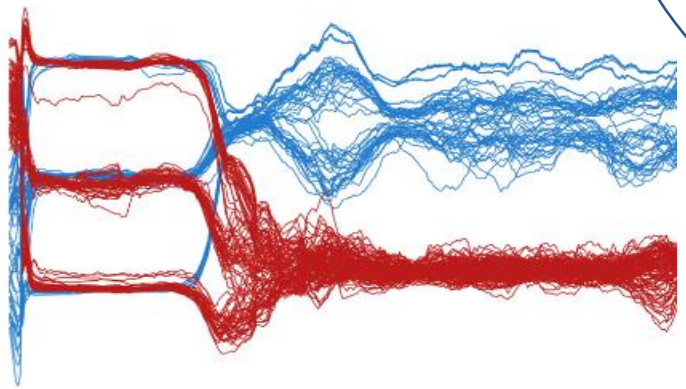
Hands-on lab on structural variants

Claire Mérot
claire.merot@univ-rennes.fr

EVOMICS 2026



Many thanks to Guy & the TAs, Raja, Milos, Flavia and Rhianah!!!

Tutorial

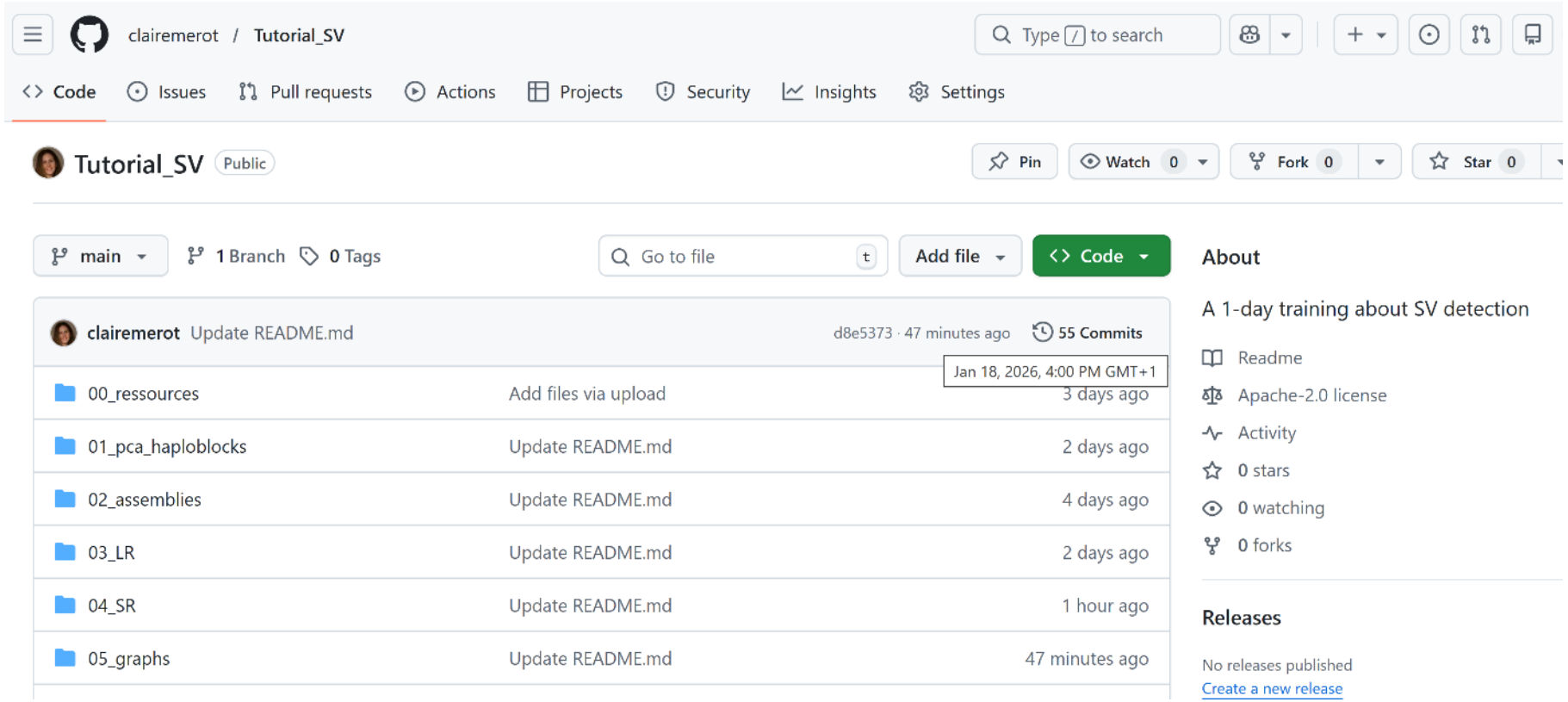
https://github.com/clairemerot/Tutorial_SV

IMPORTANT:

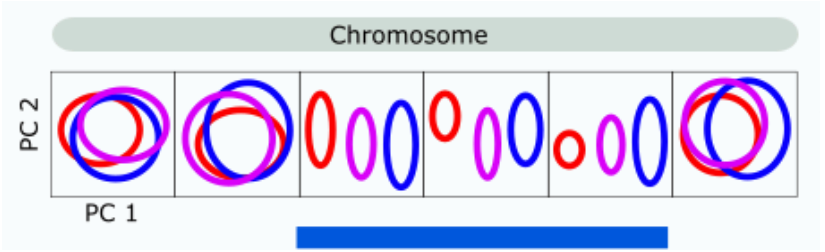
-> clone this repository in your working directory

-> open the folder Tutorial_SV

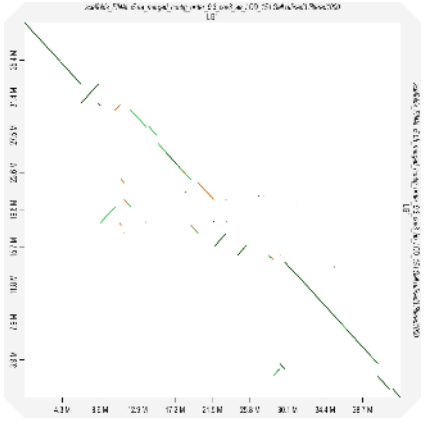
-> run everything from this location (and all paths should be ok!)



Outline

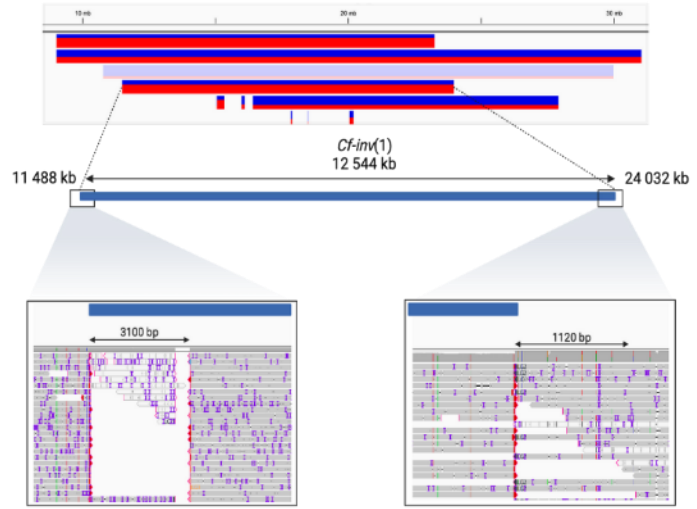


01 - Using SNPs & local PCA to detect haploblocks putatively representing large rearrangements



02 - Comparing assemblies for large rearrangements

03 - Assessing breakpoints and detecting other SVs with long-reads



04 - Using population dataset of short-reads to detect SVs

05 - Towards graph-based analysis

Data

Seaweed flies genomes
(Coelopa frigida)

Assemblies
-> 2 genomes
(ref.fasta)
(alt.fasta)

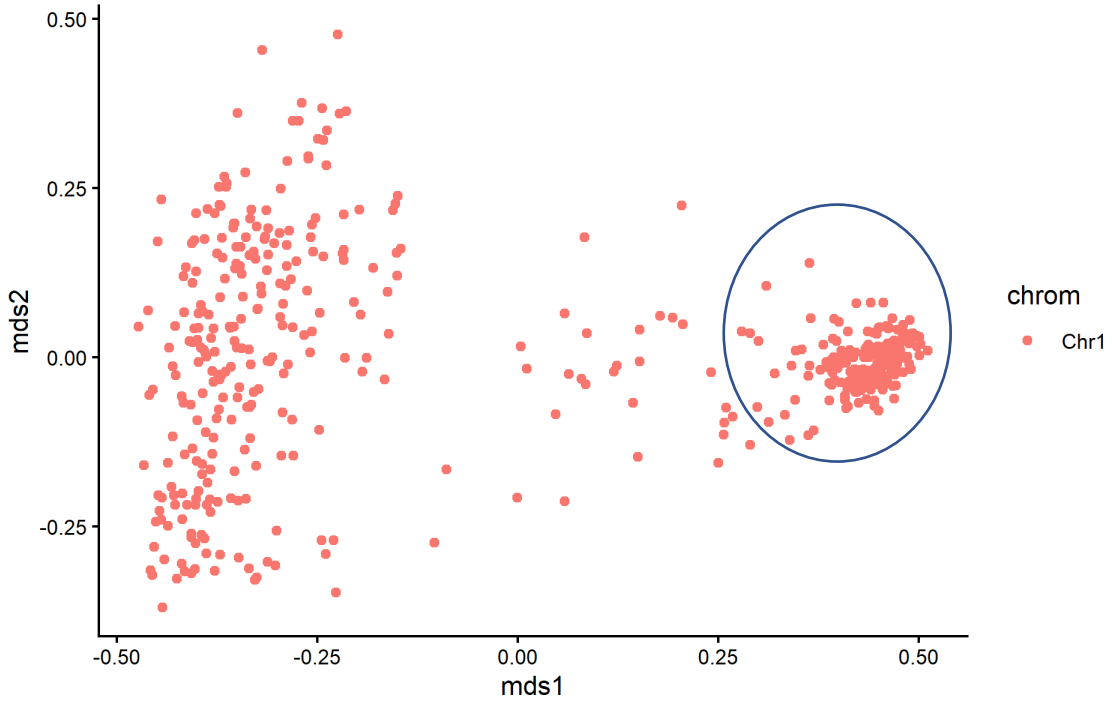
Subset of 10Mb

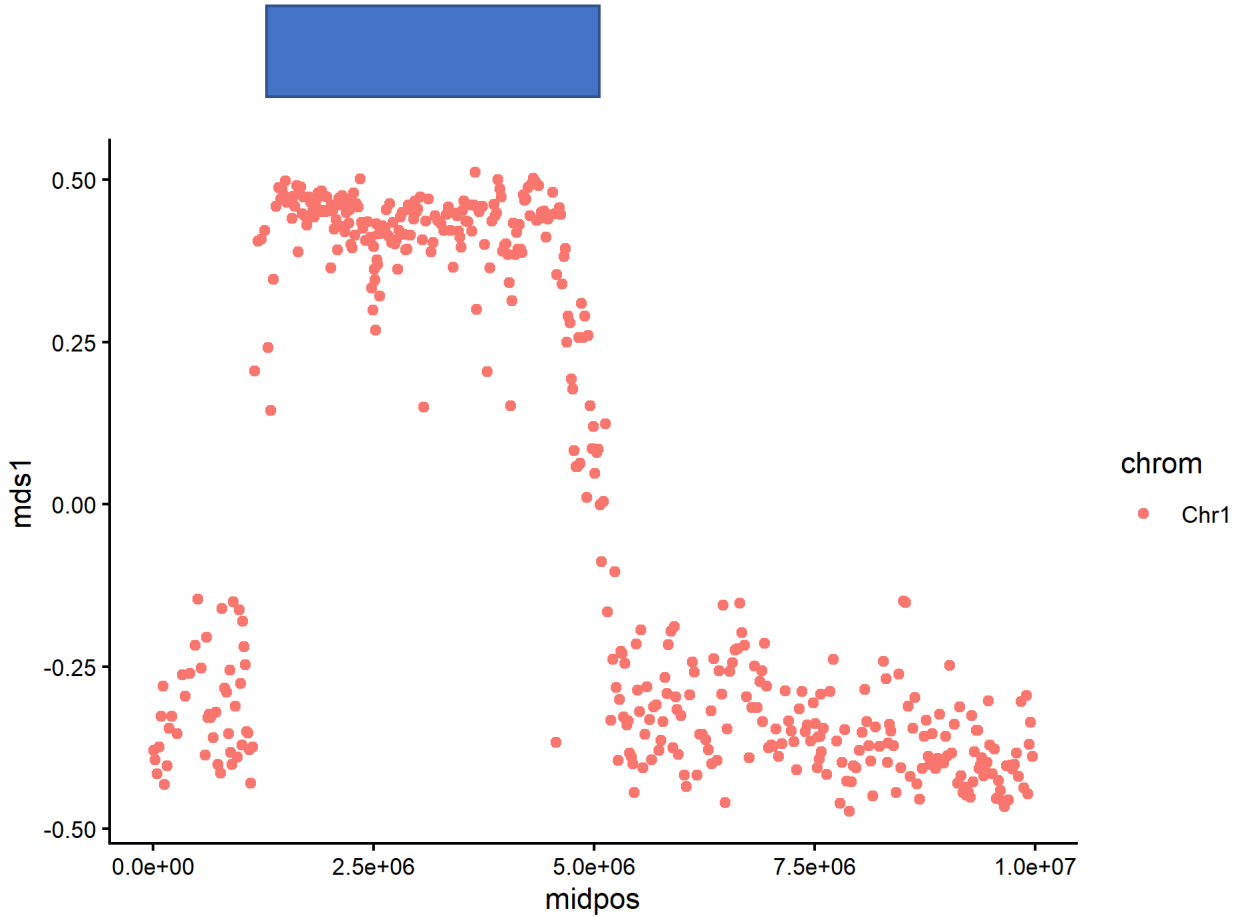
Long-reads
-> 1 sample (fri59)
ONT – Oxford Nanopore
~20X

Short-reads
-> 22 samples (illumina PE 150bp)
~10X
1 pair of fastq reads (A_R1, B_R2)
4 bamfiles (A, B, C, D)
1 vcf of SNPs from the 22 samples (A to V)

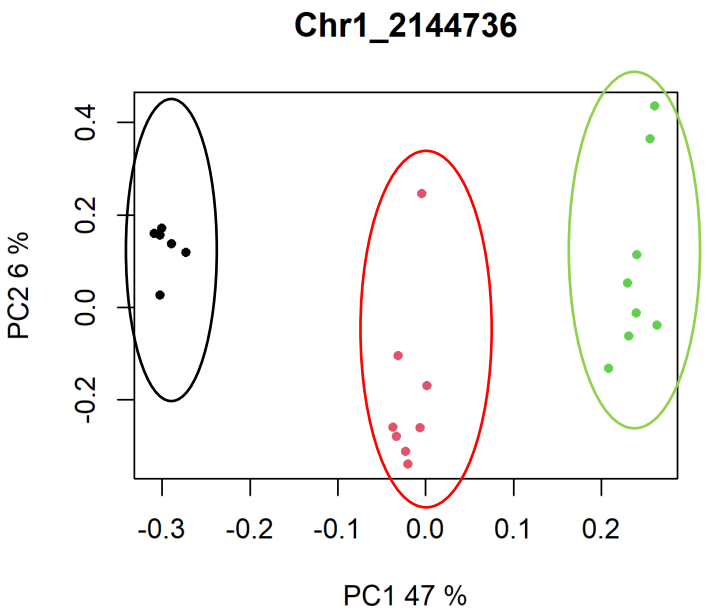
I - PCAs

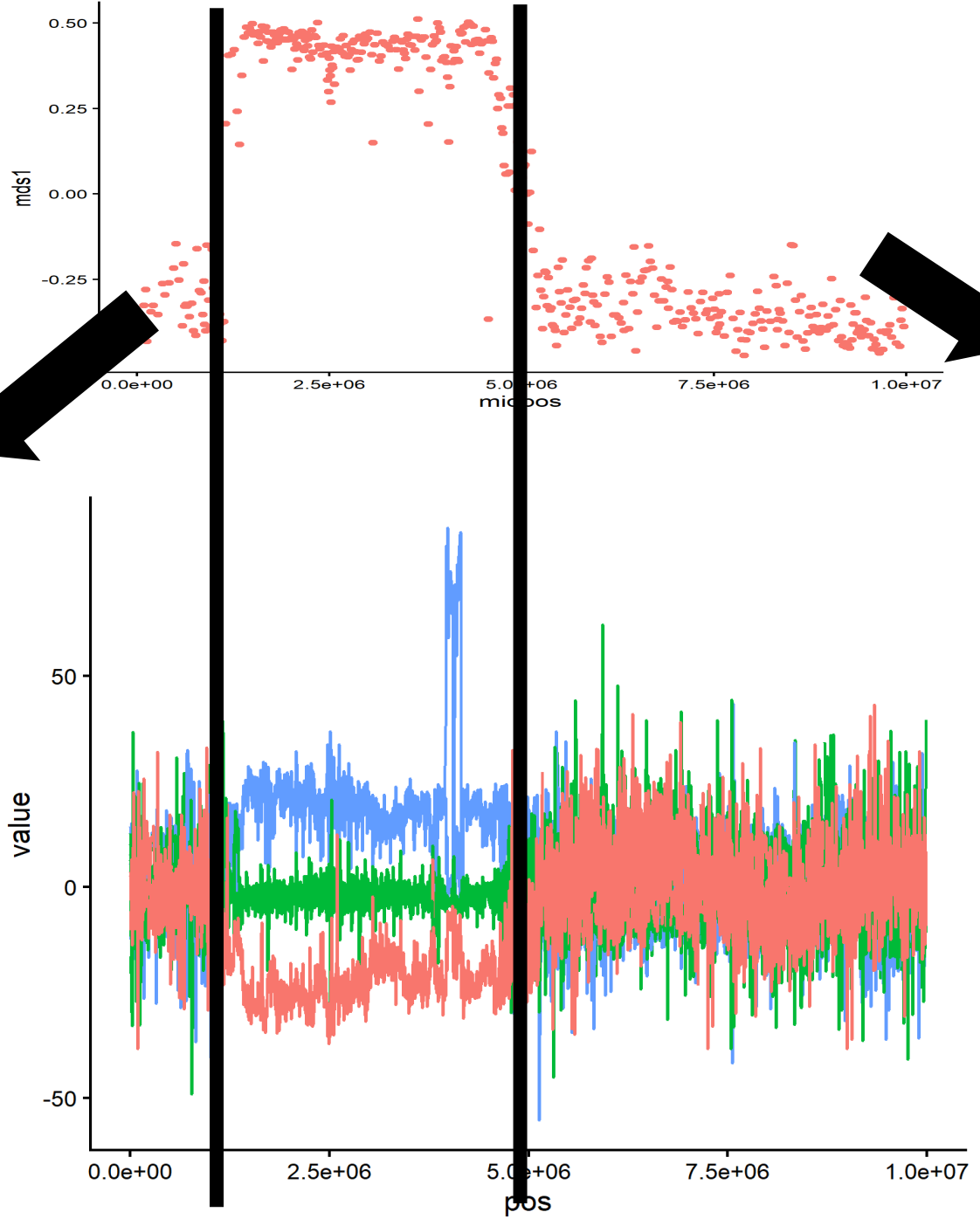
What do you think? Do we have a region where we could suspect non-recombining blocks revealing a putative inversion? How could we do to genotype our individuals for this putative variant?

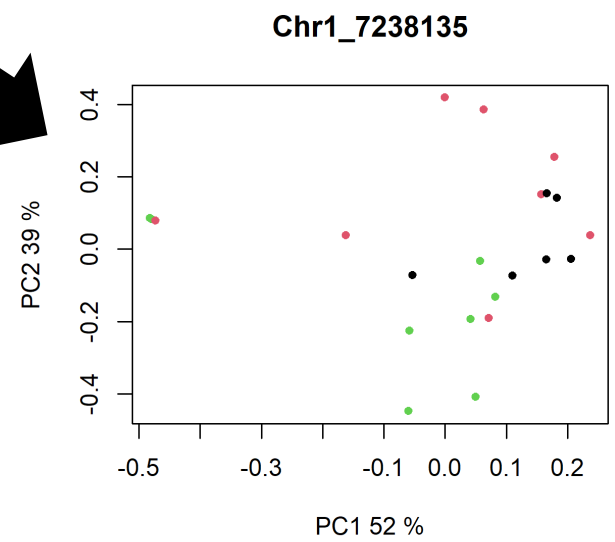




I - PCAs







Using either information from lostruct or winpca, may you identify approximative coordinates for the suspected rearrangement?

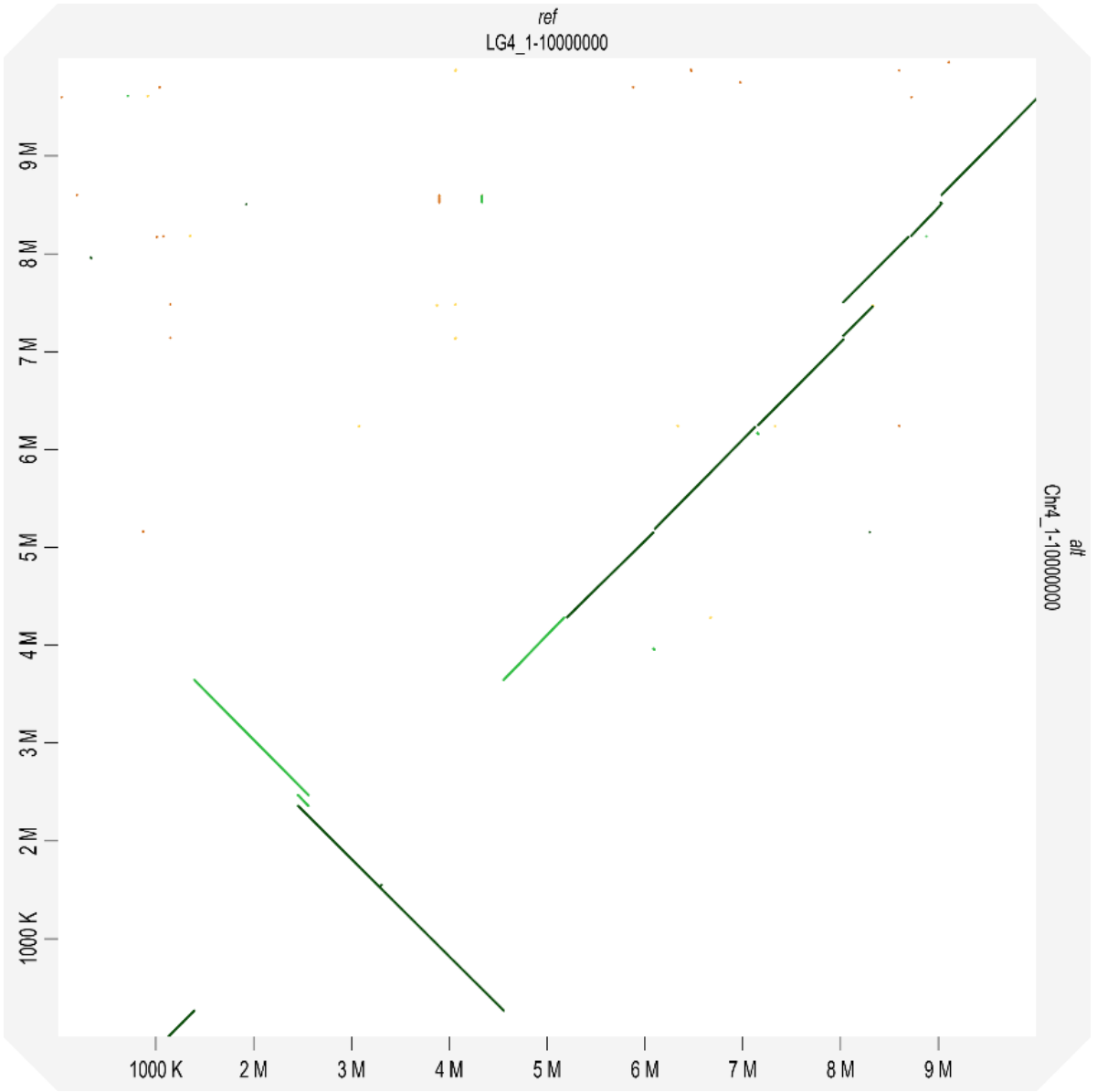
from 1Mb to 4,5Mb

2 – Assemblies comparison

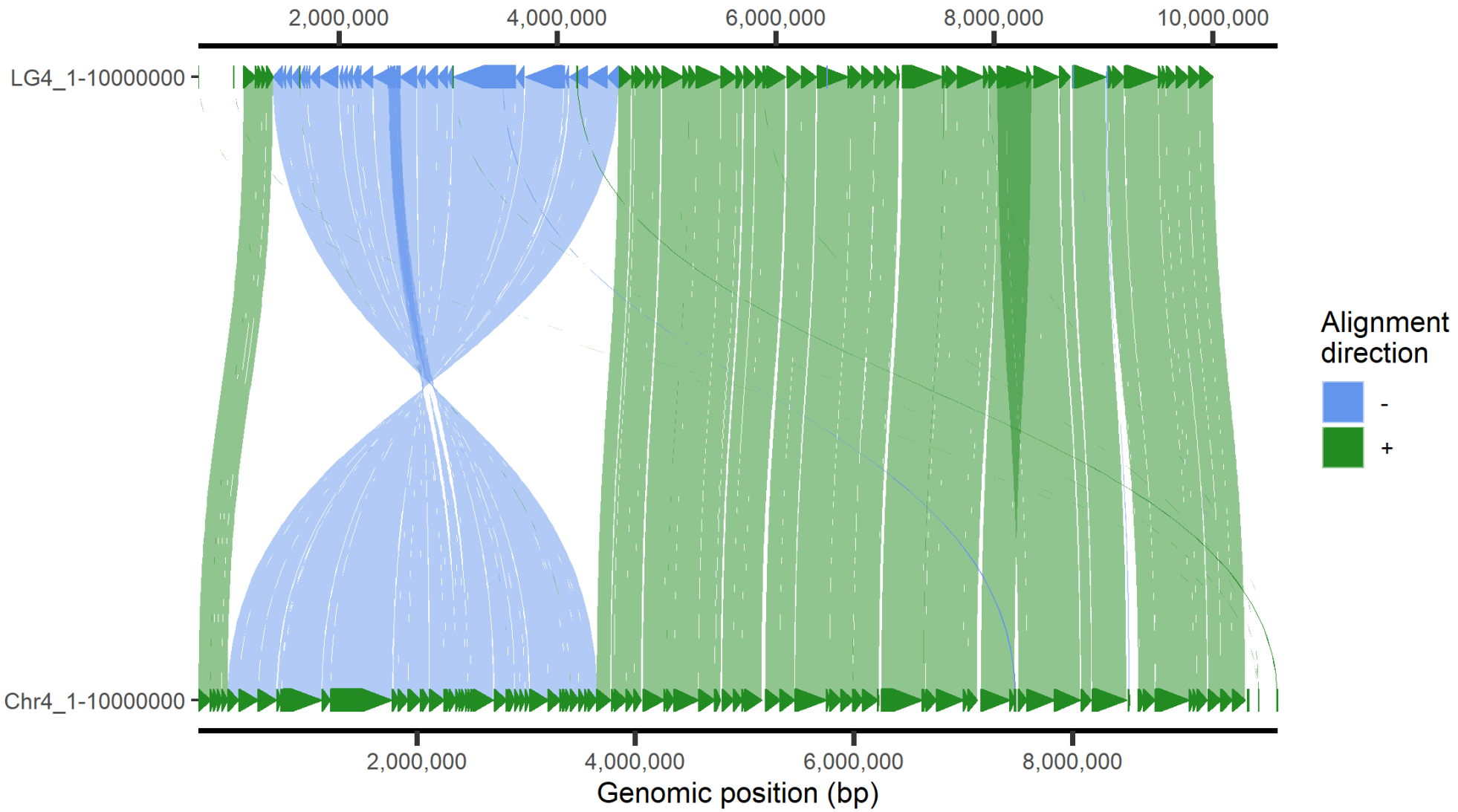
https://dgenies.toulouse.inra.fr/result/7EvbZ_20260119210131

What do you think? Do you visualise some rearrangements? Does it fit what you suspected from the PCA analysis?

Yes, a large inversion from 1Mb to 4,5Mb



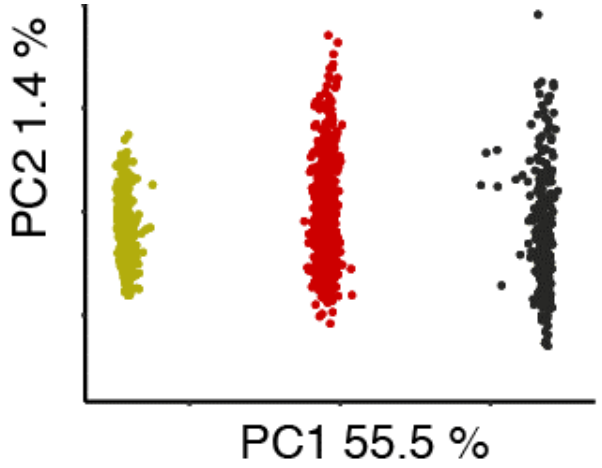
2 – Assemblies comparison

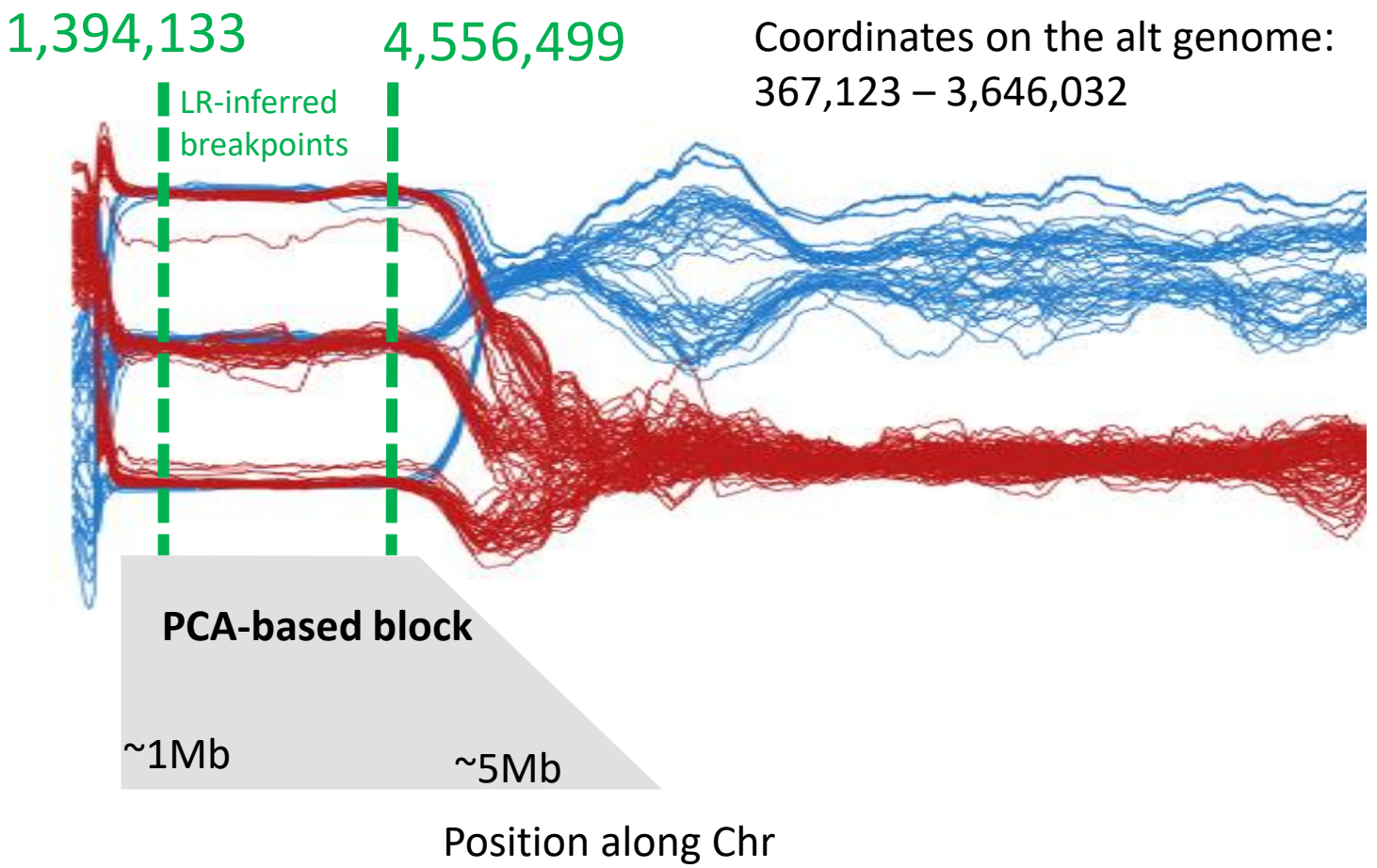


2 – Assemblies comparison

Using the paf file and/or the R package, can you extract the breakpoints of the suspected rearrangement?

How does it compare to the putative breakpoints identified by indirect PCA methods based on recombination suppression?





3 –Using long-reads

1845

-> How many SVs were detected?

-> What are the different pieces of information available for each variant?

-> How does it differ from a vcf of SNPs?

-> What kind of filters or classifications may you want to implement for subsequent analysis?

```
##FILTER=<ID=MOSAIC_VAF,Description="Mosaic variant allele fraction filter">
##FILTER=<ID=NOT_MOSAIC_VAF,Description="Variant allele fraction filter for non-mosaic">
##FILTER=<ID=ALN_NM,Description="Length adjusted mismatch filter">
##FILTER=<ID=STRAND_BND,Description="Strand support filter for BNDs">
##FILTER=<ID=STRAND,Description="Strand support filter for germline SVs">
##FILTER=<ID=STRAND_MOSAIC,Description="Strand support filter for mosaic SVs">
##FILTER=<ID=SVLEN_MIN,Description="SV length filter">
##FILTER=<ID=SVLEN_MIN_MOSAIC,Description="SV length filter for mosaic SVs">
##INFO=<ID=PRECISE,Number=0,Type=Flag,Description="Structural variation with precise breakpoints">
##INFO=<ID=IMPRECISE,Number=0,Type=Flag,Description="Structural variation with imprecise breakpoints">
##INFO=<ID=MOSAIC,Number=0,Type=Flag,Description="Structural variation classified as putative mosaic">
##INFO=<ID=SVLEN,Number=1,Type=Integer,Description="Length of structural variation">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variation">
##INFO=<ID=CHR2,Number=1,Type=String,Description="Mate chromsome for BND SVs">
##INFO=<ID=SUPPORT,Number=1,Type=Integer,Description="Number of reads supporting the structural variation">
##INFO=<ID=SUPPORT_INLINE,Number=1,Type=Integer,Description="Number of reads supporting an INS/DEL SV (non-split events only)">
##INFO=<ID=SUPPORT_LONG,Number=1,Type=Integer,Description="Number of soft-clipped reads putatively supporting the long insertion SV">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of structural variation">
##INFO=<ID=STDEV_POS,Number=1,Type=Float,Description="Standard deviation of structural variation start position">
##INFO=<ID=STDEV_LEN,Number=1,Type=Float,Description="Standard deviation of structural variation length">
##INFO=<ID=COVERAGE,Number=.,Type=Float,Description="Coverages near upstream, start, center, end, downstream of structural variation">
##INFO=<ID=STRAND,Number=1,Type=String,Description="Strands of supporting reads for structural variant">
##INFO=<ID=AC,Number=.,Type=Integer,Description="Allele count, summed up over all samples">
##INFO=<ID=SUPP_VEC,Number=1,Type=String,Description="List of read support for all samples">
##INFO=<ID=CONSENSUS_SUPPORT,Number=1,Type=Integer,Description="Number of reads that support the generated insertion (INS) consensus sequence">
##INFO=<ID=RNAMES,Number=.,Type=String,Description="Names of supporting reads (if enabled with --output-rnames)">
##INFO=<ID=VAF,Number=1,Type=Float,Description="Variant Allele Fraction">
##INFO=<ID=NM,Number=.,Type=Float,Description="Mean number of query alignment length adjusted mismatches of supporting reads">
##INFO=<ID=PHASE,Number=.,Type=String,Description="Phasing information derived from supporting reads, represented as list of: HAPLOTYPE,PHASESET,HAPLOTYPE_SU
PPORT,PHASESET_SUPPORT,HAPLOTYPE_FILTER,PHASESET_FILTER">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  SAMPLE
Chr1    3564    Sniffles2.DEL.11A7S0    ACTATAATTTTCTTTCAATAGTTATTCAATTTATTTCAATATTAAATTATATAAAAATACATTTTTTTTAATTTTTTGTTTATTTAAAAAATATGTAAAATAAG        A
        60      PASS    PRECISE;SVTYPE=DEL;SVLEN=-103;END=3666;SUPPORT=27;COVERAGE=44,43,43,43,42;STRAND=+-;STDEV_LEN=0.000;STDEV_POS=0.000;VAF=0.628   GT:GQ
:DR:DV  0/1:60:16:27
Chr1    5716    Sniffles2.DEL.11A8S0    TACACATTTTCTAAAAGCTGATTAAATTTCCTTTAAATTGATATATATATGTCATATTTATATATAGAATGAA       T       60      PASS    PRECISE;SVTYP
E=DEL;SVLEN=-72;END=5787;SUPPORT=16;COVERAGE=40,37,37,40,38;STRAND=+-;STDEV_LEN=0.000;STDEV_POS=0.000;VAF=0.421 GT:GQ:DR:DV     0/1:60:22:16
Chr1    13302   Sniffles2.INS.5S0       C       CTTTAGATCAATTATTTGAGAATTTGATTATTTGAGAAATTATTATTTGAACTAAAAA      60      PASS    PRECISE;SVTYPE=INS;SVLEN=57;E
ND=13302;SUPPORT=20;COVERAGE=35,36,36,37,37;STRAND=+-;STDEV_LEN=0.000;STDEV_POS=0.000;SUPPORT_LONG=0;VAF=0.556  GT:GQ:DR:DV     0/1:60:16:20
Chr1    36928   Sniffles2.INS.FS0       A       ATTTAAGGTTGTTGAAAATTTGCTTGAATTGTGAAATTTTTGTAGAGGTCGTCACGAATTAACGTCTTTTTTCGCGAAATATCACGTATTTAGATTTCCTAAAAAATTG
CCTGTATTATGAAATTTTCGAAAATTTAAATGTCGTCACGAATTAACGTTCTTTCGCGAAATAACGCGTATTTACGGAGATATGAACGATATAAGATTTCGAAAAATTTCTGGAGCTGTAAATTTAAAAAAAAATTTAAGGGGCGTCACGAATTAA
CGTATTT 60      PASS    IMPRECISE;SVTYPE=INS;SVLEN=279;END=36928;SUPPORT=37;COVERAGE=39,39,39,40,38;STRAND=+-;STDEV_LEN=45.889;STDEV_POS=0.000;SUPPORT_LONG=0
;VAF=0.949      GT:GQ:DR:DV     1/1:60:2:37
Chr1    53053   Sniffles2.INS.13S0      T       TTTTCGATATTTTGTCAATTATTTTTTTTGTTTTCATTTTTTTTATAAAAATTTTC        60      PASS    PRECISE;SVTYPE=INS;SVLEN=56;E
ND=53053;SUPPORT=22;COVERAGE=48,46,46,46,45;STRAND=+-;STDEV_LEN=0.900;STDEV_POS=16.908;SUPPORT_LONG=0;VAF=0.478 GT:GQ:DR:DV     0/1:60:24:22
Chr1    55215   Sniffles2.INS.16S0      T       TAAAAATGTTTTATGATTCCCTAAAAAACCTCTTGTCGGTGAGATTCCCTAAGACATCGTTTGAACCCCG  60      PASS    PRECISE;SVTYPE=INS;SV
:
```

3 –Using long-reads

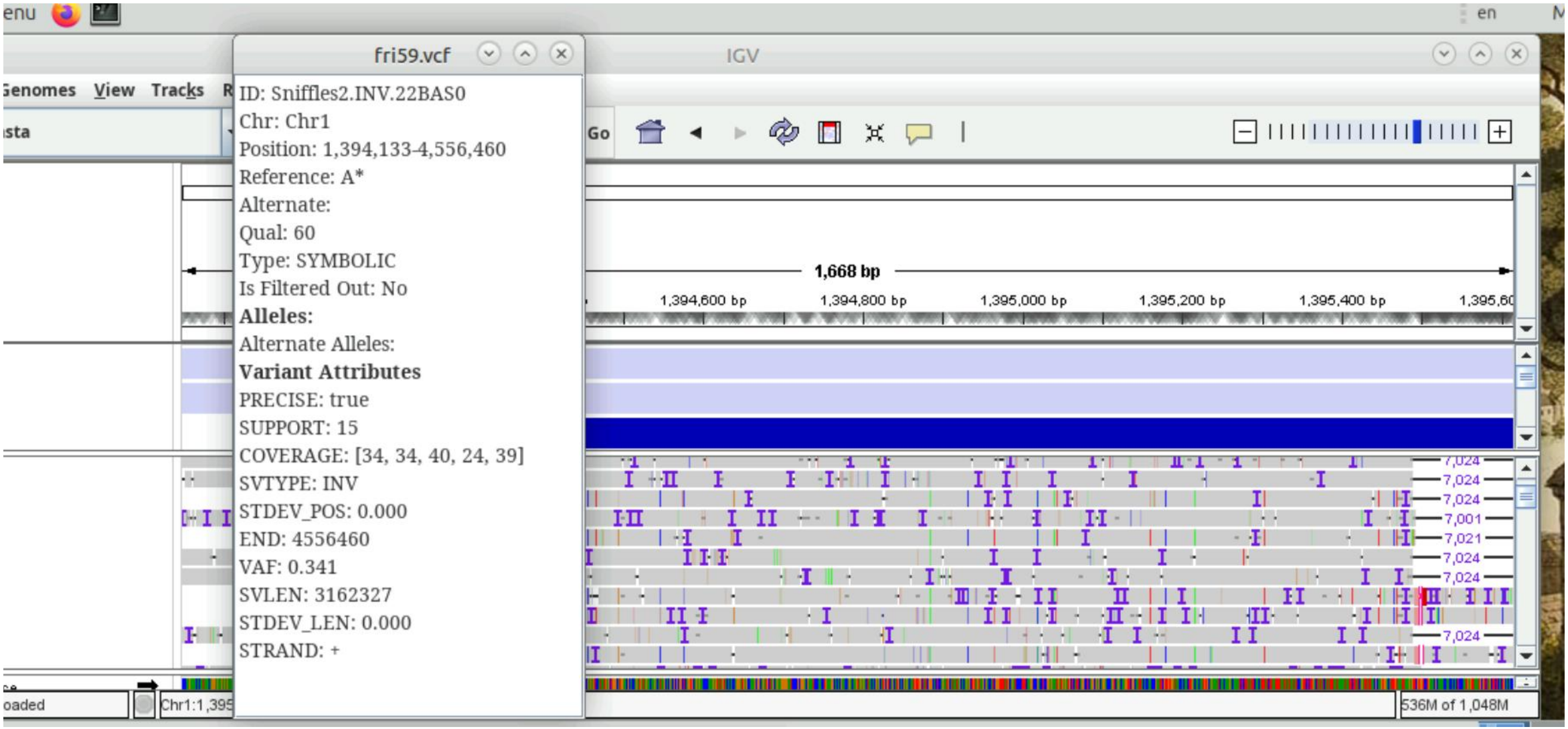
1845

-> How many SVs were detected?
-> What are the different pieces of information available for each variant?
-> How does it differ from a vcf of SNPs?
-> What kind of filters or classifications may you want to implement for subsequent analysis?

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  SAMPLE
Chr1    3564    Sniffles2.DEL.11A7S0    ACTATAATTTTCTTTCAATAGTTATTCAATTTATTTCAATATTAAATTATATAAAAATACATTTTTTTTAATTTTTTGTTTATTTAAAAAATATGTAAAATAAG        A
        60      PASS    PRECISE;SVTYPE=DEL;SVLEN=-103;END=3666;SUPPORT=27;COVERAGE=44,43,43,43,42;STRAND=+-;STDEV_LEN=0.000;STDEV_POS=0.000;VAF=0.628   GT:GQ
:DR:DV  0/1:60:16:27
Chr1    5716    Sniffles2.DEL.11A8S0    TACACATTTTCTAAAAGCTGATTAAATTTCCTTTAAATTGATATATATATGTCATATTTATATATAGAATGAA       T       60      PASS    PRECISE;SVTYP
E=DEL;SVLEN=-72;END=5787;SUPPORT=16;COVERAGE=40,37,37,40,38;STRAND=+-;STDEV_LEN=0.000;STDEV_POS=0.000;VAF=0.421 GT:GQ:DR:DV     0/1:60:22:16
Chr1    13302   Sniffles2.INS.5S0       C       CTTTAGATCAATTATTTGAGAATTTGATTATTTGAGAAATTATTATTTGAACTAAAAA      60      PASS    PRECISE;SVTYPE=INS;SVLEN=57;E
ND=13302;SUPPORT=20;COVERAGE=35,36,36,37,37;STRAND=+-;STDEV_LEN=0.000;STDEV_POS=0.000;SUPPORT_LONG=0;VAF=0.556  GT:GQ:DR:DV     0/1:60:16:20
Chr1    36928   Sniffles2.INS.FS0       A       ATTTAAGGTTGTTGAAAATTTGCTTGAATTGTGAAATTTTTGTAGAGGTCGTCACGAATTAACGTCTTTTTTCGCGAAATATCACGTATTTAGATTTCCTAAAAAATTG
CCTGTATTATGAAATTTTCGAAAATTTAAATGTCGTCACGAATTAACGTTCTTTTCGCGAAATAACGCGTATTTACGGAGATATGAACGATATAAGATTTCGAAAAATTTCTGGAGCTGTAAATTTAAAAAAAAATTTAAGGGGCGTCACGAATTAA
CGTATTT 60      PASS    IMPRECISE;SVTYPE=INS;SVLEN=279;END=36928;SUPPORT=37;COVERAGE=39,39,39,40,38;STRAND=+-;STDEV_LEN=45.889;STDEV_POS=0.000;SUPPORT_LONG=0
;VAF=0.949      GT:GQ:DR:DV     1/1:60:2:37
Chr1    53053   Sniffles2.INS.13S0      T       TTTTCGATATTTTGTCAATTATTTTTTTTGTTTTCATTTTTTTTATAAAAATTTTC        60      PASS    PRECISE;SVTYPE=INS;SVLEN=56;E
ND=53053;SUPPORT=22;COVERAGE=48,46,46,46,45;STRAND=+-;STDEV_LEN=0.900;STDEV_POS=16.908;SUPPORT_LONG=0;VAF=0.478 GT:GQ:DR:DV     0/1:60:24:22
Chr1    55215   Sniffles2.INS.16S0      T       TAAAAATGTTTTATGATTCCCTAAAAAACCTCTTGTCGGTGAGATTCCCTAAGACATCGTTTGAACCCCG  60      PASS    PRECISE;SVTYPE=INS;SV
:
```

-> select SV based on length range?
-> filter by type
-> remove SVs that do not « PASS »
-> filter based on read support
-> check IMPRECISE breakpoints….

3 –Using long-reads

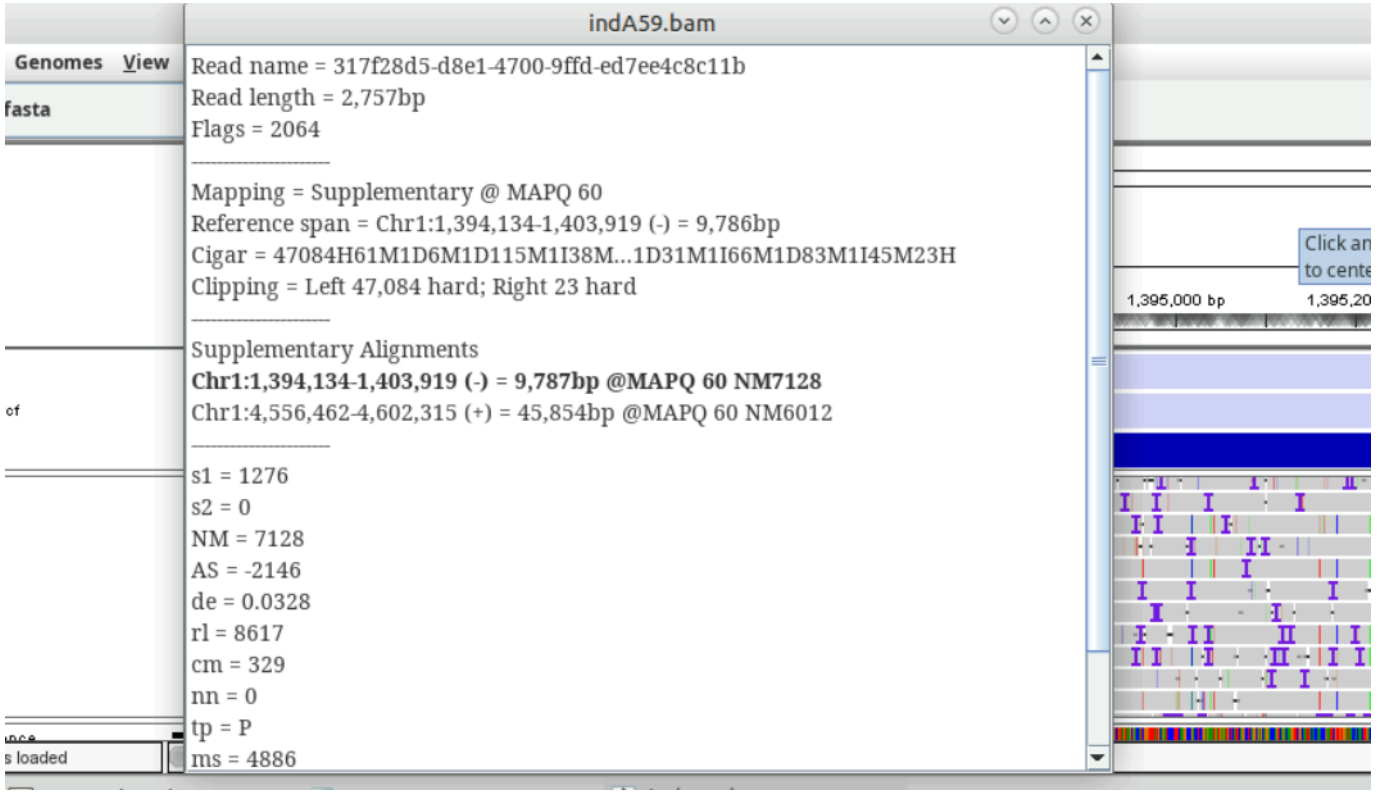


3 –Using long-reads

You can also zoom on the breakpoints previously identified with the assembly comparison for the very large rearrangement.

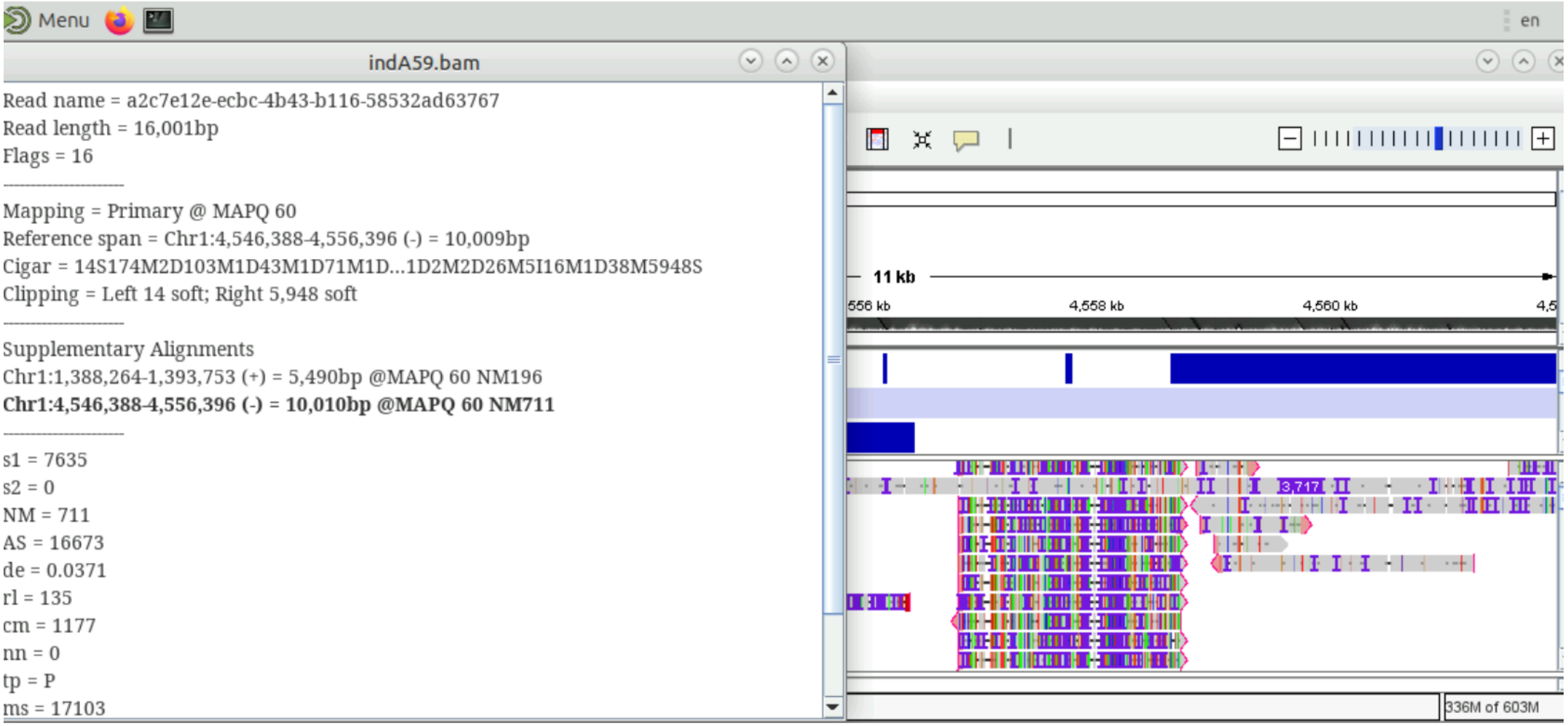
-> What do they look like here?

-> Are they supported by long-reads?



3 –Using long-reads

-> Which genotype is this individual "fri59“ for our big inversion?



3 –Using long-reads

-> Which genotype is this individual "fri59“ for our big inversion?

heterozygote

grep 'INV.22BAS0' 03_LR/fri59.vcf

Chr1 1394133 Sniffles2.INV.22BAS0 A <INV> 60 PASS PRECISE;SVTYPE=INV;SVLEN=3162327;END=4556460;SUPPORT=15;COVERAGE=34,34,40,24,39;STRAND=+;STDEV_LEN=0.000;STDEV_POS=0.000;VAF=0.341 GT:GQ:DR:DV 0/1:60:29:15

Also visible with the reads in IGV (some go over the break-points and some are split-reads

4 – SV call with short-reads

-> What do you think about the information about SVs? What differs and does not differ with the vcf produced by Sniffles?

```
##bcftools_viewCommand=view 04_SR/allSV_SR.bcf; Date=Fri Jan 16 09:35:56 2026
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  B
Chr1    519     DEL00000000     GCTTAACACAGTTCTCTTTCTTAATGTTTCATTT      G       480     PASS    PRECISE;SVTYPE=DEL;SVMETHOD=EMBL.DELLYv1.7.2;END=552;PE=0;MAP
Q=0;CT=3to5;CIPOS=-4,4;CIEND=-4,4;SRMAPQ=60;INSLEN=0;HOMLEN=4;SR=8;SRQ=0.965347;CONSENSUS=TATTTTTTTCATTTCCCGACCTTCCGTCATTATTATATTACAATCTCTGCTTGTTCCCGAGCAGCAT
TTTCTGGACGATAGTGAAAACTTCCTAAGGGAATCAATCCTTTAATCTGCTTTTTCAAATATTTTAATTTTGAAAATTGTCTTCTTTTTTATTATTTTCTCACAGTTTTTATCATTCTTTATTCTTTTCTTGCAA;CE=1.78796;CONSBP=116
        GT:GL:GQ:FT:RCL:RC:RCR:RDCN:DR:DV:RR:RV 0/1:-136.067,0,-26.8585:10000:PASS:626:862:1443:1:0:0:14:27
Chr1    958     INS00000001     A       ATCTCGCTCTATCTCGCTCTATCTCGCTCTCTCTC      240     PASS    PRECISE;SVTYPE=INS;SVMETHOD=EMBL.DELLYv1.7.2;END=958;SVLEN=34
;PE=0;MAPQ=0;CT=NtoN;CIPOS=-9,9;CIEND=-9,9;SRMAPQ=60;INSLEN=34;HOMLEN=12;SR=4;SRQ=0.978723;CONSENSUS=AACTGGGAAATTGTGAGATATTTAAATGCGACTATTAATACGTTTTAACGCTCTAT
CTCGCTCTATCTCGCTCTATCTCGCTCTCTCTCTCTCTCTCACTCTTTCACTCATTATTTTTTTATATCTTTTGTTTATTTCATTGGGGAGAGTTTTTGAATTTCATTCGTTCTTCTG;CE=1.83651;CONSBP=55     GT:GL:GQ:FT:R
CL:RC:RCR:RDCN:DR:DV:RR:RV      1/1:-248.998,-14.748,0:147:PASS:23298:47469:24171:2:0:0:0:49
Chr1    5356    DUP00000002     G       <DUP>   95      LowQual IMPRECISE;SVTYPE=DUP;SVMETHOD=EMBL.DELLYv1.7.2;END=180472;PE=2;MAPQ=59;CT=5to3;CIPOS=-529,529
;CIEND=-529,529 GT:GL:GQ:FT:RCL:RC:RCR:RDCN:DR:DV:RR:RV 0/0:0,-10.0595,-348.193:101:PASS:2476:81500:46802:3:63:2:0:0
Chr1    5356    DUP00000003     G       <DUP>   520     PASS    IMPRECISE;SVTYPE=DUP;SVMETHOD=EMBL.DELLYv1.7.2;END=826732;PE=18;MAPQ=28;CT=5to3;CIPOS=-428,42
8;CIEND=-428,428        GT:GL:GQ:FT:RCL:RC:RCR:RDCN:DR:DV:RR:RV 0/1:-40.5413,0,-188.637:10000:PASS:2476:691588:632470:2:35:22:0:0
Chr1    5382    INV00000004     G       <INV>   267     PASS    IMPRECISE;SVTYPE=INV;SVMETHOD=EMBL.DELLYv1.7.2;END=1021738;PE=10;MAPQ=29;CT=5to5;CIPOS=-375,3
75;CIEND=-375,375       GT:GL:GQ:FT:RCL:RC:RCR:RDCN:DR:DV:RR:RV 0/1:-31.3596,0,-146.018:10000:PASS:2479:1063929:385871:5:29:24:0:0
Chr1    5383    INV00000005     A       <INV>   1251    PASS    IMPRECISE;SVTYPE=INV;SVMETHOD=EMBL.DELLYv1.7.2;END=616271;PE=38;MAPQ=30;CT=5to5;CIPOS=-311,31
1;CIEND=-311,311        GT:GL:GQ:FT:RCL:RC:RCR:RDCN:DR:DV:RR:RV 0/1:-90.5894,0,-495.263:10000:PASS:2479:415898:463178:2:129:51:0:0
Chr1    5383    DUP00000006     A       <DUP>   1657    PASS    IMPRECISE;SVTYPE=DUP;SVMETHOD=EMBL.DELLYv1.7.2;END=616734;PE=55;MAPQ=28;CT=5to3;CIPOS=-168,16
8;CIEND=-168,168        GT:GL:GQ:FT:RCL:RC:RCR:RDCN:DR:DV:RR:RV 0/1:-173.782,0,-156.21:10000:PASS:2479:420550:459580:2:32:83:0:0
Chr1    5383    DUP00000007     A       <DUP>   373     PASS    IMPRECISE;SVTYPE=DUP;SVMETHOD=EMBL.DELLYv1.7.2;END=617527;PE=15;MAPQ=24;CT=5to3;CIPOS=-362,36
2;CIEND=-362,362        GT:GL:GQ:FT:RCL:RC:RCR:RDCN:DR:DV:RR:RV 0/1:-29.2598,0,-120.299:10000:PASS:2479:422663:458989:2:22:17:0:0
Chr1    5383    INV00000008     A       <INV>   181     PASS    IMPRECISE;SVTYPE=INV;SVMETHOD=EMBL.DELLYv1.7.2;END=1068736;PE=7;MAPQ=26;CT=5to5;CIPOS=-642,64
2;CIEND=-642,642        GT:GL:GQ:FT:RCL:RC:RCR:RDCN:DR:DV:RR:RV 0/0:0,-6.88546,-437.908:69:PASS:2479:1152758:338466:7:76:7:0:0
Chr1    5383    INV00000009     A       <INV>   120     LowQual IMPRECISE;SVTYPE=INV;SVMETHOD=EMBL.DELLYv1.7.2;END=3052463;PE=2;MAPQ=60;CT=5to5;CIPOS=-601,60
1;CIEND=-601,601        GT:GL:GQ:FT:RCL:RC:RCR:RDCN:DR:DV:RR:RV 0/1:-6.57117,0,-135.602:66:PASS:2479:2103493:567688:7:24:4:0:0
Chr1    5387    INV00000010     A       <INV>   157     PASS    IMPRECISE;SVTYPE=INV;SVMETHOD=EMBL.DELLYv1.7.2;END=4283625;PE=3;MAPQ=53;CT=5to5;CIPOS=-616,61
6;CIEND=-616,616        GT:GL:GQ:FT:RCL:RC:RCR:RDCN:DR:DV:RR:RV 0/1:-9.37838,0,-101.678:94:PASS:2479:2552441:827275:6:18:3:0:0
Chr1    5414    INV00000011     T       <INV>   223     PASS    IMPRECISE;SVTYPE=INV;SVMETHOD=EMBL.DELLYv1.7.2;END=871200;PE=8;MAPQ=29;CT=5to5;CIPOS=-167,167
;CIEND=-167,167 GT:GL:GQ:FT:RCL:RC:RCR:RDCN:DR:DV:RR:RV 0/1:-21.6656,0,-151.094:10000:PASS:2516:769323:576244:3:29:18:0:0
Chr1    5436    DUP00000012     A       <DUP>   70      LowQual IMPRECISE;SVTYPE=DUP;SVMETHOD=EMBL.DELLYv1.7.2;END=204296;PE=2;MAPQ=35;CT=5to3;CIPOS=-616,616
;CIEND=-616,616 GT:GL:GQ:FT:RCL:RC:RCR:RDCN:DR:DV:RR:RV 0/0:0,-2.61808,-162.885:26:PASS:2603:102684:32982:6:30:2:0:0
Chr1    5494    INV00000013     T       <INV>   51      LowQual IMPRECISE;SVTYPE=INV;SVMETHOD=EMBL.DELLYv1.7.2;END=1010012;PE=2;MAPQ=29;CT=3to3;CIPOS=-563,56
3;CIEND=-563,563        GT:GL:GQ:FT:RCL:RC:RCR:RDCN:DR:DV:RR:RV 0/0:0,-17.778,-419.003:10000:PASS:2994:1020499:419643:5:74:2:0:0
Chr1    5497    DEL00000014     A       <DEL>   138     PASS    IMPRECISE;SVTYPE=DEL;SVMETHOD=EMBL.DELLYv1.7.2;END=5278163;PE=4;MAPQ=34;CT=3to5;CIPOS=-542,54
2;CIEND=-542,542        GT:GL:GQ:FT:RCL:RC:RCR:RDCN:DR:DV:RR:RV 0/1:-5.08503,0,-129.371:51:PASS:3007:2954992:863025:7:25:4:0:0
```

Weird co-located SVs?

Lack of alt sequence for most variants

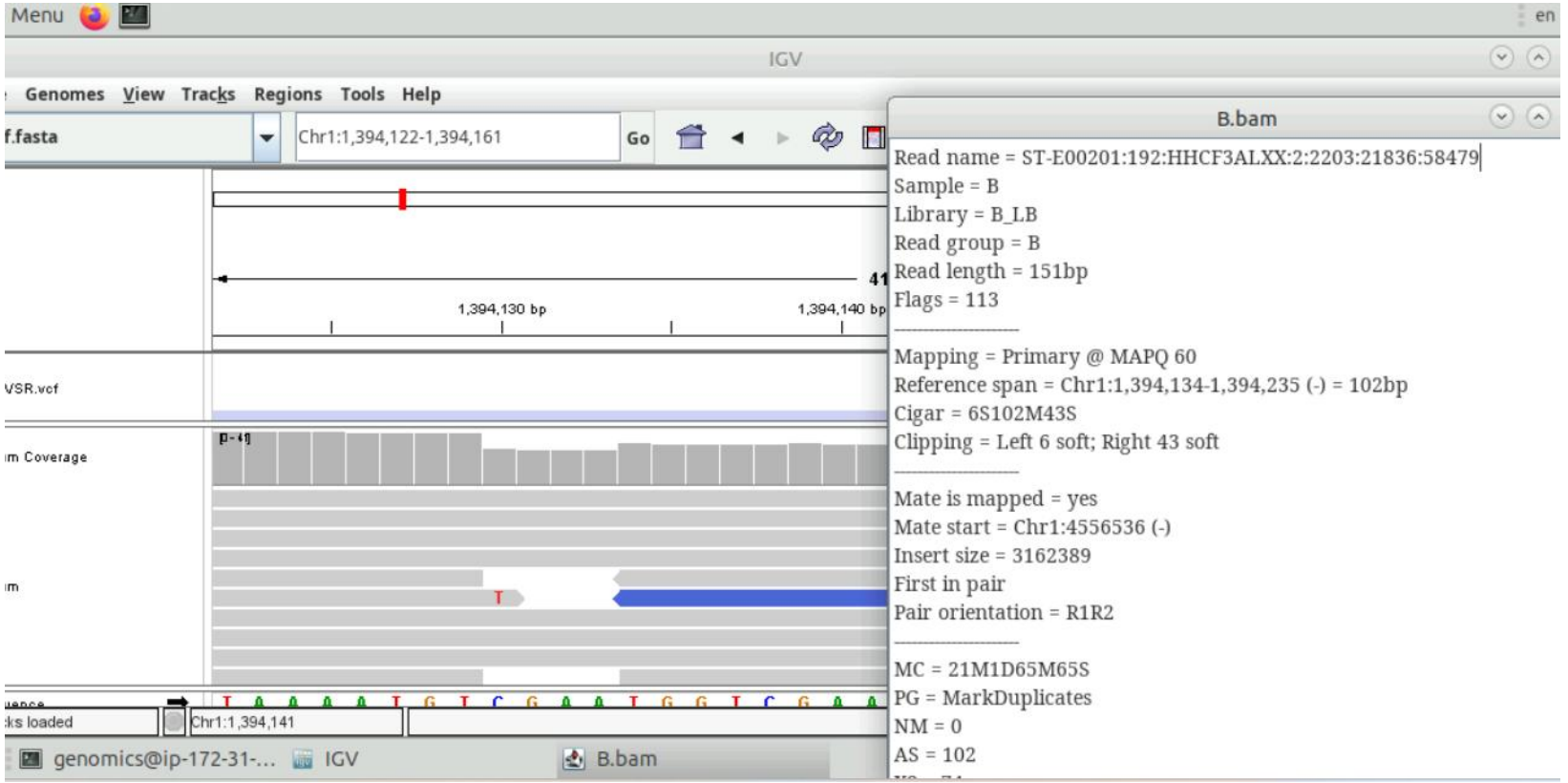
No SVLEN

4 – SV call with short-reads

-> What do you think about our large inversion? Was it detected by Delly?

Chr1 1393753 INV00011549 A <INV> 2120 PASS IMPRECISE;SVTYPE=INV;SVMETHOD=EMBL.DELLYv1.7.2;END=4556341;PE=39;MAPQ=60;
CT=5to5;CIPOS=-50,50;CIEND=-50,50 GT:GL:GQ:FT:RCL:RC:RCR:RDCN:DR:DV:RR:RV 1/1:-98.1971,-5.41565,0:54:PASS:1932910:1817073:901901:1:
0:18:0:0 0/1:-14.3866,0,-50.0868:144:PASS:1384270:1282649:621820:1:9:4:0:0 0/1:-10.7753,0,-118.775:108:PASS:1939344:1999980:
1008238:1:21:3:0:0 1/1:-100.1,-5.41844,0:54:PASS:1655494:1489965:724297:1:0:18:0:0
Chr1 1394089 INV00011550 G <INV> 1829 PASS IMPRECISE;SVTYPE=INV;SVMETHOD=EMBL.DELLYv1.7.2;END=4556153;PE=37;MAPQ=60;
CT=3to3;CIPOS=-163,163;CIEND=-163,163 GT:GL:GQ:FT:RCL:RC:RCR:RDCN:DR:DV:RR:RV 0/1:-47.1887,0,-0.689909:8:LowQual:1933048:1816911:901744
:1:1:10:0:0 0/1:-29.1784,0,-83.6784:10000:PASS:1384338:1282553:621712:1:15:6:0:0 0/1:-42.776,0,-72.8757:10000:PASS:1939476:1999794
:1008088:1:14:11:0:0 0/1:-43.2887,0,-2.68995:27:PASS:1655597:1489849:724173:1:1:10:0:0

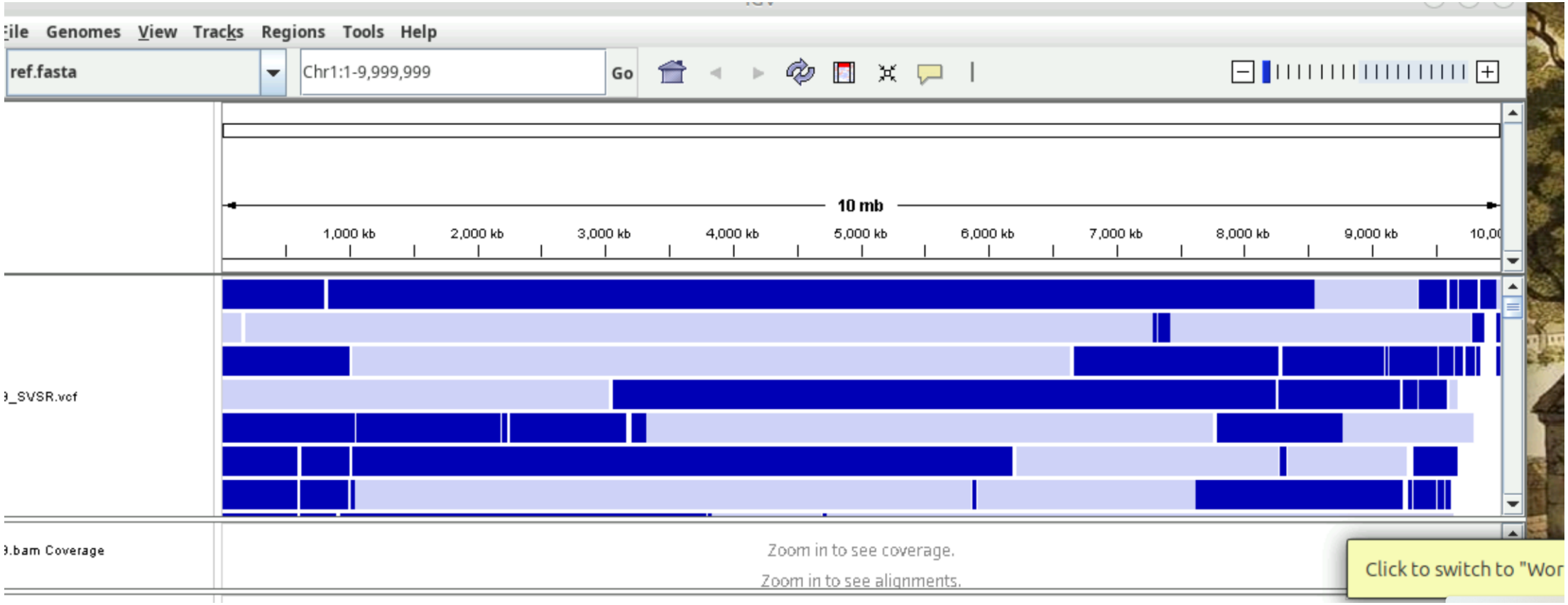
Maybe…
But it would be lost among the false positives!



4 – SV call with short-reads

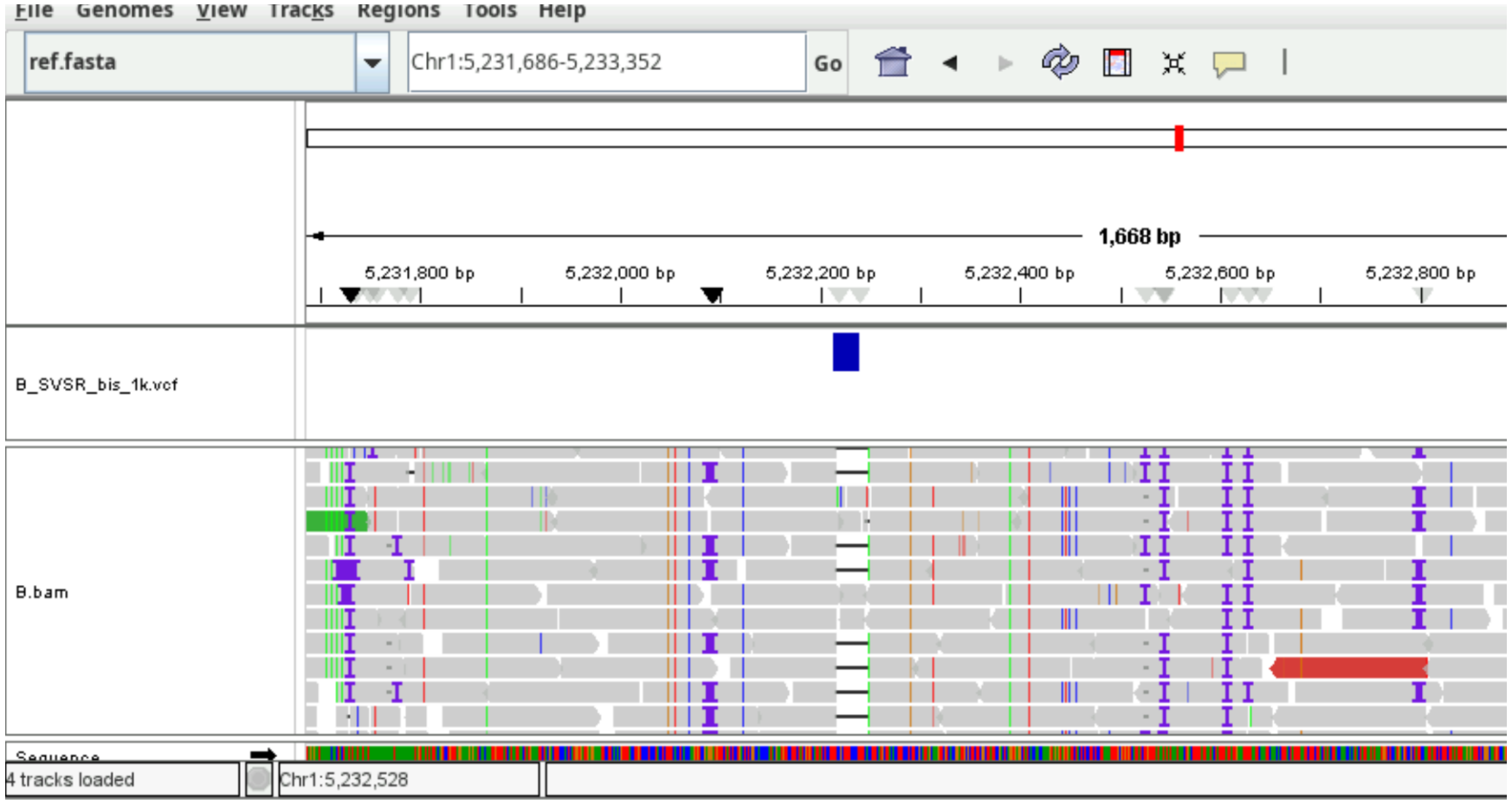
-> What do you think of the SVs? Is it readable?

Too many long SV for IGV
Lots of SVs at the same position
=> Probably a sign of lots of false positives!



4 – SV call with short-reads

-> Can you find some deletions well supported by reads ?



4 – SV call with short-reads

-> How many SVs are you analysing in total?

16330

-> What do you think of allelic frequency? And missing data?

Little missing data

BUT : lots of allelic frequency at 0 (and genotypes are 0/0) = ref/ref

⇒ Those are not even variants
⇒ Remove them

Chr1	5	INV00000000	9396125	0.75	0
Chr1	519	DEL00000001	552	0	0.125
Chr1	958	INS00000002	958	0	0
Chr1	5355	DUP00000003	826732	0	0.5
Chr1	5356	DUP00000004	180476	0	0
Chr1	5356	DUP00000005	616735	0	0.5
Chr1	5373	INV00000006	616199	0	0.5
Chr1	5375	INV00000007	894221	0	0
Chr1	5378	INV00000008	1021738	0	0.5
Chr1	5379	INV00000009	1018396	0	0
Chr1	5382	INV00000010	871395	0	0.5
Chr1	5383	DUP00000011	5547	0	0
Chr1	5383	DUP00000012	604520	0	0
Chr1	5383	DUP00000013	617527	0	0.5
Chr1	5383	INV00000014	816676	0	0
Chr1	5383	INV00000015	1024400	0	0.25
Chr1	5383	INV00000016	1068736	0	0
Chr1	5383	INV00000017	1143801	0	0
Chr1	5383	DUP00000018	1144327	0	0
Chr1	5383	INV00000019	3052463	0	0.25
Chr1	5383	INV00000020	3811013	0	0
Chr1	5383	INV00000021	8676150	0	0
Chr1	5384	DUP00000022	204296	0	0.25
Chr1	5384	INV00000023	4283625	0	0.5
Chr1	5386	INV00000024	968242	0	0
Chr1	5404	DUP00000025	1075475	0	0.125
Chr1	5405	INV00000026	1076829	0	0.125
Chr1	5425	DUP00000027	1021486	0	0.5
Chr1	5486	INV00000028	1096631	0	0
Chr1	5494	INV00000029	1010012	0	0
Chr1	5497	DEL00000030	5278163	0	0.25
Chr1	5503	INV00000031	744468	0	0.125
Chr1	5540	DEL00000032	959246	0	0
Chr1	5546	INV00000033	9567667	0	0.125
Chr1	5555	DEL00000034	8515626	0	0.125
Chr1	5557	DEL00000035	871141	0	0.5
Chr1	5621	DUP00000036	1046565	0	0
Chr1	5622	INV00000037	1046561	0	0.375
Chr1	5679	DUP00000038	6664587	0	0
Chr1	5728	INV00000039	958032	0	0.125
Chr1	5735	INV00000040	888458	0	0.5
Chr1	5757	INV00000041	1021865	0	0.5
Chr1	5776	DEL00000042	814677	0	0

4 – SV call with short-reads

-> How many SVs are you analysing in total? Same? 16330 - same

-> Do you notice differences with the earlier vcf?

	Chr	Start	ID.x	End	F_missing_before	MAF_before	ID.y	F_missing_after	MAF_after
1	Chr1	5	INV00000000	9396125	0.75	0	INV00000000	0.75	0
2	Chr1	113015	DUP00000129	2980316	0.50	0	DUP00000129	0.50	0
3	Chr1	143486	INS00000191	143486	0.75	0	INS00000191	0.75	0
4	Chr1	1060949	INV00010779	4104687	0.25	0.5	INV00010699	0.25	0.5
5	Chr1	1329380	DUP00011490	6847771	0.25	0.5	DUP00011392	0.25	0.5
6	Chr1	1419614	INS00011608	1419614	0.50	0	INS00011507	0.50	0
7	Chr1	1483914	INS00011643	1483914	0.25	0	INS00011541	0.25	0
8	Chr1	1488455	DEL00011646	1488499	0.25	0.5	DEL00011544	0.25	0.5
9	Chr1	1531849	DEL00011677	1532372	0.75	0	DEL00011574	0.75	0
10	Chr1	1537050	DEL00011680	1537103	0.25	0.166667	DEL00011577	0.25	0.166667
11	Chr1	1554194	DEL00011691	1554232	0.25	0.333333	DEL00011587	0.25	0.333333
12	Chr1	1567946	INS00011711	1567946	0.25	0	INS00011606	0.25	0
13	Chr1	1569316	DEL00011712	1569358	0.50	0	DEL00011607	0.50	0
14	Chr1	1584848	DEL00011729	1584932	0.25	0.166667	DEL00011624	0.25	0.166667
15	Chr1	1592686	DEL00011737	1706036	0.25	0.333333	DEL00011632	0.25	0.333333
16	Chr1	1607047	INV00011746	3686296	0.25	0.333333	INV00011640	0.25	0.333333
17	Chr1	1672225	INS00011781	1672225	0.25	0.333333	INS00011674	0.25	0.333333
18	Chr1	1691248	INV00011788	5153424	0.25	0.166667	INV00011681	0.25	0.166667
19	Chr1	1693676	INS00011793	1693676	0.25	0.166667	INS00011686	0.25	0.166667
20	Chr1	1733296	INV00011829	9367617	0.25	0.333333	INV00011720	0.25	0.333333
21	Chr1	1754730	DEL00011846	1754772	0.25	0.166667	DEL00011737	0.25	0.166667
22	Chr1	1796733	DEL00011874	1796793	0.25	0.333333	DEL00011765	0.25	0.333333
23	Chr1	1894880	DEL00011934	8803568	0.50	0.25	DEL00011825	0.50	0.25
24	Chr1	1899396	INS00011939	1899396	0.25	0.333333	INS00011830	0.25	0.333333

Not better for missing data

& re-genotyping has renamed the SV… (hard to compare!)

But breakpoints should have been refined (yay!)

5- Towards genome graphs

We get some warnings. Why?

The graph needs the
sequence in the ALT field…

warning:[vg::Constructor] vcflib could not canonicalize some SVs to base-level sequence; skipping variants like: Chr1 129696 Sniffles2.DUP.240S0 A <DUP> 59 PASS PRECISE;SVTYPE=DUP;SVLEN=1390;END=131086;SUPPORT=10;COVERAGE=31,23,662,25,32;STRAND=+-;STDEV_LEN=4.593;STDEV_POS=0.408;VAF=0.312
Warning: deletion END and SVLEN do not agree [canonicalize] Chr1 519626 Sniffles2.DEL.1F9S0 A <DEL> 51 PASS PRECISE;SVTYPE=DEL;SVLEN=-129928;END=649553;SUPPORT=11;COVERAGE=35,37,19,22,21;STRAND=+-;STDEV_LEN=0.000;STDEV_POS=0.000;VAF=0.423;SPAN=129928
END: 649553 SVLEN: 129928

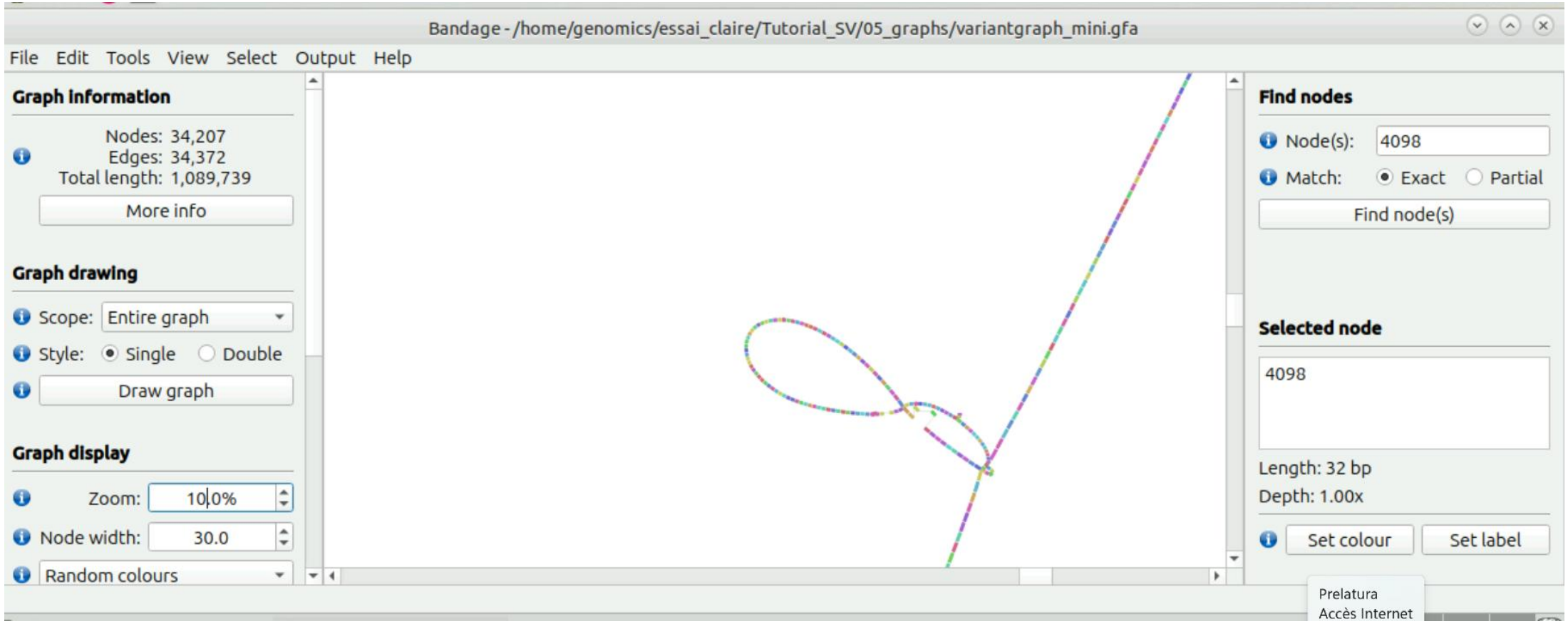
-> open the vcf: how many variants could you genotype? What does the format look like?

145 out of 170

5- Towards genome graphs

5- Towards genome graphs

Visualisation in Bandage – exemple of a bubble for a large SV



5- Towards (pan)genome graphs

-> What do you observe on the 2D representation?

Our big inversion!