Specifics of SMRT® Sequencing Data

FIND MEANING IN COMPLEXITY
PacBio technology overview

Video: http://www.pacb.com/smrt-science/smrt-sequencing/
Genetic variation and the de novo assembly of human genomes

Mark J. P. Chaisson1, Richard K. Wilson2 and Evan E. Eichler1,2

Abstract | The discovery of genetic variation and the assembly of genome sequences are both inextricably linked to advances in DNA-sequencing technology. Short-read massively parallel sequencing has revolutionized our ability to rapidly generate high-quality genome assemblies. Understanding of genetic variation is only guaranteed by complete and accurate genome assemblies, and we describe advances in both the current approaches to genome assembly and the nature of the sequence data used in the assembly process. We discuss the challenges in de novo assembly using current technologies and the potential for genome-wide studies, and we summarize recent technological advances that will enable the future understanding of genetic variation and its primary means to understanding the full range of human variation.

Contemporary genetic studies fundamentally require, for comparing the sequences of individual genomes. The dominant method for this comparison is resequencing, in which random fragments of a genome are obtained, and compared to a reference sequence. Such experiments are conducted using instruments for massively parallel sequencing (MPS), in which billions of 100–200 nucleotide sequences may be read by a single instrument in a few days. Although great advances have been made in our knowledge of diversity, cancer and genetic disease, the genetic information provided by resequencing with current technology is incomplete. There is a lack of sensitivity for detecting small insertions and deletions (indels) and structural variation74,9, there is coverage bias against particularly GC- and AT-rich DNA5, the phase of mutations over long ranges must be inferred or imputed as opposed to directly observed, and the architecture of large polymorphic copy number variations is incomplete59.

An alternative to resequencing is de novo assembly, in which the entire sequence of two haplotypes is resolved from sequence data without comparison to a reference genome sequence. Although de novo assembly is, in principle, complete and therefore the ideal for genetic variation discovery, it is still currently impossible to achieve with data generated by typical MPS resequencing projects75. There is evidence that the landscape of sequencing technology is changing in such a way that will ultimately enable more routine de novo assembly of genomes.
READ LENGTH ENABLES UNIQUE APPLICATIONS

**De novo**
- Whole genome assembly
- Transcriptome

**Why length helps**
- Span repeats
- End-to-end isoform sequencing

**Reference-based**
- Consistent coverage
- Structural variant discovery
- Haplotype phasing

- High mappability / span repeats
- Encompass variants
- Connect heterozygous sites
Highly Accurate Results

SMRT® Sequencing can achieve greater than 99.999% (QV 50) accurate sequencing results for resequencing and *de novo* applications:

1. Near perfect consensus accuracy
2. Little or no sequence context bias
3. Unambiguous mappability of sequence reads
2nd Gen Sequencing

i. Generate sequence read:

```
GTCCTGAGACACGACAGCCGACCTCTGACCGGACTCTGTTCCGGTCTTTGGACAAATCGGGATTCTGAAGCTTCCGGGGATGCTGGAGATGATAG
```

ii. Map to reference:

```
GTCCTGAGACACGACAGCCGACCTCTGACCGGACTCTGTTCCGGTCTTTGGACAAATCGGGATTCTGAAGCTTCCGGGGATGCTGGAGATGATAG
```

iii. Generate consensus (10x coverage):

```
GTCCTGAGACACGACAGCCGACCTCTGACCGGACTCTGTTCCGGTCTTTGGACAAATCGGGATTCTGAAGCTTCCGGGGATGCTGGAGATGATAG
```

Reference match

Heterozygous SNP

Reference

Homozygous SNP

Reference match

Heterozygous SNP

Reference

Homozygous SNP
SMRT® Sequencing – Errors are *Random*

i. Generate sequence read:

ii. Map to reference: [BLASR: BMC Bioinformatics 13: 238.](#)

iii. Generate consensus (10x coverage): [Quiver: SMRT Analysis](#)

Reference match  | Heterozygous SNP  | Reference  | Homozygous SNP
--- | --- | --- | ---

---

[PB PACIFIC BIOSCIENCES](#)
SMRT® Sequencing Accuracy

Concordance - Accuracy (QV)

- 99.99999% (QV 70)
- 99.9999% (QV 60)
- 99.999% (QV 50)
- 99.99% (QV 40)
- 99.9% (QV 30)
- 99% (QV 20)
- 90% (QV 10)

Coverage

Data generated with P4-C2 chemistry on PacBio® RS II; Analyzed using Quiver with 2.0.1 SMRT® Analysis
De Novo Assembly

User Group Meeting – Menlo Park, October 2014

FIND MEANING IN COMPLEXITY
Improve and Finish Genomes with the PacBio® System

De novo Assembly
Complete genomes with PacBio reads alone
Combine technologies for best of both worlds

Scaffold
Establish framework for genome and resolve ambiguities

Span Gaps
Polish genomic regions with up to 10x improvement
Subreads (purple and gold) are separated by adapter sequences (green).
Read of Insert represents the highest quality single-sequence for an insert, regardless of number of passes.
≥ 2 full passes required for CCS.
Both adapters must be detected for a read to be identified as “full pass”.
Either individual subreads, read of insert or CCS can be used for subsequent analysis depending on application needs.
Read Metrics Definitions

SMRTbell™ Template

Polymerase Read
Definition:
- Sequence of nucleotides incorporated by polymerase while reading a template
- Includes adapters
- Often called “read”
- Includes adapters
- 1 molecule, 1 pol. read

Purpose:
- QC of instrument run
- Benchmarking

Subread
Definition:
- Single pass of template
- Adapters removed
- 1 molecule, ≥1 subreads

Unique data:
- Kinetic measurements
- Rich QVs

Purpose:
- For subsequent analysis

Read of Insert
Definition:
- Represents highest-quality single-sequence for an insert, regardless of number of passes
- Generalizes CCS for <2 passes and RQ <0.9
- 1 or more passes
- 1 molecule, 1 read

Purpose:
- For Library QC
- For subsequent analysis
Basic Assembly Metrics

• Commonly used metrics include:
  - Number of contigs
  - N50: Equal to the size of the contig found if you sort contigs by size and walk to the contig that represents 50% of the total sequence
    - N50 = 10 bp
    - Mean contig length = 3 bp
  - Max contig size

• Limitation of these metrics:
  - They do not capture information about assembly accuracy!
    - Large scale misassemblies
    - Base-level errors
  - There might be more than one chromosome (plasmid, phage, and so on)
  - Contaminants may contribute to a contig (such as a cloning vector)
Scaffolds vs. Contigs Defined

- Scaffolds have Ns in them, due to links from mate pair data.
- Contigs are contiguous sequences (no Ns). PacBio® sequencing generates contigs given our continuous reads.
Hierarchical Genome Assembly Process
Hierarchical Genome Assembly Process

Short Continuous Long Reads

Pre-assembled Long Read

Construct Pre-assembled Long Reads (PLR) from CLRs

PLRs

Contig

Assemble PLRs into contigs
Pre-Assembly of Single-Pass Long Reads

Single-pass long reads

Select longest as seed reads

Pre-assembled reads

RS_PreAssembler

Map all to seed reads

Generate consensus of mapped reads
Assembly of Pre-Assembled Reads into Contigs

- Pre-assembled reads
- Identify overlaps between reads
- Unitigs
- Generate layout of overlapping reads
- Contigs of assembly
- Generate consensus
Assembly Polishing via Quiver

Single-pass long reads

Contigs

High-quality consensus

RS_Resequencing

Map to *de novo*-assembled reference

Base-quality-aware consensus of uniquely mapped reads

Quiver
Experimental Design – Ploidy

• Most Assemblers were designed for Haploid Genomes.
  – Diploid with little structural variation between the chromosomes then a haploid approach can work.
  – Structural heterozygosity appears as separate contigs.

• Select Strains to minimize heterozygosity
  – This helps facilitate assembly.
  – Use inbred lines
  – Double Haploid strains

• Diploid or multiploid genomes
  – Using a haploid assembler leads to fragmented assemblies.
  – Consider Falcon (experimental Code) or Celera Assembler can be configured to favor merging haplotypes.
Towards True Diploid Assemblies

• Truth:
  
  maternal allele
  
  paternal allele

• Current assemblers:

• New diploid/polyplid assembler:

  Associated contig 1

  Primary contig

  Associated contig 2

Keep the long range information while maintaining the relations of the alternative alleles.

FALCON assembler: https://github.com/PacificBiosciences/falcon
Sample Preparation

- Key to a successful assembly is the generation of the longest reads possible.
- Sample quality is critical to maximize potential performance.
- No amplification step during library preparation.
- Recommendations:
  - Take care during extraction to avoid gDNA damage & avoid contaminants.
  - Use extraction methods or kits that produce very high molecular weight gDNA.
  - If contaminants are present, purify starting DNA material prior to library prep.
  - Accurately quantify and qualitatively evaluate gDNA.
  - Include DNA-damage-repair step in library prep.
Iso-Seq™ Method: Sample Prep and Experimental Design for Full-Length cDNA Sequencing

Tyson A. Clark, Ph.D.  February 11, 2015
The way we do RNA-seq now is... you take the transcriptome, you blow it up into pieces and then you try to figure out how they all go back together again... If you think about it, it's kind of a crazy way to do things.”

Michael Snyder
Stanford University

Determination of Transcript Isoforms

Gene

mRNA isoforms

Short-read technologies:

Insufficient Connectivity
Splice Isoform Uncertainty

PacBio's Iso-Seq solution:

Full-length cDNA Sequence Reads
Splice Isoform Certainty – No Assembly Required
The Iso-Seq Method for High-quality, Full-length Transcripts

**Experimental Pipeline**

1. PolyA mRNA
2. cDNA synthesis with adapters
3. Size partitioning & PCR amplification
4. SMRTbell™ ligation
5. PacBio® RS II Sequencing

**Informatics Pipeline**

6. PacBio raw sequence reads
7. Clean sequence reads
8. Isoform clusters
9. Nonredundant transcript isoforms
10. Final isoforms

**Evidenced-based gene models**

[Link](https://github.com/PacificBiosciences/cDNA_primer/)
Detailed Clontech Workflow for Conversion of cDNA into SMRTbell™ Libraries

Total RNA
  ↓ Optional Poly-A Selection
polyA+ RNA
  ↓ Reverse Transcription
Full Length 1st Strand cDNA
  ↓ PCR Optimization
Amplified cDNA
  ↓ Large Scale Amplification
  ↓ Size Selection (BluePippin™, SageELF™, or gel)
    ↓ Optional 5-10 kb size fraction

1-2 kb, 2-3 kb, 3-6 kb, 5-10 kb

Re-Amplification
  ↓ 1-2 kb, 2-3 kb, 3-6 kb, 5-10 kb
SMRTbell™ Template Preparation
  ↓ 1-2 kb, 2-3 kb, 3-6 kb, 5-10 kb
Optional Size Selection (BluePippin or SageELF)
  ↓ 3-6 kb, 5-10 kb
SMRT® Sequencing
  ↓
Distribution of FL Reads from SageELF Fractionated Libraries

0.8-2 kb

2-3 kb

3-5 kb

>5 kb
Applications
Transcript Identification and Annotation

[Diagram showing gene expression patterns across scales of chr1 from 47,295,000 to 47,350,000, with annotations for Brain, Heart, and Liver.]
Identification of Alternatively Spliced Isoforms
Targeted Sequencing

<table>
<thead>
<tr>
<th>Brain</th>
<th>Heart</th>
<th>Liver</th>
</tr>
</thead>
<tbody>
<tr>
<td>74.1%</td>
<td>29.5%</td>
<td>87.9%</td>
</tr>
<tr>
<td>7.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>15.5%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.9%</td>
<td></td>
</tr>
<tr>
<td>1.1%</td>
<td>15.4%</td>
<td>2.5%</td>
</tr>
<tr>
<td>2.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.6%</td>
<td>1.4%</td>
<td>4.5%</td>
</tr>
<tr>
<td>9.8%</td>
<td></td>
<td>0.9%</td>
</tr>
</tbody>
</table>
Defining a personal, allele-specific, and single-molecule long-read transcriptome

Hagen Tiigner\textsuperscript{ab,1}, Fabian Grubert\textsuperscript{ab,1}, Donald Sharon\textsuperscript{ab,1}, and Michael P. Snyder\textsuperscript{a,2}

\textsuperscript{a}Department of Genetics, Stanford University, Yale University, New Haven, CT 06511

Edited by Sherman M. Weissman, Yale University

Personal transcriptomes in which all of an individual’s genetic variations (e.g., single nucleotide variants) and isoforms (e.g., transcript start sites, splice sites, and polyadenylation sites) are quantified for full-length transcripts are poised to revolutionize the field of genomics, for understanding individual biology and disease. However, current technologies have been described previously. To obtain such transcriptomes, we sequenced the lymphoblastoid transcriptomes of two family members (GM12878 and the parents GM12891 and GM12892) using a set of Pacific Biosciences long-read approaches, including PacBio, Canu, and HiCanu 101-bp sequencing and made the following three key observations. First, we found that reads representing all transcripts are evident for most sufficiently expressed genes, even when alleles are not included in the assembled isoforms. Second, we identified novel, previously unidentified splicing isoforms to be present in the transcriptome, thus creating the first personalized annotation for the FCRLA gene. Third, we determined SNVs in a de novo manner by mapping reads to RNA haplotypes, including HapMap haplotypes, and assigned single full-length RNA molecules to haplotypes, and demonstrated Mendelian inheritance. Fourth, we show how RNA molecules can be characterized with allele-specific expression (DAE) and differential allele expression (DAE), and the distance between exons and SNVs is influenced by the number of reads phased full-length isoform reads. The DAI reflects the number of reads, which in turn is dependent on the distance between exon and SNV. This allows the calculation of DAI and the assignment of reads to haplotypes. Overall, in addition to enabling transcriptome annotation, these results demonstrate the first large-scale and full-length personal transcriptomes.
# How Many SMRT Cells?

<table>
<thead>
<tr>
<th>Number of SMRT Cells (per sample)</th>
<th>Experimental Goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Targeted, gene-specific isoform characterization</td>
</tr>
<tr>
<td>1-8</td>
<td>General survey of full-length isoforms in a transcriptome (moderate to high expression levels) with or without size selection</td>
</tr>
<tr>
<td>12-16</td>
<td>A comprehensive survey of full-length isoforms in the transcriptome across 3-4 size fractions</td>
</tr>
<tr>
<td>&gt;16</td>
<td>Deep sequencing for comprehensive isoform discovery and identification of low abundance transcripts across 3-4 size fractions</td>
</tr>
</tbody>
</table>
Iso-Seq™ Analysis & Beyond: Advanced Bioinformatics for Transcriptome Sequencing Using Long Reads

Elizabeth Tseng, Ph.D.
Staff Scientist

FIND MEANING IN COMPLEXITY
Iso-Seq *classify* outputs Full-Length Reads

![Diagram showing the process of Iso-Seq classification](image)

- Full-Length = 5’ primer seen, polyA tail seen, 3’ primer seen
- Identify and remove primers and polyA/T tail
- Identify read stranded-ness
## Artificial Concatemers

<table>
<thead>
<tr>
<th>Cause</th>
<th>Outcome</th>
<th>Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low SMRT® adaptor concentration</td>
<td>Primer-ligated cDNA form concatemers</td>
<td>cDNA primer in the middle</td>
</tr>
</tbody>
</table>

**Artificial Concatemer**

5’ primer  Transcript 1  3’ primer  Transcript 2  5’ primer

**Iso-Seq classify** removes these.
Iso-Seq classify identifies Full-Length Reads

ReadsOfInsert

cDNA primer detection & removal

Full-Length, non-Chimeric Reads
non-Full-Length Reads
Iso-Seq *cluster* outputs High-Quality Polished Sequences

- Full-Length, non-Chimeric Reads
  - ICE
  - Unpolished Consensus

- non-Full-Length Reads
  - Quiver
  - Polished, High-Quality Consensus Sequences
  - Polished, Low-Quality Consensus Sequences
Improved Support for PacBio Long Reads in IGV

Aaron Wenger   awenger@pacb.com   8 Sept 2016
TYPICAL IGV USAGE

ILLUMINA
TYPICAL IGV USAGE
TYPICAL IGV USAGE
BASELINE IGV VIEW

PACBIO

ILLUMINA
PROPERLY SPACED READS
HIDE SMALL INDELS

PACBIO

ILLUMINA
SHOW SINGLE NUCLEOTIDE VARIANTS AT LOW ZOOM
“QUICK CONSENSUS” MODE

PACBIO

ILLUMINA
LABEL LARGE INSERTIONS

PACBIO

ILLUMINA
“QUICK PHASE” BY GROUPING BY BASEPAIR AT POSITION
IGV WITH PACBIO SUPPORT ENABLED
## READ LENGTH ENABLES UNIQUE APPLICATIONS

<table>
<thead>
<tr>
<th>Reference-based</th>
<th>Why length helps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistent coverage</td>
<td>High mappability / span repeats</td>
</tr>
<tr>
<td>Structural variant discovery</td>
<td>Span variants</td>
</tr>
<tr>
<td>Haplotype phasing</td>
<td>Connect heterozygous sites</td>
</tr>
</tbody>
</table>
CONSISTENT COVERAGE: \textit{KMT2D}

Shown are all reads from the 1000 Genomes Project PacBio and Illumina high coverage whole genome datasets for NA19240 with MAPQ > 40. Both PacBio and Illumina reads were mapped with BWA MEM to GRCh38.
STRUCTURAL VARIANT DISCOVERY: DELETION
STRUCTURAL VARIANT DISCOVERY: INSERTION

PACBIO

ILLUMINA
HAPLOTYPE PHASING: FALCON UNZIP

(a) FALCON

Initial assembly graph

SNPs   SY   SNPs   SY   SNPs

Primary contig

Associate contig 1 (Alternative allele)

Associate contig 2 (Alternative allele)

(b)

Phase heterozygous SNPs and identify the haplotype of each read

(c) FALCON-Unzip

Haplotype-resolved assembly graph

SNPs   SYs   SNPs   SYs   SNPs

Updated primary contig

haplotype 1

haplotype 2

haplotype 3

Assembly output

Variation Comparison on Chr. 4

# SNPs/50kbp

SNPs

SVs

Haplotype

Primary contig

Mbp

HAPLOTYPING PHASING: 10KB

HAPLOTYPING 1

HAPLOTYPING 2
HAPLOTYPING PHASING: 60KB
PACBIO SUPPORT IS AVAILABLE IN THE IGV DEV BUILD
PACBIO SUPPORT IS AVAILABLE IN THE IGV DEV BUILD

Utilities for preprocessing data files.
- igvtools

Other IGV Versions

Development Snapshot Build: Latest development snapshot built at least nightly.

Archived Versions

Source Code

Source code repository is hosted at GitHub. Current release branch is 2.3.x:
- https://github.com/igvteam/igv

License

Permission to use this work is granted under the MIT License

Contact

We welcome your feedback. Please send suggestions, requests, and bug reports to our help forum or open a ticket in our issue tracker.

If you are posting about a problem, please include the error message (if applicable) and your igv.log file. It will be in $<your home directory>/igv. When posting a bug report, please include:

1. The version of IGV you are using
2. What you did (e.g. I opened a bedgraph file). If the problem happens when you open a file and/or view certain genomic coordinates, attach the file (or provide a link) and specify the genomic coordinates.
3. What you expected to happen (e.g. I expected IGV to display my data)
4. What actually happened (e.g. I received an error message saying "Unknown extension:.graph")