The Resurgence of Reference Quality Genome Sequence

Michael Schatz

Jan 13, 2015
PAG XXIII

@mike_schatz / #PAGXXIII
The map-based sequence of the rice genome

International Rice Genome Sequencing Project*

Rice, one of the world’s most important food plants, has important syntonic relationships with the other cereal species, and is a model plant for Arabidopsis. In a reciprocal proteome. Twenty-nine classes of transposable elements, maize and sorghum genome nuclear chromosomes, traits. The additional stimulated improvements.

Table 2 | Size of each chromosome based on sequence data and estimated gaps

<table>
<thead>
<tr>
<th>Chr</th>
<th>Sequenced bases (bp)</th>
<th>Gaps on arm regions No.</th>
<th>Length (Mb)</th>
<th>Telomeric gaps (Mb)</th>
<th>Centromeric gaps (Mb)</th>
<th>rDNA (Mb)</th>
<th>Total (Mb)</th>
<th>Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>43,260,640</td>
<td>5</td>
<td>0.33</td>
<td>0.06</td>
<td>1.40</td>
<td></td>
<td>45.05</td>
<td>99.1</td>
</tr>
<tr>
<td>2</td>
<td>35,954,074</td>
<td>3</td>
<td>0.10</td>
<td>0.01</td>
<td>0.72</td>
<td></td>
<td>36.78</td>
<td>99.7</td>
</tr>
<tr>
<td>3</td>
<td>36,189,985</td>
<td>4</td>
<td>0.96</td>
<td>0.04</td>
<td>0.18</td>
<td></td>
<td>37.37</td>
<td>97.3</td>
</tr>
<tr>
<td>4</td>
<td>35,489,479</td>
<td>3</td>
<td>0.46</td>
<td>0.20</td>
<td></td>
<td></td>
<td>36.15</td>
<td>98.7</td>
</tr>
<tr>
<td>5</td>
<td>29,733,216</td>
<td>6</td>
<td>0.22</td>
<td>0.05</td>
<td>0.82</td>
<td></td>
<td>30.00</td>
<td>99.3</td>
</tr>
<tr>
<td>6</td>
<td>30,731,386</td>
<td>1</td>
<td>0.02</td>
<td>0.03</td>
<td></td>
<td></td>
<td>31.60</td>
<td>99.9</td>
</tr>
<tr>
<td>7</td>
<td>29,643,843</td>
<td>1</td>
<td>0.31</td>
<td>0.01</td>
<td>0.32</td>
<td></td>
<td>30.28</td>
<td>98.9</td>
</tr>
<tr>
<td>8</td>
<td>28,434,680</td>
<td>1</td>
<td>0.09</td>
<td>0.05</td>
<td></td>
<td></td>
<td>28.57</td>
<td>99.7</td>
</tr>
<tr>
<td>9</td>
<td>22,692,709</td>
<td>4</td>
<td>0.13</td>
<td>0.14</td>
<td>0.62</td>
<td>6.95</td>
<td>30.53</td>
<td>98.8</td>
</tr>
<tr>
<td>10</td>
<td>22,683,701</td>
<td>4</td>
<td>0.68</td>
<td>0.13</td>
<td>0.47</td>
<td>0.25</td>
<td>23.96</td>
<td>96.6</td>
</tr>
<tr>
<td>11</td>
<td>28,357,783</td>
<td>4</td>
<td>0.21</td>
<td>0.04</td>
<td>1.90</td>
<td>0.25</td>
<td>30.76</td>
<td>99.1</td>
</tr>
<tr>
<td>12</td>
<td>27,561,960</td>
<td>0</td>
<td>0.00</td>
<td>0.05</td>
<td>0.16</td>
<td>27.77</td>
<td>99.8</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>370,733,456</td>
<td>36</td>
<td>3.51</td>
<td>0.81</td>
<td>6.59</td>
<td>7.20</td>
<td>388.82</td>
<td>98.9</td>
</tr>
</tbody>
</table>

Contig N50: 5.1Mbp
Total projects costs: >$100M
Initial Assembly Attempts with early Illumina sequencers circa 2007-2008
(older Illumina PE76 library with small insert size ~150bp)

<table>
<thead>
<tr>
<th>Assembler</th>
<th>Data set</th>
<th>N50 contig size</th>
<th>Max contig size</th>
<th>Total assembly size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Velvet</td>
<td>25X Nipponbare</td>
<td>1049bp</td>
<td>21833bp</td>
<td>325.8 Mbp</td>
</tr>
<tr>
<td>Velvet</td>
<td>50X Nipponbare</td>
<td>411bp</td>
<td>23095bp</td>
<td>401.6 Mbp</td>
</tr>
<tr>
<td>Abyss</td>
<td>25X Nipponbare</td>
<td>1853bp</td>
<td>12688bp</td>
<td>288.4 Mbp</td>
</tr>
<tr>
<td>Abyss</td>
<td>50X Nipponbare</td>
<td>2847bp</td>
<td>34893bp</td>
<td>317.4 Mbp</td>
</tr>
</tbody>
</table>

Total costs: ~$10k
>1,000x times cheaper, but at what cost scientifically?

W.R. McCombie
Genomics Arsenal in the year 2015

Sample Preparation
Sequencing
Chromosome Mapping
Whole genome de novo assemblies of three divergent strains of rice (*O. sativa*) documents novel gene space of *aus* and *indica*

Oryza sativa Gene Diversity

- Very high quality representation of the “gene-space”
  - Overall identity ~99.9%
  - Less than 1% of exonic bases missing

- Genome-specific genes enriched for disease resistance
  - Reflects their geographic and environmental diversity

- Assemblies fragmented at (high copy) repeats
  - Difficult to identify full length gene models and regulatory features

**Overall sequence content**
In each sector, the top number is the total number of base pairs, the middle number is the number of exonic bases, and the bottom is the gene count. If a gene is partially shared, it is assigned to the sector with the most exonic bases.
Long Read Sequencing Technology

Moleculo

(PacBio RS II)

Oxford Nanopore

(Voskoboynik et al. 2013)
O. sativa pv Indica (IR64)

PacBio RS II sequencing at PacBio
- Size selection using an 10 Kb elution window on a BluePippin™ device from Sage Science

Mean: 5918bp
49.7x over 10kbp
6.3x over 20kb
Max: 54,288bp

Over 118x coverage using P5-C3 long read sequencing
**O. sativa pv Indica (IR64)**

Genome size: ~370 Mb  
Chromosome N50: ~29.7 Mbp

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Contig NG50</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MiSeq Fragments</strong></td>
<td>19 kbp</td>
</tr>
<tr>
<td>25x 456bp</td>
<td></td>
</tr>
<tr>
<td>(3 runs 2x300 @ 450 FLASH)</td>
<td></td>
</tr>
<tr>
<td><strong>“ALLPATHS-recipe”</strong></td>
<td>18 kbp</td>
</tr>
<tr>
<td>50x 2x100bp @ 180</td>
<td></td>
</tr>
<tr>
<td>36x 2x50bp @ 2100</td>
<td></td>
</tr>
<tr>
<td>51x 2x50bp @ 4800</td>
<td></td>
</tr>
<tr>
<td><strong>HGAP + CA</strong></td>
<td>4.0 Mbp</td>
</tr>
<tr>
<td>22.7x @ 10kbp</td>
<td></td>
</tr>
<tr>
<td><strong>Nipponbare</strong></td>
<td>5.1 Mbp</td>
</tr>
<tr>
<td>BAC-by-BAC Assembly</td>
<td></td>
</tr>
</tbody>
</table>

HGAP Read Lengths  
Max: 53,652bp  
22.7x over 10kbp  
*(discarded reads below 8500bp)*
S5 Hybrid Sterility Locus

S5 is a major locus for hybrid sterility in rice that affects embryo sac fertility.

- Genetic analysis of the S5 locus documented three alleles: an indica (S5-i), a japonica (S5-j), and a neutral allele (S5-n).

- Hybrids of genotype S5-i/S5-j are mostly sterile, whereas hybrids of genotypes consisting of S5-n with either S5-i or S5-j are mostly fertile.

- Contains three tightly linked genes that work together in a ‘killer-protector’-type system: ORF3, ORF4, ORF5.

- The ORF5 indica (ORF5+) and japonica (ORF5-) alleles differ by only two nucleotides.
S5 Hybrid Sterility Locus

Sanger ...ACCCTGATATTCTGAGTTACAAGGCATTCA
Illumina ...ACCCTGATATTCTGAGTTACAAGGCATTCA
PacBio ...ACCCTGATATTCTGAGTTACAAGGCATTCA

100kbp
S5 Hybrid Sterility Locus

Sanger: ...ACCCTGATATTCTGAGTTACAAGGCATTGAGCTACTGCTTGCCCACTGACGAGACC...
Illumina: ...ACCCTGATATTCTGAGTTACAAGGCATTGAGCTACTGCTTGCCCACTGACGAGACC...
PacBio: ...ACCCTGATATTCTGAGTTACAAGGCATTGAGCTACTGCTTGCCCACTGACGAGACC...
Sanger

Illumina

Pacbio

Genome size: ~370 Mb
Chromosome N50: ~29.7 Mbp
S5 Hybrid Sterility Locus
Improvements from 20kbp to 4Mbp contig N50:

- Over 20 Megabases of additional sequence
  - Extremely high sequence identity (>99.9%)
  - Thousands of gaps filled, hundreds of mis-assemblies corrected

- Complete gene models, promoter regions for nearly every gene
  - True representation of transposons and other complex features

- Opportunities for studying large scale chromosome evolution
  - Largest contigs approach complete chromosome arms
Current Collaborations

- **Pineapple**
  - UIUC

- **Human**
  - CSHL/OICR

- **Asian Sea Bass**
  - Temasek Life Sciences

- **M. ligano**
  - Hannon

- **T. vaginalis**
  - NYU
Error correction and assembly complexity of single molecule sequencing reads.
Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz, MC
http://www.biorxiv.org/content/early/2014/06/18/006395
Pan-Genome Alignment & Assembly

Time to start considering problems for which \( N \) complete genomes is the input to study the “pan-genome”
- Available today for many microbial species, near future for higher eukaryotes

SplitMEM: A graphical algorithm for pan-genome analysis with suffix skips

Extending reference assembly models
Summary & Recommendations

Reference quality genome assembly is here
- Use the longest possible reads for the analysis
- Don’t fear the error rate, coverage and algorithmics conquer most problems

Megabase N50 improves the analysis in every dimension
- Better resolution of genes and flanking regulatory regions
- Better resolution of transposons and other complex sequences
- Better resolution of chromosome organization
- Better sequence for all downstream analysis

The year 2015 will mark the return to reference quality genome sequence
Acknowledgements

**Schatz Lab**
- Rahul Amin
- Eric Biggers
- Han Fang
- Tyler Gavin

**James Gurtowski**
- Ke Jiang

**Hayan Lee**
- Zak Lemmon
- Shoshana Marcus
- Giuseppe Narzisi
- Maria Nattestad
- Aspyn Palatnick
- Srividya
- Ramakrishnan
- Rachel Sherman
- Greg Vurture
- Alejandro Wences

**CSHL**
- Hannon Lab
- Gingeras Lab
- Jackson Lab
- Hicks Lab
- Iossifov Lab
- Levy Lab
- Lippman Lab
- Lyon Lab
- Martienssen Lab

**McCombie Lab**
- Tuveson Lab
- Ware Lab
- Wigler Lab

**IT & Meetings Depts.**
- Pacific Biosciences
- Oxford Nanopore

**National Human Genome Research Institute**

**U.S. Department of Energy**

**SFARI**
- Simons Foundation Autism Research Initiative
Thank you

http://schatzlab.cshl.edu

@mike_schatz / PAGXXIII
O. sativa pv Indica (IR64)
S5 Hybrid Sterility Locus

Sanger: ACCCTGATATTCTGAGTTACAAGGCATT
Illumina: ACCCTGATATTCTGAGTTACAAGGCATT
PacBio: ACCCTGATATTCTGAGTTACAAGGCATT

100kbp

5.3 Mbp