Reducing INDEL calling errors in whole genome and exome sequencing data.

Han Fang

November 8, 2014
CSHL Biological Data Science Meeting
**Acknowledgments**

**Lyon Lab**  
Yiyang Wu  
Jason O’Rawe  
Laura J Barron  
Max Doerfel  
Constantine Hartofilis

**Schatz Lab**  
Giuseppe Narzisi  
Hayan Lee  
James Gurtowski  
Tyler Garvin  
Maria Nattestad  
Srividya Ramakrishnan

**Colleagues from Cold Spring Harbor Laboratory**  
Sara Ballouz  
Wim Verleyen  
Jesse Gillis  
Shane McCarthy  
Eric Antoniou  
Elena Ghiban  
Melissa Kramer  
Stephanie Muller  
Senem M Eskipehlivan  
Michael Wigler  
Ivan lossifov  
Michael Ronemus  
Julie Rosenbaum  
Rob Aboukhalil  
IT department
Significantly higher rates of *de novo* frame-shifts & LGDs in the affected vs. unaffected siblings

The contribution of *de novo* coding mutations to autism spectrum disorder.
Sources of INDEL calling errors?
Scalpel: Haplotype Microassembly

- Extract reads mapping within the exon including (1) well-mapped reads, (2) soft-clipped reads, and (3) anchored pairs.
- Decompose reads into overlapping $k$-mers and construct de Bruijn graph from the reads.
- Find end-to-end haplotype paths spanning the region.
- Align assembled sequences to reference to detect mutations.

Accurate de novo and transmitted indel detection in exome-capture data using microassembly.
Nature Methods. doi: [10.1038/nmeth.3069](https://doi.org/10.1038/nmeth.3069)
Scalpel INDEL Validation

- 1000 INDELs selected for validation
  - 200 Scalpel-specific
  - 200 GATK HapCaller-specific
  - 200 SOAPindel-specific
  - 200 within the intersection
  - 200 long indels (>30bp)

77% PPV
50% PPV
22% PPV

- Scalpel
- SOAPindel
- HaplotypeCaller

Venn diagram: 454 (10.3%), 239 (5.4%), 223 (5.1%), 1,397 (31.7%), 1,633 (37.1%), 304 (6.8%)
Concordance between WGS and WES data.

Reducing INDEL errors in whole genome and exome sequencing data.
Validation results

- The validation rate of WGS-WES intersection INDELs was in fact very high (95%).
- Accuracy of INDEL detection with WES is much lower than that with WGS.
- The WES-specific set had a much smaller fraction of large INDELs.

<table>
<thead>
<tr>
<th></th>
<th>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>WGS-WES 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-specific</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-specific</td>
<td>161</td>
<td>91</td>
<td>56.5%</td>
<td>1</td>
<td>1</td>
<td>100%</td>
</tr>
</tbody>
</table>
Example of WES missing a large INDEL
Coverage distributions (WGS-specific INDELs regions)

\[ C_v = 75.3\% \]

\[ C_v = 281.5\% \]

Coefficient of variation (Cv)

\[ C_v^* = \left(1 + \frac{1}{4n}\right) \cdot \frac{\bar{X}}{X} \]
Introducing the k-mer Chi-Square scores in Scalpel

The k-mer Chi-Square scores \( \chi^2 = \frac{(C_0^\text{Ref} - C_e^\text{Ref})^2}{C_e^\text{Ref}} + \frac{(C_0^\text{Alt} - C_e^\text{Alt})^2}{C_e^\text{Alt}} \), where \( C_0^\text{Ref} \) and \( C_0^\text{Alt} \) are the observed k-mer coverage for the reference and alternative alleles, \( C_e^\text{Ref} \) and \( C_e^\text{Alt} \) are the expected k-mer coverage, i.e. \( C_e^\text{Ref} = C_e^\text{Alt} = \frac{C_0^\text{Ref} + C_0^\text{Alt}}{2} \).

In a), \( C_0^\text{Ref} = 52, C_0^\text{Alt} = 48 \),
\[ \chi^2 = \frac{(52-50)^2}{50} + \frac{(48-50)^2}{50} = 0.16 \]

In b), \( C_0^\text{Ref} = 90, C_0^\text{Alt} = 10 \),
\[ \chi^2 = \frac{(90-50)^2}{50} + \frac{(10-50)^2}{50} = 64 \]

Figures are customized from http://cdn.vanillaforums.com/gatk.vanillaforums.com/FileUpload/a4/5ac06fc8af4b1b0c474f03e45f9017.png
**Benchmarking**

Effectively distinguish behaviours of problematic INDEL calls from likely true-positives. Can be easily applied to screen INDEL calls and understand their characteristics.

High quality INDELs (low error-rate - 7%):

$$\begin{align*}
\chi^2 &\leq 2.0 & \text{if } C_{o}^{\text{Alt}} &\leq 5 \\
\chi^2 &\leq 4.5 & \text{if } C_{o}^{\text{Alt}} &\leq 10 \\
\chi^2 &\leq 10.8 & \text{if } C_{o}^{\text{Alt}} &> 10
\end{align*}$$

Low quality INDELs (high error-rate - 51%):

$$\chi^2 \geq 10.8 \quad \text{if } C_{o}^{\text{Alt}} \leq 10$$
**WGS yielded more high-quality INDELs than WES.** Poly-A/T is a major contributor to the low quality INDELs, which gives rise to much more errors in the WES-specific set.
Concordance between standard WGS & PCR-free data

- Standard WGS: 538 (15.4%)
- PCR-Free: 2651 (75.8%)
- Position-match: 310 (8.8%)
PCR-free data yielded more high-quality INDELs.
PCR amplification induced many error-prone poly-A/T INDELs to the library; reducing the rate of amplification could effectively increase calling quality.
60X WGS is needed to recover 95% of INDEL. Detection of het INDELs requires higher coverage.
**60X WGS is needed to recover 95% of INDEL.**
Detection of het INDELs requires higher coverage.
Summary

• **Discussed:**
  1) Introducing a highly accurate & open-source algorithm, Scalpel (http://scalpel.sourceforge.net/)
  2) Higher accuracy of INDEL detection with WGS data than that with WES data.
  3) WES data has more false-positives, and misses a lot of large INDELs.
  4) STR regions: major sources of INDEL errors, especially near A/T homopolymers.
  5) Identify the errors introduced by PCR amplifications and caution about them.

• **Implications:**
  1) Recommend WGS data for INDEL analysis (60X PCR-free).
  2) Classification scheme of INDEL calls based off of Chi-Square scores and alternative allele coverage.