Tutorial 4 – BWA, SAMtools & BCFtools

MICB405 – BIOINFORMATICS – 2021W-T1

01 OCTOBER 2021

AXEL HAUDUC

What is SAMtools?

"Swiss army knife" of SAM/BAM file manipulation

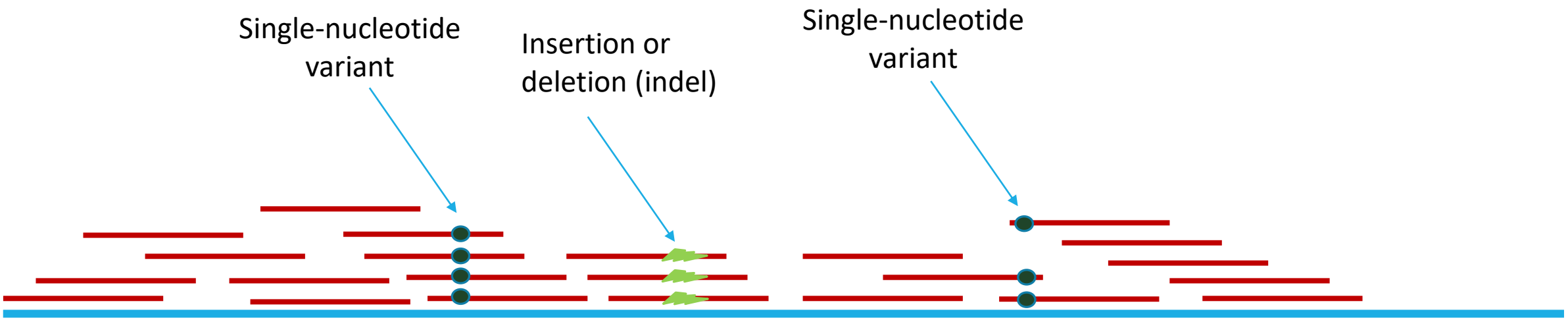
Can be used for various project-specific tasks such as filtering out unnecessary reads, and organizing reads

Here, we will be going over a SAMtools & BCFtools (more later) workflow for calling variants

Calling Variants

Before genomic data can be analyzed, it must be organized in a way that compactly represents differences between it and its reference

SAM/BAM files show sequence assemblies, but do not explicitly describe variation between the sequence of DNA



Let’s find them!

```
ahauduc_mb20@orca01:~/alignments$ samtools view bhinzii.bam | head -n 2
M01783:4:000000000-A4CKG:1:1101:1673:14600      89      NZ_CP012076.1   366929  37      250M    =       366929  0       G
TGTCGAAAGTATGGCTGATGGCCCGGGCCAGGGGCGGACTGTCGATGACCAGCCCGAGCTCGGTGTTCAGGTGGGCCGAGCGCGGGTCGAAATTGAAGGAGCCCACGAACACGCGGTGGT
CGTCCACGGCGAAGGTCTTGGCATGCAGGCTGGAGCCCGAGCTGCCGAAGGGGCCCAGGCCGCGGTGGCGCTGGACCTCGTCGCCGGCCCGGCGCATCTCGAATAGCTGCACGCCGCTGG
CCAGCAAGG       ?BFEFFFFFFBFAB:/FFFB@@@@-EAFFFF@@F@BFFB-B@FFFFFFFB?@@;@BFB?;<?FFFFFFFFFFF@FA-=B@@@@;@@EEFFFFBFFFFFFFFA;?
EFFFFF?@@=-EE?@@@AFB?@A-GBEFGGGGGGHFGFBHGHGHGHGFEC-C??CGCC@CC?EFHHCGGF/AEGCBCCCAGC/EE/E1FGFE//C@E@/>/?>A///EA/A1/C020GFG
DCF1000A0A1AF1CB>11D@1A?AA      XT:A:U  NM:i:0  SM:i:37 AM:i:0  X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:250
M01783:4:000000000-A4CKG:1:1101:1673:14600      181     NZ_CP012076.1   366929  0       *       =       366929  0       G
TGTCGAAAGTATGGCTGATGGCCCGGGCCAGGGGCGGACTGTCGATGACCAGCCCGAGCTCGGTGTTCAGGTGGGCCGAGCGCGGGTCGAAATTGAAGGAGCCCACGAACACGCGGTGGT
CGTCCACGGCGAAGGTCTTGGCATGCAGGCTGGAGCCCGAGCTGCCGAAGGGGCCCAGGCCGCGGTGGCGCTGGACCTCGTCGCCGGCCCGGCGCATCTCGAATAGCTGCACGCCGCTGG
CCAGCAAGG       ?BFEFFFFFFBFAB:/FFFB@@@@-EAFFFF@@F@BFFB-B@FFFFFFFB?@@;@BFB?;<?FFFFFFFFFFF@FA-=B@@@@;@@EEFFFFBFFFFFFFFA;?
EFFFFF?@@=-EE?@@@AFB?@A-GBEFGGGGGGHFGFBHGHGHGHGFEC-C??CGCC@CC?EFHHCGGF/AEGCBCCCAGC/EE/E1FGFE//C@E@/>/?>A///EA/A1/C020GFG
DCF1000A0A1AF1CB>11D@1A?AA
```

SAM/BAM format

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                   ATAGCT..............TCAGC
-r003                         ttagctTAGGC
-r001/2                                    CAGCGGCAT
```

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

SAM format columns

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0, 2^{16} - 1]$	bitwise FLAG
3	RNAME	String	`\*|[:rname:^*=][:rname:]*`	Reference sequence NAME[11]
4	POS	Int	$[0, 2^{31} - 1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8} - 1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[:rname:^*=][:rname:]*`	Reference name of the mate/next read
8	PNEXT	Int	$[0, 2^{31} - 1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31} + 1, 2^{31} - 1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

Decoding SAM flags

https://broadinstitute.github.io/picard/explain-flags.html

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

Alignment with **BWA MEM**

```
bwa mem \

/projects/micb405/analysis/references/ASM107827v1/GCA_001078275.1_
ASM107827v1_genomic.fna \

/projects/micb405/data/bordetella/F01_R1_1M.fastq \

/projects/micb405/data/bordetella/F01_R2_1M.fastq | \

samtools view -h -b - -o bordetella.bam
```

Removing PCR duplicates

```
# Tag mates
samtools fixmate -m bordetella.bam bordetella.fixmate.bam
# Position sort for markdup
samtools sort bordetella.fixmate.bam -o bordetella.fixmate.sort.bam
# Mark and remove duplicate reads
samtools markdup -r bordetella.fixmate.sort.bam bordetella.rmdup.bam
# Sort again
samtools sort bordetella.rmdup.bam -o bordetella.final.bam
```

Index BAM file

```
samtools index bordetella.final.bam
```

Check BAM file statistics

```
# Check flagstat for raw alignment
samtools flagstat bordetella.bam
# Check final BAM on flagstat
samtools flagstat bordetella.final.bam
```

Call variants

```
bcftools mpileup \
-f your_reference.fa \
bordetella.final.bam | \
bcftools call -m --variants-only --ploidy 1 - \
--output-type z -o bordetella.vcf.gz
```

Index VCF file

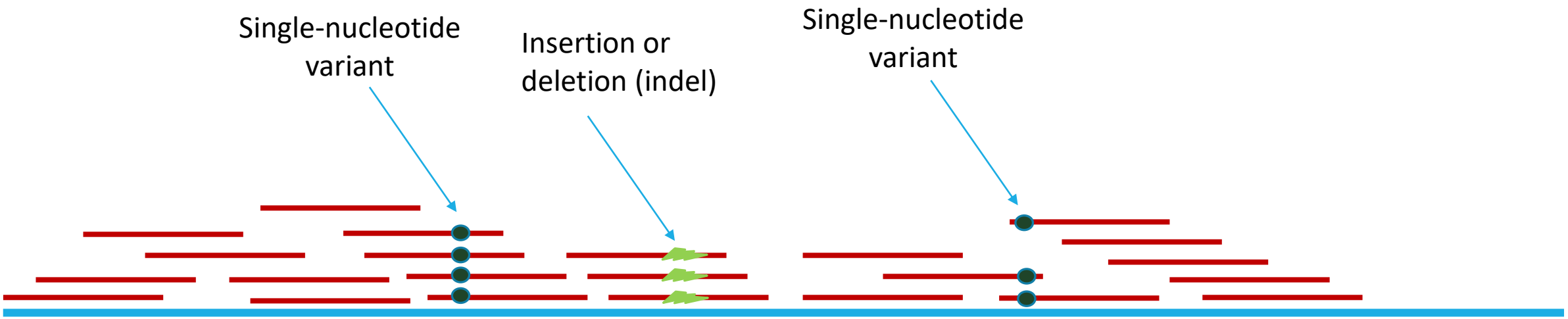
```
bcftools index -t bordetella.vcf.gz
```

```
ahauduc_mb20@orca01:~/alignments$ samtools view bhinzii.bam | head -n 2
M01783:4:000000000-A4CKG:1:1101:1673:14600      89      NZ_CP012076.1   366929  37      250M    =       366929  0       G
TGTCGAAAGTATGGCTGATGGCCCGGGCCAGGGGCGGACTGTCGATGACCAGCCCGAGCTCGGTGTTCAGGTGGGCCGAGCGCGGGTCGAAATTGAAGGAGCCCACGAACACGCGGTGGT
CGTCCACGGCGAAGGTCTTGGCATGCAGGCTGGAGCCCGAGCTGCCGAAGGGGCCCAGGCCGCGGTGGCGCTGGACCTCGTCGCCGGCCCGGCGCATCTCGAATAGCTGCACGCCGCTGG
CCAGCAAGG       ?BFEFFFFFFBFAB:/FFFB@@@@-EAFFFF@@F@BFFB-B@FFFFFFFB?@@;@BFB?;<?FFFFFFFFFFF@FA-=B@@@@;@@EEFFFFBFFFFFFFFA;?
EFFFFF?@@=-EE?@@@AFB?@A-GBEFGGGGGGHFGFBHGHGHGHGFEC-C??CGCC@CC?EFHHCGGF/AEGCBCCCAGC/EE/E1FGFE//C@E@/>/?>A///EA/A1/C020GFG
DCF1000A0A1AF1CB>11D@1A?AA      XT:A:U  NM:i:0  SM:i:37 AM:i:0  X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:250
M01783:4:000000000-A4CKG:1:1101:1673:14600      181     NZ_CP012076.1   366929  0       *       =       366929  0       G
TGTCGAAAGTATGGCTGATGGCCCGGGCCAGGGGCGGACTGTCGATGACCAGCCCGAGCTCGGTGTTCAGGTGGGCCGAGCGCGGGTCGAAATTGAAGGAGCCCACGAACACGCGGTGGT
CGTCCACGGCGAAGGTCTTGGCATGCAGGCTGGAGCCCGAGCTGCCGAAGGGGCCCAGGCCGCGGTGGCGCTGGACCTCGTCGCCGGCCCGGCGCATCTCGAATAGCTGCACGCCGCTGG
CCAGCAAGG       ?BFEFFFFFFBFAB:/FFFB@@@@-EAFFFF@@F@BFFB-B@FFFFFFFB?@@;@BFB?;<?FFFFFFFFFFF@FA-=B@@@@;@@EEFFFFBFFFFFFFFA;?
EFFFFF?@@=-EE?@@@AFB?@A-GBEFGGGGGGHFGFBHGHGHGHGFEC-C??CGCC@CC?EFHHCGGF/AEGCBCCCAGC/EE/E1FGFE//C@E@/>/?>A///EA/A1/C020GFG
DCF1000A0A1AF1CB>11D@1A?AA
```

SAM

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  bhinzii.sorted.rmdup.bam
NZ_CP012076.1   62084   .       T       C       15.8048 .       DP=252;VDB=0.0256773;SGB=-0.693139;RPB=0.298514;MQB=0.0143494;MQSB=5.37752e-08;BQB=2.10992e-09;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=72,14,21,15;MQ=54     GT:PL   0/1:51,0,255
NZ_CP012076.1   72730   .       A       C       18.7483 .       DP=248;VDB=0.0862903;SGB=-0.693136;RPB=0.712867;MQB=0.000563025;MQSB=0.66238;BQB=5.39485e-11;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=55,30,16,19;MQ=54       GT:PL   0/1:54,0,255
NZ_CP012076.1   72884   .       A       C       6.97062 .       DP=249;VDB=0.00330069;SGB=-0.693146;RPB=0.000306521;MQB=6.14155e-05;MQSB=1.35848e-15;BQB=1.55976e-12;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=75,20,13,30;MQ=49       GT:PL   0/1:41,0,255
NZ_CP012076.1   180497  .       T       G       77      .       DP=156;VDB=0.961403;SGB=-0.692067;RPB=0.0987007;MQB=0.167746;MQSB=0.12665;BQB=0.197143;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=11,3,17,3;MQ=46       GT:PL   0/1:110,0,156
NZ_CP012076.1   180632  .       A       C       18.6198 .       DP=262;VDB=0.0277801;SGB=-0.693097;RPB=0.245373;MQB=0.771275;MQSB=0.585838;BQB=5.66673e-07;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=77,7,20,10;MQ=54  GT:PL   0/1:54,0,255
NZ_CP012076.1   208525  .       G       C       4.91571 .       DP=250;VDB=0.0177041;SGB=-0.693054;RPB=0.883359;MQB=0.953977;MQSB=0.999366;BQB=2.8855e-07;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=63,28,11,17;MQ=58  GT:PL   0/1:39,0,255
NZ_CP012076.1   208530  .       G       C       29.7766 .       DP=243;VDB=0.121142;SGB=-0.693079;RPB=0.708874;MQB=0.931274;MQSB=0.977573;BQB=6.36885e-07;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=56,29,11,18;MQ=58  GT:PL   0/1:65,0,255
NZ_CP012076.1   208536  .       T       C       7.25462 .       DP=245;VDB=0.0230065;SGB=-0.693127;RPB=0.650458;MQB=0.980013;MQSB=0.99841;BQB=3.21071e-09;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=53,29,12,21;MQ=58  GT:PL   0/1:42,0,255
NZ_CP012076.1   208539  .       G       C       18.6457 .       DP=246;VDB=0.134747;SGB=-0.693021;RPB=0.566774;MQB=0.999733;MQSB=0.969107;BQB=5.74972e-06;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=55,33,9,18;MQ=58   GT:PL   0/1:54,0,255
NZ_CP012076.1   213780  .       C       T       35.7076 .       DP=161;VDB=0.0105084;SGB=-0.693147;RPB=0.813717;MQB=0.00148234;MQSB=0.0616543;BQB=1.28835e-06;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=52,15,2,51;MQ=38       GT:PL   0/1:69,0,255
NZ_CP012076.1   227404  .       G       C       14.722  .       DP=248;VDB=0.159881;SGB=-0.693079;RPB=0.999774;MQB=1;MQSB=1;BQB=5.36873e-07;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=54,38,6,23;MQ=60 GT:PL   0/1:50,0,255
NZ_CP012076.1   246361  .       A       C       3.68653 .       DP=249;VDB=0.0422406;SGB=-0.693143;RPB=0.931854;MQB=0.428684;MQSB=0.953899;BQB=1.49487e-11;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=39,48,14,24;MQ=58 GT:PL   0/1:37,0,255
NZ_CP012076.1   280905  .       A       C       45.5341 .       DP=203;VDB=0.147976;SGB=-0.693141;RPB=0.65553;MQB=0.0243711;MQSB=0.00300644;BQB=1.15583e-08;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=6,41,14,23;MQ=49 GT:PL   0/1:80,0,255
NZ_CP012076.1   280922  .       C       G       14.8967 .       DP=204;VDB=0.0419789;SGB=-0.692831;RPB=0.999864;MQB=0.0787541;MQSB=0.867229;BQB=1.08842e-06;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=12,41,3,21;MQ=53 GT:PL   0/1:50,0,255
NZ_CP012076.1   280941  .       T       C       10.9475 .       DP=224;VDB=0.189371;SGB=-0.693054;RPB=0.443031;MQB=0.368929;MQSB=0.626774;BQB=6.79775e-09;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=21,31,6,22;MQ=56   GT:PL   0/1:46,0,255
NZ_CP012076.1   281024  .       G       C       15.1545 .       DP=250;VDB=0.000501562;SGB=-0.693097;RPB=0.0797368;MQB=0.875695;MQSB=0.463424;BQB=8.48096e-06;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=47,23,14,16;MQ=54      GT:PL   0/1:50,0,255
NZ_CP012076.1   281043  .       A       G       4.3431  .       DP=250;VDB=1.62214e-05;SGB=-0.693132;RPB=0.0170843;MQB=0.493851;MQSB=0.479649;BQB=1.6574e-11;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=52,20,16,18;MQ=54       GT:PL   0/1:38,0,255
NZ_CP012076.1   285334  .       A       C       6.72407 .       DP=254;VDB=0.00854758;SGB=-0.693132;RPB=0.894356;MQB=0.805458;MQSB=0.993685;BQB=5.55742e-08;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=30,33,13,21;MQ=56       GT:PL   0/1:41,0,255
NZ_CP012076.1   295704  .       G       C       29.3095 .       DP=216;VDB=0.001811;SGB=-0.693147;RPB=0.311041;MQB=0.000220568;MQSB=2.21587e-05;BQB=8.45161e-09;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=29,46,44,3;MQ=41      GT:PL   0/1:63,0,255
NZ_CP012076.1   346097  .       A       C       39.9683 .       DP=246;VDB=0.00164725;SGB=-0.693147;RPB=0.892335;MQB=0.000430366;MQSB=9.91059e-09;BQB=1.82972e-07;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=38,45,32,15;MQ=44  GT:PL   0/1:74,0,255
NZ_CP012076.1   353202  .       G       C       12.8597 .       DP=249;VDB=0.0634319;SGB=-0.693139;RPB=0.257128;MQB=0.0854744;MQSB=0.204499;BQB=1.19043e-11;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=21,66,17,19;MQ=56       GT:PL   0/1:48,0,255
NZ_CP012076.1   367189  .       T       C       4.80995 .       DP=250;VDB=2.8507e-05;SGB=-0.69311;RPB=0.586195;MQB=0.953372;MQSB=0.699266;BQB=7.10746e-09;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=30,46,11,20;MQ=58 GT:PL   0/1:39,0,255
NZ_CP012076.1   444652  .       GTTTTTTTTTGCC   GTTTTTTTTGCC    25.4901 .       INDEL;IDV=142;IMF=0.572581;DP=248;VDB=0.665273;SGB=-0.693147;MQSB=0.960082;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=40,88,29,91;MQ=59 GT:PL   0/1:55,0,4
NZ_CP012076.1   553283  .       G       A       29.7041 .       DP=249;VDB=2.40206e-05;SGB=-0.693147;RPB=0.0834371;MQB=2.685e-07;MQSB=5.35822e-11;BQB=1.32048e-13;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=68,25,16,43;MQ=50  GT:PL   0/1:64,0,255
NZ_CP012076.1   553314  .       T       C       24.9486 .       DP=250;VDB=0.0341296;SGB=-0.693143;RPB=0.575889;MQB=0.0147348;MQSB=7.42463e-08;BQB=1.25666e-09;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=64,15,19,19;MQ=53      GT:PL   0/1:60,0,255
NZ_CP012076.1   665128  .       C       G       19.7163 .       DP=252;VDB=0.0179823;SGB=-0.693136;RPB=0.00519041;MQB=5.29875e-05;MQSB=0.000261027;BQB=3.06327e-11;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=33,30,7,28;MQ=49  GT:PL   0/1:54,0,255
NZ_CP012076.1   666034  .       T       G       25.7317 .       DP=248;VDB=0.00661159;SGB=-0.693054;RPB=0.0894573;MQB=0.18299;MQSB=0.274747;BQB=2.51206e-07;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=22,47,10,18;MQ=56       GT:PL   0/1:61,0,255
NZ_CP012076.1   717216  .       A       C       8.05014 .       DP=252;VDB=0.12407;SGB=-0.69312;RPB=0.539357;MQB=0.803192;MQSB=0.765549;BQB=1.73233e-08;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=44,33,16,16;MQ=58     GT:PL   0/1:43,0,255
NZ_CP012076.1   762441  .       G       A       45.7888 .       DP=193;VDB=0.662894;SGB=-0.693145;RPB=0.485692;MQB=5.09488e-05;MQSB=3.61198e-07;BQB=2.11327e-05;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=31,14,6,34;MQ=42     GT:PL   0/1:79,0,255
```

VCF



Now we can easily view these!

Optional: view in IGV interactive viewer!

https://software.broadinstitute.org/software/igv/download

Copy reference, bam, bam.bai, vcf.gz, and vc.gz.tbi files to your computer and load into IGV