Understanding, Curating, and Analyzing your Diploid Genome Assembly
Sarah B. Kingan, Ph.D.
Senior Scientist, Bioinformatics, PacBio Applications

East Coast UGM, Baltimore, MD
Tuesday June 27th 2017
AGENDA

- Understanding Your Diploid Assembly
  - Assembly Workflow
  - Heterozygosity and Coverage

- Curating your Assembly
  - Filtering Contigs
  - Deduplicating Haplotypes

- Submitting Your Assembly to NCBI
Understanding Your Diploid Assembly

Assembly Workflow: From Raw Reads to High Quality Reference
FALCON / HIERARCHICAL GENOME ASSEMBLY PROCESS (HGAP)

RAW READS

PREASSEMBLED READS

CONTIG ASSEMBLY

“ERROR-CORRECTED” READS

FALCON AND FALCON-UNZIP

Phased diploid genome assembly with single-molecule real-time sequencing.


ABSTRACT

While genome assembly projects have been successful in many haploid and inbred species, the assembly of noninbred or rearranged heterozygous genomes remains a major challenge. To address this challenge, we introduce the open-source FALCON and FALCON-Unzip algorithms (https://github.com/PacificBiosciences/FALCON/) to assemble long-read sequencing data into highly accurate, contiguous, and correctly phased diploid genomes. We generate new reference sequences for heterozygous samples including an F1 hybrid of Arabidopsis thaliana, the widely cultivated Vitis vinifera cv. Cabernet Sauvignon, and the coral fungus Clavicorona pyxidata, samples that have challenged short-read assembly approaches. The FALCON-based assemblies are substantially more contiguous and complete than alternate short- or long-read approaches. The phased diploid assembly enabled the study of haplotype structure and heterozygosities between homologous chromosomes, including the identification of widespread heterozygous structural variation within coding sequences.

FALCON is a diploid-aware assembler.
FALCON-Unzip module performs true phased assembly for diploid samples.

DIPLOID ASSEMBLY WITH FALCON-UNZIP

DIPLOID ASSEMBLY WITH FALCON-UNZIP

DIPLOID ASSEMBLY WITH FALCON-UNZIP

DIPLOID ASSEMBLY WITH FALCON-UNZIP

(a) FALCON

Initial assembly graph

SNPs

SV

SNPs

SV

SNPs

Primary contig

Associate contig 2
(Alternative allele)

(b)

Phase heterozygous SNPs and identify the haplotype of each read

(c) FALCON-Unzip

Haplotype-resolved assembly graph

SNPs

SVs

SNPs

SVs

SNPs

DIPLOID ASSEMBLY WITH FALCON-UNZIP

DIPLOID ASSEMBLY WITH FALCON-UNZIP

EXAMPLE ASSEMBLY OF WATER BUFFALO

<table>
<thead>
<tr>
<th></th>
<th>FALCON</th>
<th>FALCON-Unzip</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Length</td>
<td>2.66 Gb</td>
<td>2.65 Gb</td>
</tr>
<tr>
<td>Primary N50</td>
<td>18.7 Mb</td>
<td>18.8 Mb</td>
</tr>
<tr>
<td>Secondary Length</td>
<td>0.218 Gb</td>
<td>1.53 Gb</td>
</tr>
<tr>
<td>Proportion Phased</td>
<td>8.2 %</td>
<td>58 %</td>
</tr>
</tbody>
</table>

7-fold increase in haplotype phasing with Unzip module

Acknowledgements:
Tim Smith, USDA-ARS
John Williams, Lloyd Low, University of Adelaide
Paolo Ajmone-Marsan, Università Cattolica del S. Cuore
David Hume, Mick Watson, Roslin Institute

Olimpia
Phot Credit: Caterina Cambuli
INCREASED HAPLOTIG CONTIGUITY WITH FALCON-UNZIP

**CONTIG: 000078F**

<table>
<thead>
<tr>
<th></th>
<th>FALCON</th>
<th>FALCON-Unzip</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Contig Length</td>
<td>12.9 Mb</td>
<td>12.9 Mb</td>
</tr>
<tr>
<td>Number Secondary Contigs</td>
<td>30</td>
<td>34</td>
</tr>
<tr>
<td>Total Secondary Length</td>
<td>1.21 Mb</td>
<td>10.6 Mb</td>
</tr>
<tr>
<td>Secondary Contig N50</td>
<td>42.5 kb</td>
<td>470 kb</td>
</tr>
<tr>
<td>Proportion Phased</td>
<td>9.3 %</td>
<td>82%</td>
</tr>
</tbody>
</table>
POLISHING WITH ARROW: INCREASED REFERENCE QUALITY

CONSENSUS BASE ACCURACY

Consensus Base Accuracy
- Sequel 2.0 Chemistry
- Bacterial Genomes

Genome Completeness
- Avian Genome
- 50-fold Raw Coverage
- BUSCO2 analysis with eukaryota geneset

70% reduction in Fragmented Genes
15% reduction in Missing Genes
12% increase in Complete Genes

Acknowledgement:
Erich Jarvis, Rockefeller University
# Polishing with Arrow: Workflow

<table>
<thead>
<tr>
<th>Method</th>
<th>Assembly</th>
<th>Polishing</th>
</tr>
</thead>
<tbody>
<tr>
<td>HGAP4 - SMRT Link</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>FALCON</td>
<td>✓</td>
<td>resequencing pipeline from pbsmrtpipe/SMRT Link</td>
</tr>
<tr>
<td>FALCON-Unzip</td>
<td>✓</td>
<td>✓ (phased) plus optional resequencing</td>
</tr>
</tbody>
</table>

### Random Best Mapping
- Random choice of locus with equal BLASR score

### Minimum Coverage <5
- <5 reads span 500 bp window
- No consensus call
- Reference base returned as lowercase

## Diagram

```
BLASR: Map Raw Reads
     ▼
 ARROW: Consensus
      ▼
 FASTA FASTQ
```

Consensus Sequence
```
ATGCGCCGTATATGG
AAGCTAGCTAGCTAGCT
GTAGCTAGCTAGCTAGCT
GCGCCTAGATAGGG
GCCTACATAGGCCCTTTT
```
Understanding Your Diploid Assembly

Heterozygosity, Assembly Structure, and Coverage
IMPACT OF HETEROZYGOSITY ON ASSEMBLY PROCESS

Heterozygosity Level

Assembly Graph

Assembly Output

Single Primary Contig
"Collapsed Haplotype"

Haplotig and Primary Contig
"Unzipped Region"
"Phased Haplotypes"

Two Primary Contigs
"Homologous Primary Contigs"

Modified from Chin et al. 2016
IMPACT OF HETEROZYGOSITY ON ASSEMBLY PROCESS

Modified from Chin et al. 2016
IMPACT OF HETEROZYGOSITY ON ASSEMBLY PROCESS

Modified from Chin et al. 2016
IMPACT OF HETEROZYGOSITY ON ASSEMBLY PROCESS

Heterozygosity Level

Assembly Graph

Assembly Output

Single Primary Contig
“Collapsed Haplotype”

Haplotig and Primary Contig
“Unzipped Region”
“Phased Haplotypes”

Two Primary Contigs
“Homologous Primary Contigs”
“Phased Haplotypes”

Modified from Chin et al. 2016
**RAW READ COVERAGE AND ASSEMBLY STRUCTURE**

- **Heterozygosity Level**
- **Assembly Output**
- **Raw Read Coverage**

- **Single Primary Contig**
  - "Collapsed Haplotype"

- **Haplotig and Primary Contig**
  - "Unzipped Region"
  - "Phased Haplotype"

- **Two Primary Contigs**
  - "Homologous Primary Contigs"
  - "Phased Haplotype"
SMRT LINK COVERAGE REPORTS

Graphical Outputs from Resequencing Pipeline / HGAP4

COVERAGE HISTOGRAM: GENOME

COVERAGE PLOT: CONTIG
SMRT LINK COVERAGE REPORTS

Graphical Outputs from Resequencing Pipeline / HGAP4

COVERAGE HISTOGRAM: GENOME

COVERAGE PLOT: CONTIG

"haploid" coverage

"diploid" coverage

"diploid" coverage

"haploid" coverage

PRIMARY CONTIG

SECONDARY CONTIGS
SMRT LINK COVERAGE SUMMARY FILES

alignment_summary.gff: coarse coverage across all contigs

- SMRT Link job directory
  - myJob/tasks/pbreports.tasks.summarize_coverage-0/alignment_summary.gff
- File Format Specs: https://github.com/ben-lerch/SAT

<table>
<thead>
<tr>
<th>cov</th>
<th>cov2</th>
<th>gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>min</td>
<td>mean</td>
<td>number continuous gaps</td>
</tr>
<tr>
<td>median</td>
<td>s.d.</td>
<td>number gap bases</td>
</tr>
<tr>
<td>max</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TOOLS FOR CUSTOM COVERAGE ANALYSIS

- Merge BAM files (N=24)
  - pbmerge
  - samtools merge
- Coverage calculation
  - samtools depth
  - bedtools genomecov
- Visualization
  - R – text file/dataframe

CONTIG COVERAGE HISTOGRAM
bedtools genomecov, R

CONTIG COVERAGE WINDOW  bedtools genomecov/samtools depth, R
ALIGNMENTS AND VISUALIZATION

Recommended Tools for Haplotype Alignment and Analysis

- Subset Reference Sequence
  - samtools faidx

- Alignment
  - MUMmer (v4, multi-threaded support)
  - NUCmer, delta-filter, show-coords, show-snps, etc

- Visualization
  - mummerplot
  - assemblytics
  - gepard (alignment + vis)

- FALCON Assembly Tools
  - https://github.com/PacificBiosciences/apps-scripts/
  - FALCONAssemblyTools repo

HAPLOTIGS TO PRIMARY CONTIG DOTPLOT
alignHaplotigs2Primary.sh
Assemblytics
Assembly Finishing
Filtering, Circularizing, Haplotype Deduplication
GUIDELINE FOR CONTIG FILTERING

https://github.com/PacificBiosciences/apps-scripts/tree/master/FALCONAssemblyTools

Circularize organelle
- Identify
  - high coverage
  - “circular ctg” FALCON annotation
  - blast hit to organelle
- Circularize and polish
  - minimus2, circulator

Remove low quality contigs
- Filter out contigs with >50% unpolished bases (lowercase)

Remove nested haplotigs
- Short haplotigs that align within longer haplotigs

HAPLOTIGS ALIGNED TO PRIMARY
DEDUPLICATING PRIMARY CONTIGS

- Heterozygosity Level
- Assembly Output
- Raw Read Coverage

Single Primary Contig
"Collapsed Haplotype"

Haplotig and Primary Contig
"Unzipped Region"
"Phased Haplotypes"

Two Primary Contigs
"Homologous Primary Contigs"
DOES MY ASSEMBLY HAVE HOMOLOGOUS PRIMARY CONTIGS?

**Primary** assembly length is longer than haploid genome size

- Inbred individual: diploid assembly: assembly length = 1N
- F1 hybrid: haploid assembly: assembly length = 2N

Haploid coverage on primary contigs in regions *without haplotigs*

BUSCO analysis on **primary** contigs indicates widespread duplicated genes
METHODS TO IDENTIFY HOMOLOGOUS PRIMARY CONTIGS

BUSCO/Gene Annotation
- Pros: simple, works for highly divergent haplotypes
- Cons: unannotated contigs excluded
- Usage Case: high contiguity assembly, highly divergent haplotypes

All-By-All Alignments
- Pros: simple
- Cons: high compute time/manual curation
- Usage Case: small genome (<1 Gb)

Purge Haplotypes Pipeline (Mike Roach)
- Pros: uses coverage and pairwise identity
- Cons: some manual curation
- Usage Case: many
EXAMPLE: AEDES MOSQUITO FALCON-UNZIP ASSEMBLY

- Expected Genome Size: ~1.3 Gb
- Primary Contig Length: 1.69 Gb

BUSCO ANALYSIS:
ARTHROPOD GENESET (N = 2675)

<table>
<thead>
<tr>
<th>ASSEMBLY</th>
<th>Aedes PACBIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPLETE</td>
<td>98%</td>
</tr>
<tr>
<td>MISSING</td>
<td>2%</td>
</tr>
<tr>
<td>FRAGMENTED</td>
<td>10%</td>
</tr>
<tr>
<td>DUPLICATED</td>
<td>32%</td>
</tr>
</tbody>
</table>

Acknowledgement:
Aedes Genome Working Group
Leslie Vosshall, Ben Matthews,
Rockefeller University
BUSCO METHOD

github.com/skingan/HomolContigsByAnnotation

Input: BED-formatted Gene Annotation

Output: list of contig pairs that share genes

User Curation of Contig Pairs

Convert BUSCO Output to BED
ALL-BY-ALL ALIGNMENTS

https://github.com/PacificBiosciences/apps-scripts/blob/master/FALCONAssemblyTools/get_homologs.py

Identify homologous relationships by alignments

- Each contig aligned to all shorter contigs using NUCmer
  - uses multi-threaded version of MUMmer4
- High quality alignments filtered
- Multi-sequence alignment visualized in mummerplot
- Manual curation of plots

OUTPUT EXAMPLE
PURGE HAPLOTIGS
MIKE ROACH, AUSTRALIAN WINE RESEARCH INSTITUTE

https://bitbucket.org/mroachawri/purge_haplotigs/

Semi-automated pipeline to remove haplotigs from primary contigs

- Input: BAM of mapped PacBio reads to primary contigs
- Output: curated haploid representation of assembly
  - Record of association between excluded and retained primary contigs

Pipeline
1. coverage histogram and user-defined coverage cut offs
2. contig-specific coverage analysis to identify candidate haplotigs
3. alignments and iterative purging of candidates
PURGE HAPLOTIGS: EXAMPLE FROM BARBERRY BUSH

1. PRIMARY CONTIG COVERAGE HISTOGRAM

2. INDIVIDUAL CONTIG COVERAGE ASSESSMENT
   - <80% contig length has diploid coverage
   - 4470 / 4672 contigs flagged as “suspect”

3. ITERATIVE REASSIGNMENT
   - All-by-all BLAST to find two best hits of ”suspect” contigs
   - NUCmer alignment and summary stats
   - Categorization as “repeat” or “haplotig”

OUTPUT: CURATED ASSEMBLY
- Revised haploid genome
- Log file of reassignment

Acknowledgement: Iago Hale, UNH
METHODS TO IDENTIFY HOMOLOGOUS PRIMARY CONTIGS

BUSCO/Gene Annotation
- github.com/skingan/HomolContigsByAnnotation

All-By-All Alignments
- https://github.com/PacificBiosciences/apps-scripts/blob/master/FALCONAssemblyTools/get_homologs.py

Purge Haplotypes Pipeline (Mike Roach)
- https://bitbucket.org/mroachawri/purge_haplotigs/
Sharing Your Assembly

Diploid Assembly Submission to NCBI
ASSEMBLY SUBMISSION TYPES

SIMPLE: no complex molecules
- Complete replicon
- WGS contigs

COMPLEX: scaffolds or chromosomes
- Haploid Only
- Haploid + Alts
- Diploid/Polyploid (chromosome-level)

Diploid
ASSEMBLY SUBMISSION TYPES

SIMPLE:
no complex molecules

Complete replicon

WGS contigs

COMPLEX:
scaffolds or chromosomes

Haploid Only

Haploid + Alts

Diploid/
Polyploid
(chromosome-level)

optional scaffolding

phased scaffolding

http://biorxiv.org/content/early/2017/02/02/103911
**ASSEMBLY SUBMISSION TYPES**

**SIMPLE:**
- no complex molecules
  - Complete replicon
    - WGS contigs

**COMPLEX:**
- scaffolds or chromosomes
  - Haploid Only
  - Haploid + Alts
    - optional scaffolding
    - phased scaffolding
    - Diploid/Polyploid (chromosome-level)

**Examples of Haploid + Alts**
- GCA_001753755.2
  - Arabidopsis thaliana F1 from Chin et al. 2016
- GCA_002008985.2
  - zebra finch Korlach et al. 2017

[http://biorxiv.org/content/early/2017/02/02/103911](http://biorxiv.org/content/early/2017/02/02/103911)
ASSEMBLY UNITS: PRIMARY AND ALT

Zebra finch: GCA_002008985.2

- FALCON-Unzip:
  - Primary Contigs and Haplotigs
- FALCON:
  - Primary Contigs and Associated Contigs
REGIONS

Zebra finch: GCA_002008985.2

REGIONS: where single haplotigs or overlapping haplotigs map to primary assembly unit
ALTERNATE LOCUS PLACEMENT FILE

- Required for haploid + alts submission
- Details placement of alt sequences relative to primary assembly

<table>
<thead>
<tr>
<th>HEADER</th>
<th>EXAMPLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>alt_asm_name</td>
<td>haplotigs</td>
</tr>
<tr>
<td>prim_asm_name</td>
<td>Primary Assembly</td>
</tr>
<tr>
<td>alt_scaf_name</td>
<td>000123F_001</td>
</tr>
<tr>
<td>parent_type</td>
<td>SCAFFOLD</td>
</tr>
<tr>
<td>parent_name</td>
<td>000123F</td>
</tr>
<tr>
<td>ori</td>
<td>+</td>
</tr>
<tr>
<td>alt_scaf_start</td>
<td>90</td>
</tr>
<tr>
<td>alt_scaf_stop</td>
<td>41595</td>
</tr>
<tr>
<td>parent_start</td>
<td>8663681</td>
</tr>
<tr>
<td>parent_stop</td>
<td>8708427</td>
</tr>
<tr>
<td>alt_start_tail</td>
<td>89</td>
</tr>
<tr>
<td>alt_stop_tail</td>
<td>2</td>
</tr>
</tbody>
</table>
PLACEMENT FILE TOOLS
https://github.com/skingan/NCBI_DiploidAssembly

generate_placement.py
- Written by Jason Chin
- Runs NUCmer and generates placement file

nucmer2ncbiPlacement.py
- Written by Sarah Kingan
- generates placement file from directory of filtered NUCmer alignments
- Contains suggested MUMmer commands
RESOURCES

FALCON
- http://pb-falcon.readthedocs.io/
- https://github.com/PacificBiosciences/FALCON-integrate
- https://github.com/PacificBiosciences/apps-scripts/tree/master/FALCONAssemblyTools
- Unzip binary to be released with SMRT Analysis 5.0 in Q3

SMRT Analysis
- http://www.pacb.com/support/software-downloads/