

# GigaScience

## SOAPnuke: A MapReduce Acceleration supported Software for integrated Quality Control and Preprocessing of High-Throughput Sequencing Data --Manuscript Draft--

<b>Manuscript Number:</b>	GIGA-D-17-00173R1	
<b>Full Title:</b>	SOAPnuke: A MapReduce Acceleration supported Software for integrated Quality Control and Preprocessing of High-Throughput Sequencing Data	
<b>Article Type:</b>	Technical Note	
<b>Funding Information:</b>	Science & Technology Bureau of Fujian Province (2013YZ0002-2)	Mr. Lin Fang
	the Joint Project of Natural Science and health Foundation of Fujian Province (2015J01397)	Ms. Chunmei Shi
<b>Abstract:</b>	<p>Quality Control (QC) and preprocessing are essential steps for sequencing data analysis to ensure the accuracy of results. However, existing tools cannot provide a satisfying solution with integrated comprehensive functions, proper architectures and highly-scalable acceleration. In this article, we demonstrate SOAPnuke as a tool with abundant functions for a 'QC-Preprocess-QC' workflow and MapReduce acceleration framework. Four modules with different preprocessing functions are designed for processing datasets from genomic, small RNA (sRNA), Digital Gene Expression (DGE) and metagenomic experiments respectively. As a workflow-like tool, SOAPnuke centralizes processing functions in one executable and predefine their order to avoid the necessity of reformatting different files when switching tools. Furthermore, the MapReduce framework enables large scalability to distribute all the processing works to an entire compute cluster.</p> <p>We conducted a benchmarking where SOAPnuke and other tools are used to preprocess ~30x NA12878 dataset published by GIAB. The standalone operation of SOAPnuke struck a balance between resource occupancy and performance. When accelerated on 16 working nodes with MapReduce, SOAPnuke achieved ~5.7 times of the fastest speed of other tools.</p>	
<b>Corresponding Author:</b>	Lin Fang  CHINA	
<b>Corresponding Author Secondary Information:</b>		
<b>Corresponding Author's Institution:</b>		
<b>Corresponding Author's Secondary Institution:</b>		
<b>First Author:</b>	Yuxin Chen	
<b>First Author Secondary Information:</b>		
<b>Order of Authors:</b>	Yuxin Chen	
	Yongsheng Chen	
	Chunmei Shi	
	Zhibo Huang	
	Yong Zhang	
	Shengkang Li	
	Yan Li	
	Jia Ye	
	Chang Yu	

	Zhuo Li
	Xiuqing Zhang
	Jian Wang
	Huanming Yang
	Lin Fang
	Qiang Chen
<b>Order of Authors Secondary Information:</b>	
<b>Response to Reviewers:</b>	<p>Dear Reviewer:</p> <p>Thank you for your letter. We were pleased to know that our manuscript was rated as potentially acceptable for publication in GigaScience, subject to adequate revision and response to your comments.</p> <p>We revised the manuscript following the instructions provided in your letter. To address the first question, two sentences has been added to the DISCUSSION section: 'To users without computing cluster, SOAPnuke might not be an optimal tool in terms of overall performance. Thus, we are performing refactoring to increase the standalone processing speed'. As suggested in the second question, Variant calling results and corresponding description have been added to RESULT section. We also greatly appreciate the suggestions offered in third question and have changed the wording and bottom-half content. In terms of language editing, we have fixed the problems you indicated and re-collated the whole manuscript.</p> <p>We would like to express our sincere gratitude to you for improving the quality of our manuscript with helpful suggestions.</p> <p>Thank you.</p> <p>Best Regards, Magic Fang</p>
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
Are you submitting this manuscript to a special series or article collection?	No
<b>Experimental design and statistics</b>	Yes
<p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	
<b>Resources</b>	Yes
A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely	

<p>identified, should be included in the Methods section. Authors are strongly encouraged to cite <a href="#">Research Resource Identifiers</a> (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	
<p><b>Availability of data and materials</b></p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in <a href="#">publicly available repositories</a> (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.</p> <p>Have you have met the above requirement as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	<p>Yes</p>

1  
2  
3  
4 1 *Technical Note:*

5  
6  
7 2 **SOAPnuke: A MapReduce Acceleration supported Software for**  
8 3 **integrated Quality Control and Preprocessing of**  
9 4 **High-Throughput Sequencing Data**

10 5 *Yuxin Chen<sup>1†</sup>, Yongsheng Chen<sup>2†</sup>, Chunmei Shi<sup>3,4,5†</sup>, Zhibo Huang<sup>1</sup>, Yong Zhang<sup>1,6</sup>, Shengkang*  
11 6 *Li<sup>1,6</sup>, Yan Li<sup>1</sup>, Jia Ye<sup>1</sup>, Chang Yu<sup>7</sup>, Zhuo Li<sup>8,9</sup>, Xiuqing Zhang<sup>1</sup>, Jian Wang<sup>1,10</sup>, Huanming Yang<sup>1,10</sup>,*  
12 7 *Lin Fang<sup>1,6\*</sup> and Qiang Chen<sup>3,4,5\*\*</sup>*

13  
14  
15  
16  
17  
18 8 *1 BGI-Shenzhen, Shenzhen 518083; 2 Geneplus-Beijing, Beijing 102206; 3 Department of Oncology,*  
19 9 *Fujian Medical University Union Hospital, Fuzhou 350001; 4 Fujian Key Laboratory of Translational*  
20 10 *Cancer Medicine, Fuzhou 350014; 5 Department of Stem Cell Research Institute, Fujian Medical*  
21 11 *University Stem Cell Research Institute, Fuzhou 350000; 6 Collaborative Innovation Center of High*  
22 12 *Performance Computing, National University of Defense Technology, Changsha 410073; 7 Intel China*  
23 13 *Ltd., Shanghai 200336; 8 Guangdong Provincial Hospital of Chinese Medicine, Guangzhou 510120; 9*  
24 14 *Department of Surgery, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong; 10*  
25 15 *James D. Watson Institute of Genome Sciences, Hangzhou 310058, China*

26  
27 16 Yuxin C [chenyuxin@genomics.cn](mailto:chenyuxin@genomics.cn), ORCID: [0000-0002-9246-1829](https://orcid.org/0000-0002-9246-1829); Yongsheng C  
28 17 [chenys@geneplus.org.cn](mailto:chenys@geneplus.org.cn); CS [scmfz@qq.com](mailto:scmfz@qq.com); ZH [huangzhibo@genomics.cn](mailto:huangzhibo@genomics.cn), ORCID:  
29 18 [0000-0002-2750-6517](https://orcid.org/0000-0002-2750-6517); YZ [zhangyong2@genomics.cn](mailto:zhangyong2@genomics.cn); SL [lishengkang@genomics.cn](mailto:lishengkang@genomics.cn), ORCID:  
30 19 [0000-0002-6864-5644](https://orcid.org/0000-0002-6864-5644); YL [liyana@genomics.cn](mailto:liyana@genomics.cn); JY [yejia@genomics.cn](mailto:yejia@genomics.cn); CY [chang.yu@intel.com](mailto:chang.yu@intel.com); ZL  
31 20 [lzgdphcm@163.com](mailto:lzgdphcm@163.com); XZ [zhangxq@genomics.cn](mailto:zhangxq@genomics.cn); JW [wangjian@genomics.cn](mailto:wangjian@genomics.cn); HY [yanghm@genomics.cn](mailto:yanghm@genomics.cn),  
32 21 ORCID: [0000-0002-0858-3410](https://orcid.org/0000-0002-0858-3410); LF [fangl@genomics.cn](mailto:fangl@genomics.cn), ORCID: [0000-0002-5954-3435](https://orcid.org/0000-0002-5954-3435); QC  
33 22 [cqiang8@189.cn](mailto:cqiang8@189.cn)

34  
35  
36  
37  
38 23 †Contributed equally

39 24 \*First corresponding author

40 25 \*\*Second corresponding author

41  
42  
43  
44 26 **ABSTRACT**

45  
46  
47  
48 27 Quality Control (QC) and preprocessing are essential steps for sequencing data analysis to  
49 28 ensure the accuracy of results. However, existing tools cannot provide a satisfying solution with  
50 29 integrated comprehensive functions, proper architectures and highly-scalable acceleration. In  
51 30 this article, we demonstrate SOAPnuke as a tool with abundant functions for a  
52 31 ‘QC-Preprocess-QC’ workflow and MapReduce acceleration framework. Four modules with  
53 32 different preprocessing functions are designed for processing datasets from genomic, small RNA  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

33 (sRNA), Digital Gene Expression (DGE) and metagenomic experiments respectively. As a  
34 workflow-like tool, SOAPnuke centralizes processing functions in one executable and predefine  
35 their order to avoid the necessity of reformatting different files when switching tools.  
36 Furthermore, the MapReduce framework enables large scalability to distribute all the processing  
37 works to an entire compute cluster.

38 We conducted a benchmarking where SOAPnuke and other tools are used to preprocess ~30x  
39 NA12878 dataset published by GIAB. The standalone operation of SOAPnuke struck a balance  
40 between resource occupancy and performance. When accelerated on 16 working nodes with  
41 MapReduce, SOAPnuke achieved ~5.7 times of the fastest speed of other tools.

42 **KEYWORDS:** High-throughput sequencing, Quality control, Preprocessing,  
43 MapReduce

44 **BACKGROUND**

45 High-throughput sequencing (HTS) instruments have enabled many large-scale studies and  
46 generated enormous amount of data [1-3]. However, the presence of low-quality bases, sequence  
47 artifacts and sequence contamination can introduce serious negative impact on downstream  
48 analyses. Thus, QC and preprocessing of raw data serve as the critical steps to initiate analysis  
49 pipelines [4, 5]. QC investigates several statistics of datasets to ensure data quality, and  
50 preprocessing trims off undesirable terminal fragments and filters out substandard reads [6].

51 We have conducted a survey on existing 31 tools and widely shared functions are listed in  
52 Supplementary Material 1.

53 Existing tools for QC and preprocessing can be divided into two categories according to their  
54 structures: toolkit and workflow. Toolkit-like software provides multiple executables such as  
55 statistics computer, clipper and filtrator [7-15]. In practice, raw data is processed by a few

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

56 individual executables in sequence. Comparatively, workflow-like software offers an integral  
57 workflow where functions are performed in predefined order [6, 16-27].

58 However, both categories have their own demerits. When using toolkit-like software, it is  
59 complex and error-prone to write additional scripts to wrap executables. Moreover, it consumes  
60 much time to generate and read intermediate files, which is hard for acceleration. Besides, the  
61 same variables could possibly be computed repetitively. For instance, average quality score of  
62 each read is necessary for counting quality score distribution by reads, and filtering reads based  
63 on average quality scores. It has to be counted twice if these two functions are implemented by  
64 different toolkits.

65 For workflow-like tools, an optimal architecture is required since the orders of functions are  
66 fixed. Most of existing tools successively perform QC and preprocessing without complete  
67 statistics of preprocessed datasets. If the preprocessing operation is not suitable for a given  
68 dataset, the problem can only be revealed by downstream analyses.

69 Datasets sequenced from various samples may require different processing functions or  
70 parameters. Existing workflow-like tools mostly support genomics data processing, only a few of  
71 them are developed for other types of studies, such as RNA-seq and metagenomics data. For  
72 example, RObiNA [22] provides four modules for different RNA sequencing experiments.  
73 PrinSeq [6] offers a QC stat, dinucleotide odds ratios, to show how the dataset might be related  
74 to other viral/microbial metagenomes. However, there is still no single tool supporting multiple  
75 data types.

76 Several tools have made certain progress in overcoming the limitations mentioned above.  
77 Galaxy [37] is a web-based platform incorporating various existing toolkit-like software. Users  
78 can conveniently concatenate tools into a pipeline on the web interface. NGS QC toolkit [16]

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

79 offers a workflow with QC on both raw and preprocessed datasets, though the preprocessing  
80 functions are too few.

81 In terms of software acceleration, only multi-threading is adopted by existing tools [14-16,  
82 24-28]. This approach only works for standalone operation and is limited by the maximum  
83 number of processors in one computer server. It may be incompetent when dealing with huge  
84 present and potential volume of sequencing datasets.

85 To solve these problems, we have developed a workflow-like tool, SOAPnuke, for integrated QC  
86 and preprocessing of large HTS datasets. Similar to NGS QC toolkit, SOAPnuke performs  
87 two-step QC. Trimming, filtering and other frequently used functions are integrated in our  
88 program. Four modules are designed to handle genomic, metagenomic, DGE and sRNA datasets  
89 respectively. In addition, SOAPnuke is extended to multiple working nodes for parallel  
90 computing using Hadoop MapReduce framework.

91 **METHODS**

92 **QC & PREPROCESSING**

93 SOAPnuke (SOAPnuke, RRID:SCR\_015025) was developed to summarize statistics of both raw  
94 and preprocessed data. Basic statistics are comprised of the number of sequences and bases,  
95 base composition, Q20 and Q30, and filtering information. Complex statistics include the  
96 distribution of quality score and base composition distribution for each position. For the quality  
97 score distribution, Q20 and Q30 for each position are plotted in line chart and the quantiles of  
98 the quality are represented in a boxplot. And for the base composition distribution, an  
99 overlapping histogram is used to display base composition distribution for each position. These  
100 calculations are conducted by C++ and the plots are generated by R 3.3.2 [38]. An example of  
101 the two plots are shown in Fig.1. A comprehensive list of statistics available in SOAPnuke is

1  
2  
3  
4 102 included in Additional File 2. Statistics of preprocessed data are compared with some preset  
5  
6 103 thresholds. A warning message will be issued if median score of any position in per-base quality  
7  
8 104 distribution is lower than 25 and a failure will be issued if lower than 20. For per-base base  
9  
10 105 composition, a warning will be raised if difference between A and T, or G and C in any position  
11  
12 106 is greater than 10% or a failure will be issued if greater than 20%.

13  
14  
15  
16 107 Fig.1. An example of QC complex statistics. (a) per-base quality distribution of raw paired-end  
17  
18 108 reads. (b) per-base Q20 and Q30 of raw and preprocessed paired-end reads. (c) per-base base  
19  
20 109 composition distribution of raw paired-end reads.

21  
22  
23  
24 110 In the step of preprocessing, those undesirable terminal fragments are trimmed off, substandard  
25  
26 111 reads are filtered out, and certain transform operations are applied. On both ends of reads,  
27  
28 112 bases of assigned number or of quality lower than threshold will be trimmed off. Sequencing  
29  
30 113 adapters can be aligned, where mismatch is supported while no INDEL is tolerated, and cut to 3'  
31  
32 114 end. Filtering can be performed on reads with adapter, short length, too many ambiguous bases,  
33  
34 115 low average quality or too many low-quality bases. The sequencing batches, such as tile of  
35  
36 116 Illumina sequencer[39] and fov (field of view) of BGI sequencer[40], with unfavorable  
37  
38 117 sequencing quality can be assigned so the corresponding sequences will be discarded. In  
39  
40 118 addition, reads with identical nucleotides can be deduplicated to keep only one copy.  
41  
42 119 Transformation comprises quality system conversion, interconversion between DNA and RNA,  
43  
44 120 and compression of output with gzip, etc. Additional File 3 lists the above preprocessing  
45  
46 121 functions and their parameters.

## 47 48 49 50 51 122 MODULES DESIGN

52  
53  
54  
55 123 To improve processing performance of different types of data, four modules are specialized in  
56  
57 124 SOAPnuke, including General, DGE, sRNA and Meta modules. (1) General module can handle  
58  
59 125 most of the DNA re-sequencing datasets, as described in the section of QC & PROCESSING. (2)

60  
61  
62  
63  
64  
65



1  
2  
3  
4 126 DGE Profiling generates single-end read which has a 'CATG' segment neighboring targeted  
5  
6 127 sequences of 17 base pairs[41]. By default, DGE module will find the targeted segment and trim  
7  
8 128 off other parts. Moreover, reads with ambiguous bases will be filtered. (3) sRNA module  
9  
10 129 incorporates filtering of poly-A tags as polyadenylation is a feature of mRNA data and sRNA  
11  
12 130 sequences can be contaminated by mRNA during sample preparation[42]. (4) Metagenomics  
13  
14 131 preprocessing module customizes a few functions from General module for trimming adapters  
15  
16 132 and low-quality bases on both ends, dropping reads with too short length or too many  
17  
18 133 ambiguous bases. Detailed parameters settings can be accessed in Additional File 3.  
19  
20  
21  
22

## 23 134 SOFTWARE FEATURES

24  
25  
26 135 SOAPnuke is written by C++ for good scalability and performance and it can be run on both  
27  
28 136 Linux and Windows platforms.  
29  
30  
31

32 137 Two paralleled strategies are implemented for acceleration. Multi-threading is developed for  
33  
34 138 standalone operation. Data is cut into blocks of fixed size, and each block is processed by one  
35  
36 139 thread. This design utilizes multiple cores in a working node. In SOAPnuke, the creation and  
37  
38 140 allocation of threads are managed by threadpool library, which decreases the overhead of  
39  
40 141 creating and destroying threads. More importantly, Hadoop MapReduce is applied to achieve  
41  
42 142 rapid processing in multi-node cluster for ultra-large-scale data. In the mapping phase, each  
43  
44 143 read is kept as a key-value pair, where key is readID and value is sequence and quality scores. In  
45  
46 144 shuffle phase, the key-value pairs are sorted, and each pair of paired-end reads is gathered.  
47  
48 145 During the reducing phase, blocks of fixed size are processed by various threads of multiple  
49  
50 146 nodes, and each block generates an individual result. After that, it is optional to merge the  
51  
52 147 results into integrated fastq file(s).  
53  
54  
55  
56

57 148 To prove the effectiveness of the acceleration design, we have conducted a performance tests on  
58  
59 149 SOAPnuke and other alternative tools. A ~30x human genome dataset published by GIAB [43]  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 150 were extracted as testing data(see Addition File 4) . In terms of computing environment, up to  
5  
6 151 16 nodes were used, each of which has 24 cores of Intel(R) Xeon(R) CPU E5-2620 v4 @  
7  
8 152 2.10GHz and RAM of 128G. SOAPnuke operations for testing were set as described in published  
9  
10 153 manuscripts(see the reference list in Additional File 5). Trimming adapters and filtering on  
11  
12 154 length and quality were selected for their universality. We chose other workflow-like tools  
13  
14 155 capable of performing these functions, which are Trimmomatic (Trimmomatic,  
15  
16 156 RRID:SCR\_011848)[27], AfterQC [30], BBDuk [31] and AlignTrimmer [36]. The parameter  
17  
18 157 setting is also available in Addition File 4.  
19  
20  
21  
22

## 23 158 **RESULTS**

26 159 In the performance test, we chose three indexes for evaluation: elapsed time, CPU usage and  
27  
28 160 maximum RAM usage. As shown in Table.1, AfterQC is the tool occupying least resources.  
29  
30 161 However, its processing time is too long for practical usage, especially considering we ran the  
31  
32 162 program with pypy, which is announced to be 3 times as fast as standard python. Among the  
33  
34 163 remaining tools, SOAPnuke struck an appropriate balance between resources occupancy and  
35  
36 164 performance. Furthermore, users can choose to run SOAPnuke on multiple nodes with  
37  
38 165 MapReduce framework if high throughput performance is demanded. In our testing, 16 nodes  
39  
40 166 can achieve ~32 times acceleration compared to standalone operation, which is 5.37 times faster  
41  
42 167 than the highest speed of four tested tools.  
43  
44  
45  
46

Tools \ Indexes	Time (min)	Throughput (read/s)	CPU	Max RAM ( GB )
SOAPnuke (1 node 1 thread)	302.7	33947.8	250%	0.62
SOAPnuke (16 nodes)	9.4	1093191.1	640%	50.10
Trimmomatic (1 thread)	84.7	121380.1	75%	2.98
Trimmomatic (24 threads)	50.5	203582.1	239%	10.28
BBDuk	57.2	162230.2	259%	11.40

61  
62  
63  
64  
65

AlienTrimmer	530.2	19076.1	99%	0.54
AfterQC (pypy)	2482.7	4319.1	99%	0.21

168 Table.1. Evaluation of the data processing performance across SOAPnuke and four other tools.  
169 Time, throughput, CPU and maximum memory occupation are presented. For CPU usage, 100%  
170 means full load of a single CPU core. Maximum RAM usage means the highest occupancy of  
171 RAM during the whole processing.

172 After the preprocessing, downstream analyses were performed with GATK (GATK,  
173 RRID:SCR\_001876) best practice pipeline (see the description of GATK best practices [44]).  
174 Data was processed by alignment, rmDup, baseRecal, bamSort and haplotypeCaller modules in  
175 order. For the haplotypeCaller, GIAB high-confidence small variant and reference calls v3.3.2  
176 [45] were used as gold standard. Details of this testing is available in Additional File 4.

Tools \ Indexes	SNPs Precision	SNPs Sensitivity	SNPs F-measure	INDELs Precision	INDELs Sensitivity	INDELs F-measure
SOAPnuke	0.9967	0.9811	0.9888	0.9806	0.9575	0.9689
Trimmomatic	0.9966	0.9811	0.9888	0.9806	0.9575	0.9689
BBDuk	0.9966	0.9797	0.9881	0.9698	0.9184	0.9434
AlienTrimmer	0.9954	0.9810	0.9882	0.9792	0.9540	0.9665
AfterQC	0.9968	0.9811	0.9889	0.9811	0.9586	0.9697

177 Table.2. Variant calling result of SOAPnuke and other 4 tools. F-measure is a measure  
178 considering both the precision and recall of the variant calling result. SNP and INDEL are two  
179 main categories of variants.

180 As seen in the Table.2, AfterQC achieves best variant calling result. The F-measures of  
181 SOAPnuke and Trimmomatic are the same, which are slightly lower than that of AfterQC.  
182 AlienTrimmer performs slightly worse, and BBDuk has the worst result whose INDEL calling  
183 result differs greatly from that of other tools. In summary, though the variant calling result of  
184 AfterQC is optimal, it is not worth considering for its long processing time. Among the  
185 remaining tools, SOAPnuke and Trimmomatic tie for first place.

## DISCUSSION AND CONCLUSION

Data quality is critical to downstream analysis, which makes it important to use reliable tools for preprocessing. To omit unnecessary input/output and computation, workflow-like structure is adopted in SOAPnuke, where QC and preprocessing functions are integrated within an executable program. Compared to most of workflow-like tools, such as PrinSeq [6] and RObiNA [26], SOAPnuke adds statistics of preprocessed data for better understanding of data. To cope with datasets generated from different experiments, four modules are predefined with tailored functions and parameters. In terms of acceleration approach, multi-threading is the sole method adopted by existing tools [14-16, 24-28] but only applicable to single-node operations.

SOAPnuke utilizes MapReduce to realize concurrent execution on multi-node operations, where CPU cores of multiple nodes can be involved in a single task. It improves the scalability of parallel execution and the applicability to mass data. SOAPnuke also include multi-threading for standalone computing. Our test results indicate that SOAPnuke can achieve ~5.37 times faster than the maximum speed of other tools with multi-threading. It is worth mentioning that processing speed is not directly proportional to the number of working nodes, because some procedures like initialization of MapReduce cannot be accelerated as nodes increase, and the burden of communication between nodes aggravates as well.

For the future works, we will continue adding functions to feature modules. For example, in preprocessing of DGE datasets, filtering out singleton reads is frequently included [46-48]. For sRNA module, screening out reads based on alignment with noncoding RNA databases (such as tRNA, rRNA and snoRNA) [49,50] is under development. It is also considerable to add statistics such as per-read quality distribution and length distribution. To users without computing cluster, SOAPnuke might not be an optimal tool in terms of overall performance. Thus, we are performing refactoring to increase the standalone processing speed.

1  
2  
3  
4 210 However, we have found two problems worth exploring regarding QC and preprocessing. Firstly,  
5  
6 211 in terms of preprocessing, it is difficult to choose optimal parameters for a specific dataset.  
7  
8 212 Datasets from the same experiments and sequencers tend to share features, so users always  
9  
10 213 select the same parameters for those similar data. The parameters are initially defined based on  
11  
12 214 experiments on a specific dataset or just experience, which may already introduces some error  
13  
14 215 and bias. Moreover, even if the parameters are optimal for the tested dataset, they are possibly  
15  
16 216 inappropriate for other data because of random factors. Thus, the current method is a  
17  
18 217 compromise. However, it might be a considerable solution that preprocessing settings are  
19  
20 218 automatically adjusted during the processing. Secondly, some of the QC statistics are of limited  
21  
22 219 help to judge the availability of data. For example, as the threshold of filtering out low-quality  
23  
24 220 reads is increased from 0 to 40, the mean quality of all reads or each position will rise  
25  
26 221 accordingly, and the result of variant calling will be improved at the very beginning but then gets  
27  
28 222 worse. It is because preprocessing is a procedure required to strike a balance between removing  
29  
30 223 noise and keeping useful information, while single QC statistics cannot reflect the global balance.  
31  
32 224 A comprehensive list of QC statistics in SOAPnuke can help solve the problem since raising the  
33  
34 225 threshold of mean quality after the balance alone might make other irrelevant statistics worse.  
35  
36 226 Thus, it is worthwhile to explore ways to comprehensively analyze all statistics to evaluate the  
37  
38 227 effect of preprocessing. Currently, this procedure is performed empirically by users. In our  
39  
40 228 future work, these two problems will be considered for the development of updated versions.  
41  
42  
43  
44  
45  
46

## 47 229 **Availability and requirements**

48  
49  
50 230 Project name: SOAPnuke

51  
52  
53 231 Project home page: <https://github.com/BGI-flexlab/SOAPnuke>

54  
55  
56 232 RRID: SCR\_015025  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

233 Operating system(s): Linux, Windows

234 Programming language: C++

235 Requirements: libraries: boost, zlib, log4cplus and openssl; R

236 License: GPL

237

238 **Availability of supporting data**

239 Snapshots of the code and test data are also stored in the *GigaScience* repository, GigaDB [51].

240

241 **Abbreviations**

242 QC, quality control; HTS, high throughput sequencing; DGE, digital gene expression; sRNA,  
243 small RNA

244 **Declarations**

245 **ACKNOWLEDGEMENTS**

246 This research was supported by Collaborative Innovation Center of High Performance

247 Computing, Critical Patented Project of the Science & Technology Bureau of Fujian Province,

248 China (Grant No. 2013YZ0002-2) and the Joint Project of Natural Science and health

249 Foundation of Fujian Province, China (Grant No.2015J01397).

250 **AUTHORS' CONTRIBUTIONS**

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

251 LF and QC conceived the project. Yuxin C and CS conducted the survey on existing tools for QC  
252 and preprocessing. Yuxin C, Yongsheng C, CS, ZH, YZ, SL, JY, ZL, XZ, JW, HY, LF, QC provided  
253 feedback on features and functionality. YongSheng C, ZH and SL wrote the standalone version  
254 of SOAPnuke. Yuxin C wrote the MapReduce version of SOAPnuke. Yuxin C and ZH performed  
255 the above-mentioned test. Yuxin C, YL, CY and LF wrote the manuscript. All authors read and  
256 approved the final manuscript.

## 257 COMPETING INTERESTS

258 The authors declare that they have no competing interests.

## 259 OPEN ACCESS

260 This article is distributed under the terms of the Creative Commons Attribution 4.0  
261 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits  
262 unrestricted use, distribution, and reproduction in any medium, provided you give appropriate  
263 credit to the original author(s) and the source, provide a link to the Creative Commons license,  
264 and indicate if changes were made. The Creative Commons Public Domain Dedication waiver  
265 (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in  
266 this article, unless otherwise stated.

## 267 ADDITIONAL FILES

268 Supplementary Material 1: Comparison of features and functions of various tools for QC and  
269 preprocessing. (XLSX 41kb)

270 Supplementary Material 2: Details of QC in SOAPnuke. (PDF 304kb)

271 Supplementary Material 3: Details of preprocessing in SOAPnuke. (PDF 1.6mb)

- 1  
2  
3  
4 272 Supplementary Material 4: Details of preprocessing performance test and downstream analyses.  
5  
6 273 (DOCX 38kb)  
7  
8  
9  
10 274 Supplementary Material 5: Details of researches involving SOAPnuke. (XLSX 12kb)  
11

12  
13 275 REFERENCES  
14

- 15 276 1. Fox S, Filichkin S, Mockler TC. **Applications of ultra-high-throughput sequencing.**  
16  
17 277 Methods Mol Biol. 2009;553:79-108.  
18  
19 278 2. Soon WW, Hariharan M, Snyder MP. **High-throughput sequencing for biology and**  
20  
21 **medicine.** Mol Syst Biol. 2013;9:640.  
22 279  
23  
24 280 3. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. **Big Data:**  
25  
26 281 **Astronomical or Genomical?** PLoS Biol. 2015;13(7):e1002195.  
27  
28 282 4. Guo Y, Ye F, Sheng Q, Clark T, Samuels DC. **Three-stage quality control strategies for**  
29  
30 **DNA re-sequencing data.** Brief Bioinform. 2014;15(6):879-89.  
31 283  
32  
33 284 5. Zhou X, Rokas A. **Prevention, diagnosis and treatment of high-throughput**  
34  
35 285 **sequencing data pathologies.** Mol Ecol. 2014;23(7):1679-700.  
36  
37 286 6. Schmieder R, Edwards R. **Quality control and preprocessing of metagenomic**  
38  
39 287 **datasets.** Bioinformatics. 2011;27(6):863-4.  
40  
41 288 7. Moxon S, Schwach F, Dalmay T, Maclean D, Studholme DJ, Moulton V. **A toolkit for**  
42  
43 **analysing large-scale plant small RNA datasets.** Bioinformatics. 2008;24(19):2252-3.  
44 289  
45  
46 290 8. Gordon A, Hannon GJ. **Fastx-toolkit. FASTQ/A short-reads preprocessing tools**  
47  
48 291 **(unpublished)** [http://hannonlab.cshl.edu/fastx\\_toolkit](http://hannonlab.cshl.edu/fastx_toolkit). Accessed 1 Nov 2017.  
49  
50 292 9. Cox MP, Peterson DA, Biggs PJ. **SolexaQA: At-a-glance quality assessment of**  
51  
52 **Illumina second-generation sequencing data.** BMC bioinformatics. 2010 Sep  
53 293  
54  
55 294 27;11(1):485.  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

10. Zhang T, Luo Y, Liu K, Pan L, Zhang B, Yu J, et al. **BIGpre: a quality assessment package for next-generation sequencing data.** *Genomics Proteomics Bioinformatics.* 2011;9(6):238-44.

11. Aronesty E. **ea-utils: Command-line tools for processing biological sequencing data.** *Expression Analysis*, Durham, NC. 2011.

12. Yang X, Liu D, Liu F, Wu J, Zou J, Xiao X, et al. **HTQC: a fast quality control toolkit for Illumina sequencing data.** *BMC Bioinformatics.* 2013;14:33.

13. Li H. **seqtk: Toolkit for processing sequences in FASTA/Q formats.** <https://github.com/lh3/seqtk>. Accessed 1 Mar 2017.

14. Zhou Q, Su X, Wang A, Xu J, Ning K. **QC-Chain: fast and holistic quality control method for next-generation sequencing data.** *PLoS One.* 2013;8(4):e60234.

15. Zhou Q, Su X, Jing G, Ning K. **Meta-QC-Chain: comprehensive and fast quality control method for metagenomic data.** *Genomics Proteomics Bioinformatics.* 2014;12(1):52-6.

16. Patel RK, Jain M. **NGS QC Toolkit: a toolkit for quality control of next generation sequencing data.** *PLoS One.* 2012;7(2):e30619.

17. Simon A. **FastQC: a quality control tool for high throughput sequence data.** <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> Accessed 1 Nov 2017

18. Schmieder R, Lim YW, Rohwer F, Edwards R. **TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets.** *BMC Bioinformatics.* 2010;11:341.

19. Falgueras J, Lara AJ, Fernandez-Pozo N, Canton FR, Perez-Trabado G, Claros MG. **SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read.** *BMC Bioinformatics.* 2010;11:38.

- 1  
2  
3  
4 319 20. St John, J. **SeqPrep: Tool for stripping adaptors and/or merging paired reads**  
5  
6 320 **with overlap into single reads.** <https://github.com/jstjohn/SeqPrep> Accessed 1 Nov  
7  
8 321 2017
- 10 322 21. Kong Y. **Btrim: a fast, lightweight adapter and quality trimming program for**  
11  
12 **next-generation sequencing technologies.** *Genomics*. 2011;98(2):152-3.  
13 323  
14
- 15 324 22. Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, et al. **RobiNA: a**  
16  
17 **user-friendly, integrated software solution for RNA-Seq-based transcriptomics.**  
18  
19 326 *Nucleic Acids Res*. 2012;40(Web Server issue):W622-7.  
20
- 22 327 23. Martin M. **Cutadapt removes adapter sequences from high-throughput**  
23  
24 328 **sequencing reads.** *EMBnet. journal*. 2011 May 2;17(1):pp-10.  
25
- 26 329 24. Schubert M, Lindgreen S, Orlando L. **AdapterRemoval v2: rapid adapter trimming,**  
27  
28 **identification, and read merging.** *BMC Res Notes*. 2016;9:88.  
29 330
- 30 331 25. Dodt M, Roehr JT, Ahmed R, Dieterich C. **FLEXBAR-Flexible Barcode and Adapter**  
31  
32 **Processing for Next-Generation Sequencing Platforms.** *Biology (Basel)*.  
33 332  
34 2012;1(3):895-905.  
35 333  
36
- 37 334 26. Li YL, Weng JC, Hsiao CC, Chou MT, Tseng CW, Hung JH. **PEAT: an intelligent and**  
38  
39 **efficient paired-end sequencing adapter trimming algorithm.** *BMC Bioinformatics*.  
40 335  
41 2015;16 Suppl 1:S2.  
42 336  
43
- 44 337 27. Bolger AM, Lohse M, Usadel B. **Trimmomatic: a flexible trimmer for Illumina**  
45  
46 338 **sequence data.** *Bioinformatics*. 2014;30(15):2114-20.  
47
- 48 339 28. Sturm M, Schroeder C, Bauer P. **SeqPurge: highly-sensitive adapter trimming for**  
49  
50 **paired-end NGS data.** *BMC Bioinformatics*. 2016;17:208.  
51 340  
52
- 53 341 29. Jiang H, Lei R, Ding SW, Zhu S. **Skewer: a fast and accurate adapter trimmer for**  
54  
55 342 **next-generation sequencing paired-end reads.** *BMC Bioinformatics*. 2014;15:182.  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

- 1  
2  
3  
4 343 30. Chen S, Huang T, Zhou Y, Han Y, Xu M, Gu J. **AfterQC: automatic filtering, trimming,**  
5  
6 344 **error removing and quality control for fastq data.** BMC Bioinformatics.  
7  
8 345 2017;18(Suppl 3):80.  
9  
10 346 31. BUSHNELL, Brian. **BBMap: A Fast, Accurate, Splice-Aware Aligner.** Ernest Orlando  
11  
12 Lawrence Berkeley National Laboratory, Berkeley, CA (US), 2014.  
13 347  
14  
15 348 32. Joshi NA, Fass JN. **Sickle: A sliding-window, adaptive, quality-based trimming**  
16  
17 349 **tool for FastQ files.** <https://github.com/najoshi/sickle>. Accessed 1 Nov 2017.  
18  
19 350 33. Pertea, G. **fqtrim: trimming&filtering of next-gen reads.**  
20  
21 <https://ccb.jhu.edu/software/fqtrim/>. Access 1 Nov 2017.  
22 351  
23  
24 352 34. Vince B. **Scythe: A Bayesian adapter trimmer.** <https://github.com/vsbuffalo/scythe>  
25  
26 353 Access 1 Mar 2017.  
27  
28 354 35. Leggett RM, Clavijo BJ, Clissold L, Clark MD, Caccamo M. **NextClip: an analysis and**  
29  
30 355 **read preparation tool for Nextera Long Mate Pair libraries.** Bioinformatics.  
31  
32 356 2014;30(4):566-8.  
33  
34  
35 357 36. Criscuolo A, Brisse S. **AlienTrimmer: a tool to quickly and accurately trim off**  
36  
37 358 **multiple short contaminant sequences from high-throughput sequencing reads.**  
38  
39 359 Genomics. 2013;102(5-6):500-6.  
40  
41 360 37. Goecks J, Nekrutenko A, Taylor J, Galaxy T. **Galaxy: a comprehensive approach for**  
42  
43 361 **supporting accessible, reproducible, and transparent computational research**  
44  
45 362 **in the life sciences.** Genome Biol. 2010;11(8):R86.  
46  
47  
48 363 38. Team RC. **R: A language and environment for statistical computing.** R Foundation  
49  
50 364 for Statistical Computing, Vienna, Austria. 2013.  
51  
52 365 39. Illumina. NextSeq 500 System Overview.  
53  
54 [https://support.illumina.com/content/dam/illumina-support/courses/nextseq-system-over](https://support.illumina.com/content/dam/illumina-support/courses/nextseq-system-overview/story_content/external_files/NextSeq500_System_Overview_narration.pdf)  
55 366 [view/story\\_content/external\\_files/NextSeq500\\_System\\_Overview\\_narration.pdf](https://support.illumina.com/content/dam/illumina-support/courses/nextseq-system-overview/story_content/external_files/NextSeq500_System_Overview_narration.pdf) Accessed  
56  
57 367 1 Nov 2017.  
58  
59 368  
60  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

369 40. Huang J, Liang X, Xuan Y, Geng C, Li Y, Lu H, Qu S, Mei X, Chen H, Yu T, Sun N. A  
reference human genome dataset of the BGISEQ-500 sequencer. *GigaScience*. 2017;6(5):1-9.

370  
371 41. Zhang X, Hao L, Meng L, Liu M, Zhao L, Hu F, et al. Digital gene expression tag profiling  
analysis of the gene expression patterns regulating the early stage of mouse spermatogenesis.  
372 *PLoS One*. 2013;8(3):e58680.

373  
374 42. Tam S, Tsao MS, McPherson JD. Optimization of miRNA-seq data preprocessing. *Brief*  
*Bioinform*. 2015;16(6):950-63.

375  
376 43. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, et al. **Extensive sequencing of**  
**seven human genomes to characterize benchmark reference materials**. *Sci Data*.  
377 2016;3:160025.

378  
379 44. GATK best practices <http://www.broadinstitute.org/gatk/guide/best-practices>. Access 1<sup>st</sup>  
380 Nov 2017

381 45. GIAB high-confidence small variant and reference calls v3.3.2. GIAB. 2017  
382 [ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878\\_HG001/NISTv3.3.2/](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.3.2/). Access 1  
383 Nov 2017.

384 46. Zhang X, Hao L, Meng L, Liu M, Zhao L, Hu F, et al. **Digital gene expression tag**  
**profiling analysis of the gene expression patterns regulating the early stage of**  
385 **mouse spermatogenesis**. *PLoS One*. 2013;8(3):e58680.

386  
387 47. Zhou L, Chen J, Li Z, Li X, Hu X, Huang Y, et al. **Integrated profiling of microRNAs**  
**and mRNAs: microRNAs located on Xq27.3 associate with clear cell renal cell**  
388 **carcinoma**. *PLoS One*. 2010;5(12):e15224.

389  
390 48. Han Y, Zhang X, Wang W, Wang Y, Ming F. **The suppression of WRKY44 by**  
**GIGANTEA-miR172 pathway is involved in drought response of Arabidopsis**  
391 **thaliana**. *PLoS One*. 2013;8(11):e73541.

392

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

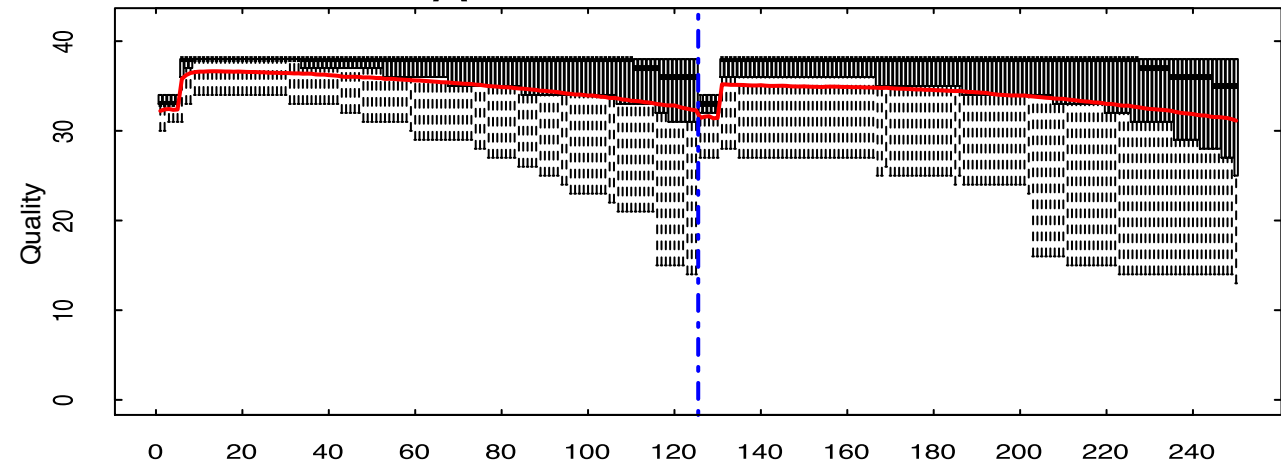
393 49. Hall AE, Lu WT, Godfrey JD, Antonov AV, Paicu C, Moxon S, et al. **The cytoskeleton**  
394 **adaptor protein ankyrin-1 is upregulated by p53 following DNA damage and**  
395 **alters cell migration.** Cell Death Dis. 2016;7:e2184.

396 50. Surbanovski N, Brillì M, Moser M, Si-Ammour A. **A highly specific**  
397 **microRNA-mediated mechanism silences LTR retrotransposons of strawberry.**  
398 Plant J. 2016;85(1):70-82.

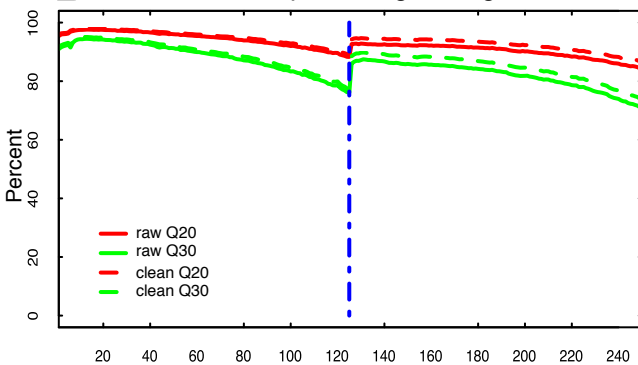
399 51. Chen, Y; Chen, Y; Shi, C; Huang, Z; Zhang, Y; Li, S; Li, Y; Ye, J; Yu, C; Li, Z; Zhang, X; Wang,  
400 J; Yang, H; Fang, L; Chen, Q (2017): **Supporting data for "SOAPnuke: A MapReduce**  
401 **Acceleration supported Software for integrated Quality Control and**  
402 **Preprocessing of High-Throughput Sequencing Data"** GigaScience Database.  
403 <http://dx.doi.org/10.5524/100373>

Figure.1 [Click here to download Figure Fig.1.pdf](#)

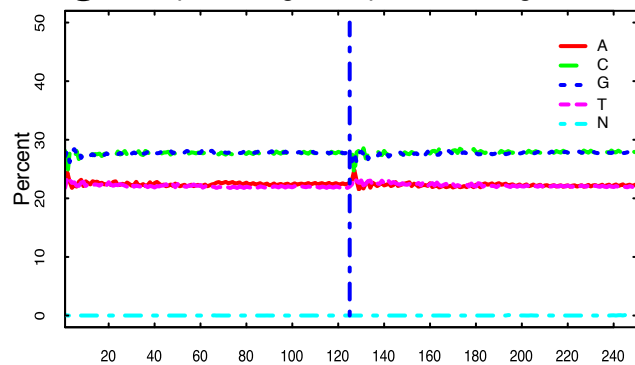
### A Base quality along reads



### B Q20 Q30 base percentage along reads




### C Base percentage composition along reads



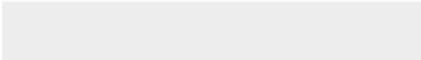



Click here to access/download  
**Supplementary Material**  
SM1.xlsx

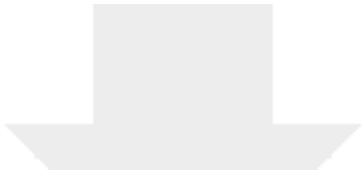




Click here to access/download  
**Supplementary Material**  
SM2.pdf

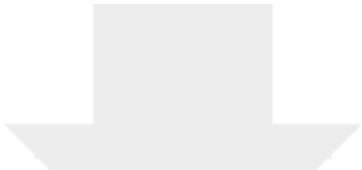







Click here to access/download  
**Supplementary Material**  
SM3.pdf





Click here to access/download  
**Supplementary Material**  
SM4.docx





Click here to access/download  
**Supplementary Material**  
SM5.xlsx

