Introductions

Srividya “Sri” Ramakrishnan
DOE Systems Biology Knowledgebase
Worlds fastest genomics pipelines

Tyler Garvin
WSBS
Interactive CNV and QC of single cell sequencing

Greg Vurture
CSHL URP / NYU
Mathematics of genomic architecture and heterozygosity

Aspyn Palatnick
CSH HS iGenomics
Outline

1. Read length & assembly complexity
2. Single molecule assembly of rice
3. De novo indel mutations in autism
Assembling a Genome

1. Shear & Sequence DNA

2. Construct assembly graph from overlapping reads
   ...AGCCTAGGGATGCGCGACACGT
   GGATGCGCGACACGTGCATATCCGGTTTGGTCAACCTCGGACGGAC
   CAACCTCGGACGGACCTCAGCGAA...

3. Simplify assembly graph
Assembly Complexity
Reducing Complexity

Longer reads span more repeats, simplifying the assembly problem

- Idealized assembly of *B. anthracis* reduces to a single contig with 5kb reads
- Exact improvement depends on the specific genome

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

![Diagram showing N50 size calculation](image)

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. Read length & assembly complexity
2. Single molecule assembly of rice
3. De novo indel mutations in autism
Population structure of *Oryza sativa*

3 varieties selected for *de novo* sequencing

Garris et al. (2005)
*Genetics* 169: 1631–1638
New whole genome de novo assemblies of three divergent strains of rice documents novel gene space of Aus and Indica subpopulations
Single Molecule Sequencing Technology

PacBio RS II

Moleculo

Oxford Nanopore

Clive G. Brown @Clive_G_Brown
I’ve reluctantly rejoined twitter purely so that I can make one tweet - when the appropriate time arises...

Expand
SMRT Sequencing Data

TTGTAAGCAGTTGAAAACTATGTGTGATTTAGAATAAGAAGCATGAAAG
TTGTAAGCAGTTGAAAAACTATGTGTGATTTAGAATAAGAAGCATGAAAG
ATTATAAA-CAGTTGATCCATT-AGAAGA-AAAAGCGCAAAGCGGCTAGGG
ATTATAAA-TCAGTTGATCCATT-AGAAACGC-AAAAGGC-GCTAGGG
CAACCTTGAAATGTAATCGCACTTGAAGAAAGATTATTATCTCGCGCCCG
CAACCTTGAAATGTAATCGCACTTGAAGAAAGATTATTATCTCGCGCCCG
TAACGAATCAAGATTTCTGAAAACACATATAACAACCTCCAAAA-CACAA
TAACGAATCAAGATTTCTGAAAACACATATAACAACCTCCAAAA-CACAA
-AGGAGGGAAAGGGGGAATATCT-ATAAAAGATTTGAAATAAGTGA
-AGGAGGGAAAGGGGGAATATCT-ATAAAAGATTTGAAATAAGTGA
ACT-AATTCAACAATA-ATAAACACTTTTA-ACGAATGGT-AGAA-GTT
ACT-AATTCAACAATA-ATAAACACTTTTA-ACGAATGGT-AGAA-GTT
TCGGAGAGATCCAAACAATTGACC-ATCGCCCTTGA-GTTACC-AATCAA
TCGGAGAGATCCAAACAATTGACC-ATCGCCCTTGA-GTTACC-AATCAA
TC-GAGATCAGGAAACATG-GGCACTC-CTTTGACGTACTAAATCAA
TC-GAGATCAGGAAACATG-GGCACTC-CTTTGACGTACTAAATCAA
ATCCAGTGAAATATATTATTTATGCATCAGGSAGGTACCATTACATTAG
ATCCAGTGAAATATATTATTTATGCATCAGGSAGGTACCATTACATTAG
ATCCAGTGAAATATATTATTTATGCATCAGGSAGGTACCATTACATTAG
ATCCAGTGAAATATATTATTTATGCATCAGGSAGGTACCATTACATTAG

<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>

Sample of 100k reads aligned with BLASR requiring >100bp alignment
PacBio Error Correction: HGAP

- With 50-100x of Pacbio coverage, virtually all of the errors can be eliminated
  - Works well for Microbial genomes: single contig per chromosome routinely achieved
  - Difficult to scale up for use with eukaryotic genomes

**Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data**
Hybrid Sequencing

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

**Pacific Biosciences**
*SMRT Sequencing*
- Lower throughput (1Gbp/day)
- Lower accuracy (~85%)
- Long reads (5kbp+)
1. Correction Pipeline
   1. Map short reads to long reads
   2. Trim long reads at coverage gaps
   3. Compute consensus for each long read

2. Error corrected reads can be easily assembled, aligned

Hybrid error correction and de novo assembly of single-molecule sequencing reads.

Hybrid Error Correction: PacBioToCA
http://wgs-assembler.sf.net
Improved Gene Reconstruction

FOXP2 assembled in a single contig in the PacBio parrot assembly

Hybrid error correction and de novo assembly of single-molecule sequencing reads.
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. (2013) PNAS doi: 10.1073/pnas.1220766110
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>HiSeq Fragments</td>
<td>3,925</td>
</tr>
<tr>
<td>50x 2x100bp @ 180</td>
<td></td>
</tr>
<tr>
<td>MiSeq Fragments</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>“ALLPATHS-recipe”</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>PBeCR Reads</td>
<td>50,995</td>
</tr>
<tr>
<td>19x @ 3500 ** MiSeq for correction</td>
<td></td>
</tr>
</tbody>
</table>

In collaboration with McCombie & Ware labs @ CSHL
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**

Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz MC et al. (2013) *In preparation*
Enhanced PacBio Error Correction

PacBioToCA fails in complex regions
1. Simple Repeats – Kmer Frequency Too High to Seed Overlaps
2. Error Dense Regions – Difficult to compute overlaps with many errors
3. Extreme GC – Lacks Illumina Coverage
Error Correction with pre-assembled Illumina reads

https://github.com/jgurtowski/pbtools

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

Unitigs:
High quality contigs formed from unambiguous, unique overlaps of reads
Each read is placed into a single unitig

Can Help us overcome:
1. Simple Repeats – Kmer Frequency Too High to Seed Overlaps
2. Error Dense Regions – Difficult to compute overlaps with many errors
Preliminary Rice Assemblies

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Contig NG50</th>
</tr>
</thead>
<tbody>
<tr>
<td>HiSeq Fragments</td>
<td>3,925</td>
</tr>
<tr>
<td>50x 2x100bp @ 180</td>
<td></td>
</tr>
<tr>
<td>MiSeq Fragments</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>“ALLPATHS-recipe”</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>PBeCR Reads</td>
<td>50,995</td>
</tr>
<tr>
<td>19x @ 3500 ** MiSeq for correction</td>
<td></td>
</tr>
<tr>
<td>Enhanced PBeCR</td>
<td>155,695</td>
</tr>
<tr>
<td>19x @ 3500 ** MiSeq for correction</td>
<td></td>
</tr>
</tbody>
</table>

In collaboration with McCombie & Ware labs @ CSHL
Early chemistries

<table>
<thead>
<tr>
<th>Year</th>
<th>Read Length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>453</td>
</tr>
<tr>
<td>2009</td>
<td>1012</td>
</tr>
<tr>
<td>2010</td>
<td>1734</td>
</tr>
</tbody>
</table>

Throughput: ~ 350-400 Mb

- P4–C2
- C2–C2
- ECR2
- FCR

P5–C3
- 8,500 bp

Read Length (bp)

2008 2009 2010 2011 2012 2013

Throughput: ~ 350-400 Mb
De novo assembly of Arabidopsis

A. thaliana Ler-0 sequenced at PacBio

- Sequenced using the latest 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 >100x

Genome size: 124.6 Mb
GC content: 33.92%
Raw data: 11 Gb
Assembly coverage: 15x over 9kbp

Sum of Contig Lengths: 149.5 Mb
Number of Contigs: 1788
Max Contig Length: 12.4 Mb
N50 Contig Length: 8.4 Mb
Assembly Complexity of Long Reads
Outline

1. Read length & assembly complexity
2. Single molecule assembly of rice
3. De novo indel mutations in autism
Variation Detection Complexity

SNPs + Short Indels
High precision and sensitivity

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

```
| AGAATAGGCGAG |
```

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

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

```
| TTTAGAATAGGC |
```

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

Sens: 48%
FDR: 38%
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

**SCALPEL:** Micro-assembly approach to accurately detect de novo and transmitted indel mutations within exome-Capture data

Scalpel Pipeline

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

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

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

Align assembled sequences to reference to detect mutations.
Repeats in the Exome

Specificity Challenge: 30% of exons have a perfect 10bp or larger repeat
Compute an on-the-fly analysis of repeat composition

![Graph showing the distribution of repeats in the exome. The x-axis represents K-mer size in base pairs (bp), and the y-axis represents the number of repeat sequences of a given length. The graph includes different colors for 0-mismatch, 1-mismatch, 2-mismatch, and 3-mismatch repeats, with exact repeats shown in blue.](chart.png)
Simulation Analysis

Simulated 10,000 indels in a exome from a known log-normal distribution
Experimental Analysis & Validation

Selected one deep coverage exome for deep analysis
- Individual was diagnosed with ADHD (See Gholson for details)
- 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 Indel Discovery
Scalpel Indel Discovery

SOAPindel: ABC'B'C'B'D
Scalpel: ABC'B'D
Scalpel Indel Discovery
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 reported strong enrichment for de novo gene killing mutations (nonsense, frameshift, splice site mutations)
- 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
Revised Analysis of the SSC

Constructed database of >1M transmitted and de novo indels
Strengthened enrichment for de novo frameshift mutations (35:16)
Many new gene candidates identified, population analysis underway
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
Schatz Lab
Giuseppe Narzisi
Shoshana Marcus
James Gurtowski
Srividya
Ramakrishnan
Hayan Lee
Rob Aboukhalil
Mitch Bekritsky
Charles Underwood
Tyler Gavin
Alejandro Wences
Greg Vurtle
Eric Biggers
Aspyn Palatnick

CSHL
Hannon Lab
Gingeras Lab
Jackson Lab
Iossifov Lab
Levy Lab
Lippman Lab
Lyon Lab
Martienissen Lab
McCombie Lab
Ware Lab
Wigler Lab

IT Department
See you at

Genome Informatics

Oct 30 – Nov 2

http://schatzlab.cshl.edu
@mike_schatz