Single Cell and Single Molecule Analysis of Cancer

Michael Schatz

April 20, 2015
Laufer Center Retreat
Schatzlab Overview

**Human Genetics**
- Role of mutations in disease
- Narzisi et al. (2014)
- Iossifov et al. (2014)

**Plant Biology**
- Genomes & Transcriptomes
- Schatz et al. (2014)
- Ming et al. (2013)

**Algorithmics & Systems Research**
- Ultra-large scale biocomputing
- Marcus et al. (2014)
- Schatz et al. (2013)

**Single Cell & Single Molecule**
- CNVs, SVs, & Cell Phylogeny
- Garvin et al. (2014)
- Roberts et al. (2013)
Sequence Assembly Problem

1. Shear & Sequence DNA

2. Construct assembly graph from overlapping reads

3. Simplify assembly graph

On Algorithmic Complexity of Biomolecular Sequence Assembly Problem
Assembly Complexity
The advantages of SMRT sequencing
Genomics Arsenal in the year 2015

Sample Preparation  Sequencing  Chromosome Mapping
PacBio SMRT Sequencing

Imaging of fluorescently phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).

Single Molecule Sequences
“Corrective Lens” for Sequencing
Consensus Accuracy and Coverage

Coverage can overcome random errors

- Dashed: error model from binomial sampling
- Solid: observed accuracy

\[ CNS Error = \sum_{i=\left[\frac{c}{2}\right]}^{c} \binom{c}{i} (e)^i (1-e)^{n-i} \]

Nature Biotechnology. 30:693–700
PacBio Assembly Algorithms

PBJelly

Gap Filling and Assembly Upgrade

English et al (2012)
PLOS One. 7(11): e47768

PacBioToCA & ECTools

Hybrid/PB-only Error Correction

Nature Biotechnology. 30:693–700

HGAP & Quiver

PB-only Correction & Polishing

Chin et al (2013)
Nature Methods. 10:563–569

< 5x PacBio Coverage > 50x
3rd Gen Long Read Sequencing

PacBio RS II

CSHL/PacBio
3rd Gen Long Read Sequencing

PacBio RS II

CSHL/PacBio

0 10k 20k 30k 40k

Genome of the 3rd Gen Long Read Sequencing technology, showing applications in various organisms.
3rd Gen Long Read Sequencing

PacBio RS II

CSHL/PacBio

2.5 Mbp

4.0 Mbp

1.4 Mbp

4.5 Mbp

4.6 Mbp
SK-BR-3

Most commonly used Her2-amplified breast cancer cell line

*Can we resolve the complex structural variations, especially around Her2?*

Ongoing collaboration between CSHL and OICR to *de novo* assemble the complete cell line genome with PacBio long reads

(Davidson et al, 2000)
PacBio read length distribution

- Mean: 9kb
- 72.6X coverage
- 49.3X coverage over 10kb
- 12.0X coverage over 20kb
- Max: 71kb
Structural variant discovery with long reads

1. Alignment-based split read analysis: Efficient capture of most events
   BWA-MEM + Lumpy

2. Local assembly of regions of interest: In-depth analysis with base-pair precision
   Localized HGAP + Celera Assembler + MUMmer

3. Whole genome assembly: In-depth analysis including novel sequences
   DNAnexus-enabled version of Falcon

Total Assembly: 2.64Gbp  Contig N50: 2.56 Mbp  Max Contig: 23.5Mbp
PacBio

Chr 17: 83 Mb
Confirmed both known gene fusions in this region
Confirmed both known gene fusions in this region
Joint coverage and breakpoint analysis to discover underlying events
By comparing the proportion of reads that are spanning or split at breakpoints we can begin to infer the history of the genetic lesions.

1. Healthy diploid genome
2. Original translocation into chromosome 8
3. Duplication, inversion, and inverted duplication within chromosome 8
4. Final duplication from within chromosome 8
Cancer lesion Reconstruction

Available today under the Toronto Agreement:
• Fastq & BAM files of aligned reads
• Interactive Coverage Analysis with BAM.IOBIO
• Whole genome assembly & alignment

Available soon
• Whole genome methylation analysis
• Full length cDNA transcriptome analysis
• Comparison to single cell analysis of >100 individual cells

Go see Maria’s poster!

4. Final duplication from within chromosome 8
**Single-Cell Copy Number Analysis**

- Extremely low coverage sequencing (~1x) from amplified cells is sufficient to determine large copy number changes (>50kbp)

- Use this technique to discover CNVs in multiple cells from the same tumor to map its progress

- Implemented a new analysis suite (Ginkgo) to carry out the highly specialized processing

*Interactive analysis and quality assessment of single-cell copy-number variations.*
What should we expect from an assembly?

Summary & Recommendations

Long reads are the key to high quality genome assemblies:

- < 100 Mbp: HGAP/PacBio2CA @ 100x PB C3-P5, expect near perfect chromosome arms.
- < 1GB: HGAP/PacBio2CA @ 100x PB C3-P5, high quality assembly: contig N50 over 1Mbp.
- > 1GB: hybrid/gap filling, expect contig N50 to be 100kbp – 1Mbp.

Caveats:

- Model only as good as the available references (esp. haploid sequences).
- Technologies are quickly improving, exciting new scaffolding technologies!

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
Fritz Sedlacek
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

Cornell
Susan McCouch
Lyza Maron
Mark Wright

OICR
John McPherson
Karen Ng
Timothy Beck
Yogi Sundaravadanam

NBACC
Adam Phillippy
Serge Koren

SBU
Skiena Lab
Patro Lab

NSF
National Human Genome Research Institute

U.S. Department of Energy

SFARI
SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

Oxford Nanopore Technologies

Pacific Biosciences

Alfred P. Sloan Foundation
Thank you
http://schatzlab.cshl.edu
@mike_schatz