GeneTEFlow: A Nextflow-based pipeline for analysing gene and transposable elements expression from RNA-Seq data

Xiaochuan Liu[1], Jadwiga R Bienkowska[2], and Wenyan Zhong[1]

[1]Oncology Research & Development, Pfizer Worldwide Research and Development, Pearl River, NY 10965, USA

[2]Oncology Research & Development, Pfizer Worldwide Research and Development, San Diego, CA 92121, USA

Corresponding authors:

wenyan.zhong@pfizer.com

Jadwiga.R.Bienkowska@pfizer.com

Abstract

Transposable elements (TEs) are mobile genetic elements in eukaryotic genomes. Recent research highlights the important role of TEs in the embryogenesis, neurodevelopment, and immune functions. However, there is a lack of a one-stop and easy to use computational pipeline for expression analysis of both genes and locus-specific TEs from RNA-Seq data. Here, we present GeneTEFlow, a fully automated, reproducible and platform-independent workflow, for the comprehensive analysis of gene and locus-specific TEs expression from RNA-Seq data employing Nextflow and Docker technologies. This application will help researchers more easily perform integrated analysis of both gene and TEs expression, leading to a better understanding of roles of gene and TEs regulation in human diseases. GeneTEFlow is freely available at https://github.com/zhongw2/GeneTEFlow.

Introduction

Transposable elements (TEs) are mobile DNA sequences which have the capacity to move from one location to another on the genome[1]. TEs make up a considerable fraction of most eukaryotic genomes and can be classified into retrotransposons and DNA transposons according to their different mechanisms of transposition and chromosomal integration[2, 3]. Retrotransposons are made of Long Terminal Repeats (LTRs) and non-LTRs that include long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) that mobilize via a RNA intermediate, while DNA transposons mobilize and function through a DNA intermediate[4-6]. TEs can be transcribed from the genome[7] and have been demonstrated to play important roles in the mammalian embryogenesis[8, 9], neurodevelopment[10, 11], and immune functions[12, 13]. Furthermore, aberrant expressions of TEs have been linked to

cancers[14-16], neurodegenerative disorders[17, 18], and immune-mediated inflammation[19, 20]. Therefore, it has become increasingly important to explore biological roles of TEs expression. However, genome-wide analysis of TEs expression from high throughput RNA sequencing data has been a challenging computational problem. TEs contain highly repetitive sequence elements, making it arduous to unambiguously assign reads to the correct genomic location and accurately quantitate their expression level. Several bioinformatics tools have been developed to address this challenge with relatively good success [16, 21-23]. Recently, SQuIRE was reported to have the capability to quantify locus-specific expression of TEs from RNA-Seq data[23]. In addition, RNA-Seq data has long been used to detect dysregulated genes between different disease and/or drug treatment conditions to help understand disease mechanisms and/or drug response mechanisms. Therefore, it is of great interest to quantify both TEs and gene expression to elucidate contribution of both to disease mechanisms. Although many open source software and tools exist for analysing gene [24-26] and TEs expression, there are considerable challenges to efficiently apply these tools. In general, these multi-step data processing pipelines use many different tools. Correct versions of each tool need to be installed separately, and appropriate options, parameters, different reference genome and gene annotation files have to be set at each step. This can be quite tedious and challenging especially for non-computational users. Additionally, to ensure reproducibility of the analysis results, it is critical to capture analysis parameters from each step of the process. Equally important, to enable general use of the pipeline, the pipeline should be platform agnostic. Thus far, a one-stop computational framework for the comprehensive analysis of gene and locus-specific TEs expression from RNA-Seq data does not exist.

To address this need, we developed GeneTEFlow, a reproducible and platform-independent workflow, for the comprehensive analysis of gene and locus-specific TEs expression from RNA-Seq data using Nextflow[27] and Docker[28] technologies. GeneTEFlow provides several features and advantages for integrated gene and TEs transcriptomic analysis. First, by employing Docker technology, GeneTEFlow encapsulates bioinformatics tools and applications of specific versions into Docker containers enabling tracking, eliminating the need for software installation by users, and ensuring portability of the pipeline on multiple computing platforms including stand-alone workstations, high-performance computing (HPC) clusters, and cloud computing systems. Second, GeneTEflow uses Nextflow to define the computational workflows, not only enabling parallelization and complete automation of the analysis, but also providing capability to track analysis parameters. Thus, GeneTEFlow allows users to generate reproducible analysis results through utilization of both Docker and Nextflow in a platform independent manner. Lastly, GeneTEFlow has modular architecture, and modules in GeneTEFlow can be turned on or off, providing developers with flexibility to build extensions tailored to specific analysis needs.

Implementation

The GeneTEFlow pipeline was developed using Nextflow, a portable, flexible, and reproducible workflow management system, and Docker technology, a solution to securely build and run applications on multiple platforms. To build the GeneTEFlow pipeline, a series of bioinformatics tools (S1 Table) were selected for QC, quantitation and differential expression analysis of genes and TEs from RNA-Seq data. These bioinformatics tools and custom scripts were built into four Docker containers to ensure portability of the workflow on different computational platforms. Data processing and analysis steps were implemented by modules

using Nextflow. Modules are connected through channels and can be run in parallel. Each module in GeneTEFlow can include any executable Linux scripts such as Perl, R, or Python. Parameters for each module are defined in a configuration file.

A conceptual workflow of GeneTEFlow is illustrated in Fig 1. The workflow includes four major inputs: raw sequence files in fastq format, a sample meta data file in excel format, reference genome and gene annotation files, and a Nextflow configuration file. The sample meta data file contains detailed sample information and the design of group comparisons between different experimental conditions. Human reference genome UCSC hg38 with the gene annotation (.gtf) was downloaded from Illumina iGenomes collections[29] and used by the bioinformatics tools included in GeneTEFlow. Scheduling of computational resources for each application module is defined in the configuration file.

Fig 1. Illustration of GeneTEFlow: a Nexflow-based pipeline for identification of differentially expressed genes and locus specific transposable elements from RNA-Seq data.

GeneTEFlow analysis is performed in following steps: QC, expression quantification, differential expression and down-stream analysis. First, adapter sequences are trimmed off from the Illumina raw reads using Trimmomatic(v0.36)[30] for single-end or paired-end reads, and low-quality reads are filtered out. Next, FastQC(v0.11.7)[31] is executed to survey the quality of sequencing reads, and report is generated to help identify any potential issues of the high throughput sequencing data. Reference genome index for mapping sequencing reads to mRNA genes is built using “rsem-prepare-reference” of RSEM (v.1.3.0). Reads remaining after the pre-processing step are mapped to the reference genome using STAR(v2.6.0c)[32]. Gene level

123 expression is quantitated as expected counts and transcripts per million (TPM) using "rsem-
124 calculate-expression" of RSEM(v1.3.0) with default parameters [33]. Custom Perl scripts were
125 developed to aggregate data from each sample into a single data matrix for expected counts and
126 TPM values respectively. The expression quantification of locus-specific TEs is performed by
127 SQuIRE[23].

128 In addition, we also implemented quality control measures after reads alignment step to
129 detect potential outlier samples resulted from experimental errors. Boxplot and density plot are
130 used to evaluate the overall consistency of the expression distribution for each sample. Sample
131 correlation analysis is performed with Pearson method using TPM values to assess the
132 correlation between biological replicates from each sample group. Principal component analysis
133 (PCA) is employed to identify potential outlier samples and to investigate relationships among
134 sample groups.

135 Differential expression analysis of genes and transposable elements is performed using
136 DESeq2(v1.18.1) package[34]. Significantly up-regulated and down-regulated genes and TEs are
137 summarized in a table. To analyse overlap among significantly regulated genes and TEs from
138 pair-wise comparisons between different sample groups we use Venn diagrams. We perform
139 hierarchical clustering of significantly dysregulated genes or TEs using R package
140 "ComplexHeatmap" [35] with euclidean distance and average linkage clustering parameters.
141 Gene set enrichment analysis (GSEA, v3.0) [36] is conducted using collections from the
142 Molecular Signatures Database (MSigDB) [37]. The outputs (S2 Table) from GeneTEFlow are
143 organized into several folders predefined in a GeneTEFlow configuration file. A tutorial with
144 detailed instructions on how to set up and run GeneTEFlow is provided at
145 https://github.com/zhongw2/GeneTEFlow

Application of GeneTEFlow

We applied GeneTEFlow to a public dataset from Brawan's study [38] investigating tissue-specific expression changes of genes and transposable elements. Human RNA-Seq data from brain, heart and testis tissues were downloaded from GEO (accession number: GSE30352) (S3 Table). Expression analysis of genes and TEs were performed using GeneTEFlow and results are shown in Fig 2. Gene expression analysis was performed using RSEM and DESeq2 modules while TEs expression analysis was conducted using SQuIRE and DESeq2 modules within GeneTEFlow. Significantly regulated genes were identified with FDR less than 0.05 and fold change greater than 2. Significantly regulated locus-specific transposable elements were identified with FDR less than 0.05 and fold change greater than 1.5. The number of significantly regulated genes and transposable elements were summarized into two tables respectively (Fig 2, top panels). Using GeneTEFlow, we detected genes and TE differentially expressed between different tissue types (brain vs heart tissues: 6,264 genes and 1,277 TEs; testis vs heart tissues: 7,066 genes and 595 TE; brain vs testis tissues: 8,125 genes and 1,297 TEs) with most significant gene and TE expression differences observed being between brain and testis tissues. Our analysis identified large number of both genes and TEs with tissue specific patterns (Fig 2, middle panels and bottom panels). More in depth analysis to include additional tissue types would be required to fully understand the tissue specific gene and TEs expression and their relationship. GeneTEFlow is a computational solution to facilitate such studies.

Fig 2. Differential expression analysis results of genes and transposable elements from GeneTEFlow. Left panels: gene results; right panels: TEs results. Top panels: number of significantly regulated genes or TEs in each sample group comparison. Significance was defined

as following: FDR ≤ 0.05 and fold change ≥ 2 for gene expression analysis; FDR ≤ 0.05 and fold change ≥ 1.5 for TEs expression analysis. Middle panels: overlaps of significantly regulated genes or TEs amongst sample group comparisons. Bottom panels: hierarchical clustering of significantly regulated genes or TEs.

In addition to quantification of TEs expression, SQuIRE provides quantification of gene expression. Therefore, we compared gene level expression quantification between RSEM and SQuIRE (S1 Fig). The results showed high concordance (correlation coefficient: ~97%) of the gene level expression quantification between the two methods (S1 Fig, highlighted in red box) suggesting a robust measurement for both gene and TEs expression by SQuIRE.

Conclusions

In conclusion, we have developed and made available an automated pipeline to comprehensively analyse both gene and locus-specific TEs expression from RNA-Seq data. Taking advantage of the advanced functionalities provided by Nextflow and Docker, GeneTEFlow allows users to run analysis reproducibly on different computing platforms without the need for individual tool installation and manual version tracking. We believe this pipeline will be of great help to further our understanding of roles of both gene and TEs regulation in human diseases. This pipeline is flexible and can be easily extended to include additional types of analysis such as alternative splicing, fusion genes, and so on.

Competing interests

WZ and JRB are employees of Pfizer Inc.
XL was contractor of Pfizer Inc. when the work was being conducted.

Funding

Not applicable.

Authors' contributions

WZ conceptualized the work. XL and WZ designed and implemented the pipeline. XL, JRB and WZ drafted and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We gratefully acknowledge inputs and support from our colleagues: Jeremy Myers, Keith Ching, Corey Dasilva and Da Tse.

References

1. Biémont C, Vieira C. Junk DNA as an evolutionary force. Nature. 2006;443(7111):521-4. doi: 10.1038/443521a.
2. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics. 2007;8(12):973-82. doi: 10.1038/nrg2165.
3. Feschotte C. Transposable elements and the evolution of regulatory networks. Nature Reviews Genetics. 2008;9(5):397-405. doi: 10.1038/nrg2337.
4. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. Genome Biology. 2018;19(1):199. doi: 10.1186/s13059-018-1577-z.
5. Lanciano S, Mirouze M. Transposable elements: all mobile, all different, some stress responsive, some adaptive? Current Opinion in Genetics & Development. 2018;49:106-14. doi: https://doi.org/10.1016/j.gde.2018.04.002.
6. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nature Reviews Genetics. 2017;18(2):71-86. doi: 10.1038/nrg.2016.139.
7. Rebollo R, Romanish MT, Mager DL. Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. Annual Review of Genetics. 2012;46(1):21-42. doi: 10.1146/annurev-genet-110711-155621. PubMed PMID: 22905872.

8. Percharde M, Lin C-J, Yin Y, Guan J, Peixoto GA, Bulut-Karslioglu A, et al. A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. Cell. 2018;174(2):391-405.e19. doi: https://doi.org/10.1016/j.cell.2018.05.043.
9. Garcia-Perez JL, Widmann TJ, Adams IR. The impact of transposable elements on mammalian development. Development. 2016;143(22):4101-14. doi: 10.1242/dev.132639.
10. Sun W, Samimi H, Gamez M, Zare H, Frost B. Pathogenic tau-induced piRNA depletion promotes neuronal death through transposable element dysregulation in neurodegenerative tauopathies. Nature Neuroscience. 2018;21(8):1038-48. doi: 10.1038/s41593-018-0194-1.
11. Guo C, Jeong H-H, Hsieh Y-C, Klein H-U, Bennett DA, De Jager PL, et al. Tau Activates Transposable Elements in Alzheimer's Disease. Cell Reports. 2018;23(10):2874-80. doi: https://doi.org/10.1016/j.celrep.2018.05.004.
12. Colombo AR, Elias HK, Ramsingh G. Senescence induction universally activates transposable element expression. Cell Cycle. 2018;17(14):1846-57. doi: 10.1080/15384101.2018.1502576.
13. Koonin EV, Krupovic M. Evolution of adaptive immunity from transposable elements combined with innate immune systems. Nature Reviews Genetics. 2015;16(3):184-92. doi: 10.1038/nrg3859.
14. Colombo AR, Triche T, Ramsingh G. Transposable Element Expression in Acute Myeloid Leukemia Transcriptome and Prognosis. Scientific Reports. 2018;8(1):16449. doi: 10.1038/s41598-018-34189-x.
15. Burns KH. Transposable elements in cancer. Nature Reviews Cancer. 2017;17(7):415-24. doi: 10.1038/nrc.2017.35.
16. Criscione SW, Zhang Y, Thompson W, Sedivy JM, Neretti N. Transcriptional landscape of repetitive elements in normal and cancer human cells. BMC Genomics. 2014;15(1):583. doi: 10.1186/1471-2164-15-583.
17. Krug L, Chatterjee N, Borges-Monroy R, Hearn S, Liao W-W, Morrill K, et al. Retrotransposon activation contributes to neurodegeneration in a Drosophila TDP-43 model of ALS. PLOS Genetics. 2017;13(3):e1006635. doi: 10.1371/journal.pgen.1006635.
18. Tam OH, Ostrow LW, Gale Hammell M. Diseases of the nERVous system: retrotransposon activity in neurodegenerative disease. Mobile DNA. 2019;10(1):32. doi: 10.1186/s13100-019-0176-1.
19. De Cecco M, Ito T, Petrashen AP, Elias AE, Skvir NJ, Criscione SW, et al. L1 drives IFN in senescent cells and promotes age-associated inflammation. Nature. 2019;566(7742):73-8. doi: 10.1038/s41586-018-0784-9.
20. Colombo AR, Elias HK, Ramsingh G. Senescence induction universally activates transposable element expression. Cell cycle (Georgetown, Tex). 2018;17(14):1846-57. Epub 08/16. doi: 10.1080/15384101.2018.1502576. PubMed PMID: 30080431.
21. Jin Y, Tam OH, Paniagua E, Hammell M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics. 2015;31(22):3593-9. doi: 10.1093/bioinformatics/btv422.
22. Lerat E, Fablet M, Modolo L, Lopez-Maestre H, Vieira C. TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. Nucleic Acids Research. 2016;45(4):e17-e. doi: 10.1093/nar/gkw953.
23. Yang WR, Ardeljan D, Pacyna CN, Payer LM, Burns KH. SQuIRE reveals locus-specific regulation of interspersed repeat expression. Nucleic Acids Research. 2019;47(5):e27-e. doi: 10.1093/nar/gky1301.

24. Varet H, Brillet-Guéguen L, Coppée J-Y, Dillies M-A. SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data. PLOS ONE. 2016;11(6):e0157022. doi: 10.1371/journal.pone.0157022.
25. Finotello F, Di Camillo B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. Briefings in Functional Genomics. 2014;14(2):130-42. doi: 10.1093/bfgp/elu035.
26. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nature Biotechnology. 2020;38(3):276-8. doi: 10.1038/s41587-020-0439-x.
27. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nature Biotechnology. 2017;35(4):316-9. doi: 10.1038/nbt.3820.
28. Merkel D. Docker: lightweight Linux containers for consistent development and deployment. Linux J. 2014;2014(239):Article 2.
29. iGenomes : https://support.illumina.com/sequencing/sequencing_software/igenome.html.
30. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20. Epub 04/01. doi: 10.1093/bioinformatics/btu170. PubMed PMID: 24695404.
31. FastQC : https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
32. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2012;29(1):15-21. doi: 10.1093/bioinformatics/bts635.
33. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12(1):323. doi: 10.1186/1471-2105-12-323.
34. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15(12):550. doi: 10.1186/s13059-014-0550-8.
35. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32(18):2847-9. doi: 10.1093/bioinformatics/btw313.
36. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;102(43):15545-50. doi: 10.1073/pnas.0506580102.
37. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739-40. doi: 10.1093/bioinformatics/btr260.
38. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011;478(7369):343-8. doi: 10.1038/nature10532.

Supporting information

S1 Fig. Comparison of gene expression quantification by RSEM and SQuIRE. Gene expression (total 22,955 genes) of samples from brain tissues (left), heart tissues (middle), and testis tissues (right) was calculated by both RSEM and SQuIRE. Lower diagonal panels: pairwise comparisons using log2(TPM + 1) of 22,955 genes. Upper diagonal panels: correlation coefficient of each comparison. Panels highlighted in red: correlation coefficient of comparisons between RSEM and SQuIRE gene expression quantification of the same sample. Rep_: replicate, _RSEM: quantification performed by RSEM, _SQuIRE: quantification performed by SQuIRE.

S1 Table. Major bioinformatics tools installed in GeneTEFlow

S2 Table. Major outputs from GeneTEFlow

S3 Table. Human RNA-Seq data used in the example application of GeneTEFlow

S1_File. Supplemental tables: S1-S3 Tables.

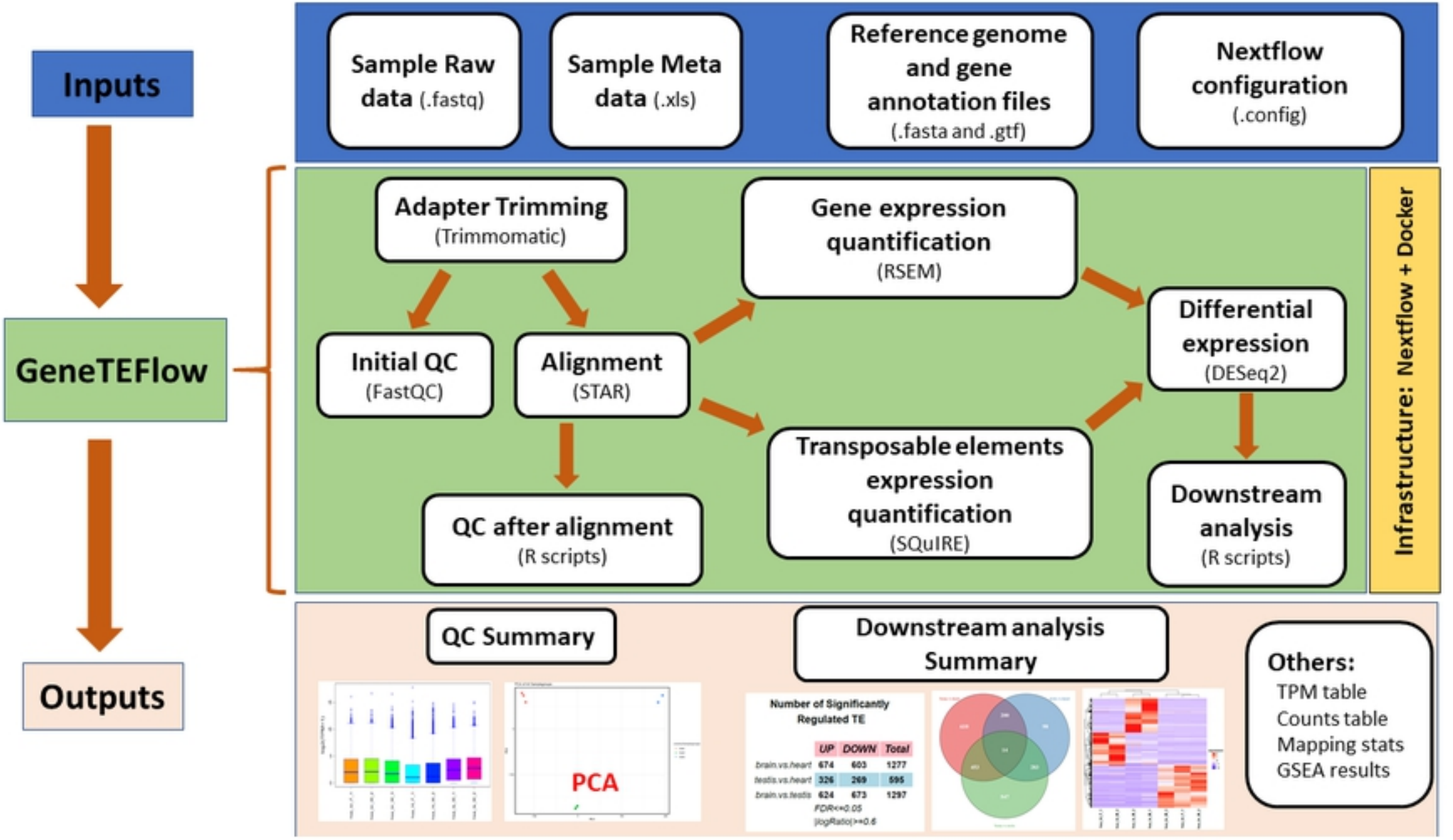

Inputs
Sample Raw data (.fastq)
Sample Meta data (.xls)
Reference genome and gene annotation files (.fasta and .gtf)
Nextflow configuration (.config)
GeneTEFlow
Adapter Trimming (Trimmomatic)
Initial QC (FastQC)
Alignment (STAR)
QC after alignment (R scripts)
Gene expression quantification (RSEM)
Transposable elements expression quantification (SQuIRE)
Differential expression (DESeq2)
Downstream analysis (R scripts)
Infrastructure: Nextflow + Docker
Outputs
QC Summary
PCA
Downstream analysis Summary
Number of Significantly Regulated TE
UP DOWN Total
brain.vs.heart 674 603 1277
testis.vs.heart 326 269 595
brain.vs.testis 624 673 1297
FDR<=0.05
|logRatio|>=0.6
Others:
TPM table
Counts table
Mapping stats
GSEA results

Fig 1

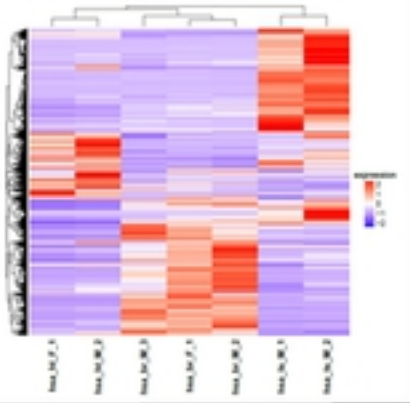
Genes

Number of Significantly Regulated Genes

	UP	DOWN	Total
brain.vs.heart	3374	2890	6264
testis.vs.heart	4127	2939	7066
brain.vs.testis	3802	4413	8215

FDR<=0.05

|logRatio|>=1



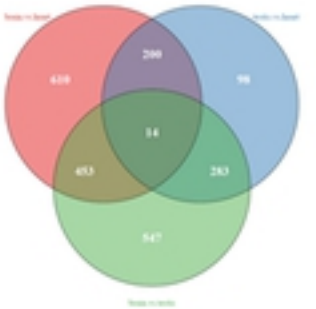
Transposable Elements

Number of Significantly Regulated TE

	UP	DOWN	Total
brain.vs.heart	674	603	1277
testis.vs.heart	326	269	595
brain.vs.testis	624	673	1297

FDR<=0.05

|logRatio|>=0.6



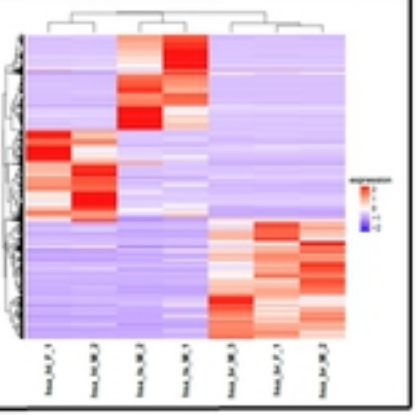


Fig 2