Cold Spring Harbor Laboratory

Hybrid De Novo Assembly of Eukaryotic Genomes

James Gurtowski

Dr. Michael Schatz

Cold Spring Harbor Laboratory

Assembly of Rice
Genome Size: 400MB

Assembler	Sequencing Data	N50 Contig
ALLPATHS	60x 101bp Illumina PE (180bp insert) 2k jump 5k jump	21kb
Celera	28x Flashed MiSeq 250 PE	4.5kb
Celera	19x Error Corrected Pacbio -pacbioToCA with flashed MiSeq	34kb

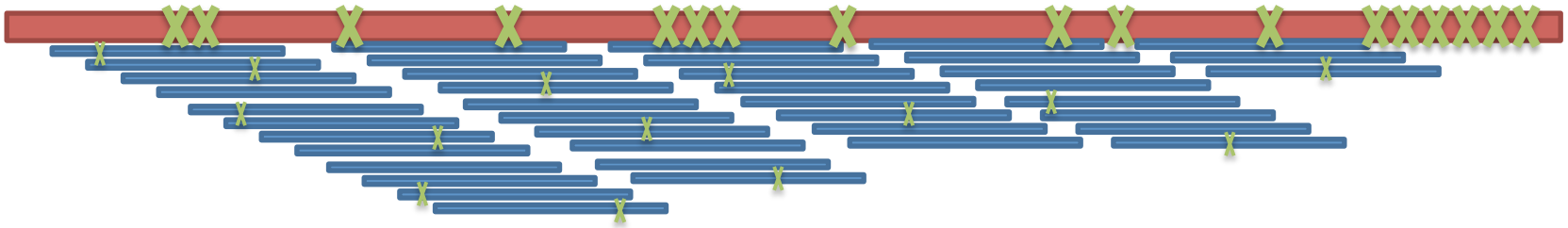
Collaboration with McCombie Lab (CSHL) and Pacific Biosciences

PacBio Error Correction

http://wgs-assembler.sf.net

1. Correction Pipeline
    1. Map short reads to long reads
    2. Trim long reads at coverage gaps
    3. Compute consensus for each long read

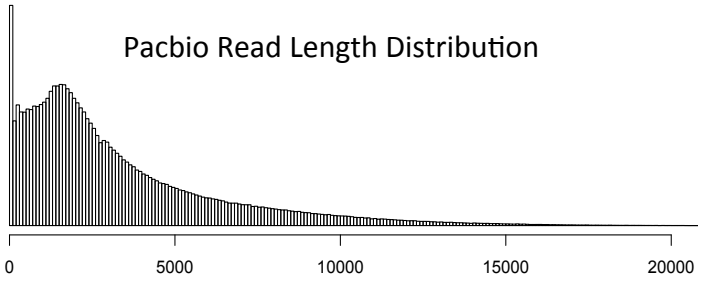
2. Error corrected reads can be easily assembled, aligned

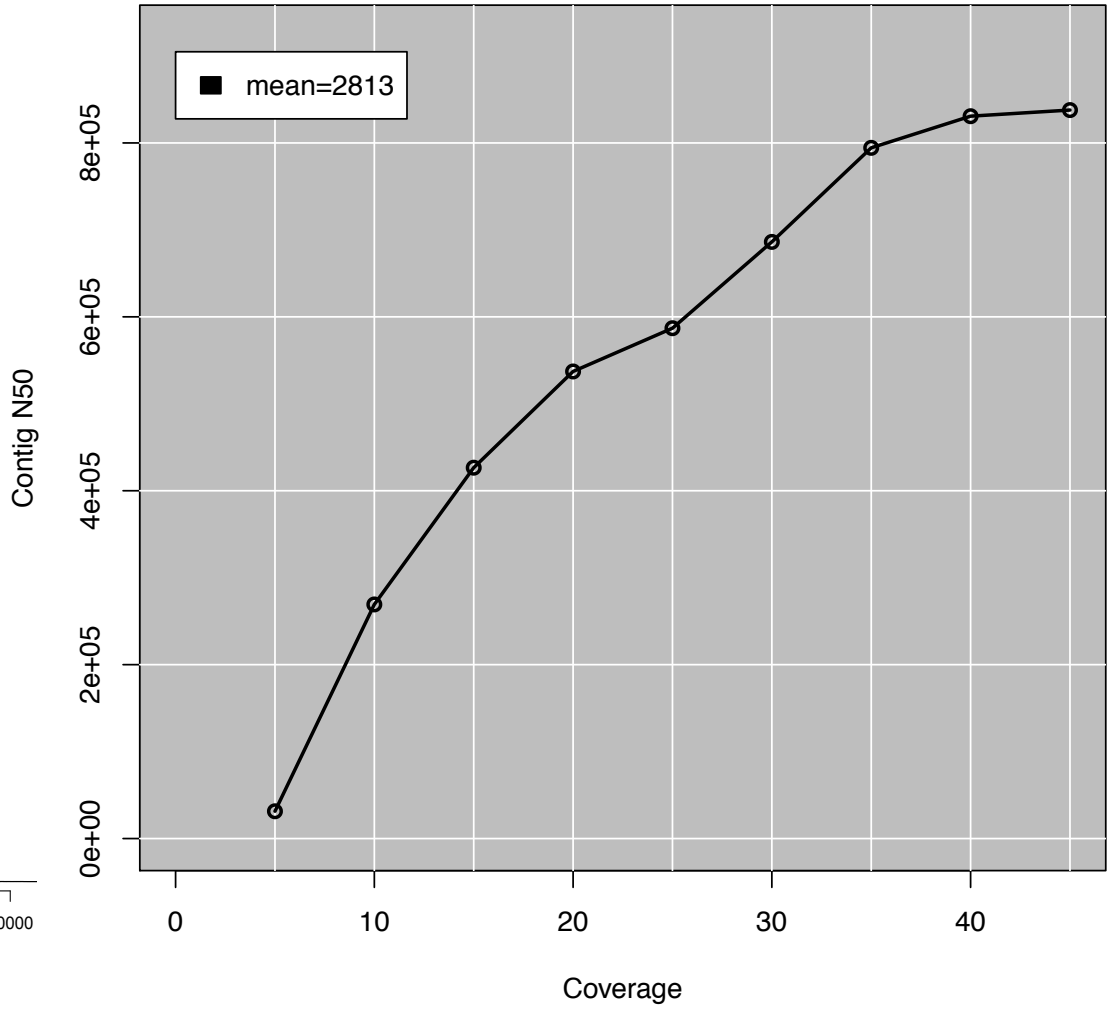


Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Koren, S, Schatz, MC, et al. (2012) *Nature Biotechnology*. doi:10.1038/nbt.2280

Assembly Coverage Model in Rice

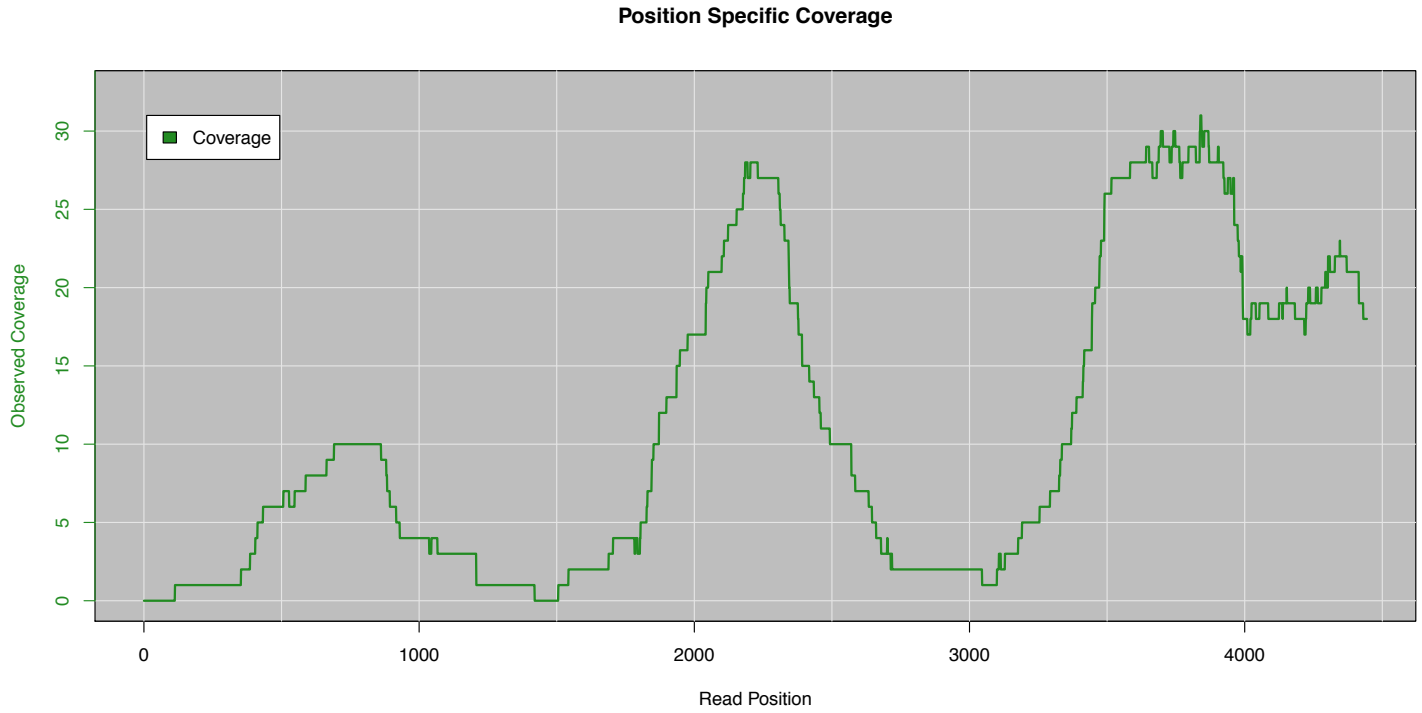
Simulate perfect reads from reference genome with Pacbio read length distribution





Assembly complexity of long read sequencing
Marcus, S, Lee, H, et al. (2013) *In preparation*

PacbioToCA Splits Reads



PacbioToCA splits reads at low coverage assuming they are adapters left over from the primary analysis pipeline

Has MaxGap Parameter: Do not split reads at low coverage. Suggested setting: 1500bp

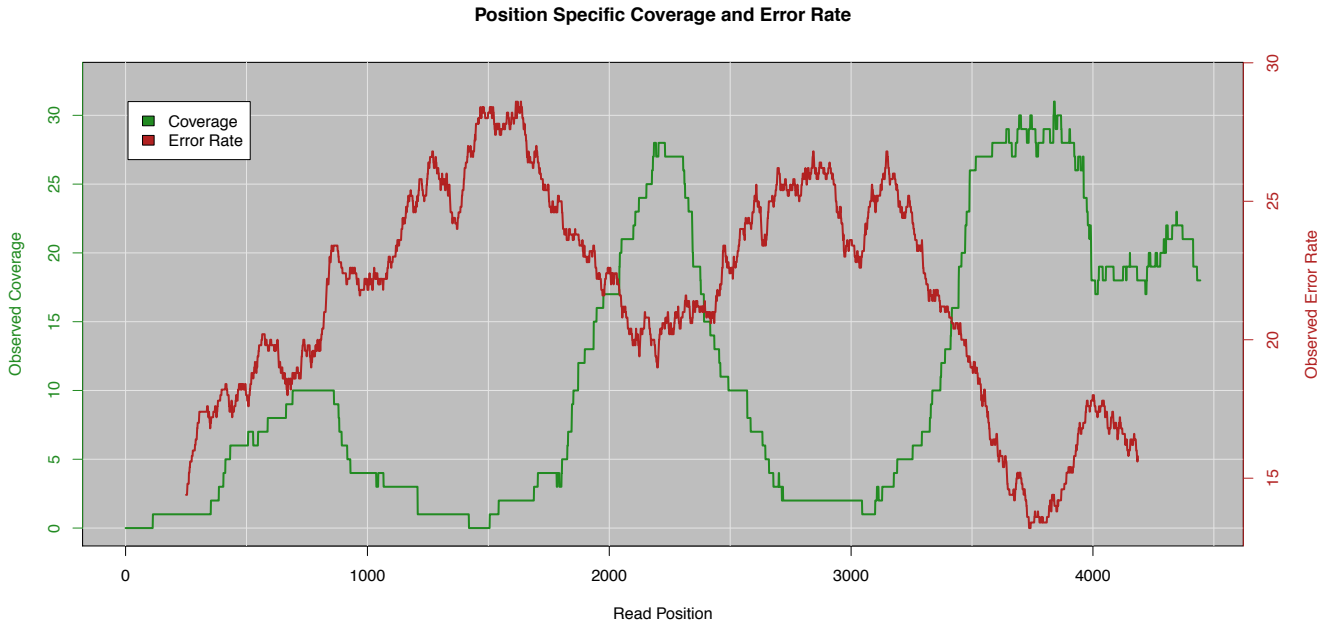
Rice Assembly

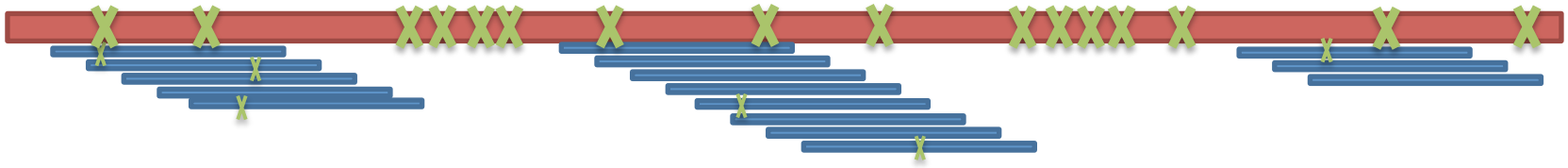
Assembler	Sequencing Data	N50 Contig
ALLPATHS	60x 101bp Illumina PE (180bp insert) 2k jump 5k jump	21kb
Celera	28x Flashed MiSeq 250 PE	4.5kb
Celera	19x Error Corrected Pacbio -pacbioToCA with flashed MiSeq -maxGap 1500	~~34kb~~ 58kb

What are these low coverage regions?

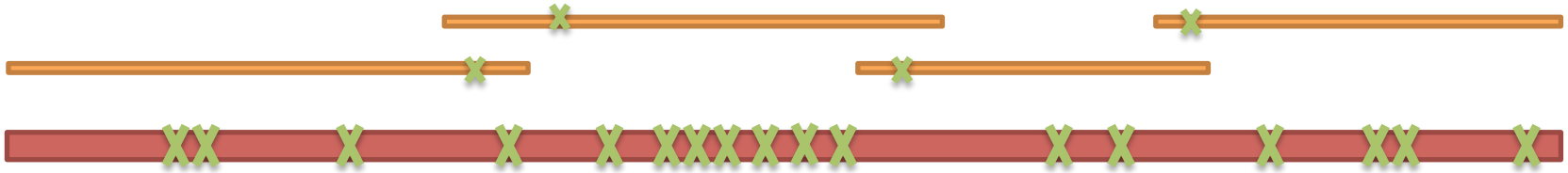
Low Coverage Regions

1. Simple Repeats – Kmer Frequency Too High to Seed Overlaps
2. GC Rich Regions – Known Illumina Bias
3. **Error Dense Regions – Difficult to compute overlaps with many errors**





Correction with Unitigs



Unitigs: High quality contigs formed from unambiguous, unique overlaps of reads

Can Help us overcome:

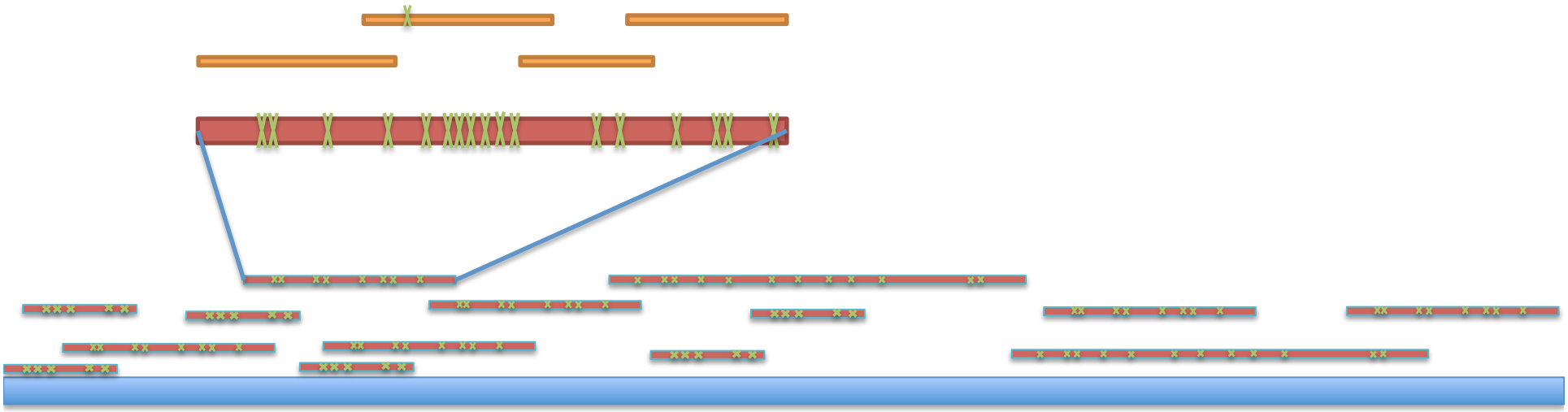
1. **Simple Repeats – Kmer Frequency Too High to Seed Overlaps**
2. GC Rich Regions – Known Illumina Bias
3. **Error Dense Regions – Difficult to compute overlaps with many errors**

Celera guarantees all reads are incorporated into a unitig

How Can We Align Unitigs?

Most Aligners Ask:

Where in the genome did this read come from? Blasr, Bowtie2, BWA, BLAST



We are looking for the set of Unitigs that came from the same genomic position as the Pacbio Read.

MUMmer/Nucmer

1. Features All-vs-All Alignments

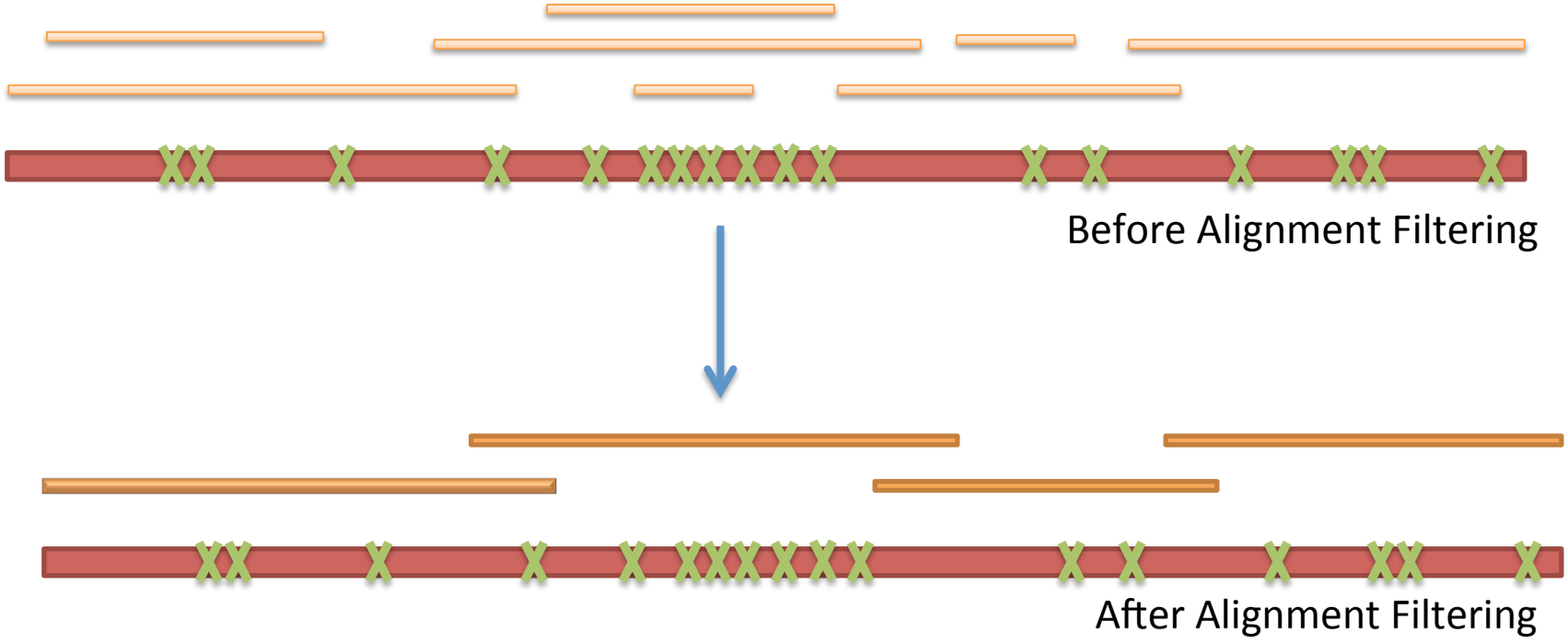
2. Delta-Filter
   - Dynamic Programming To Optimize layout of Unitigs with respect to PB Read

3. Show-Snps
   - Outputs differences between Pacbio Read and Unitig

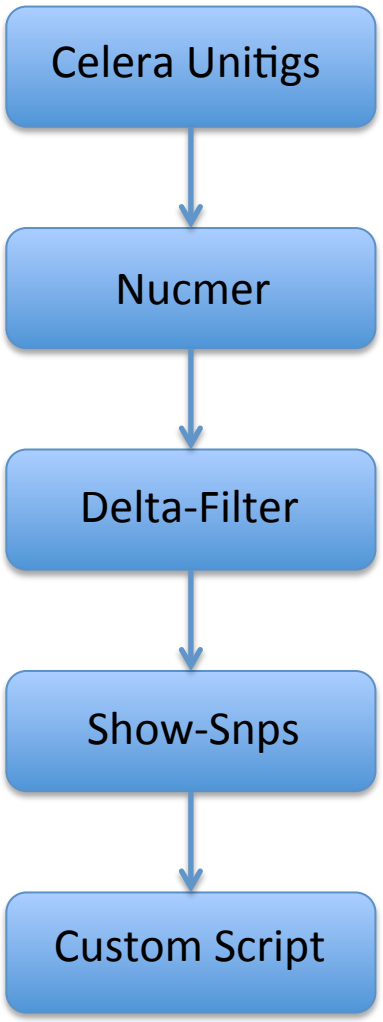
Fast algorithms for large-scale genome alignment and comparison.
Delcher AL, Phillippy A, Carlton J, Salzberg SL, Nucleic Acids Res. 2002 Jun 1;30(11):2478-83.

Delta-Filter

Uses Dynamic Programming (Longest Increasing Subset) to find the longest mutually consistent subset of unitigs with respect to the Pacbio Read



Pipeline Workflow



Generate Unitigs with Celera Using Illumina or another high identity sequencing technology

Align Unitigs to Pacbio Reads With Nucmer

Use Delta-Filter to Generate Unitig Layout

Show-Snps shows differences between trusted Illumina Unitig Sequence and Pacbio Read

Script to “Correct” Pacbio Read

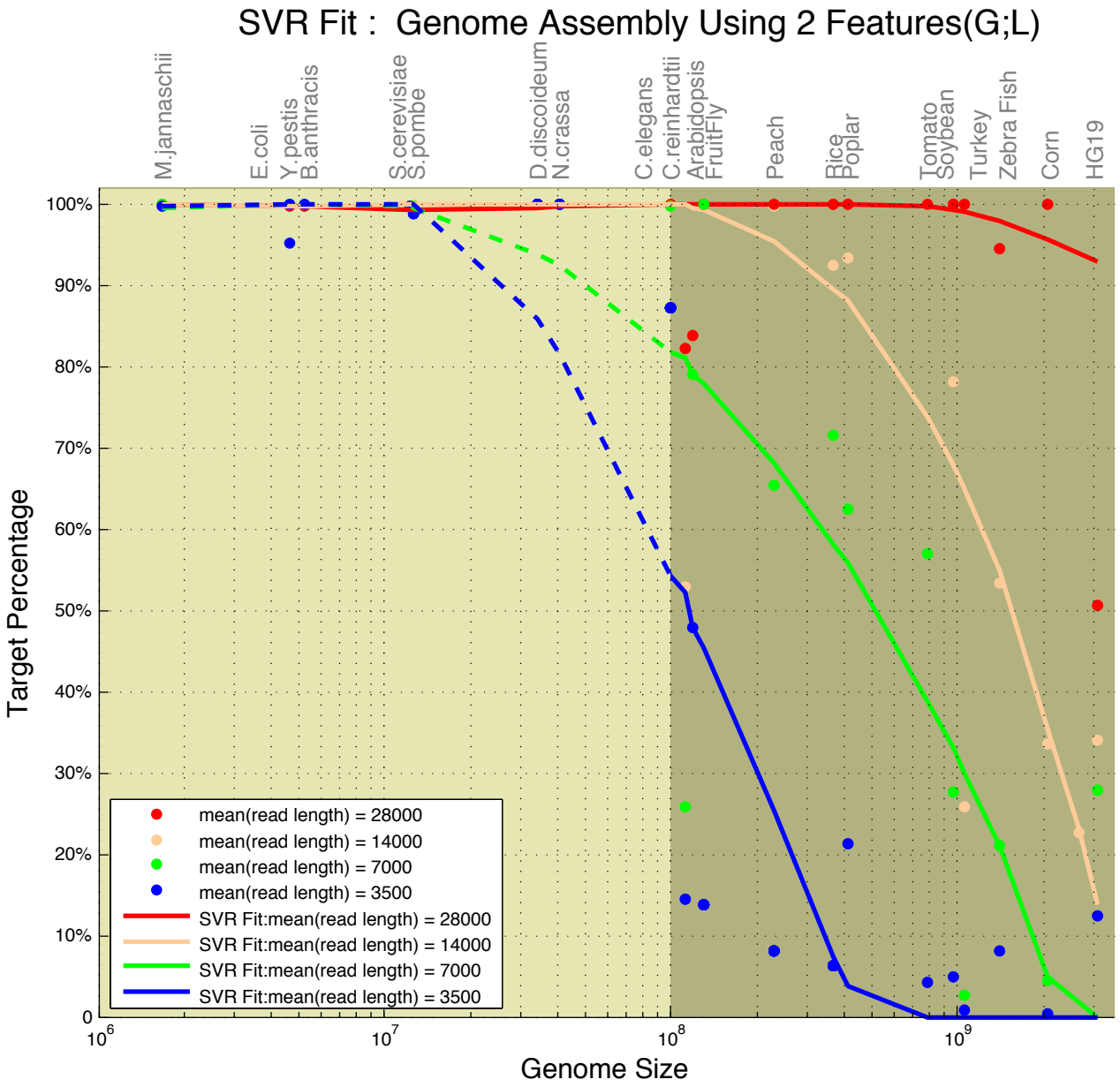
Note: Reads are never split or trimmed

Results

Assembler	Sequencing Data	N50 Contig
ALLPATHS	101bp Illumina PE (180bp insert) 2k jump 5k jump	21kb
Celera	28x Flashed MiSeq 250 PE	4.5kb
Celera	19x Error Corrected Pacbio -pacbioToCA pipeline with -flashed MiSeq library -maxGap=1500	~~34kb~~ 58kb
Celera	19x Error Corrected Pacbio -New Pipeline -flashed MiSeq for Unitig generation	155kb

Take Home Points from new Pipeline

- Aligning unitigs from Illumina assemblies instead of raw Illumina reads.
  - Repeats are compressed in Unitigs
    - Alignment more tractable
  - Unitigs help span clusters of errors
- Reads are never split
  - Leverage Celera Assembler
    - Overlap Based Trimming
    - Chimera Detection and Splitting



Assembly complexity of long read sequencing

Marcus, S, Lee, H, et al. (2013) *In preparation*

Acknowledgments

Schatz Lab

Michael Schatz
Shoshana Marcus
Hayan Lee
Alejandro Wences

CSHL

McCombie Lab