GigaScience

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

--Manuscript Draft--

Manuscript Number:	GIGA-D-17-00173
Full Title:	SOAPnuke: A MapReduce Acceleration supported Software for integrated Quality Control and Preprocessing of High-Throughput Sequencing Data
Article Type:	Technical Note
Funding Information:	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
Abstract:	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 an abundant-function tool with 'QC-Preprocess-QC' architecture 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 the entire computer cluster. 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.
Corresponding Author:	Lin Fang CHINA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Yuxin Chen
First Author Secondary Information:	
Order of Authors:	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
Order of Authors Secondary Information:	
Opposed Reviewers:	
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using	Yes

a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	

Technical Note:

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

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

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

Yuxin C chenyuxin@genomics.cn; Yongsheng C chenys@geneplus.org.cn; CS scmfz@qq.com; ZH huangzhibo@genomics.cn; YZ zhangyong2@genomics.cn; SL lishengkang@genomics.cn; YL liyan@genomics.cn; JY yejia@genomics.cn; CY chang.yu@intel.com;ZL lzgdphcm@163.com; XZ zhangxq@genomics.cn; JW wangjian@genomics.cn;HY yanghm@genomics.cn;LF fangl@genomics.cn;QC cqiang8@189.cn

[†]Contributed equally

*First corresponding author

**Second corresponding author

ABSTRACT

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 an abundant-function tool with 'QC-Preprocess-QC' architecture 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 the entire computer cluster.

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.

KEYWORDS: High-throughput sequencing, Quality control, Preprocessing, MapReduce

BACKGROUND

High-throughput sequencing (HTS) instruments have enabled many large-scale studies and generated enormous amount of data [1-3]. However, the presence of low-quality bases, sequence artifacts and sequence contamination can introduce serious negative impact on downstream analyses. Thus, QC and preprocessing of raw data serve as the critical steps to initiate analysis pipelines [4, 5]. QC investigates several statistics of datasets to ensure data quality, and preprocessing trims off undesirable terminal fragments and filters out substandard reads [6]. We have conducted a survey on existing 31 tools and widely shared functions are listed in Supplementary Material 1.

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

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

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

For workflow-like tools, an optimal architecture is required since the orders of functions are fixed. Most of existing tools successively perform QC and preprocessing without complete statistics of preprocessed datasets. If the preprocessing operation is not suitable for a given dataset, the problem will not be revealed until alignment finishes.

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

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

offers a workflow with QC on un-preprocessed and preprocessed datasets while the preprocessing functions are too few.

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

To solve the problems mentioned above, we developed a workflow-like tool, SOAPnuke, for integrated QC and preprocessing of large HTS datasets. Similar to NGS QC toolkit, SOAPnuke performs two-step QC. Frequently used functions of similar tools were integrated in our program, including summary statistics calculation, trimming, filtering and other operations. Four modules are designed to handle genomic, metagenomic, DGE (Digital Gene Expression profiling) and sRNA (small RNA) datasets. In addition, SOAPnuke was extended to multiple working nodes for parallel computing using Hadoop MapReduce framework.

METHODS

QC & PREPROCESSING

We developed a two-step procedure to calculate summary statistics of un-preprocessed and preprocessed data. Basic statistics are comprised of the number of sequences and bases, base composition, Q20 and Q30, and filtering information. Complex statistics include the distribution of quality score and base composition distribution for each position. For the former distribution, Q20 and Q30 for each position were plotted in line chart and the quantiles of the quality were represented in a boxplot. And for the latter distribution, an overlapping histogram is used to display base composition distribution for each position. The calculations are conducted by C++ and the figures are generated by R 3.3.2 [38]. An example of the two figures

are provided in fig.1. A comprehensive list of statistics available in SOAPnuke is included in Additional File 2. Stats of preprocessed data are compared with some preset thresholds. For per-base quality distribution, a warning will be issued if median score in any position is less than 25, furthermore, a failure will be issued if less than 20. For per-base base composition, a warning will be issued if difference between A and T, or G and C in any position is greater than 10% or a failure will be issued if greater than 20%.

Fig.1. An example of QC complex stats. (a) per-base quality distribution of un-preprocessed paired-end reads. (b) per-base Q20&Q30 of un-preprocessed and preprocessed paired-end reads. (c) per-base base composition distribution of un-preprocessed paired-end reads.

In the step of preprocessing, those undesirable terminal fragments are trimmed off, substandard reads are filtered out, and some transform operations are applied. On both ends of reads, bases of assigned number or of quality lower than threshold will be trimmed off. Sequencing adapters can be aligned, where mismatch is supported while no indel is tolerated, and cut to 3' end. Filtering can be performed on reads with adapter, short length, too many ambiguous bases, low average quality or too many low-quality bases. Those sequencing batches, such as tile for Illumina sequencer and fov for BGI sequencer, with unfavorable sequencing quality can be assigned so the corresponding sequences will be discarded. In addition, reads with identical nucleotides can be deduplicated to keep only one copy. Transformation comprises quality system conversion, interconversion between DNA and RNA, and compression of output with gzip, etc. A comprehensive list of functions mentioned above with their parameters is available in Additional File 3.

MODULES DESIGN

In SOAPnuke, four modules with specified preprocessing functions are set to process different types of data, including General, DGE, sRNA and Meta modules. The previous section shows

preprocessing functions of General module which can handle most of the DNA re-sequencing datasets.

Digital Gene Expression Profiling (DGE) generates single-end read which has a 'CATG' segment neighboring targeted sequences of 17 base pairs. By default, DGE module will find the targeted segment and trim off other parts. Moreover, reads with ambiguous bases will be filtered. sRNA sequences can be contaminated by mRNA during sample preparation. Thus, filtering based on Poly-A tag is added to mRNA module as polyadenylation is a feature of mRNA data. Metagenomics preprocessing needs less functions than General module. It basically contains trimming adapters and low-quality bases on both ends, dropping reads with too short length or too many ambiguous bases. Detailed parameters settings can be accessed in Additional File 3.

SOFTWARE FEATURES

SOAPnuke is written by C++ for good scalability and performance. In terms of cross-platform compatibility, it can be run on both Linux and Windows.

Two paralleled strategies are implemented for acceleration. Multi-threading is developed for stand-alone operation. Data is cut into blocks of fixed size, and each block is processed by one thread. This design utilizes multiple cores in a working node. In SOAPnuke, the creation and allocation of threads are managed by threadpool library, which decreases the overhead of creating and destroying threads. Hadoop MapReduce is applied to achieve rapid processing in multi-node cluster. In the mapping phase, each read is kept as a key-value pair, where key is readID and value is sequence and quality scores. In shuffle phase, the key-value pairs are sorted, and each pair of paired-end reads is gathered. During the reducing phase, blocks of fixed size are processed by various threads of multiple nodes, and each block generates an individual result. After that, it is optional to merge the results into two fastq files.

To prove the effectiveness of the acceleration design, we have conducted a performance tests on SOAPnuke and other tools. The datasets are published by GIAB [39] and we extracted ~30x for testing. The access to testing data is available in Addition File 4. In terms of testing environment, up to 16 nodes were used, each of which has 24 cores of Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz and RAM of 128G. To choose operations for testing, we referred to those manuscripts using SOAPnuke in their works, which are included in Additional File 5. Trimming adapters and filtering on length and quality were selected for their universality. We chose workflow-like tools capable of performing these functions, which are Trimmomatic [27], AfterQC [30], BBDuk [31] and AlignTrimmer [36]. The parameter setting is also available in Addition File 4.

RESULTS

In the performance test, we chose three indexes for evaluation: elapsed time, CPU usage and maximum RAM usage. As shown in table.1, AfterQC is the tool occupying least resources. However, its processing time is too long for practical usage, especially considering we ran the program with pypy, which is announced to be as 3 times fast as standard python. Among the remaining tools, SOAPnuke struck an appropriate balance between resources occupancy and performance. If high performance is demanded, users can choose to run SOAPnuke on multiple nodes with MapReduce framework. In our testing, 16 nodes can achieve ~32 times acceleration compared to standalone operation, which is 5.37 times faster than the highest speed of other tools.

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

Table.1. Evaluation of the data processing performance across SOAPnuke and other 4 tools. Time, throughput, CPU and maximum memory occupation are presented. For CPU usage, 100% means full load of a single CPU core. Maximum RAM usage means the highest occupancy of RAM during the whole processing.

After the preprocessing, we have conducted the downstream analyses with GATK best practice pipeline (see the description of GATK best practices [http://www.broadinstitute.org/gatk/guide/best-practices]). The result shows that the operations of SOAPnuke have similar impact on downstream analysis and details are available in Additional File 4.

DISCUSSION AND CONCLUSION

Quality of data is critical to downstream analysis, which makes it important to use reliable tools for preprocessing. To omit unnecessary IO and computation, workflow-like structure is adopted in SOAPnuke, where QC and preprocessing functions are integrated within an executable. 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 sequenced 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. 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 various modules. For example, in preprocessing of DGE datasets, filtering out singleton reads is always included [40-42]. For sRNA module, screening out reads based on alignment with some RNA databases (such as tRNA, rRNA and snoRNA) [43,44] is under development. It is also considerable to add statistics like per-read quality distribution and length distribution. In terms of program structure, we are performing refactoring to increase the standalone processing speed.

However, we have found two intractable problems with QC and preprocessing respectively among the existing tools. Firstly, in terms of preprocessing, it is difficult to choose optimal parameters for a specific dataset. Datasets from the same experiments and sequencers are tend to share features, so users always select the same parameters for those similar data. The parameters are initially defined based on experiments on a specific dataset or just experience, which already introduces some error. Moreover, even if the parameters are optimal for the tested dataset, they are possibly inappropriate for other data because of other random factors. Thus, the current method is a compromise on processing time. However, it might be a considerable solution that preprocessing settings are automatically adjusted during the processing. Secondly, the QC statistics are of limited help to judge the availability of data. For example, as the threshold of filtering out low-quality reads gets from 0 to 40, all quality-related QC stats will persist rising, but the result of variant calling will be improved then get worse. It is because preprocessing is a procedure required to strike a balance between removing noise and

keeping useful information, while the balance cannot be reflected by the QC statistics. QC stats can only indicate whether the filtering condition is too loose instead of too harsh. As a potential solution, empirical models on similar datasets can be referred to for judging whether the preprocessing conditions are too harsh. These two problems will also be considered during the development of following versions.

Availability and requirements

Project name: SOAPnuke

Project home page: https://github.com/BGI-flexlab/SOAPnuke

Operating system(s): Linux, Windows

Programming language: C++

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

License: GPL

Abbreviations

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

Declarations

ACKNOWLEDGEMENTS

This research was supported by Collaborative Innovation Center of High Performance Computing, Critical Patented Project of the Science & Technology Bureau of Fujian Province,

China (Grant No. 2013YZ0002-2) and the Joint Project of Natural Science and health Foundation of Fujian Province, China (Grant No.2015J01397).

AUTHORS' CONTRIBUTIONS

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

COMPETING INTERESTS

The authors declare that they have no competing interests.

OPEN ACCESS

ADDITIONAL FILES

Additional file 1: Comparison of features and functions of various tools for QC and preprocessing. (XLSX 41kb)

Additional file 2: Details of QC in SOAPnuke. (PDF 304kb)

Additional file 3: Details of preprocessing in SOAPnuke. (PDF 1.6mb)

Additional file 4: Details of preprocessing performance test and downstream analyses. (DOCX 37kb)

Additional file 5: Details of researches involving SOAPnuke. (XLSX 12kb)

REFERENCES

1. Fox S, Filichkin S, Mockler TC. **Applications of ultra-high-throughput sequencing**. Methods Mol Biol. 2009;553:79-108.
2. Soon WW, Hariharan M, Snyder MP. **High-throughput sequencing for biology and medicine.** Mol Syst Biol. 2013;9:640.
3. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. **Big Data: Astronomical or Genomical?** PLoS Biol. 2015;13(7):e1002195.
4. Guo Y, Ye F, Sheng Q, Clark T, Samuels DC. **Three-stage quality control strategies for DNA re-sequencing data.** Brief Bioinform. 2014;15(6):879-89.
5. Zhou X, Rokas A. **Prevention, diagnosis and treatment of high-throughput sequencing data pathologies.** Mol Ecol. 2014;23(7):1679-700.
6. Schmieder R, Edwards R. **Quality control and preprocessing of metagenomic datasets.** Bioinformatics. 2011;27(6):863-4.
7. Moxon S, Schwach F, Dalmay T, Maclean D, Studholme DJ, Moulton V. **A toolkit for analysing large-scale plant small RNA datasets.** Bioinformatics. 2008;24(19):2252-3.
8. Gordon A, Hannon GJ. **Fastx-toolkit. FASTQ/A short-reads preprocessing tools (unpublished)** http://hannonlab. cshl. edu/fastx_toolkit. Accessed 1 Mar 2017.

9. Cox MP, Peterson DA, Biggs PJ. **SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data.** BMC bioinformatics. 2010 Sep 27;11(1):485.
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 Mar 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.
20. St John, J. **SeqPrep: Tool for stripping adaptors and/or merging paired reads with overlap into single reads.** https://github.com/jstjohn/SeqPrep Accessed 1 Mar 2017
21. Kong Y. **Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies.** Genomics. 2011;98(2):152-3.
22. Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, et al. **RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics**. Nucleic Acids Res. 2012;40(Web Server issue):W622-7.
23. Martin M. **Cutadapt removes adapter sequences from high-throughput sequencing reads**. EMBnet. journal. 2011 May 2;17(1):pp-10.
24. Schubert M, Lindgreen S, Orlando L. **AdapterRemoval v2: rapid adapter trimming, identification, and read merging.** BMC Res Notes. 2016;9:88.
25. Dodt M, Roehr JT, Ahmed R, Dieterich C. **FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms**. Biology (Basel). 2012;1(3):895-905.
26. Li YL, Weng JC, Hsiao CC, Chou MT, Tseng CW, Hung JH. **PEAT: an intelligent and efficient paired-end sequencing adapter trimming algorithm**. BMC Bioinformatics. 2015;16 Suppl 1:S2.
27. Bolger AM, Lohse M, Usadel B. **Trimmomatic: a flexible trimmer for Illumina sequence data**. Bioinformatics. 2014;30(15):2114-20.
28. Sturm M, Schroeder C, Bauer P. **SeqPurge: highly-sensitive adapter trimming for paired-end NGS data**. BMC Bioinformatics. 2016;17:208.

29. Jiang H, Lei R, Ding SW, Zhu S. **Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads**. BMC Bioinformatics. 2014;15:182.

30. Chen S, Huang T, Zhou Y, Han Y, Xu M, Gu J. **AfterQC: automatic filtering, trimming, error removing and quality control for fastq data.** BMC Bioinformatics. 2017;18(Suppl 3):80.

31. BUSHNELL, Brian. **BBMap: A Fast, Accurate, Splice-Aware Aligner.** Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA (US), 2014.

32. Joshi NA, Fass JN. **Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files**. https://github.com/najoshi/sickle. Accessed 1 Mar 2017.

33. Pertea, G. **fqtrim: trimming&filtering of next-gen reads.** https://ccb.jhu.edu/software/fqtrim/. Access 1 Mar 2017.

34. Vince B. **Scythe: A Bayesian adapter trimmer**. https://github.com/vsbuffalo/scythe Access 1 Mar 2017.

35. Leggett RM, Clavijo BJ, Clissold L, Clark MD, Caccamo M. **NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries.** Bioinformatics. 2014;30(4):566-8.

36. Criscuolo A, Brisse S. **AlienTrimmer: a tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads.** Genomics. 2013;102(5-6):500-6.

37. Goecks J, Nekrutenko A, Taylor J, Galaxy T. **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences**. Genome Biol. 2010;11(8):R86.

38. Team RC. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. 2013.

39. 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. 2016;3:160025.

40. 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.** PLoS One. 2013;8(3):e58680.

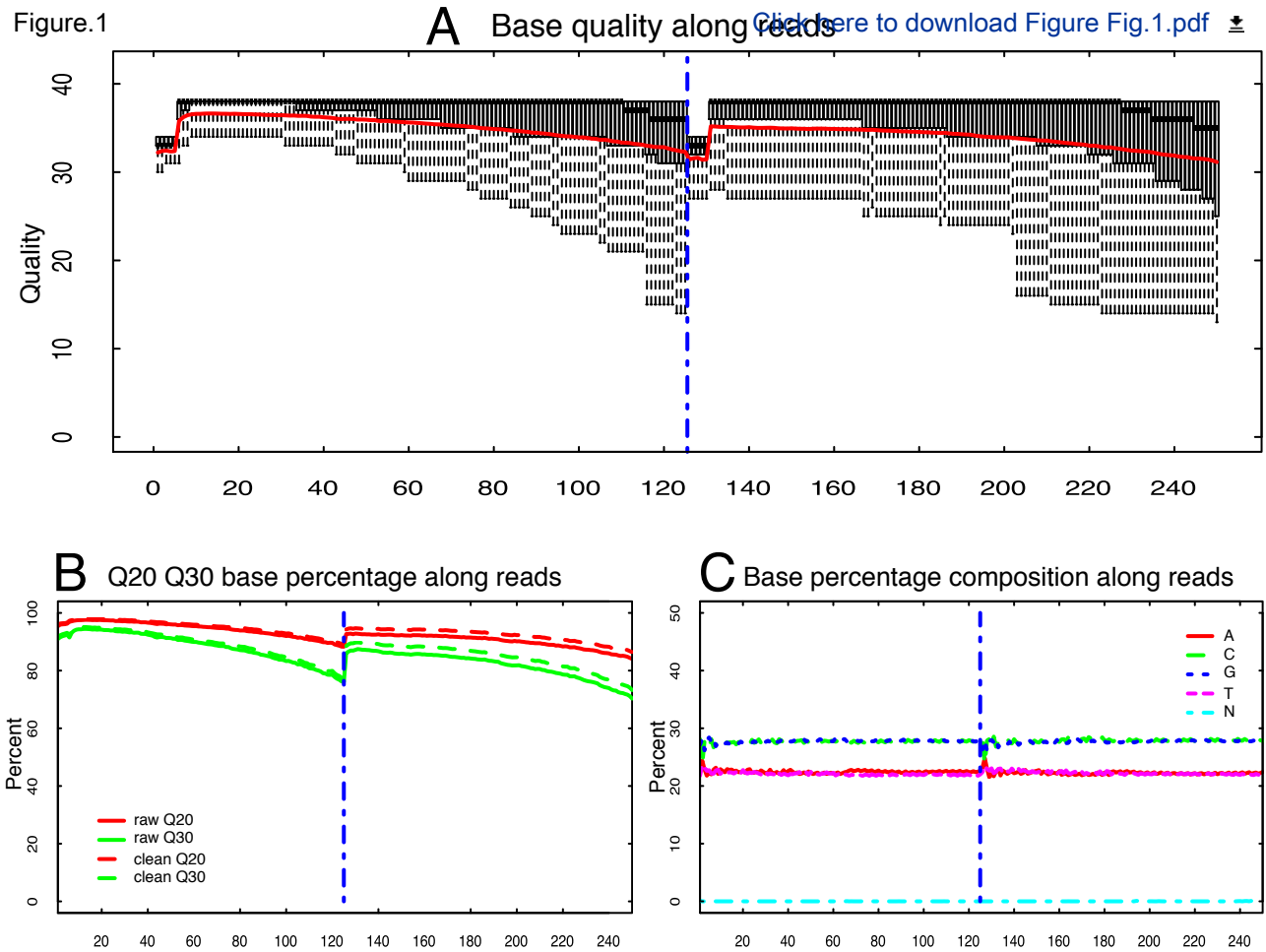
41. Zhou L, Chen J, Li Z, Li X, Hu X, Huang Y, et al. **Integrated profiling of microRNAs and mRNAs: microRNAs located on Xq27.3 associate with clear cell renal cell carcinoma.** PLoS One. 2010;5(12):e15224.

42. 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 thaliana**. PLoS One. 2013;8(11):e73541.

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

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

Figure.1



Additional File 1

Additional File 2

Additional File 3

Additional File 4

Additional File 5