Title
Resequencing: The Untold Story - Recognizing False Positives, False Negatives and Structural Variation in User Data

Permalink
https://escholarship.org/uc/item/8rt0j45

Author
Lipzen, Anna

Publication Date
2014-04-21
Resequencing: The Untold Story – Recognizing False Positives, False Negatives and Structural Variation in user Data

Anna Lipzen\textsuperscript{1Δ}, Wendy Schackwitz \textsuperscript{1}, Joel Martin\textsuperscript{1}, Len A. Pennacchio\textsuperscript{1}

\textsuperscript{1} Genomics Division, Lawrence Berkeley National Lab, Berkeley, CA / Department of Energy Joint Genome Institute, Walnut Creek, CA

ΔTo whom correspondence may be addressed. E-mail: alipzen@lbl.gov

March 25, 2012

ACKNOWLEDGMENTS:

The work conducted by the US Department of Energy (DOE) Joint Genome Institute is supported by the Office of Science of the DOE under Contract Number DE-AC02-05CH11231. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government, or any agency thereof, or the Regents of the University of California.

DISCLAIMER:

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.
Resequencing: The Untold Story
Recognizing False Positives, False Negatives and Structural Variation in User Data
Anna Lipzen\textsuperscript{1}, Wendy Schackwitz\textsuperscript{1}, Joel Martin\textsuperscript{1}, Len A. Penncachio\textsuperscript{1}
DOE Joint Genome Institute, Walnut Creek, USA

One of the biggest values our team brings is the 20 years of combined experience analyzing Re-Seq data. Additionally, the JGI has worked on a huge variety of projects, giving us unmatched exposure to Re-Seq data. This experience is used to assist the collaborator with interpreting their results. Below are several examples of false calls that we can identify. Common sources of false positives include: edges of structural variation, Illumina sequence specific errors, collapsed repeats & ambiguously mapped reads. Sources of false negatives include: library bias and sequence divergence.

**False Positives**
- **False Positives**
- **False Negatives**

<table>
<thead>
<tr>
<th>Source</th>
<th>Haploid - divergent</th>
<th>Diploid - conserved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected</td>
<td>Found: 98%</td>
<td>False Positive: 5%</td>
</tr>
<tr>
<td>Misalignments</td>
<td></td>
<td>False Negative: 10%</td>
</tr>
<tr>
<td>Ambiguously mapped</td>
<td>90%</td>
<td></td>
</tr>
<tr>
<td>Low depth</td>
<td>94%</td>
<td></td>
</tr>
<tr>
<td>Structural variants</td>
<td>95%</td>
<td></td>
</tr>
</tbody>
</table>

**Sequencer-originated miscalcs**
- Certain sequence context can make reads prone to Illumina sequence-specific errors.
- This error results in strand-biased false calls.

**Ambiguously mapped reads**
- This was a multi-allelic call in a haploid genome. This is likely a real variant and incorrect call is due to reads mapping ambiguously in a repetitive region.

**Structural Variation**
We use several methods for detecting structural variants. 
\textsuperscript{1} BrockSunce and \textsuperscript{2} Proteo\textsuperscript{2} compute the SV breakpoints based on read mapping results and the reference genome. For projects with overall high sequence coverages, low depth regions and regions where no reads begin (\textsuperscript{3}monocenters\textsuperscript{3}) often flag certain SV events. Some tools are quite good at identifying that SV exists, but they are unable to pin point the precise location of the event. We manually examine these calls to attempt to give an exact result.

**Success rate of SV discovery varies by detection method employed**

---

1. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.