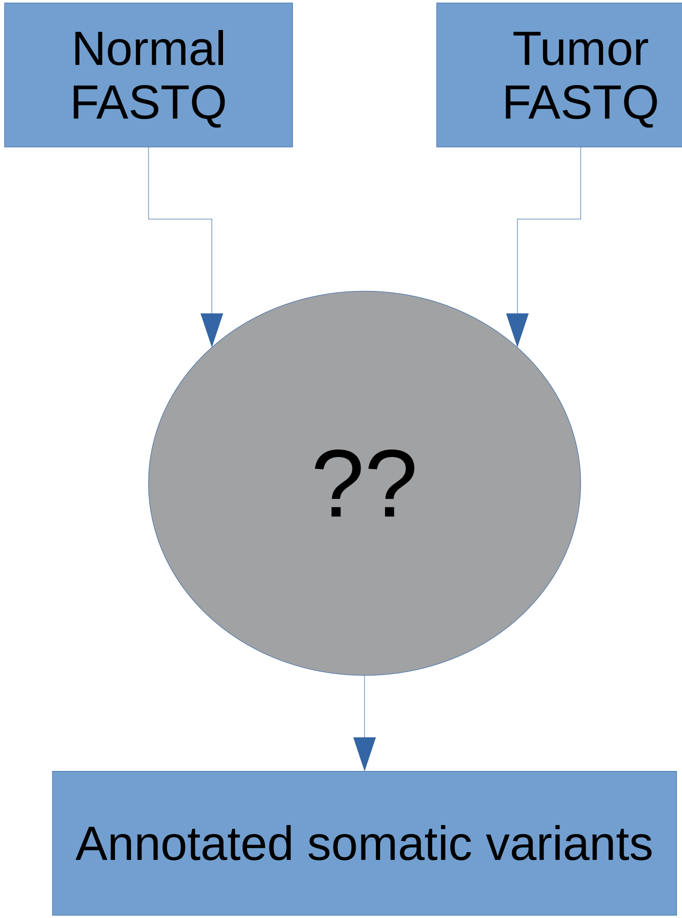
Somatic mutation calling pipelin for a matched sample pair

Willie Yu 17 Jan 2022
willie.yu@duke-nus.edu.sg

Given

- Short read DNA sequences in FASTQ format from
  - Patient Tumor
  - Patient Matched Normal

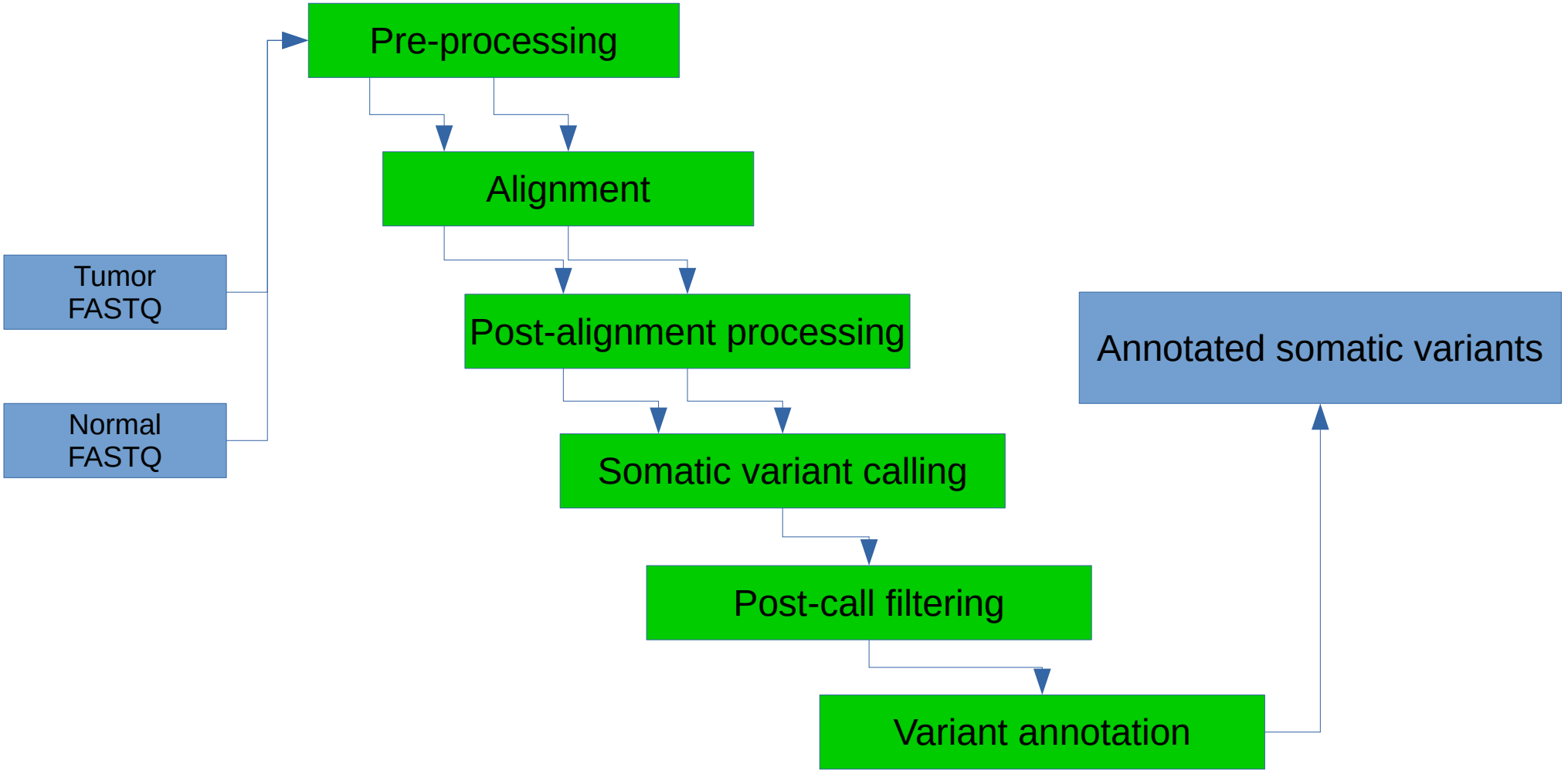
```
@A00609:341:HK33WDSX2:3:1101:1235:1000 1:N:0:GTAGCTGA+AGTAGGTC
NGCTCCTTTTTGTCCTCTTTTGCTCTTGCTTGGTGTGACATGCCTCTTCTGCAACTGCTGAAGTAATTCACCGCACGTTTTCACTCTTTCCTCATTTTTTCCACCTGGCAGAAGCGTGAAGAAAGATTACAGACACATGGGATACCTCTG
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:F,FFFFFFFFFFFFFFFFF:F:,FFFFF,FF,FFFFFFFFFFFFFFFF,FFFFFFFFF:FFFFFFF::FFFFFF:FF
@A00609:341:HK33WDSX2:3:1101:2392:1000 1:N:0:GTAGCTGA+AGTAGGTC
NCCAGCCTGGCCAACAAGGTGAAACCCCATCTCTACTACAAATACAAAAACTAGCCAGGTGTGGTGGCACACGCCTGTAATCCCAGCTACTAGGGGAGGCTGAGGCAAGAGAATCGCTTGAACCTGGGAGGCAGAGGTTGAAGTGAGCTG
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFF,FF,FFFFFFFF:FFF:FFFFFFFFFFFFF:FFFFFFFFFFFFFFF,FFFFFFFFFFF::FFFFF:FFFFFFF:FFFFF,,FFFFFFFF
@A00609:341:HK33WDSX2:3:1101:2627:1000 1:N:0:GTAGCTGA+AGTAGGTC
NAAAGCTGGATGGAGAATGACTTTGACGAGCTAAGAGGAGAAGGCTTCAGACGATCAAATTACTCTGAGCTACAGGAGGACATTCAAACCAAAGGCAAACAAGTTGAAAACTTTGAAAAAAATTTAGAAGAATGTATAACTAGAATAACC
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFF
@A00609:341:HK33WDSX2:3:1101:3242:1000 1:N:0:GTAGCTGA+AGTAGGTC
NTATTTTCTCAGTGCGCAAAAATTCAATCAACAAAATATATTTAATGTCTAGAATAAGTATCAAAGTAACAGATTCAGAAAGTAAAAATTCAATCAACCTATTATTTTTGTATATAAGACATGGAACTGGTTTATTGACTTCACAGGCTT
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FF:FFFFFFFF:FFFFFFFFFFF:
@A00609:341:HK33WDSX2:3:1101:3477:1000 1:N:0:GTAGCTGA+AGTAGGTC
NTCTCAGGTTCCAGGAATTAAGACGTAGACATCTCTGGGGGAAGAGGGTGGGTGGTTTACTATTCTGCCTACCACTCTAGTCTTAACTTTGTTATTAGGGCCCAATTATTATTCCAATATGCTGAATTAAATAGATTTCAGGAAAACTGA
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFF:FF:FFF:FFFF
@A00609:341:HK33WDSX2:3:1101:3857:1000 1:N:0:GTAGCTGA+AGTAGGTC
NAACTGAATTTATTCTCAGCATAGATGAGCCTGGTGCCACCTTGAATATCCACCATTCACAGACTGCCACTTTGGTCATGGGAACTCACATGTGCAACTTGTTTCCATGACAGCCTGTAAATTACAGGTCCCAAAGTCAGCTGCAAGGAA
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00609:341:HK33WDSX2:3:1101:3965:1000 1:N:0:GTAGCTGA+AGTAGGTC
NTTAATTTATTTAAAGATAATTTTTAAAAACTGAGGCAAATCAGTCATGGTGTTAAATTTATGTGACAAAAAATATTTTGTCAGTATTATTTGAAGTATTTTCTAATTTTATTGATATAAAATCAAACTTATATTGGTAAATTCTCTTTC
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFF,FFFFFF,FFFFFF
```



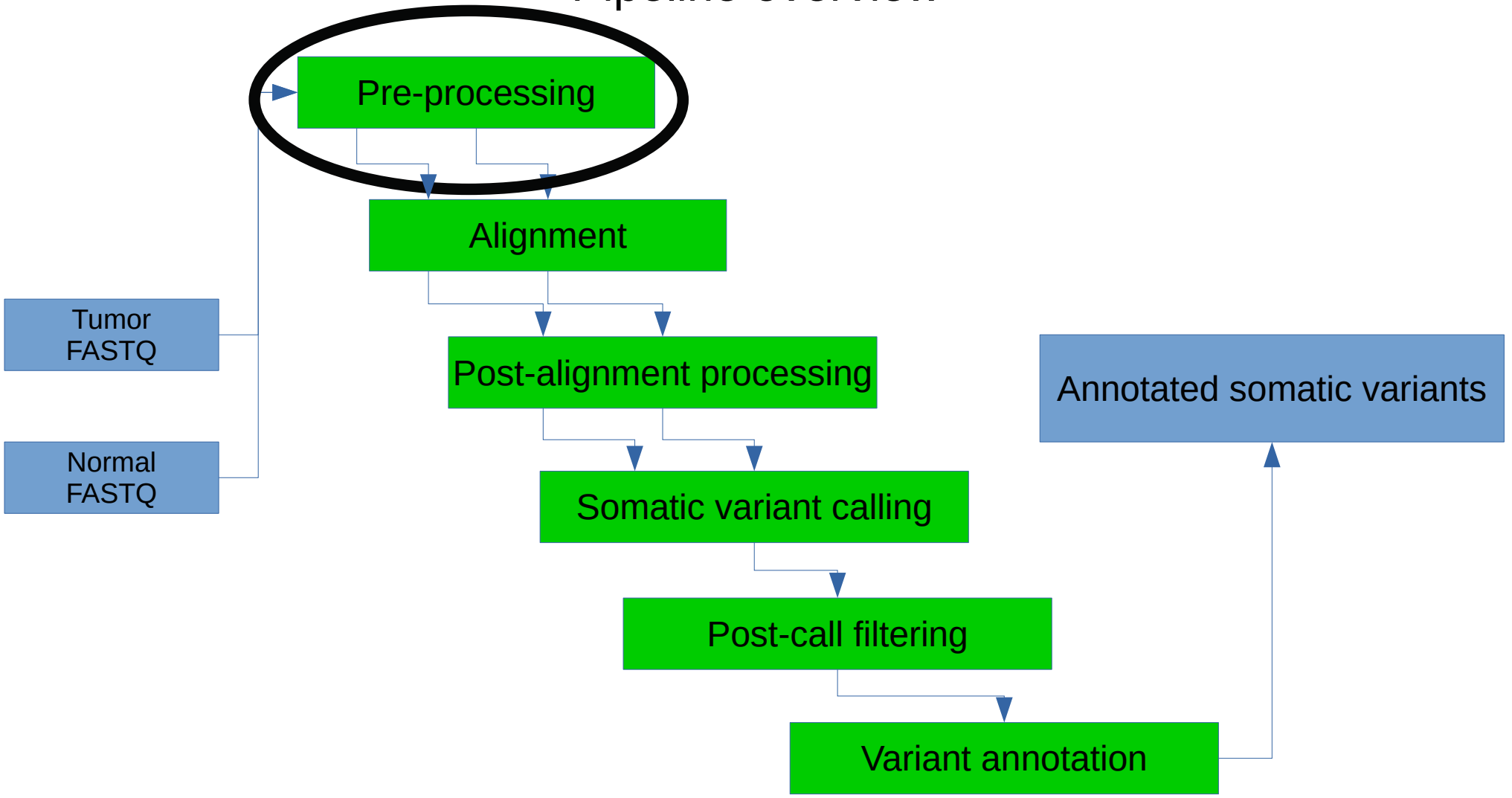

```
@A00609:341:HK33WDSX2:3:1101:1235:1000 1:N:0:GTAGCTGA+AGTAGGTC
NGCTCCTTTTTGTCCTCTTTTGCTCTTGCTTGGTGTGACATGCCTCTTCTGCAACTGCTGAAGTAATTCACCGCACGTTTTCACTCTTTCCTCATTTTTTCCACCTGGCAGAAGCGTGAAGAAAGATTACAGACACATGGGATACCTCTG
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:F,FFFFFFFFFFFFFFFFFF:F:,FFFFF,FF,FFFFFFFFFFFFFFFF,FFFFFFFFF:FFFFFFF::FFFFFF:FF
@A00609:341:HK33WDSX2:3:1101:2392:1000 1:N:0:GTAGCTGA+AGTAGGTC
NCCAGCCTGGCCAACAAGGTGAAACCCCATCTCTACTACAAATACAAAAACTAGCCAGGTGTGGTGGCACACGCCTGTAATCCCAGCTACTAGGGGAGGCTGAGGCAAGAGAATCGCTTGAACCTGGGAGGCAGAGGTTGAAGTGAGCTG
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFF,FF,FFFFFFFF:FFF:FFFFFFFFFFFFF:FFFFFFFFFFFFFFF,FFFFFFFFFFF::FFFFF:FFFFFFF:FFFFF,,FFFFFFFF
@A00609:341:HK33WDSX2:3:1101:2627:1000 1:N:0:GTAGCTGA+AGTAGGTC
NAAAGCTGGATGGAGAATGACTTTGACGAGCTAAGAGGAGAAGGCTTCAGACGATCAAATTACTCTGAGCTACAGGAGGACATTCAAACCAAAGGCAAACAAGTTGAAAACTTTGAAAAAAATTTAGAAGAATGTATAACTAGAATAACC
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFF
@A00609:341:HK33WDSX2:3:1101:3242:1000 1:N:0:GTAGCTGA+AGTAGGTC
NTATTTTCTCAGTGCGCAAAAATTCAATCAACAAAATATATTTAATGTCTAGAATAAGTATCAAAGTAACAGATTCAGAAAGTAAAAATTCAATCAACCTATTATTTTTGTATATAAGACATGGAACTGGTTTATTGACTTCACAGGCTT
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FF:FFFFFFFF:FFFFFFFFFFF:
@A00609:341:HK33WDSX2:3:1101:3477:1000 1:N:0:GTAGCTGA+AGTAGGTC
NTCTCAGGTTCCAGGAATTAAGACGTAGACATCTCTGGGGGAAGAGGGTGGGTGGTTTACTATTCTGCCTACCACTCTAGTCTTAACTTTGTTATTAGGGCCCAATTATTATTCCAATATGCTGAATTAAATAGATTTCAGGAAAACTGA
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFF:FF:FFF:FFFF
@A00609:341:HK33WDSX2:3:1101:3857:1000 1:N:0:GTAGCTGA+AGTAGGTC
NAACTGAATTTATTCTCAGCATAGATGAGCCTGGTGCCACCTTGAATATCCACCATTCACAGACTGCCACTTTGGTCATGGGAACTCACATGTGCAACTTGTTTCCATGACAGCCTGTAAATTACAGGTCCCAAAGTCAGCTGCAAGGAA
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00609:341:HK33WDSX2:3:1101:3965:1000 1:N:0:GTAGCTGA+AGTAGGTC
NTTAATTTATTTAAAGATAATTTTTAAAAACTGAGGCAAATCAGTCATGGTGTTAAATTTATGTGACAAAAAATATTTTGTCAGTATTATTTGAAGTATTTTCTAATTTTATTGATATAAAATCAAACTTATATTGGTAAATTCTCTTTC
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFF,FFFFFF,FFFFFF
```


Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	GeneDetail.refGene
10	63211352	63211352	G	T	intronic	JMJD1C	.
10	111455769	111455769	T	C	intergenic	HEAT2;GPAM	dist=337173;dist=694096
11	4130486	4130486	A	G	intronic	RRM1	.
12	120420659	120420659	C	A	intergenic	MSI1;COX6A1	dist=51495;dist=17454
16	2226050	2226050	C	G	intronic	E4F1	.
16	85499571	85499571	A	G	intergenic	MIR5093;GSE1	dist=193246;dist=111838
17	82626591	82626591	C	A	intronic	WDR45B	.
19	6797171	6797171	G	A	intronic	VAV1	.
19	13950460	13950460	G	A	intronic	PODNL1	.
3	73247693	73247693	C	T	intergenic	PPP4R2;PDZRN3	dist=178492;dist=134737

Pipeline overview

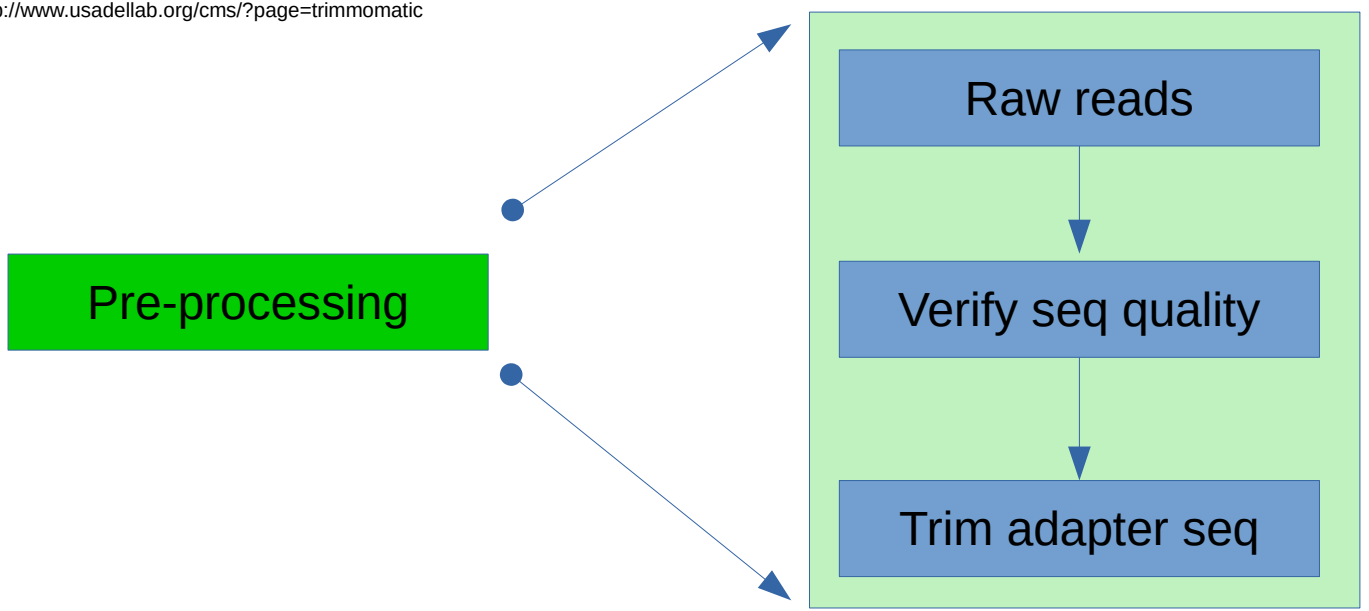


Pipeline overview



Pre-processing

- Purpose
  - Check and verify sequence qualities
    - FastQC
    - https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
  - Trimming of adapter sequences
    - Trimmomatic
    - http://www.usadellab.org/cms/?page=trimmomatic



Pre-processing: FastQC

file:///home/willie/temp_folder/AB21Jul3_DDHG210006186-1a_HJYLKDSX2_L2_2_fastqc.html 120%

Most Visited GLIDR Fedora Docs Fedora Project User Communities Red Hat Free Content Error: RStudio Sign ... RKI - Coronavirus S... OpenBSD manual p... Other Bookmarks

FastQC Report

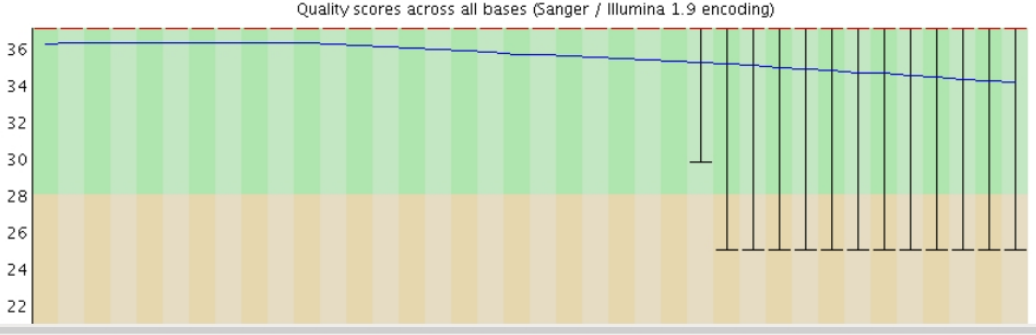
Wed 22 Dec 2021
AB21Jul3_DDHG210006186-1a_HJYLKDSX2_L2_2.fq.gz

Summary

Basic Statistics

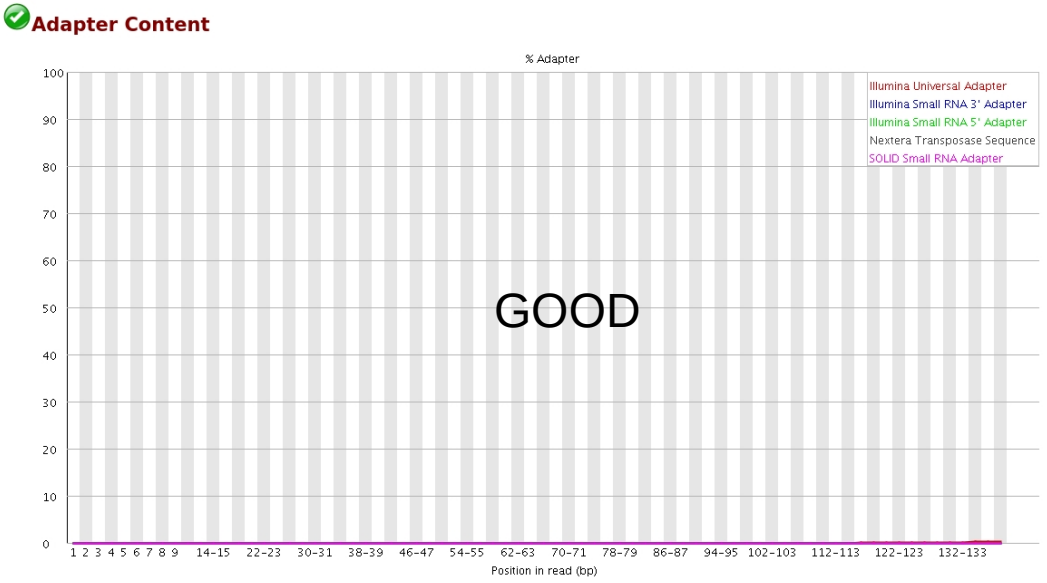
Measure	Value
Filename	AB21Jul3_DDHG210006186-1a_HJYLKDSX2_L2_2.fq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	35085362
Sequences flagged as poor quality	0
Sequence length	150
%GC	44

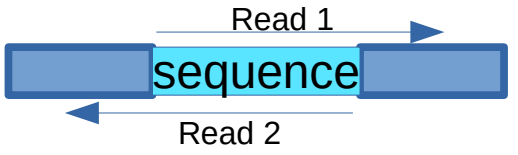
Per base sequence quality

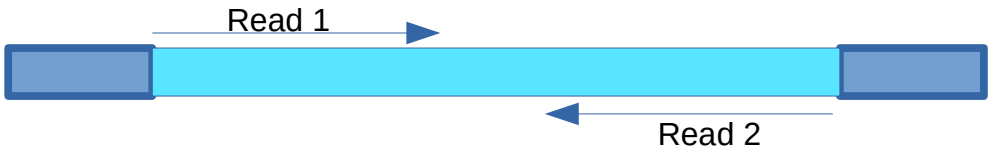


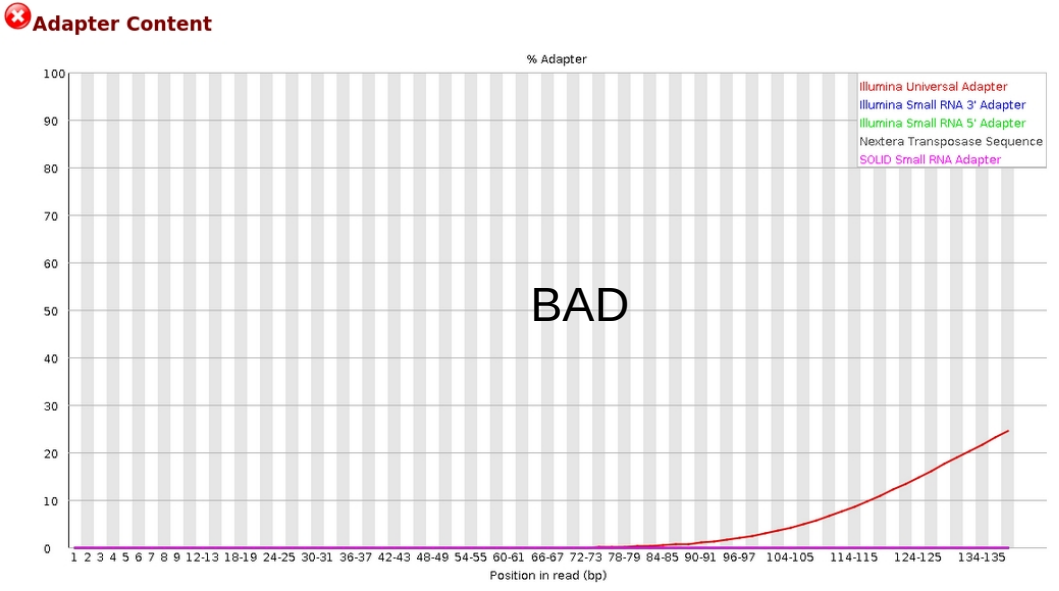
Produced by FastQC (version 0.11.9)

Pre-processing: FastQC



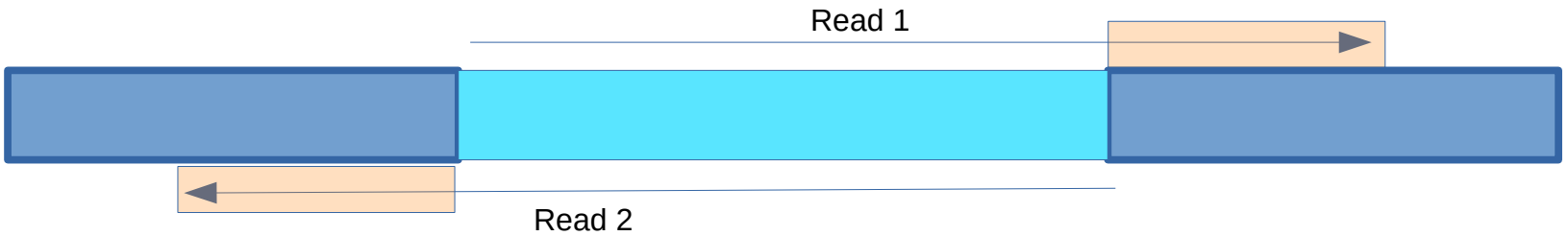






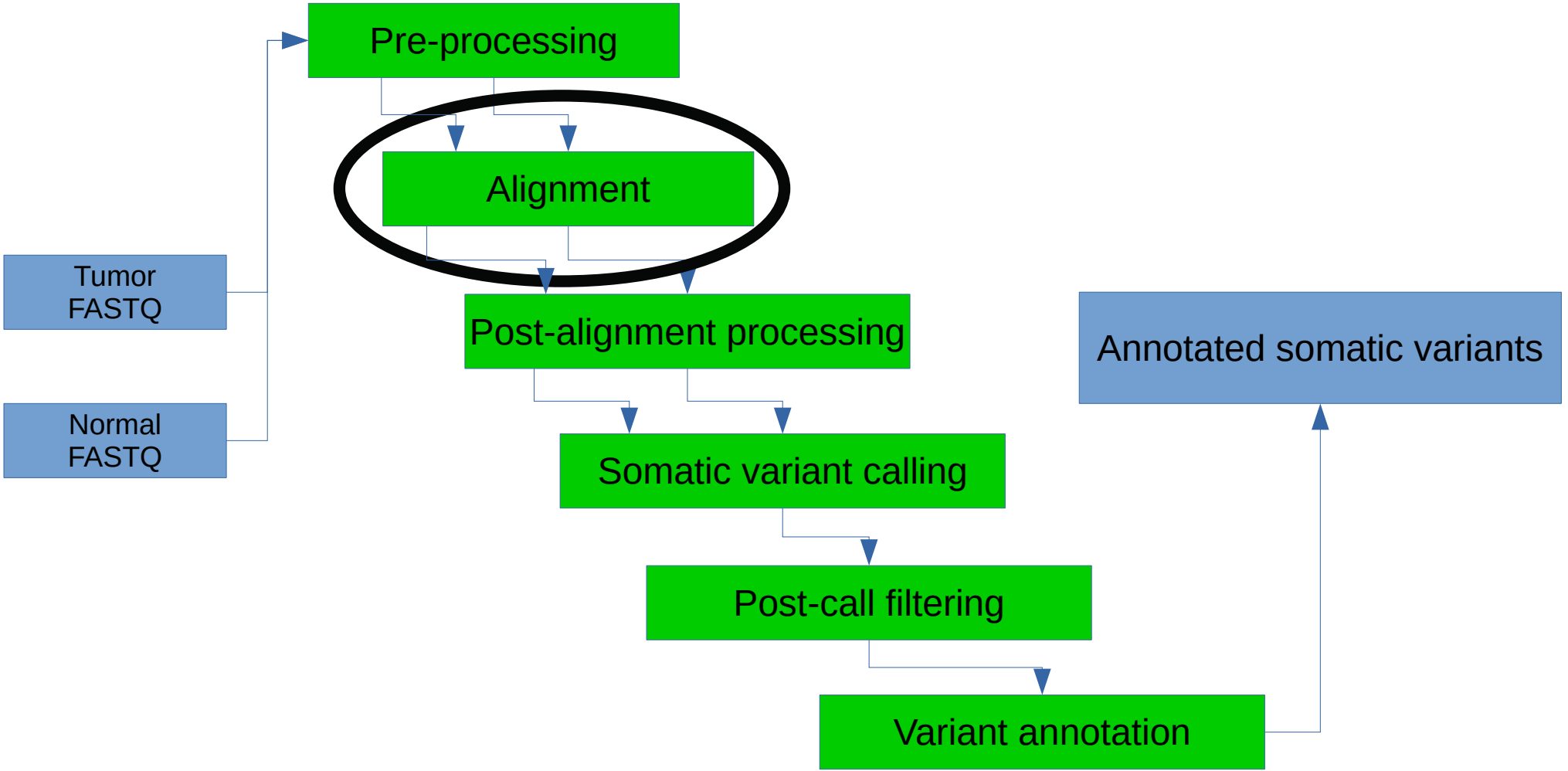
Adapter trimming

Cut out the part of the DNA insert matching the adapter sequence



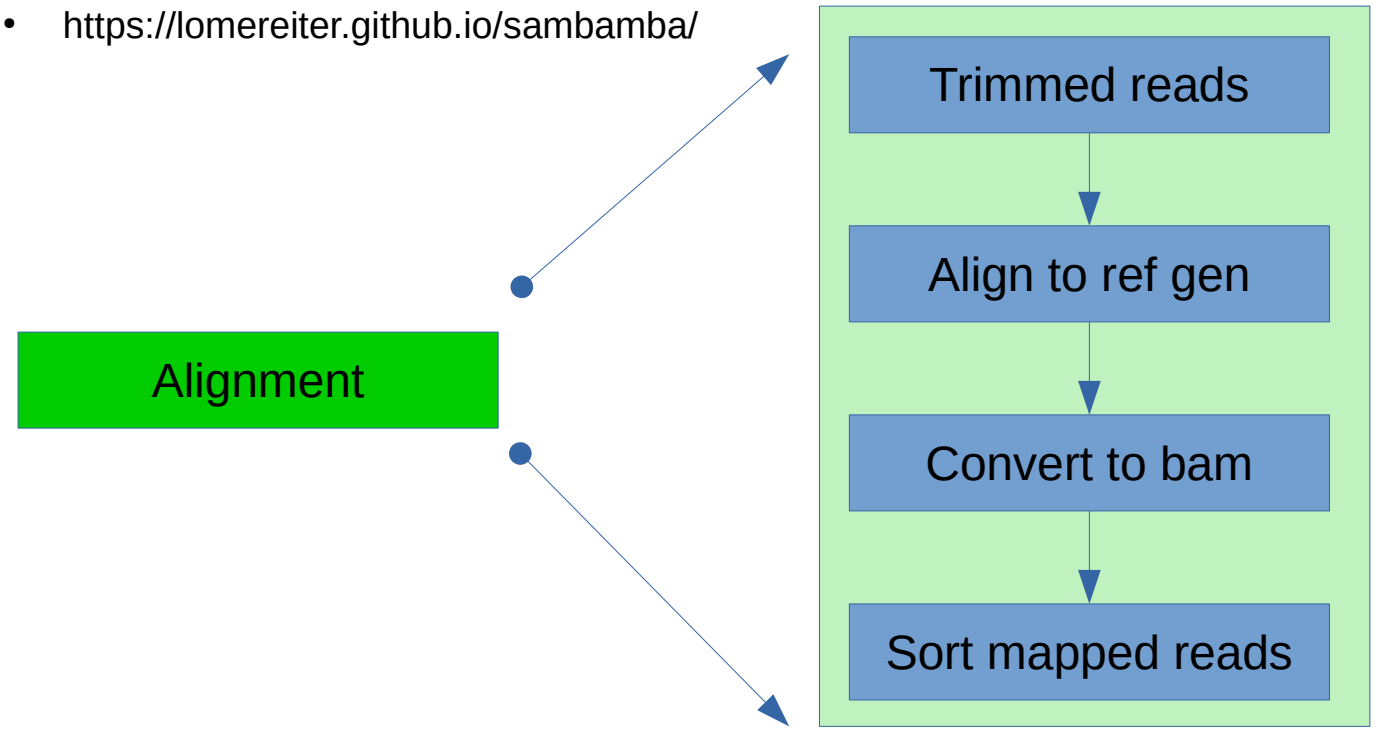
Trimmomatic
http://www.usadellab.org/cms/?page=trimmomatic

Pipeline overview

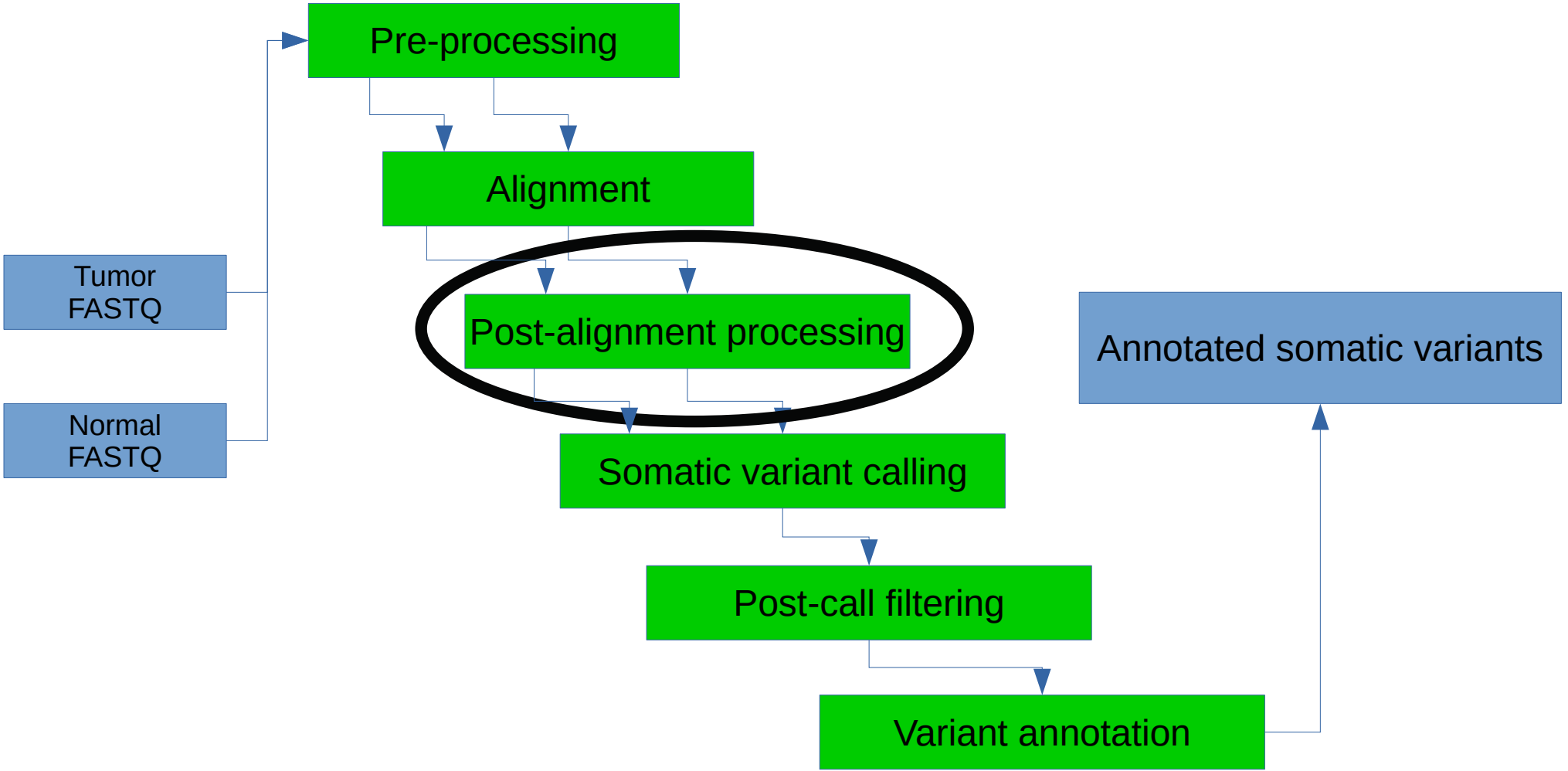


Alignment

- Goal: Mapping each DNA insert against a reference genome
  - BWA-MEM2: mapping DNA sequences against a large reference genome
    - https://github.com/bwa-mem2/bwa-mem2
  - Sambamba:convert aligned output into bam format and sort mapped reads by chr
    - https://lomereiter.github.io/sambamba/

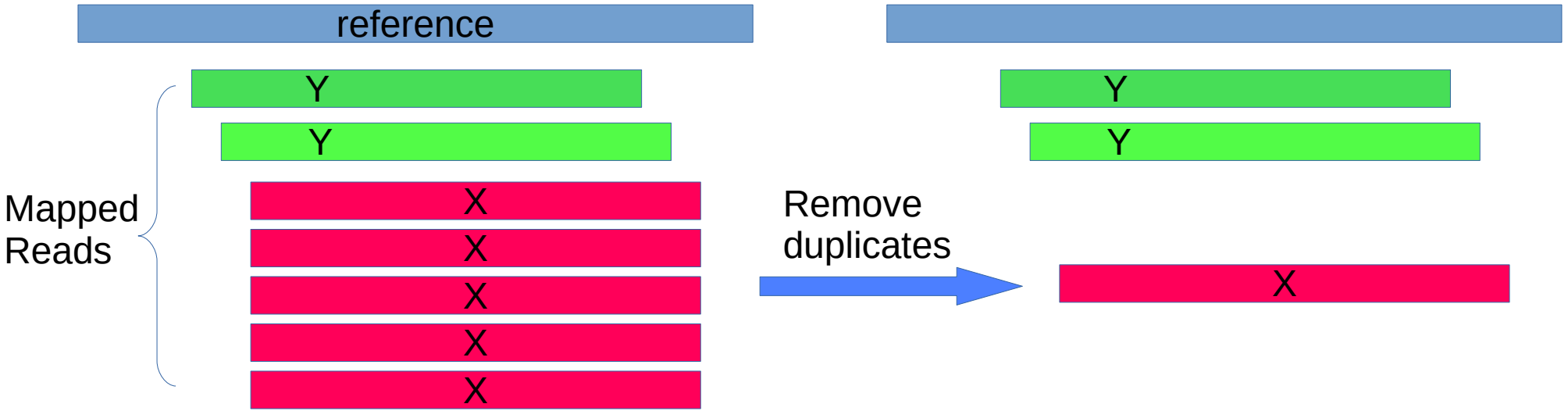


Pipeline overview

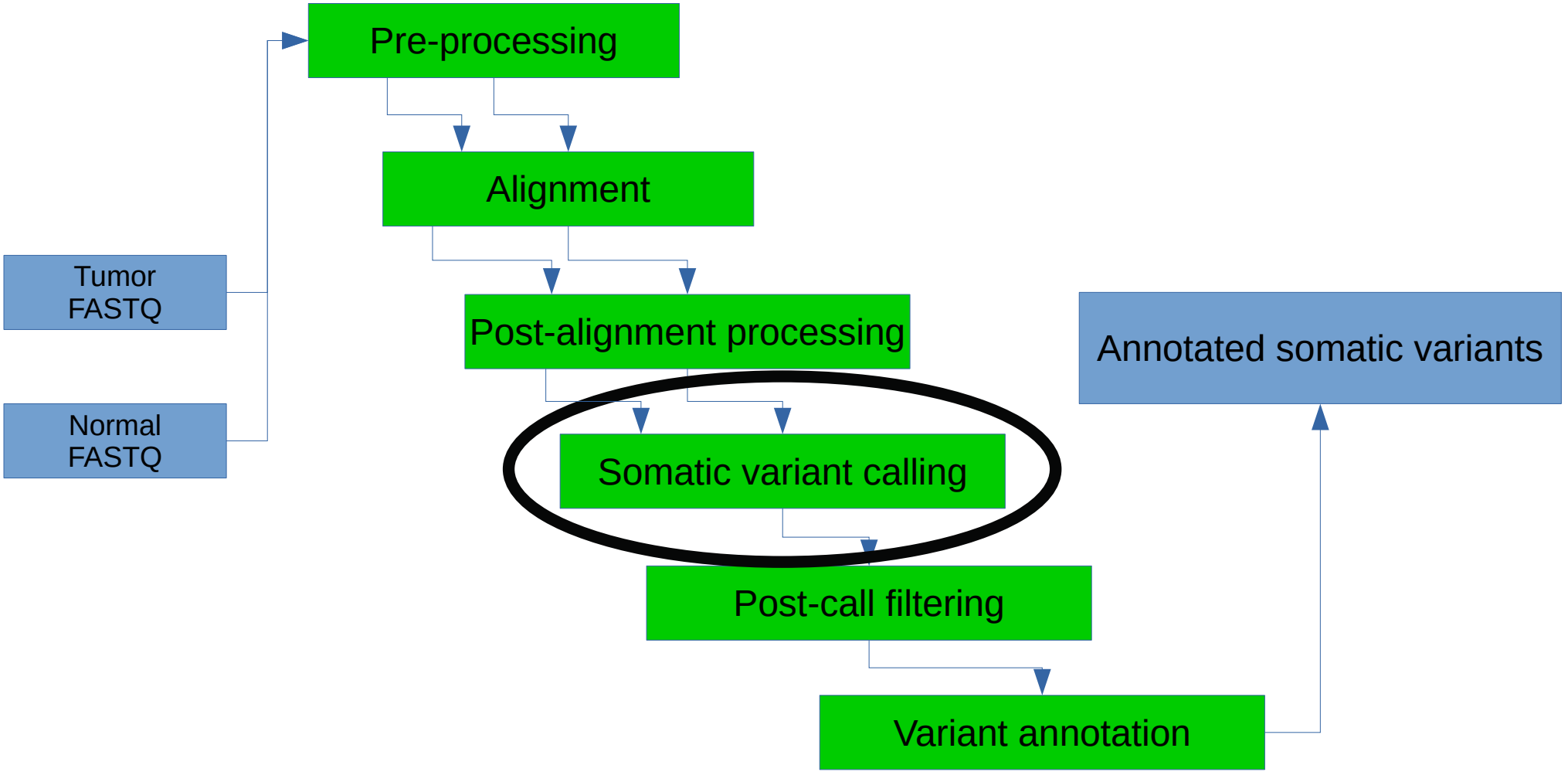


Post-alignment processing

- PCR duplicate removal:
  - Sambamba: duplicate removal
- Why duplicate removal?
  - Minimize sequencing error propagation
  - Reduce false positives in variant calling

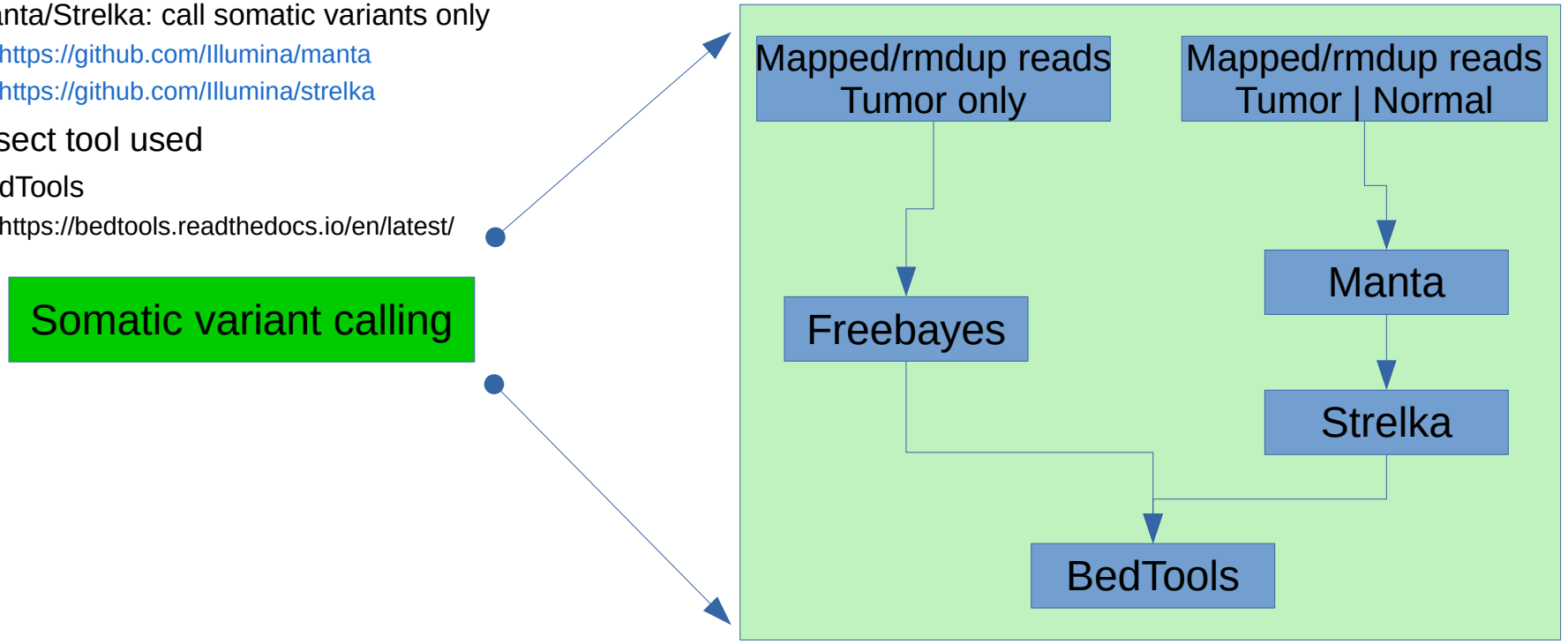


Pipeline overview

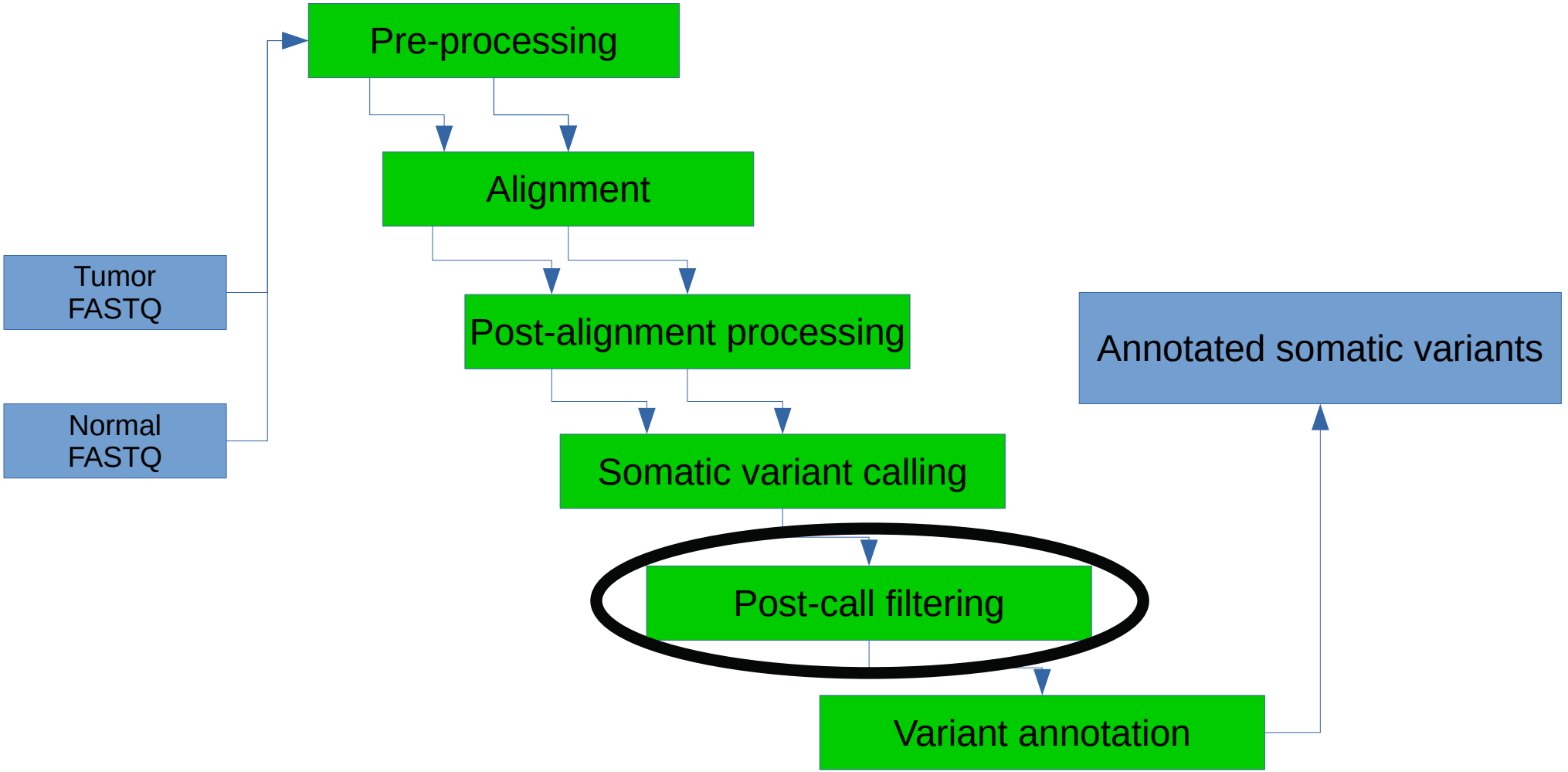


Somatic variant calling

- All variant callers generates artifacts/false positive calls
  - Solution: Intersect variants with at least 2 callers
- Variant callers used
  - Freebayes: call all variants
    - https://github.com/freebayes/freebayes
  - Manta/Strelka: call somatic variants only
    - https://github.com/Illumina/manta
    - https://github.com/Illumina/strelka
- Intersect tool used
  - BedTools
    - https://bedtools.readthedocs.io/en/latest/



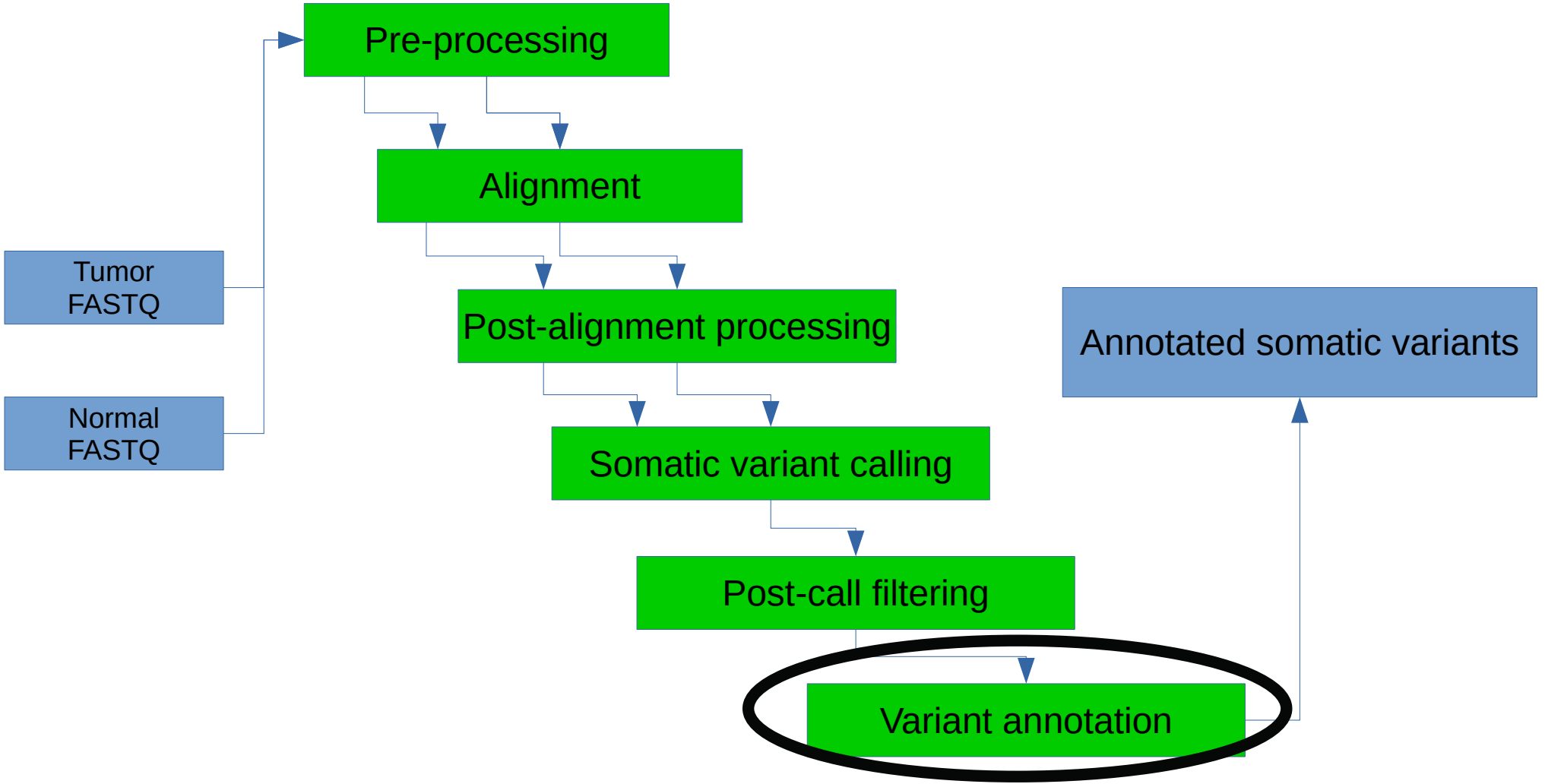
Pipeline overview



Post-call filtering

- Caller specific artifacts removal still insufficient
- False positive calls still in high numbers due to
  - Germline variants leak
  - Variant calls in highly variable / repetitive regions of the genome
- Solution
  - Create blacklists of genomic positions / regions
  - Use BedTools to filter against called variants

Pipeline overview



Variant Annotation

Filtered somatic variant list → ? → Annotated somatic variant list

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  NORMAL  TUMOR
1       1591986 .       G       A       .       PASS    SOMATIC;QSS=23;TQSS=1;NT=ref;QSS_NT=23;
,0:19,19:0,0      16:0:0:0:5,5:0,0:11,11:0,0
1       3605430 .       T       G       .       PASS    SOMATIC;QSS=19;TQSS=1;NT=ref;QSS_NT=19;
,0:0,0:17,17      22:1:0:0:0,0:0,0:7,7:14,15
1       3827839 .       G       A       .       PASS    SOMATIC;QSS=20;TQSS=1;NT=ref;QSS_NT=20;
:0,0      15:0:0:0:4,4:0,0:11,11:0,0
1       5491320 .       G       C       .       PASS    SOMATIC;QSS=36;TQSS=1;NT=ref;QSS_NT=36;
,0:24,24:0,0      24:0:0:0:0,0:10,10:14,14:0,0
1       7985092 .       G       A       .       PASS    SOMATIC;QSS=16;TQSS=1;NT=ref;QSS_NT=16;
:0,0      21:0:0:0:9,9:0,0:12,12:0,0
1       17562546        .       A       T       .       PASS    SOMATIC;QSS=41;TQSS=1;NT=ref;QS
:0:34,35:0,0:0,0:0,0      32:0:0:0:22,22:0,0:0,0:10,10
1       18141927        .       G       A       .       PASS    SOMATIC;QSS=19;TQSS=1;NT=ref;QS
:0:0,0:0,0:18,19:0,0      18:0:0:0:8,8:0,0:10,10:0,0
1       21283788        .       T       C       .       PASS    SOMATIC;QSS=16;TQSS=1;NT=ref;QS
:0:0,0:0,0:0,0:15,15      23:0:0:0:0,0:7,7:0,0:16,16
1       24992764        .       G       A       .       PASS    SOMATIC;QSS=31;TQSS=1;NT=ref;QS
:0:0,0:0,0:22,22:0,0      21:0:0:0:9,9:0,0:12,12:0,0
1       25949960        .       G       A       .       PASS    SOMATIC;QSS=18;TQSS=1;NT=ref;QS
,0:15,15:0,0      27:0:0:0:7,7:0,0:20,20:0,0
1       26044713        .       C       T       .       PASS    SOMATIC;QSS=15;TQSS=1;NT=ref;QS
:0:0,0:16,16:0,0:0,0      31:0:0:0:0,0:19,19:0,0:12,12
1       27747855        .       G       A       .       PASS    SOMATIC;QSS=31;TQSS=1;NT=ref;QS
:0:0,0:0,0:22,22:0,0      24:0:0:0:10,10:0,0:14,14:0,0
1       28029437        .       A       G       .       PASS    SOMATIC;QSS=34;TQSS=1;NT=ref;QS
:0:23,24:0,0:0,0:0,0      25:0:0:0:15,15:0,0:10,10:0,0
1       28393932        .       T       G       .       PASS    SOMATIC;QSS=33;TQSS=1;NT=ref;QS
:0:0,0:0,0:0,0:23,23      27:0:0:0:0,0:0,0:13,13:14,14
```

Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	GeneDetail.refGene
1	1591986	1591986	G	A	intergenic	SSU72;FNDC10	dist=17123;dist=6026
1	3605430	3605430	T	G	intronic	MEGF6	.
1	3827839	3827839	G	A	intronic	CEP104	.
1	5491320	5491320	G	C	intergenic	AJAP1;MIR4689	dist=698798;dist=371352
1	7985092	7985092	G	A	UTR3	PARK7	NM_001123377:c.*38G>A;NM_007262:c.*38G>A
1	17562546	17562546	A	T	intronic	ARHGEF10L	.
1	18141927	18141927	G	A	intronic	IGSF21	.
1	21283788	21283788	T	C	intronic	ECE1	.
1	24992764	24992764	G	A	intergenic	RUNX3;MIR4425	dist=27754;dist=30739
1	25949960	25949960	G	A	intergenic	STMN1;PAFAH2	dist=43083;dist=9807
1	26044713	26044713	C	T	intronic	SLC30A2	.
1	27747855	27747855	G	A	intronic	FAM76A	.
1	28029437	28029437	A	G	intronic	EYA3	.
1	28393932	28393932	T	G	intronic	PHACTR4	.

Solution: ANNOVAR
https://annovar.openbioinformatics.org/en/latest/

Pipeline overview