MPG NGS workshop I: SNP calling

Mark DePristo

Manager, Medical and Population Genetic Analysis
Genome Sequencing and Analysis Group
Medical and Population Genetics Program
Broad Institute of Harvard and MIT
02/04/10
Three slide background on SNP calling in the GATK
SNP calling workflow

**Data input and output**

- **Call-ready BAM files**
  - (cleaned, dedupped, recalibrated, with well-formatted header)
  - **200Gb**

- **Raw variants (VCF)**
  - (all sites confidently containing non-reference bases; with genotypes)
  - **1 Gb**

- **Filtered variants (VCF)**
  - (separate true segregating variation from machine artifacts)
  - **1 Gb**

**Processing tools**

- GATK unified genotyper
- GATK variant analysis
- Expert user judgement
- GATK variant filtration

**Runtime**

- **GATK unified genotyper**: 10 hrs
- **GATK variant analysis**: Instant
- **Expert user judgement**: Days**
- **GATK variant filtration**: 30 min

**Ease of use**

- Very easy
- Tools are easy to use but parameter selection requires significant expertise and judgement

* Runtime and file sizes are for a single sample 30x whole genome BAM
** Potentially requires many rounds of experimentation and evaluation
GATK single sample genotype likelihoods

\[
L(G \mid D) = P(G) P(D \mid G) = \prod_{b \in \{\text{good\_bases}\}} P(b \mid G)
\]

- Priors applied during multi-sample calculation; \(P(G) = 1\)
- Likelihood of data computed using pileup of bases and associated quality scores at given locus
- Only “good bases” are included: those satisfying minimum base quality, mapping read quality, pair mapping quality, NQS
- \(P(b \mid G)\) uses platform-specific confusion matrices
- \(L(G\mid D)\) computed for all 10 genotypes

We apply a generalization of the single sample SNP caller to Pilot 1

- This approach allows us to combine weak single sample calls to discover variation among samples with high confidence

Making raw variant calls with the GATK unified genotyper
Running the Unified Genotyper

```java
java -Xmx2048m -jar GenomeAnalysisTK.jar
   -R /broad/1KG/reference/human_b36_both.fasta
   -T UnifiedGenotyper
   -D dbsnp_129_b36.rod
   -varout NA19240.raw.vcf
   -confidence 50
   --heterozygosity 1.000000e-03
   -I NA19240.SLX.bam
```

Minimum phred-scaled confidence required to emit a SNP

1 het per 1000 reference bases on average for a Yoruban

BAM file containing NA19240 SLX reads

Raw VCF calls (NA19240.raw.vcf)

<table>
<thead>
<tr>
<th>#CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
<th>INFO</th>
<th>FORMAT</th>
<th>NA19240</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36496</td>
<td>.</td>
<td>T</td>
<td>A</td>
<td>53.13</td>
<td>0</td>
<td>&lt;ATTRIBUTES&gt;</td>
<td>GT:DP:GQ</td>
<td>1/0:6:84.70</td>
</tr>
<tr>
<td>1</td>
<td>45162</td>
<td>rs10399749</td>
<td>C</td>
<td>T</td>
<td>331.37</td>
<td>0</td>
<td>&lt;ATTRIBUTES&gt;</td>
<td>GT:DP:GQ</td>
<td>0/1:27:99.00</td>
</tr>
<tr>
<td>1</td>
<td>48677</td>
<td>.</td>
<td>G</td>
<td>A</td>
<td>399.86</td>
<td>0</td>
<td>&lt;ATTRIBUTES&gt;</td>
<td>GT:DP:GQ</td>
<td>1/0:25:99.00</td>
</tr>
</tbody>
</table>

Long string of variant annotations (more info in a few slides)

SNP calling artifacts

• SNP calls are generally infested with false positives
  – From systematic machine artifacts, mismapped reads, aligned indels/CNV
  – Raw SNP calls might have between 5-20% FPs among novel calls

• Separating true variation from artifacts depends very much on the particulars of one’s data and project goals
  – Whole genome deep data, WG low-pass, hybrid capture, pooled PCR are have significantly different error modes
Filtering artifacts out of your SNP calls

• The GATK uses a three pass approach
  – First emit all sites potentially containing a true variant
  – Aggregate SNP covariates in the raw VCF to determine the relationship between each covariate and error
    [warning: requires user expertise]
  – Finally, apply these filters to the raw VCF using the GATK VariantFiltration tool

• We are currently working on a robust, easy-to-use automated tool
Variant annotations and filters

VCF record for an A/G SNP at 22:49582364

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Position</th>
<th>Genotype</th>
<th>AC</th>
<th>AN</th>
<th>AF</th>
<th>AB</th>
<th>MQ</th>
<th>MQ0</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>49582364</td>
<td>0/1:12:99.00</td>
<td>3</td>
<td>6</td>
<td>0.50</td>
<td>0.67</td>
<td>71.31</td>
<td>22</td>
</tr>
</tbody>
</table>

INFO field

- **AC**: No. chromosomes carrying alt allele
- **AN**: Total no. of chromosomes
- **AF**: Allele frequency
- **AB**: Allele balance of ref/alt in hets
- **DP**: Depth of coverage
- **Dels**: No. of MAPQ 0 reads at locus
- **HRun**: Length of longest contiguous homopolymer
- **MQ**: RMS MAPQ of all reads
- **MQ0**: No. of MAPQ 0 reads at locus
- **QD**: QUAL score over depth
- **SB**: Estimated SB score
- **GT:DP:GQ**: 0/1:12:99.00; 0/1:11:89.43; 0/1:28:37.78

Heterozygous genotype A/G in all three individuals

Selecting filtering thresholds

Selected filters are: AB > 0.75 || DP > 300 || MQ0 > 40 || SB > -0.10 || 3 snps within 10bp

Running Variant Filtration

```
java -Xmx2048m -jar GenomeAnalysisTK.jar
    -R /broad/1KG/reference/human_b36_both.fasta
    -T VariantFiltration
    -B variant,VCF,NA19240.raw.vcf
    -D dbsnp_129_b36.rod
    --clusterWindowSize 10
    --filterExpression "AB > 0.75 || DP > 300 || MQ0 > 40 || SB > -0.10"
    -l INFO
    -o NA19240.filtered.vcf
```

Filters out any group of 3 SNPs within 10 bp of each other

Expression describing SNPs that should be filtered out

Filtered VCF calls (NA19240.filtered.vcf)

<table>
<thead>
<tr>
<th>#CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
<th>INFO</th>
<th>FORMAT</th>
<th>NA19240</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36496</td>
<td>.</td>
<td>T</td>
<td>A</td>
<td>53.13</td>
<td>GATK_FILTER</td>
<td></td>
<td>GT:DP:GQ</td>
<td>1/0:6:84.70</td>
</tr>
<tr>
<td>1</td>
<td>45162</td>
<td>rs10399749</td>
<td>C</td>
<td>T</td>
<td>331.37</td>
<td>0</td>
<td>&lt;ATTRIBUTES&gt;</td>
<td>GT:DP:GQ</td>
<td>0/1:27:99.00</td>
</tr>
<tr>
<td>1</td>
<td>48677</td>
<td>.</td>
<td>G</td>
<td>A</td>
<td>399.86</td>
<td>0</td>
<td>&lt;ATTRIBUTES&gt;</td>
<td>GT:DP:GQ</td>
<td>1/0:25:99.00</td>
</tr>
</tbody>
</table>

SNPs with poor characteristics have their FILTER field filled in
Raw and filtered autosomal calls for YRI daughter and trio

<table>
<thead>
<tr>
<th>Call set</th>
<th>Callable bases(^1)</th>
<th># variants</th>
<th>dbSNP%</th>
<th>Ti/Tv (Est. FP rate(^2))</th>
<th>Hapmap 3 Sensitivity(^3)</th>
<th>Hapmap 3 Concordance(^3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single individual calls from the GATK</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw NA19240</td>
<td>2.70B (89%)</td>
<td>4.52M</td>
<td>77.83</td>
<td>2.07 (1.9%)</td>
<td>1.81 (18.1%)</td>
<td>99.41</td>
</tr>
<tr>
<td>Filtered NA19240</td>
<td>4.26M</td>
<td>80.42</td>
<td></td>
<td>2.10 (~0.0%)</td>
<td>2.01 (5.6%)</td>
<td>99.14</td>
</tr>
<tr>
<td><strong>Daughter + parents multi-sample calls from the GATK</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw YRI trio together</td>
<td>2.5B (81%)</td>
<td>6.24M</td>
<td>71.65</td>
<td>2.07 (1.9%)</td>
<td>1.80 (18.8%)</td>
<td>99.62</td>
</tr>
<tr>
<td>Filtered YRI trio together</td>
<td>5.60M</td>
<td>74.86</td>
<td></td>
<td>2.11 (~0.0%)</td>
<td>2.02 (5.0%)</td>
<td>99.29</td>
</tr>
</tbody>
</table>

1. % of all 3.1B bases of the B36 human genome called with at least Q50 confidence
2. Calculated as 1 - (titv_Observed - 0.5) / (titv_Expected - 0.5) with titv_Expected of 2.1
3. NA19240 sensitivity and concordance results
Example novel variant

Chr1:67634785 in 3’ untranslated region
Example scripts

• 1000 Genomes SLX YRI BAM files:
  – Locally available at:
    /humgen/gsa-hpprojects/1kg/1kg_pilot2/
    useTheseBamsForAnalyses/<sample>.SLX.bam
  – Available for download at 1000genomes.org

• Scripts and VCF files:
  – /humgen/gsa-scr1/pub/tutorials/MPG_workshop
Appendix
Choosing a minimum confidence score for a SNP

- Each point on plot includes ~3000 SNPs from NA19240
- The density of points across the confidence interval indicates the number of SNPs
- ~0.5% of SNPs have Q < 100, and only 2% are less than Q < 200
- The default Q50 threshold results in an highly sensitive call set