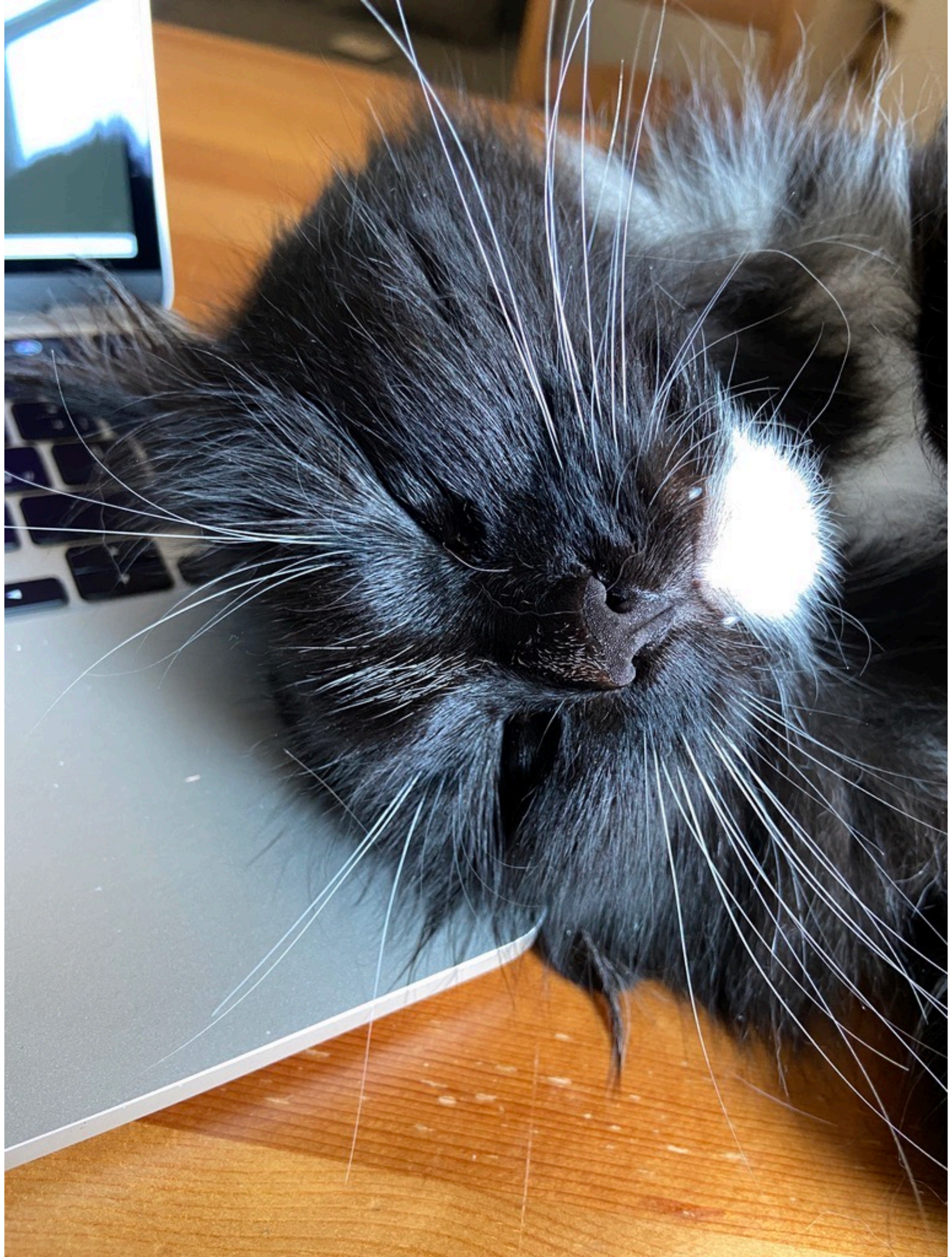
# Structural variant activities!!

1. SV quiz
2. SV calling tutorial

SVs are awesome and fun

QUIZ!

With prizes ;)



# SV classification (aka SV quiz)

You can find all the files necessary to answer the questions in the folder “SV_quiz”. Within this folder you will find a subfolder corresponding to each of the questions. The tools we suggest to use are all freely available (IGV, MAFFT or LAST online) but Censor/Repbase which has a limited free use. Nonetheless, the limits of Censor should allow all of you to answer these questions. If you have any problem with Censor please ask me (Vale) or Alex :)

Visualise BAM files

IGV: Amazon instance or download on your computer (igvteam/igv)

Make dotplots

MAFFT: https://mafft.cbrc.jp/alignment/server/index.html

LAST: http://lastweb.cbrc.jp

Repeat database Censor (sequence homology to Repbase repeat database): https://www.girinst.org/censor/index.php

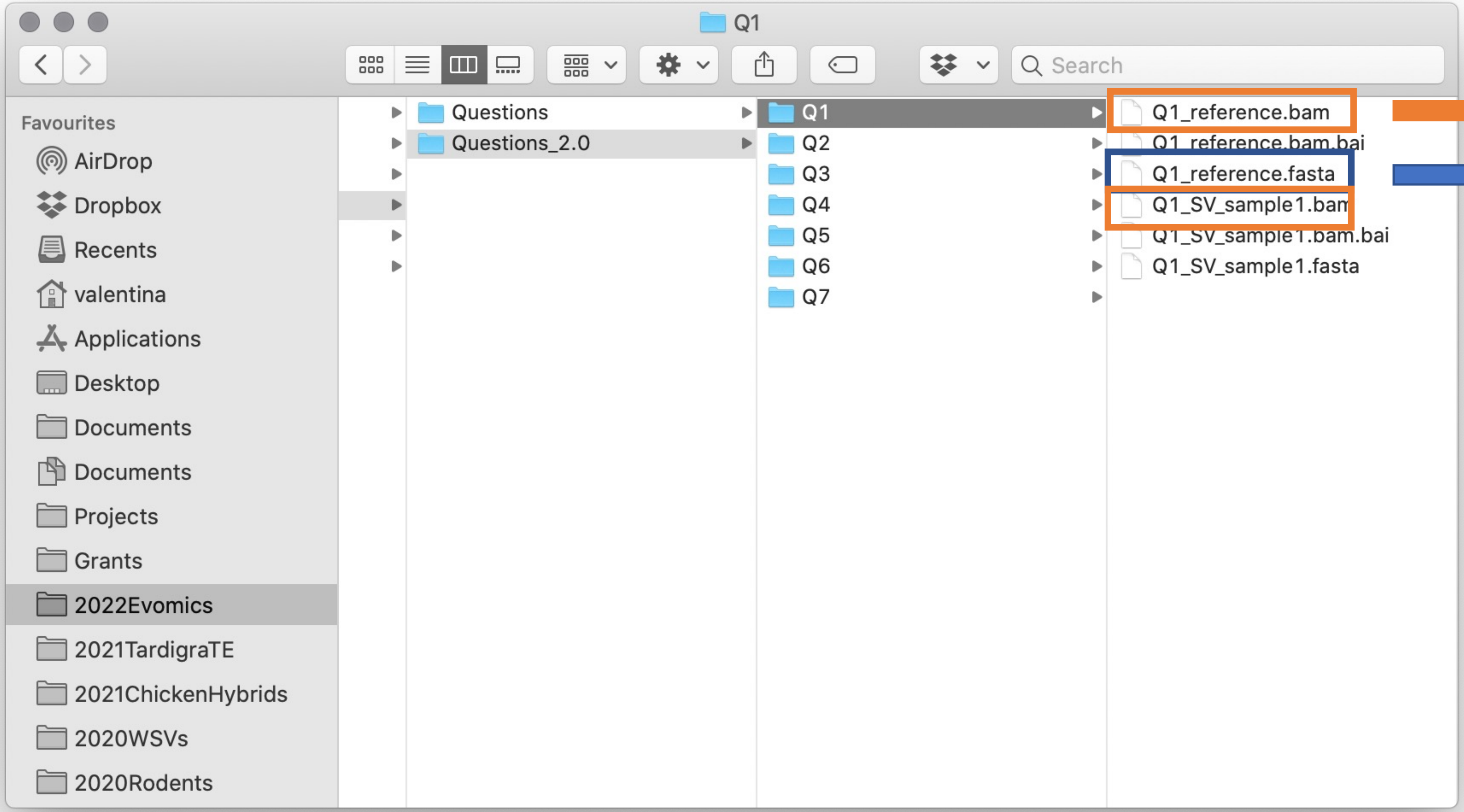
# Q1: What is it?

## Hint: Is there any signature of incongruent read mapping along the sequence of interest?

- Open the sequence and respective BAM files in IGV.

# Data for the quiz

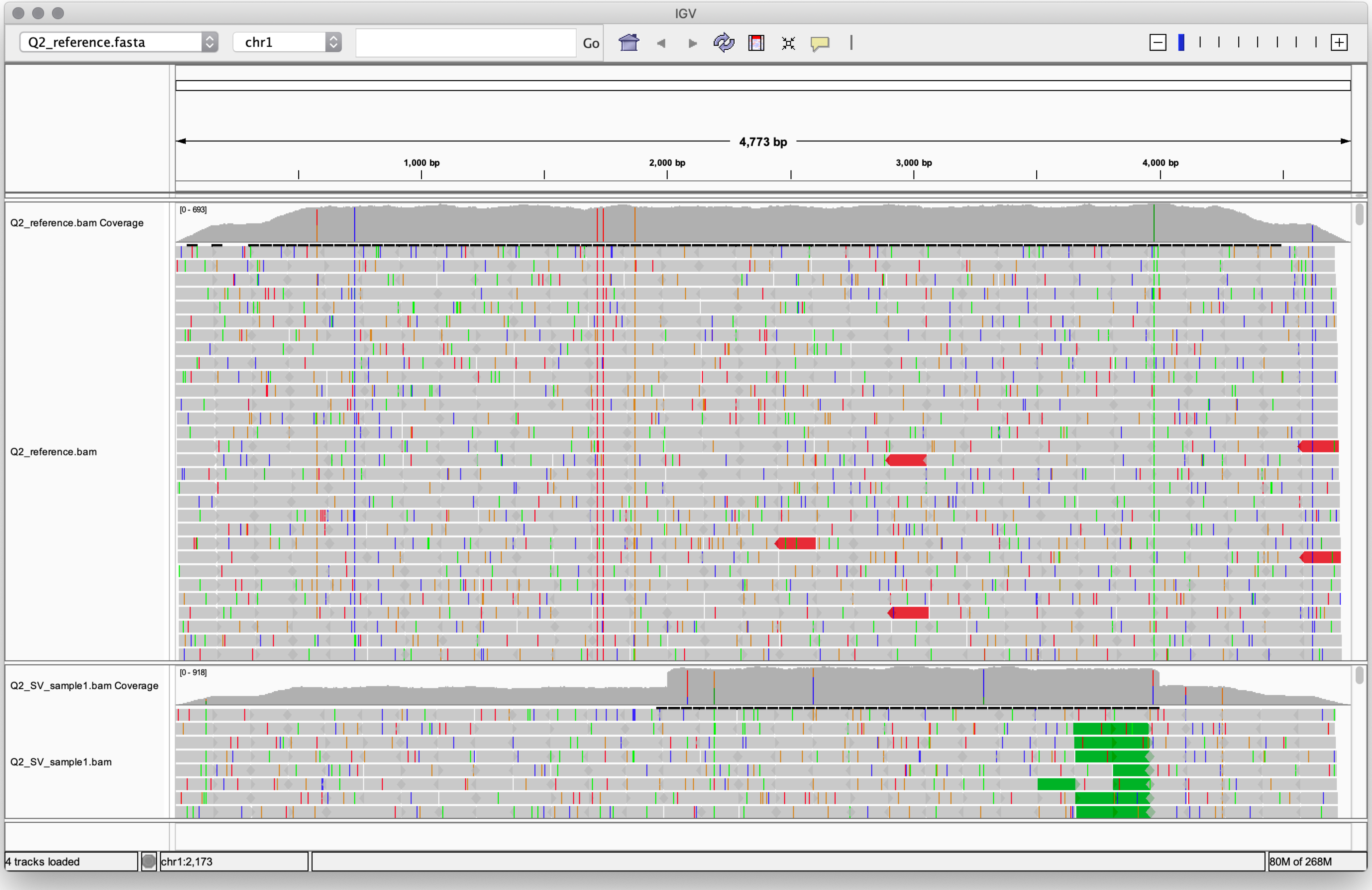
**Reference** fasta to be uploaded on IGV through the menu Genomes > Load Genome from File
BAM file of reads from reference mapped to reference: File > Load from File
BAM file of reads from samples mapped to reference: File > Load from File



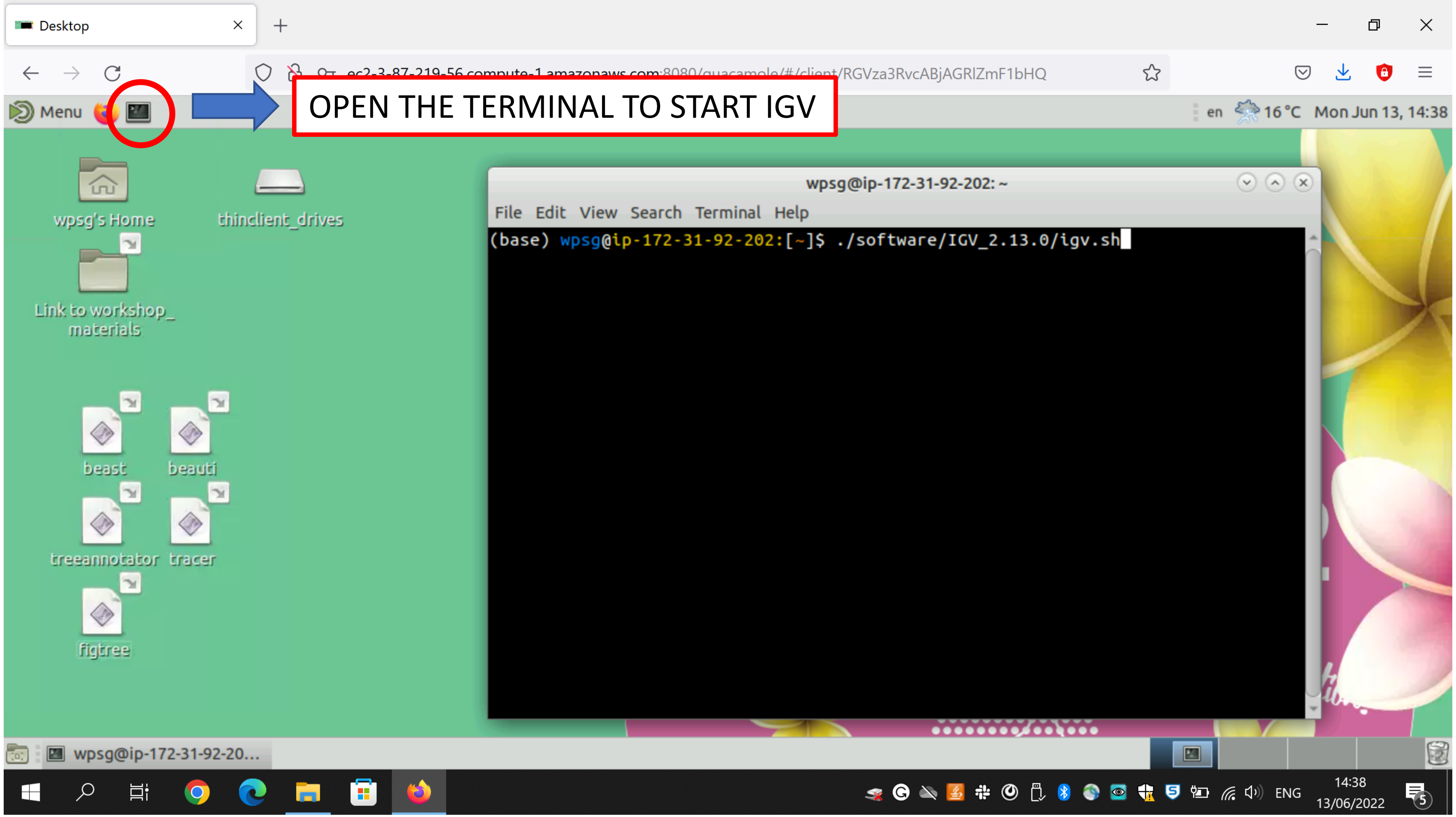
Upload as “File”

Upload as “Genome”

Reference fasta always named in the form:
Q*_reference.fasta
BAM reference: Q*_reference.bam
BAM sample(s): Q*_sample*.bam


IGV
Q2_reference.fasta
chr1
Go
4,773 bp
1,000 bp
2,000 bp
3,000 bp
4,000 bp
[0 - 693]
Q2_reference.bam Coverage
Q2_reference.bam
[0 - 918]
Q2_SV_sample1.bam Coverage
Q2_SV_sample1.bam
4 tracks loaded
chr1:2,173
80M of 268M

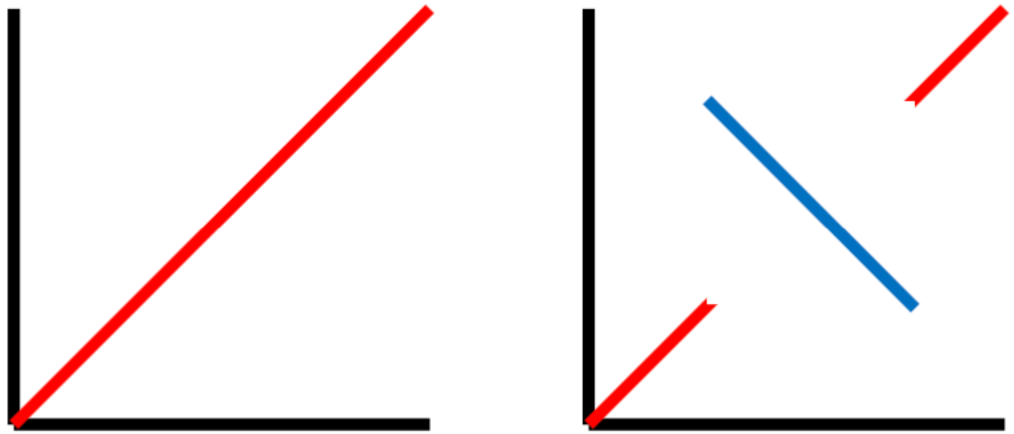
# Open IGV in Guacamole



OPEN THE TERMINAL TO START IGV

```
(base) wpsg@ip-172-31-92-202:[~]$ ./software/IGV_2.13.0/igv.sh
```

# Dot plot

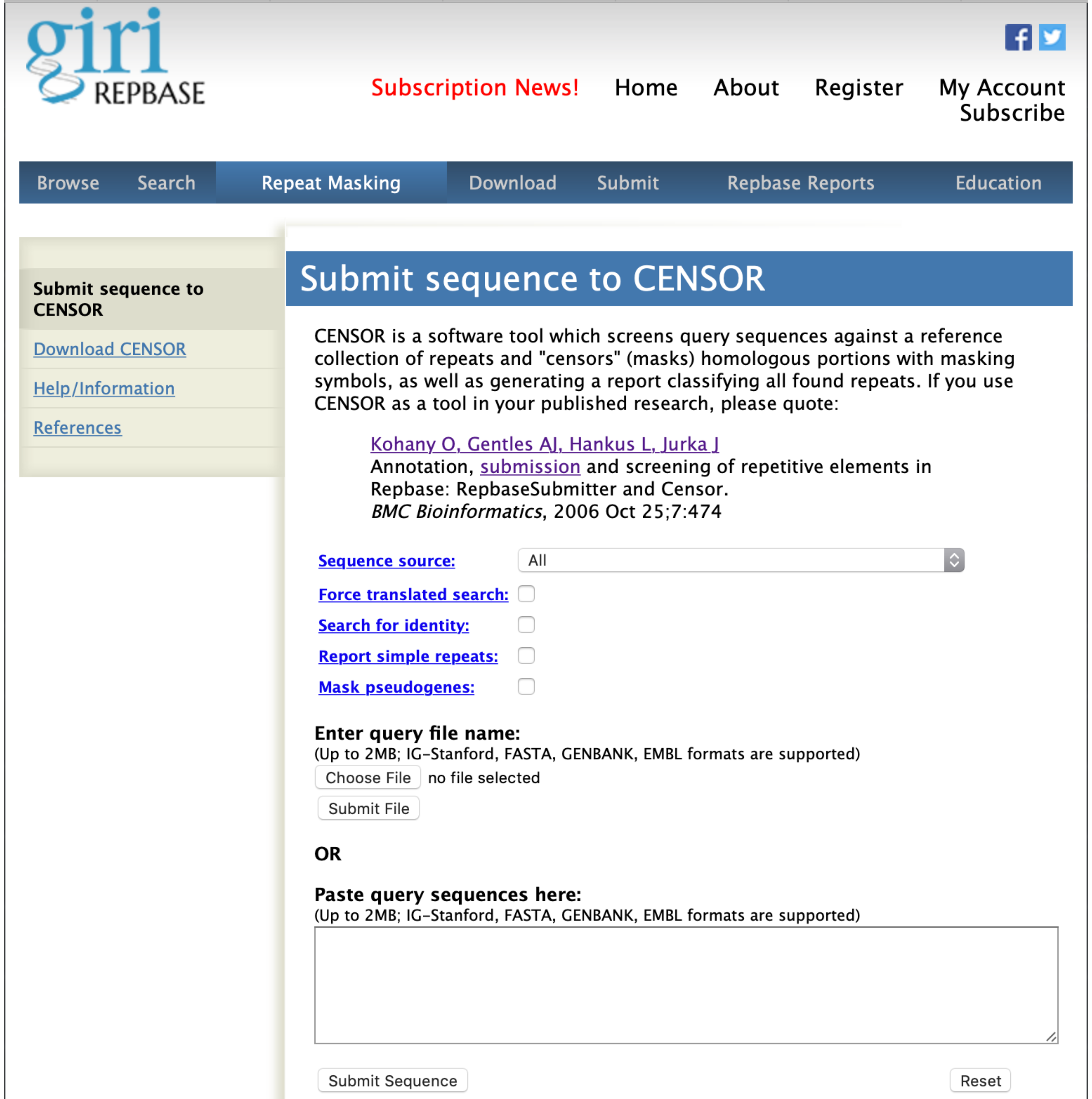


You can get important information by aligning sequences, not only reads

Copy reference and sample sequences into LAST or MAFFT to get a dot plot

Is your SV a transposon????

Go to Repbase, click on Repeat Masking menu to run Censor on your sequences and find homologies with known transposable elements



## Submit sequence to CENSOR

CENSOR is a software tool which screens query sequences against a reference collection of repeats and "censors" (masks) homologous portions with masking symbols, as well as generating a report classifying all found repeats. If you use CENSOR as a tool in your published research, please quote:

> Kohany O, Gentles AJ, Hankus L, Jurka J
> Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor.
> *BMC Bioinformatics*, 2006 Oct 25;7:474

Sequence source: All

Force translated search: □

Search for identity: □

Report simple repeats: □

Mask pseudogenes: □

**Enter query file name:**
(Up to 2MB; IG-Stanford, FASTA, GENBANK, EMBL formats are supported)

Choose File no file selected

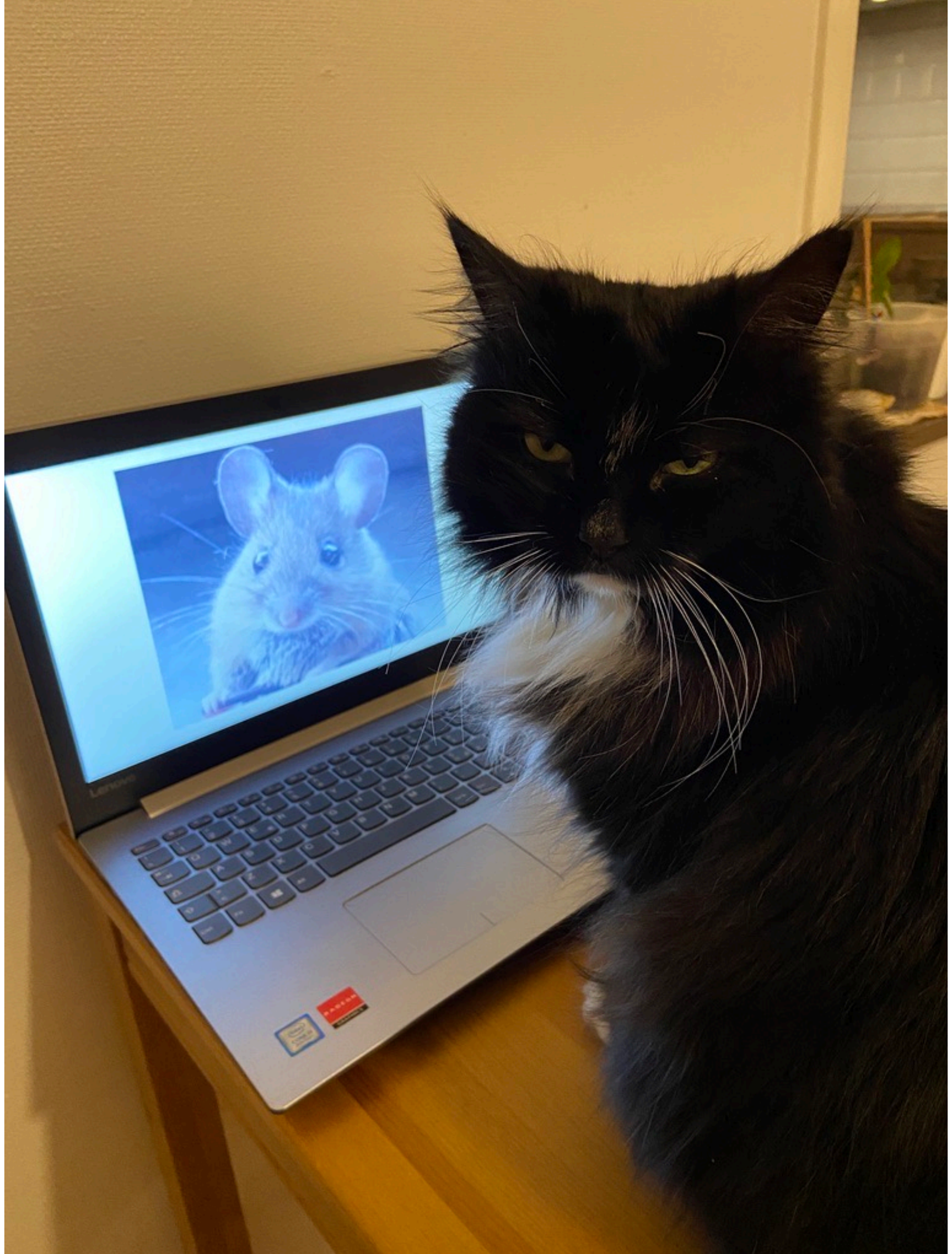
Submit File

**OR**

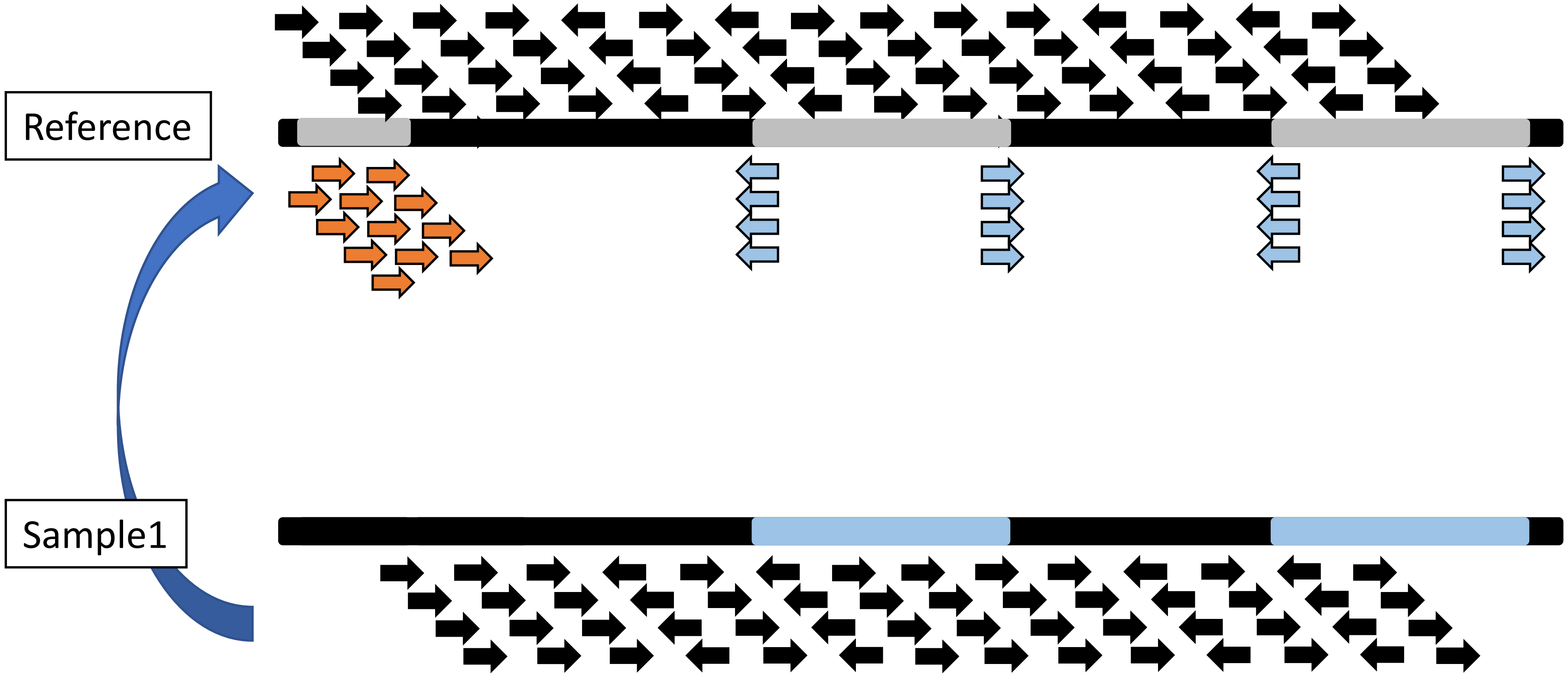
**Paste query sequences here:**
(Up to 2MB; IG-Stanford, FASTA, GENBANK, EMBL formats are supported)

Submit Sequence Reset

SVs are awesome
but painful



# SV simulation



# SV simulation

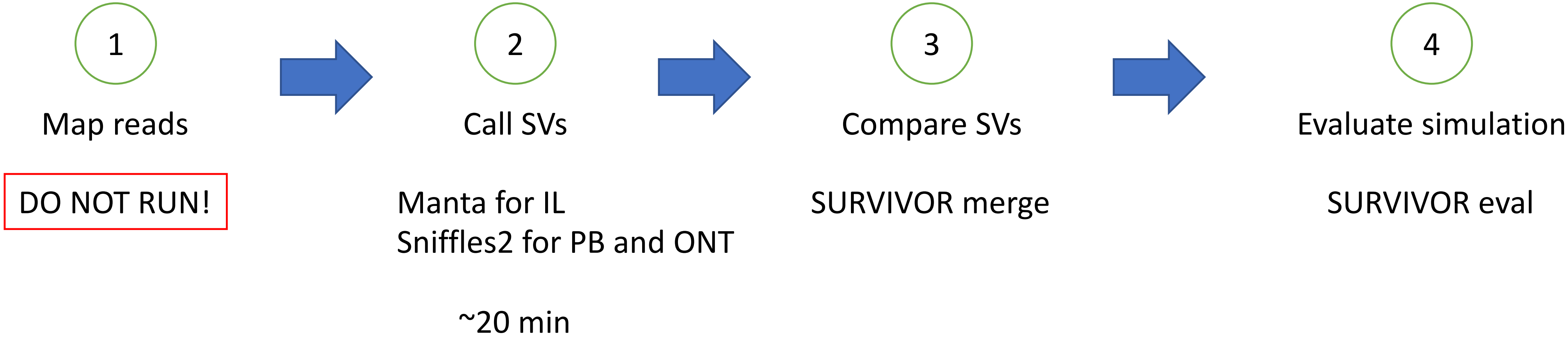
Simulated libraries 30X

| | |
|---|---|
| Illumina paired-end reads | 150 bp x 2, 500 bp insertion size |
| PacBio reads | 6 kb |
| Nanopore reads | 5.5 kb |

Two conda envs!

One only for Manta

The second (SV_Env) for all the other analysis



1 Map reads

DO NOT RUN!

2 Call SVs

Manta for IL
Sniffles2 for PB and ONT

~20 min

3 Compare SVs

SURVIVOR merge

4 Evaluate simulation

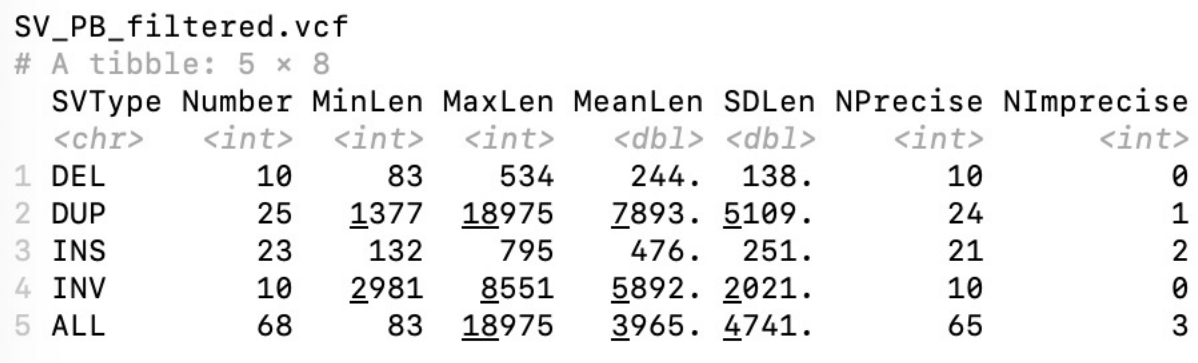
SURVIVOR eval

# WRAP-UP

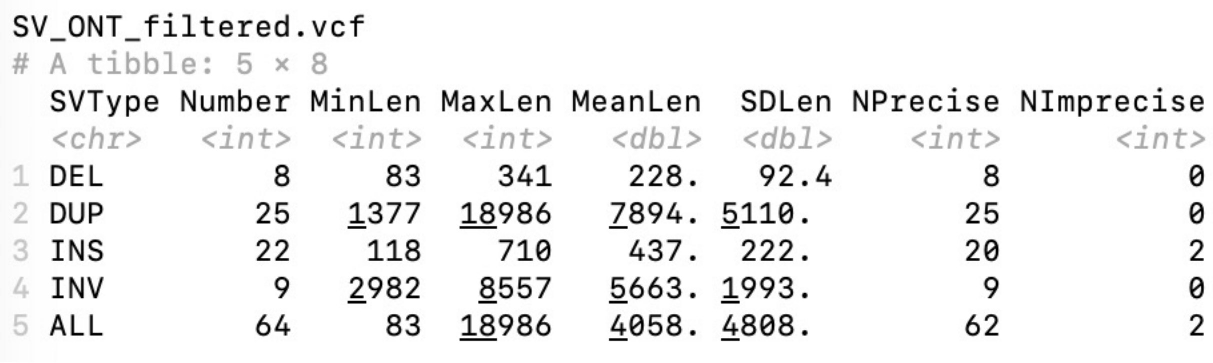
2 SV calling tutorial

# SV simulation

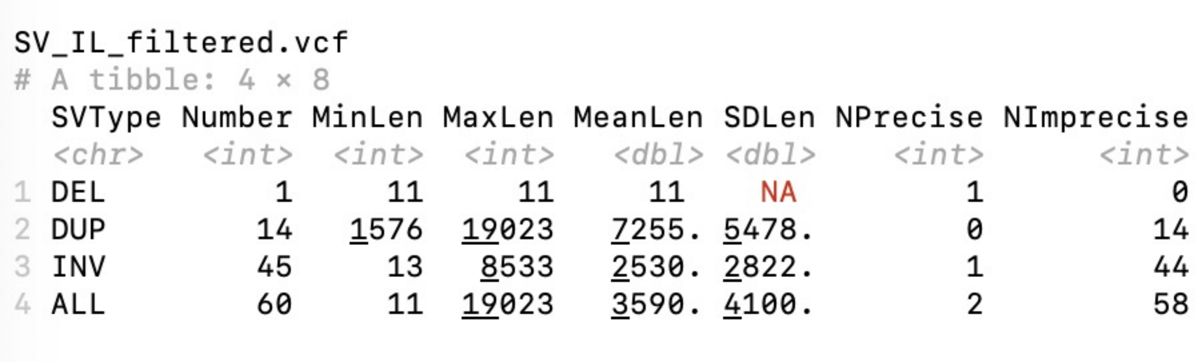
SV CALLING



```
SV_PB_filtered.vcf
# A tibble: 5 × 8
  SVType Number MinLen MaxLen MeanLen SDLen NPrecise NImprecise
  <chr>   <int>  <int>  <int>   <dbl> <dbl>    <int>      <int>
1 DEL        10     83    534    244.  138.       10          0
2 DUP        25   1377  18975   7893. 5109.       24          1
3 INS        23    132    795    476.  251.       21          2
4 INV        10   2981   8551   5892. 2021.       10          0
5 ALL        68     83  18975   3965. 4741.       65          3
```

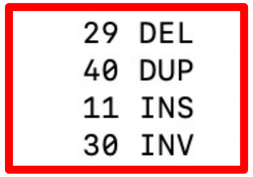


```
SV_ONT_filtered.vcf
# A tibble: 5 × 8
  SVType Number MinLen MaxLen MeanLen  SDLen NPrecise NImprecise
  <chr>   <int>  <int>  <int>   <dbl>  <dbl>    <int>      <int>
1 DEL         8     83    341    228.   92.4        8          0
2 DUP        25   1377  18986   7894. 5110.        25          0
3 INS        22    118    710    437.  222.        20          2
4 INV         9   2982   8557   5663. 1993.         9          0
5 ALL        64     83  18986   4058. 4808.        62          2
```



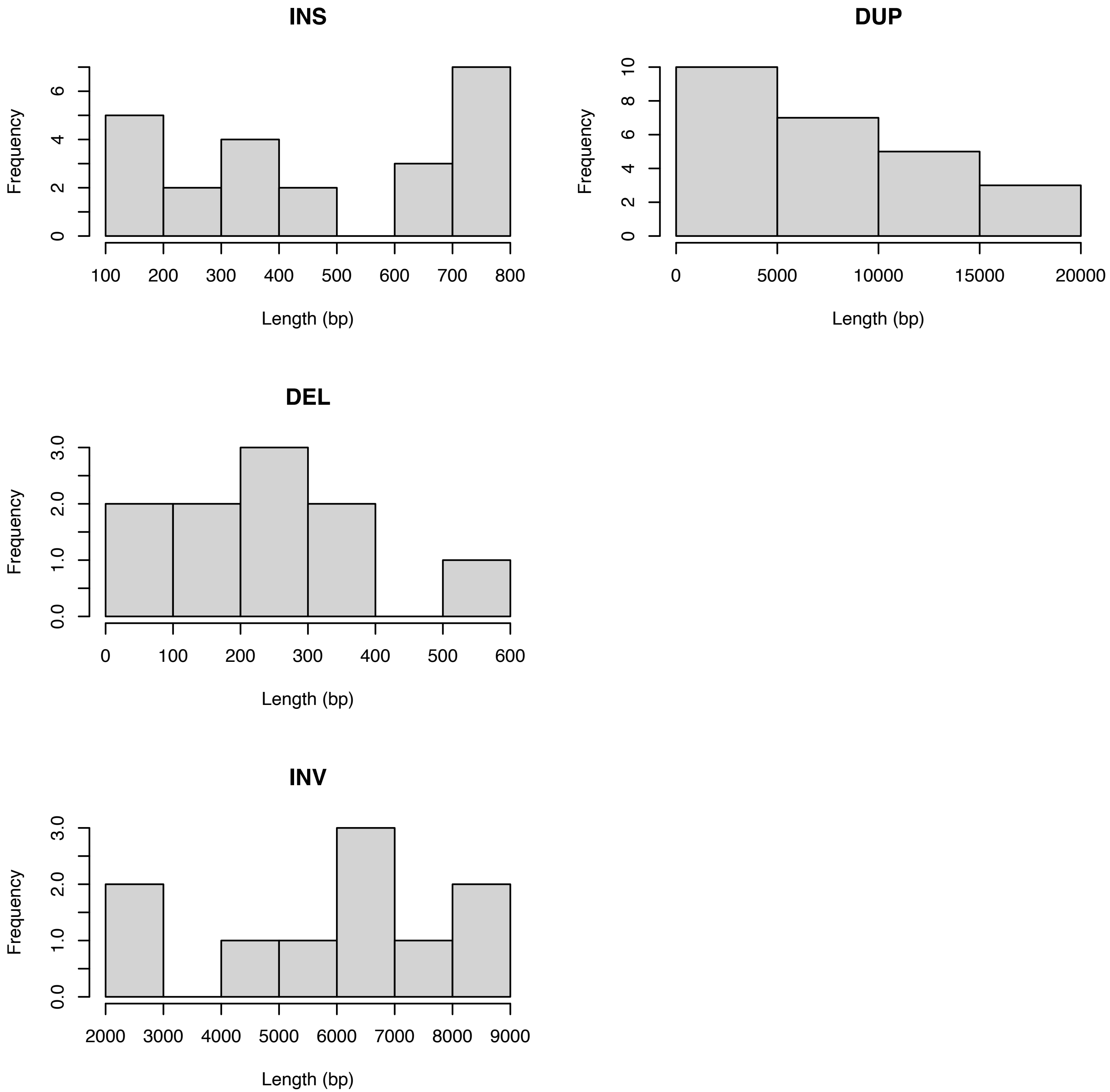
```
SV_IL_filtered.vcf
# A tibble: 4 × 8
  SVType Number MinLen MaxLen MeanLen SDLen NPrecise NImprecise
  <chr>   <int>  <int>  <int>   <dbl> <dbl>    <int>      <int>
1 DEL         1     11     11      11    NA        1          0
2 DUP        14   1576  19023   7255. 5478.        0         14
3 INV        45     13   8533   2530. 2822.        1         44
4 ALL        60     11  19023   3590. 4100.        2         58
```

SIMULATED

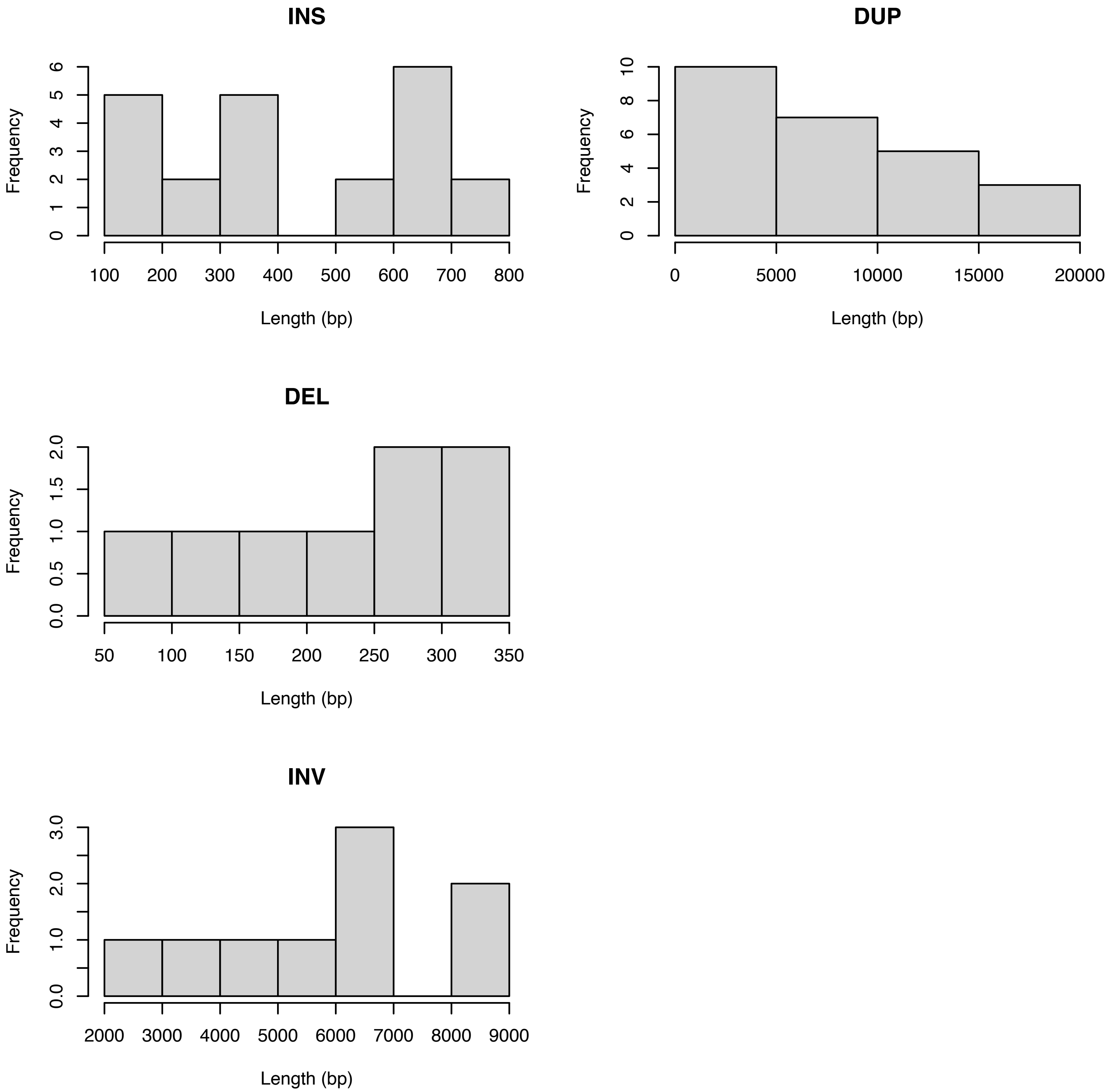


```
29 DEL
40 DUP
11 INS
30 INV
```

PACBIO

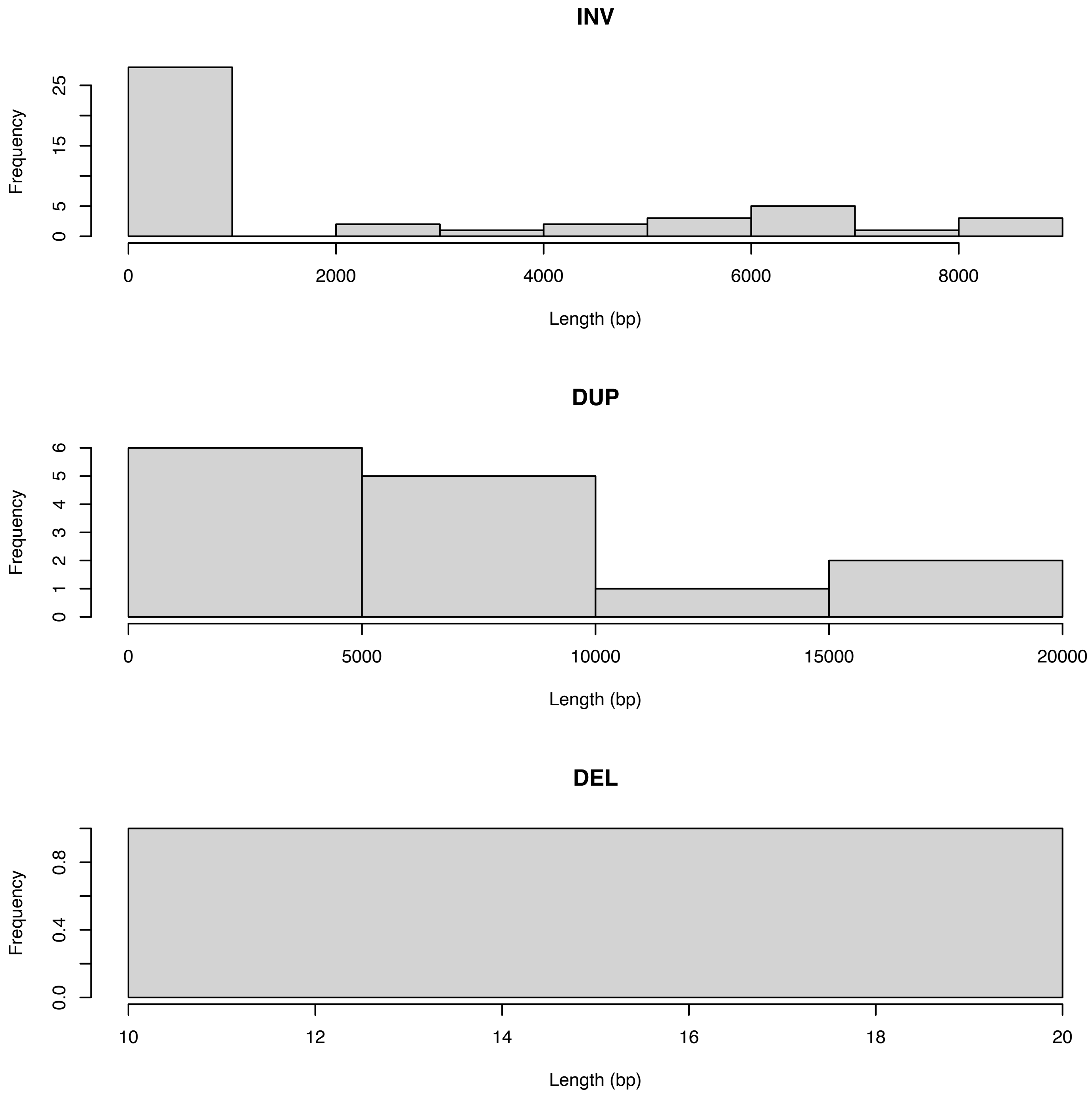

INS
Frequency
0
2
4
6
100
200
300
400
500
600
700
800
Length (bp)
DUP
Frequency
0
2
4
6
8
10
0
5000
10000
15000
20000
Length (bp)
DEL
Frequency
0.0
1.0
2.0
3.0
0
100
200
300
400
500
600
Length (bp)
INV
Frequency
0.0
1.0
2.0
3.0
2000
3000
4000
5000
6000
7000
8000
9000
Length (bp)

ONT



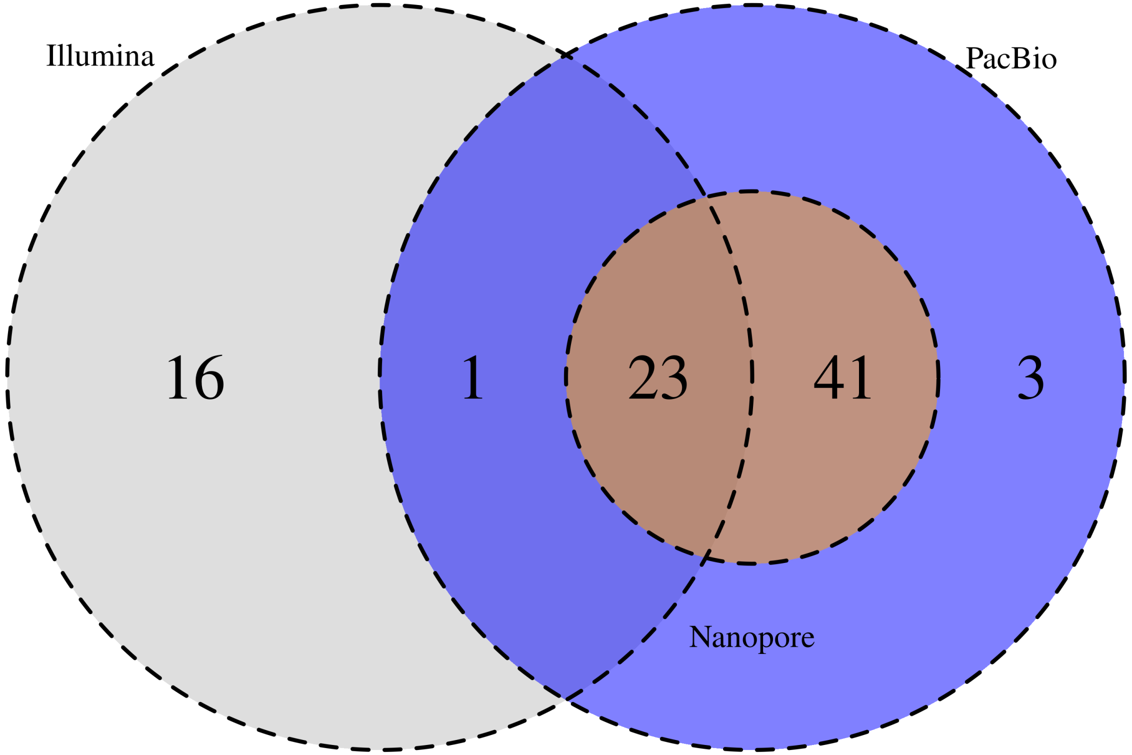
INS
Frequency
0
1
2
3
4
5
6
100
200
300
400
500
600
700
800
Length (bp)
DUP
Frequency
0
2
4
6
8
10
0
5000
10000
15000
20000
Length (bp)
DEL
Frequency
0.0
0.5
1.0
1.5
2.0
50
100
150
200
250
300
350
Length (bp)
INV
Frequency
0.0
1.0
2.0
3.0
2000
3000
4000
5000
6000
7000
8000
9000
Length (bp)

IL


INV
Frequency
0
5
15
25
0
2000
4000
6000
8000
Length (bp)
DUP
Frequency
0
1
2
3
4
5
6
0
5000
10000
15000
20000
Length (bp)
DEL
Frequency
0.0
0.4
0.8
10
12
14
16
18
20
Length (bp)

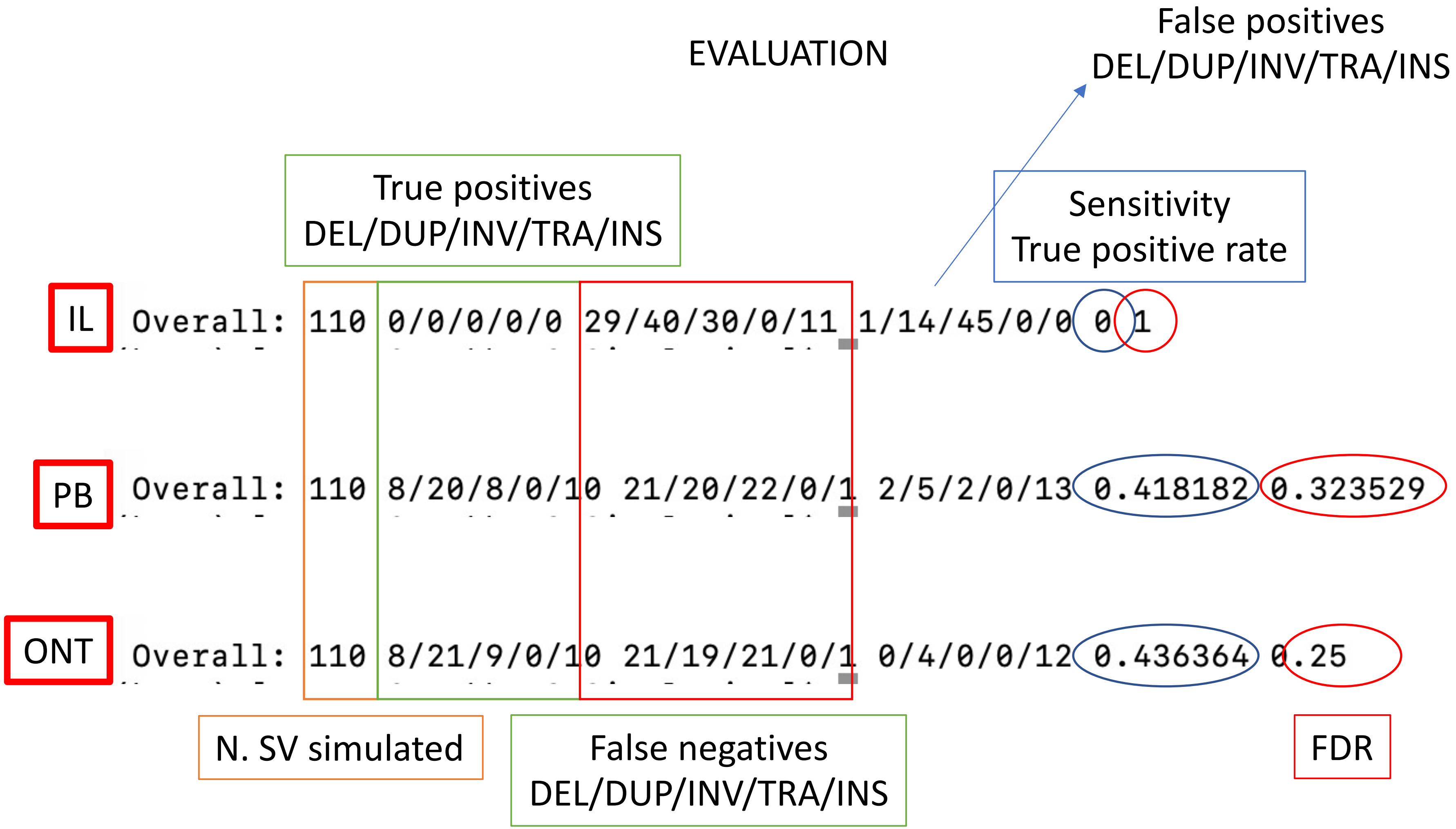
# SV simulation

COMPARISON



# SV simulation

EVALUATION



```
IL   Overall: 110 0/0/0/0/0 29/40/30/0/11 1/14/45/0/0 0 1

PB   Overall: 110 8/20/8/0/10 21/20/22/0/1 2/5/2/0/13 0.418182 0.323529

ONT  Overall: 110 8/21/9/0/10 21/19/21/0/1 0/4/0/0/12 0.436364 0.25
```

True positives
DEL/DUP/INV/TRA/INS

False positives
DEL/DUP/INV/TRA/INS

Sensitivity
True positive rate

N. SV simulated

False negatives
DEL/DUP/INV/TRA/INS

FDR

# So what????

Quality control and correction of reads
Check for mappability
Test more tools
Simulate on your data and reference to find FDR
Increase in read length
Assembly–based approach

AJHG

Volume 108, Issue 5, 6 May 2021, Pages 919-928

Report

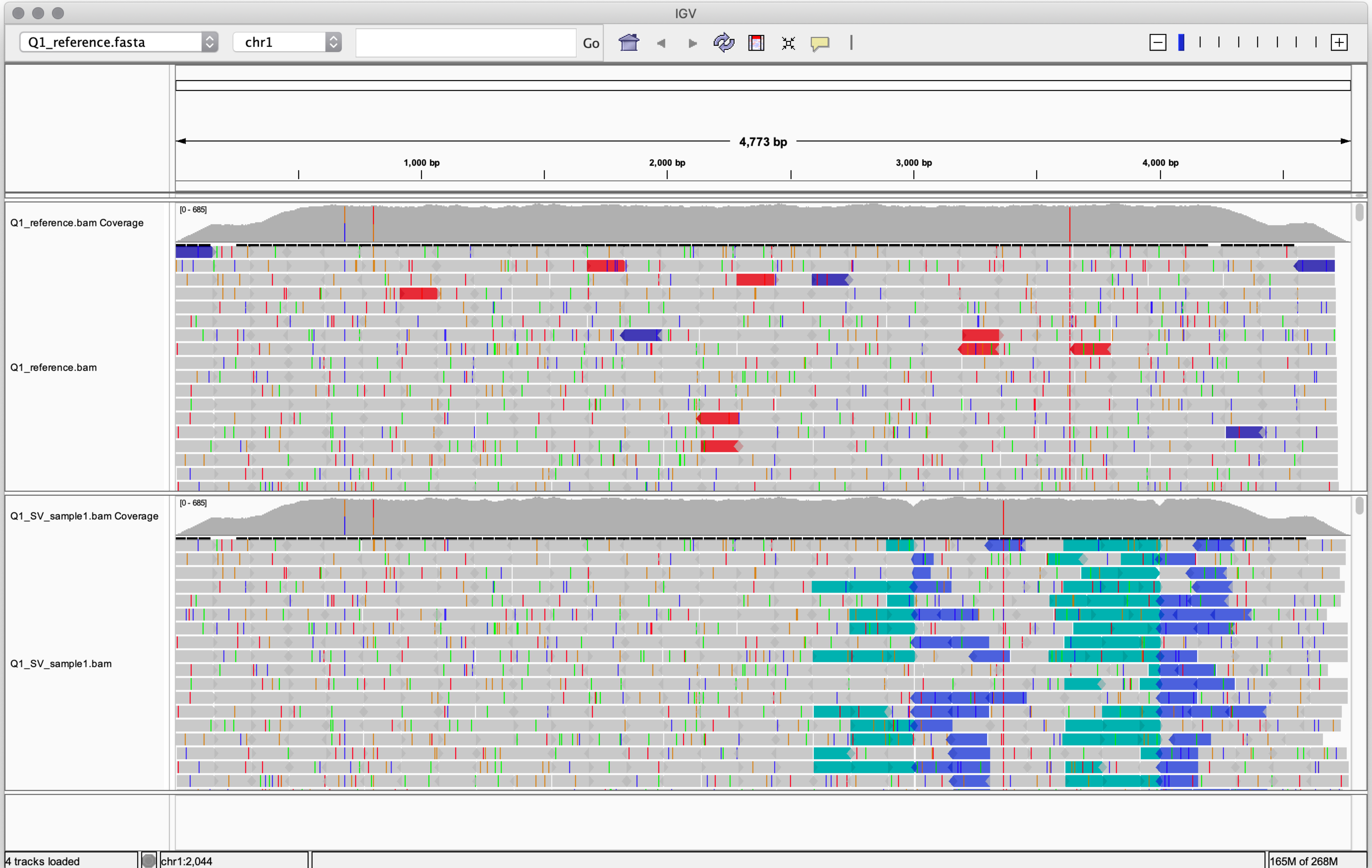
# Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies

Xuefang Zhao [1, 2, 3], Ryan L. Collins [1, 2, 4], Wan-Ping Lee [5], Alexandra M. Weber [6, 7], Yukyung Jun [5], Qihui Zhu [5], Ben Weisburd [2], Yongqing Huang [8], Peter A. Audano [9], Harold Wang [1, 2], Mark Walker [2, 3], Chelsea Lowther [1, 2, 3], Jack Fu [1, 2, 3], Human Genome Structural Variation Consortium, Mark B. Gerstein [10], Scott E. Devine [11], Tobias Marschall [12], Jan O. Korbel [13, 14] … Michael E. Talkowski [1, 2, 3, 4]

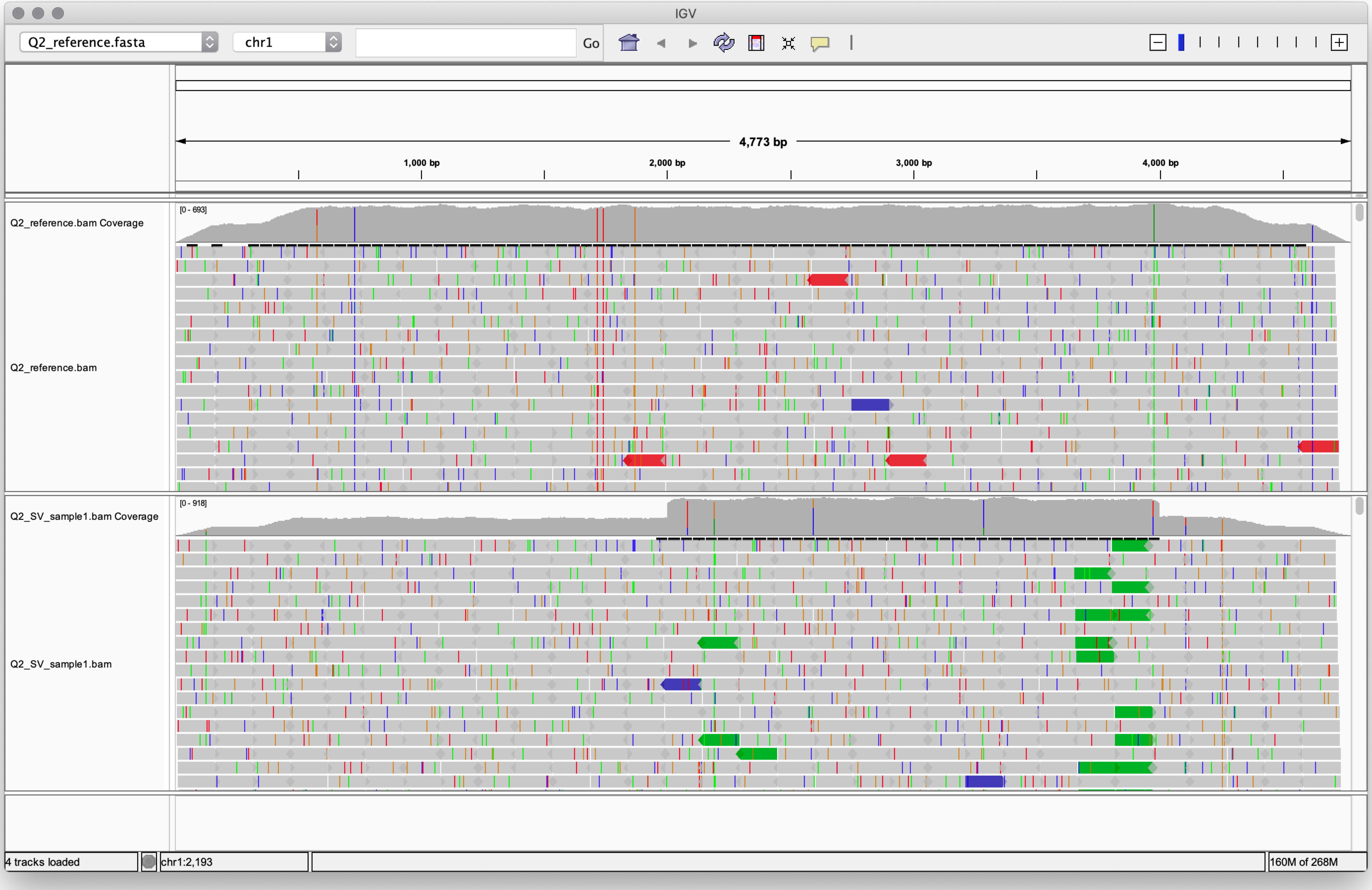
Finally, we explored the concordance of SV detection for a class of SVs that is strongly enriched for pathogenic variation and appears to be a significant blind spot for long-read assembly technologies: large CNVs captured by depth-based analyses from srWGS. Our initial analyses suggested that lrWGS assembly methods failed to capture all but one of the small number of large (>5 kb) CNVs that could be detected by srWGS read-depth methods in three probands (average size = 14.7 kb). Recognizing the limitation of read-depth analyses to

# 1 SV quiz → ANSWERS and PRIZES

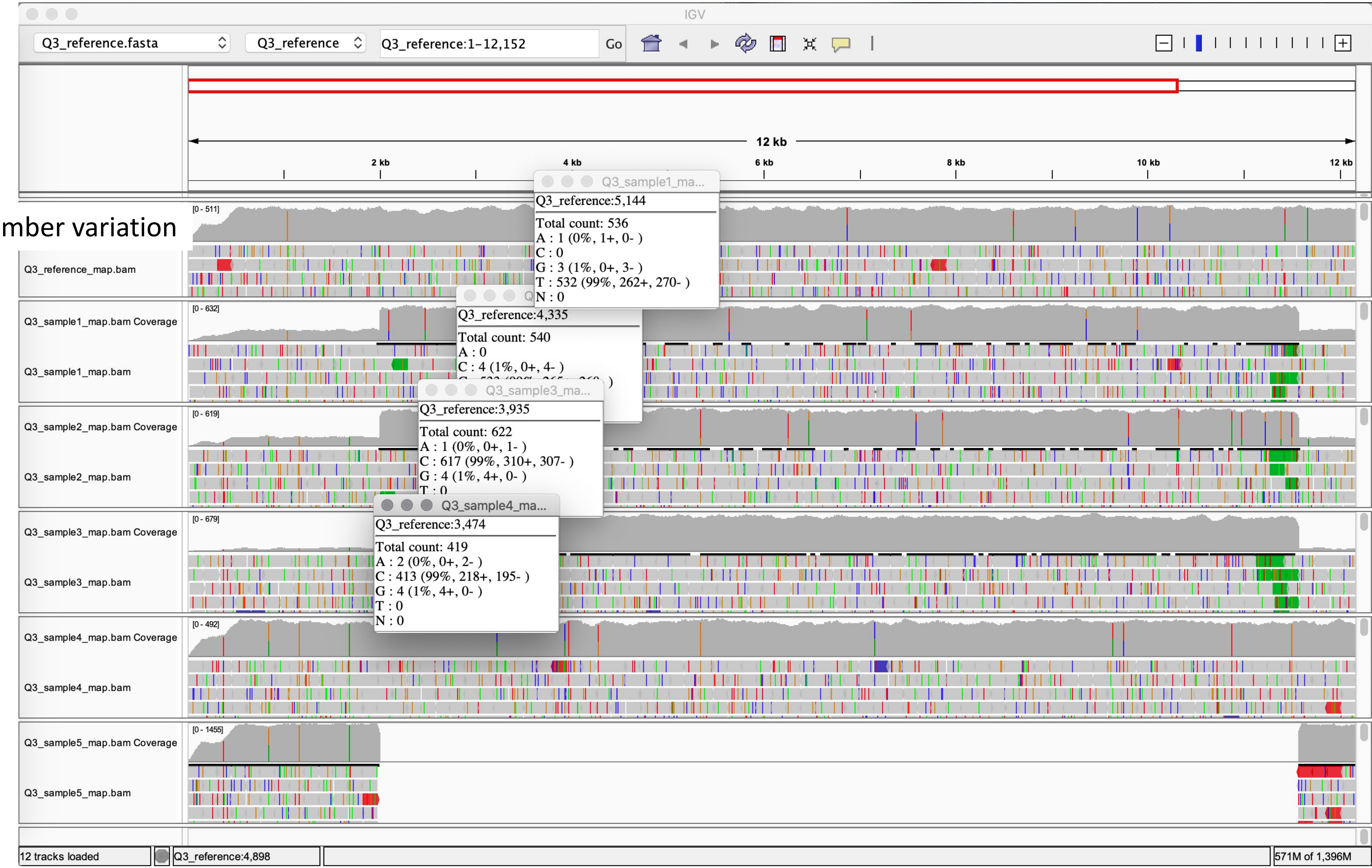
Q1: inversion



Q2: duplication



## Q3: copy number variation



Q3_reference:5,144

Total count: 536
A : 1 (0%, 1+, 0- )
C : 0
G : 3 (1%, 0+, 3- )
T : 532 (99%, 262+, 270- )
N : 0

Q3_reference:4,335

Total count: 540
A : 0
C : 4 (1%, 0+, 4- )
G : 532 (99%, 265+, 268- )

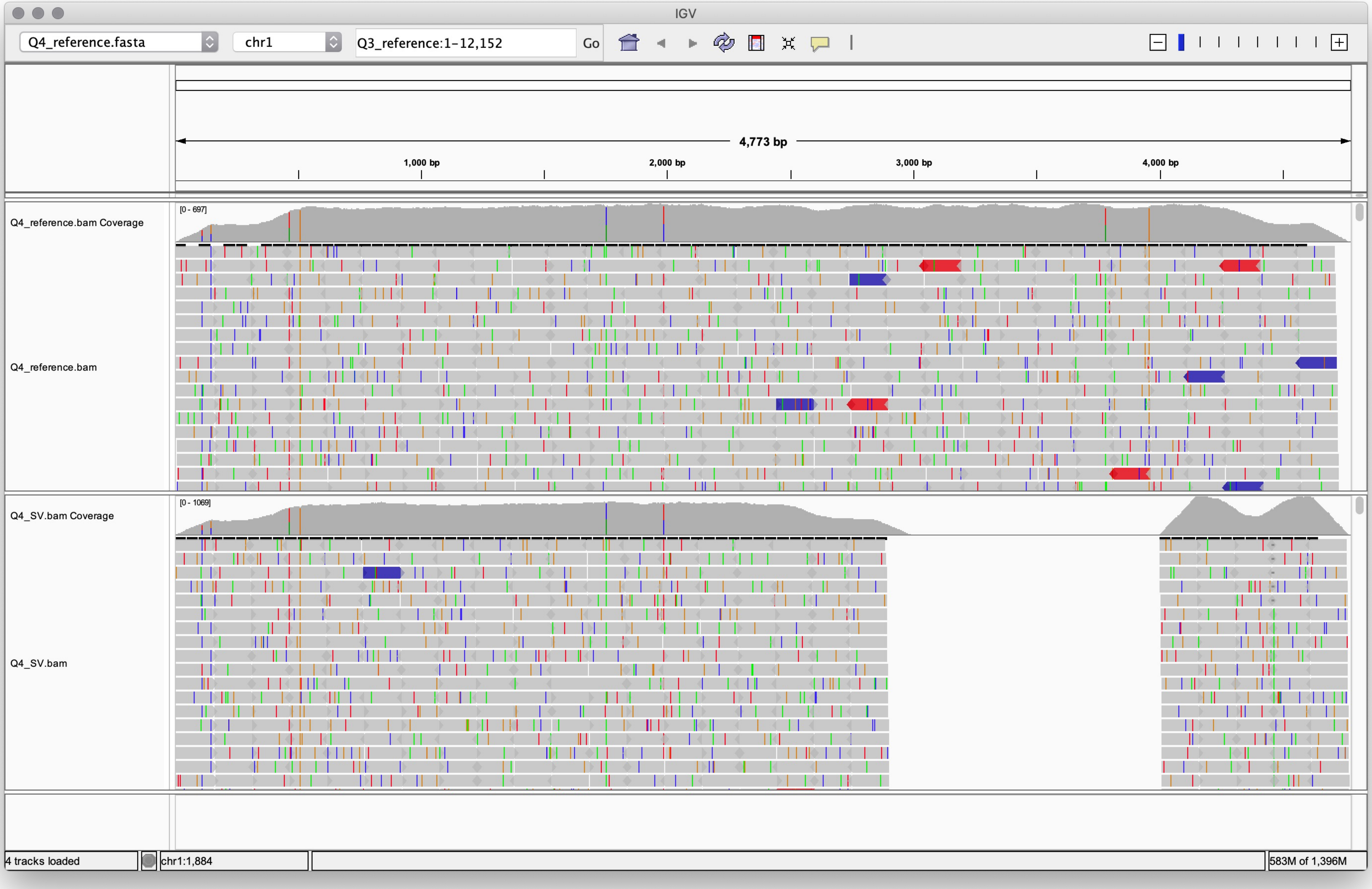
Q3_reference:3,935

Total count: 622
A : 1 (0%, 0+, 1- )
C : 617 (99%, 310+, 307- )
G : 4 (1%, 4+, 0- )
T : 0

Q3_reference:3,474

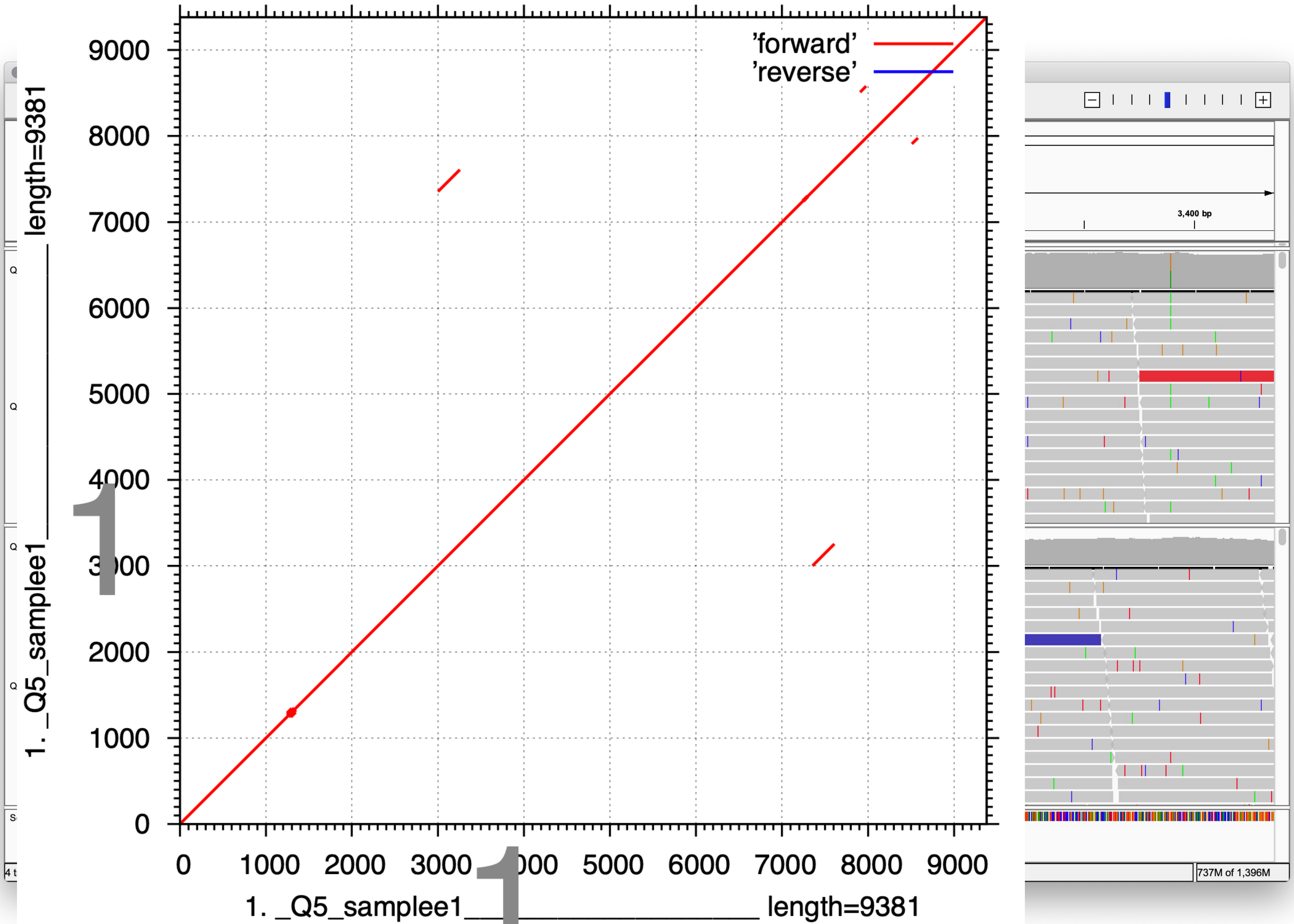
Total count: 419
A : 2 (0%, 0+, 2- )
C : 413 (99%, 218+, 195- )
G : 4 (1%, 4+, 0- )
T : 0
N : 0

Q4: deletion

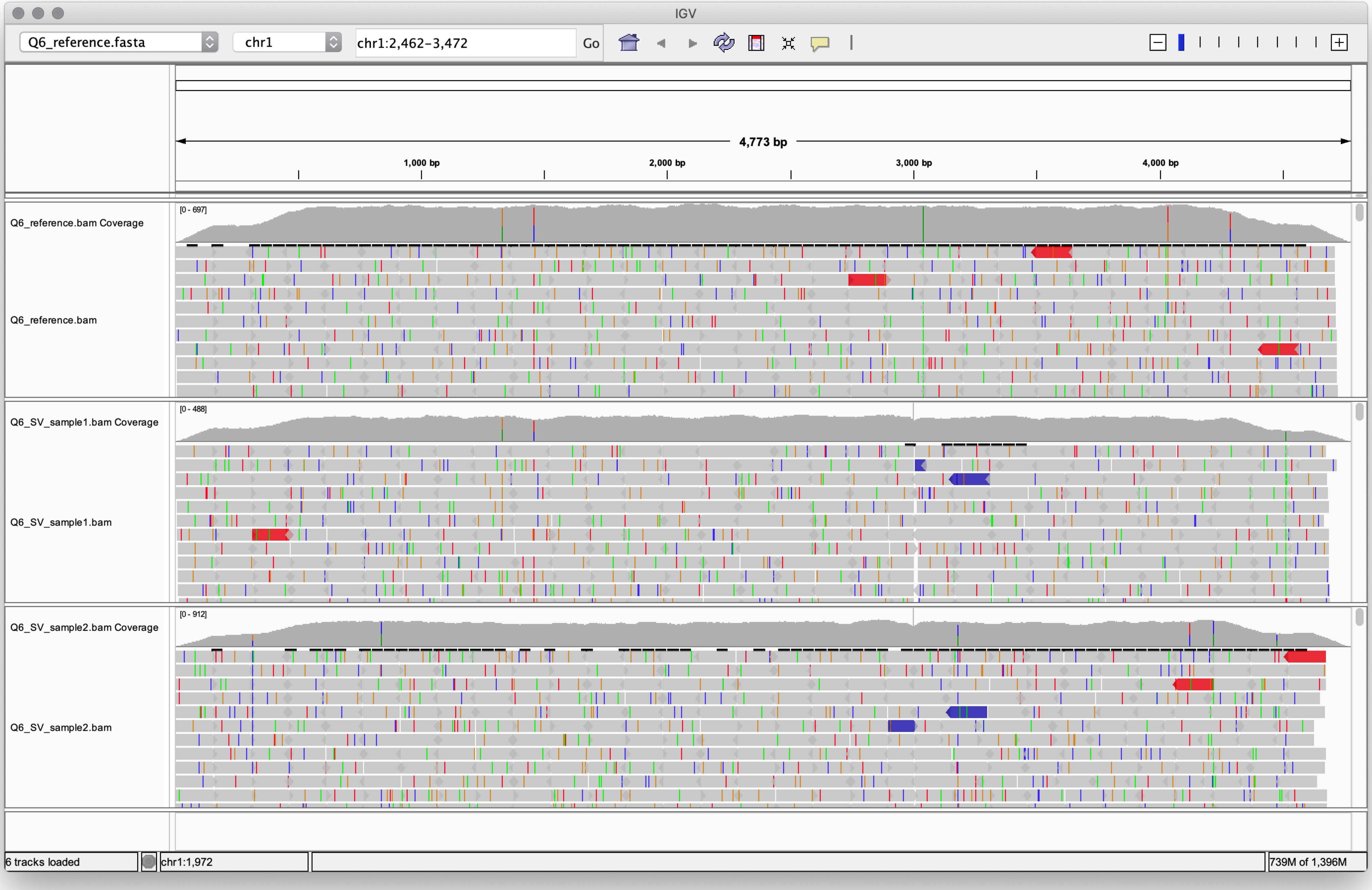


IGV
Q4_reference.fasta
chr1
Q3_reference:1-12,152
Go
4,773 bp
1,000 bp
2,000 bp
3,000 bp
4,000 bp
Q4_reference.bam Coverage
[0 - 697]
Q4_reference.bam
Q4_SV.bam Coverage
[0 - 1069]
Q4_SV.bam
4 tracks loaded
chr1:1,884
583M of 1,396M

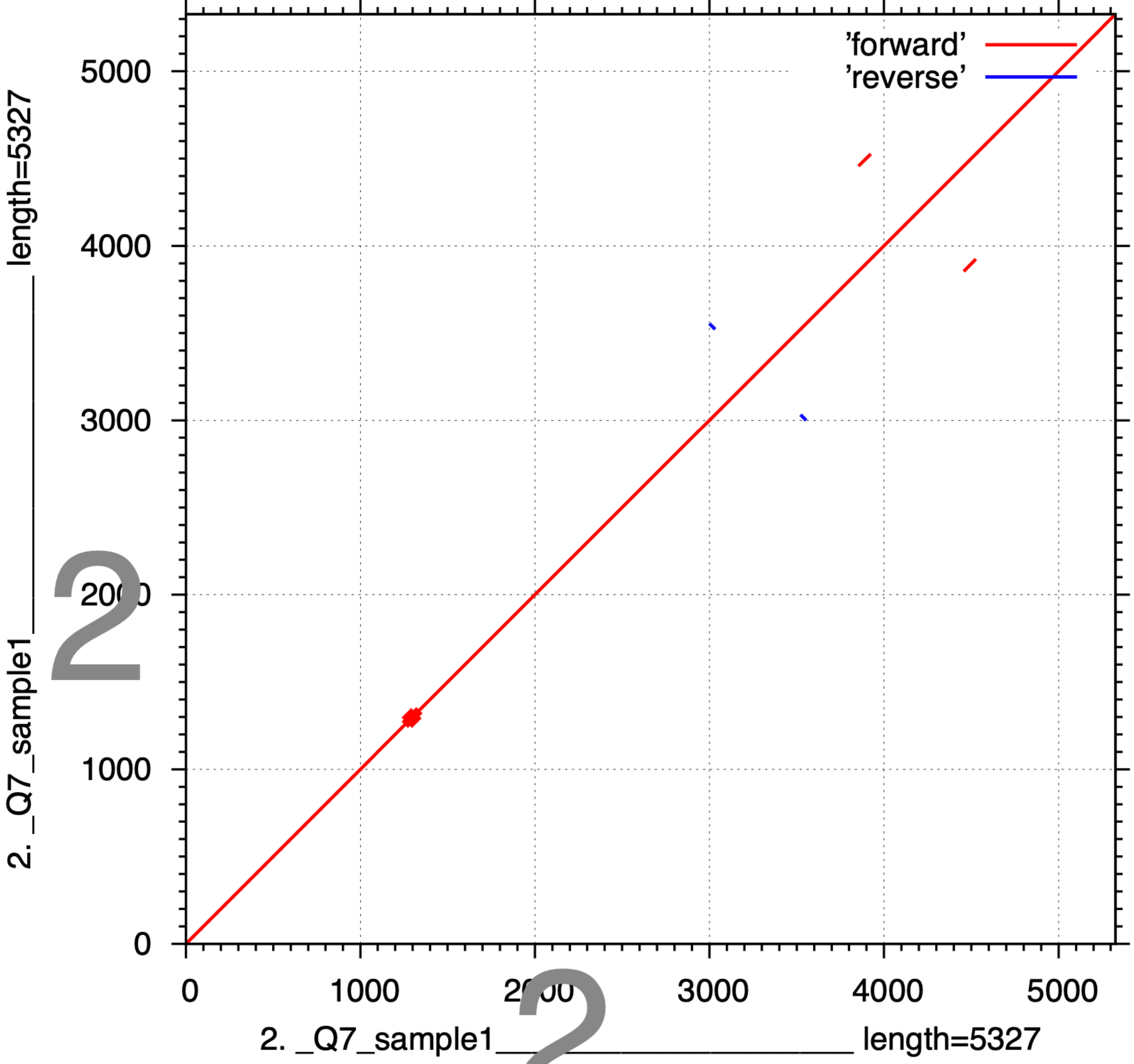
Q5: LTR insertion



Q6: LTR insertion
**soloLTR**



Q7: DNA TE


'forward'
'reverse'
5000
4000
3000
2000
1000
0
0
1000
2000
3000
4000
5000
2. _Q7_sample1____________________ length=5327
2. _Q7_sample1____________________ length=5327
2
2