Diploid Genome Assembly and Comprehensive Haplotype Sequence Reconstruction

Jason Chin, Paul Peluso, David Rank, Fritz Sedlazeck, Maria Nattestad, Michael Schatz, Greg Concepcion, Alicia Clum, Kerrie Barry, Alex Copeland, Ronan O’Malley
Acknowledgments

- All PacBio Colleagues

- Ronan O’Malley, Chongyuan Luo, Joseph Ecker (HHMI / The Salk Institute)

- Alicia Clum, Kerrie Barry, Alex Copeland (Joint Genome Institute)

- Maria Nattestad, Fritz Sedlazeck, Michael Schatz (CSHL)

- Open source toolsets
  - Daligner (https://dazzlerblog.wordpress.com), Gene Myers
  - BLASR (https://github.com/PacificBiosciences/blasr), Mark Chaisson
  - Python, NetworkX for rapid algorithm prototyping
  - Gephi, Graphviz for graph visualization
SOLVING THE DIPLOID ASSEMBLY PROBLEM

- Falcon (a polyploid-aware assembler): generating the contigs through the bubbles
- Falcon Unzip: identifying smaller variants and using them to separate the haplotypes

- Bubbles = big variants between the haplotypes
- Collapsed Path = smaller variants between the haplotypes
WHY DO WE SEE BUBBLES?

In most OLC assembler design, the overlapper does not catch differences at SNP level but structural variations are naturally segregated.
THE FALCON UNZIP PROCESS

FALCON-Unzip

Augmented with haplotype information of each reads

Updated primary contig + “associate haplotigs”

FALCON

Primary contig

Associate contig 1 (Alternative allele)

Associate contig 2 (Alternative allele)
PHASING READ INTO HAPLOTYPE GROUPS

Align SMRT reads to the initial primary contig

Identify het-SNPs → Phase het-SNPs → Group reads with phased SNPs → Reconstruct haplotypes

Haplotypes 0 and 1

More het-SNPs in longer reads: 8% to 15% sequence error rate is not an issue given enough long read coverage for phasing.
QUESTION: HOW TO RESOLVE STRUCTURAL VARIATIONS & HET-SNPS PHASING AT ONCE

<table>
<thead>
<tr>
<th>Information Sources</th>
<th>Pros &amp; Cons</th>
<th>Assembly graph features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structural Variations</td>
<td>✓ Overlap-layout process catches SV haplotypes</td>
<td>✓ Nearby SVs may be phased automatically</td>
</tr>
<tr>
<td></td>
<td>✗ Collapsed paths when there is no SV</td>
<td>✗ Haplotype-fused paths</td>
</tr>
<tr>
<td>het-SNP</td>
<td>✓ Easy to group SNPs/reads into different haplotypes</td>
<td>✓ Haplotype-specific paths</td>
</tr>
<tr>
<td></td>
<td>✗ No phasing information associated with SVs</td>
<td>✗ More fragmented contigs</td>
</tr>
</tbody>
</table>
MERGE HAPLOTYPE INFORMATION AND “UNZIP”

Tiling path of haplotype 0

Remove edges connecting different haplotypes

Tiling path of haplotype 1
PUT EVERYTHING TOGETHER

4 major haplotype phased blocks determined by het-SNPs

Un-phased region

~ 4.80 Mb

Add missing haplotype specific nodes & edges

Remove edges that connect different haplotypes

The final graph comprises a primary contig (blue), a major haplotig (red) and other smaller haplotigs.
POLISHING: ALLELE-SPECIFIC ALIGNMENT FOR FINAL CONSENSUS

“Augmented alignment”: Each read has extra attribute (e.g., contig identifier, phasing block, haplotype phase), an aligner uses those information to place the read to specific reference sequence or regions.
CONSTRUCT ARABIDOPSIS THALIANA COL-0 X CVI-0 DIPLOID F1 LINE

- Two inbred lines sequenced in 2013 (P4 chemistry), assembled as haploid genomes
- F1 line constructed and sequenced in 2015 (P6 chemistry), assembled with FALCON and FALCON-Unzip

• Primary contigs ~ 1n representation of the genome
• Haplotigs ~ phased sequences from where the homologous chromosomes are distinguishable
# Arabidopsis thaliana F1 Diploid Assembly Statistics

## Assembly Size (Mb)

<table>
<thead>
<tr>
<th>Strain</th>
<th>Inbred Col-0</th>
<th>Inbred Cvi-0</th>
<th>Col-0 x Cvi-0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Assembler</strong></td>
<td>CA/HGAP</td>
<td>CA/HGAP</td>
<td>FALCON primary contigs</td>
</tr>
<tr>
<td>Assembly Size (Mb)</td>
<td>126</td>
<td>119</td>
<td>143</td>
</tr>
<tr>
<td># contigs</td>
<td>1325</td>
<td>194</td>
<td>426</td>
</tr>
<tr>
<td>N50 size (Mb)</td>
<td>6.210</td>
<td>4.79</td>
<td>7.92</td>
</tr>
<tr>
<td>Max Contig size (Mb)</td>
<td>10.25</td>
<td>11.25</td>
<td>13.39</td>
</tr>
</tbody>
</table>

## N50 size (Mb)

- F1 Unzip haplotigs: 6.92
- F1 Unzip p-contigs: 7.96
- F1 FALCON p-contigs: 7.92
- Inbred Cvi-0: 4.79
- Inbred Col-0: 6.21
EVALUATE THE DIPLOID ASSEMBLY RESULT

Haploid-like contig in the inbred-line assemblies

Col-0 assembly

Cvi-0 assembly

Col-0 x Cvi-0 assembly

By aligning the haplotigs to the parental genome assemblies, we can evaluate the haplotigs’ quality, e.g. haplotyping accuracy and CDS prediction consistency.
COMPARE F1 ASSEMBLY TO THE INBRED ASSEMBLIES

- We call the SNP and SVs against the parental inbred assemblies for all primary contigs and haplotigs.
- Most haplotigs can be fully assigned to one of the parental haplotypes.
COMPARE F1 ASSEMBLY TO THE INBRED ASSEMBLIES

- We call the SNP and SVs against the parental inbred assemblies for all primary contigs and haplotigs.
- Most haplotigs can be fully assigned to one of the parental haplotypes.
ANNOTATION COMPARISON

Compare *de novo* gene prediction (with AUGUSTUS (Stanke 2003)) between different assemblies

<table>
<thead>
<tr>
<th>Assemblies</th>
<th>TAIR 10</th>
<th>Col-0</th>
<th>Cvi-0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of predicted CDS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Col-0</td>
<td>30006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Col-0 x Cvi-0</td>
<td>56775</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100% indel-free full length overlaps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Col-0</td>
<td>25,966</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Col-0 x Cvi-0</td>
<td>25,865</td>
<td>26,537</td>
<td>27,370</td>
</tr>
<tr>
<td></td>
<td>(92.9%)</td>
<td>(88.4%)</td>
<td>(99.9%)</td>
</tr>
</tbody>
</table>
# Other Smaller and Larger Diploid Genomes

<table>
<thead>
<tr>
<th>FALCON-Unzip Results:</th>
<th>Clavicorona pyxidata (Coral Fungus)</th>
<th>Cabernet Sauvignon**</th>
<th>Human*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haploid Genome Size:</td>
<td>~ 44 Mb</td>
<td>~ 500 Mb</td>
<td>~ 3 Gb</td>
</tr>
<tr>
<td>Primary contig size</td>
<td>41.9 Mb</td>
<td>591.0 Mb</td>
<td>2.76 Gb</td>
</tr>
<tr>
<td>Primary contig N50</td>
<td>1.5 Mb</td>
<td>2.2 Mb</td>
<td>22.9 Mb</td>
</tr>
<tr>
<td>Haplotig size</td>
<td>25.5 Mb</td>
<td>372.2 Mb</td>
<td>2.0 Gb</td>
</tr>
<tr>
<td>Haplotig N50</td>
<td>872 kb</td>
<td>767 kb</td>
<td>330 kb</td>
</tr>
</tbody>
</table>

*Led by Cantu lab, UC Davis and Cramer lab, UN Reno

**Preliminary results. Fast file system and efficient computational infrastructure are currently needed for large genomes.
SUMMARY

- Single data type for routine diploid assembly

- Large genomes are more computationally challenging but it is mostly an engineering problem now:
  - Haplotype phasing improvement, incorporate 3rd party phasing code
  - Develop a sequence aligner for “augmented alignment” for faster Quiver consensus process

- FALCON-Unzip code: (No code, No truth!!) if you like to hack it for now, email me (jchin@pacb.com)

- Want to attack the algorithm problem for polyploid assembly? Let us help you!

Thanks for your attention!