Informatics on High-throughput Sequencing Data

(Summer Course 2020)

Day 16

Variant Calling

- Variant calling entails identifying single nucleotide polymorphisms (SNPs) and small insertions and deletion (indels) from next generation sequencing data

AGTTTGTTTGAAAGTCGT	Ref. (Healthy Tissue)
AGTTTGTCTGAAAGTCGT	(Diseased Tissue)
AGTTTGTTTGAAAGTCGT	Ref. (Healthy Tissue)
AGTTTGTTTG--AGTCGT	(Diseased Tissue)
AGTTTG TGAAAGTCGT	Ref. (Healthy Tissue)
AGTTTGTTTGAAAGTCGT	(Diseased Tissue)

Variant Calling

- Variant calling entails identifying single nucleotide polymorphisms (SNPs) and small insertions and deletion (indels) from next generation sequencing data

AGTTTGTTTGAAAGTCGT Ref. (Healthy Tissue)

AGTTT GTCGT (Diseased Tissue)

AGTTTGTTTGAAAG TCGT Ref. (Healthy Tissue)

AGTTTGTTTGAAAGGAAAGTCGT (Diseased Tissue)

samtools mpileup

- Pileup format is a text-based format for summarizing the base calls of aligned reads to a reference sequence.

- This format facilitates SNP/indel calling and brief alignment viewing by eyes.

- Pileup format consists of TAB-separated lines, with each line representing the pileup of reads at a single genomic position.

samtools mpileup

```
Chr5   12266268   A   18   ..................        CCCCCCC5C>CC?CBACC
Chr5   12266269   G   18   ..................        CCCCCCC9C/CC;C<8CC
Chr5   12266270   C   18   ..................        CCCCCCC>C@CC?CBACC
Chr5   12266271   T   19   ..................^!.     CCCCCCCBC@CCBCBACCC
Chr5   12266272   A   20   ...................^!.    CCCCCCC>C?CCBCBACCCD
Chr5   12266273   C   20   .$...................     CCCCCCC?C>CC@CBACCCC
Chr5   12266274   T   19   ...................       CCCCCC@C7CC?C@ACCCC
Chr5   12266275   G   17   .................         CCCCCC1CCCCBACCCC
Chr5   12266276   A   18   ..................        BCCCCB<CCC8CBACCCC
Chr5   12266277   G   20   ...................^!.    CCCCCC7C0CC=C@ACCCC<
Chr5   12266278   A   20   .......$.............     CCCCCC>C9CC@C>ACCCC<
Chr5   12266279   G   19   .$..A......A.AA.....      CCCCCCC5CC@CBACCCC:
Chr5   12266280   T   18   .$......$...........      CCCDCC9CD?B@ACCCC4
```

Chromosome name : Ref. name

samtools mpileup

```
Chr5    12266268    A    18    ..................       CCCCCCC5C>CC?CBACC
Chr5    12266269    G    18    ..................       CCCCCCC9C/CC;C<8CC
Chr5    12266270    C    18    ..................       CCCCCCC>C@CC?CBACC
Chr5    12266271    T    19    ..................^!.    CCCCCCCBC@CCBCBACCC
Chr5    12266272    A    20    ...................^!.   CCCCCCC>C?CCBCBACCCD
Chr5    12266273    C    20    .$...................    CCCCCCC?C>CC@CBACCCC
Chr5    12266274    T    19    ...................      CCCCCC@C7CC?C@ACCCC
Chr5    12266275    G    17    .................        CCCCCC1CCCCBACCCC
Chr5    12266276    A    18    ..................       BCCCCB<CCC8CBACCCC
Chr5    12266277    G    20    ...................^!.   CCCCCC7C0CC=C@ACCCC<
Chr5    12266278    A    20    .......$.............    CCCCCC>C9CC@C>ACCCC<
Chr5    12266279    G    19    .$..A......A.AA.....     CCCCCCC5CC@CBACCCC:
Chr5    12266280    T    18    .$......$...........     CCCDCC9CD?B@ACCCC4
```

1-based position on the chromosome (Ref.).

samtools mpileup

```
Chr5    12266268    A    18    ..................       CCCCCCC5C>CC?CBACC
Chr5    12266269    G    18    ..................       CCCCCCC9C/CC;C<8CC
Chr5    12266270    C    18    ..................       CCCCCCC>C@CC?CBACC
Chr5    12266271    T    19    ..................^!.    CCCCCCCBC@CCBCBACCC
Chr5    12266272    A    20    ...................^!.   CCCCCCC>C?CCBCBACCCD
Chr5    12266273    C    20    .$...................    CCCCCCC?C>CC@CBACCCC
Chr5    12266274    T    19    ...................      CCCCCC@C7CC?C@ACCCC
Chr5    12266275    G    17    .................        CCCCCC1CCCCBACCCC
Chr5    12266276    A    18    ..................       BCCCCB<CCC8CBACCCC
Chr5    12266277    G    20    ...................^!.   CCCCCC7C0CC=C@ACCCC<
Chr5    12266278    A    20    .......$.............    CCCCCC>C9CC@C>ACCCC<
Chr5    12266279    G    19    .$..A......A.AA.....     CCCCCCC5CC@CBACCCC:
Chr5    12266280    T    18    .$......$...........     CCCDCC9CD?B@ACCCC4
```

Reference base at this position.

samtools mpileup

```
Chr5    12266268    A    18    ..................      CCCCCCC5C>CC?CBACC
Chr5    12266269    G    18    ..................      CCCCCCC9C/CC;C<8CC
Chr5    12266270    C    18    ..................      CCCCCCC>C@CC?CBACC
Chr5    12266271    T    19    ..................^!.   CCCCCCCBC@CCBCBACCC
Chr5    12266272    A    20    ...................^!.  CCCCCCC>C?CCBCBACCCD
Chr5    12266273    C    20    .$...................   CCCCCCC?C>CC@CBACCCC
Chr5    12266274    T    19    ...................     CCCCCC@C7CC?C@ACCCC
Chr5    12266275    G    17    .................       CCCCCC1CCCCBACCCC
Chr5    12266276    A    18    ..................      BCCCCB<CCC8CBACCCC
Chr5    12266277    G    20    ...................^!.  CCCCCC7C0CC=C@ACCCC<
Chr5    12266278    A    20    .......$.............   CCCCCC>C9CC@C>ACCCC<
Chr5    12266279    G    19    .$..A......A.AA.....    CCCCCCC5CC@CBACCCC:
Chr5    12266280    T    18    .$......$...........    CCCDCC9CD?B@ACCCC4
```

Number of reads covering this position.

samtools mpileup

```
Chr5    12266268    A    18    ..................        CCCCCCC5C>CC?CBACC
Chr5    12266269    G    18    ..................        CCCCCCC9C/CC;C<8CC
Chr5    12266270    C    18    ..................        CCCCCCC>C@CC?CBACC
Chr5    12266271    T    19    ..................^!.     CCCCCCCBC@CCBCBACCC
Chr5    12266272    A    20    ...................^!.    CCCCCCC>C?CCBCBACCCD
Chr5    12266273    C    20    .$...................     CCCCCCC?C>CC@CBACCCC
Chr5    12266274    T    19    ...................       CCCCCC@C7CC?C@ACCCC
Chr5    12266275    G    17    .................         CCCCCC1CCCCBACCCC
Chr5    12266276    A    18    ..................        BCCCCB<CCC8CBACCCC
Chr5    12266277    G    20    ...................^!.    CCCCCC7C0CC=C@ACCCC<
Chr5    12266278    A    20    .......$.............     CCCCCC>C9CC@C>ACCCC<
Chr5    12266279    G    19    .$..A......A.AA.....      CCCCCCC5CC@CBACCCC:
Chr5    12266280    T    18    .$......$...........      CCCDCC9CD?B@ACCCC4
```

For each read covering the position, this column contains:

If this is the first position covered by the read, a “^” character followed by the alignment's mapping quality encoded as an ASCII character. !:33

samtools mpileup

```
Chr5    12266268    A    18    ..................        CCCCCCC5C>CC?CBACC
Chr5    12266269    G    18    ..................        CCCCCCC9C/CC;C<8CC
Chr5    12266270    C    18    ..................        CCCCCCC>C@CC?CBACC
Chr5    12266271    T    19    ..................^!.     CCCCCCCBC@CCBCBACCC
Chr5    12266272    A    20    ...................^!.    CCCCCCC>C?CCBCBACCCD
Chr5    12266273    C    20    .$...................     CCCCCCC?C>CC@CBACCCC
Chr5    12266274    T    19    ...................       CCCCCC@C7CC?C@ACCCC
Chr5    12266275    G    17    .................         CCCCCC1CCCCBACCCC
Chr5    12266276    A    18    ..................        BCCCCB<CCC8CBACCCC
Chr5    12266277    G    20    ...................^!.    CCCCCC7C0CC=C@ACCCC<
Chr5    12266278    A    20    .......$.............     CCCCCC>C9CC@C>ACCCC<
Chr5    12266279    G    19    .$..A......A.AA.....      CCCCCCC5CC@CBACCCC:
Chr5    12266280    T    18    .$......$...........      CCCDCC9CD?B@ACCCC4
```

For each read covering the position, this column contains:

A single character indicating the read base and the strand to which the read has been mapped:
^!.
.$

Char	Meaning
. (dot)	Match, forward
, (comma)	Match, Reverse
ACGTN	Mismatch, forward
acgtn	Mismatch, Reverse
^	Beginning of read
$	End of read
+[0-9]+[ACGTNactgn]+	Insertion (i.e. +3ACC)
-[0-9]+[ACGTNactgn]+	Deletion (i.e. -2GG)
>	Reference Skip

SAM tools

```
./samtools mpileup -f /Users/sarael-metwally/Documents/Summer/bwa/wu_0.v7.fas sample.sorted.bam > sample.mpileup
```

```
awk ‘{if($4 > 12) print $0;}’ sample.mpileup > results.txt
```

Thanks!

// | ?