Digital Expression Explorer 2: a repository of 4.5 trillion uniformly processed RNA-seq reads and counting

Mark Ziemann*[12], Antony Kaspi[2], Assam El-Osta[23] * m.ziemann@deakin.edu.au

(1) Deakin University, Geelong, Australia, School of Life and Environmental Sciences. (2) Department of Diabetes, Monash University Central Clinical School, The Alfred Medical Research and Education Precinct, Melbourne, Vic, Australia. (3) Hong Kong Institute of Diabetes and Obesity, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong SAR.

Introduction

Data from tens of thousands of transcriptome sequencing projects have been deposited to NCBI GEO [1], however comparing two or more projects is challenging due to the inconsistent processing pipelines used.

To alleviate these obstacles we developed DEE2. It provides expression data in the form of counts that are compatible with many types of downstream analysis. DEE2 covers more projects, experiments and runs than other repositories and will be a valuable resource for meta-analysis of transcriptome data.

Dataset

DEE2 hosts over 470,000 RNA-seq data sets (SRA runs) derived from nine species. DEE2 consists of over 4.5 trillion assigned sequence reads (Table 1). The data provided include gene-wise expression counts and transcript-wise counts alongside QC metrics and detailed analysis logs.

Table 1. Hosted gene expression data as of Nov 2018.

Species	Projects	Experiments	Runs	Assigned reads STAR	Assigned reads Kallisto
A. thaliana	924	14514	23183	2.38E+11	2.43E+11
C. elegans	285	5491	7423	8.09E+10	7.32E+10
D. melanogaster	620	13850	18118	1.67E+11	1.79E+11
D. rerio	423	23979	25692	9.26E+10	9.15E+10
E. coli	176	1467	1617	1.25E+10	9.35E+09
H. sapiens	5787	151120	175049	1.79E+12	1.97E+12
M. musculus	5737	174576	208579	1.64E+12	1.85E+12
R. norvegicus	337	4144	5046	5.01E+10	5.37E+10
S. cerevisiae	440	10236	11366	7.40E+10	7.32E+10
Total	14729	399377	476073	4.14E+12	4.55E+12

Pipeline features

The pipeline is implemented as a shell script in a docker container. It obtains raw data from NCBI SRA and performs quality controls including adapter detection and removal before mapping with STAR and Kallisto to generate gene and transcript counts (Fig 1). The pipeline can even be used to process private fastq files.

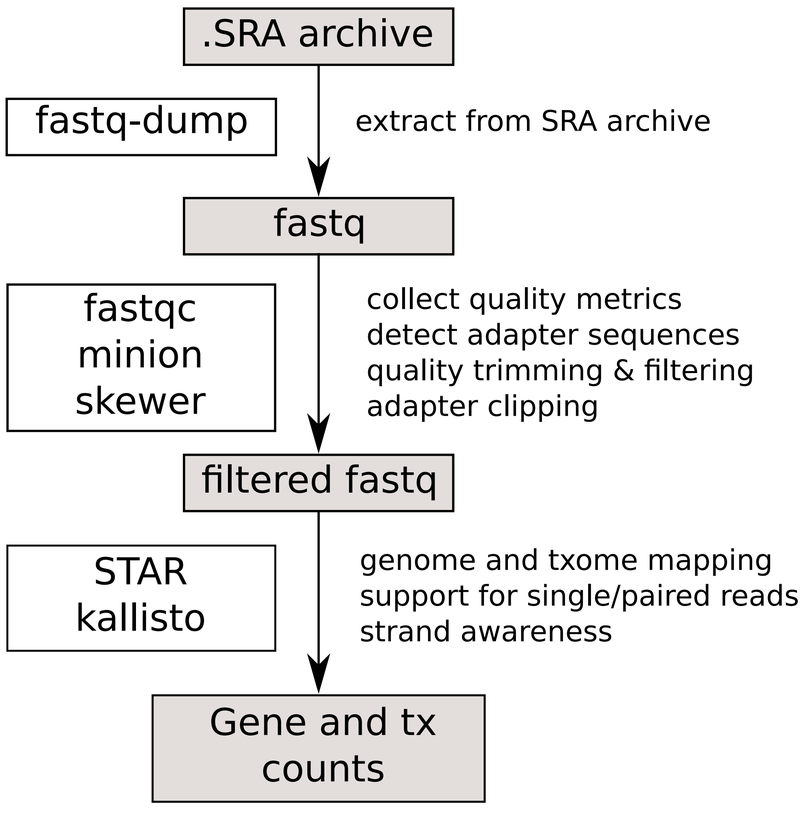

Fig 1. Overview of steps in the RNA-seq data processing pipeline.

Quality control

To demonstrate the accuracy of the pipeline, we performed a simulation study. Synthetic Illumina HiSeq RNA-seq data were generated from Ensembl transcripts and processed with the pipeline. The reads per million (RPM) values were compared between the simulated and processed data and Spearman correlation coefficients (ρ) were calculated; showing that kallisto gene level counts were more accurate than STAR gene level counts (Fig 2).

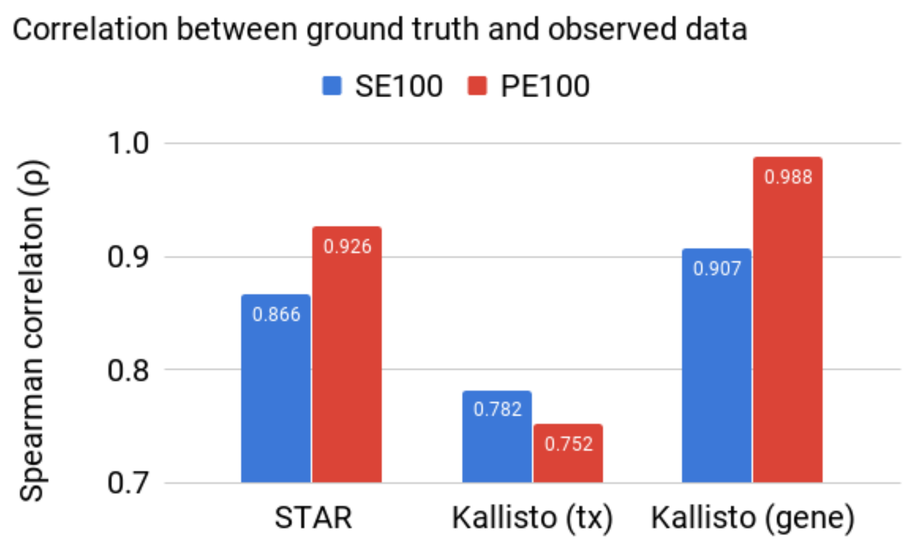

Fig 2. Simulated human RNA-seq reads processed with the DEE2 pipeline.

Conclusion

There is in excess of US$136 million worth of RNA-seq data deposited to NCBI SRA but its reuse potential is yet to be met. DEE2 and other aggregation initiatives such as ARCHS4 [4] and Recount2 [5] will ensure that these uniformly processed data are freely available.

How to access the data & code

Point your web browser to **http://dee2.io** and provide GEO or SRA accession numbers or keywords to identify datasets of interest.

The website also hosts bulk dumps of all datasets:
http://dee2.io/bulk.html

A custom R script is provided to enable programmatic access to datasets that is documented on the GitHub repo (it also contains all other source code): **https://github.com/markziemann/dee2**

The pipeline is available as a Docker image:
https://hub.docker.com/r/mziemann/tallyup/

License: GPL v3.

Acknowledgements

This research was made possible by use of the Multi-modal Australian ScienceS Imaging and Visualisation Environment (**MASSIVE**) and **NeCTAR Research Cloud**, both supported by the Australian National Collaborative Research Infrastructure Strategy (NCRIS). This work was supported by Deakin eResearch and Monash eResearch Centres. We thank Dr Ross Lazarus and Dr Haloom Rafehi for bioinformatics expertise and helpful discussions. We thank Julian Vreugdenburg for technical support. We thank the many users that have provided feedback on earlier versions of DEE2.

References

[1] Barrett et al, 2013. DOI: 10.1093/nar/gks1193
[2] Dobin et al, 2013. DOI: 10.1093/bioinformatics/bts635
[3] Bray et al, 2016. DOI: 10.1038/nbt.3519
[4] Lachmann et al, 2018. DOI: 10.1038/s41467-018-03751-6
[5] Collado-Torres et al, 2017. DOI: 10.1038/nbt.3838