Algorithms for de novo genome assembly and disease analytics

Michael Schatz
Outline

1. De novo assembly by analogy
2. Long Read Assembly
3. Disease Analytics
Shredded Book Reconstruction

• Dickens accidentally shreds the first printing of *A Tale of Two Cities*
  – Text printed on 5 long spools

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, …

It was the best of times, it was the best of times, it was the age of wisdom, it was the age of foolishness, …

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, …

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, …

• How can he reconstruct the text?
  – 5 copies × 138,656 words / 5 words per fragment = 138k fragments
  – The short fragments from every copy are mixed together
  – Some fragments are identical
Greedy Reconstruction

The repeated sequence make the correct reconstruction ambiguous

- It was the best of times, it was the [worst/age]

Model the assembly problem as a graph problem
de Bruijn Graph Construction

\[ D_k = (V, E) \]
- \( V = \) All length-\( k \) subfragments (\( k < l \))
- \( E = \) Directed edges between consecutive subfragments
  - Nodes overlap by \( k-1 \) words

- Locally constructed graph reveals the global sequence structure
  - Overlaps between sequences implicitly computed

It was the best of

Original Fragment  Directed Edge

It was the best  was the best of

de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001
de Bruijn Graph Assembly

It was the best
was the best of
the best of times,
best of times, it
of times, it was
times, it was the
it was the age
the age of foolishness
was the age of
the age of wisdom,
age of wisdom, it
of wisdom, it was
wisdom, it was the

After graph construction, try to simplify the graph as much as possible
It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness.

After graph construction, try to simplify the graph as much as possible.
The full tale

... it was the **best** of times it was the **worst** of times ...
... it was the **age of wisdom** it was the **age of foolishness** ...
... it was the **epoch of belief** it was the **epoch of incredulity** ...
... it was the **season of light** it was the **season of darkness** ...
... it was the **spring of hope** it was the **winter of despair** ...

---

**age of**

**wisdom**

**foolishness**

**best**

**epoch of**

**belief**

**winter of despair**

**worst**

**spring of hope**

**light**

**epoch of incredulity**

**darkness**

**winter of despair**
N50 size

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

Example: 1 Mbp genome

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

Note:
A “good” N50 size is a moving target relative to other recent publications. 10-20kbp contig N50 is currently a typical value for most “simple” genomes.
Outline

1. De novo assembly by analogy

2. Long read assembly

3. Disease Analytics
De Novo Genome Assembly

• Novel genomes

• Metagenomes

• Sequencing assays
  – Structural variations
  – Transcript assembly
  – …
Assembling a Genome

1. Shear & Sequence DNA

2. Construct assembly graph from overlapping reads
   ...
   AGCCTAGGGATGC
   GGATGCGACACGT
   CGCATATCGGTGGTCTAGCGAA

3. Simplify assembly graph
Assembly Complexity

A
R
B
R
C
R
Assembly Complexity
Long Read Sequencing Technology

PacBio RS II

Molecule

Oxford Nanopore

CSHL/PacBio

(Voskoboynik et al. 2013)

Broad/OxNano @ AGBT
SMRT Sequencing

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

SMRT Sequencing Data

Sample of 100k reads aligned with BLASR requiring >100bp alignment

<table>
<thead>
<tr>
<th>Match</th>
<th>83.7%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insertions</td>
<td>11.5%</td>
</tr>
<tr>
<td>Deletions</td>
<td>3.4%</td>
</tr>
<tr>
<td>Mismatch</td>
<td>1.4%</td>
</tr>
</tbody>
</table>
Consensus Accuracy and Coverage

Coverage can overcome random errors

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

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

*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*

---

**PacBio Coverage**

< 5x  

> 50x
S. cerevisiae W303

PacBio RS II sequencing at CSHL by Dick McCombie
- Size selection using an 7 Kb elution window on a BluePippin™ device from Sage Science

Over 175x coverage in 16 SMRTcells / 2 days using P5-C3

Mean: 5910
83x over 10kbp
8.7x over 20kb
Max: 36,861bp
S. cerevisiae W303

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

PacBio assembly using HGAP + Celera Assembler
- 12.4Mbp; 21 non-redundant contigs; N50: 811kbp; >99.8% id
S. cerevisiae W303
S288C Reference sequence
• 12.1Mbp; 16 chromo + mitochondria; N50: 924kbp

PacBio assembly using HGAP + Celera Assembler
• 12.4Mbp; 21 non-redundant contigs; N50: 811kbp; >99.8% id

Near-perfect assembly:
All but 1 chromosome assembled as a single contig
S. pombe dg21

PacBio RS II sequencing at CSHL
- Size selection using an 7 Kb elution window on a BluePippin™ device from Sage Science

Over 275x coverage in 5 SMRTcells / 1 afternoon using P5-C3

Max: 35,415bp

Mean: 5170

103x over 10kbp

7.6x over 20kb
S. pombe dg21
ASM294 Reference sequence
• 12.6Mbp; 3 chromo + mitochondria; N50: 4.53Mbp
PacBio assembly using HGAP + Celera Assembler
• 12.7Mbp; 13 non-redundant contigs; N50: 3.83Mbp; >99.9% id

Near perfect assembly:
Chr1: 1 contig
Chr2: 2 contigs
Chr3: 2 contigs
MT: 1 contig
A. thaliana Ler-0 sequenced at PacBio

- Sequenced using the previous P4 enzyme and C2 chemistry
- Size selection using an 8 Kb to 50 Kb elution window on a BluePippin™ device from Sage Science
- Total coverage >119x

Genome size: 124.6 Mbp
Chromosome N50: 23.0 Mbp
Corrected coverage: 20x over 10kb

Sum of Contig Lengths: 149.5Mb
N50 Contig Length: 8.4 Mb
Number of Contigs: 1788

High quality assembly of chromosome arms
Assembly Performance: 8.4Mbp/23Mbp = 36%
MiSeq assembly: 63kbp/23Mbp = .2%
ECTools: Error Correction with pre-assembled reads

https://github.com/jgurtowski/ectools

Short Reads -> Assemble Unitigs -> Align & Select - > Error Correct

Can Help us overcome:
1. Error Dense Regions – Longer sequences have more seeds to match
2. Simple Repeats – Longer sequences easier to resolve

However, cannot overcome Illumina coverage gaps & other biases
A. thaliana Ler-0

**O. sativa pv Indica (IR64)**

PacBio RS II sequencing at PacBio
- Size selection using an 10 Kb elution window on a BluePippin™ device from Sage Science

Over 14.1x coverage in 47 SMRTcells using P5-C3

- 12.34x over 10kb
- Mean: 10,232bp
- 4.1x over 20kb
- Max: 54,288bp
O. sativa pv Indica (IR64)

Genome size: ~370 Mb
Chromosome N50: ~29.7 Mbp

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Contig NG50</th>
</tr>
</thead>
<tbody>
<tr>
<td>MiSeq Fragments</td>
<td>19,078</td>
</tr>
<tr>
<td>25x 456bp</td>
<td></td>
</tr>
<tr>
<td>(3 runs 2x300 @ 450 FLASH)</td>
<td></td>
</tr>
<tr>
<td>&quot;ALLPATHS-recipe&quot;</td>
<td>18,450</td>
</tr>
<tr>
<td>50x 2x100bp @ 180</td>
<td></td>
</tr>
<tr>
<td>36x 2x50bp @ 2100</td>
<td></td>
</tr>
<tr>
<td>51x 2x50bp @ 4800</td>
<td></td>
</tr>
<tr>
<td>ECTools</td>
<td>271,885</td>
</tr>
<tr>
<td>10.7x @ 10kbp</td>
<td></td>
</tr>
</tbody>
</table>

ECTools Read Lengths
Mean: 9,348
Max: 54,288bp
10.75x over 10kbp
What should we expect from an assembly?

https://en.wikipedia.org/wiki/Genome_size
Assembly Complexity of Long Reads

**Assembly complexity of long read sequencing**

Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz MC et al. (2014) *In preparation*
Assembly Complexity of Long Reads

Assembly complexity of long read sequencing
Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz MC et al. (2014) In preparation
Assembly Complexity of Long Reads

Assembly complexity of long read sequencing
Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz MC et al. (2014) In preparation
Assembly Recommendations

• Long read sequencing of eukaryotic genomes is here

• Recommendations
  
  < 100 Mbp: HGAP/PacBio2CA @ 100x PB C3-P5
  expect near perfect chromosome arms

  < 1GB: HGAP/PacBio2CA @ 100x PB C3-P5
  expect high quality assembly: contig N50 over 1Mbp

  > 1GB: hybrid/gap filling
  expect contig N50 to be 100kbp – 1Mbp

  > 5GB: Email mschatz@cshl.edu

• Caveats
  – Model only as good as the available references (esp. haploid sequences)
  – Technologies are quickly improving, exciting new scaffolding technologies
Pan-Genome Alignment & Assembly

Time to start considering problems for which N complete genomes is the input to study the “pan-genome”
- 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?
- What are the pan-genome network properties?

Rapid pan genome analysis with augmented suffix trees
Marcus, S, Schatz, MC (2014) *In preparation*
Outline

1. De novo assembly by analogy
2. Long Read Assembly
3. Disease Analytics
Variation Detection Complexity

SNPs + Short Indels

High precision and sensitivity

```
. . TTTAGAATAG-CGAGTGC . .
AGAATAGGCGAG
```

“Long” Indels (>5bp)

Reduced precision and sensitivity

```
. . TTTAG----------AGTGC . .
TTTAGAATAGGC
ATAGGCGAGTGC
```

Analysis confounded by sequencing errors, localized repeats, allele biases, and mismapped reads
Scalpel: Haplotype Microassembly

DNA sequence **micro-assembly** pipeline for accurate detection and validation of *de novo* mutations (SNPs, indels) within exome-capture data.

Features

1. Combine **mapping** and **assembly**
2. Exhaustive search of **haplotypes**
3. **De novo mutations**

**Accurate detection of de novo and transmitted INDELs within exome-capture data using micro-assembly**

Scalpel Pipeline

1. Extract reads mapping within the exon including (1) well-mapped reads, (2) soft-clipped reads, and (3) anchored pairs.

2. Decompose reads into overlapping $k$-mers and construct de Bruijn graph from the reads.

3. Find end-to-end haplotype paths spanning the region.

4. Align assembled sequences to reference to detect mutations.

- deletion
- insertion
Simulation Analysis

Simulated 10,000 indels in an exome from a known log-normal distribution
Selected one deep coverage exome for deep analysis
- Individual was diagnosed with ADHD and turrets syndrome
- 80% of the target at >20x coverage
- Evaluated with Scalpel, SOAPindel, and GATK Haplotype Caller

1000 indels selected for validation
- 200 Scalpel
- 200 GATK Haplotype Caller
- 200 SOAPindel
- 200 within the intersection
- 200 long indels (>30bp)
Scalpel Indel Discovery

Scalpel

SOAPindel

HaplotypeCaller
Scalpel Indel Discovery

Scalpel

SOAPindel

HaplotypeCaller
Scalpel Indel Discovery

SOAPindel: ABC'BCB'D
Scalpel: ABC'B'D
Exome sequencing of the SSC

Last year saw 3 reports of >593 families from the Simons Simplex Collection

• Parents plus one child with autism and one non-autistic sibling

• All attempted to find mutations enriched in the autistic children

• Iossifov (343) and O’Roak (50) used GATK, Sanders (200) didn’t attempt to identify indels

De novo gene disruptions in children on the autism spectrum

De novo mutations revealed by whole-exome sequencing are strongly associated with autism

Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations
De novo mutation discovery and validation

**Concept:** Identify mutations not present in parents.

**Challenge:** Sequencing errors in the child or low coverage in parents lead to false positive de novos

**Reference:** ...TCAAATCCTTTTTAATAAAAGAAGAGCTGACA...

Father: ...TCAAATCCTTTTTAATAAAAGAAGAGCTGACA...
Mother: ...TCAAATCCTTTTTAATAAAAGAAGAGCTGACA...
Sibling: ...TCAAATCCTTTTTAATAAAAGAAGAGCTGACA...
Proband (1): ...TCAAATCCTTTTTAATAAAAGAAGAGCTGACA...
Proband (2): ...TCAAATCCTTATTTAAT***AAGAGCTGACA...

4bp heterozygous deletion at chr15:93524061 CHD2
De novo Genetics of Autism

• In 593 family quads so far, we see significant enrichment in de novo *likely gene killers* in the autistic kids
  – Overall rate basically 1:1
  – 2:1 enrichment in nonsense mutations
  – 2:1 enrichment in frameshift indels
  – 4:1 enrichment in splice-site mutations
  – Most de novo originate in the paternal line in an age-dependent manner (56:18 of the mutations that we could determine)

• Observe strong overlap with fragile X protein (FMPR) network
  – Related to neuron development and synaptic plasticity
  – Also strong overlap with chromatin remodelers

**Accurate detection of de novo and transmitted INDELs within exome-capture data using micro-assembly**
Summary

New Biotechnology
- Sequencing: Pacific Biosciences, Moleculo, Oxford Nanopore
- Mapping: Hi-C interactions, BioNanoGenomics, OpGen
- Faster/Cheaper/Better assemblies

Algorithmics
- Indexing and compressing of very large datasets
- Improved error correction, large graph analysis
- Networks and populations of genomes

Annotation & Comparative Genomics
- Identifying functional elements
- Cross species comparisons, models of evolution
- Identifying mutations responsible for disease and other traits
Acknowledgements

Schatz Lab
James Gurtowski
Hayan Lee
Giuseppe Narzisi

Shoshana Marcus
Srividya
Ramakrishnan
Rob Aboukhalil
Mitch Bekritsky
Charles Underwood
Tyler Gavin
Alejandro Wences
Greg Vupture
Eric Biggers
Aspyn Palatnick

CSHL
McCombie Lab
Wigler Lab

Hannon Lab
Gingeras Lab
Jackson Lab
Hicks Lab
Iossifov Lab
Levy Lab
Lippman Lab
Lyon Lab
Martienssen Lab
Tuveson Lab
Ware Lab

Pacific Biosciences
Thank you
http://schatzlab.cshl.edu
@mike_schatz