HYBRID ASSEMBLY OF THE ANCESTRAL WHEAT Ae. tauschii 4.25Gb GENOME

Aleksey Zimin*, Guillaume Marcais, Daniela Puiu, Steven Salzberg and James Yorke, et al.
University of Maryland, Johns Hopkins University

Aegilops tauschii is one of the progenitors of common wheat

- Aegilops tauschii sequence provides a reference for study of polyploid genome evolution by facilitating comparison of the wheat D-genome and Ae. tauschii genomic sequences

Dominant Sequencing technologies

- **Illumina**
  - Cheap: as low as $5K for de novo mammalian genome
  - Accurate: 1-2% error rate
  - Only 150 to 350 bp reads, almost all paired

- **PacBio**
  - Inexpensive: $100k+ for a mammalian genome
  - High error rate ~15%
  - Random sequence insertions, chimeric reads
  - We get on average ~10000 bp reads

Data for A. tauschii assembly

- WGS Illumina data and Pacbio

  - Data used for the hybrid assembly:
    - 62x WGS coverage by 2x250bp Illumina Paired end reads
    - 450bp fragment size
    - 35x WGS P6C4 Pacbio reads, ~10Kb N50 size

Definitions

- **Read**, **super-read**, **mega-read**

  - **Read**
    - a fragment of genome sequence read by a sequencing machine
    - 100-250bp long for Illumina sequencing
  
  - **Super-read**
    - a synthetic read produced by extending illumina read(s) by using k-mer graph;
    - typically several reads extend to the same super-read
    - 400-2000 bp average length

  - **Mega-read**
    - a synthetic read produced by merging super-reads with exact sequence overlaps guided by a template long read (Pacbio read)
    - 5000-8000 bp average length

Advantages of our hybrid approach

- We aim at producing long near-perfect “mega-reads” from the Pacbio SMRT reads
- We pre-process the Illumina reads to form Super-reads:
  - Much longer - average >500bp
  - 2-3x overlapping genome coverage
  - Exact k-overlaps (usually min 69+bp) known
- Require only ~10-20x coverage of Pacbio reads and 50x-100x Illumina 150-250bp reads
- Can preserve and resolve haplotype information based on the accurate Illumina and long Pacbio reads
- Relatively inexpensive computationally (less than 1 month on 48-core 512Gb computer for 3Gb genome).
Preliminary Assembly results

<table>
<thead>
<tr>
<th>The First Preliminary Result</th>
<th>My Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembled sequence</td>
<td>3.98Gbp</td>
</tr>
<tr>
<td></td>
<td>4.2Gbp</td>
</tr>
<tr>
<td>N50 contig size</td>
<td>242Kbp</td>
</tr>
<tr>
<td></td>
<td>~400Kbp</td>
</tr>
<tr>
<td>N50 scaffold size</td>
<td>252Kbp</td>
</tr>
</tbody>
</table>

In every project we always run assembly more than once to achieve the best result.

Overview of the MaSuRCA mega-reads technique

Efficient error correction of the PacBio reads

- We use every single PacBio read as a template
- Find the best tiling of each PacBio read with the super reads to produce accurate mega-reads
- Assemble the mega-reads + the other data with Celera Assembler 8.3
- Optional: Post-process to resolve (unzip) haplotypes

MaSuRCA mega-reads 3.2.x

- The mega-reads technique was developed by our group at the University of Maryland and added to the MaSuRca assembler. Available now for beta-testing from us [http://www.genome.umd.edu](http://www.genome.umd.edu), by request.
- Design aimed at large genomes
- The latest version can assemble mammalian genome in ~1 month on one 48-core computer
- We are using it to assemble 22Gbp genome of Loblolly pine, 2.8Gbp cow genome, and 1Gb Manakin genome (in collaboration with Smithsonian Institution)

MaSuRCA performance on PacBio Hybrid data

Results on S. cerevisiae W303 data set

<table>
<thead>
<tr>
<th>Assembler</th>
<th>Input data</th>
<th>Aligned NGA50 Contig Kb</th>
<th>Structural mis-assemblies</th>
</tr>
</thead>
<tbody>
<tr>
<td>REFERENCE</td>
<td>40x 454 Sequencing</td>
<td>924</td>
<td>N/A</td>
</tr>
<tr>
<td>HGAP</td>
<td>237x PacBio</td>
<td>809</td>
<td>4</td>
</tr>
<tr>
<td>ECTools</td>
<td>20x PacBio + 100x MiSeq</td>
<td>401</td>
<td>9</td>
</tr>
<tr>
<td>PacBioToCA</td>
<td>20x PacBio + 100x MiSeq</td>
<td>320</td>
<td>15</td>
</tr>
<tr>
<td>MaSuRCA</td>
<td>20x PacBio + 100x MiSeq</td>
<td>804</td>
<td>3</td>
</tr>
</tbody>
</table>

Mega-read sizes

Three hybrid data sets

<table>
<thead>
<tr>
<th>Data set</th>
<th>Pacbio N50 sub-read size (Kb)</th>
<th>Mega-read N50 read size (Kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ae. tauschii, 35x</td>
<td>11.5</td>
<td>9.2</td>
</tr>
<tr>
<td>S. cerevisiae (yeast), 20x</td>
<td>11.6</td>
<td>9.3</td>
</tr>
<tr>
<td>Drosophila pseudoobscura, 20x</td>
<td>8.4</td>
<td>6.9</td>
</tr>
<tr>
<td>Manacus vitellinus, 10x</td>
<td>18.0</td>
<td>12.3</td>
</tr>
</tbody>
</table>

MaSuRCA performance on PacBio Hybrid data

Faux haplotype experiment - in development

- We created faux haplotype data set of 100x Illumina + 20x Pacbio for the yeast (S. cerevisiae)
- Create modified version of the finished sequence by introducing SNPs
- Split Illumina and PacBio reads into 2 groups and introduce SNPs into one group based on alignment to the finished sequence (where the matches are)
- Assemble, and then separate (unzip) the haplotypes
- Use for development and validation

<table>
<thead>
<tr>
<th>Assembler</th>
<th>Haplotype difference rate</th>
<th>Aligned NGA50 Contig Kb</th>
</tr>
</thead>
<tbody>
<tr>
<td>REFERENCE</td>
<td>N/A</td>
<td>924</td>
</tr>
<tr>
<td>MaSuRCA</td>
<td>0.1%</td>
<td>250</td>
</tr>
<tr>
<td>MaSuRCA</td>
<td>1%</td>
<td>268</td>
</tr>
<tr>
<td>MaSuRCA</td>
<td>0.1%</td>
<td>793</td>
</tr>
</tbody>
</table>
Super reads
A key idea used in the assembly

- Based on the observation that most of the sequence in genomes is locally unique - branches are relatively rare

- Consider 10-mers (we use much longer k of course):
  AGCTGACTGACTGGTAACAA
  AGCTGACTGA
  GCTGACTGAC

- The idea is to make reads longer instead of breaking them into k-mers.

Super reads
Extending a read to become a super-read

- Consider a read - can its ends be extended uniquely?
  ACTGACCATGACCATGACCATGACATG
  extend 5 ACTGACCTGG
  extend 3 CGACTGATGG

- Typically Illumina sequencing projects generate data with high coverage (>50x). With 100bp reads this implies that a new read starts on average at least every other base:
  read R extended to super read S (red)
  Many other reads extend to the same S as well

Super reads
Extending a read to become a super-read

- Consider a read - can its ends be extended uniquely?
  ACTGACCATGACCATGACCATGACATG
  extend 5 ACTGACCTGG
  extend 3 CGACTGATGG

- Typically Illumina sequencing projects generate data with high coverage (>50x). With 100bp reads this implies that a new read starts on average at least every other base:
  read R extended to super read S (red)
  Many other reads extend to the same S as well

Super reads
We can keep Extending on the left

- Consider a read
  ACTGACCATGACCATGACCATGACATG
  extend 5 ACTGACCTGG
  extend 3 CGACTGATGG

- Typically Illumina sequencing projects generate data with high coverage (>50x). With 100bp reads this implies that a new read starts on average at least every other base:
  read R extended to super read S (red)
  Many other reads extend to the same S as well

Super reads
Extend, stopping at the next branch (or where there is no data)

- Consider a read
  CGACTGACCATGACCATGACATG
  extend 5 GACTGACTGG
  extend 3 CGACTGATGG

- Typically Illumina sequencing projects generate data with high coverage (>50x). With 100bp reads this implies that a new read starts on average at least every other base:
  read R extended to super read S (red)
  Many other reads extend to the same S as well
Overview of the mega-reads technique

Efficient error correction of the PacBio reads

- We use every single PacBio read as a template
  - Map (approximately) super-reads to PacBio reads
  - Exact overlaps between the super reads confirmed by mapping = proper overlaps
  - Mega-read is a properly overlapping contig of super-reads that matches the PacBio read best
  - If more than one mega-read tiling the pacbio read then join (or not) with pacbio sequence
- Assemble the corrected reads + the other data with Celera Assembler 8.3
- Post-process to resolve (unzip) haplotypes

Matching super reads, letters represent segments of k-mer graph, path indicated in red

C D F H I J
E D F H I K M N O P
B C D E F I H K L M N
A B C F G H
PacBio read yields mega-read
A B C D E F I H K L M N O P

More than one mega-read tiling a Pacbio read

- Some Pacbio reads yield more than one covering mega-read with a gap in corrected coverage
- Insertions in Pacbio reads
- Chimeric Pacbio reads
- Repeats
- Missing Illumina coverage
- We use super reads to decide whether we can use raw Pacbio sequence to join the covering mega-reads
- Each join must be in 2+ reads -- same flanking super-reads

Progress towards WGS PacBio/Illumina hybrid assembly of the Chinese Spring genome

- Doubled haploid Chinese Spring wheat (accession Dv418)
- 33X wheat genome equivalents of PacBio WGS long and superlong reads (Dv418)
- 50X wheat genome equivalents of Hiseq 3000 150bp pairend reads (Dv418)
- 50X wheat genome equivalents of Hiseq 2500 250bp pairend reads (Dv418)

Summary

- Mega-reads benefit from the accuracy of Illumina and the length of the Pacbio reads
- The goal is to assemble haplotype-resolved mammalian-size genome in 2-3 weeks on a single 64-core computer
- Up to 30Gbp genome on a computer with 1Tb of RAM
- MaSuRCA 3.2.x to be released in 2016
Acknowledgements

Funding agencies
- USDA
- NIH
- NSF