Nanopore Sequenced Transcriptome Analysis Resource

NanoSTAR QC Report

Report created: 18 September 2021

Study design

This NanoSTAR report was generated automatically based on user input, reflected below.

The principles of reproducible research have been implemented, and a full list of software used (including version numbers) is included at the end of this document.

Table 1: Groups, sequence files, and md5 sum scores.

	File	Group	md5
A1	RawData/barcode01.fastq	Group A	e9e159e3971f9f763982171240f7ae85
A2	RawData/barcode02.fastq	Group A	f45d800194c22121840dcb253b98fd9e
A3	RawData/barcode03.fastq	Group A	3d663742aac25fbca6783fae93f880d3
A4	RawData/barcode04.fastq	Group A	50eeacde8fe39bbf0b315fa2424b1ce9
A5	RawData/barcode08.fastq	Group A	fe6df1a73099a57bd313d4c34b5a3b71
A6	RawData/barcode11.fastq	Group A	657cc7e63c84c7bbb8a534b418f1ca67
A7	RawData/2barcode05.fastq	Group A	9c37a209e5e6e4ff08727ad4178ebbb6
B1	RawData/barcode05.fastq	Group B	6c6510785c6cb1eb605d63041a0e4d5b
B2	RawData/barcode06.fastq	Group B	4412c9ed26b0df1867e6111d6ae9b4a4
B3	RawData/barcode10.fastq	Group B	4811b3f4acdcee4b6f4cb2609f69b0c5
B4	RawData/2barcode03.fastq	Group B	5d45750d1214e7dc8234a5d371e495aa
B5	RawData/2barcode06.fastq	Group B	062d50cbfded5dbd1befeef13cdc2648
B6	RawData/2barcode07.fastq	Group B	bc8e984c544e23340fb78420945b00be
B7	RawData/2barcode08.fastq	Group B	8cf5114c86c390095afc57654b2c317b
B8	RawData/2barcode10.fastq	Group B	b8664e3be186444edb9ea008fc66876f
B9	RawData/2barcode11.fastq	Group B	140e60c5e4f59e1400d9880e130616dc
C1	RawData/barcode07.fastq	Group C	1b01583fcbc63c5a6b07014509088de5
C2	RawData/barcode09.fastq	Group C	83bee191a1ae5b5315f4f07a3ce72ff3
C3	RawData/barcode12.fastq	Group C	32a0b527dac319fa6469b861f78794d9
C4	RawData/2barcode02.fastq	Group C	f1db901c049b795ed17b0a45eb2760d0
C5	RawData/2barcode01.fastq	Group C	cb1fb3137a5f47502b54e1c7424f450f
C6	RawData/2barcode04.fastq	Group C	b9dedec80587f6a58f5b19ddf6ae1733
C7	RawData/2barcode09.fastq	Group C	dfb5ca9b428ea9e93391ed9bd8de14cb
C8	RawData/2barcode12.fastq	Group C	dcf8cdd3a54cddc3c34e5437a171ec62

Violin plots of sequence characteristics

Review violin plots of sequence lengths and quality characteristics to assess the sample integrity and quality, and to ensure that the quality, depth and characteristics are suitable. Considerable variability within and between the experimental samples compromises the ability to identify differentially expressed genes and transcripts.

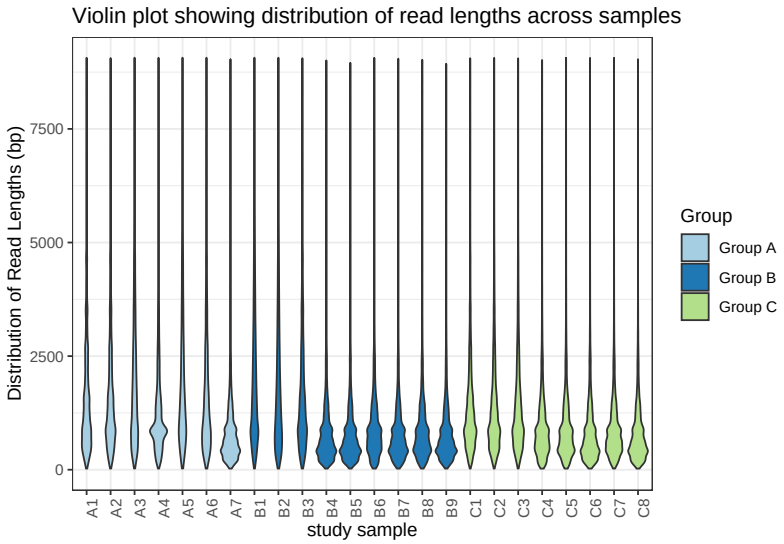

Figure 1: Violin plot of read lengths

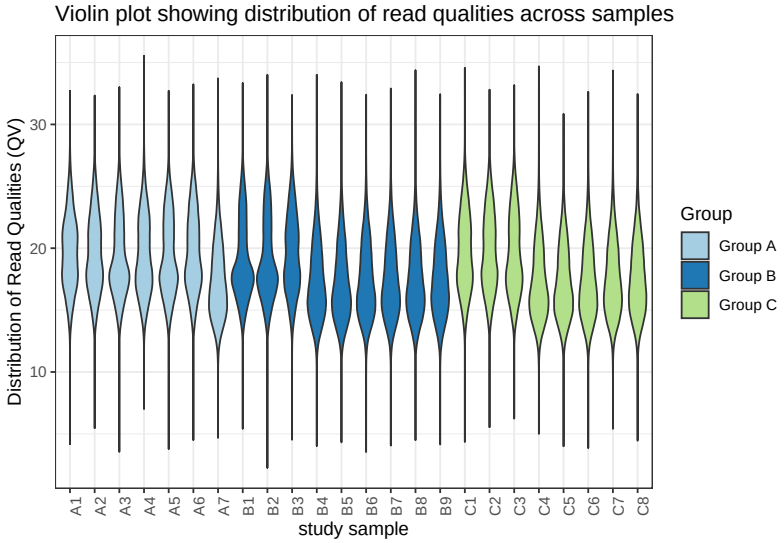

Figure 2: Violin plot of read quality scores

Review of raw cDNA sequences

The table below, detailing raw fastq files, has been produced using the R **ShortRead** package (Morgan, Lawrence, and Anders (2018)).

Table 2: Summary statistics from Samtools flagstat function.

	A1	A2	A3	A4	A5	A6	A7	B1	B2	B3	B4	B5	B6	B7	B8	B9	C1	C2	C3	C4	C5	C6	C7	C8
reads	137,997	117,275	143,506	113,816	68,871	67,271	101,685	178,058	184,998	73,827	90,802	49,792	48,070	61,425	39,413	72,656	120,195	109,618	72,976	32,893	32,992	73,496	53,323	49,338
mbs	278.6	233.2	400.2	214	273.9	182.2	92.6	500.9	532.6	168	79.8	41.3	48.8	55.3	37.2	63.2	180.4	175	121.7	34.1	34.3	67.3	52.6	43.8
min	25	26	27	25	27	25	25	25	25	25	25	25	25	25	25	25	25	25	26	25	25	25	25	25
max	57161	42519	48138	48913	55234	55482	28637	46922	72194	54889	26862	18982	23173	24486	23438	23017	30823	41419	39705	22030	26039	34606	44971	27833
mean	2018.7	1988.7	2788.7	1880.2	3976.6	2708.3	910.3	2812.9	2879.2	2275.2	879.3	830.2	1014.7	901	944.1	870.2	1500.9	1596.1	1667.8	1037.3	1039.4	916	986.8	888.7
median	1301	1336	1774	1036	1996	1446	660	1811	1702	1390	637	608	802	668	727	637	1097	1080	1147	799	769	650	753	641
qval	19.7	19.5	19.3	19.4	19.4	19.6	17.3	19	19.1	19.5	17.3	17.1	17.3	17.4	17.4	17.5	19.7	19.7	19.5	17.2	17.2	17.3	17.5	17.5
gc	53.9	52.9	47.4	52.4	50.3	50.2	53.9	45.3	45.5	53.5	54.2	54	52.6	53.3	53.1	53.7	53.1	52.5	51.6	53.1	53.6	53.7	53.4	53.2
n50	3037	2945	4454	3063	8576	4868	1250	4476	4911	3567	1220	1126	1335	1210	1261	1176	2010	2277	2346	1382	1408	1298	1286	1240
l50	23822	22163	24099	17291	9395	8781	20992	31387	28670	12030	18647	10488	10946	12974	8741	15432	25850	21106	14309	7144	6743	14558	11767	10217
n90	924	929	1312	831	1672	1157	449	1303	1318	1016	433	417	547	451	485	431	775	782	817	545	516	443	514	436
l90	88123	77234	88234	74117	38387	39156	69041	111305	109788	46560	61303	34083	33269	41993	27214	50040	83632	74105	49540	22499	22364	48985	37218	33345

Review of cDNA read mapping

These mapping statistics were produced by **samtools flagstat** (Li et al. (2009)), written to files in the `Analysis/samtools` folder.

Table 3: Summary statistics from the minimap2 long read mapping of reads to the reference transcriptome

	A1	A2	A3	A4	A5	A6	A7	B1	B2	B3	B4	B5	B6	B7	B8	B9	C1	C2	C3	C4	C5	C6	C7	C8
nreads	137997	117275	143506	113816	68871	67271	101685	178058	184998	73827	90802	49792	48070	61425	39413	72656	120195	109618	72976	32893	32992	73496	53323	49338
Read mappings	354292	317994	550031	303719	368224	255594	167689	714770	716560	222112	145273	78672	83833	102916	67438	120609	266534	266219	178686	56845	56127	121957	91078	81942
Secondary	178534	167712	339446	158471	243489	154619	55904	451574	443151	122334	45660	24673	30264	35394	23572	40674	121310	129588	86886	20148	19657	41296	31883	27613
Supplementary	37761	33007	67079	31432	55864	33704	10100	85138	88411	25951	8811	4207	5499	6097	4453	7279	25029	27013	18824	3804	3478	7165	5872	4991
Duplicates	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
%mapping	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

The read mapping statistics shown in the table above are derived from the **bam** files produced by **minimap2** (Li (2018)). **Read mappings** corresponds to the number of unique read mappings. **Secondary** corresponds to the number of secondary alignment for a read that can be mapped to multiple locations. **Supplementary** alignments could correspond to chimeric alignments, or represent a structural variant or complex fusion event. **Duplicate** reads are unlikely but would represent multiple *identical* reads at the same locus that are likely the result of an amplification during the library preparation. **%Mapping** corresponds to the percentage fraction of sequence reads that are mapped to the reference genome - this is calculated as the number of primary mappings against the total number of sequence reads observed in the starting fastq files.

Reproducible research

This analysis used publicly available **Linux** software, which are listed below with their version numbers.

```
# packages in environment at /home/chris/miniconda3:
#
# Name                    Version                   Build  Channel
bioconductor-annotationdbi 1.54.0           r41hdfd78af_0    bioconda
bioconductor-deseq2       1.32.0            r41h399db7b_0    bioconda
bioconductor-dexseq       1.38.0            r41hdfd78af_0    bioconda
bioconductor-drimseq      1.20.0            r41hdfd78af_0    bioconda
bioconductor-edger        3.34.0            r41h399db7b_0    bioconda
bioconductor-genomicfeatures 1.44.0           r41hdfd78af_0    bioconda
bioconductor-pcamethods   1.84.0            r41h399db7b_0    bioconda
bioconductor-shortread    1.50.0            r41h399db7b_0    bioconda
bioconductor-stager       1.14.0            r41hdfd78af_0    bioconda
filtlong                  0.2.1                h9a82719_0    bioconda
porechop                  0.2.4            py39h7cff6ad_2    bioconda
r-devtools                2.4.2             r41hc72bb7e_0    conda-forge
r-digest                  0.6.27            r41h03ef668_0    conda-forge
r-dplyr                   1.0.7             r41h03ef668_0    conda-forge
r-ggplot2                 3.3.5             r41hc72bb7e_0    conda-forge
r-gplots                  3.1.1             r41hc72bb7e_0    conda-forge
r-gridextra               2.3             r41hc72bb7e_1003    conda-forge
r-kableextra              1.3.4             r41hc72bb7e_0    conda-forge
r-pheatmap                1.0.12            r41hc72bb7e_2    conda-forge
r-plotrix                 3.8_2             r41hc72bb7e_0    conda-forge
r-reshape2                1.4.4             r41h03ef668_1    conda-forge
r-rstudioapi              0.13              r41hc72bb7e_0    conda-forge
r-tidyr                   1.1.3             r41h03ef668_0    conda-forge
r-tidyverse               1.3.1             r41hc72bb7e_0    conda-forge
r-viridis                 0.6.1             r41hc72bb7e_1    conda-forge
r-viridislite             0.4.0             r41hc72bb7e_0    conda-forge
r-writexl                 1.4.0             r41hcfec24a_0    conda-forge
r-yaml                    2.2.1             r41hcfec24a_1    conda-forge
Parsing /var/lib/dpkg/status... completed.
apt-show-versions:all/focal 0.22.11 uptodate
minimap2:amd64/focal 2.17+dfsg-2 uptodate
pandoc:amd64/focal 2.5-3build2 uptodate
salmon:amd64/focal 0.12.0+ds1-1 uptodate
samtools:amd64/focal 1.10-3 uptodate
texlive-fonts-recommended:all/focal 2019.20200218-1 uptodate
texlive-latex-base:all/focal 2019.20200218-1 uptodate
texlive-latex-extra:all/focal 2019.202000218-1 uptodate
texlive-latex-recommended:all/focal 2019.20200218-1 uptodate
```

This report has been created for reproducibility, using **Rmarkdown**, publicly available **R** packages, and the **LaTeX** document typesetting software. Packages and their version numbers are listed below.

```
R version 4.1.1 (2021-08-10)
Platform: x86_64-conda-linux-gnu (64-bit)
Running under: Ubuntu 20.04.3 LTS

Matrix products: default
BLAS/LAPACK: /home/chris/miniconda3/lib/libopenblasp-r0.3.17.so

attached base packages:
[1] stats4    parallel  stats     graphics  grDevices utils     datasets  methods   base

loaded via a namespace (and not attached):
 [1] bitops_1.0-7           fs_1.5.0               RColorBrewer_1.1-2     webshot_0.5.2
 [5] httr_1.4.2             rprojroot_2.0.2        tools_4.1.1            utf8_1.2.2
 [9] R6_2.5.1               DBI_1.1.1              colorspace_2.0-2       withr_2.4.2
[13] tidyselect_1.1.1       prettyunits_1.1.1      processx_3.5.2         compiler_4.1.1
[17] cli_3.0.1              rvest_1.0.1            xml2_1.3.2             DelayedArray_0.18.0
[21] desc_1.3.0             labeling_0.4.2         scales_1.1.1           callr_3.7.0
[25] systemfonts_1.0.2      stringr_1.4.0          rmarkdown_2.10         svglite_2.0.0
[29] jpeg_0.1-9             pkgconfig_2.0.3        htmltools_0.5.2        sessioninfo_1.1.1
[33] fastmap_1.1.0          rlang_0.4.11           rstudioapi_0.13        farver_2.1.0
[37] generics_0.1.0         hwriter_1.3.2          RCurl_1.98-1.4         magrittr_2.0.1
[41] GenomeInfoDbData_1.2.6 Matrix_1.3-4           Rcpp_1.0.7             munsell_0.5.0
[45] fansi_0.5.0            lifecycle_1.0.0        stringi_1.7.4          zlibbioc_1.38.0
[49] pkgbuild_1.2.0         plyr_1.8.6             grid_4.1.1             crayon_1.4.1
[53] lattice_0.20-44        locfit_1.5-9.4         knitr_1.34             ps_1.6.0
[57] pillar_1.6.2           pkgload_1.2.2          glue_1.4.2             evaluate_0.14
[61] latticeExtra_0.6-29    remotes_2.4.0          png_0.1-7              vctrs_0.3.8
[65] testthat_3.0.4         gtable_0.3.0           purrr_0.3.4            assertthat_0.2.1
[69] cachem_1.0.6           xfun_0.25              viridisLite_0.4.0      tibble_3.1.4
[73] memoise_2.0.0          ellipsis_0.3.2
```

The session data produced in the production of this report, which can be used for further analysis of the dataset, can be found here:

```
Analysis/Results/NanoSTAR_QC_Report.Rdata
```

References and citations

Li, Heng. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." *Bioinformatics* 34 (18): 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Morgan, Martin, Michael Lawrence, and Simon Anders. 2018. *ShortRead: FASTQ Input and Manipulation.*