Algorithms for single cell
and single molecule biology

Michael Schatz

March 27, 2015
Biotech Symposium / Simons Foundation
Schatzlab Overview

**Human Genetics**
- Autism, Cancer, Psychiatric Disorders
  - Narzisi et al. (2014)
  - Iossifov et al. (2014)

**Informatics**
- Ultra-large scale biocomputing
  - Blood et al. (2014)
  - Schatz et al. (2013)

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

**Biotechnology**
- Single Cell & Single Molecule Analysis
  - Garvin et al. (2014)
  - Roberts et al. (2013)
Outline

1. **Single Molecule Sequencing**
   *Long read sequencing of a breast cancer cell line*

2. **Single Cell Copy Number Analysis**
   *Intra-tumor heterogeneity and metastatic progression*
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
Often an astronomical number of possible assemblies
- Value computed by application of the BEST theorem (Hutchinson, 1975)

\[
\mathcal{W}(G, t) = (\det L) \left\{ \prod_{u \in V} (r_u - 1)! \right\} \left\{ \prod_{(u,v) \in E} a_{uv}! \right\}^{-1}
\]

\(L = n \times n\) matrix with \(r_u - a_{uu}\) along the diagonal and \(-a_{uv}\) in entry \(uv\)
\(r_u = d^+(u) + 1\) if \(u=t\), or \(d^+(u)\) otherwise
\(a_{uv} = \) multiplicity of edge from \(u\) to \(v\)

**Assembly Complexity of Prokaryotic Genomes using Short Reads.**
Assembly Complexity
Assembly Complexity

The advantages of SMRT sequencing
N50 size

Def: 50% of the genome is in contigs as large as the N50 value

Example: 1 Mbp genome 50%

N50 size = 30 kbp
(300k+100k+45k+45k+30k = 520k >= 500kbp)

A larger N50 is indicative of improvement in every dimension:
• Better resolution of genes and flanking regulatory regions
• Better resolution of transposons and other complex sequences
• Better resolution of chromosome organization
Single Molecule Sequencing

PacBio RS II

Oxford Nanopore

CSHL/PacBio

CSHL/ONT
PacBio SMRT Sequencing

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

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

Coverage can overcome random errors

- Dashed: error model; Solid: observed accuracy

\[ CNS \text{ Error } = \sum_{i=\lceil c/2 \rceil}^{c} \binom{c}{i} (e)^i (1-e)^{n-i} \]

Nature Biotechnology. 30:693–700
PacBio Assembly Algorithms

PBJelly

Gap Filling

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

PacBioToCA & ECTools

Hybrid Error Correction

Nature Biotechnology. 30:693–700

HGAP & Quiver

PB-only Correction

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

< 5x PacBio Coverage > 50x
PacBio Sequencing

PacBio RS II

CSHL/PacBio

0 10k 20k 30k 40k

CSHL/PacBio

PacBio Sequencing

Circuits
PacBio Sequencing

PacBio RS II

CSHL/PacBio

2.5 Mbp

4.0 Mbp

1.4 Mbp

4.5 Mbp

4.6 Mbp
Her2 amplified breast cancer

Breast cancer

- About 12% of women will develop breast cancer during their lifetimes
- ~230,000 new cases every year (US)
- ~40,000 deaths every year (US)

Her2 amplified breast cancer

- 20% of breast cancers
- 2-3X recurrence risk
- 5X metastasis risk

Statistics from American Cancer Society and Mayo Clinic.

(Adapted from Slamon et al, 1987)
SK-BR-3

Most commonly used Her2+ 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)
Improving SMRTcell Performance

mean: 6.2kb  
yield: 213Mbp/SMRT cell

mean: 8.3kb  
yield: 620 Mbp/SMRT cell

mean: 11.3kb  
yield: 1031 Mbp/SMRT cell
PacBio read length distribution

- Mean: 9kb
- 72.6X coverage
- 49.3X coverage over 10kb
- 12.0X coverage over 20kb
- Max: 71kb
Genome-wide alignment coverage average is around 54X. Coverage per chromosome greatly varies.
Her2

Chr 17: 83 Mb

8 Mb
Her2

PacBio

Her2

PacBio
PacBio and Illumina coverage values are highly correlated but Illumina shows greater variance because of poorly mapping reads.
PacBio
67X @ 10kb

Illumina
120X @ 100bp

Repeats
21-mers

Her2

8 Mb
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
Green arrow indicates an inverted duplication. False positive and missing Illumina calls due to mis-mapped reads (especially low complexity).
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
SKBR3 Oncogene Analysis

Known missense mutation in p53: **R175H**

**Oncogene amplifications**

<table>
<thead>
<tr>
<th></th>
<th>Reference</th>
<th>Illumina</th>
<th>PacBio</th>
</tr>
</thead>
<tbody>
<tr>
<td>ErbB2 (Her2)</td>
<td>ATCTGAGCAGCGCTCATGGTGTCGGGGGCAGGGCCTCACAACCTCCGTCATGTGCTGTGACTGCTTT</td>
<td>ATCTGAGCAGCGCTCATGGTGTCGGGGGCAGGGCCTCACAACCTCCGTCATGTGCTGTGACTGCTTT</td>
<td>ATCTGAGCAGCGCTCATGGTGTCGGGGGCAGGGCCTCACAACCTCCGTCATGTGCTGTGACTGCTTT</td>
</tr>
<tr>
<td>MYC</td>
<td>≈27X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MET</td>
<td>≈8X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Gene fusions**

<table>
<thead>
<tr>
<th>Known Gene fusions</th>
<th>Confirmed by PacBio reads?</th>
</tr>
</thead>
<tbody>
<tr>
<td>TATDN1</td>
<td>Yes</td>
</tr>
<tr>
<td>GSDMB</td>
<td></td>
</tr>
<tr>
<td>RARA</td>
<td>Yes</td>
</tr>
<tr>
<td>PKIA</td>
<td></td>
</tr>
<tr>
<td>ANKHD1</td>
<td>Yes</td>
</tr>
<tr>
<td>PCDH1</td>
<td></td>
</tr>
<tr>
<td>CCDC85C</td>
<td>Yes</td>
</tr>
<tr>
<td>SETD3</td>
<td></td>
</tr>
<tr>
<td>SUMF1</td>
<td>Yes</td>
</tr>
<tr>
<td>LRRFIP2</td>
<td></td>
</tr>
<tr>
<td>WDR67 (TBC1D31)</td>
<td>Yes</td>
</tr>
<tr>
<td>ZNF704</td>
<td></td>
</tr>
<tr>
<td>DHX35</td>
<td>Yes</td>
</tr>
<tr>
<td>ITC1</td>
<td></td>
</tr>
<tr>
<td>NFS1</td>
<td>Yes *read-through transcription</td>
</tr>
<tr>
<td>PREX1</td>
<td></td>
</tr>
<tr>
<td>CYTH1</td>
<td>Yes *nested inside 2 translocations</td>
</tr>
<tr>
<td>EIF3H</td>
<td></td>
</tr>
</tbody>
</table>

**Genetic Lesion History Analysis Underway**
SK-BR-3 Her2+ Breast Cancer Reference Genome

Released all data pre-publication to accelerate breast cancer research:
http://schatzlab.cshl.edu/data/skbr3/

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

Available soon
• Whole genome methylation analysis
• Full-length cDNA Transcriptome analysis
• Comparison to single cell analysis of >100 individual cells
SK-BR-3 Her2+ Breast Cancer Reference Genome

Release

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

Available soon:
• Whole genome methylation analysis
• Full-length cDNA Transcriptome analysis
• Comparison to single cell analysis of >100 individual cells
Outline

1. **Single Molecule Sequencing**
   *Long read sequencing of a breast cancer cell line*

2. **Single Cell Copy Number Analysis**
   *Intra-tumor heterogeneity and metastatic progression*
Single Cell Sequencing

Recombination / Crossover in germ cells

Circulating tumor cells

Neuronal mosaicism

Clonal Evolution in tumors
Tumour evolution inferred by single-cell sequencing

Nicholas Navin¹,², Jude Kendall¹, Jennifer Troge³, Peter Andrews¹, Linda Rodgers¹, Jeanne McIndoo¹, Kerry Cook³, Asya Stepansky¹, Dan Levy¹, Diane Esposito¹, Lakshmi Muthuswamy³, Alex Krasnitz³, W. Richard McCombie³, James Hicks¹ & Michael Wigler³
Copy-number Profiles
Whole Genome Amplification

Brian Owens, Nature News 2012
Whole Genome Amplification

ONE GENOME FROM MANY
Sequencing the genomes of single cells is similar to sequencing those from multiple cells — but errors are more likely.

DNA is broken into fragments and then sequenced.

The sequences are assembled to give a common, 'consensus' sequence.

Errors introduced in earlier steps make sequence assembly difficult; the final sequence can have gaps.

Brian Owens, Nature News 2012
Whole Genome Amplification Techniques

DOP-PCR (Degenerate Oligonucleotide Primed PCR)

MDA (Multiple Displacement Amplification)

MALBAC (Multiple Annealing and Looping Based Amplification Cycles)

Interactive Analysis and Quality Assessment of Single Cell Copy Number Variations
Garvin, T., Aboukhalil, R. et al. (2014) Under review
Data are noisy

- Potential for biases at every step
  - WGA: Non-uniform amplification
  - Library Preparation: Low complexity, read duplications, barcoding
  - Sequencing: GC artifacts, short reads
  - Computational analysis: mappability, GC correction, segmentation, tree building

Coverage is too sparse and noisy for SNP analysis, requires special processing
1) Binning

Single Cell CNV analysis

- Divide the genome into “bins” with ~50 – 100 reads / bin
- Map the reads and count reads per bin

*Use uniquely mappable bases to establish bins*
1) Binning

Single Cell CNV analysis
- Divide the genome into “bins” with ~50 – 100 reads / bin
- Map the reads and count reads per bin

*Use uniquely mappable bases to establish bins*
Single Cell CNV analysis

- Divide the genome into “bins” with \(~50 – 100\) reads / bin
- Map the reads and count reads per bin

*Use uniquely mappable bases to establish bins*
2) Normalization

Also correct for mappability, GC content, amplification biases
3) Segmentation

Circular Binary Segmentation (CBS)
4) Estimating Copy Number

\[ CN = \text{argmin} \left\{ \sum_{i,j} (\hat{Y}_{i,j} - Y_{i,j}) \right\} \]
5) Cells to Populations
Gingko
http://qb.cshl.edu/ginkgo

Interactive Single Cell CNV analysis & clustering
- Easy-to-use, web interface, parameterized for binning, segmentation, clustering, etc
- Per cell through project-wide analysis in any species

Compare MDA, DOP-PCR, and MALBAC
- DOP-PCR shows superior resolution and consistency

Available for collaboration
- Extending clustering methods, prototyping scRNA
Gingko
http://qb.cshl.edu/ginkgo

Interactive Single Cell CNV analysis & clustering
– Easy-to-use, web interface, parameterized for binning, segmentation, clustering, etc
– Per cell through project-wide analysis in any species
– Compare MDA, DOP-PCR, and MALBAC
– DOP-PCR shows superior resolution and consistency

Available for collaboration
– Extending clustering methods, prototyping

Interactive analysis and quality assessment of single-cell copy-number variations
Tyler Garvin, Robert Aboukhalil, Jude Kendall, Timour Baslan, Gurinder S Atwal, James Hicks, Michael Wigler, Michael Schatz
doi: http://dx.doi.org/10.1101/011346
Pan-Genome Alignment & Assembly

Time to start focusing on problems studying populations of complete genomes
• Available today for many microbial species, near future for higher eukaryotes

Pan-genome colored de Bruijn graph
• Encodes all the sequence relationships between the genomes
• How well conserved is a given sequence?

SplitMEM: A graphical algorithm for pan-genome analysis with suffix skips

Extending reference assembly models
Understanding Genome Structure & Function

**Single Molecule Sequencing**
Now have the ability to **perfectly assemble** microbes and many small eukaryotes, **reference quality** assemblies of larger eukaryotes

**Single Cell Sequencing**
Exciting technologies to probe the genetic and molecular composition of complex environments

*These advances give us incredible power to study how genomes mutate and evolve*
Largely limited by our quantitative power to make comparisons and find patterns
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

**OICR**
- Karen Ng
- Timothy Beck
- Yogi Sundaravadanam
- John McPherson

**NBACC**
- Adam Phillippy
- Serge Koren

**Pacific Biosciences**
- Oxford Nanopore

**SFARI**

**ALFRED P. SLOAN FOUNDATION**
Thank you

http://schatzlab.cshl.edu
@mike_schatz