PACIFIC BIOSCIENCES®

Advances in Sequence Consensus and Clustering Algorithms for Effective *De Novo* Assembly and Haplotyping Applications with SMRT® Sequencing

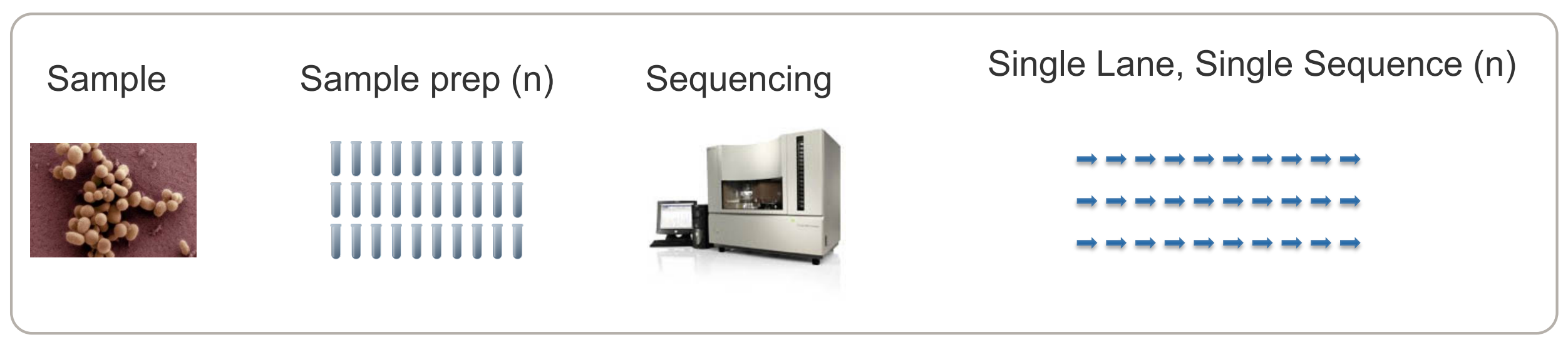
Jason Chin, Patrick Marks, David H. Alexander, Aaron Klammer, Michael Brown, Cheryl Heiner
Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025

Introduction

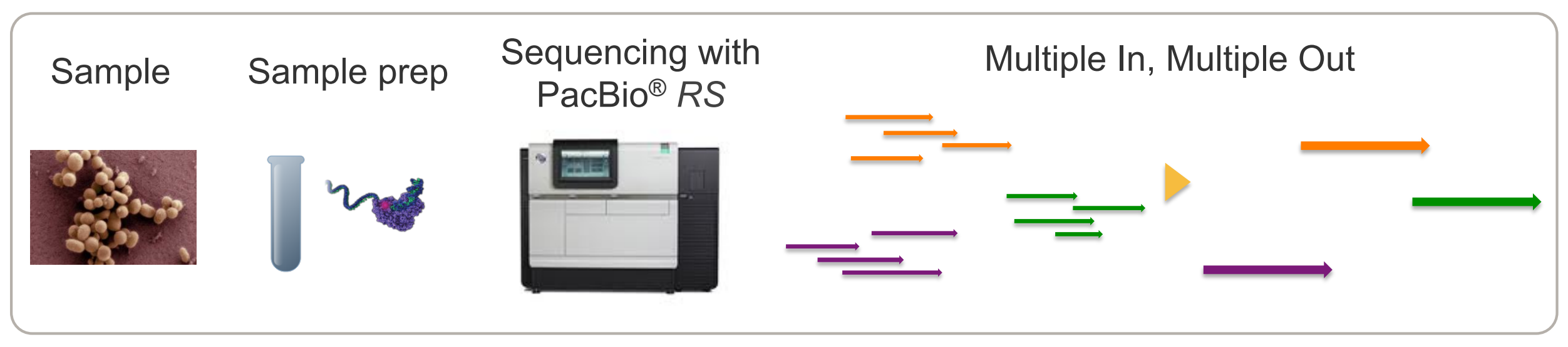
One of the major applications of DNA sequencing technology is to bring together information that is distant in sequence space so that understanding genome structure and function becomes easier on a large scale. The Single Molecule Real Time (SMRT®) sequencing platform provides direct sequencing data that can span several thousand bases to tens of thousands of bases in a high-throughput fashion. In contrast to solving genomic puzzles by patching together smaller piece of information, long sequence reads can decrease potential computation complexity by reducing combinatorial factors significantly. We demonstrate algorithmic approaches to construct accurate consensus when the differences between reads are dominated by insertions and deletions. High-performance implementations of such algorithms allow more efficient *de novo* assembly with a pre-assembly step that generates highly accurate, consensus-based reads which can be used as input for existing genome assemblers. In contrast to recent hybrid assembly approach, only a single ~10 kb or longer SMRTbell™ library is necessary for the hierarchical genome assembly process (HGAP). Meanwhile, with a sensitive read-clustering algorithm with the consensus algorithms, one is able to discern haplotypes that differ by less than 1% different from each other over a large region. One of the related applications is to generate accurate haplotype sequences for HLA loci. Long sequence reads that can cover the whole 3 kb to 4 kb diploid genomic regions will simplify the haplotyping process. These algorithms can also be applied to resolve individual populations within mixed pools of DNA molecules that are similar to each, e.g., by sequencing viral quasi-species samples.

Generate High-Throughput, High-Quality, Long *De Novo* Consensus Sequences

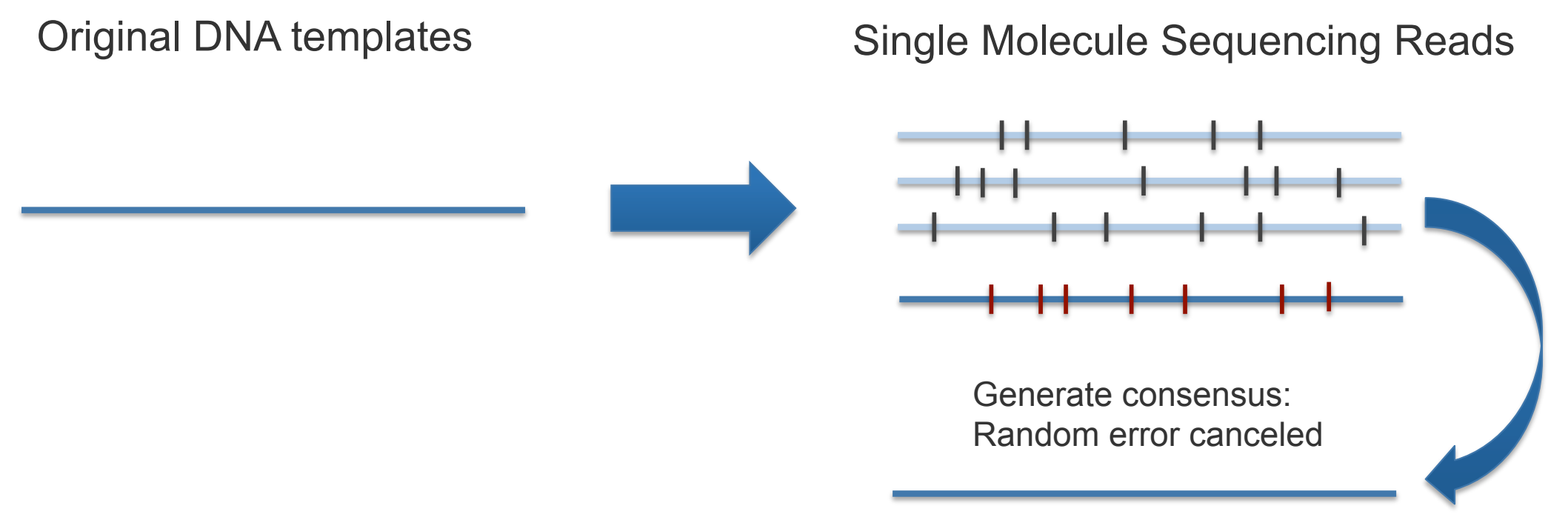
Sample prep / de-mixing (with molecular biology)



Bioinformatics clustering and consensus generates equivalent or better and longer final results in a high throughput way

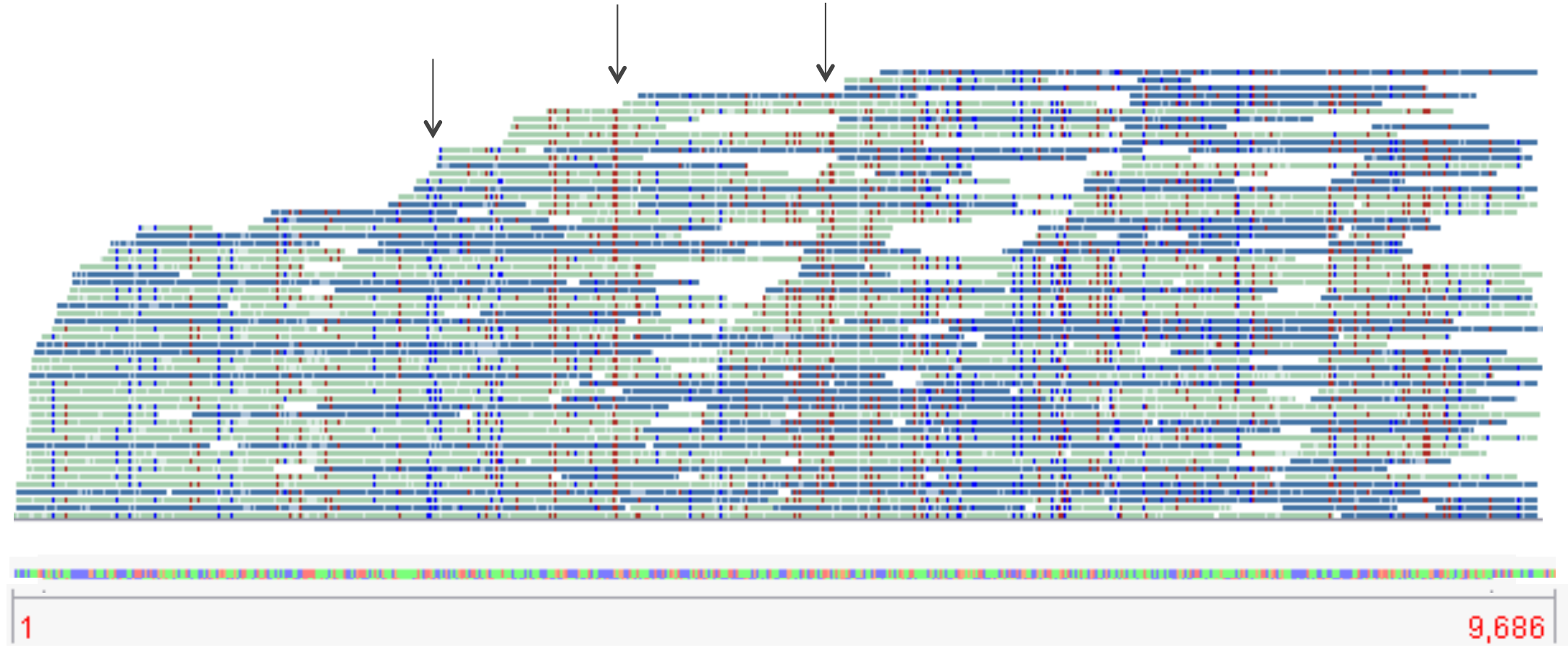


Errors Are Random in SMRT® Sequencing, Not Correlated with Real Variants



Example: Generate Highly Accurate Consensus Reads with a Seed Read

Inconsistent sites removed through a consensus steps

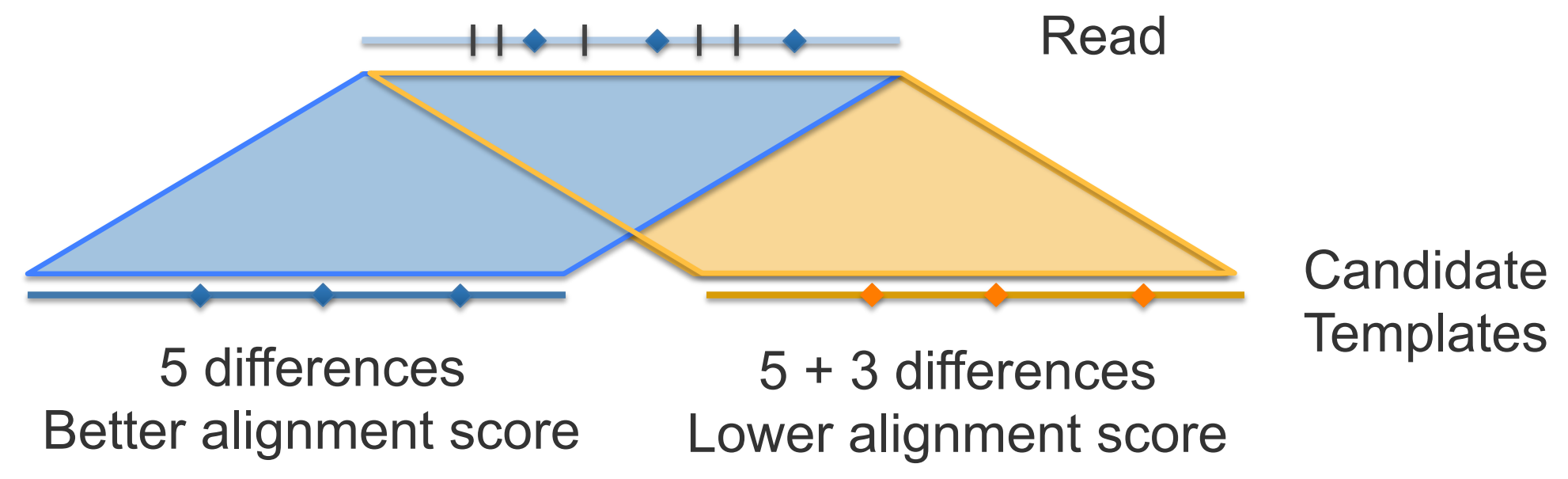


1. Start with 9.7 kb seed read
2. Align other reads to the seed read for construct mini-assembly
3. Construct accurate pre-assembled consensus sequence

- Utilizes every bit of data:
  - Longest reads for continuity
  - Shorter reads used for consensus accuracy
- Sequence Identity to the reference: 85.7% (seed read) ➔ 99.3% (pre-assembled long read), 9089 bp
- Chimera / low quality regions can be filtered out early
- Accurate long consensus reads easier to assemble

Clustering And Consensus

Long Reads Provide Accurate Mapping

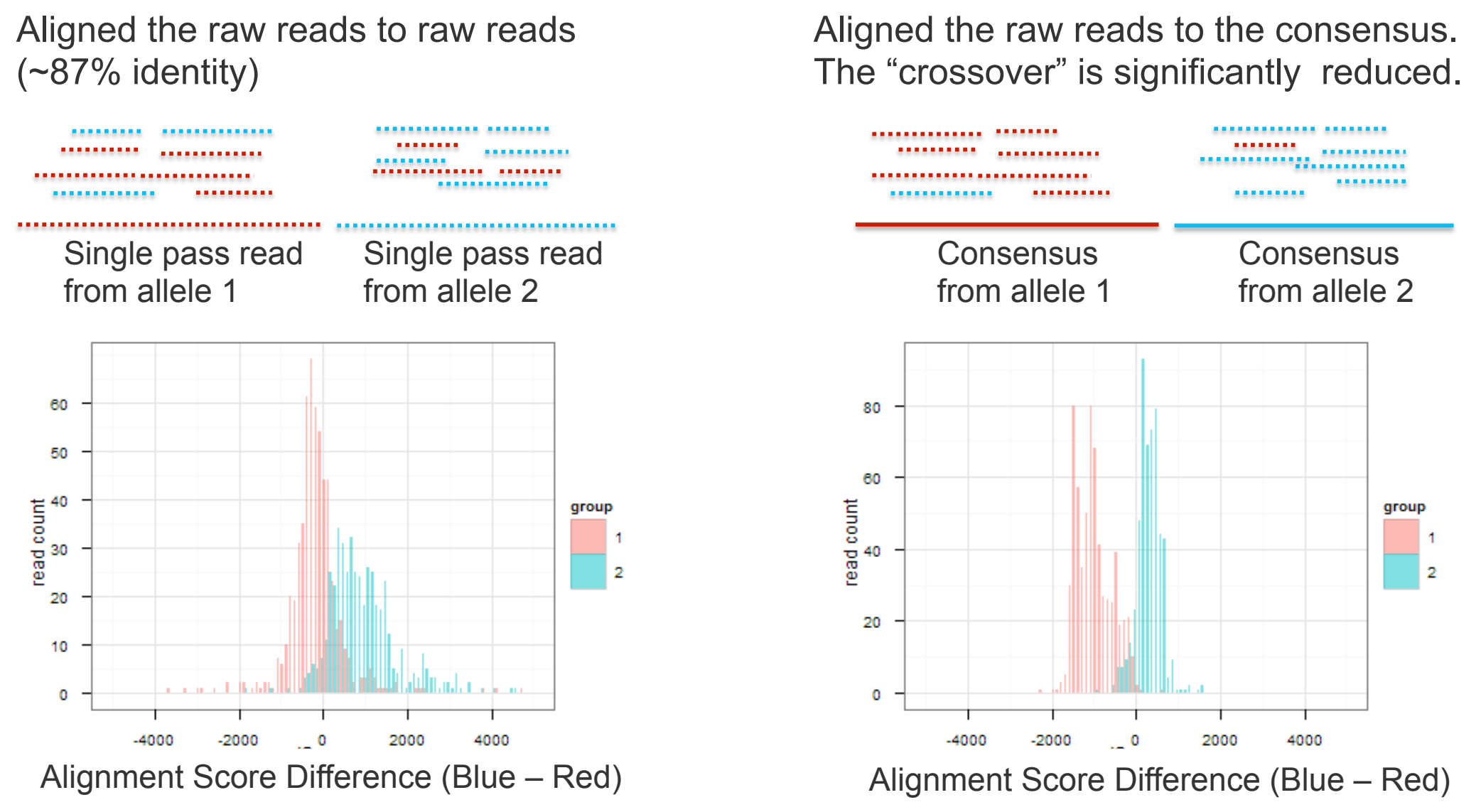


The real variants are not correlated with the errors. No ambiguity to map the reads correctly.

The longer the reads, the more useful information comes from the real variants!!

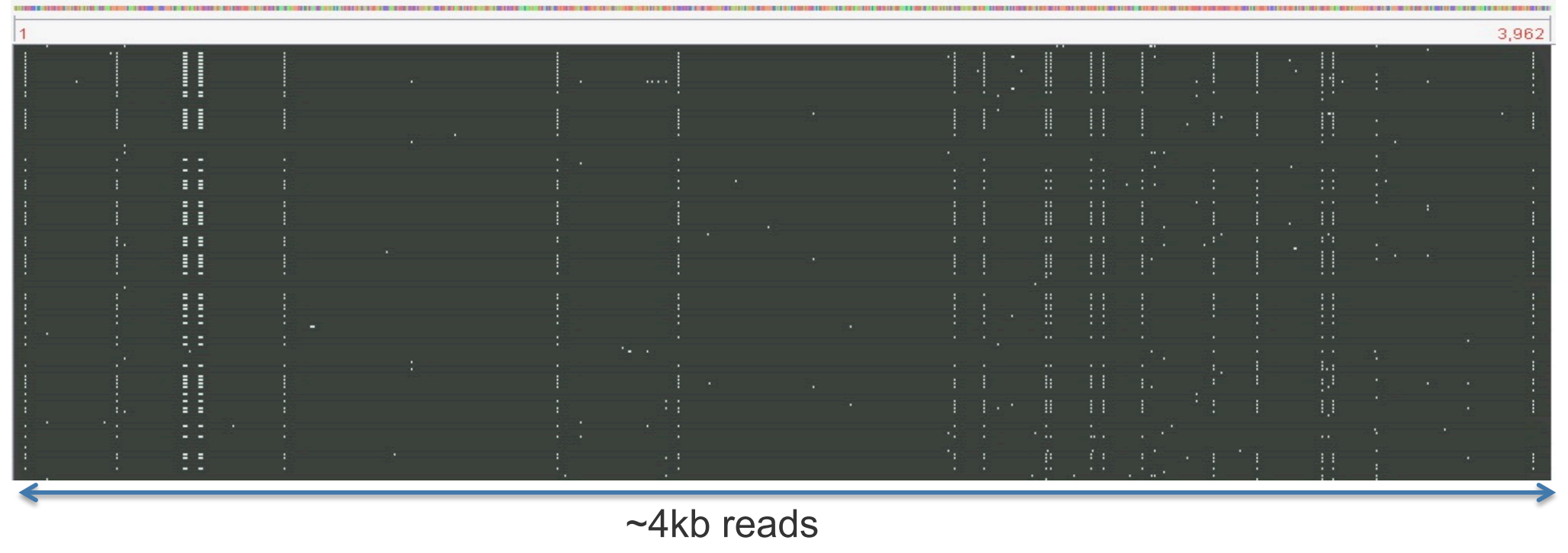
Iteratively Improving Clustering and Consensus in *De Novo* Fashion

Two HLA-C alleles as templates, 98% identical to each other over 4.1 kb regions.


Aligned the raw reads to raw reads (~87% identity)

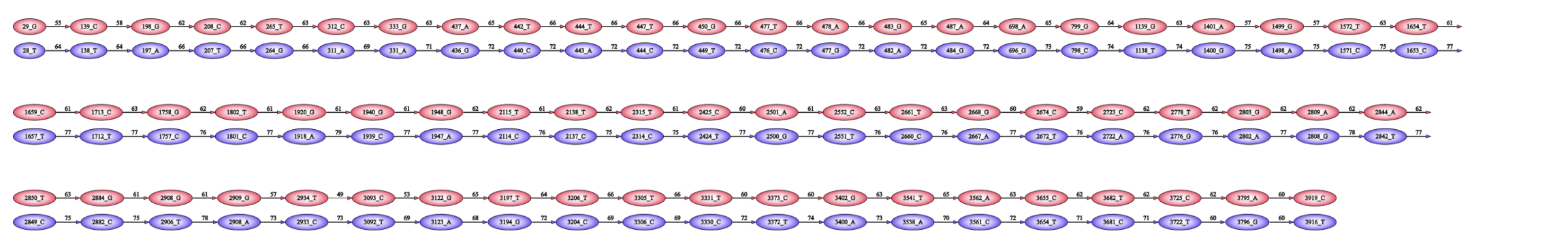
Aligned the raw reads to the consensus. The “crossover” is significantly reduced.

~4 kb Consensus Reads Constructed for Direct High Resolution HLA Haplotyping



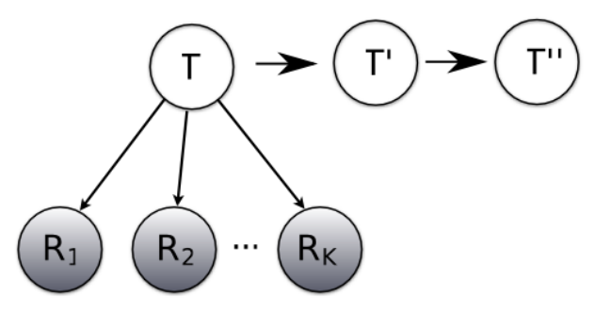
Two haplotype clusters can be visualized directly without a haplotype assembly process from mixed reads

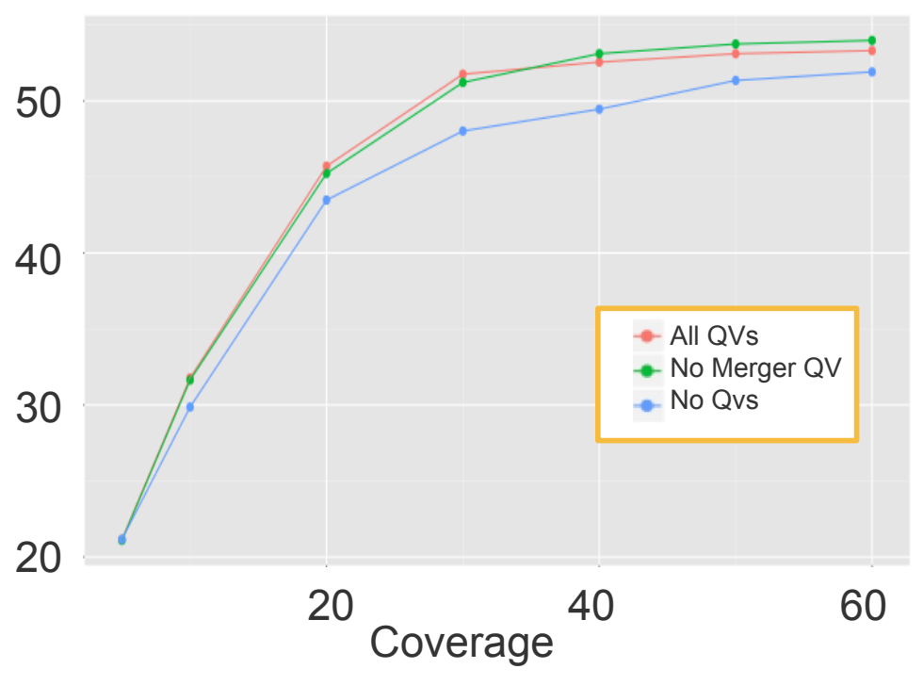
Full phased variants



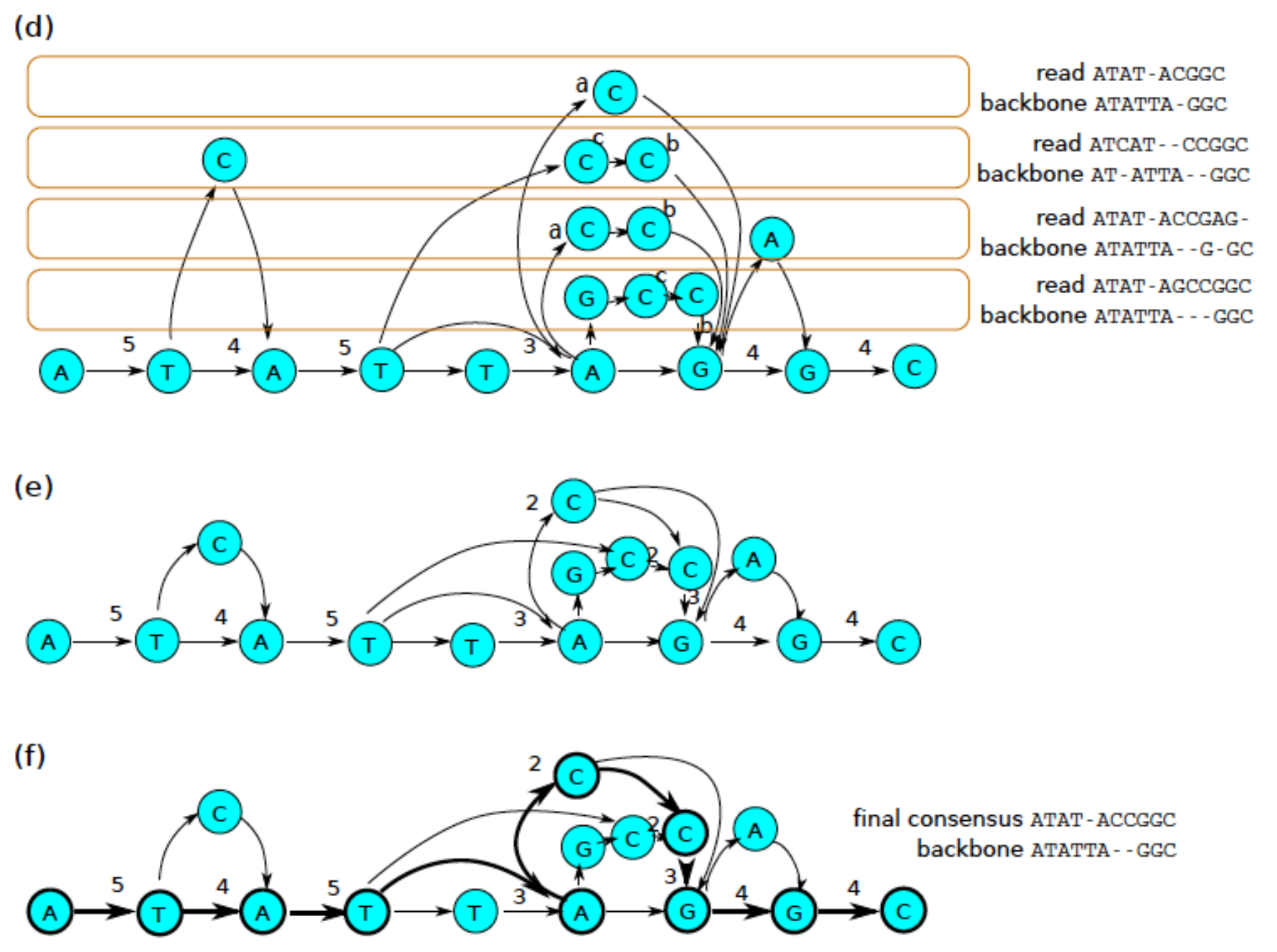
Quiver: A Graphical Model Consensus Calling Algorithm for High Finishing Quality

```
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
  ▸ $\hat{T} \leftarrow \mathrm{PoaConsensus}(\mathbf{R})$
  ▸ Repeat until convergence:
        $\hat{T} \leftarrow \hat{T} + \mu$
    where $\mu$ is a single base mutation with
        $\Pr(\mathbf{R} \mid \hat{T} + \mu) > \Pr(\mathbf{R} \mid \hat{T})$
```





Fast Consensus Construction Using Directed Acyclic Graphs for Pre-assembly

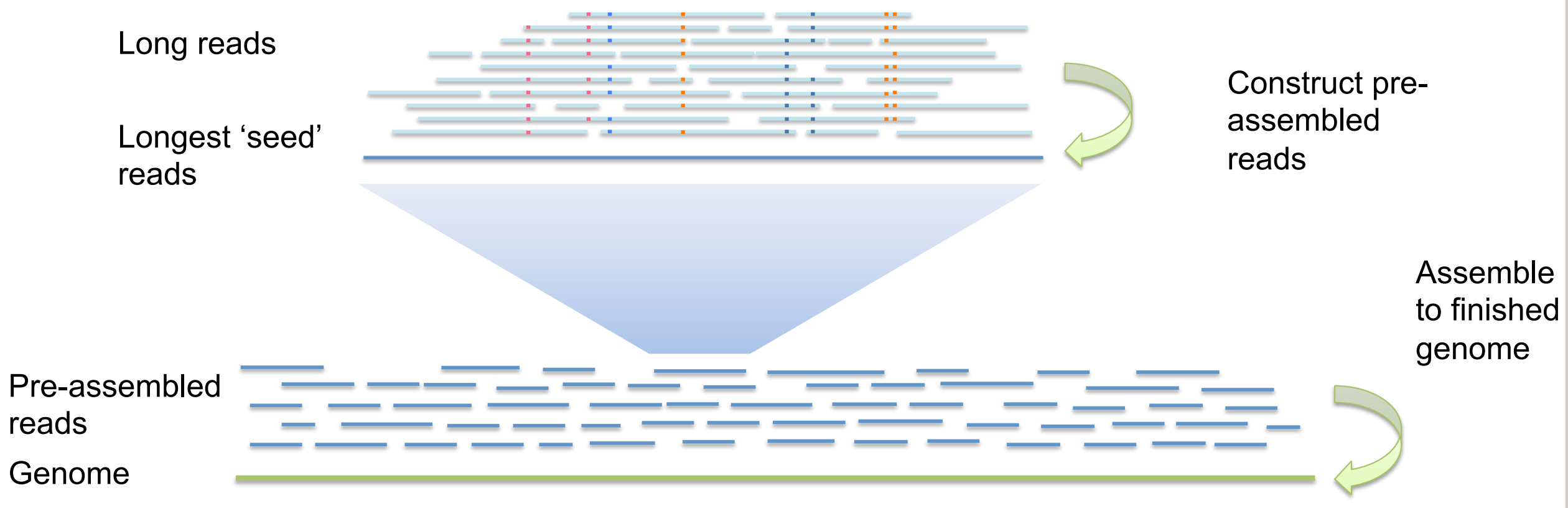


read ATAT-ACGGC
backbone ATATTA-GGC
read ATCAT--CCGGC
backbone AT-ATTA--GGC
read ATAT-ACCGAG-
backbone ATATTA--G-GC
read ATAT-AGCCGGC
backbone ATATTA---GGC

final consensus ATAT-ACCGGC
backbone ATATTA--GGC

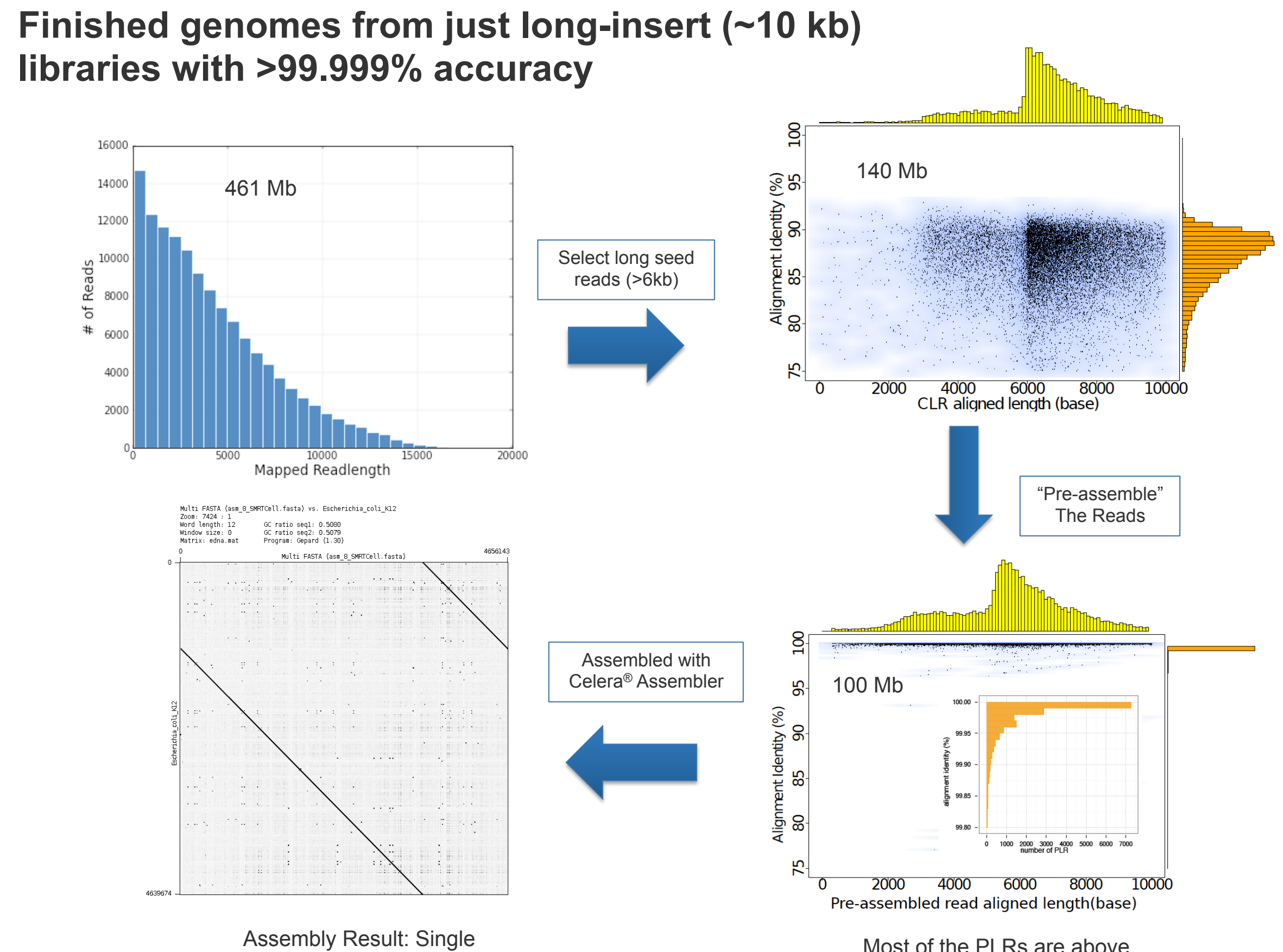
Hierarchical Genome Assembly Process

Overview of the Hierarchical Genome Assembly Process (HGAP)



Bacteria Assembly with HGAP

Finished genomes from just long-insert (~10 kb) libraries with >99.999% accuracy



Assembly Result: Single contig spans the whole chromosome

Most of the PLRs are above 99.95% accuracy using only PacBio long reads.

Escherichia coli (K12 MG 1655) Assembly Results

SMRT® Cells	CLR bases	CLR Cov.	Seed read Cov.	PLR Cov.	PLR nReads	PLR mean read length	Assembly size	# of contigs >10 kb (all)	Genome covered	N50	Concordance with Sanger reference	QV	% full-length matched ORF predicted
8	460M	99.4	30.2	21.5	17,232	5,777	4.66M	1(2)	100.3%	4,65M	99.99951%	53.1	99.8%
6	340M	73.4	22.6	15.7	13,090	5,566	4.70M	10(14)	101.3%	1,16M	99.99938%	52.1	100.0%
4	232M	50.0	14.9	10.1	8,610	5,422	4.69M	17(21)	101.1%	0.39M	99.99876%	49.1	98.8%

- The final consensus uses Quiver to reduce residual errors.
- 21.1X PLRs with average length of 5.7 kb resolve all ~5.5 kb rRNA repeats to give (almost)-single contig assembly
- With 4 SMRT Cells, we reach 98.8% ORF prediction concordance with an assembly of N50 =390k, 21 contigs

Plasmodium falciparum Assembly

Plasmodium falciparum 3D7 genome (~80% AT rich, 23.3 Mb genome) HGAP results **from single standard 10 kb library**:

	22 SMRT Cells	30 SMRT Cells
Raw Sequence Coverage	114x	~155x
# of contigs	**135**	**98**
Contig N50	**918 kb**	**1,242 kb**
Largest Contig	1.79 Mb	2.535 Mb
Assembly Size	23.7 Mb	23.7 Mb
SNP or discordant sites identified	5,547*	5,112*

The assembled contigs size is about >300x larger than previous published assemblies with next generation sequencing for similar genomes.

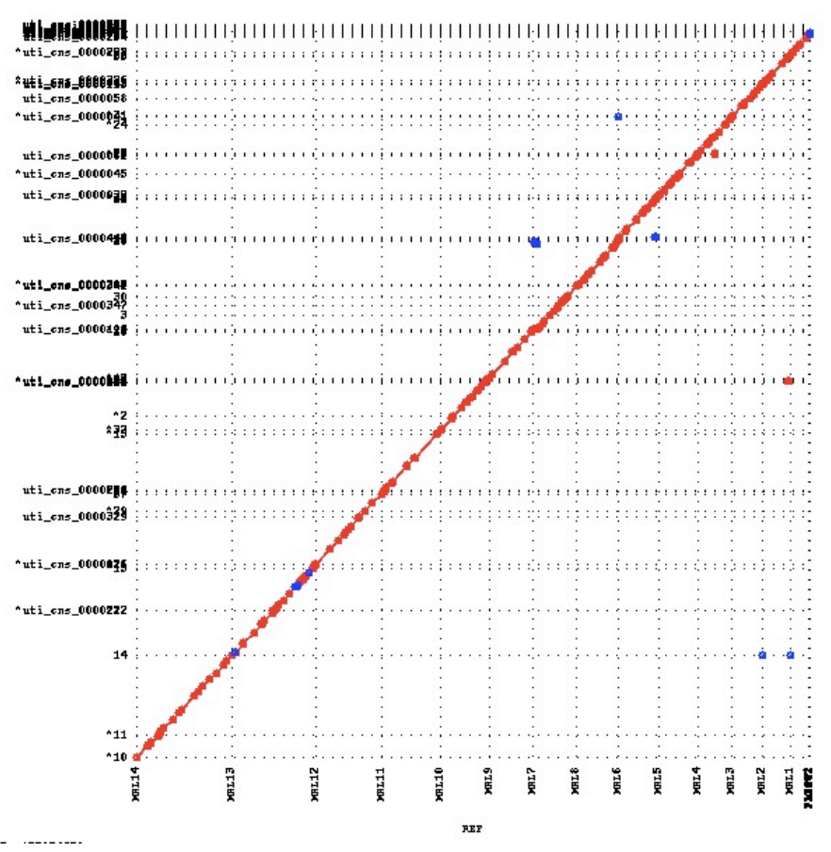
*Preliminary Results

Earlier *Plasmodium* assembly Results for Reference

	454® Prosequencing		Sanger Sequencing		Illumina® Sequencing	
	7C126	SC05	Dd2	HB3	NP-3D7-S	NP-3D7-L
Raw Sequence Coverage	33x	36x	7.8x	7.1 x	43x	64x
# of contigs	**9,452**	**9,597**	**4,511**	**2,971**	**26,920**	**22,839**
Contig N50	**3.3 kb**	**3.3 kb**	**11.6 kb**	**20.6 kb**	**1.5 kb**	**1.6 kb**
Largest Contig	36.7 kb	34.4 kb	79.2 kb	111.9 kb	29.1 kb	24.0 kb
Assembly Size	20.8 Mb	21.1 Mb	19.5 Mb	23.4 MB	19.0 Mb	21.1 Mb

Data from Upeka Samarakoon, et al. BMC Genomics. 2011; 12: 116

Assembled 98 Contigs Aligned to the Reference



Acknowledgement

We thank Stanford Genome Technology Center for providing the HLA-C DNA sample and Carsten Russ (Broad Institute) and Sarah Volkmann (Harvard School of Public Health) for providing the *Plasmodium falciparum* 3D7 genomic DNA. We would also like to thank the Joint Genome Institute for DNA samples and discussions for developing HGAP.