PROGRESS TOWARD A LOW BUDGET REFERENCE GRADE GENOME ASSEMBLY

Ben Rosen
USDA-ARS AGIL
January 9, 2016

USDA Goat Genome Assembly Project

“Improving Livestock Productivity through Enhanced Breeding Programs”
- Goats benefit small holders
- ~90% of global goat population is in developing countries
- Improve native goats
- A proper reference genome is necessary to construct DNA-based breeding tools for genetic improvement.

Reference animal selection
- Screened US goat breeds with Illumina 50k Goat SNP chip
- Average inbreeding coefficient in San Clemente breed above 0.40
- The San Clemente breed is a “rescued” feral population
  - Two severe bottlenecks since 1875
  - Papadum has an inbreeding coefficient of 0.42 and a Y-Chromosome

Sequencing and Assembly
- PacBio Sequencing (P4 + P5)
  - Human sized genome (~3Gbp)
  - 194Gbp in SRR reads (~70x, 3.110 mean, 40.830 max)
- Pre-assembly — PBcR (PacBio corrected Reads)
- pbdialog (Directed Acyclic Graph Consensus)
- Assembly — Celera Assembler
- MHAP: Minkhash Alignment Process
  - The “AltaVista” algorithm - to detect duplicate web pages
- Applied to DNA sequencing matching and alignment
- Error correction
  - Quiver
  - Pilon
  - 25x illumina reads

Overview
- Sequencing and Assembly
  - PacBio sequencing (P4-P5)
  - De novo assembly
    - Celera Assembler PBcR
      - Sergey Koren & Adam Phillippy (NHGRI)
- Scaffolding 1
  - BioNano Irys System
  - Single molecule NanoChannel optical mapping
    - Alex Hastie (BioNano Genomics)
- Scaffolding 2
  - Hi-C/LACHESIS
  - Chromosome scale scaffolding based on chromatin interactions
    - Shawn Sullivan and Ivan Liachko (UW)
- Reconciliation
  - Radiation hybrid mapping
    - X-ray breakage of chromosomes
    - Brian Sayre (VSU)
- Annotation
  - EVidenceModeler
  - Weighted evidence combiner
    - Ben Rosen (ARS)

Assembly Statistics
- CHIR 1.0 (SHORT READ + OPTICAL MAP)
  - 2.6 Gb total length
  - 140 Mb gap length
  - 71,438 scaffolds
  - 14.4 Mb scaffold N50
  - 337,495 contigs
  - 18.9 Kb contig N50

- PAPADUM V3 (PACBIO)
  - 2.9 Gb total
  - 3110 contigs
  - 4.7 Mb N50
  - 35.6 Mb max
  - Plus “degenerate”
  - 2.3 Gb total
  - 33767 contigs
  - 3.6 Mb N50
More sequence in the PacBio assembly

PacBio Assembly Contigs

Pseudo-chromosome 1

PacBio Assembly Contigs

Pseudo-chromosome 28

Major Structural Inconsistencies

Sergey Koren
Adam Phillippy

Current CHIR reference genome

Current CHI reference genome

Scaffolding

Hi-C/LACHESIS

Chromatin interactions

BioNano Genome Maps

NanoChannel optical mapping

BioNano Scaffolding

A

PacBio Read

Irys Contigs

B

C

Conflict Resolution

+BioNano Assembly Statistics

<table>
<thead>
<tr>
<th></th>
<th>PAPADUM V3</th>
<th>BIONANO</th>
<th>HYBRID</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.6 Gb total</td>
<td>2.7 Gb total</td>
<td>2.6 Gb total</td>
<td></td>
</tr>
<tr>
<td>3110 contigs</td>
<td>533 scaffolds</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.7 Mb N50</td>
<td>8.8 Mb max</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35.6 Mb max</td>
<td>76.3 Mb max</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

HYBRID + PAPADUM V3

2.7 Gb total
1576 scaffolds
22.8 Mb N50
Complete Scaffolding of Chr 20

<table>
<thead>
<tr>
<th>Chr</th>
<th># contigs</th>
<th>length (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>chrX</td>
<td>19</td>
<td>387</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>289</td>
</tr>
<tr>
<td>chr29</td>
<td>25</td>
<td>28</td>
</tr>
<tr>
<td>chr28</td>
<td>27</td>
<td>51.2</td>
</tr>
<tr>
<td>chr27</td>
<td>28</td>
<td>9</td>
</tr>
<tr>
<td>chr26</td>
<td>21</td>
<td>23</td>
</tr>
<tr>
<td>chr25</td>
<td>29</td>
<td>65.2</td>
</tr>
<tr>
<td>chr24</td>
<td>23</td>
<td>37</td>
</tr>
<tr>
<td>chr23</td>
<td>29</td>
<td>69.8</td>
</tr>
<tr>
<td>chr22</td>
<td>22</td>
<td>24</td>
</tr>
<tr>
<td>chr21</td>
<td>18</td>
<td>26</td>
</tr>
<tr>
<td>chr20</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>chr19</td>
<td>22</td>
<td>62.4</td>
</tr>
<tr>
<td>chr18</td>
<td>20</td>
<td>67</td>
</tr>
<tr>
<td>chr17</td>
<td>17</td>
<td>24</td>
</tr>
<tr>
<td>chr16</td>
<td>15</td>
<td>25</td>
</tr>
<tr>
<td>chr15</td>
<td>14</td>
<td>21</td>
</tr>
<tr>
<td>chr14</td>
<td>10</td>
<td>31</td>
</tr>
<tr>
<td>chr13</td>
<td>13</td>
<td>18</td>
</tr>
<tr>
<td>chr12</td>
<td>12</td>
<td>37</td>
</tr>
<tr>
<td>chr11</td>
<td>8</td>
<td>26</td>
</tr>
<tr>
<td>chr10</td>
<td>9</td>
<td>28</td>
</tr>
<tr>
<td>chr9</td>
<td>11</td>
<td>32</td>
</tr>
<tr>
<td>chr8</td>
<td>6</td>
<td>33</td>
</tr>
<tr>
<td>chr7</td>
<td>7</td>
<td>23</td>
</tr>
<tr>
<td>chr6</td>
<td>5</td>
<td>32</td>
</tr>
<tr>
<td>chr5</td>
<td>4</td>
<td>41</td>
</tr>
<tr>
<td>chr4</td>
<td>2</td>
<td>24</td>
</tr>
<tr>
<td>chr3</td>
<td>3</td>
<td>56</td>
</tr>
<tr>
<td>chr2</td>
<td>1</td>
<td>31</td>
</tr>
<tr>
<td>chr1</td>
<td>0</td>
<td>33</td>
</tr>
</tbody>
</table>

Hi-C/LACHESIS

Building chromosome sized scaffolds

<table>
<thead>
<tr>
<th></th>
<th>Papadum V3</th>
<th>Bionano</th>
<th>Hi-C/Lachesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.6 Gb total</td>
<td>2.7 Gb total</td>
<td>2.6 Gb total</td>
<td></td>
</tr>
<tr>
<td>3110 contigs</td>
<td>1575 scaffolds</td>
<td>596 scaffolds</td>
<td></td>
</tr>
<tr>
<td>4.7 Mb N50</td>
<td>28.4 Mb N50</td>
<td>41.7 Mb N50</td>
<td></td>
</tr>
<tr>
<td>35.6 Mb max</td>
<td>76.3 Mb max</td>
<td>187.3 Mb max</td>
<td></td>
</tr>
</tbody>
</table>

X-chromosome

Assembly Statistics

RH map to finalize assembly
Annotation pipeline

Gather evidence
- RNAseq
  - 19 tissues
  - Tophat
  - Stringtie
  - PASA

Ab initio predictions
- Augustus
- Genemark-ET
- BRAKER1

• cDNA alignment
- Sheep
- Pig
- Cow
- Human
- Mouse

Combine
- EVIDenceModeler (EVM)

Refine/add alt-splicing
- PASA
- Program to Assemble Spliced Alignments

Annotation comparisons

CHIR 1.0 annotation LiftOver
- Look at CHIR1.0 genes split in our assembly

<table>
<thead>
<tr>
<th>Gene type</th>
<th>CHIR 1.0</th>
<th>Sheep</th>
<th>Antelope</th>
<th>Cow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Mappings</td>
<td>9,534</td>
<td>9,534</td>
<td>9,534</td>
<td></td>
</tr>
<tr>
<td>Unmapped</td>
<td>564</td>
<td>9,534</td>
<td>9,140</td>
<td></td>
</tr>
<tr>
<td>Split Exons</td>
<td>592</td>
<td>1,930</td>
<td>1,528</td>
<td></td>
</tr>
</tbody>
</table>

Comparative mapping with sheep

Acknowledgements

USDA ARS
- Derek Bickhart
- Curt Van Tassell
- Tim Smith
- Steven Schroeder
- NHGRI
- Adam Phillippy
- Sergey Koren
- Recombinetics
- Tad Sonstegard

University of Washington
- Shawn Sullivan
- Ivan Lischko
- Jay Shendure
- Josh Burton
- BioNano Genomics
- Alex Hastie
- Boudewijin Ten Hallers
- Saki Chan
- Virginia State University
- Brian Sayre

NIFA/AFRI
- USDA USAID – Feed the future.

USDA does not endorse specific products and services, and is an equal opportunity employer.