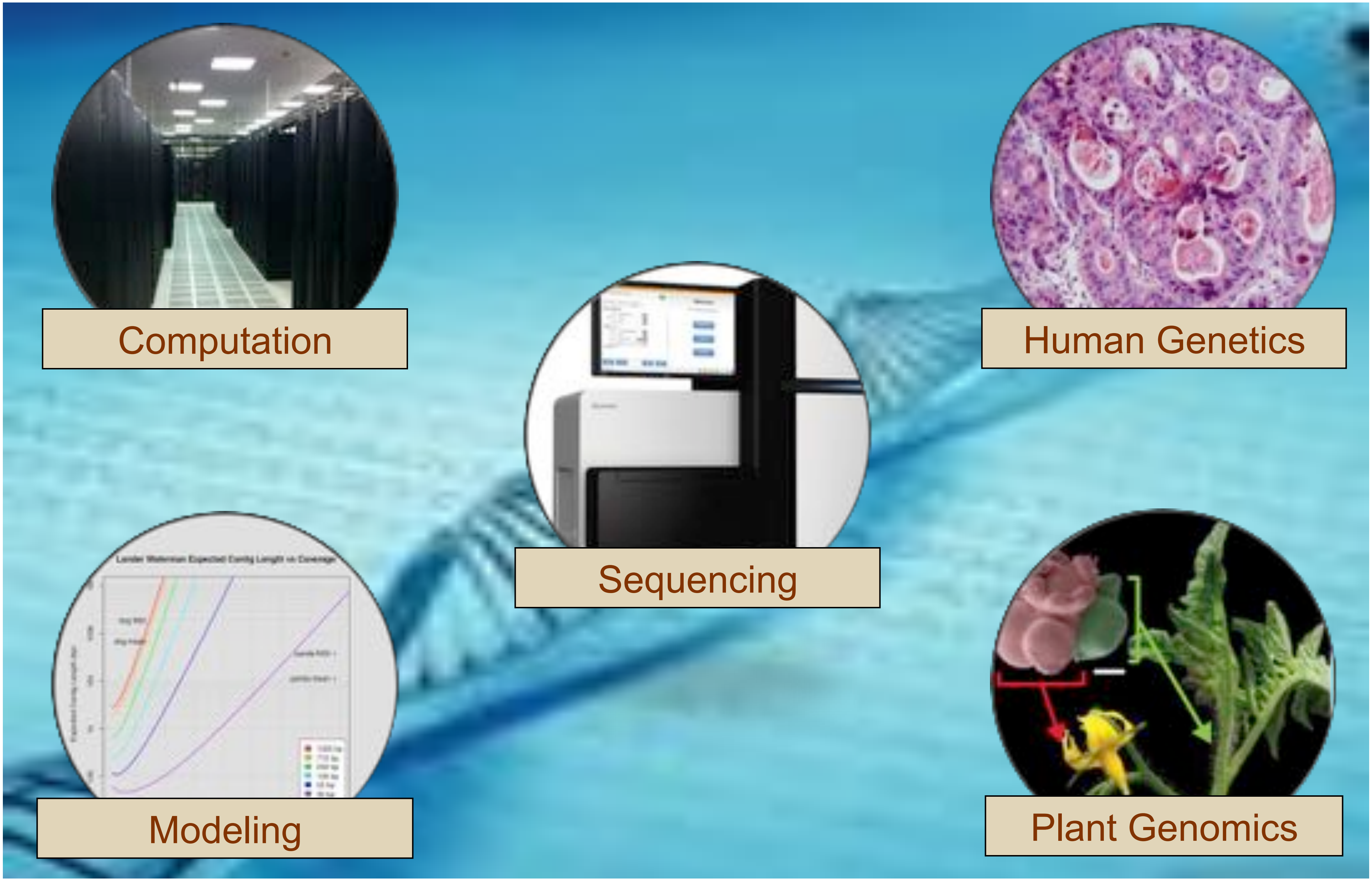
De novo assembly of complex genomes

Michael Schatz

Sept 18, 2012

Statistical Bioinformatics, Purdue University

Schatz Lab Overview



Computation

Human Genetics

Sequencing

Modeling

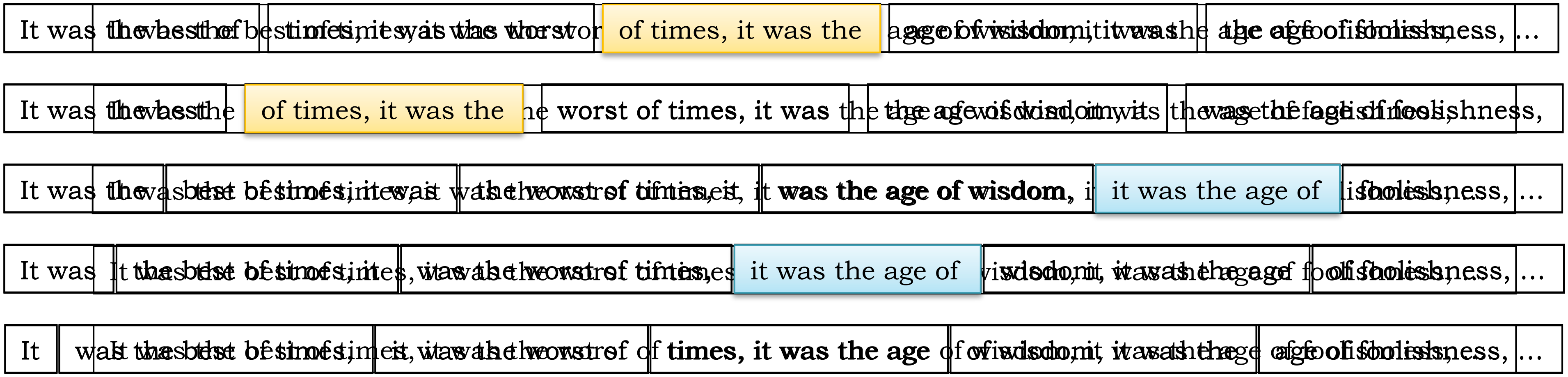
Plant Genomics

Outline

1. Genome assembly by analogy
2. Hybrid error correction and assembly
3. Very recent sequencing results

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,

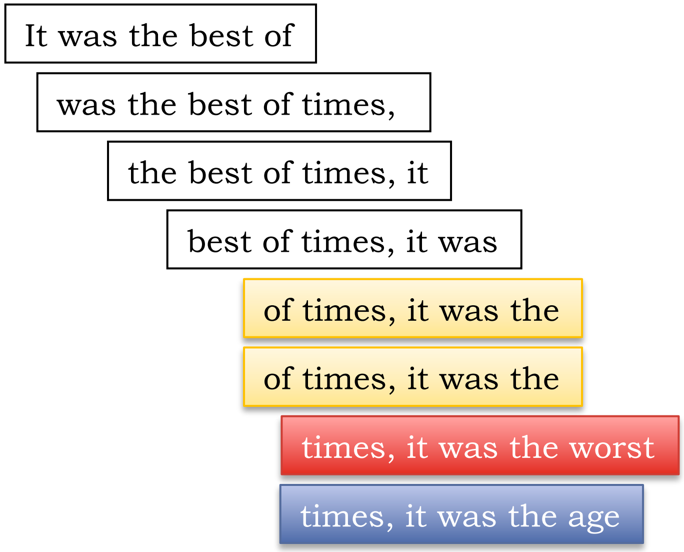
was the age of foolishness,

was the best of times,

was the worst of times,

wisdom, it was the age

worst of times, it was



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]

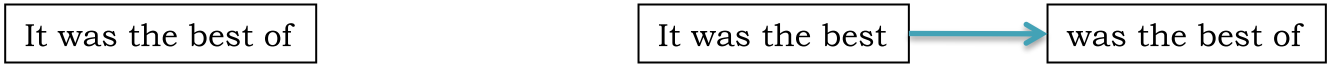
Model sequence reconstruction as a graph problem.

de Bruijn Graph Construction

- $G_k = (V,E)$
  - 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

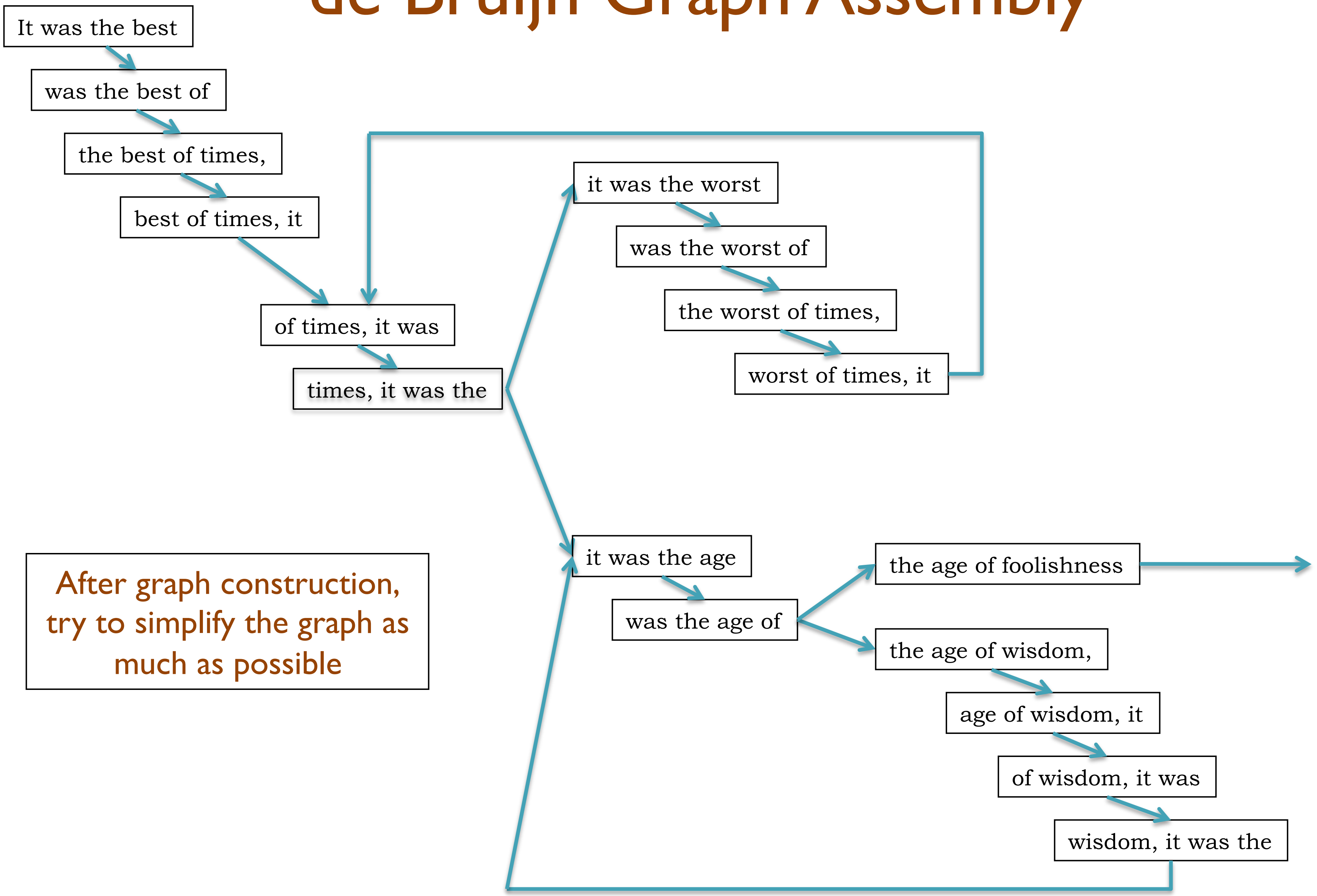
It was the best

was the best of

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

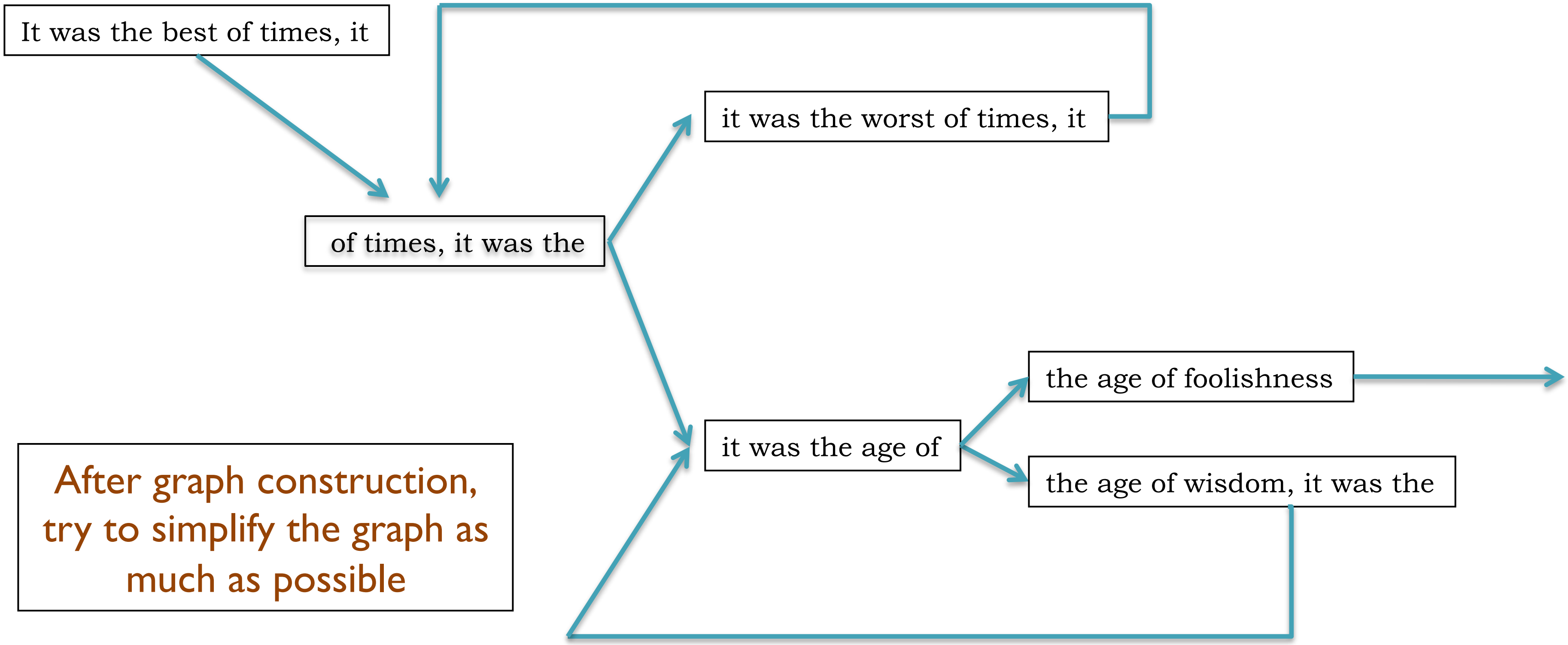
de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001

de Bruijn Graph Assembly



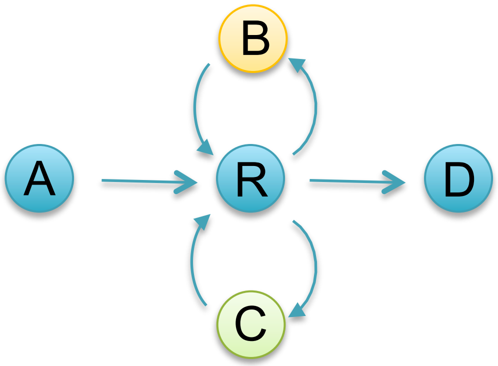
After graph construction, try to simplify the graph as much as possible

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

Counting Eulerian Tours



ARBRCRD
or
ARCRBRD

Generally an exponential number of compatible sequences

– Value computed by application of the BEST theorem (Hutchinson, 1975)

$$\mathcal{W}(G,t) = (\det L)\left\{\prod_{u\in V}(r_u-1)!\right\}\left\{\prod_{(u,v)\in E} a_{uv}!\right\}^{-1}$$

L = $n \times n$ matrix with r_u-a_{uu} along the diagonal and -a_{uv} in entry uv

$r_u = d^+(u)+1$ if u=t, or $d^+(u)$ otherwise

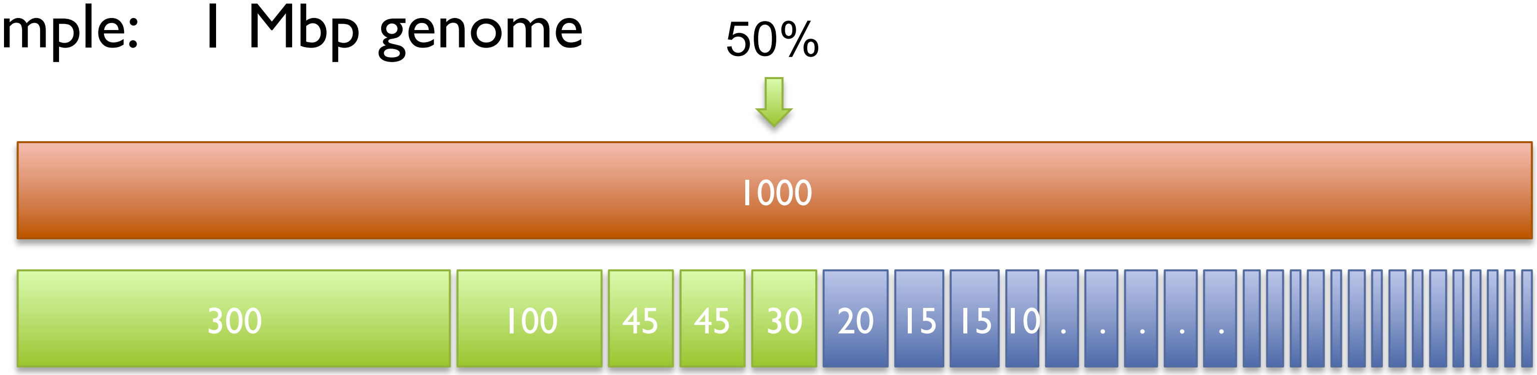
a_{uv} = multiplicity of edge from u to v

Assembly Complexity of Prokaryotic Genomes using Short Reads.
Kingsford C, Schatz MC, Pop M (2010) *BMC Bioinformatics*.

N50 size

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

Example: 1 Mbp genome



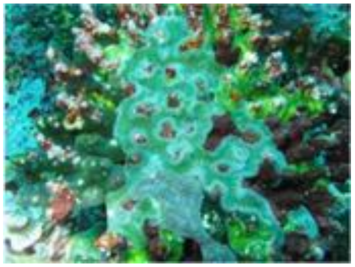
N50 size = 30 kbp
(300k+100k+45k+45k+30k = 520k >= 500kbp)

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

Assembly Applications

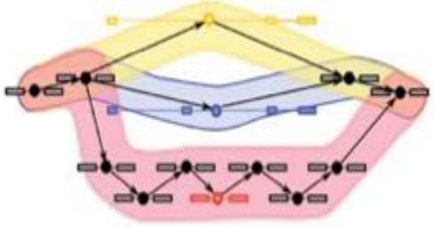
Novel genomes

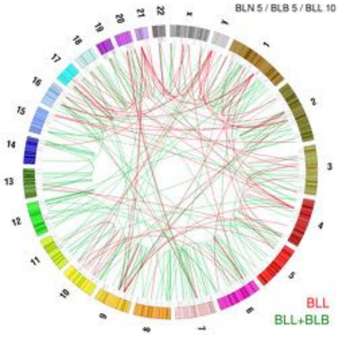
Metagenomes



Sequencing assays

- Transcript assembly
- Structural variations
- Haplotype analysis
- …





Why are genomes hard to assemble?

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

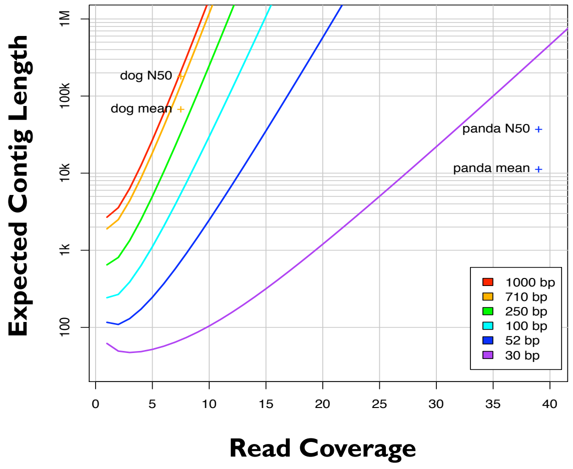
2. ***Sequencing***:
   - (Very) large genomes, imperfect sequencing

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

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

Ingredients for a good assembly

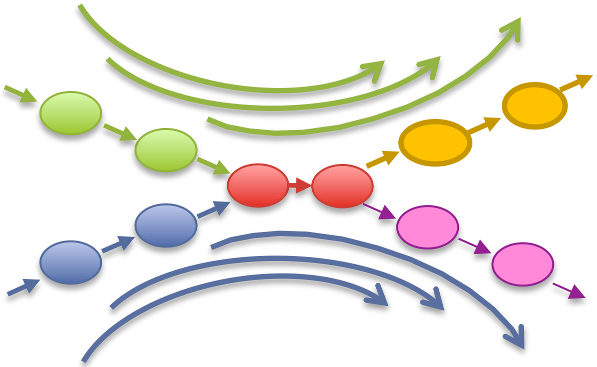
Coverage



High coverage is required

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

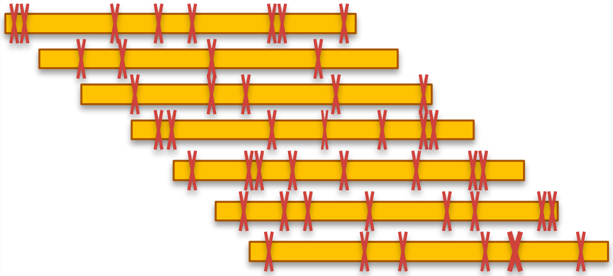
Read Length



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

Hybrid Sequencing

Illumina
Sequencing by Synthesis

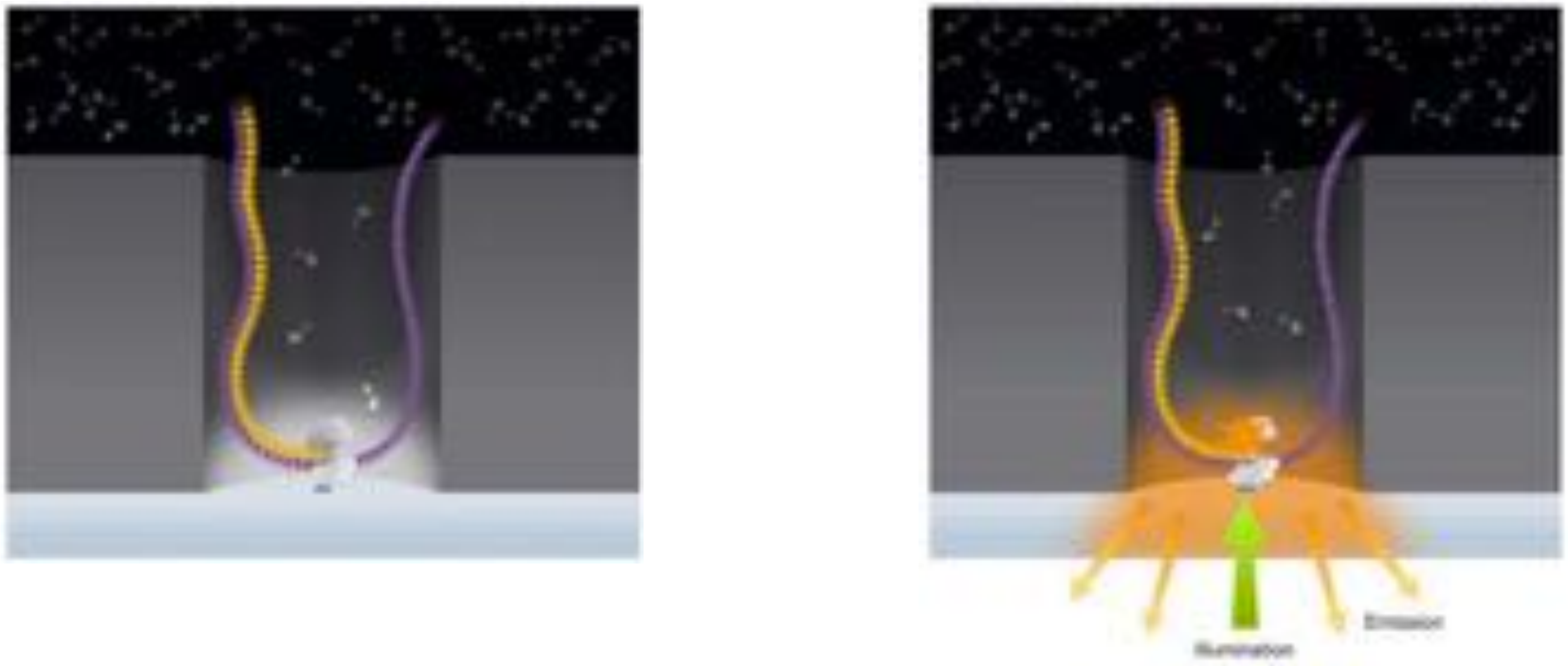
High throughput (60Gbp/day)
High accuracy (~99%)
Short reads (~100bp)

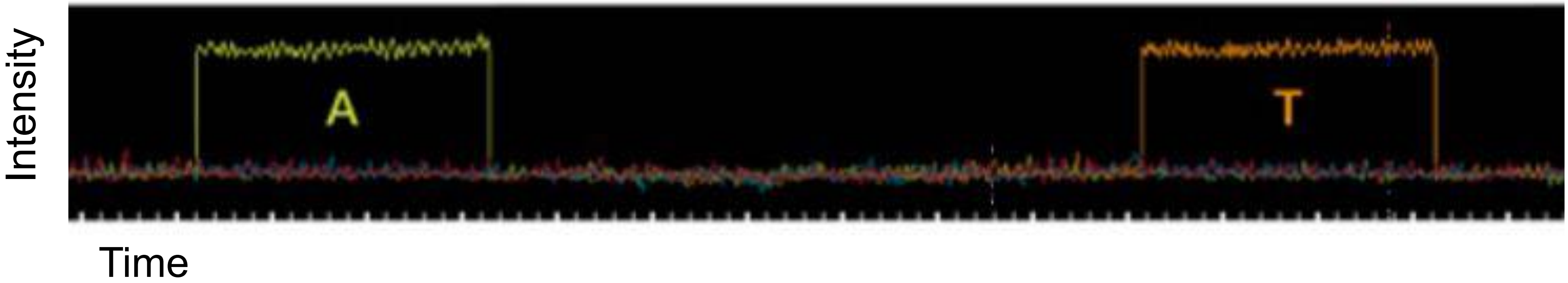
Pacific Biosciences
SMRT Sequencing

Lower throughput (600Mbp/day)
Lower accuracy (~85%)
Long reads (1-2kbp+)

SMRT Sequencing

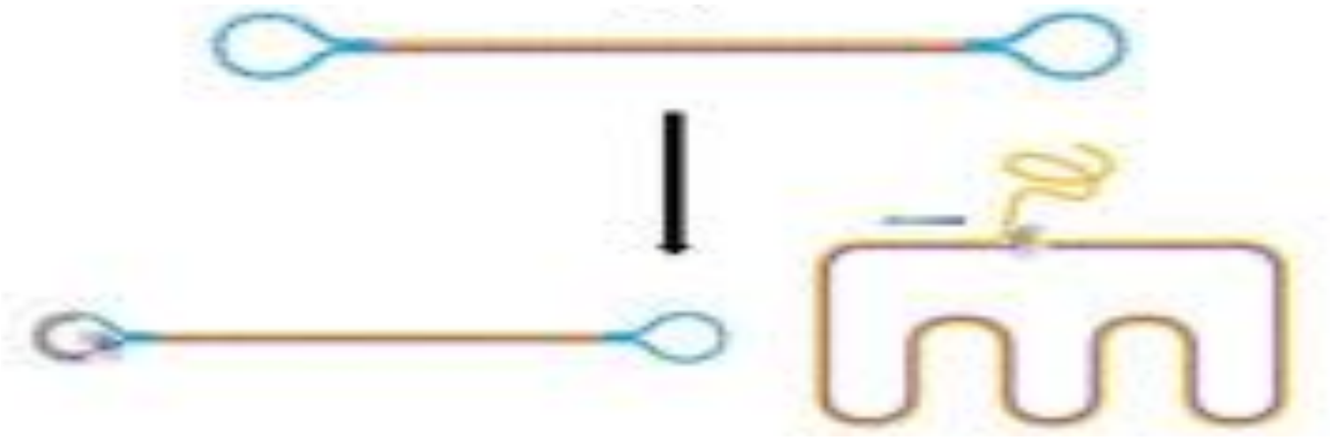
Imaging of florescent phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).





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

SMRT Read Types



- ***Standard sequencing***
  - Long inserts so that the polymerase can synthesize along a single strand

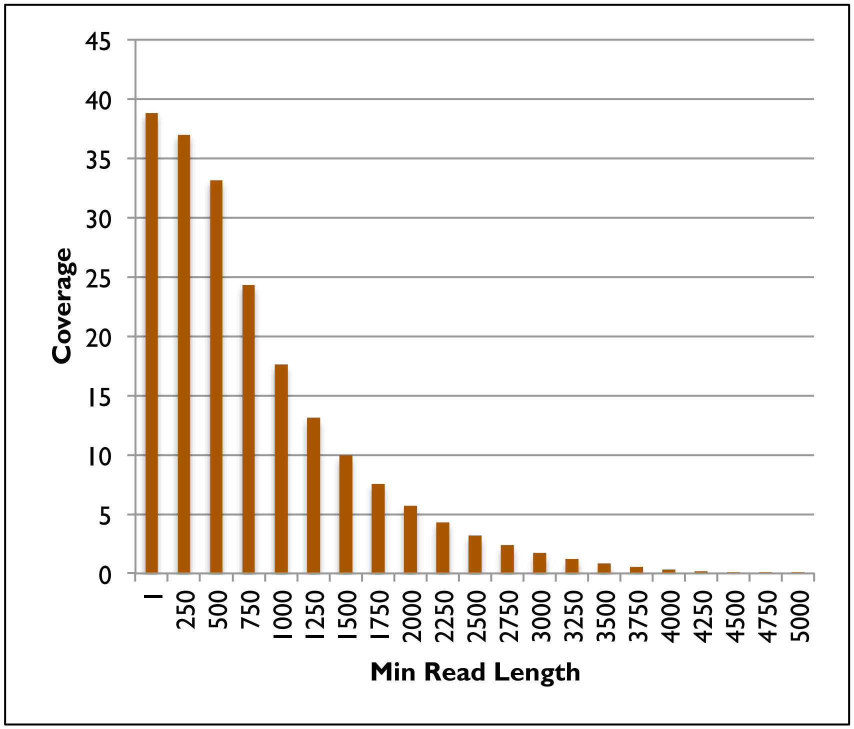
- ***Circular consensus sequencing***
  - Short inserts, so polymerase can continue around the entire SMRTbell multiple times and generate multiple sub-reads from the same single molecule.

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

SMRT Sequencing Data

Yeast
(Pre-release Chemistry / 2010)

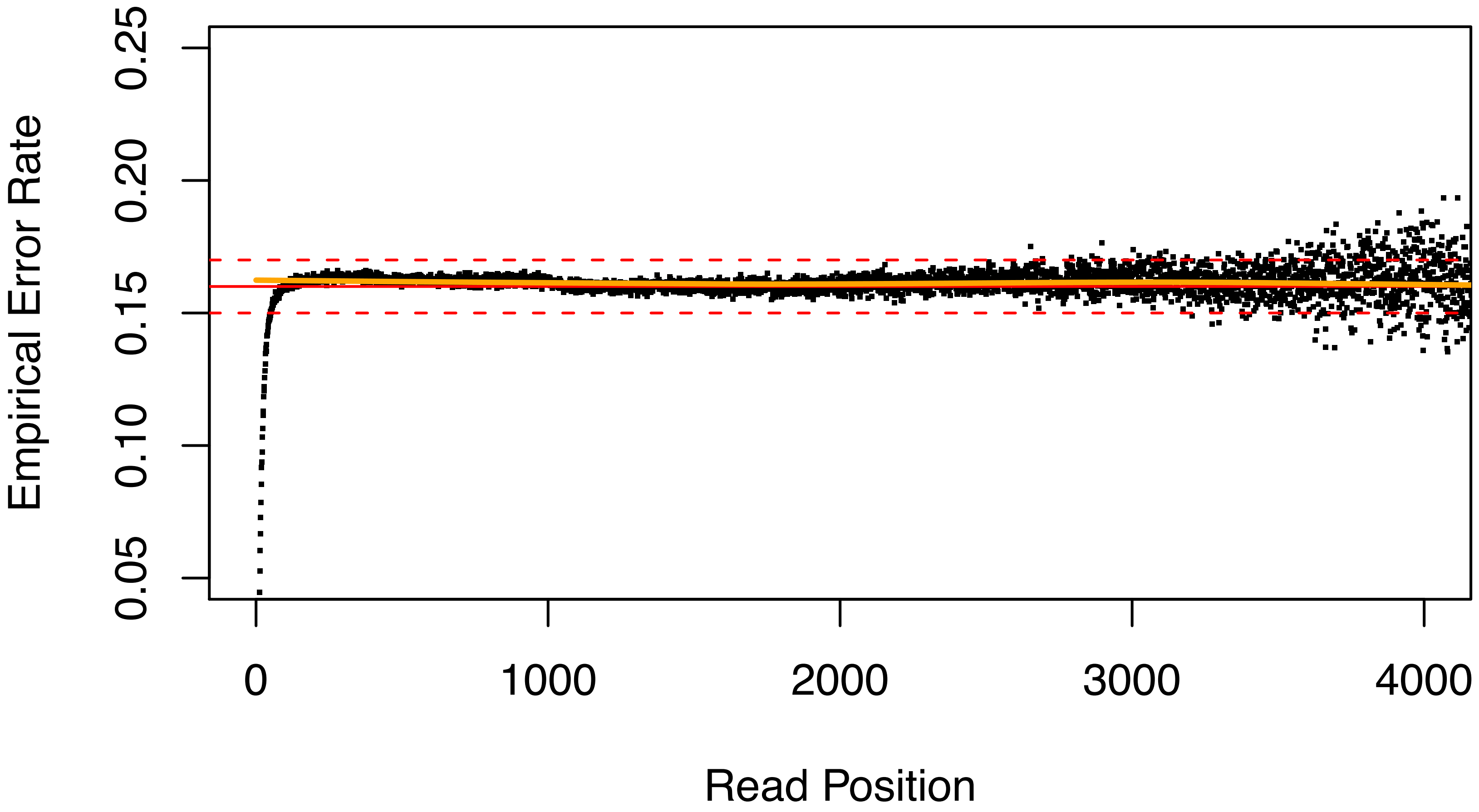
65 SMRT cells
734,151 reads after filtering
Mean: 642.3 +/- 587.3
Median: 553 Max: 8,495



```
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
Average overall accuracy: 83.7%, 11.5% insertions, 3.4% deletions, 1.4% mismatch

Read Quality

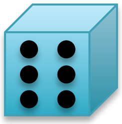
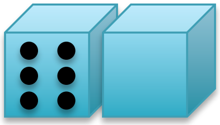
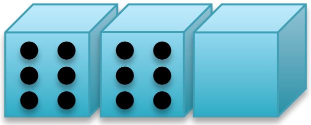
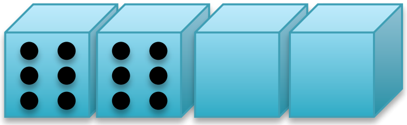
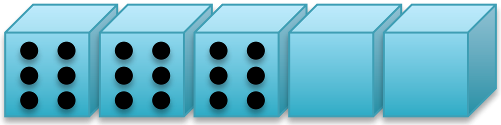


Consistent quality across the entire read

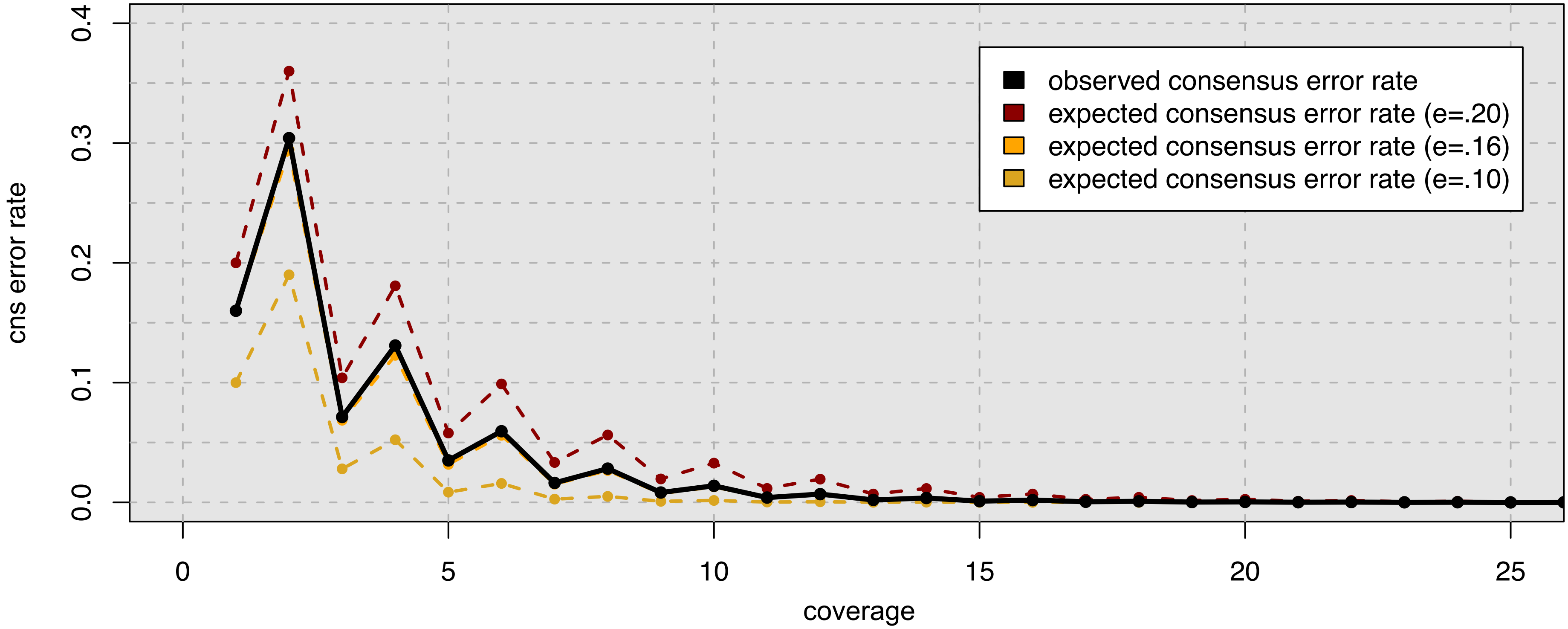
- Uniform error rate, no apparent biases for GC/motifs
- Sampling artifacts at beginning and ends of alignments

Consensus Quality: Probability Review

Roll *n* dice => What is the probability that at least half are 6's

n	*Min to Lose*	*Losing Events*	*P(Lose)*
1		*1/6*	*16.7%*
2		*P(1of 2) + P(2 of 2)*	*30.5%*
3		*P(2 of 3) + P(3 of 3)*	*7.4%*
4		*P(2 of 4) + P(3 of 4) + P(4 of 4)*	*13.2%*
5		*P(3 of 5) + P(4 of 5) + P(5 of 5)*	*3.5%*
n	*ceil(n/2)*	$\sum_{i=\lceil n/2\rceil}^{n} P(i\ \ of\ \ n) = \sum_{i=\lceil n/2\rceil}^{n}\binom{n}{i}(p)^{i}(1-p)^{n-i}$	

Consensus Accuracy and Coverage



Coverage can overcome random errors

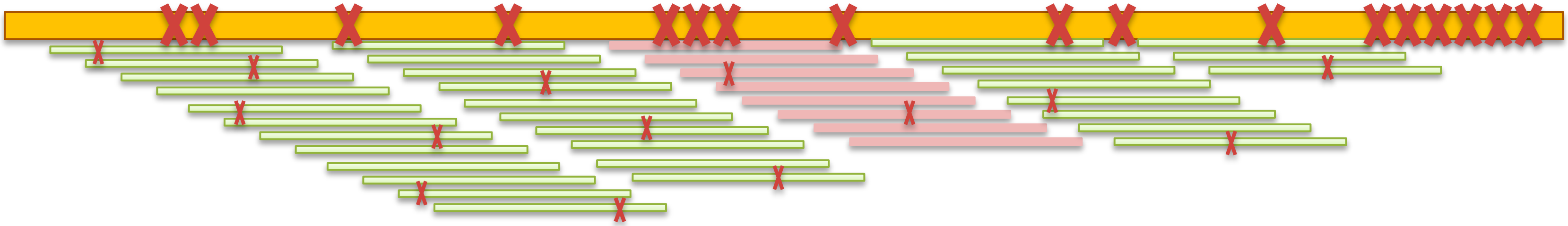
- Dashed: error model from binomial sampling; solid: observed accuracy
- For same reason, CCS is extremely accurate when using 5+ subreads

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

PacBio Error Correction

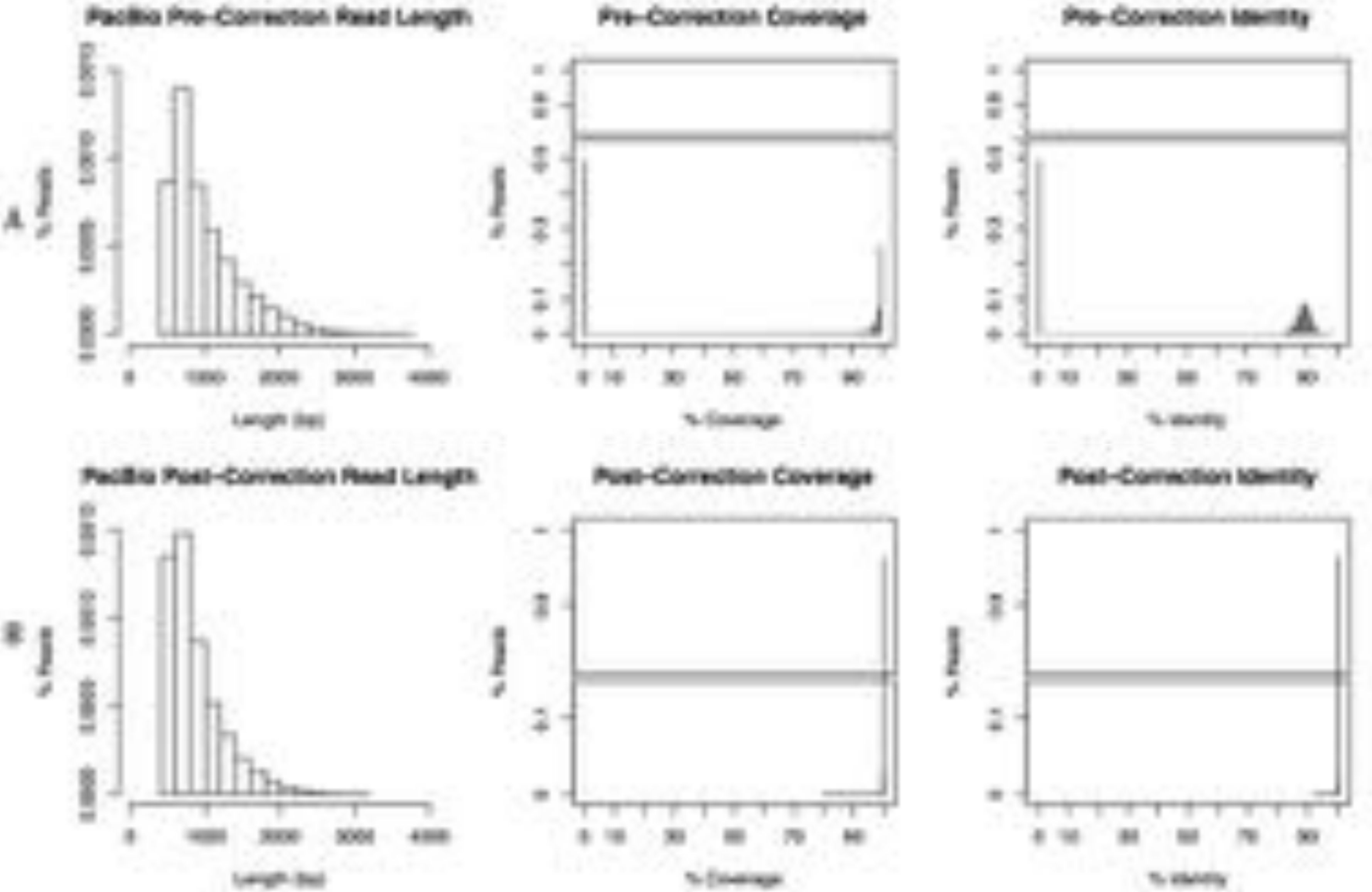
http://wgs-assembler.sf.net

1. Correction Pipeline
   1. Map short reads (SR) to long reads (LR)
   2. Trim LRs at coverage gaps
   3. Compute consensus for each LR
2. Error corrected reads can be easily assembled, aligned



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

Error Correction Results

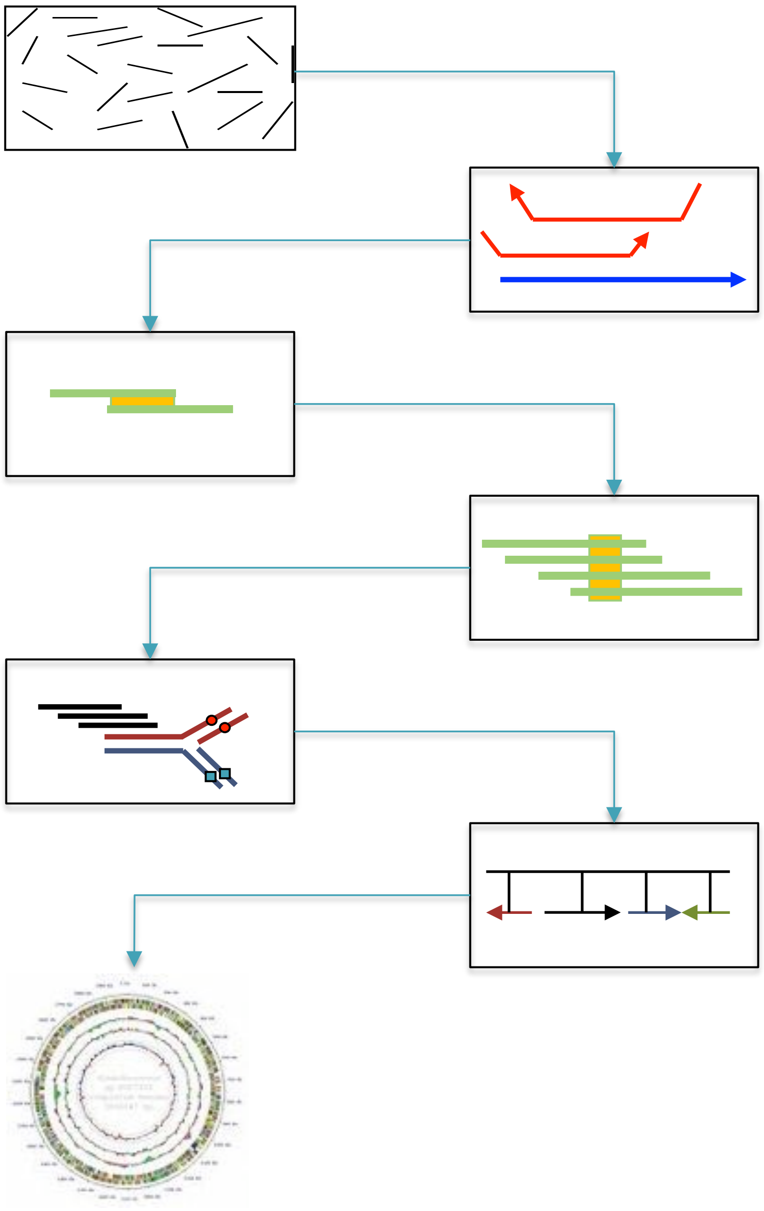


Correction results of 20x PacBio coverage of E. coli K12 corrected using 50x Illumina

Celera Assembler

http://wgs-assembler.sf.net

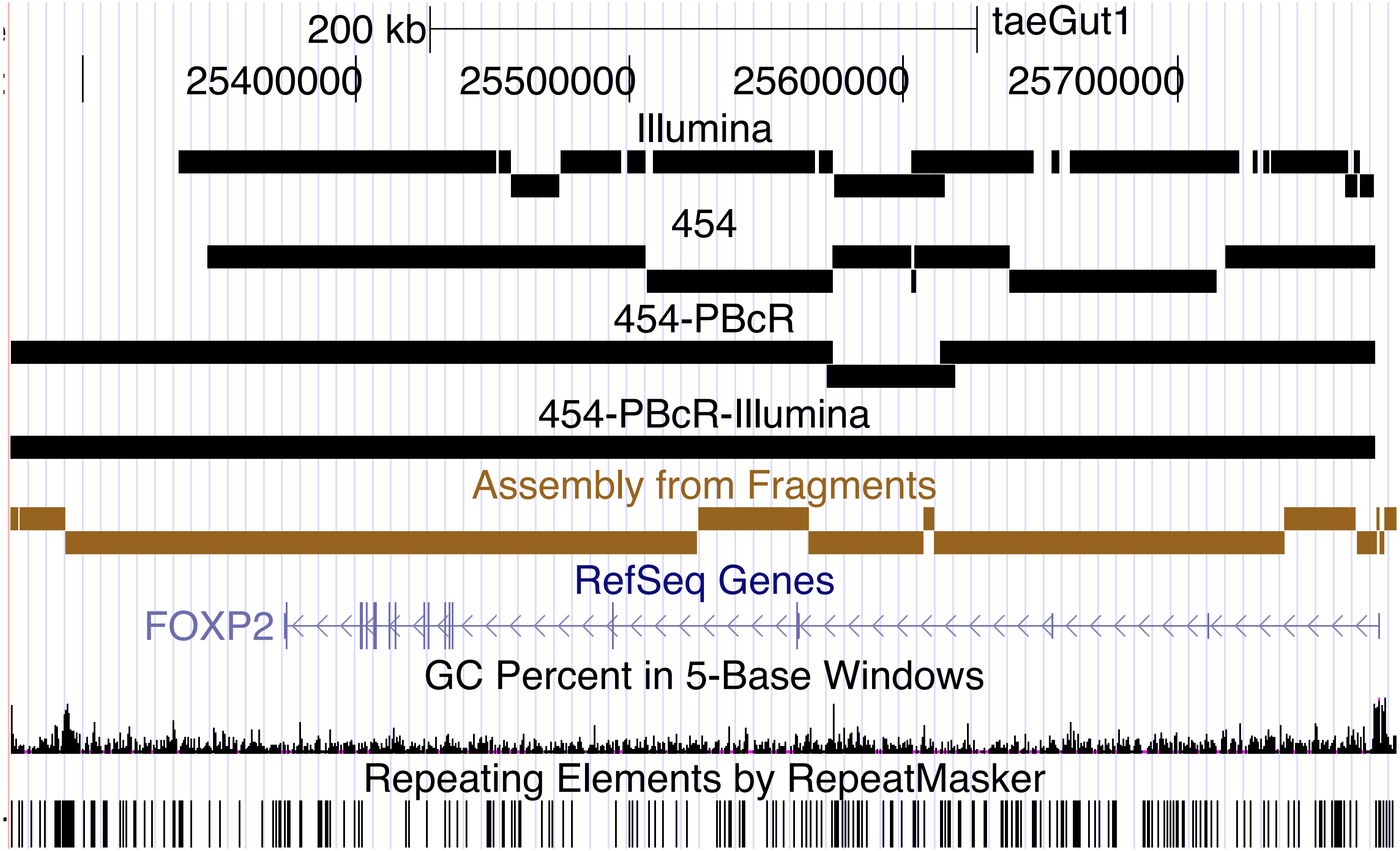
1. Pre-overlap
   - Consistency checks
2. Trimming
   - Quality trimming & partial overlaps
3. Compute Overlaps
   - Find high quality overlaps
4. Error Correction
   - Evaluate difference in context of overlapping reads
5. Unitigging
   - Merge consistent reads
6. Scaffolding
   - Bundle mates, Order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



SMRT-Assembly Results

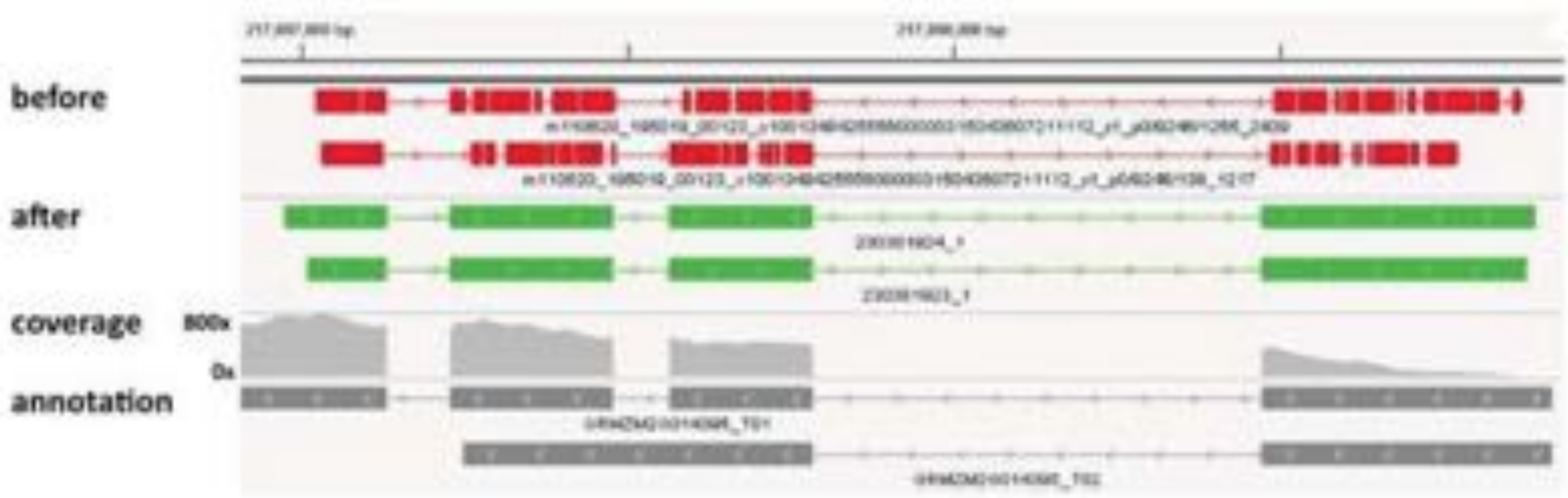
Hybrid assembly results using error corrected PacBio reads
Meets or beats Illumina-only or 454-only assembly in every case

Improved Gene Reconstruction



FOXP2 assembled on a single contig

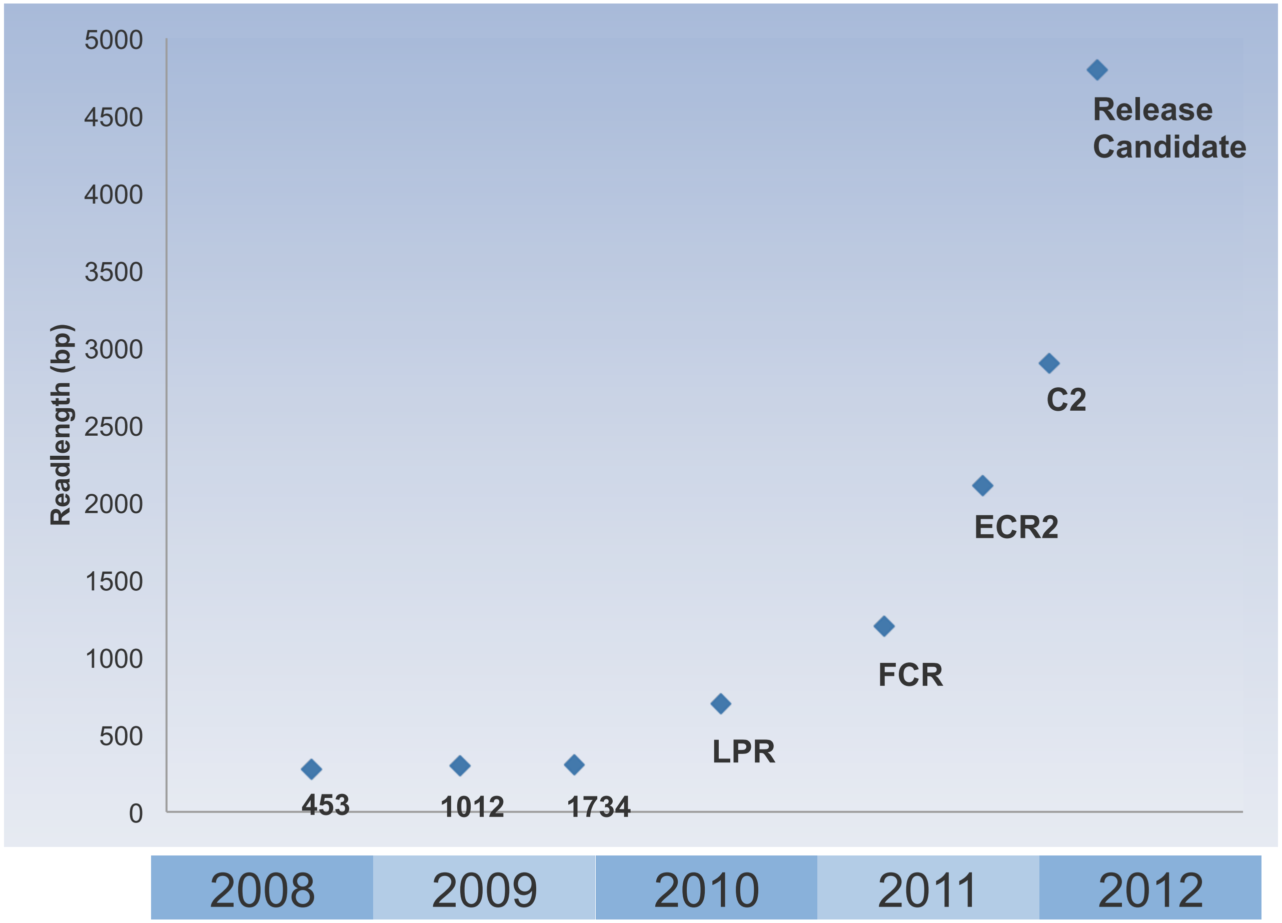
Transcript Alignment



- Long-read single-molecule sequencing has potential to directly sequence full length transcripts
  - Raw reads and raw alignments (red) have many spurious indels inducing false frameshifts and other artifacts
  - Error corrected reads almost perfectly match the genome, pinpointing splice sites, identifying alternative splicing

- New collaboration with Gingeras Lab looking at splicing in human

PacBio Technology Roadmap



Internal Roadmap has made steady progress towards improving read length and throughput

Very recent improvements:

1. Improved enzyme:

   Maintains reactions longer

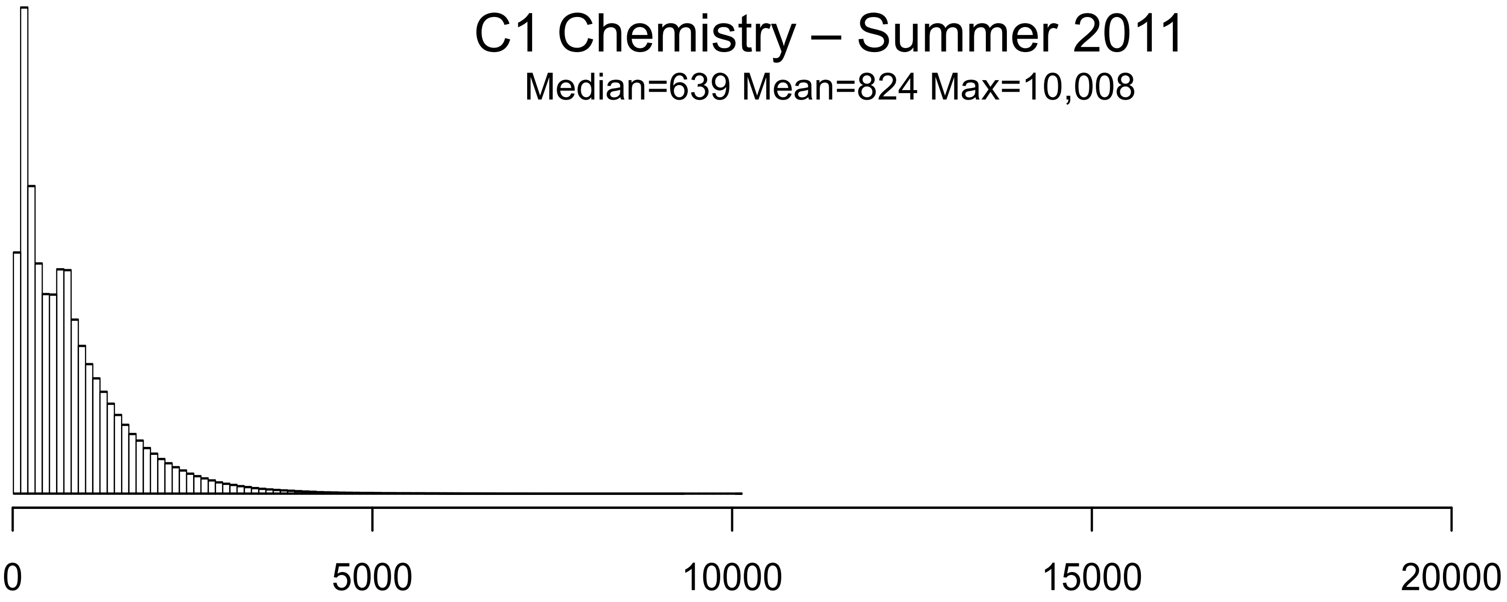
2. “Hot Start” technology:

   Maximize subreads

3. MagBead loading:

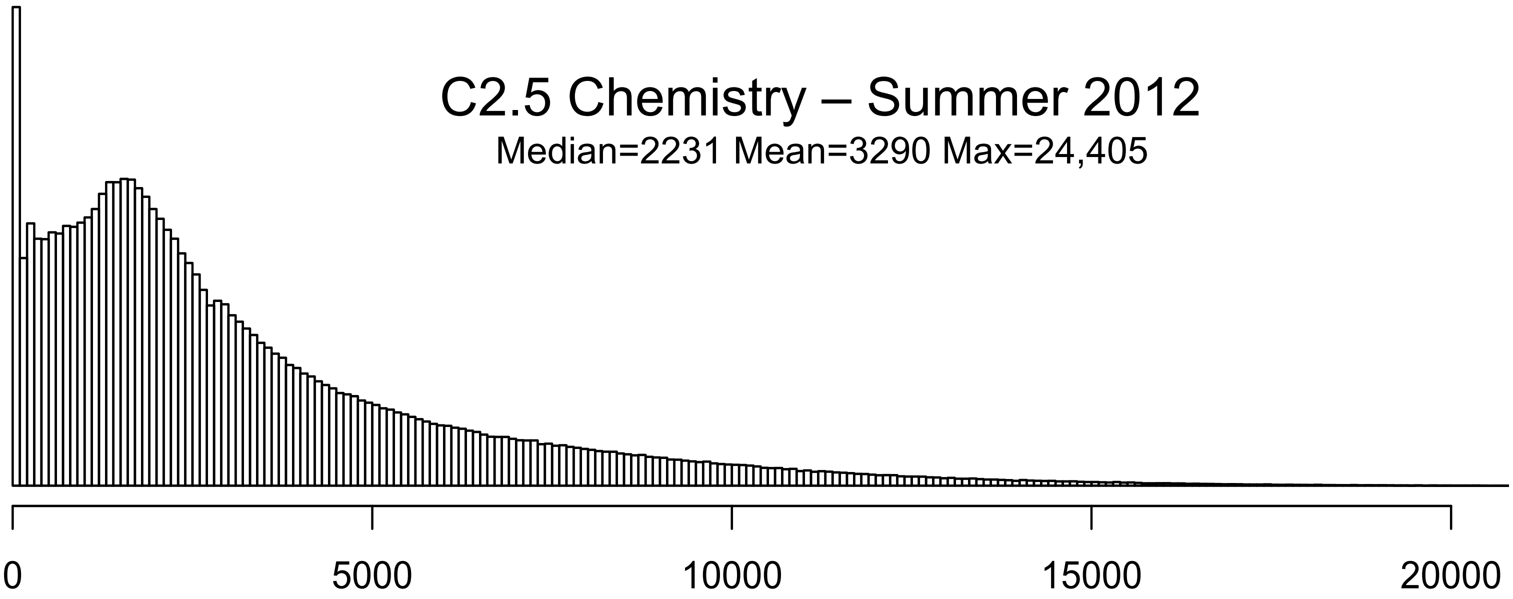
   Load longest fragments

PacBio Rice Sequencing



C1 Chemistry – Summer 2011

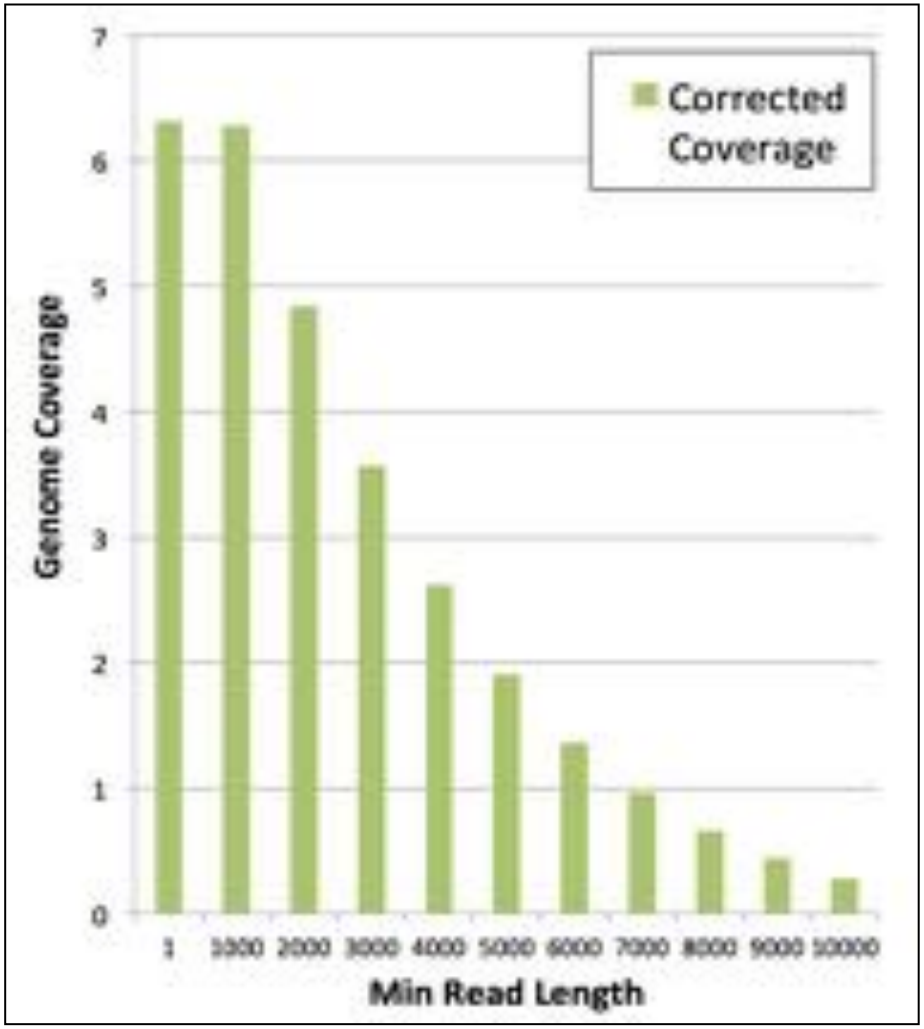
Median=639 Mean=824 Max=10,008



C2.5 Chemistry – Summer 2012

Median=2231 Mean=3290 Max=24,405

Preliminary Rice Assemblies



Assembly	Contig N50
Illumina Fragments 50x 2x100bp @ 180	3925
Illumina Mates 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	13696
MiSeq Fragments 23x 459bp 8x 2x251bp @ 450	6444
PBeCR Reads 6.3x 2146bp ** MiSeq for correction	13600
PBeCR + Mates 6.3x 2146bp ** MiSeq for correction 51x 2x50bp @ 4800	In Progress

In collaboration with McCombie & Ware labs @ CSHL

Single Molecule Sequencing Summary

PacBio RS has capabilities not found in any other technology

- Substantially longer reads -> span repeats
- Unbiased sequence coverage -> close sequencing gaps
- Single molecule sequencing -> haplotype phasing, alternative splicing

Long reads enables highest quality de novo assembly

- Longer reads have more information than shorter reads
- Because the errors are random we can compensate for them
- One chromosome, one contig achieved in microbes

Exciting developments on the horizon

- Longer reads, higher throughput PacBio
- Nanopore Sequencing

Acknowledgements

Schatz Lab

Giuseppe Narzisi
Shoshana Marcus
Rob Aboukhalil
Mitch Bekritsky
Charles Underwood
James Gurtowski
Alejandro Wences

Hayan Lee
Rushil Gupta
Avijit Gupta
Shishir Horane
Deepak Nettem
Varrun Ramani
Eric Biggers

CSHL

Hannon Lab
Iossifov Lab
Levy Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Ware Lab
Wigler Lab

NBACC

Adam Phillippy
Sergey Koren

JHU/UMD

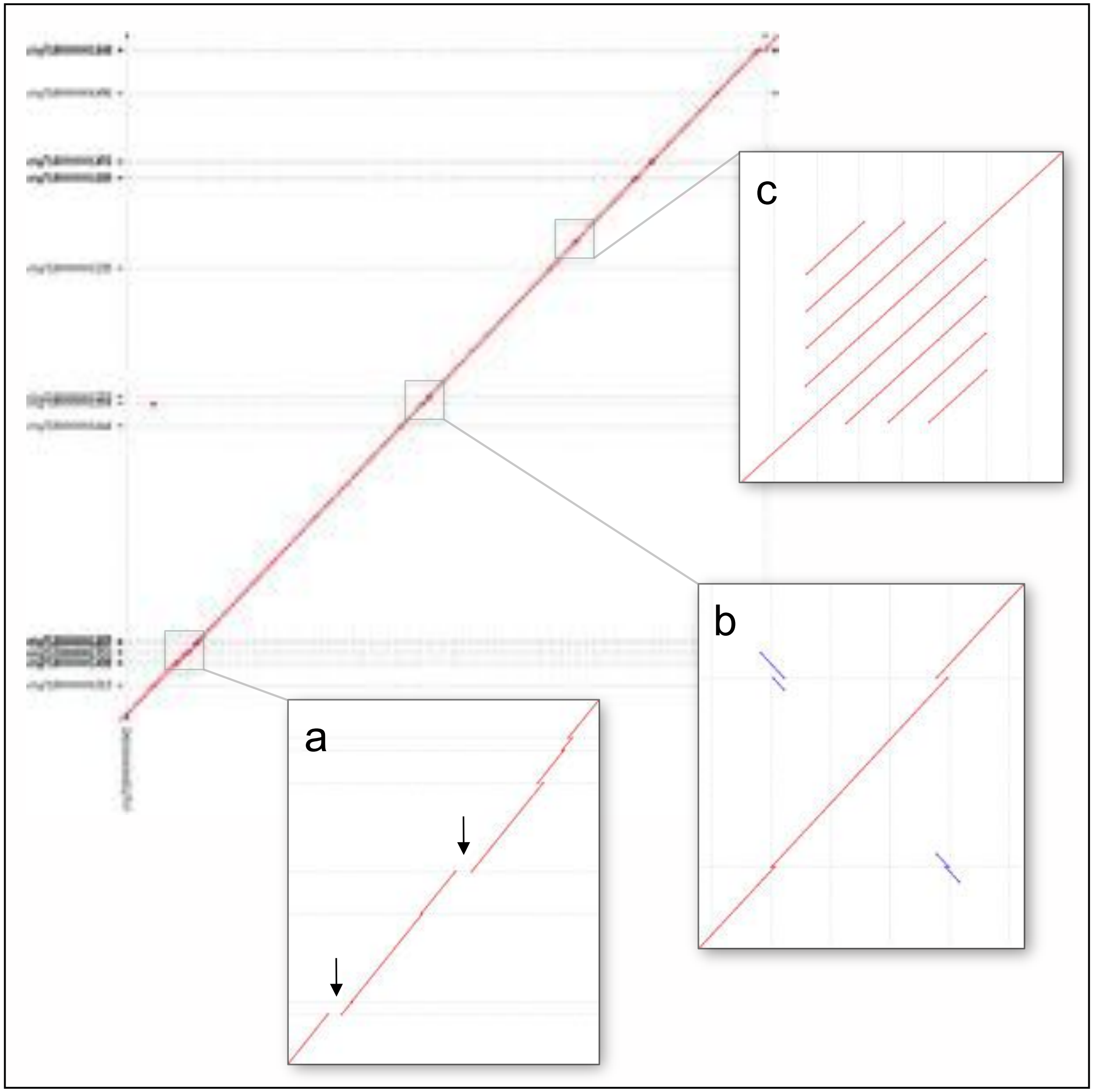
Steven Salzberg
Mihai Pop
Ben Langmead
Cole Trapnell

Thank You!

Want to push the frontier of bioinformatics, biotechnology, & genetics?

http://schatzlab.cshl.edu/apply/

Long Read Advantages

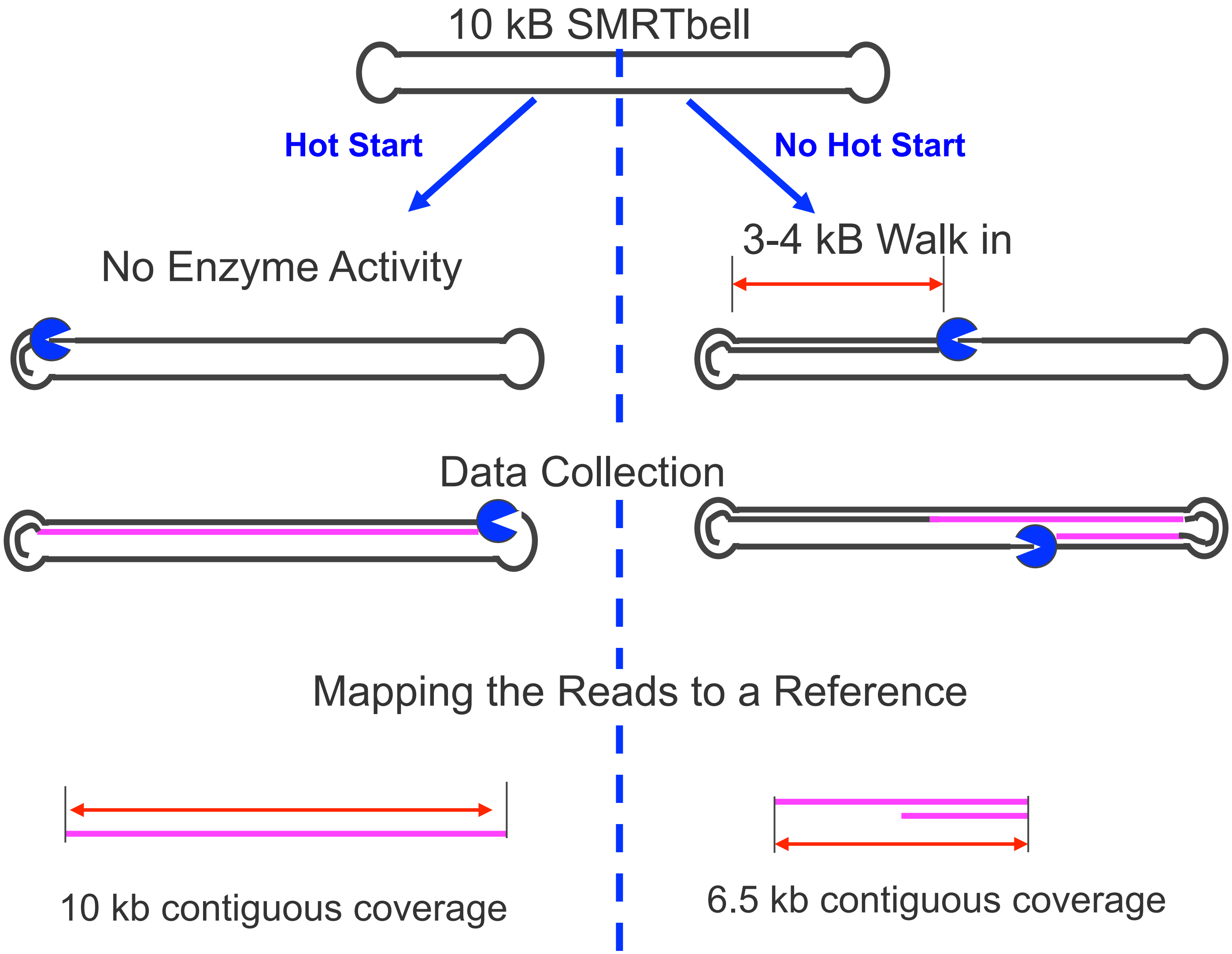


(a) Long reads close sequencing gaps

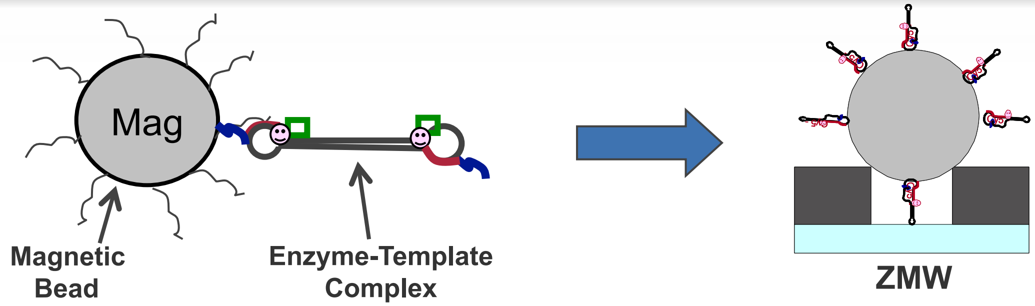
(b) Long reads assemble across long repeats

(c) Long reads span complex microsatellites

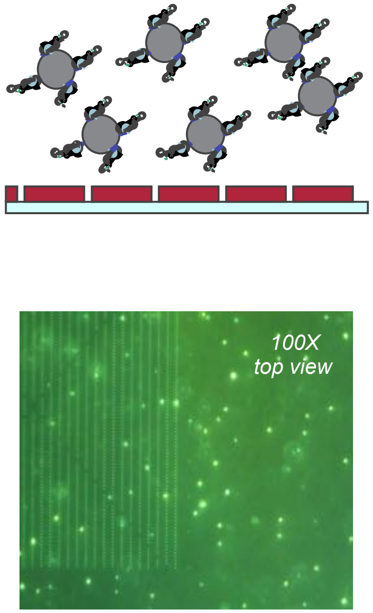
Theoretical Benefits of Hot Start Sequencing


10 kB SMRTbell
Hot Start
No Hot Start
No Enzyme Activity
3-4 kB Walk in
Data Collection
Mapping the Reads to a Reference
10 kb contiguous coverage
6.5 kb contiguous coverage

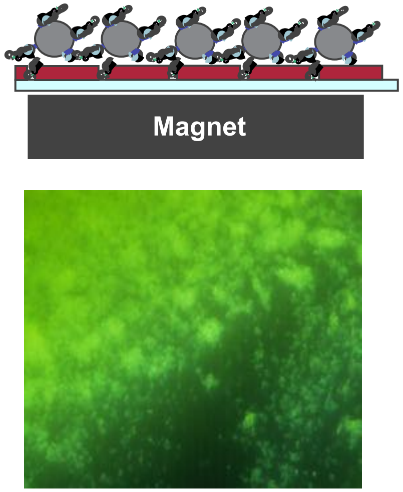
Magnetic Bead Enzyme-Template Complex Loading



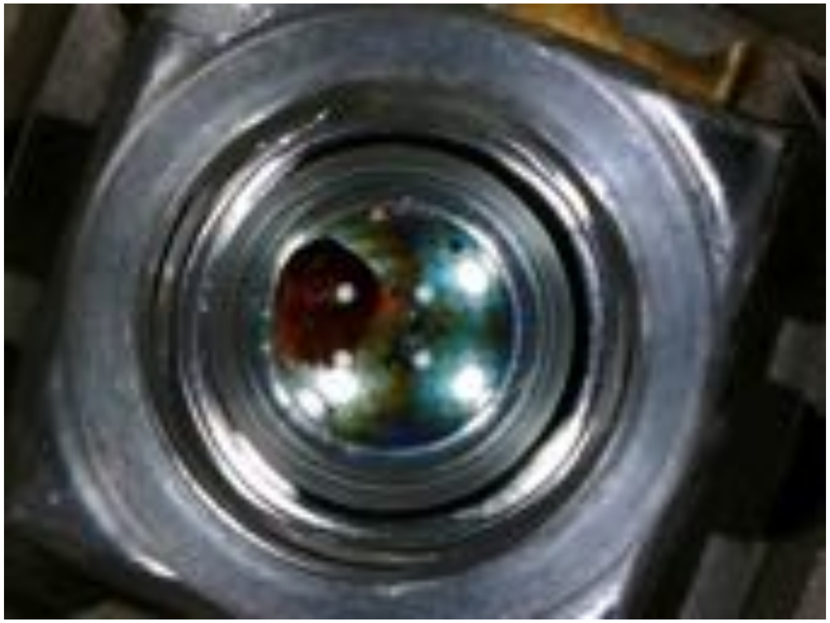
Multiple complexes attached to magnetic beads that are much larger than individual ZMWs



(I)
Pre-Deposition: Complex loaded beads in solution

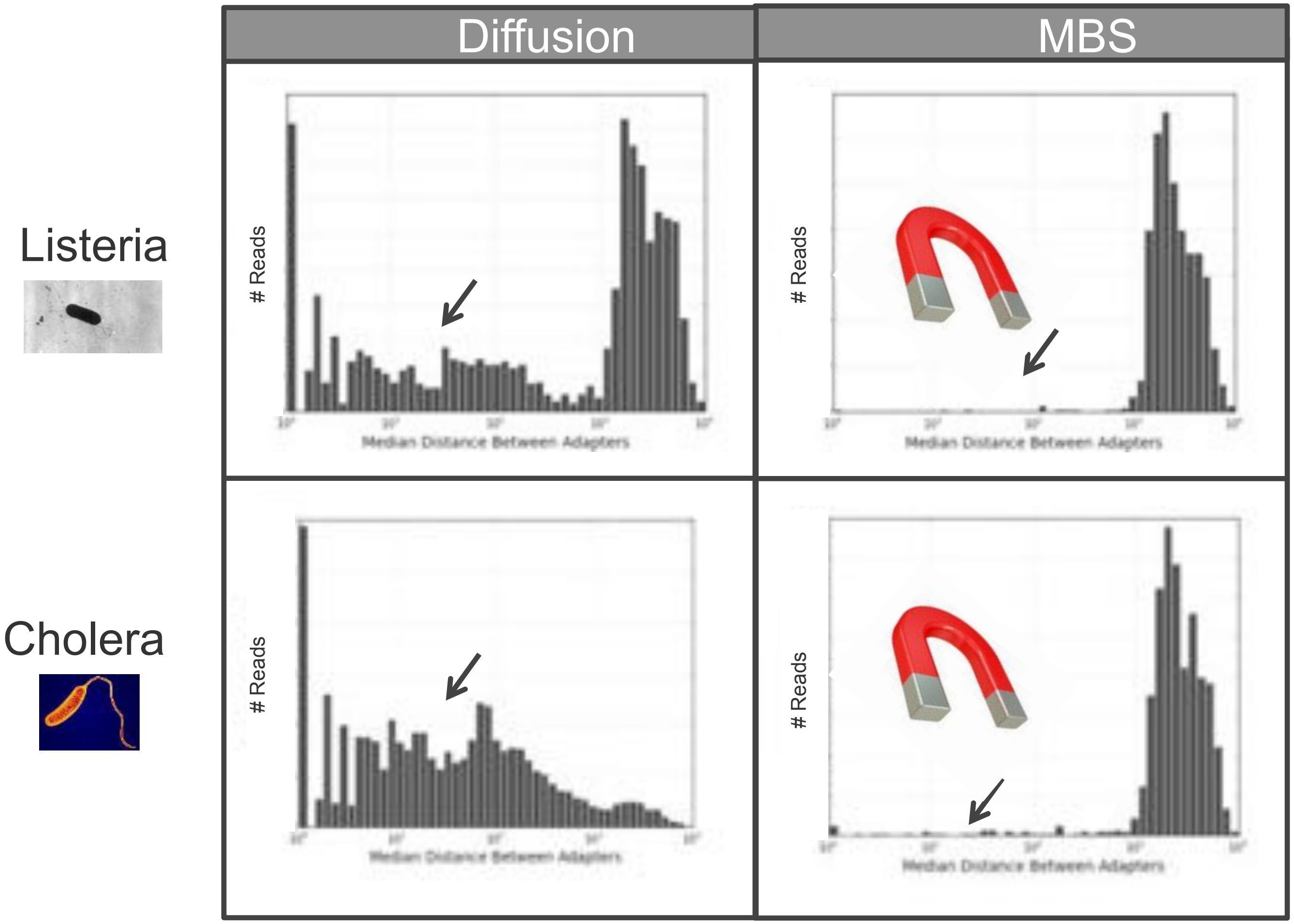


(II)
Introduce magnet: Bead complexes pulled to chip surface



Rotate magnet to evenly disperse beads across entire chip surface

MBS (MagBead Station)



Improvements to Sample Prep