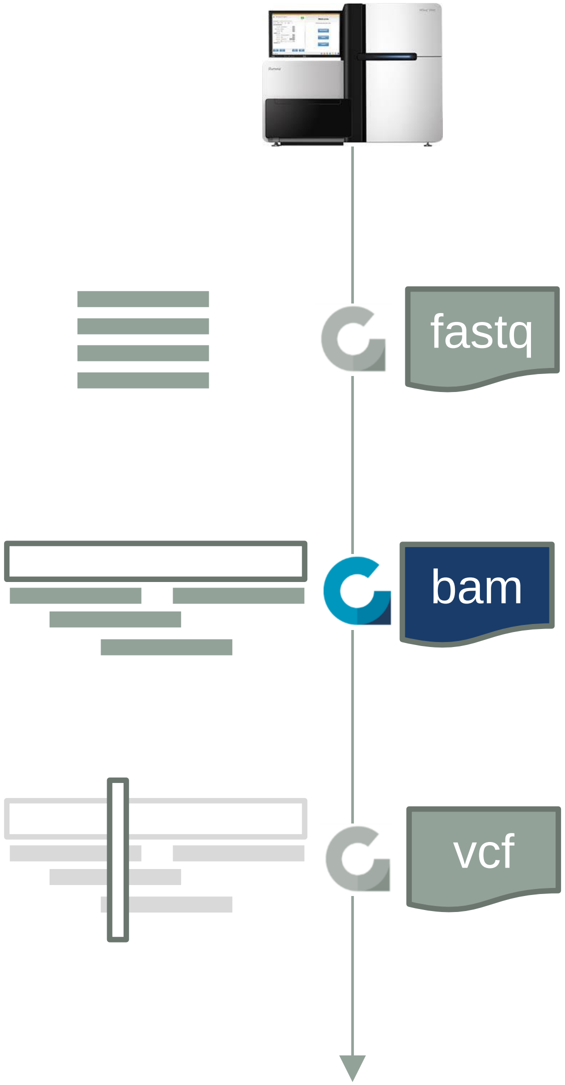
Read mapping and alignment: SAM format, alignment, IGV

GENOMICS CORELEUVEN

Luc Dehaspe
Erika Souche

Overview



Sequencing

Sequence of nucleotides within a read?

Genomic location of a read?

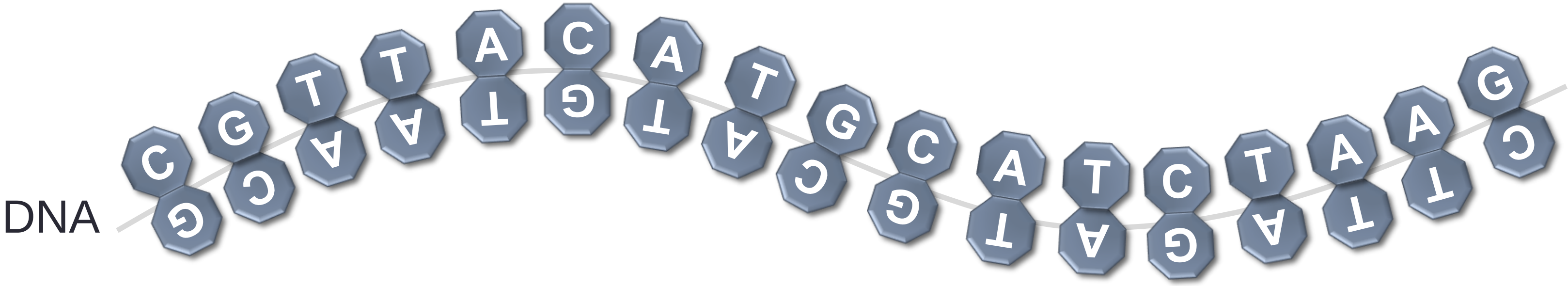
Genotype of a genomic location?

Sequence Alignment & Mapping

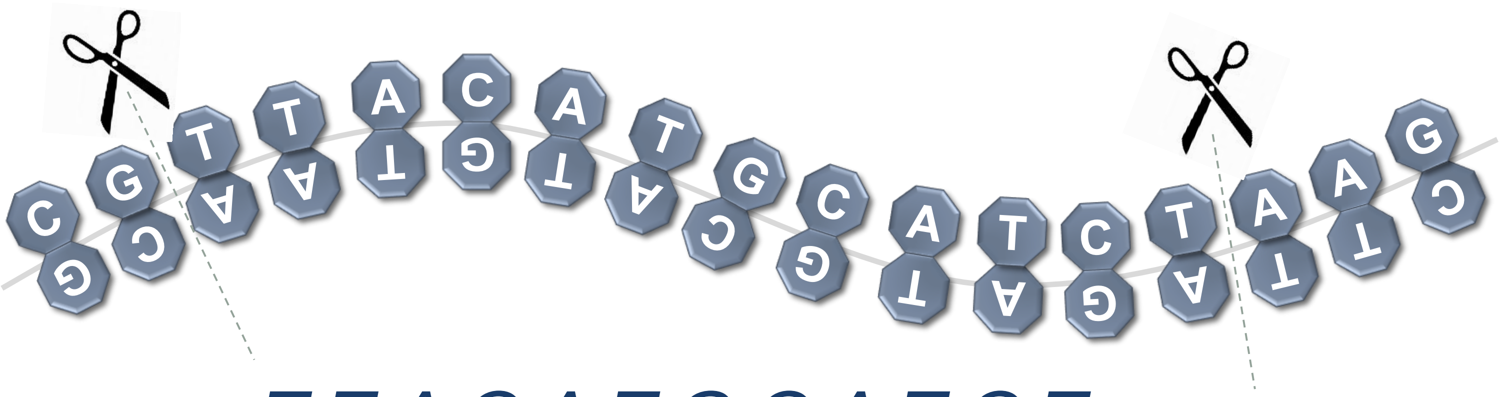
look up read in reference sequence ….

… allowing mismatches

Sequence Alignment & Mapping
look up read in reference sequence


DNA
C G T T A C A T G C A T C T A A G
G C A A T G T A C G T A G A T T C

Sequence Alignment & Mapping
look up read in reference sequence

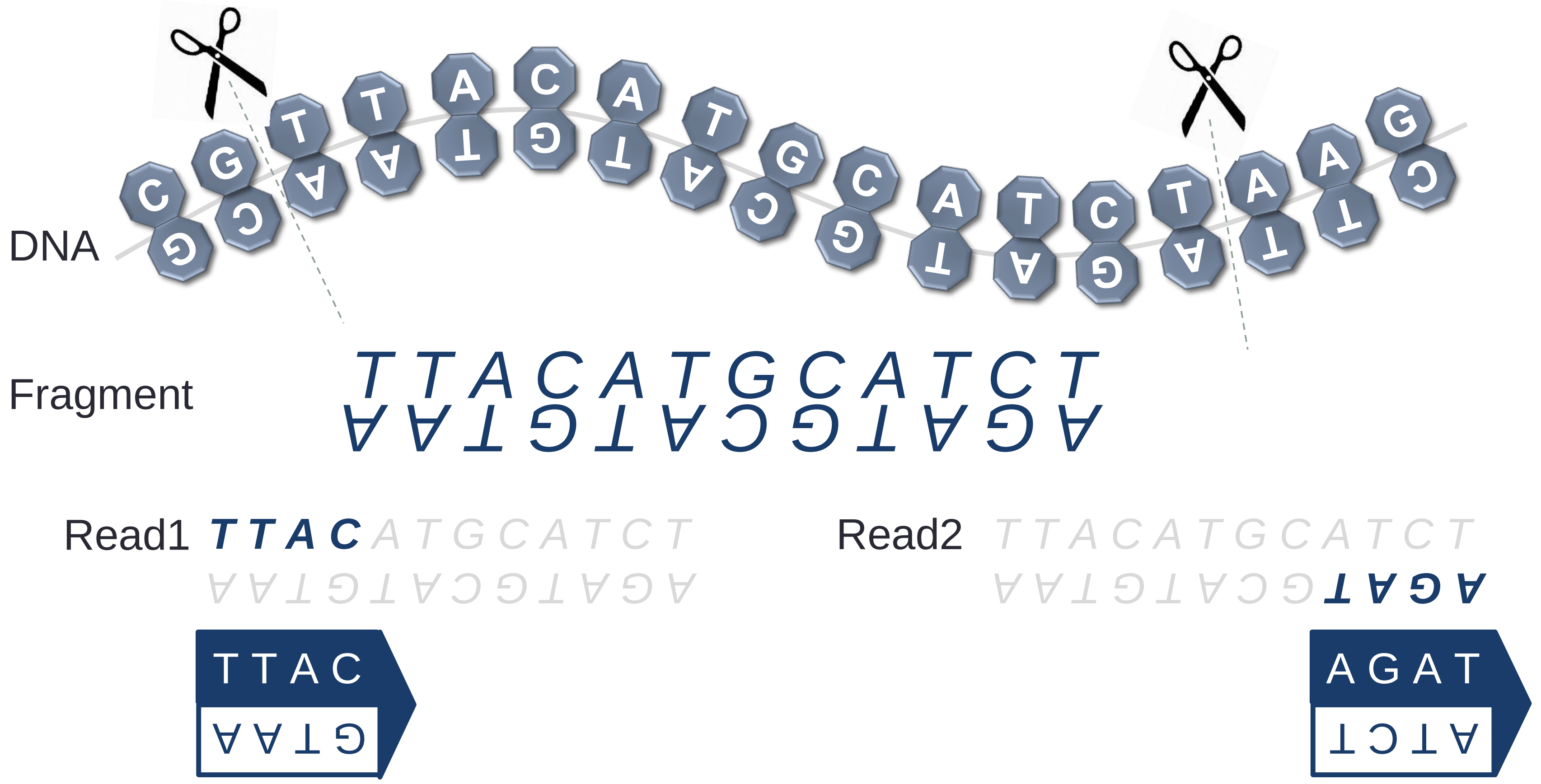

C G T T A C A T G C A T C T A A G
G C A A T G T A C G T A G A T T C

DNA

Fragment

T T A C A T G C A T C T
A A T G T A C G T A G A

Sequence Alignment & Mapping
look up read in reference sequence


C G T T A C A T G C A T C T A A G
G C A A T G T A C G T A G A T T C

DNA

Fragment
T T A C A T G C A T C T
A A T G T A C G T A G A

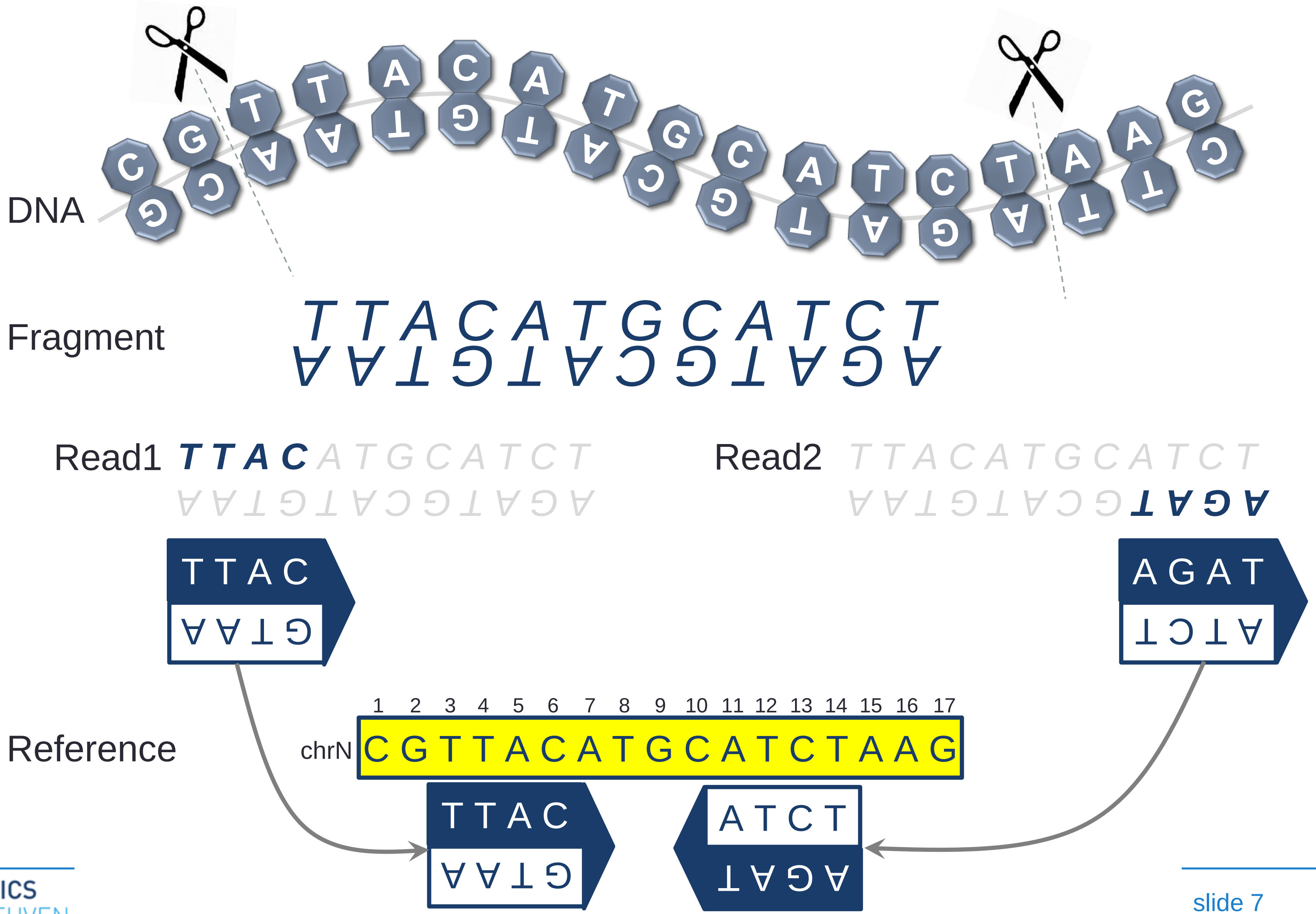
Read1 **T T A C** A T G C A T C T
A A T G T A C G T A G A

Read2 T T A C A T G C A T C T
A A T G T A C G **T A G A**

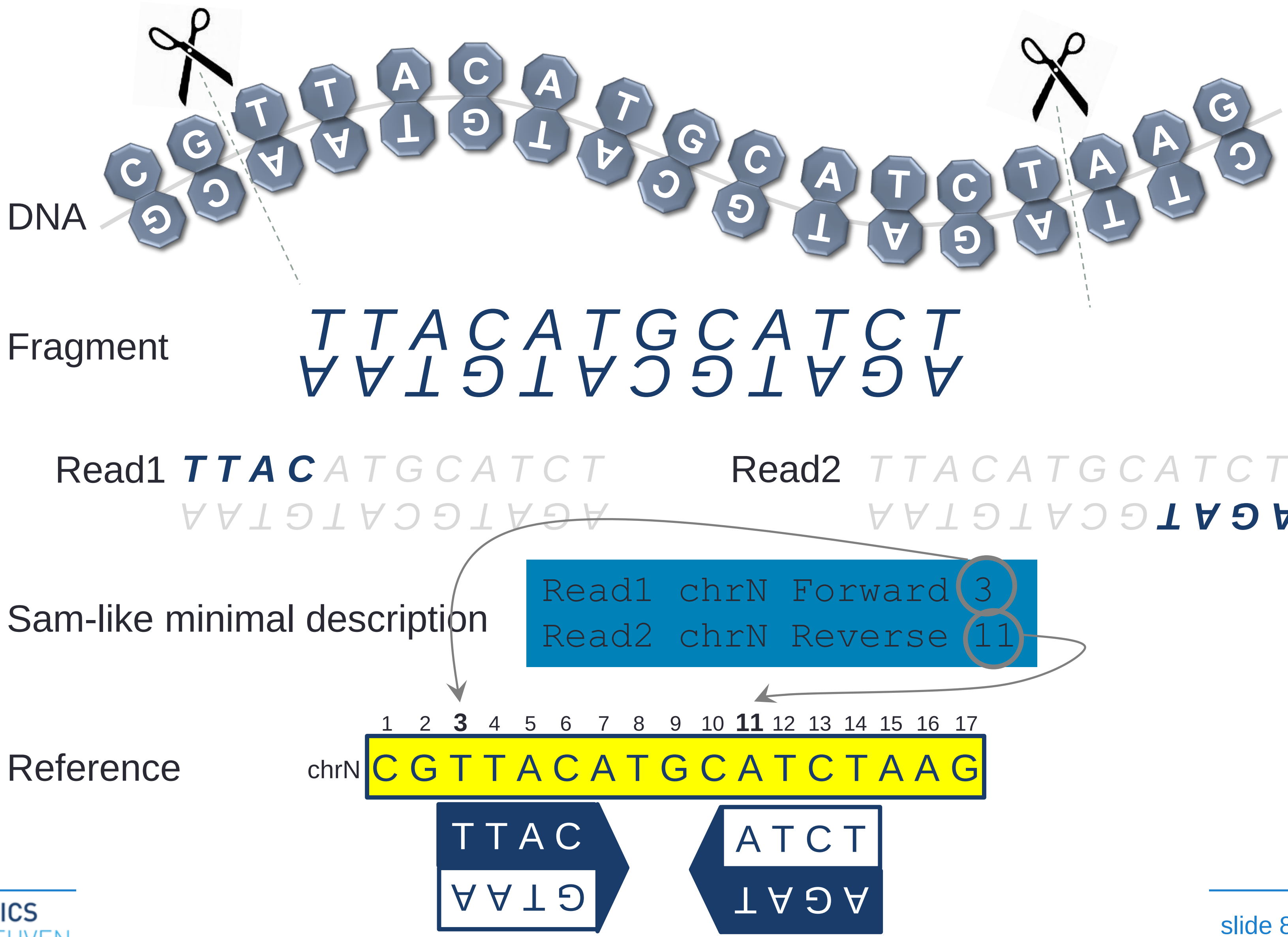
T T A C
A A T G

A G A T
T C T A

Sequence Alignment & Mapping
look up read in reference sequence



Sequence Alignment & Mapping
look up read in reference sequence



DNA

C G T T A C A T G C A T C T A A G
G C A A T G T A C G T A G A T T C

Fragment

Read1 **T T A C** A T G C A T C T

Read2 T T A C A T G C A T C T

Sam-like minimal description

```
Read1  chrN  Forward  3
Read2  chrN  Reverse  11
```

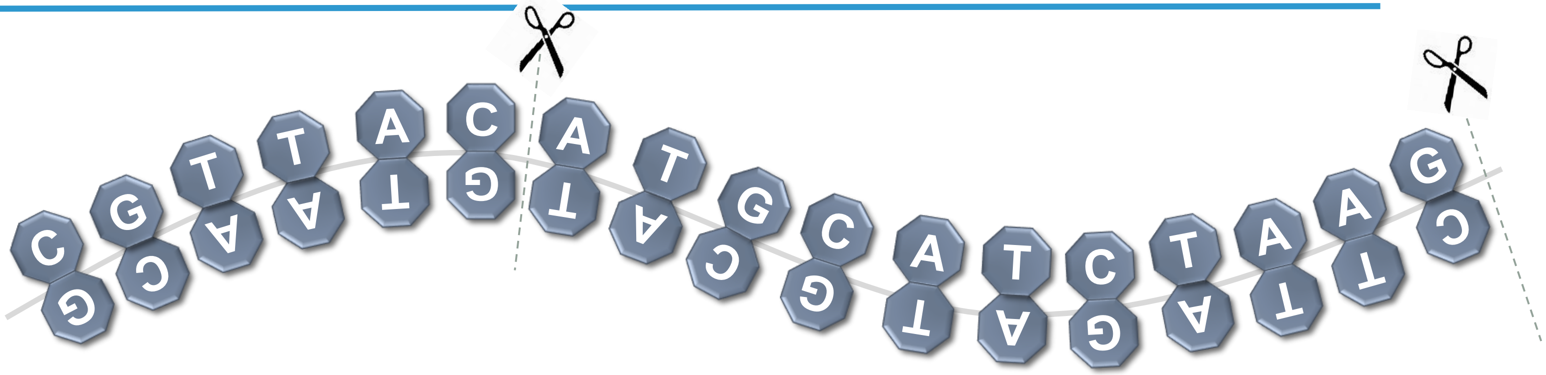
Reference

1	2	**3**	4	5	6	7	8	9	10	**11**	12	13	14	15	16	17
C	G	T	T	A	C	A	T	G	C	A	T	C	T	A	A	G

chrN

Sequence Alignment & Mapping
look up read in reference sequence

DNA


C G T T A C A T G C A T C T A A G
G C A A T G T A C G T A G A T T C

Fragment

A T G C A T C T A A G
T A C G T A G A T T C

Read3 **_A T G C_** _A T C T A A G_
_T A C G T A G A T T C_

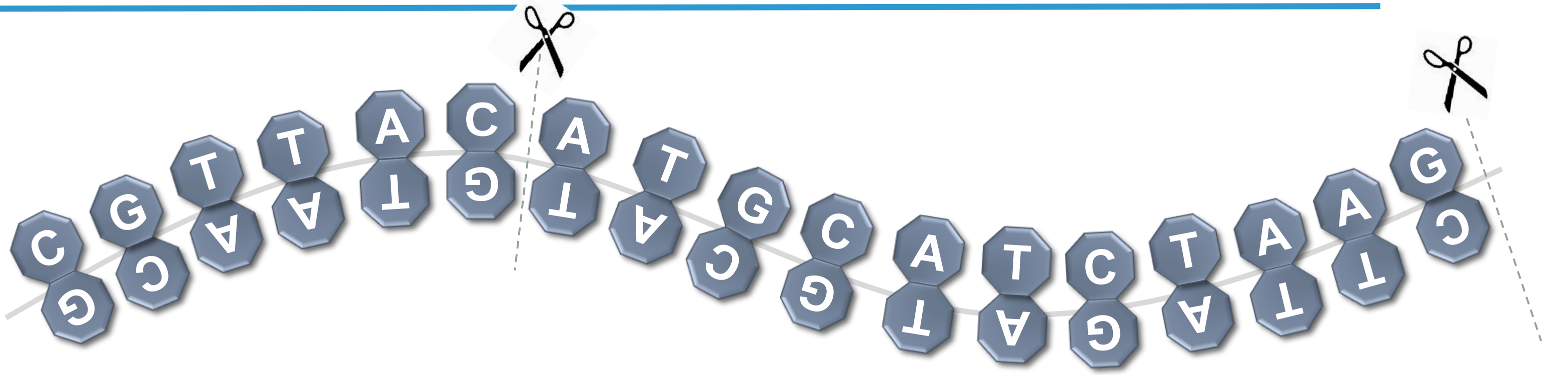
A T G C
T A C G

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

Reference chrN C G T T A C A T G C A T C T A A G

Sequence Alignment & Mapping
look up read in reference sequence

DNA



C G T T A C A T G C A T C T A A G
G C A A T G T A C G T A G A T T C

Fragment A T G C A T C T A A G
T A C G T A G A T T C

Read3 **A T G C** A T C T A A G
T A C G T A G A T T C

```
Read3 chrN  Forward 7
```

Or ?

```
Read3 chrN  Reverse 9
```

1 2 3 4 5 6 **7** 8 **9** 10 11 12 13 14 15 16 17

Reference chrN C G T T A C A T G C A T C T A A G

No unique mapping

```
Read3 chrN  Forward 7 lowMQual
```

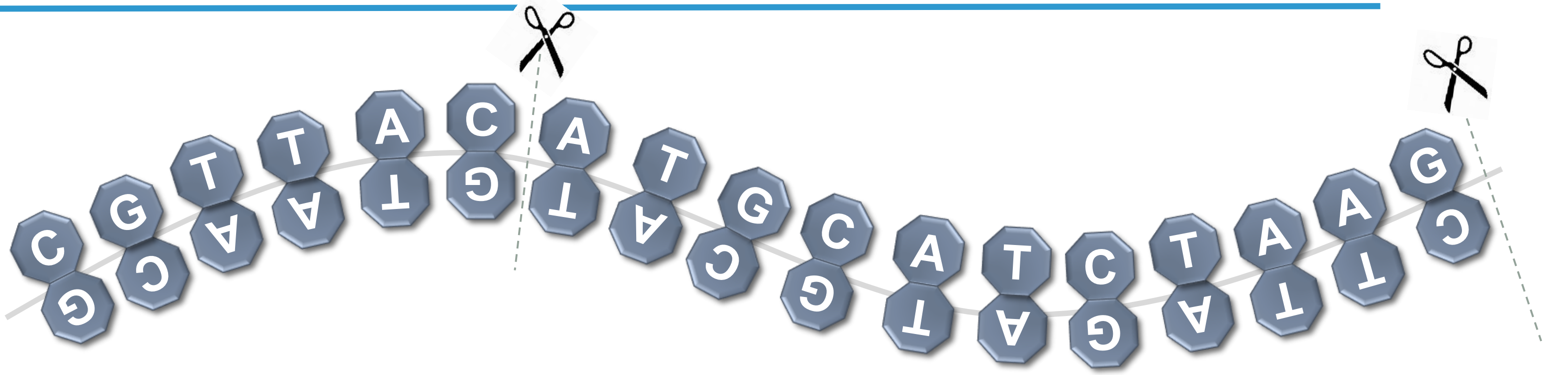
How can we improve mapping quality?

A T G C
T A C G

G C A T
C G T A

Sequence Alignment & Mapping
look up read in reference sequence

DNA



C G T T A C A T G C A T C T A A G
G C A A T G T A C G T A G A T T C

Fragment *A T G C A T C T A A G*
T A C G T A G A T T C

Read4 ***A T G C A T C*** *T A A G*
T A C G T A G A T T C

A T G C A T C
T A C G T A G

Reference chrN

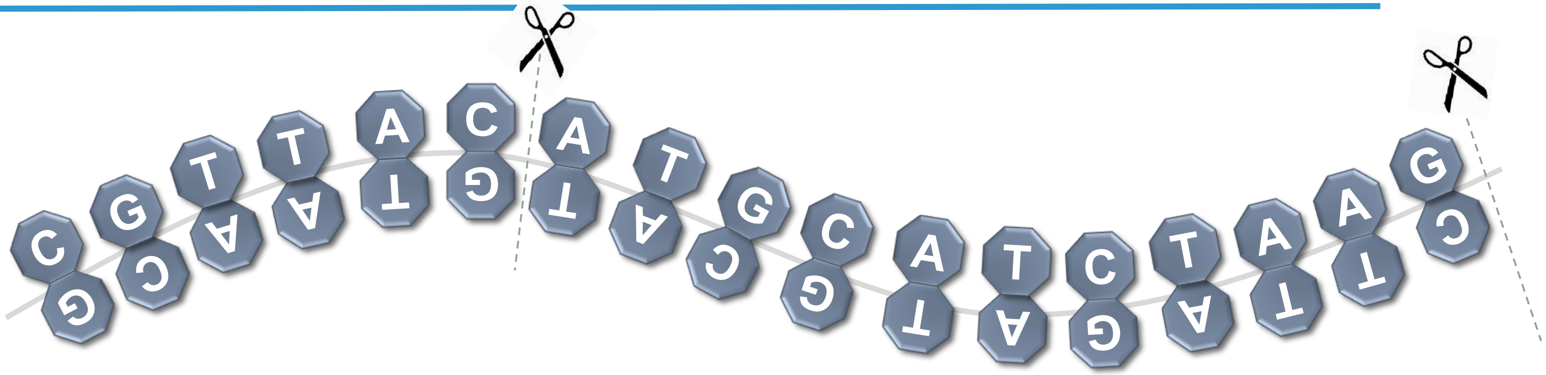
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
C	G	T	T	A	C	A	T	G	C	A	T	C	T	A	A	G

How can we improve mapping quality?

Strategy 1: READ ON

Sequence Alignment & Mapping
look up read in reference sequence

DNA

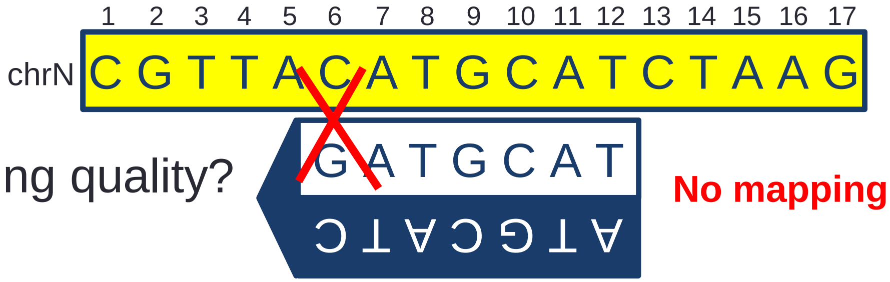

C G T T A C A T G C A T C T A A G
G C A A T G T A C G T A G A T T C

Fragment

A T G C A T C T A A G
T A C G T A G A T T C

Read4 ***A T G C <u>A T C</u>*** *T A A G*
T A C G T A G A T T C

Reference



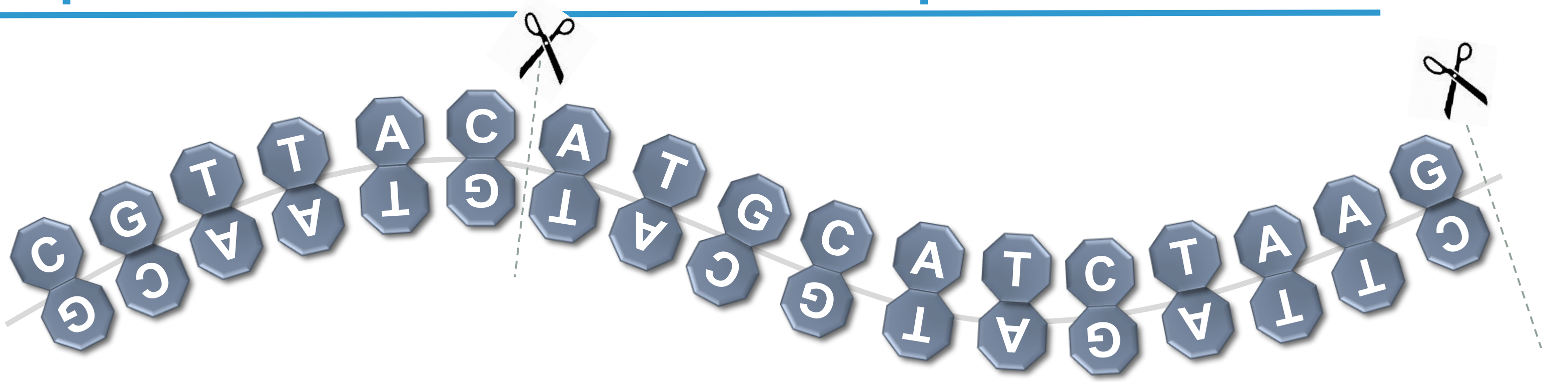
No mapping

How can we improve mapping quality?

Strategy 1: READ ON

Sequence Alignment & Mapping
look up read in reference sequence

DNA


C G T T A C A T G C A T C T A A G
G C A A T G T A C G T A G A T T C

Fragment

ATGCATCTAAG
TACGTAGATTC

Read4 **_ATGCATC_**TAAG
TACGTAGATTC

1 2 3 4 5 6 **7** 8 9 10 11 12 13 14 15 16 17

Reference chrN CGTTACATGCATCTAAG

ATGCATC
TACGTAG

Unique mapping

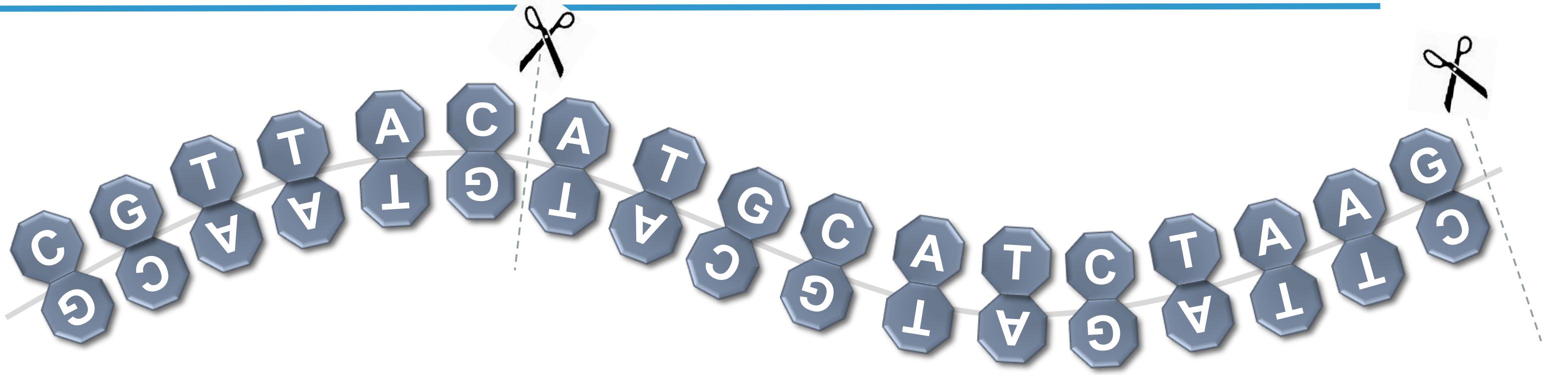
How can we improve mapping quality?

Strategy 1: READ ON

```
Read3 chrN  Forward 7 highMQual
```

Sequence Alignment & Mapping
look up read in reference sequence

DNA



C G T T A C A T G C A T C T A A G
G C A A T G T A C G T A G A T T C

Fragment

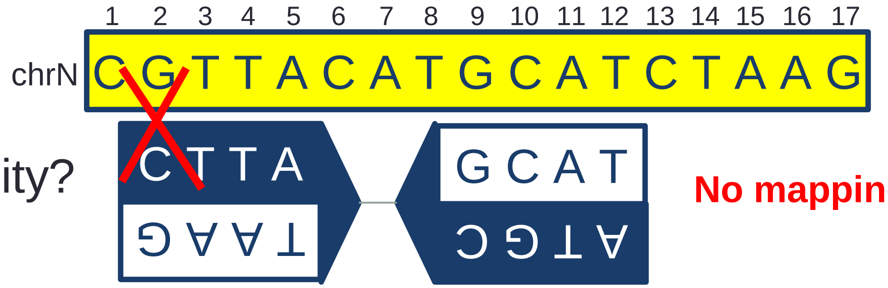
A T G C A T C T A A G
T A C G T A G A T T C

Read5a ***A T G C*** A T C T A A G

Read5b T A C G T A G ***A T T C***

Reference

How can we improve mapping quality?

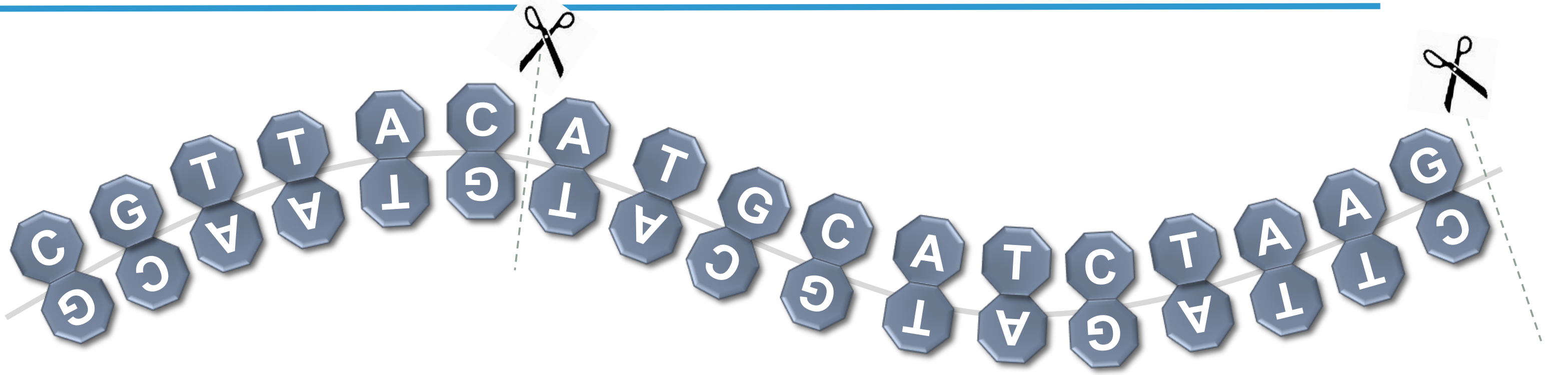


No mapping

Strategy 2: READ BOTH SIDES

Sequence Alignment & Mapping
look up read in reference sequence

DNA



C G T T A C A T G C A T C T A A G
G C A A T G T A C G T A G A T T C

Fragment

A T G C A T C T A A G
T A C G T A G A T T C

Read5a **A T G C** A T C T A A G
Read5b T A C G T A G **A T T C**

Reference

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
chrN	C	G	T	T	A	C	A	T	G	C	A	T	C	T	A	A	G

How can we improve mapping quality?

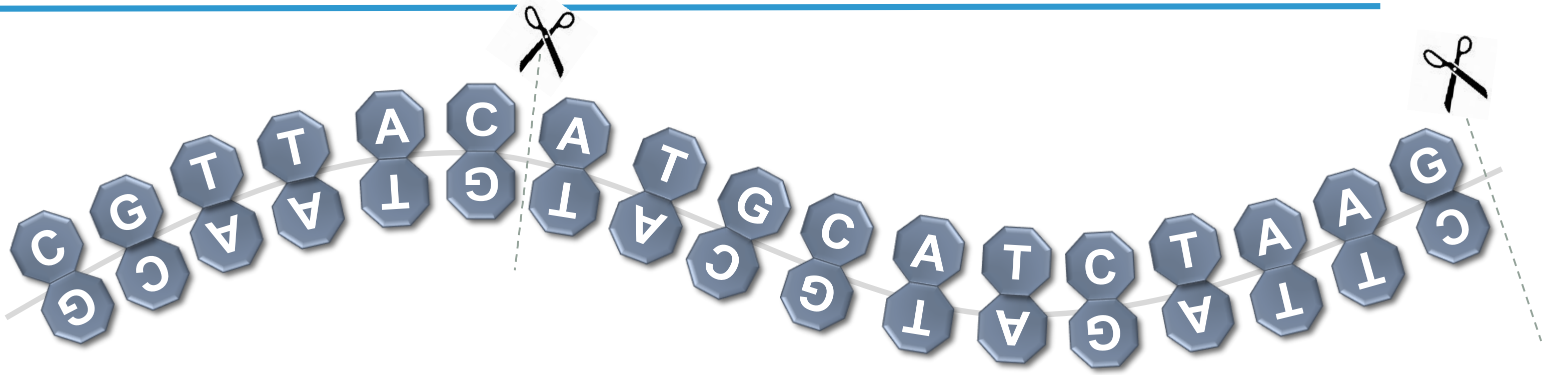
A T G C   T A A G
T A C G   A T T C

Unique mapping

Strategy 2: READ BOTH SIDES

Sequence Alignment & Mapping
look up read in reference sequence

DNA

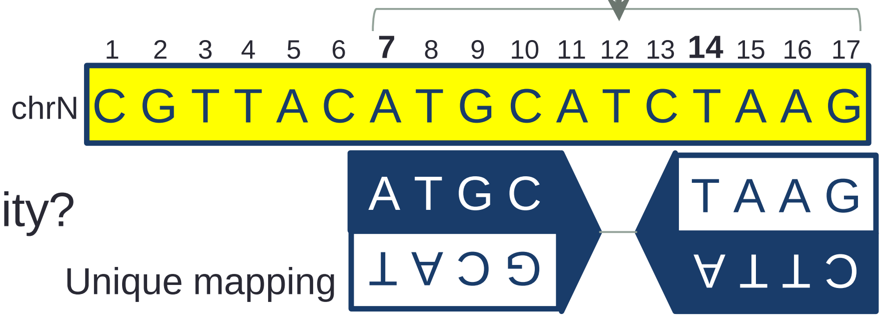

C G T T A C A T G C A T C T A A G
G C A A T G T A C G T A G A T T C

Fragment ATGCATCTAAG
TACGTAGATTC

Read5a ***ATGC***ATCTAAG
Read5b TACGTAG***ATTC***

Insert size = fragment size

Reference

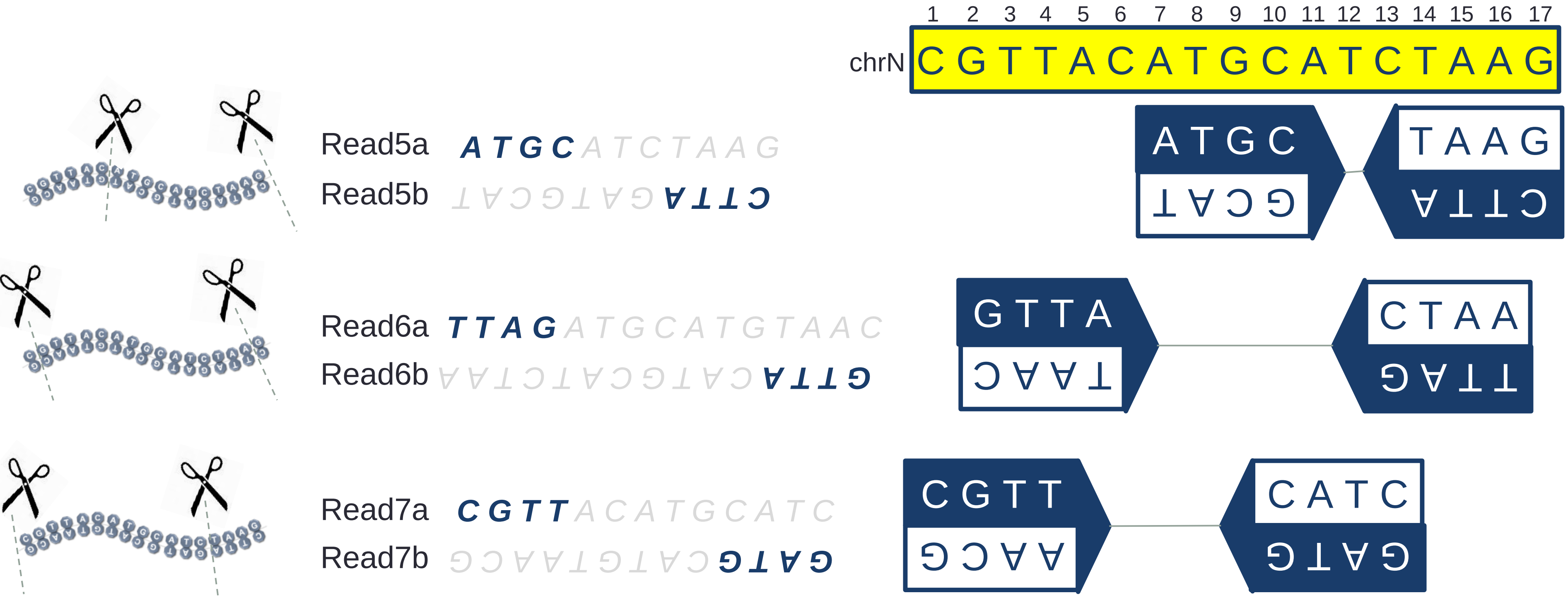


How can we improve mapping quality?

Strategy 2: READ BOTH SIDES

```
Read5a chrN F,Paired,PairMapped,First   7 hiMQ chrN 14  11
Read5b chrN R,Paired,PairMapped,Second 14 hiMQ chrN  7 -11
```

Sequence Alignment & Mapping
look up read in reference sequence



Read5a **ATGC**ATCTAAG

Read5b ꓕ∀Ɔפꓕ∀פ**∀ꓕꓕƆ**

Read6a **TTAG**ATGCATGTAAC

Read6b ∀∀ꓕƆꓕ∀Ɔפꓕ∀Ɔ**∀ꓕꓕפ**

Read7a **CGTT**ACATGCATC

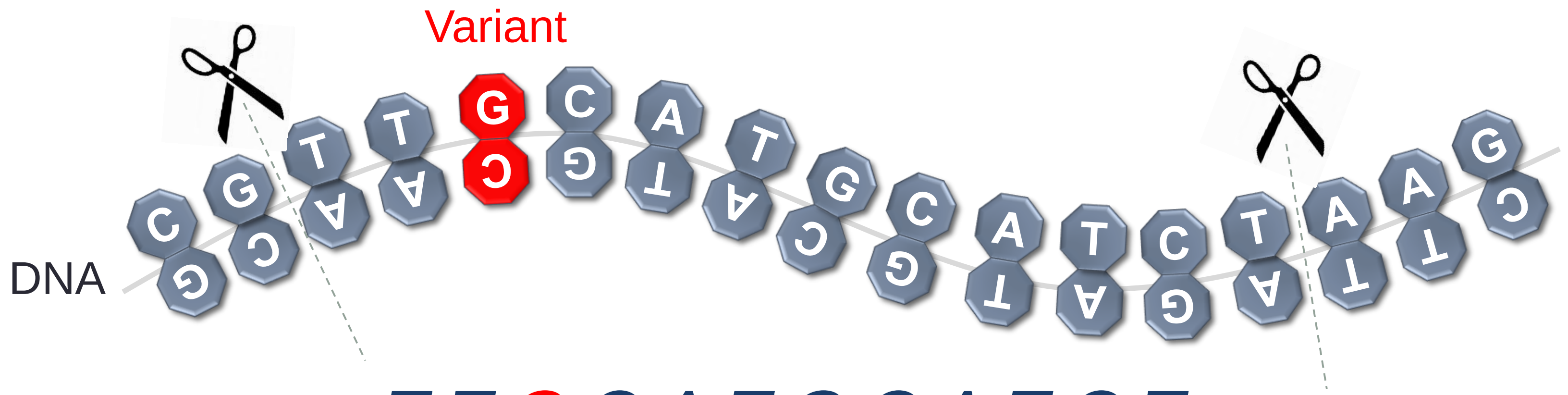
Read7b פƆ∀∀ꓕפꓕ∀Ɔ**פꓕ∀פ**

```
Read5a  chrN  F,Paired,PairMapped,First     7  hiMQ  chrN  14   11
Read5b  chrN  R,Paired,PairMapped,Second  14  hiMQ  chrN   7  -11

Read6a  chrN  R,Paired,PairMapped,Second  13  hiMQ  chrN   2  -15
Read6b  chrN  F,Paired,PairMapped,First     2  hiMQ  chrN  13   15

Read7a  chrN  F,Paired,PairMapped,First     1  hiMQ  chrN  10   13
Read7b  chrN  R,Paired,PairMapped,Second  10  hiMQ  chrN   1  -13
```

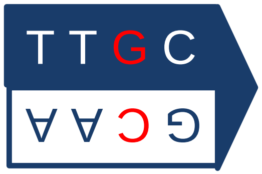
Sequence Alignment & Mapping allowing mismatches



Fragment *T T G C A T G C A T C T*
A A C G T A C G T A G A

Read8 ***T T G C*** *A T G C A T C T*
A A C G T A C G T A G A

	1	2	**3**	4	5	6	7	8	9	10	11	12	13	14	15	16	17
chrN	C	G	T	T	A	C	A	T	G	C	A	T	C	T	A	A	G



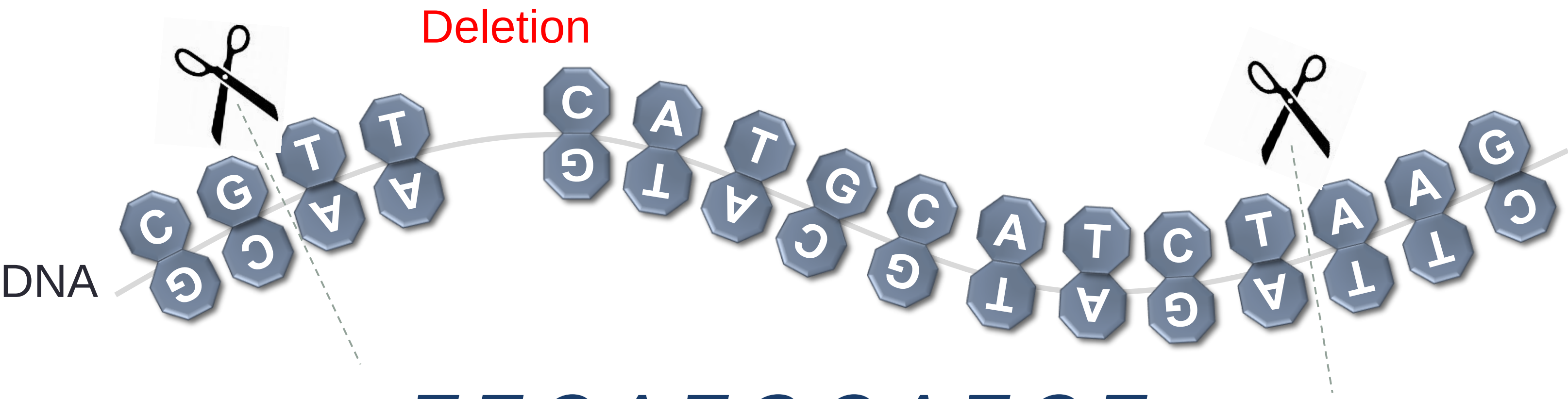
4 Matches or Mismatches

```
Read8 chrN  Forward 3 hiMQ 4M 3A>G
```

CIGAR string

Compact Ideosyncratic Gapped Alignment Report

Sequence Alignment & Mapping allowing mismatches



Fragment

T T C A T G C A T C T

A A G T A C G T A G A

Read9 ***T T C A*** T G C A T C T

A A G T A C G T A G A

1 2 **3** 4 5 6 7 8 9 10 11 12 13 14 15 16 17

chrN C G T T A C A T G C A T C T A A G

T T - C A

A A - G T

2 Matches or Mismatches

1 Deletion

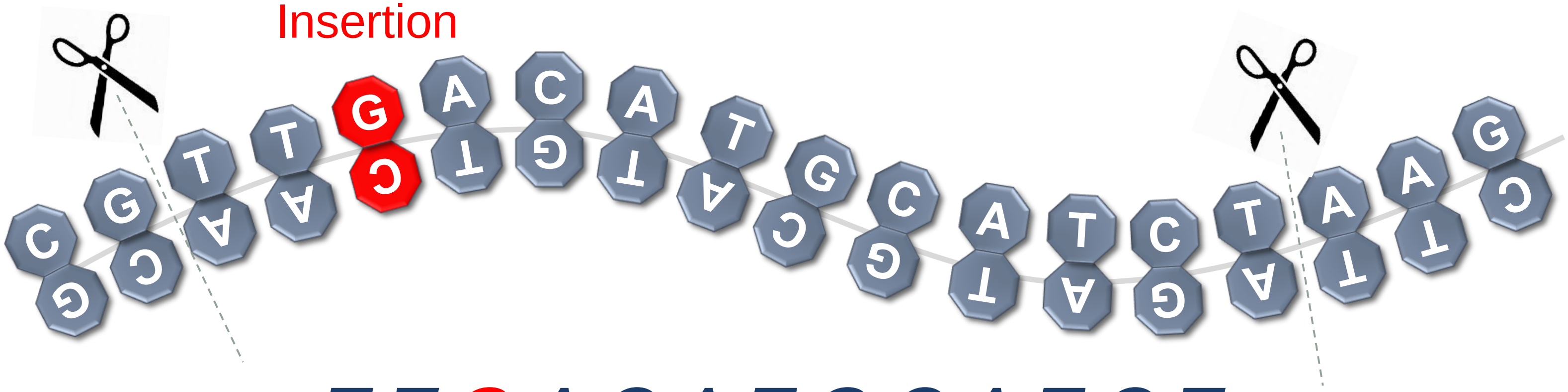
2 Matches or Mismatches

```
Read9 chrN  Forward 3 hiMQ 2M1D2M
```

CIGAR

Sequence Alignment & Mapping allowing mismatches

Insertion



DNA

Fragment TTGACATGCATCT
AACTGTACGTAGA

Read10 TTGACATGCATCT
AACTGTACGTAGA

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

chrN CGTT-ACATGCATCTAAG

TTGA
AACT

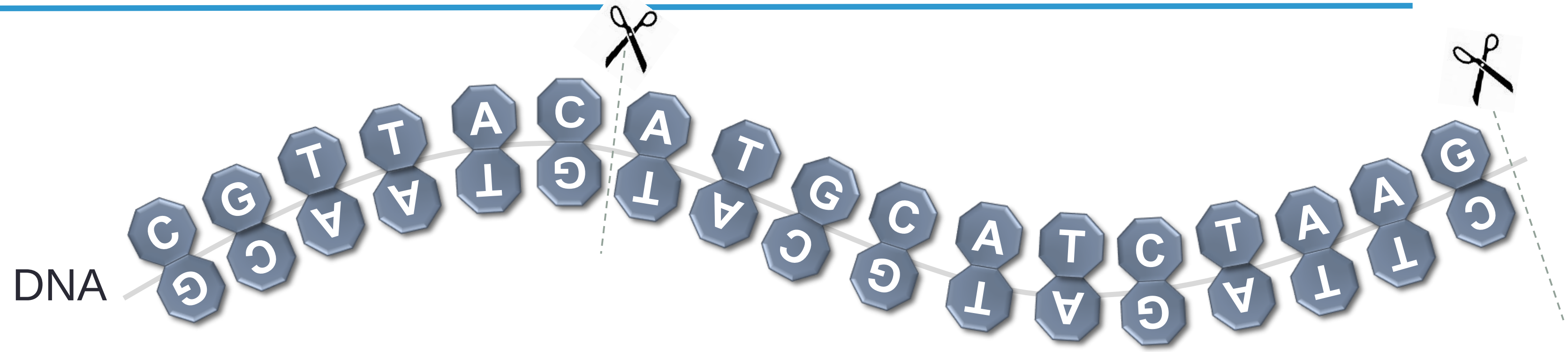
2 Matches or Mismatches
1 Insertion
1 Match or Mismatch

```
Read10 chrN  Forward 3 hiMQ 2M1I1M
```

CIGAR

GENOMICS CORELEUVEN

Sequence Alignment & Mapping allowing mismatches


DNA
C G T T A C A T G C A T C T A A G
G C A A T G T A C G T A G A T T C

Fragment with sequencing artefact (e.g; adapter)

G G G G A T G C A T C T A A G
C C C C T A C G T A G A T T C

Read11

G G G G A T G C A T C T A A G
C C C C T A C G T A G A T T C

1 2 3 4 5 6 **7** 8 9 10 11 12 13 14 15 16 17

Reference chrN C G T T A C A T G C A T C T A A G

G G G G A T G C
C C C C T A C G

4 Soft clipped
4 Matches or Mismatches

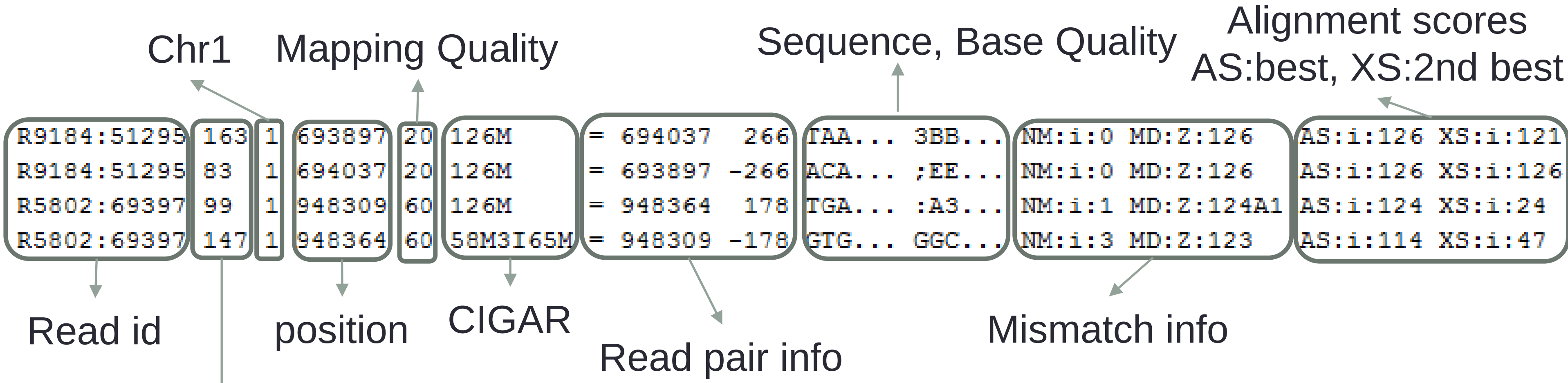
```
Read11  chrN  Forward 7 hiMQ 4S4M
```

CIGAR

Sequence Alignment & Mapping
SAM – BAM - CRAM

- BAM and CRAM files are compressed SAM files

  Not human readable, convert to SAM or use viewer (eg,IGV)
- CRAM smaller than BAM (40%-70%) but takes longer to read
- SAM format :



```
R9184:51295 163 1 693897 20 126M      = 694037  266 TAA... 3BB... NM:i:0 MD:Z:126    AS:i:126 XS:i:121
R9184:51295 83  1 694037 20 126M      = 693897 -266 ACA... ;EE... NM:i:0 MD:Z:126    AS:i:126 XS:i:126
R5802:69397 99  1 948309 60 126M      = 948364  178 TGA... :A3... NM:i:1 MD:Z:124A1  AS:i:124 XS:i:24
R5802:69397 147 1 948364 60 58M3I65M  = 948309 -178 GTG... GGC... NM:i:3 MD:Z:123    AS:i:114 XS:i:47
```

Chr1

Mapping Quality

Sequence, Base Quality

Alignment scores
AS:best, XS:2nd best

Read id

position

CIGAR

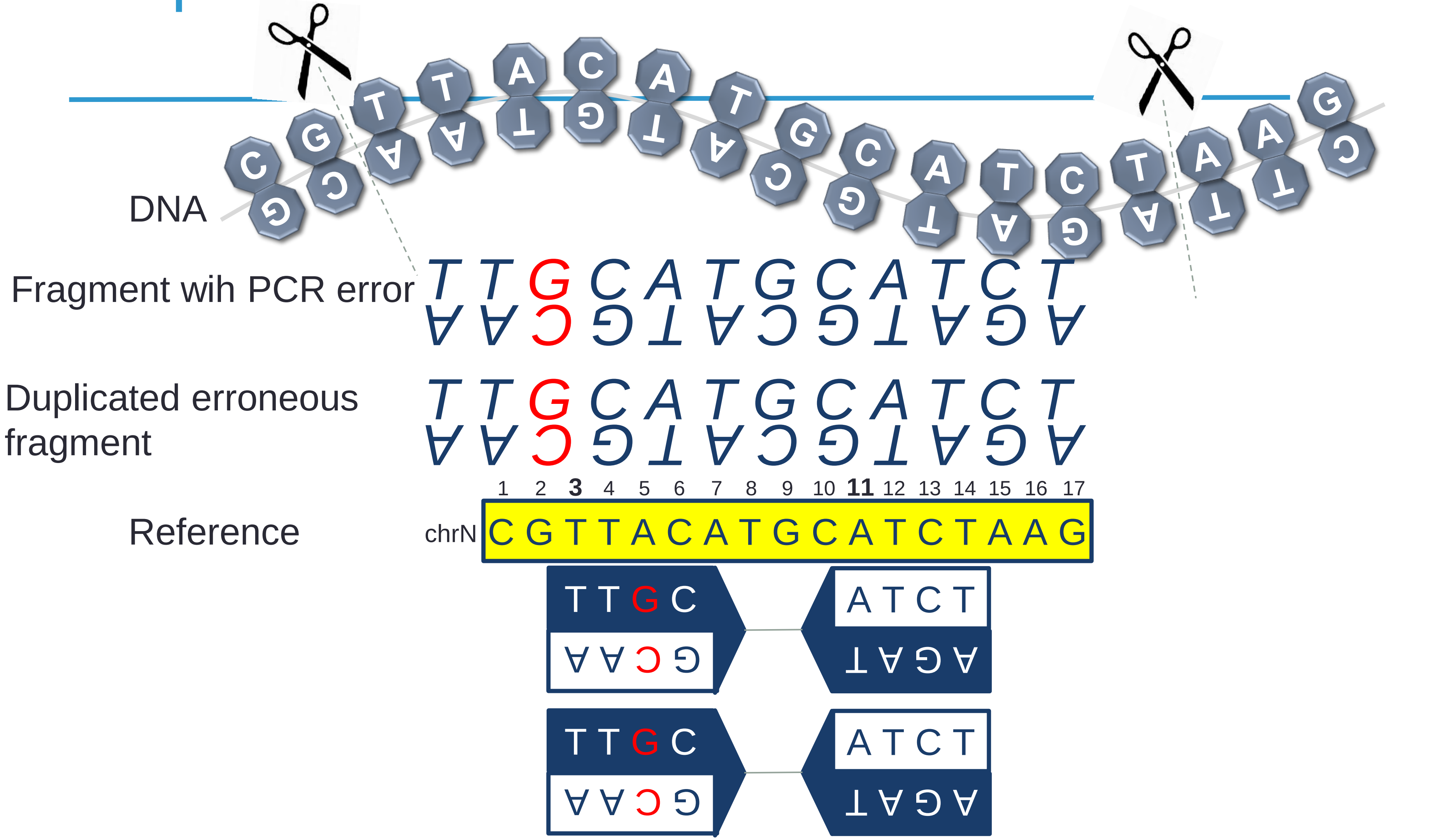
Read pair info

Mismatch info

163: Forward,Paired,PairMapped,Second
83: Reverse,Paired,PairMapped,First
99: Forward,Paired,PairMapped,First
147: Reverse,Paired,PairMapped,Second
https://broadinstitute.github.io/picard/explain-flags.html

bio-bwa.sourceforge.net
samtools.sourceforge.net

Duplicate removal



Read8a chrN F,Paired,PairMapped,First 3 hiMQ chrN 11 12
Read8b chrN R,Paired,PairMapped,Second 11 hiMQ chrN 3 -12

~~Read9a chrN F,Paired,PairMapped,First 3 hiMQ chrN 11 12~~
~~Read9b chrN R,Paired,PairMapped,Second 11 hiMQ chrN 3 -12~~

Base quality score recalibration (BQSR)

- Corrects systematic errors made by the sequencer when it estimates the quality score of each base call
- Performed on BAM, not on FASTQ
  - Requires genomic location of base
  - Ignores genomic locations where variants known to occur frequently
  - Considers each remaining variant an error
- Uses machine learning to characterize regions where more/less errors found than predicted by sequencer
- Example: *any base call that comes after AA in a read should have its quality score reduced by 1%*

Source: https://gatkforums.broadinstitute.org/gatk/discussion/44/base-quality-score-recalibration-bqsr

Viewing a BAM file

- Without graphical user interface
  - Samtools – suite of tools for handling SAM, BAM, CRAM
- With graphical user interface
  - IGV – Integrative Genome Viewer

https://github.com/samtools/samtools
http://software.broadinstitute.org/software/igv/

Samtools

- View
  - Read mapping information

```
samtools view sample.bam 1:11131116-11133317 | less

D00210:1282:CD2J0ANXX:5:1303:1159:51350 163     1       11131905
60      126M    =       11132084        305
GACTGCCTTCTCCAACCACCAACGAGACAGCTACAGCACCTCCAGCACTCCCCACCAATCTCTCTGCACAGCACCTGC
TGCCATCTGCCAGGATAGATACTGATTGCCCACCATCCCTCAGCAGAA
@=>BBECDCFEADBCFDDFDD@F@CGCFDEFEAFBHFBFDFFDDGFDGECDADDFADCBFEFEFEHFDGBGCAGEEIG
FHFDDBFEIGEDGFBAAGCBAFEIBCBIFDDDFDDBGEDFFAGDAC?B  MC:Z:126M
BD:Z:MMNNNNNNMLMMMNNMMNNLNNLLLMMNMKNONMNLONNMNMMNNOONMNMNJJNMNNMMNMMMMMNNNMLOON
MNMNNNNNNNNNNMNNNNOOMNLNNNLNNNNNMMNNJNMNNNNNNJMMNOONONM     MD:Z:126
BI:Z:QQRRRQORQQQQQRRQQRQQRRQQQQRRQRRRQRQRRQQQRQQRRRQQRQQPPRQQRRRRQQQQQRQQQQRRQ
QQRRQRRQORRRQRQORRQRRQQRRQRRRQRRQQOPRQQRRRQPRQRRRQRRQ     NM:i:0  MQ:i:60
AS:i:126        XS:i:19 RG:Z:GC065340.run.181130.HiSeq2500.FCA.lane5-389E55F4-
49A02F07  PG:Z:MarkDuplicates-746E271E
```

Samtools

- Pileup
    - Bases observed at each position

```
samtools mpileup sample.bam -f genome.fa -r 1:11131116-11133317

1     11132152    T     56
.$..............,...,,..,.,..........,..,.,....,,...,..^].^].
CEAgFEFgE;EggKgChEFCCEEDECEEBEE=EEEECEECECEEEECCEE6CAAA@

1     11132153    C     57
..............,...,,..,.,..........,..,.,....,,...,....^].^].
GFoGGHmGDFoo^mHnFHHHGFHGHEGFDF/FFGGAGGHGHGFHGHHGGDHCCAAAA

1     11132154    A     57
..............,...,,..,.,..........,..,.,....,,...,......
DDgDDDhD2DgiUhEiDEEEECEDEDDEDE<CADE?2DFDEDDDDEEDDDED>??AA
```

Samtools

- Pileup
  - Bases observed at each position

```
samtools mpileup sample.bam -f genome.fa -r 1:11131116-11133317

1     11132152    T     56
.$..............,...,,..,.,..........,..,.,....,,...,..^].^].
CEAgFEFgE;EggKgChEFCCEEDECEEBEE=EEEECEECECEEEECCEE6CAAA@
```

Chromosome

Samtools

- Pileup
  - Bases observed at each position

```
samtools mpileup sample.bam -f genome.fa -r 1:11131116-11133317

1     11132152    T     56
.$..............,...,,..,.,..........,..,.,....,,...,..^].^].
CEAgFEFgE;EggKgChEFCCEEDECEEBEE=EEEECEECECEEEEECCEE6CAAA@
```

Position

Chromosome

Samtools

- Pileup
    - Bases observed at each position

```
samtools mpileup sample.bam -f genome.fa -r 1:11131116-11133317

1     11132152    T    56
.$..............,...,,..,.,..........,..,.,....,,...,..^].^].
CEAgFEFgE;EggKgChEFCCEEDECEEBEE=EEEECEECECEEEECCEE6CAAA@
```

Reference

Position

Chromosome

Samtools

- Pileup
  - Bases observed at each position

```
samtools mpileup sample.bam -f genome.fa -r 1:11131116-11133317

1     11132152    T     56
.$..............,...,,..,.,..........,..,.,....,,...,..^].^].
CEAgFEFgE;EggKgChEFCCEEDECEEBEE=EEEECEECECEEEECCEE6CAAA@
```

Depth

Reference

Position

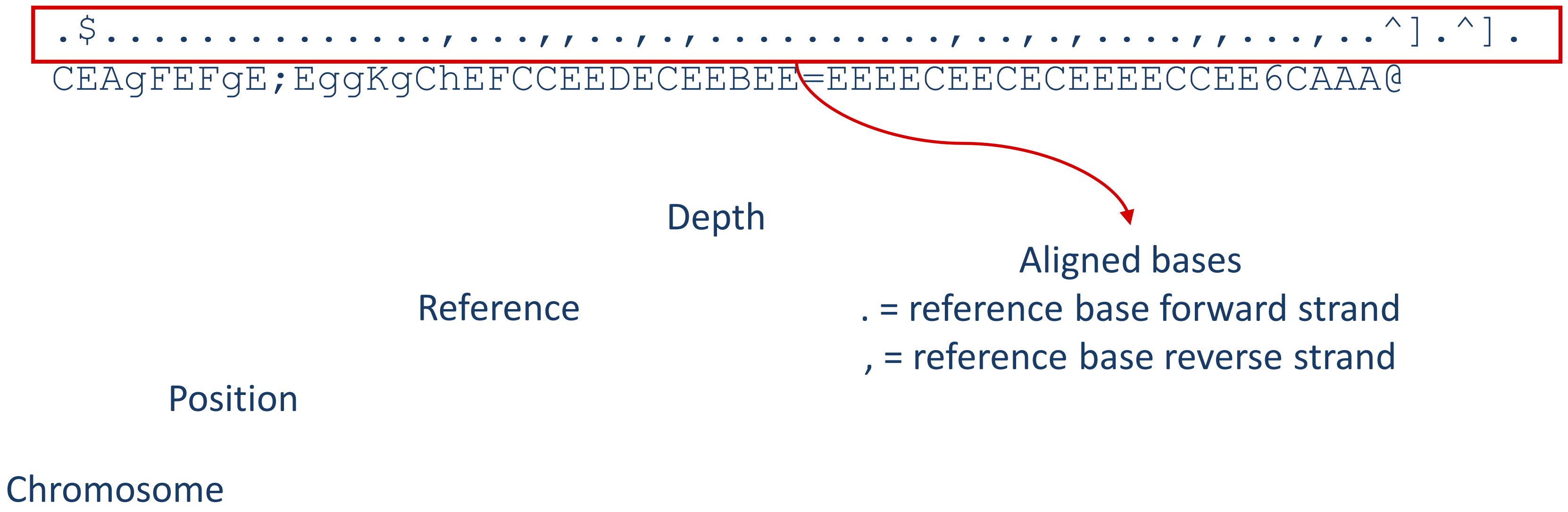
Chromosome

Samtools

- Pileup
  - Bases observed at each position

```
samtools mpileup sample.bam -f genome.fa -r 1:11131116-11133317

1      11132152    T      56
.$..............,...,,..,.,..........,..,.,....,,...,..^].^].
CEAgFEFgE;EggKgChEFCCEEDECEEBEE=EEEECEECECEEEECCEE6CAAA@
```



Depth

Aligned bases
. = reference base forward strand
, = reference base reverse strand

Reference

Position

Chromosome

Samtools

- Pileup
  - Bases observed at each position

```
samtools mpileup sample.bam -f genome.fa -r 1:11131116-11133317

1      11132152    T      56
.$..............,...,,..,.,..........,..,.,....,,...,..^].^].
CEAgFEFgE;EggKgChEFCCEEDECEEBEE=EEEECEECECEEEECCEE6CAAA@
```

Depth

Aligned bases

. = reference base forward strand

, = reference base reverse strand

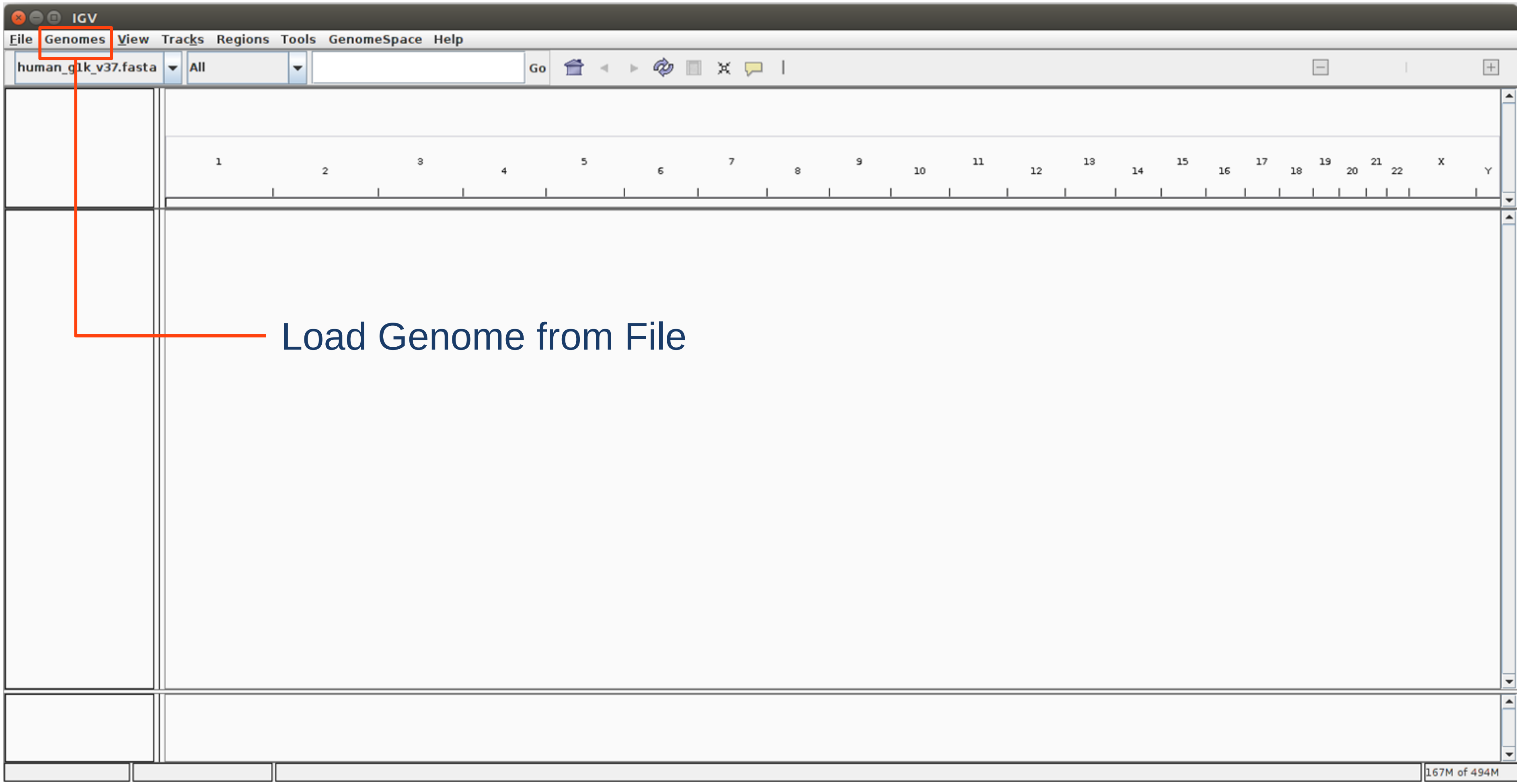
Reference

Position

Chromosome

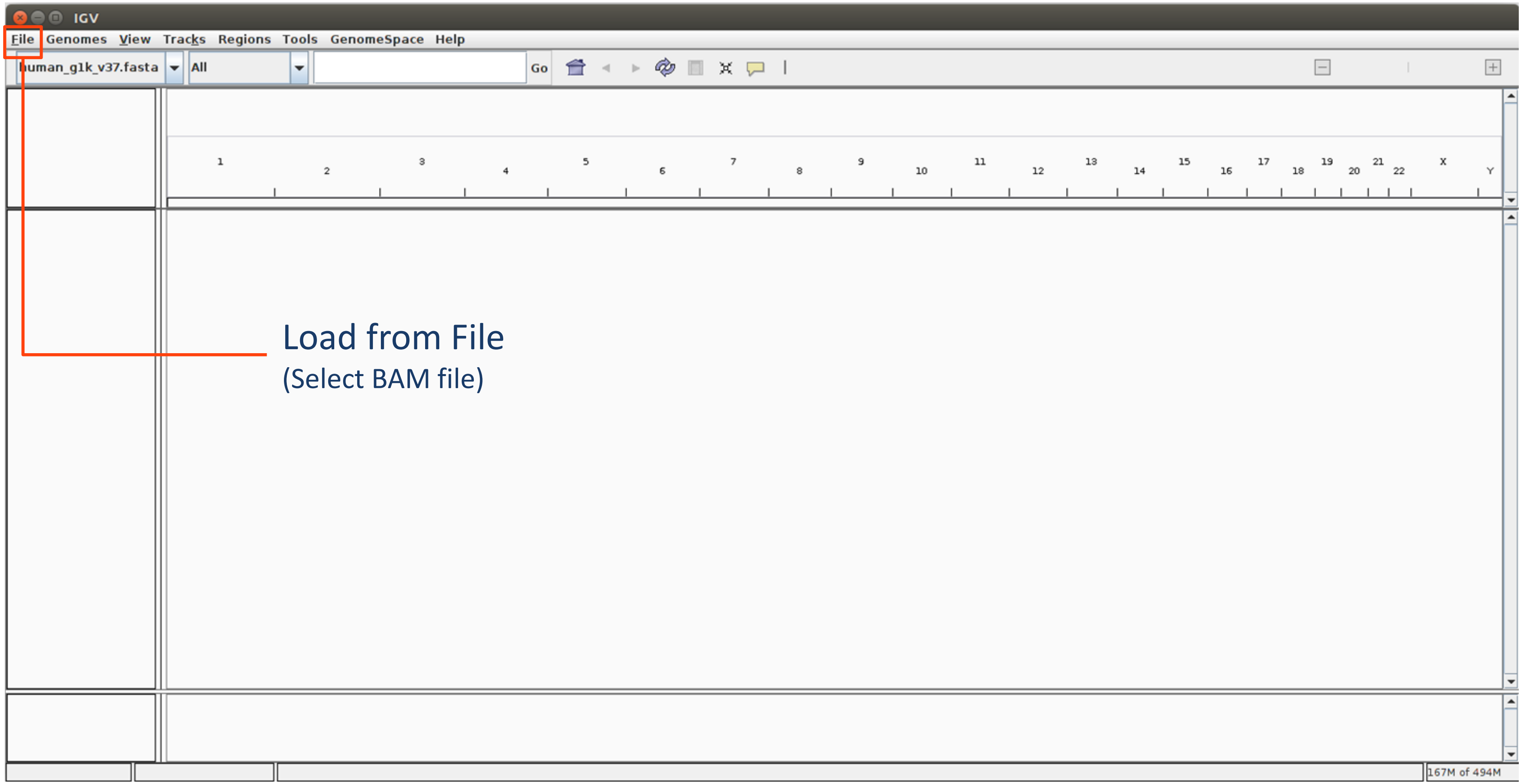
Base qualities

IGV


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
human_g1k_v37.fasta
All
Go
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
167M of 494M

Load Genome from File

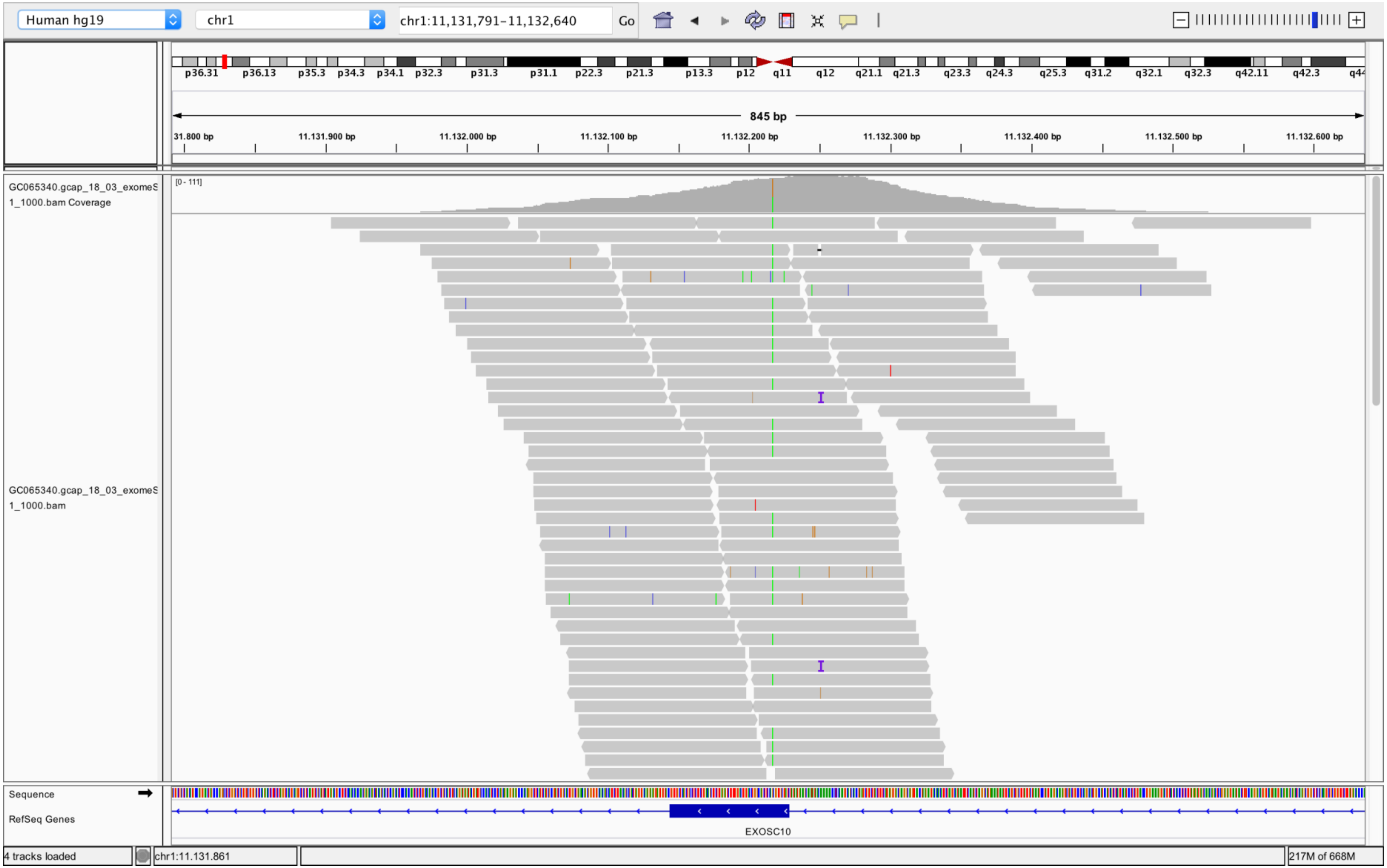
IGV



Load from File

(Select BAM file)

Example 1



Human hg19

chr1

chr1:11,131,791-11,132,640

Go

GC065340.gcap_18_03_exomeS1_1000.bam Coverage

GC065340.gcap_18_03_exomeS1_1000.bam

Sequence

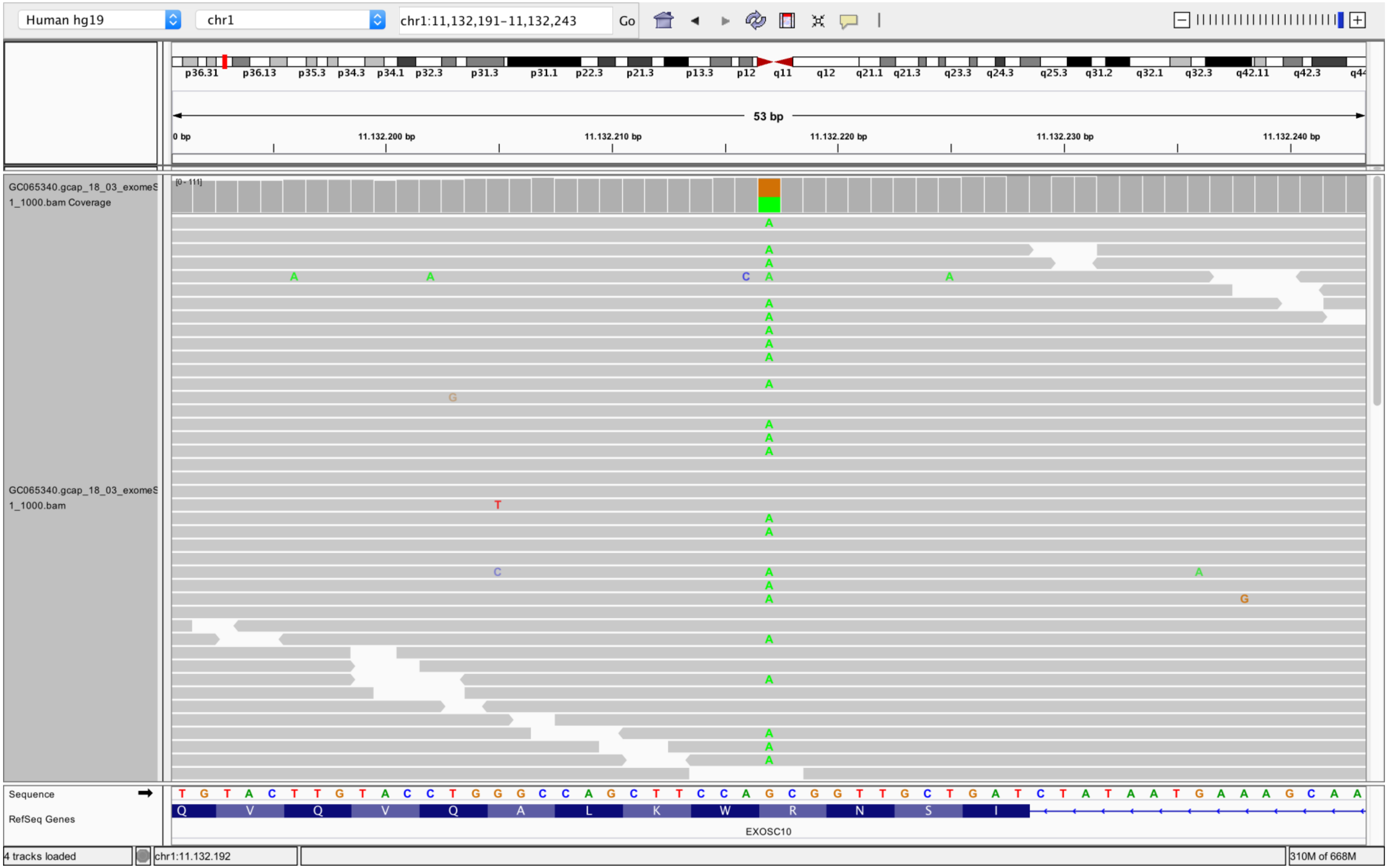
RefSeq Genes

4 tracks loaded

chr1:11.131.861

217M of 668M

Example 1 – targeted sequencing



Human hg19

chr1

chr1:11,132,191-11,132,243

Go

GC065340.gcap_18_03_exomeS1_1000.bam Coverage

[0 - 111]

GC065340.gcap_18_03_exomeS1_1000.bam

Sequence

T G T A C T T G T A C C T G G G C C A G C T T C C A G C G G T T G C T G A T C T A T A A T G A A A G C A A

RefSeq Genes

Q V Q V Q A L K W R N S I

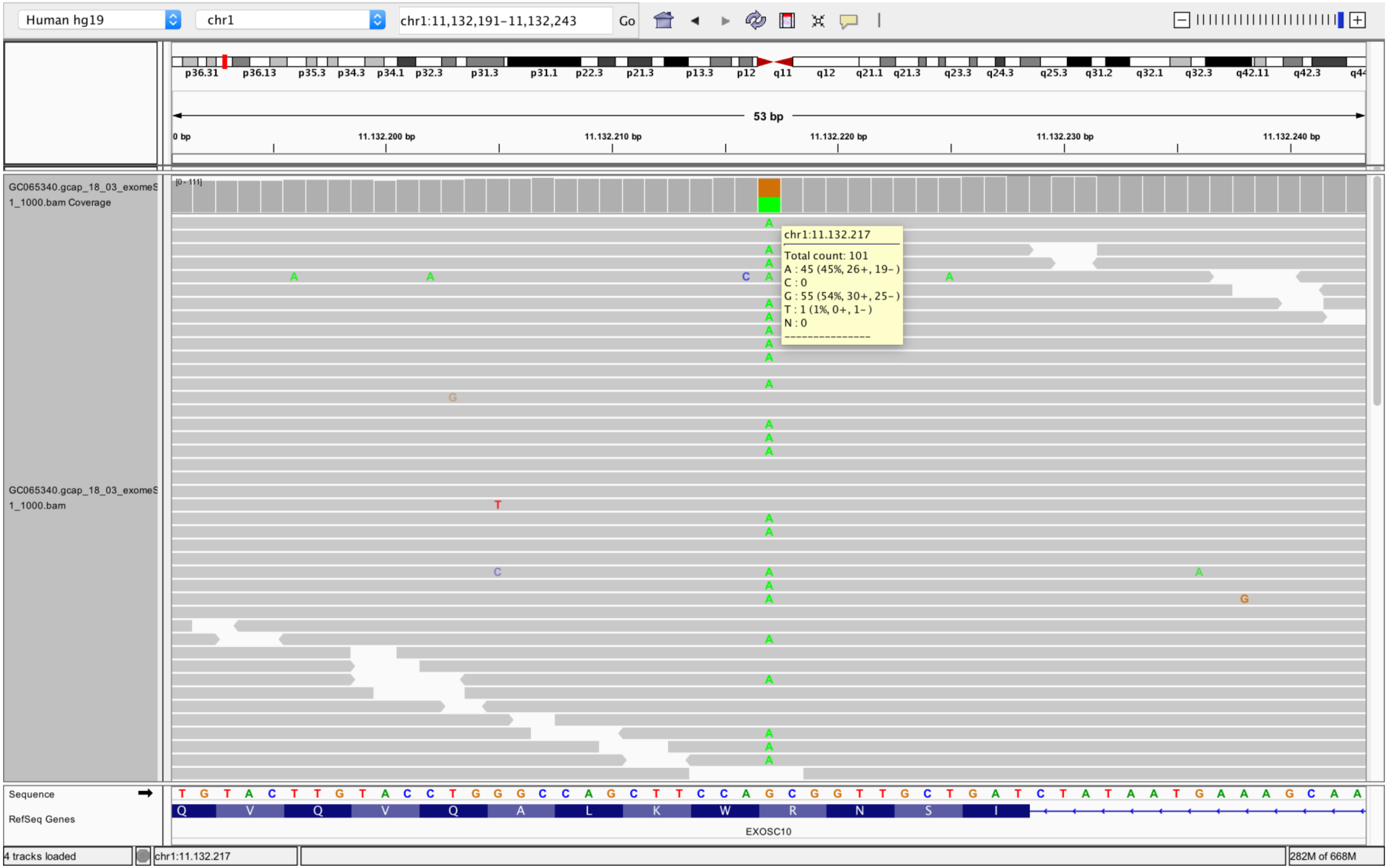
EXOSC10

4 tracks loaded

chr1:11.132.192

310M of 668M

Example 1 – targeted sequencing



Human hg19

chr1

chr1:11,132,191-11,132,243

GC065340.gcap_18_03_exomeS1_1000.bam Coverage

GC065340.gcap_18_03_exomeS1_1000.bam

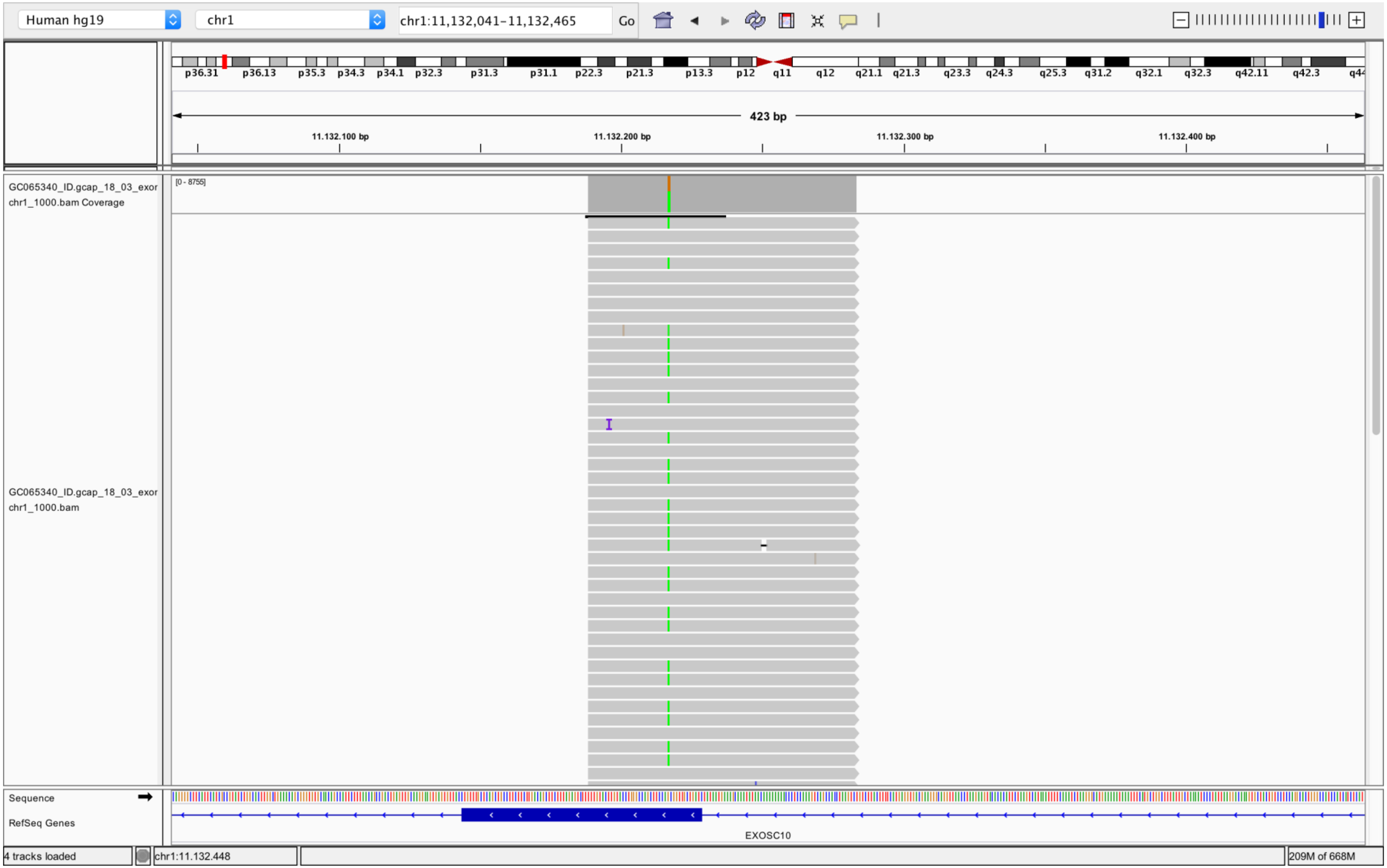
chr1:11.132.217
Total count: 101
A : 45 (45%, 26+, 19-)
C : 0
G : 55 (54%, 30+, 25-)
T : 1 (1%, 0+, 1-)
N : 0
----

4 tracks loaded

chr1:11.132.217

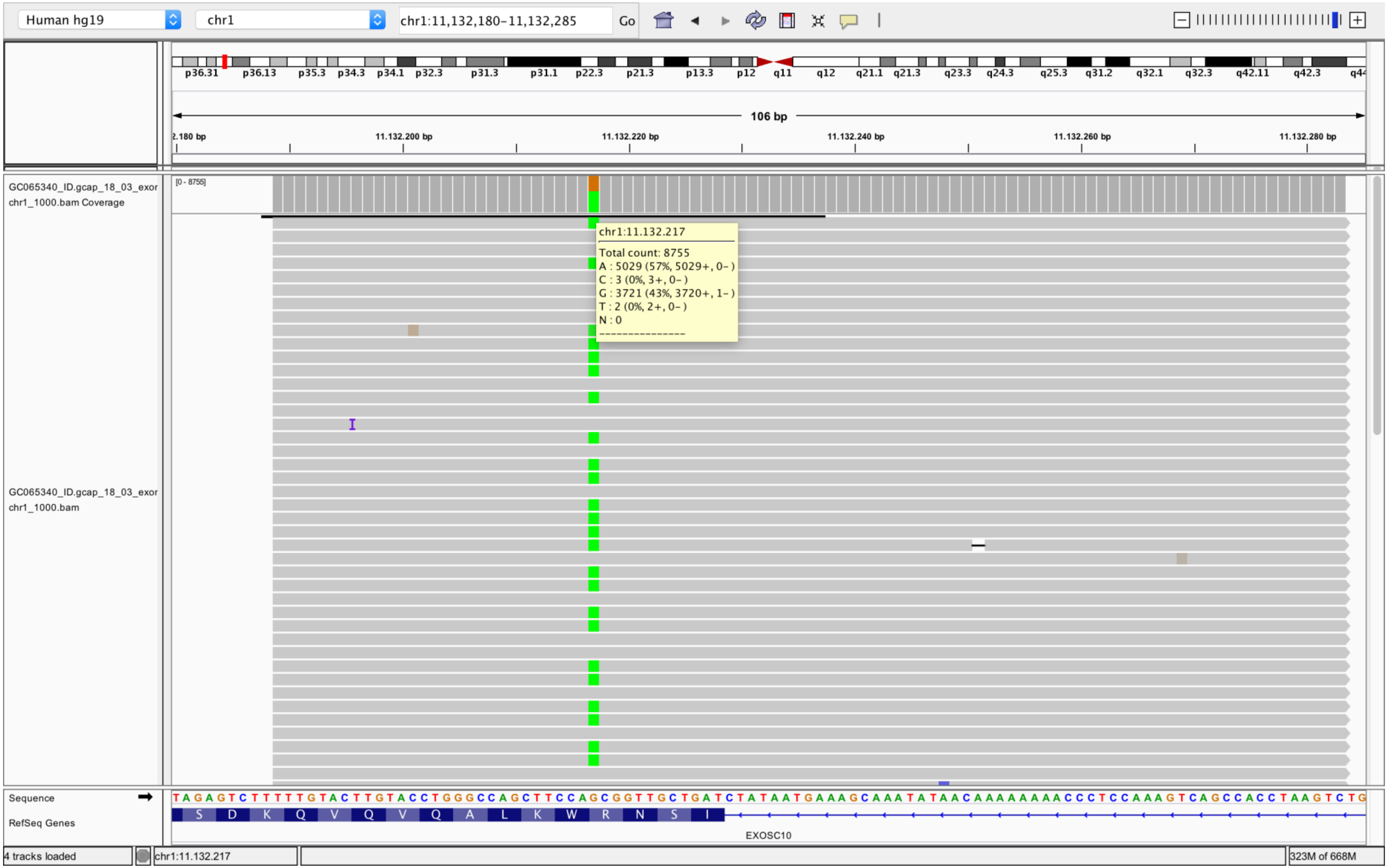
282M of 668M

Example 2



Human hg19
chr1
chr1:11,132,041-11,132,465
Go
p36.31 p36.13 p35.3 p34.3 p34.1 p32.3 p31.3 p31.1 p22.3 p21.3 p13.3 p12 q11 q12 q21.1 q21.3 q23.3 q24.3 q25.3 q31.2 q32.1 q32.3 q42.11 q42.3 q44
423 bp
11.132.100 bp
11.132.200 bp
11.132.300 bp
11.132.400 bp
GC065340_ID.gcap_18_03_exor chr1_1000.bam Coverage
[0 - 8755]
GC065340_ID.gcap_18_03_exor chr1_1000.bam
Sequence
RefSeq Genes
EXOSC10
4 tracks loaded
chr1:11.132.448
209M of 668M

Example 2 – amplicon sequencing



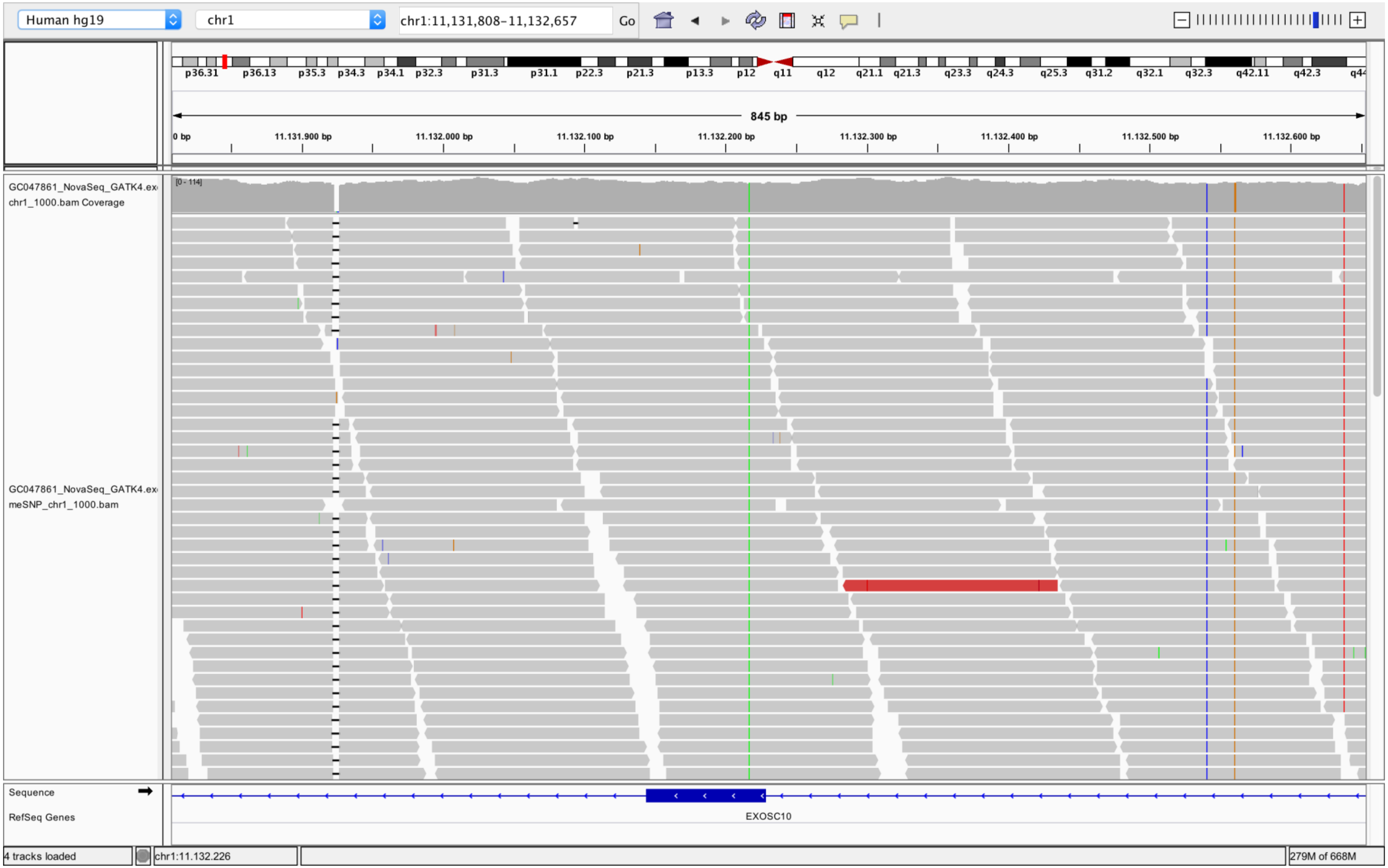
chr1:11.132.217

Total count: 8755
A : 5029 (57%, 5029+, 0-)
C : 3 (0%, 3+, 0-)
G : 3721 (43%, 3720+, 1-)
T : 2 (0%, 2+, 0-)
N : 0

TAGAGTCTTTTTGTACTTGTACCTGGGCCAGCTTCCAGCGGTTGCTGATCTATAATGAAAGCAAATATAACAAAAAAAACCCTCCAAAGTCAGCCACCTAAGTCTG

chr1:11.132.217

Example 3



Human hg19 | chr1 | chr1:11,131,808-11,132,657 | Go

GC047861_NovaSeq_GATK4.ex chr1_1000.bam Coverage

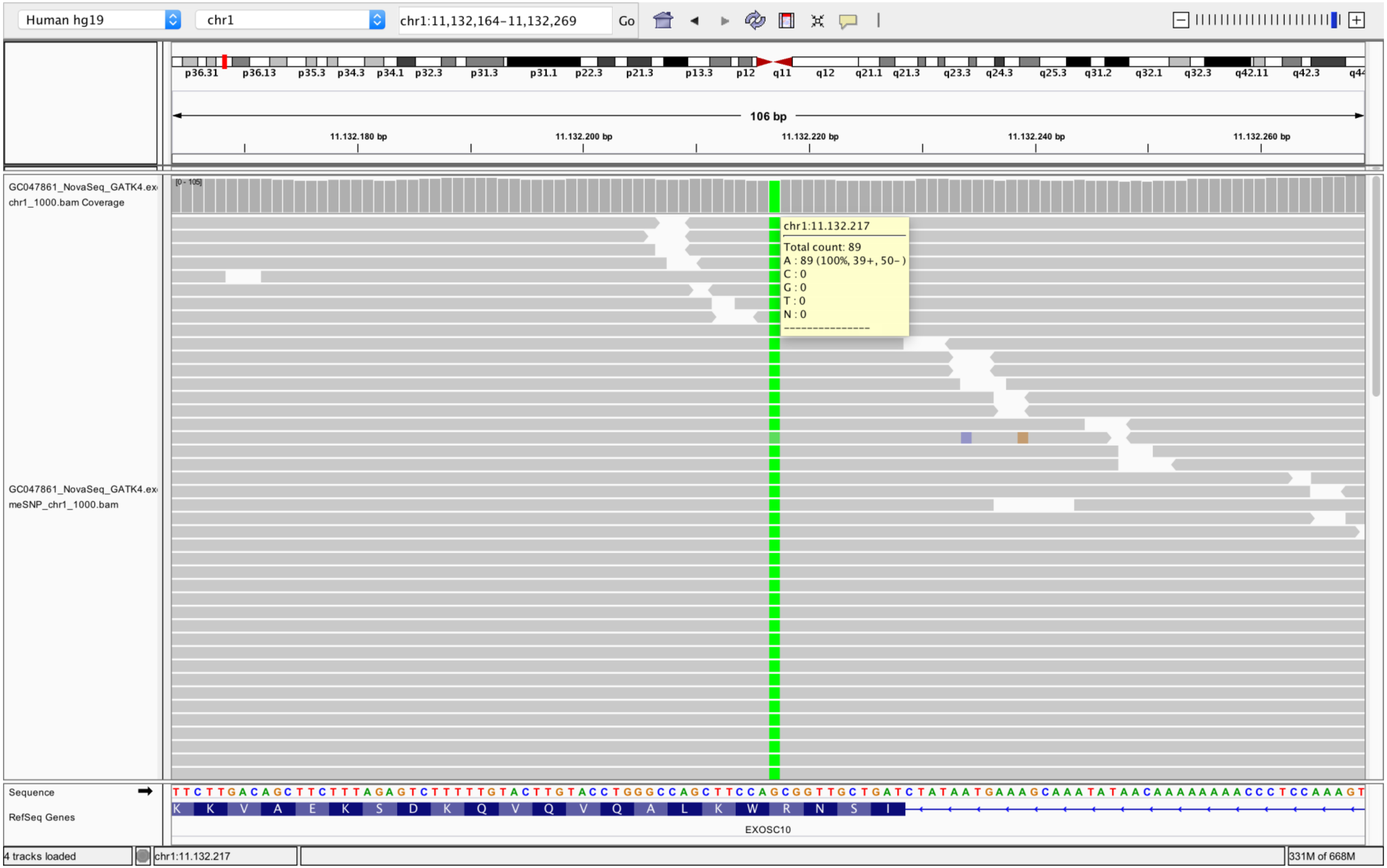
GC047861_NovaSeq_GATK4.ex meSNP_chr1_1000.bam

Sequence

RefSeq Genes

4 tracks loaded | chr1:11.132.226 | 279M of 668M

Example 3 – WGS



Human hg19

chr1

chr1:11,132,164-11,132,269 Go

chr1:11.132.217

Total count: 89
A : 89 (100%, 39+, 50-)
C : 0
G : 0
T : 0
N : 0
----

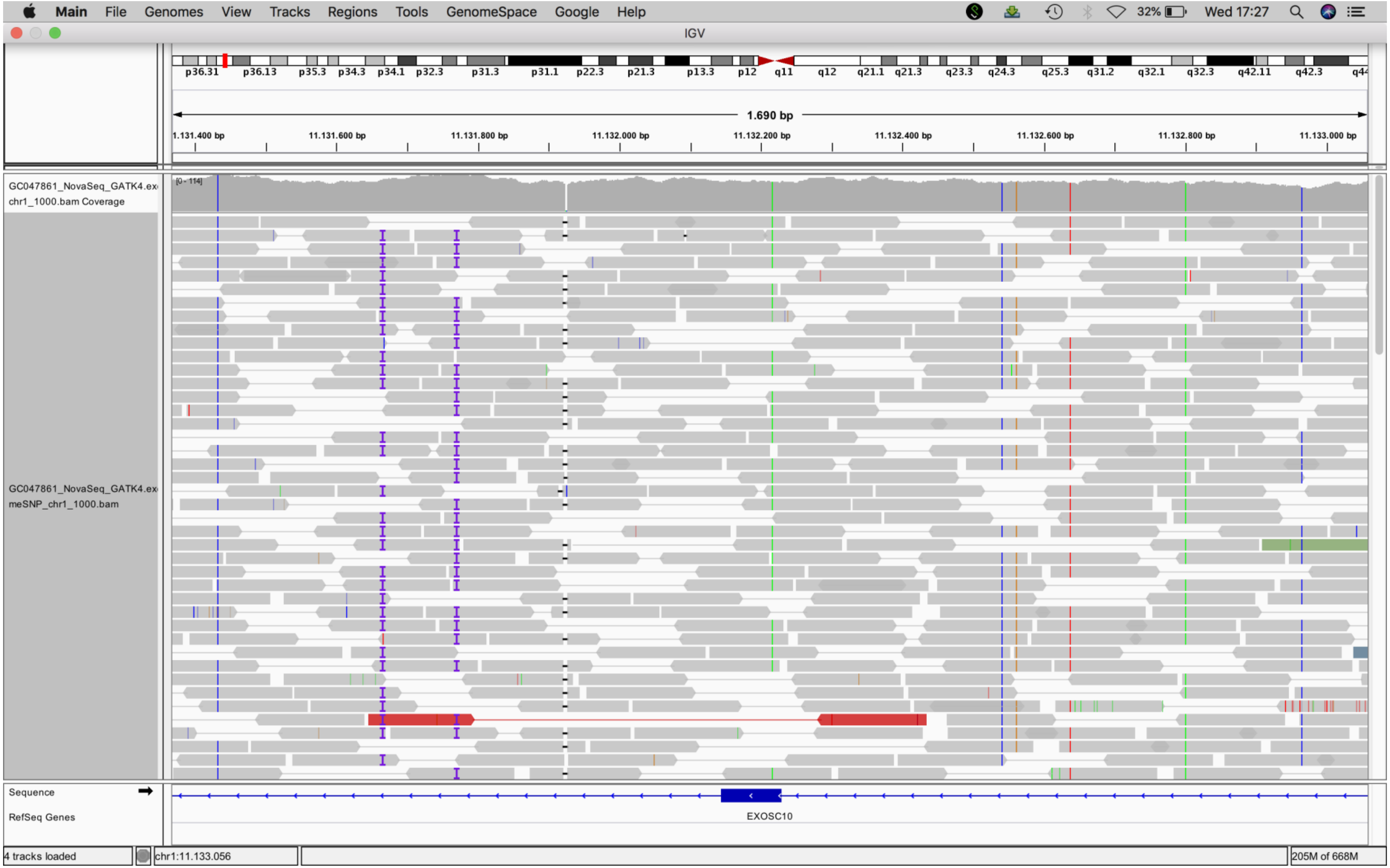
4 tracks loaded

chr1:11.132.217

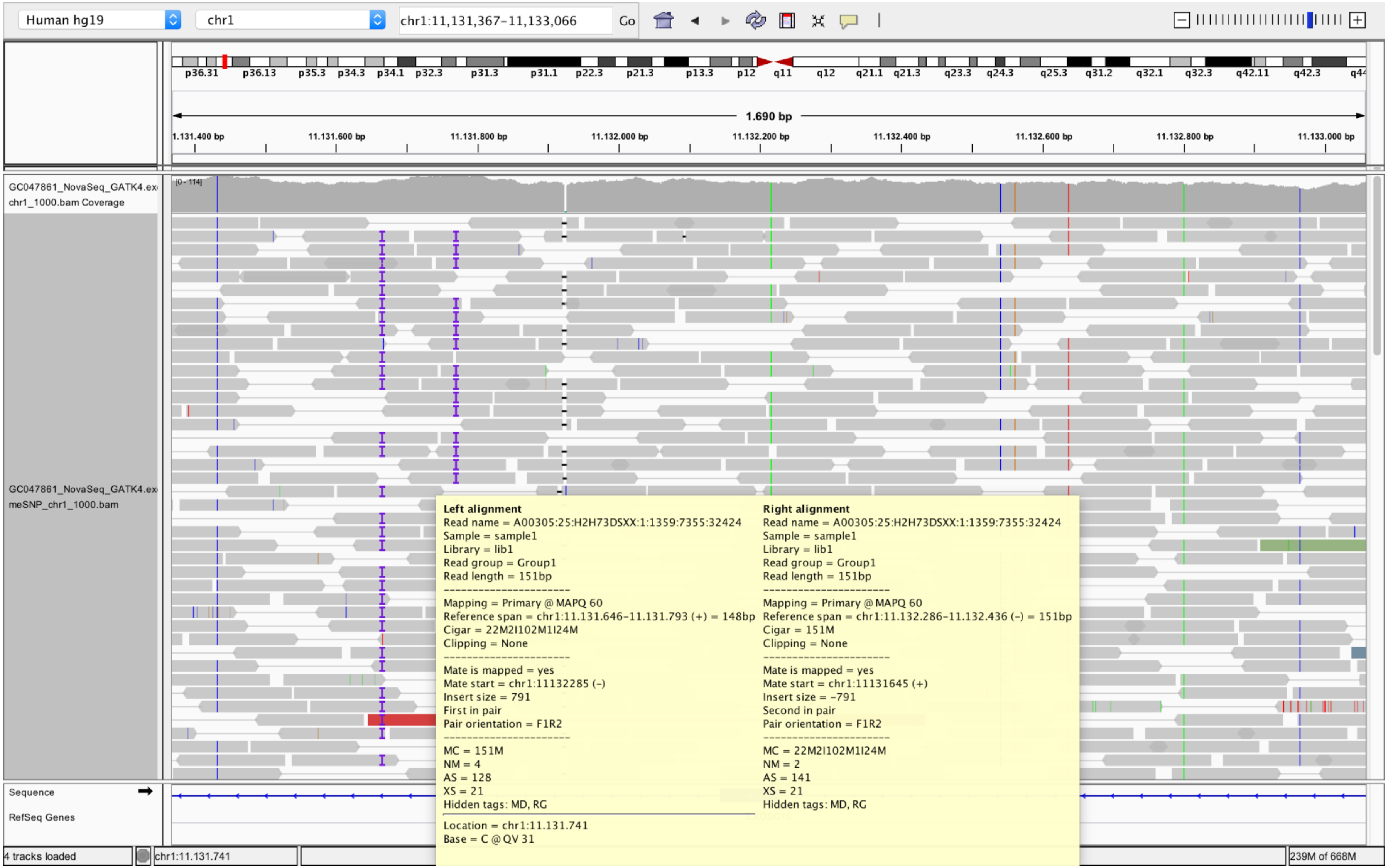
331M of 668M

p36.31
p36.13
p35.3
p34.3
p34.1
p32.3
p31.3
p31.1
p22.3
p21.3
p13.3
p12
q11
q12
q21.1
q21.3
q23.3
q24.3
q25.3
q31.2
q32.1
q32.3
q42.11
q42.3
q44
106 bp
11.132.180 bp
11.132.200 bp
11.132.220 bp
11.132.240 bp
11.132.260 bp
GC047861_NovaSeq_GATK4.ex chr1_1000.bam Coverage
[0 - 105]
GC047861_NovaSeq_GATK4.ex meSNP_chr1_1000.bam
Sequence
RefSeq Genes
TTCTTGACAGCTTCTTTAGAGTCTTTTTGTACTTGTACCTGGGCCAGCTTCCAGCGGTTGCTGATCTATAATGAAAGCAAATATAACAAAAAAAACCCTCCAAAGT
K K V A E K S D K Q V Q V Q A L K W R N S I
EXOSC10

Example 3 – WGS


Main File Genomes View Tracks Regions Tools GenomeSpace Google Help
32% Wed 17:27
IGV
p36.31 p36.13 p35.3 p34.3 p34.1 p32.3 p31.3 p31.1 p22.3 p21.3 p13.3 p12 q11 q12 q21.1 q21.3 q23.3 q24.3 q25.3 q31.2 q32.1 q32.3 q42.11 q42.3 q44
1.690 bp
1.131.400 bp 11.131.600 bp 11.131.800 bp 11.132.000 bp 11.132.200 bp 11.132.400 bp 11.132.600 bp 11.132.800 bp 11.133.000 bp
GC047861_NovaSeq_GATK4.ex chr1_1000.bam Coverage
[0 - 114]
GC047861_NovaSeq_GATK4.ex meSNP_chr1_1000.bam
Sequence
RefSeq Genes
EXOSC10
4 tracks loaded
chr1:11.133.056
205M of 668M

Example 3 – WGS



```
Left alignment
Read name = A00305:25:H2H73DSXX:1:1359:7355:32424
Sample = sample1
Library = lib1
Read group = Group1
Read length = 151bp
----------------------
Mapping = Primary @ MAPQ 60
Reference span = chr1:11.131.646-11.131.793 (+) = 148bp
Cigar = 22M2I102M1I24M
Clipping = None
----------------------
Mate is mapped = yes
Mate start = chr1:11132285 (-)
Insert size = 791
First in pair
Pair orientation = F1R2
----------------------
MC = 151M
NM = 4
AS = 128
XS = 21
Hidden tags: MD, RG

Location = chr1:11.131.741
Base = C @ QV 31
```

```
Right alignment
Read name = A00305:25:H2H73DSXX:1:1359:7355:32424
Sample = sample1
Library = lib1
Read group = Group1
Read length = 151bp
----------------------
Mapping = Primary @ MAPQ 60
Reference span = chr1:11.132.286-11.132.436 (-) = 151bp
Cigar = 151M
Clipping = None
----------------------
Mate is mapped = yes
Mate start = chr1:11131645 (+)
Insert size = -791
Second in pair
Pair orientation = F1R2
----------------------
MC = 22M2I102M1I24M
NM = 2
AS = 141
XS = 21
Hidden tags: MD, RG
```

BAM file quality control

- Did I select/sequence what I wanted to ?
- What is the mean coverage ?
- How much of my region of interest is
  - Covered at 30X ?
  - Not covered at all ?
- Is the coverage even ?
- …

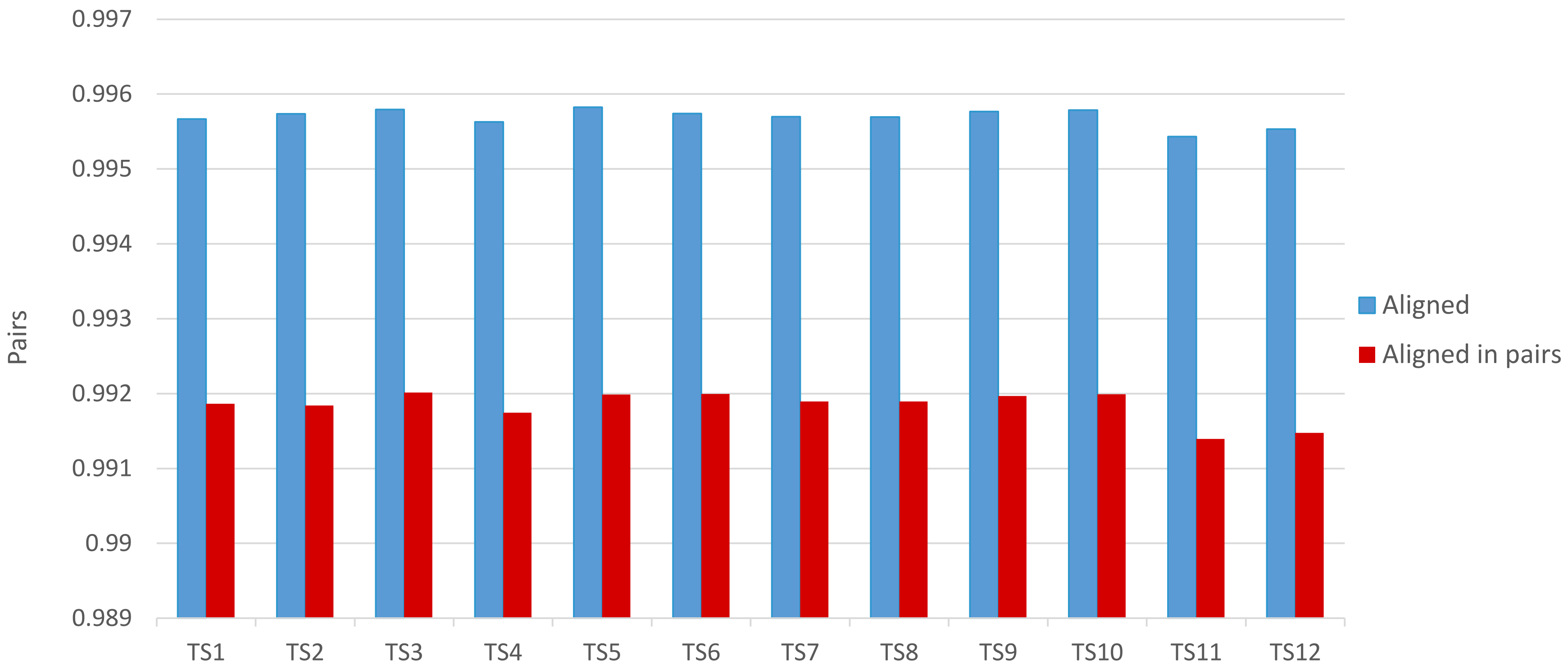
Picard tools

- Set of (command line) tools to
  - Manipulate NGS data
    - SAM/BAM/CRAM
    - VCF/BCF
  - Compute metrics

https://broadinstitute.github.io/picard/

Alignment metrics

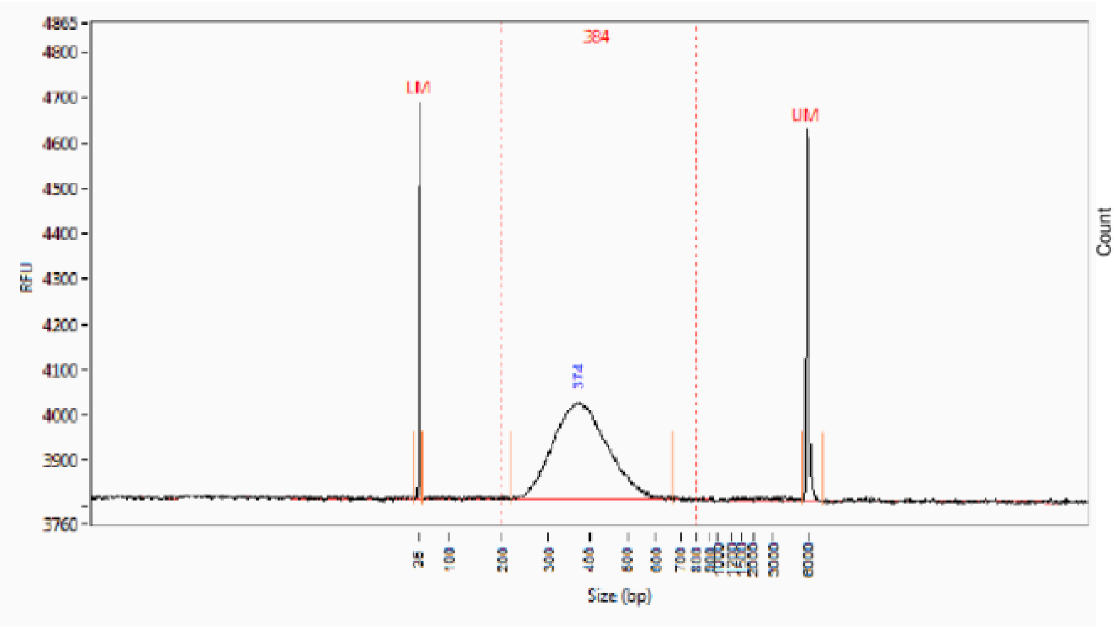
- Picard CollectAlignmentSummaryMetrics

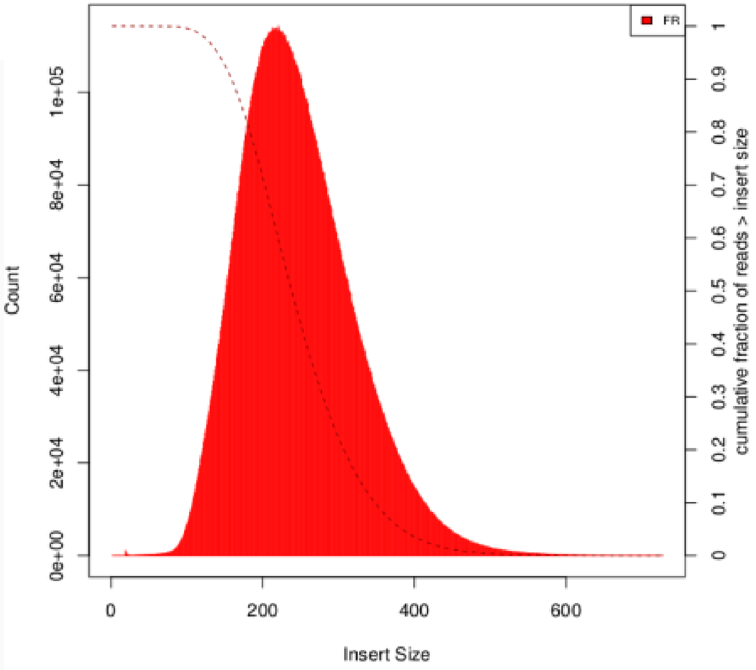


```
java -jar ~/bin/picard.jar CollectAlignmentSummaryMetrics
I=sample.bam O=sample_sum_metrics.txt R=ref.fasta
```

Insert size

- Picard CollectInsertSizeMetrics

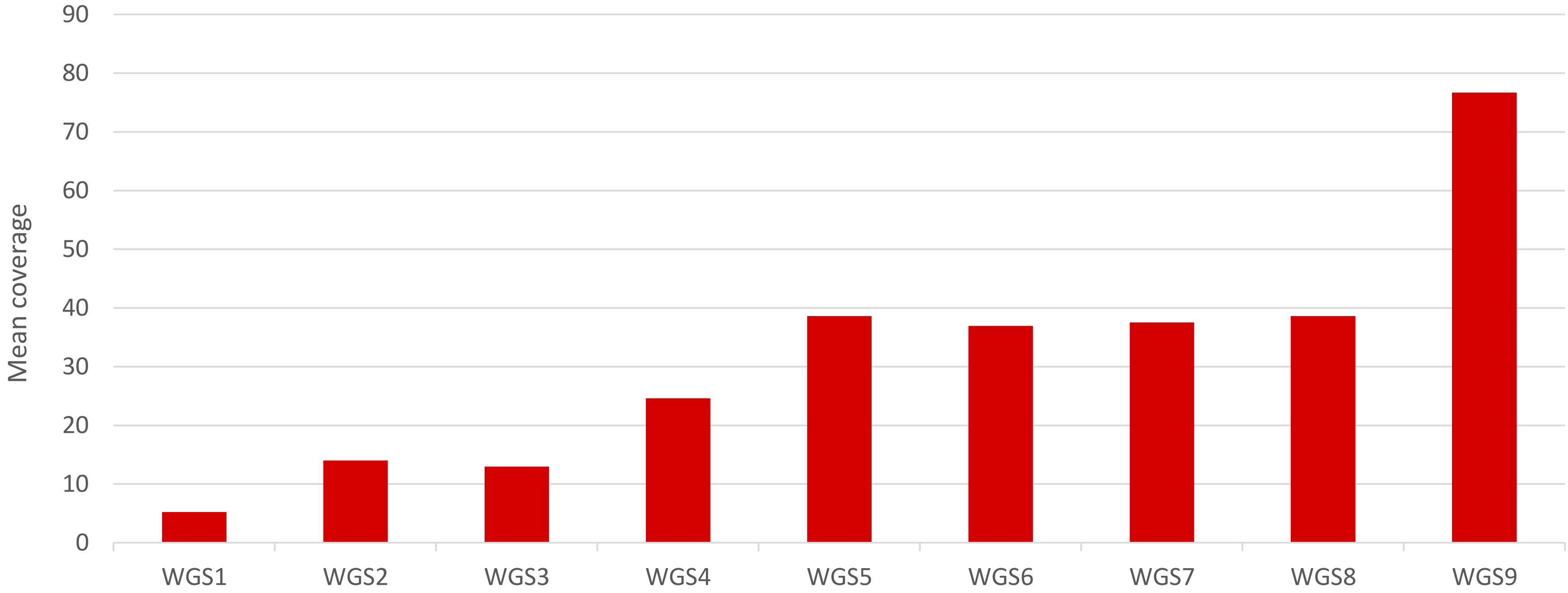




```
java -jar ~/bin/picard.jar CollectInsertSizeMetrics I=sample.bam
O=sample_insert_size_metrics.txt H=sample_insert_size_histogram.pdf
```

WGS metrics

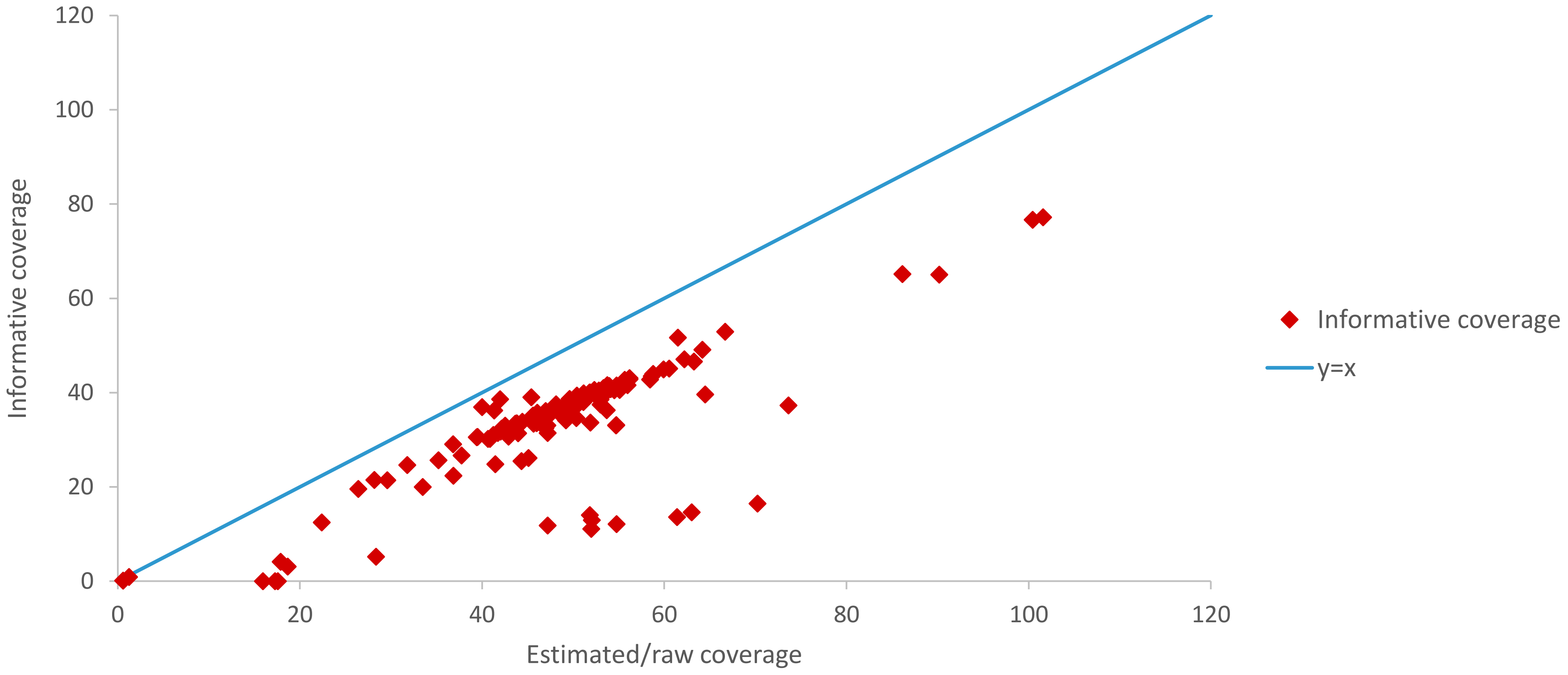
- CollectWgsMetrics
  - Mean coverage



```
java -jar picard.jar CollectWgsMetrics I=sample.bam
O=sample_wgs_metrics.txt R=reference_sequence.fasta
```

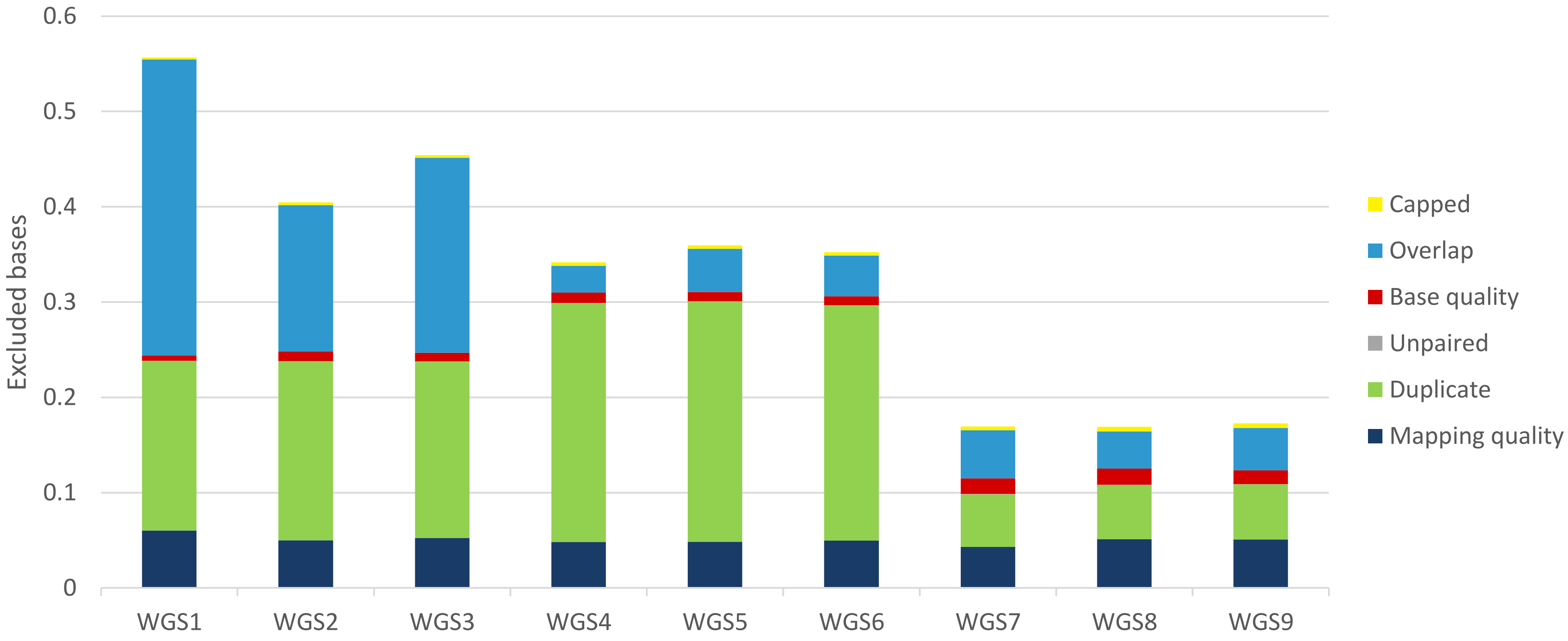
WGS metrics

- CollectWgsMetrics
  - Raw vs informative coverage



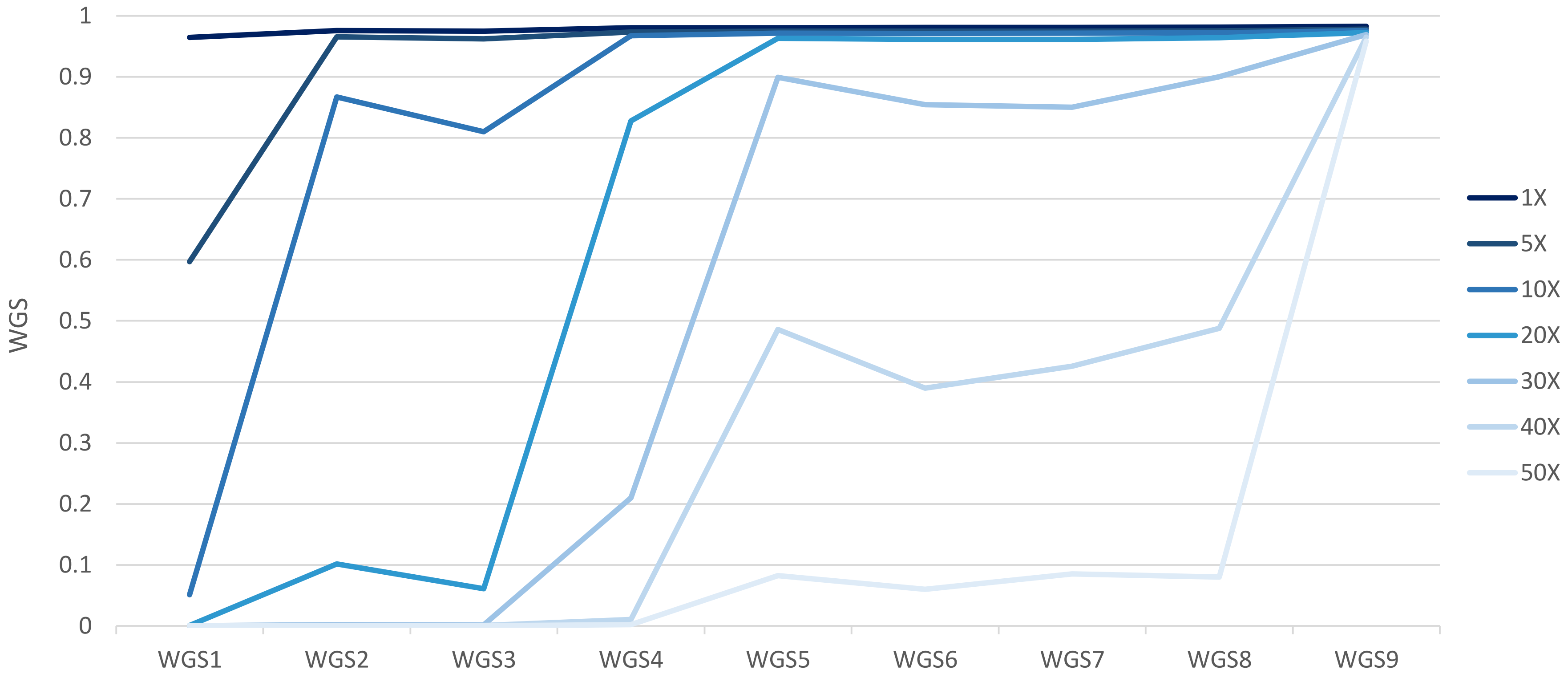
WGS metrics

- CollectWgsMetrics
  - Excluded bases



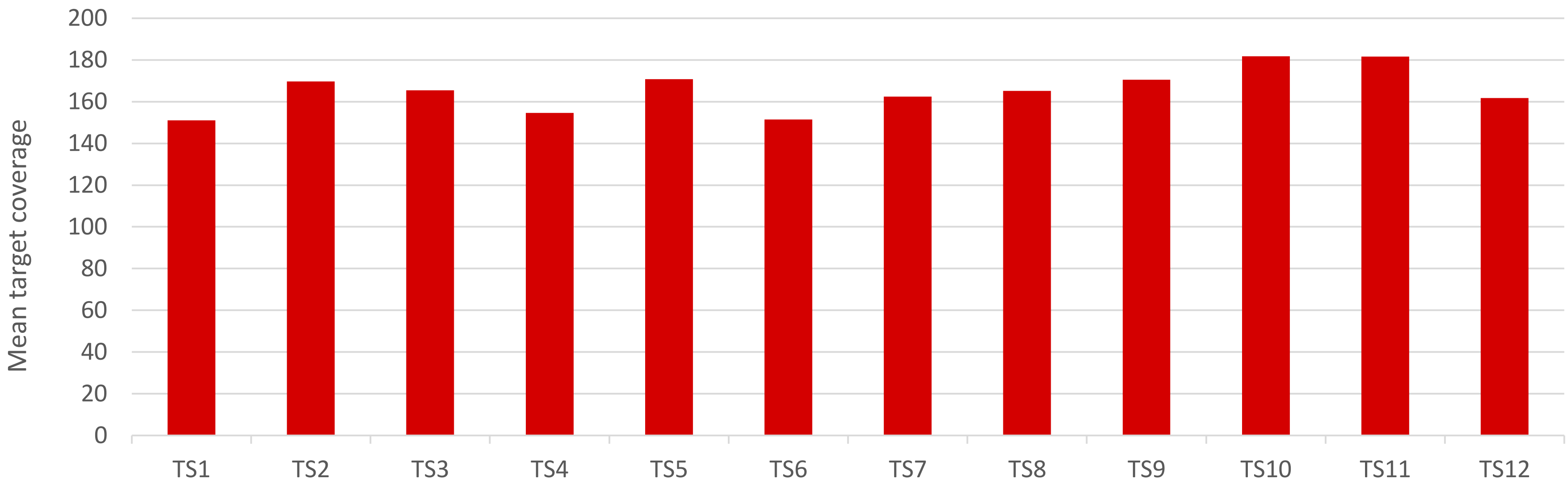
WGS metrics

- CollectWgsMetrics
  - Proportion of WGS covered at 1-50X



Targeted metrics

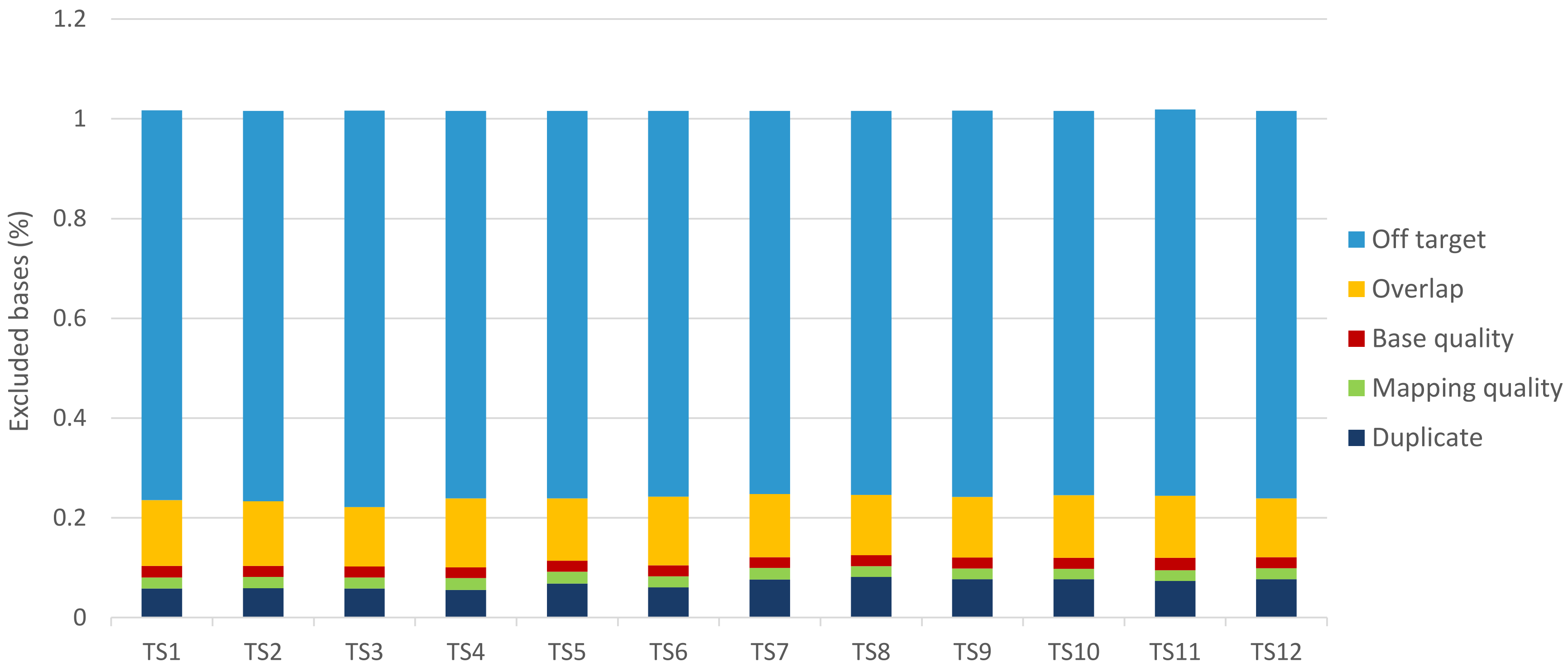
- CollectHsMetrics
  - Mean target coverage



```
java -jar picard.jar CollectHsMetrics I=sample.bam
O=sample_hs_metrics.txt R=reference_sequence.fasta
BAIT_INTERVALS=bait.interval_list
TARGET_INTERVALS=target.interval_list
```

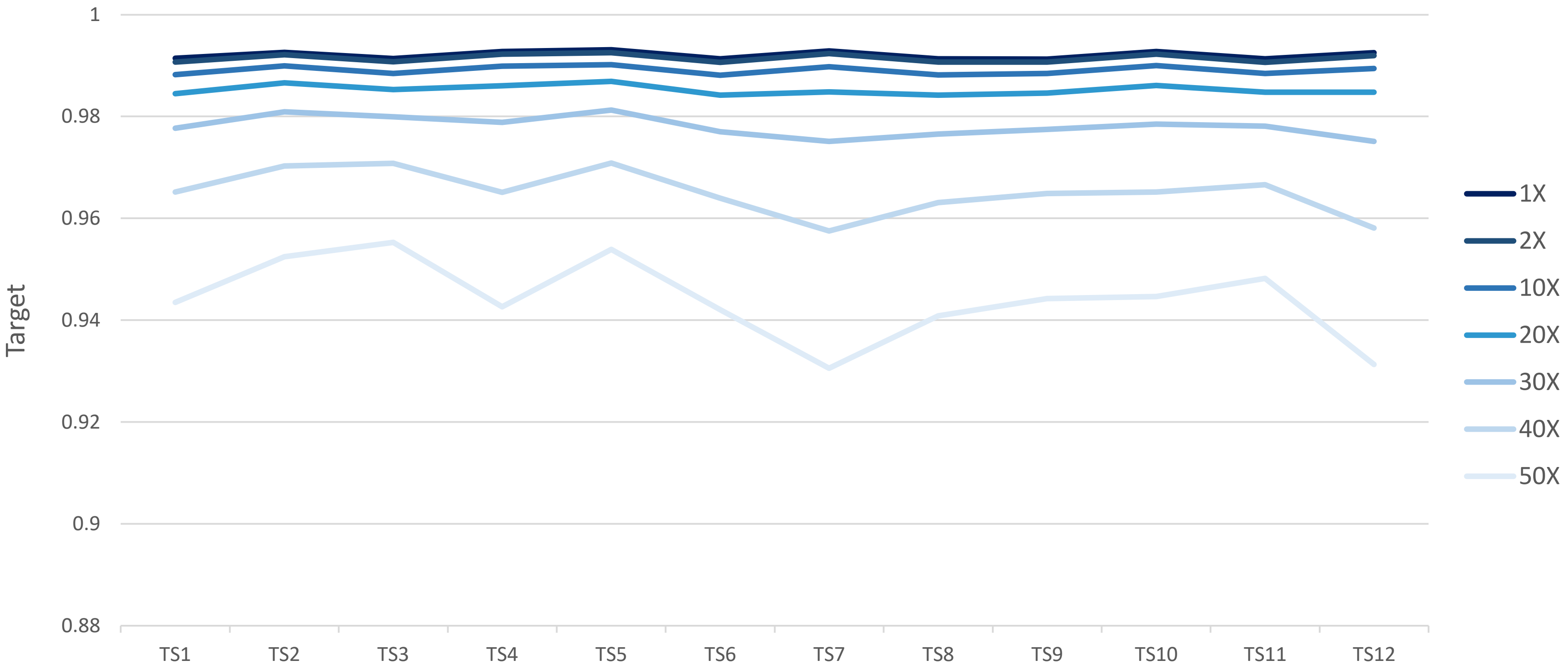
Targeted metrics

- CollectHsMetrics
  - Excluded bases



Targeted metrics

- CollectHsMetrics
  - ○ Proportion of target covered at 1-50X



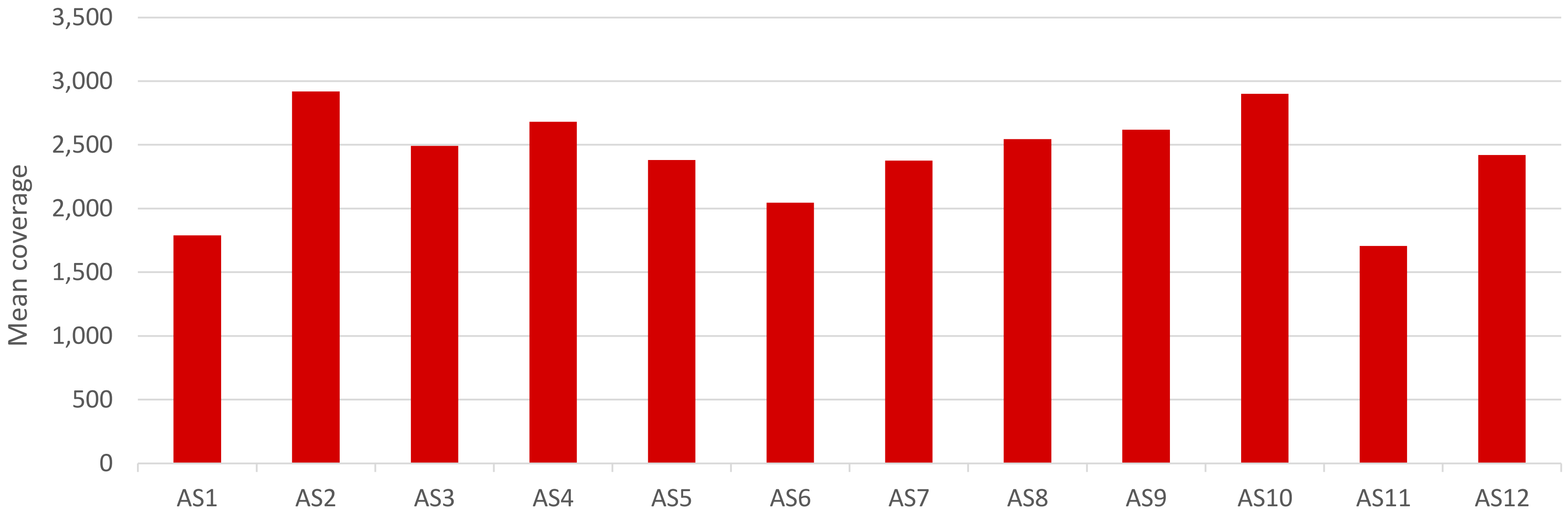
Targeted metrics

- CollectHsMetrics
  - Capture efficiency

Sample	Selected bases	Fold enrichment	Fold_80_base penalty
TS1	0.95	101.21	1.86
TS2	0.94	100.95	1.91
TS3	0.94	100.03	1.84
TS4	0.94	101.63	1.89
TS5	0.94	100.50	1.88
TS6	0.94	101.53	1.87
TS7	0.95	102.72	2.08
TS8	0.95	101.88	2.01
TS9	0.95	102.28	2.03
TS10	0.95	102.85	2.09
TS11	0.95	102.67	2.06
TS12	0.95	101.78	2.07

Amplicon metrics

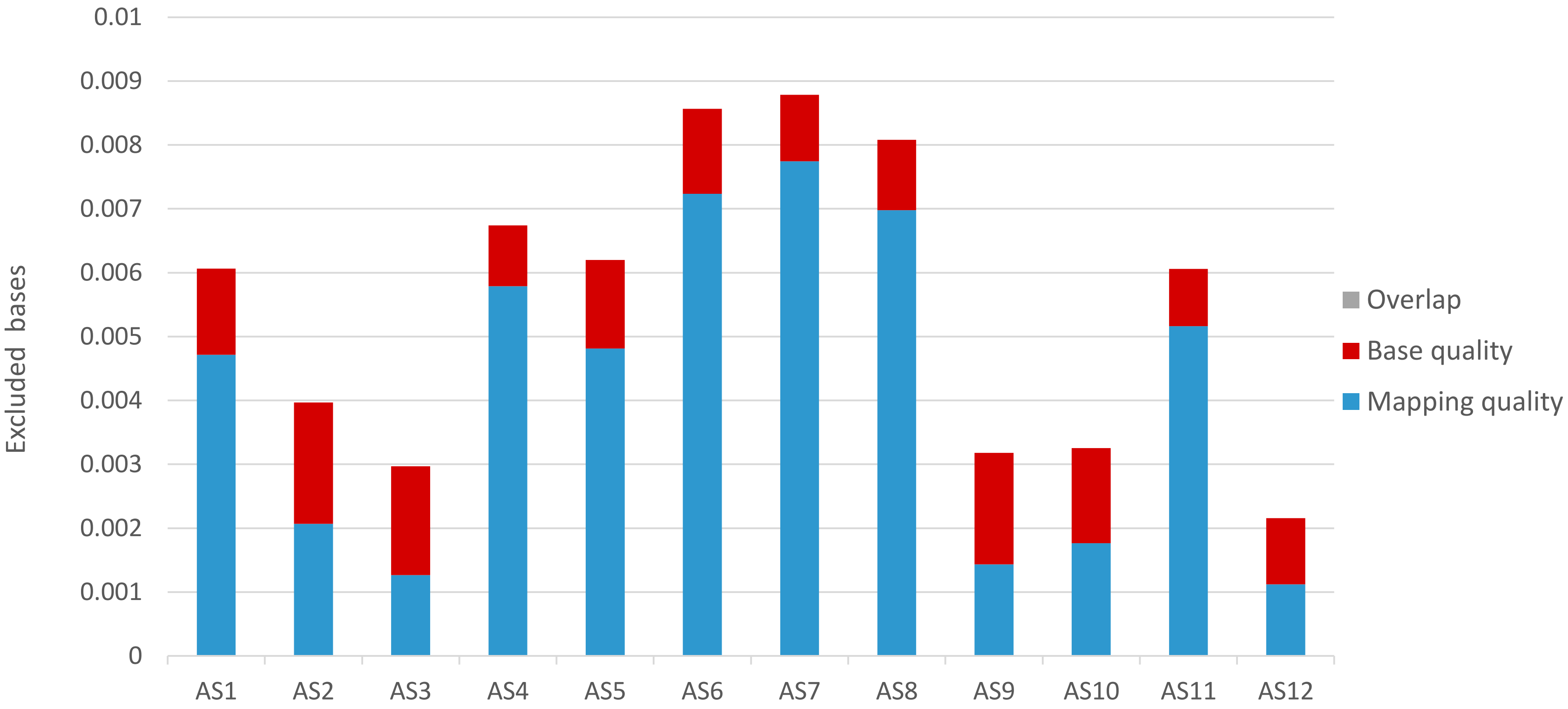
- CollectTargetedPcrMetrics
  - Mean target coverage



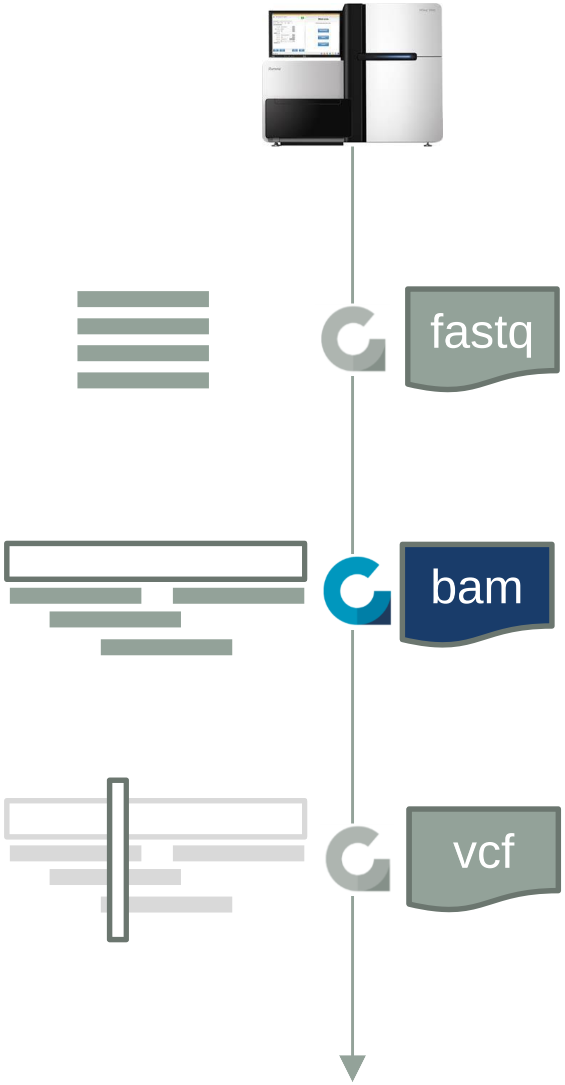
```
java -jar picard.jar CollectTargetedPcrMetrics I=input.bam
O=sample_pcr_metrics.txt R=reference_sequence.fasta
AMPLICON_INTERVALS=amplicon.interval_list
TARGET_INTERVALS=targets.interval_list
```

Amplicon metrics

- CollectTargetedPcrMetrics
  - Excluded bases



Overview



Sequencing

Sequence of nucleotides within a read?

Genomic location of a read?

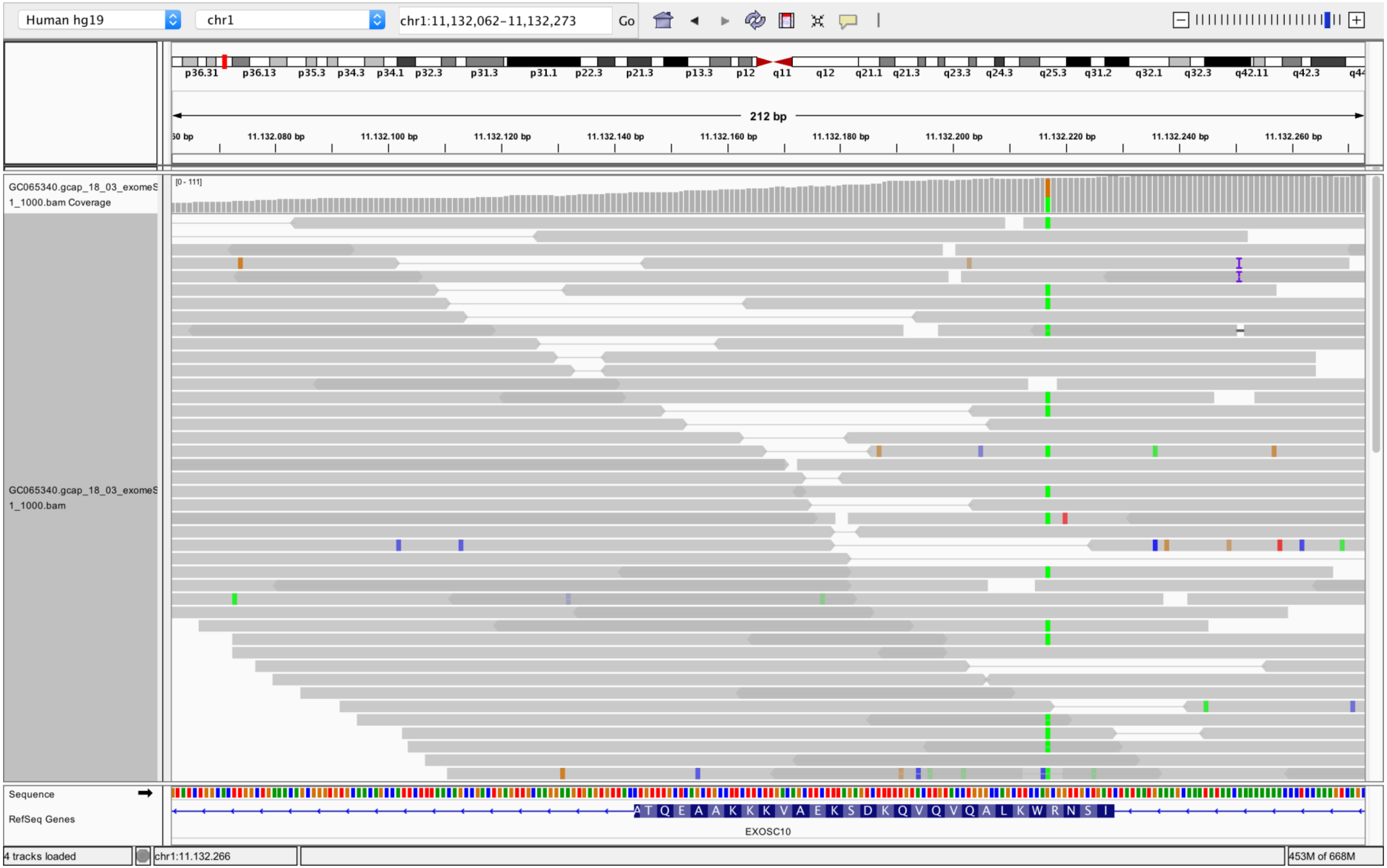
Genotype of a genomic location?

Questions?

GENOMICS
CORELEUVEN

Luc Dehaspe
Erika Souche

Example 1



Human hg19

chr1

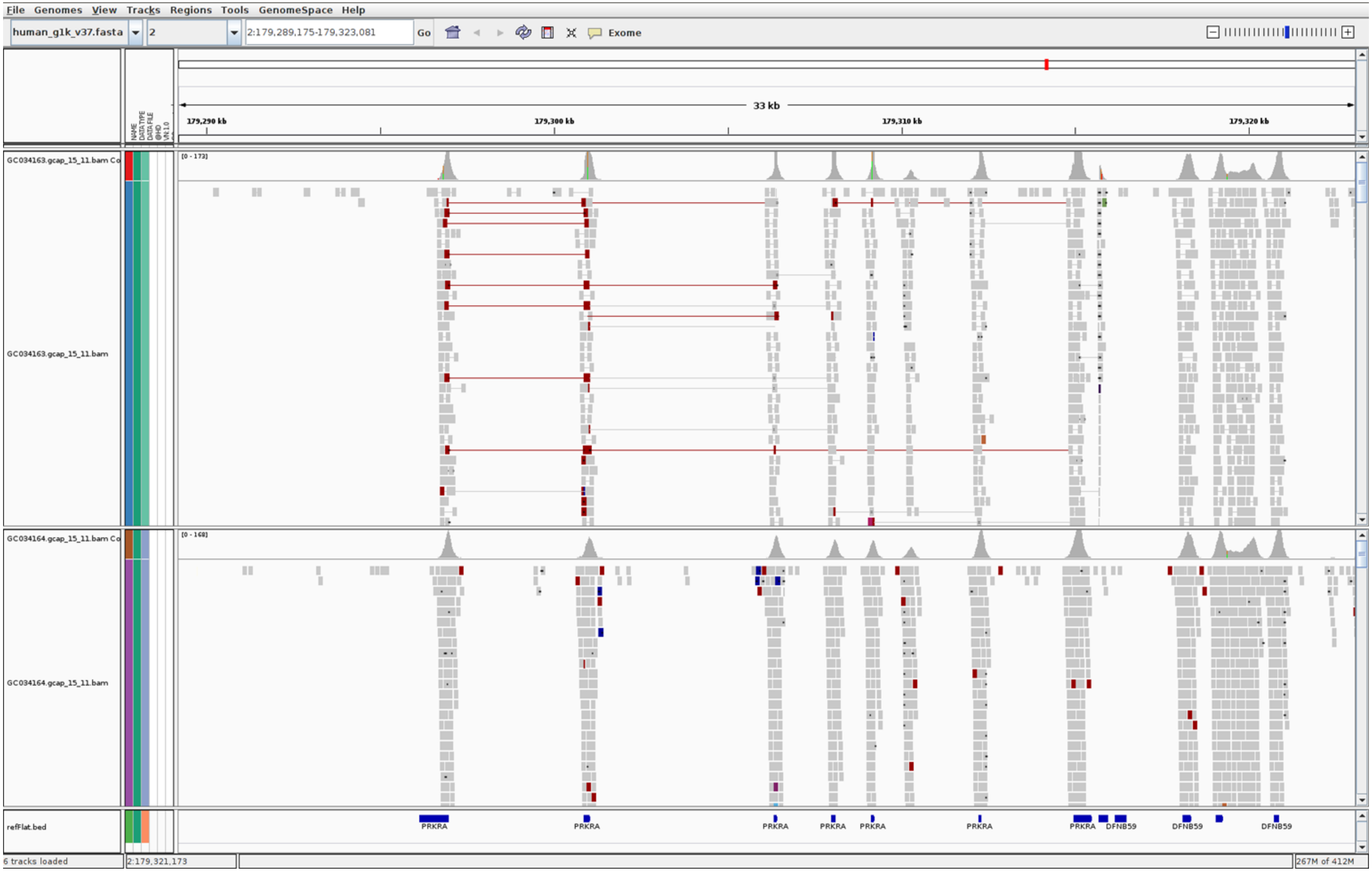
chr1:11,132,062-11,132,273

4 tracks loaded

chr1:11.132.266

453M of 668M

Special cases



human_g1k_v37.fasta

2

2:179,289,175-179,323,081

GC034163.gcap_15_11.bam Coverage

GC034163.gcap_15_11.bam

GC034164.gcap_15_11.bam Coverage

GC034164.gcap_15_11.bam

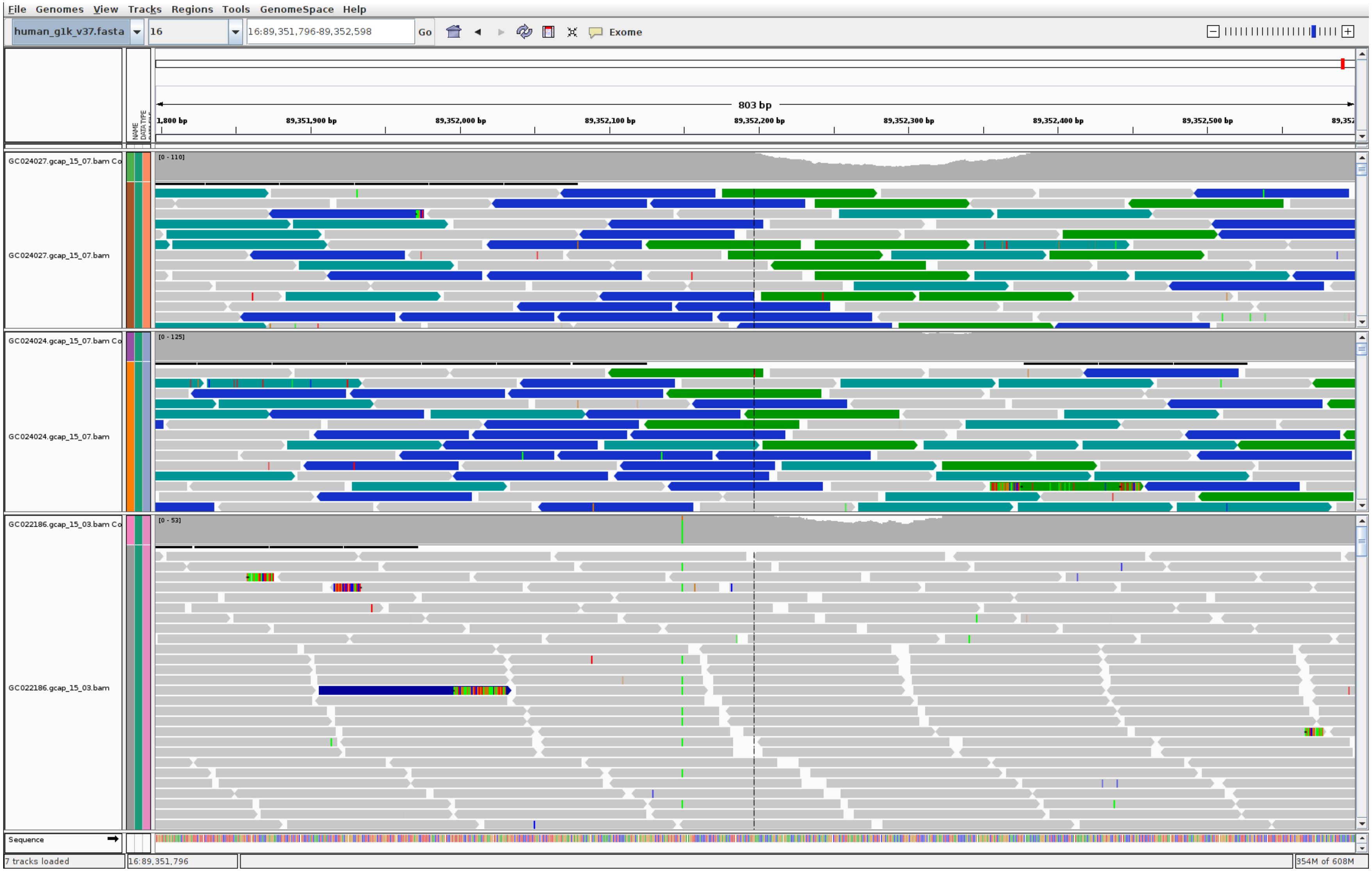
refFlat.bed

6 tracks loaded

2:179,321,173

267M of 412M

Special cases


File Genomes View Tracks Regions Tools GenomeSpace Help
human_g1k_v37.fasta
16
16:89,351,796-89,352,598
Go
Exome
803 bp
1,800 bp
89,351,900 bp
89,352,000 bp
89,352,100 bp
89,352,200 bp
89,352,300 bp
89,352,400 bp
89,352,500 bp
89,352
NAME
DATA TYPE
GC024027.gcap_15_07.bam Co
[0 - 110]
GC024027.gcap_15_07.bam
GC024024.gcap_15_07.bam Co
[0 - 125]
GC024024.gcap_15_07.bam
GC022186.gcap_15_03.bam Co
[0 - 53]
GC022186.gcap_15_03.bam
Sequence
7 tracks loaded
16:89,351,796
354M of 608M