**A - First type of wrong reads**

- reads with unusual distance between blaSHV genes

- consist of several DNA fragments that ligated together

- several of these DNA fragments carry different amplified regions

- presence of several red or green regions

- these reads should not be kept in the analysis

**CNV12**

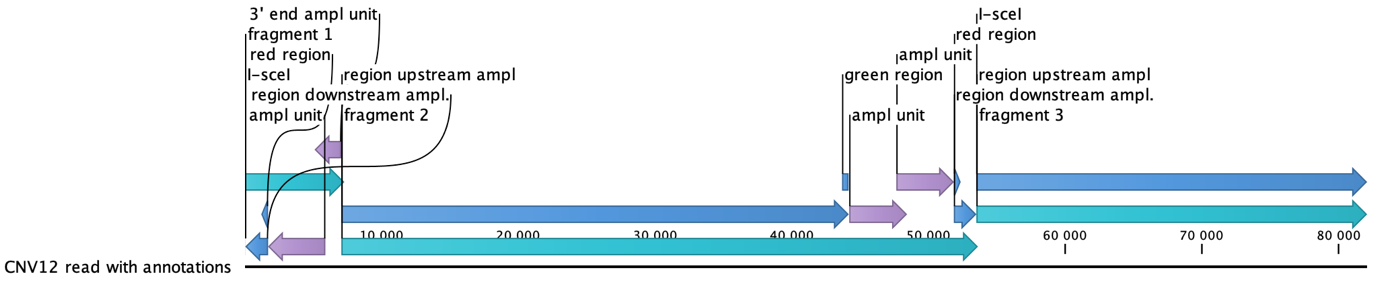
query subject identity length mismatch gaps start.query end.query start.subject end.subject e.value bit.score

blaSHV-1 21f4029e-dc29-4720-984b-9fb49cffe090 93.707 874 28 24 1 861 3574 4433 0 1284

blaSHV-1 21f4029e-dc29-4720-984b-9fb49cffe090 98.077 52 1 0 1 52 6996 7047 1.84e-20 91.6

blaSHV-1 21f4029e-dc29-4720-984b-9fb49cffe090 93.915 871 26 22 3 861 46495 45640 0 1290

blaSHV-1 21f4029e-dc29-4720-984b-9fb49cffe090 94.253 870 27 20 1 861 49952 49097 0 1308



**CNV9**

query subject identity length mismatch gaps start.query end.query start.subject end.subject e.value bit.score

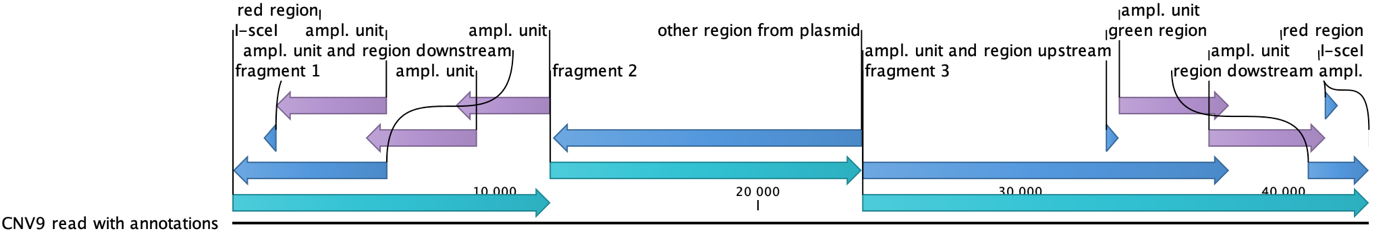
blaSHV-1 be55166a-1415-4f64-b234-029d76f27bb1 92.571 875 34 28 1 861 3565 4422 0 1227

blaSHV-1 be55166a-1415-4f64-b234-029d76f27bb1 89.217 881 49 39 1 859 6976 7832 0 1059

blaSHV-1 be55166a-1415-4f64-b234-029d76f27bb1 92.36 877 26 25 1 861 10401 11252 0 1210

blaSHV-1 be55166a-1415-4f64-b234-029d76f27bb1 91.657 875 34 31 1 861 36061 35212 0 1175

blaSHV-1 be55166a-1415-4f64-b234-029d76f27bb1 92.06 869 43 23 1 861 39421 38571 0 1199



**CNV\_5\_1**

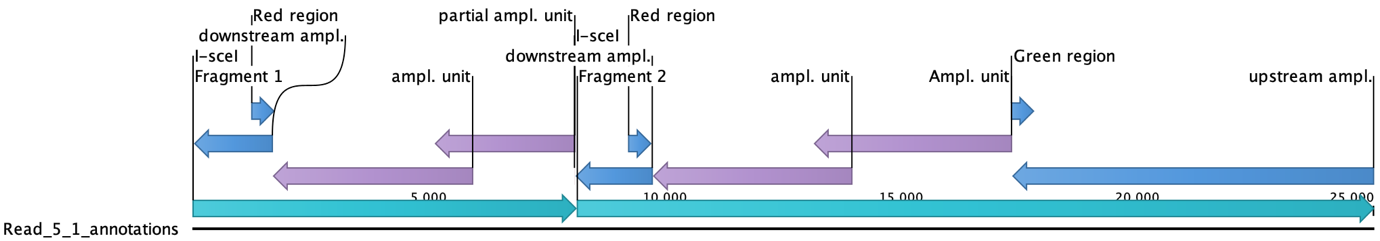
query subject i dentity length mismatch gaps start.query end.query start.subject end.subject e.value bit.score

blaSHV-1 cf14f1cd-68b2-4993-9bd1-3098614683a1 94.948 871 19 20 1 861 3604 4459 0 1341

blaSHV-1 cf14f1cd-68b2-4993-9bd1-3098614683a1 95.091 876 19 18 1 861 7017 7883 0 1358

blaSHV-1 cf14f1cd-68b2-4993-9bd1-3098614683a1 92.069 870 39 26 1 861 11633 12481 0 1197

blaSHV-1 cf14f1cd-68b2-4993-9bd1-3098614683a1 93.037 876 26 32 1 861 15049 15904 0 1247



**CNV\_5\_2**

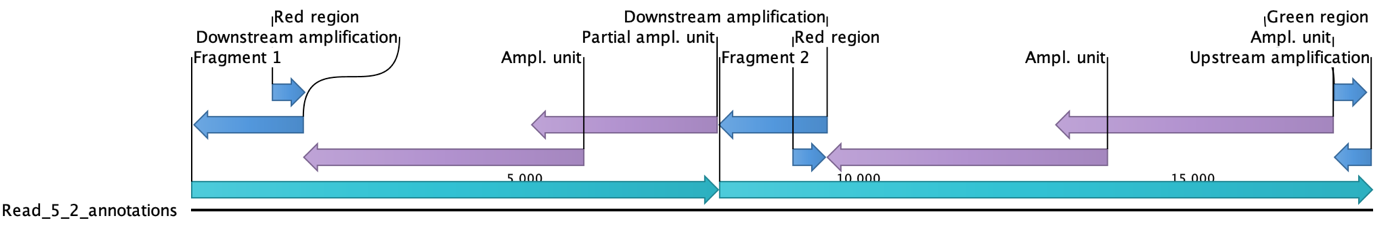
query subject identity length mismatch gaps start.query end.query start.subject end.subject e.value bit.score

blaSHV-1 948c7886-16fb-419f-acac-8155bc32222c 89.977 878 40 43 1 859 3587 4435 0 1090

blaSHV-1 948c7886-16fb-419f-acac-8155bc32222c 91.106 877 35 33 4 861 6979 7831 0 1147

blaSHV-1 948c7886-16fb-419f-acac-8155bc32222c 89.761 879 39 37 1 861 11428 12273 0 1077

blaSHV-1 948c7886-16fb-419f-acac-8155bc32222c 94.24 868 24 21 1 861 14791 15639 0 1303

****

**CNV\_5\_3**

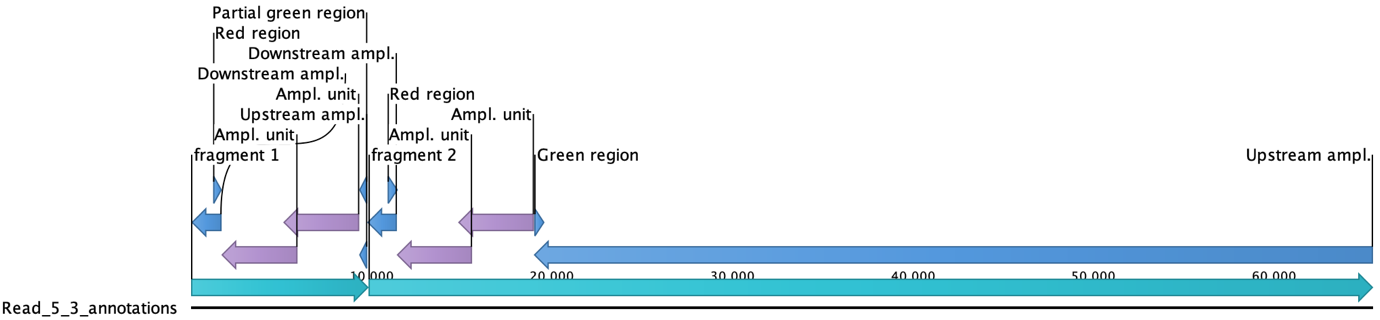
query subject identity length mismatch gaps start.query end.query start.subject end.subject e.value bit.score

blaSHV-1 7009605b-a57f-4ae2-8aeb-59e4822ffa7f 93.273 877 29 26 1 861 3575 4437 0 1266

blaSHV-1 7009605b-a57f-4ae2-8aeb-59e4822ffa7f 93.19 881 21 32 1 861 7004 7865 0 1258

blaSHV-1 7009605b-a57f-4ae2-8aeb-59e4822ffa7f 92.466 876 25 29 1 861 13280 14129 0 1214

blaSHV-1 7009605b-a57f-4ae2-8aeb-59e4822ffa7f 93.226 871 28 30 1 861 16701 17550 0 1253

****

**CNV6\_1**

query subject identity length mismatch gaps start.query end.query start.subject end.subject e.value bit.score

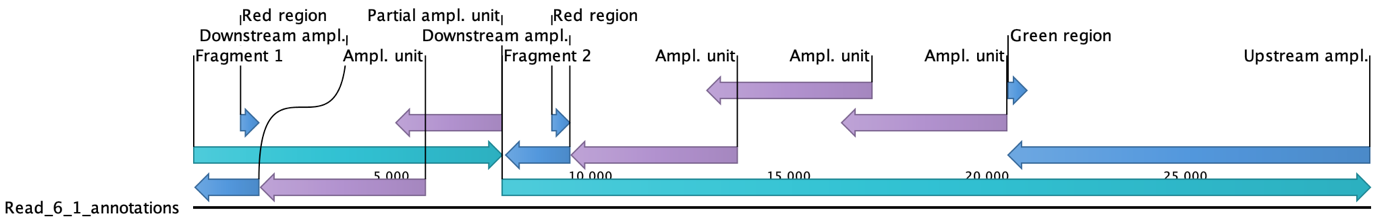
blaSHV-1 368708bd-a4cc-49db-acdf-b3960889697a 92.898 873 35 26 1 861 3561 4418 0 1243

blaSHV-1 368708bd-a4cc-49db-acdf-b3960889697a 93.875 849 26 21 1 842 6948 7777 0 1256

blaSHV-1 368708bd-a4cc-49db-acdf-b3960889697a 91.959 883 37 29 1 861 11379 12249 0 1206

blaSHV-1 368708bd-a4cc-49db-acdf-b3960889697a 92.255 878 22 33 1 861 14794 15642 0 1203

blaSHV-1 368708bd-a4cc-49db-acdf-b3960889697a 93.341 871 31 22 1 861 18197 19050 0 1262



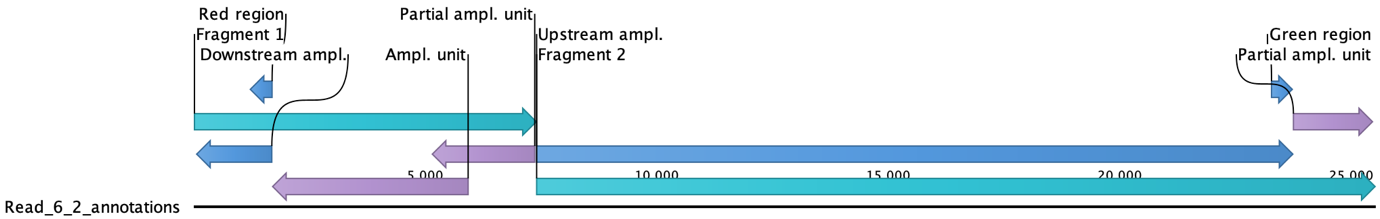
**CNV\_6\_2**

query subject identity length mismatch gaps start.query end.query start.subject end.subject e.value bit.score

blaSHV-1 965d93aa-a6af-4dda-ac06-f53bf9af4a27 90.465 881 44 34 1 861 3608 4468 0 1125

blaSHV-1 965d93aa-a6af-4dda-ac06-f53bf9af4a27 94.722 360 11 8 1 356 7032 7387 6.07e-160 553

blaSHV-1 965d93aa-a6af-4dda-ac06-f53bf9af4a27 90.391 281 13 13 590 861 25463 25188 5.11e-101 357

****

**CNV\_9**

query subject identity length mismatch gaps start.query end.query start.subject end.subject e.value bit.score

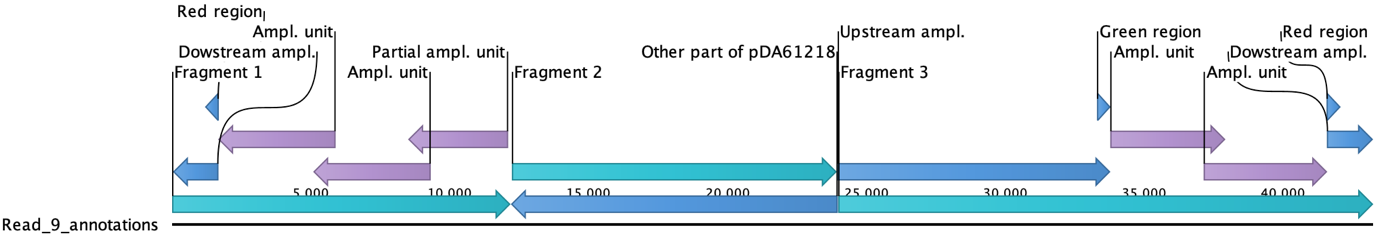
blaSHV-1 be55166a-1415-4f64-b234-029d76f27bb1 92.571 875 34 28 1 861 3565 4422 0 1227

blaSHV-1 be55166a-1415-4f64-b234-029d76f27bb1 89.217 881 49 39 1 859 6976 7832 0 1059

blaSHV-1 be55166a-1415-4f64-b234-029d76f27bb1 92.36 877 26 25 1 861 10401 11252 0 1210

blaSHV-1 be55166a-1415-4f64-b234-029d76f27bb1 91.657 875 34 31 1 861 36061 35212 0 1175

blaSHV-1 be55166a-1415-4f64-b234-029d76f27bb1 92.06 869 43 23 1 861 39421 38571 0 1199

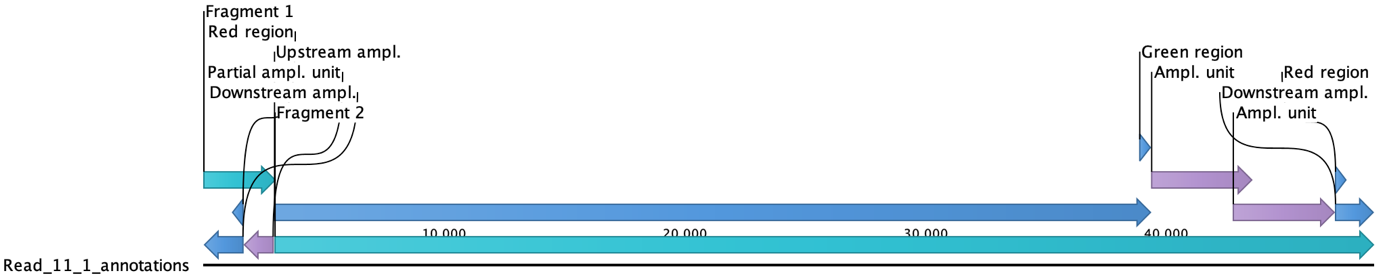
****

**CNV\_11\_1**

query subject identity length mismatch gaps start.query end.query start.subject end.subject e.value bit.score

blaSHV-1 c812976f-85a9-4122-922a-93755e9bc9f9 91.116 878 37 36 1 861 41661 40808 0 1151

blaSHV-1 c812976f-85a9-4122-922a-93755e9bc9f9 91.913 878 35 29 1 861 45079 44221 0 1195

****

**CNV\_12**

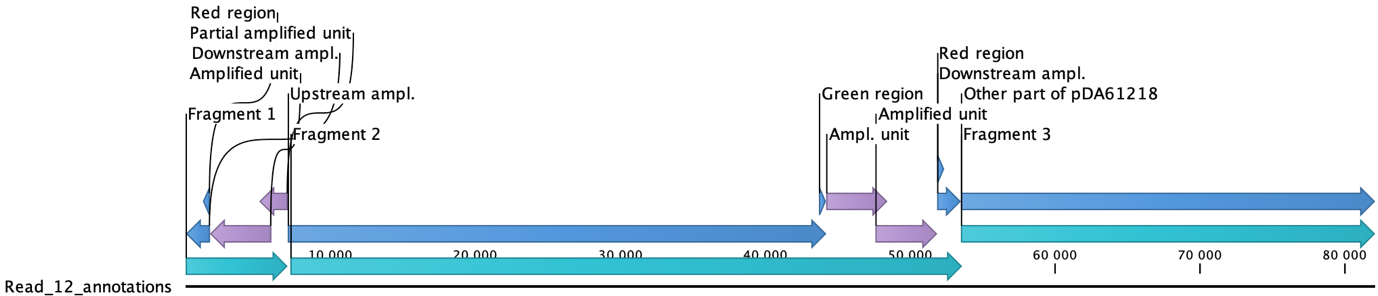
query subject i dentity length mismatch gaps start.query end.query start.subject end.subject e.value bit.score

blaSHV-1 21f4029e-dc29-4720-984b-9fb49cffe090 93.707 874 28 24 1 861 3574 4433 0 1284

blaSHV-1 21f4029e-dc29-4720-984b-9fb49cffe090 98.077 52 1 0 1 52 6996 7047 1.84e-20 91.6

blaSHV-1 21f4029e-dc29-4720-984b-9fb49cffe090 93.915 871 26 22 3 861 46495 45640 0 1290

blaSHV-1 21f4029e-dc29-4720-984b-9fb49cffe090 94.253 870 27 20 1 861 49952 49097 0 1308

****

**CNV\_13\_1**

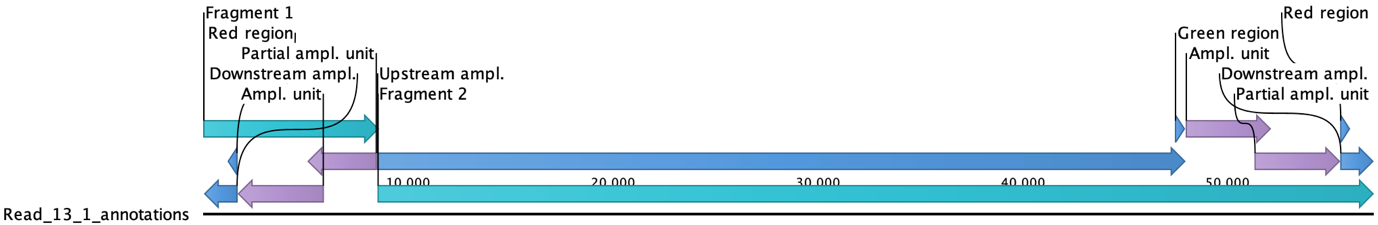
query subject identity length mismatch gaps start.query end.query start.subject end.subject e.value bit.score

blaSHV-1 9d4dac8e-a042-4de7-9d39-e30866c9765c 91.938 707 29 25 159 859 3732 4416 0 965

blaSHV-1 9d4dac8e-a042-4de7-9d39-e30866c9765c 93.7 873 26 27 1 861 6957 7812 0 1280

blaSHV-1 9d4dac8e-a042-4de7-9d39-e30866c9765c 89.425 870 50 29 1 859 50220 49382 0 1059

blaSHV-1 9d4dac8e-a042-4de7-9d39-e30866c9765c 92.906 874 25 29 1 861 53630 52781 0 1236

****

**B- Second type of wrong reads**

- reads with unusual distance between blaSHV genes

- consist of several DNA fragments that ligated together

- several DNA fragments carry different amplified regions

- presence of a single red and a single green region

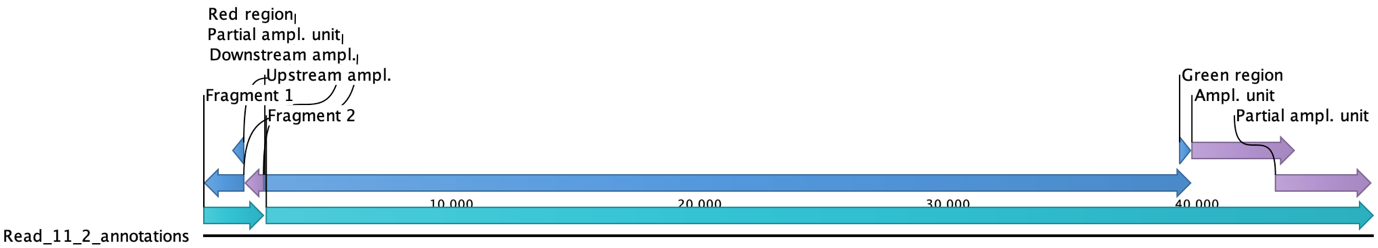
- these reads should not be kept in the analysis

**CNV\_11\_2**

query subject i dentity length mismatch gaps start.query end.query start.subject end.subject e.value bit.score

blaSHV-1 3c6176f6-ccb9-4961-bfcc-adc009458e2e 91.419 874 36 31 1 861 42070 41223 0 1162

blaSHV-1 3c6176f6-ccb9-4961-bfcc-adc009458e2e 91.638 873 39 28 1 861 45449 44599 0 1177

****

**C- Reads with numerous amplified copies that seemed correct according to Andrei**

Reads are indeed correct and should be kept in the analysis.

**CNV\_5\_4**

query subject identity length mismatch gaps start.query end.query start.subject end.subject e.value bit.score

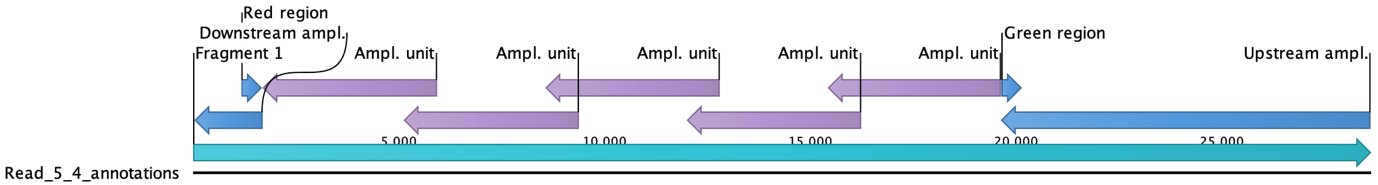
blaSHV-1 4cb0481e-2af5-4e6a-9e91-2700d0c481a0 93.714 875 27 24 1 861 3591 4451 0 1286

blaSHV-1 4cb0481e-2af5-4e6a-9e91-2700d0c481a0 95.183 872 19 16 1 861 7022 7881 0 1356

blaSHV-1 4cb0481e-2af5-4e6a-9e91-2700d0c481a0 94.145 871 24 24 1 861 10470 11323 0 1301

blaSHV-1 4cb0481e-2af5-4e6a-9e91-2700d0c481a0 95.632 870 12 22 1 861 13889 14741 0 1373

blaSHV-1 4cb0481e-2af5-4e6a-9e91-2700d0c481a0 94.527 877 20 22 1 861 17317 18181 0 1328

****

**CNV\_7**

query subject identity length mismatch gaps start.query end.query start.subject end.subject e.value bit.score

blaSHV-1 1e27948e-4554-4533-be59-16e1154978fd 93.265 876 27 30 1 861 3569 4427 0 1262

blaSHV-1 1e27948e-4554-4533-be59-16e1154978fd 93.607 876 31 24 1 861 6973 7838 0 1284

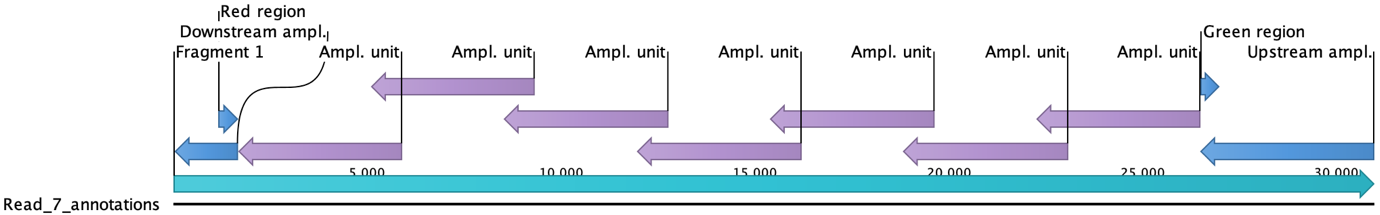
blaSHV-1 1e27948e-4554-4533-be59-16e1154978fd 94.737 874 22 18 1 861 10410 11272 0 1338

blaSHV-1 1e27948e-4554-4533-be59-16e1154978fd 93.28 878 22 29 1 861 13851 14708 0 1260

blaSHV-1 1e27948e-4554-4533-be59-16e1154978fd 93.8 871 26 24 1 861 17287 18139 0 1284

blaSHV-1 1e27948e-4554-4533-be59-16e1154978fd 94.971 875 22 18 1 861 20707 21573 0 1352

blaSHV-1 1e27948e-4554-4533-be59-16e1154978fd 94.539 879 16 25 1 861 24160 25024 0 1328



**(Updated) possible approach to solve the problem:**

Step 1- Correctly detect red and green regions. This was not the case for these reads, so there is a need to adjust the BLAST parameters or the size of the red and /or green regions used for the BLAST analysis. All reads with more than one red or more than one green region have to be trashed (reads presented in A). This is an easy first step to get rid of most problematic reads in the wrole of the data.

Step 2- Some of the reads with only one red and one green region might still be wrong (e.g., see read presented in B). Such reads will not be trashed following step 1 and they will lead to wrongly determined numbers of amplified units. These rare reads are likely only affecting the data for rarely detected events (events in the lower or higher tail of the distribution of copy numbers of amplified units detected) and should therefore be sorted out only for these rarely detected events. For all reads with rare events in term of amplification copy numbers (e.g., we could set a threshold at events detected in maximum 1% of the bacteria for example, or even lower), BLAST blaSHV to detect how many repeats are present (only use fully detected blaSHV genes, that is results matching over ±861nt in length). If the number of repeats does not fit with the distance between the green and red regions, trash the read. If the number of repeats is compatible with the distance between the red and green regions, keep the read.

**Remark 1:** When we will sequence populations where the average copy number of amplified units is higher, we should expect the frequency of reads of type B to drastically increase (since ligation of reads that ended in the amplified unit is more likely). Thus, for these populations it might be important to figure-out how to detect (and remove from the analysis) such reads by analysing all the reads that passed the Step 1 above. This will likely depend on what is reasonable to compute on a large number of reads.

**A third type of wrong reads might also be present** in the data collected from populations with high average copy numbers of amplified units. These wrong reads will result from two (or more) fragments that ligated to form what could appear as a "normal" amplified region, but in reality is only the result of the ligation. Such events could be detected using the fact that one amplified unit would likely not be of the expected correct size. We should determine if such analysis is possible (that is, if it is computationally reasonable) on a large scale, or what compromise method of analysis one could use (for example, only analyzing a subset of the reads displaying each of the CN present in the population distribution and extrapolating the final results using that analysis).

**Remark 2:** The fact that red regions were sometimes missed might also indicate that in the initial screen for reads with green and red regions might have gotten rid of good reads where the region might have been missed by the BLAST analysis. Thus, it is important that the blast of the red and green regions are optimized to pick up the largest number of correct reads.

**Remark on the size of the reads to keep for analysis.** If using reads that are, for example, ≥6.5kb -long (that is, reads that carry a single copy + the red and the green regions), the correct frequency of bacteria with no amplified unit (deletion of the amplified unit) and the frequency of bacteria with a single amplified unit can be correctly determined. The frequency will be the number of reads with 0 or 1 copy of amplified units divided by the total number of reads analyzed (that is, the total number of reads of ≥6.5kb with both single red and single green regions detected). However, the frequency of bacteria carrying >1 amplified units cannot be determined correctly with reads of ≥6.5 kb since some reads with >1 copy of amplified units will be kept (those where both the green and red sequences are found) while some will be trashed (reads where the green region is not detected because the read ends within an amplified unit and does not cover the whole amplified region). To correctly determine the frequency of bacteria with >1 amplified units, longer reads have to be kept. For example, reads of ≥8.3kb, with both single red and single green regions detected, can be used to precisely determine the frequency of reads with 2 copies of the amplified units. Analysis of these reads should also give again the correct frequencies for 1 and 0 amplified units, assuming that there are enough reads to measure these frequencies. Hence, calculating the frequency of bacteria with different copy numbers of amplified units might require performing several analysis using reads of different minimum sizes. For highest CN of amplified units, there might not be enough reads to detect their frequency precisely and data might have to be extrapolated from the CN distribution curve.