RNAsik: A Pipeline for complete and reproducible RNA-seq analysis that runs anywhere with speed and ease

6 February 2018

Summary

RNA sequencing (RNA-seq) is one of many applications of high throughput sequencing, in which millions of short sequence reads, typically around 100 bases long, are produced from RNA samples with the aim of characterising entire transcriptomes. In order to analyse RNA-seq data, multiple bioinformatics tools are collected together into a pipeline, in which each tool accepts processed data from the previous tool as its input. The RNAsik pipeline streamlines processing of RNA-seq data and facilitates the production of reproducible results. This pipeline can run standalone on workstations, cloud instances, or on High Performance Computing (HPC) clusters. A single RNAsik run gives a comprehensive overview of the experiment and produces output suitable for Differential Gene Expression (DGE) analysis.

With an alignment based approach, RNAsik incorporates two main steps: 1) alignment of short reads from FASTQ files to a reference genome or transcriptome; and 2) counting the number of reads mapped to annotated genomic features (such as genes). The table of counts generated by RNAsik can be further analysed with any contemporary count-based DGE tools/packages - one such package particularly suited to the task is Degust (Powell 2015), a powerful front-end and user friendly tool for DGE data analysis, visualisation and exploration. Additionally RNAsik can produce a number of different quality control (QC) metrics, such as sequencing quality metrics reported using FastQC (Bioinformatics 2011), intra- and inter-genic mapping rates estimation using QualiMap (Okonechnikov, Conesa, and García-Alcalde 2016), and sequencing library size and GC bias estimation using Picard Tools (Broadinstitute, n.d.). These QC metrics are automatically summarised into a single, dynamic, HTML report generated by MulitQC (Ewels et al. 2016).

Other features of the RNAsik pipeline include the ability to mark duplicated reads

using Picard Tools (Broadinstitute, n.d.), sorting and indexing of alignments using Samtools (Li et al. 2009) to enable viewing in genome browser applications such as IGV [Robinson2011-du], an enhanced table of counts with additional meta-information about each gene (e.g. biotype and human readable gene names), and ready-to-use coverage plots for every sample using bedtools2 (Quinlan and Hall 2010) and UCSC tools (Raney et al. 2014). The RNAsik pipeline logs every step of processing including the number of samples and associated FASTQ files, software tool versions and sequencing strand information.

RNAsik is written in BigDataScript (BDS) (Cingolani, Sladek, and Blanchette 2015), which is a domain-specific language (DSL). BDS generates an additional HTML report alongside a typical RNAsik analysis. In addition to RNAsik internal logging, this report holds system information such as run-time information and the exit status for every tool. RNAsik employs many other useful features within BDS such as inbuilt checkpointing for retries on failure and ability to talk to an HPC cluster queue directly.

RNAsik incorporates commonly used tools such as STAR aligner (Dobin et al. 2013), featureCounts (Liao, Smyth, and Shi 2014) and samtools (Li et al. 2009). However, one can extend RNAsik with other new tools and features. Recently, RNAsik has been extended to include two other aligners, Hisat2 (Kim, Langmead, and Salzberg 2015) and BWA-MEM (Li 2013). This broadens the scope of RNAsik to bacterial RNA-seq analysis and improves diversity. RNAsik is an open-source project under Apache License 2.0, and contributions are welcome. In the near future, there are plans to extend RNAsik in several directions including the incorporation of alignment-free read quantification and an RNA-seq variant calling option (Sun et al. 2016). RNAsik simplifies and speeds up RNA-seq analysis and automates many of the QC steps that are important but often overlooked.

References

Bioinformatics, Babraham. 2011. "FastQC: A Quality Control Tool for High Throughput Sequence Data." *Cambridge, UK: Babraham Institute*.

Broadinstitute. n.d. "Picard-Tools." broadinstitute. http://broadinstitute.github.io/picard.

Cingolani, Pablo, Rob Sladek, and Mathieu Blanchette. 2015. "BigDataScript: A Scripting Language for Data Pipelines." *Bioinformatics* 31 (1): 10–16. http://dx.doi.org/10.1093/bioinformatics/btu595.

Dobin, Alexander, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. 2013. "STAR: Ultrafast Universal RNA-seq Aligner." *Bioinformatics* 29 (1): 15–21.

http://dx.doi.org/10.1093/bioinformatics/bts635.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. "MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report." *Bioinformatics* 32 (19): 3047–8. http://dx.doi.org/10.1093/bioinformatics/btw354.

Kim, Daehwan, Ben Langmead, and Steven L Salzberg. 2015. "HISAT: A Fast Spliced Aligner with Low Memory Requirements." *Nat. Methods* 12 (4): 357–60. http://dx.doi.org/10.1038/nmeth.3317.

Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM," March. http://arxiv.org/abs/1303.3997.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–9. http://dx.doi.org/10.1093/bioinformatics/btp352.

Liao, Yang, Gordon K Smyth, and Wei Shi. 2014. "FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features." *Bioinformatics* 30 (7): 923–30. http://dx.doi.org/10.1093/bioinformatics/btt656.

Okonechnikov, Konstantin, Ana Conesa, and Fernando García-Alcalde. 2016. "Qualimap 2: Advanced Multi-Sample Quality Control for High-Throughput Sequencing Data." *Bioinformatics* 32 (2). academic.oup.com: 292–94. http://dx.doi.org/10.1093/bioinformatics/btv566.

Powell, David. 2015. "Degust: Powerfull and User Friendly Front-End Data Analsysis, Visualisation and Exploratory Tool for Rna-Sequencing." github. http://degust.erc.monash.edu.

Quinlan, Aaron R, and Ira M Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42. http://dx.doi.org/10.1093/bioinformatics/btq033.

Raney, Brian J, Timothy R Dreszer, Galt P Barber, Hiram Clawson, Pauline A Fujita, Ting Wang, Ngan Nguyen, et al. 2014. "Track Data Hubs Enable Visualization of User-Defined Genome-Wide Annotations on the UCSC Genome Browser." *Bioinformatics* 30 (7). academic.oup.com: 1003–5. http://dx.doi.org/10.1093/bioinformatics/btt637.

Sun, Zhifu, Aditya Bhagwate, Naresh Prodduturi, Ping Yang, and Jean-Pierre A Kocher. 2016. "Indel Detection from RNA-seq Data: Tool Evaluation and Strategies for Accurate Detection of Actionable Mutations." *Brief. Bioinform.*, July. http://dx.doi.org/10.1093/bib/bbw069.