Algorithms for studying the structure and function of genomes

Michael Schatz
Schatzlab Overview

**Human Genetics**
Role of mutations in disease
Narzisi et al. (2014)
lossifov et al. (2014)

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

**Algorithmics & Systems Research**
Ultra-large scale biocomputing
Blood et al. (2014)
Schatz et al. (2013)

**Single Cell & Single Molecule**
CNVs, SVs, & Cell Phylogenetics
Garvin et al. (2014)
Roberts et al. (2013)
Genome Structure & Function

1. **Structure: Sequencing and Assembly**
   *Long Read Single Molecule Sequencing*

2. **Function: Disease Analytics**
   *The role of indels in autism spectrum disorders*
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 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, ...

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

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

Original Fragment

```
It was the best of
```

Directed Edge

```
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
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.
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 ...
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 greater 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
• Better sequence for all downstream analysis
Assembly Applications

• Novel genomes

• Metagenomes

• Sequencing assays
  – Structural variations
  – Transcript assembly
  – ...

Like Dickens, we must computationally reconstruct a genome from short fragments
Genomics Across the Tree of Life
The map-based sequence of the rice genome

Table 2: Size of each chromosome based on sequence data and estimated gaps

<table>
<thead>
<tr>
<th>Chr</th>
<th>Sequenced bases (bp)</th>
<th>Gaps on arm regions</th>
<th>Telomeric gaps (Mb)</th>
<th>Centromeric gap (Mb)</th>
<th>rDNA (Mb)</th>
<th>Total (Mb)</th>
<th>Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>43,260,640</td>
<td>5</td>
<td>0.33</td>
<td>0.06</td>
<td>1.40</td>
<td>45.05</td>
<td>99.1</td>
</tr>
<tr>
<td>2</td>
<td>35,954,074</td>
<td>3</td>
<td>0.10</td>
<td>0.01</td>
<td>0.72</td>
<td>36.38</td>
<td>99.7</td>
</tr>
<tr>
<td>3</td>
<td>36,189,985</td>
<td>4</td>
<td>0.96</td>
<td>0.04</td>
<td>0.18</td>
<td>37.37</td>
<td>97.3</td>
</tr>
<tr>
<td>4</td>
<td>35,489,479</td>
<td>3</td>
<td>0.46</td>
<td>0.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>29,733,216</td>
<td>6</td>
<td>0.22</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>30,731,386</td>
<td>1</td>
<td>0.02</td>
<td>0.03</td>
<td>0.82</td>
<td>31.60</td>
<td>99.8</td>
</tr>
<tr>
<td>7</td>
<td>29,643,843</td>
<td>1</td>
<td>0.31</td>
<td>0.01</td>
<td>0.32</td>
<td>30.28</td>
<td>98.9</td>
</tr>
<tr>
<td>8</td>
<td>28,434,680</td>
<td>1</td>
<td>0.09</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>22,692,709</td>
<td>4</td>
<td>0.13</td>
<td>0.34</td>
<td>0.62</td>
<td>28.57</td>
<td>99.7</td>
</tr>
<tr>
<td>10</td>
<td>22,683,701</td>
<td>4</td>
<td>0.68</td>
<td>0.13</td>
<td>0.47</td>
<td>23.96</td>
<td>96.6</td>
</tr>
<tr>
<td>11</td>
<td>28,357,783</td>
<td>4</td>
<td>0.21</td>
<td>0.04</td>
<td>0.25</td>
<td>30.76</td>
<td>99.1</td>
</tr>
<tr>
<td>12</td>
<td>27,561,960</td>
<td>0</td>
<td>0.00</td>
<td>0.05</td>
<td>0.16</td>
<td>27.77</td>
<td>99.8</td>
</tr>
<tr>
<td>All</td>
<td>370,733,456</td>
<td>36</td>
<td>3.51</td>
<td>0.81</td>
<td>6.59</td>
<td>388.82</td>
<td>98.9</td>
</tr>
</tbody>
</table>

Contig N50: 5.1Mbp
Total projects costs: >$100M
Total costs: ~$10k
>1,000x times cheaper, but at what cost scientifically?

W.R. McCombie
Genomics Arsenal in the year 2015

Sample Preparation  Sequencing  Chromosome Mapping
Whole genome de novo assemblies of three divergent strains of rice (O. sativa) documents novel gene space of aus and indica
**Oryza sativa Gene Diversity**

- Very high quality representation of the “gene-space”
  - Overall identity ~99.9%
  - Less than 1% of exonic bases missing

- Genome-specific genes enriched for disease resistance
  - Reflects their geographic and environmental diversity

- Assemblies fragmented at (high copy) repeats
  - Difficult to identify full length gene models and regulatory features

**Overall sequence content**
In each sector, the top number is the total number of base pairs, the middle number is the number of exonic bases, and the bottom is the gene count. If a gene is partially shared, it is assigned to the sector with the most exonic bases.
Long Read Sequencing Technology

- **Moleculo**
- **PacBio RS II**
- **Oxford Nanopore**

(Voskoboynik et al. 2013)
PacBio SMRT Sequencing

Imaging of fluorescently phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).
Single Molecule Sequences
“Corrective Lens” for Sequencing
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)

**PacBio Coverage**
- < 5x
- > 50x
Consensus Accuracy and Coverage

Coverage can overcome random errors

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

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

*Nature Biotechnology.* 30:693–700
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 118x coverage using P5-C3 long read sequencing

Mean: 5918bp
49.7x over 10kbp
6.3x 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 kbp</td>
</tr>
<tr>
<td>25x 456bp</td>
<td></td>
</tr>
<tr>
<td>(3 runs 2x300 @ 450 FLASH)</td>
<td></td>
</tr>
<tr>
<td>“ALLPATHS-recipe”</td>
<td>18 kbp</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>HGAP + CA</td>
<td>4.0 Mbp</td>
</tr>
<tr>
<td>22.7x @ 10kbp</td>
<td></td>
</tr>
<tr>
<td>Nipponbare</td>
<td>5.1 Mbp</td>
</tr>
<tr>
<td>BAC-by-BAC Assembly</td>
<td></td>
</tr>
</tbody>
</table>

**HGAP Read Lengths**  
Max: 53,652bp  
22.7x over 10kbp  
*(discarded reads below 8500bp)*
S5 Hybrid Sterility Locus

**S5 is a major locus for hybrid sterility in rice that affects embryo sac fertility.**

- Genetic analysis of the S5 locus documented three alleles: an indica (S5-i), a japonica (S5-j), and a neutral allele (S5-n)
- Hybrids of genotype S5-i/S5-j are mostly sterile, whereas hybrids of genotypes consisting of S5-n with either S5-i or S5-j are mostly fertile.
- Contains three tightly linked genes that work together in a ‘killer-protector’-type system: ORF3, ORF4, ORF5
- The ORF5 indica (ORF5+) and japonica (ORF5-) alleles differ by only **two nucleotides**
S5 Hybrid Sterility Locus

Sanger     ...ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC...
Illumina   ...ACCCTGATATTCTGAGTTACAAGGCATTCACTAGCTACTGCTTGCCCACTGACGAGACC...
PacBio     ...ACCCTGATATTCTGAGTTACAAGGCATTCACTAGCTACTGCTTGCCCACTGACGAGACC...

100kbp
S5 Hybrid Sterility Locus

Sanger  ...ACCCTGATATTCTGAGTTACAAGGCATT\textcolor{red}{CAGCTACTGCTTGCCCACTGACGAGACC}...
Illumina  ...ACCCTGATATTCTGAGTTACAAGGCATT\textcolor{red}{CAGCTACTGCTTGCCCACTGACGAGACC}...
PacBio  ...ACCCTGATATTCTGAGTTACAAGGCATT\textcolor{red}{CAGCTACTGCTTGCCCACTGACGAGACC}...
Sanger

Illumina

Pacbio

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

S5 Hybrid Sterility Locus
**Improvements from 20kbp to 4Mbp contig N50:**

- Over 20 Megabases of additional sequence
  - Extremely high sequence identity (>99.9%)
  - Thousands of gaps filled, hundreds of mis-assemblies corrected
- Complete gene models, promoter regions for nearly every gene
  - True representation of transposons and other complex features
- Opportunities for studying large scale chromosome evolution
  - Largest contigs approach complete chromosome arms
Current Collaborations

Pineapple
UIUC

Human
CSHL/OICR

Asian Sea Bass
Temasek Life Sciences

C. glabrata
JHU

T. vaginalis
NYU
Long Read Sequencing of SK-BR-3

**Long read PacBio sequencing of SK-BR-3 breast cancer cell line**
- Her2+ breast cancer is one of the most deadly forms of the disease
  - SK-BR-3 is one of the most important models, known to have widespread CNVs
- Currently have 60x coverage with long read PacBio sequencing (mean: ~10kbp)
  - Discovered a complex series of nested duplications and translocations around HER2
  - Currently analyzing breakpoints in an attempt to infer the mutation history

In collaboration with McCombie (CSHL) and McPherson (OICR) labs

(Wen-Sheng et al, 2009)
(Navin et al, 2011)
PacBio® Advances in Read Length

- Early PacBio chemistries
  - 380
  - 1012-1734
  - LPR
  - K6A
  - K6A2
- P6-C1
- P6-C2
- P6-C4

Read Length (bp)

Year

Error correction and assembly complexity of single molecule sequencing reads.
Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz, MC
http://www.biorxiv.org/content/early/2014/06/18/006395
Tomorrow at Noon

Oxford Nanopore Sequencing

Pan-Genomics

Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome

SplitMEM: A graphical algorithm for pan-genome analysis with suffix skips
1. **Structure: Sequencing and Assembly**
   *Long Read Single Molecule Sequencing*

2. **Function: Disease Analytics**
   *The role of indels in autism spectrum disorders*
Genetic Basis of Autism Spectrum Disorders

**Complex disorders of brain development**
- Characterized by difficulties in social interaction, verbal and nonverbal communication and repetitive behaviors.
- Have their roots in very early brain development, and the most obvious signs of autism and symptoms of autism tend to emerge between 2 and 3 years of age.

**U.S. CDC identify around 1 in 68 American children as on the autism spectrum**
- Ten-fold increase in prevalence in 40 years, only partly explained by improved diagnosis and awareness.
- Studies also show that autism is four to five times more common among boys than girls.
- Specific causes remain elusive

**What is Autism?**
http://www.autismspeaks.org/what-autism
Searching for the genetics behind human disorders and plant phenotypes

**Search Strategy**

- Currently uses WGS or WES short read resequencing for economic reasons

- Collaborate with Lyon, McCombie, Tuveson, and Wigler labs to examine the genetic basis of cancer, ASD, and other psychiatric disorders

- Also collaborating with the Lippman, Ware, and Gingeras labs to study high value crops

*Are there any genetic variants present in affected individuals, that are not present or are present at a substantially reduced rate in their relatives?*
Exome sequencing of the SSC

The year 2012 was an exciting year for autism genetics

- 3 reports of ~600 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
- All used poor or no tools for indels:
  - Iossifov (343 families) and O’Roak (50 families) used GATK UnifiedGenotype
  - Sanders (200 families) didn’t attempt

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
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 de novo and transmitted indel detection in exome-capture data using microassembly.**
Scalpel Algorithm

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
Experimental Analysis & Validation

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)
Refined indel analysis

Examine sources of indel errors

- Experimental validation of indels called from 30x whole genome vs. 110x whole exome
- Most of the errors due to microsatellite slippage introduced during exome capture, also missing most long indels
- Recommend PCR-free WGS if at all possible

<table>
<thead>
<tr>
<th></th>
<th>All INDELs</th>
<th>Valid</th>
<th>PPV</th>
<th>INDELs &gt;5bp</th>
<th>Valid (&gt;5bp)</th>
<th>PPV (&gt;5bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersection</td>
<td>160</td>
<td>152</td>
<td>95.0%</td>
<td>18</td>
<td>18</td>
<td>100%</td>
</tr>
<tr>
<td>WGS</td>
<td>145</td>
<td>122</td>
<td>84.1%</td>
<td>33</td>
<td>25</td>
<td>75.8%</td>
</tr>
<tr>
<td>WES</td>
<td>161</td>
<td>91</td>
<td>56.5%</td>
<td>1</td>
<td>1</td>
<td>100%</td>
</tr>
</tbody>
</table>

Reducing INDEL calling errors in whole-genome and exome sequencing data
De novo Genetics of Autism

• In 2,500 family quads we see significant enrichment in de novo likely gene disruptions (LGDs) in the autistic kids
  – Overall rate basically 1:1
  – 2:1 enrichment in frameshift indels
  – Confirmed trends observed in previous studies, contributed dozens of new autism candidate genes.

The burden of de novo coding mutations in autism spectrum disorders.
What’s next?

Giuseppe Narzisis
Somatic mutation detection
Coding and non-coding mutations in cancer and autism

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

Maria Nattestad
Hi-C Chromatin Interactions
Plant Assembly & Analysis

Tyler Garvin
Single Cell CNV
Tumor and Somatic Heterogeneity
Understanding Genome Structure & Function

Reference quality genome assembly is here
- Use the longest possible reads for the analysis
- Don’t fear the error rate
  • Coverage and algorithmics conquer random errors

Population analysis
- Large scale sequencing give us new insights into the origins of disease, the processes of development, and the forces of evolution
- See similar trends in the population analysis of many cells, integration of multiple assays

Also very interested in teaching the next generation of undergraduate and graduate students
## Acknowledgements

<table>
<thead>
<tr>
<th>Schatz Lab</th>
<th>CSHL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rahul Amin</td>
<td>Hannon Lab</td>
</tr>
<tr>
<td>Eric Biggers</td>
<td>Gingeras Lab</td>
</tr>
<tr>
<td>Han Fang</td>
<td>Jackson Lab</td>
</tr>
<tr>
<td>Tyler Gavin</td>
<td>Hicks Lab</td>
</tr>
<tr>
<td>James Gurtowski</td>
<td>Iossifov Lab</td>
</tr>
<tr>
<td>Ke Jiang</td>
<td>Levy Lab</td>
</tr>
<tr>
<td>Hayan Lee</td>
<td>Lippman Lab</td>
</tr>
<tr>
<td>Zak Lemmon</td>
<td>Lyon Lab</td>
</tr>
<tr>
<td>Shoshana Marcus</td>
<td>Martienssen Lab</td>
</tr>
<tr>
<td>Giuseppe Narzisi</td>
<td>McCombie Lab</td>
</tr>
<tr>
<td>Maria Nattestad</td>
<td>Tuveson Lab</td>
</tr>
<tr>
<td>Aspyn Palatnick</td>
<td>Ware Lab</td>
</tr>
<tr>
<td>Srividya Ramakrishnan</td>
<td>Wigler Lab</td>
</tr>
<tr>
<td>Rachel Sherman</td>
<td>IT &amp; Meetings Depts.</td>
</tr>
<tr>
<td>Greg Vurture</td>
<td>Pacific Biosciences</td>
</tr>
<tr>
<td>Alejandro Wences</td>
<td>Oxford Nanopore</td>
</tr>
</tbody>
</table>

**Logos:**
- National Human Genome Research Institute
- U.S. Department of Energy
- SFARI (SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE)
Thank you

http://schatzlab.cshl.edu

@mike_schatz