Comparison and Validation of Somatic Mutation Callers

Andrey Sivachenko
TCGA Symposium, November 17-18, 2011
SNVs are Defined Simply...

- (single nucleotide) differences from the reference
- Ideally: resequence and read the results out
- If only everything looked as in the example below
... but SNVs Can be Hard to Call

- Multiple issues in library preparation, sequencing and data processing (base calling, alignments) can result in a spectrum of SNV-like events, from good to terrible
- Need to watch for:
  - Alignment quality around the event
  - “Strandness” – orientation of supporting reads
  - Position in read
  - Sufficient coverage (both in tumor and normal)
  - Sequence context
  - Potential tumor contamination in normal
  - …
Specificity → Need to protect against two types of errors

**Signal:** ~1 somatic mutation per Mb
**Goal:** >95% validation rate and ideally approach 100%
→ Need error rate to be ≤ 0.05 errors/Mb! 99.9999% is not good enough

**Noise:** Two types of false positives

1. **NO EVENT**
   - **At risk:** Every base
   - **Source:** Misread bases, Sequencing artifacts, Misaligned reads

2. **GERMLINE EVENT (in T+N)**
   - **At risk:** ~1000 germline / Mb (known)
     10-20 rare germline / Mb (novel)
   - **Source:** Low coverage in normal
The project initiated with the goal of comparing, evaluating, and improving mutation calling algorithms

- Select a set of reference samples
- Call mutations using different algorithms & compare

Comparison alone allows only to contrast the callers against each other

- If caller A makes a call and caller B does not, it is helpful to characterize the difference
- Is there a difference in heuristics involved?
- Is there a difference in some statistics of such caller-specific SNVs
- Ultimately, one needs the ground truth (validation data)
Data

• For this round of the analysis, the subset of data from Phase III of the project was used
  – 20 Lung Squamous TCGA samples sequenced at Broad (whole-exome)
  – Same sequencing data (distributed between centers as aligned bam files) were called at 4 centers using different algorithms
    • Broad
    • Washington University, Saint Louis
    • UCSC
    • Baylor College of Medicine
  – Resulting callsets shared between the centers for comparison
• In addition, for this work we use RNA-Seq data as a validation dataset
  – Sequenced at UNC for TCGA
Simple Characterization of Mutation Callers

- Look at shared vs center-specific events
  - There is a large overlap, but there are still many calls made by each center alone
  - The center-specific calls have, in general, different properties
  - Are these specific false-positive modes of each caller or specific strength?

TCGA-33-4532

![Venn diagram showing shared and center-specific calls](image-url)

![Scatter plot showing allelic fraction vs coverage in tumor](image-url)
• Tendency to call center-specific events at coverages different from where shared events are located

**Broad-only vs ALL**

**WUSTL-only vs ALL**

**UCSC-only vs ALL**

**BCM-only vs ALL**
Calls vs Allelic Fraction

- Allelic fraction distribution of center-specific calls differs from that of shared calls
**Calls vs Call Quality**

- How do callers qualify their own unique calls – are reported qualities meaningful/reliable?

![Broad-only vs ALL](chart1)

![WUSTL-only vs ALL](chart2)

![UCSC-only vs ALL](chart3)
• Some center-specific calls are questionable upon “manual review” (examples follow)
• Many, however, are convincing
Center-specific call, questionable

- Broad-only, single event at coverage ~1000
  - Questionable alignments in the region; no support in RNA-Seq (all RNA-Seq reads are 0 mapping quality)
Center-Specific Call, questionable

- WUSTL at coverage 5, allelic fraction 0.67
  - Likely, a germline event
Center-Specific Call, questionable

• BCM in TCGA-66-2777
  – Clearly a germline event
WE NEED A LOT OF VALIDATION DATA TO COMPARE THE TOOLS
Using RNA-Seq as Validation Set

- Independent library construction
- Different protocol
- Same sequencing technology
- It is possible to call mutations (de-novo) from aligned RNA-Seq data
  - Likely a too conservative approach
- Assume that de-novo DNA-Seq mutation calling is sufficiently conservative
  - Weaker evidence from RNA-Seq (than what would be required for a stand-alone de-novo call) can be considered as validation
Sensitivity -- depends on coverage and allelic-fraction

Kristian Cibulskis
Is Allelic Fraction an Issue?

- Original calls have a range of allelic fractions
- Is it safe to ask for fixed (low) number of observations in RNA-Seq
  - In general, NO
  - However: AF in RNA-Seq and DNA-Seq strongly correlate
Looking for SNV in RNA-Seq

- Consider every called mutation site with coverage in RNA-Seq above N as “covered”
- If covered site has at least two reads with alt. allele in RNA-Seq, consider it “validated”

<table>
<thead>
<tr>
<th>center</th>
<th>n.calls</th>
<th>covered</th>
<th>validated</th>
<th>validated.pct.covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>BI</td>
<td>405</td>
<td>186</td>
<td>152</td>
<td>81.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BI</td>
<td>405</td>
<td>150</td>
<td>131</td>
<td>87.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

• A framework is established within TCGA for evaluating and improving mutation calling algorithms

• We are working on validating mutations:
  – Using additional experiments in the sequencing centers (but this may be only partial validation)
  – based on RNA-seq after correcting for the power to detect the mutation
Acknowledgments

Broad
Gad Getz
Kristian Cibulskis
Rui Jing
Alex Ramos
Carrie Sougnez
Peter Hammerman
Scott Carter

WUSTL
Li Ding
Mike McLellan
Ken Chen
Xian Fan

UCSC
David Haussler
Chris Wilks
Singer Ma
Zack Sanborn
Rachel Harte
Daniel Zerbino
Jing Zhu

Baylor
David Wheeler
Jennifer Drummond
Kyle Chang

UNC
Neil Hayes
Matthew Wilkerson

TCGA research Network