Detecting Structural Variants in PacBio Reads – Tools and Applications

Aaron Wenger
STRUCTURAL VARIANT = DIFFERENCE ≥50 BP

- **Deletion**
- **Insertion**
- **Duplication**
- **Inversion**
- **Tandem Repeat**
- **Translocation**
## Assays to Detect Structural Variants

<table>
<thead>
<tr>
<th>Technologies</th>
<th>Cytogenetics</th>
<th>Microarray</th>
<th>Sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>karyotype, FISH</td>
<td>oligo aCGH, SNP aCGH</td>
<td>short read, PacBio long read</td>
</tr>
<tr>
<td>Resolution</td>
<td>&gt;5 Mb, &gt;100 kb</td>
<td>&gt;50 kb</td>
<td>basepair</td>
</tr>
<tr>
<td>Limitation</td>
<td>marker density, microscope resolution</td>
<td>probe density</td>
<td>read length (mappability and spanning), systematic bias</td>
</tr>
</tbody>
</table>
STRUCTURAL VARIANTS DETECTED IN A HUMAN GENOME

PacBio

20,000

Short reads

4,000 repeats + GC-rich + large insertions

VARIATION BETWEEN TWO HUMAN GENOMES

variants

5 × 10^6

basepairs affected

5 Mb

3 Mb

10 Mb

SNVs

indels

structural variants

THE HUMAN GENOME IS COMPARATIVELY EASY

<table>
<thead>
<tr>
<th></th>
<th>structural variants per Mb</th>
<th>percent of genome affected by SV</th>
<th>reported short-read sensitivity for SV</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>7</td>
<td>0.3%</td>
<td>20%</td>
</tr>
<tr>
<td>fruit fly</td>
<td>36</td>
<td>1.0%</td>
<td>23%</td>
</tr>
<tr>
<td>zebrafish</td>
<td>162</td>
<td>10.1%</td>
<td>1%</td>
</tr>
</tbody>
</table>


Zebrafish image courtesy of Lizzy Griffiths
EACH CLADE HAS ITS OWN STRUCTURAL VARIANT PROFILE

human

fruit fly

zebrafish
Any study of genetic variation is incomplete until structural variation is detected with PacBio sequencing.
Any study of genetic variation is incomplete until structural variation is detected with PacBio sequencing.
Any study of genetic variation is incomplete until structural variation is detected with PacBio sequencing.
CLINICAL CASE HISTORY

7 yrs  left atrial myxoma resection, atrial repair
10 yrs  testicular mass, right orchiectomy
13 yrs  pituitary tumor
16 yrs  recurrence of myxomata, resection, adrenal microadenoma
18 yrs  recurrence of ventricular myxomata, resection, VT
19 yrs  ACTH-independent Cushing’s disease, thyroid nodules
21 yrs  transphenoidal resection of pituitary
present (26 yrs)  recurrence of myxomata, consideration for heart transplant

genetics suggests Carney complex
PRKAR1A testing negative

short-read whole genome sequencing negative

EVALUATING STRUCTURAL VARIANTS

Sequel System

SEQUENCE (8-FOLD)

Deletions ≥50 bp

Initial call set 6,971

NGM-LR

MAP READS

Insertions ≥50 bp

6,821

PBHoney

CALL VARIANTS

## EVALUATING STRUCTURAL VARIANTS

<table>
<thead>
<tr>
<th>Sequel System</th>
<th>NGM-LR</th>
<th>PBHoney</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SEQUENCE (8-FOLD)</strong></td>
<td><strong>MAP READS</strong></td>
<td><strong>CALL VARIANTS</strong></td>
</tr>
<tr>
<td>Deletions ≥50 bp</td>
<td>Insertions ≥50 bp</td>
<td></td>
</tr>
<tr>
<td>Initial call set</td>
<td>6,971</td>
<td>6,821</td>
</tr>
<tr>
<td>Not in segdup</td>
<td>5,893</td>
<td>6,254</td>
</tr>
<tr>
<td>Not in NA12878 “healthy” control</td>
<td>2,476</td>
<td>3,171</td>
</tr>
<tr>
<td>Overlaps RefSeq coding exon</td>
<td>39</td>
<td>16</td>
</tr>
<tr>
<td>Gene linked to some disease in OMIM</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

HETEROZYGOUS 2.2 KB DELETION IN *PRKAR1A*

chr17: 66,510,000

Deletion chr17:66,510,475-66,512,658

PRKAR1A

RNA-SEQ SHOWS REDUCED EXPRESSION AND NOVEL EXON-EXON JUNCTION

DELETION CLASSIFIED AS PATHOGENIC

Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology

PVS1 null variant (nonsense, frameshift, canonical splice sites, initiation codon, single or multiexon deletion) in a gene where loss of function is a known mechanism of disease.

PS2 de novo (both maternity and paternity confirmed) in a patient with the disease and no family history

CLINICAL CASE HISTORY

7 yrs  left atrial myxoma resection, atrial repair

10 yrs testicular mass, right orchiectomy

13 yrs pituitary tumor

16 yrs recurrence of myxomata, resection, adrenal microadenoma

18 yrs recurrence of ventricular myxomata, resection, VT

19 yrs ACTH-independent Cushing’s disease, thyroid nodules

21 yrs transphenoidal resection of pituitary

present (26 yrs) recurrence of myxomata, consideration for heart transplant

---

genetics suggests Carney complex
PRKAR1A testing negative

short-read whole genome sequencing negative

PacBio sequencing identifies causative structural variant in PRKAR1A

Long-read genome sequencing identifies causal structural variation in a Mendelian disease

Jason D. Merker, MD, PhD,1,2 Aaron M. Wenger, PhD,3 Tam Sneddon, DPhil,2 Megan Grove, MS, LCGC,2 Zachary Zappala, PhD,1,4 Laure Fresard, PhD,1,4 Daryl Waggott, MSc,5,6 Sowmi Utiramerur, MS,2 Yanli Hou, PhD,1 Kevin S. Smith, PhD,1, Stephen B. Montgomery, PhD,1,4 Matthew Wheeler, MD, PhD,5,6 Jillian G. Buchan, PhD,1,2, Christine C. Lambert, BA,3 Kevin S. Eng, MS,3 Luke Hickey, BS,3 Jonas Kornlach, PhD,3, James Ford, MD,4,5,7 and Euan A. Ashley, MRCP, DPhil,2,4,5,6

Purpose: Current clinical genomics assays primarily utilize short-read sequencing (SRS), but SRS has limited ability to evaluate repetitive regions and structural variants. Long-read sequencing (LRS) has complementary strengths, and we aimed to determine whether LRS could offer a means to identify overlooked genetic variation in patients undiagnosed by SRS.

Methods: We performed low-coverage genome LRS to identify structural variants in a patient who presented with multiple neoplasia and cardiac myxomata, in whom the results of targeted clinical testing and genome SRS were negative.

Results: This LRS approach yielded 6,971 deletions and 6,821 insertions > 50 bp. Filtering for variants that are absent in an unrelated control and overlap a disease gene coding exon identified three deletions and three insertions. One of these, a heterozygous 2,184 bp deletion, overlaps the first coding exon of PRKAR1A, which is implicated in autosomal dominant Carney complex. RNA sequencing demonstrated decreased PRKAR1A expression. The deletion was classified as pathogenic based on guidelines for interpretation of sequence variants.

Conclusion: This first successful application of genome LRS to identify a pathogenic variant in a patient suggests that LRS has significant potential for the identification of disease-causing structural variation. Larger studies will ultimately be required to evaluate the potential clinical utility of LRS.

Genet Med advance online publication 22 June 2017

Key Words: Carney complex; long-read sequencing; PacBio; PRKAR1A; structural variant
Any study of genetic variation is incomplete until structural variation is detected with PacBio sequencing.
SMRT LINK 5.0 INTRODUCES PBSV

`pbsv` command line utility for top-level commands

`pbsvutil` command line utility for detailed commands

SMRT Link web interface
PBSV: SMRT LINK STRUCTURAL VARIANT CALLER

SMRT Analysis

Create New Analysis - Settings

Name *
HG00733 10-fold SV

Analysis Application *
Structural Variant Calling [Beta]

References *
hg38

Data Sets

Name
HG00733_Subreads

Structural Variants

minimum reads that support variant (count)
2

minimum reads that support variant (percent)
0.2
## PBSV: SMRT LINK STRUCTURAL VARIANT CALLER

### SMRT Analysis

<table>
<thead>
<tr>
<th>File</th>
<th>Size</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Log</td>
<td>0 bytes</td>
<td>log</td>
</tr>
<tr>
<td>Structural variants</td>
<td>12,426,862 bytes</td>
<td>vcf</td>
</tr>
<tr>
<td>Structural variants</td>
<td>6,823,102 bytes</td>
<td>bed</td>
</tr>
<tr>
<td>Aligned reads</td>
<td>18,542,771,404 bytes</td>
<td>bam</td>
</tr>
<tr>
<td>Master Log</td>
<td>633 bytes</td>
<td>log</td>
</tr>
</tbody>
</table>
### PBSV: SMRT LINK STRUCTURAL VARIANT CALLER

#### SMRT Analysis

<table>
<thead>
<tr>
<th>File</th>
<th>Size</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Log</td>
<td>0 bytes</td>
<td>log</td>
</tr>
<tr>
<td>Structural variants</td>
<td>12,426,862 bytes</td>
<td>vcf</td>
</tr>
<tr>
<td>Structural variants</td>
<td>6,823,102 bytes</td>
<td>bed</td>
</tr>
<tr>
<td>Aligned reads</td>
<td>18,542,771,404 bytes</td>
<td>bam</td>
</tr>
<tr>
<td>Master Log</td>
<td>633 bytes</td>
<td>log</td>
</tr>
</tbody>
</table>

**chr1**

**904490**

ACGCGGCCCGCTCCTCTCCTCCGAACGTGCGCTCTCGAACGCGGCCCGCTCCTCTCCTCCGAACGCGGCCCGCTCCTCTCCTCCGA

**PASS**

**IMPRECISE; SVTYPE=DEL; END=904587; SVLEN=-97; SVANN=TANDEM**

**GT:AD:DP**

0/1:9:15
PBSV: SMRT LINK STRUCTURAL VARIANT CALLER

SMRT Analysis

<table>
<thead>
<tr>
<th>File</th>
<th>Size</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Log</td>
<td>0 bytes</td>
<td>log</td>
</tr>
<tr>
<td>Structural variants</td>
<td>12,426,862 bytes</td>
<td>vcf</td>
</tr>
<tr>
<td>Structural variants</td>
<td>6,823,102 bytes</td>
<td>bed</td>
</tr>
<tr>
<td>Aligned reads</td>
<td>18,542,771,404 bytes</td>
<td>bam</td>
</tr>
<tr>
<td>Master Log</td>
<td>633 bytes</td>
<td>log</td>
</tr>
</tbody>
</table>

chr1 904490 904587 Deletion -97 . GT:AD:DP 0/1:9:15 SVANN=TANDEM
## PBSV: SMRT LINK STRUCTURAL VARIANT CALLER

### SMRT Analysis

<table>
<thead>
<tr>
<th>File</th>
<th>Size</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Log</td>
<td>0 bytes</td>
<td>log</td>
</tr>
<tr>
<td>Structural variants</td>
<td>12,426,862 bytes</td>
<td>vcf</td>
</tr>
<tr>
<td>Structural variants</td>
<td>6,823,102 bytes</td>
<td>bed</td>
</tr>
<tr>
<td><strong>Aligned reads</strong></td>
<td><strong>18,542,771,404 bytes</strong></td>
<td><strong>bam</strong></td>
</tr>
<tr>
<td>Master Log</td>
<td>633 bytes</td>
<td>log</td>
</tr>
</tbody>
</table>
# PBSV: SMRT LINK STRUCTURAL VARIANT CALLER

## SMRT Analysis

<table>
<thead>
<tr>
<th>Count by Annotation</th>
<th>Insertions (count)</th>
<th>Insertions (total bp)</th>
<th>Deletions (count)</th>
<th>Deletions (total bp)</th>
<th>All Variants (count)</th>
<th>All Variants (total bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Length Histogram (&lt;1 kb)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tandem Repeat</td>
<td>7,483</td>
<td>2,742,385</td>
<td>4,210</td>
<td>1,247,305</td>
<td>11,693</td>
<td>3,989,690</td>
</tr>
<tr>
<td>Alu</td>
<td>1,236</td>
<td>398,032</td>
<td>1,177</td>
<td>367,370</td>
<td>2,413</td>
<td>765,402</td>
</tr>
<tr>
<td>L1</td>
<td>44</td>
<td>244,741</td>
<td>83</td>
<td>444,636</td>
<td>127</td>
<td>689,376</td>
</tr>
<tr>
<td>SVA</td>
<td>18</td>
<td>31,987</td>
<td>29</td>
<td>51,831</td>
<td>47</td>
<td>83,818</td>
</tr>
<tr>
<td>Unannotated</td>
<td>4,344</td>
<td>2,007,459</td>
<td>2,661</td>
<td>2,803,452</td>
<td>7,005</td>
<td>4,810,911</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>13,125</td>
<td>5,424,604</td>
<td>8,160</td>
<td>4,914,593</td>
<td>21,285</td>
<td>10,339,197</td>
</tr>
</tbody>
</table>
PBSV: SMRT LINK STRUCTURAL VARIANT CALLER

SMRT Analysis

Length Histogram (<1 kb)

Count by Annotation

Length Histogram (<1 kb)
Length Histogram (>1 kb)

![Length Histogram Graph](image)
PBSV: SMRT LINK STRUCTURAL VARIANT CALLER

SMRT Analysis

Length Histogram (≥1 kb)

Count by Annotation
- Length Histogram (<1 kb)
- Length Histogram (≥1 kb)

Graph showing counts of deletions and insertions by variant length (kb).
HOW MUCH TO SEQUENCE?

Human HG00733 Sequel System 211 Gb (70-fold)

Standard is full dataset Subsample / titrate to lower coverage Evaluate overlap with standard call set

![Graphs showing % SVs detected versus coverage for Het and Hom](image-url)

**Coverage**

- Human HG00733 Sequel System: 211 Gb (70-fold)
- Standard is full dataset
- Subsample / titrate to lower coverage
- Evaluate overlap with standard call set
SUMMARY

SNVs  indels  structural variants

5 Mb  3 Mb  10 Mb

PacBio
20,000

Short reads
4,000
repeats + GC-rich + large insertions

BRIEF REPORT Genetics in Medicine

Long-read genome sequencing identifies causal structural variation in a Mendelian disease

Jason D. Merker, MD, PhD,1,2, Aaron M. Wenger, PhD,2, Tam Sneddon, DPhil,2, Megan Grove, MS, LCGC,2, Zachary Zappala, PhD,1,4, Laure Fresard, PhD,1, Daryl Waggott, MS,2,5,6, Sowmi Ultramur, MS,2, Yanli Hou, PhD,1, Kevin S. Smith, PhD,1, Stephen B. Montgomery, PhD,1,4, Matthew Wheeler, MD, PhD2,6, Jillian G. Buchan, PhD,1,2, Christine C. Lambert, BA,2, Kevin S. Eng, MS,2, Luke Hickey, BS,2, Jonas Korlach, PhD,2, James Ford, MD,2,5,7 and Evan A. Ashley, MRCP, DPhil,2,6,7

Purpose: Current clinical genomics assays primarily utilize short-read sequencing (SRS), but SRS has limited ability to evaluate repetitive regions and structural variants. Long-read sequencing (LRS) has complementary strengths, and we aimed to determine whether LRS could offer a means to identify overlooked genetic variation in patients undiagnosed by SRS.

Methods: We performed low-coverage genome LRS to identify structural variants in a patient with multiple neoplasia and cardiac myxomas, in whom the results of targeted clinical testing and genome SRS were negative.

Results: This LRS approach yielded 6,971 deletions and 6,821 insertions >50 bp. Filtering for variants that are absent in an unrelated control and overlap a disease gene coding exon identified three deletions and three insertions. One of these, a heterozygous 2,016 bp deletion, overlaps the first coding exon of PRRKRA, which is implicated in autosomal dominant Carney complex. RNA sequencing demonstrated decreased PRRKRA expression. The deletion was classified as pathogenic based on guidelines for interpretation of sequence variants.

Conclusions: This first successful application of genome LRS to identify a pathogenic variant in a patient suggests that LRS has significant potential for the identification of disease-causing structural variation. Larger studies will ultimately be required to evaluate the potential clinical utility of LRS.

Key Words: Carney complex; long-read sequencing; PacBio; PRRKRA; structural variant
ACKNOWLEDGMENTS

Stanford Univ
- Euan Ashley
- Jason Merker
- Sowmi Utiramerur

PacBio
- Yuan Li
- Chris Dunn
- Ben Lerch
- Jim Drake
- Nat Echols
- Aaron Klammer
- Christine Lambert

Schatz Lab
- Michael Schatz
- Philipp Rescheneder
- Fritz Sedlazeck

NGM-LR

- penalty
- gap size
- convex
- errors
- indels

Long-read genome sequencing identifies causal structural variation in a Mendelian disease

Prominent clinical phenotypes are primarily caused by single gene mutations, which can be the result of structural variability in a Mendelian pattern. Long-read sequencing provides high-resolution genome-wide structural variation assessment. Long reads are accurately aligned based on resampling for improved genome assembly. More accurate genome assembly allows for the identification of structural variation in a patient with multiple genetic abnormalities. Long reads from long reads are used to construct a complete picture of the full genome sequence. Long reads from the full genome sequence are used to construct a complete picture of the full genome sequence.