Less is more: filtering and trimming ONT sequencing data

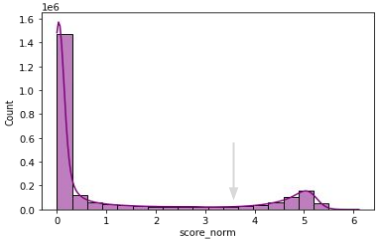
Natalia Nenasheva[1,2,3*], Valery Ilyinsky[3], Vsevolod Makeev[1,2]

[1] *Moscow Institute of Physics and Technology, Moscow, Russia*
[2] *Vavilov Institute of General Genetics, Moscow, Russia*
[3] *Medical genetic center Genotek, Moscow, Russia*

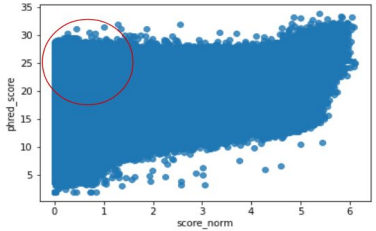
Moscow 2022

Data filtering criteria developed

- The cause of the emergence of reads with a high quality indicator, but at the same time poorly aligned to the genome, has been identified



Distribution of alignment scores normalized to the length of each read

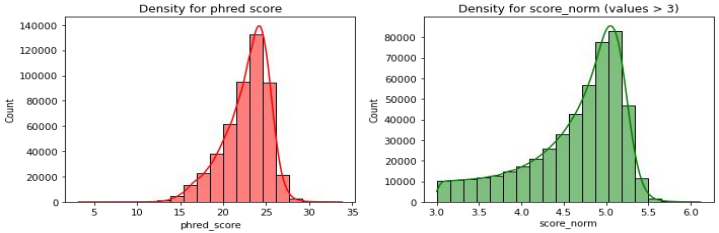


Reads with phred score > 25 and normalized score score_norm < 1

- Revealed multiple alignment of reads with phred score > 25 (high quality) and normalized score score_norm < 1

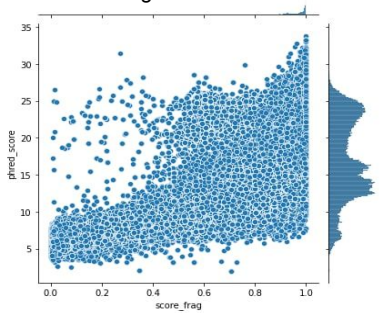
Sequences producing significant alignments:	Score (Bits)
AC093414.3 Homo sapiens chromosome 3 clone RP11-193I22, complete...	60859
AC012590.5 Homo sapiens BAC clone CTD-2375H4 from 7, complete se...	1657
AC079809.4 Homo sapiens BAC clone RP11-958M14 from 7, complete s...	1657
AC019051.8 Homo sapiens BAC clone RP11-92L24 from 2, complete se...	1657
AC018646.3 Homo sapiens chromosome 7 clone RP11-344L16, complete...	1655
NG_030347.1 Homo sapiens neuron navigator 2 (NAV2), RefSeqGene o...	1652
AF104455.1 Homo sapiens chromosome 7qtelo BAC F6, complete sequence	1652
AC023950.6 Homo sapiens chromosome 11, clone RP11-808N1, complet...	1652
AC190206.3 Pan troglodytes BAC clone CH251-639P15 from chromosom...	1650
AC191237.2 Pan troglodytes BAC clone CH251-271O14 from chromosom...	1650
AP023475.1 Homo sapiens DNA, chromosome 15, nearly complete genome	1646
AC147322.2 Pan troglodytes BAC clone RP43-41H22 from chromosome ...	1646
AC279844.1 Pongo abelii chromosome unknown clone CH276-312I17, c...	1644
AC192128.3 Pan troglodytes BAC clone CH251-610J6 from chromosome...	1644

- Selected reads with the best alignment score and the best quality



Density distribution of phred_score and score_norm for data with score_norm > 3

- It has been established that the level of its fragmentation should also be taken into account as a parameter for filtering reads.



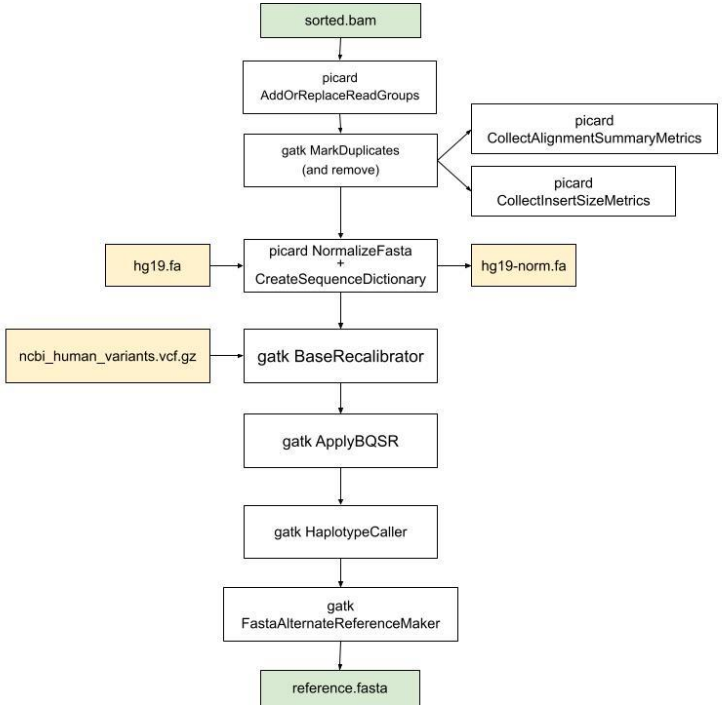
Dependence of the read quality score phred_score on the normalized read fragmentation score score_frag

Data filtering criteria:

1) *Value of the alignment score (from last tool), normalized to the length of the read, should be higher than 3*
2) *Value of the nucleotide quality parameter (phred score) should be higher than 10*
3) *Value of the read fragmentation parameter should be higher than 0.8*

Reference sequence compiled and data alignment performed

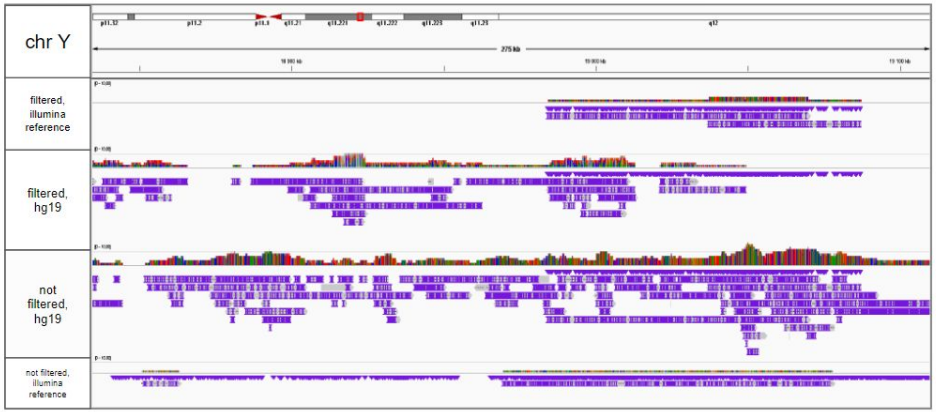
- A reference sequence assembly algorithm was applied based on Illumina data from one of the samples



- Alignment of data (initial and passed through the filtering step) for reference (standard assembly and collected according to illumina data)

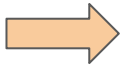
Sample	Mean coverage (hg19)	Mean coverage (illumina fasta)
fk4034	6,51	5,16
kz5281	5,45	4,24
mc5290	7,23	5,14
sk6940	6,18	3,27
uo8780	5,21	4,16
yo4642	6,34	5,29

Sample ***filtered***	Mean coverage (hg19)	Mean coverage (illumina fasta)
fk4034	1,21	1,13
kz5281	1,02	1,2
mc5290	1,12	1,05
sk6940	1,09	1,19
uo8780	0,86	0,78
yo4642	1,14	1,07



Conclusion

sample	mean phred score	mean read length	mean fragmentation score	% of b.p with good quality (Q>20)
fk4034	23,6	10122,65	0,93	55,5
kz5281	22,1	9813,88	0,88	58,0
mc5290	17,9	9191,1	0,84	54,5
sk6940	20,7	8156,07	0,8	39,7
uo8780	16,9	9625,83	0,91	38,3
yo4642	18,2	9038,37	0,86	52,8



- After applying the filtering criteria for all samples, ~ 55-65% of reads remained
- Average phred score is above 16
- The length of the read being mapped is at least 8000 bp.
- Data fragmentation score is above 0.8
- At least 38% of the selected nucleotides are of very high quality
- Using the filtering criteria and aligning the data to the reference build obtained from the illumin data, we cut off about 80% of the reads, leaving the best

Thank you for your attention!