Variant Filtering

Ecole de Bioinformatique – Roscoff
octobre 2015
Workflow

1. Raw reads (Fastq)
2. Mapping on the reference genome
3. Mapping Post-processing
4. Variant Calling Pre-processing
5. Variant Calling
6. Variant Filtering & Annotation
Why filter?

Some use cases:

- Extract a subset of variants (localization, type)
- Combine variants from several analyses
- Compare obtained variants from several data types (RNA-Seq, Exome-Seq, Whole Genome)
- Identify new variants compare to a reference list
- Apply specific filter for Chip design
- ...
Use specific tools to rewrite / annotate VCF File.

Reminder (VCF Format):

```plaintext
##fileformat=VCFv4.1
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">
##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Deletions">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most two segregating haplotypes">
##FILTER=<ID=LowQual,Description="Low quality">
```

...
Which Tools?

**VcfTools** (http://vcftools.sourceforge.net/)

- C++ /Perl
- Tool / Tool suite
  - Filter
  - Convert
  - Compare
  - Stats (LD estimate, ...)
  - Merge
  - Annotate

- Available via Toolshed (limited implementation)
**Which Tools?**

**BcfTools** ([http://www.htslib.org/man/bcftools/](http://www.htslib.org/man/bcftools/)) well suited to Samtools

- C++

- Tool suite:
  - Annotate (edit, add annotations)
  - Concat (concatenate data from same sample)
  - Filter
  - Merge
  - Stats
  - View (subset, filter, convert)
  - Call (variant calling)

- Available via Toolshed (limited implementation)
Which Tools?

**vcflib** ([https://github.com/ekg/vcflib](https://github.com/ekg/vcflib)), well suited to FreeBayes

- C++ library
- Tool suite:
  - Comparison (union, intersection, combine vcf files)
  - Format conversion (export to tsv, SQLite, Bed file)
  - Filtering (filter with expression, subsample, variant types)
  - Annotation (from other VCF File, Bed file, nearest variant)
  - Variant representation (complex variant, multiallelic)
  - Genotype manipulation (remove aberrant genotype, provide GL → genolikelihood)
  - Classification of variants (heterozygosity, by annotation, pcr primers)

- Available via Toolshed (well maintained)
Which Tools?

**GATK** ([https://www.broadinstitute.org/gatk/](https://www.broadinstitute.org/gatk/))

- Java
- Tool suite:
  - Filtering (filter with expression, ...)
  - Selection (criteria, reference file)
  - Annotation (reference file, VCF specific tag)
  - Comparison (union, intersection, combine vcf files)
  - Evaluation (report, stratification)
  - ...

- Available via Toolshed (well maintained)
HOWTO?

- Remove variant entry or add Filter info (Hard Filtering):

**Before Filtering**

<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>Pickrell</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr12</td>
<td>406292</td>
<td>rs2229351</td>
<td>G</td>
<td>A</td>
<td>994.77</td>
<td></td>
<td>.</td>
<td>AC=1;AF=0.500;AN=2;DB;DP=69;Dels=0.00;FS=4.853;MQ=37.00;MQ0=0;GT:AD:DP:GQ:PL</td>
<td>0/1:33,36:66:99:1023,0,994</td>
</tr>
<tr>
<td>Chr12</td>
<td>416046</td>
<td>rs35042439</td>
<td>C</td>
<td>CT</td>
<td>391.73</td>
<td></td>
<td>.</td>
<td>AC=1;AF=0.500;AN=2;DB;DP=46;FS=0.000;MQ=37.49;MQ0=0;QD=8.52;GT:AD:DP:GQ:PL</td>
<td>0/1:22,17:46:99:429,0,521</td>
</tr>
</tbody>
</table>

**After Filtering**

Filter: **MQ < 30.0**

<table>
<thead>
<tr>
<th>#FILTER=&lt;ID=LowQual,Description=&quot;Low quality, mapping &lt; 30.0&quot;&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>#CHROM</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>Chr12</td>
</tr>
<tr>
<td>Chr12</td>
</tr>
</tbody>
</table>

Filter: **PASS** (Not Filtered), **LowQual** (Filtered)
HOWTO?

• Understand VCF Format File
• Identify specific tags
• Fix Thresholds
• Find external resources (dbSNP) to exclude / keep known Variants (other VCF File)
• Limit analysis to specific genomic locations (BED File)
Which Criteria?

It depends on:

- Variant caller: methods, available info, VCF specific tags
- Data Type: DNA-Seq, Exome-Seq, RNA-Seq,
- Sequencing Technology: (depth, protocol)
- Reference genome: reliability of the reference sequence
- Studied species: Genome features (Transposable Elements, Tandem Repeats)
- Available resources: reference variant sets
Which Criteria ?

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*

- **Depth (DP):** Min / Max => \( d \pm 4\sqrt{d} \), \( d \) = average Read Depth

 Nb SNPs

![Graph showing coverage distribution across genomes with different depths.](image)

- Low DP : mapping errors, sequencing errors
- High DP : CNVs or Repeat Regions, mapping errors
- Reliable with High coverage > 40X
- DNA-Seq OK, Exome-Seq NOK
Which Criteria?

Li, H. (2014). *Bioinformatics*
- **Low Complexity regions** (LC) : exclude variants located in LC regions
- **Variant Quality** (QU) : exclude variants with low quality
- **Double Strand filter** (DS) : exclude variants with number of reads (ALT allele) below a defined Threshold on reverse or forward strand
- **Fisher Strand filter** (FS) : reference / no-reference reads highly correlated with strand.
- **Allele Balance** (AB) : HET > 30%

Meynert et al (2014). *BMC Bioinformatics*
- Coverage uniformity vs coverage depth (whole genome vs exome-seq) -> critical for heterozygous sites
Which Criteria?

Li, H. (2014). *Bioinformatics*, misc = AB, DS, FS
GATK: VCF Format

**FORMAT Tags**

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT</td>
<td>Genotype, 0/0, 0/1, 1/1</td>
</tr>
<tr>
<td>GQ</td>
<td>Genotype Quality (Highest value = 99)</td>
</tr>
<tr>
<td>AD / DP</td>
<td>Depth per Allele / Depth = global coverage</td>
</tr>
<tr>
<td>PL</td>
<td>Genotype Likelihoods, max 0 (Phred Score)</td>
</tr>
</tbody>
</table>

**INFO Tags**

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC, AF, AN</td>
<td>(AC) Alleles Count, and (AF) Allele Frequency for each ALT allele, (AN)Total number of allele</td>
</tr>
<tr>
<td>DB</td>
<td>If present, then the variant is in dbSNP.</td>
</tr>
<tr>
<td>DP</td>
<td>Coverage (reads that passed quality metrics)</td>
</tr>
<tr>
<td>DS</td>
<td>Were any of the samples downsampled because of too much coverage?</td>
</tr>
<tr>
<td>MQ and MQ0</td>
<td>Root Mean Square Mapping Quality and Mapping Quality Zero total count</td>
</tr>
<tr>
<td>BaseQualityRankSum</td>
<td>Test: quality of Reference reads vs ALT reads</td>
</tr>
<tr>
<td>MappingQualityRankSum</td>
<td>Test: Mapping quality of Reference reads vs ALT reads</td>
</tr>
<tr>
<td>ReadPosRankSum</td>
<td>Test: Distance of ALT reads from the end of the reads</td>
</tr>
<tr>
<td>HaplotypeScore</td>
<td>Consistency of the site with at most two segregating haplotype</td>
</tr>
<tr>
<td>QD</td>
<td>Variant Quality / depth of non-ref samples</td>
</tr>
<tr>
<td>FS</td>
<td>Test (Fisher): Phred score p-value for strand bias</td>
</tr>
<tr>
<td>InbreedingCoeff</td>
<td>Inbreeding coefficient as estimated from the genotype likelihoods per-sample when compared against the Hardy-Weinberg expectation</td>
</tr>
</tbody>
</table>

http://gatkforums.broadinstitute.org/discussion/1268/how-should-i-interpret-vcf-files-produced-by-the-gatk
GATK: recommended filters

Use case: non-reference variant db, GATK recommended filters for recalibration

- **SNPs:**
  - $QD < 2.0$ (Variant Quality / depth of non-ref samples)
  - $MQ < 40.0$ (Mapping Quality)
  - $FS > 60.0$ (Phred score Fisher’s test p-value for strand bias)
  - $HaplotypeScore > 13.0$ (Consistency of the site with at most two segregating haplotype)
  - $MQRankSum < -12.5$ (Mapping quality of Reference reads vs ALT reads)
  - $ReadPosRankSum < -8.0$ (Distance of ALT reads from the end of the reads)

- **INDELs:**
  - $QD < 2.0$ (Variant Quality / depth of non-ref samples)
  - $ReadPosRankSum < -20.0$ (Distance of ALT reads from the end of the reads)
  - $InbreedingCoeff < 0.8$
  - $FS > 200.0$ (Phred score Fisher’s test p-value for strand bias)

**Tutorial**

- **GATK : Variant Filtration**
- **GATK : Select Variants**
- **GATK : Combine Variants**
- **GATK : Variant Annotator**
- **GATK : Eval Variants**

**Exome-Seq (GATK calling)**
- Apply filters on GATK available tags
- Extract filtered variants
- Combine / Merge results in one file
- Annotate with external resources
- Variant Evaluation on specific criteria / data / caller

**RNA-Seq (Varscan calling)**
- Apply filters on Varscan available tags
- Extract filtered variants
GATK: Variant Filtration

- Modify FILTER column (Hard Filtering)
- Criteria on INFO Tags
- Criteria on FORMAT Tags
- Handle missing Values

https://www.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walkers_filters_VariantFiltration.php
GATK: Select Variants

- Direct selection (exclude filtered variants)
- Criteria on INFO Tags
- Criteria on FILTER Tags
- No Criteria on FORMAT Tags
- Intersection / Union with other VCF Files
- Exclude / Include samples
- Selected genomic regions (BED File)

https://www.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walkers_variantutils_SelectVariants.php
Criteria: JEXL expression

JEXL = Java Expression Language

- Key, value
- Case-sensitive (Uppercase, Lowercase, MQ ≠ mq)
- Type-sensitive:
  - ##FORMAT=<ID=AD,Number=.,Type=Integer,Description=".."
  - Integer = 2
  - Float = 2.0
  - String = "two"

- Operators:
  - Relational: ==, !=, <,>,<=, >=
  - Logical: && (AND) , || (OR)

http://gatkforums.broadinstitute.org/discussion/1255/what-are-jexl-expressions-and-how-can-i-use-them-with-the-gatk
GATK: Combine Variants

- Combine Variant from VCF files:
  - Several samples
  - Different methods

- Add "set" tag in INFO column
  - set = file1, unique to file1
  - set = Intersection, found in all files

Calculate quality control metrics

- 1 / multiple vcf
- Stratification modules (Novelty, Sample,...)
- Metrics modules:
  - CompOverlap: overlap between eval and comp sites
  - CountVariants: counts of variant classes in the sample
  - TiTvVariantEvaluator: Transition/Transversion Variant Evaluator

https://www.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walker_s_varianteval_VariantEval.php
GATK: Eval Variants

**Example:** Human expected Ti/Tv:

- Ti (A\textbf{->}G, C\textbf{->}T) twice as frequently as Tv (A\textbf{->}C, A\textbf{->}T, G\textbf{->}C,G\textbf{->}T)
- Ti/Tv > 3 in coding regions (exome)

http://genome.sph.umich.edu/wiki/SNP_Call_Set_Properties

**Eval Report:**

<table>
<thead>
<tr>
<th>TiTvVariantEvaluator</th>
<th>CompRod</th>
<th>EvalRod</th>
<th>JexlExpression</th>
<th>Novelty</th>
<th>nTi</th>
<th>nTv</th>
<th>tiTvRatio</th>
<th>nTiInComp</th>
<th>nTvInComp</th>
<th>TiTvRatioStandard</th>
</tr>
</thead>
<tbody>
<tr>
<td>TiTvVariantEvaluator</td>
<td>dbsnp</td>
<td>input_0</td>
<td>Gatk</td>
<td>all</td>
<td>640</td>
<td>241</td>
<td>2.66</td>
<td>634</td>
<td>241</td>
<td>2.63</td>
</tr>
<tr>
<td>TiTvVariantEvaluator</td>
<td>dbsnp</td>
<td>input_0</td>
<td>Gatk</td>
<td>known</td>
<td>636</td>
<td>239</td>
<td>2.66</td>
<td>633</td>
<td>241</td>
<td>2.63</td>
</tr>
<tr>
<td>TiTvVariantEvaluator</td>
<td>dbsnp</td>
<td>input_0</td>
<td>Gatk</td>
<td>novel</td>
<td>4</td>
<td>2</td>
<td>2.00</td>
<td>1</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>TiTvVariantEvaluator</td>
<td>dbsnp</td>
<td>input_0</td>
<td>Intersection</td>
<td>all</td>
<td>95</td>
<td>30</td>
<td>3.17</td>
<td>88</td>
<td>29</td>
<td>3.03</td>
</tr>
<tr>
<td>TiTvVariantEvaluator</td>
<td>dbsnp</td>
<td>input_0</td>
<td>Intersection</td>
<td>known</td>
<td>95</td>
<td>30</td>
<td>3.17</td>
<td>88</td>
<td>29</td>
<td>3.03</td>
</tr>
<tr>
<td>TiTvVariantEvaluator</td>
<td>dbsnp</td>
<td>input_0</td>
<td>Intersection</td>
<td>novel</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>TiTvVariantEvaluator</td>
<td>dbsnp</td>
<td>input_0</td>
<td>Varscan</td>
<td>all</td>
<td>192</td>
<td>53</td>
<td>3.62</td>
<td>162</td>
<td>45</td>
<td>3.60</td>
</tr>
<tr>
<td>TiTvVariantEvaluator</td>
<td>dbsnp</td>
<td>input_0</td>
<td>Varscan</td>
<td>known</td>
<td>172</td>
<td>44</td>
<td>3.91</td>
<td>162</td>
<td>44</td>
<td>3.68</td>
</tr>
<tr>
<td>TiTvVariantEvaluator</td>
<td>dbsnp</td>
<td>input_0</td>
<td>Varscan</td>
<td>novel</td>
<td>20</td>
<td>9</td>
<td>2.22</td>
<td>0</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>TiTvVariantEvaluator</td>
<td>dbsnp</td>
<td>input_0</td>
<td>none</td>
<td>all</td>
<td>927</td>
<td>324</td>
<td>2.86</td>
<td>1398900</td>
<td>692871</td>
<td>2.02</td>
</tr>
<tr>
<td>TiTvVariantEvaluator</td>
<td>dbsnp</td>
<td>input_0</td>
<td>none</td>
<td>known</td>
<td>903</td>
<td>313</td>
<td>2.88</td>
<td>883</td>
<td>314</td>
<td>2.81</td>
</tr>
<tr>
<td>TiTvVariantEvaluator</td>
<td>dbsnp</td>
<td>input_0</td>
<td>none</td>
<td>novel</td>
<td>24</td>
<td>11</td>
<td>2.18</td>
<td>1398017</td>
<td>692557</td>
<td>2.02</td>
</tr>
</tbody>
</table>
PRACTICAL!
Appendix: Phred Score

\[ Q = -10 \log_{10} P \quad \text{and} \quad P = 10^{\frac{-Q}{10}} \]

<table>
<thead>
<tr>
<th>Phred Quality Score</th>
<th>Probability of incorrect base call</th>
<th>Base call accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1000</td>
<td>99.9%</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10,000</td>
<td>99.99%</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100,000</td>
<td>99.999%</td>
</tr>
<tr>
<td>60</td>
<td>1 in 1,000,000</td>
<td>99.9999%</td>
</tr>
</tbody>
</table>
Appendix: filters

• Note that the InbreedingCoeff statistic is a population-level calculation that is only available with 10 or more samples. If you have fewer samples you will need to omit that particular filter statement.

• **For shallow-coverage (<10x): you cannot use filtering to reliably separate true positives from false positives. You must use the protocol involving variant quality score recalibration.**

• The maximum DP (depth) filter only applies to whole genome data, where the probability of a site having exactly N reads given an average coverage of M is a well-behaved function. First principles suggest this should be a binomial sampling but in practice it is more a Gaussian distribution. Regardless, the DP threshold should be set a 5 or 6 sigma from the mean coverage across all samples, so that the DP > X threshold eliminates sites with excessive coverage caused by alignment artifacts. Note that **for exomes, a straight DP filter shouldn't be used** because the relationship between misalignments and depth isn't clear for capture data.

• That said, all of the caveats about determining the right parameters, etc, are annoying and are largely eliminated by variant quality score recalibration.

• [https://www.broadinstitute.org/gatk/guide/article?id=3225](https://www.broadinstitute.org/gatk/guide/article?id=3225)