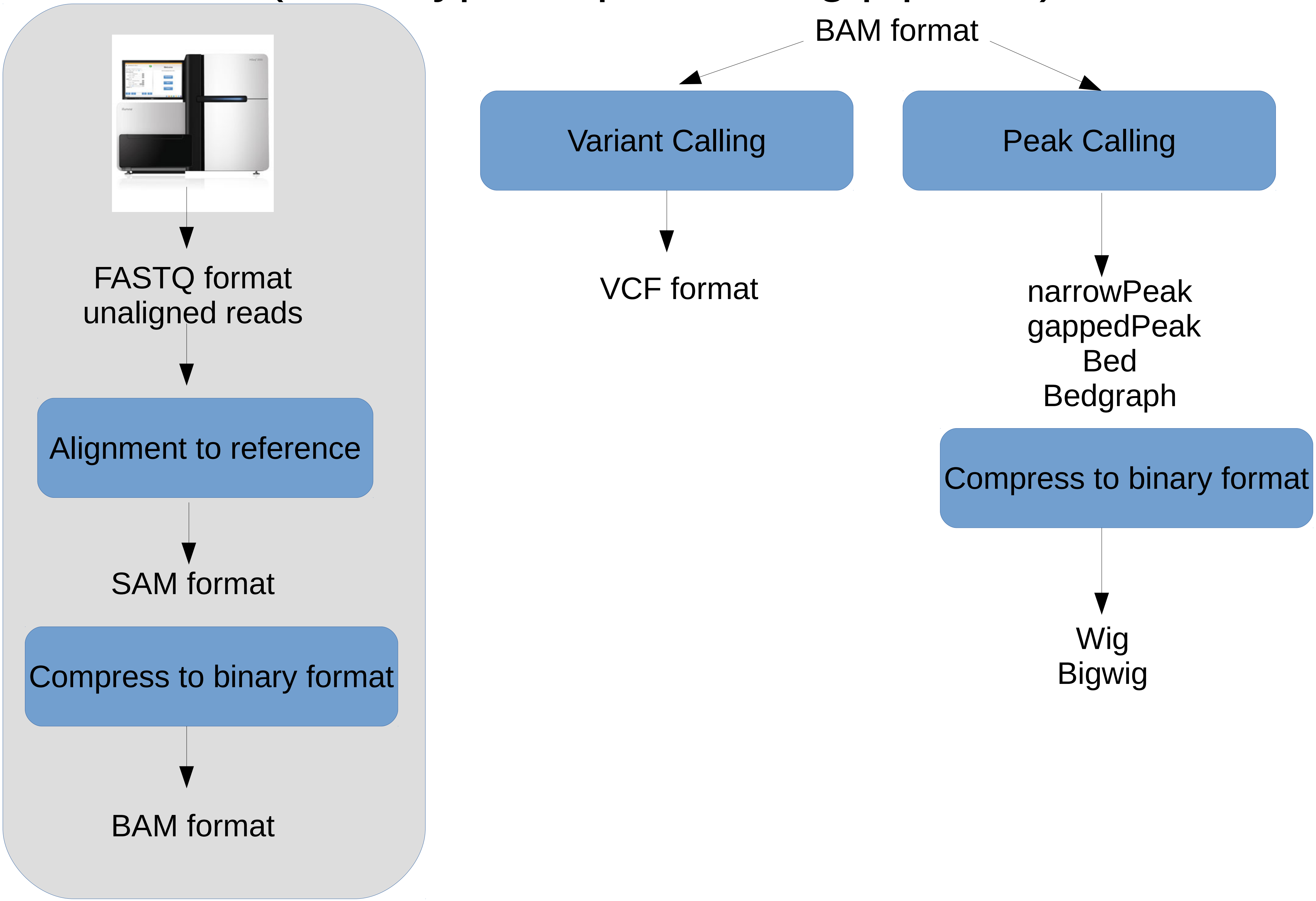
Data Formats Cheat Sheet
(for a “typical” processing pipeline)



FASTQ format
unaligned reads

Alignment to reference

SAM format

Compress to binary format

BAM format

BAM format

Variant Calling

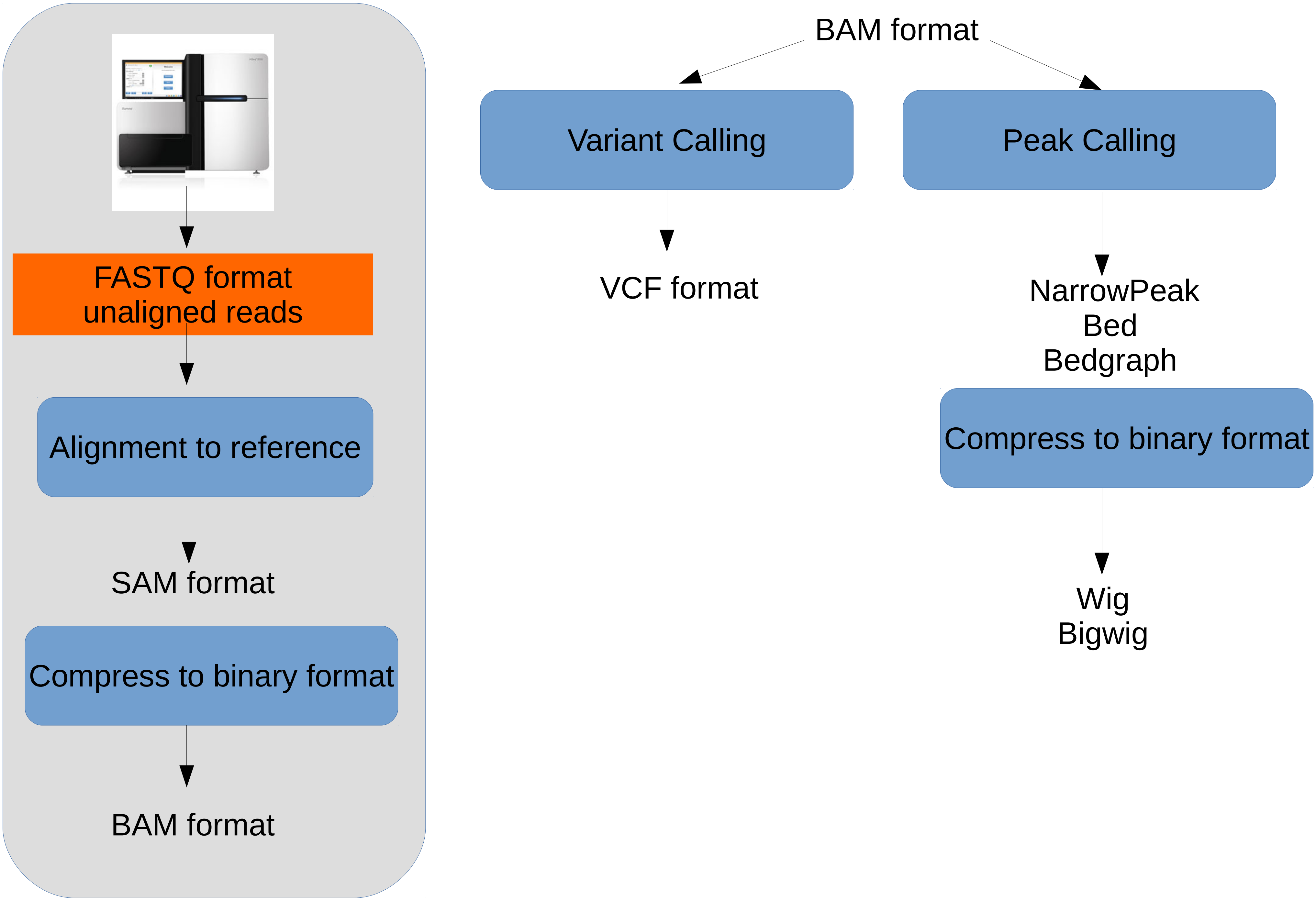
VCF format

Peak Calling

narrowPeak
gappedPeak
Bed
Bedgraph

Compress to binary format

Wig
Bigwig



FASTQ format
unaligned reads

Alignment to reference

SAM format

Compress to binary format

BAM format

BAM format

Variant Calling

VCF format

Peak Calling

NarrowPeak
Bed
Bedgraph

Compress to binary format

Wig
Bigwig

FASTA and FASTQ Files

FASTQ FORMAT:

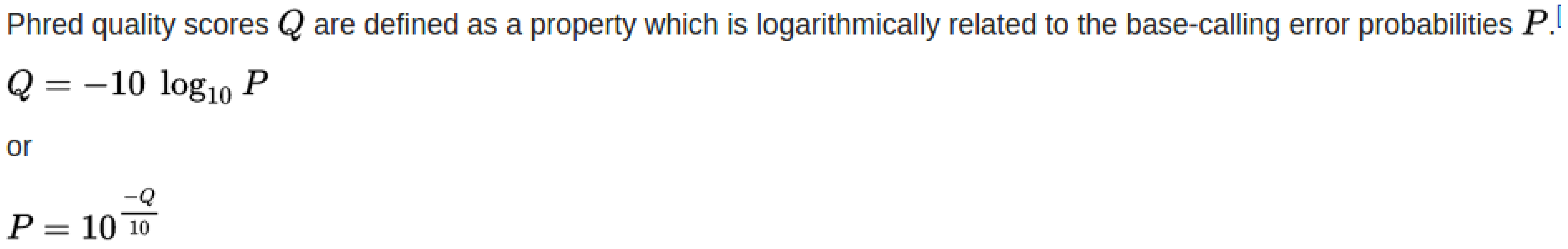
```
@J00118:203:HFL3VBBXX:4:1101:23399:1349 1:N:0:TAAGGCGA+TCTACTCT
GNTCTAGGGTGTAGCCTGAGAATAGGGGAAATCAGTGAATGCTGTCTCTTATACACATCTCCGAGCCCACGAGACT
+
A#AFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJ<FJJJJJJJJJJJJJJJJJJJJFJJJJJ
@J00118:203:HFL3VBBXX:4:1101:24292:1349 1:N:0:TAAGGCGA+TCTACTCT
GNCTGAGAGGGCCCCTGTTAGGGGTCATGGGCTGGGTTTTACTATATGATAGGCATGTGATTGGTGGGTCATTATG
+
A#AFFJJJJJJJJJJJJ<JJJJJJJJJJJJJJJJJJ<FJJFJFJJJFFFFJJFFJ<AFJJJJJJAFFJ<JJFJJJJ
```

- Read name
  - Preceded by @
  - Unique identifier for each read in the dataset

- Base sequence (A,T,C,G,N)

- + sign

- Phred quality score
  - ASCII encoding



Phred quality scores Q are defined as a property which is logarithmically related to the base-calling error probabilities P.[

$$Q = -10 \log_{10} P$$

or

$$P = 10^{\frac{-Q}{10}}$$

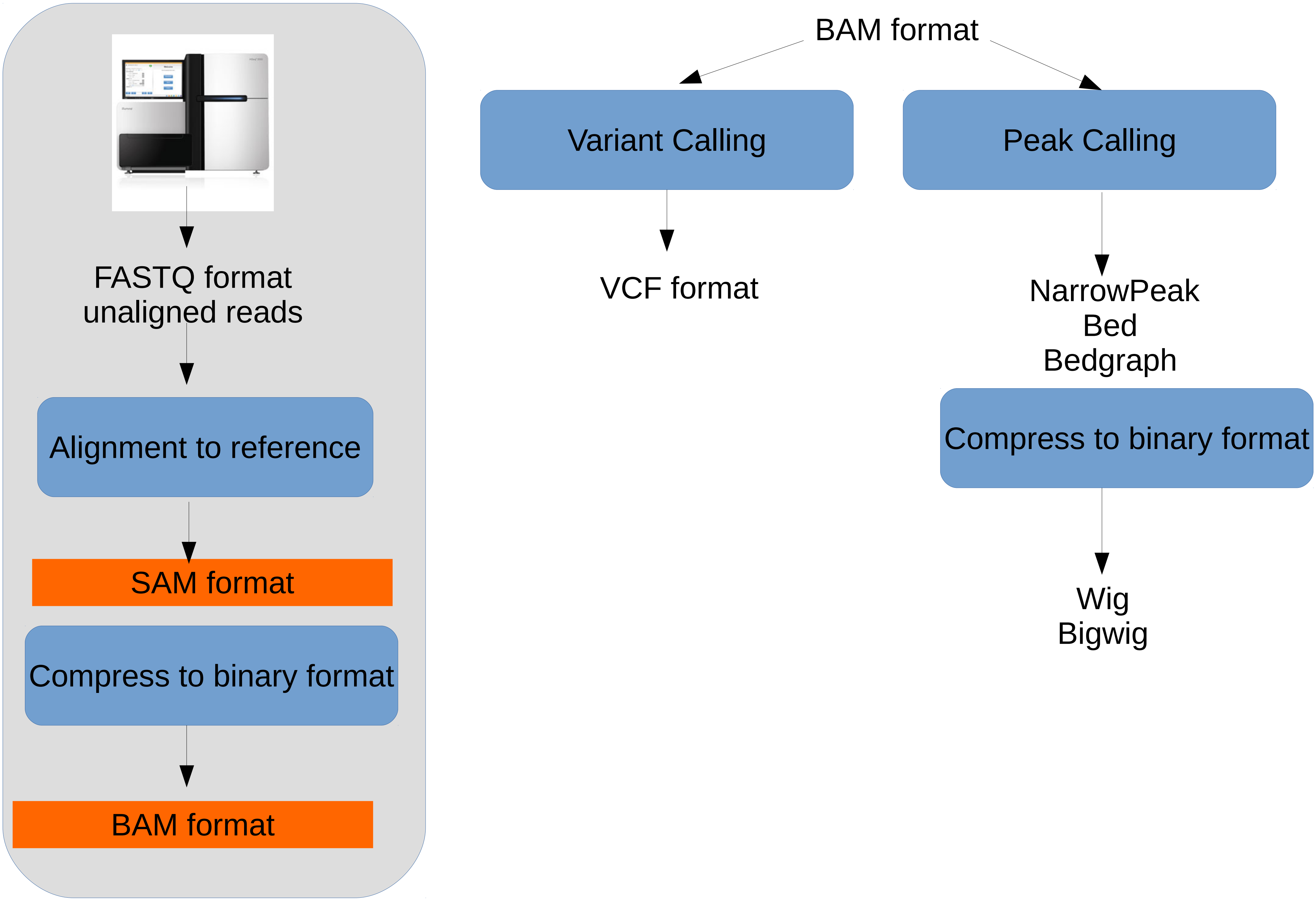
FASTA and FASTQ Files

FASTA FORMAT:

```
>J00118:203:HFL3VBBXX:4:1101:23399:1349 1:N:0:TAAGGCGA+TCTACTCT
GNTCTAGGGTGTAGCCTGAGAATAGGGGAAATCAGTGAATGCTGTCTCTTATACACATCTCCGAGCCCACGAGACT
>J00118:203:HFL3VBBXX:4:1101:24292:1349 1:N:0:TAAGGCGA+TCTACTCT
GNCTGAGAGGGCCCCTGTTAGGGGTCATGGGCTGGGTTTTACTATATGATAGGCATGTGATTGGTGGGTCATTATG
```

- Read name
  - Preceded by >
  - Unique identifier for each read in the dataset

- Base sequence (A,T,C,G,N)



FASTQ format
unaligned reads

Alignment to reference

SAM format

Compress to binary format

BAM format

BAM format

Variant Calling

VCF format

Peak Calling

NarrowPeak
Bed
Bedgraph

Compress to binary format

Wig
Bigwig

SAM and BAM format

- Alignment algorithms such as BWA and Bowtie2 will accept input data in FASTA format and will generate aligned output in SAM or BAM format.

- BAM format stores data in a compressed, indexed, binary format.

- SAM is the human readable version of BAM – **never store data in SAM format! It is much larger than BAM and contains the same information.**

- Use the **samtools** program to work with bam files – sort, index, look for variants, and so much more
  - **http://www.htslib.org/doc/samtools.html**

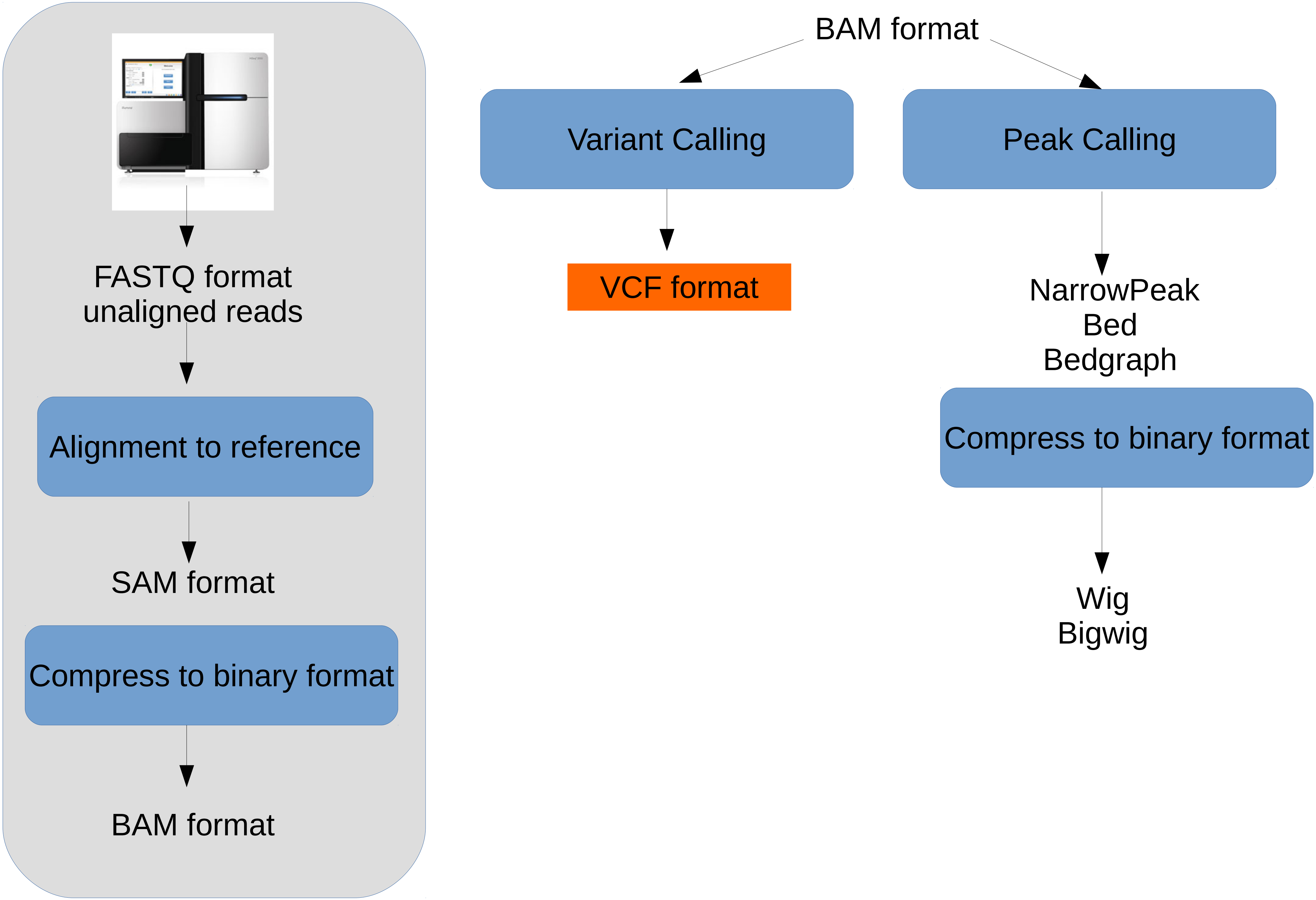
Samtools is your friend!
(essential commands for viewing/working with (s/b)am files)

```
samtools view -bt ref_list.txt -o aln.bam aln.sam.gz

samtools sort -T /tmp/aln.sorted -o aln.sorted.bam aln.bam

samtools index aln.sorted.bam

samtools stats aln.sorted.bam

samtools bedcov aln.sorted.bam

samtools depth aln.sorted.bam

samtools view aln.sorted.bam chr2:20,100,000-20,200,000
```

So what's actually in a bam (sam) file?

“Row” for a single read:

J00118:203:HFL3VBBXX:4:2205:13626:34495	Read name
99	a bitwise set of information describing the alignment, FLAG
chr10	Chromosome where the read aligned
60573	Starting position of alignment
1	Phred-scaled quality score
76M	
=	
60621	
124	
GTTATTAGATGATTCAAATATGAAGTGCTGTTATGCCAAACAATGAATCTTTGTGTTATACA	Sequence
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ	Quality
AS:i:-6	
XS:i:-6	
XN:i:0	
XM:i:1	
XO:i:0	
XG:i:0	
NM:i:1	
MD:Z:24A51	
YS:i:0	
YT:Z:CP	



FASTQ format
unaligned reads

Alignment to reference

SAM format

Compress to binary format

BAM format

BAM format

Variant Calling

VCF format

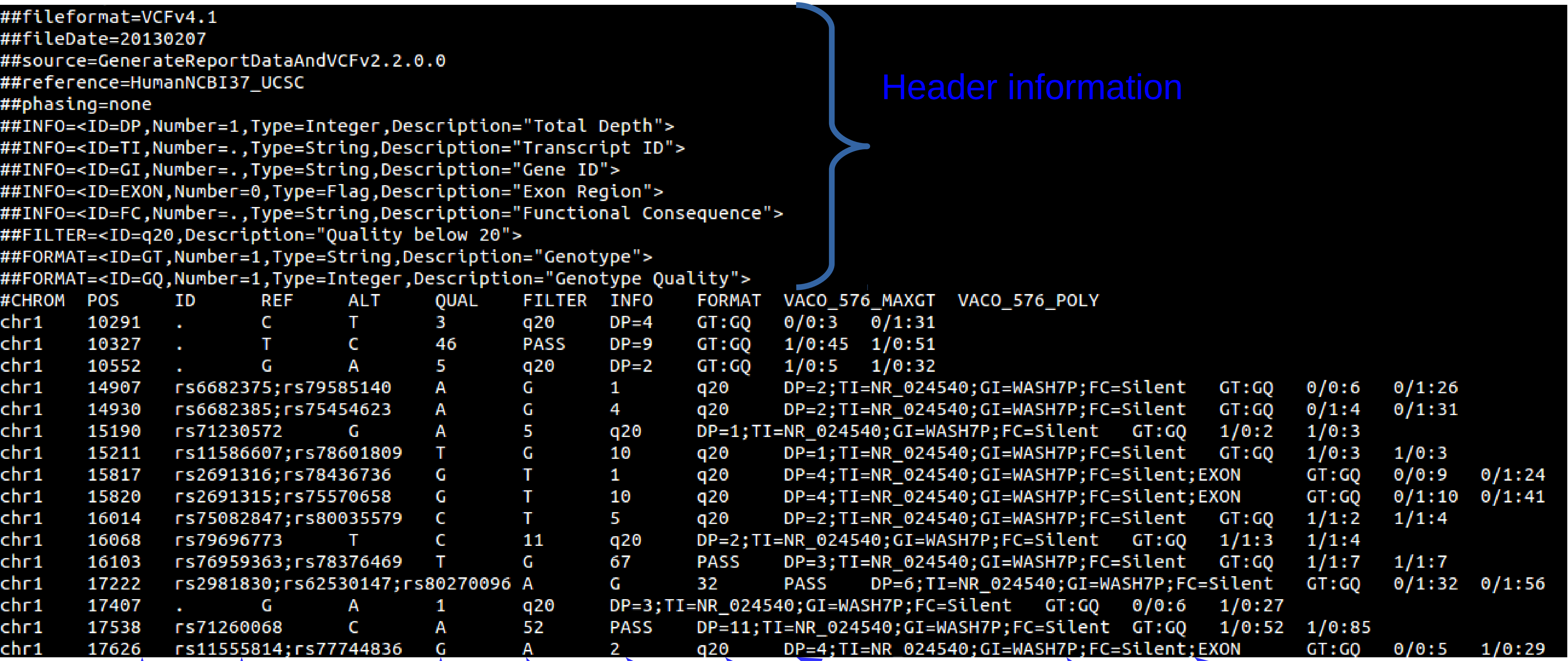
Peak Calling

NarrowPeak
Bed
Bedgraph

Compress to binary format

Wig
Bigwig

VCF
(Variant Call Format)



```
##fileformat=VCFv4.1
##fileDate=20130207
##source=GenerateReportDataAndVCFv2.2.0.0
##reference=HumanNCBI37_UCSC
##phasing=none
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=TI,Number=.,Type=String,Description="Transcript ID">
##INFO=<ID=GI,Number=.,Type=String,Description="Gene ID">
##INFO=<ID=EXON,Number=0,Type=Flag,Description="Exon Region">
##INFO=<ID=FC,Number=.,Type=String,Description="Functional Consequence">
##FILTER=<ID=q20,Description="Quality below 20">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  VACO_576_MAXGT  VACO_576_POLY
chr1    10291   .       C       T       3       q20     DP=4    GT:GQ   0/0:3   0/1:31
chr1    10327   .       T       C       46      PASS    DP=9    GT:GQ   1/0:45  1/0:51
chr1    10552   .       G       A       5       q20     DP=2    GT:GQ   1/0:5   1/0:32
chr1    14907   rs6682375;rs79585140    A       G       1       q20     DP=2;TI=NR_024540;GI=WASH7P;FC=Silent   GT:GQ   0/0:6   0/1:26
chr1    14930   rs6682385;rs75454623    A       G       4       q20     DP=2;TI=NR_024540;GI=WASH7P;FC=Silent   GT:GQ   0/1:4   0/1:31
chr1    15190   rs71230572      G       A       5       q20     DP=1;TI=NR_024540;GI=WASH7P;FC=Silent   GT:GQ   1/0:2   1/0:3
chr1    15211   rs11586607;rs78601809   T       G       10      q20     DP=1;TI=NR_024540;GI=WASH7P;FC=Silent   GT:GQ   1/0:3   1/0:3
chr1    15817   rs2691316;rs78436736    G       T       1       q20     DP=4;TI=NR_024540;GI=WASH7P;FC=Silent;EXON      GT:GQ   0/0:9   0/1:24
chr1    15820   rs2691315;rs75570658    G       T       10      q20     DP=4;TI=NR_024540;GI=WASH7P;FC=Silent;EXON      GT:GQ   0/1:10  0/1:41
chr1    16014   rs75082847;rs80035579   C       T       5       q20     DP=2;TI=NR_024540;GI=WASH7P;FC=Silent   GT:GQ   1/1:2   1/1:4
chr1    16068   rs79696773      T       C       11      q20     DP=2;TI=NR_024540;GI=WASH7P;FC=Silent   GT:GQ   1/1:3   1/1:4
chr1    16103   rs76959363;rs78376469   T       G       67      PASS    DP=3;TI=NR_024540;GI=WASH7P;FC=Silent   GT:GQ   1/1:7   1/1:7
chr1    17222   rs2981830;rs62530147;rs80270096 A       G       32      PASS    DP=6;TI=NR_024540;GI=WASH7P;FC=Silent   GT:GQ   0/1:32  0/1:56
chr1    17407   .       G       A       1       q20     DP=3;TI=NR_024540;GI=WASH7P;FC=Silent   GT:GQ   0/0:6   1/0:27
chr1    17538   rs71260068      C       A       52      PASS    DP=11;TI=NR_024540;GI=WASH7P;FC=Silent  GT:GQ   1/0:52  1/0:85
chr1    17626   rs11555814;rs77744836   G       A       2       q20     DP=4;TI=NR_024540;GI=WASH7P;FC=Silent;EXON      GT:GQ   0/0:5   1/0:29
```

Header information

chromosome

position

variant id (if known)

Reference allele

Alternate allele

Quality score

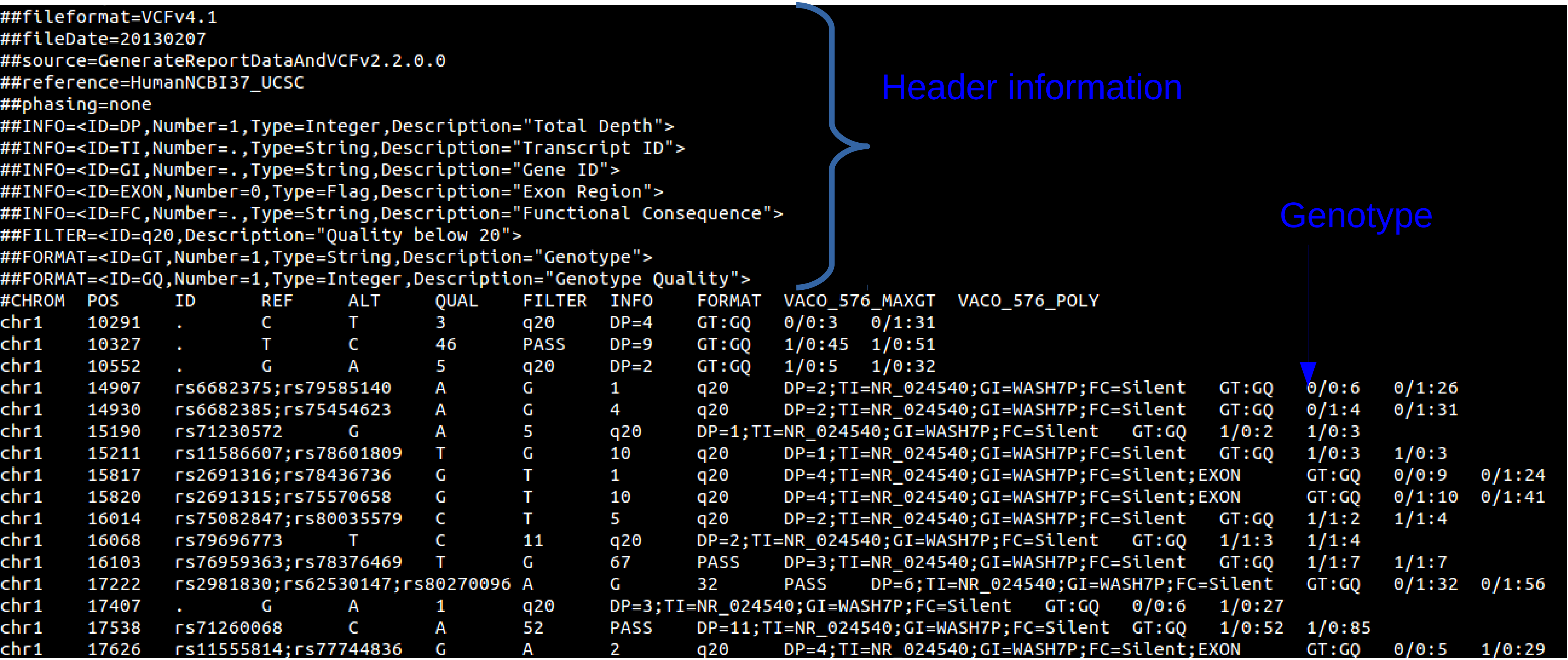
Quality threshold used to filter variants

Depth – how many reads aligned to the position

Transcript ID/Gene ID

Functional consequence

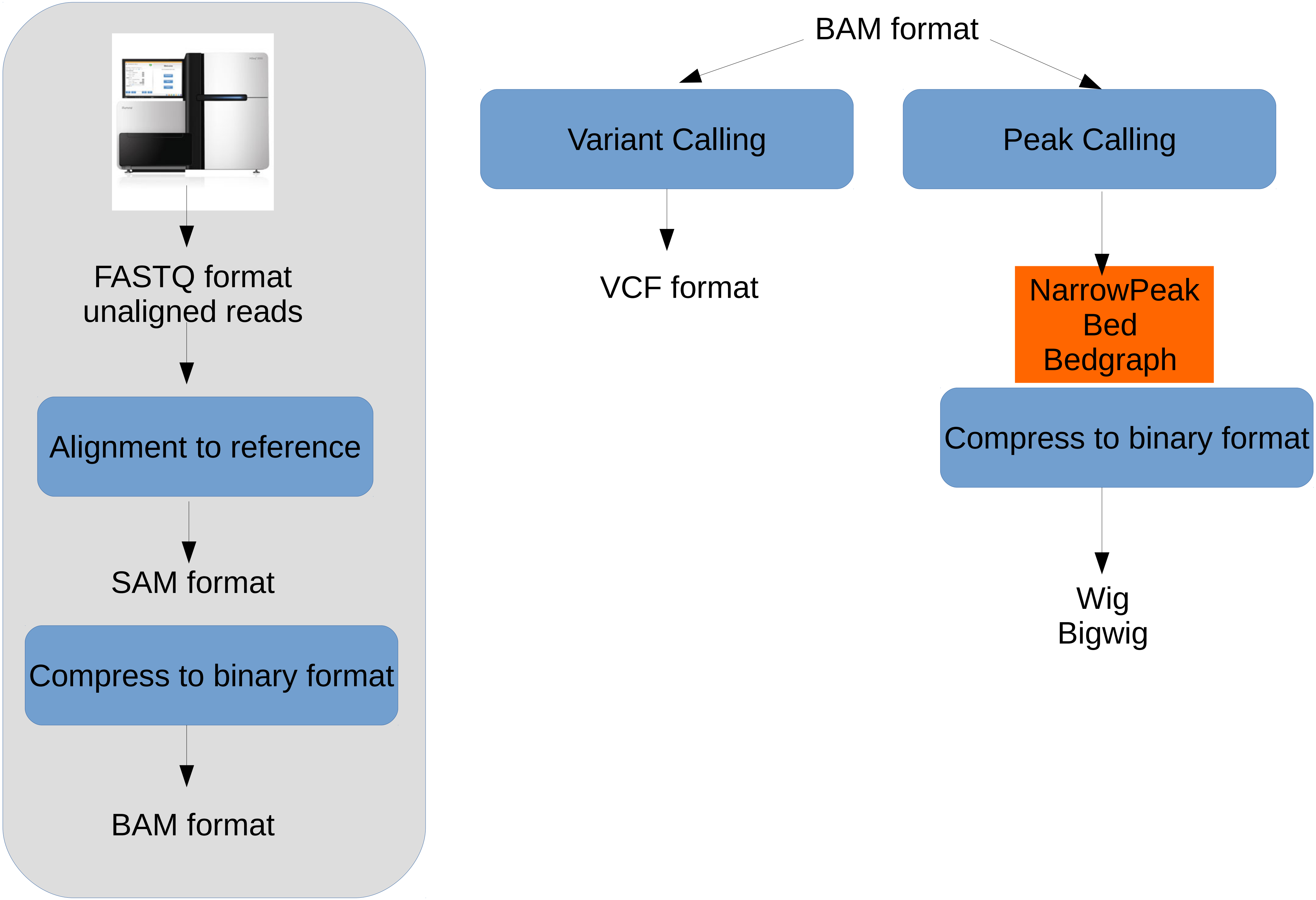
VCF
(Variant Call Format)



```
##fileformat=VCFv4.1
##fileDate=20130207
##source=GenerateReportDataAndVCFv2.2.0.0
##reference=HumanNCBI37_UCSC
##phasing=none
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=TI,Number=.,Type=String,Description="Transcript ID">
##INFO=<ID=GI,Number=.,Type=String,Description="Gene ID">
##INFO=<ID=EXON,Number=0,Type=Flag,Description="Exon Region">
##INFO=<ID=FC,Number=.,Type=String,Description="Functional Consequence">
##FILTER=<ID=q20,Description="Quality below 20">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  VACO_576_MAXGT  VACO_576_POLY
chr1    10291   .       C       T       3       q20     DP=4    GT:GQ   0/0:3   0/1:31
chr1    10327   .       T       C       46      PASS    DP=9    GT:GQ   1/0:45  1/0:51
chr1    10552   .       G       A       5       q20     DP=2    GT:GQ   1/0:5   1/0:32
chr1    14907   rs6682375;rs79585140    A       G       1       q20     DP=2;TI=NR_024540;GI=WASH7P;FC=Silent   GT:GQ   0/0:6   0/1:26
chr1    14930   rs6682385;rs75454623    A       G       4       q20     DP=2;TI=NR_024540;GI=WASH7P;FC=Silent   GT:GQ   0/1:4   0/1:31
chr1    15190   rs71230572      G       A       5       q20     DP=1;TI=NR_024540;GI=WASH7P;FC=Silent   GT:GQ   1/0:2   1/0:3
chr1    15211   rs11586607;rs78601809   T       G       10      q20     DP=1;TI=NR_024540;GI=WASH7P;FC=Silent   GT:GQ   1/0:3   1/0:3
chr1    15817   rs2691316;rs78436736    G       T       1       q20     DP=4;TI=NR_024540;GI=WASH7P;FC=Silent;EXON      GT:GQ   0/0:9   0/1:24
chr1    15820   rs2691315;rs75570658    G       T       10      q20     DP=4;TI=NR_024540;GI=WASH7P;FC=Silent;EXON      GT:GQ   0/1:10  0/1:41
chr1    16014   rs75082847;rs80035579   C       T       5       q20     DP=2;TI=NR_024540;GI=WASH7P;FC=Silent   GT:GQ   1/1:2   1/1:4
chr1    16068   rs79696773      T       C       11      q20     DP=2;TI=NR_024540;GI=WASH7P;FC=Silent   GT:GQ   1/1:3   1/1:4
chr1    16103   rs76959363;rs78376469   T       G       67      PASS    DP=3;TI=NR_024540;GI=WASH7P;FC=Silent   GT:GQ   1/1:7   1/1:7
chr1    17222   rs2981830;rs62530147;rs80270096 A       G       32      PASS    DP=6;TI=NR_024540;GI=WASH7P;FC=Silent   GT:GQ   0/1:32  0/1:56
chr1    17407   .       G       A       1       q20     DP=3;TI=NR_024540;GI=WASH7P;FC=Silent   GT:GQ   0/0:6   1/0:27
chr1    17538   rs71260068      C       A       52      PASS    DP=11;TI=NR_024540;GI=WASH7P;FC=Silent  GT:GQ   1/0:52  1/0:85
chr1    17626   rs11555814;rs77744836   G       A       2       q20     DP=4;TI=NR_024540;GI=WASH7P;FC=Silent;EXON      GT:GQ   0/0:5   1/0:29
```

0/0 → 2 reference alleles
0/1 → 1 reference allele, 1 alternate allele
1/1 → 2 alternate alleles

MAXGT – genotype compared to reference sequence (i.e. hg19)
POLY – genotype for a polymorphic site



FASTQ format
unaligned reads

Alignment to reference

SAM format

Compress to binary format

BAM format

BAM format

Variant Calling

VCF format

Peak Calling

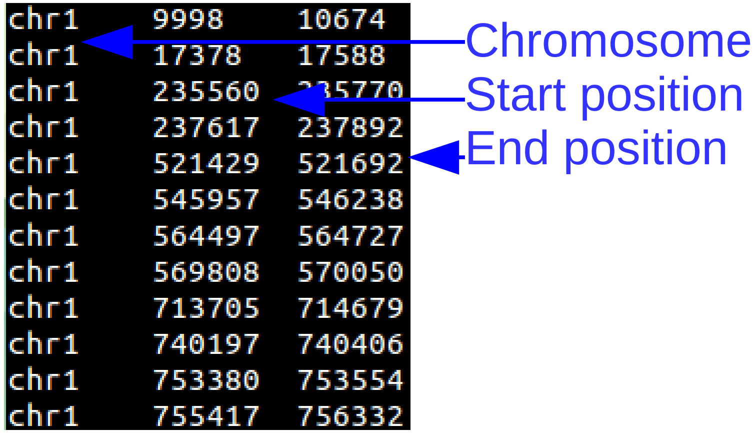
NarrowPeak
Bed
Bedgraph

Compress to binary format

Wig
Bigwig

BED files (and variations)

Bed files denote genome regions/ positions of interest



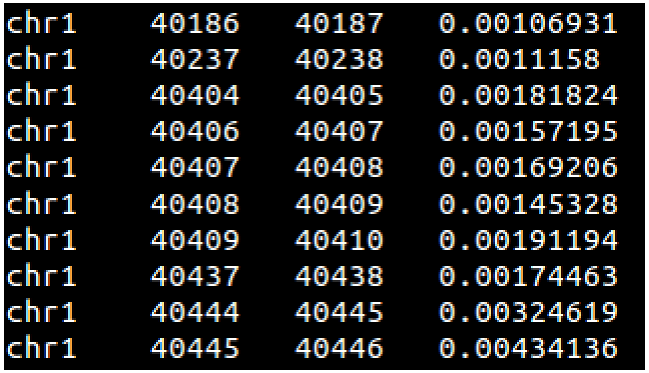
```
chr1     9998     10674
chr1     17378    17588
chr1     235560   235770
chr1     237617   237892
chr1     521429   521692
chr1     545957   546238
chr1     564497   564727
chr1     569808   570050
chr1     713705   714679
chr1     740197   740406
chr1     753380   753554
chr1     755417   756332
```

Chromosome

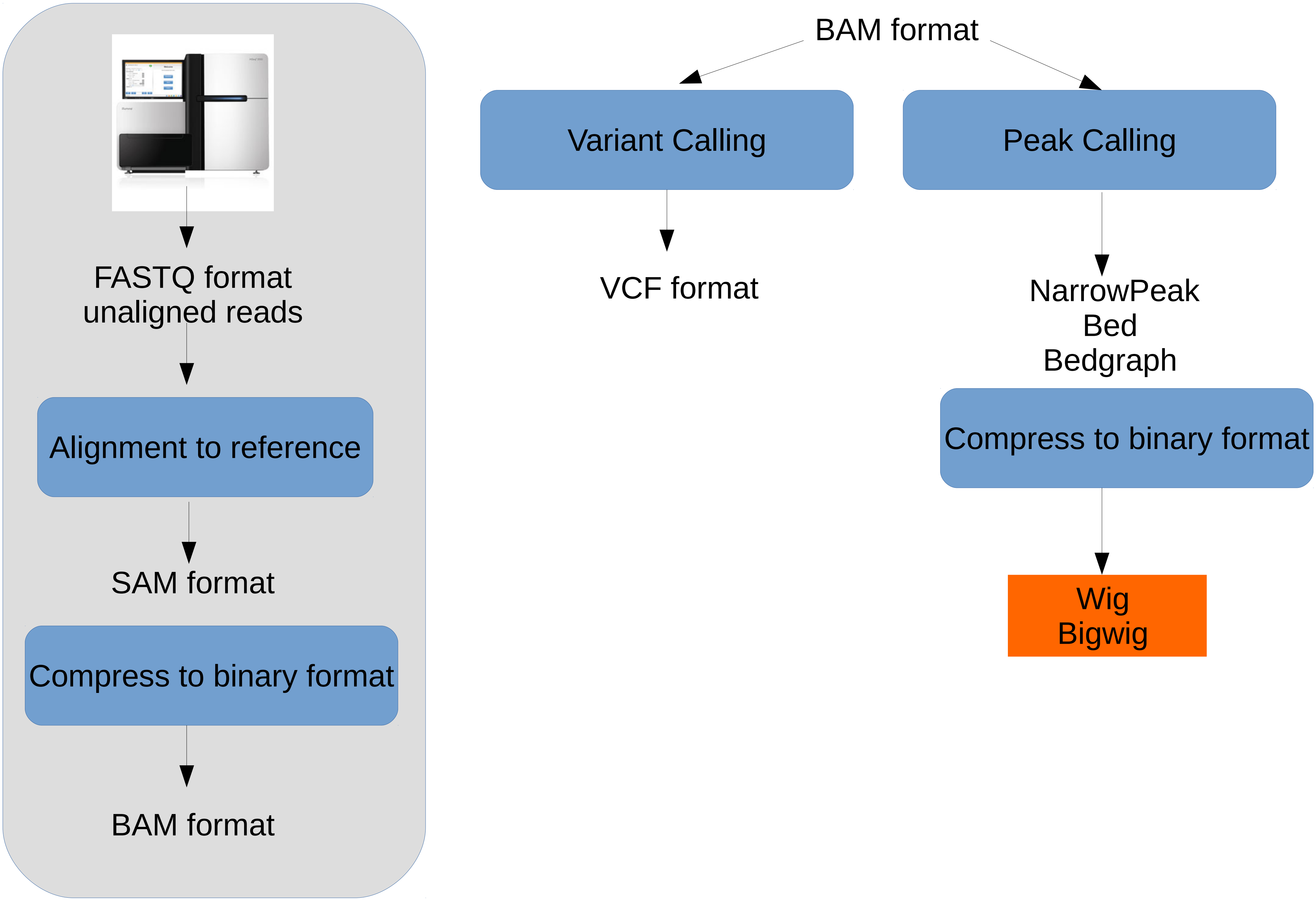
Start position

End position

Bedgraph files are typically used to denote genome coverage in an added 4th column



```
chr1     40186    40187    0.00106931
chr1     40237    40238    0.0011158
chr1     40404    40405    0.00181824
chr1     40406    40407    0.00157195
chr1     40407    40408    0.00169206
chr1     40408    40409    0.00145328
chr1     40409    40410    0.00191194
chr1     40437    40438    0.00174463
chr1     40444    40445    0.00324619
chr1     40445    40446    0.00434136
```



FASTQ format
unaligned reads

Alignment to reference

SAM format

Compress to binary format

BAM format

BAM format

Variant Calling

VCF format

Peak Calling

NarrowPeak
Bed
Bedgraph

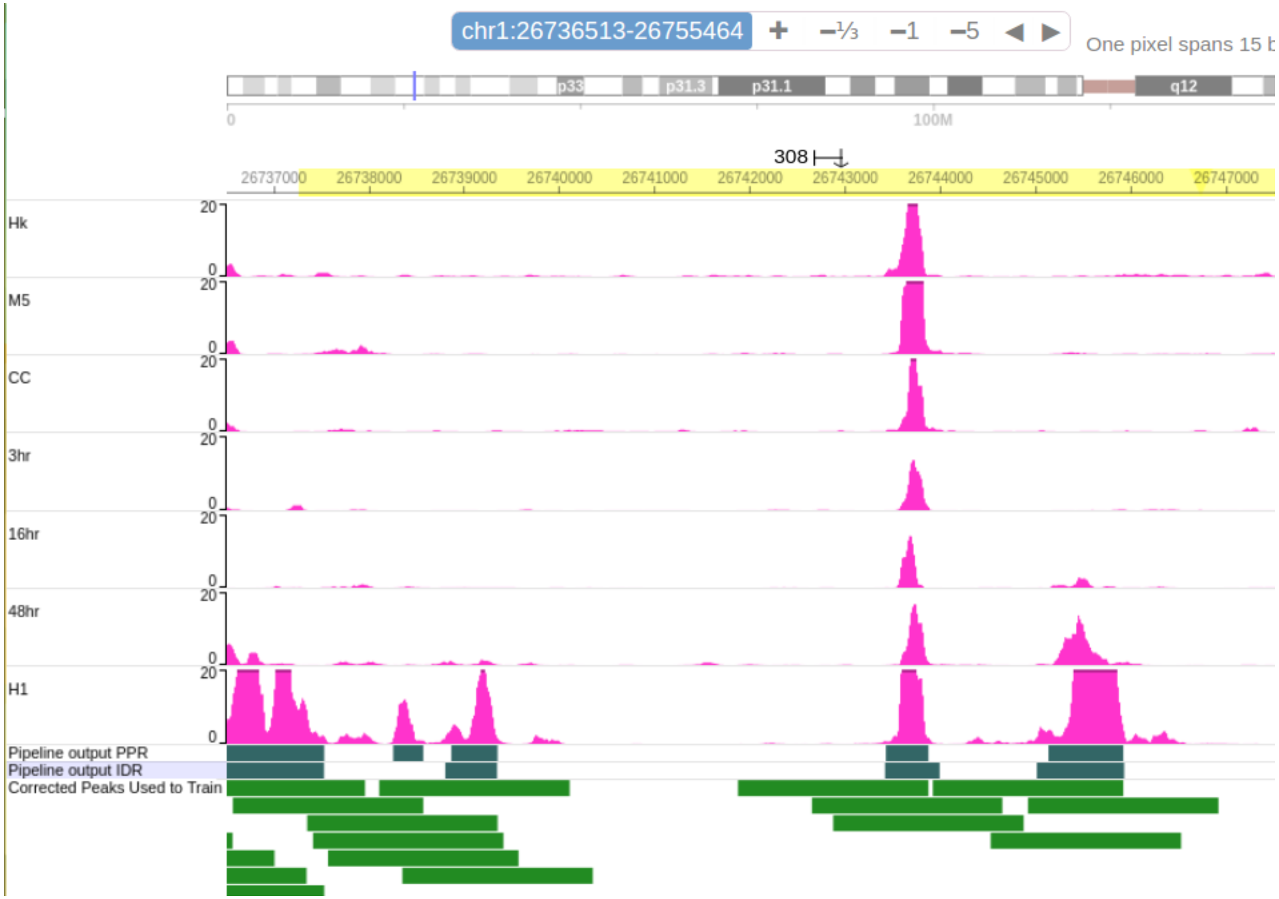
Compress to binary format

Wig
Bigwig

Browser Track Formats: BigWig and Hammock

- **Allow for visualization of large datasets in a browser such as UCSC or Washu**

- **BigWigs are binary files**
- **Hammock tracks are human-readable**
- **Washu browser demo**


chr1:26736513-26755464
+
−⅓
−1
−5
◀
▶
One pixel spans 15 b
p33
p31.3
p31.1
q12
0
100M
308
26737000
26738000
26739000
26740000
26741000
26742000
26743000
26744000
26745000
26746000
26747000
Hk
M5
CC
3hr
16hr
48hr
H1
20
0
Pipeline output PPR
Pipeline output IDR
Corrected Peaks Used to Train