QUESTIONS?

- PacBio read error is predominantly deletions
- There is no magic formula to estimate the best parameters for running FALCON to get the optimum assembly out of your data.
  - The understanding of your target genome, repeat structures, GC contents
  - The understanding of your own data, length distribution, coverage depth
- DALIGNER and String graph algorithms were developed by Eugene Myers
- DALIGNER is 20 to 40 times faster than BLASER
- Quiver polishing of the draft assembly is run outside FALCON
  - Requires .bax.h5 files
  - Might have to split data for alignment step, then merge all alignment files (cmp.h5) and sort it before Quiver step.

FALCON_Unzip

Jie (Jessie) Li
PhD
Bioinformatics Analyst
Bioinformatics Core
UCD

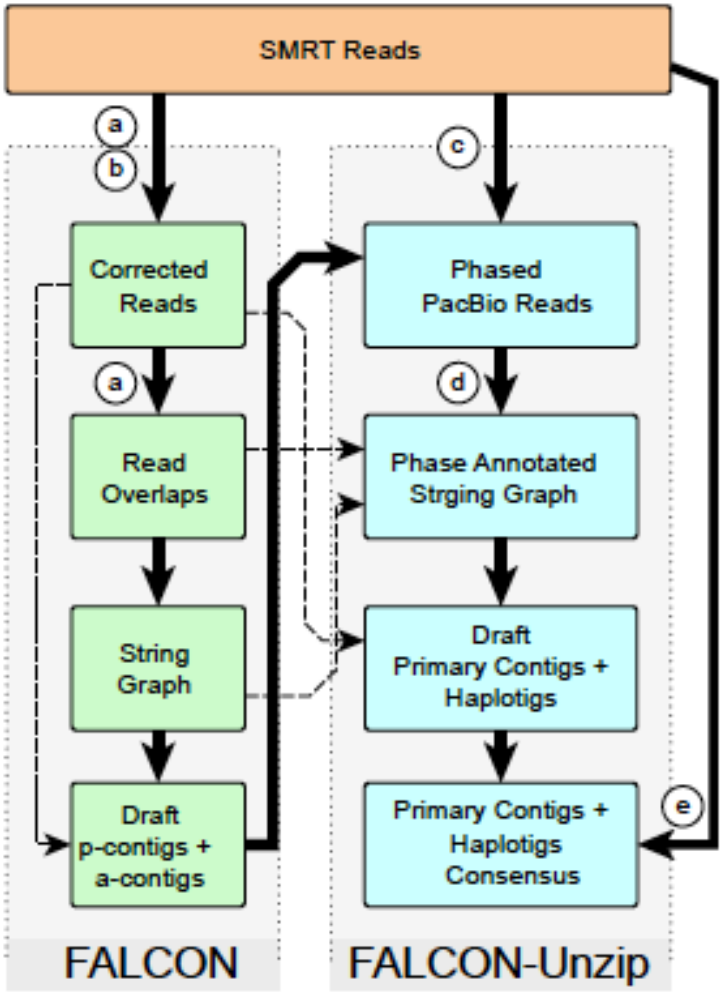
- Copy necessary files for running falcon_unzip to your own directory

```
srun --reservation=workshop --pty /bin/bash

mkdir falcon_unzip

cd falcon_unzip

cp /share/biocore/workshops/Genome-Assembly-Workshop/examples/falcon_unzip/input* .
```

Ploidy in Genome Assembly

- Most assemblers are designed for haploid genomes.
  - Works well for a diploid genome with little structural variation between the chromosome copies, with occasional structural heterozygosity appearing as separate contigs.
  - In diploid genomes with larger structural variation or multiploid genomes, assemblies are more fragmented.
- Most available methods proposed for diploid assembly tend to produce highly fragmented results, with contigs averaging just a few hundred bases to several kilobases in length.
- Sequencing both parents and offsprings to infer haplotypes, but requires sequencing additional samples and is fundamentally limited in contiguity of the initial assemblies.
- Pooled clonal fosmid sequencing produces diploid sequences but is expensive, labor intensive and the assembly contiguity is limited by the clonability of the source DNA, and the size and quality of the sequenced fosmids.

FALCON_Unzip

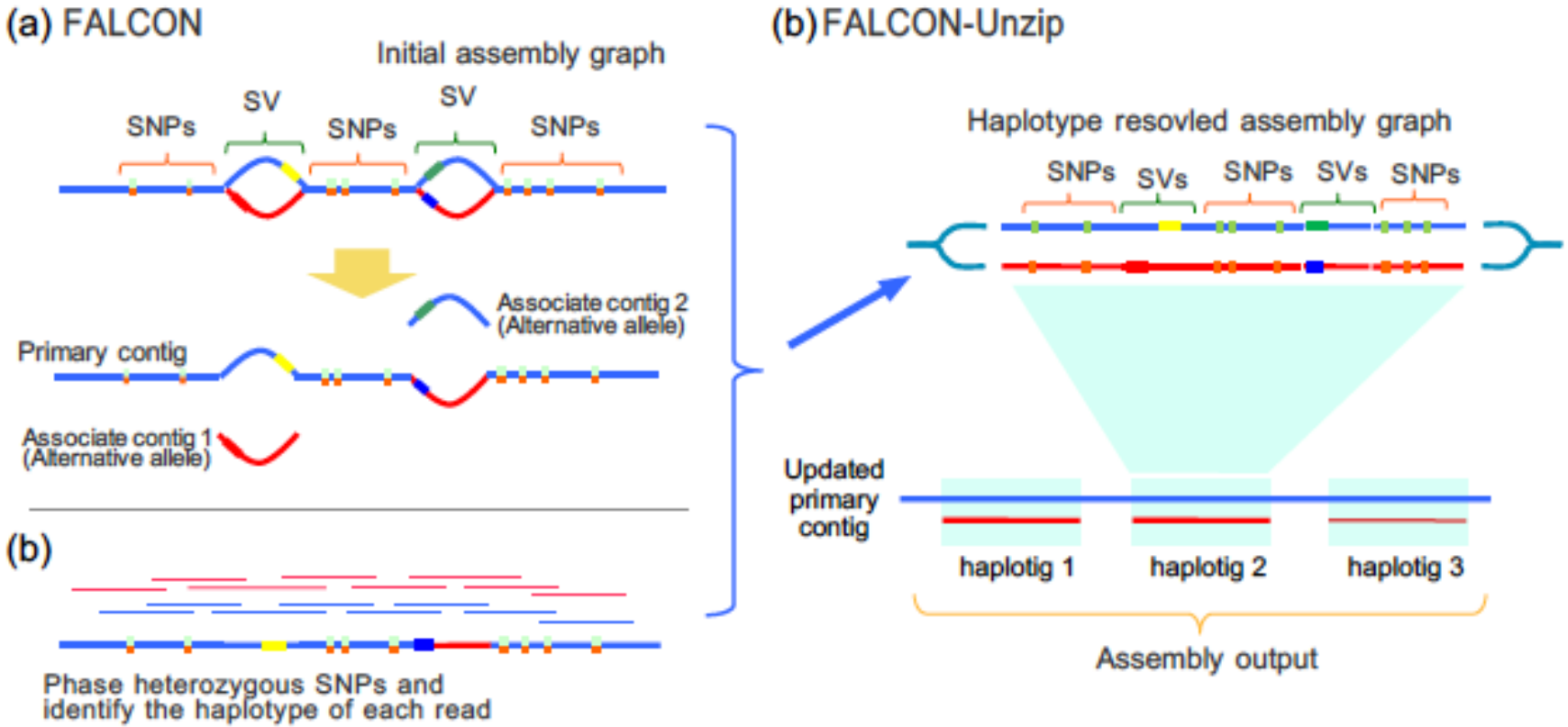


External code and internal modules used in FALCON and FALCON-Unzip

- (a) Daligner
- (b) Consensus Module (FALCON-sense)
- (c) Phasing Module (FALCON-phasing)
- (d) Graph "Unzip" Module
- (e) BLASR Alignment+ Quiver Consensus Module

http://dx.doi.org/10.1101/056887

FALCON_Unzip



http://dx.doi.org/10.1101/056887

FALCON_Unzip

- Step 1: track PacBio reads to each primary contig and generate alignments between the reads and the contigs.
  - Examine the tiling path of each of the contigs. The reads corresponding to the tiling path are assigned to the contig.
  - Examine the overlapping data generated from the assembly process. Score each pair overlap by the overlap length. If the best pair overlap read already has been assigned to a contig, the unassigned query read will be assigned to the same contig.
  - The reads are grouped into sets that are associated with each primary contig. Each contig and its associated reads are aligned independently using BLASR.

http://dx.doi.org/10.1101/056887

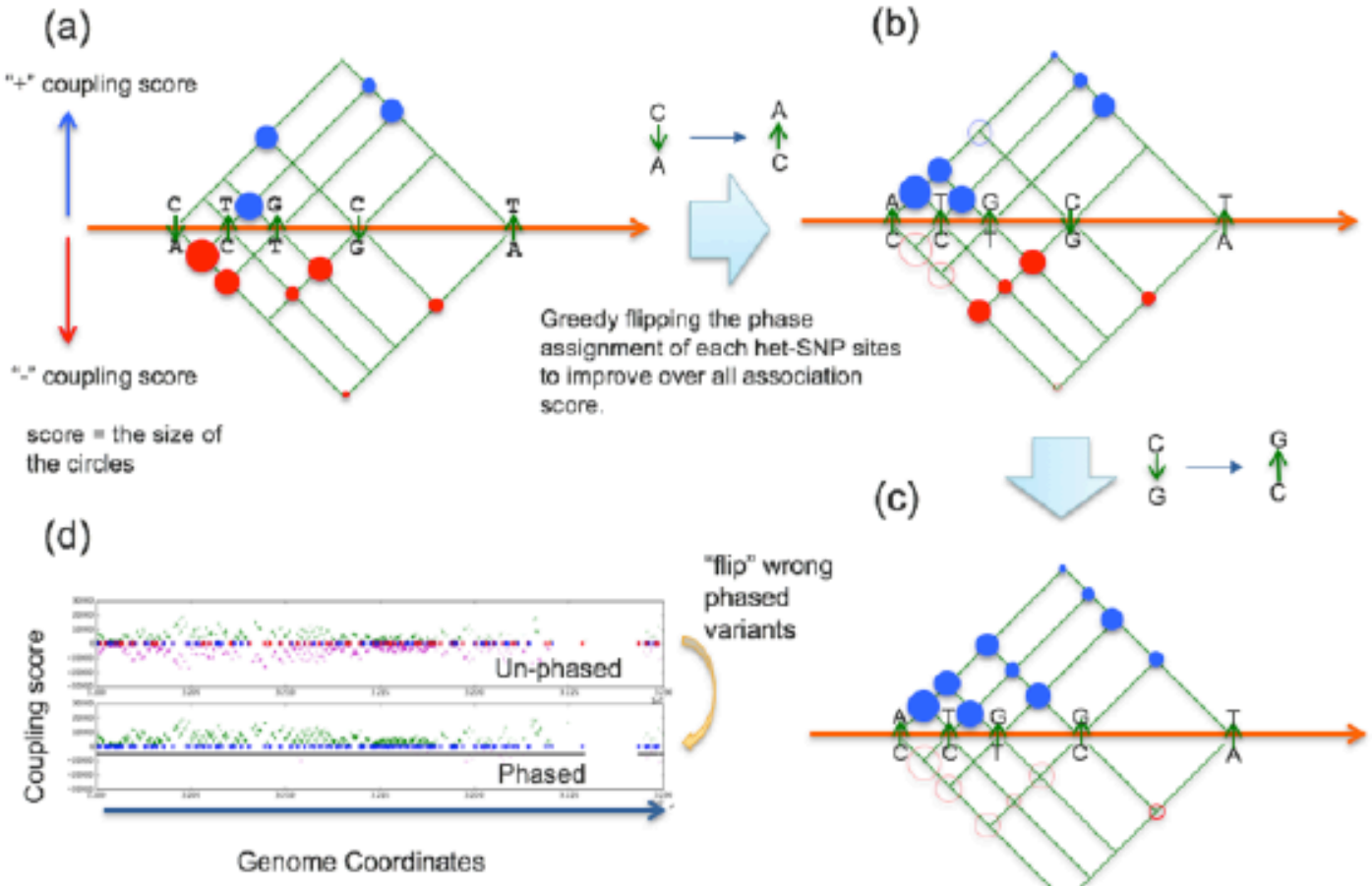
FALCON_Unzip

- Step 2: call heterozygous SNPs.
  - Focus on SNPs and ignor SNPs that have insertions or deletions nearby.
  - For each base, count the number of each A, C, G, T from aligned reads.
  - If the highest count is less than 75% and the second highest count is more than 25%, call it a heterozygous SNP site and will be used for phasing downstream.

http://dx.doi.org/10.1101/056887

FALCON_Unzip

- Step 3: phase heterozygous SNPs for each contig.



http://dx.doi.org/10.1101/056887

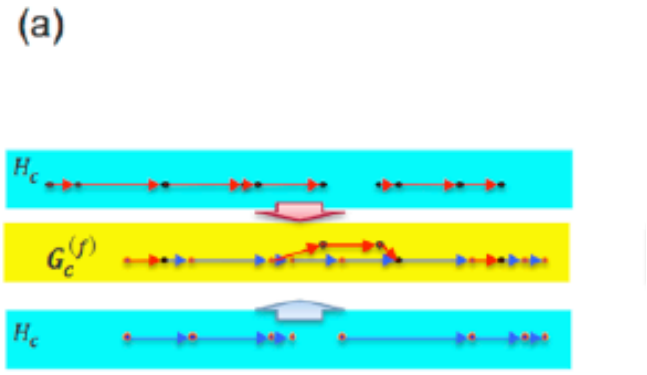
FALCON_Unzip

- Step 4: assign phase to the raw reads.
  - Examine the alignment to the reference.
  - Count which phases it agrees the most and assign “block identifier” and “phase identifier” to the read.

http://dx.doi.org/10.1101/056887

FALCON_Unzip

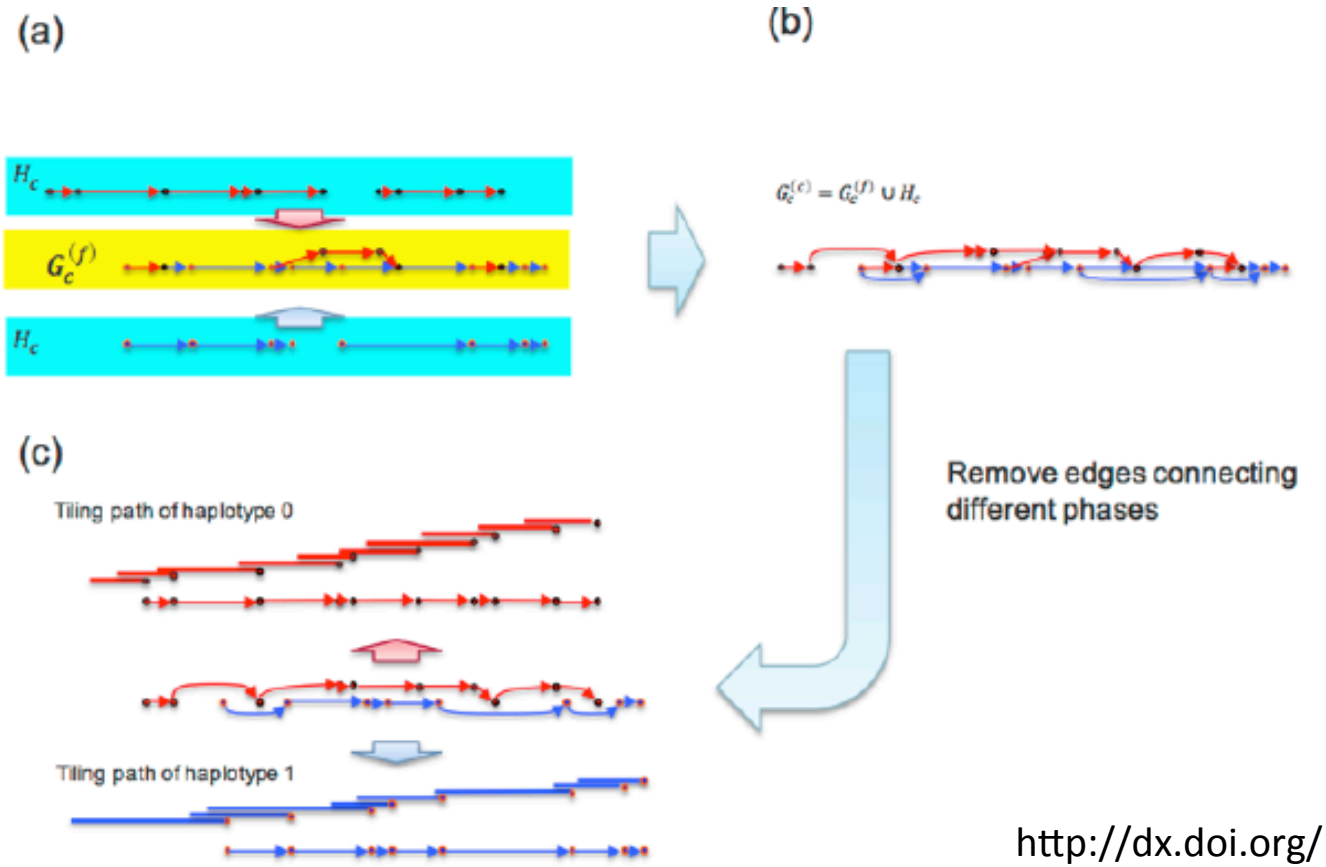
- Step 5: construct a haplotype-specific assembly graph from all reads that mapped to it (H_c), ignoring the overlaps between any two reads from the same block but different phases.



http://dx.doi.org/10.1101/056887

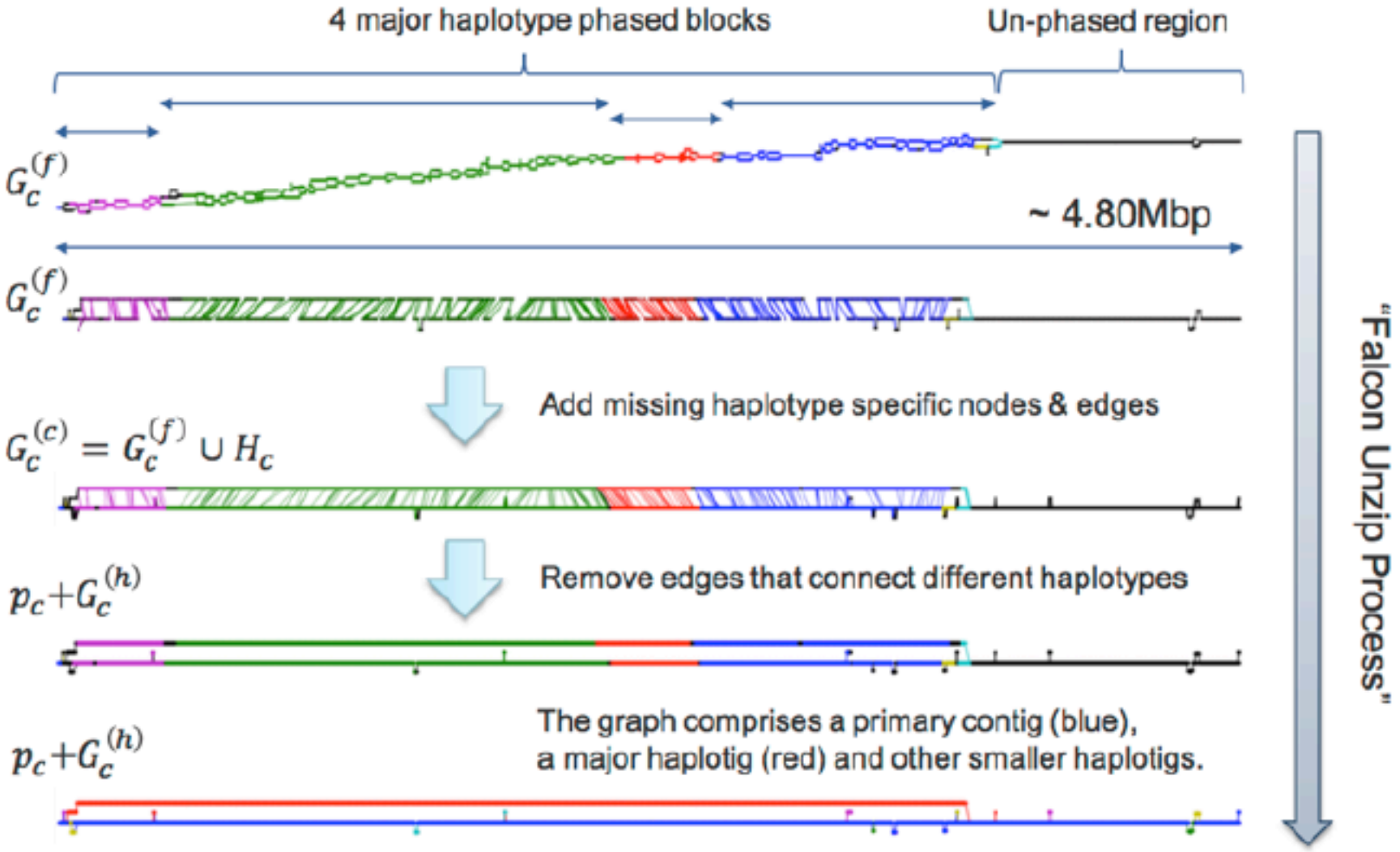
FALCON_Unzip

- Step 6: incorporate phased reads to haplotype-fused string graph for unzipping collapsed paths.
  - Produces primary contigs and haplotigs



http://dx.doi.org/10.1101/056887

FALCON_Unzip



http://dx.doi.org/10.1101/056887

FALCON_Unzip

- Step 7: assign reads to specific primary contigs or haplotigs for generating the final consensus with Quiver.
  - "block identifier" and "phase identifier" are used.

http://dx.doi.org/10.1101/056887

FALCON_Unzip

- Input
  - Results from FALCON assembly
  - Raw sequencing data with “signal pulse” information in bam format (.bax.h5 to .bam)
- Output
  - 3-unzip
  - 4-quiver
    - cns_output
- https://github.com/PacificBiosciences/FALCON_unzip/wiki

FALCON_Unzip

- Performance
  - Idealy a diploid genome produces primary contigs and haplotigs in similar sizes.

	HGAP inbreds, Col-0 vs. Cvi-0		Falcon Unzip haplotigs vs primary contigs	
Variant Type	events	Affected Bases	events	Affected Bases
SNP Count	501,243	1,002,486	430,043	860,086
indel > 50 bp	1,051	882,736	966	798,438
repeat contration/expansion > 50 bp	1,670	3,746,572	1,481	3,130,205
tandem contration/expansion > 50 bp	73	97,319	65	85,495
total SV > 5bp detected	2,794	4,726,627	2,512	4,014,138
predicted CDS	Col-0:28176, Cvi-0:27797		p:31658, h:25117	
Aligned CDS pairs	27,424		24,808	
predicted coding sequence SNPs	183,942	367,884	147,811	295,622
other predicted coding sequence variants	16,748	153,260	15,151	136,245
local in-frame variants	5,135	82,929	4,090	66,681
local non in-frame variants	11,613	70,331	11,061	69,564

http://dx.doi.org/10.1101/056887

Running FALCON_Unzip

- Convert .bax.h5 to .bam

export LD_LIBRARY_PATH=/home/jfass/tools/pitchfork/deployment/lib/:$LD_LIBRARY_PATH

module unload smrtanalysis/2.3.0-140936.p0

/home/jfass/tools/pitchfork/deployment/bin/bax2bam input.bax.h5 –o input

Running FALCON_Unzip

- Edit fc_unzip.cfg file

```
[General]
# the job_type is not used for now.
job_type=local

[Unzip]
input_fofn=input.fofn
input_bam_fofn=input_bam.fofn

smrt_bin=/home/jfass/tools/pitchfork/deployment/bin/temp
sge_phasing= -pe smp 2 --reservation=workshop
sge_quiver= -pe smp 4 --reservation=workshop
unzip_concurrent_jobs = 4
quiver_concurrent_jobs = 2
```

Running FALCON_Unzip

- Set proper environment (bash inputcommands.sh):

source /home/msettles/Python_venv/bin/activate

export LD_LIBRARY_PATH=/usr/lib/x86_64-linux-gnu/:
$LD_LIBRARY_PATH

export PATH=$PATH:/home/jfass/tools/pitchfork/deployment/bin:/
software/python/2.7.6/x86_64-linux-ubuntu14.04:/home/msettles/opt/src/
FALCON-integrate/fc_env/bin/:/home/msettles/Python_venv/bin:/usr/
local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin

export PATH=/home/jfass/tools/MUMmer3.23:$PATH:

Running FALCON_Unzip

fc_unzip.py fc_unzip.cfg >& unzip.out.log

fc_quiver.py fc_unzip.cfg >& quiver.out.log