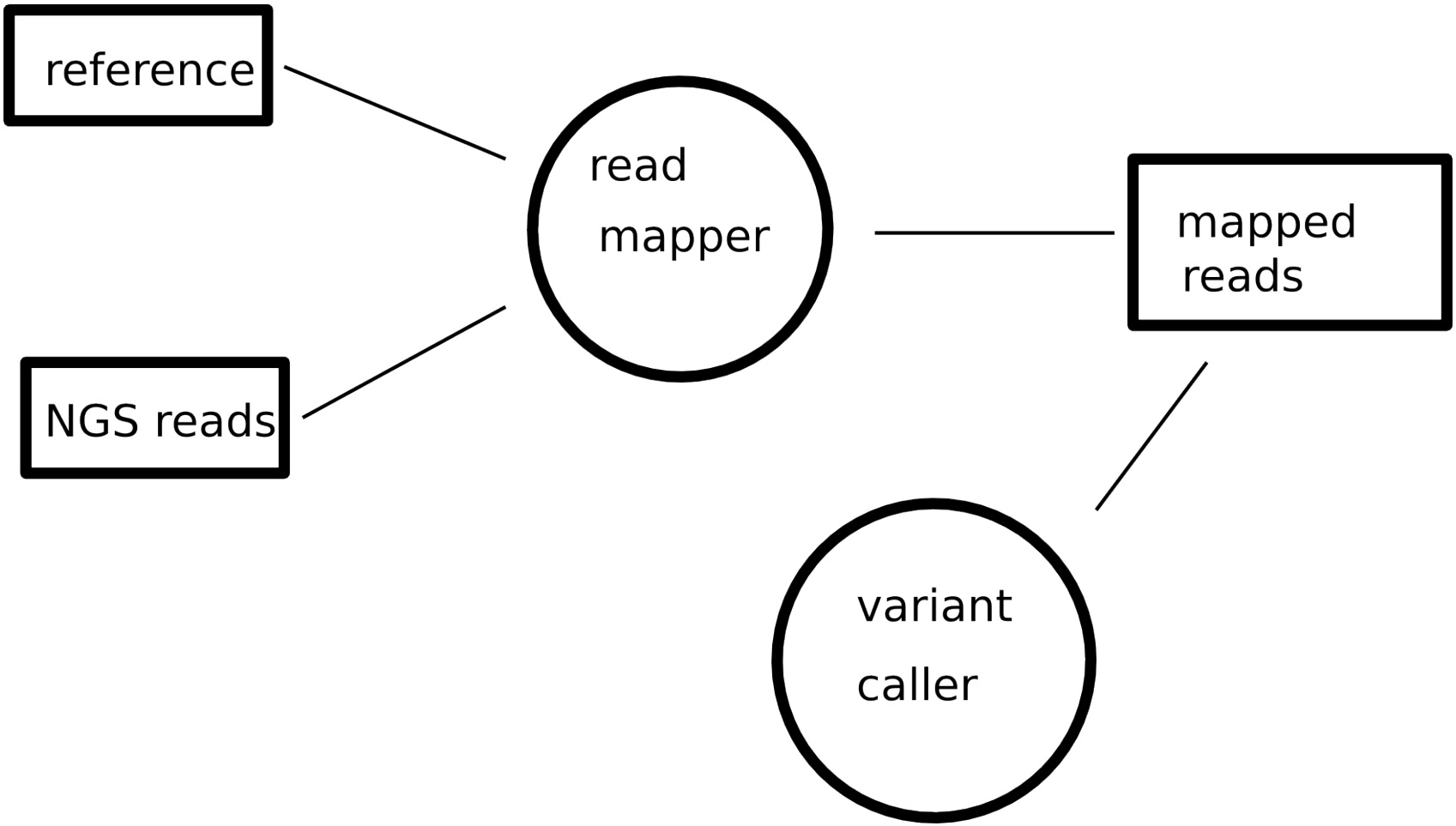
Pipeline overview



Read mapping

- Index reference genome (need to do it only once per reference per mapping tool)

```
tmap index -f reference.fasta
```

Read mapping

- Index reference genome (need to do it only once per reference per mapping tool)

```
tmap index -f reference.fasta
```

- Map reads against reference genome

```
tmap map3 -f reference.fasta \
-r reads.fastq \
-i fastq \
-o 2 \
-s mapped_reads.bam
```

BAM file

- Let's see what is inside the bam file

```
samtools view mapped_reads.bam | head -n 20
```

```
SFHMR:03755:02039	4	*	0	0	*	*	0	0	GATACGTCGTTGACCGATGACCCCGGTTCAGGCTTCACCACAGTGTGGAACGCGGTCGTCTCCGAACTTAACGGCGACCCTAAGGT
TGACGACGGACCCAGCAGTGATGCTAATCTCAGCGCTCCGCTGACCCCTCAGCAAAGGGCTTGGCTCAA	CCCBBBCABA;BAB;BAAA????5?<D8>6:/99/9=@9???AAACC<@<@BA@9@:;;;AA9=>96708<BB;@???A6;;4:;A9@<<<@D@6444+5<=
=>?>>AA??:8@??<8<?:@D?4::8;----'.;;...5/59/==7A:@?@@@	RG:Z:NOID	PG:Z:tmap
SFHMR:01788:02543	16	gi|385235550|ref|NC_017387.1|	2594310	1	131S8M	*	0	0	CAATGGCTGGACGAGATTGAGCCAAGCCCTTTGCTGAGGGGTCAGCGGAGCGCTGAGATTAG
CATCACTGCTGGGTCCGTCGTCAACCTTAGGGTCGCCGTTAAGTTCGGAGACGACCGCGTTCCACACTGTGGTGAAG	>:BA;AAA;B@@>@@@8BBA?;D<D?8??7AAABCCC8DCCBBCBGADDBBBBBCCCC@BBCAAA775==>7+541263334661>9@9A9)99
9>?9?A:9199;??5AAB99:C=CCBC<H=AAACA?>>:DCA<AA	RG:Z:NOID	PG:Z:tmap	MD:Z:8	NM:i:0	AS:i:8	XA:Z:map3-1	XS:i:-2147483647	XT:i:1
SFHMR:00676:00709	4	*	0	0	*	*	0	0	CGCATCGCTCCGCCGGCGACCGACGAAGCCGACGACACTACCGTGCCGCCTTCCGAAAATCCTGCTACCACATCGCCAGACACCAC
AACCGACAACGACGAGATTGATGACAGCGCTGCGGCACGGGGCGATA ??@@CAAADC<BB<C=BAAA=BBB??:@B=BCBAABBFBBB<@@@@:>=7>:C:BCDC4CC<B@@@@A:A779?99148==8>;@@@:A9@@>>:>???@D@@@9?>>>AAAAA>??A??9>>>@A
B/??@@A	RG:Z:NOID	PG:Z:tmap
SFHMR:01673:02496	4	*	0	0	*	*	0	0	ACCGACGAAGCCGACGACACTACCGTGCCGCCTTCCGAAAATCCTGCTACCACATCGCCAGACACCACAACCGACAACGACGAGAT
TGATGACAGCGCTGCGGCACGGGGCGATAACCAGCACAGTTGGCCAAGTTACTTC	BB:AAB@?8?@2996@@:::???:>>>A>DMAJ>D>BAAA7@??CCBBC;=//////8BBBBCE8ACD<B7>>EC<999=7:7<=:AA>>@CD?>?ABBAC;<>:===2>@@>G9A;
=>>??A@A5;4;7=37<2>?A;A	RG:Z:NOID	PG:Z:tmap
SFHMR:01579:01306	4	*	0	0	*	*	0	0	CTACCACATCGCCAGACACACACCACACACGAGCTT	:;=B576;=9554966666666/)////////555/	RG:Z:N
OID	PG:Z:tmap
SFHMR:00598:01413	16	gi|385235550|ref|NC_017387.1|	2603163	1	136S13M1D2M	*	0	0	GGTGTAGCGACGGTTAAGGCTGGTTACGCCAGCGGTAGCGGAATCGGTATTGTG
CGGGCGCTCGGTGAAGTAACTTGGCCAACTGTGCTGGTTATCGCCCCGTGCCGCAGCGCTGTCATCAATCTCGTCGTTGTCGGTTGTGGTGTCTGGC	C>CCBBABBBAA<A=C<B=CBB=B=AAAC=DFGD>EB@@@@J?BAB<BBB=BAA???9??????;@FD=C
BB<AB<A<B=B<CCCEA>?B9?8?@@AAAA6ABCD:@AFDBABBABBBCEC>EICBBBDCC=CCBA;A:AAA<BBBBBA<C	RG:Z:NOID	PG:Z:tmap	MD:Z:13^A2	NM:i:1	AS:i:8	XA:Z:map3-1	XS:i:-
2147483647	XT:i:1
SFHMR:00701:00509	4	*	0	0	*	*	0	0	ACGACGAGATTGATGACAGCGCTGCGGCACGGGGCGATAACCAGCACAGTTGGCCAAGTTACTTCACCGAGCGCCCGCACAATACC
GATTCCGCTACCGCTGGCGTAACCAGCCTTAACCGTCGCTACACCTTTGAT	??AAAC??@<A?CAC@AA@BBBBCB<BAD?3????@AA;A;ABA@@BCB=B;B;@<A>9?=@;ABA<A@>>>?7@A9::A<AAA=CBC>C=CBBCD?DCB?9CC::9E>DCC>B=C>C
@???@BAAABB;@9@@@@A	RG:Z:NOID	PG:Z:tmap
SFHMR:02998:02581	16	gi|385235550|ref|NC_017387.1|	941779	68	109S19M	*	0	0	TATCAAAGGTGTAGCGACGGTTAAGGCTGGTTACGCCAGCGGTAGCGGAATCGGTATTGTGC
GGGCGCTCGGTGAAGTAACTTGGCCAACTGTGCTGGTTATCGCCCCGTGCCGCAGCGCTGTCATCA	=CCDDD<B<BBBBBAABBB=B<B<B=CCC>C=BBBA<AAAA=@??@A8A<@AC<C776:BBBBB6@@@@D@A8?::4<@@9>8386B:?8>@AAA@@A8>8>@
=::::-?AAD<D@@AA@<<>@A<<@:	RG:Z:NOID	PG:Z:tmap	MD:Z:13A3G1	NM:i:2	AS:i:11	XA:Z:map3-1	XS:i:-2147483647	XT:i:1
SFHMR:01284:00299	4	*	0	0	*	*	0	0	CCCCCTGGTTCCATCTGGGGCGAGTCCGGTCCGTCGGCAAGACACACCTGCTGACACGCGGCAGGCAATCCTATGCCCAACGGTTG
TTCCCGGGAATCGCAGG	(33333-3/<8:::;8,86869663'-(.-2....(..689:;///6::8<888887><656/2663//(//><>4BA8<:46/6656*57*>>75/////)/	RG:Z:NOID	PG:Z:tmap
SFHMR:03138:02989	4	*	0	0	*	*	0	0	GGTCTCGGCAAGACACACCTGCTACACGCGGCAGGCAACTATGCCCAACGGTTGTTCCCGGGAATGCAGGTCAAATATGTCTCCAC
CGAGGAATTCACCAACGACTTCATTAACTCGCTCCGCGATGACCGCAAGGTCGCATTCAAACGCAGCTACCG	E>EDBBB<BC<@?@CBCF=9??ACBB?@@B=BBCAFD@CCBCBBC<B=??9?<@@;BB:BB:B=ECAAC=C>=>5>>>>>?>=8>B9?@@;A3;
3;96171.--11-2BB<C<BBCDMCC@BEDDDCCF:??A<A<BBAABC=AAA9AA???AB??9@	RG:Z:NOID	PG:Z:tmap
SFHMR:02085:01984	4	*	0	0	*	*	0	0	TCGCTCCGCGATGACCGCAAGGTCGCATTCAAACGCAGCTACCGCGACGTAGACGTGCTGTTGGTCGACGACATCCAATTCATTGA
AGGCAAAGAGGGTATTCAAGAGGAGTTCTTCCACACCTTCAACACCTTGCACAATGCCAACAAGCAAATCG	>=???9@@??776:2:::6=6<<;7659;;2D@@@@BB?@@:@;;;A?>?@>>>??8661/4-98888:>>A@=;@7E=A;:3A?9>7>=4????5A959=C
C:>?A:A@@:AB;B;CBBB<@<A@<CBB<B<BBBBB<BCC>B=CGCG?:3:96;<	RG:Z:NOID	PG:Z:tmap
SFHMR:00470:00449	0	gi|385235550|ref|NC_017387.1|	3708242	50	18M123S	*	0	0	ACTGACGTACAACCACCCGAGCTGGAGACCCGCATCGCCATCTTGCGCAAGAAAGCACAGAT
GGAACGGCTCGCGGTCCCCGACGATGTCCTCGAACTCATCGCCAGCAGTATCGAACGCAATATCCGTTAGAAGCTCGAG	?CBBBABBB883:2;:>4<;<@:95:BBBE<B@?@BBB<B???:?BCCC<CCC<CBBBBCBCB<E:??@ABBBB<999929>A???BBF<AAB
B;BBDDBBBB<ABBBBBBBBB>:?>?B=B=873,,(,,212,666//	RG:Z:NOID	PG:Z:tmap	MD:Z:1T0G15	NM:i:2	AS:i:10	XA:Z:map3-1	XS:i:-2147483647	XT:i:1
SFHMR:03248:01652	4	*	0	0	*	*	0	0	TGTCCTCGAACTCATCGCCAGCAGTATCGAACGCAATATCCGTGAACTCGAGGGCGCGCTGATCCGGGTCACCGCGTTCGCCTCAT
TGAACAAAACACCAATCGACAAAGCGCTGGCCGAGATTGTGCTTCGCGATCTGATCGCCGACGCCAACACCATGCAAATCA	<ACK>BBBB>CCBBBBBB=FCCB@@AAABB?BB@@:B9::=BBDE=ECCCBBB<CBBCDA@<<<2771:AAA<CAAA<ACCACBBC
>CC<AACC7AAB<A;>;<>:::,:EEECC=C=DCCCC<BBBBB:???AAABBBBCCDC=BCB9959<AAA:??BBBB<CCC	RG:Z:NOID	PG:Z:tmap
SFHMR:02004:01353	16	gi|385235550|ref|NC_017387.1|	900816	96	130S13M	*	0	0	GCTCTTCGACGGTAGTGTCGAAGTATTCGGCGGTGGCAGCATGATCGTCGCCGCTGCTGATT
TGCATGGTGTTGGCGTCGGCGATCAGATCGCGAAGCACAATCTCGGCCAGCGCTTTGTCGTTGGTGTTTTGTTCAATGAGG	46671===<;6=::////A:2:;B@;::3??9??;C:9///:::://:9457??426>2443>>8>;;3?:;4=5==?=@:BAAA?9
99===>751@?AA?7>>C>@:C?CAA>AB:BA=<<73:2?>>/=???;@B<CCCC=C	RG:Z:NOID	PG:Z:tmap	MD:Z:13	NM:i:0	AS:i:13	XA:Z:map3-1	XS:i:-2147483647	XT:i:1
```

BAM file format
The header

- Let's look at the bam header

```
samtools view -H mapped_reads.bam
```

```
@HD     VN:1.4
@SQ     SN:gi|385235550|ref|NC_017387.1|        LN:4138388
@RG     ID:NOID PG:tmap SM:NOSM
@PG     ID:tmap CL:map3 -f reference.fasta -r reads.fastq -i fastq -o 2 -s mapped_reads.bam     VN:3.4.1
```

BAM file header modifying

- Let's change the header so that it includes our sample name!

```
samtools view -H mapped_reads.bam \
| sed 's/SM:NOSM/SM:Sample1/' \
| samtools reheader - mapped_reads.bam \
> mapped_reads.reheaded.bam
```

Getting already modified header

- Alternatively, we can specify the sample name during mapping process

```
tmap map3 -f reference.fasta \
-r reads.fastq \
-R "@RG\tID:SomeID\tSM:Sample1" \
-i fastq \
-o 2 \
-s mapped_reads.bam
```

Variant calling

- Variant callers need sorted bam file

```
samtools sort mapped_reads.reheaded.bam \
mapped_reads.reheaded.sorted
```

Variant calling

```
samtools mpileup -uf reference.fasta \
mapped_reads.reheaded.sorted.bam \
| bcftools call -cv \
| vcfutils.pl varFilter > variants.samtools.vcf
```

Variant call format (VCF)

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##samtoolsVersion=1.0-11-geeb4b22+htslib-1.0-2-g1903fd4
##samtoolsCommand=samtools mpileup -uf reference.fasta mapped_reads.sorted.bam
##reference=file://reference.fasta
##contig=<ID=gi|385235550|ref|NC_017387.1|,length=4138388>
##ALT=<ID=X,Description="Represents allele(s) other than observed.">
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO=<ID=IDV,Number=1,Type=Integer,Description="Maximum number of reads supporting an indel">
##INFO=<ID=IMF,Number=1,Type=Float,Description="Maximum fraction of reads supporting an indel">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias for filtering splice-site artefacts in RNA-seq data (bigger is better)",Version=3>
##INFO=<ID=RPB,Number=1,Type=Float,Description="Mann-Whitney U test of Read Position Bias (bigger is better)">
##INFO=<ID=MQB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality Bias (bigger is better)">
##INFO=<ID=BQB,Number=1,Type=Float,Description="Mann-Whitney U test of Base Quality Bias (bigger is better)">
##INFO=<ID=MQSB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality vs Strand Bias (bigger is better)">
##INFO=<ID=SGB,Number=1,Type=Float,Description="Segregation based metric.">
##INFO=<ID=MQ0F,Number=1,Type=Float,Description="Fraction of MQ0 reads (smaller is better)">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=AF1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele frequency (assuming HWE)">
##INFO=<ID=AF2,Number=1,Type=Float,Description="Max-likelihood estimate of the first and second group ALT allele frequency (assuming HWE)">
##INFO=<ID=AC1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele count (no HWE assumption)">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Root-mean-square mapping quality of covering reads">
##INFO=<ID=FQ,Number=1,Type=Float,Description="Phred probability of all samples being the same">
##INFO=<ID=PV4,Number=4,Type=Float,Description="P-values for strand bias, baseQ bias, mapQ bias and tail distance bias">
##INFO=<ID=G3,Number=3,Type=Float,Description="ML estimate of genotype frequencies">
##INFO=<ID=HWE,Number=1,Type=Float,Description="Chi^2 based HWE test P-value based on G3">
##INFO=<ID=DP4,Number=4,Type=Integer,Description="Number of high-quality ref-forward , ref-reverse, alt-forward and alt-reverse bases">
##bcftools_callVersion=1.0-26-g53fcd5c+htslib-1.0-6-g546bfcb
##bcftools_callCommand=call -cv
#CHROM  POS  ID  REF  ALT  QUAL  FILTER  INFO  FORMAT  NOSM
gi|385235550|ref|NC_017387.1|  5461  .  T  G  20.8005  .  DP=4;VDB=0.14;SGB=-0.453602;MQ0F=0;AF1=1;AC1=2;DP4=0,0,0,2;MQ=60;FQ=-32.988  GT:PL  1/1:52,6,0
gi|385235550|ref|NC_017387.1|  14453  .  A  C  20.8005  .  DP=2;VDB=0.02;SGB=-0.453602;MQ0F=0;AF1=1;AC1=2;DP4=0,0,0,2;MQ=47;FQ=-32.988  GT:PL  1/1:52,6,0
gi|385235550|ref|NC_017387.1|  51777  .  TAA  T  14.1422  .  INDEL;IDV=3;IMF=0.75;DP=4;VDB=0.0381671;SGB=-0.556411;MQ0F=0;AF1=1;AC1=2;DP4=0,0,0,4;MQ=25;FQ=-46.52
1:54,12,0
gi|385235550|ref|NC_017387.1|  69668  .  G  C  17.8363  .  DP=4;VDB=0.02;SGB=-0.453602;MQ0F=0;AF1=1;AC1=2;DP4=0,0,0,2;MQ=60;FQ=-32.988  GT:PL  1/1:49,6,0
gi|385235550|ref|NC_017387.1|  108655  .  G  C  3.01703  .  DP=4;VDB=0.1;SGB=-0.453602;RPB=0.5;MQB=0.75;BQB=0.5;MQ0F=0.25;AF1=0.500542;AC1=1;DP4=0,2,0,2;MQ=45;FQ
=1,0.352957,1,1  GT:PL  0/1:30,0,25
```

Summary of used commands

- Reference genome indexing

```
tmap index -f reference.fasta
```

- Read mapping against reference genome

```
tmap map3 -f reference.fasta \
-r reads.fastq \
-i fastq \
-o 2 \
-s mapped_reads.bam
```

- Bam header modifying

```
samtools view -H mapped_reads.bam \
| sed 's/SM:NOSM/SM:Sample1/' \
| samtools reheader - mapped_reads.bam \
> mapped_reads.reheaded.bam
```

Summary of used commands

- Mapped read sorting

```
samtools sort mapped_reads.reheaded.bam \
mapped_reads.reheaded.sorted
```

- Variant calling

```
samtools mpileup -uf reference.fasta \
mapped_reads.reheaded.sorted.bam \
| bcftools call -cv \
| vcfutils.pl varFilter > variants.samtools.vcf
```

Let's build pipeline!

```
tmap map3 -f reference.fasta \
-r reads.fastq \
-i fastq \
-o 2 \
-s mapped_reads.bam

samtools view -H mapped_reads.bam \
| sed 's/SM:NOSM/SM:Sample1/' \
| samtools reheader - mapped_reads.bam \
> mapped_reads.reheaded.bam

samtools sort mapped_reads.reheaded.bam \
mapped_reads.reheaded.sorted
samtools mpileup -uf reference.fasta \
mapped_reads.reheaded.sorted.bam \
| bcftools call -cv \
| vcfutils.pl varFilter > variants.samtools.vcf
```

```
  GNU nano 2.2.6                File: my_pipeline.sh                      Modified

#Read mapping
tmap map3 -f reference.fasta \
-r reads.fastq \
-i fastq \
-o 2 \
-s mapped_reads.bam

#Header modifying
samtools view -H mapped_reads.bam \
| sed 's/SM:NOSM/SM:Sample1/' \
| samtools reheader - mapped_reads.bam \
> mapped_reads.reheaded.bam

#Bam file sorting
samtools sort mapped_reads.reheaded.bam mapped_reads.reheaded.sorted
#Variant detection
samtools mpileup -uf reference.fasta mapped_reads.reheaded.sorted.bam \
| bcftools call -cv \
| vcfutils.pl varFilter > variants.samtools.vcf

^G Get Help    ^O WriteOut    ^R Read File    ^Y Prev Page    ^K Cut Text    ^C Cur Pos
^X Exit        ^J Justify     ^W Where Is     ^V Next Page    ^U UnCut Text  ^T To Spell
```

To run pipeline, just type:

```
bash my_pipeline.sh
```

Your task

Perform variant calling using following tools:

- For read mapping - BWA (also remember to index reference genome)
- Samtools for bam sorting