Error Correction and Assembly of Oxford Nanopore Sequencing

James Gurtowski
Assembly Complexity
The advantages of SMRT sequencing
Oxford Nanopore MinION

- Thumb drive sized sequencer powered over USB
- Senses DNA by measuring changes to ion flow
- Reads both DNA Strands (2D)
Nanopore Basecalling

Basecalling currently performed at Amazon with frequent updates to algorithm

- Hidden Markov model
- Only four options per transition
- Pore type = distinct kmer length

Form probabilistic path through measured states, currents, and transitions
- e.g. Viterbi algorithm
Our Data - Yeast W303

Mean Read Length:
~6kb (10kb shear)

Best: 446Mb

New Flowcells
7k Avg Read Length

Oxford Flowcell Yields

Yield (Megabases) vs. Date

Flowcell Yield vs. Read Length
Nanopore Alignments

**Alignment Statistics (BLASTN)**
- Mean alignment length at ~7kbp
- Shearing targeted 10kbp
- 255k reads align (64%)
- 174x coverage

**7kb mean alignment length**

68% mean identity
Nanopore Accuracy

Alignment Quality (BLASTN)
Of reads that align, average ~65% identity
“2D base-calling” improves to ~77% identity

![Graph showing frequency distribution of percent identity with average ID of 65% and 77%]

- 65% AVG ID
- 77% AVG ID
64% of the data map using BLASTN
Long Read Correction Algorithms

**PBJelly**
Gap Filling and Assembly Upgrade
English *et al* (2012)  
PLOS One. 7(11): e47768

**PacBioToCA**  
Hybrid Error Correction
Nature Biotechnology. 30:693–700

**HGAP & Quiver**
LR-only Correction & Polishing
Chin *et al* (2013)  
Nature Methods. 10:563–569

< 5x Long Read Coverage > 50x
NanoCorr: Nanopore-Illumina Hybrid Error Correction

1. BLAST Miseq reads to all raw Oxford Nanopore reads

2. Select non-repetitive alignments
   ○ First pass scans to remove “contained” alignments
   ○ Second pass uses Dynamic Programming (LIS) to select set of high-identity alignments with minimal overlaps

3. Compute consensus of each Oxford Nanopore read
   ○ Currently using Pacbio’s pbdagcon

https://github.com/jgurtowski/nanocorr
Post Correction Identity

Mean: 97%

Mean: 68%
ONT vs Illumina Assembly

Oxford N50 : 585kb

Illumina N50 : 58kb
ONT Assembly Completeness

<table>
<thead>
<tr>
<th>Genomic Feature</th>
<th>Average Length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDS</td>
<td>1282</td>
</tr>
<tr>
<td>gene</td>
<td>1344</td>
</tr>
<tr>
<td>rRNA</td>
<td>1393</td>
</tr>
<tr>
<td>gene cassette</td>
<td>2951</td>
</tr>
<tr>
<td>transposable element</td>
<td>4396</td>
</tr>
<tr>
<td>telomere</td>
<td>5836</td>
</tr>
</tbody>
</table>

Bar chart showing frequency of various genomic features for different sequencing technologies:

- S288C
- Nanopore
- Miseq
Long Read Assembly

S288C Reference sequence
• 12.1Mbp; 16 chromo + mitochondria; N50: 924kbp

**Illumina MiSeq**
30x, 300bp PE (Flashed)
Celera Assembler
• 6953 non-redundant contigs
• N50: 59kb >99.9% id

**Oxford Nanopore**
28x corrected reads > 7kb
NanoCorr + Celera Assembler
• 95 non-redundant contigs
• N50: 585kbp >99.78% id

**Pacific Biosciences**
25x corrected reads > 10kb
HGAP + Celera Assembler
• 21 non-redundant contigs
• N50: 811kb >99.8% id
E. Coli K12 Single Contig Assembly with MinION

Nanocor Correction Results
145x Oxford Nanopore X 35x MiSeq

Single Contig Assembly
99.99% Identity (Pilon polishing)

A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer
Joshua Quick, Aaron R Quinlan and Nicholas J Loman
Future of Oxford Nanopore

Zamin Iqbal and 5 others retweeted

GenomeWeb InSequence @InSequence · Oct 20
Oxford Nanopore shows off Promethion at ASHG. #ASHG14 #nanopore
Acknowledgements

Michael Schatz
Dick McCombie
Sara Goodwin
Schatz Lab

Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome
Sara Goodwin, James Gurtowski, Scott Ethe-Sayers, Panchajanya Deshpande, Michael Schatz, W Richard McCombie
doi: http://dx.doi.org/10.1101/013490