Workflow

Raw reads (Fastq) → Mapping on the reference genome → Mapping Post-processing → Variant Calling Pre-processing → Variant Calling → Variant Filtering & Annotation
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
- ...
Howto?

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=DB,Number=0,Type=Flag,Description="dbSNP Membership">
##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">
...
#contig=<id=chr12,length=133851895>
#reference=file://tmp/13905.1.galaxy.q/tmp-gatk-MPG57G/gatk_input.fasta

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT Pickrell
Chr12 406292 rs2229351 G A 994.77 . 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 0/1:33,36:66:99:1023,0,994
Chr12 416046 rs35042439 C CT 391.73 . 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 0/1:22,17:46:99:429,0,521
```
# Tools

<table>
<thead>
<tr>
<th>tool</th>
<th>suited to</th>
<th>link</th>
<th>Galaxy availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>VcfTools</td>
<td>-</td>
<td><a href="https://vcftools.github.io/index.html">https://vcftools.github.io/index.html</a></td>
<td>limited</td>
</tr>
<tr>
<td>vcflib</td>
<td>FreeBayes</td>
<td><a href="https://github.com/ekg/vcflib">https://github.com/ekg/vcflib</a></td>
<td>good</td>
</tr>
<tr>
<td>GATK</td>
<td>GATK</td>
<td><a href="https://software.broadinstitute.org/gatk/">https://software.broadinstitute.org/gatk/</a></td>
<td>good</td>
</tr>
</tbody>
</table>
Methods

Remove variant entry (Hard Filtering) or add Filter info (Soft 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>
</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>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>AC=1;AF=0.500;AN=2;DB;DP=46;FS=0.000;MQ=27.49;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>

Before Filtering

After Filtering

Filter : MQ < 30.0

Filter : **PASS** (Not Filtered), . (no data, filtering not performed), **LowQual** (Filtered)

By default, filtering is applied at variant level (INFO), genotype filtering (FORMAT) possible
Howto?

- Understand VCF Format File
- Identify specific tags
- Fix Thresholds
- Find external ressources (dbSNP) to exclude / keep known Variants (other VCF File)
- Limit analysis to specific genomic locations (BED File)
It depends of:

- 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
**Criteria?**

**Depth (DP): Min / Max => d ± 4√d, d = average Read Depth**

- 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

---

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*
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
Li, H. (2014). *Bioinformatics*  
misc = AB, DS, FS
“The key difference between a regular VCF and a gVCF is that the gVCF has records for all sites, whether there is a variant call there or not. The goal is to have every site represented in the file in order to do joint analysis of a cohort in subsequent steps. The records in a gVCF include an accurate estimation of how confident we are in the determination that the sites are homozygous-reference or not.”

“The term GVCF is sometimes used simply to describe VCFs that contain a record for every position in the genome (or interval of interest) regardless of whether a variant was detected at that site or not.”

Pros:
- Keep trace of rare variants
- Non-covered genome regions (structural variant)
- Addition of new samples

Cons:
- Computational time
- File size

https://software.broadinstitute.org/gatk/guide/article?id=4017
https://sites.google.com/site/gvcftools/home/about-gvcf/gvcf-conventions
genome VCF: gVCF and all sites file

* Some tools may output an all-sites VCF that looks like what you can get using HC with -ERC BP_RESOLUTION but they do not provide an accurate estimate of reference confidence.

https://software.broadinstitute.org/gatk/guide/article?id=4017
Tutorial

GATK : Variant Filtration

GATK : Select Variants

GATK : Combine Variants

Exome-Seq (GATK calling)

Apply filters on GATK available tags

Extract filtered variants

Combine / Merge results in one file

Extract union/intersection/specific variants

RNA-Seq (Varscan calling)

Apply filters on Varscan available tags

Extract filtered variants
GATK Tools

Variant Filtration:

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

Select Variants:

- Direct selection (exclude filtered variants = Hard Filtering)
- 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)
GATK Tools

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
Tutorial
## 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
Appendix: 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)

http://gatkforums.broadinstitute.org/discussion/3225/how-can-i-filter-my-callset-if-i-cannot-use-vqsr-recalebrate-variants
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)