In pursuit of perfect genome sequencing
Michael Schatz

June 28, 2017
PacBio Users Meeting
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. However, most traits remain only partially explained or not at all

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

Signatures of mutational processes in human cancer
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?
1. Correctness:
Is the genome faithfully represented?
1. 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
Genotyping Theory

• 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

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?
“88% of GWAS SNPs are intronic or intergenic of unknown function”
ENCODE Consortium (2012)
Resolving the complexity of the human genome using single-molecule sequencing

Mark J. P. Chaisson¹, John Huddleston¹⁻², Megan Y. Dennis¹, Peter H. Sudmant¹, Maika Malig¹, Fereydoun Hormozdari¹, Francesca Antonacci³, Urvashi Surti¹, Richard Sandstrom¹, Matthew Boitano¹, Jane M. Landolin⁵, John A. Stamatoyannopoulos¹, Michael W. Hunkapiller⁵, Jonas Korlach⁵ & Evan E. Eichler¹⁻²

The human genome is arguably the most complete mammalian reference assembly¹⁻³, yet more than 160 euchromatic gaps remain⁴⁻⁶ 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 sequencing⁸. 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.

for recruiting additional sequence reads for assembly (Supplementary Information). Using this approach, we closed 50 gaps and extended into 40 others (60 boundaries), adding 398 kb and 721 kb of novel sequence to the genome, respectively (Supplementary Table 4). The closed gaps in the human genome were enriched for simple repeats, long tandem repeats, and high (G+C) content (Fig. 1) but also included novel exons (Supplementary Table 20) and putative regulatory sequences based on DNase I hypersensitivity and chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq) analysis (Supplementary Information). We identified a significant 15-fold enrichment of short tandem repeats (STRs) when compared to a random sample ($P < 0.00001$) (Fig. 1a). A total of 78% (39 out of 50) of the closed gap sequences were composed of 10% or more of STRs. The STRs were frequently embedded in longer, more complex, tandem arrays of degenerate repeats reaching up to 8,000 bp in length (Extended Data Fig. 1a–c), some of which bore resemblance to sequences known to be toxic to Escherichia coli¹⁰. Because most human reference sequences¹¹,¹² have been derived from clones propagated in E. coli, it is perhaps not surprising that the application of a long-read sequence technology to uncloned DNA would resolve such gaps. Moreover, the length and complex degeneracy of these
Long-read genome sequencing identifies causal structural variation in a Mendelian disease

Accurate detection of complex structural variations using single molecule sequencing
NGMLR + Sniffles

Accurate detection of complex structural variations using single molecule sequencing
Accurate detection of complex structural variations using single molecule sequencing
No more false positives!

Accurate detection of complex structural variations using single molecule sequencing
No more false positives!

Accurate detection of complex structural variations using single molecule sequencing
No more false positives!

Accurate detection of complex structural variations using single molecule sequencing
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
Assembly Complexity

Short contigs & Incomplete genes

Variants hidden in repeats

Gap in coverage

Unresolved Heterozygosity
Assembly Complexity

The diagram illustrates the assembly complexity of different components, labeled A, B, C, and D. The complexity is represented by the number of connections and interactions among these components. The diagram shows a network of interactions, indicating the complexity of assembling these parts.
Assembly Complexity

The advantages of SMRT sequencing
(A few) Recent PacBio Assemblies

- 7.0 Mbp
- 4.0 Mbp
- 4.5 Mbp
- 4.6 Mbp
- 1.4 Mbp

#1mbctgclub
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

Fragment Length (kbp)

60x Paired End
All 4 samples

μ = 350bp

10X Genomics

Molecule Length (kbp)

35x Linked Reads
All 4 samples

μ = 117kbp

PacBio

Read Length (kbp)

55x Long Reads
*Only ENC-002

μ = 7.5kbp
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
Structural Variations Concordance

Main Diagonal
- Calls per tool

Outer triplets
- Concordance by Technology

Inner triplets
- Concordance by Assembly
- Concordance by Mappers

Overall:
- Longnnnnng reads give the best concordance 😊
In pursuit of perfect genome sequencing

1. Why “Perfect”?

2. What is “Perfect”?

3. How will we achieve it?
   Lonnnnnng reads :-)  

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

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

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

**PacBio Sequel**

- $350k instrument cost
- $30k/human @ 50x

**SMRTcell v2**

- 1M Zero Mode Waveguides
- ~15kb average read length
PacBio Roadmap

Maize Sequel Sequencing
56x coverage
10,935 bp N50 read length
In pursuit of perfect genome sequencing

- **Three C’s of Genome Quality: Correctness, Completeness & Contiguity**
  - The key for perfect genomes is lonnnnnnnng reads 😊
  - 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

---

http://schatz-lab.org

---

**Tools**

- FALCON
- LRSim
- SURVIVOR
- NGM+Sniffles
- Ribbon
- Assemblytics
Thank you!
@mike_schatz

Looking for a postdoc?
http://schatz-lab.org/apply/