nanopore infosheet

December 2018

1 Adapters

Apparently there's a way MinKNOW recognizes the adapter sequences during the sequencing procedure even if the basecalling is switched off (because there's a pore status "Adapter" meaning that right now an adapter is going through the pore) but the link from the Nanopore Community explaining the procedure does not work (`github.com/nanopore-wgs-consortium/NA12878/blob/master/rna_slides/Slide09.png`).

The trimming of adapters during the minKNOW sequencing without basecalling is unknown. Approximately 20% of the reads processed with Albacore will still contain adapters, from Oxford Nanopore Community users experience. One can trim 30-60 bases from each end to trim most of the adapter sequences, or use specific instruments like *Porechop*. On *Porechop* github page in the file `adapters.py` all the adapter sequences are contained.

Direct RNA (SQK-RNA001 and 002) adapter sequences:
RTA
Top:
5' - GGCTTCTTCTTGCTCTTAGGTAGTAGGTTC - 3'
Bottom:
5' - GAGGCGAGCGGTCAATTTTCCTAAGAGCAAGAAGAAGC-CTTTTTTTTTT - 3'
RMX
Top:
5' - TGATGATGAGGGATAGACGATGGTTGTTTCTGTTGGTGCT-GATATTGCTTTTTTTTTTTTTATGATGCAAGATACGCAC - 3'
Bottom:
5' - GAGGCGAGCGGTCAATTTGCAATATCAGCACCAACAGAAACAAC-CATCGTCTATCCCTCATCATCAGAACCTACTA - 3'

Rapid DNA (SQK-RAD004) sequencing kit adapter:
5' - GTTTTCGCATTTATCGTGAAACGCTTTCGCGTTTTTCGT-GCGCCGCTTCA - 3'

2 Raw-raw data

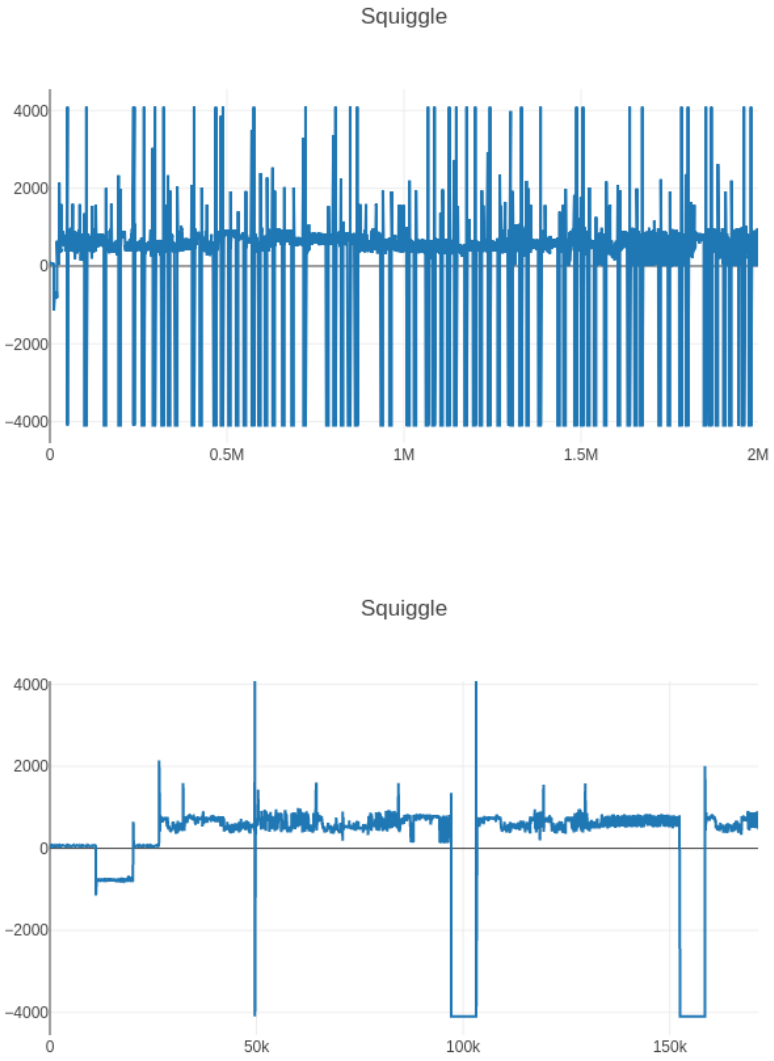


Figure 1: Long RNA5 (G-quadruplex forming) raw data from channel 100, pA vs samples (1/3024 s). Every 20-50k samples (around 10-20 s) dramatic pore voltage reversals occur. Top: full view of the raw data; bottom: more detailed view of the raw data, but still not cut into reads.

3 Quality score and read classification

Phred quality score represents the probability of error in one base. If $P = 1$, the base is 100% incorrect; if $P = 0$, it is 100% correct. The quality score is calculated as:

$$Q = -10log_{10}P, \qquad \text{or} \qquad P = 10^{-Q/10}. \tag{1}$$

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

Figure 2: Phred quality score symbols and respective error probabilities.

4 RNA3 and RNA9 short sequences analysis

The RNA3 and RNA9 reads were performed with same flow cell FAH54029 (it is indicated in the bulk filename). RNA3, G-quadruplex forming sequence, was sequenced first, and then RNA9, mutated sequence. RNA3 might have contaminated RNA9, which was observed in practice: around 0.5% of the reads contained G-quadruplex forming sequence from RNA3, and some of them cause blockage not typical for RNA9. The most convenient way to perform manual sequence analysis is through the .fastq file of the basecalled reads after Albacore processing; it generates one .fastq file containing all the sequences.

RNA3
5' - GGAUAAAGAGUUGGGUGGGUGGGUGGGUCAGAAGCACAGAUAU
CCUGGUUUGCUCCUGAGGAUCAUGGAUACGGUACUGAAGUUUCUA
CUAAGAACACACCAUGCUCAGAGAACAAACUUGACAUCCAGGA - 3'
RNA9
5' - GGAUAAAGAGUUGAGUGAGUGAGUGAGUCAGAAGCACAGAUAU
CCUGGUUUGCUCCUGAGGAUCAUGGAUACGGUACUGAAGUUUCUA
CUAAGAACACACCAUGCUCAGAGAACAAACUUGACAUCCAGGA - 3'

We're able to perform visualization of the bulk files with overlaid reads (pass+fail) and to plot the detailed passed reads with overlaid sequence separately (custom script). The new software `bulkvis` has been released in Nov. 2018 which allows to visualize bulk files with some flags like adaptor onset, pore unblocking etc. The sampling rate is currently ./3012 for RNA and ./6024

for DNA. Depending on the voltage applied, ambient temperature and electrochemical resistance of both common and individual electrodes (might degrade over time), the open pore current will have values between 1200-1600 pA, and plateau levels corresponding to pore stalling prior to sequencing at 700-900 pA. Voltage drift happens in course of hours, about 3% in 2 hours of sequencing, according to graphs in nanopore community, and os corrected dynamically each 2 hours; for short sessions it's not observable.

Some examples of the reads below. It was shown by authors [1] that some reads are cut for no particular reason while the molecule might still be in pore, if you look at the bulk signal; this happens with G-quadruplex forming molecules as well due to blockage the molecules cause not recognized as a normal signal. Software applies voltage reversal in order to unblock the pore, and initially 2 seconds of reverse voltage was applied statically, now the progressive flickering is implemented (one quick flick and a few longer flicks).

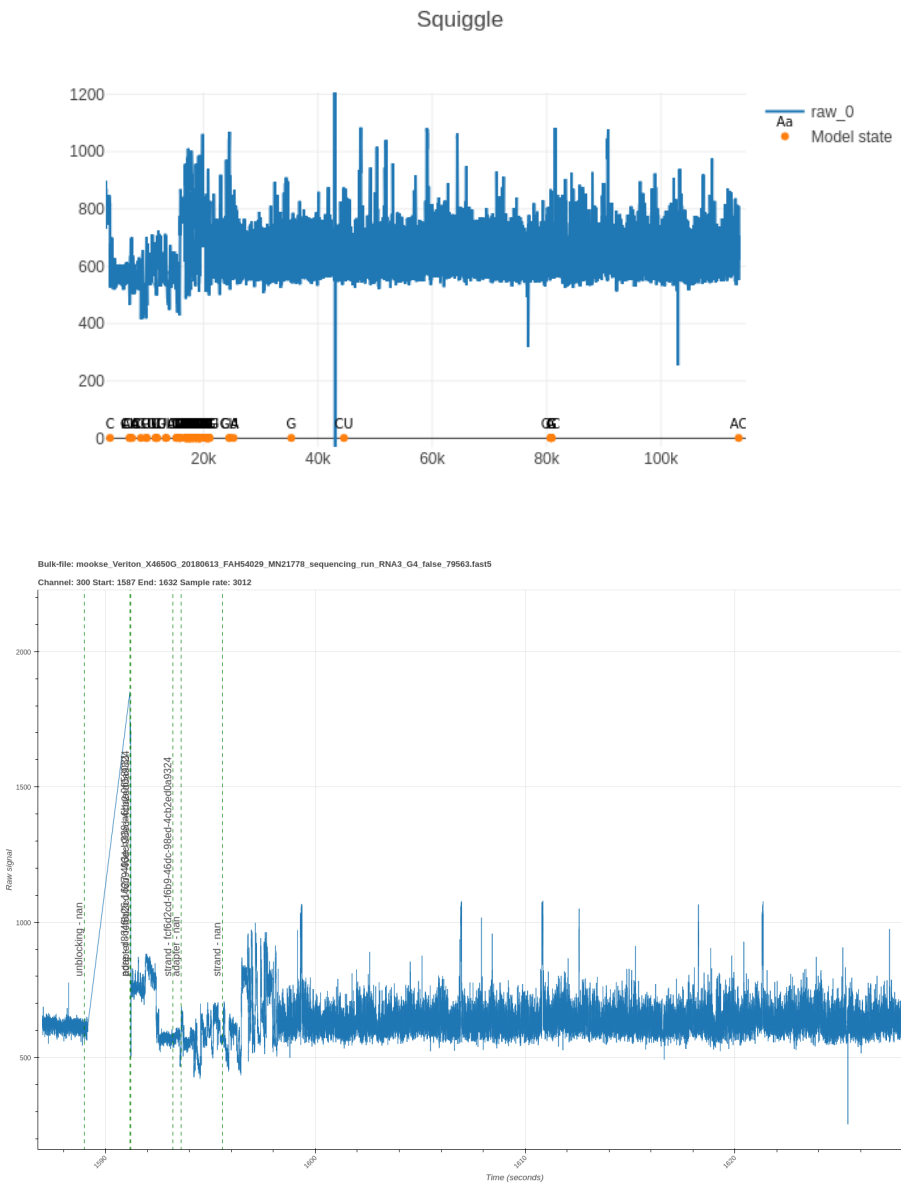


Figure 3: Top: RNA3, read 654 from channel 300: typical blockage for G-quadruplex forming sequence, sequence preceding G-rich part has been found in the basecalled fastq. Bottom: the same read visualized with `bulkvis`.

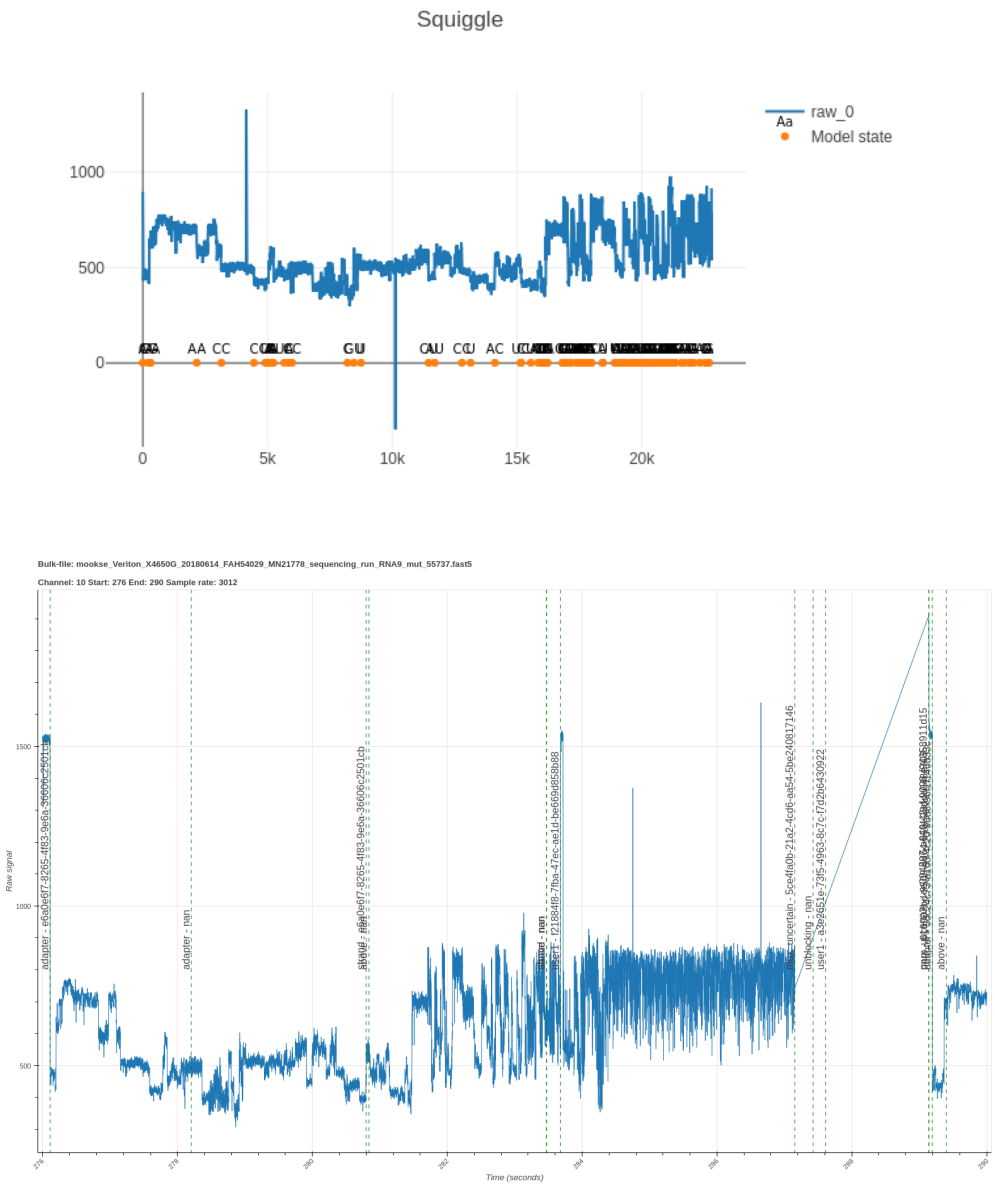


Figure 4: Top: RNA9, read 131 from channel 10: nontypical blockage for mutated sequence. Bottom: the same read visualized with `bulkvis`. The blockage part is not considered as a part of read by MinKNOW, and is probably caused by some other contaminants as this read coincides with RNA9 sequence, no RNA3 presence found.

Further sequence analysis shows that although G-quadruplex sequence is, in fact, disrupted improportionally compared to plain sequence and the bases are read incorrectly in this part, QS is retained high almost always, meaning that G-quadruplex presence changes the signal such that the base is recognized incorrectly. For instance, here in red the bases which are incorrectly read but still have a high (8 and above) quality score are indicated:

UUUGGUGGUGGGUGUCAGAAGCACAGA – from the read 691

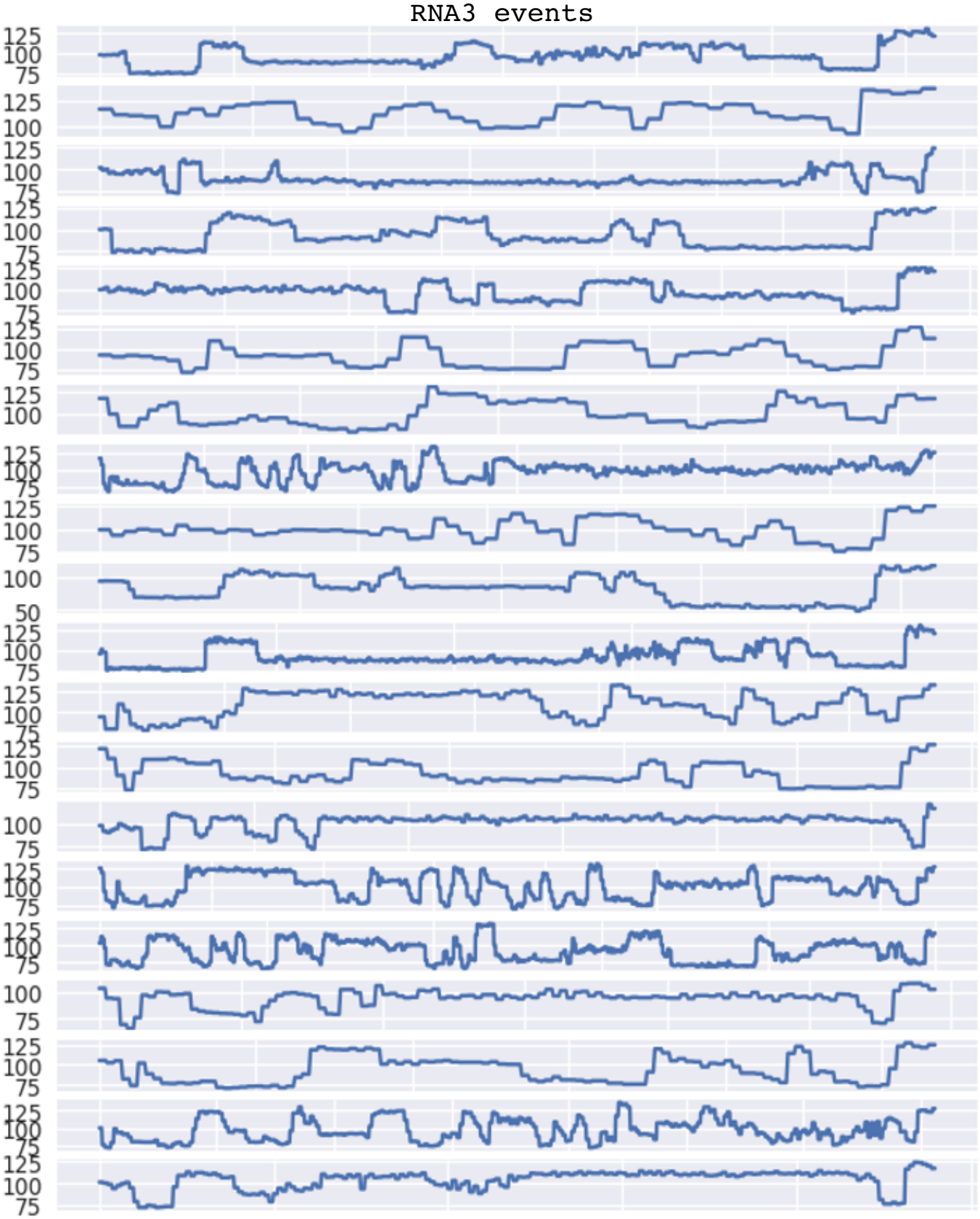
GGUGGGUGGGUGGGUCAGAAGCACAGA – known sequence

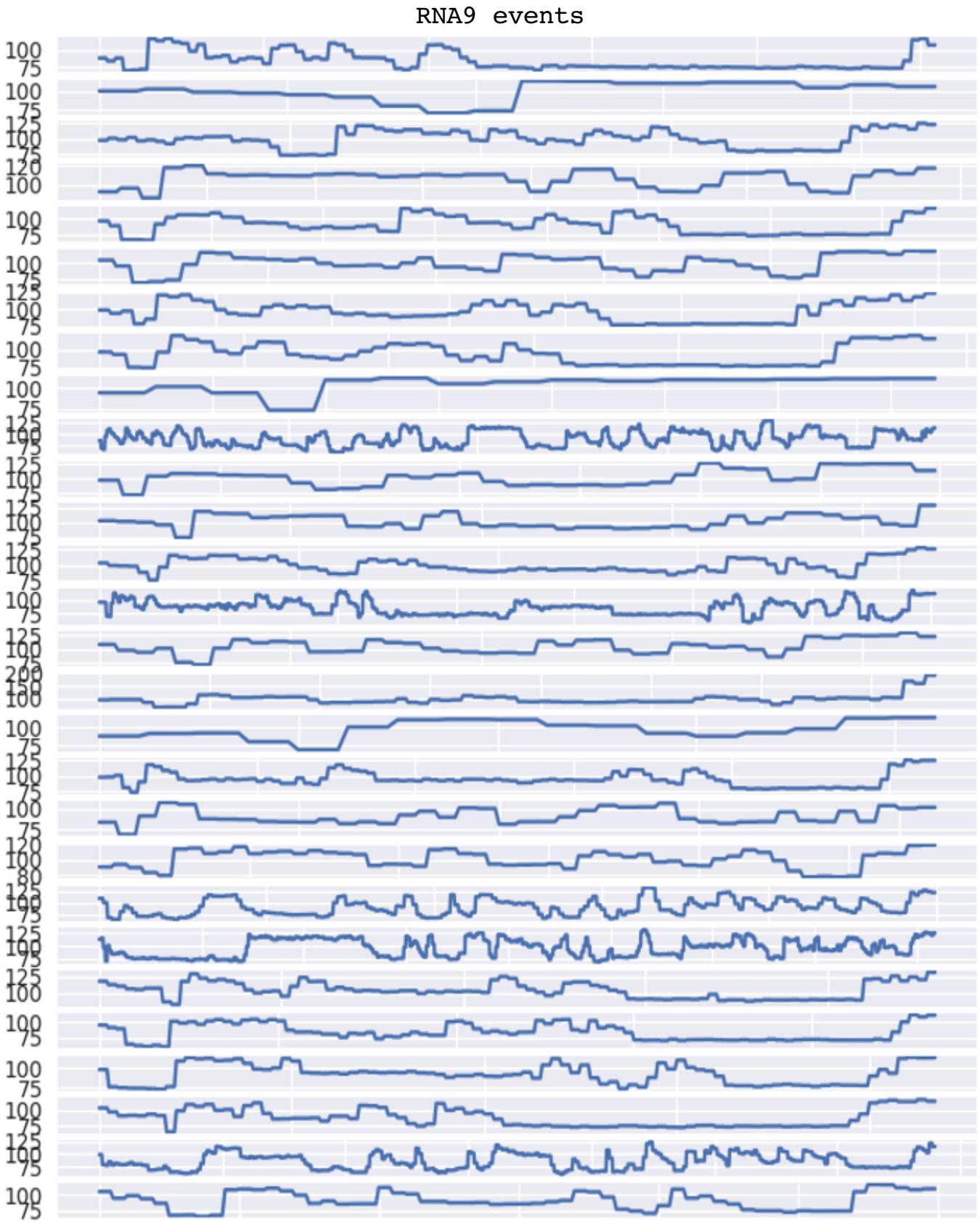
The rest of the incorrect bases above have low quality score. In the example below all the highlighted bases from the read have exceptionally high quality scores (16 and above).

UGGUUGGUUGGUUGGUACAGAAGCACAGAUA – from the read 483

UGGGUGGGUGGGUGGGUCAGAAGCACAGAUA – known sequence

In the figures below the regions corresponding to G-quadruplex forming sequence in RNA3 and respective mutated sequence in RNA9 are shown, extracted as signal/events between the preceding and following sequences. There's no clear pattern observed.


RNA3 events
125
100
75
125
100
125
100
75
125
100
75
125
100
75
125
100
75
125
100
125
100
75
125
100
75
100
50
125
100
75
125
100
75
125
100
75
100
75
125
100
75
125
100
75
100
75
125
100
75
125
100
75


RNA9 events

References

[1] Alexander Payne, Nadine Holmes, Vardhman Rakyan, Matthew Loose; BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files, Bioinformatics, bty841, 2018.