Callset Evaluation

Comparing statistics between your callset and a truth set
You are here in the GATK Best Practices workflow for germline variant discovery.
Where are you on this spectrum?

<table>
<thead>
<tr>
<th>IDEAL</th>
<th>OKAY</th>
<th>TERRIBLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your variant calls perfectly match the underlying biological truth</td>
<td>You found many real variants and called few false positives</td>
<td>You didn’t find any real variants and only called artifacts!</td>
</tr>
</tbody>
</table>

= What callset evaluation methods aim to determine

*(not veracity of individual variant calls)*
How do I figure out how good/bad my callset is?

MyCallset.vcf

Specs:

<table>
<thead>
<tr>
<th># SNPs</th>
<th># Indels</th>
<th>TiTv Ratio</th>
<th>Indel Ratio</th>
<th>Concordance</th>
</tr>
</thead>
</table>

Guiding principle: divergence is indicative of error
Key assumption: truth set is representative / comparable

- Important to match dataset properties!
  - Population ethnicity (European, African, etc.)
  - Sequencing / exp. design (WGS vs. WES)
  - Cohort size

http://www.nature.com/nature/journal/v526/n7571/full/nature15393.html
Ethnicity affects many variant call metrics
Older populations tend to display more heterogeneity
If possible, use truth sets generated with orthogonal methods

**Sequencing**
- Sanger sequencing
- Other HTS technologies

**Probe/Array-based**
- GeneChip
- Microarrays
Commonly used truth sets

- dbSNP
  All previously reported variation (lots of junk!)

- Sample-matched genotyping chip
  Awesome! But adds cost & limited to known variants

- HapMap
  Highly validated common human variants

- OMNI
  Common variation validated by array

- NIST Genomes in a Bottle (single sample evaluation)
  Consensus callsets from common benchmarking samples
Recommended metrics for callset evaluation

**Number of Indels & SNPs**

**Indel Ratio**

**Genotype Concordance**

**TiTv Ratio**
Number of Indels & SNPs

- Variants = Indels + SNPs
- Useful for order-of-magnitude sanity check

<table>
<thead>
<tr>
<th>Sequencing Type</th>
<th># of Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>WGS</td>
<td>~4.4 M</td>
</tr>
<tr>
<td>WES</td>
<td>~41 k</td>
</tr>
</tbody>
</table>
• If random: expect ratio of 0.5
  Twice as many possible transversions vs transitions!

• Low TiTv ratio indicates high rate of false positives

<table>
<thead>
<tr>
<th>Sequencing Type</th>
<th>TiTv Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>WGS</td>
<td>2.0-2.1</td>
</tr>
<tr>
<td>WES</td>
<td>3.0-3.3</td>
</tr>
</tbody>
</table>
• Ratio of insertions to deletions
• Varies by type of study
e.g. rare variant association vs common variant association
• Most appropriate truth set is genotyping chip for same sample
• % Genotype calls in callset matching GT calls in truth set
• Unmatched variants considered false positives
Cheat sheet of concordance metrics

**SENSITIVITY vs. FALSE DISCOVERY RATE**

- **my callset**: 12
- **gold standard**: 10

- **false positives (FP)**: 5
- **false negatives (FN)**: 3

- **true positives (TP)**: 7

**SENSITIVITY**

\[
\text{SENSITIVITY} = \frac{TP}{TP + FN} = \frac{7}{7 + 3} = 70\%
\]

**FALSE DISCOVERY RATE**

\[
\text{FALSE DISCOVERY RATE} = \frac{FP}{FP + TP} = \frac{5}{5 + 7} = 42\%
\]

**GENOTYPE CONCORDANCE**

- **gold standard**: ★★★★★★
- **my callset**: ★★★★★★

<table>
<thead>
<tr>
<th>matches (4)</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
</table>

- ★ heterozygous (0/1)
- ★★ homozygous-variant (1/1)

**GT CONCORDANCE**

\[
\text{GT CONCORDANCE} = \frac{\sum \text{matches}}{TP} = \frac{4}{7} = 57\%
\]
So how do I get these metrics?

<table>
<thead>
<tr>
<th>Tool</th>
<th>Variant Level Evaluation</th>
<th>Genotype Level Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GATK</strong></td>
<td><strong>VariantEval</strong></td>
<td><strong>GenotypeConcordance</strong></td>
</tr>
<tr>
<td></td>
<td>java -jar GenomeAnalysisTK.jar \</td>
<td>java -jar GenomeAnalysisTK.jar \</td>
</tr>
<tr>
<td></td>
<td>-T VariantEval \</td>
<td>-T GenotypeConcordance \</td>
</tr>
<tr>
<td></td>
<td>-R reference.b37.fasta \</td>
<td>-R reference.b37.fasta \</td>
</tr>
<tr>
<td></td>
<td>-eval callset.vcf \</td>
<td>--comp truthset.vcf \</td>
</tr>
<tr>
<td></td>
<td>-D truthset.vcf \</td>
<td>--eval callset.vcf \</td>
</tr>
<tr>
<td></td>
<td>-o results.eval.grp</td>
<td>-o results.grp</td>
</tr>
<tr>
<td><strong>Picard</strong></td>
<td><strong>CollectVariantCallingMetrics</strong></td>
<td><strong>GenotypeConcordance</strong></td>
</tr>
<tr>
<td></td>
<td>java -jar picard.jar \</td>
<td>java -jar picard.jar \</td>
</tr>
<tr>
<td></td>
<td>CollectVariantCallingMetrics</td>
<td>CollectVariantCallingMetrics</td>
</tr>
<tr>
<td></td>
<td>INPUT=callset.vcf \</td>
<td>INPUT=callset.vcf \</td>
</tr>
<tr>
<td></td>
<td>DBSNP=truthset.vcf \</td>
<td>DBSNP=truthset.vcf \</td>
</tr>
<tr>
<td></td>
<td>OUTPUT=results</td>
<td>OUTPUT=results</td>
</tr>
</tbody>
</table>
Which variant-level evaluator should I use?

**GATK VariantEval**
- More detailed analysis
- More options for stratification
- Ability to compare to multiple truth sets

**Picard CollectVariantCallingMetrics**
- Best performance & speed on very large callsets
- Few options beyond the metrics discussed here
You are here in the GATK Best Practices workflow for germline variant discovery.
Further reading

http://www.broadinstitute.org/gatk/guide/article?id=6308
http://www.nature.com/nature/journal/v526/n7571/full/nature15393.html