Analysis of Structural Variants using 3rd generation Sequencing

Michael Schatz

January 12, 2016
Bioinformatics / PAG XXIV

@mike_schatz / #PAGXXIV
Analysis of Structural Variants using 3rd generation Sequencing

Michael Schatz

January 12, 2016
Bioinformatics / PAG XXIV

@mike_schatz / #PAGXXIV
# The Resurgence of Reference Quality Genomes

Michael Schatz & Daniel Rokhsar  
Tuesday, January 12, 2016 @ 4pm – 6pm  
Town & Country - Pacific Salon 1

<table>
<thead>
<tr>
<th>Time</th>
<th>Title</th>
<th>Speaker(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4:00pm</td>
<td>The Resurgence of Reference Quality Genomes</td>
<td>Michael Schatz, CSHL + JHU</td>
</tr>
<tr>
<td>4:20pm</td>
<td>High Quality, Highly Contiguous Genome Assemblies Now</td>
<td>Richard Green, Dovetail Genomics</td>
</tr>
<tr>
<td>4:40pm</td>
<td>Scalable Parallel Algorithms for de novo Assembly of Complex Genomes</td>
<td>Aydin Buluc, Lawrence Berkeley National Laboratory</td>
</tr>
<tr>
<td>5:00pm</td>
<td>Using PacBio Long Reads to Generate a High Quality Reference for the Allotetraploid Coffea arabica and its Maternal Diploid Ancestor Coffea eugeniodes</td>
<td>Marcela Yepes, Cornell University</td>
</tr>
<tr>
<td>5:20pm</td>
<td>MaSuRCA Mega-Reads Assembly Technique for Haplotype Resolved Genome Assembly of Hybrid PacBio and Illumina Data</td>
<td>Aleksey Zimin, University of Maryland</td>
</tr>
<tr>
<td>5:40pm</td>
<td>How to Compare and Cluster Every Known Genome in about an Hour</td>
<td>Sergey Koren, NHGRI</td>
</tr>
</tbody>
</table>
Structural Variations

Any mutation >50bp

Profound impact on genome structure and function

Genome structural variation discovery and genotyping
## Structural Variation Sequence Signatures

<table>
<thead>
<tr>
<th>SV classes</th>
<th>Read pair</th>
<th>Read depth</th>
<th>Split read</th>
<th>Assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deletion</td>
<td><img src="image1.png" alt="Diagram" /></td>
<td><img src="image2.png" alt="Diagram" /></td>
<td><img src="image3.png" alt="Diagram" /></td>
<td><img src="image4.png" alt="Diagram" /></td>
</tr>
<tr>
<td>Novel sequence insertion</td>
<td><img src="image5.png" alt="Diagram" /></td>
<td><img src="image6.png" alt="Diagram" /></td>
<td><img src="image7.png" alt="Diagram" /></td>
<td><img src="image8.png" alt="Diagram" /></td>
</tr>
<tr>
<td>Mobile-element insertion</td>
<td><img src="image9.png" alt="Diagram" /></td>
<td><img src="image10.png" alt="Diagram" /></td>
<td><img src="image11.png" alt="Diagram" /></td>
<td><img src="image12.png" alt="Diagram" /></td>
</tr>
<tr>
<td>Inversion</td>
<td><img src="image13.png" alt="Diagram" /></td>
<td><img src="image14.png" alt="Diagram" /></td>
<td><img src="image15.png" alt="Diagram" /></td>
<td><img src="image16.png" alt="Diagram" /></td>
</tr>
<tr>
<td>Interspersed duplication</td>
<td><img src="image17.png" alt="Diagram" /></td>
<td><img src="image18.png" alt="Diagram" /></td>
<td><img src="image19.png" alt="Diagram" /></td>
<td><img src="image20.png" alt="Diagram" /></td>
</tr>
<tr>
<td>Tandem duplication</td>
<td><img src="image21.png" alt="Diagram" /></td>
<td><img src="image22.png" alt="Diagram" /></td>
<td><img src="image23.png" alt="Diagram" /></td>
<td><img src="image24.png" alt="Diagram" /></td>
</tr>
</tbody>
</table>
**Structural Variation Sequence Signatures**

<table>
<thead>
<tr>
<th>SV classes</th>
<th>Read pair</th>
<th>Read depth</th>
<th>Split read</th>
<th>Assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deletion</td>
<td><img src="image" alt="Diagram" /></td>
<td><img src="image" alt="Diagram" /></td>
<td><img src="image" alt="Diagram" /></td>
<td><img src="image" alt="Diagram" /></td>
</tr>
</tbody>
</table>

**PacBio Sequel**

- >10kbp Mean Read Lengths
- ~$15k / Mammalian-sized genome

**Single Molecule Sequencing**
- No amplification artifacts
- More uniform coverage
- Essentially no GC biases

**Long read lengths**
- Improved mappability
- More likely to span breakpoints
- More robust split read analysis
- More robust assemblies

**Basepair resolution for 50bp through 50Mbp events**
SK-BR-3

Most commonly used Her2-amplified breast cancer cell line

Highly-rearranged Mammalian genome
80 chromosomes instead of 46
Numerous chromosome fusions, rearrangements, other SVs

(Davidson et al, 2000)
PacBio Long-Read Sequencing

mean read length: 9 kb
max read length: 71 kb

72X overall coverage

Genome-wide coverage averages around 54X
Coverage per chromosome varies greatly as expected from previous karyotyping results
Genome structural analysis

Assembly-based

- Assembly with Falcon on DNAnexus
  - Alignment with MUMmer
    - Call variants between consecutive alignments with ABVC
    - Call variants within alignments with ABVC

~ 11,000 local variants
50 bp < size < 10 kbp

Alignment-based

- Alignment with BWA-MEM
  - Copy number analysis
  - SV-calling from split reads with Sniffles
    - Validations
    - SplitThreader
      - Detailed analysis of Her2 amplifications

661 long-range variants
(>10kb distance)
Genome structural analysis

Assembly-based

- Assembly with Falcon on DNAnexus
  - Alignment with MUMmer
    - Call variants between consecutive alignments with ABVC
    - Call variants within alignments with ABVC

~ 11,000 local variants
50 bp < size < 10 kbp

Alignment-based

- Alignment with BWA-MEM
  - Copy number analysis
  - SV-calling from split reads with Sniffles
    - Validations
    - SplitThreader
      - Detailed analysis of Her2 amplifications

661 long-range variants
(>10kb distance)
Assembly using PacBio yields far better contiguity

Number of sequences: 13,532  
Total sequence length: 2.97 Gb  
Mean: 266 kb  
Max: 19.9 Mb  
N50: 2.46 Mb

Relative to a genome size of 3 Gb

Number of sequences: 748,955  
Total sequence length: 2.07 Gb  
Mean: 2.8 kb  
Max: 61 kb  
N50: 3.3 kb
ABVC: Assembly-Based Variant-Caller

Defined point

Overlapping alignments suggest tandem repeat

Gap where sequences do not align uniquely suggests a repeat

Insertion

Deletion

Tandem Expansions

Tandem Contractions

Repeat Expansions

Repeat Contractions
Assembly-based analysis highly effective for local SVs (<10kbp)

- ~11,000 SVs between 50bp and 10kbp in size, totaling >10Mbp of variation
- Essentially perfect positive predictive value

Alignment artifacts confound larger events (>10kbp)

- WGA alignments confused by large repetitive elements near SVs
- SV breakpoints may be poorly spanned by a contig
  - ~100bp on one side, 1Mbp on the other
Genome structural analysis

Assembly-based

- Assembly with Falcon on DNAnexus
- Alignment with MUMmer
  - Call variants between consecutive alignments with ABVC
  - Call variants within alignments with ABVC

~11,000 local variants
50 bp < size < 10 kbp

Alignment-based

- Alignment with BWA-MEM
  - Copy number analysis
  - SV-calling from split reads with Sniffles
    - Validations
    - SplitThreader
      - Detailed analysis of Her2 amplifications

661 long-range variants
(>10kb distance)
Long Read Structural Variation Analysis

Split-read analysis greatly improved by long reads
- Improved chances of spanning event, including nested events
- However, many SVs lost due to poor alignments and poor PacBio support
  - LUMPY fails on reads that span more than 1 breakpoint, poor localization

New methods in development: NGM-LR + Sniffles
- **NGM-LR**: Improve mapping of noisy long reads
- **Sniffles**: Integrates SV evidence from split alignments, alignment fidelity (CIGAR string and MD tag)
Mapping a ~500bp deletion

Similar issues for insertions, inversions; or Nanopore sequencing
Improved seeding, improved gap scoring: convex instead of affine
Long-range structural variants found by Sniffles

661 long-range variants (>10kb distance)
Long-range structural variants found by Sniffles
Long-range structural variants found by Sniffles
SplitThreader
Threading SV breakpoints to infer the history of rearrangements in complex genomes

CHR 1
ATCGCCTA

CHR 2
GTCCATAG

CHR 1
ATCG
CCGA
ATAG

CHR 2
GTCC
ATAG

20
80
100
1. Healthy chromosome 17 & 8
2. Translocation into chromosome 8
3. Translocation within chromosome 8
4. Complex variant and inverted duplication within chromosome 8
5. Translocation within chromosome 8
1. Healthy chromosome 17 & 8
2. Translocation into chromosome 8
3. Translocation within chromosome 8
4. Complex variant and inverted duplication within chromosome 8
5. Translocation within chromosome 8

Inferring the evolution of genome structure
**Summary & Acknowledgements**

**ABVC + SplitThreader by Maria Nattestad**
- Assembly-based variant analysis is efficient and accurate
  - 10s of thousands variants present in mammalian-sized genomes
- SplitThreader infers the evolution to genome structure
  - Additional context as genes are moved next to new promoters and other regulatory elements

**NGM-LR + Sniffles by Fritz Sedazeck**
- Correct long read mapping is essential for SV analysis
  - Design the mapping strategy for the error model of the data
- Integrate all available information for robust SV calling
  - Currently extending to other long-range mapping technologies: Oxford Nanopore, BioNano, 10X Genomics

**Special thanks to Dick McCombie (CSHL), John McPherson (OICR), PacBio**
**Funding by NSF, NIH, DOE, Sloan Foundation**
Thank you!

The Resurgence of Reference Quality Genomes

Michael Schatz & Daniel Rokhsar
Tuesday January 12, 2016 @ 4pm – 6pm
Town & Country - Pacific Salon 1