Bioinformatics, 33(11), 2017, 1723–1724
doi: 10.1093/bioinformatics/btx038
Advance Access Publication Date: 27 January 2017
Applications Note

OXFORD

Genome analysis

ANAQUIN: a software toolkit for the analysis of spike-in controls for next generation sequencing

Ted Wong[1], Ira W. Deveson[1,2], Simon A. Hardwick[1,3] and Tim R. Mercer[1,3,*]

[1]Genomics and Epigenetics Division, Garvan Institute of Medical Research, Sydney, NSW, Australia, [2]Faculty of Science, School of Biotechnology and Biomolecular Sciences, UNSW, Sydney, NSW, Australia and [3]Faculty of Medicine, St Vincents Clinical School, UNSW, Sydney, NSW, Australia

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on September 26, 2016; revised on January 5, 2017; editorial decision on January 18, 2017; accepted on January 23, 2017

Abstract

Summary: Spike-in controls are synthetic nucleic-acid sequences that are added to a user's sample and constitute internal standards for subsequent steps in the next generation sequencing workflow.

The Anaquin software toolkit can be used to analyze the performance of spike-in controls at multiple steps during RNA sequencing or genome sequencing analysis, providing useful diagnostic statistics, data visualization and sample normalization.

Availability and Implementation: The software is implemented in C++/R and is freely available under BSD license. The source code is available from github.com/student-t/Anaquin, binaries and user manual from www.sequin.xyz/software and R package from bioconductor.org/packages/Anaquin

Contact: anaquin@garvan.org.au or t.mercer@garvan.org.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Next-generation sequencing (NGS) is widely used in biological research and is being increasingly used for clinical diagnosis. However, NGS experiments are confounded by technical variation, biases and artifacts that arise during library preparation, sequencing and subsequent bioinformatic analysis.

Spike-in controls are RNA or DNA molecules that can be directly added to a user's sample prior to sequencing (Deveson *et al.*, 2016; Hardwick *et al.*, 2016; Jiang *et al.*, 2011; Zook *et al.*, 2012). Spike-in controls are typically synthetic sequences that can be distinguished from the natural RNA/DNA sequences in the sample. This enables spike-ins to be analyzed in parallel to the accompanying natural sample, acting as internal quantitative and qualitative controls.

The analysis of spike-in controls enables an assessment of multiple steps during the NGS workflow (see Fig. 1). This includes measuring features of the NGS library (such as library complexity, quality and sequencing error), determining diagnostic statistics (such as sensitivity and specificity) and for quality-control and trouble-shooting purposes.

There are a range of statistical and bioinformatic strategies to analyze spike-in controls. The *erccdashboard* software provides easy analysis and visualization of ERCC RNA spike-in controls that are commonly used in microarray and RNA sequencing experiments (Munro *et al.*, 2014). However, as spike-ins are being increasingly adopted in genome sequencing and metagenomics, there is a growing need for the analysis of spike-ins in diverse experimental contexts.

To facilitate the analysis of spike-in controls for NGS, we have developed a software toolkit, termed *Anaquin*. This toolkit allows users to evaluate the performance of spike-in controls and the accompanying RNA/DNA sample. This toolkit is compatible with most common bioinformatics tools and data formats, and is easily integrated into standard NGS workflows.

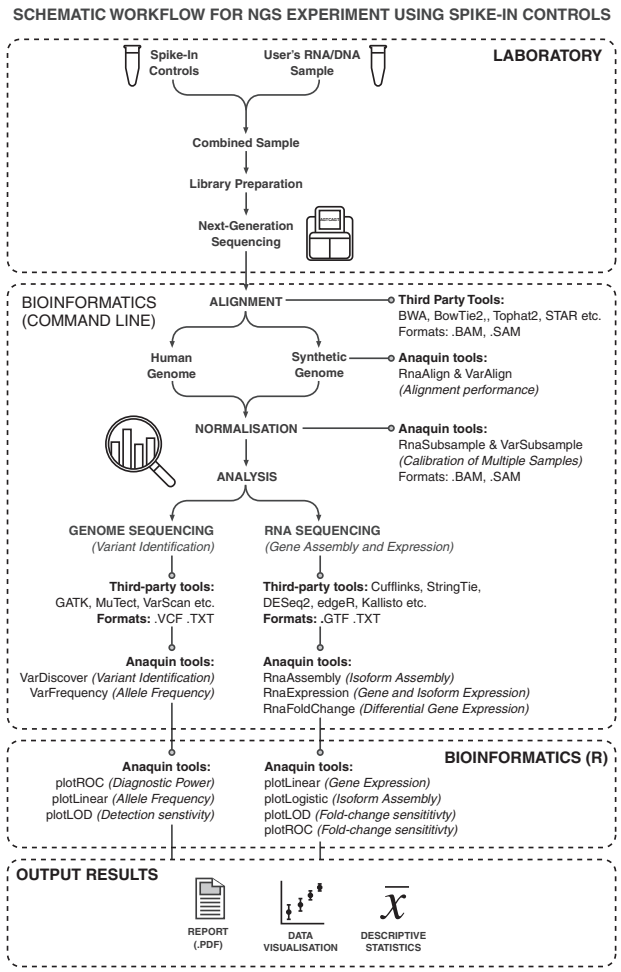

Fig. 1. Schematic overview of next-generation sequencing workflow, with analytical steps using Anaquin and third-party tools indicated

2 Results

2.1 Implementation

Anaquin is implemented in both C++ and R programming language. The C++ command-line software is run in a UNIX environment, and is useful for intensive computation analysis (eg. analyzing large .BAM alignment files), and integration with other command line tools and data formats. The related R-package is distributed by Bioconductor and is ideal for data visualization and statistical analysis.

Anaquin has been designed for integration with NGS bioinformatics pipelines of third-party software. Accordingly, the software supports the use of standard data formats (such as SAM, BAM, BED, VCF, GTF etc.) and has been tested in conjunction with popular third-party software (such Cufflinks, DESeq2, TopHat2, STAR, GATK, VarScan, etc.). Users can also convert non-supported data formats into simple tab-delimited text formats that are supported by Anaquin. Where appropriate, Anaquin also supports multiple replicate input files.

Anaquin may also require reference information files to help with the analysis of spike-in controls. Common examples include (i) mixture files that indicate the concentration of spike-in controls in a mixture, (ii) sequence files that provide the sequence of the spike-ins or an artificial in silico chromosome or genome sequences to which spike-ins align or (iii) annotation files that indicate the coordinates of spike-in controls with respect to the aforementioned in silico chromosome or genome.

2.2 Tools

Anaquin toolkit is organized in a hierarchal fashion, with a range of different tools that can be used to assess the performance of spike-in controls at several steps during the user's NGS workflow of third party software. For example, in RNA sequencing experiments, Anaquin can be used to assess split-read alignments (RnaAlign), isoform assembly (RnaAssembly), gene and isoform quantification (RnaExpression). In addition, Anaquin enables normalization between multiple samples (RnaSubsample) and can be used to assess differential gene expression between libraries (RnaFoldChange). For DNA sequencing experiments, Anaquin provides tools to assess the performance of read alignment (VarAlign) and variant identification (VarDiscover), and calibrate sequencing coverage between samples or replicates (VarSubsample).

Anaquin calculates a range of summary statistics derived from spike-in controls within an NGS library. For example, Anaquin can measure the minimal expression sufficient for the de novo assembly of RNA isoforms or assess the accuracy of alternative isoform measurements. Anaquin also provides detailed statistics on individual spike-in controls in CSV format that can be easily exported for further investigation.

Finally, Anaquin also generates template code for R. This enables the easy import of spike-in data into R for further analysis with the wide range of statistical and bioinformatics tools available through the Bioconductor project. This includes the ability to quickly visualize spike-in data using scatter-plots (to investigate dependence between variables) and receiver operating characteristic curves (ROC) plots (to assess diagnostic performance; see Supplementary Data **S1** for examples plots provided by Anaquin). Notably, the assessment of spike-ins enables users to optimize input parameters for third-party tools and/or set filtering criteria in order to maximize the performance of their bioinformatic workflow.

It is important to note that the range of possible analysis with spike-in controls is diverse and will continue to expand. Spike-in controls allow empirical evaluation of almost any aspect of the NGS workflow and can inform novel statistical analyses yet to be developed. Accordingly, we anticipate that additional tools will be added to Anaquin in conjunction with continued research and development of spike-in controls.

Funding

The authors would like to thank the following funding sources: T.W is supported by Paramor Family fellowship. I.W.D. and S.A.H. are supported by Australian Postgraduate Award scholarships. T.R.M. is supported by an Australian National Health and Medical Research Council (NHMRC) fellowship (APP1062470).

Conflict of Interest: Garvan Institute of Medical Research has filed patent applications on aspects of spike-in design.

References

Deveson,I.W. *et al.* (2016) Representing genetic variation with synthetic DNA standards. *Nat. Methods*, **13**, 784–791.

Hardwick,S.A. *et al.* (2016) Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat. Methods*, **13**, 792–798.

Jiang,L. *et al.* (2011) Synthetic spike-in standards for RNA-seq experiments. *Genome Res.*, **21**, 1543–1551.

Munro,S.A. *et al.* (2014) Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. *Nat. Commun.*, **5**, 5125.

Zook,J.M. *et al.* (2012) Synthetic spike-in standards improve run-specific systematic error analysis for DNA and RNA sequencing. *PloS One*, **7**, e41356.