LA-UR-12-10336

Title: Genome Improvement with PacBio sequencing technology

Author(s):
Han, Shunsheng
Chertkov, Olga
Daligault, Hajnalka E.
Davenport, Karen Walston
Detter, John C.
Munk, A. Christine
Reitenga, Krista G.
Zhang, Xiaojing

Intended for:
DOE
AGBT, 2012-02-15/2012-02-18 (Marco Island, Florida, United States)
NMED
US EPA
Biological resources
Reading Room
RCRA

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer,is operated by the Los Alamos National Security, LLC for the National NuclearSecurity Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By acceptance of this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Departmentof Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

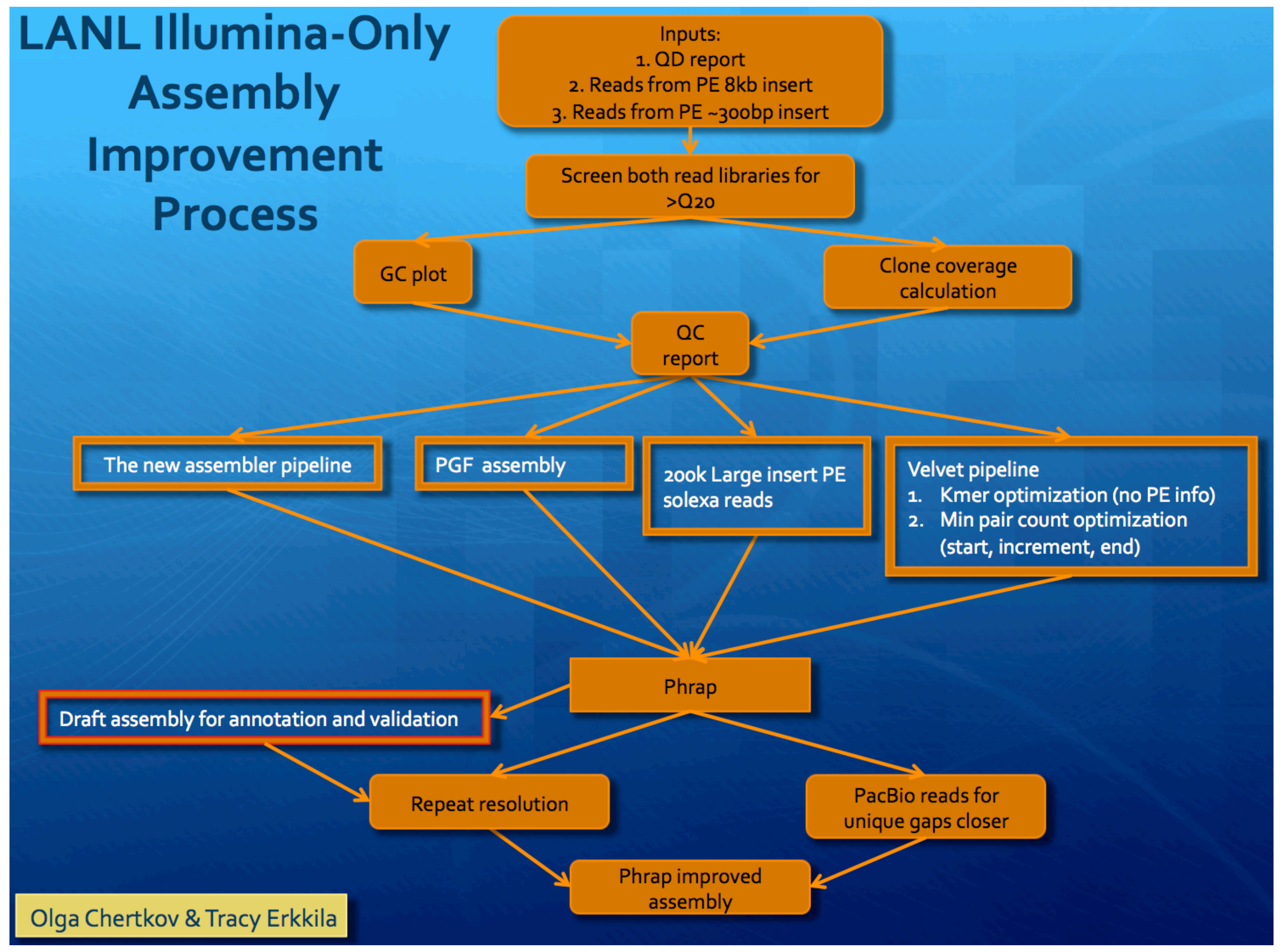
Genome Improvement with PacBio sequencing technology

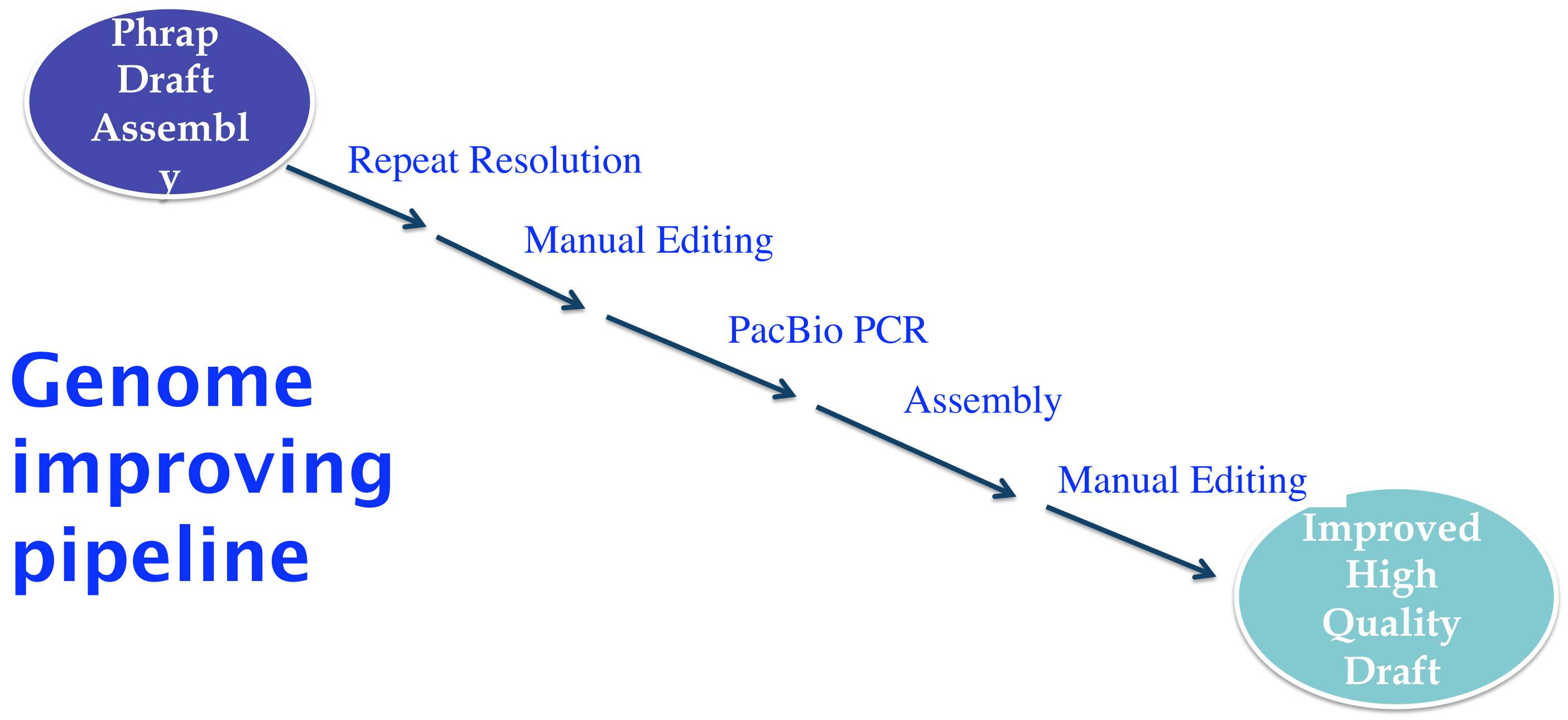
Cliff S. Han[1], Lucy (Xiaojing) Zhang[1], Krista Reitenga[1], Primo Baybayan[2], Susana Wang[2],
Karen W. Davenport[1], Hajnalka E. Daligault[1], Olga Chertkov[1], Chris Munk[1], and Chris Detter[1]

1 The DOE Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, NM 87545; 2 Pacific Bioscience, 1380 willow Rd. Menlo park, CA 94025.

LA-UR 12-01419

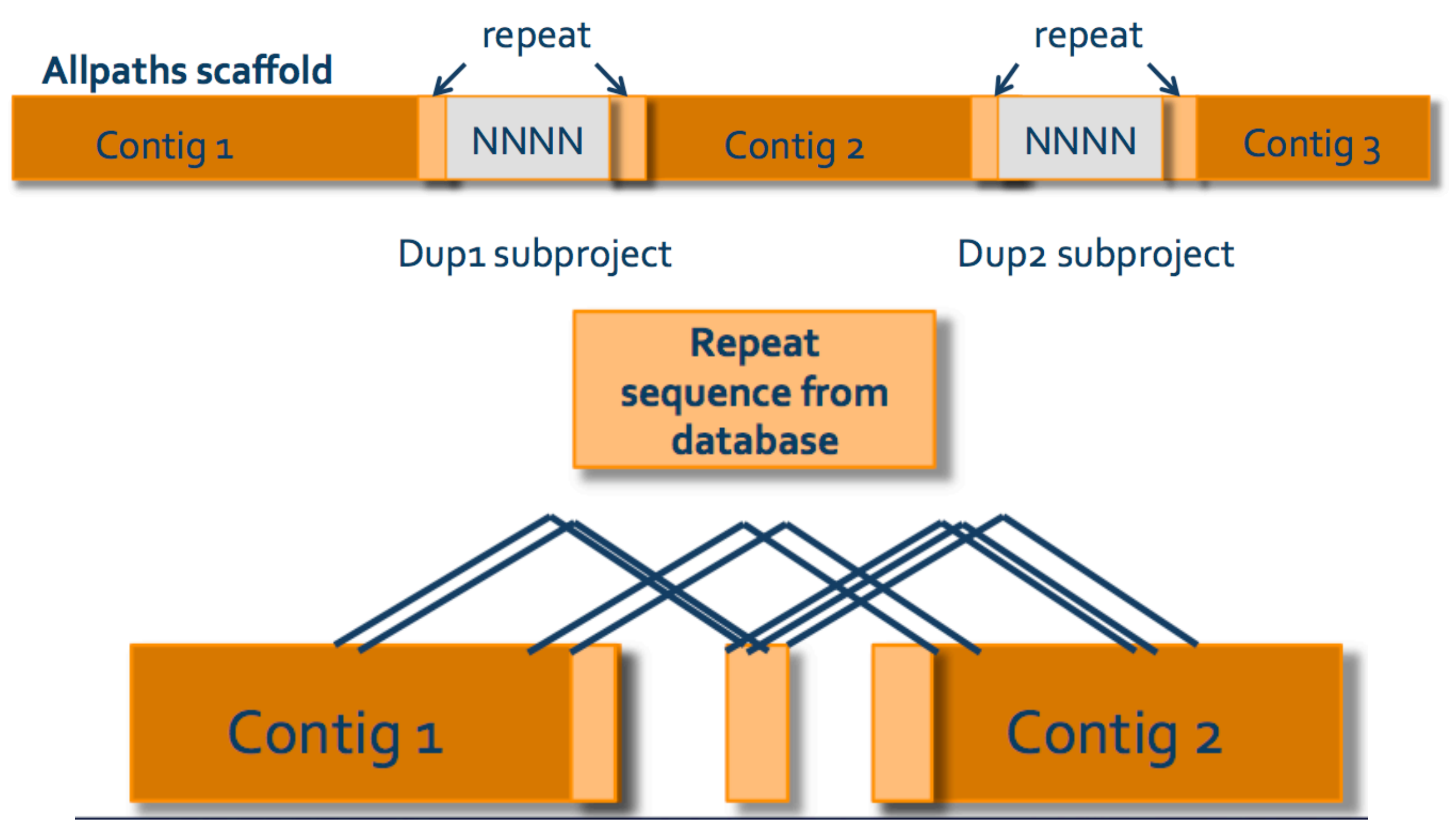
Abstract: As massively parallelized sequencing technologies enable draft genome sequencing at minimal cost, the cost for closing gaps in the draft genome with current Sanger methods becomes disproportionally high. The DOE Joint Genome Institute sequences hundreds of bacterial genomes each year. There could be dozens to hundreds of contigs in draft assemblies produced with only Illumina data, which at JGI include short (~300 bases) and medium (~8 Kb) insert libraries. A significant of them need to be finished/improved to a better level of continuity. Data from PacBio has been proven to improve continuity of assembly as it has better coverage in high or low GC regions comparing to Illumina technology. Unfortunately, the throughput of the RS currently does not compare to that of our Illumina pipeline yet. An alternate approach is to sequence the gaps and repetitive regions of genomes drafted with Illumina technology. Amplicon sequencing with PacBio RS machine has the potential to replace our current Sanger based genome finishing process. We used two finished genomes with high and medium GC content as test cases. Illumina data was reproduced with current technology with libraries of short and medium insert sizes. The two assemblies of the Illumina-only data produced 98 and 105 contigs in a limited number of scaffolds. PCR products from those gaps were pooled and sequenced with a PacBio machine. Subreads were aligned to known sequences next to PCR primer regions. Subreads that belong to a single PCR amplicon were assembled separately. The assembled sequences were aligned with the primer sequences again. Sequences between the primer pair are used to close gaps in the main project. Hard stops with 30 bp hairpin structure can be sequenced with PacBio without problem.


LANL Illumina-Only Assembly Improvement Process
Inputs:
1. QD report
2. Reads from PE 8kb insert
3. Reads from PE ~300bp insert
Screen both read libraries for >Q20
GC plot
Clone coverage calculation
QC report
The new assembler pipeline
PGF assembly
200k Large insert PE solexa reads
Velvet pipeline
1. Kmer optimization (no PE info)
2. Min pair count optimization (start, increment, end)
Phrap
Draft assembly for annotation and validation
Repeat resolution
PacBio reads for unique gaps closer
Phrap improved assembly
Olga Chertkov & Tracy Erkkila



Genome improving pipeline

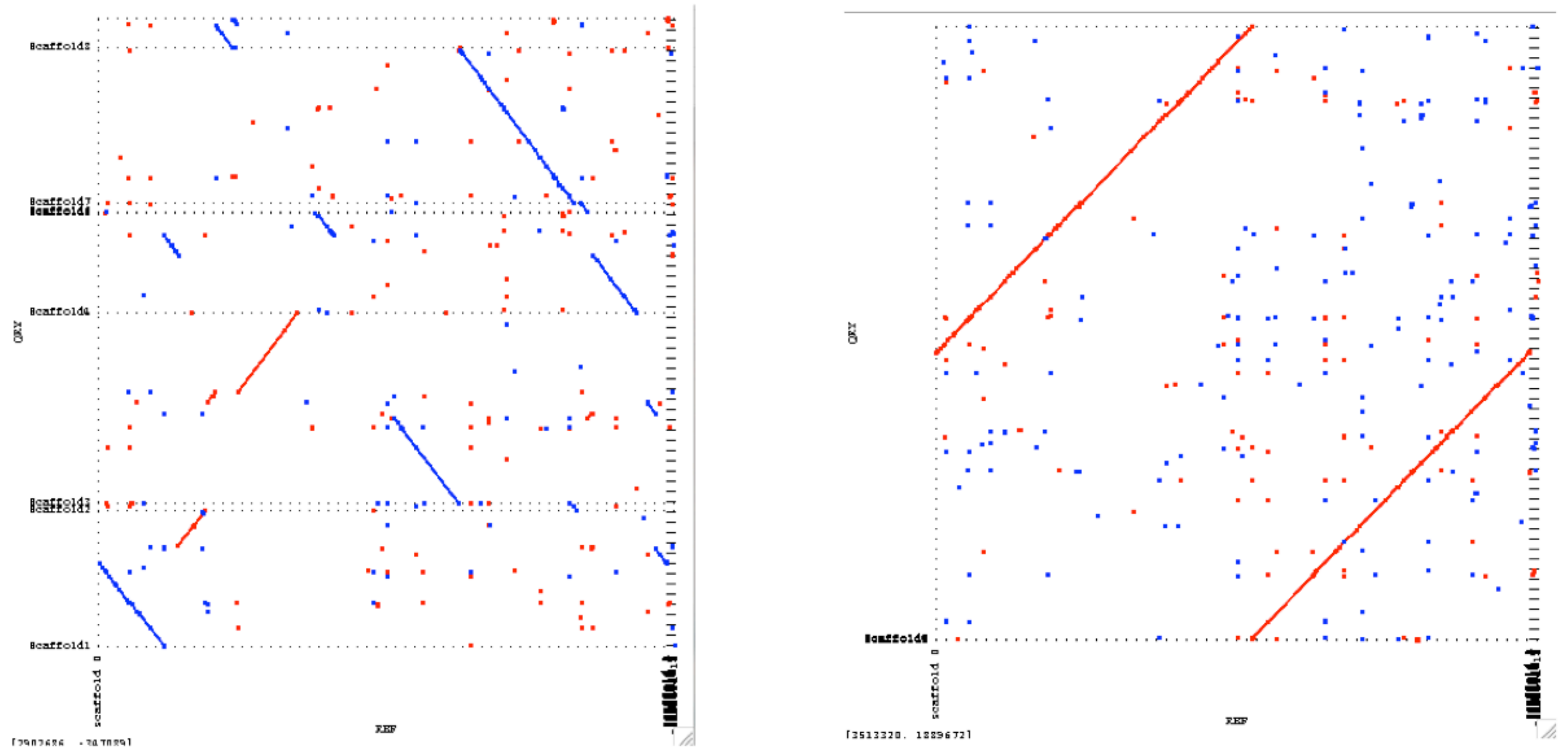
Repeat Resolution software



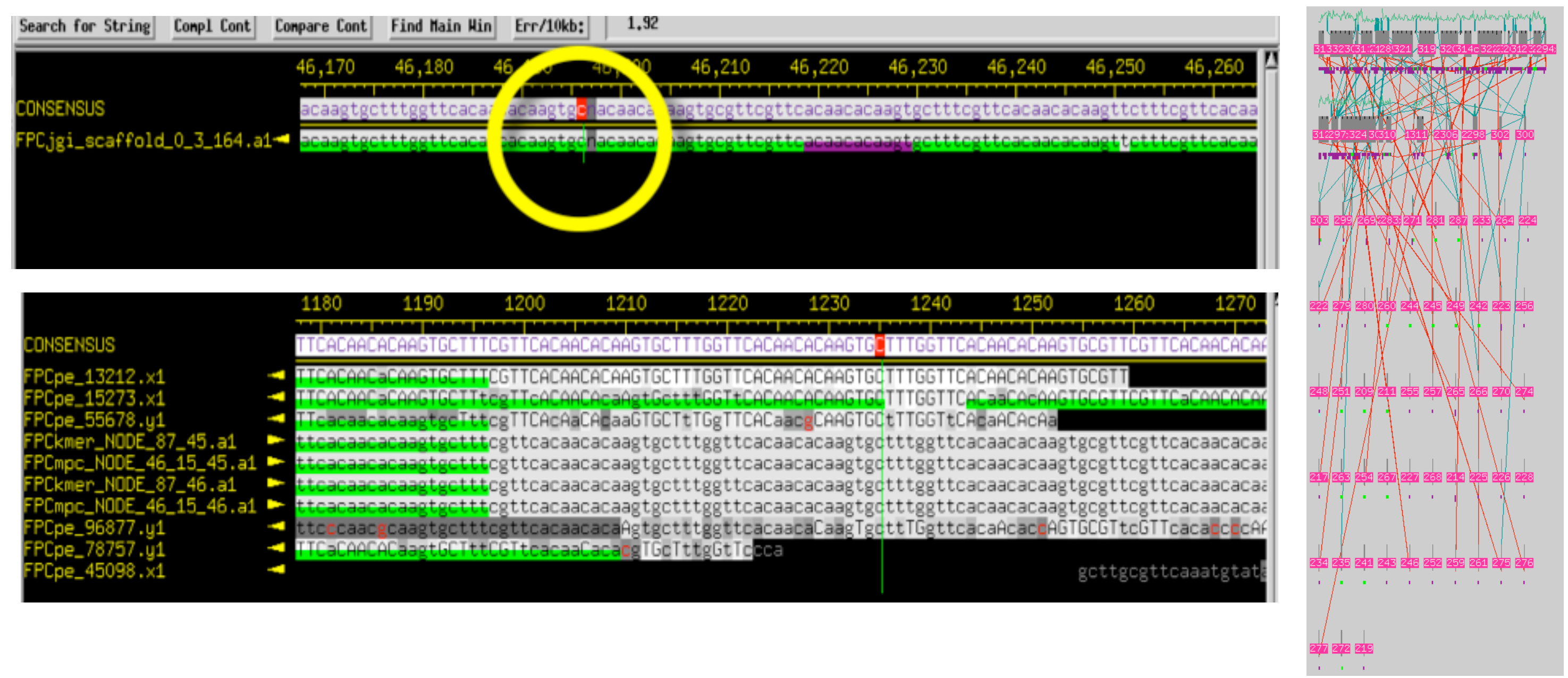
Scaffolds: Allpaths is more accurate

Allpaths draft vs. misassembled Startup draft

Allpaths draft vs. improved Intermediate draft

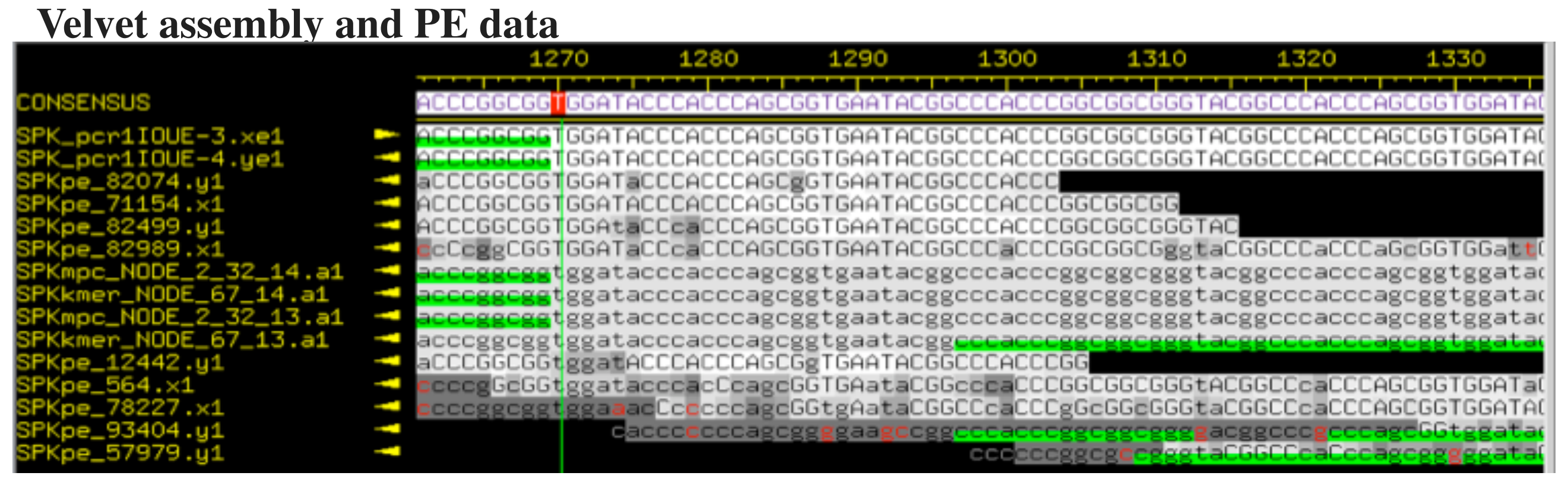


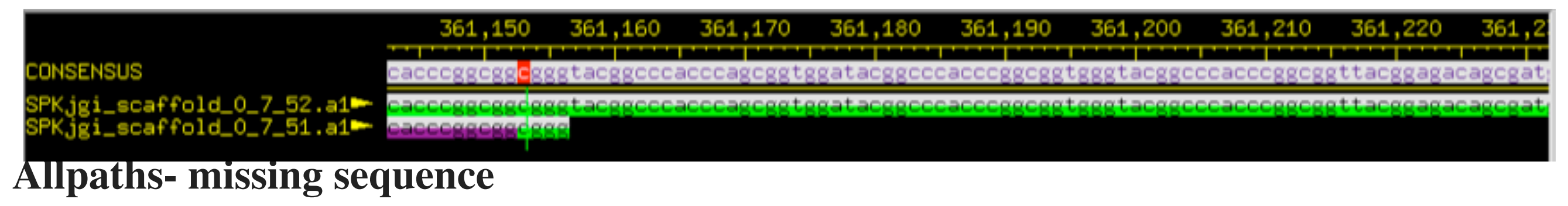
Sequence: Velvet is more accurate



Missing sequence in Allpaths assembly

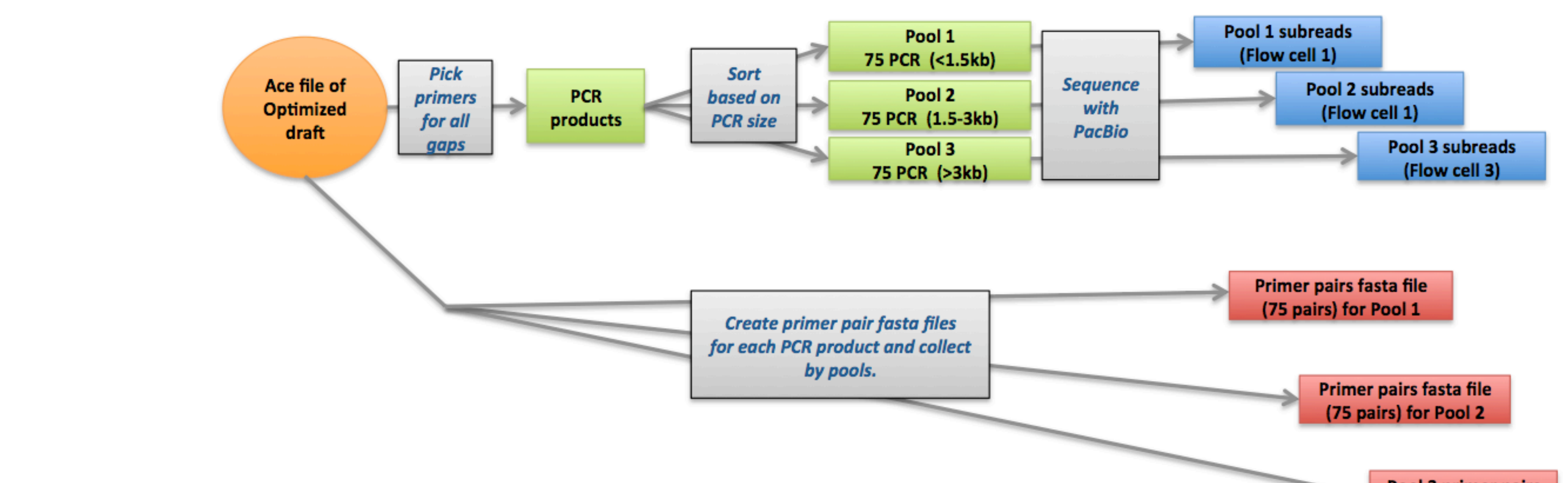
Velvet assembly and PE data

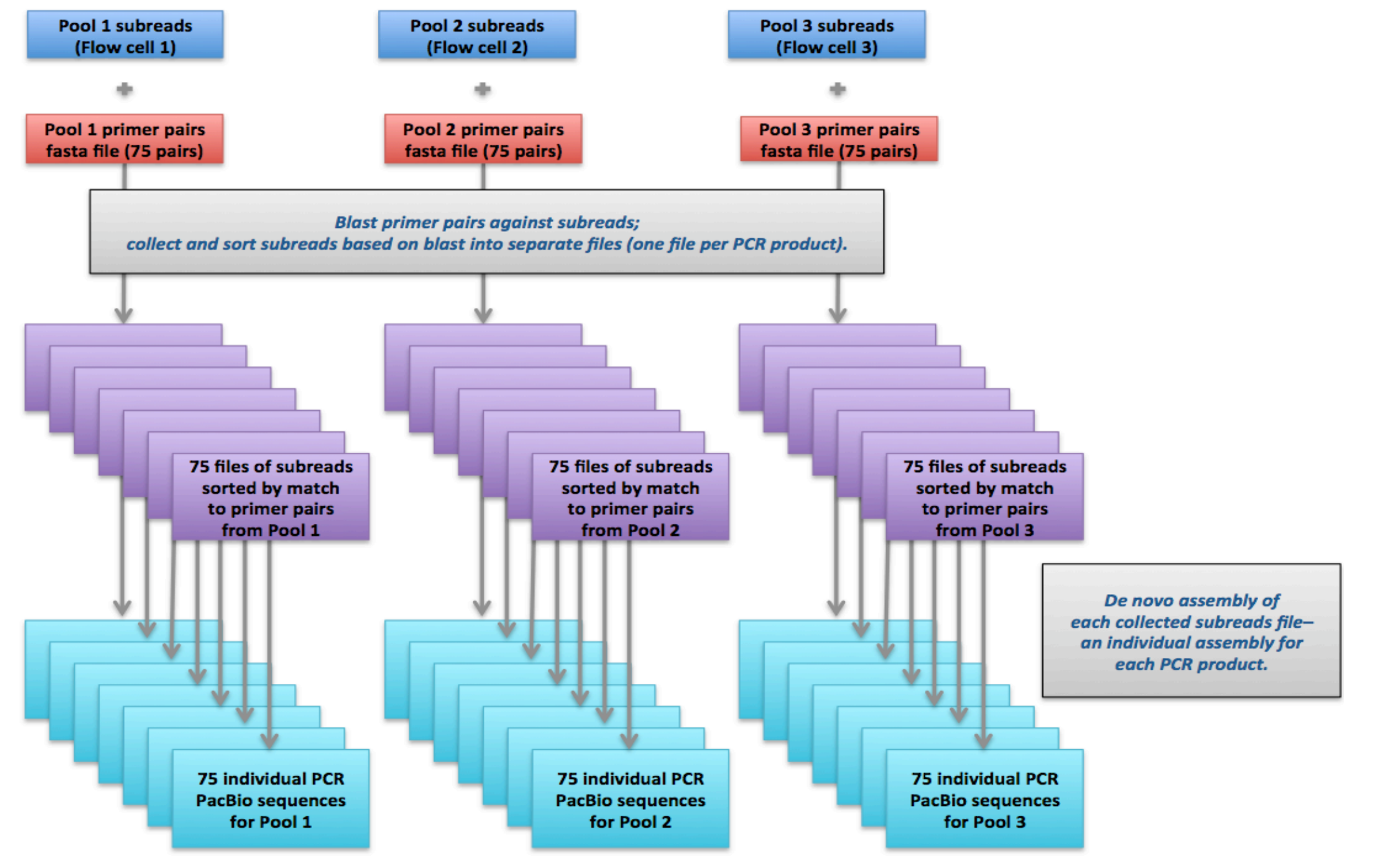


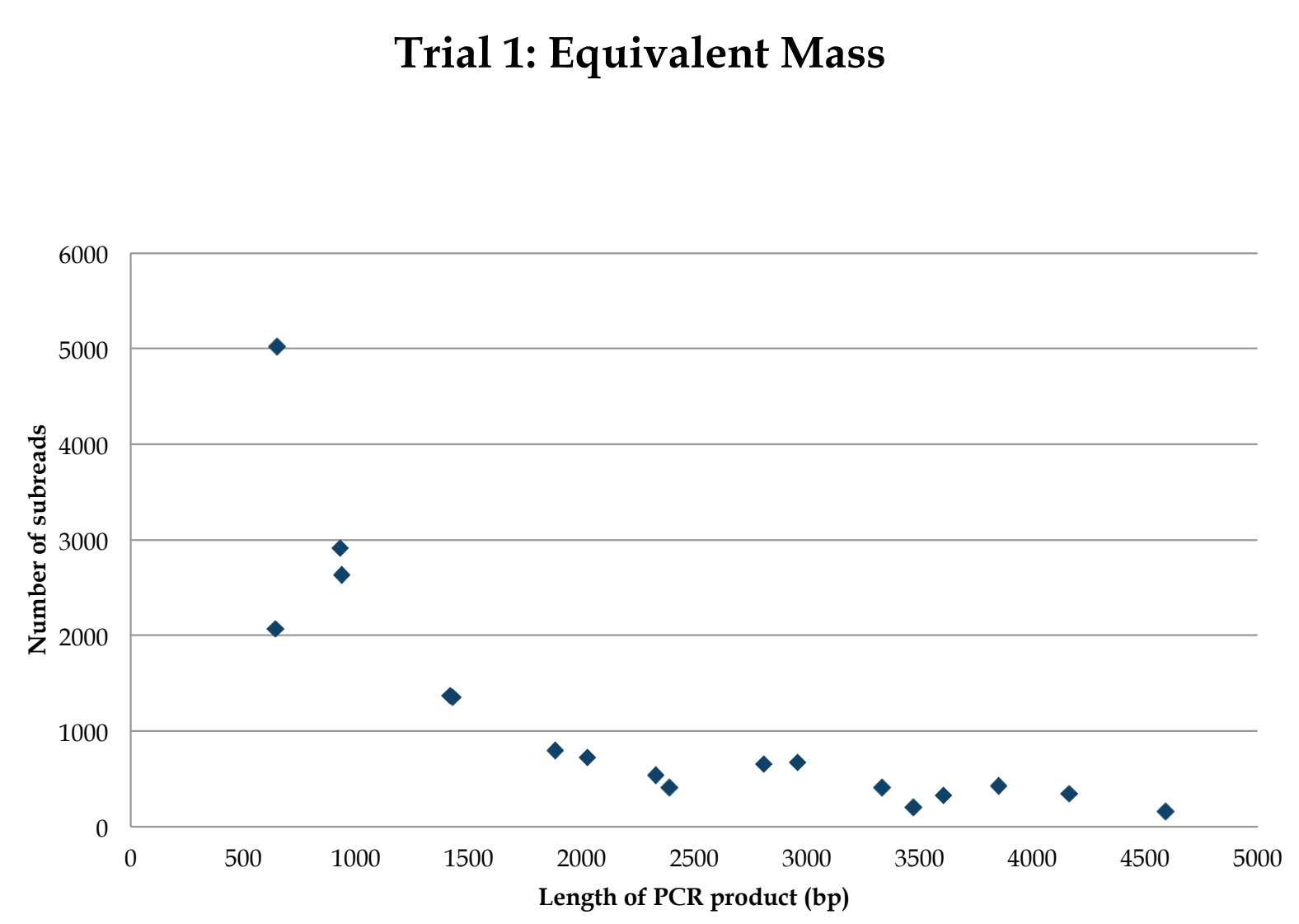


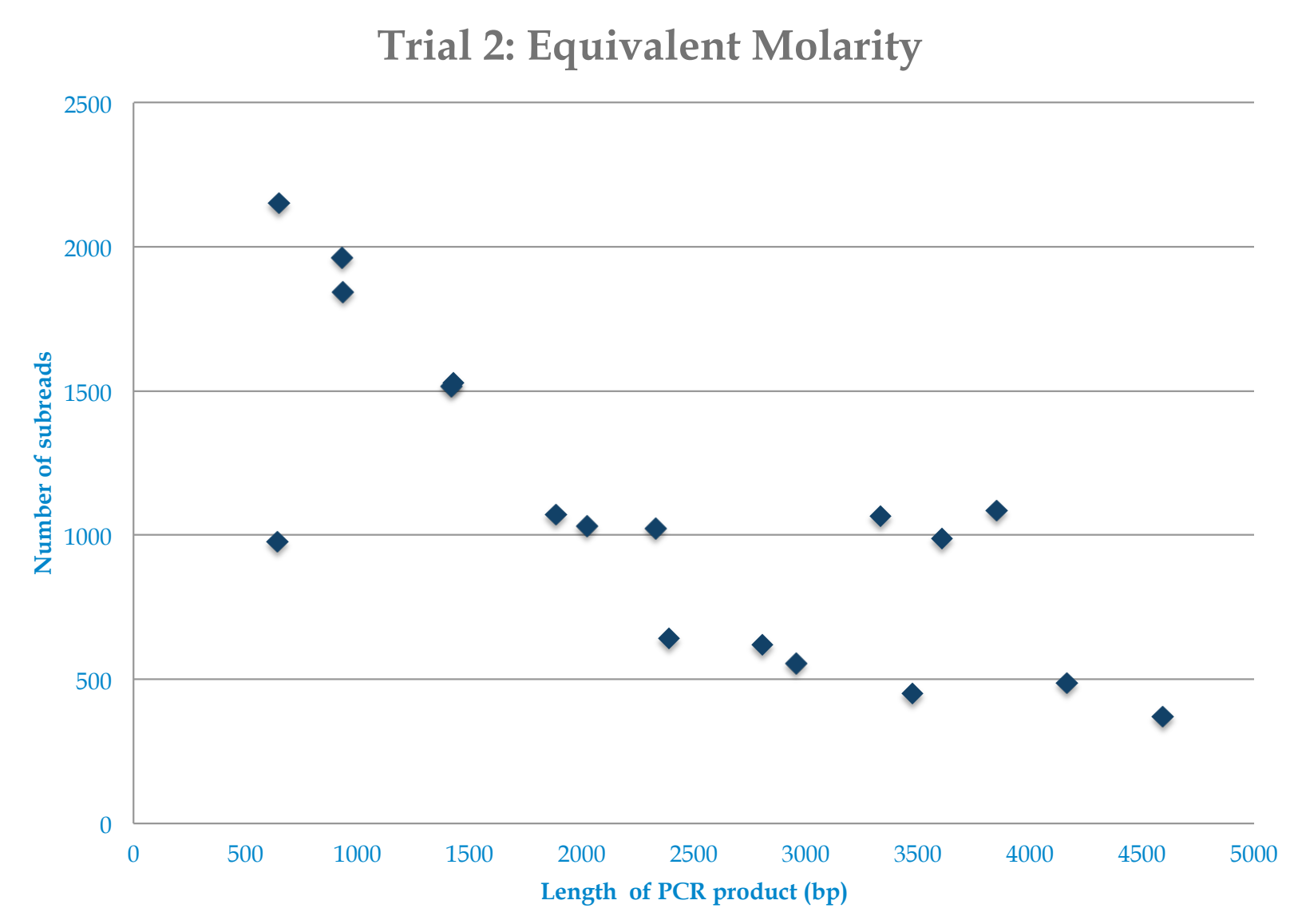
Allpaths- missing sequence

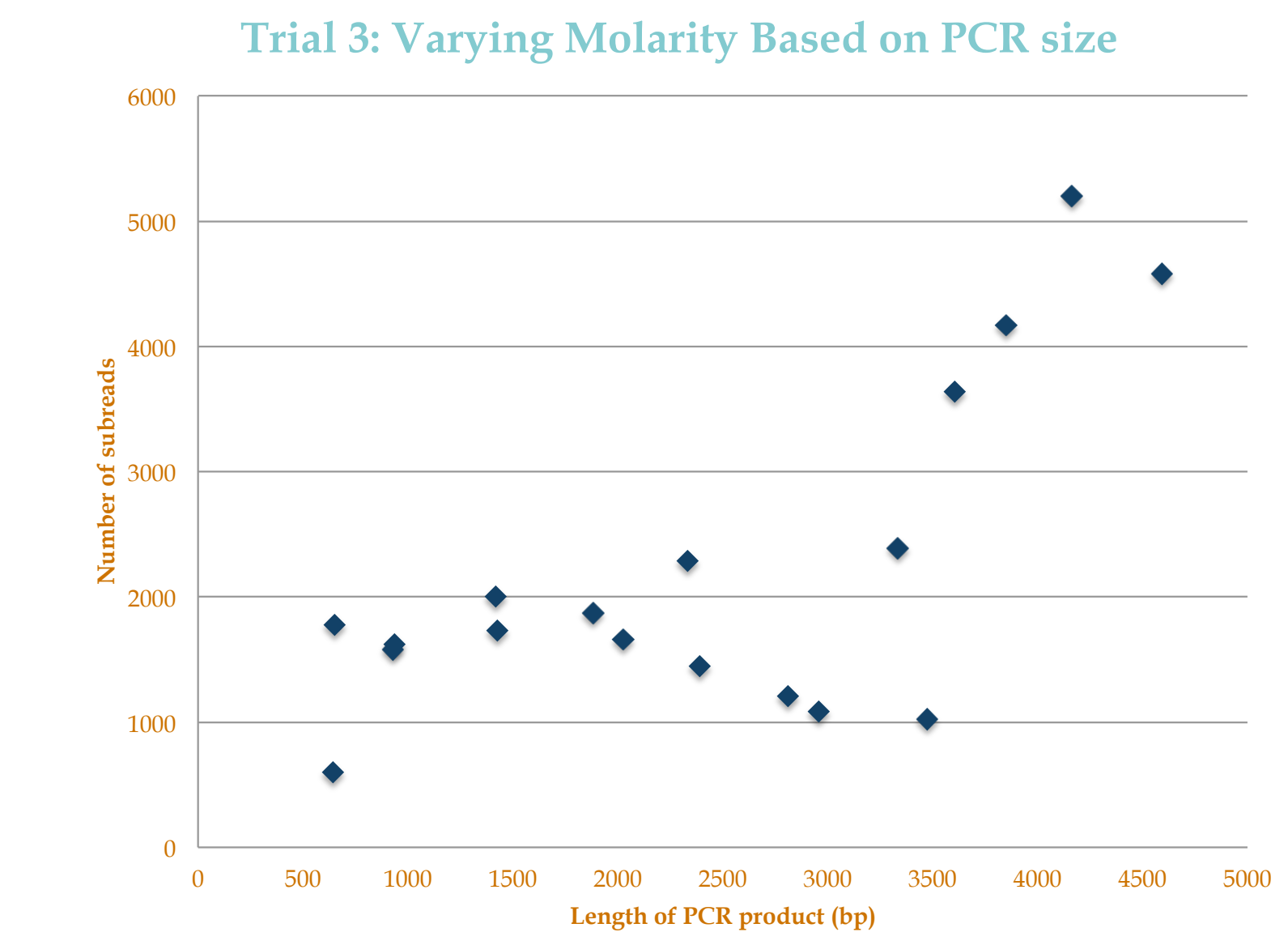
PacBio Sequencing of PCR for Gap Closure


Ace file of Optimized draft
Pick primers for all gaps
PCR products
Sort based on PCR size
Pool 1 75 PCR (<1.5kb)
Pool 2 75 PCR (1.5-3kb)
Pool 3 75 PCR (>3kb)
Sequence with PacBio
Pool 1 subreads (Flow cell 1)
Pool 2 subreads (Flow cell 1)
Pool 3 subreads (Flow cell 3)
Create primer pair fasta files for each PCR product and collect by pools.
Primer pairs fasta file (75 pairs) for Pool 1
Primer pairs fasta file (75 pairs) for Pool 2
Pool 3 primer pairs fasta file (75 pairs)

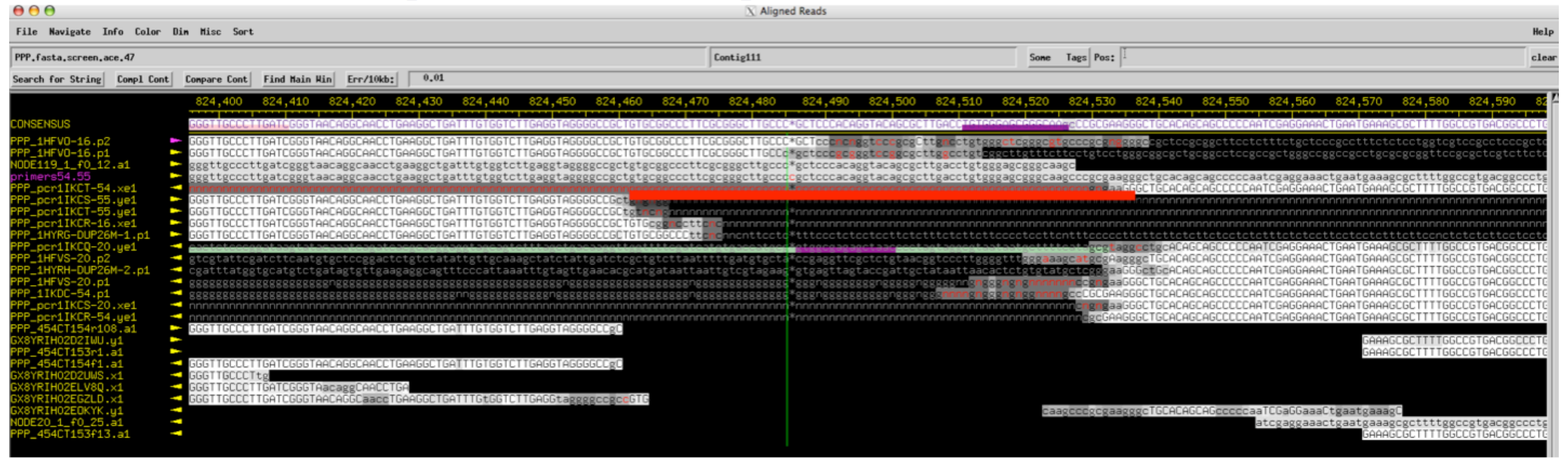

Pool 1 subreads (Flow cell 1)
Pool 2 subreads (Flow cell 2)
Pool 3 subreads (Flow cell 3)
Pool 1 primer pairs fasta file (75 pairs)
Pool 2 primer pairs fasta file (75 pairs)
Pool 3 primer pairs fasta file (75 pairs)
Blast primer pairs against subreads; collect and sort subreads based on blast into separate files (one file per PCR product).
75 files of subreads sorted by match to primer pairs from Pool 1
75 files of subreads sorted by match to primer pairs from Pool 2
75 files of subreads sorted by match to primer pairs from Pool 3
De novo assembly of each collected subreads file– an individual assembly for each PCR product.
75 individual PCR PacBio sequences for Pool 1
75 individual PCR PacBio sequences for Pool 2
75 individual PCR PacBio sequences for Pool 3







Hard stops are no problem with PacBio



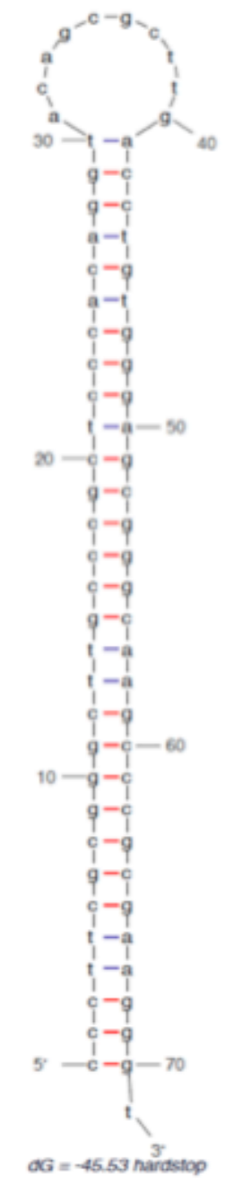
cccttcgcgggcttgcccgctcccacaggtacagcgcttgacctgtgggagcgggcaagcccgcgaaggg

30bp long

30bp long

hairpin loop

An example of a sequence that creates a hairpin loop in the secondary structure which has been very difficult to complete when sequencing a genome.



Shown in the sequences above, only the PacBio read (underlined in red) goes through the hard stop region. Illumina, 454, and Sanger (PCR) reads didn't.