De novo assembly of complex genomes
Michael Schatz

April 10, 2013
CPHG, University of Virginia
Schatz Lab Overview

Computation

Human Genetics

Sequencing

Modeling

Plant Genomics
Outline

1. Genome assembly by analogy
2. Hybrid error correction and assembly
3. De novo mutations in autism
Outline

1. Genome assembly by analogy
2. Hybrid error correction and assembly
3. De novo mutations in autism
Shredded Book Reconstruction

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

<table>
<thead>
<tr>
<th>It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...</td>
</tr>
<tr>
<td>It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...</td>
</tr>
<tr>
<td>It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...</td>
</tr>
<tr>
<td>It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...</td>
</tr>
</tbody>
</table>

- 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 sequence reconstruction as a graph problem.
de Bruijn Graph Construction

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

Original Fragment  Directed Edge

It was the best of  It was the best  was the best of

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

de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001
After graph construction, try to simplify the graph as much as possible
de Bruijn Graph Assembly

It was the best of times, it of times, it was the

it was the worst of times, it

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 winder of despair …
Generally an exponential number of compatible sequences

- Value computed by application of the BEST theorem

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

\[ L = n \times n \text{ matrix with } r_u - a_{uu} \text{ along the diagonal and } -a_{uv} \text{ in entry } uv \]

\[ r_u = d^+(u) + 1 \text{ if } u = t, \text{ or } d^+(u) \text{ otherwise} \]

\[ a_{uv} = \text{ multiplicity of edge from } u \text{ to } v \]

---

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:
N50 values are only meaningful to compare when base genome size is the same in all cases
Outline

1. Genome assembly by analogy

2. Hybrid error correction and assembly

3. De novo mutations in autism
Assembly Applications

Novel genomes

Metagenomes

Sequencing assays
- Transcript assembly
- Structural variations
- Haplotype analysis
- ...

Genome 10K

HMP

...
Why are genomes hard to assemble?

1. **Biological:**
   - (Very) High ploidy, heterozygosity, repeat content

2. **Sequencing:**
   - (Very) large genomes, imperfect sequencing

3. **Computational:**
   - (Very) Large genomes, complex structure

4. **Accuracy:**
   - (Very) Hard to assess correctness
Ingredients for a good assembly

**Read Length**

- Reads & mates must be longer than the repeats
  - Short reads will have false overlaps forming hairball assembly graphs
  - With long enough reads, assemble entire chromosomes into contigs

**Coverage**

- High coverage is required
  - Oversample the genome to ensure every base is sequenced with long overlaps between reads
  - Biased coverage will also fragment assembly

**Quality**

- Errors obscure overlaps
  - Reads are assembled by finding kmers shared in pair of reads
  - High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Current challenges in de novo plant genome sequencing and assembly
Hybrid Sequencing

**Illumina**  
*Sequencing by Synthesis*  
High throughput (60Gbp/day)  
High accuracy (~99%)  
Short reads (~100bp)

**Pacific Biosciences**  
*SMRT Sequencing*  
Lower throughput (600Mbp/day)  
Lower accuracy (~90%)  
Long reads (2-5kbp+)
SMRT Sequencing

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

SMRT Read Types

- **Standard sequencing**
  - Long inserts so that the polymerase can synthesize along a single strand

- **Circular consensus sequencing**
  - Short inserts, so polymerase can continue around the entire SMRTbell multiple times and generate multiple sub-reads from the same single molecule.
  - Barbell sequence: ATCTCTCTTttttctccctccgcgtttggtttGAGAGAGAT

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 Quality: Probability Review

Roll $n$ dice => What is the probability that at least half are 6’s?

<table>
<thead>
<tr>
<th>$n$</th>
<th>Min to Lose</th>
<th>Losing Events</th>
<th>$P$(Lose)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>$1/6$</td>
<td>16.7%</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>$P(1 \text{ of } 2) + P(2 \text{ of } 2)$</td>
<td>30.5%</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>$P(2 \text{ of } 3) + P(3 \text{ of } 3)$</td>
<td>7.4%</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>$P(2 \text{ of } 4) + P(3 \text{ of } 4) + P(4 \text{ of } 4)$</td>
<td>13.2%</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>$P(3 \text{ of } 5) + P(4 \text{ of } 5) + P(5 \text{ of } 5)$</td>
<td>3.5%</td>
</tr>
</tbody>
</table>

For $n \geq 2$, the probability that at least half are 6’s is given by the formula:

$$
\sum_{i=\lceil n/2 \rceil}^{n} P(i \text{ of } n) = \sum_{i=\lceil n/2 \rceil}^{n} \binom{n}{i} (p)^i (1-p)^{n-i}
$$
**Consensus Accuracy and Coverage**

Coverage can overcome random errors

- Dashed: error model from binomial sampling; solid: observed accuracy
- For same reason, CCS is extremely accurate when using 5+ subreads

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

• Motivations
  – 15 crops provide 90% of the world’s food
  – Responsible for maintaining the balance of the carbon cycles, soil from erosion
  – Promising sources of renewable energy
  – Plant byproducts used in many medicines
  – Model organisms for studying biological systems

• Challenges
  – Very large genomes, some many times larger than human
  – High repeat content, especially high copy retrotransposons
  – High ploidy, high heterozygosity
Population structure in *Oryza sativa*

3 varieties selected for de novo sequencing

- High quality BAC-by-BAC reference
  - ~370 Mbp genome in 12 chromosomes
  - About 40% repeats:
    - Many 4-8kbp repeats
    - 300kbp max high identity repeat (99.99%)
  - Useful model for other cereal genomes

Garris et al. (2005)
*Genetics* 169: 1631–1638
PacBio Long Read Rice Sequencing

C1 Chemistry – Summer 2011
Median=639 Mean=824 Max=10,008

C2XL Chemistry – Summer 2012
Median=2231 Mean=3290 Max=24,405
## Preliminary Rice Assemblies

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Contig NG50</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HiSeq Fragments</strong></td>
<td>3,925</td>
</tr>
<tr>
<td>50x 2x100bp @ 180</td>
<td></td>
</tr>
<tr>
<td><strong>MiSeq Fragments</strong></td>
<td>6,332</td>
</tr>
<tr>
<td>23x 459bp</td>
<td></td>
</tr>
<tr>
<td>8x 2x251bp @ 450</td>
<td></td>
</tr>
<tr>
<td><strong>“ALLPATHS-recipe”</strong></td>
<td>18,248</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><strong>PBeCR Reads</strong></td>
<td>50,995</td>
</tr>
<tr>
<td>7x @ 3500 ** MiSeq for correction</td>
<td></td>
</tr>
<tr>
<td><strong>PBeCR + Illumina Shred</strong></td>
<td>59,695</td>
</tr>
<tr>
<td>7x @ 3500 ** MiSeq for correction</td>
<td></td>
</tr>
<tr>
<td>5x @ 3000bp shred</td>
<td></td>
</tr>
</tbody>
</table>

In collaboration with McCombie & Ware labs @ CSHL
Improved Gene Reconstruction

FOXP2 assembled in a single contig in the PacBio parrot assembly

Long Read CNV Analysis

Aluminum tolerance in maize is important for drought resistance and protecting against nutrient deficiencies

- Segregating population localized a QTL on a BAC, but unable to genotype with Illumina sequencing because of high repeat content and GC skew
- Long read PacBio sequencing corrected by CCS reads revealed a triplication of the ZnMATE1 membrane transporter

A rare gene copy-number variant that contributes to maize aluminum tolerance and adaptation to acid soils
Maron, LG et al. (2012) PNAS. doi: 10.1073/pnas.1220766110
Transcript Alignment

- Long-read single-molecule sequencing has potential to directly sequence full length transcripts
  - Raw reads and raw alignments (red) have many spurious indels inducing false frameshifts and other artifacts
  - Error corrected reads almost perfectly match the genome, pinpointing splice sites, identifying alternative splicing

- New collaboration with Gingeras Lab looking at splicing in human

Assembly Coverage Model

Simulate PacBio-like reads to predict how the assembly will improve as we add additional coverage.

Only 8x coverage is needed to sequence every base in the genome, but 40x improves the chances repeats will be spanned by the longest reads.

Assembly complexity of long read sequencing
Internal Roadmap has made steady progress towards improving read length and throughput

Very recent improvements:

1. Improved enzyme:
   Maintain reactions longer

2. “Hot Start” technology:
   Maximize subreads

3. MagBead loading:
   Load longest fragments

PacBio Users Meeting, June 18, Frederick MD
Speculation for 2014

Doubling the average read length dramatically improves the assembly quality

• Able to span a larger repeats and lock contigs together

Expect to see contig N50 values over 1Mbp very soon, even in very complicated plant and animal species

• Megabase contig N50 already routine in microbial assembly with PacBio sequencing
Speculation for 2014

With PacBio-like reads averaging 11.2kbp (4x current), we should be able to assemble almost every chromosome arm of rice into single contigs
  • The 300kbp near perfect repeat is the only exception

Even with the current assembly, we are seeing new genes and other sequences missing in the “high quality” BAC-by-BAC reference genome.
Speculation for 2015

For human, it will still take a few more rounds of read length doubling before we should expect to see single contig chromosome arms.

However, we can still learn a lot of interesting biology about the ~13% of the human genome that is currently inaccessible.

Outline

1. Genome assembly by analogy

2. Hybrid error correction and assembly

3. De novo mutations in autism
Variation Detection Complexity

**SNPs + Short Indels**
High precision and sensitivity

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

```
TTTAGAATAGGCAG
```

**“Long” Indels (>5bp)**
Reduced precision and sensitivity

```
..TTTAG-------AGTGC...
```

```
TTTAGAATAGGC
ATAGGCGAGTGC
```

Analysis confounded by sequencing errors, localized repeats, allele biases, and mismapped reads

Sens: 86%
FDR: .19%
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

NRXN1 de novo SNP
(auSSC12501 chr2:50724605)
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.
Repeats in the Genome

Specificity Challenge: 30% of exons have a perfect 10bp or larger repeat

Reference Exon: Localized repeat sequence

Variant Read: Large deletion or critical snp?
Scalpel Indel Discovery

Detection of de novo mutations in exome-capture data using micro-assembly
Narzisi et al. (2013) In preparation
Scalpel Indel Discovery

Detection of de novo mutations in exome-capture data using micro-assembly
Narzisi et al. (2013) In preparation
Scalpel Indel Discovery

Detection of de novo mutations in exome-capture data using micro-assembly
Narzisi et al. (2013) In preparation
Scalpel Indel Discovery

Detection of de novo mutations in exome-capture data using micro-assembly
Narzisi et al. (2013) In preparation
Exome sequencing of the SSC

Sequencing of 343 families from the Simons Simplex Collection
- Parents plus one child with autism and one non-autistic sibling
- Enriched for higher-functioning individuals

Families prepared and captured together to minimize batch effects
- Exome-capture performed with NimbleGen SeqCap EZ Exome v2.0 targeting 36 Mb of the genome.
- ~80% of the target at >20x coverage with ~93bp reads

De novo gene disruptions in children on the autism spectrum
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.

Ref: `...TCAGAACAGCTGGATGAGATCTTAGGGAACCTACCAGGAGATTGTCTTTGCCCAGGA...`

Father: `...TCAGAACAGCTGGATGAGATCTTAGGGAACCTACCAGGAGATTGTCTTTGCCCAGGA...`

Mother: `...TCAGAACAGCTGGATGAGATCTTAGGGAACCTACCAGGAGATTGTCTTTGCCCAGGA...`

Sib: `...TCAGAACAGCTGGATGAGATCTTAGGGAACCTACCAGGAGATTGTCTTTGCCCAGGA...`

Aut(1): `...TCAGAACAGCTGGATGAGATCTTAGGGAACCTACCAGGAGATTGTCTTTGCCCAGGA...`

Aut(2): `...TCAGAACAGCTGGATGAGATCTTAGGGAACCTACCAGGAGATTGTCTTTGCCCAGGA...`

6bp heterozygous deletion at chr13:25280526 ATP12A
De novo Genetics of Autism

- In 343 family quads so far, we see significant enrichment in de novo likely gene killers in the autistic kids
  - Overall rate basically 1:1 (432:396)
  - 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 the 842 genes known to be associated with fragile X protein FMRP
  - Related to neuron development and synaptic plasticity
  - Also strong overlap with chromatin remodelers

De novo gene disruptions in children on the autism spectrum
Summary

• Hybrid assembly let us combine the best characteristics of 2\textsuperscript{nd} and 3\textsuperscript{rd} gen sequencing
  – Long reads and good coverage are the keys to a good de novo assembly
  – Single contig de novo assemblies of entire microbial chromosomes are now routine; Single contig de novo assemblies of entire plant and animal chromosomes on the horizon

• Assembly is the missing link towards high accuracy indel mutation discovery
  – Allows the algorithm to break free from the expectations of the reference
  – Pinpointing de novo mutations require both high sensitivity and specificity

• We are starting to apply these technologies to discover significant biology that is otherwise impossible to measure
## Acknowledgements

<table>
<thead>
<tr>
<th>Schatz Lab</th>
<th>CSHL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Giuseppe Narzisi</td>
<td>Hannon Lab</td>
</tr>
<tr>
<td>Shoshana Marcus</td>
<td>Gingeras Lab</td>
</tr>
<tr>
<td>James Gurtowski</td>
<td>Iossifov Lab</td>
</tr>
<tr>
<td>Alejandro Wences</td>
<td>Levy Lab</td>
</tr>
<tr>
<td>Hayan Lee</td>
<td>Lippman Lab</td>
</tr>
<tr>
<td>Rob Aboukhalil</td>
<td>Lyon Lab</td>
</tr>
<tr>
<td>Mitch Bekritsky</td>
<td>Martienssen Lab</td>
</tr>
<tr>
<td>Charles Underwood</td>
<td>McCombie Lab</td>
</tr>
<tr>
<td>Rushil Gupta</td>
<td>Ware Lab</td>
</tr>
<tr>
<td>Avijit Gupta</td>
<td>Wigler Lab</td>
</tr>
<tr>
<td>Shishir Horane</td>
<td></td>
</tr>
<tr>
<td>Deepak Nettem</td>
<td>IT Department</td>
</tr>
<tr>
<td>Varrun Ramani</td>
<td></td>
</tr>
<tr>
<td>Piyush Kansal</td>
<td>NBACC</td>
</tr>
<tr>
<td>Eric Biggers</td>
<td>Adam Phillippy</td>
</tr>
<tr>
<td>Aspy Palatnick</td>
<td>Sergey Koren</td>
</tr>
</tbody>
</table>
Thank You!

Michael Schatz @mike_schatz
Can you assemble genomes, find mutations, and decode secret messages? Get ready for the #DNA60IFX challenge! bit.ly/16VKqsG

Expand