DE NOVO WHOLE GENOME ASSEMBLY AND SEQUENCING OF THE SUPERB FAIRYWREN (Malurus cyaneus)

JOSHUA PEÑALBA
LEO JOSEPH
CRAIG MORITZ
ANDREW COCKBURN
2014

2015

2016

2017

Synthetic Long-read (Moleculo)
Synthetic Long-read (Moleculo)

CHiCago

Short insert shot-gun Mate-pair
2014

2015

2016

2017

illuminai

PACBIO

RSII

Sequel

CHiCago

Synthetic Long-read (Moleculo)

Dovetail Genomics
Synthetic Long-read (Moleculo)  
CHiCago  
Short insert shot-gun Mate-pair  
RSII  
Linkage map 26K SNPs  
Sequel  
PACBIO®

2014  
2015  
2016  
2017
2014

Synthetic Long-read (Moleculo)

2015

illumina

RSII

PACBIO

Sequel

26K SNPs

2016

Linkage map

2017

Chromium

10X GENOMICS

CHiCago
Synthetic Long-read (Moleculo)
CHiCago
illuminar
RSII
Sequel
Linkage map 26K SNPs
Chromium
Chromium
10X Genomics
illumina
PACBIO
Diversity Arrays Technology
Synthetic Long-read (Moleculo)

CHiCago

Short insert shot-gun Mate-pair

RSII

Sequel

Chromium

Linkage map 26K SNPs

10x GENOMICS

Chromium

Saphyr

HiC
Historic DNA
Historic DNA

Adaptation

Structural variation
SO, YOU WANT TO SEQUENCE A REFERENCE GENOME?
insights from de novo sequencing and assembly of the superb fairywren genome (Malurus cyaneus)

JOSHUA V. PENALBA*  , LEO JOSEPH  , CRAIG MORITZ  , ANDREW COCKBURN

1. Ecology & Evolution, Australian National University, Canberra 2. Centre for Biodiversity Analysis, Canberra 3. Australian National Wildlife Collection, CSIRO, Canberra

@joshpenalba
josh.penalba@gmail.com
http://joshuapenalba.com

Why are you sequencing the fairywren genome?
- Speciation genomics
- Long-term study
- Phylogenetic gap
- Additional resource

The broader aim of my PhD thesis is to understand speciation processes using genomic data. This genome will serve as a reference for studying genomic introgression across a bird suture zone.

If a genome was a jigsaw puzzle...

SEQUENCING TECH & TYPES OF DATA

Illumina shotgun
- High quality short fragments
- DNA insert size > 150
- Yields highly fragmented assemblies
- NOT suitable for assembly unless paired with a different technology

Illumina mate pair
- Jumping libraries linking across long distances across the chromosome
- Yields genome assemblies
- CAN be used for scaffolding long range assemblies
- Often needed as at least a different fragment size

PacBio
- Single molecules, long read sequence with high error rate
- Insert sizes: > 75kb & improving
- Requires MiSeq for error correction & scaffolding

BioNano
- High quality DNA input can yield short reads associated to longer reads
- Great for guiding assemblies

10X Genomics
- Short reads associated to longer reads
- DNA insert sizes: > 300kb
- General quality DNA that yield quality de novo assemblies
- Requires Hi-C coverage

HC / CHICAGO
- Chromosome conformation information
- Association between pairs of reads based on proximity in genome
- HC - hi res - robust orientation
- CHICAGO - in vitro - dependent on DNA
- Can link scaffold into chromosome-scale assemblies

Linkage map
- Generic markers associated with a known chromosome
- Quality of assembly depends on density of markers and number of subreads
- Can build linkage groups associated with chromosomes

Where do I start?
- Budget
- Genome size
- Heterozygosity
- Input DNA?

What about annotation?

Annotation is a beast in its own right.
Proper annotation first requires RNA sequencing from a range of tissue types. The sequencing is then followed by an incredibly time intensive computational pipeline.
Although it’s not ideal, quick and dirty annotation can be achieved by using gene sets from existing databases.
Don’t forget that repeats, not just genes, need to be annotated too! Annotation still in planning stage...
N50

- contigs: 14 Kb
- scaffolds: 6 Mb
<table>
<thead>
<tr>
<th>N50</th>
<th>contigs</th>
<th>scaffolds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>illumina</td>
<td>14 Kb</td>
</tr>
<tr>
<td></td>
<td>PacBio</td>
<td>467 Kb</td>
</tr>
<tr>
<td></td>
<td>10x Genomics</td>
<td>467 Kb</td>
</tr>
<tr>
<td>N50</td>
<td>contigs</td>
<td>scaffolds</td>
</tr>
<tr>
<td>-----</td>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>illumina</td>
<td>14 Kb</td>
<td>6 Mb</td>
</tr>
<tr>
<td>PacBio</td>
<td>467 Kb</td>
<td>8 Mb</td>
</tr>
<tr>
<td>10x Genomics</td>
<td>467 Kb</td>
<td>10 Mb</td>
</tr>
<tr>
<td>Diversity Arrays Technology</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>