New RNA-seq Bioinformatics Best Practice

Introduction

We process more RNA-seq samples at the NGI than any other type of experiment. The sequencing data from all RNA projects is passed through an analysis pipeline to check quality control metrics. If the user has requested "best practice analysis", this data is then delivered as a starting point for their downstream analysis.

The existing RNA-seq analysis pipeline has been used at the NGI Stockholm since 2012. It has proved to be robust and reliable and has been used by many research groups across Sweden. Aside from incremental software updates, it has remained largely unchanged as the NGI has grown around it.

In April 2017 NGI Stockholm is moving to a brand new RNA-seq analysis pipeline, rebuilt from the ground up. The reasons for doing this were twofold: first, the RNA field had progressed since the pipeline was written and there are newer tools are now considered to be “best practice”. We have heard anecdotal evidence that many groups were re-running data through their own processing pipelines because of this. Secondly, the old pipeline was very slow. Projects typically took several days to run and used a large amount of compute power. By swapping a few steps with newer, faster alternatives we can decrease the compute requirements considerably.

Pipeline tools

The steps in the old and new pipelines are shown in Table 1. Some are unchanged, some have been swapped for alternative tools and some are new.

Migrating to the new pipeline

Whilst many of the new tools run similar tasks to those that they replace, results may not be directly comparable. In our validations of the new pipeline, we found that the alignments and gene counts were highly correlated between the old and new pipelines. However, FPKM results from the new StringTie software are not directly comparable to those generated previously with Cufflinks. For this reason, we recommend that old data is reprocessed with the new pipeline before comparison.

Alternatively, you can run Cufflinks v2.1.1 on data from the new pipeline with the following command (where `alignments.bam` are your aligned reads in the results/STAR directory):

```
cufflinks \
  -p 8 \
  --library-type fr-firststrand \
  -G genes.gtf \
  -o cufflinks_out_[sampleID] \
  alignments.bam
```

Old Pipeline	New Pipeline	Description
FastQC	FastQC	*Quality control (raw)*
-	Trim Galore!	*Adapter & quality trimming*
TopHat	STAR	*Alignment*
HTSeq	featureCounts	*Gene counts*
Cufflinks	StringTie	*Normalised FPKM*
RSeQC	RSeQC	*Quality control (alignments)*
-	dupRadar	*Quality control (duplication)*
Preseq	Preseq	*Library complexity*
Pheatmap (R)	Heatmap (edgeR)	*Sample similarity*
-	edgeR	*MDS plot & distance tree*
NGI Reports	MultiQC	*Reporting*

Table 1. *Old and new pipeline tools, along with their role in the pipeline. Note that new RSeQC commands have been added in the new pipeline. See References below for more information.*

Workflow software

The old pipeline was run using a custom Python script which is firmly tied to the NGI computational architecture and workflow. This means that it is essentially impossible for anyone else to run the workflow. For the new pipeline, we have used Nextflow to handle all of the background pipeline management. This tool allows much greater flexibility and ease of use - anyone can now run our best-practice analysis pipeline themselves on virtually any compute infrastructure.

By default, the new pipeline has been written to run on UPPMAX HPC systems such as *milou* or *bianca*. However, with a single command line addition, the pipeline can run locally instead. There is also built-in support for Docker, meaning that all dependencies can be automatically handled in a highly reproducible manner. Custom configuration files can be used to get the pipeline to run on virtually any compute cluster software, or even in the cloud. Please see the documentation for more information.

More information

For more information about the new pipeline, you can find it and its documentation on GitHub: https://github.com/SciLifeLab/NGI-RNAseq.

A HTML file is delivered with every project describing the output generated by the pipeline (also available directly on GitHub).

If you have any questions, please do not hesitate to get in touch at support@ngisweden.se

References

Useful links:

- NGI-RNAseq pipeline code
  - https://github.com/SciLifeLab/NGI-RNAseq
- Open-source software at SciLifeLab
  - http://opensource.scilifelab.se/
- NGI order portal
  - https://ngisweden.scilifelab.se/

Tools used:

- FastQC
  - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- Trim Galore!
  - https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
- STAR
  - https://github.com/alexdobin/STAR
  - DOI:10.1093/bioinformatics/bts635
- Subread featureCounts
  - http://bioinf.wehi.edu.au/featureCounts/
  - DOI: 10.1093/bioinformatics/btt656
- RSeQC
  - http://rseqc.sourceforge.net/
  - DOI: 10.1093/bioinformatics/bts356
- dupRadar
  - https://bioconductor.org/packages/dupRadar/
- Preseq
  - http://smithlabresearch.org/software/preseq/
- edgeR
  - https://bioconductor.org/packages/edgeR/
  - DOI: 10.1093/bioinformatics/btp616
- MultiQC
  - http://multiqc.info/
  - DOI: 10.1093/bioinformatics/btw354
- Nextflow
  - https://www.nextflow.io/
- Docker
  - https://www.docker.com/

NGI-RNAseq