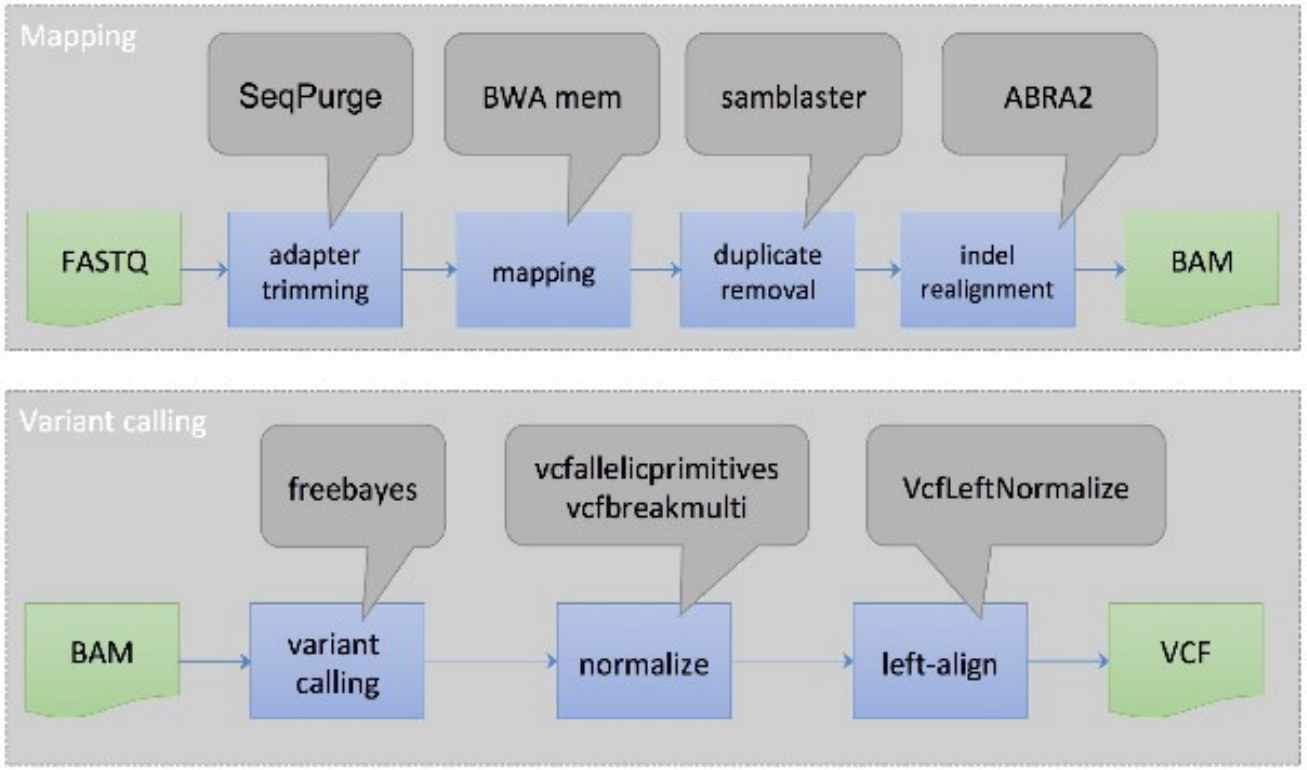
NGS Data-Analysis

Ablauf

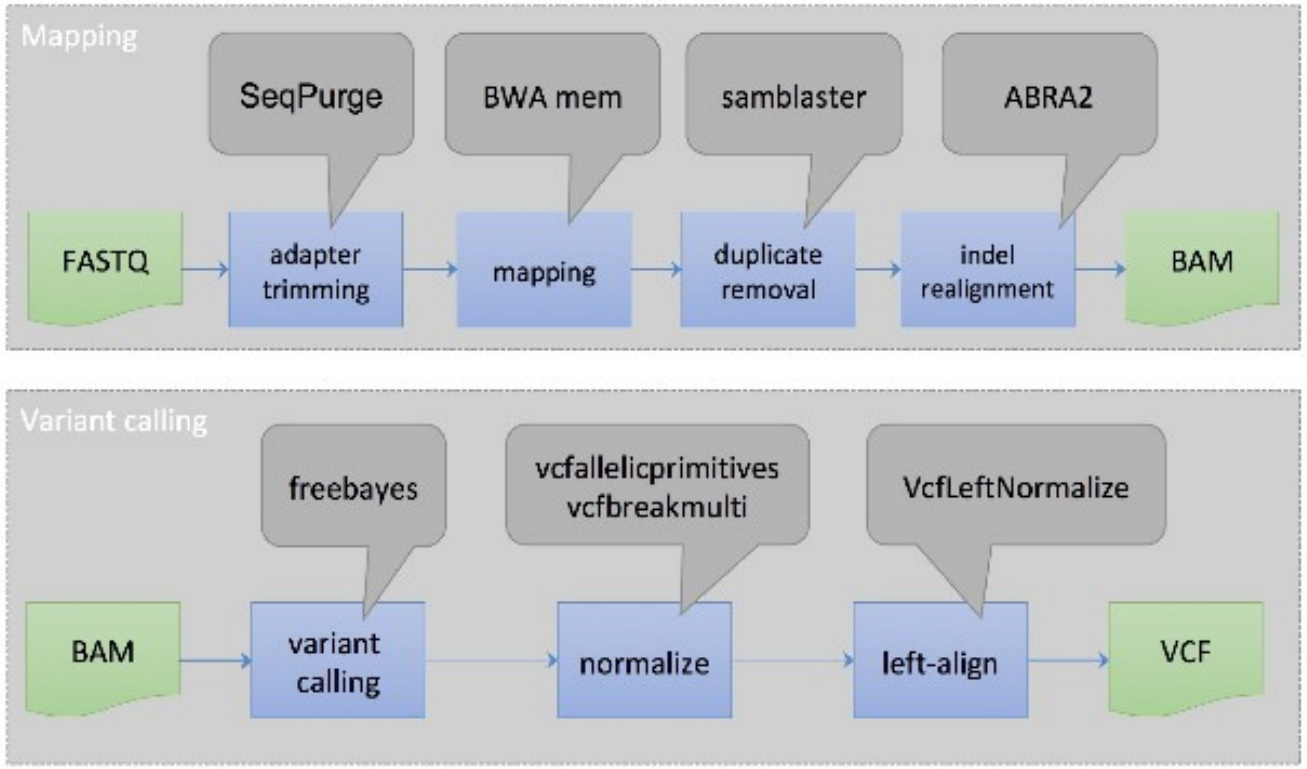

Mapping
SeqPurge
BWA mem
samblaster
ABRA2
FASTQ
adapter trimming
mapping
duplicate removal
indel realignment
BAM
Variant calling
freebayes
vcfallelicprimitives vcfbreakmulti
VcfLeftNormalize
BAM
variant calling
normalize
left-align
VCF

Daten-Typen

- FastQ
- Sam/Bam (Sequence / Binary Alignment Map)
- VCF (Variant Calling Format)

```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEE
```

```
#CHROM POS      ID         REF ALT    QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370    rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330    .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTCT   G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```


Mapping
SeqPurge
BWA mem
samblaster
ABRA2
FASTQ
adapter trimming
mapping
duplicate removal
indel realignment
BAM
Variant calling
freebayes
vcfallelicprimitives vcfbreakmulti
VcfLeftNormalize
BAM
variant calling
normalize
left-align
VCF

Adapter Trimming mit SeqPurge

- Enfernung der Adapter Sequencen
- Input: Forward-Read und Reverse-Read als FastQ-Dateien (gzipped)
- Output: 2 FastQ-Dateien (gzipped)

Mapping mit BWA mem

- Allignment der Reads auf ein Referenzgenom.
- Input: FastQ-Reads (gzipped) und Referenzgenom als Fasta (gzipped)
- Output: Allignment als SAM-Datei (gzipped)
- Zuerst Erstellung von FM-Index für Referenz-Genom

Duplicate Removal mit SamBlaster

- Entfernung von Duplikaten
- Input: SAM-Datei
- Output: SAM-Datei gzipped
- Zur Entfernung der Duplikate wird –removeDups Option verwendet

Indel Reallignment mit ABRA2

- Realignment zur Verbesserung der Genauigkeit
- Verbessert Genauigkeit besonders im Bereich von InDel’s
- Input: Allignment als Bam-File , Referenzgenom als Fasta
- Output: Bam-File

- Output von SamBlaster ist Sam Datei => Umwandlung zu Bam-Datei mit SamTools nötig

Variant Calling mit freebayes

- Finden von wahrscheinlichsten genetischen Varianten
- Input: Alignment BAM-Datei, Referenzgenom als Fasta
- Output: Genetische Varianten als VCF-Datei

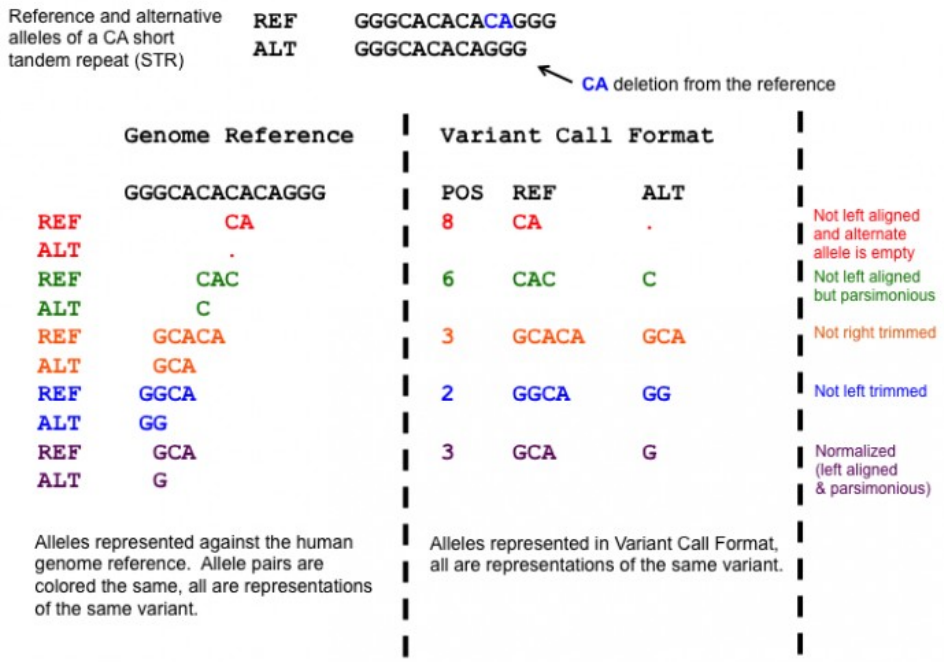
Normalizing mit vcfallelicprimitives/vcfbreakmulti

- Vereinheitlichung von Varianten (tab-delimiter) und aufteilen von Records die mehrere Allele enthalten.
- Chrom 55 <id> AT CG,AG

- Chrom 55 <id> AT CG
- Chrom 56 <id> C G

Left-Align mit vcfleftnormalize

- Verschiebung von Indel’s nach links. (Komplexe Indel’s werden übersprungen)
- Input: VCF und Referenzgenom (Fasta)
- Output:VCF



Zusammenfassung

- NGS Data Analysis sorgt für Aufbereitung, Alignment und Finden von Varianten.
- Erhöhung der Qualität und Vereinheitlichung der Daten um weitere Verwendung und lesbarkeit zu vereinfachen