In pursuit of perfect genome sequencing

Michael Schatz

May 22, 2017
World Metrology Day @ JIMB
Schatzlab Overview

**Human Genetics**
Role of mutations in disease
- Feigin et al. (2017)
- Narzisi et al. (2015)

**Agricultural Genomics**
Genomes & Transcriptomes
- Lemmon et al. (2016)
- Ming et al. (2015)

**Algorithmics & Systems Research**
Ultra-large scale biocomputing
- Stevens et al. (2015)
- Marcus et al. (2014)

**Biotechnology Development**
Single Cell & Single Molecule Sequencing
- Chin et al. (2016)
- Garvin et al. (2015)
In pursuit of perfect genome sequencing

1. Why “Perfect”? 

2. What is “Perfect”? 

3. How will we achieve it? 

4. When will we achieve it?
In pursuit of perfect genome sequencing

1. Why “Perfect”? 
2. What is “Perfect”? 
3. How will we achieve it? 
4. When will we achieve it?
Genetic Origins of Human Diversity

GWAS Catalog contains 33,674 unique SNP-trait associations. OMIM contains records for more than 5000 traits with known molecular basis.

http://www.ebi.ac.uk/gwas/diagram
Somatic Mutations In Cancer

Signatures of mutational processes in human cancer
Mammalian Evolution

Digits and fin rays share common developmental histories
“Needles in a stack of needles”

Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data
In pursuit of perfect genome sequencing

1. **Why “Perfect”?**
   
   *Because it is important, complex, and diffuse*

2. **What is “Perfect”?**

3. **How will we achieve it?**

4. **When will we achieve it?**
In pursuit of perfect genome sequencing

1. Why “Perfect”? 
2. What is “Perfect”? 
3. How will we achieve it? 
4. When will we achieve it?
I. Correctness:
Is the genome faithfully represented?
I. Correctness:
Is the genome faithfully represented?

Sample of 100k reads aligned with BLASR requiring >100bp alignment
Average overall accuracy: 83.7%, 11.5% insertions, 3.4% deletions, 1.4% mismatch
If there were no sequencing errors, identifying SNPs would be trivial:
  - Any time a read disagrees with the reference, it must be a variant!

A single read of many differing from the reference is probably just an error, but it becomes more likely to be real as we see it multiple times
  - Use binomial test to evaluate prob. of heterozygosity vs. prob of error
  - Coverage (oversampling) is our main tool to improve accuracy
Consensus Accuracy and Coverage

Coverage can overcome random errors

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

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

Hybrid error correction and de novo assembly of single-molecule sequencing reads.
"The overall base-to-base concordance rate is about 99.99% (QV40 in Phred scale) in the F1 FALCON-Unzip assembly. The insertion and deletion (indel) concordances to the parental lines were lower (about QV40) than the SNP concordance rate (about QV50), with most residual errors concentrated in long homopolymer sequences"
2. Completeness:
How much of the genome is present?
2. Completeness: How much of the genome is present?

“88% of GWAS SNPs are intronic or intergenic of unknown function”
ENCODE Consortium (2012)
LETTER

Resolving the complexity of the human genome using single-molecule sequencing

Mark J. P. Chaisson¹, John Huddleston¹⁻², Megan Y. Dennis³, Peter H. Sudmant⁴, Maika Malig⁵, Fereydoun Hormozdiari¹, Francesca Antonacci³, Urvashi Surti³, Richard Sandstrom¹, Matthew Boitano⁵, Jane M. Landolin⁵, John A. Stamatoyannopoulos¹, Michael W. Hunkapiller³, Jonas Korlach⁵ & Evan E. Eichler¹,²

The human genome is arguably the most complete mammalian reference assemblies¹⁻³, yet more than 160 euchromatic gaps remain⁴⁻⁵ and aspects of its structural variation remain poorly understood ten years after its completion⁷⁻⁹. To identify missing sequence and genetic variation, here we sequence and analyse a haploid human genome (CHM1) using single-molecule, real-time DNA sequencingⁱ⁰. We close or extend 55% of the remaining interstitial gaps in the human GRCh37 reference genome—78% of which carried long runs of degenerate short tandem repeats, often several kilobases in length, embedded within (G+C)-rich genomic regions. We resolve the complete sequence of 26,079 euchromatic structural variants at the base-pair level, including inversions, complex insertions and long tracts of tandem repeats.

Most have not been previously reported, with the greatest increases in sensitivity occurring for events less than 5 kilobases in size. Compared to the human reference, we find a significant insertional bias (3:1) in regions corresponding to complex insertions and long short tandem repeats. Our results suggest a greater complexity of the human genome in the form of variation of longer and more complex repetitive DNA that can now be largely resolved with the application of this longer-read sequencing technology.
3. Contiguity
How much context is available?
3. Contiguity

How much context is available?

*If you have 99% completeness, are you missing 1% of every gene or are the missing sequences localized to certain regions?*

*How far can you go until you hit a gap in resolution?*
Assembly Complexity

A R B R C R D

A → B → R → C → D

A → R → C → D → A
Assembly Complexity

Short contigs & Incomplete genes

Gap in coverage

Variants hidden in repeats

Unresolved Heterozygosity
Assembly Complexity
The advantages of SMRT sequencing
Recent Long Read Assemblies

**Human Analysis N50 Sizes**

<table>
<thead>
<tr>
<th>Technology</th>
<th>N50 Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina Discovar (contig asm)</td>
<td>0</td>
</tr>
<tr>
<td>Moleculo Prism (phasing)</td>
<td>5,000,000</td>
</tr>
<tr>
<td>10X GemCode Long Ranger (phasing)</td>
<td>10,000,000</td>
</tr>
<tr>
<td>PacBio FALCON (contig asm)</td>
<td>15,000,000</td>
</tr>
</tbody>
</table>

**Structural Variants in CHM1**

- Deletions
- Insertions
- Repeat Contraction
- Repeat Expansion
- Tandem Contraction
- Tandem Expansion

---

*Third-generation sequencing and the future of genomics*

Lee et al (2016) *bioRxiv*

doi: http://dx.doi.org/10.1101/048603

*Assemblytics: a web analytics tool for the detection of variants from an assembly*

Nattestad & Schatz (2016) *Bioinformatics.*

doi: 10.1093/bioinformatics/btw369
In pursuit of perfect genome sequencing

1. Why “Perfect”? 

2. What is “Perfect”? 
   
   100% correct, complete, & contiguous 

3. How will we achieve it? 

4. When will we achieve it?
In pursuit of perfect genome sequencing

1. Why “Perfect”? 

2. What is “Perfect”? 

3. How will we achieve it? 

4. When will we achieve it?
Genomics Arsenal in the year 2017

- Sample Preparation
- Sequencing
- Chromosome Mapping
Genomic Sequencing Data

Illumina

60x Paired End
All 4 samples

µ=350bp

Fragment Length (kbp)

10X Genomics

35x Linked Reads
All 4 samples

µ=117kbp

Molecule Length (kbp)

PacBio

55x Long Reads
*Only ENC-002

µ=7.5kbp

Read Length (kbp)
Assembly Contiguity

GRC38 Reference
- Includes alt sequences

10X Genomics/SuperNova
- 21 Mbp scaffold N50
- 162 Mbp in scaffold gaps

PacBio/Falcon-unzip
- 7.0 Mbp contig N50

10X Genomics/SuperNova
- 50 kbp contig N50

Illumina/MegaHit
- 13 kbp contig N50
Missing Insertions from Short and Linked Read?

Illumina

10X Genomics

PacBio

Variants 50 to 500 bp

Count

Variant size

Count

Variant size

Count

Variant size
Structural Variations Concordance

Sniffles: 17,139
Falcon: 7,857, 12,241
LongRanger: 2,823, 1,946, 3,785
SuperNova: 3,394, 2,837, 1,486, 18,862
SURVIVOR2: 3,291, 2,163, 2,274, 1,646, 6,631
MegaHit: 1,851, 1,529, 569, 1,378, 687, 3,855
Structural Variations Concordance

### Main Diagonal
- Calls per tool

<table>
<thead>
<tr>
<th>Tool</th>
<th>Sniffles</th>
<th>Falcon</th>
<th>LongRanger</th>
<th>SuperNova</th>
<th>SURVIVOR2</th>
<th>MegaHit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sniffles</td>
<td>17,139</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Falcon</td>
<td></td>
<td>12,241</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LongRanger</td>
<td>2,823</td>
<td></td>
<td>3,785</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SuperNova</td>
<td>3,394</td>
<td>2,837</td>
<td>1,486</td>
<td>18,862</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SURVIVOR2</td>
<td>3,291</td>
<td>2,163</td>
<td>2,274</td>
<td>1,646</td>
<td>6,631</td>
<td></td>
</tr>
<tr>
<td>MegaHit</td>
<td>1,858</td>
<td>1,529</td>
<td>569</td>
<td>1,378</td>
<td>687</td>
<td>3,855</td>
</tr>
</tbody>
</table>
Structural Variations Concordance

Main Diagonal
• Calls per tool

Outer triplets
• Concordance by Technology

<table>
<thead>
<tr>
<th>Tool</th>
<th>Sniffles</th>
<th>Falcon</th>
<th>LongRanger</th>
<th>SuperNova</th>
<th>SURVIVOR2</th>
<th>MegaHit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sniffles</td>
<td>17,139</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Falcon</td>
<td>7,857</td>
<td>12,241</td>
<td>3,785</td>
<td>18,862</td>
<td>6,631</td>
<td></td>
</tr>
<tr>
<td>LongRanger</td>
<td>2,823</td>
<td>1,946</td>
<td>3,785</td>
<td>2,837</td>
<td>6,631</td>
<td>3,855</td>
</tr>
<tr>
<td>SuperNova</td>
<td>3,394</td>
<td>2,837</td>
<td>1,486</td>
<td>18,862</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SURVIVOR2</td>
<td>3,291</td>
<td>2,163</td>
<td>2,274</td>
<td>1,646</td>
<td>1,378</td>
<td>687</td>
</tr>
<tr>
<td>MegaHit</td>
<td>1,858</td>
<td>1,529</td>
<td>569</td>
<td>1,378</td>
<td>687</td>
<td>3,855</td>
</tr>
</tbody>
</table>

- PacBio
- 10X Genomics
- Illumina
Structural Variations Concordance

Main Diagonal
- Calls per tool

Outer triplets
- Concordance by Technology

Inner triplets
- Concordance by Assembly

<table>
<thead>
<tr>
<th>Tool</th>
<th>Sniffles</th>
<th>Falcon</th>
<th>LongRanger</th>
<th>SuperNova</th>
<th>SURVIVOR2</th>
<th>MegaHit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sniffles</td>
<td>17,139</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Falcon</td>
<td>7,857</td>
<td>12,241</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LongRanger</td>
<td>2,823</td>
<td>1,946</td>
<td>3,785</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SuperNova</td>
<td>3,394</td>
<td>2,837</td>
<td>1,486</td>
<td>18,862</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SURVIVOR2</td>
<td>3,291</td>
<td>2,163</td>
<td>2,274</td>
<td>1,646</td>
<td>6,631</td>
<td></td>
</tr>
<tr>
<td>MegaHit</td>
<td>1,858</td>
<td>1,529</td>
<td>569</td>
<td>1,378</td>
<td>687</td>
<td>3,855</td>
</tr>
</tbody>
</table>

10X Genomics

PacBio

Illumina
Structural Variations Concordance

- **Main Diagonal**
  - Calls per tool

- **Outer triplets**
  - Concordance by Technology

- **Inner triplets**
  - Concordance by Assembly
  - Concordance by Mappers

<table>
<thead>
<tr>
<th>Tool</th>
<th>Sniffies</th>
<th>Falcon</th>
<th>LongRanger</th>
<th>SuperNova</th>
<th>SURVIVOR2</th>
<th>MegaHit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sniffies</td>
<td>17,139</td>
<td>7,857</td>
<td>2,823</td>
<td>3,394</td>
<td>3,291</td>
<td>1,858</td>
</tr>
<tr>
<td>Falcon</td>
<td>12,241</td>
<td>1,946</td>
<td>1,237</td>
<td>1,486</td>
<td>2,163</td>
<td>1,529</td>
</tr>
<tr>
<td>LongRanger</td>
<td>3,785</td>
<td>1,466</td>
<td>1,646</td>
<td>2,837</td>
<td>2,274</td>
<td>569</td>
</tr>
<tr>
<td>SuperNova</td>
<td>18,862</td>
<td>3,785</td>
<td>1,946</td>
<td>1,237</td>
<td>1,646</td>
<td>1,378</td>
</tr>
<tr>
<td>SURVIVOR2</td>
<td>6,631</td>
<td>1,646</td>
<td>2,837</td>
<td>1,466</td>
<td>3,291</td>
<td>687</td>
</tr>
<tr>
<td>MegaHit</td>
<td>3,855</td>
<td>2,274</td>
<td>1,237</td>
<td>1,378</td>
<td>1,858</td>
<td>1,529</td>
</tr>
</tbody>
</table>

- **10X Genomics**
- **Illumina**
Structural Variations Concordance

Main Diagonal
- Calls per tool

Outer triplets
- Concordance by Technology

Inner triplets
- Concordance by Assembly
- Concordance by Mappers

Overall:
- We need multiple technologies and approaches
In pursuit of perfect genome sequencing

1. Why “Perfect”? 

2. What is “Perfect”? 

3. How will we achieve it?
   *Combinations of technologies*

4. When will we achieve it?
In pursuit of perfect genome sequencing

1. Why “Perfect”?  
2. What is “Perfect”?  
3. How will we achieve it?  
4. When will we achieve it?
Consensus Accuracy and Coverage

Coverage can overcome **random** errors

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

\[
\text{CNS Error} = \sum_{i=0}^{c} \binom{c}{i} e^i (1-e)^{n-i}
\]

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

**Illumina Novaseq**
- $850k instrument cost
- $1k / human @ 50x
- Short reads, high throughput

**10X Chromium**
- $125k instrument costs
- $2k / human
- Linked reads, medium throughput
PacBio Roadmap

**PacBio Sequel**

- $350k instrument cost
- ~$30k / human @ 50x
- Long reads, Medium throughput

**SMRTcell v2**

- 1M Zero Mode Waveguides
- ~15kb average read length
- ~$1000 / SMRTcell
Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome
In pursuit of perfect genome sequencing

- **Three C’s of Genome Quality: Correctness, Completeness & Contiguity**
  - Very excited for combinations of long reads + Hi-C based scaffolding
  - Expect new insights on the causes of diseases, forces of evolution

- **Multiple sequencing technologies & approaches needed**
  - *PacBio*: Best Resolution of SVs
  - *10X/HIC*: Best Phasing
  - *De novo*: Best Resolution of small SVs
  - *Mapping*: Best resolution of large SVs

- **We have just begun to explore the universe of variants present**
  - Tens of thousands of SVs per person, many megabases of variation
  - Also need to push these ideas into single cell and population scale analysis
Acknowledgements

**Schatz Lab**
Charlotte Darby
Han Fang
Tyler Gavin
James Gurtowski
Sam Kovaka
Laurent Luo
Maria Nattestad
Srividya
Ramakrishnan
Fritz Sedlazeck

**CSHL**
Gingeras Lab
Jackson Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Tuveson Lab
Ware Lab
Wigler Lab

**JHU**
Langmead Lab
Salzberg Lab
Timp Lab
Wheelan Lab

**Cornell**
Susan McCouch
Lyza Maron
Mark Wright

**SBU**
Skiena Lab
Patro Lab

**OICR**
John McPherson
Karen Ng
Timothy Beck
Yogi Sundaravadanam

**NYU**
Jane Carlton
Elodie Ghedin
Thank you

http://schatz-lab.org

@mike_schatz

Now recruiting postdocs!