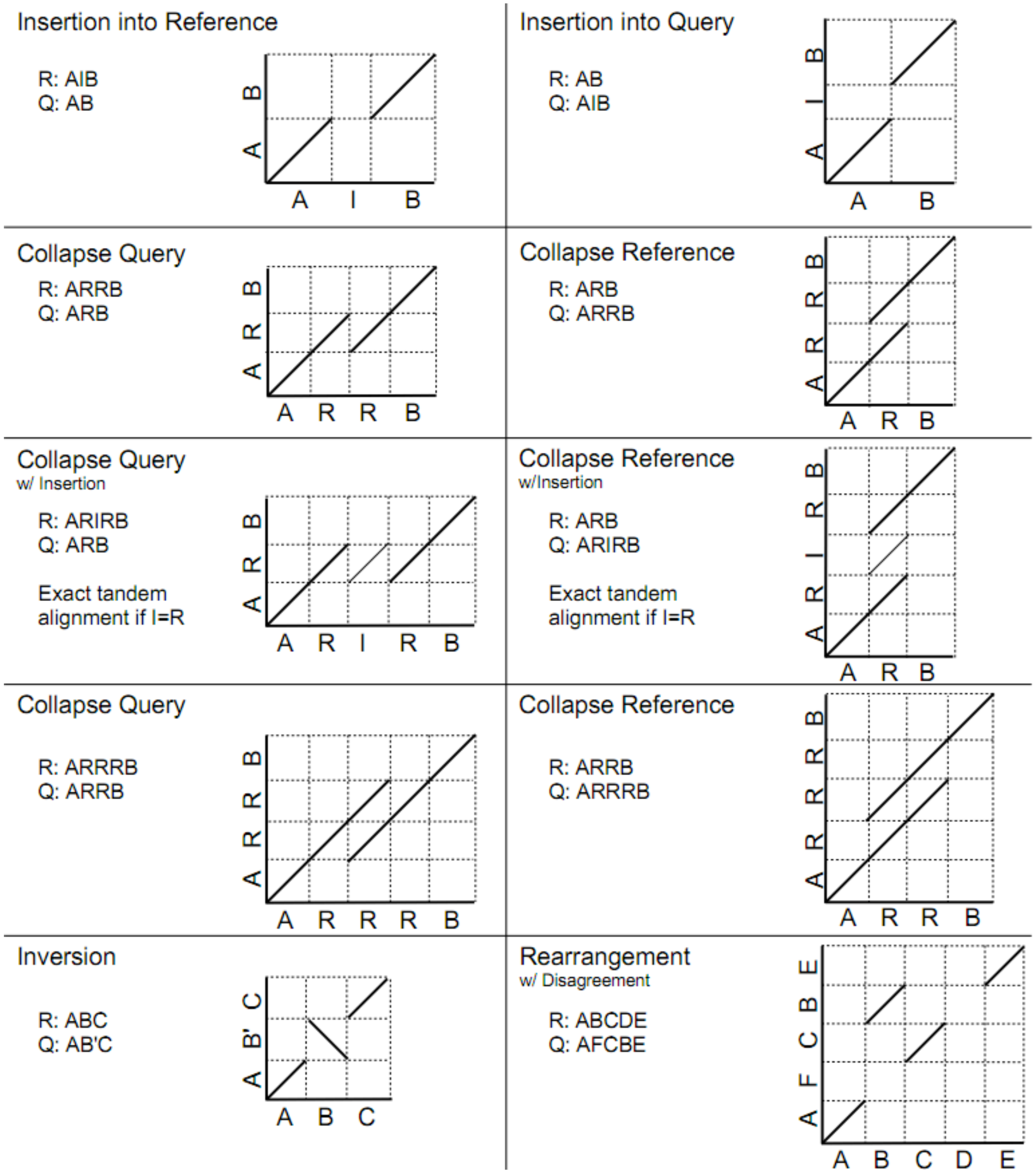
- With an alignment dot plot
  - *N* x *M* matrix
    - Let *i* = position in genome *A*
    - Let *j* = position in genome *B*
    - Fill cell *(i,j)* if A_i shows similarity to B_j



  - A perfect alignment between *A* and *B* would completely fill the positive diagonal

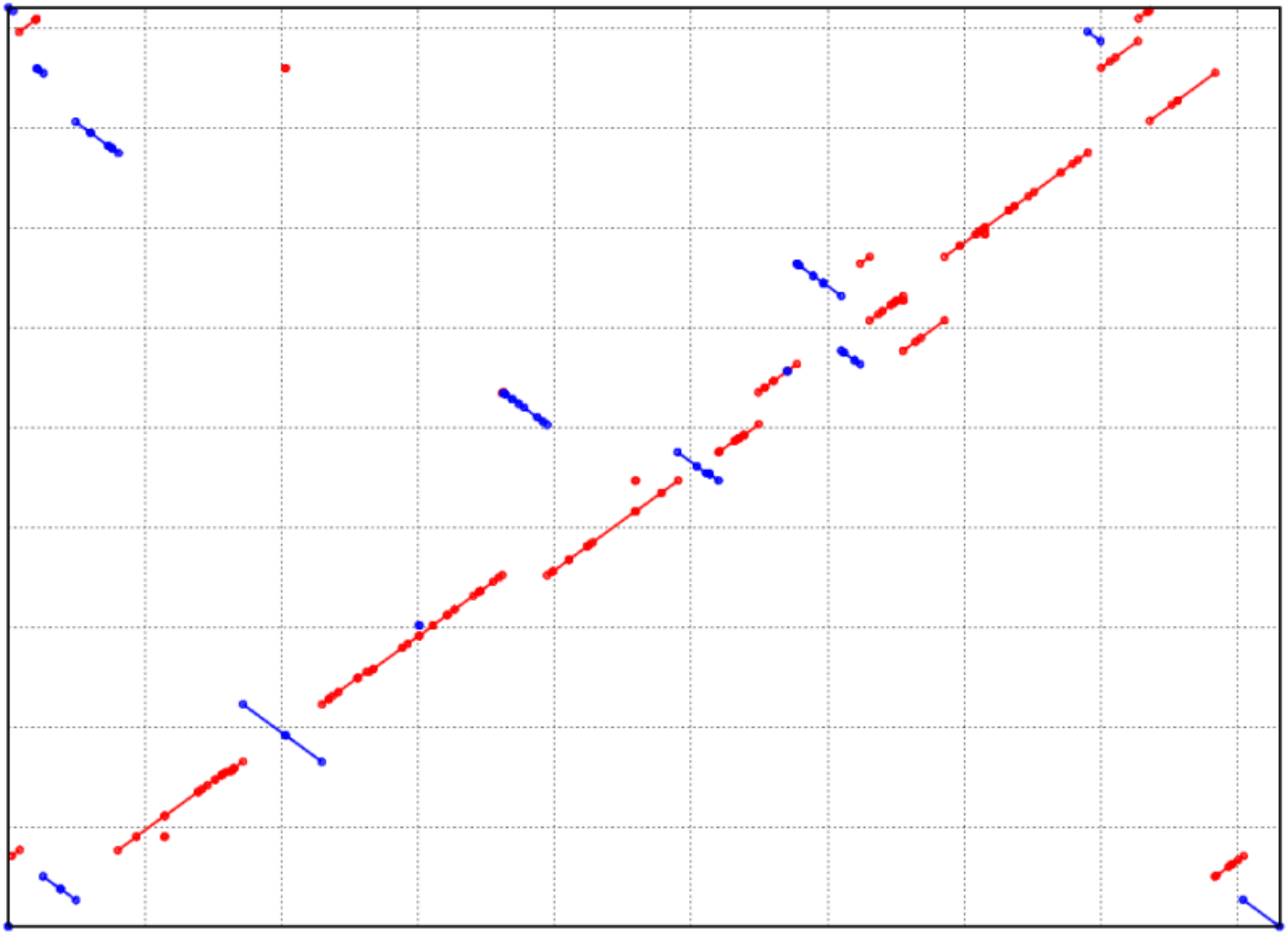

Translocation
Inversion
Insertion
B
A
B
A

SV Types



- Different structural variation types / misassemblies will be apparent by their pattern of breakpoints
- Most breakpoints will be at or near repeats
- Things quickly get complicated in real genomes

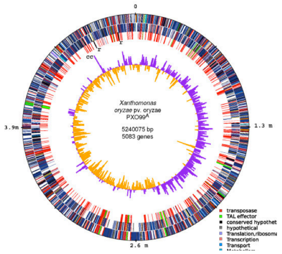
http://mummer.sf.net/manual/AlignmentTypes.pdf



Alignment of 2 strains of *Y. pestis*

http://mummer.sourceforge.net/manual/

Assembly Summary



Assembly quality depends on

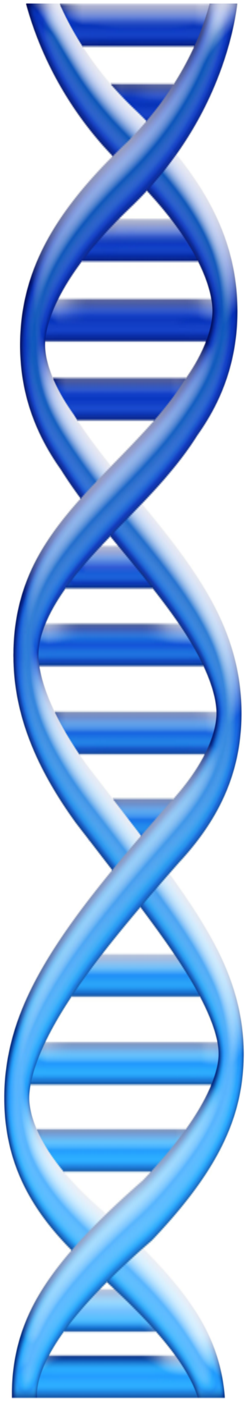
1. ***Coverage***: low coverage is mathematically hopeless
2. ***Repeat composition***: high repeat content is challenging
3. ***Read length***: longer reads help resolve repeats
4. ***Error rate***: errors reduce coverage, obscure true overlaps

- 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

Break

PHASE TWO : INTERPRETATION
SHENEMAN The Star-Ledger
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME

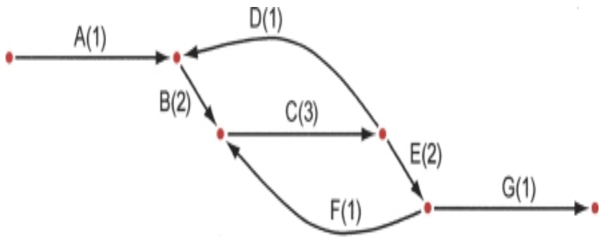


Outline

1. Assembly theory
   1. Assembly by analogy
   2. De Bruijn and Overlap graph
   3. Coverage, read length, errors, and repeats

2. Whole Genome Alignment
   1. Aligning & visualizing with MUMmer

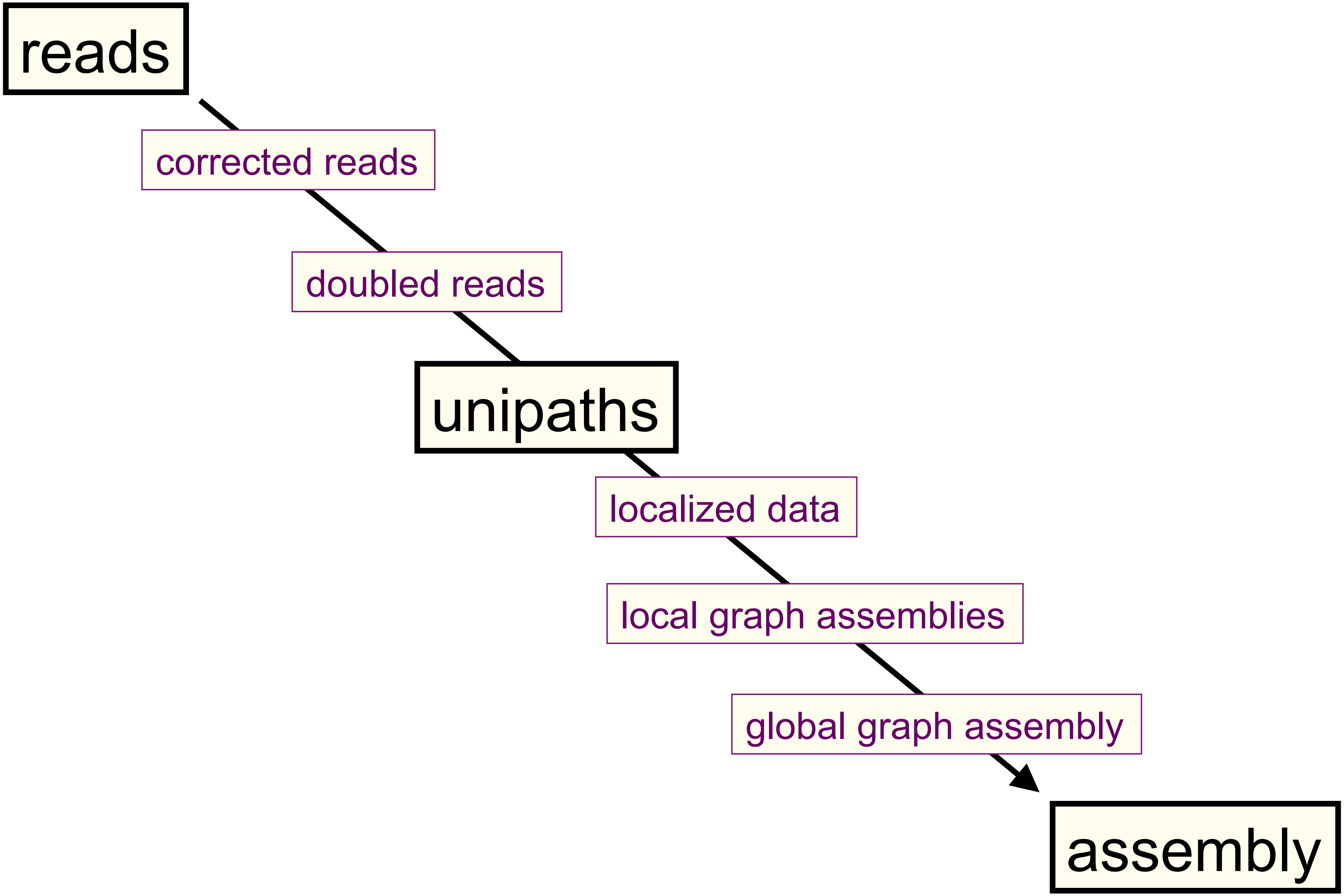
3. Genome assemblers
   1. ALLPATHS-LG: recommended for Illumina-only projects
   2. Celera Assembler: recommended for PacBio projects



Genome assembly with ALLPATHS-LG
Iain MacCallum

BROAD INSTITUTE

How ALLPATHS-LG works



reads

corrected reads

doubled reads

unipaths

localized data

local graph assemblies

global graph assembly

assembly

ALLPATHS-LG sequencing model

Libraries (insert types)	Fragment size (bp)	Read length (bases)	Sequence coverage (x)	Required
Fragment	180*	≥ 100	45	yes
Short jump	3,000	≥ 100 preferable	45	yes
Long jump	6,000	≥ 100 preferable	5	no**
Fosmid jump	40,000	≥ 26	1	no**

*See next slide.

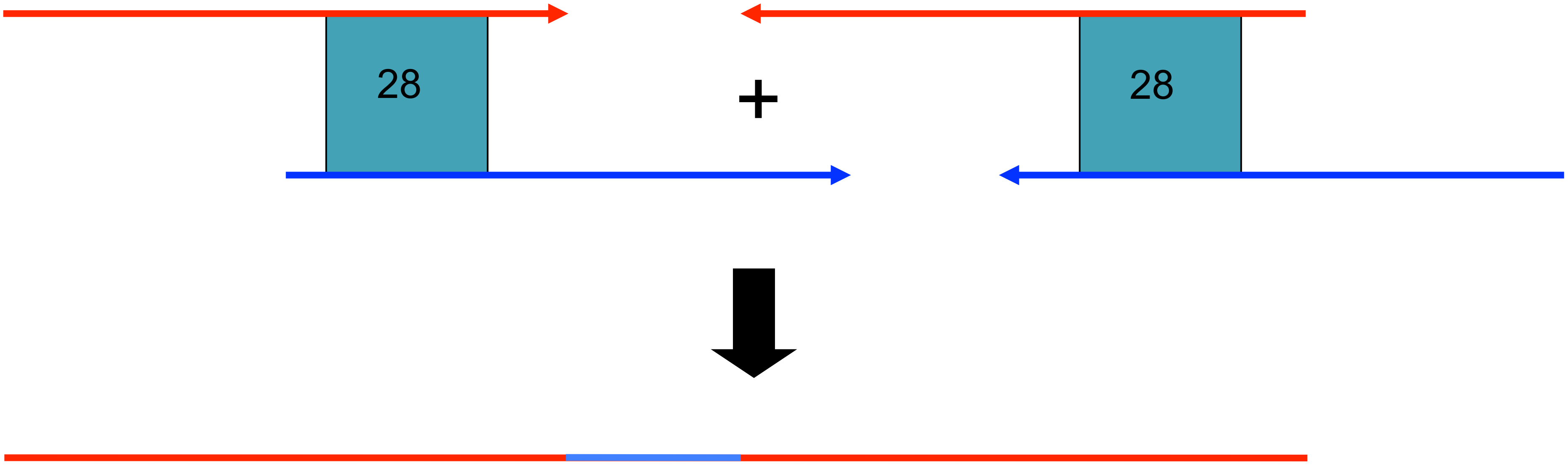
**For best results. Normally not used for small genomes.
However essential to assemble long repeats or duplications.

Cutting coverage in half still works, with some reduction in quality of results.

All: protocols are either available, or in progress.

Read doubling

To close a read pair (red), we require the existence of another read pair (blue), overlapping perfectly like this:



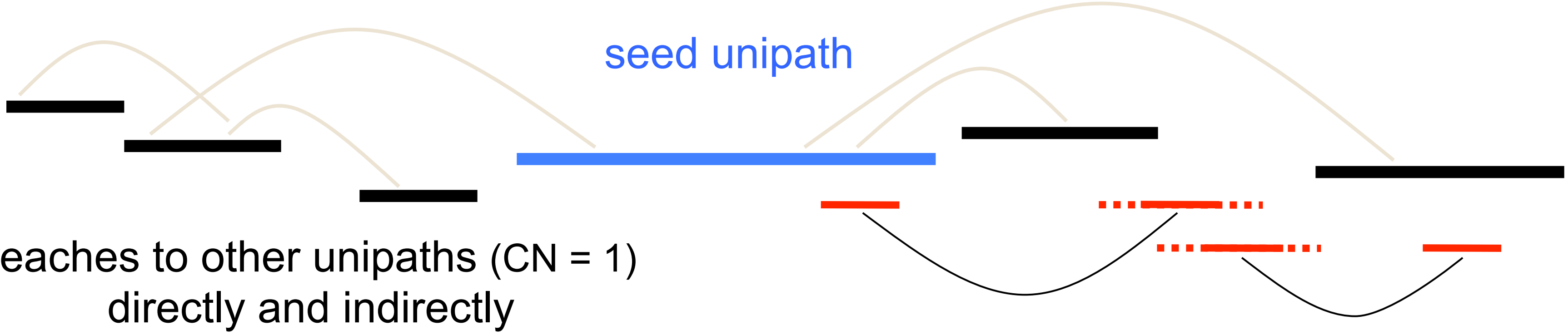
More than one closure allowed (but rare).

Localization

I. Find ‘seed’ unipaths, evenly spaced across genome

(ideally long, of copy number CN = 1)

II. Form neighborhood around each seed



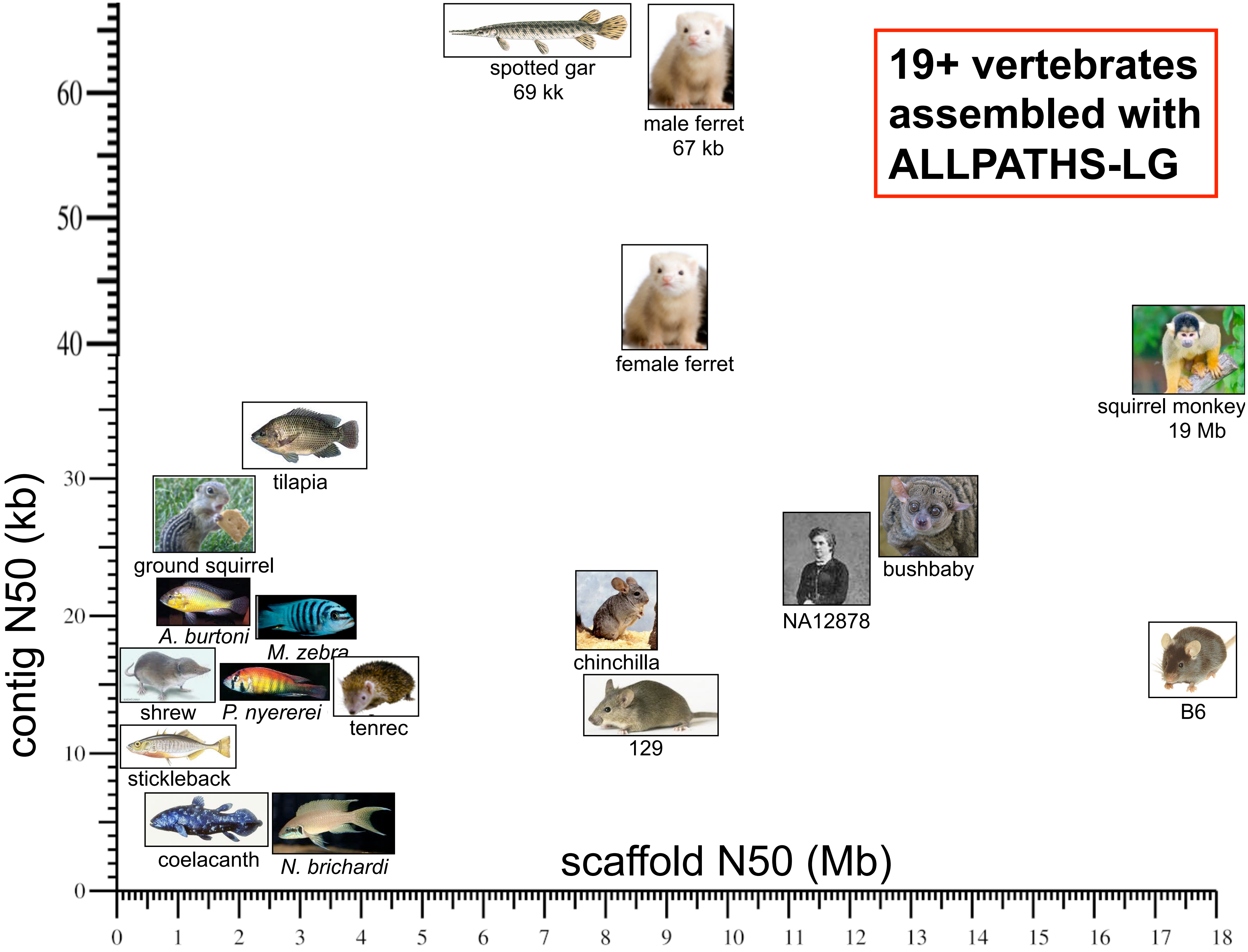
seed unipath

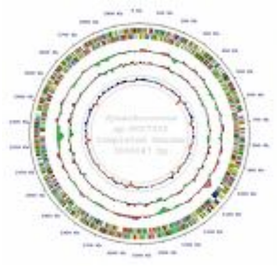
reaches to other unipaths (CN = 1)
directly and indirectly

read pairs reach into repeats

and are extended by other
unipaths

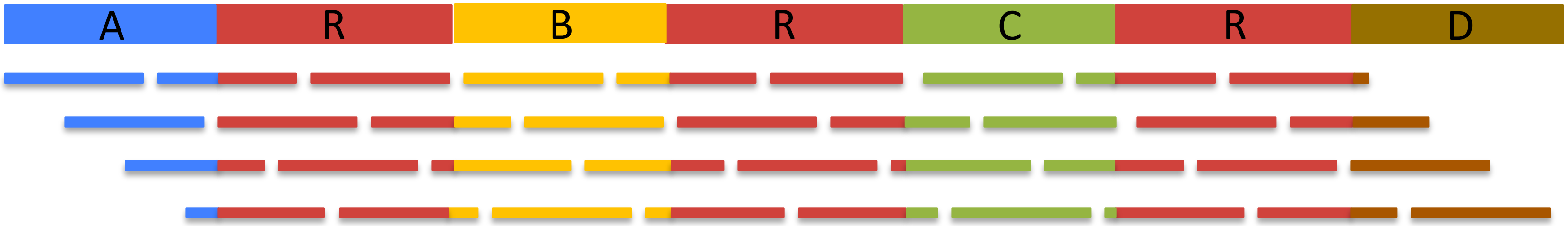
19+ vertebrates assembled with ALLPATHS-LG

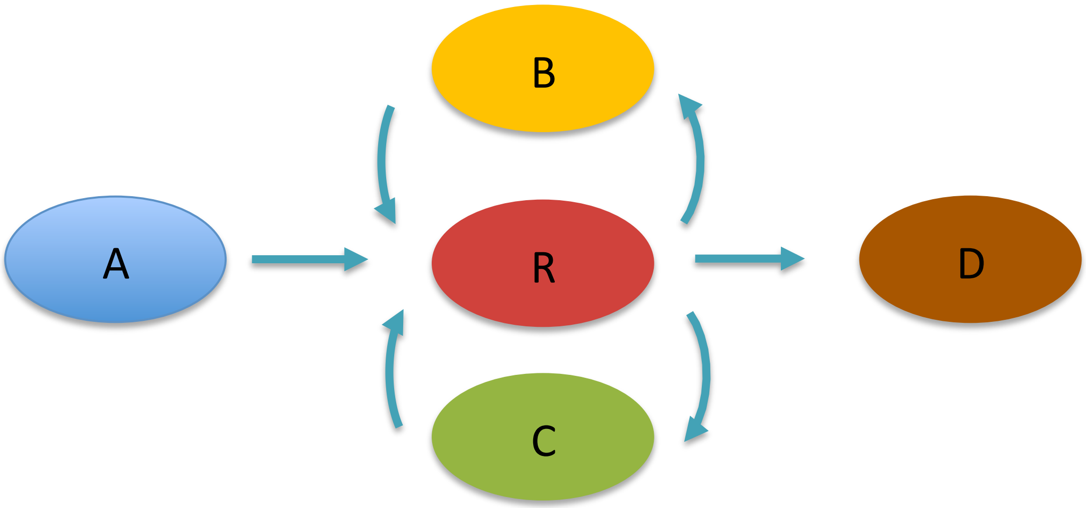




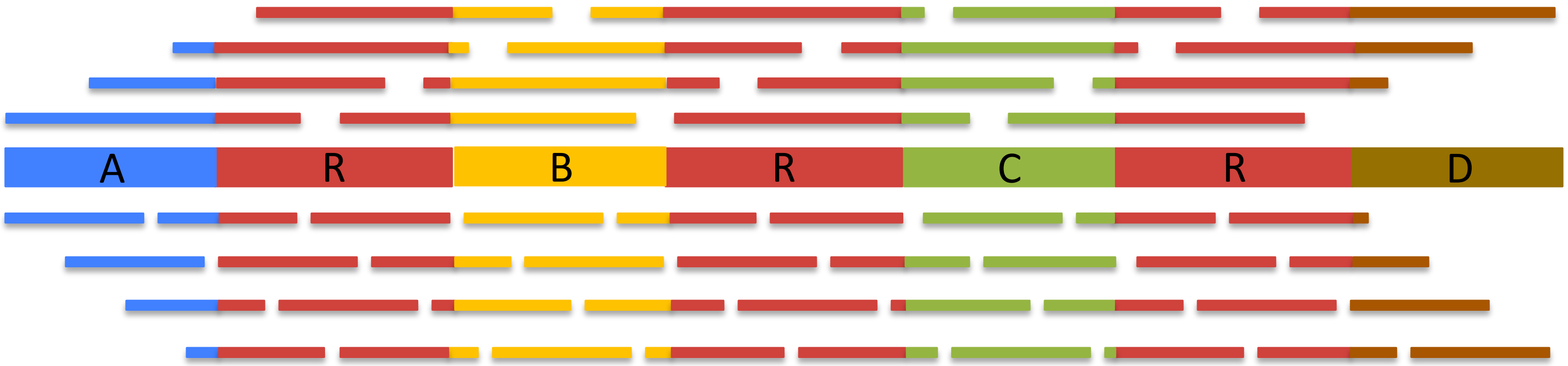
Genome assembly with the Celera Assembler

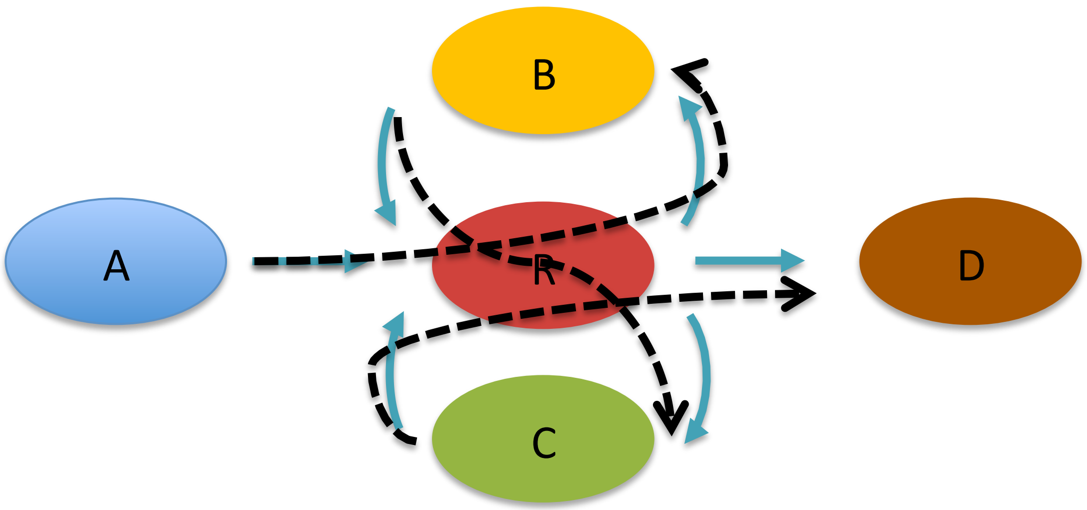
Assembly Complexity



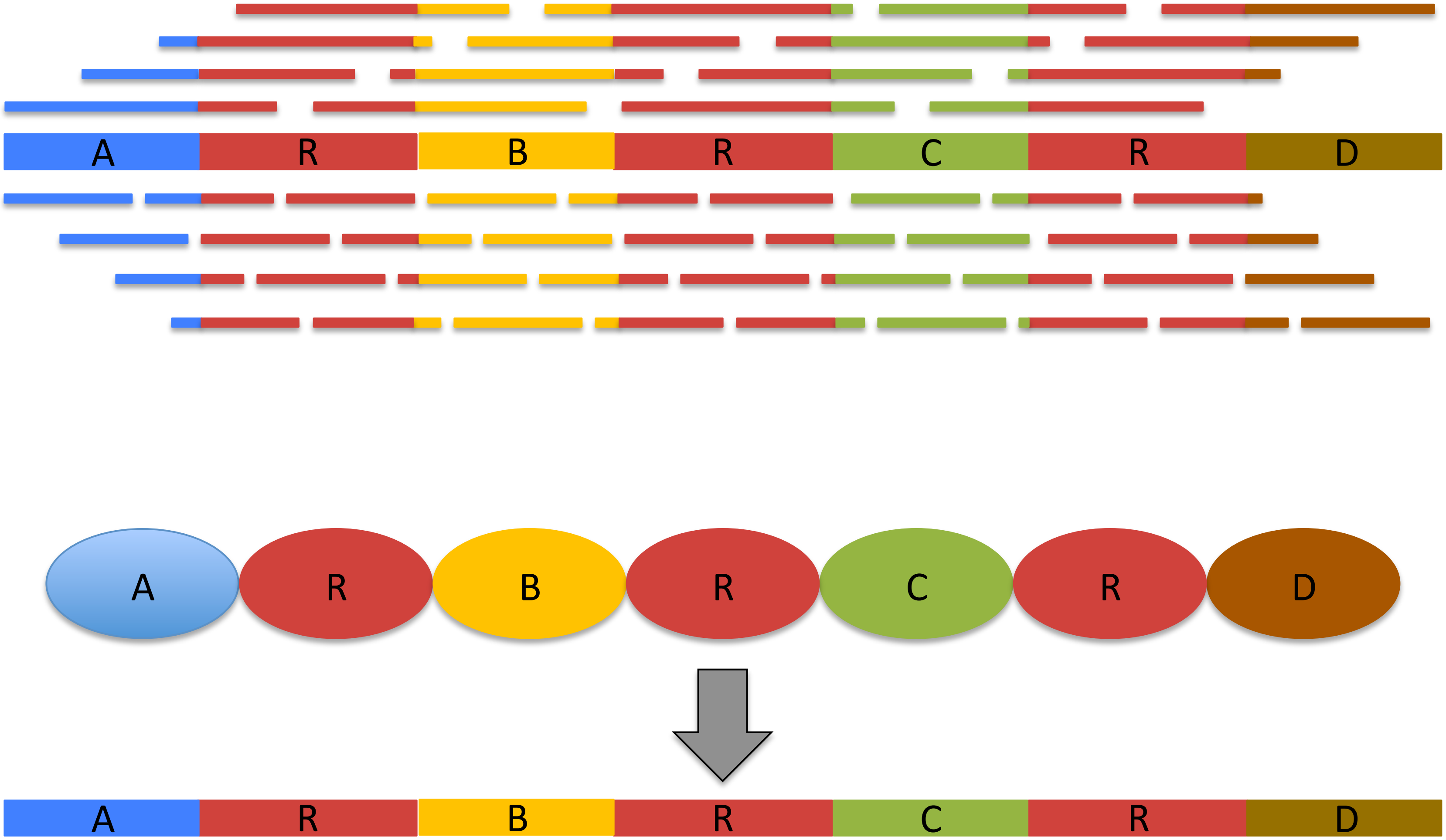


Assembly Complexity





Assembly Complexity



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

Long Read Sequencing Technology

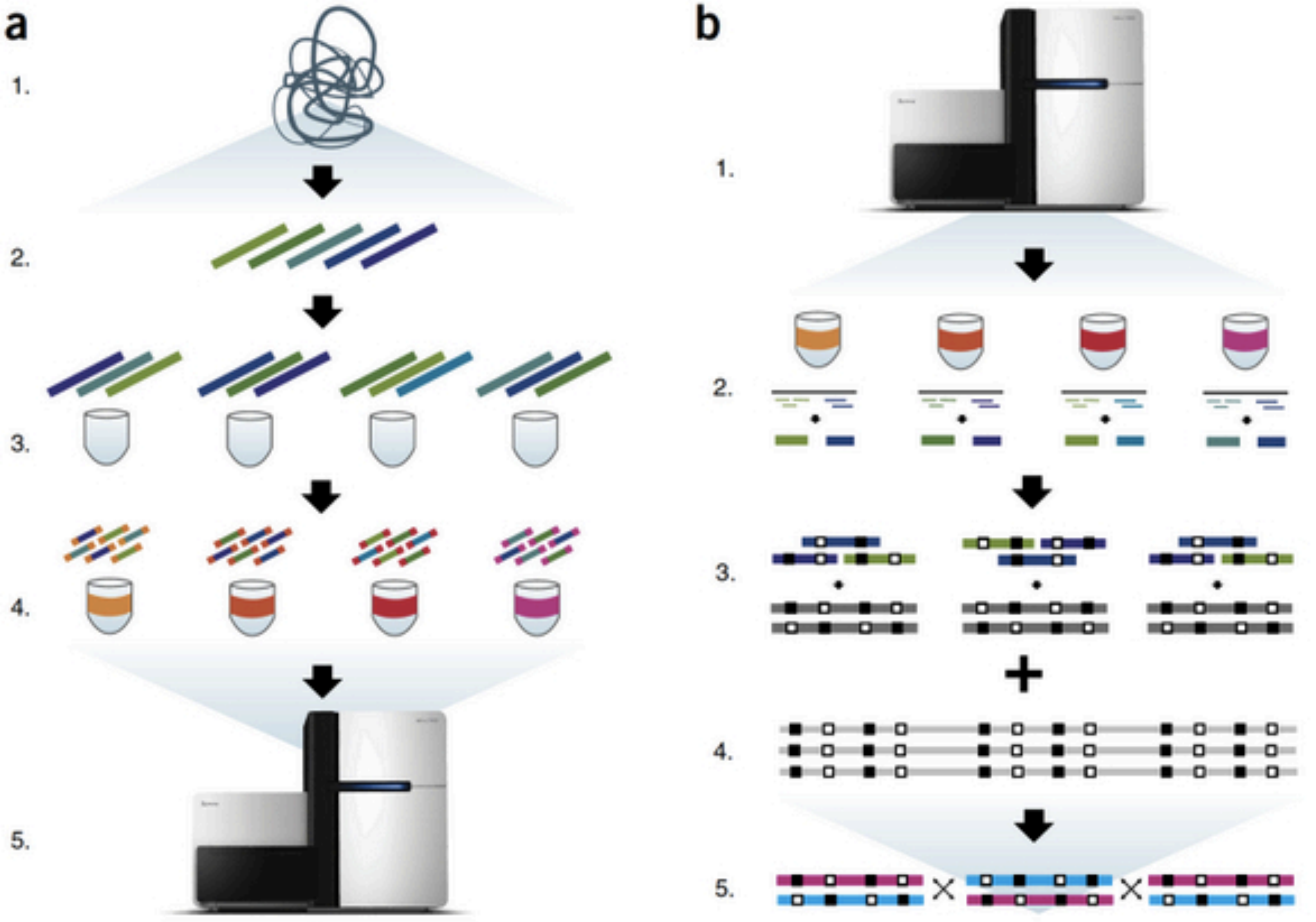
Moleculo

PacBio RS II

Oxford Nanopore

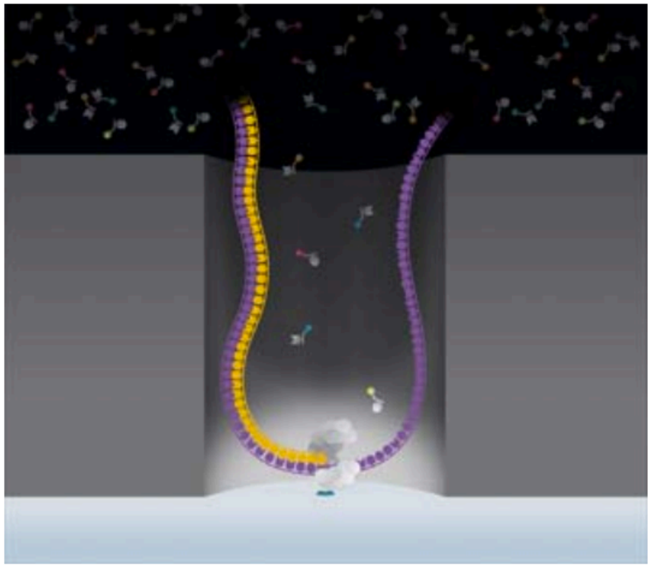
Moleculo Sequencing

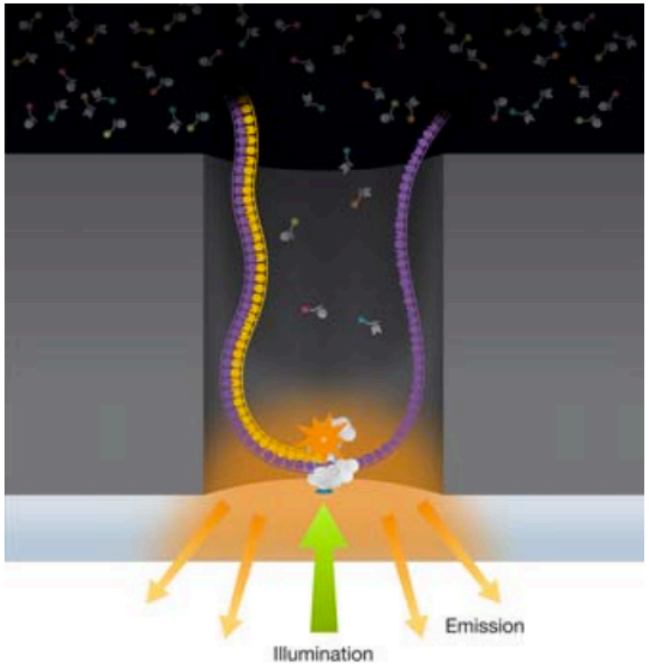
Clever library preparation technique to turn a short read sequencer into a quazi-long read sequencer

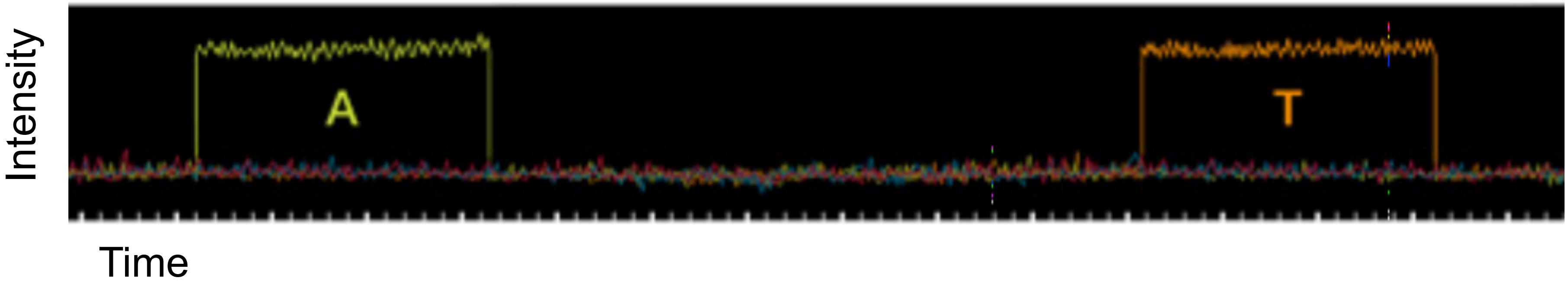


PacBio SMRT Sequencing

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



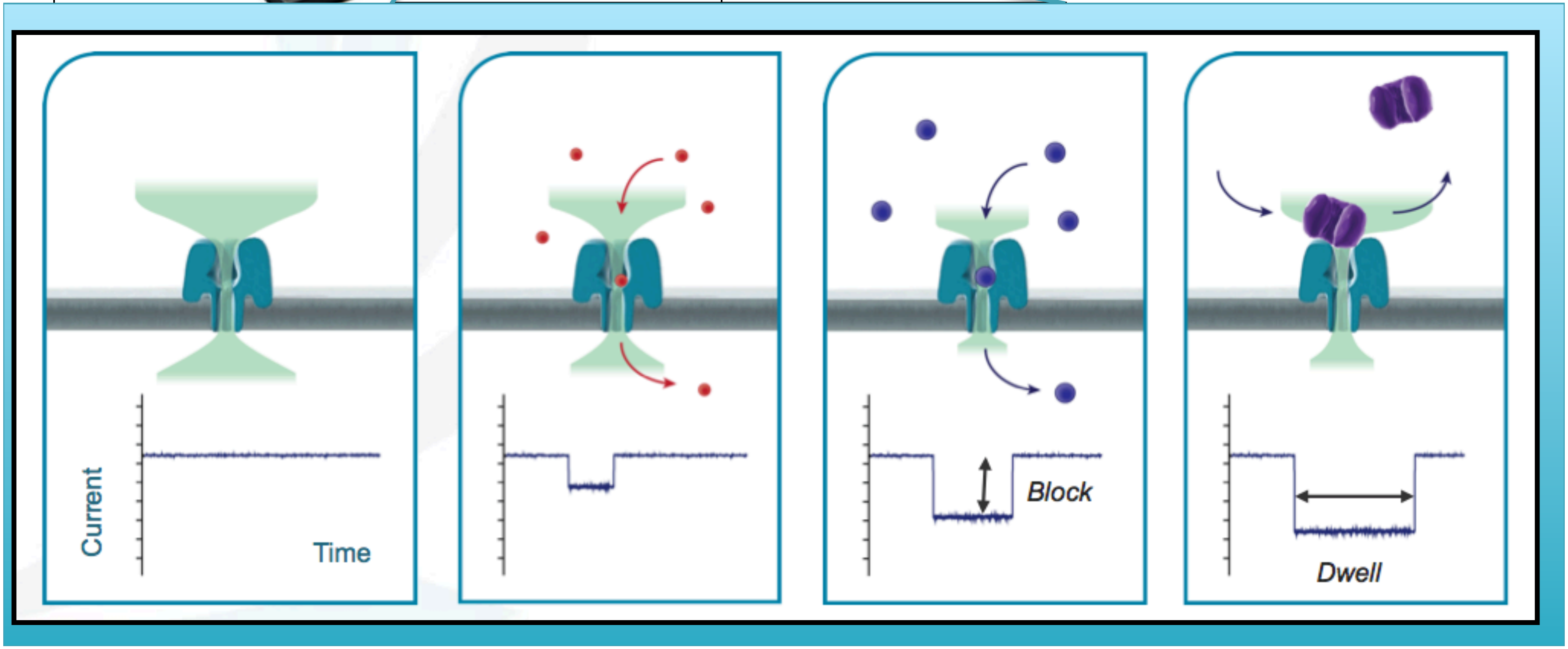




http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf

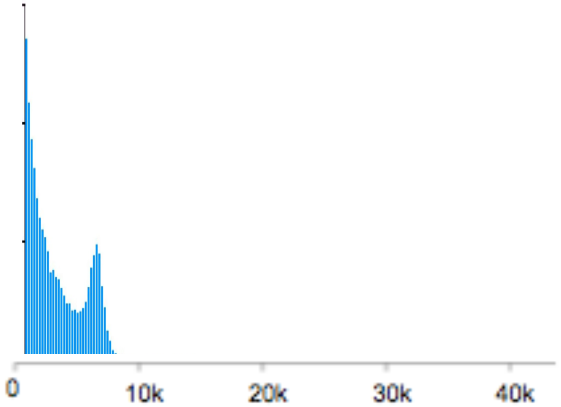
Oxford Nanopore MinION

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



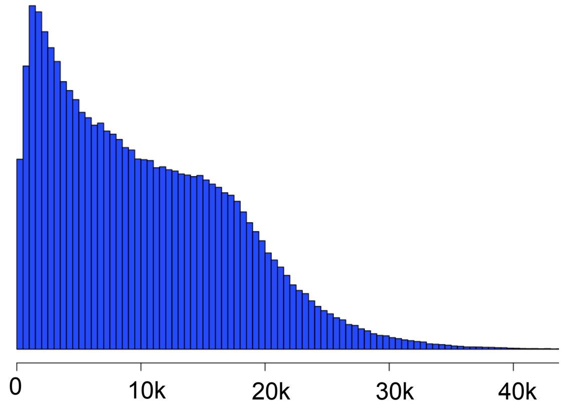
Long Read Sequencing Technology

Moleculo



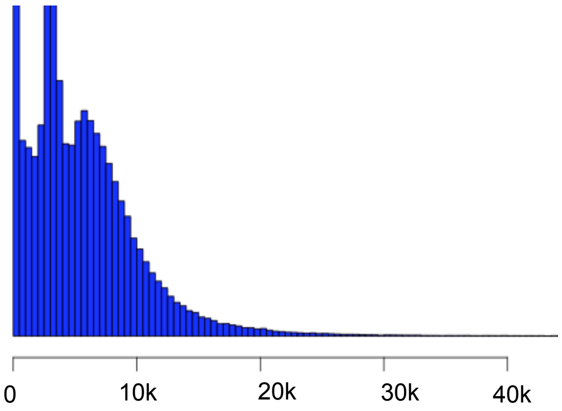
(Voskoboynik et al. 2013)

PacBio RS II



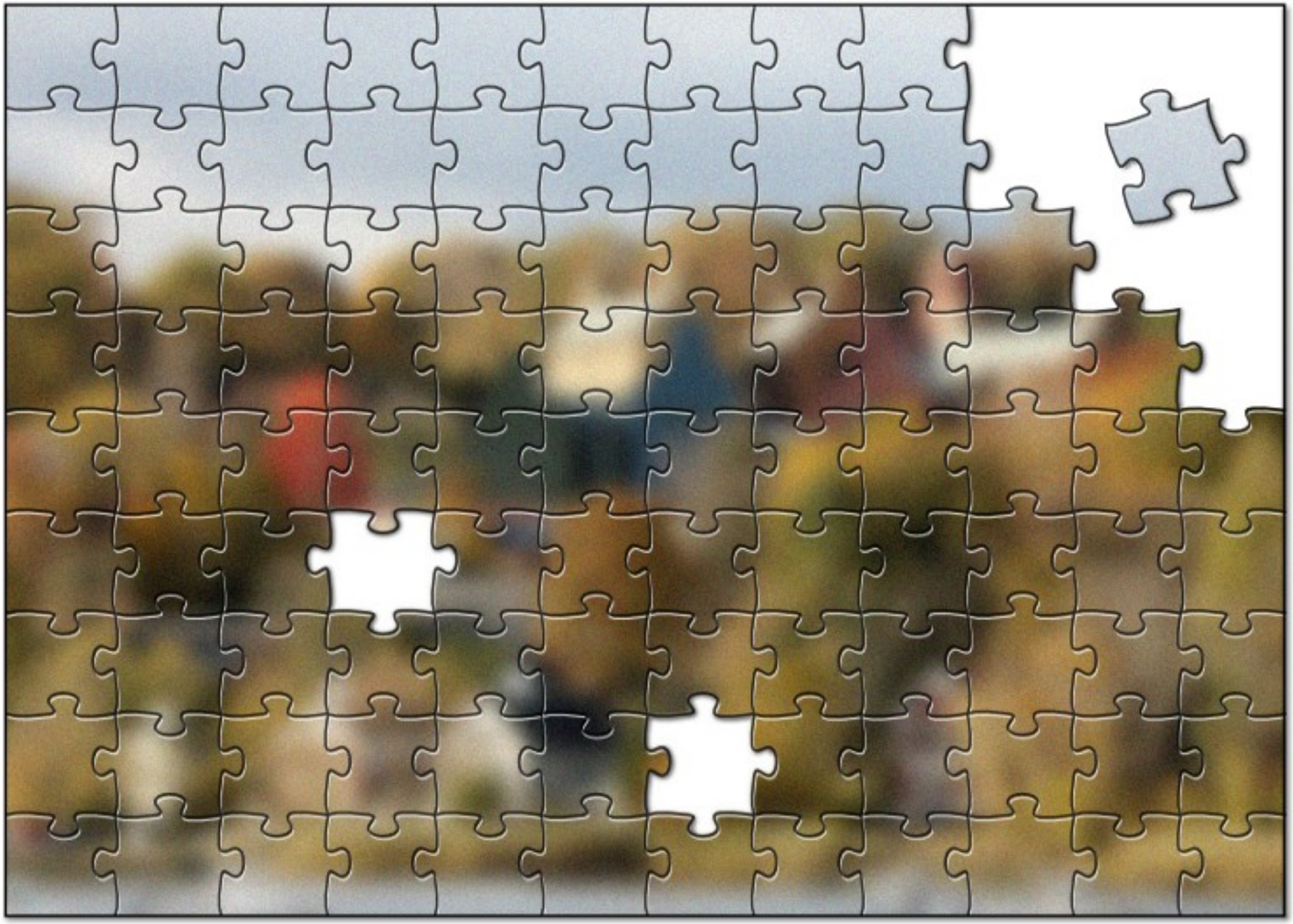
CSHL/PacBio

Oxford Nanopore

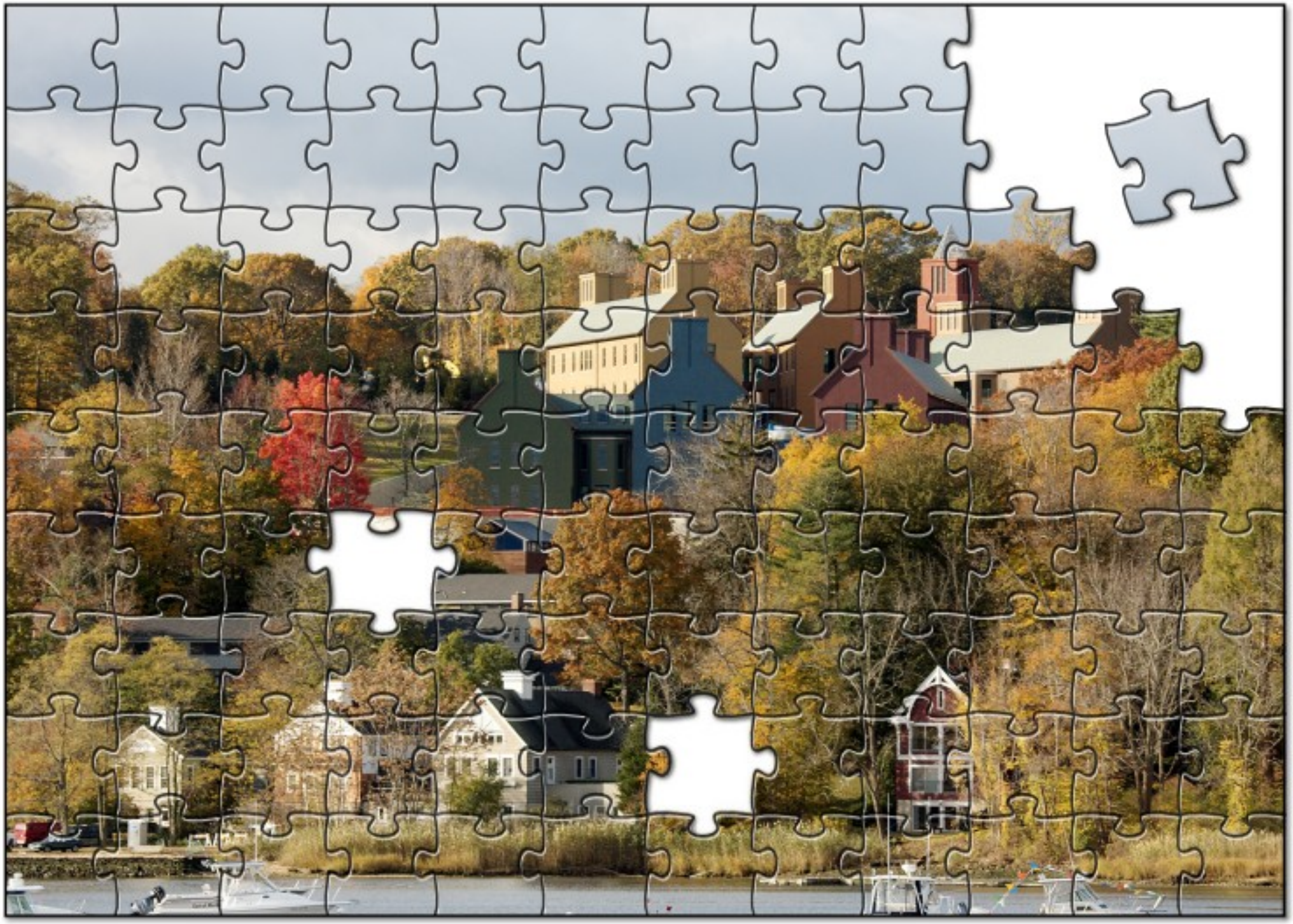


CSHL/ONT

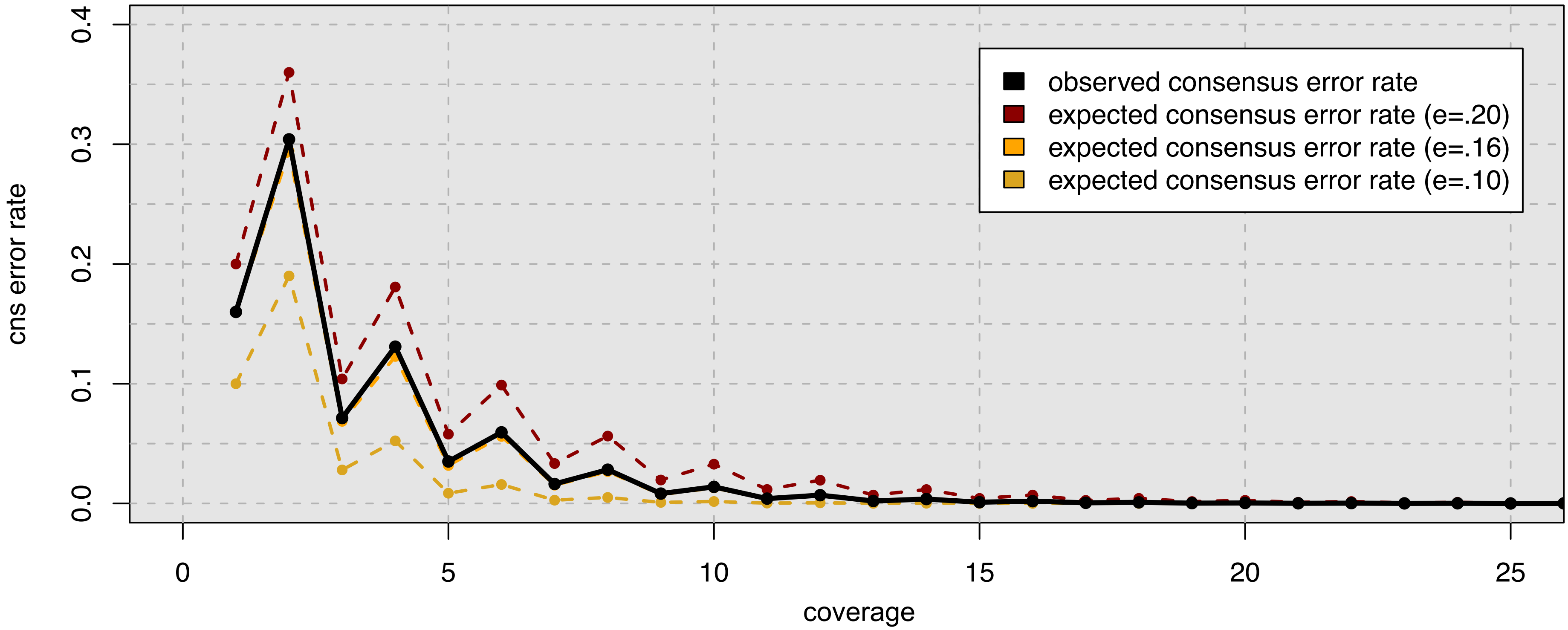
Single Molecule Sequences



“Corrective Lens” for Sequencing



Consensus Accuracy and Coverage



Coverage can overcome random errors

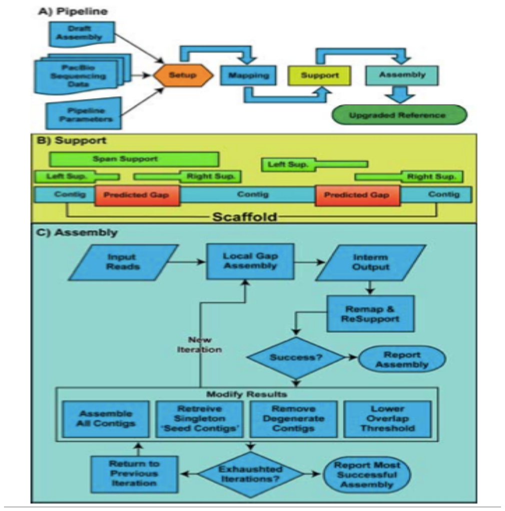
- Dashed: error model from binomial sampling
- Solid: observed accuracy

Koren, Schatz, *et al* (2012)
Nature Biotechnology. 30:693–700

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

PacBio Assembly Algorithms

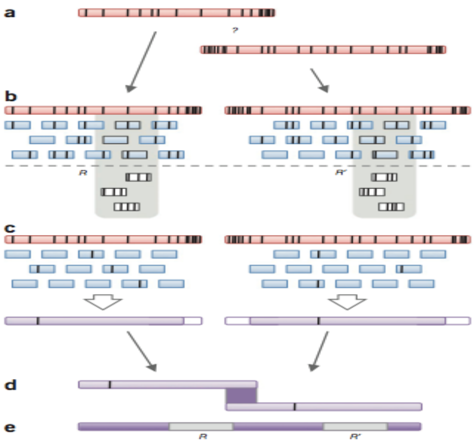
PBJelly



Gap Filling and Assembly Upgrade

English *et al* (2012)
PLOS One. 7(11): e47768

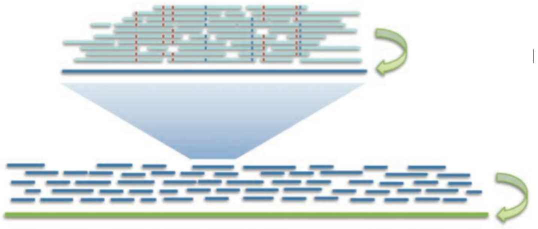
PacBioToCA & ECTools



Hybrid/PB-only Error Correction

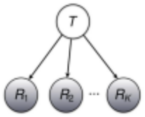
Koren, Schatz, et al (2012)
Nature Biotechnology. 30:693–700

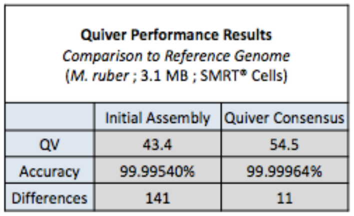
HGAP & Quiver



$\Pr(\mathbf{R} \mid T)$

$$\Pr(\mathbf{R} \mid T) = \prod_k \Pr(R_k \mid T)$$





Quiver Performance Results
Comparison to Reference Genome
(*M. ruber* ; 3.1 MB ; SMRT® Cells)

	Initial Assembly	Quiver Consensus
QV	43.4	54.5
Accuracy	99.99540%	99.99964%
Differences	141	11

PB-only Correction & Polishing

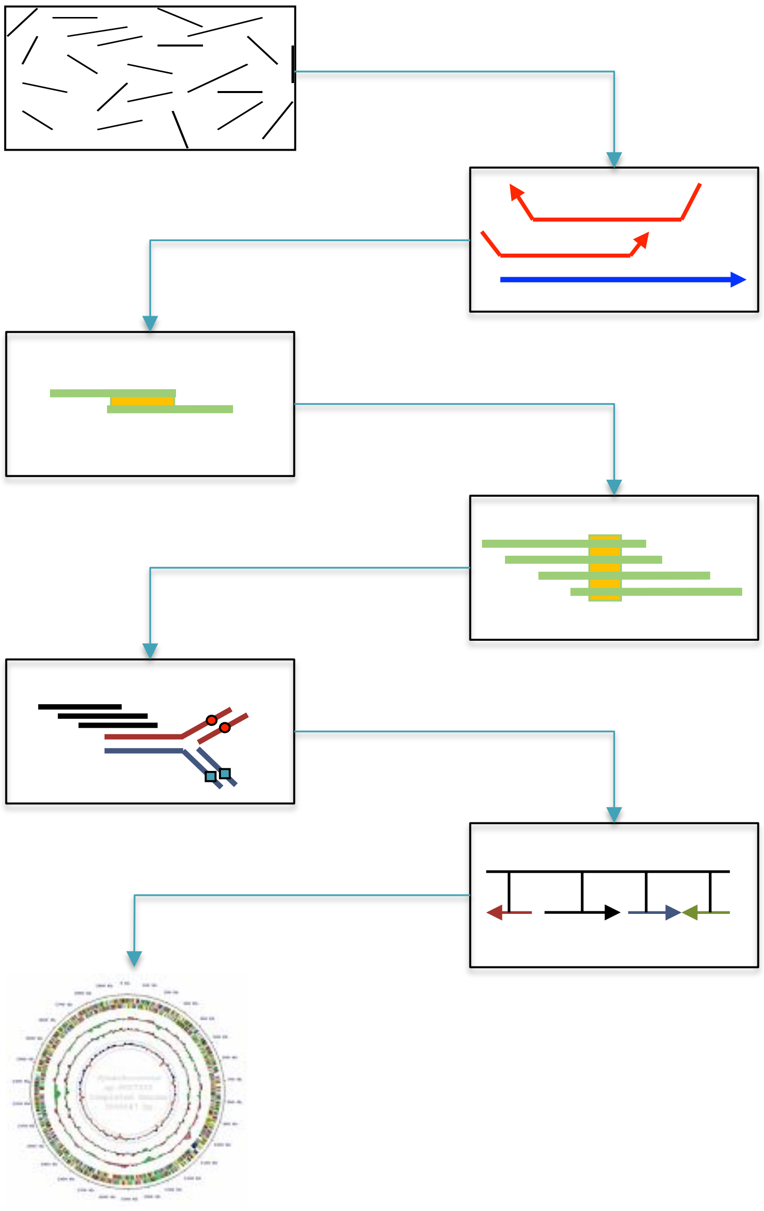
Chin *et al* (2013)
Nature Methods. 10:563–569

< 5x PacBio Coverage > 50x

Celera Assembler

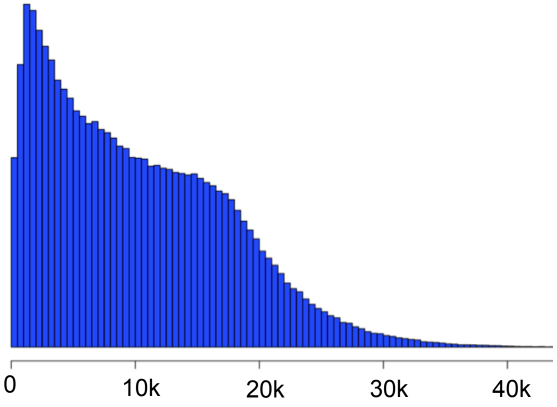
http://wgs-assembler.sf.net

1. Pre-overlap
   - Consistency checks
2. Trimming
   - Quality trimming & partial overlaps
3. Compute Overlaps
   - Find high quality overlaps
4. Error Correction
   - Evaluate difference in context of overlapping reads
5. Unitigging
   - Merge consistent reads
6. Scaffolding
   - Bundle mates, Order & Orient
7. Finalize Data
   - Build final consensus sequences

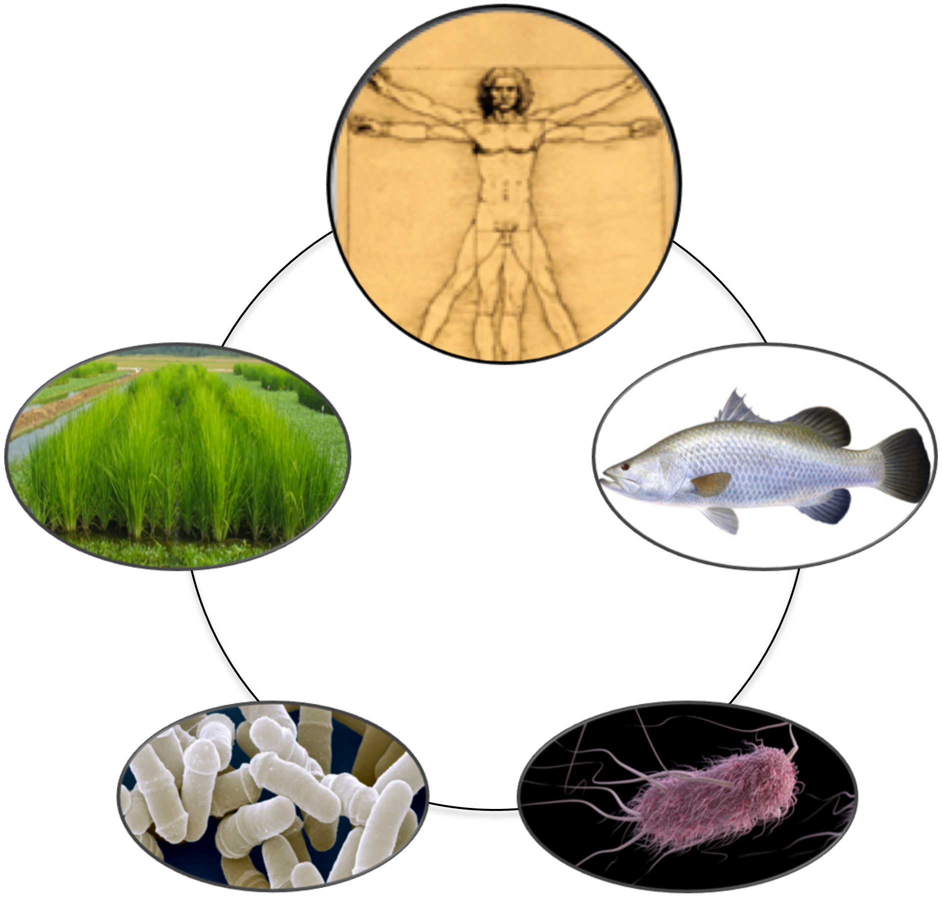


3rd Gen Long Read Sequencing

PacBio RS II

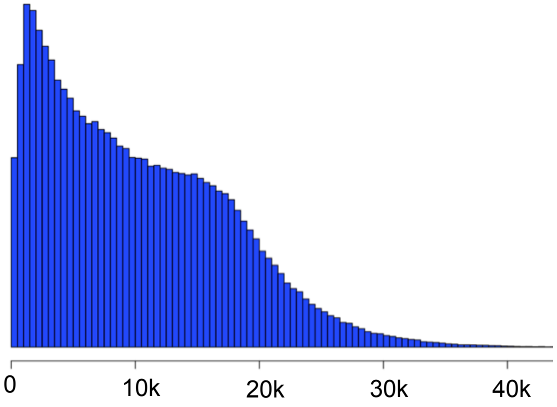


CSHL/PacBio

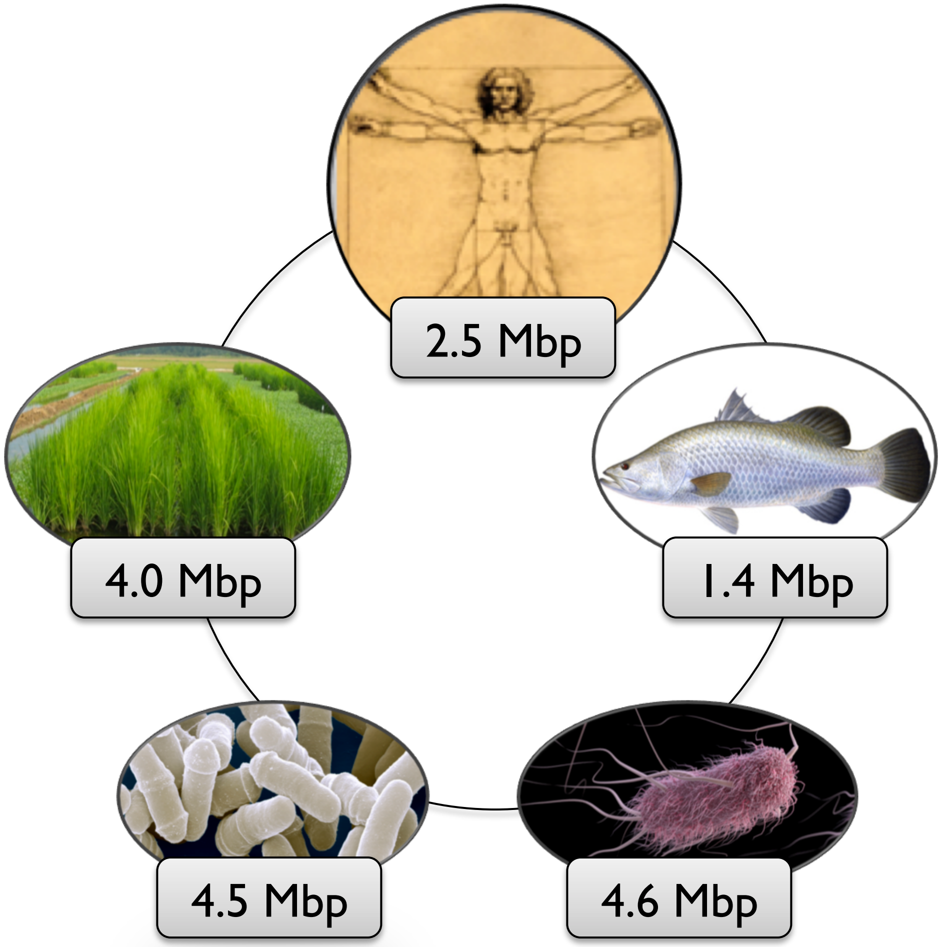


3rd Gen Long Read Sequencing

PacBio RS II



CSHL/PacBio



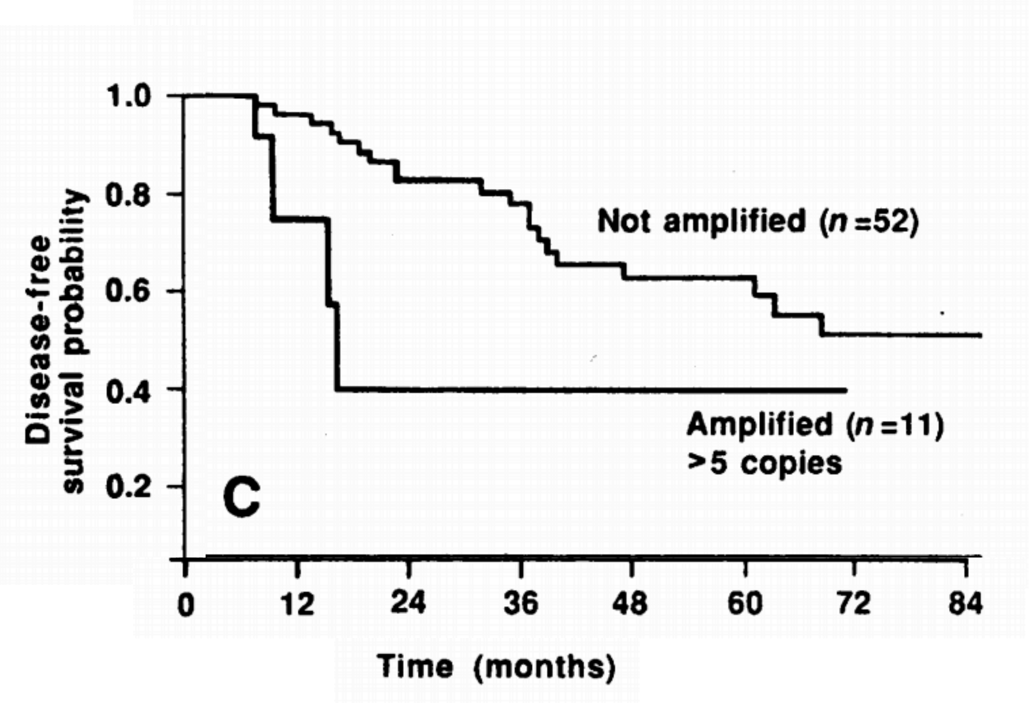
Her2 amplified breast cancer

Breast cancer

- About 12% of women will develop breast cancer during their lifetimes
- ~230,000 new cases every year (US)
- ~40,000 deaths every year (US)

Her2+ breast cancer

- 20% of breast cancers
- 2-3X recurrence risk
- 5X metastasis risk



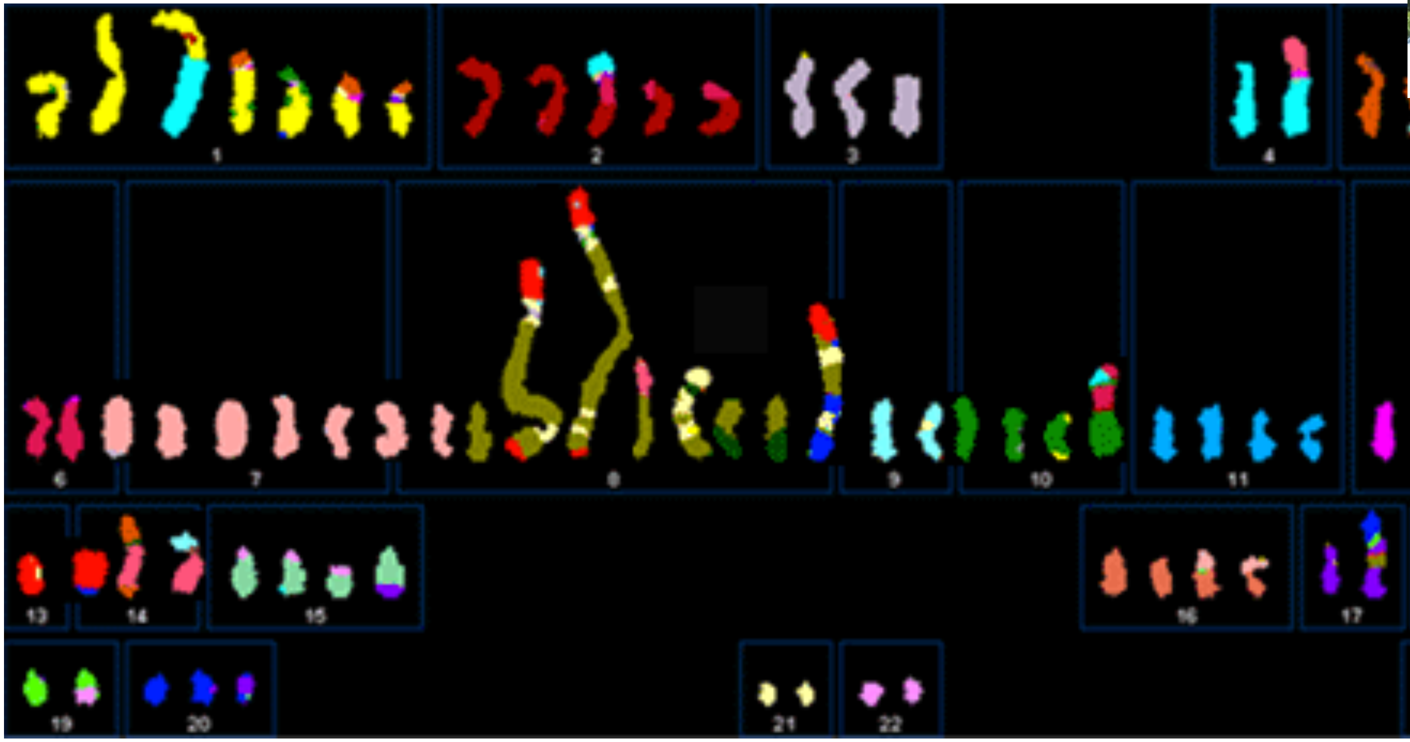
(Adapted from Slamon et al, 1987)

Statistics from American Cancer Society and Mayo Clinic.

Recurrence and metastasis from Gonzalez-Angulo, et al, 2009.

SK-BR-3

Most commonly used Her2-amplified breast cancer ce

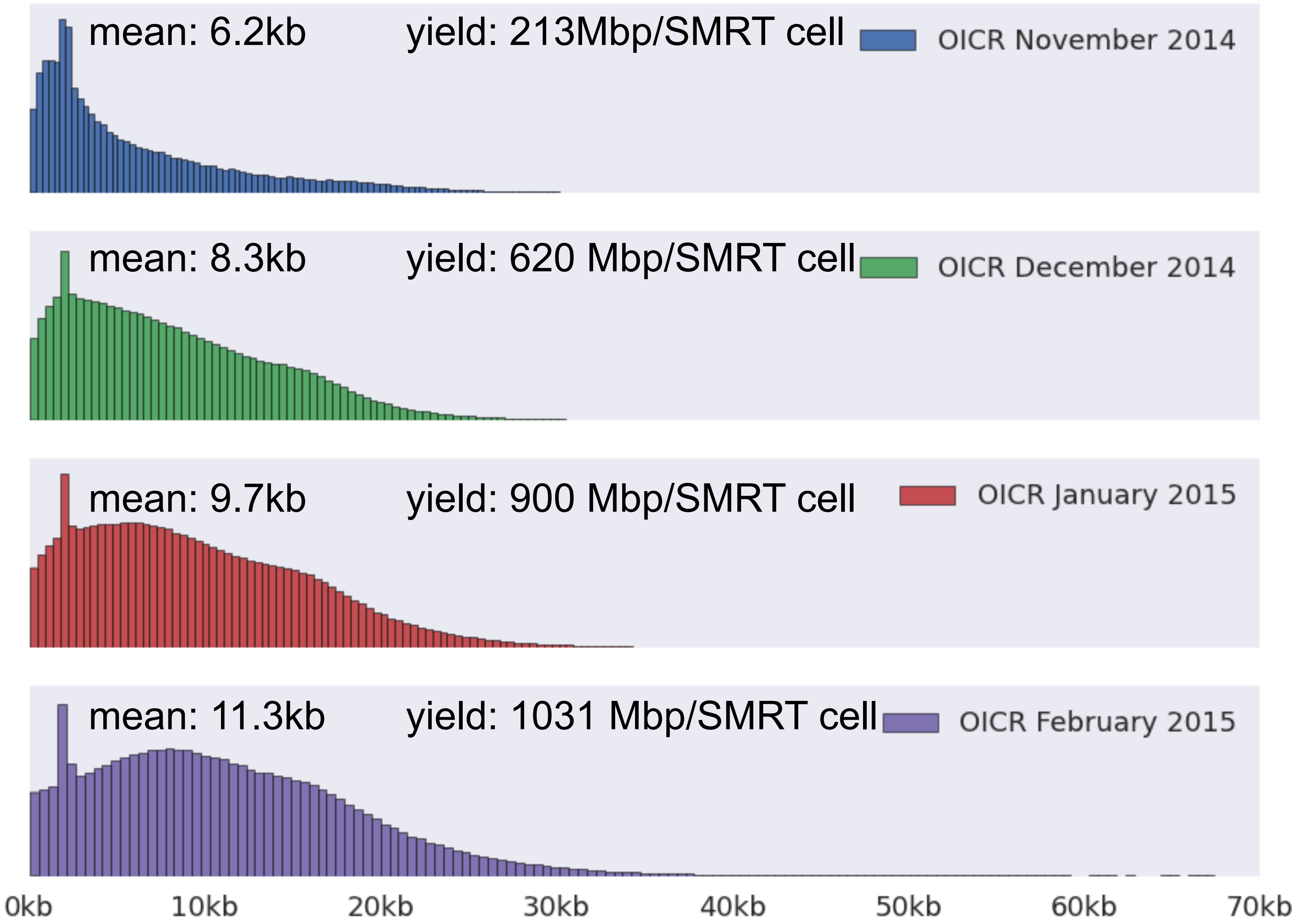


(Davidson et al, 2000)

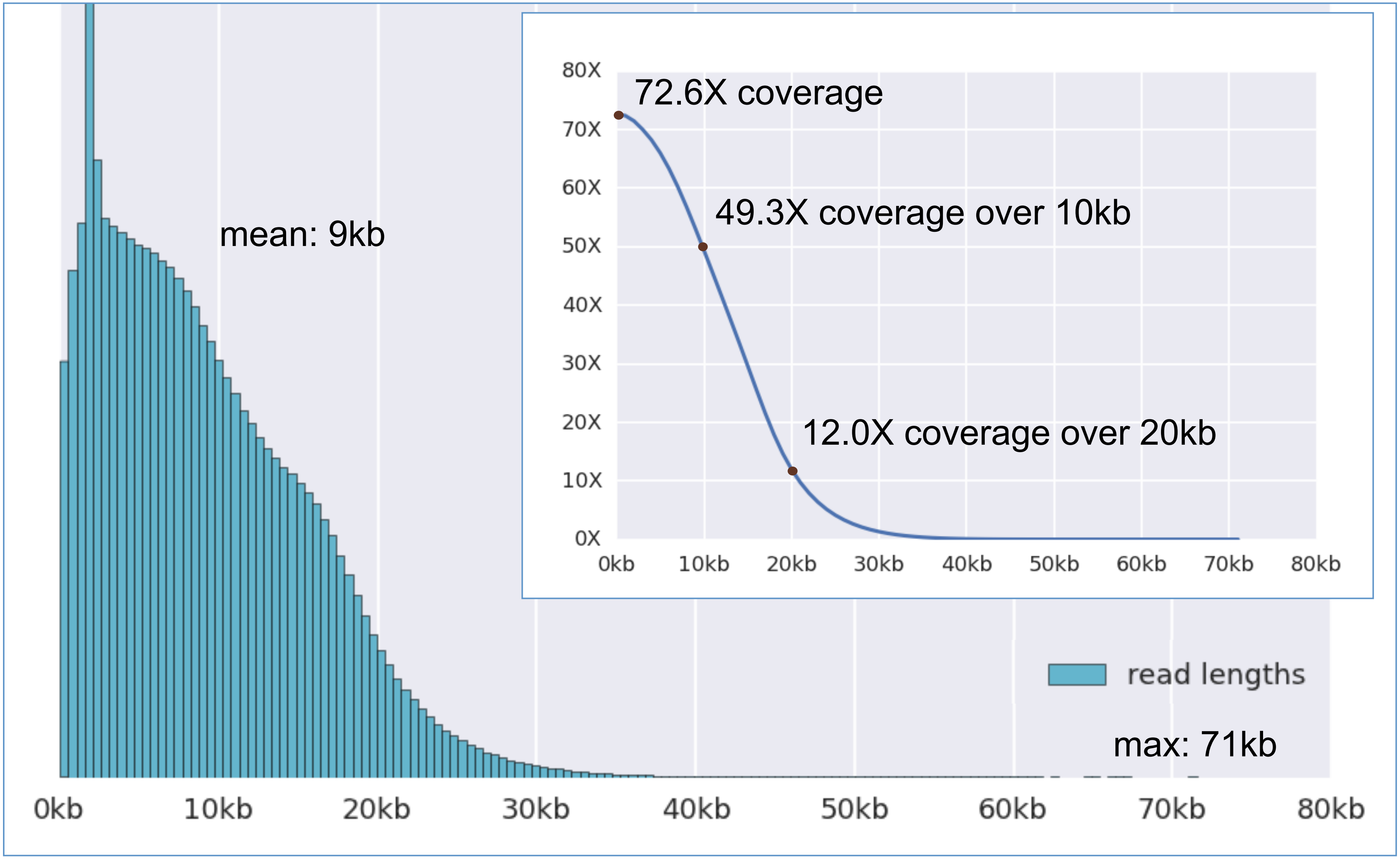
Can we resolve the complex structural variations, especially around Her2?

Ongoing collaboration between CSHL and OICR to *de novo* assemble the complete cell line genome with PacBio long reads

Improving SMRTcell Performance



PacBio read length distribution



mean: 9kb

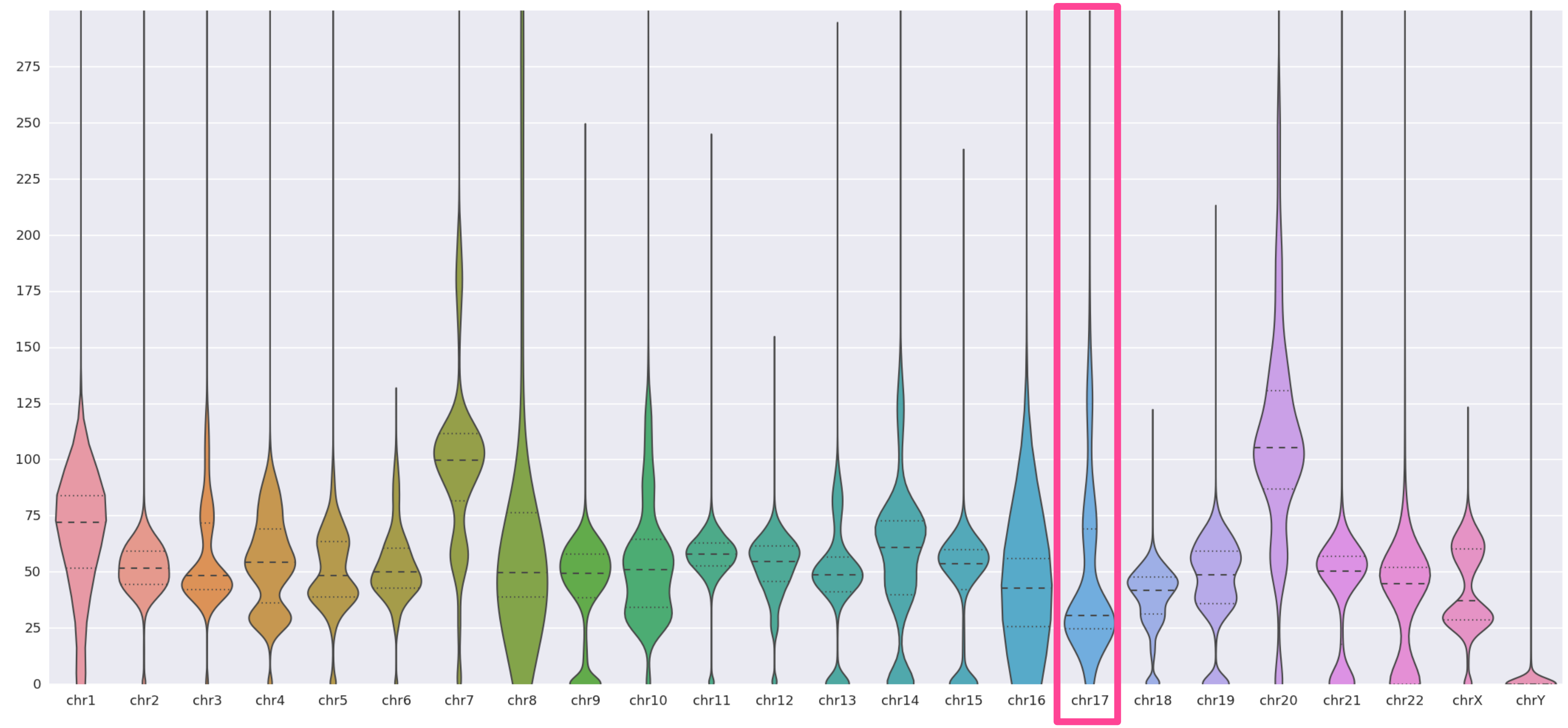
max: 71kb

72.6X coverage

49.3X coverage over 10kb

12.0X coverage over 20kb

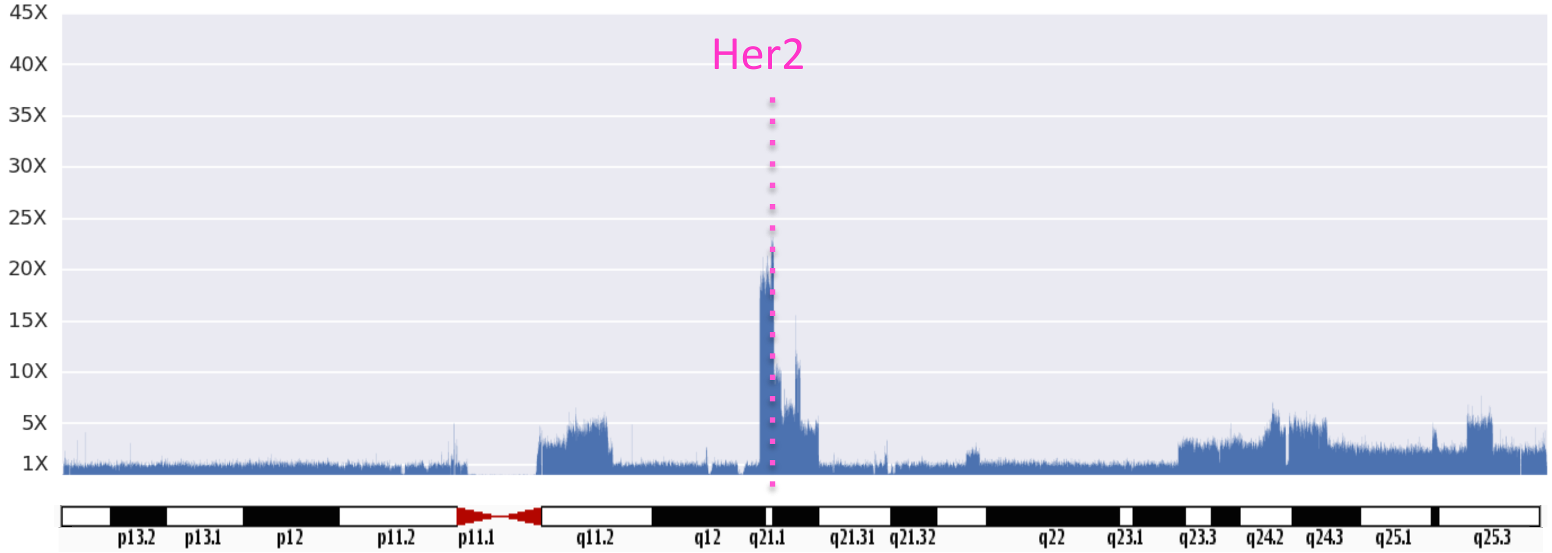
Genome-wide alignment coverage



Genome-wide coverage averages around 54X

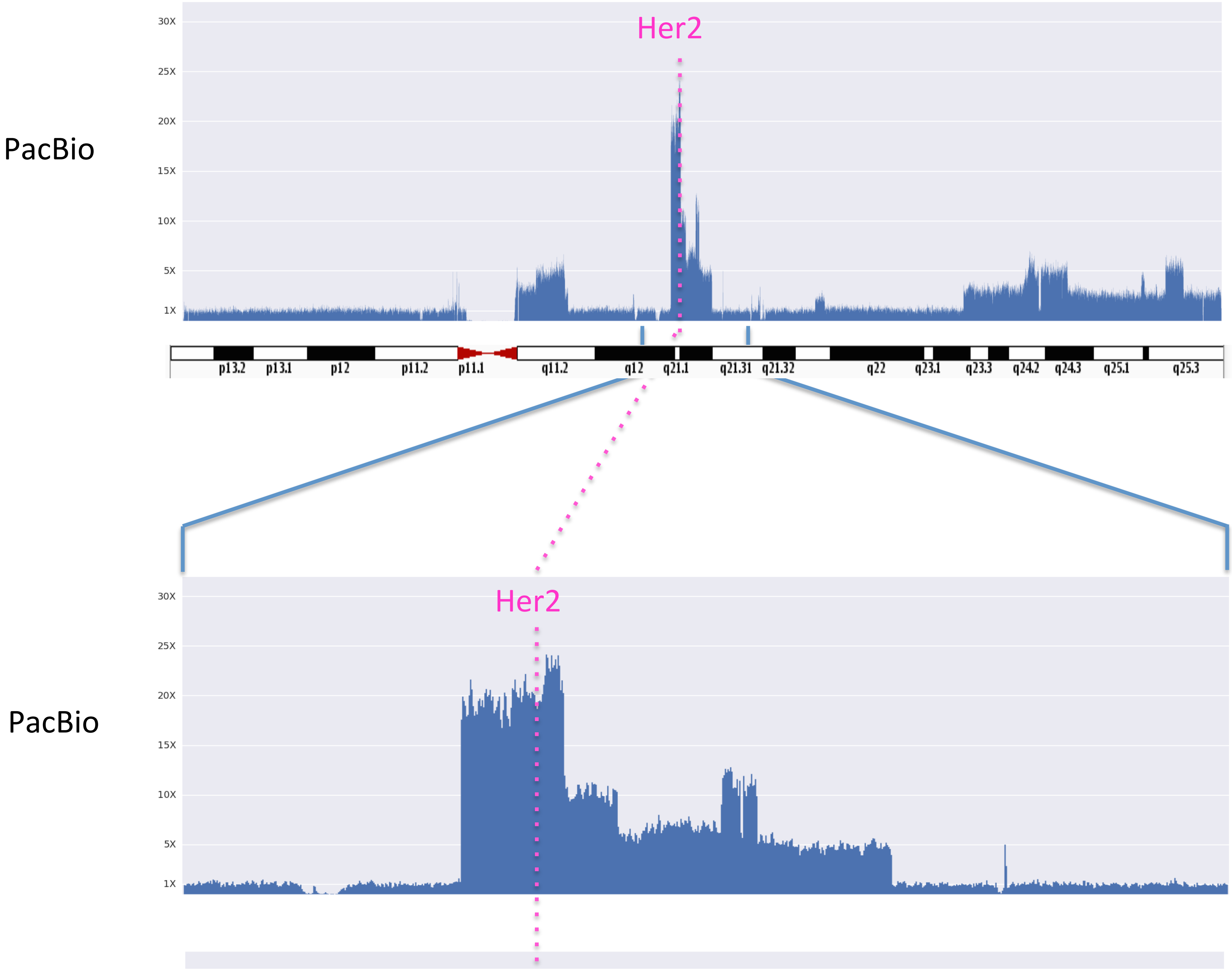
Coverage per chromosome varies greatly as expected from previous karyotyping results

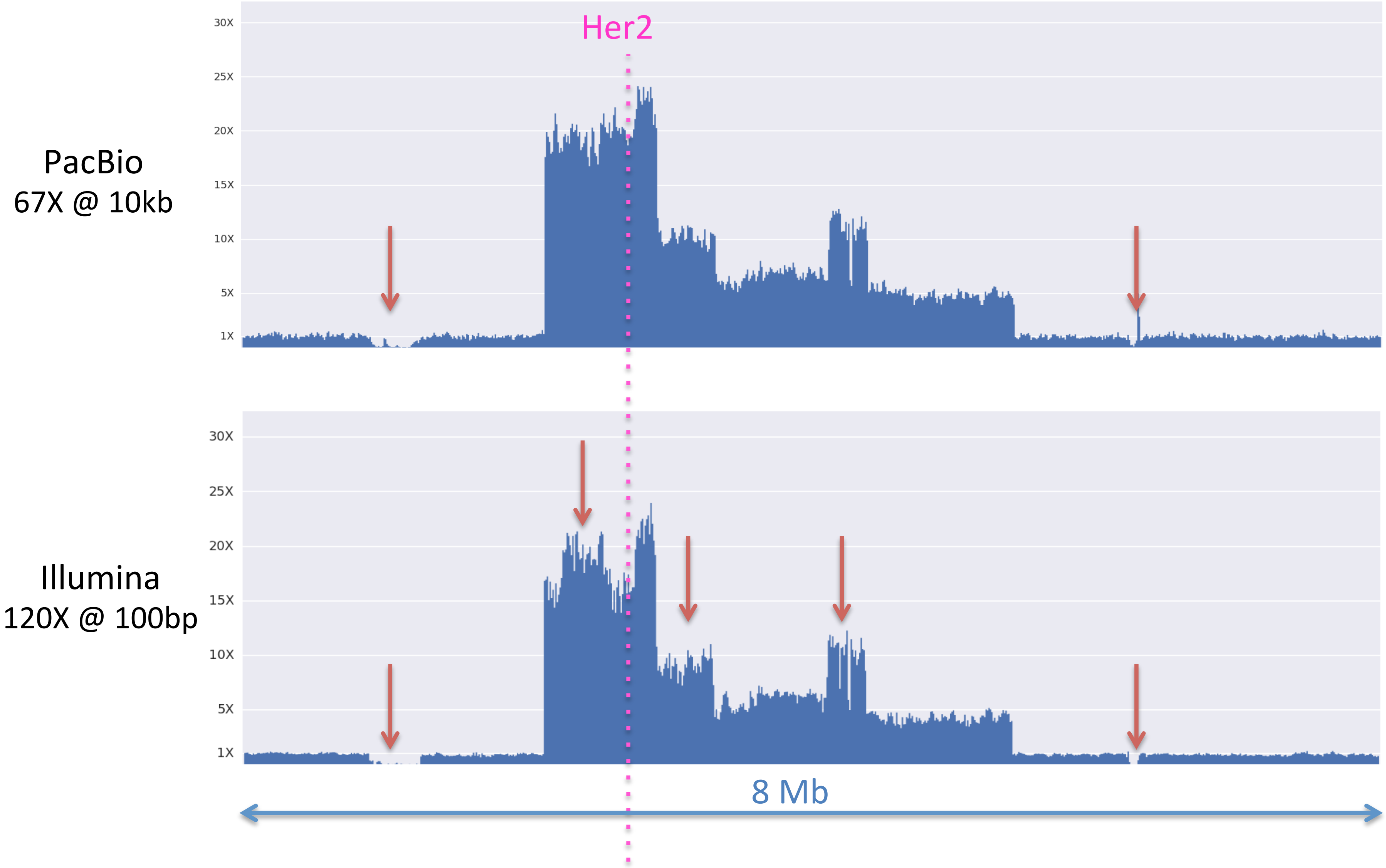
PacBio



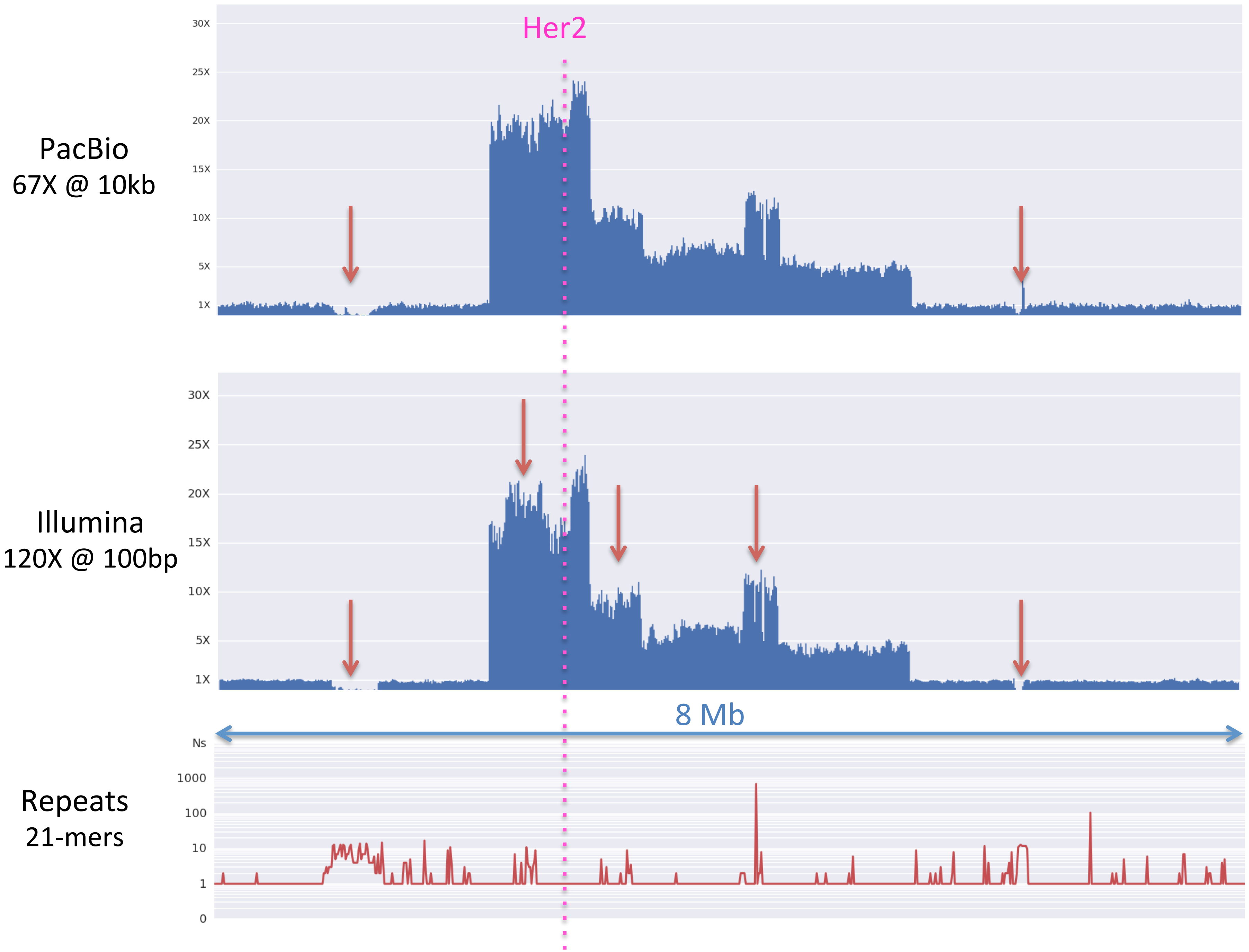
Chr 17: 83 Mb

8 Mb

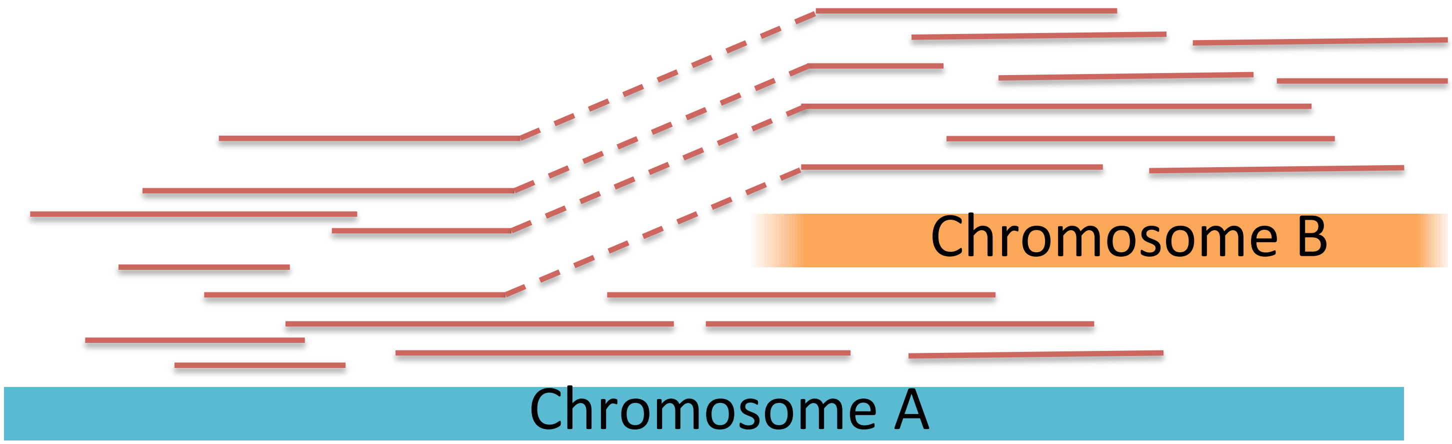

PacBio
30X
25X
20X
15X
10X
5X
1X
Her2
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.32
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.3
PacBio
30X
25X
20X
15X
10X
5X
1X
Her2



PacBio and Illumina coverage values are highly correlated
but Illumina shows greater variance because of poorly mapping reads

Her2
PacBio
67X @ 10kb
30X
25X
20X
15X
10X
5X
1X
Illumina
120X @ 100bp
30X
25X
20X
15X
10X
5X
1X
8 Mb
Repeats
21-mers
Ns
1000
100
10
1
0

Structural variant discovery with long reads



1. Alignment-based split read analysis: Efficient capture of most events

BWA-MEM + Lumpy

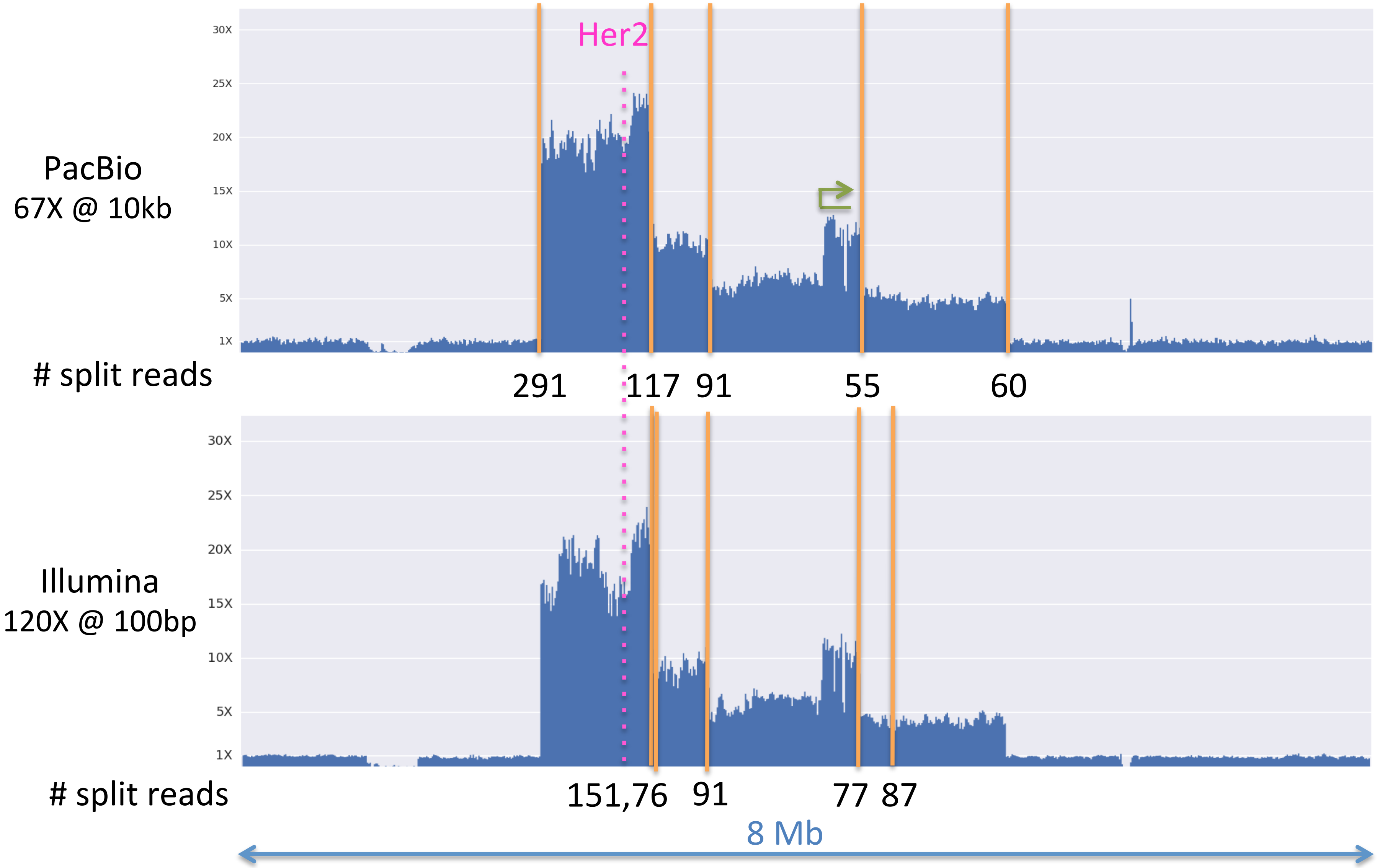
2. Local assembly of regions of interest: In-depth analysis with *base-pair precision*

Localized HGAP + Celera Assembler + MUMmer

3. Whole genome assembly: In-depth analysis including *novel sequences*

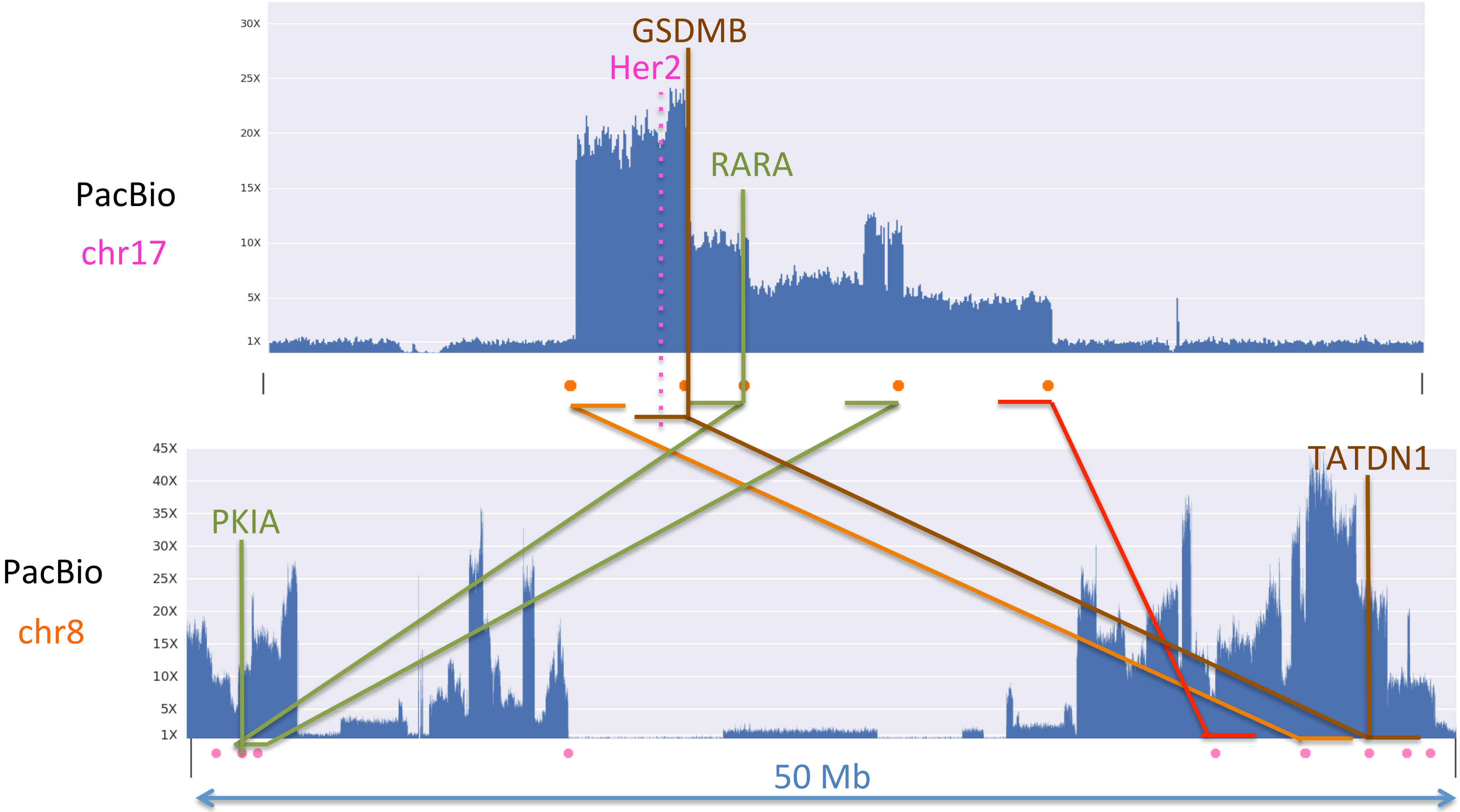
DNAnexus-enabled version of Falcon

Total Assembly: 2.64Gbp **Contig N50: 2.56 Mbp** **Max Contig: 23.5Mbp**

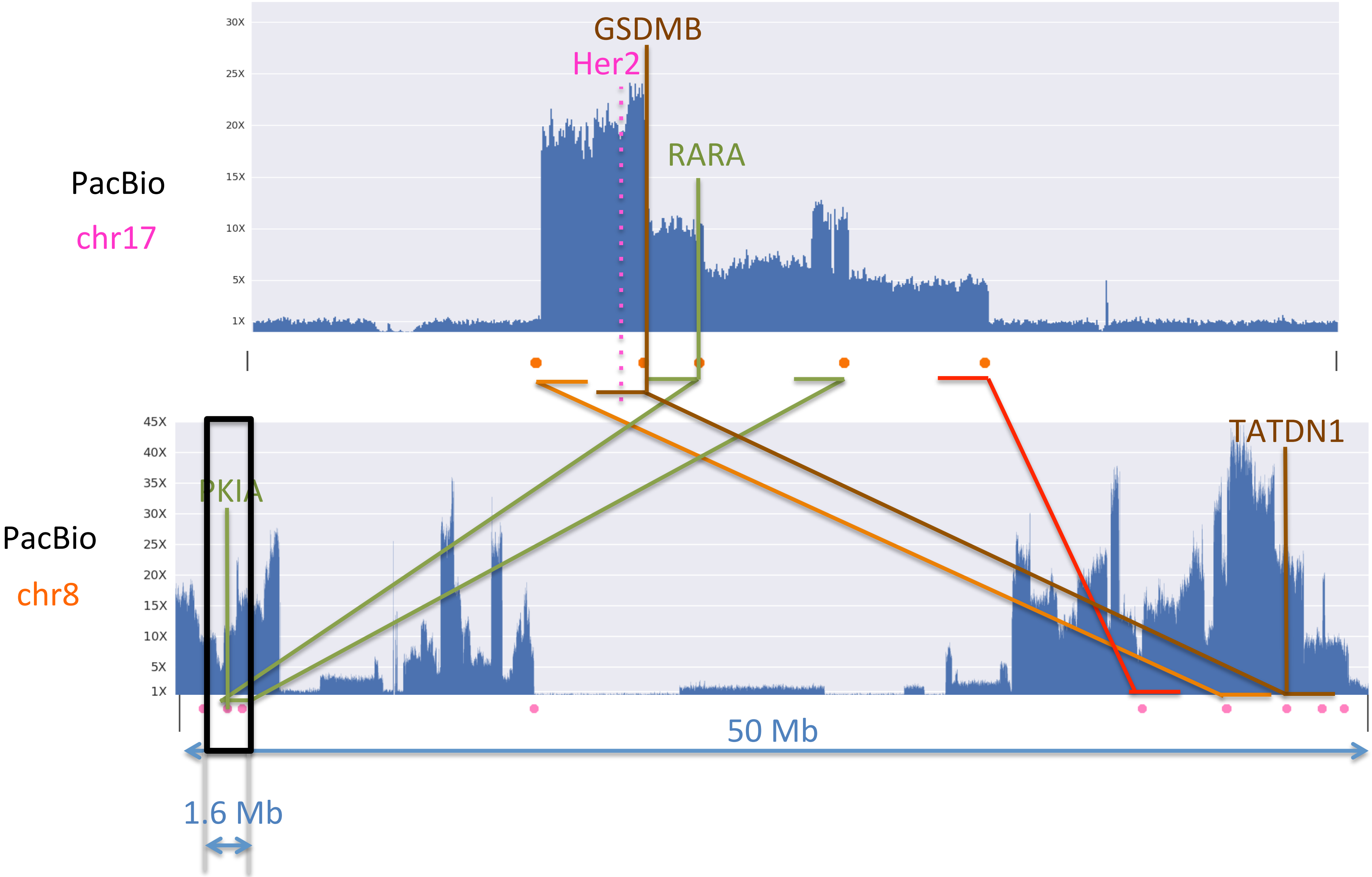


Green arrow indicates an inverted duplication.

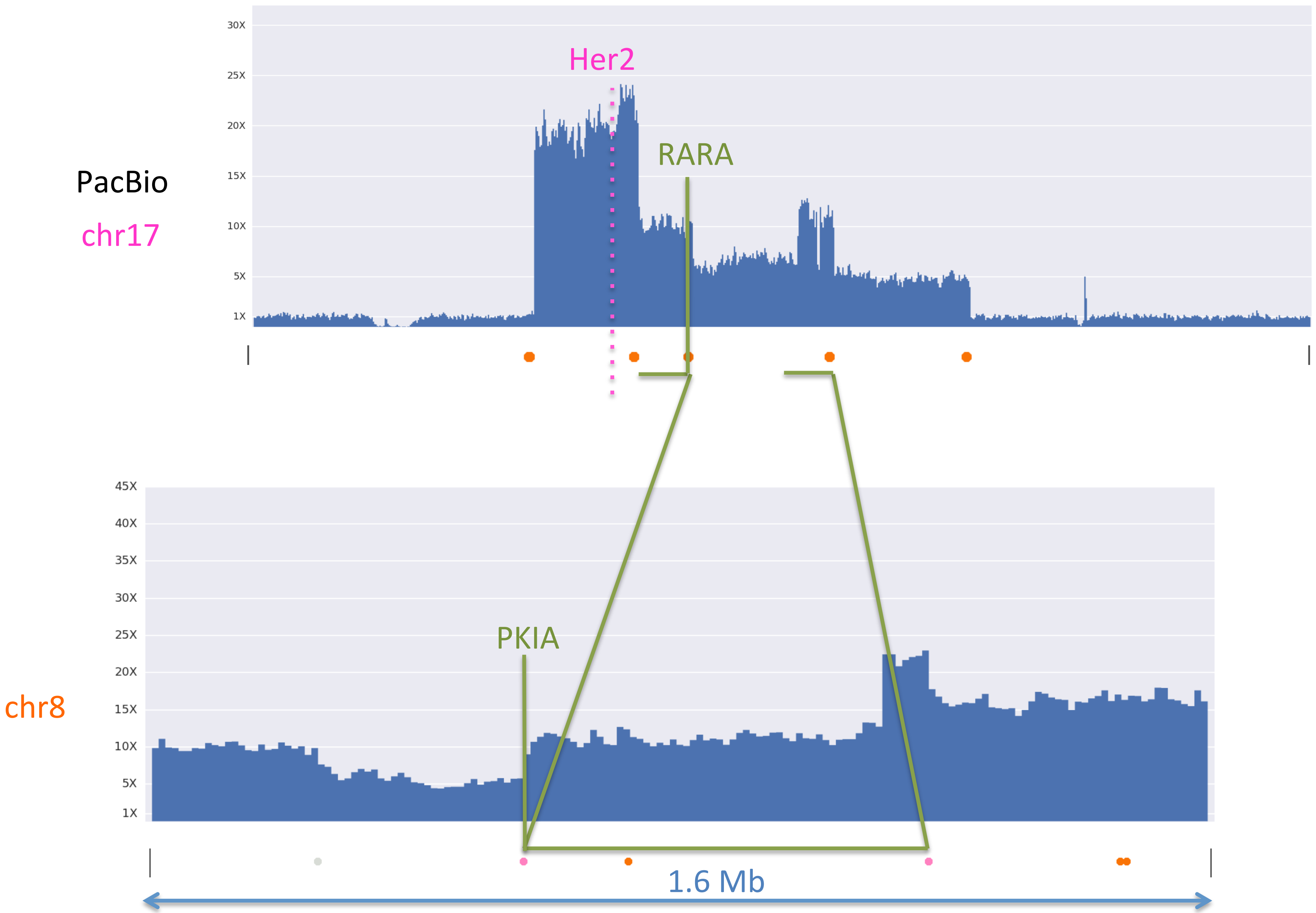
False positive and missing Illumina calls due to mis-mapped reads (especially low complexity).



Confirmed both known gene fusions in this region

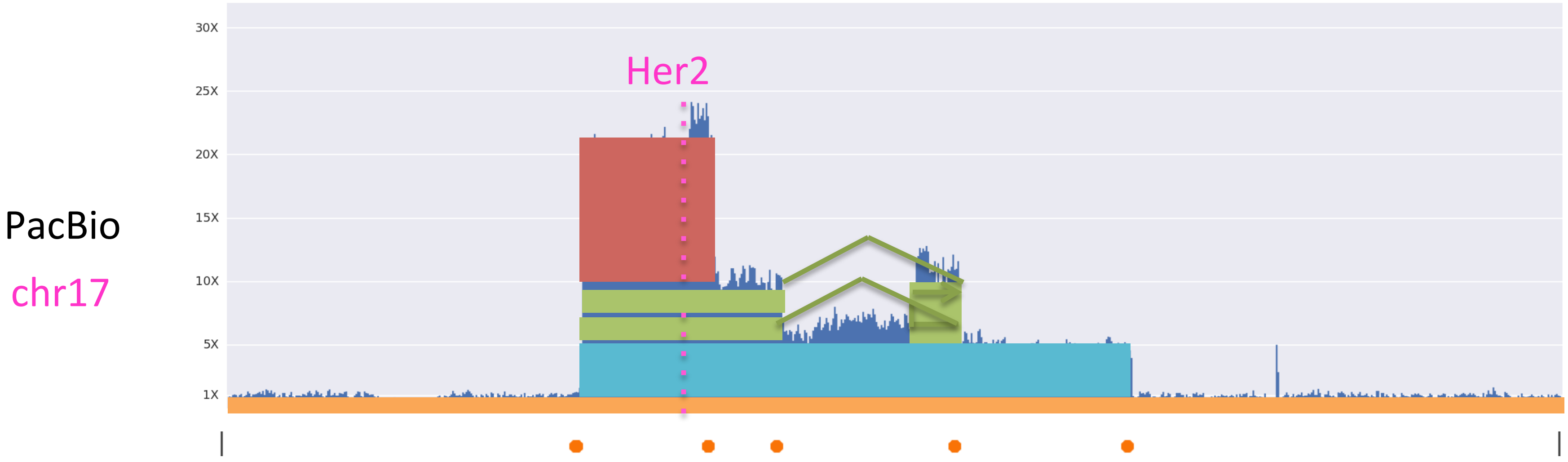


Confirmed both known gene fusions in this region



Joint coverage and breakpoint analysis to discover underlying events

Cancer lesion Reconstruction



By comparing the proportion of reads that are spanning or split at breakpoints we can begin to infer the history of the genetic lesions.

1. Healthy diploid genome

2. Original translocation into chromosome 8

3. Duplication, inversion, and inverted duplication within chromosome 8

4. Final duplication from within chromosome 8

SKBR3 Oncogene Analysis

Known missense mutation in p53: **R175H**

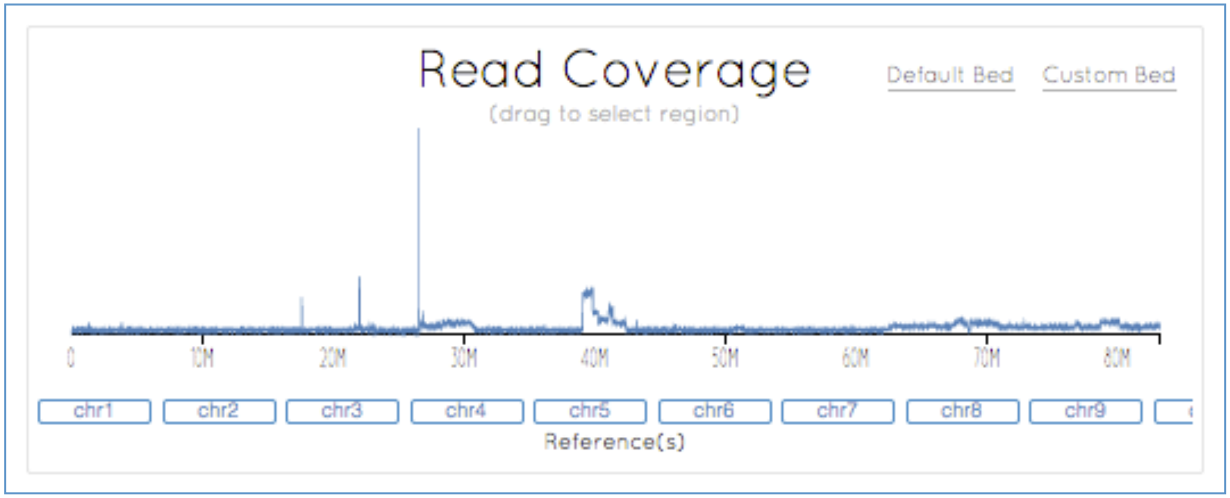
```
                                         Arg
Reference     ATCTGAGCAGCGCTCATGGTGGGGGCAGCGCCTCACAACCTCCGTCATGTGCTGTGACTGCTT
Illumina      ATCTGAGCAGCGCTCATGGTGGGGGCAGTGCCTCACAACCTCCGTCATGTGCTGTGACTGCTT
PacBio        ATCTGAGCAGCGCTCATGGTGGGGGCAGTGCCTCACAACCTCCGTCATGTGCTGTGACTGCTT
                                         His
```

Oncogene amplifications	
ErbB2 (Her2/neu)	≈20X
MYC	≈27X
MET	≈8X

Genetic Lesion History Analysis Underway

Known Gene fusions		Confirmed by PacBio reads?
TATDN1	GSDMB	**Yes**
RARA	PKIA	**Yes**
ANKHD1	PCDH1	**Yes**
CCDC85C	SETD3	**Yes**
SUMF1	LRRFIP2	**Yes**
WDR67 (TBC1D31)	ZNF704	**Yes**
DHX35	ITCH	**Yes**
NFS1	PREX1	**Yes *read-through transcription**
CYTH1	EIF3H	**Yes *nested inside 2 translocations**

Her2+ Breast Cancer Reference Genome



Available *today* under the Toronto Agreement:

- Fastq & BAM files of aligned reads
- Interactive Coverage Analysis with BAM.IOBIO
- Whole genome assembly

Available soon

- Whole genome methylation analysis
- Full length cDNA transciptome analysis
- Comparison to single cell analysis of >100 individual cells

http://schatzlab.cshl.edu/data/skbr3/

What should we expect from an assembly?

The resurgence of reference quality genomes

Summary & Recommendations

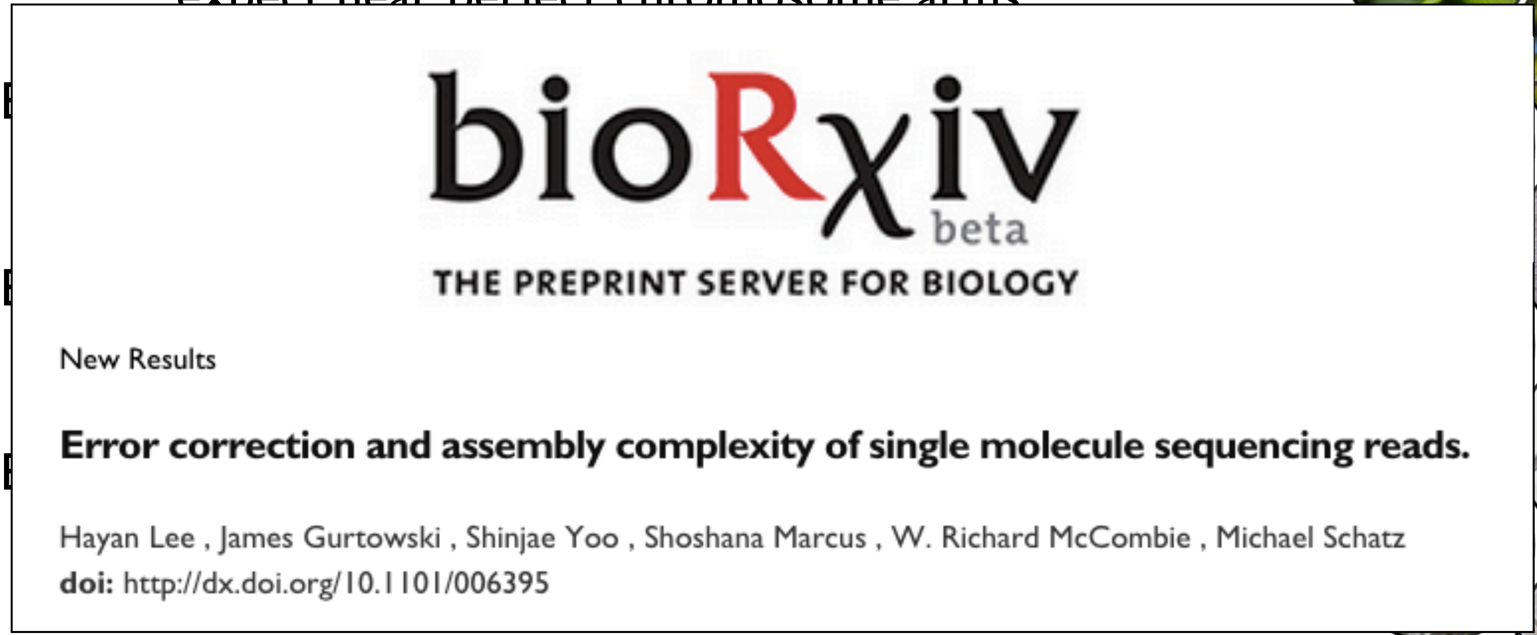
< 100 Mbp: HGAP/PacBio2CA @ 100x PB C3-P5

expect near perfect chromosome arms

< 1G

> 1G

> 5G



bioRχiv beta

THE PREPRINT SERVER FOR BIOLOGY

New Results

Error correction and assembly complexity of single molecule sequencing reads.

Hayan Lee , James Gurtowski , Shinjae Yoo , Shoshana Marcus , W. Richard McCombie , Michael Schatz

doi: http://dx.doi.org/10.1101/006395

Caveats

Model only as good as the available references (esp. haploid sequences)

Technologies are quickly improving, exciting new scaffolding technologies

Acknowledgements

Schatz Lab

Rahul Amin
Eric Biggers
Han Fang
Tyler Gavin
James Gurtowski
Ke Jiang
Hayan Lee
Zak Lemmon
Shoshana Marcus
Giuseppe Narzisi
Maria Nattestad
Aspyn Palatnick
Srividya Ramakrishnan
Fritz Sedlazeck
Rachel Sherman
Greg Vurture
Alejandro Wences

CSHL

Hannon Lab
Gingeras Lab
Jackson Lab
Hicks Lab
Iossifov Lab
Levy Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Tuveson Lab
Ware Lab
Wigler Lab

OICR

Karen Ng
Timothy Beck
Yogi Sundaravadanam
John McPherson

NBACC

Adam Phillippy
Serge Koren

Oxford NANOPORE Technologies™

PACIFIC BIOSCIENCES®

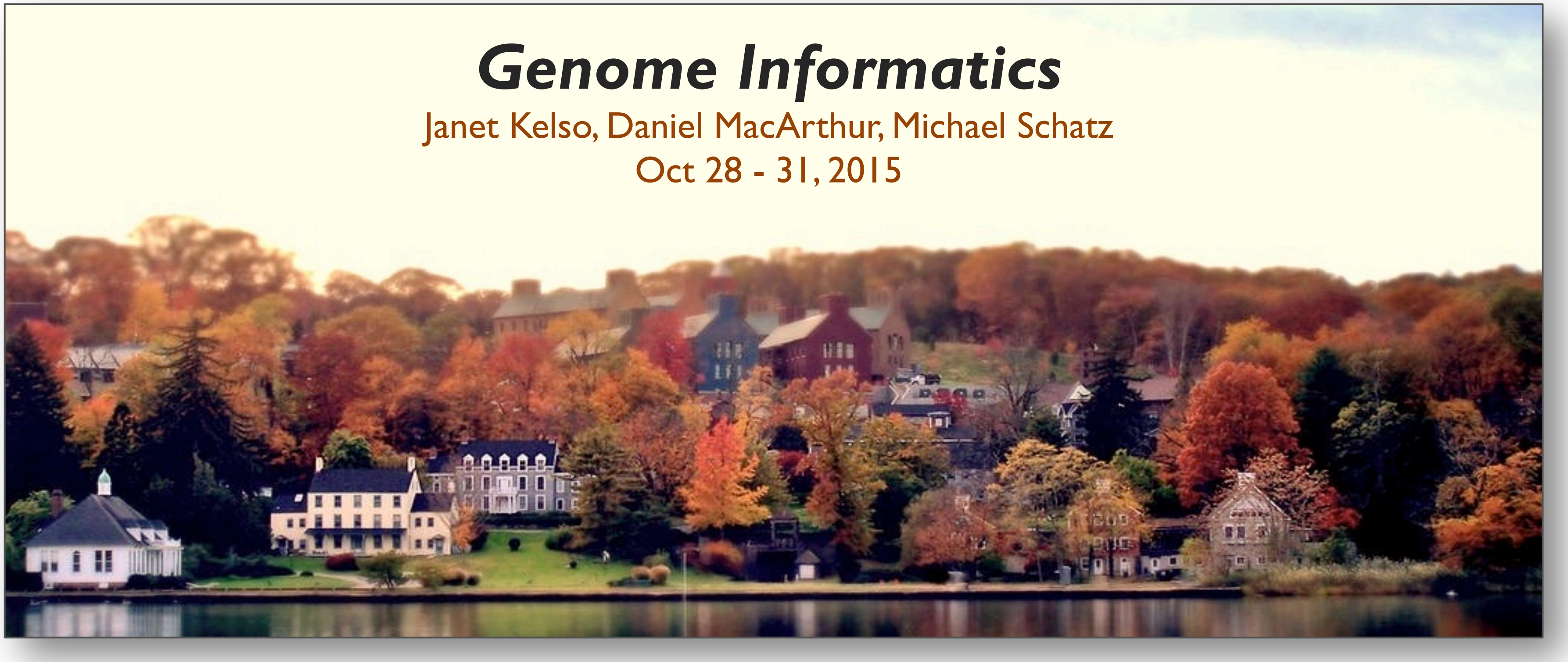
NSF

National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

SFARI
SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

ALFRED P. SLOAN FOUNDATION



Genome Informatics

Janet Kelso, Daniel MacArthur, Michael Schatz

Oct 28 - 31, 2015

Thank you

http://schatzlab.cshl.edu

@mike_schatz