Available online at www.sciencedirect.com
ScienceDirect
www.sciencedirect.com/science/journal/16720229
GENOMICS PROTEOMICS & BIOINFORMATICS

Application Note

BIGpre: A Quality Assessment Package for Next-Generation Sequencing Data

Tongwu Zhang[1,2], Yingfeng Luo[1], Kan Liu[1], Linlin Pan[1], Bing Zhang[1], Jun Yu[1], and Songnian Hu[1*]

[1] *CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China;*
[2] *James D. Watson Institute of Genome Sciences, Zhejiang University, Hangzhou 31007, China.*

Genomics Proteomics Bioinformatics 2011 Dec; 9(6): 238-244 DOI: 10.1016/S1672-0229(11)60027-2

Received: Sep 04, 2011; Accepted: Nov 23, 2011

Abstract

The emergence of next-generation sequencing (NGS) technologies has significantly improved sequencing throughput and reduced costs. However, the short read length, duplicate reads and massive volume of data make the data processing much more difficult and complicated than the first-generation sequencing technology. Although there are some software packages developed to assess the data quality, those packages either are not easily available to users or require bioinformatics skills and computer resources. Moreover, almost all the quality assessment software currently available didn't taken into account the sequencing errors when dealing with the duplicate assessment in NGS data. Here, we present a new user-friendly quality assessment software package called BIGpre, which works for both Illumina and 454 platforms. BIGpre contains all the functions of other quality assessment software, such as the correlation between forward and reverse reads, read GC-content distribution, and base Ns quality. More importantly, BIGpre incorporates associated programs to detect and remove duplicate reads after taking sequencing errors into account and trimming low quality reads from raw data as well. BIGpre is primarily written in Perl and integrates graphical capability from the statistics package R. This package produces both tabular and graphical summaries of data quality for sequencing datasets from Illumina and 454 platforms. Processing hundreds of millions reads within minutes, this package provides immediate diagnostic information for user to manipulate sequencing data for downstream analyses. BIGpre is freely available at http://bigpre.sourceforge.net/.

Key words: next-generation sequencing, quality assessment, duplicate reads, sequencing error

Introduction

Next-generation sequencing (NGS) technology has demonstrated its capacity to produce an enormous volume of data cheaply at an unprecedented speed. The variety of NGS features makes these platforms coexist in the marketplace, with some having clear advantages for particular applications over others (*1*). At present, five NGS platforms are available, including the Roche GS-FLX 454 Genome Sequencing (also referred as 454 sequencing), the Illumina Genome Analyzer (also referred as Solexa sequencing), the ABI SOLiD analyzer, Polonator G.007 and the Helicos HeliScope platforms (*1-4*). The new Illumina HiSeq 2000 Genome Analyzer is capable of producing single reads of 2X100 base pairs (bp) and gener-

*Corresponding author.
E-mail: husn@big.ac.cn

ates about 200 gigabase (Gb) of short sequences per run. The current model, SOLiD 4.0 analyzer, has a read length of up to 50 bp and can produce 80-100 Gb of mappable sequences per run. Additional platforms with faster sequencing speed and lower reagents cost have become available recently (such as Ion Torrent and PacBio) (*5*).

The giant datasets generated by the NGS platforms present the big challenges and opportunities for software and algorithm development (*6*). During the past several years, a large number of new software applications and algorithms have been developed for sequence alignment or assembly but only a few for quality assessment and visualization, such as TileQC (*7*), SolexaQA (*8*), and PIQA (*9*). Meanwhile, with the deep concern of NGS raw data, the duplicate reads are the major problems for the subsequent analysis in both Illumina and 454 platforms (*10, 11*). The duplicate reads are mainly caused by emulsion PCR, especially with insufficient sample.

Until now, there are few types of software designed to detect or preprocess the duplicate reads. Most of them are implemented in the pipeline of other softwares (e.g. SAMtools (*12*) and GATLK (*13*)), which cannot be used directly to raw reads. The two widely used quality control packages for duplicate read removal is available in rmdup in SAMtools and FASTQ/A Collapser in the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). The rmdup in SAMtools is used to remove potential PCR duplicates in short read mapping project and takes the read mapping data (such as SAM or BAM format) but not raw reads as the input. Moreover, it recognizes the duplicates with the identical external coordinates, therefore, the sequencing errors, such as homopolymer in 454 sequencing reads, are not taken into account in SAMtools. Another program FASTQ/A Collapser just simply collapses identical sequences (in a FASTA file) into a single sequence and does not allow sequencing errors and sequencing length differences, either.

Given the fact that sequencing errors are common in NGS platforms and the length of reads varies after trimming low quality or in 454 sequencing platform, BIGpre uses the mismatch setting to get more precise detection of duplicates and takes into account the different sequencing errors along the read length in different sequencing platforms. Here, we present BIGpre, a user-friendly and free software package which can provide rapid, simple and comprehensive assessment of read quality and duplication rate for both Illumina and 454 platforms.

Implementation

BIGpre is command-line executable package written in perl and available as source code on multiple Linux platforms (e.g. Fedora and Ubuntu). The current version (BIGpre 2.02) has been tested on Linux (2.6.18-164.el5, x86_64) (BIGpre 3.0, a higher version, is under test now and will be available next year). BIGpre, along with an implementation file of the described method and example can be downloaded at the project website http://bigpre.sourceforge.net/. The package requires the programming environment R (Version 2.5 or higher). The R software is available at the website "The R Project for Statistical Computing" (http://www.r-project.org/). The memory requirement is equal to the size of the input data.

Results and Discussion

As we know, low quality reads often result in the high mis-assembly rate of *de novo* sequencing and the high false SNP calling of re-sequencing, whereas presence of duplicate reads often leads to overestimation of the sequencing coverage (*14*). So the quality control of the raw data is very important to ensure the data reliability for further analysis. BIGpre is a useful package including several programs designed to assess read quality in generality and flexibility for both Illumina and 454 sequencing data. All the program codes are written with perl scripts and the package can display the results graphically with the statistics package R. To decrease memory exhaustion, BIGpre reads data from disk at request instead of keeping data cached.

Here, we introduce the key features of BIGpre. A detailed documentation including screen shots is also available at the software web site listed above. The following datasets are used as example. Dataset from one lane (paired-end) and three lanes (mate-paired) for whole genome sequencing of *Litopenaeus van-*

namei, respectively, was used for Illumina quality control and duplication analysis. In addition, dataset used for 454 quality control and duplication analysis is 1/4 run of transcriptome sequencing of *Bemisia tabaci*. All the datasets are sequenced in Beijing Institute of Genomics (Zhang, T., *et al*. unpublished).

Sequencing quality assessment

The program *solqs* in BIGpre is used to assess the quality of Illumina sequencing data. In Illumina sequencing, the read quality assessment is based on each cycle along the read including undefined base (named as "N") (**Figure 1A-C**). And the phred value Q20 is an important quality score to evaluate the sequencing error rate. Generally, the mean GC content in all the forward/reverse reads will be coincident with the genome GC content (gGC) in genome sequencing (**Figure 1D**), but the read mean quality distribution often shows the bias in some regions with extreme high/low GC content. Also, the correlation between forward and reverse reads was affected by the machine running time (**Figure 1E**). *Solqs* can evaluate the correlation between forward and reverse reads and analyse the four nucleotide ration in each cycle along the read and count the undefined base. In addition, s*olqs* can also show the boundary effects (*8, 15*) and the data production in each tile (**Figure 1F, G**).

The program *454QC* in BIGpre is designed to evaluate the quality of reads sequenced by 454 platform. Similar to *solqs*, *454QC* can provide the information of read length and base quality distribution, GC content statistics, and undefined base "N" analysis (**Figure 2A**). Besides, this program also introduces a new assessment to poly(N), which is the main quality concern of 454 data (*16*), including ploy(N) length and base quality of different position in poly(N) (**Figure 2B-E**).

Duplicate read identification

Two programs *soldup* and *454dup* in BIGpre can identify and count duplicate reads by comparing the read consistency. The duplicate reads in NGS data are usually caused by either the failure of the emulsion PCR step to match on sequence to one bead, or the insufficient starting material. Large insert libraries, such as 20 kb mate-paired library in 454 platform are prone to generate duplicate reads. With insufficient starting material, PCR copies of the same target are sequenced multiple times. Removing duplicate from raw reads is very important, because the duplicates can cause genomic and transcriptomic mis-assembly, mislead variant basecalling and alignment algorithms (*17*), and lead to incorrect interpretation of the abundance of species and genes in metagenomic study (*11*).

There are several principles to detect duplicate reads. First, the detection should be mapping-free, which means it should be independent of reference genome (*17*). Second, in order to reduce the false duplicates, the detection should be done after trimming of low quality reads. Third, the high raw cluster density can generate the artificial reads and result in the higher duplication ratio (*14*). Therefore, these reads should be removed as duplicate reads from raw data. Fourth, for the data generated from paired-end or mate-paired libraries, both ends should be taken into consideration.

For Illumina platform data, the program *soldup* analyses the duplication ratio with different duplicate length sets in each sequencing library and provides the useful unique reads (**Figure 3A**) and data useful ratio when adding a new lane (**Figure 3B**). Moreover, the user can select a maximal length as the criterion for duplicate length and remove all the duplicates with the one selected copy or coverage value copy left.

For 454 platform data, the situation is much more complicated. The 454 duplicate reads include both artificial and natural duplicates, and can make up 4-44% of total reads in metagenomic samples (*11*). Separating artificial duplicates from natural duplicates is very difficult. Removing all the duplicates will result in the underestimation of sequencing coverage, since majority of duplicates in 454 sequencing platforms are low depth with two or three copy (**Figure 4**). The program *454dup* is designed to group reads by constructing consensus sequences as the duplicates. In those duplicate groups, the read is sorted by the read length and quality. Also, the ploy (N) mis-match in each duplicate group is treated as the sequencing error. In the tested dataset, we found that about 5% raw reads were grounded as duplicates and most of the

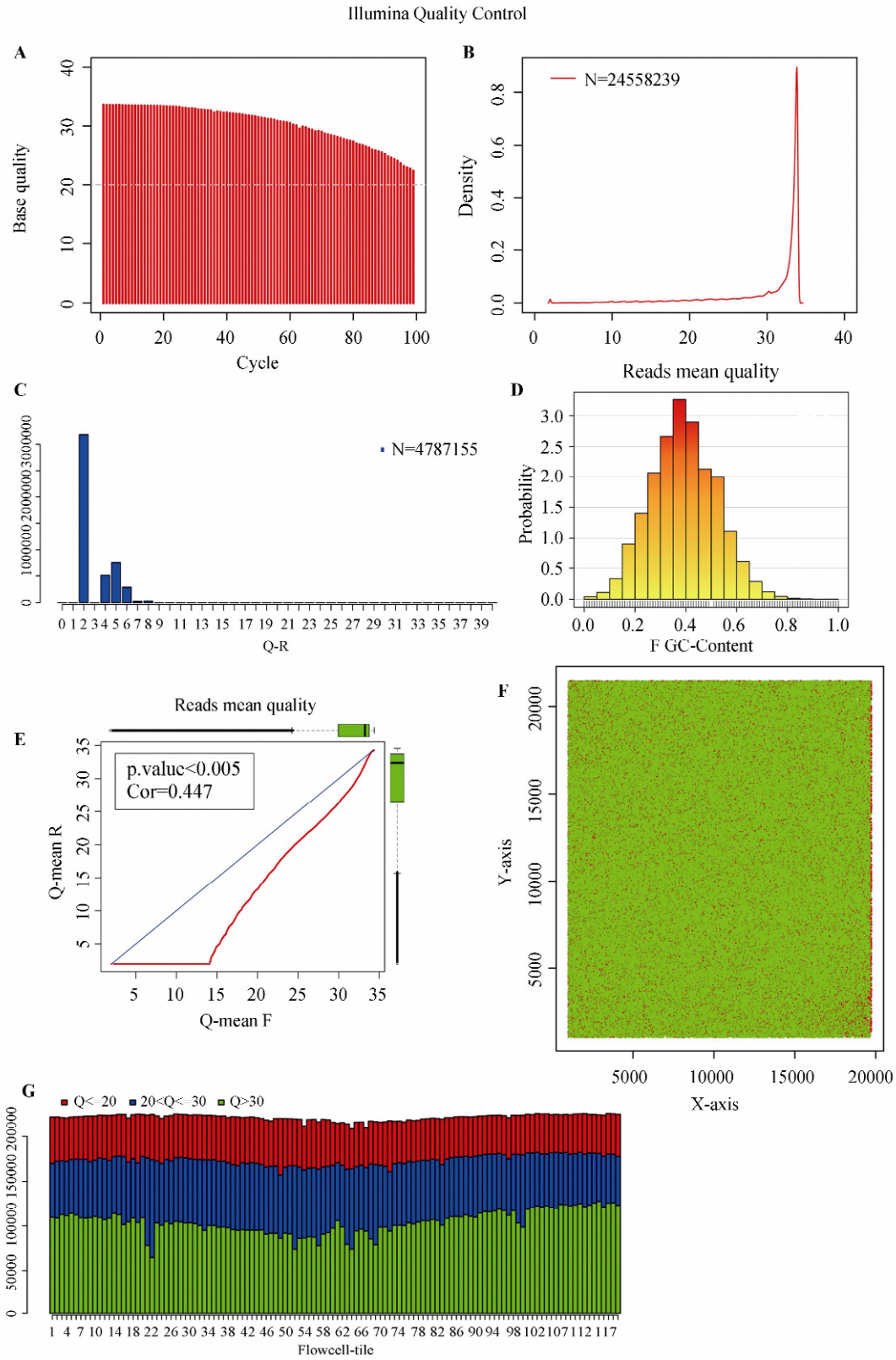

Figure 1 The main read quality control in Illumina sequencing platform. **A.** Per base sequence quality distribution. **B.** Per sequence quality score distribution. **C.** Per base "N" quality score distribution. **D.** Correlation between forward and reverse sequences. **E.** Reads quality score distribution in one Illumina tile. Q>20 in green and Q<=20 in red. **F.** Per sequence GC content distribution. **G.** The distribution of data production in all Illumina tiles. Q>30 in green, Q>20 && Q<=30 in blue and Q<=20 in red.

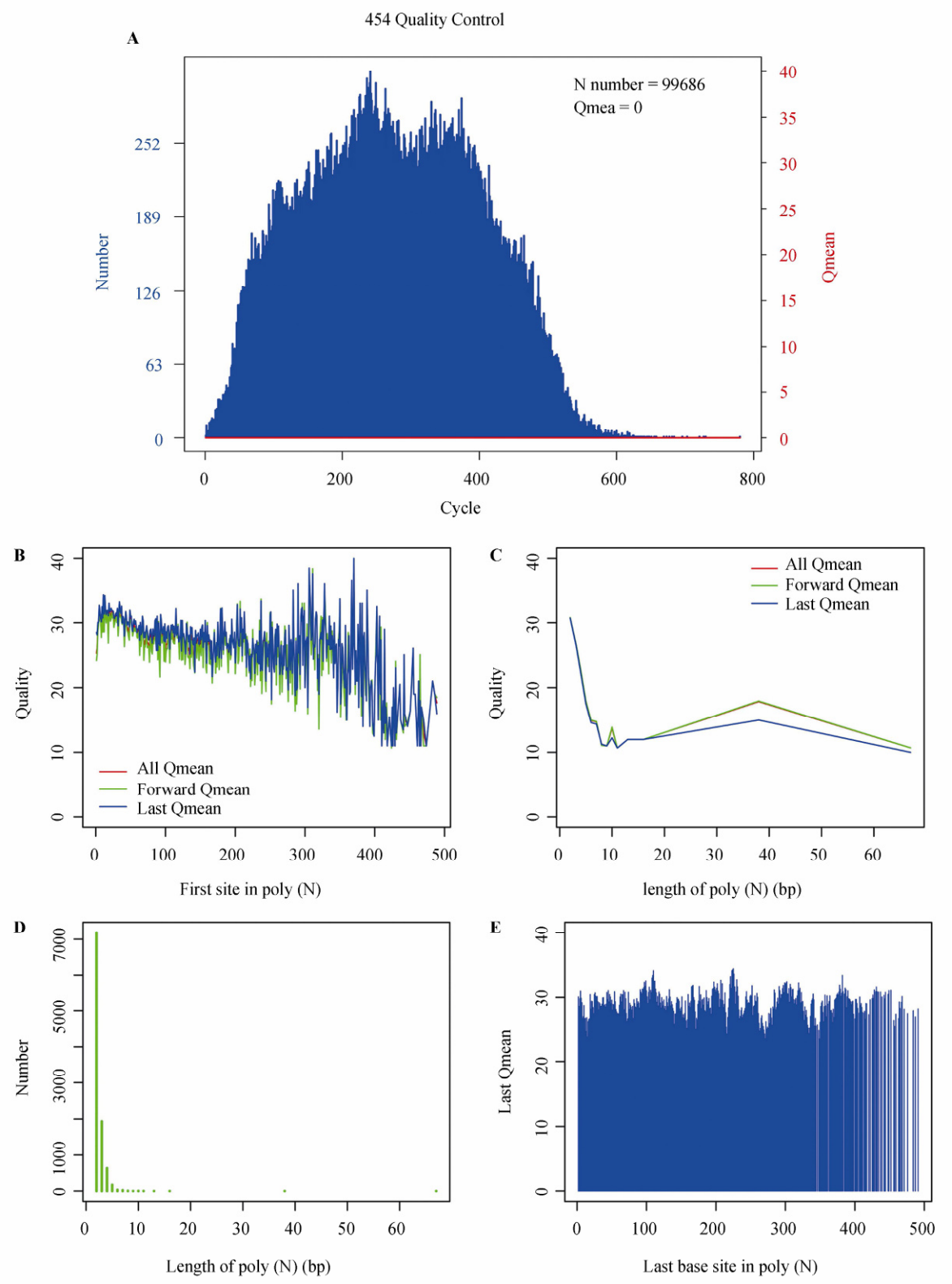

Figure 2 The main read quality control in 454 sequencing platform. **A.** Per base "N" sequence quality and number distributions. **B.** per poly(N) sequence quality distribution at different positions. Forward Qmean indicates the poly(N) mean quality excluding last base; Last Qmean indicates the last base quality in poly(N). **C.** Per poly(N) sequence quality distribution for different length. **D.** Distribution of the length of poly(N). **E.** Last base quality distribution in poly(N) at different positions.

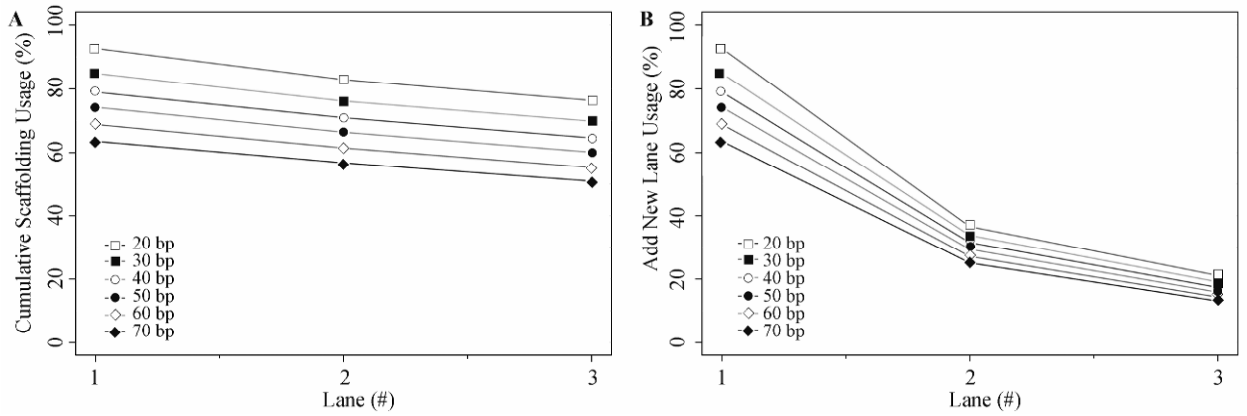

Figure 3 Duplicate analysis in Illumina paired-end library. **A**. Cumulate scaffolding usage ratio (also cumulate useful reads ratio) in Illumina mate-pair library (insert size: 5 kb) with different duplicate length sets. **B**. New lane usage ratio with different length sets.

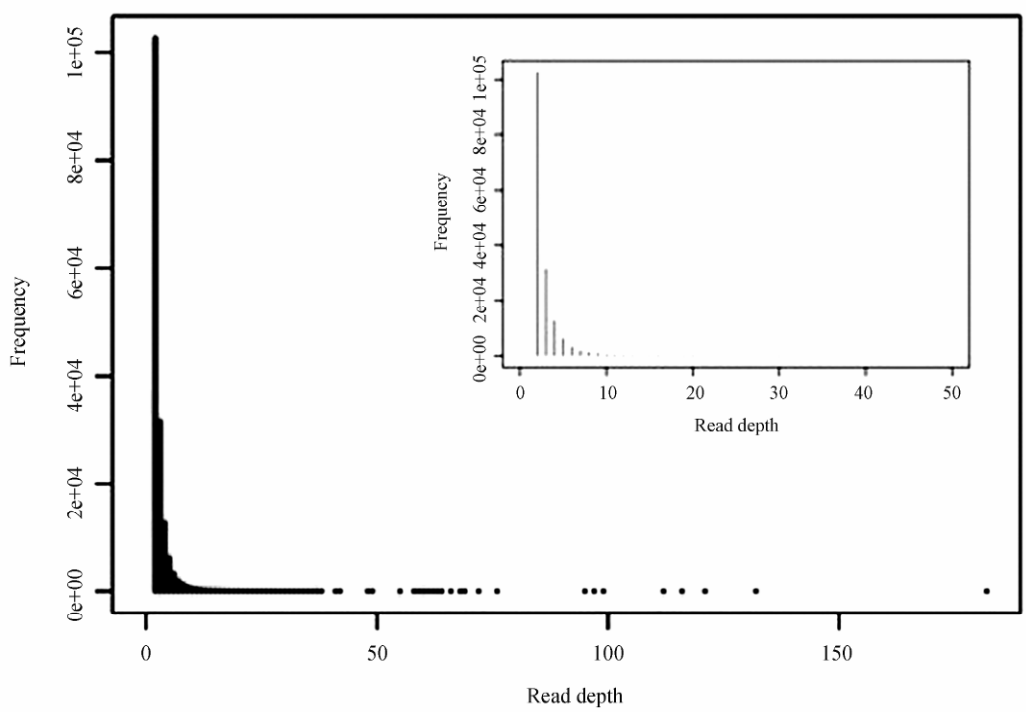

Figure 4 Frequency of duplicate read depth in 454 transcriptome sequencing. The read depth is indicated by the number of reads in each duplicate group.

duplicates were artificial due to the low sequencing coverage. Because the reads in the same group have the same sequence, the user can choose the criterion to filter the duplicates in each group.

Additional preprocess tools for Illumina data

BIGpre also includes some tools for preprocessing Illumina data and all those tools are organized into one program called *solapt*. The program *solsize* is designed to analyse the true library insert size after mapping the reads to reference, and can be used to evaluate the library quality. The program *soljoin* takes the sequencing direction of insert library into account and is designed specifically to join paired-end reads into a single longer read, when the library insert size is smaller than total length of paired end reads. These manually prolonged reads would be very helpful to the genome and transcriptome *de novo* assembly. The program *solfiter* can be used to filter and trim raw

data into high quality fastq reads to remove low quality bases while the program *solin* can remove the internal adapter from the mate-paired read and produce two new paired-end reads.

Conclusion

The BIGpre package provides both tabular and graphical summaries of data quality for both Illumina and 454 platforms. Compared to other available tools, the BIGpre package is designed for the data preprocessing and quality control with easy access and providing more information. It integrates several functions into one package to insure that only the high quality reads are used for subsequent analysis: (i) assesses the read quality with rapid, simple and effective measures for two platforms. (ii) detects and analyses the duplicate reads and duplication depth. (iii) preprocesses the sequencing data, such as joining the small paired-end reads into longer single reads, removing the internal adapter sequences, and trimming raw data into high quality sequences. This package produces standardized outputs within minutes, thus facilitates the reads quality comparison in each machine runs, and provides library quality and complexity assessment in a very intuitive manner.

Acknowledgements

The authors wish to acknowledge Dr. Qing Zhou and Dr. Xiaomin Yu for their helpful comments. This work was supported by the National Natural Science Foundation of China (Grant No. 31000561 and 30900825) and the Knowledge Innovation Program of the Chinese Academy of Sciences (Grant No. KSCX2-EW-R-01-04).

Authors' contributions

TZ drafted the manuscript and developed the software. YL and KL participated in the software design and manuscript writing. LP and BZ participated in the initial design. JY and SH proposed the idea of the software and revised the manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors have no competing interests to declare.

References

1 Metzker, M.L. 2010. Sequencing technologies - the next generation. *Nat Rev. Genet.* 11: 31-46.
2 Ng, P.C. and Kirkness, E.F. 2010. Whole genome sequencing. *Methods Mol. Biol.* 628: 215-226.
3 Schuster, S.C. 2008. Next-generation sequencing transforms today's biology. *Nat. Methods* 5: 16-18.
4 Tucker, T.*, et al.* 2009. Massively parallel sequencing: the next big thing in genetic medicine. *Am. J. Hum. Genet.* 85: 142-154.
5 Schadt, E.E.*, et al.* 2010. A window into third-generation sequencing. *Hum. Mol. Genet.* 19: R227-240.
6 Bateman, A. and Quackenbush, J. 2009. Bioinformatics for next generation sequencing. *Bioinformatics* 25: 429.
7 Dolan, P.C. and Denver, D.R. 2008. TileQC: a system for tile-based quality control of Solexa data. *BMC Bioinformatics* 9: 250.
8 Cox, M.P.*, et al.* 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 11: 485.
9 Martinez-Alcantara, A.*, et al.* 2009. PIQA: pipeline for Illumina G1 genome analyzer data quality assessment. *Bioinformatics* 25: 2438-2439.
10 Kozarewa, I.*, et al.* 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods* 6: 291-295.
11 Niu, B.*, et al.* 2010. Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics* 11: 187.
12 Li, H.*, et al.* 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078-2079.
13 McKenna, A.*, et al.* 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20: 1297-1303.
14 Li, R.*, et al.* 2010. The sequence and *de novo* assembly of the giant panda genome. *Nature* 463: 311-317.
15 Rougemont, J.*, et al.* 2008. Probabilistic base calling of Solexa sequencing data. *BMC Bioinformatics* 9: 431.
16 Quinlan, A.R.*, et al.* 2008. Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat. Methods* 5: 179-181.
17 Levin, J.Z.*, et al.* 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* 7: 709-715.