Highly Accurate Consensus Sequence and Genome Assembly using PacBio® *RS* Sequence Data

David H. Alexander, Ph.D.
Pacific Biosciences

November 29, 2012

Outline

- History of DNA sequencing
- PacBio SMRT® sequencing in a nutshell
- A genome assembly primer
- HGAP assembly procedure
- High accuracy consensus: the Quiver algorithm

Section 2

History of DNA sequencing

Generation 1: Sanger sequencing

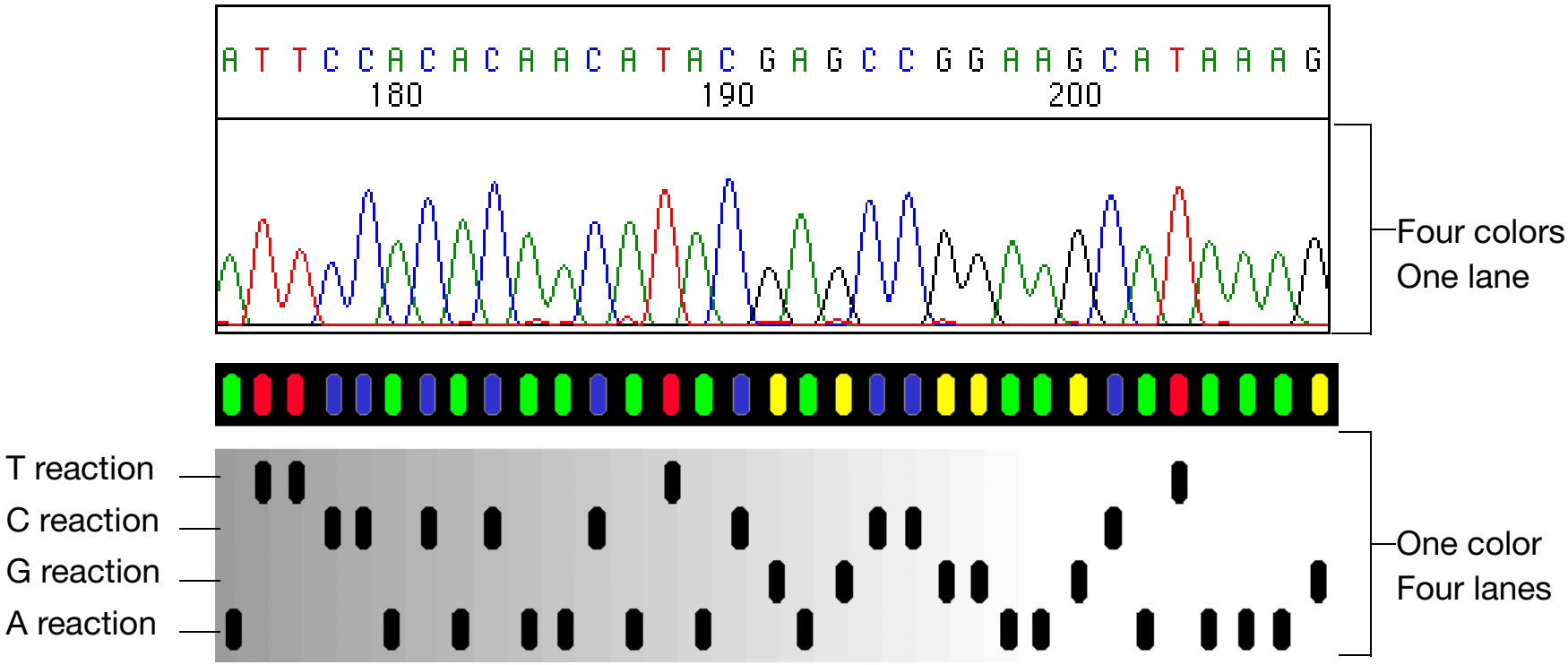
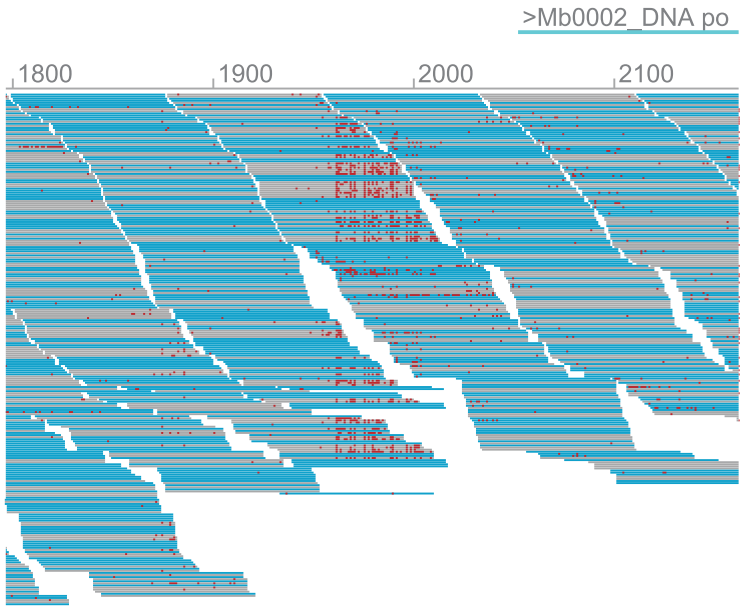
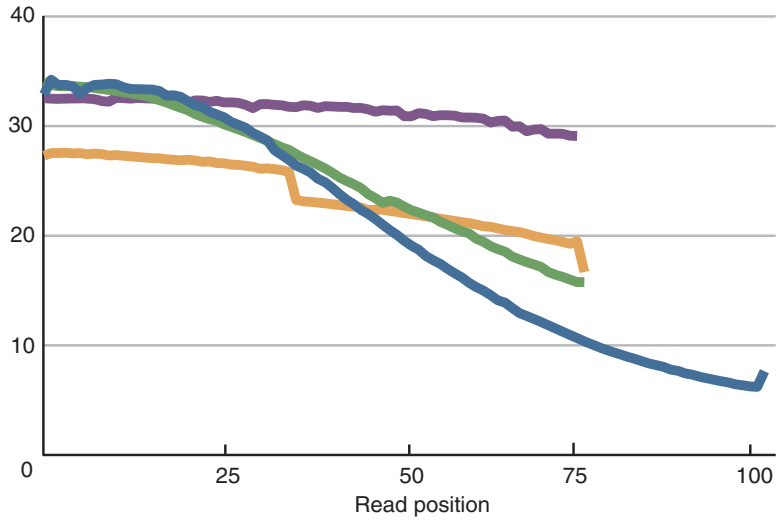


Figure 8 Fluorescent sequencing compared with radioactive sequencing

- Dideoxy chain termination
- Longish reads (~ 700 bp). High accuracy. Low throughput

Generation 2: High-throughput short-read technologies

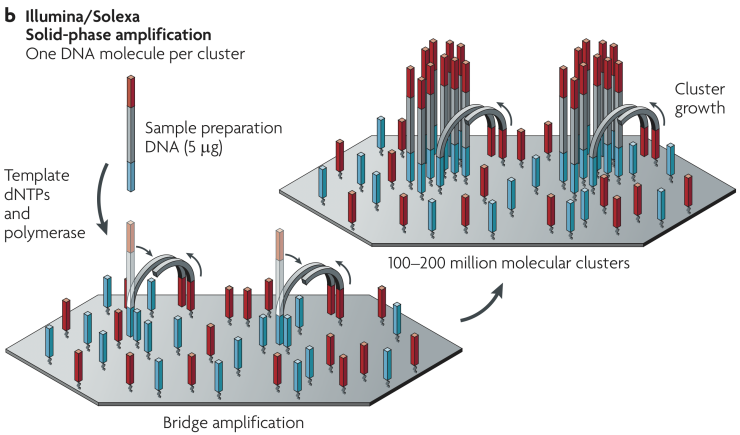


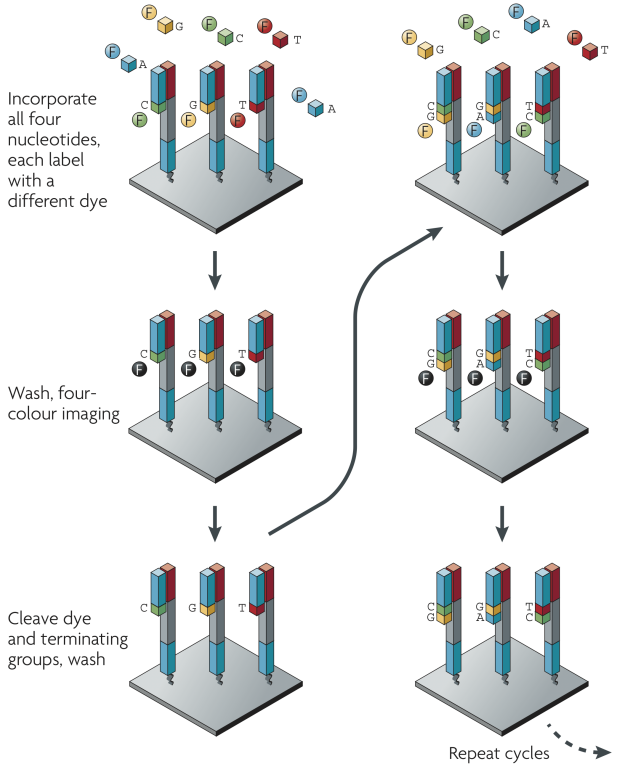


[Nakamura et al.]

- Short-reads (~ 100 bp), paired-end
  - Read length limited by phase coherence of amplicon colonies
- Mate-pair libraries give potential for long-range information
- **Very** high throughput

Generation 2 example: Illumina





- Bridge PCR method for clonal amplification
- Cyclic, reversible-terminator based chemistry

Generation 3: High-throughput long-read technologies

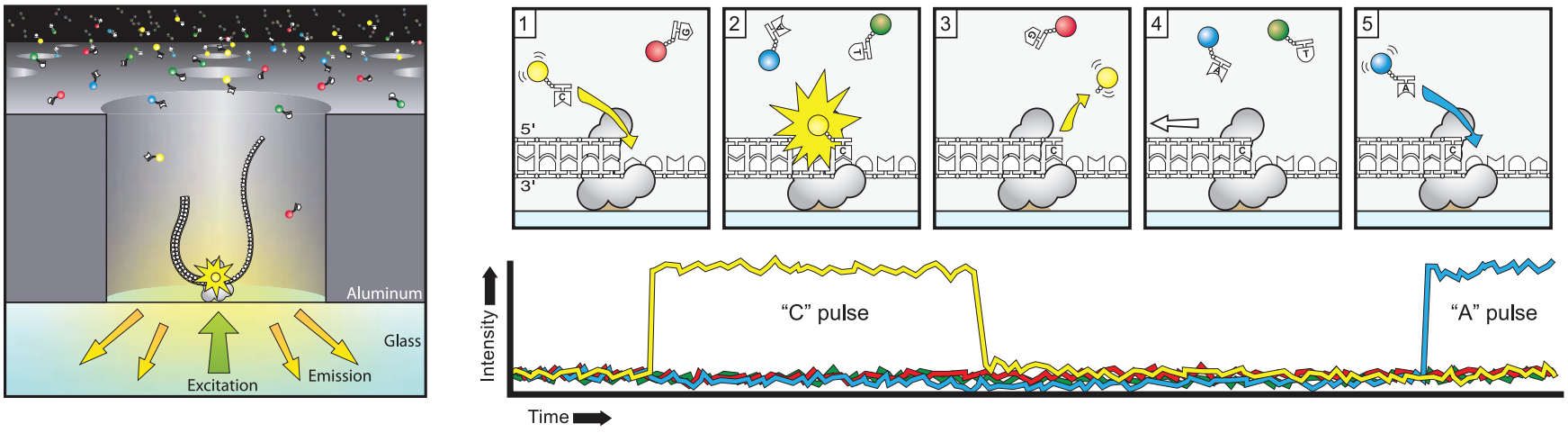
PacBio SMRT® sequencing

- Very long reads (C2 chemistry mean > 3000bp; XL: > 5000bp)
- High throughput

Section 3

PacBio SMRT®Sequencing

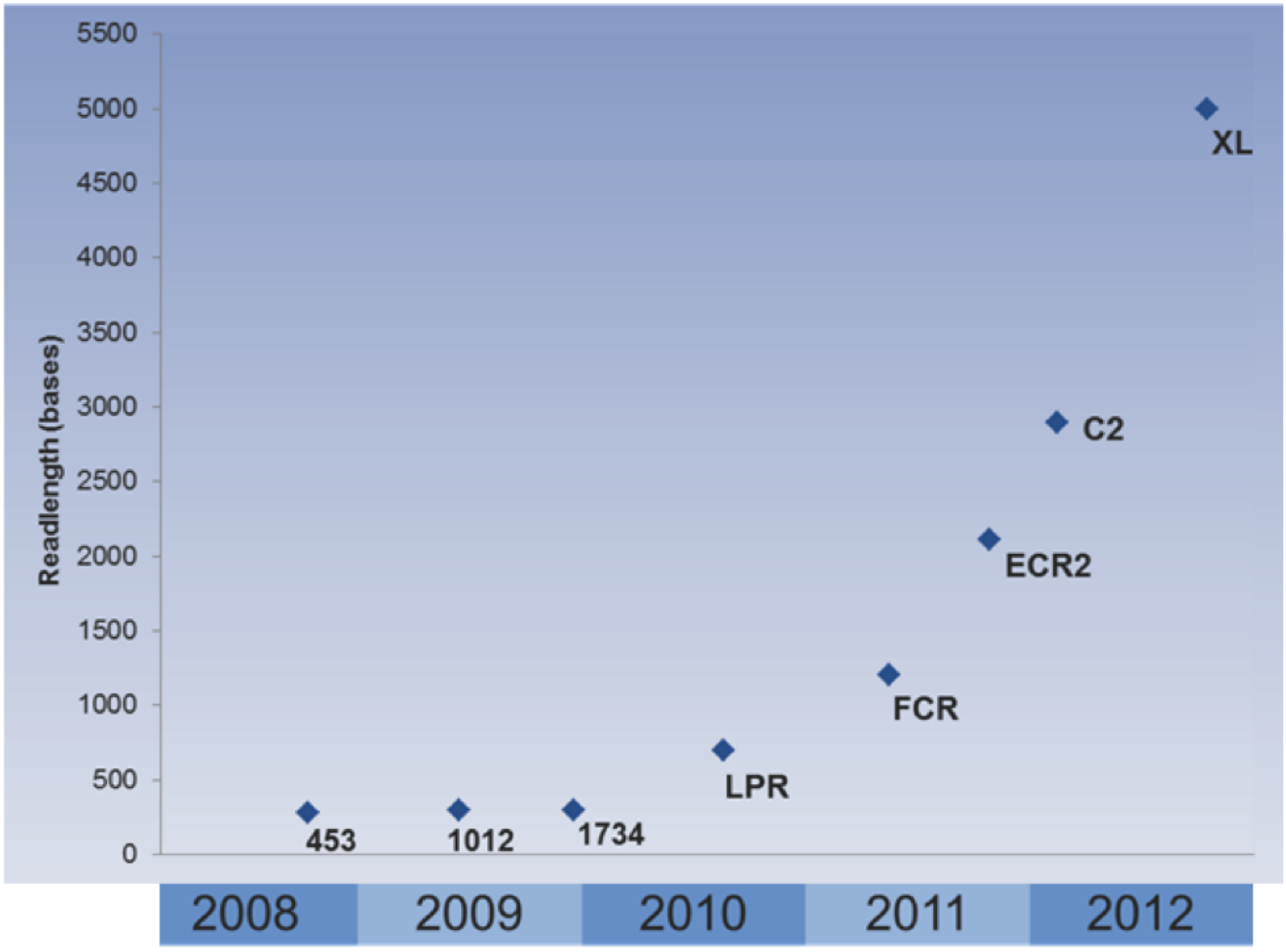
SMRT®= “single molecule, real time”



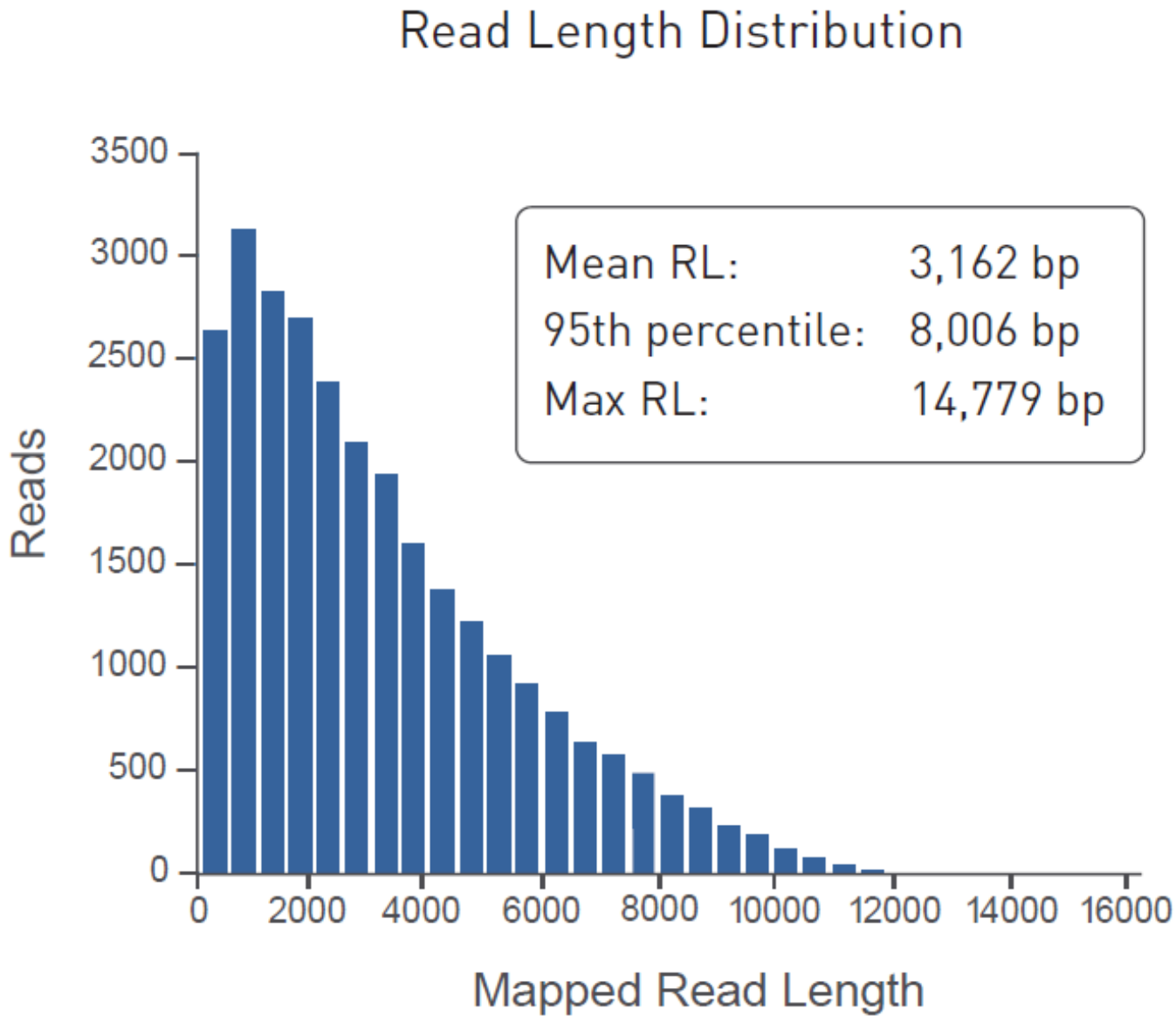
[Eid, 2009]

- Polymerases immobilized in Zero-Mode Waveguides
- DNA sequencing reaction watched in real time using fluorescent-labeled nucleotides

Long, and very long, reads: history

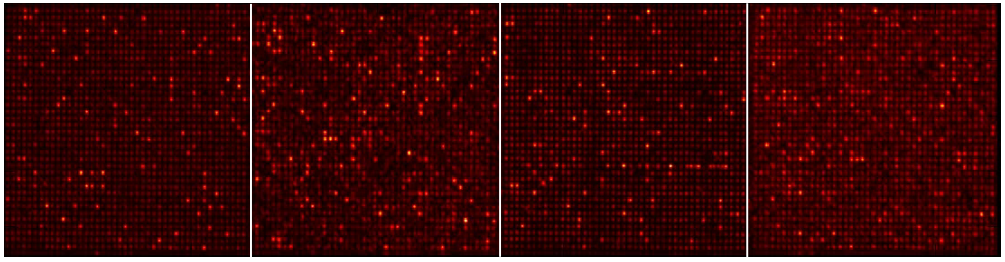


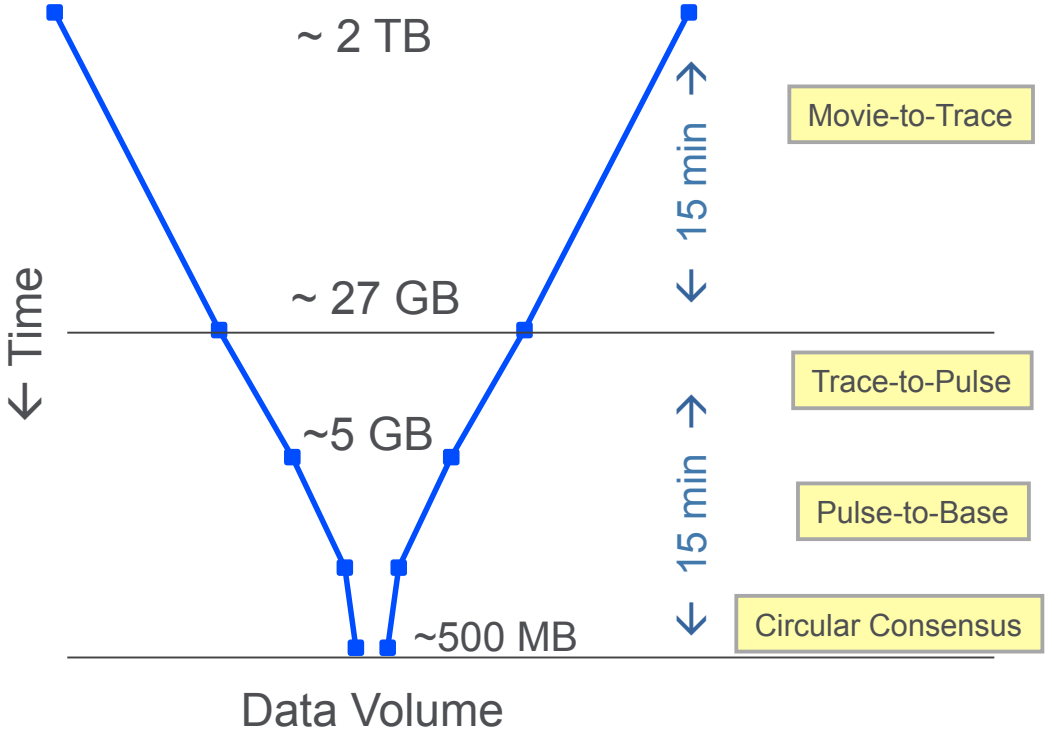
Long, and very long, reads: C2

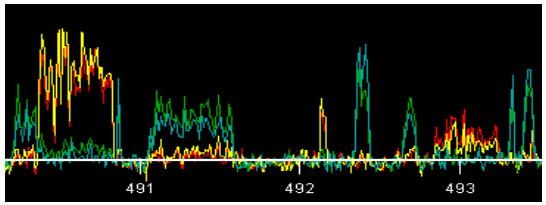


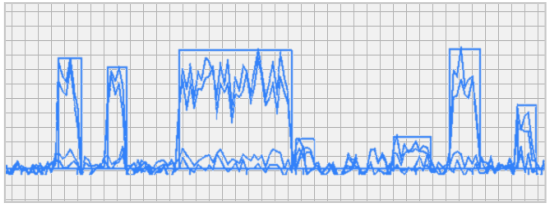
From photon counts to base calls: **primary** analysis

Reduce photon counts to base sequences in soft real-time







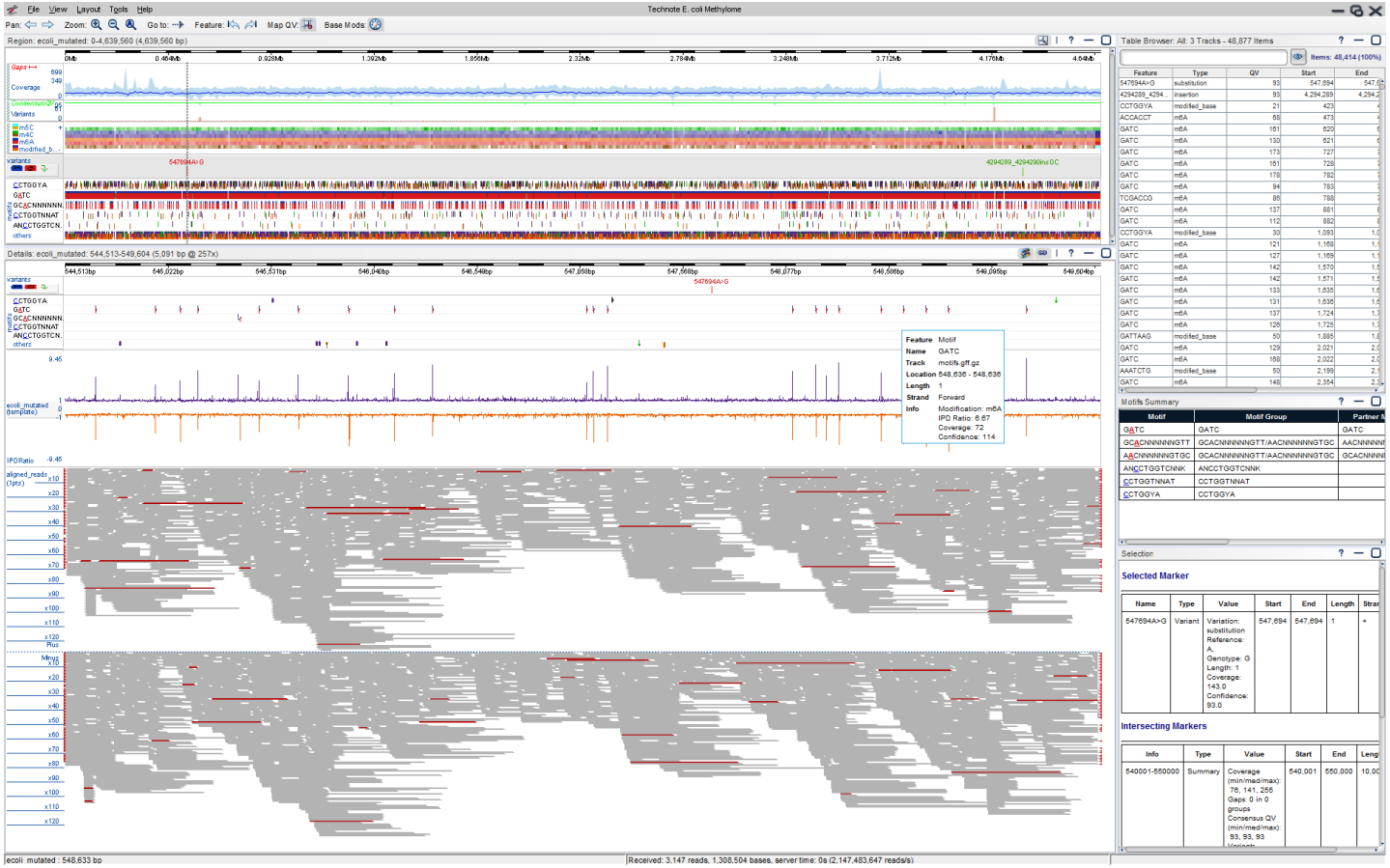


GCAACGATCACCTAAA…
GCAACGATCACCTAAA…
GCAACGATCACCTAAA…
GCAACGATCACCTAAA…

ACGATCACCTAAA…

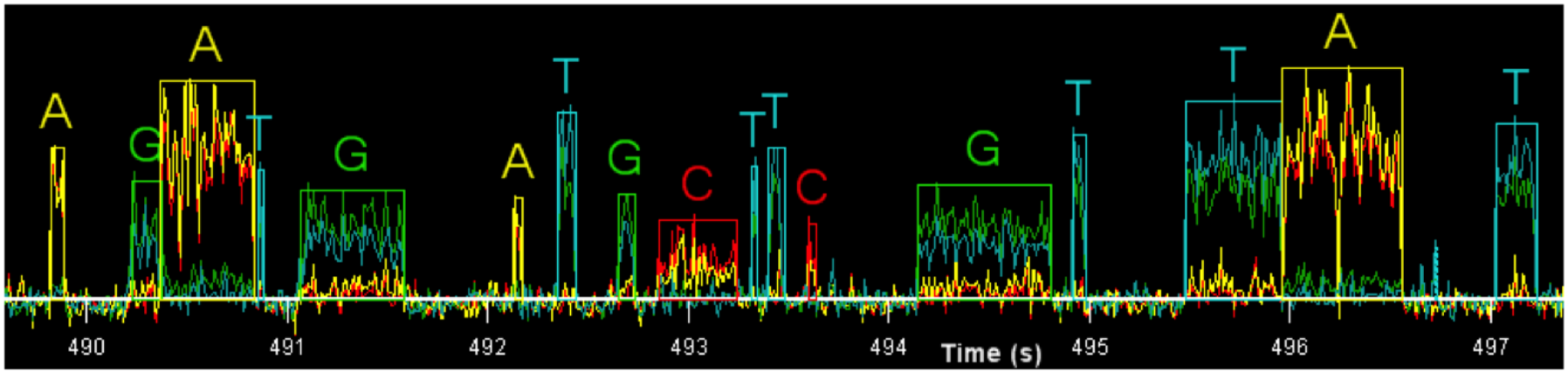
[Credit: Jim Labrenz]

From bases to actionable information: **secondary** analysis



- Variant calling
- Phasing
- *De novo* genome assembly
- Methylation / base modification detection
- Transcript analysis

PacBio data: error model



- Errors dominated by indels
  - Cognate extras (homopolymer expansion)
  - Noncognate extras
  - Dark pulses
  - Pulse merging (homopolymer contraction)
- *Essentially no substitutions*

PacBio data: data format

`bas.h5` file contains a lot more than just basecalls. Basecaller provides metrics indicating areas of uncertainty. “Breadcrumbs”

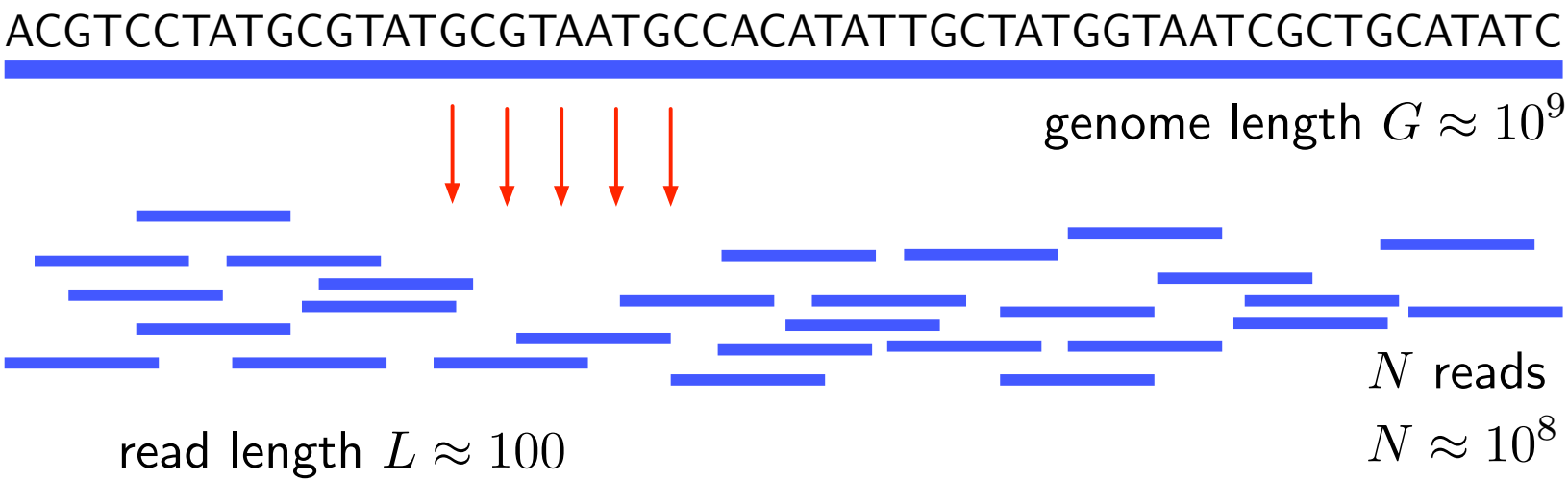
Base	Insertion QV	Substitution QV	Deletion QV	Deletion Tag	Merge QV
A	8	12	16	N	14
T	2	12	5	T	100
T	11	30	4	G	25
G	12	30	11	A	11
G	3	30	16	N	27
C	6	30	16	N	19
C	3	19	3	C	21
G	2	21	4	G	22

$$QV = -10 \log_{10} p_{error}$$

Section 4

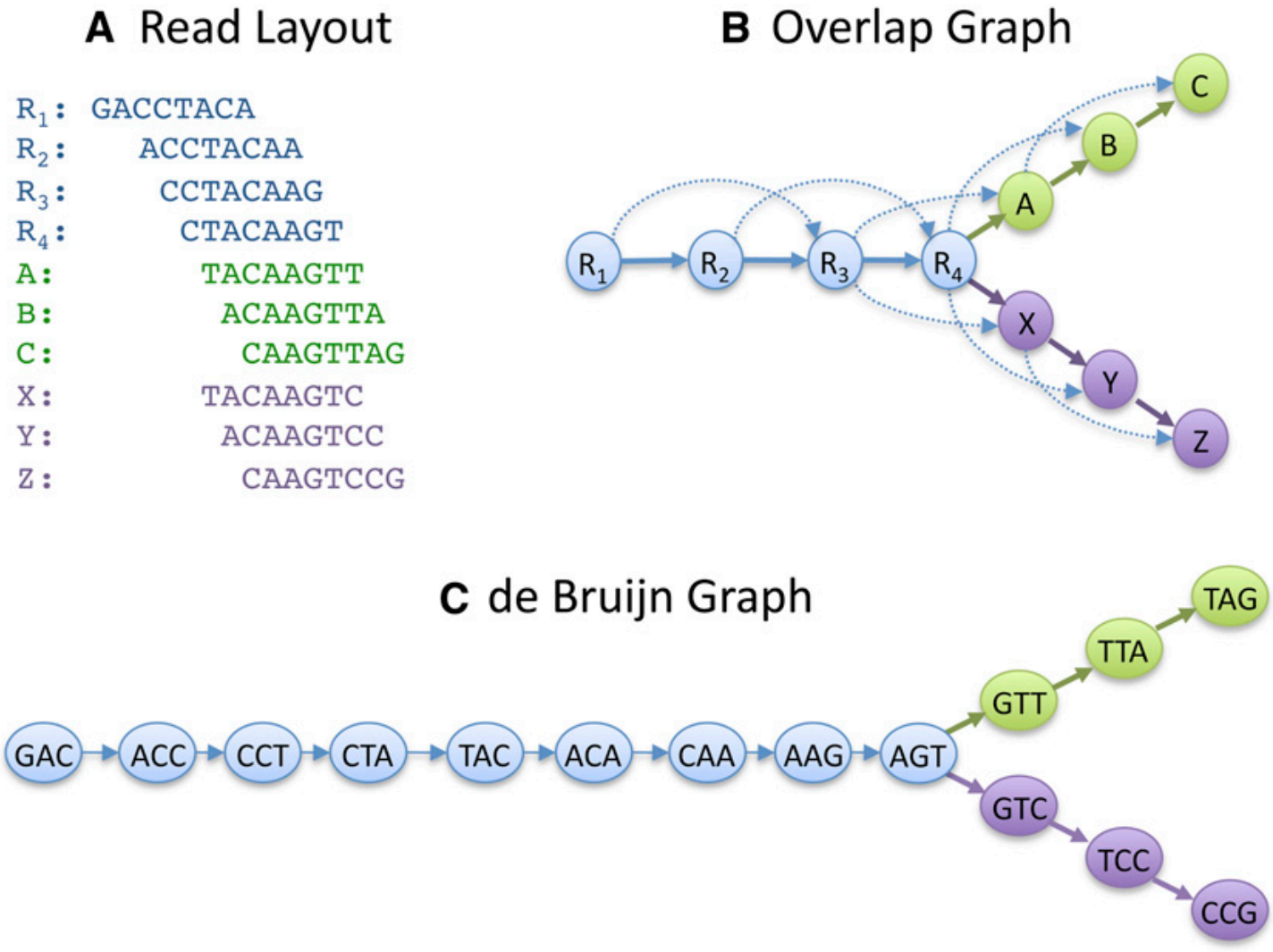
A genome assembly primer

Shotgun sequencing and assembly



[Credit: David Tse]

Shotgun sequencing and assembly



[Credit: Mike Schatz]

Lander-Waterman model (1988)

Idealization of genome assembly. Letting L denote readlength, G genome size, and C coverage, main results:

Expected number of contigs (ideal)

$$\mathrm{E}[\mathrm{NumContigs}] = L^{-1} G C e^{-C}$$

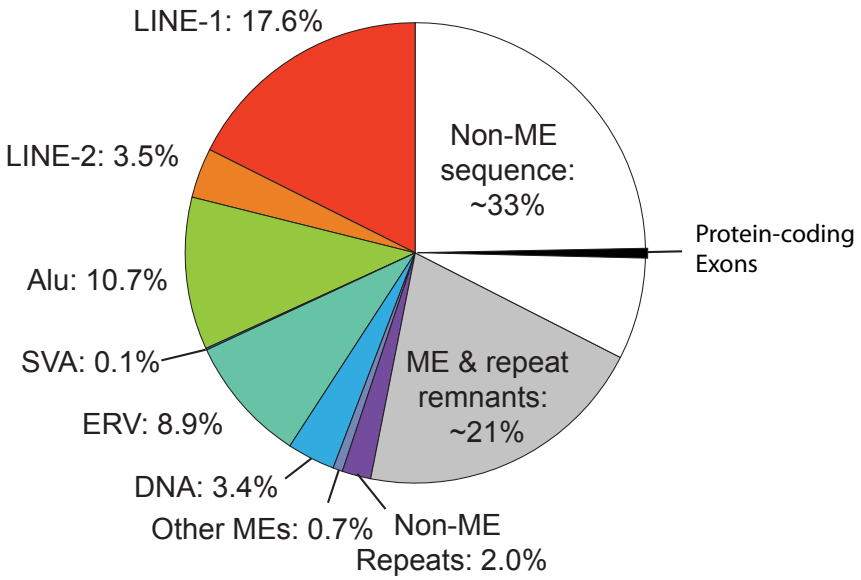
Expected contig length (ideal)

$$\mathrm{E}[\mathrm{ContigLength}] = L C^{-1} (e^{C} - 1)$$

Predicts that long reads are better, but sufficient short read coverage would be just as good. **However...**

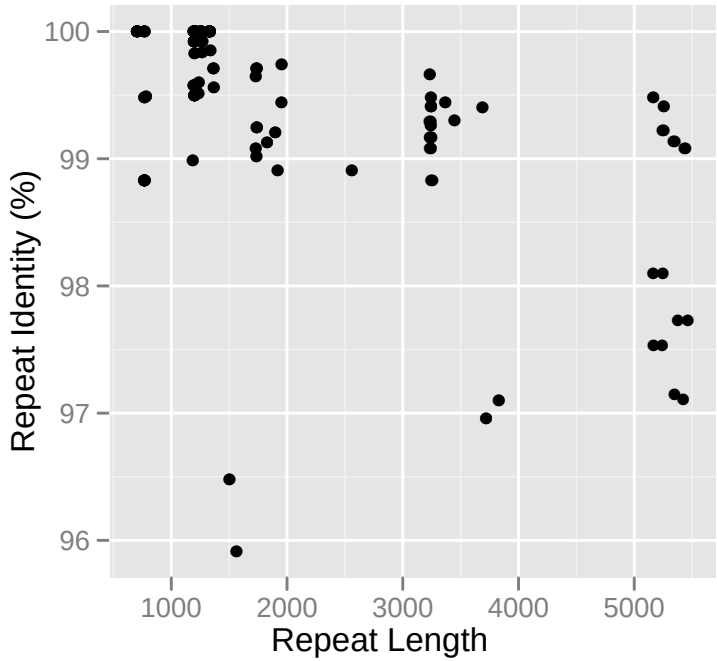
Long repeats are everywhere

Human mobile elements



[Credit: David Witherspoon]

E Coli long repeats



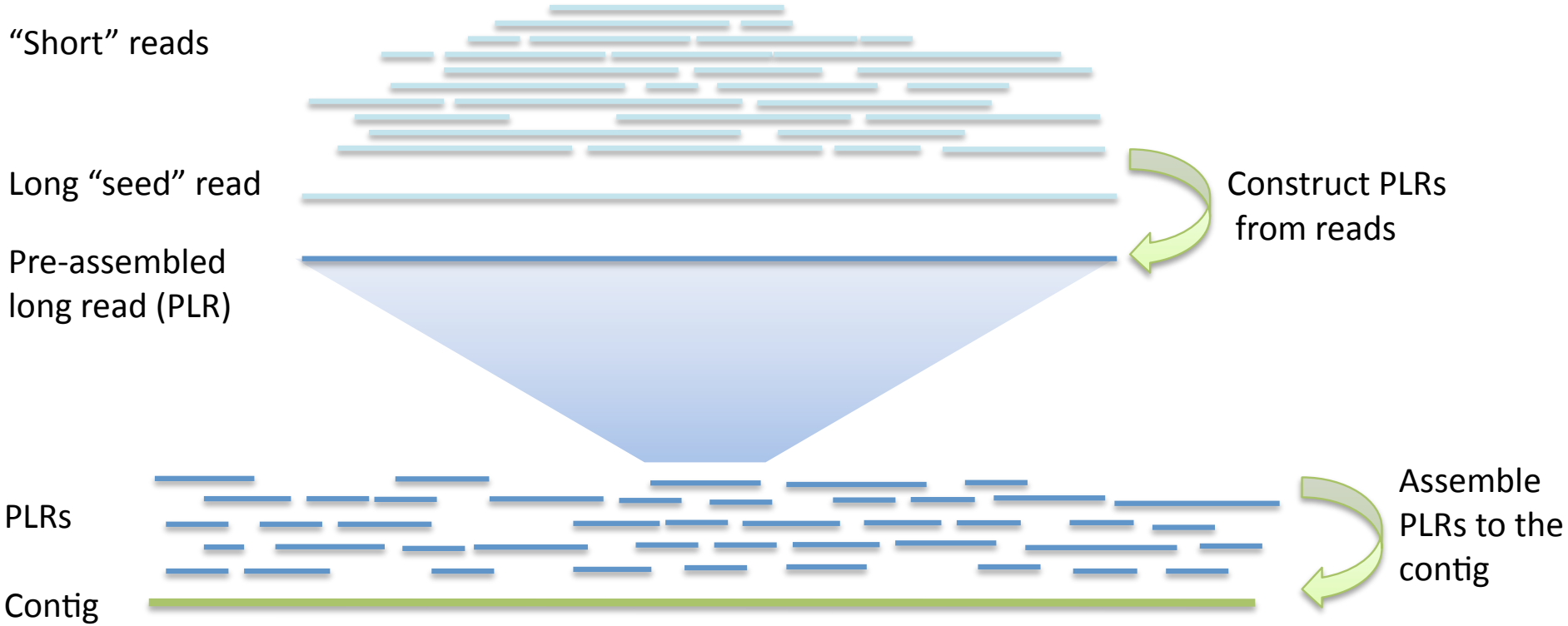
Long repeats stymie short read assembly:

- False edges introduced into overlap graph
- *Misassembly*
- Workarounds—mate-pair, clone libraries—require more laborious sample prep

Section 5

HGAP assembly procedure

Hierarchical genome assembly procedure (HGAP)



[Credit: Jason Chin]

HGAP application: *Meiothermus ruber* (JGI/DOE)

10kb SMRTbell library

4 SMRTCells - 330 Mb

Long seed reads (>5kb) - 92 Mb

pre-assembly

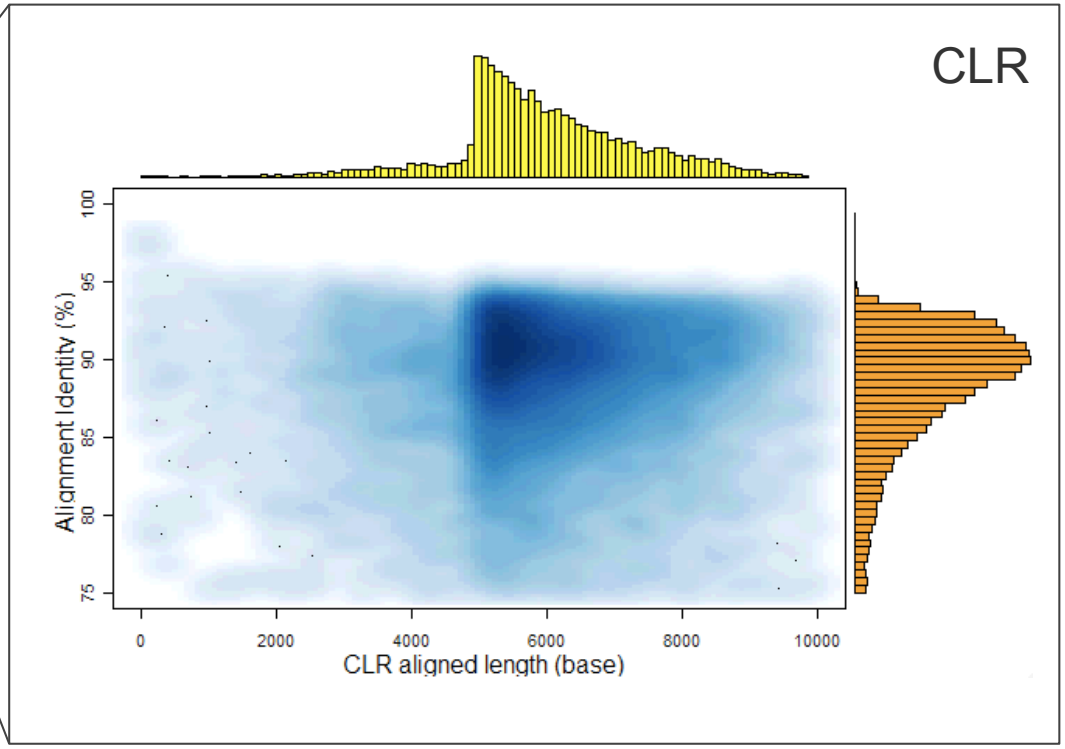
Pre-assembled long reads - 61 Mb

Celera Assembler

5 contigs

clean-up (Minimus2)

1 contig



[Slide: Jonas Korlach]

HGAP application: *Meiothermus ruber* (JGI/DOE)

10kb SMRTbell library

4 SMRTCells - 330 Mb

Long seed reads (>5kb) - 92 Mb

pre-assembly

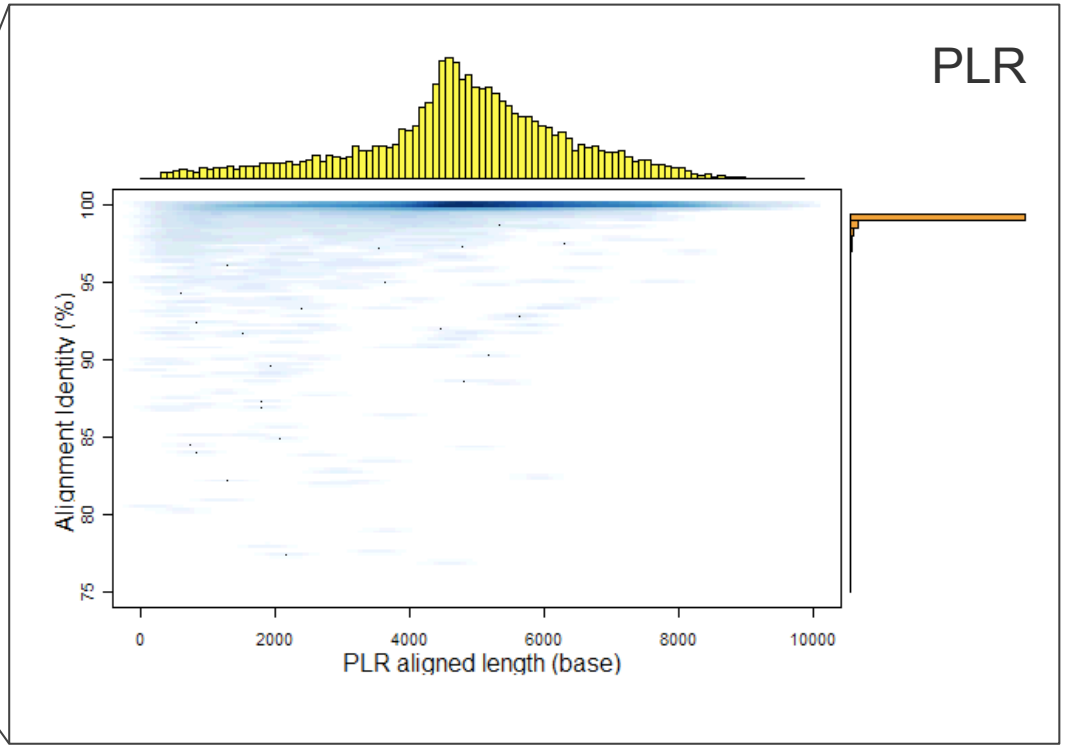
Pre-assembled long reads - 61 Mb

Celera Assembler

5 contigs

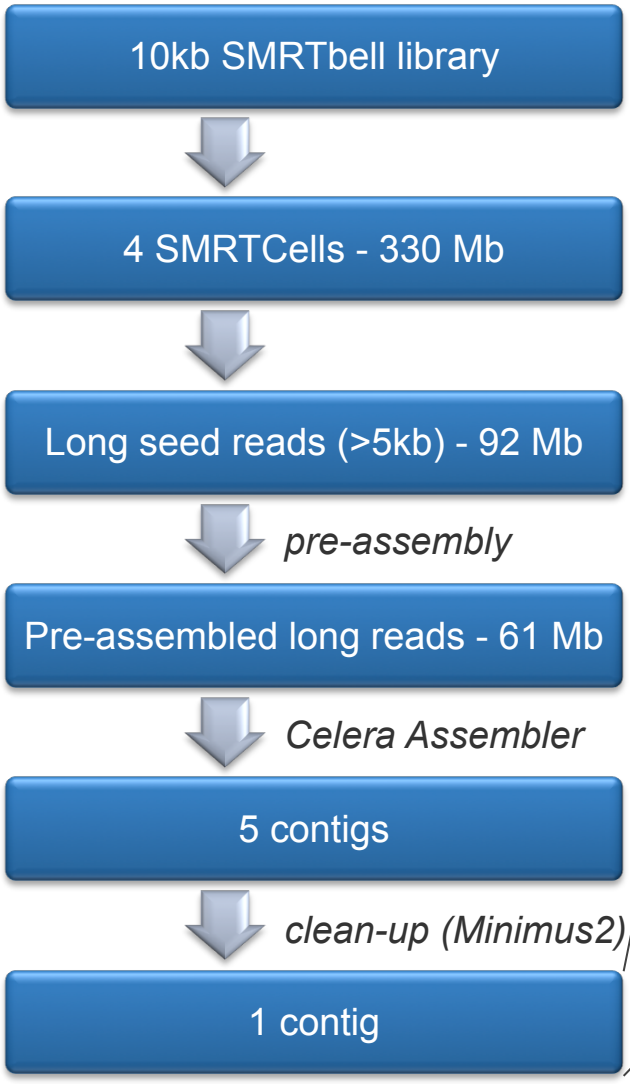
clean-up (Minimus2)

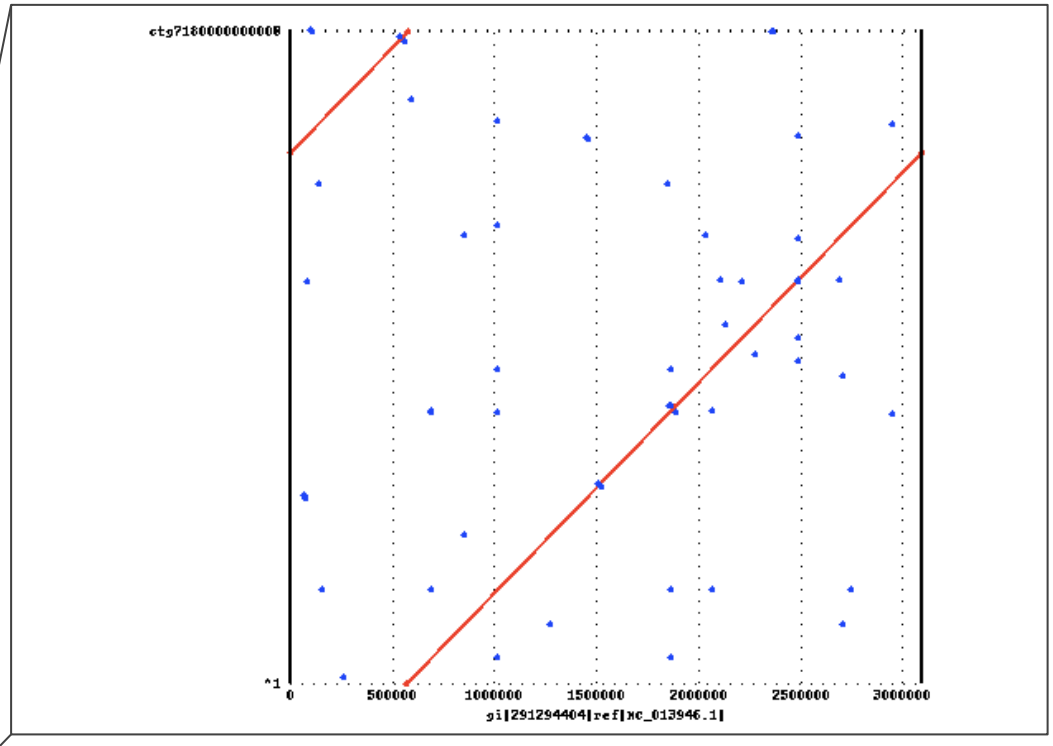
1 contig



[Slide: Jonas Korlach]

HGAP application: *Meiothermus ruber* (JGI/DOE)





single contig spans the entire reference

[Slide: Jonas Korlach]

Section 6

High accuracy consensus: the Quiver algorithm

The consensus problem

- *Given*: A sequence of reads $\mathbf{R} = \{R^1, R^2, \ldots R^K\}$ (encompassing basecalls *and* pulse metric features)
- *Desired*: A consensus sequence $\widehat{T}$ that is, in some sense, a "best" estimate of the underlying true template sequence T that was present in the ZMW.

Applications:

Variant calling, finishing assemblies, circular consensus sequencing...

Example

Template	GATTACA
Read 1	GATTCA
Read 2	GATTTACA
Read 3	GATACA

Multiple-alignment consensus approach

Do an MSA and let each column vote.

Read 1	`GA-TT-CA`
Read 2	`GATTTACA`
Read 3	`GA--TACA`
Plurality	`GA-TTACA`

Problems:

- No notion of template vs observations–no clean way to represent our error model.
- No way to take advantage of QV information

Quiver: a model-based approach

- Encode our sequencing error model as $\Pr(\mathbf{R} \mid T)$
  - Dynamic programming model that can be formalized as a *pair HMM* (actually a *pair CRF*).
- Use a greedy algorithm to maximize the likelihood $\Pr(\mathbf{R} \mid T)$ in the unknown template T.

Quiver algorithm overview

QuiverConsensus for reference window W: (*Rough sketch*)

- Use reference alignment to identify reads $\mathbf{R} = \{R^1, R^2, \ldots R^K\}$ corresponding to W
- *Throw away reference—not used in computing consensus*
- $\widehat{T} \leftarrow \mathsf{PoaConsensus}(\mathbf{R})$
- Repeat until convergence:

$$\widehat{T} \leftarrow \hat{T} + \{\text{single base mutations } \mu \mid \Pr(\mathbf{R} \mid \widehat{T} + \mu) > \Pr(\mathbf{R} \mid \widehat{T})\}$$

How to compute $\Pr(\mathbf{R} \mid T)$?

1. Reads are assumed independent, so

$$\Pr(\mathbf{R} \mid T) = \prod_{k=1}^{K} \Pr(R^k \mid T)$$

2. For PacBio, indels are the rule, so consider the possible *alignments*—the ways T can be construed to have generated R^k:

$$\Pr(R \mid T) = \sum_{\mathscr{A}} \Pr(R, \mathscr{A} \mid T)$$

Computed efficiently using a standard Sum-Product dynamic programming approach.

Sketch of dynamic programming

- Sum-Product definition:

  $A_{ij} \doteq$ marginal prob. of an alignment of $R_{0..(i+1)}$ to $T_{0..(j+1)}$

  $B_{ij} \doteq$ marginal prob. of an alignment of $R_{i..I}$ to $T_{j..J}$

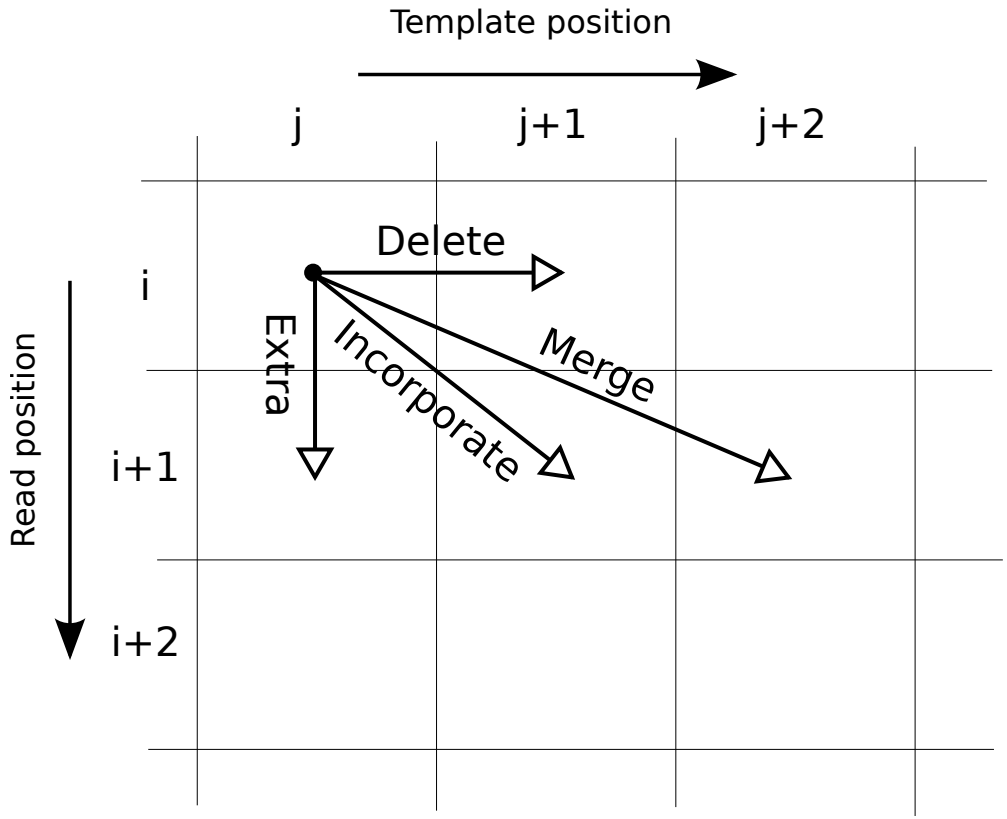
- Sum-Product recursion:

$$A_{ij} = \sum_{m:(i',j')\rightarrow(i,j)} (A_{i'j'} \times \text{moveScore}(m))$$

$$B_{ij} = \sum_{m:(i,j)\rightarrow(i',j')} (\text{moveScore}(m) \times B_{i'j'})$$

- For Viterbi approximation, replace *marginal* by *maximum*, replace *sum* by *max*.

Alignment moves



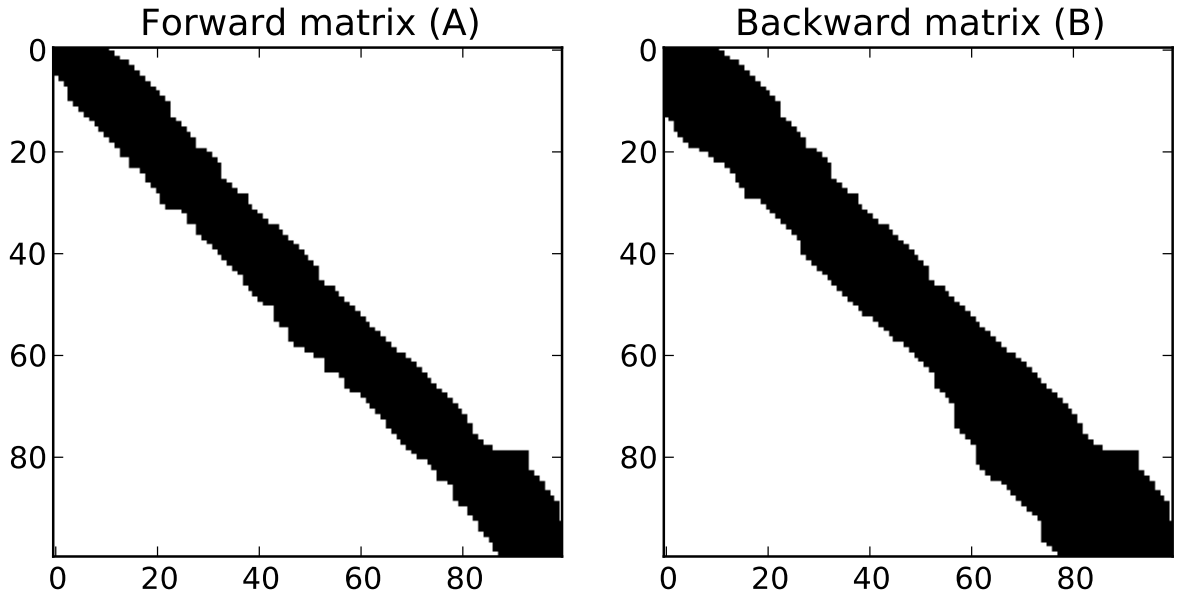
- Additional “merge” move helps better account for pulse merging
- Move scores determined by pulse metrics (QVs)

Efficiently computing $\Pr(R \mid T + \mu)$

- Need to compute score of mutation μ quickly as this is the *rate-limiting operation* in computing the consensus.
- Do not refill entire A, B matrices–we just recalculate two columns of A and join with one column of B.
- Exploit forward-backward identity

$$\begin{aligned}\Pr(R \mid T) =& A_{IJ} = B_{00} \\ =& \sum_{m:(i',j')\to(i,j)} A_{i'j'} \times B_{ij}, \text{ for } \textbf{any } j\end{aligned}$$

Banding for memory and CPU efficiency



- Optimization 1: *banded dynamic programming*: only compute a narrow band of high-scoring rows within each column.
- Optimization 2: Only *store* the bands.

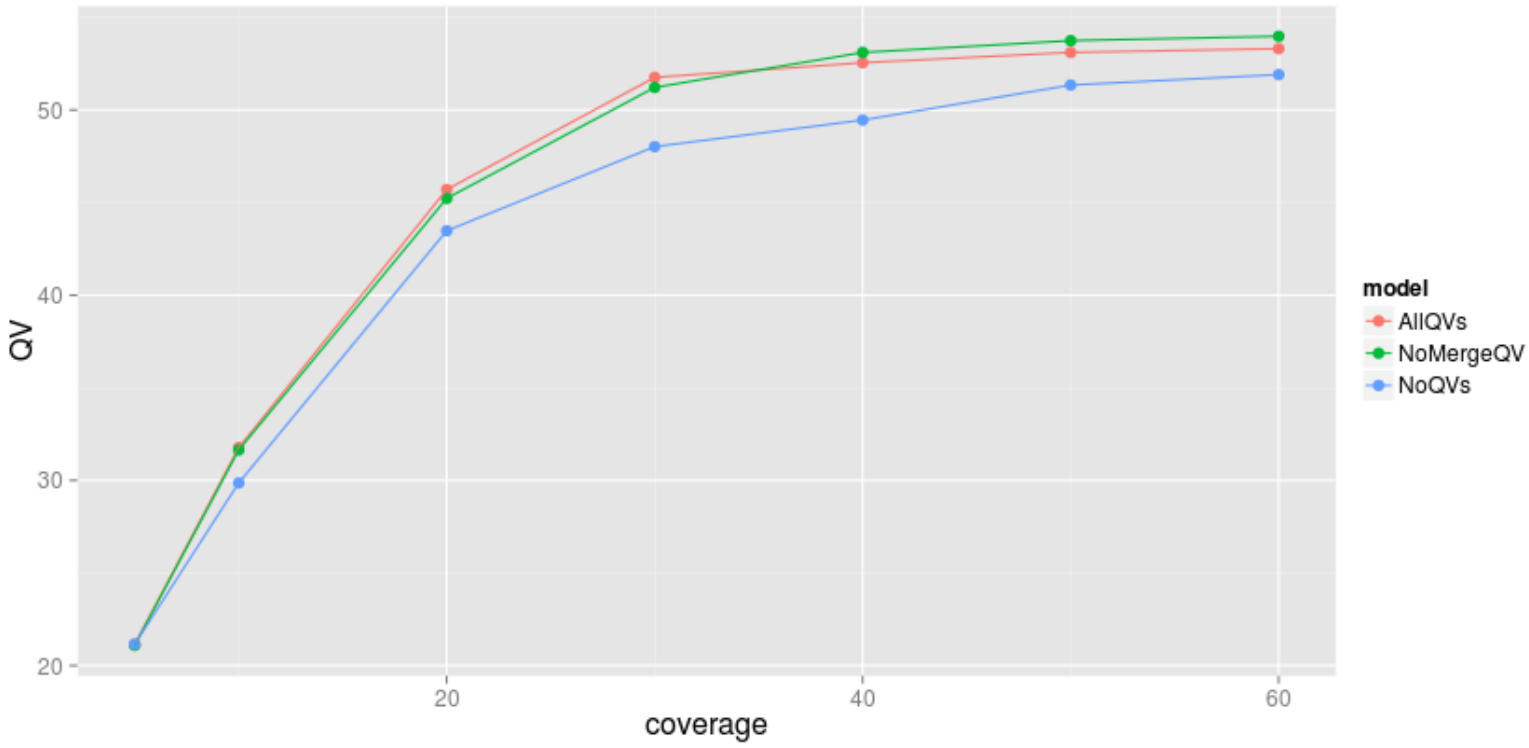
	Naive	Banded
Initial computation of A, B	$O(L^3)$	$O(L^2)$
Computation of mutation score	$O(L)$	$O(1)$
Storage space for A, B	$O(L^2)$	$O(L)$

Quiver application: polishing *M. ruber*

Comparison of assembly to Sanger reference:

Celera assembler "make-consensus"	Q43.4	99.9954%
Quiver	**Q54.5**	**99.99964%**

Quiver accuracy



- With accuracy > Q50, we are now confronting pesky biological facts!
  - Frequent mutations to our in-house *E. Coli* strain
  - Nonclonal samples
- We are also starting to find errors in Sanger-based references!

Quiver extentions

- Reducing coverage requirements for high accuracy
- Diploid/polyploid/sample mixtures

Section 7

Conclusions

Conclusions

The PacBio® *RS* has become a powerful platform for assembly and highly accurate sequencing, and is getting better by the week!

Acknowledgements

- Jonas Korlach
- Pat Marks (Quiver co-author)
- Jason Chin
- Aaron Klammer
- Collaborators at DOE-JGI, Eichler lab