Hybrid Error Correction and De Novo Assembly with Oxford Nanopore

Michael Schatz

Jan 13, 2015
PAG Bioinformatics

@mike_schatz / #PAGXXIII
Oxford Nanopore MinION

- Thumb drive sized sequencer powered over USB
- Capacity for 512 reads at once
- Senses DNA by measuring changes to ion flow
Nanopore Sequencing

- DNA sample
- Constriction
- 2-minute centrifugation
- G-tube sense/antisense
- End-repair and dA-tailing
- Adaptor ligation
- Motor attachment
- High molecular weight genomic DNA
- Current graph
Nanopore Basecalling

Basecalling currently performed at Amazon with frequent updates to algorithm
Nanopore Readlengths

Oxford Nanopore Sequencing at CSHL
30 runs, 267k reads, 122x total coverage
Between 11 and 73k reads per run!
Mean flow cell: 50 Mbp in 2 days
Max flow cell: 446Mbp in 2 days

Mean: 5473bp
41x over 10kbp
8x over 20kb
Max: 146,992bp

Spike-in
noise
Nanopore Sequences
“Corrective Lens” for Sequencing
Nanopore Accuracy

Alignment Quality (BLASTN)
Of reads that align, average ~64% identity
“2D base-calling” improves to ~70% identity
Nanopore Accuracy

Alignment Quality (BLASTN)
Of reads that align, average ~64% identity
“2D base-calling” improves to ~70% identity
Nanopore Accuracy

Alignment Quality (BLASTN)
Of reads that align, average ~64% identity
“2D base-calling” improves to ~70% identity
NanoCorr: Nanopore-Illumina Hybrid Error Correction

https://github.com/jgurtowski/nanocorr

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

---

Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome
Long Read Assembly

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

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

**Oxford Nanopore**
NanoCorr + Celera Assembler
• 214 non-redundant contigs
• N50: 472kbp >99.78% id
Advantages of Long Reads

In yeast, Nanopore-based assembly is ~10x more contiguous
In E. coli, Nanopore-based assembly is basically perfect

Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome.
Genomic Futures?

@InSequence · Oct 20
Oxford Nanopore shows off Promethion at ASHG. #ASHG14 #nanopore
Genomic Futures?
iGenomics: Mobile Sequence Analysis
Asyn Palatnick, Elodie Ghedin, Michael Schatz

The worlds first genomics analysis app for iOS devices

First application:
• Handheld diagnostics and therapeutic recommendations for influenza infections
• In a few seconds, iGenomics tells you which antivirals to take or avoid
• Coming soon to the App Store

Future applications
• Pathogen detection
• Food safety
• Biomarkers
• etc..
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

Trends in Algorithmics
– Exciting developments in the future for mobile and remote analysis
– Now is the time to start thinking about pan-genome analysis over a large number of genomes

The resurgence of reference quality genome sequence
Michael Schatz, Ian Korf, Dan Rokhsar
Tuesday @ 4pm, Pacific Salon 1
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

Pacific Biosciences
Oxford Nanopore

NSF
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