## Variant Calling Tools:

<table>
<thead>
<tr>
<th>Caller</th>
<th>Version</th>
<th>Parameters</th>
</tr>
</thead>
</table>
| VarDict (Java)  | 1.5.5   | **vardict**  
|                 |         |  -C (indicate chromosomes by numbers)  
|                 |         |  -f 0.01 (threshold for allele frequency)  
|                 |         |  -h (print a header row)  
|                 |         |  -c 1 (column for chromosome)  
|                 |         |  -S 2 (column for region start)  
|                 |         |  -E 3 (column for region end)  
|                 |         |  -g 4 (column for gene name) |
| LoFreq          | 2.1.2   | **lofreq** call  
|                 |         |  --call-indels  |
| GATK*           | 3.5     | **gatk**  --T BaseRecalibrator  
|                 |         |  --maximum_cycle_value 1500  
|                 |         |  --covariates:  
|                 |         |  ContextCovariate, CycleCovariate, QualityScoreCovariate, ReadGroupCovariate  
|                 |         |  --knownSites:  
|                 |         |  dbSNP 138.b37.vcf, Mills_and_1000G_gold_standard.indels.b37.vcf,  
|                 |         |  1000G_phase1.indels.b37.vcf  
|                 |         |  -nct 1  
|                 |         | **gatk**  --T PrintReads  
|                 |         |  --BQSR  
|                 |         | **gatk**  --T HaplotypeCaller  
|                 |         |  --standard_min_confidence_threshold_for_calling 30.0  
|                 |         |  --standard_min_confidence_threshold_for_emitting 10.0  
|                 |         |  --downsample_to_coverage 1500  
|                 |         |  --max_alternate_alleles 9  
|                 |         |  --dbsnp dbsnp_129.b37.vcf  
|                 |         |  --num_cpu_threads_per_data_thread 1  |
| samtools        | 1.3     | **samtools** mpileup  
|                 |         |  --min-MQ 1  
|                 |         |  --BCF  
|                 |         |  --uncompressed  
|                 |         |  --output $(output).bcf $(input).bam  
|                 |         | **bcftools** call  
|                 |         |  --variants-only  
|                 |         |  --multiallelic-caller  
|                 |         |  --output-type v (uncompressed vcf)  
|                 |         |  --output $(output).vcf $(input).bcf  |
| VarScan         | 2.4.0   | **samtools** mpileup --output $(output).bcf  
|                 |         | **varsan** mpileup2snp $(input).bcf  
|                 |         | **varsan** mpileup2indel $(input).bcf  |
| FreeBayes       | 1.0.2-6 | **freebayes**  
|                 |         |  --min-alternate-fraction 0.01  |
| SNVer           | 0.5.3   | **snver**  
|                 |         |  -b 0.01 (discard locus with ratio of alt/ref below threshold)  |
| Platypus        | 0.8.1   | **platypus** callVariants  
|                 |         |  --filterDuplicates=0  
|                 |         |  --minFlank=0  |

* GATK is not delivered with the software due to license restrictions.
## Annotation databases

<table>
<thead>
<tr>
<th>Database</th>
<th>Version</th>
<th>Release date</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>COSMIC*</td>
<td>v86</td>
<td>2018-08-01</td>
<td><a href="https://cancer.sanger.ac.uk/cosmic">https://cancer.sanger.ac.uk/cosmic</a></td>
</tr>
<tr>
<td>1000 Genomes Project</td>
<td>Phase 3</td>
<td>2013-05-02</td>
<td><a href="http://www.internationalgenome.org">http://www.internationalgenome.org</a></td>
</tr>
<tr>
<td>ExAC</td>
<td>v0.3</td>
<td>2015-11-29</td>
<td><a href="http://exac.broadinstitute.org">http://exac.broadinstitute.org</a></td>
</tr>
<tr>
<td>PROVEAN scores (v1.1)</td>
<td>v1.1</td>
<td></td>
<td><a href="http://provean.jcvi.org">http://provean.jcvi.org</a></td>
</tr>
<tr>
<td>Transcripts and Exons for GRCh37</td>
<td></td>
<td>Mar 2014</td>
<td><a href="https://www.ensembl.org/index.html">https://www.ensembl.org/index.html</a></td>
</tr>
<tr>
<td>SNP eff</td>
<td>v4.2</td>
<td></td>
<td><a href="http://snpeff.sourceforge.net">http://snpeff.sourceforge.net</a></td>
</tr>
</tbody>
</table>

* COSMIC is not delivered with the software due to license restrictions.
**Basic filtering step**

Filter out variants with...

- Low BaseQuality of alternative allele (threshold: $15$)
  
  AND:
  
  (reads with reference allele are present) OR (BaseQuality of reference allele $>15$)

- BaseQuality of reference allele by $7$ higher than of alternative allele

- Low BaseQuality of reference allele ($\leq 15$)
  
  AND high number of reads with reference allele ($> 30$)
Calculation of artifact / polymorphism score

### Occurrence in other samples
- **Art**: No occurrence in any other sample (NrSamples = 1)
- **Art**: (no Hotspot) AND (same variant in > 50% of all samples) +2
- **Art**: Nr of samples with same variant > 3 +2 +1
- **Art**: and 90% of these samples have VAF > 0.85

### Allelic Frequency / Prediction
- **Poly**: NOT previous AND "unplausible" allelic frequency < 0 OR between 0.35 AND 0.65 OR > 0.85 +1
- **Poly**: NOT previous AND Provean score ≥ -1.5 +1
- **Poly**: Provean score ≥ -1.5 +1
- **Poly**: Provean score ≤ -4 -1 +1
- **Poly**: Variant Allelic Frequency (VAF) < 0.02 +2

### Type of Variant
- **Art**: "stop_gained" mutations (stop_gained suchen) -1
- **Art**: "inframe" mutations, but not "stop_gained" +1

### Insertions / Deletions
- **Art**: Variant is an insertion / deletion / complex indel +1
- **Art**: Different variants at same locus found in other samples? +1
- **Art**: VAF < 0.05 +1

### Strand Bias
- **Art**: Small StrandBias (p ≥ 0.001):
  - alternative on Forward strand ≤ 2 and reference ≥ 15 +1
  - alternative on Reverse strand ≤ 2 and reference ≥ 15 +1
- **Art**: Large StrandBias (p < 0.001):
  - alternative on both strands ≥ 10 -1
  - alternative on Forward strand ≤ 2 and reference < 15 -1
  - alternative on Reverse strand ≤ 2 and reference < 15 -1

### Callers
- **Art**: Variant found by only one caller +1
- **Art**: Variant found by 4 callers -1
- **Art**: Variant found by 5 callers -2
- **Art**: Variant found by ≥ 6 callers -3 +1
- **Art**: Called by LoFreq and FreeBayes and VarDict -3
Databases

Databases to be used:
- COSMIC *with not SNP and "haematopoietic and lymphoid tissue" > 20
- ClinVar *with significance rating as "(likely) pathogenic"
- dbSNP *with PM_flag or not in v129
- dbSNP v129 (wird in NrAnyDBs zweifach gezählt!)
- 1000Genomes, threshold: > 0.001
- ESP6500, threshold: > 0.03
- ExAC, threshold: > 0.000

Variant not present in any of the previous databases
- and VAF < 0.1 +1
- and same variant in > 50% of all samples +1
- Variant matching thresholds in no (non-clinical*) database -1
- Variant matching thresholds in 2 or 3 (non-clinical*) databases +1
- Variant matching thresholds in ≥ 4 (non-clinical*) databases +2
- Variant found in ≥ 2 disease associated DBs* -1
- Variant identified as "Precious mutation" (PM) by dbSNP -2

Known Hotspots
- Variant in a known hotspot mutation site -3
- Not a known hotspot, but "Precious mutation" by dbSNP -1

Finally: Exclude improbable polymorphisms

High Polymorphism score (≥ 2), no hotspot AND:
- VAF ≤ 0.1 +5
- VAF ≤ 0.2 +2
- frameshift_mutation +2

Classification

<table>
<thead>
<tr>
<th>Artifact</th>
<th>Artifact score ≥ 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>likely Polymorphism</td>
<td>(no hotspot, no frameshift and high VAF) AND Polymorphism score ≥ 2</td>
</tr>
<tr>
<td>Polymorphism</td>
<td>(no hotspot, no frameshift and high VAF) AND Polymorphism score ≥ 3 OR (Polymorphism score ≥ 2) AND (Cosmic NrHaemato ≤ 100)</td>
</tr>
<tr>
<td>Probably True</td>
<td>None of the above</td>
</tr>
</tbody>
</table>