*Briefings in Bioinformatics*, 2017, 1–5

**doi: 10.1093/bib/bbw143**
Paper

# hppRNA—a Snakemake-based handy parameter-free pipeline for RNA-Seq analysis of numerous samples

Dapeng Wang

Corresponding author: Dapeng Wang, Department of Plant Sciences, University of Oxford, S Parks Rd, Oxford OX1 3RB, UK. E-mail: dapeng.wang@plants.ox.ac.uk

## Abstract

RNA-Seq technology has been gradually becoming a routine approach for characterizing the properties of transcriptome in terms of organisms, cell types and conditions and consequently a big burden has been put on the facet of data analysis, which calls for an easy-to-learn workflow to cope with the increased demands from a large number of laboratories across the world. We report a one-in-all solution called hppRNA, composed of four scenarios such as pre-mapping, core-workflow, post-mapping and sequence variation detection, written by a series of individual Perl and R scripts, counting on well-established and preinstalled software, irrespective of single-end or paired-end, unstranded or stranded sequencing method. It features six independent core-workflows comprising the state-of-the-art technology with dozens of popular cutting-edge tools such as Tophat–Cufflink–Cuffdiff, Subread–featureCounts–DESeq2, STAR–RSEM–EBSeq, Bowtie–eXpress–edgeR, kallisto–sleuth, HISAT–StringTie–Ballgown, and embeds itself in Snakemake, which is a modern pipeline management system. The core function of this pipeline is turning the raw fastq files into gene/isoform expression matrix and differentially expressed genes or isoforms as well as the identification of fusion genes, single nucleotide polymorphisms, long noncoding RNAs and circular RNAs. Last but not least, this pipeline is specifically designed for performing the systematic analysis on a huge set of samples in one go, ideally for the researchers who intend to deploy the pipeline on their local servers. The scripts as well as the user manual are freely available at https://sourceforge.net/projects/hpprna/

**Key words:** RNA-Seq; pipeline; a large number of samples; gene expression profiling; sequence variation

## Introduction

With the advancement of sequencing technology and big-data analysis approach, RNA-Seq tends to be more prevalent and important in the biological laboratory in the current era and deems one of the most dominant and efficient methodologies in the measurement of gene expression [1, 2]. RNA-Seq has brought about not only the novel chances of detection of low-expressed genes and discrimination of isoforms belonging to the same gene by means of unlimited in-depth sequencing, but also the notable challenges in terms of both data structure and analysis such as high-volume data, high-performance computing capacity and tricky processing [3, 4]. A typical RNA-Seq analysis involves multiple steps and hence the difficulty also arises from the fact that it is imperative to enable consistency of the data formats between the tools used in the adjacent steps. Apart from this, researchers have to write scripts of their own to interpret and parse the intermediate outputs derived from a variety of software.

To overcome the barrier, lots of pipeline programs for RNA-Seq analysis have been developed, including types of remotely hosted and web-based servers and locally installed packages based on a wide variety of programming or coding systems, each of which has its particular strength and advantage. For instance, some tools accept complementary data formats as input, such as Chipster, wapRNA, PRADA, RseqFlow and RobiNA [5–9]. Besides, RSEQtools establishes a model to protect the individual privacy by splitting the total information of alignments into two parts and building a relationship between them for the following in-depth analysis [10]. TRAPLINE constructs the networks of protein–protein and miRNA–target interactions [11]. TCW supports the cross-species transcriptome analysis through the integration of a few evolutionary tools. ArrayExpressHTS and easyRNASeq are chiefly implemented in R language and both of them use R objects to store intermediate data and call the other programs that are inside or outside R environment [12, 13]. NGSUtils, ViennaNGS and S-MART provide a collection of next generation sequencing (NGS) relevant tools that are

**Dapeng Wang**, Department of Plant Sciences, University of Oxford. He is working on the construction and application of a series of pipelines for next-generation-sequencing analysis.

**Submitted:** 19 September 2016; **Received (in revised form):** 12 December 2016

capable of manipulating, transforming, analyzing and/or displaying specialized data files containing reads, alignments or annotations [14–16]. GENE-Counter addresses the issue caused by insufficient number of samples by means of a modified negative binomial distribution to enhance the statistical power [17]. RAP and FX work with the advanced cloud computing technology to perform the analysis and have some refined strategies in aligning of reads or detection of alternative splicing events [18, 19]. BioWardrobe is a system that can be installed in a core server and offers the service to all the biologists in a user-friendly graphical interface, with a selection of simplified parameters [20]. QuickNGS allows the analysis across a large number of species covering a great variety of taxonomic categories and uses a MySQL database to organize and manage the data to promote the efficiency of data usage [21]. ExpressionPlot brings the functionality of data presentation in a comparative fashion and interactive operation between the tabular data and graphic data via hyperlinks [22]. GeneProf maintains a webserver and a database simultaneously so that users would be able to contrast the analytical results from their own data with those from public data repositories collected from large projects or certain sets of samples from multiple experiments in the context of the same workflows [23, 24]. GenomeSpace comes with two conspicuous features which are easy format conversion between tools and flexible sub-modules together with detailed instructions and examples that serve as the basic or minimum unit for the formation of complicated tasks or high-level analysis [25]. Galaxy is a popular web-based workbench that combines an abundance of tools that work for specific steps involved in the global analysis and enable the recording, sharing and reusing of the workflows that have been executed [26].

These currently available programs or servers have made immense contributions to the end users in the global community of biology, accelerating the emergence and development of more novel and powerful tools. However, some of the pipelines require data transferring that is time-consuming and allow restricted working space, or only offer one or a small number of package choices for each step, or ask the users to decide on the redundant parameters step by step, which is not suitable for the analysis of a large data set at a time. For large-scale applications, we present a comprehensive pipeline implemented in the local server, which only requires a sample table specifying the description of each sample, a bunch of compared groups and a limited number of parameters for the experimental design, generating all important consequences after one round. The most valuable part of this pipeline is that it includes six categories of core-workflows and all the codes are programmed in Snakemake pipeline management system. Besides, several monitor points have been set up in several important steps to assess and evaluate the quality of the analysis, such as in the levels of reads, alignments and samples. We speculate that this work offers an alternative choice for the researchers who are keen to take the full advantage of NGS technology in their projects with less concern of the trivial steps.

## Methods

Genomic sequences in fasta format and gene annotations in gtf format for two species such as human (hg19) and mouse (mm10) were gleaned from iGenomes database provided by illumina corporation (ftp://ussd-ftp.illumina.com/). Only genes that are defined as protein-coding and located in the complete chromosomes could be retained. Another file composed of the relationship between gene name and transcript name was created for further use. To reduce the total amount of resource data, index files for mapping will be produced only after respective workflow is evoked and will be stored within the workflow folder. Known long noncoding RNA (lncRNA) and repeat annotation files were collected from GENCODE database [27] and UCSC Table Browser [28], respectively.

Initially, the pipeline precedes with the raw fastq files generated from the sequencer or commercial repository such as BaseSpace. The quality of reads is evaluated by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) for both the raw fastq files and processed fastq files. Base quality filter is done by means of PRINSEQ-lite [29] or FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) for paired-end or single-end reads, respectively. Adaptor sequences, where applicable, could be trimmed by Cutadapt [30] in light of library construction methods. The number of unique reads is estimated by fastx_collapser (as part of FASTX-Toolkit) based on sequence identity.

In the main module of this pipeline, there are six categories of typical workflows to work on the three steps including mapping, quantification and differentially expressed genes (DEGs) detection: (1) Tophat–Cufflink–Cuffdiff [31–33]; (2) Subread–featureCounts–DESeq2 [34–36]; (3) STAR–RSEM–EBSeq [37–39]; (4) Bowtie–eXpress–edgeR [40–42]; (5) kallisto–sleuth [43, 44]; (6) HISAT–StringTie–Ballgown [45–47] (Figure 1). With the guidance of the methodological designs from different workflows, genome sequences are fed to the mappers for workflow (1), (2), (3) and (6), and transcript sequences are offered to those for workflow (4) and (5). In all cases, the BAM files, either in genomic coordinate or in transcript coordinate, would be generated by the mappers, of which the former is preferred. Two matrix files contain the expression quantities in the level of gene or transcript/isoform where various defined measurements such as fragments per kilobase million (FPKM), reads per kilobase million (RPKM) or transcripts per million (TPM) will be used under different circumstances, which is dependent on the type of software or algorithm. In particular, the strategy of considering the sum of the values of transcripts from the identical gene as the final value for this gene will be adopted unless such kind of results can be directly outputted through the in-built module of this software. Likewise, the DEG results will be presented in two files in the two levels of gene and transcript, which comprise all DEGs/transcripts in all possibilities of compared groups and corrected *P*-values and fold-changes based on the statistical testing with optional approaches.

For the simplicity and compactness, the produced bam file is further processed with SAMtools [48], in terms of sorting, indexing, transforming the formats between BAM and SAM files and assessing potential polymerase chain reaction (PCR) duplicate level. If possible, unique alignment is selected by BlackOPs [49] with the help of identification of tag 'NH:i:1'. To control the mapping efficiency, RNA quality and balance of read distribution are examined by plotting the read density across the consensus gene model with ngs.plot.r [50]. Wig files are created by SAMtools and BigWig files that would be able to display the tracks in species-oriented genome browser are further obtained by wigToBigWig provided by UCSC genome browser team. A set of genes with highly confident expression are selected after filtering out the genes that have expression values $<1$ in all samples. R packages are responsible for the graphical illustration of the global tendency and relationship between and among all the samples by analyzing the expression profiling for all genes, resulting in a group of nice images for heatmap and principle component analysis as well as some useful text files including the expression matrix with selected genes and the matrix of

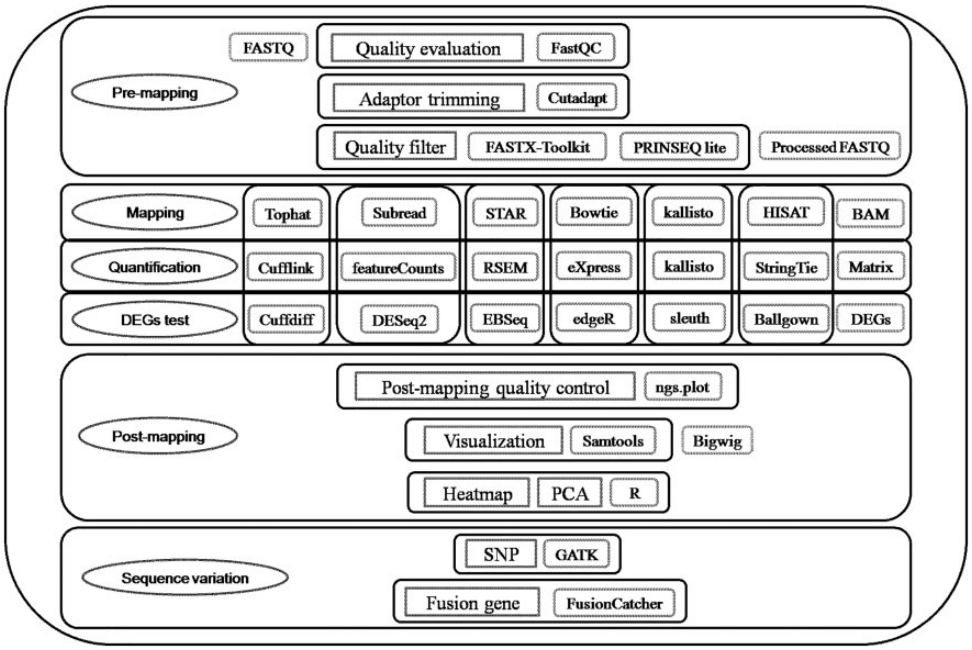


**Figure 1.** The flowchart for hppRNA.

variable loadings for understanding the original contribution of certain gene to certain principle, allowing users to explore how the replicates are clustered together or how the samples from different conditions are scattered.

Fusion genes are inspected by FusionCatcher [51] based on the proof of physical spanning reads from paired-end or long-read single-end data. For single nucleotide polymorphism (SNP) detection, reads are mapped onto the genome through STAR [39] 2-pass aligning and duplicates are marked by means of Picard (http://broadinstitute.github.io/picard/). In the end, reliable SNPs are reported by GATK [52] via post-processing, variant calling and filtering of mutations from each sample according to a combination of thresholds.

This pipeline is capable of performing lncRNAs analysis by means of two ways such as known lncRNAs quantification and novel lncRNAs identification. All the well-annotated known lncRNAs have been integrated into the above-mentioned six main workflows and as a result, users could be able to check the expression profiling for mRNAs and lncRNAs at a time. In addition, this package offers the functional module for identifying the unknown lncRNAs contained in the analyzed samples. In detail, workflow 1 is used to align reads and assemble transcripts under the guidance of a combined annotation set of known mRNAs and lncRNAs. One consensus transcript set is produced from assemblies of all samples via Cuffmerge [31] and the potential novel transcripts are determined by comparing the new set with the known gene set through Cuffcompare [31] and selecting the specific class codes such as 'i', 'j', 'o', 'u' and 'x' [53]. Subsequently, iSeeRNA [54] distinguishes noncoding genes from coding ones, and all the newly discovered lncRNAs are added to the known gene sets, which can be imported to workflow 5 for any further analysis.

To perform the circular RNA (circRNA) analysis, STAR [39] is chosen to do the mapping with different strategies for single-end and paired-end data, leading to the generation of chimeric alignments. Afterward, DCC [55] is in charge of discovering and quantifying circRNAs, and CircTest [55] works to examine the difference between pairwise conditions in terms of relative abundance of circRNAs through comparison of expressions between circRNAs and their host genes.

## Implementation

The entire pipeline is written by a variety of programming languages such as Perl, R and Python, where Perl and R take charge of streamlining the whole process of the pipeline for multiple samples simultaneously, connecting the output with input in the adjoining steps and customizing the diversified parameters provided by users. The pipeline management system is made following modern Snakemake [56] framework together with ordinary Python codes, which facilitate the error handling, parallel operation and file self-control. The parameters or arguments for each software involved can be classified into three groups such as default, modified and user-provided setting where the third one specifies the situations of fragment length and sequencing orientation, allowing this pipeline to be suited for various categories of library construction and sequencing strategies, either unstranded or stranded, either single-end or paired-end, which shows the potential to have widespread effective applications in general circumstances. To run the whole pipeline properly, only three steps should be taken one by one. First, a shell script is executed to collect the genome and annotation data, install the requisite publicly available software/tools as well as gather homemade Perl and R scripts. Second, the raw fastq files, single-end or paired-end, for all samples in an analysis should be stored in a folder. Lastly, a main Snakemake file should be compiled through performing a Perl script with an input table file including the detailed description for the project and experimental design and the specific explanation for each sample and the condition of replicates as well as which comparisons the user would like to make.

## Discussion

We aim to develop a smart and standard pipeline that makes it possible to carry out the analysis of RNA-Seq from the very beginning to the very end in massive samples without a heavy burden in compiling the scripts, satisfying the common requirements from public research community. In particular, we endeavor to extract information as much as possible to fully understand the gene expression profiling in RNA level and sequence variation and point mutation in DNA level from RNA-Seq data. Furthermore, we focus our attention on human and mouse genomes in the first version of this pipeline, which is easy to be expanded to other well-annotated genomes of model species as we use an extremely extendable and flexible style of formulation and organization of each module in writing the scripts, especially rendering the formats of data sets in a unified manner. We will regularly add more novel workflows which consist of newly developed calculation tools as anything new emerges. We welcome all the feedback from users regarding our pipeline and are always waiting at some point to improve and update the modules to meet the specific demands from them and hope to assist in making full use of the merit of RNA-Seq technology as it goes.

**Key Points**

- This pipeline incorporates six prevailing alternative workflows to align, quantify and analyze RNA-Seq data, which gives more options for users to choose or easily do some comparison analysis for distinct approaches.
- It is wrapped in the Snakemake pipeline management system.
- It handles an unlimited number of samples and is fitted for various types of experimental designs.
- It is user-friendly and all the installation for dozens of software is automated.
- It provides analyses for mRNAs, lncRNAs and circRNAs and deals with the detection of sequence variations such as fusion genes and SNPs from RNA-Seq data.

## References


1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63.
2. Brawand D, Soumillon M, Necsulea A, *et al*. The evolution of gene expression levels in mammalian organs. *Nature* 2011;**478**:343–8.
3. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* 2011;**12**:87–98.
4. Conesa A, Madrigal P, Tarazona S, *et al*. A survey of best practices for RNA-seq data analysis. *Genome Biol* 2016;**17**:13.
5. Kallio MA, Tuimala JT, Hupponen T, *et al*. Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics* 2011;**12**:507.
6. Torres-Garcia W, Zheng S, Sivachenko A, *et al*. PRADA: pipeline for RNA sequencing data analysis. *Bioinformatics* 2014;**30**:2224–6.
7. Lohse M, Bolger AM, Nagel A, *et al*. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Res* 2012;**40**:W622–7.
8. Wang Y, Mehta G, Mayani R, *et al*. RseqFlow: workflows for RNA-Seq data analysis. *Bioinformatics* 2011;**27**:2598–600.
9. Zhao W, Liu W, Tian D, *et al*. wapRNA: a web-based application for the processing of RNA sequences. *Bioinformatics* 2011;**27**:3076–7.
10. Habegger L, Sboner A, Gianoulis TA, *et al*. RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* 2011;**27**:281–3.
11. Wolfien M, Rimmbach C, Schmitz U, *et al*. TRAPLINE: a standardized and automated pipeline for RNA sequencing data analysis, evaluation and annotation. *BMC Bioinformatics* 2016;**17**:21.
12. Delhomme N, Padioleau I, Furlong EE, *et al*. easyRNASeq: a bioconductor package for processing RNA-Seq data. *Bioinformatics* 2012;**28**:2532–3.
13. Goncalves A, Tikhonov A, Brazma A, *et al*. A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics* 2011;**27**:867–9.
14. Breese MR, Liu Y. NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics* 2013;**29**:494–6.
15. Zytnicki M, Quesneville H. S-MART, a software toolbox to aid RNA-Seq data analysis. *PLoS One* 2011;**6**:e25988.
16. Wolfinger MT, Fallmann J, Eggenhofer F, *et al*. ViennaNGS: a toolbox for building efficient next- generation sequencing analysis pipelines. *F1000Res* 2015;**4**:50.
17. Cumbie JS, Kimbrel JA, Di Y, *et al*. GENE-counter: a computational pipeline for the analysis of RNA-Seq data for gene expression differences. *PLoS One* 2011;**6**:e25279.
18. Hong D, Rhie A, Park SS, *et al*. FX: an RNA-Seq analysis tool on the cloud. *Bioinformatics* 2012;**28**:721–3.
19. D'Antonio M, D'Onorio De Meo P, Pallocca M, *et al*. RAP: RNA-Seq analysis pipeline, a new cloud-based NGS web application. *BMC Genomics* 2015;**16**:S3.
20. Kartashov AV, Barski A. BioWardrobe: an integrated platform for analysis of epigenomics and transcriptomics data. *Genome Biol* 2015;**16**:158.
21. Wagle P, Nikolic M, Frommolt P. QuickNGS elevates next-generation sequencing data analysis to a new level of automation. *BMC Genomics* 2015;**16**:487.
22. Friedman BA, Maniatis T. ExpressionPlot: a web-based framework for analysis of RNA-Seq and microarray gene expression data. *Genome Biol* 2011;**12**:R69.
23. Halbritter F, Kousa AI, Tomlinson SR. GeneProf data: a resource of curated, integrated and reusable high-throughput genomics experiments. *Nucleic Acids Res* 2014;**42**:D851–8.
24. Halbritter F, Vaidya HJ, Tomlinson SR. GeneProf: analysis of high-throughput sequencing experiments. *Nat Methods* 2011;**9**:7–8.
25. Qu K, Garamszegi S, Wu F, *et al*. Integrative genomic analysis by interoperation of bioinformatics tools in GenomeSpace. *Nat Methods* 2016;**13**:245–7.
26. Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 2010;**11**:R86.
27. Harrow J, Frankish A, Gonzalez JM, *et al*. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 2012;**22**:1760–74.
28. Speir ML, Zweig AS, Rosenbloom KR, *et al*. The UCSC genome browser database: 2016 update. *Nucleic Acids Res* 2016;**44**:D717–25.
29. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 2011;**27**:863–4.
30. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 2011;**17**:10–12.

31. Trapnell C, Williams BA, Pertea G, *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010;**28**:511–15.
32. Kim D, Pertea G, Trapnell C, *et al*. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 2013;**14**:R36.
33. Trapnell C, Hendrickson DG, Sauvageau M, *et al*. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* 2013;**31**:46–53.
34. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**:923–30.
35. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.
36. Liao Y, Smyth GK, Shi W. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res* 2013;**41**:e108.
37. Leng N, Dawson JA, Thomson JA, *et al*. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* 2013;**29**:1035–43.
38. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 2011;**12**:323.
39. Dobin A, Davis CA, Schlesinger F, *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21.
40. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40.
41. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**:357–9.
42. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods* 2013;**10**:71–3.
43. Pimentel HJ, Bray N, Puente S, *et al*. Differential analysis of RNA-Seq incorporating quantification uncertainty. *bioRxiv* 2016:058164.
44. Bray NL, Pimentel H, Melsted P, *et al*. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016;**34**:525–7.
45. Frazee AC, Pertea G, Jaffe AE, *et al*. Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat Biotechnol* 2015;**33**:243–6.
46. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015;**12**:357–60.
47. Pertea M, Pertea GM, Antonescu CM, *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 2015;**33**:290–5.
48. Li H, Handsaker B, Wysoker A, *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.
49. Cabanski CR, Wilkerson MD, Soloway M, *et al*. BlackOPs: increasing confidence in variant detection through mappability filtering. *Nucleic Acids Res* 2013;**41**:e178.
50. Shen L, Shao N, Liu X, *et al*. ngs.plot: quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics* 2014;**15**:284.
51. Nicorici D, Satalan M, Edgren H, *et al*. FusionCatcher-a tool for finding somatic fusion genes in paired-end RNA-sequencing data, *bioRxiv* 2014:011650.
52. McKenna A, Hanna M, Banks E, *et al*. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**:1297–303.
53. Sun L, Zhang Z, Bailey TL, *et al*. Prediction of novel long non-coding RNAs based on RNA-Seq data of mouse Klf1 knockout study. *BMC Bioinformatics* 2012;**13**:331.
54. Sun K, Chen X, Jiang P, *et al*. iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC Genomics* 2013;**14(Suppl 2)**:S7.
55. Cheng J, Metge F, Dieterich C. Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics* 2016;**32**:1094–6.
56. Koster J, Rahmann S. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics* 2012;**28**:2520–2.