Genome Sequencing & Assembly

Michael Schatz

Nov. 18, 2015

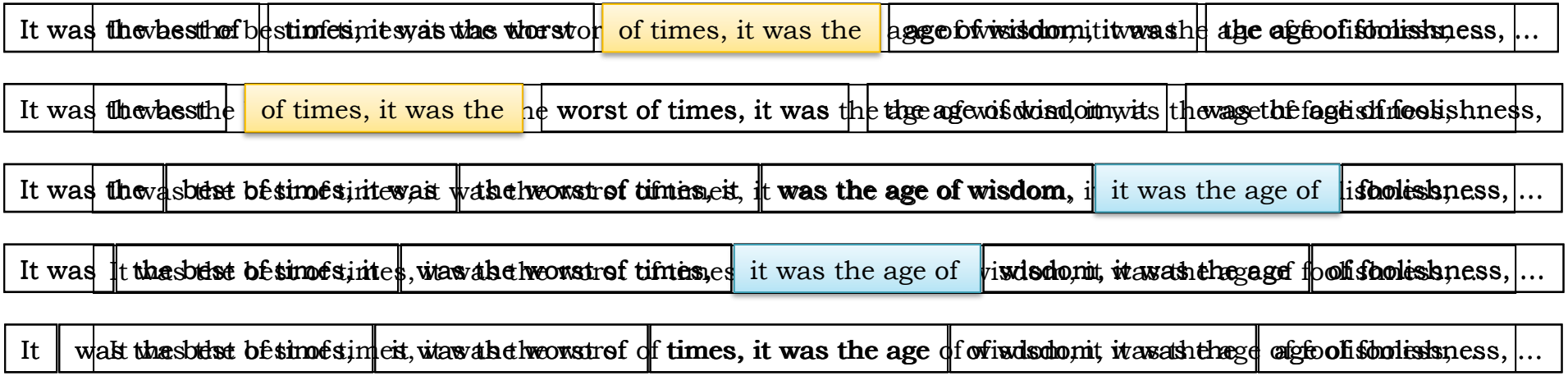
CSHL Adv. Sequencing Course

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 long read projects
4. Summary & Recommendations

Shredded Book Reconstruction

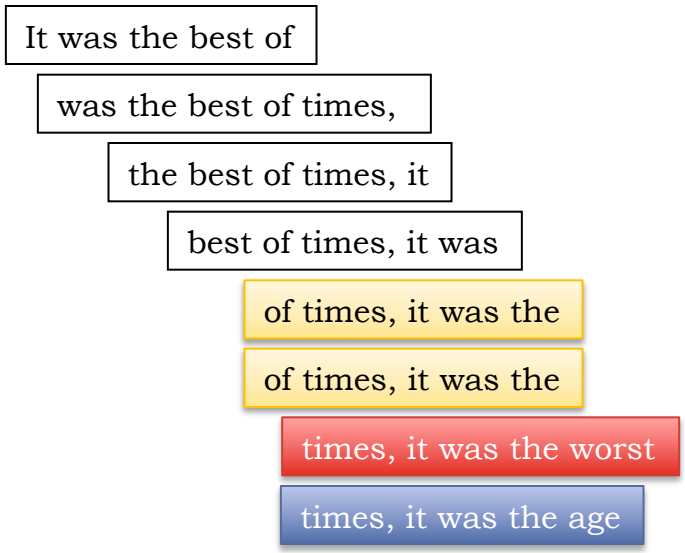
- Dickens accidentally shreds the first printing of A Tale of Two Cities
  - Text printed on 5 long spools



- 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



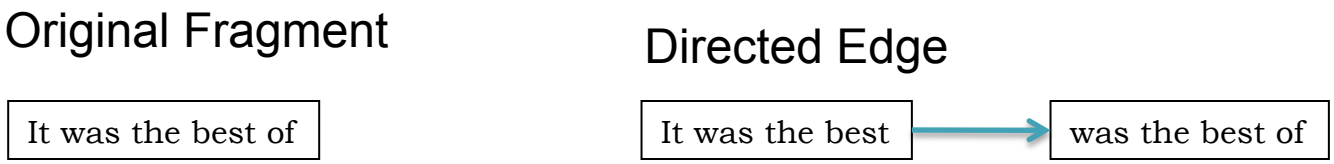
The repeated sequence make the correct reconstruction ambiguous

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

Model the assembly problem as a graph problem

de Bruijn Graph Construction

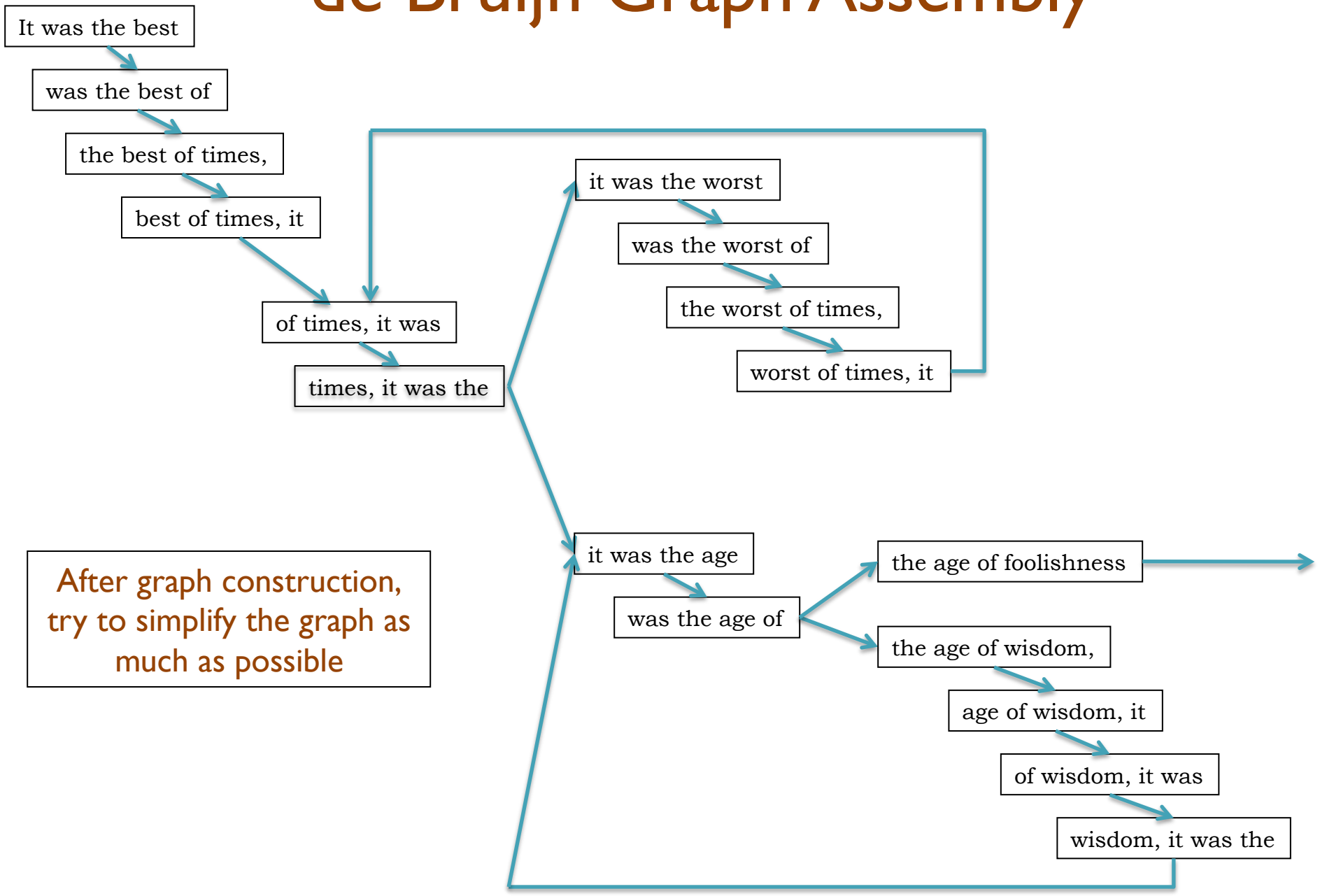
- $D_k = (V,E)$
  - V = All length-k subfragments (k < l)
  - E = Directed edges between consecutive subfragments
    - Nodes overlap by k-1 words



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

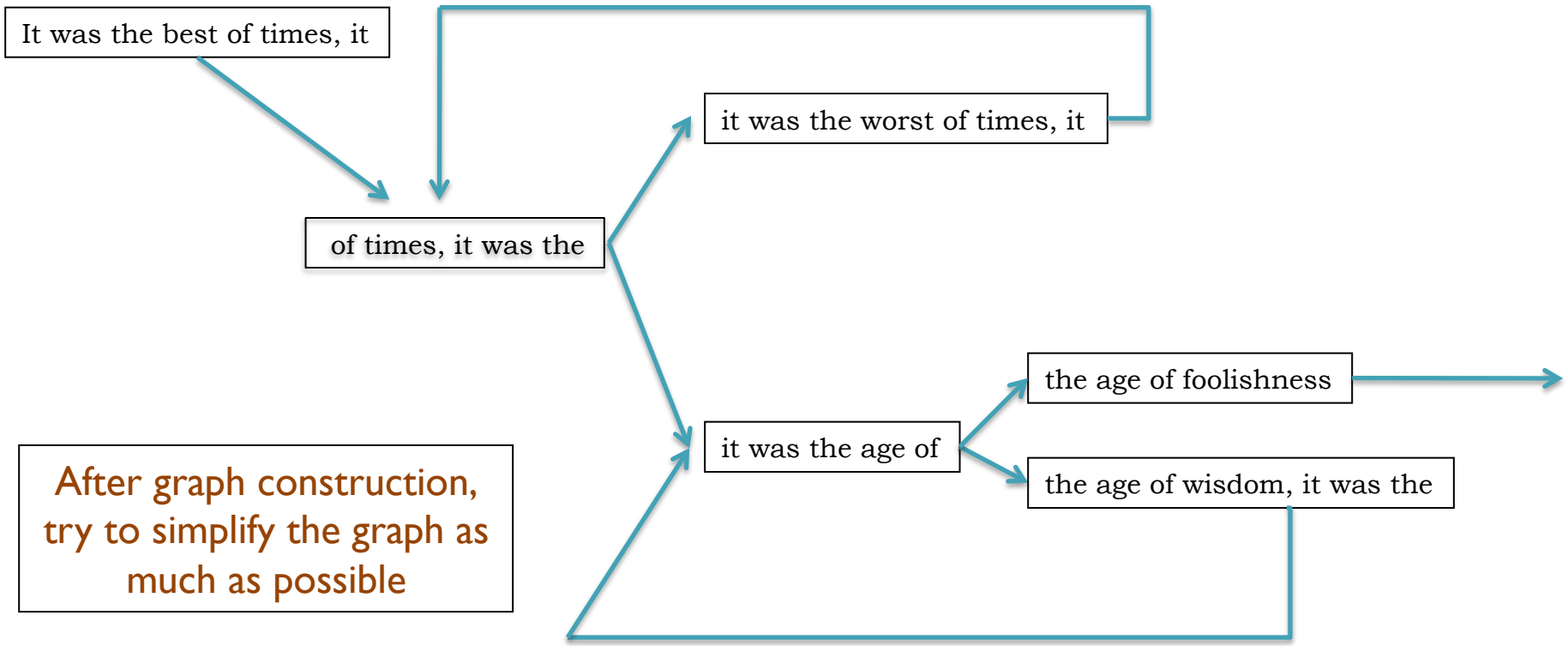
de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001

de Bruijn Graph Assembly



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

de Bruijn Graph Assembly



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

The full tale

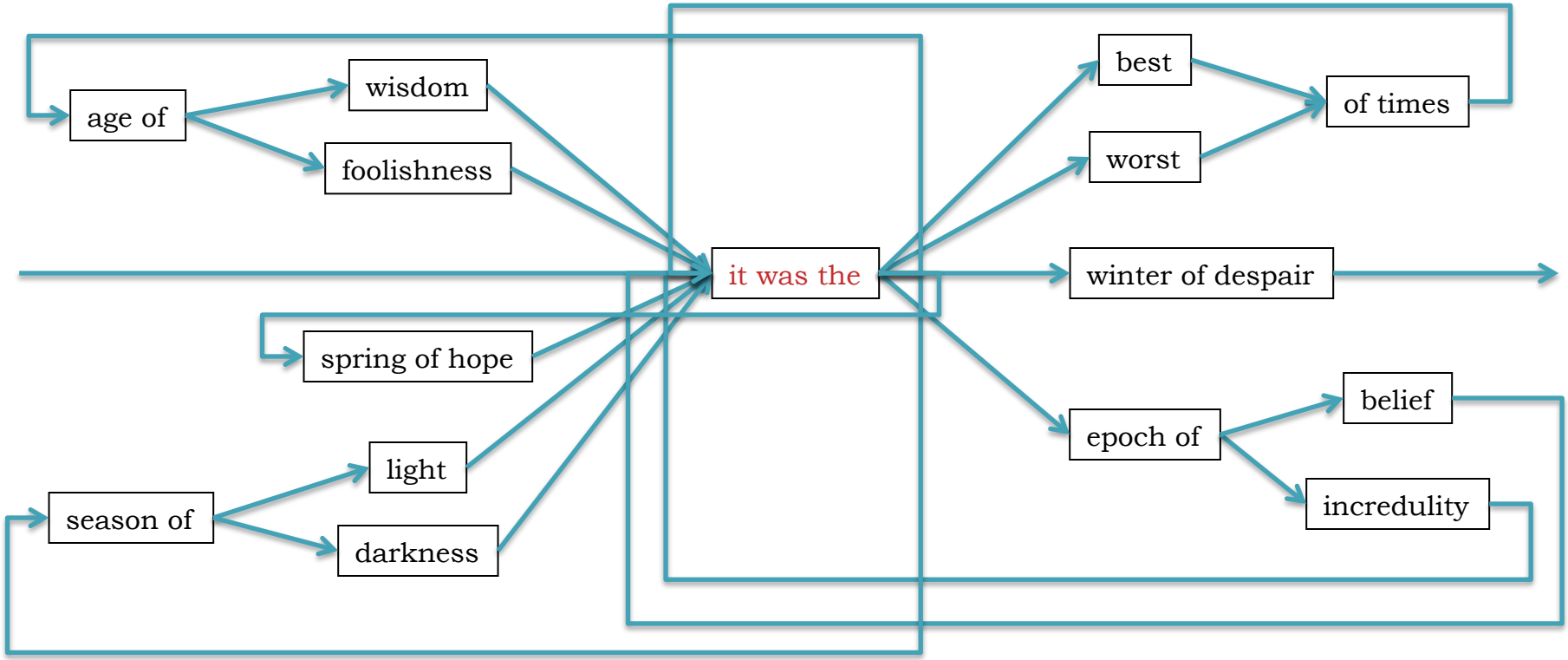
… it was the best of times it was the worst of times …

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

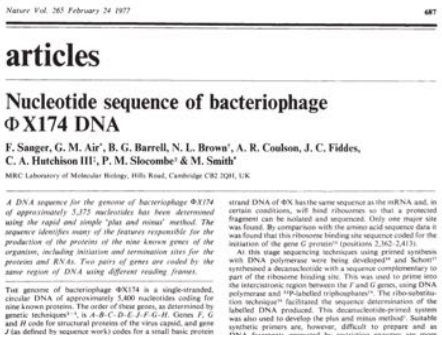
… it was the epoch of belief it was the epoch of incredulity …

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

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



Milestones in Genome Assembly



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

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

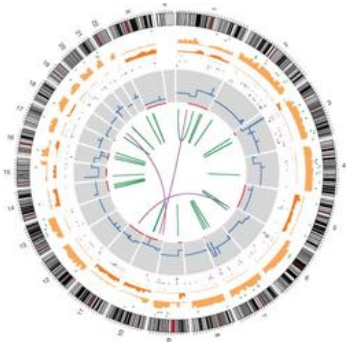
Like Dickens, we must computationally reconstruct a genome from short fragments

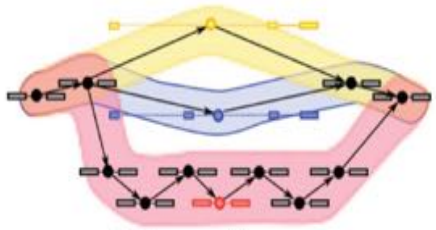
Assembly Applications

- Novel genomes

- Metagenomes

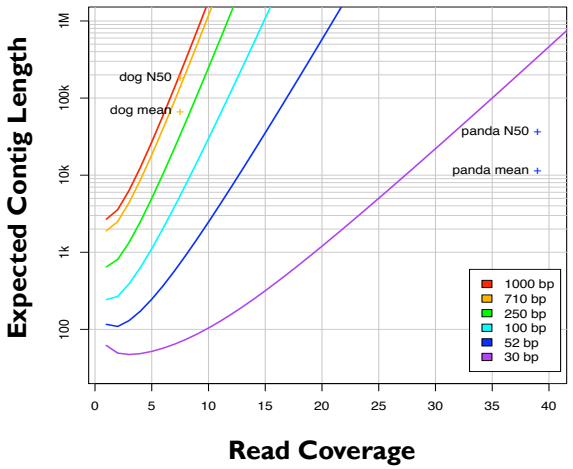
- Sequencing assays
  - Structural variations
  - Transcript assembly
  - …





Ingredients for a good assembly

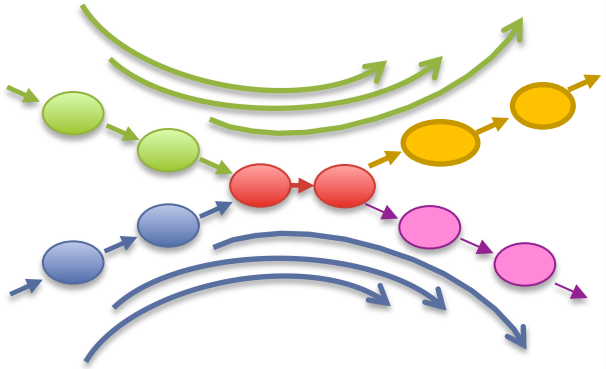
Coverage



High coverage is required

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

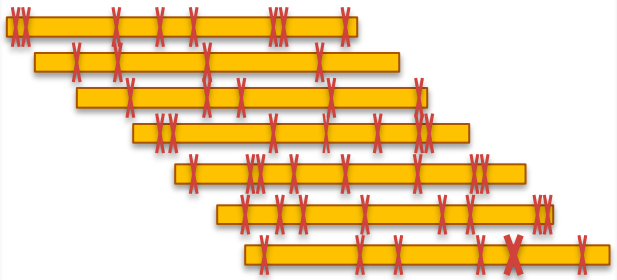
Read Length



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

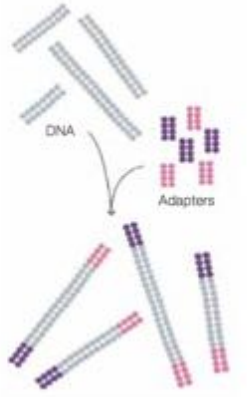


Errors obscure overlaps

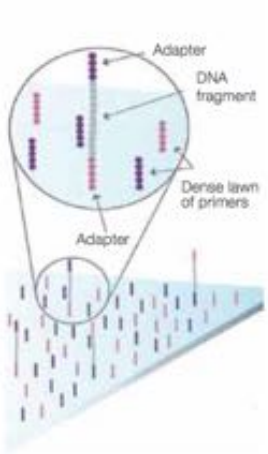
- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

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

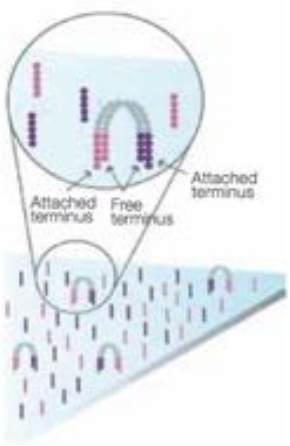
Illumina Sequencing by Synthesis



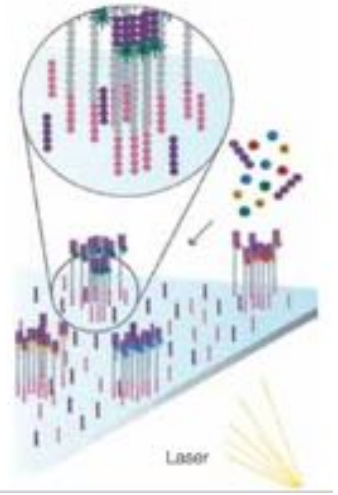
1. Prepare



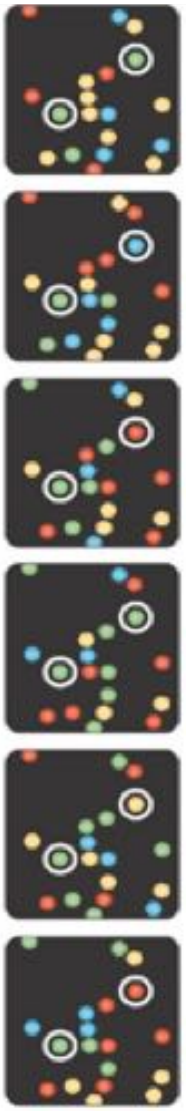
2. Attach



3. Amplify



4. Image

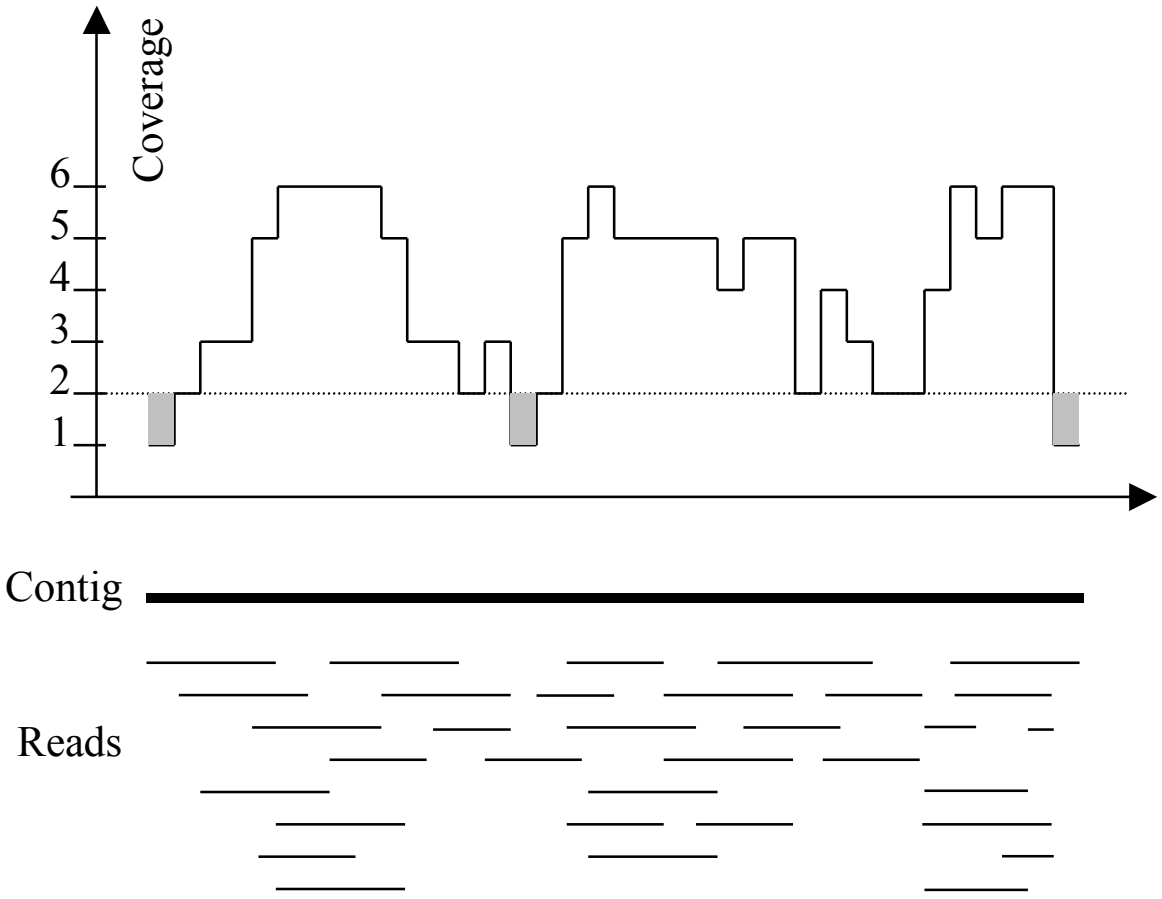


5. Basecall

Metzker (2010) Nature Reviews Genetics 11:31-46

http://www.youtube.com/watch?v=l99aKKHcxC4

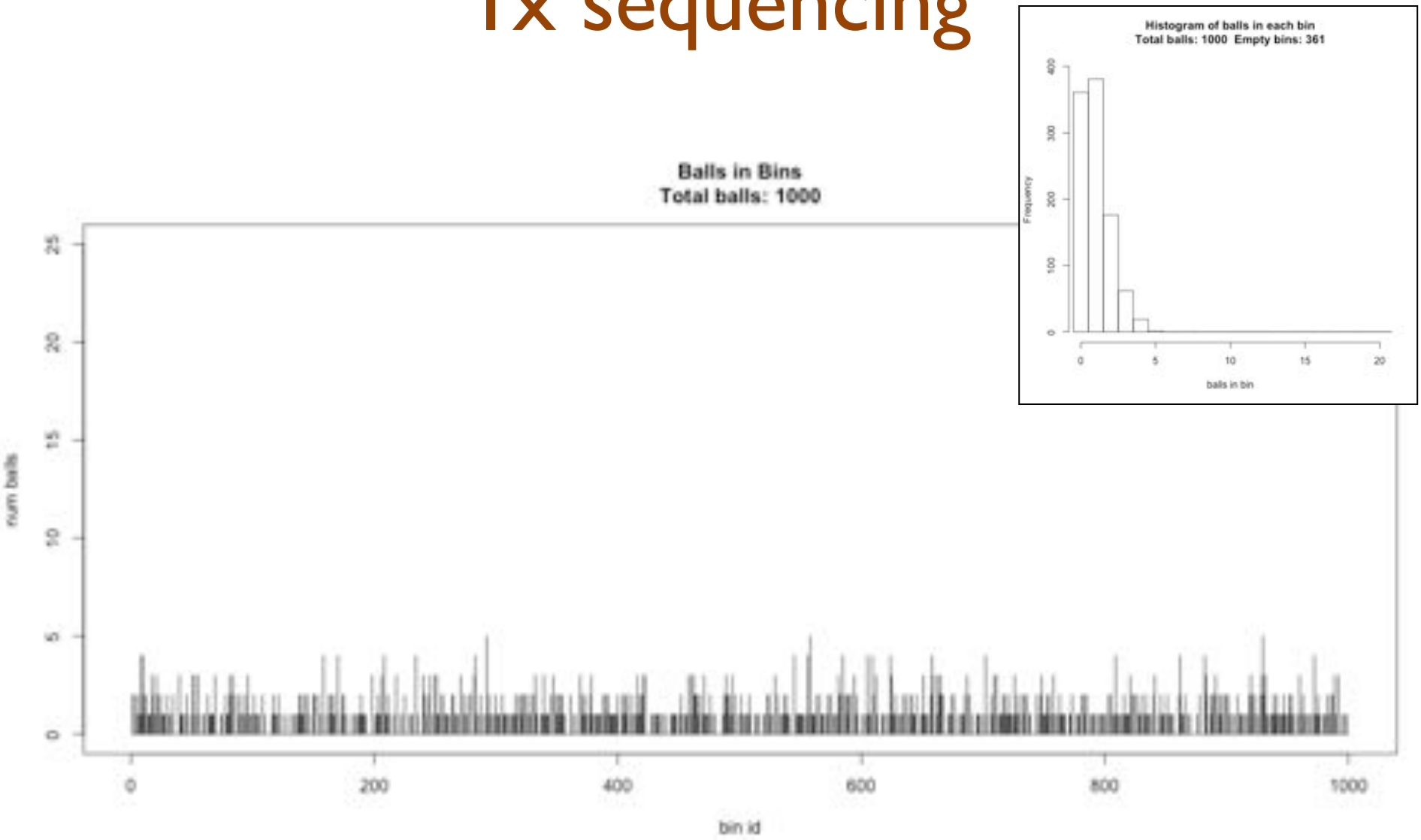
Typical sequencing coverage



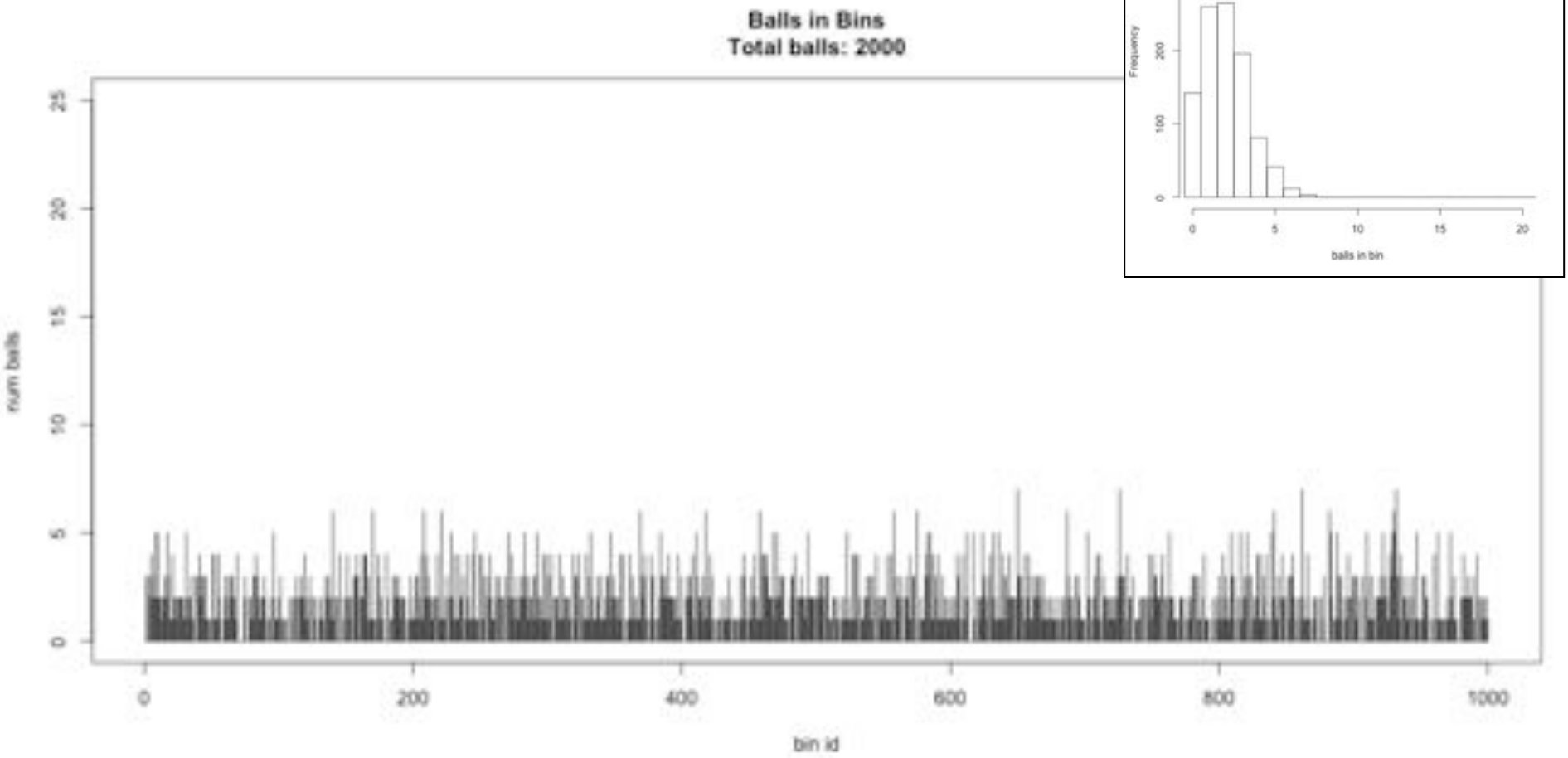
Imagine raindrops on a sidewalk

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

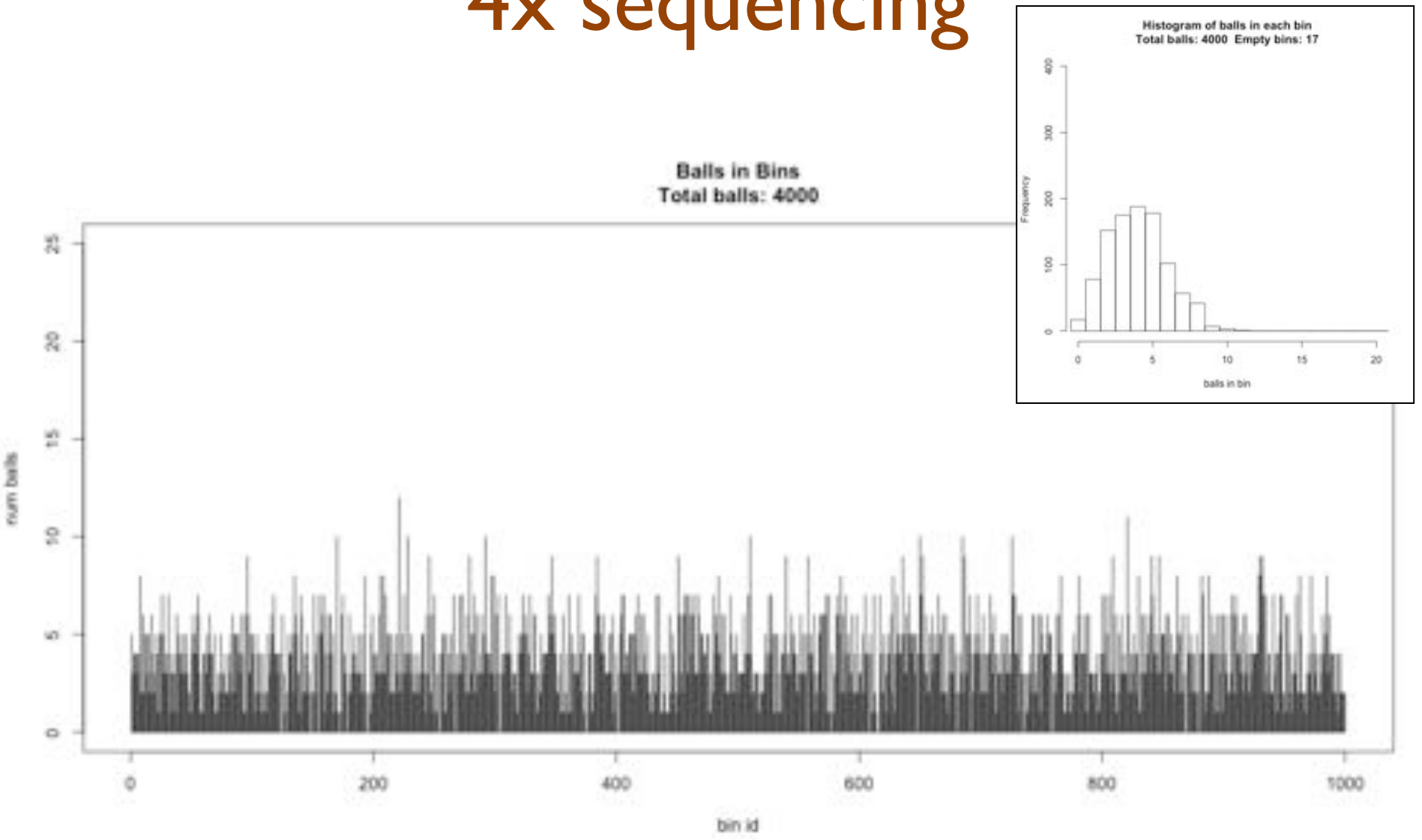
1x sequencing



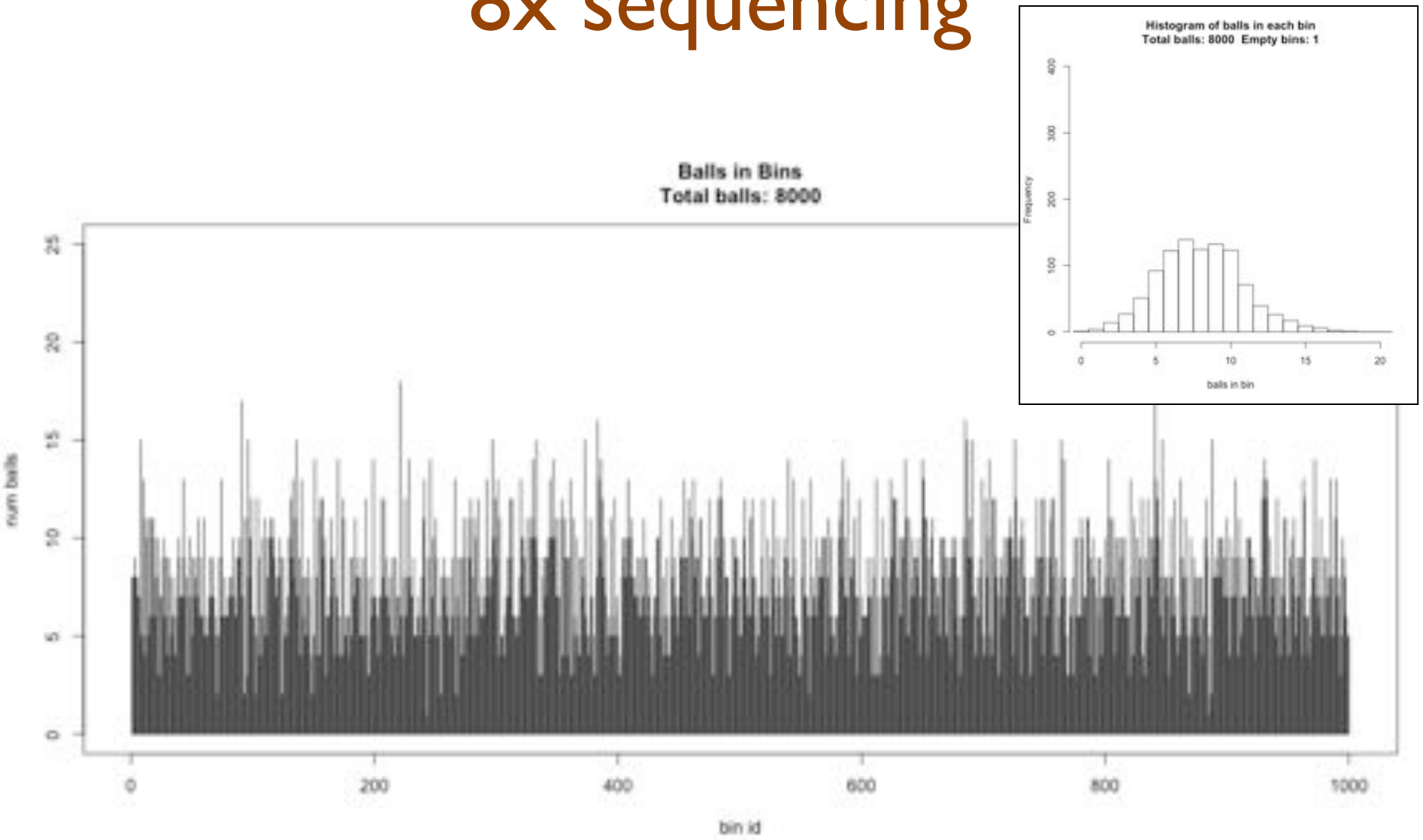
2x sequencing



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.

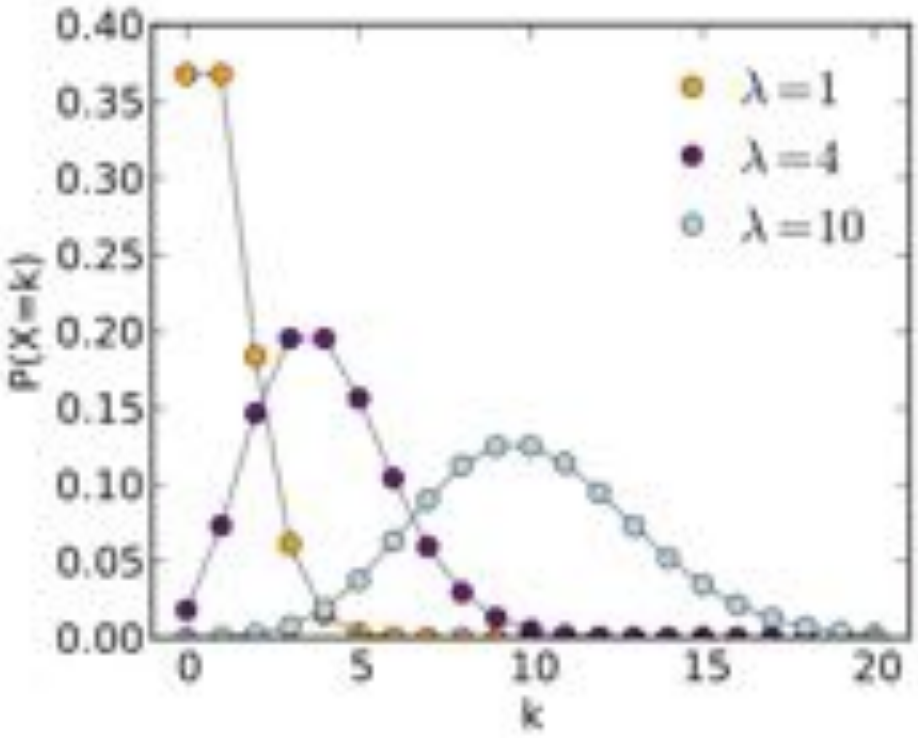
$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Formulation comes from the limit of the binomial equation

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

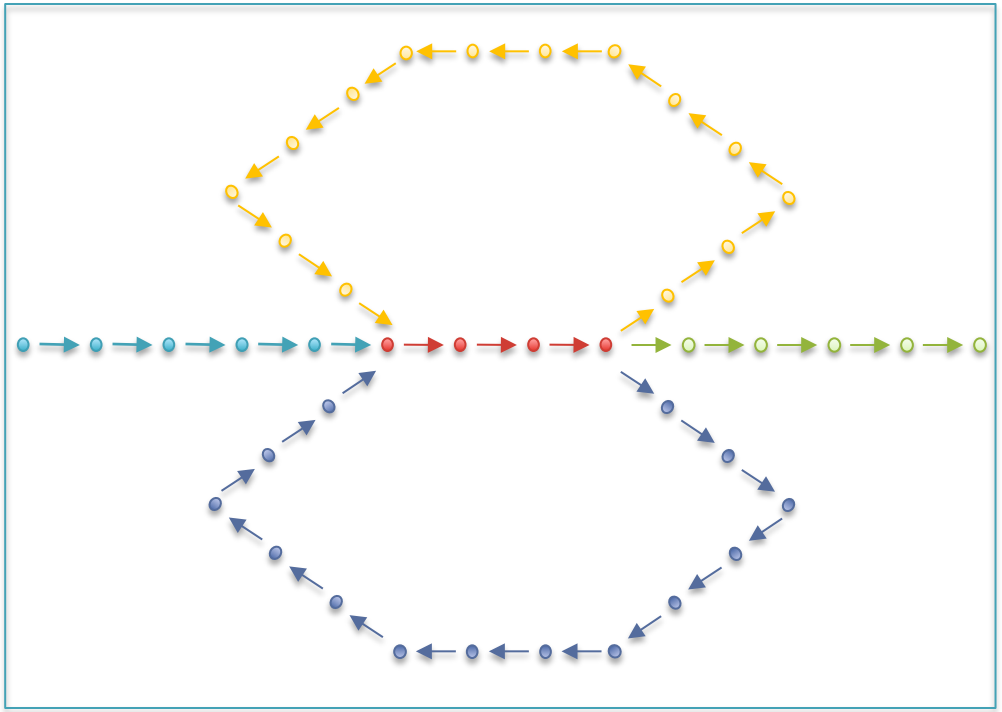
Key property:

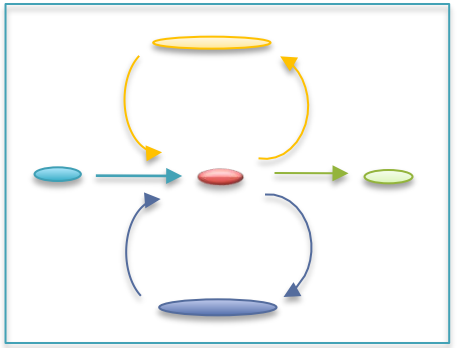
- ***The standard deviation is the square root of the mean.***



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, (3) heterozygosity, and (4) repeats



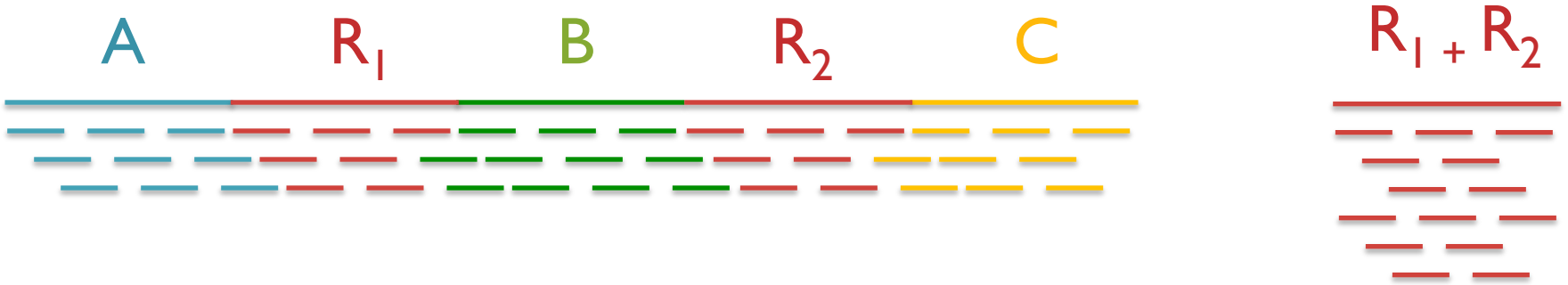


Repetitive regions

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.

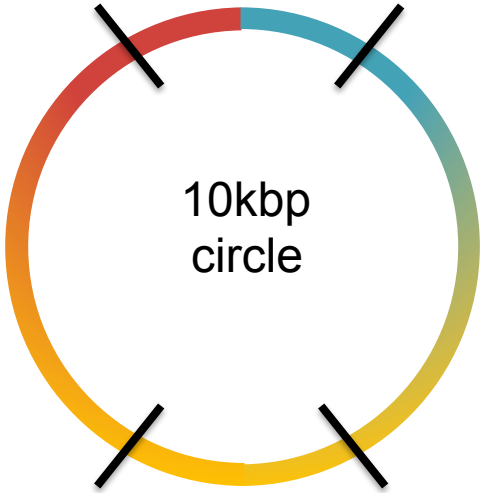
Paired-end and Mate-pairs

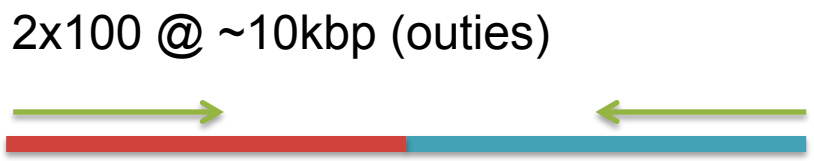
Paired-end sequencing

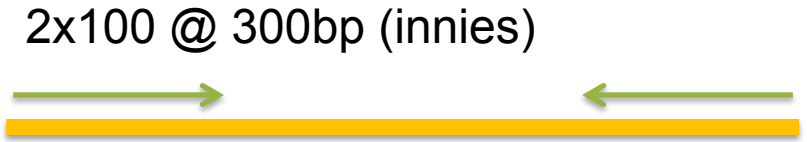
- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation

Mate-pair sequencing

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

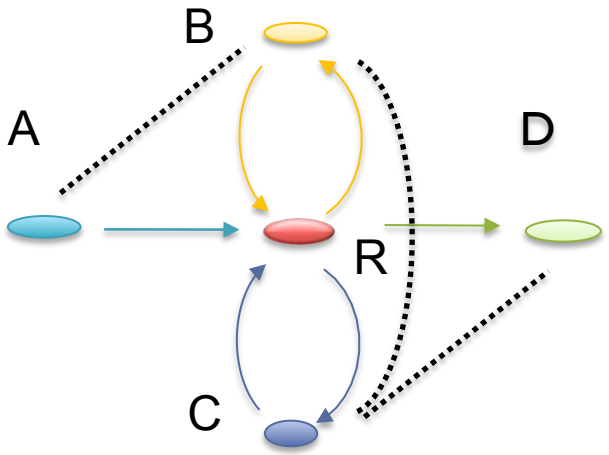




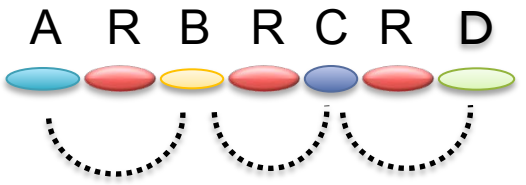


Scaffolding

- 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

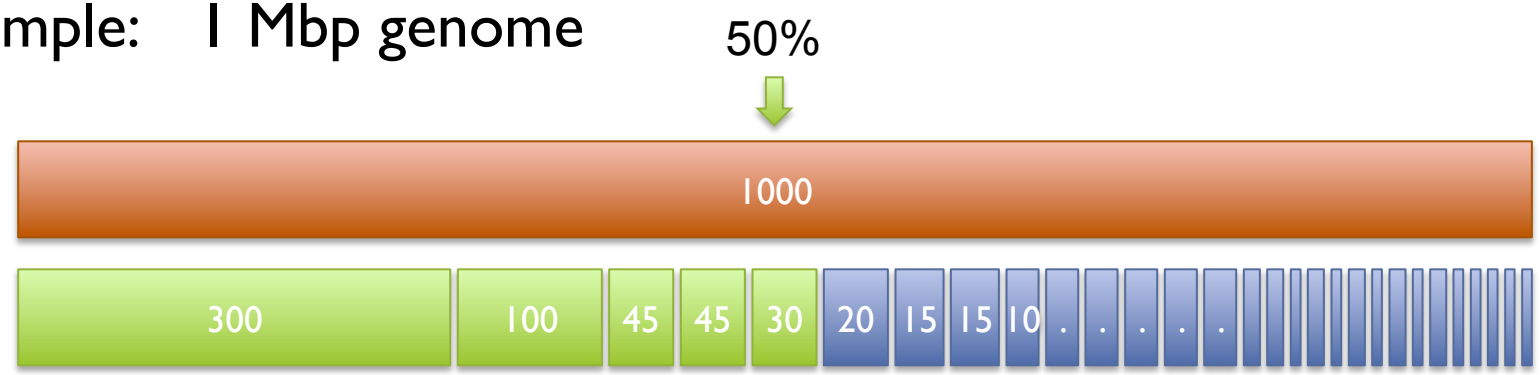


- Final scaffold may have internal gaps called sequencing gaps
  - We know the order, orientation, and spacing, but just not the bases. Fill with Ns instead

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)

A greater N50 is indicative of improvement 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

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 long read projects

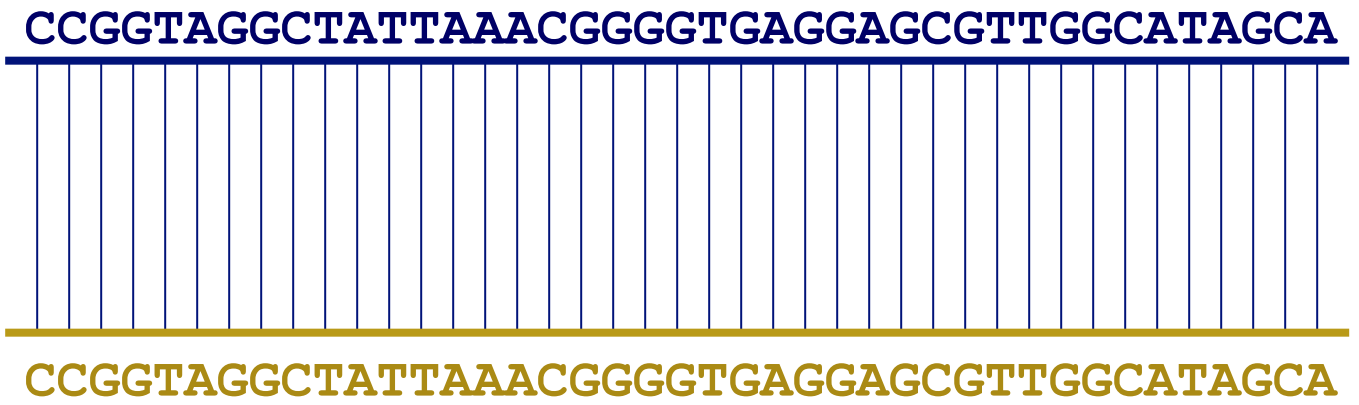
4. Summary and Recommendations

Whole Genome Alignment with MUMmer

Slides Courtesy of Adam M. Phillippy
University of Maryland

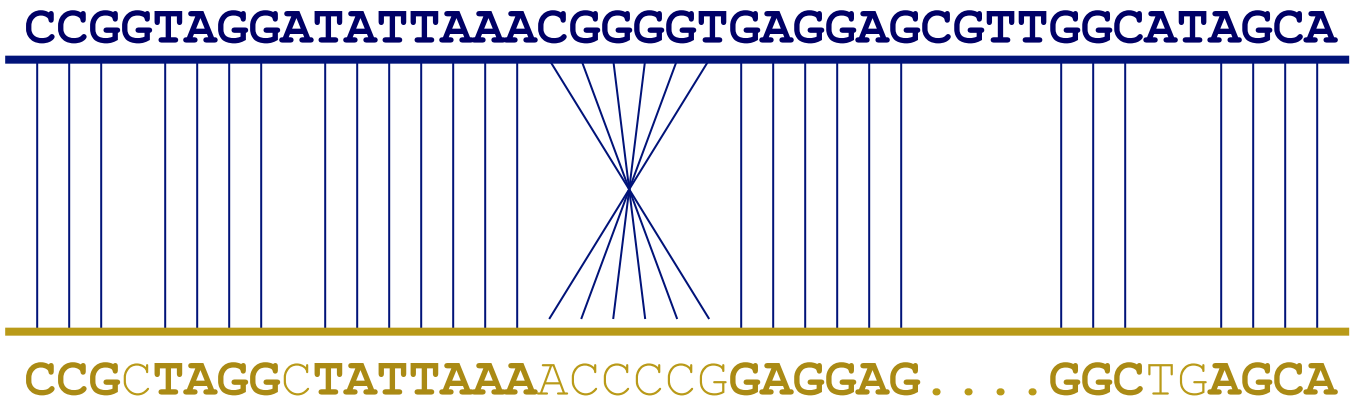
Goal of WGA

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



Not so fast...

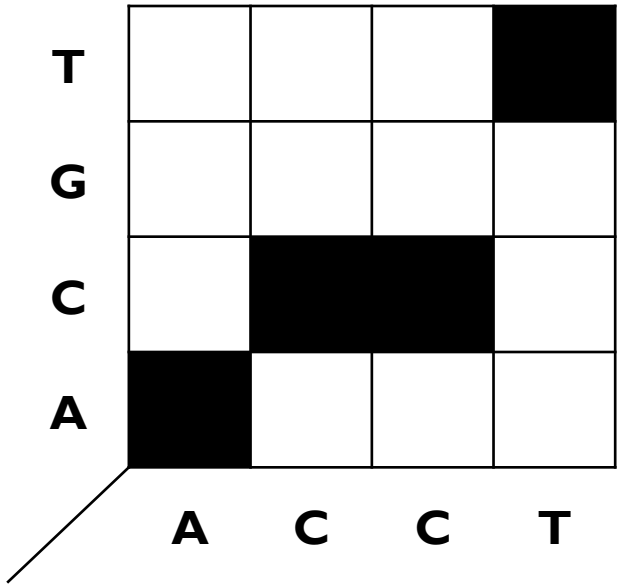
- Genome *A* may have insertions, deletions, translocations, inversions, duplications or SNPs with respect to *B* (sometimes all of the above)



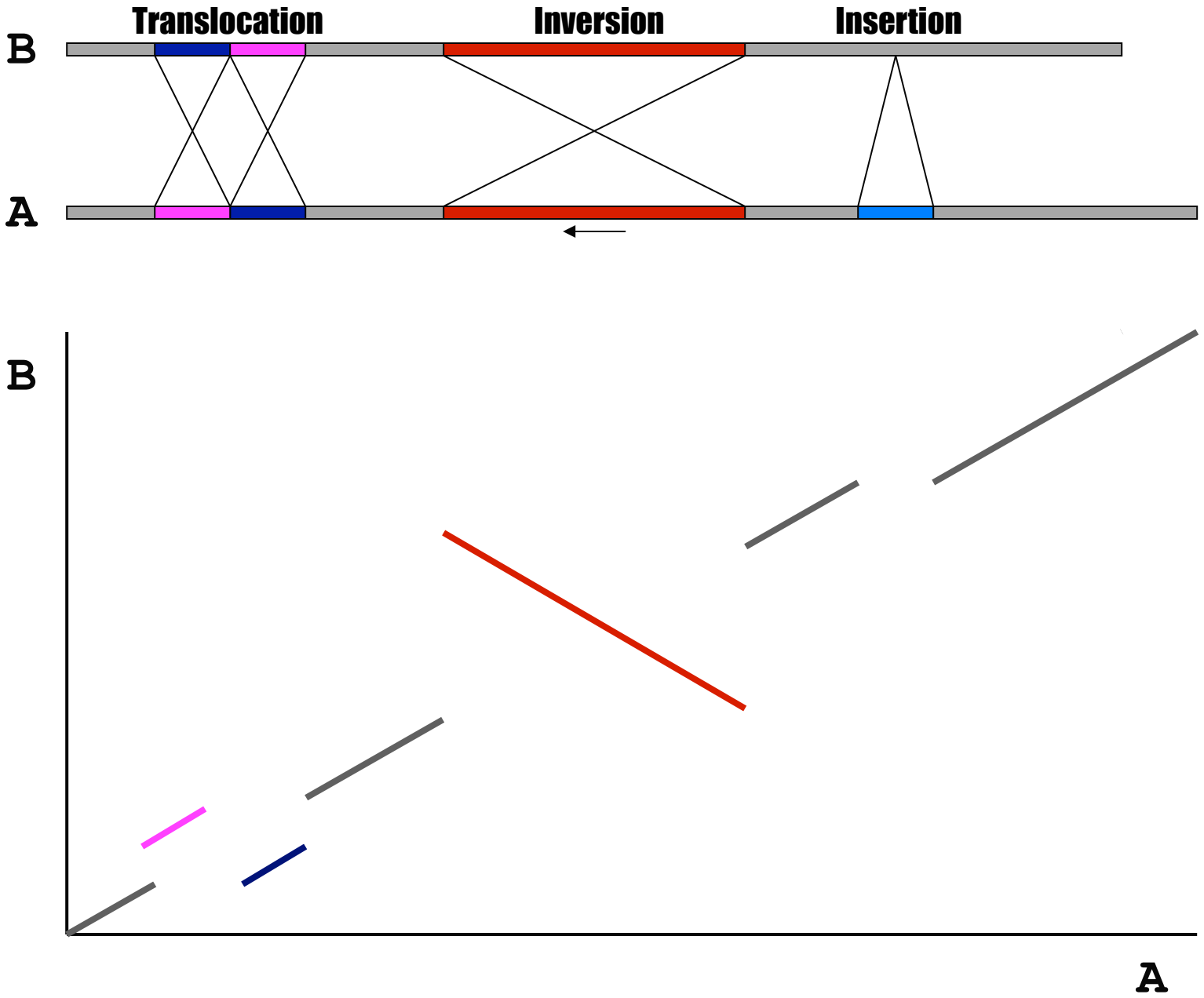
WGA visualization

- How can we visualize *whole* genome alignments?

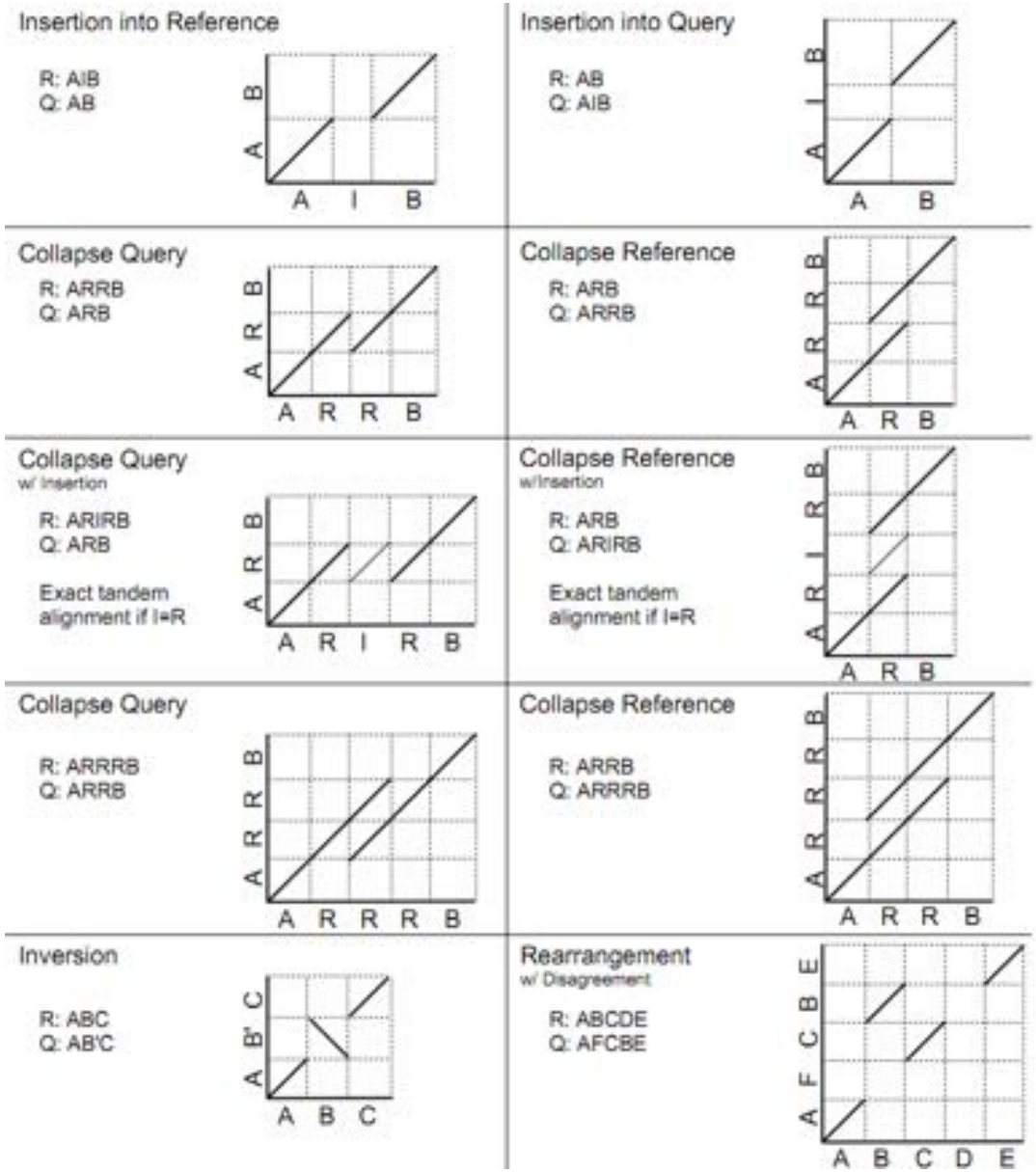
- With an alignment dot plot
  - *N* x *M* matrix
    - Let *i* = position in genome *A*
    - Let *j* = position in genome *B*
    - Fill cell *(i,j)* if A_i shows similarity to B_j



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

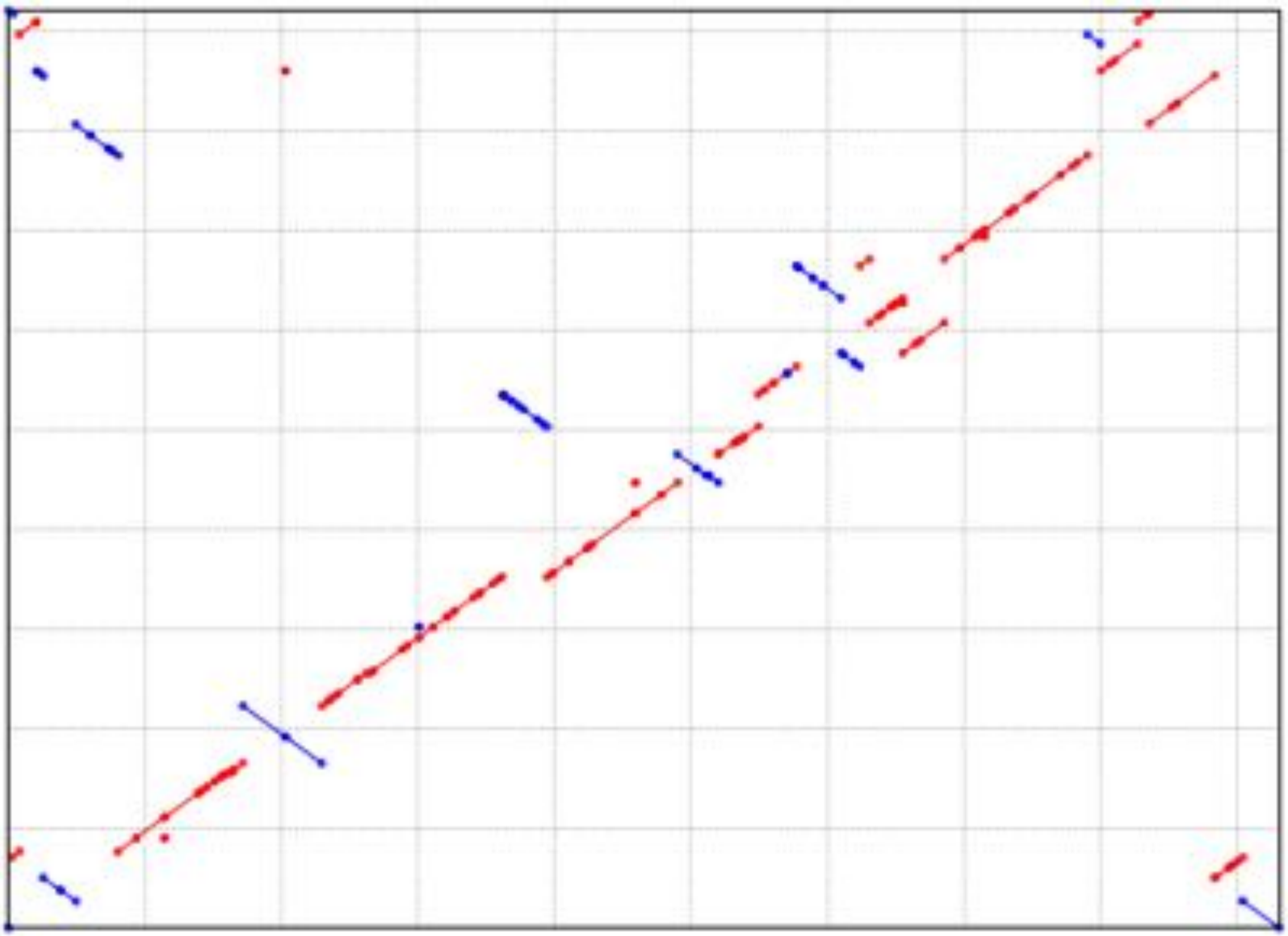

Translocation
Inversion
Insertion
B
A
B
A

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/

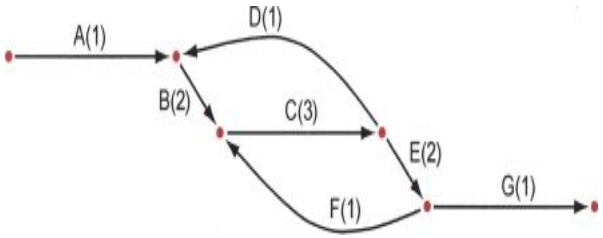
Outline

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 long read projects

4. Summary and Recommendations

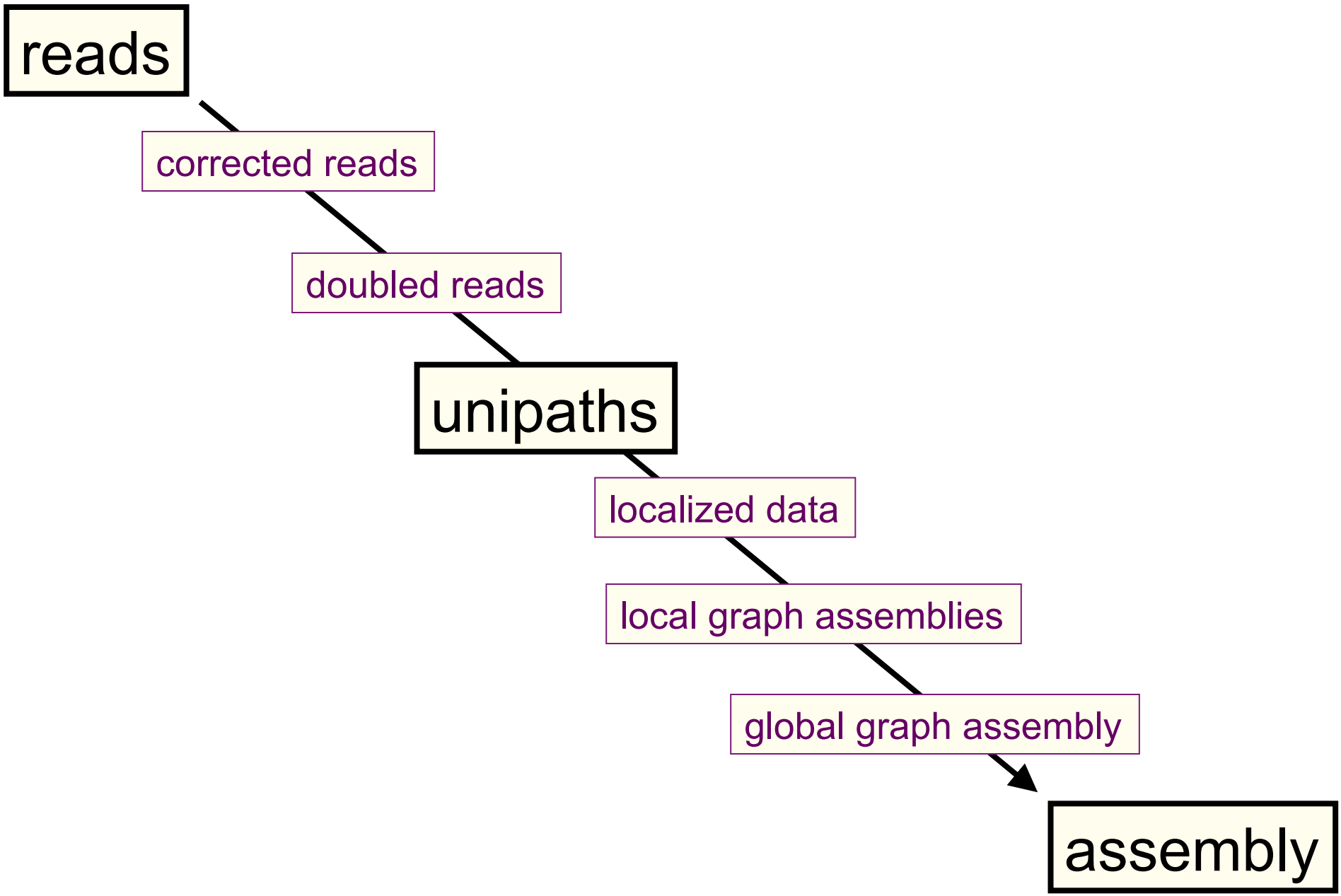


Short read assembly with ALLPATHS-LG

Iain MacCallum

BROAD INSTITUTE

How ALLPATHS-LG works



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.

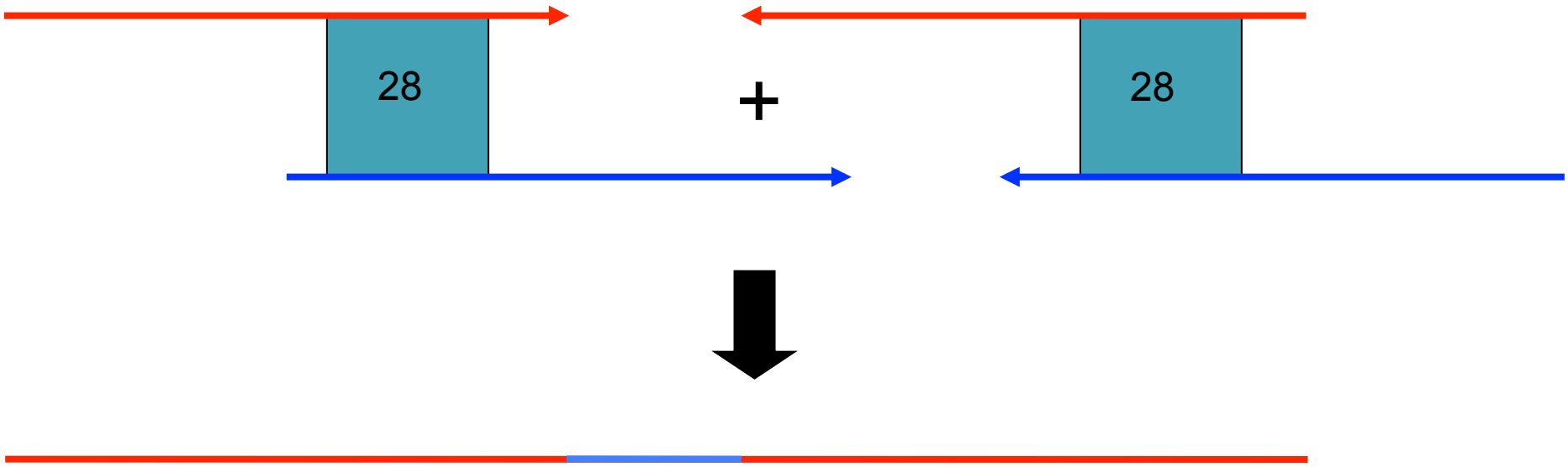
**For best results. Normally not used for small genomes.
However essential to assemble long repeats or duplications.

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:

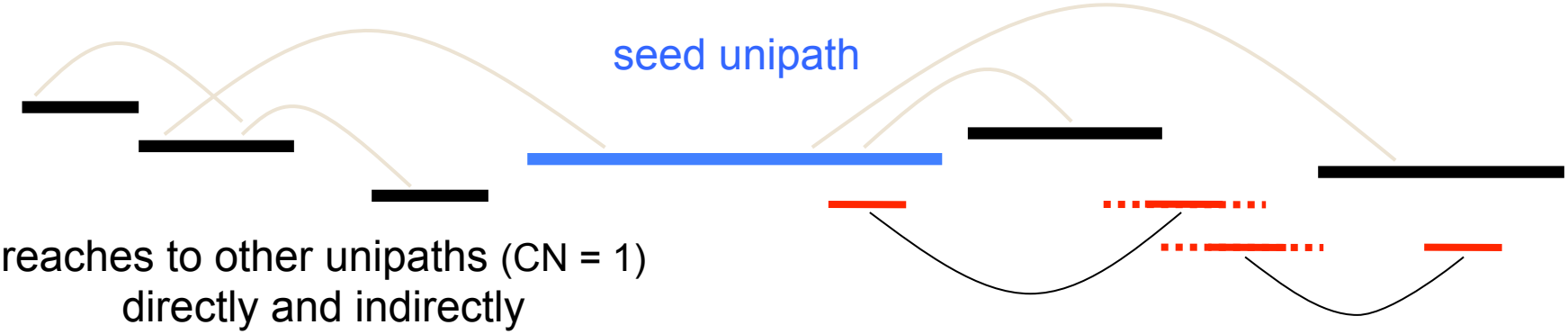


More than one closure allowed (but rare).

Localization

I. Find 'seed' unipaths, evenly spaced across genome
(ideally long, of copy number CN = 1)

II. Form neighborhood around each seed

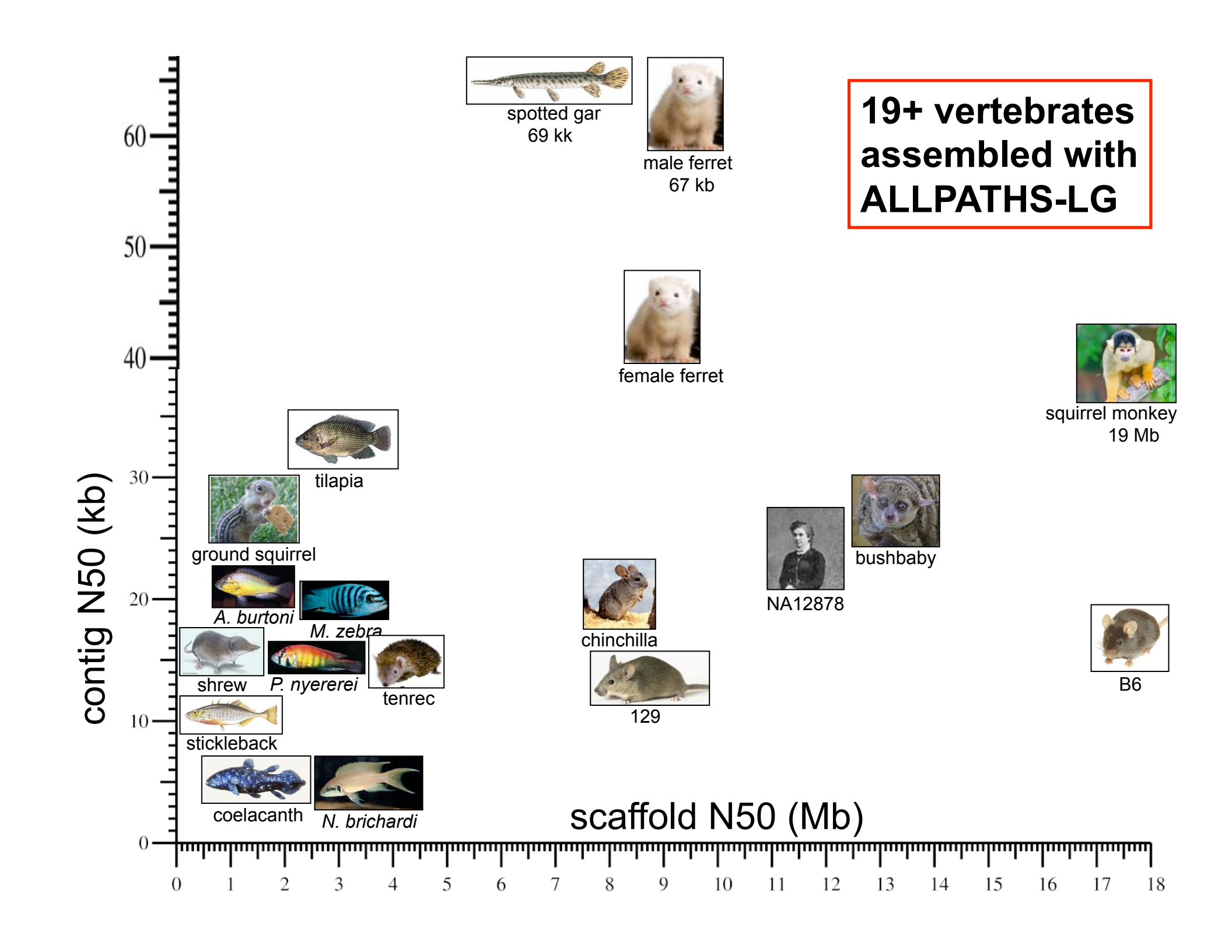


reaches to other unipaths (CN = 1)
directly and indirectly

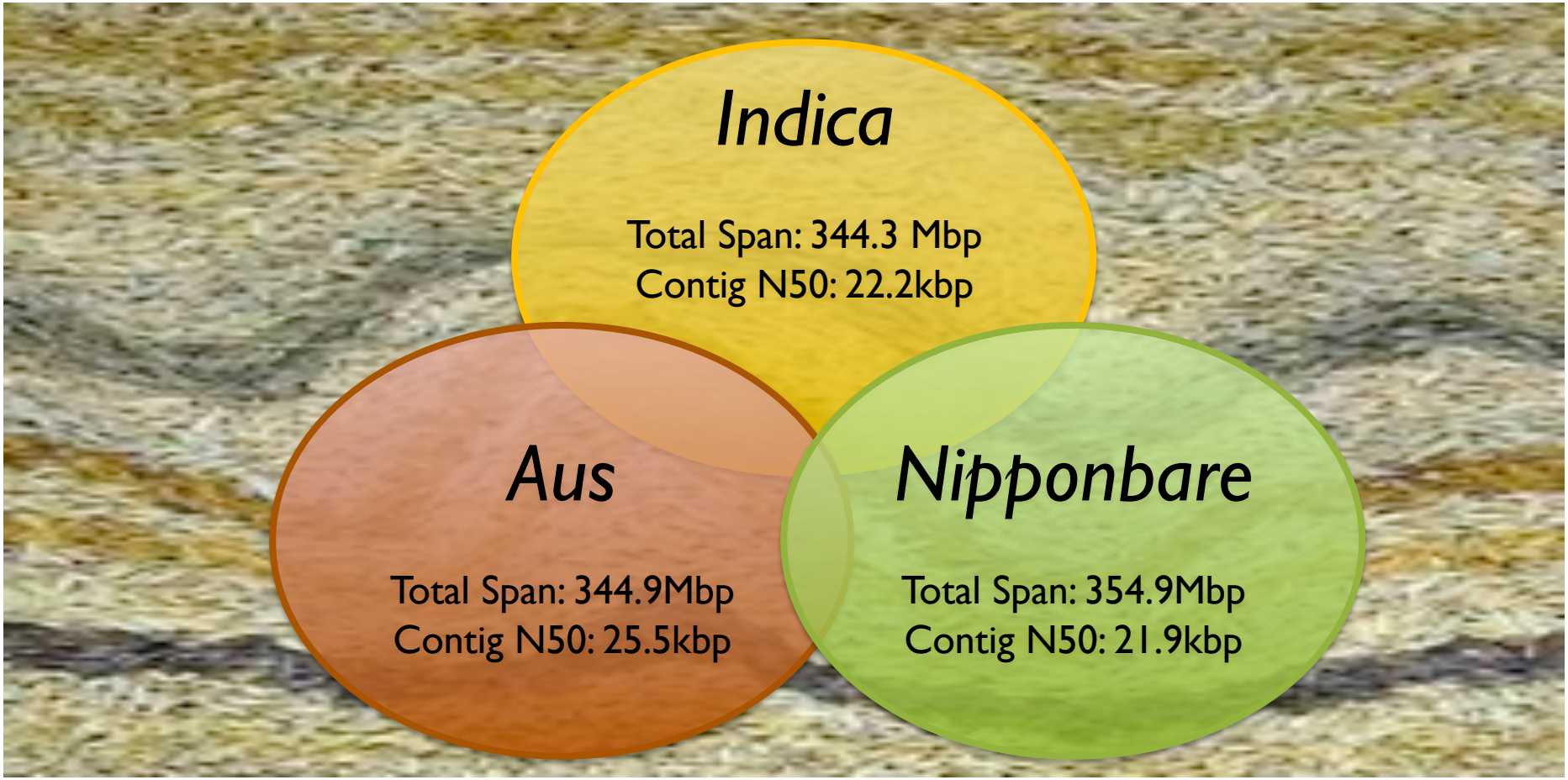
read pairs reach into repeats

and are extended by other unipaths

.....


19+ vertebrates assembled with ALLPATHS-LG
spotted gar
69 kk
male ferret
67 kb
female ferret
squirrel monkey
19 Mb
tilapia
ground squirrel
A. burtoni
M. zebra
shrew
P. nyererei
tenrec
stickleback
coelacanth
N. brichardi
chinchilla
129
NA12878
bushbaby
B6
contig N50 (kb)
scaffold N50 (Mb)
0
10
20
30
40
50
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

Population structure of *Oryza sativa*



Whole genome de novo assemblies of three divergent strains of rice (*O. sativa*) documents novel gene space of *aus* and *indica*

Schatz, MC, Maron, L, Stein, *et al* (2014) *Genome Biology.* 15:506 doi:10.1186/s13059-014-0506-z

Pan-genomics of draft assemblies

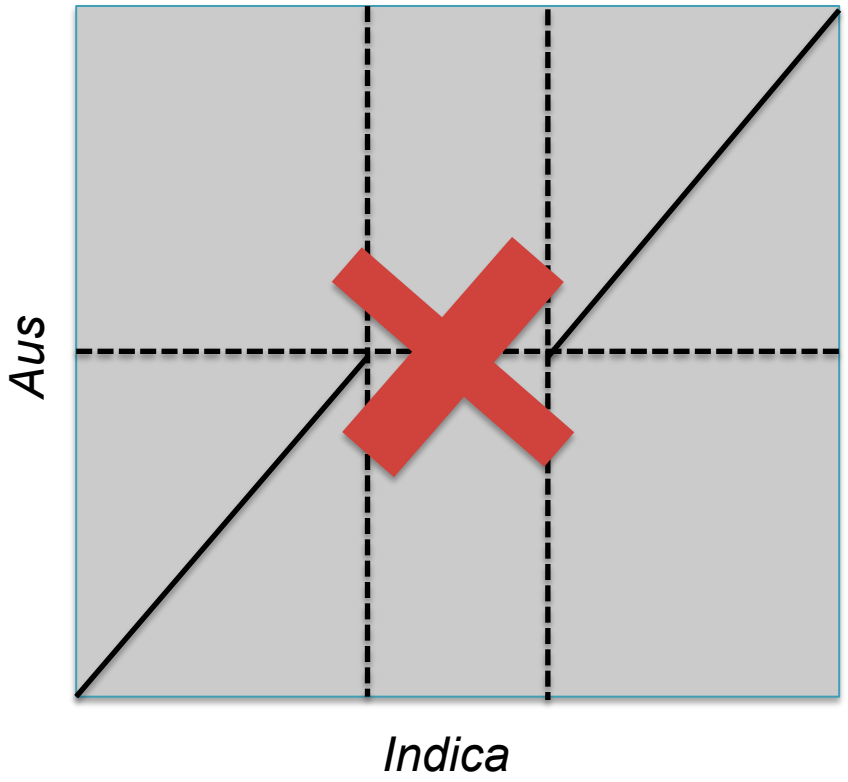
Strategy:

1. Align the genomes to each other (MUMmer)
2. Identify segments of genome A that do not align anywhere to genome B (BEDTools)

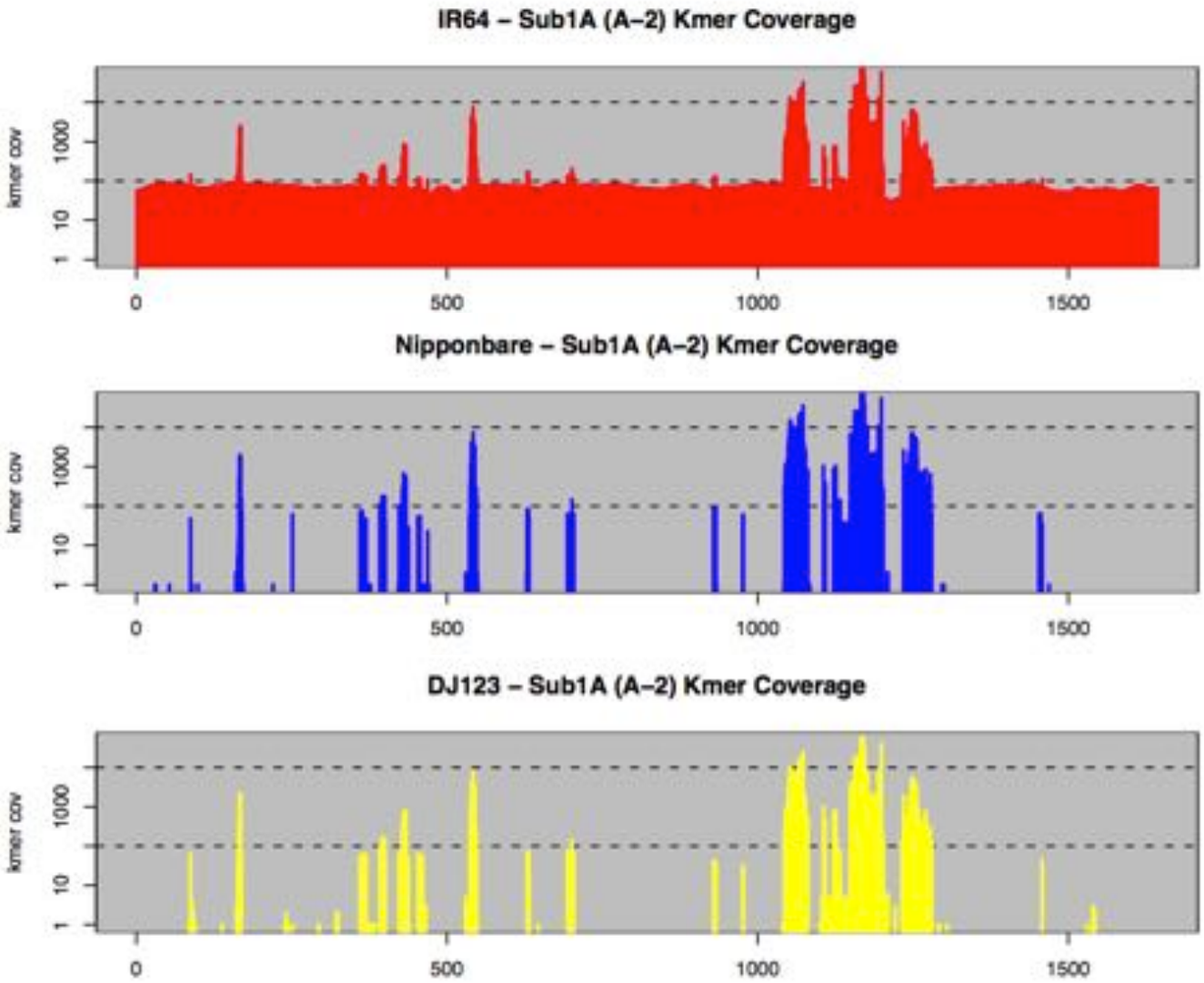
→ Megabases specific to each genome!!!!

3. Screen regions that fail to align with their k-mer frequencies (jellyfish)
   - In reality, "Genome specific regions" averaged over 10,000x kmer coverage while unique regions were ~50x

→ 100s of KB specific to each genome!!!



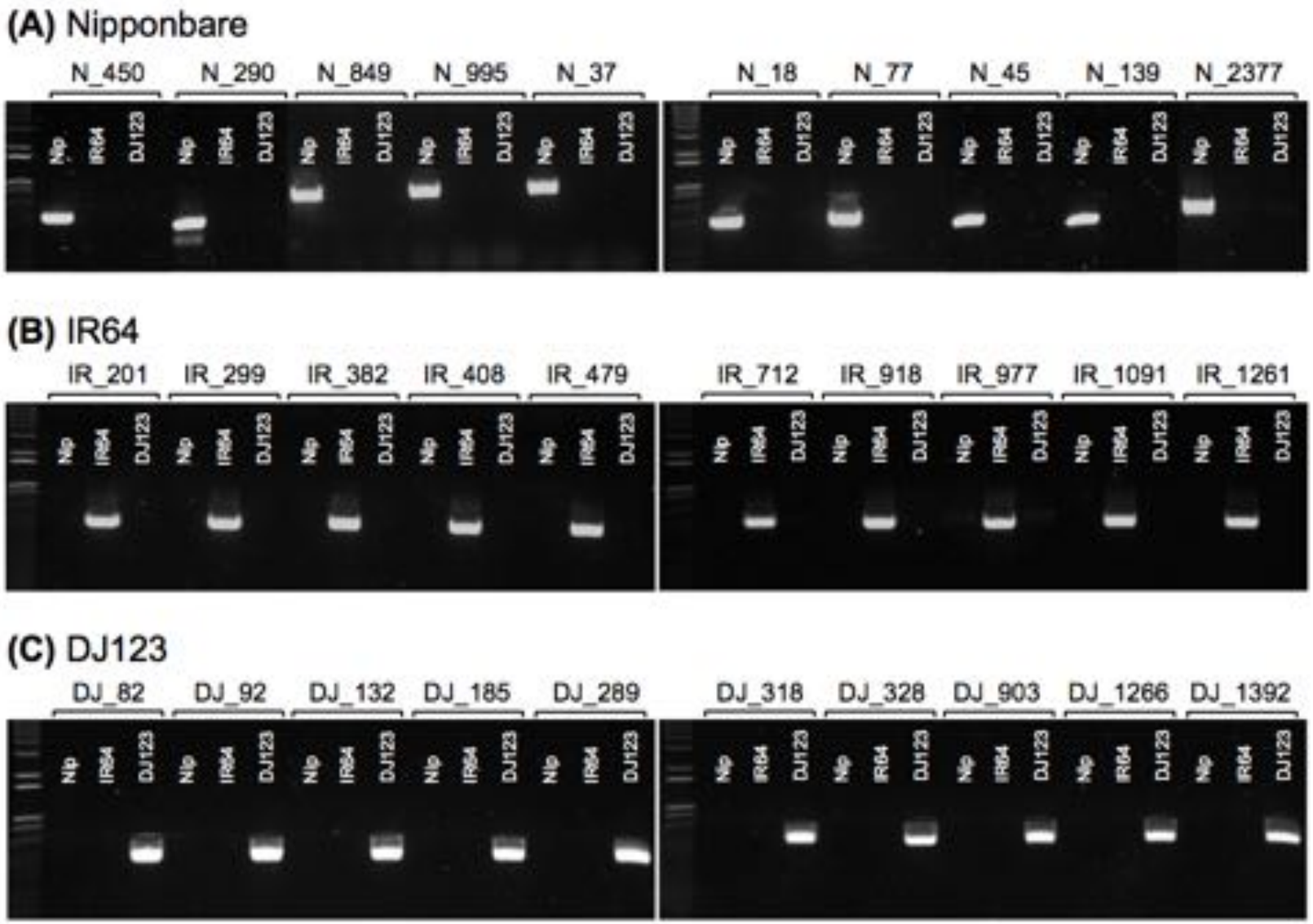
Reference-free kmer analysis



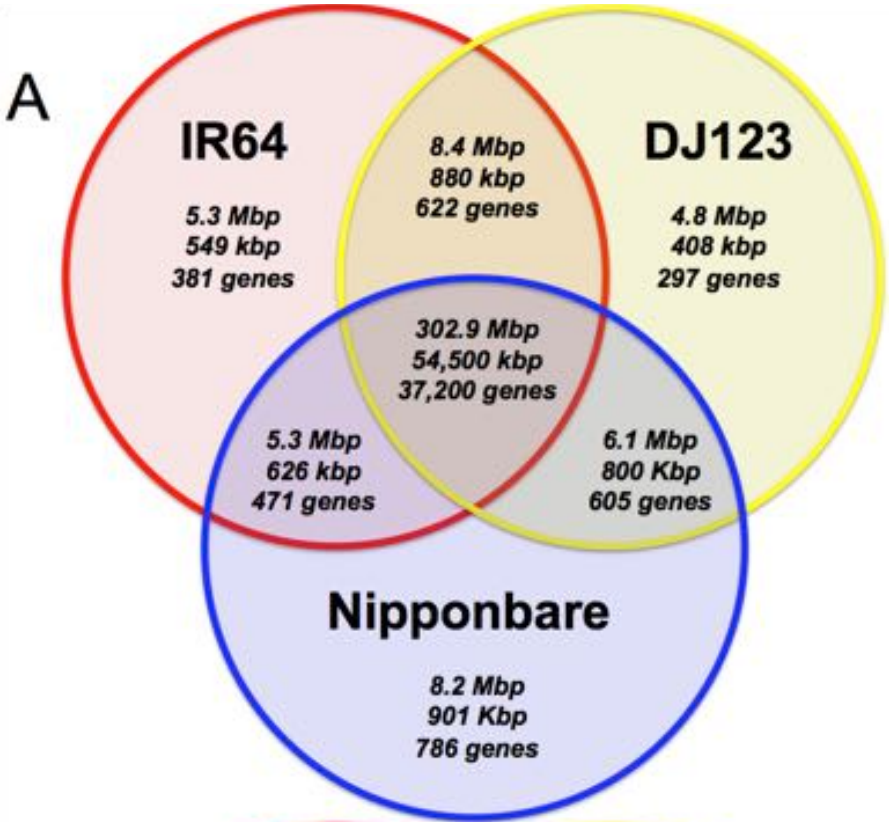
Draft assemblies are difficult to conclusively analyze to determine if a given sequence is truly specific to one genome or another

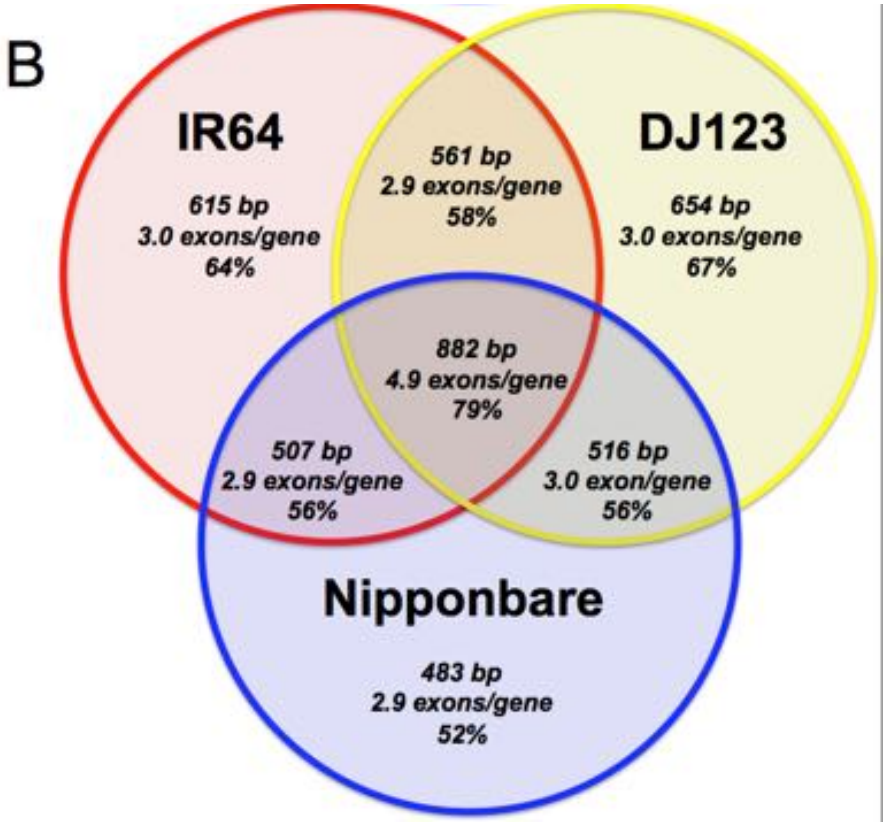
- The sequence may be mis-assembled (or incompletely assembled in the other genome)
- Use k-mer analysis to rule out mis-assemblies
- Here we see the *Sub1A (A-2)* locus present only in IR64

Strain specific regions



Oryza sativa Gene Diversity



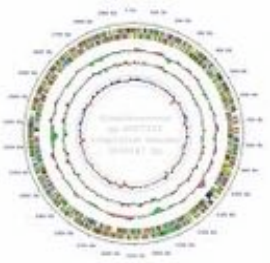


Overall sequence content
In each sector, the top number is the total number of base pairs, the middle number is the number of exonic bases, and the bottom is the gene count. If a gene is partially shared, it is assigned to the sector with the most exonic bases.

Genic content
In each sector, the top number is the median CDS length, the middle number is the average number of exons per gene, and the bottom is the percentage InterPro/homology.

Strain specific regions

- Very high quality representation of the "gene-space"
  - Overall identity ~99.9%
  - Less than 1% of exonic bases missing
- Genome-specific genes enriched for disease resistance
  - Reflects their geographic and environmental diversity
  - Detailed analysis of the *S5* hybrid sterility locus, the *Sub1* submergence tolerance locus, the *LRK* gene cluster associated with improved yield, and the *Pup1* cluster associated with phosphorus deficiency
- Assemblies fragmented at (high copy) repeats
  - Missing regions have mean k-mer coverage >10,000x
  - Difficult to identify full length gene models and regulatory features



Long read assembly with the Celera Assembler

Vol 436|11 August 2005|doi:10.1038/nature03895 nature

ARTICLES

The map-based sequence of the rice genome

International Rice Genome Sequencing Project*

Rice, one of the world's most important food plants, has important synteny relationships with the other cereal species and is a model plant fo
389 Mb genome, inclu
transposable-element
Arabidopsis. In a recip
proteome. Twenty-nin
classes of transposabl
maize and sorghum ge
nuclear chromosomes.
traits. The additional s
accelerate improveme

Table 2 | Size of each chromosome based on sequence data and estimated gaps

Chr	Sequenced bases (bp)	Gaps on arm regions No.	Gaps on arm regions Length (Mb)	Telomeric gaps* (Mb)	Centromeric gap† (Mb)	rDNA‡ (Mb)	Total (Mb)	Coverage§ (%)
1	43,260,640	5	0.33	0.06	1.40		45.05	99.1
2	35,954,074	3	0.10	0.01	0.72		36.78	99.7
3	36,189,985	4	0.96	0.04	0.18		37.37	97.3
4	35,489,479	3	0.46	0.20			36.15	98.7
5	29,733,216	6	0.22	0.05			30.00	99.3
6	30,731,386	1	0.02	0.03	0.82		31.60	99.8
7	29,643,843	1	0.31	0.01	0.32		30.28	98.9
8	28,434,680	1	0.09	0.05			28.57	99.7
9	22,692,709	4	0.13	0.14	0.62	6.95	30.53	98.8
10	22,683,701	4	0.68	0.13	0.47		23.96	96.6
11	28,357,783	4	0.21	0.04	1.90	0.25	30.76	99.1
12	27,561,960	0	0.00	0.05	0.16		27.77	99.8
All	370,733,456	36	3.51	0.81	6.59	7.20	388.82	98.9

Contig N50: 5.1Mbp
Total projects costs: >$100M

Initial Assembly Attempts with early Illumina sequencers circa 2007-2008

(older Illumina PE76 library with small insert size ~150bp)

Assembler	Data set	N50 contig size	Max contig size	Total assembly size
Velvet	25X Nipponbare	1049bp	21833bp	325.8 Mbp
Velvet	50X Nipponbare	411bp	23095bp	401.6 Mbp
Abyss	25X Nipponbare	1853bp	12688bp	288.4 Mbp
Abyss	50X Nipponbare	2847bp	34893bp	317.4 Mbp

Total costs: ~$10k
>1,000x times cheaper, but at what cost scientifically?

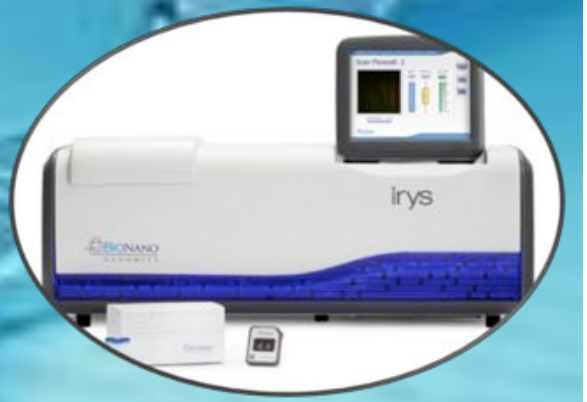
W.R. McCombie

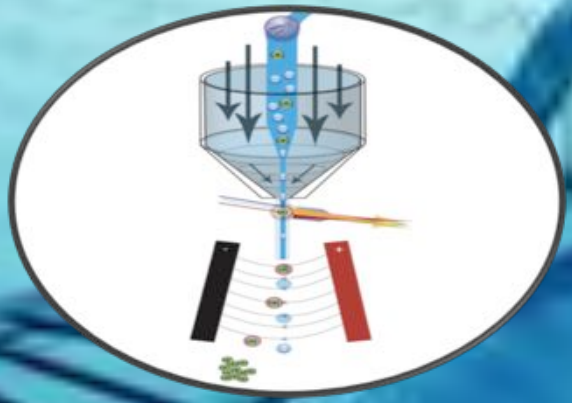
Genomics Arsenal in the year 2015

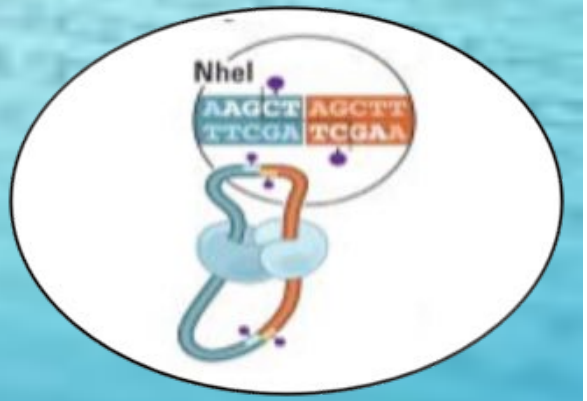
Sample Preparation

Sequencing

Chromosome Mapping

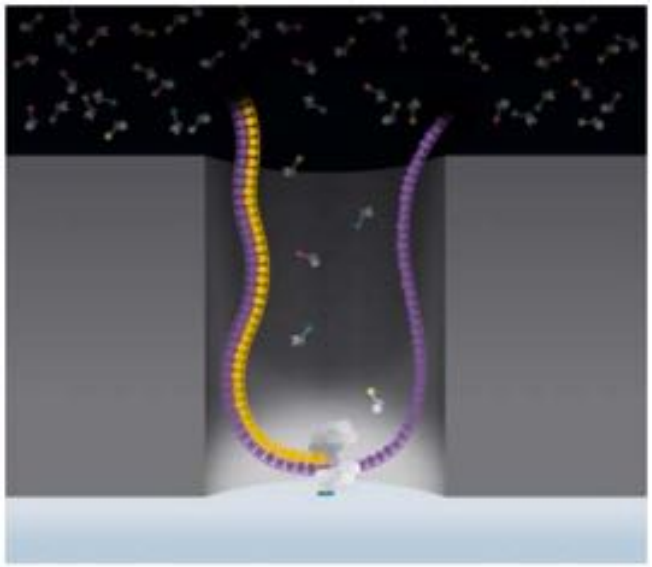


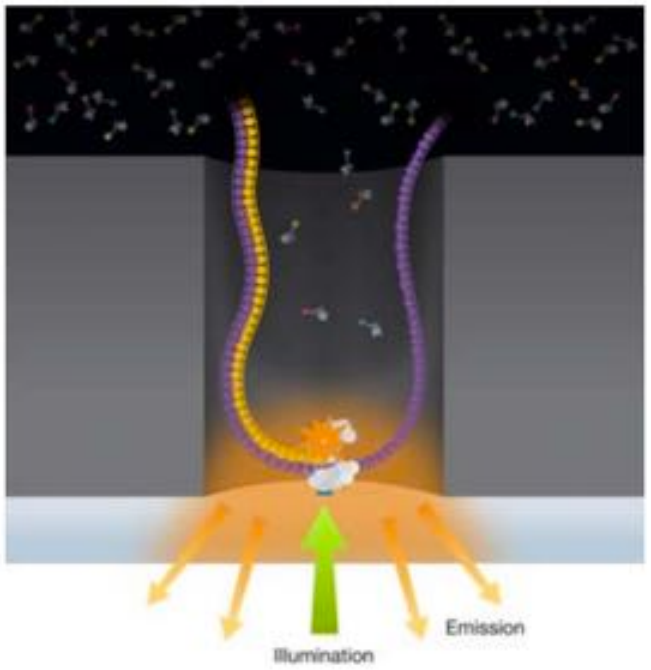


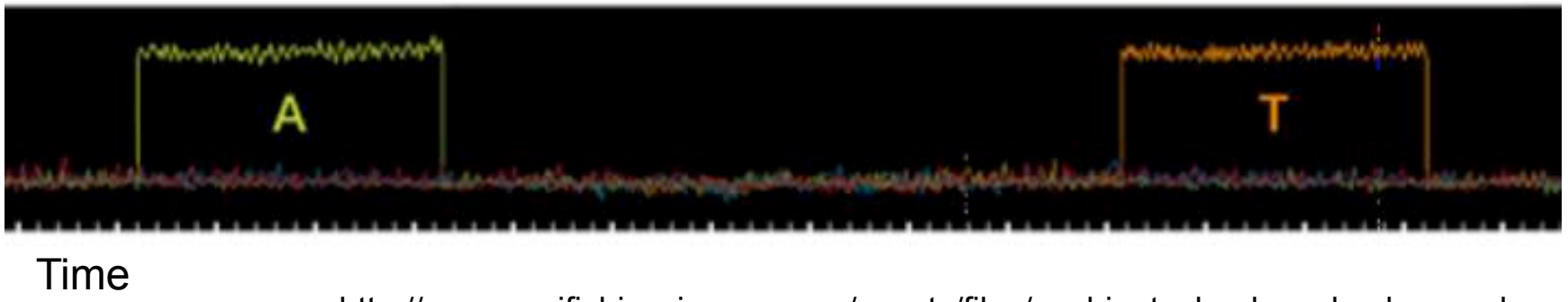


PacBio SMRT Sequencing

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

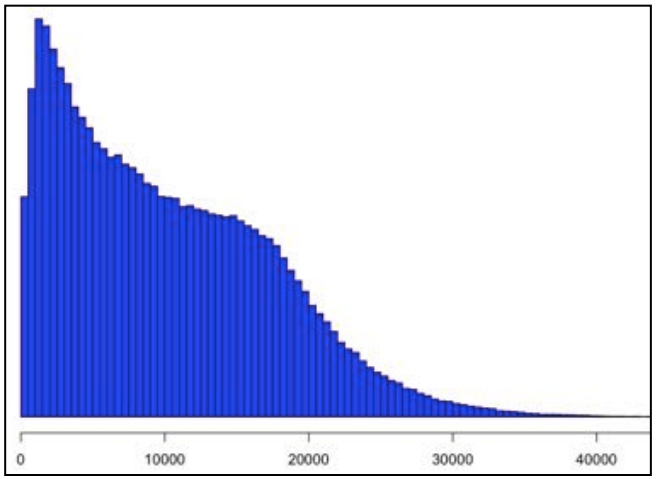






http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf

SMRT Sequencing Data

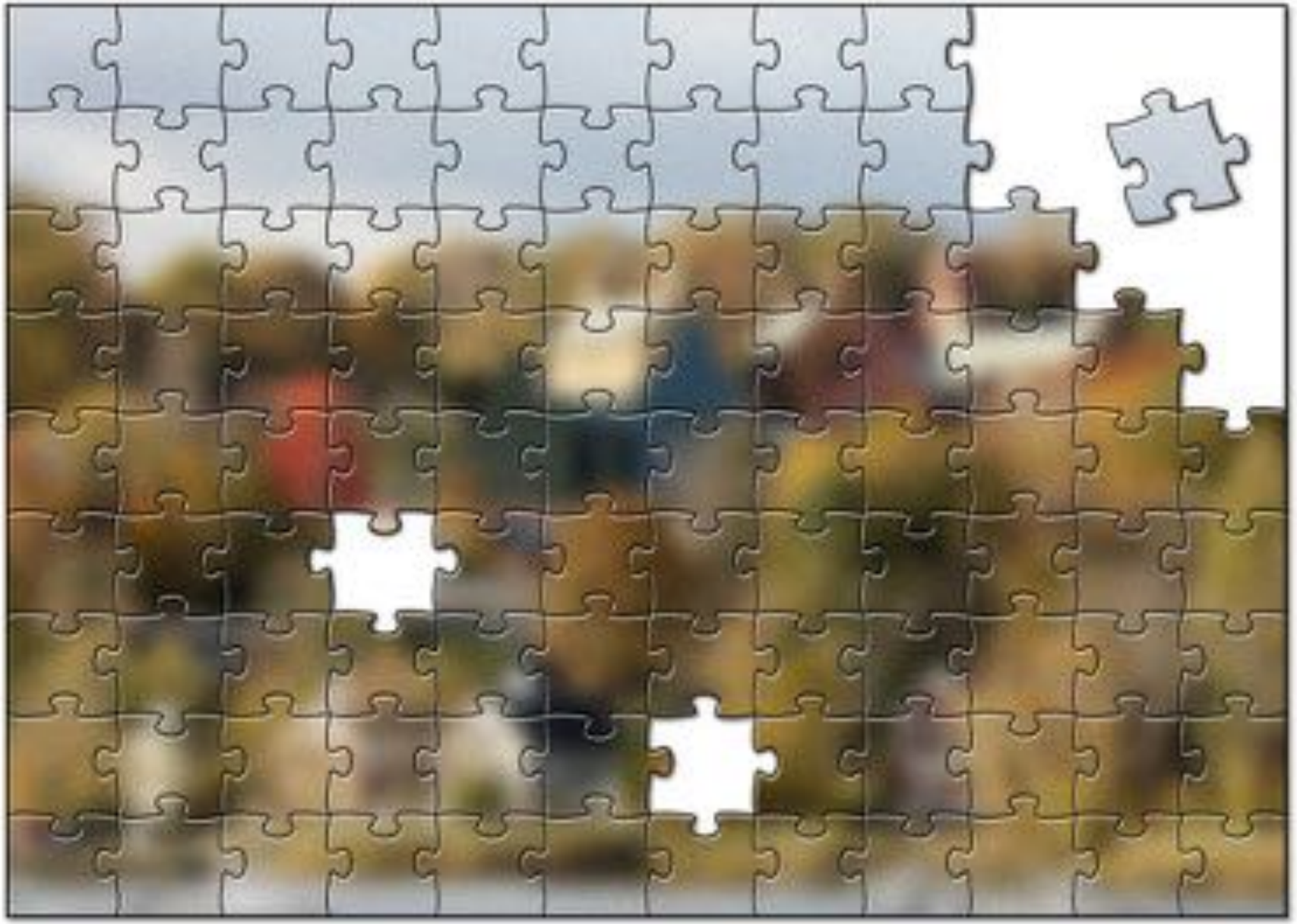


Match	83.7%
Insertions	11.5%
Deletions	3.4%
Mismatch	1.4%

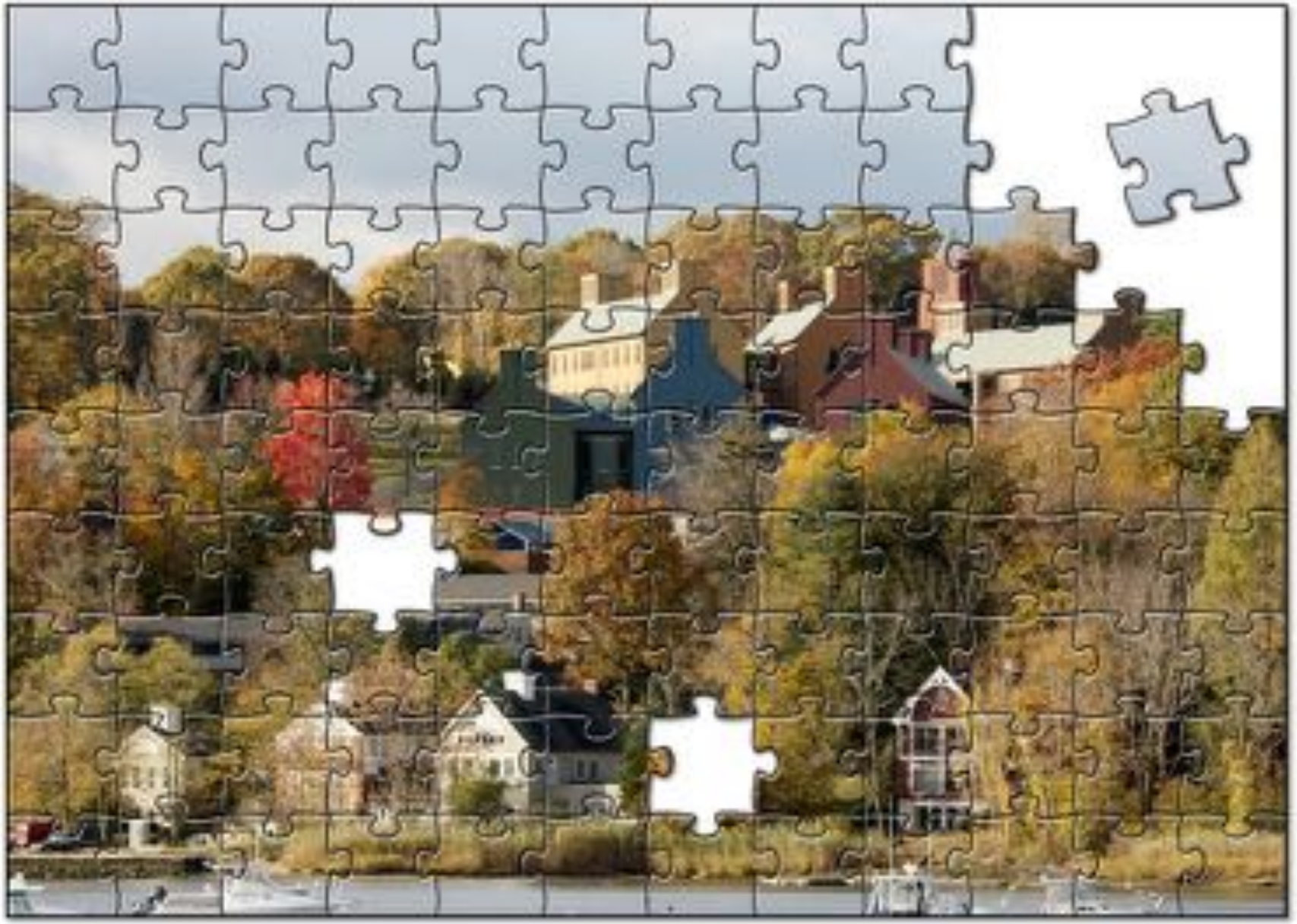
```
TTGTAAGCAGTTGAAAACTATGTGTGGATTTAGAATAAAGAACATGAAAG
|||||||||||||||||||||||||| ||||||| ||||||||||||| |||
TTGTAAGCAGTTGAAAACTATGTGT-GATTTAG-ATAAAGAACATGGAAG

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

ATCCAGTGGAAAATATAATTTATGCAATCCAGGAACTTATTCACAATTAG
||||||| ||||||||  |||||| ||||| ||||||||||||||||||
ATCCAGT-GAAAATATA--TTATGC-ATCCA-GAACTTATTCACAATTAG
```

Sample of 100k reads aligned with BLASR requiring >100bp alignment

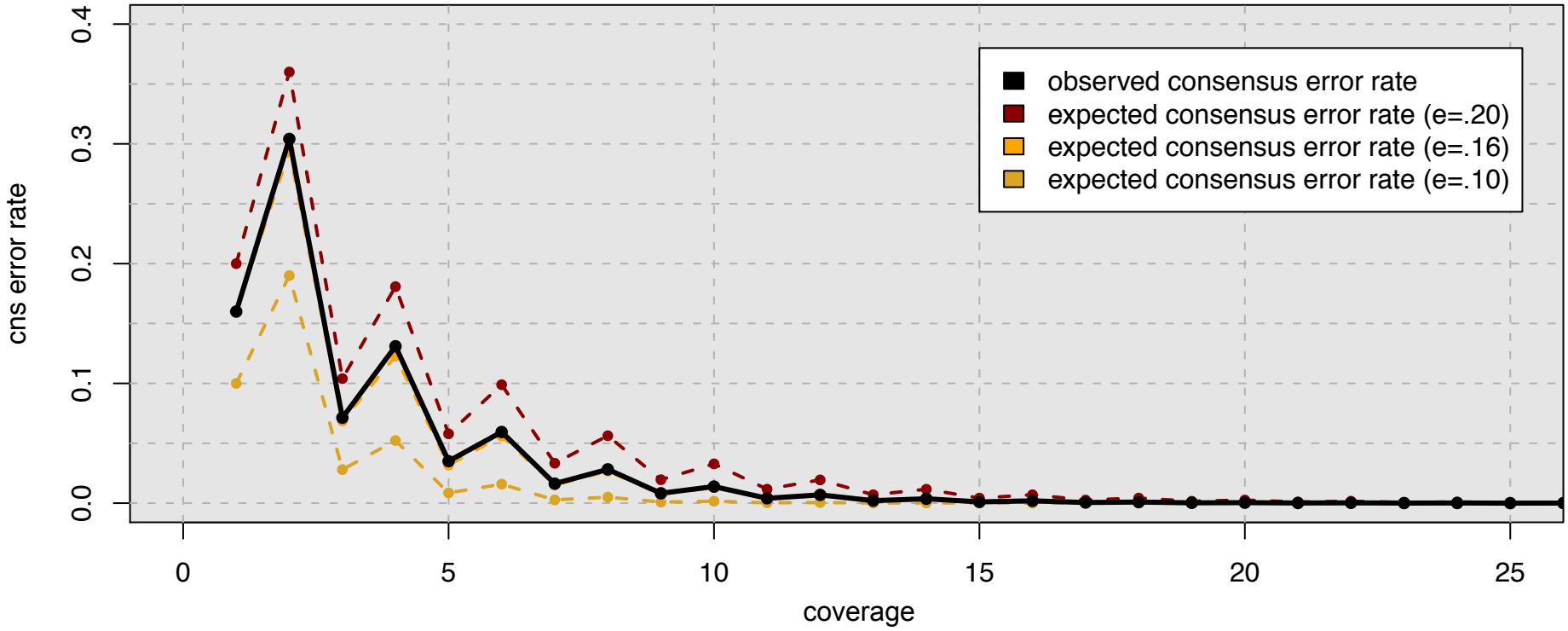
Single Molecule Sequences



“Corrective Lens” for Sequencing



Consensus Accuracy and Coverage



Coverage can overcome random errors

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

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

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

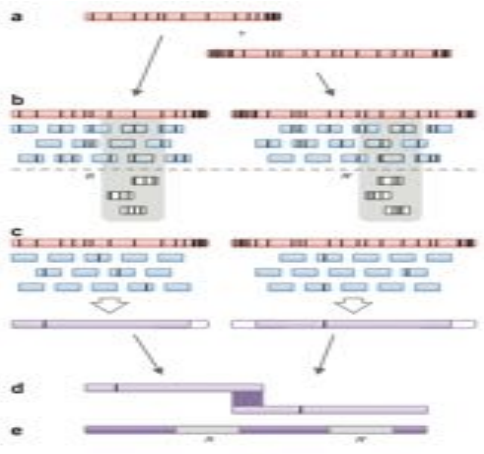
PacBio Assembly Algorithms

PBJelly

Gap Filling and Assembly Upgrade

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

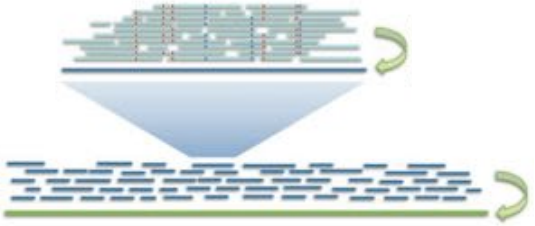
PacBioToCA & ECTools



Hybrid/PB-only Error Correction

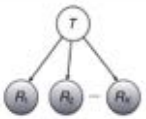
Koren, Schatz, *et al* (2012)
Nature Biotechnology. 30:693–700

HGAP/MHAP & 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

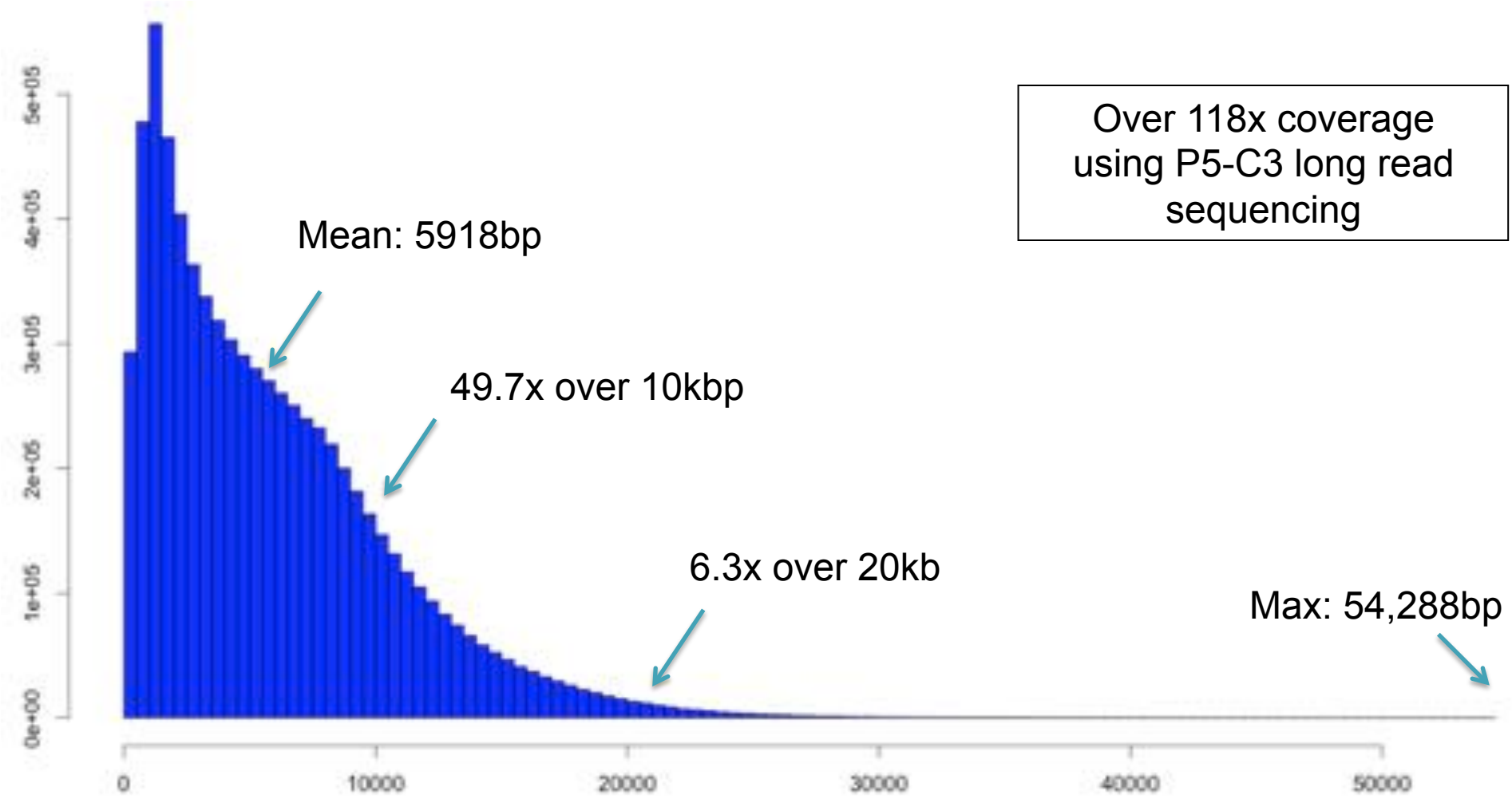
Chin *et al* (2013)
Nature Methods. 10:563–569

< 5x PacBio Coverage > 50x

O. sativa pv Indica (IR64)

PacBio RS II sequencing at PacBio

- Size selection using an 10 Kb elution window on a BluePippin™ device from Sage Science

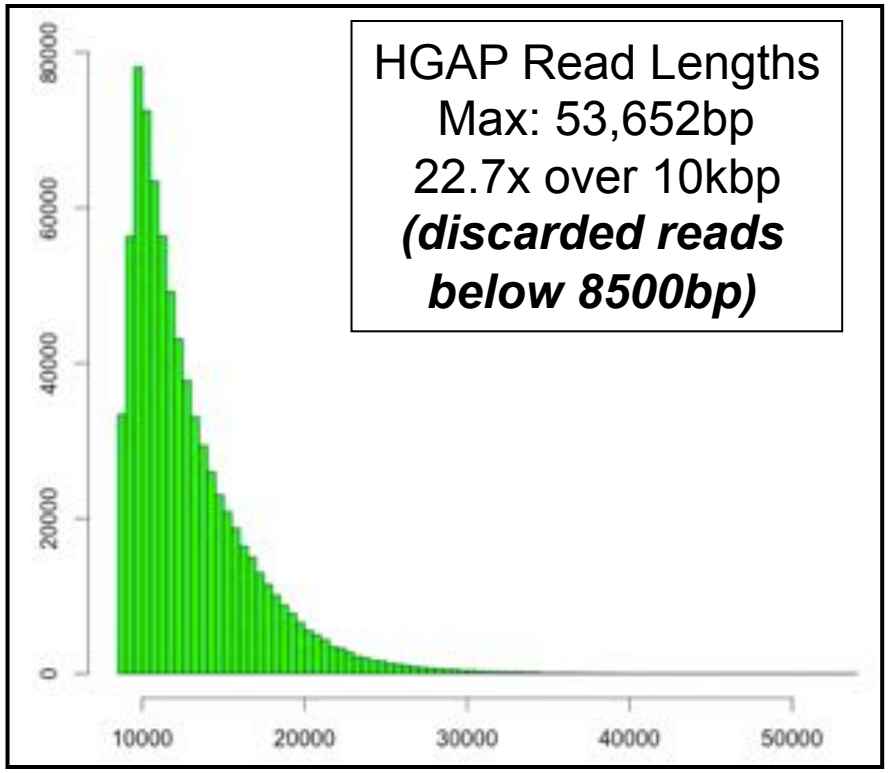


Over 118x coverage using P5-C3 long read sequencing

O. sativa pv Indica (IR64)

Genome size: ~370 Mb
Chromosome N50: ~29.7 Mbp

Assembly	Contig NG50
MiSeq Fragments 25x 456bp (3 runs 2x300 @ 450 FLASH)	19 kbp
"ALLPATHS-recipe" 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18 kbp
HGAP + CA 22.7x @ 10kbp	4.0 Mbp
Nipponbare BAC-by-BAC Assembly	5.1 Mbp



S5 Hybrid Sterility Locus

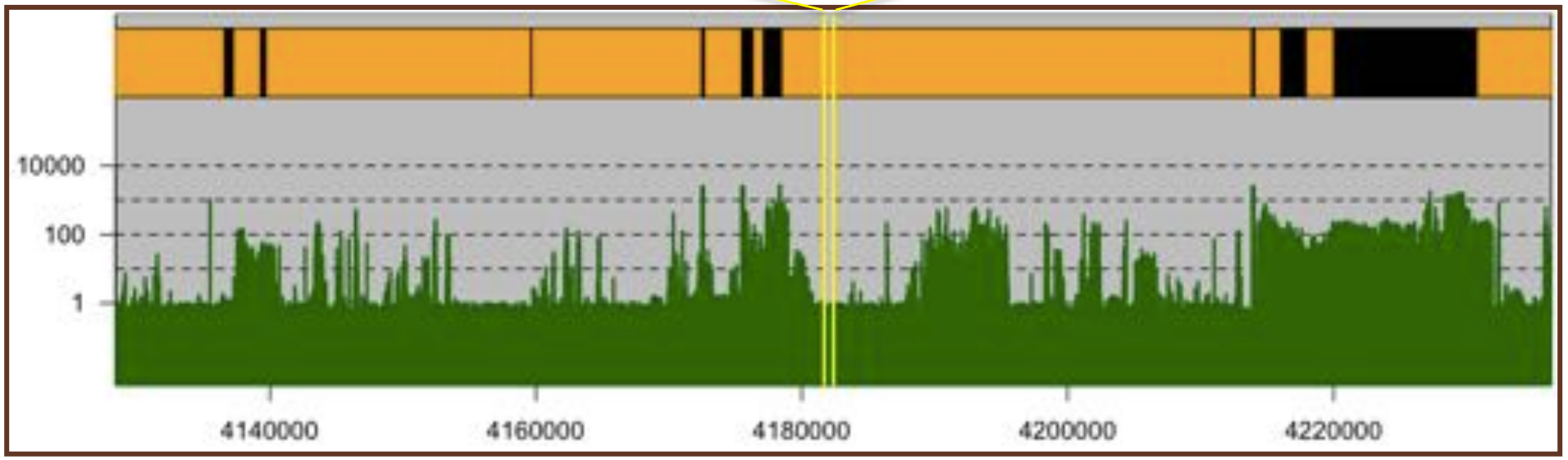
```
Sanger      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
Illumina    …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
PacBio      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
```

S5 is a major locus for hybrid sterility in rice that affects embryo sac fertility.

- Genetic analysis of the S5 locus documented three alleles: an indica (S5-i), a japonica (S5-j), and a neutral allele (S5-n)
- Hybrids of genotype S5-i/S5-j are mostly sterile, whereas hybrids of genotypes consisting of S5-n with either S5-i or S5-j are mostly fertile.
- Contains three tightly linked genes that work together in a 'killer-protector'-type system: ORF3, ORF4, ORF5
- The ORF5 indica (ORF5+) and japonica (ORF5-) alleles differ by only **two nucleotides**

S5 Hybrid Sterility Locus

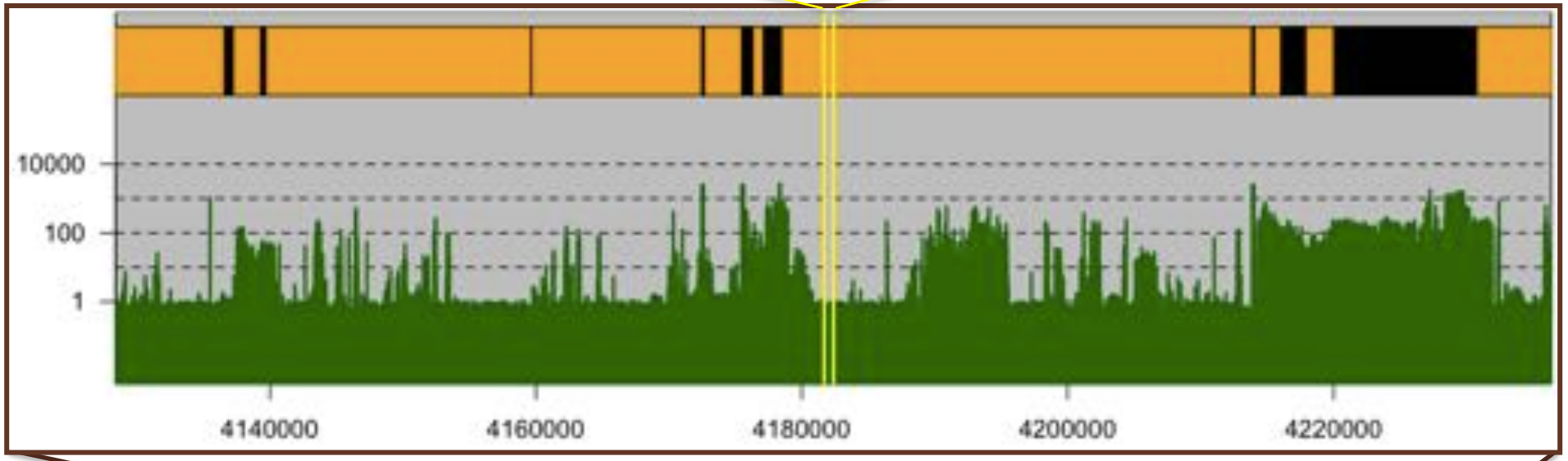
```
Sanger      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
Illumina    …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
PacBio      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
```

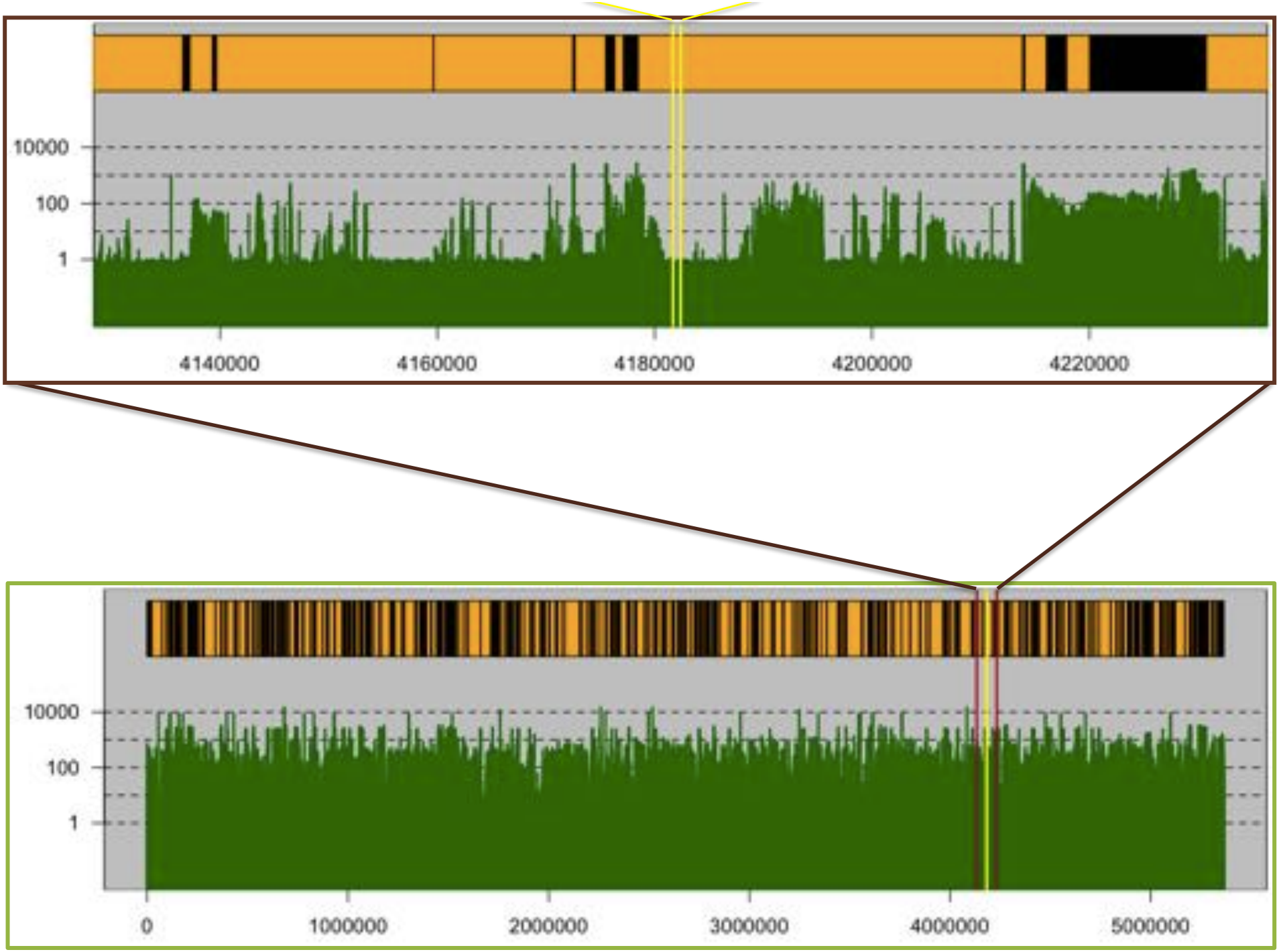


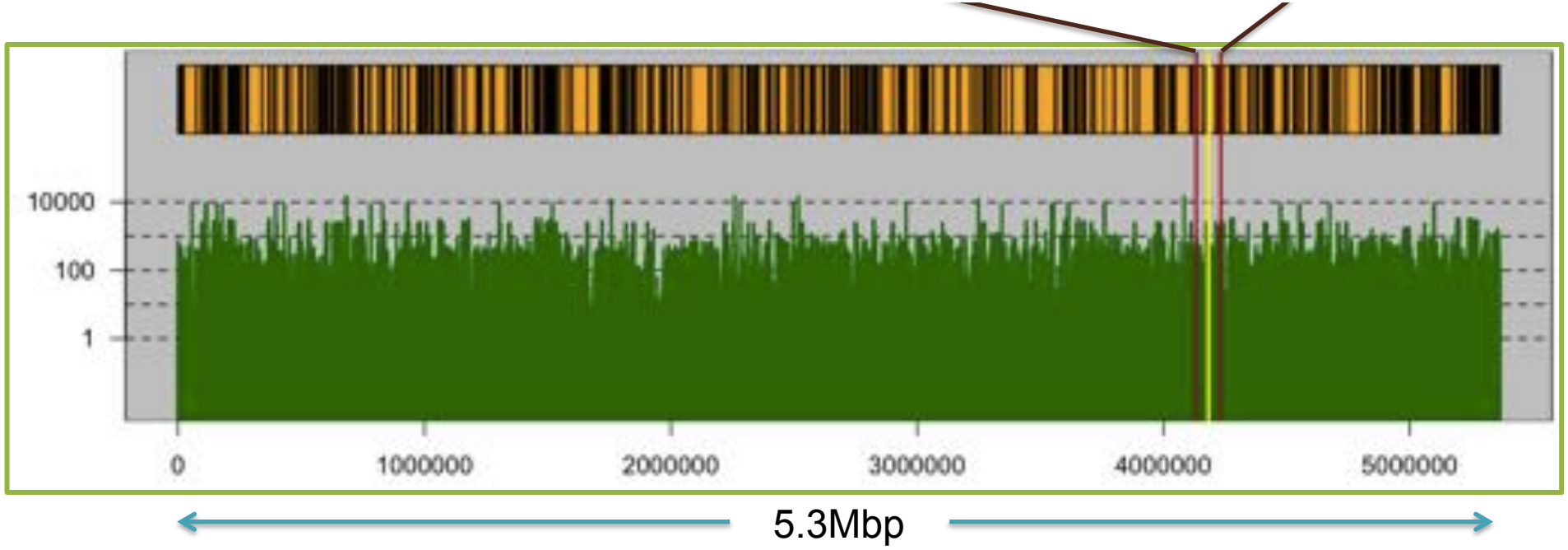
100kbp

S5 Hybrid Sterility Locus

```
Sanger      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
Illumina    …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
PacBio      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
```




10000
100
1
4140000
4160000
4180000
4200000
4220000
10000
100
1
0
1000000
2000000
3000000
4000000
5000000



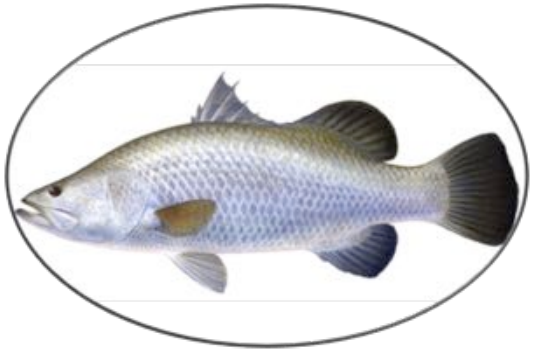
Improvements from 20kbp to 4Mbp contig N50:

- Over 20 Megabases of additional sequence
  - Extremely high sequence identity (>99.9%)
  - Thousands of gaps filled, hundreds of mis-assemblies corrected
- Complete gene models, promoter regions for nearly every gene
  - True representation of transposons and other complex features
- Opportunities for studying large scale chromosome evolution
  - Largest contigs approach complete chromosome arms

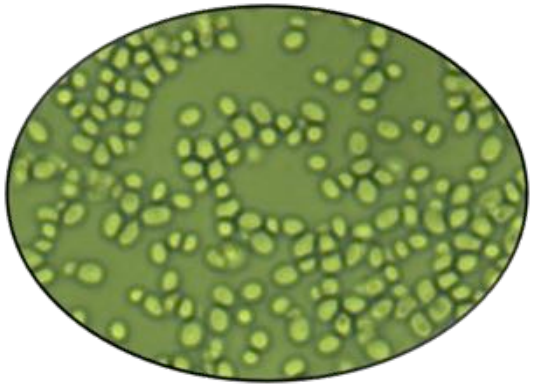
Current Collaborations

Pineapple
UIUC

Human
CSHL/OICR



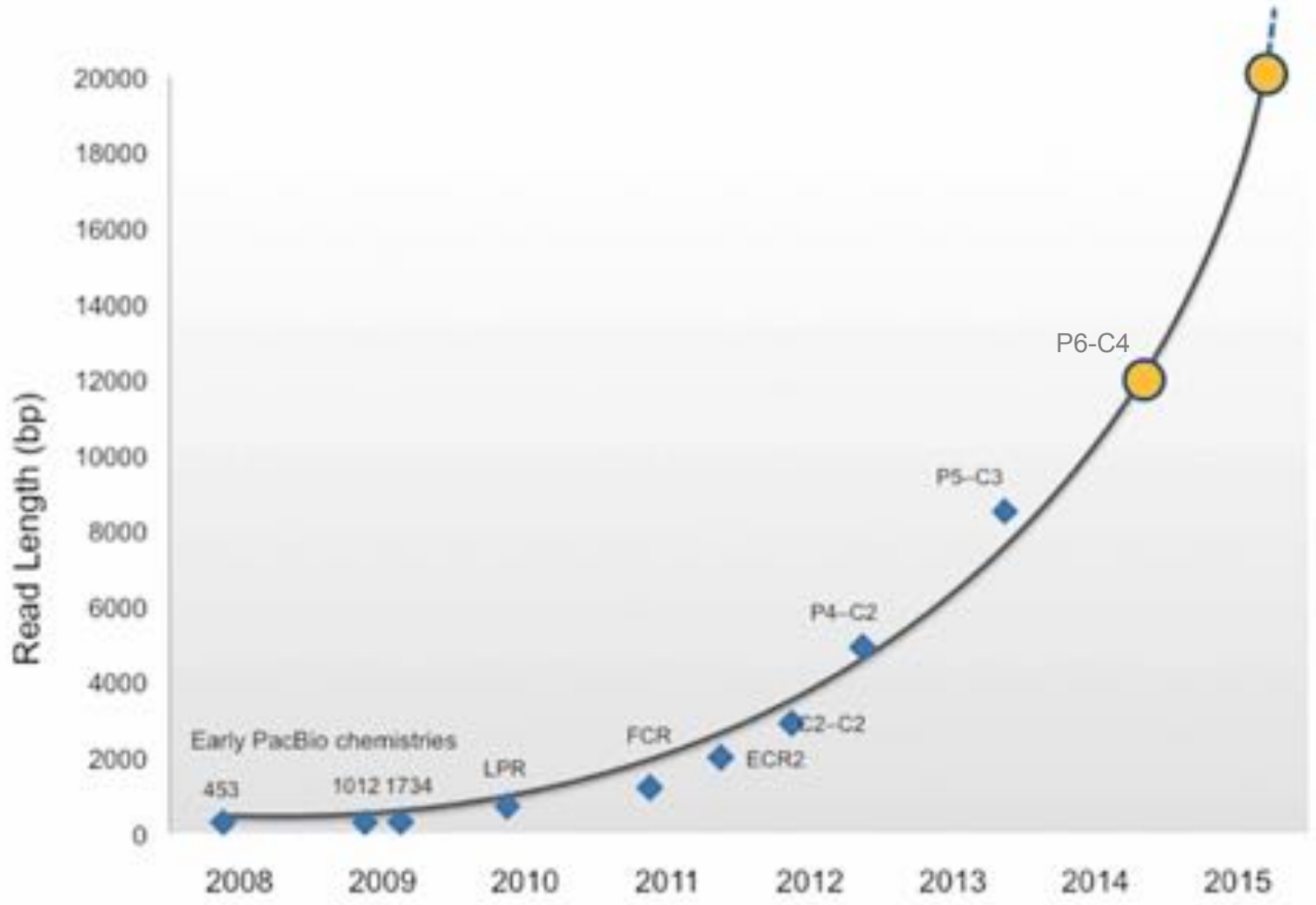
Asian Sea Bass
Temasek Life Sciences



C. glabrata
JHU

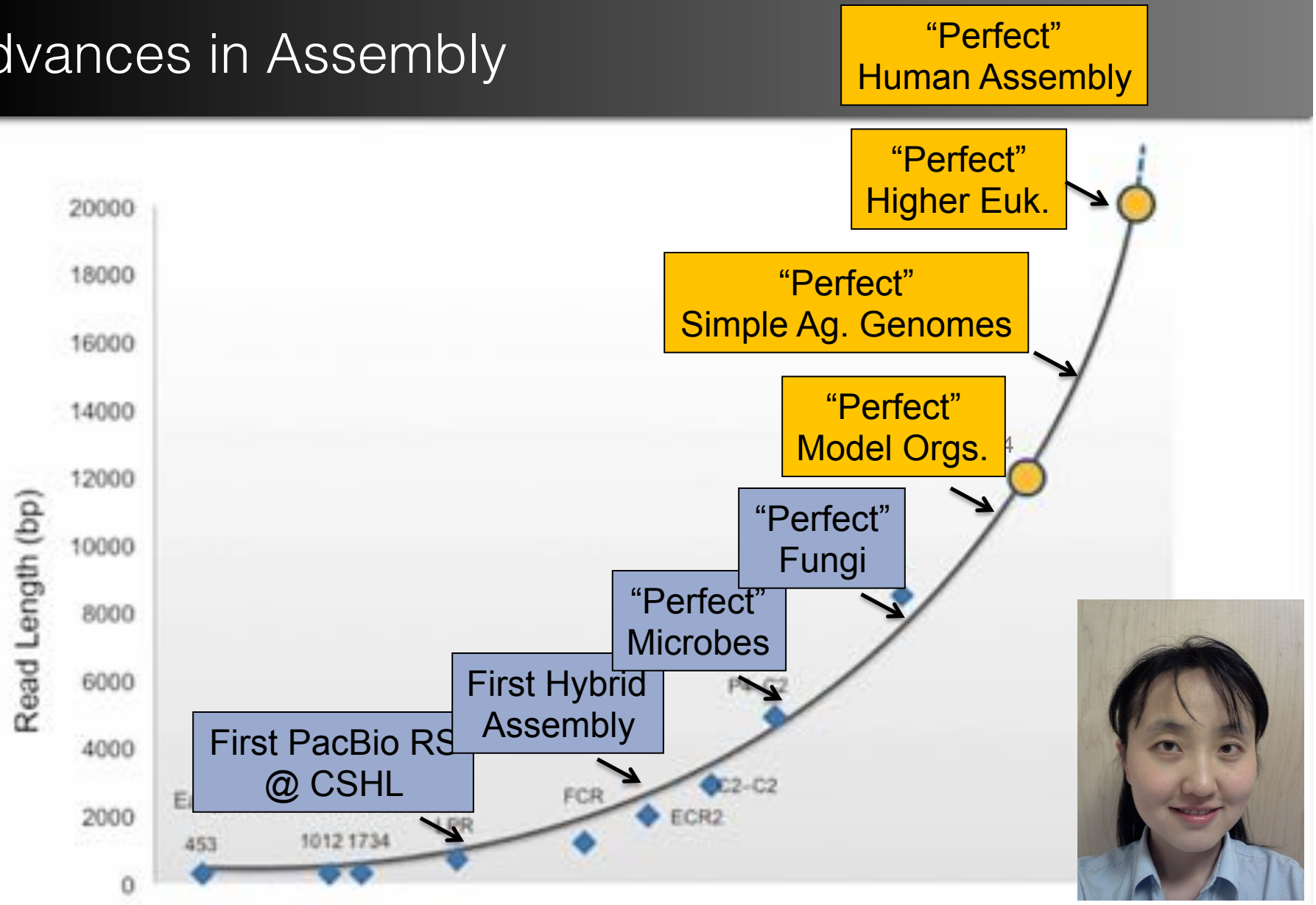
T. vaginalis
NYU

PacBio® Advances in Read Length



PACIFIC BIOSCIENCES

Advances in Assembly



Error correction and assembly complexity of single molecule sequencing reads.
Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz, MC
http://www.biorxiv.org/content/early/2014/06/18/006395

What should we expect from an assembly?

Analysis of dozens of genomes from across the tree of life with real and simulated data

Summary & Recommendations

< 100 Mbp: HGAP/PacBio2CA @ 100x PB C3-P5
expect near perfect chromosome arms

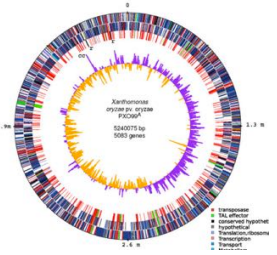
< 1GB: HGAP/PacBio2CA @ 100x PB C3-P5
high quality assembly: contig N50 over 1Mbp

> 1GB: hybrid/gap filling
expect contig N50 to be 100kbp – 1Mbp

> 5GB: Email mschatz@cshl.edu

Error correction and assembly complexity of single molecule sequencing reads.
Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz, MC
http://www.biorxiv.org/content/early/2014/06/18/006395

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

Acknowledgements

Schatz Lab
Rahul Amin
Han Fang
Tyler Gavin
James Gurtowski
Hayan Lee
Zak Lemmon
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

SBU
Skiena Lab
Patro Lab

Cornell
Susan McCouch
Lyza Maron
Mark Wright

OICR
John McPherson
Karen Ng
Timothy Beck
Yogi Sundaravadanam

NYU
Jane Carlton
Elodie Ghedin

NSF

National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

SFARI
SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

Oxford NANOPORE Technologies

PACIFIC BIOSCIENCES®

ALFRED P. SLOAN FOUNDATION

Thank you

http://schatzlab.cshl.edu

@mike_schatz