Dave Adelson Lab Placement
Complete Notebook

PDF Version generated by

Ha Tran (a1743091@adelaide.edu.au)

on

Nov 13, 2021 @12:32 AM AEDT

Table of Contents

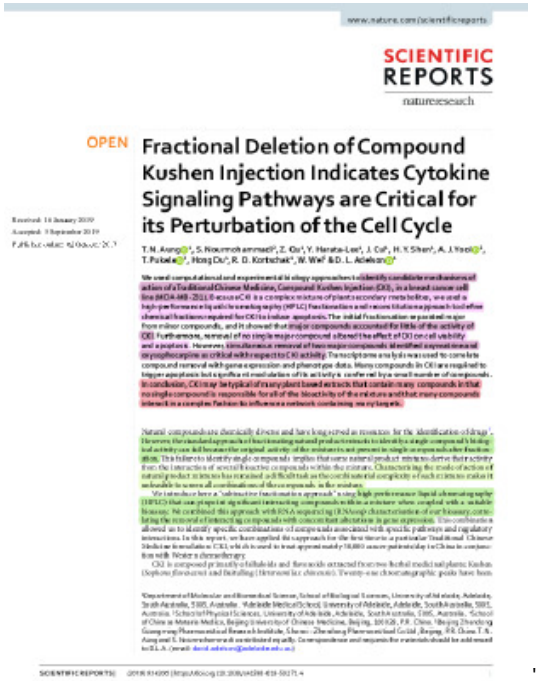
Aung et al., 2019. Fractional Deletion of Compound Kushen Injection Indicates Cytokine signalling Pathways are Critical for its Perturbation of the Cell Cycle

Ha Tran (a1743091@adelaide.edu.au) - Aug 31, 2021, 3:00 AM AEST

Aung,T N; Nourmohammadi,S; Qu,Z; Harata-Lee,Y; Cui,J; Shen,H Y; Yool,A J; Pukala,T; Du,Hong; Kortschak,R D; Wei,W; Adelson,D L (2019). Fractional Deletion of Compound Kushen Injection Indicates Cytokine Signaling Pathways are Critical for its Perturbation of the Cell Cycle. *Sci Rep*, 9(1) 14200. **Abstract**- We used computational and experimental biology approaches to identify candidate m **more...**

View in : ▪ PubMed ▪ Sci Rep ▪ PubMed Central

Ha Tran (a1743091@adelaide.edu.au) - Aug 31, 2021, 2:54 AM AEST



www.nature.com/scientificreports

SCIENTIFIC REPORTS

nature research

OPEN

Fractional Deletion of Compound Kushen Injection Indicates Cytokine Signaling Pathways are Critical for its Perturbation of the Cell Cycle

_Aung_et_al._2019_._Fractional_Deletion_of_COmpound_Kushen_Injection_Indicates_Cytokine_SIgnaling_Pathways_are_Critical_for_its_Perturbation_of_the_Cell_Cycle.pdf(2.3 MB) - download

(Qu, Z et al., 2016). Identification of candidate anti-cancer molecular mechanisms of Compound Kushen Injection using functional genomics.

edgeR

Ha Tran (a1743091@adelaide.edu.au) - Nov 12, 2021, 7:29 AM AEDT

RNA seq two source of variation
Biological variation and measurement errors

Negative binomial distribution to account for the two sources of errors

Dealing with a small number of samples, solution was empirical bayes statistical model

QLF method control FDR a lot better and a thus recommended

IMPOSING A FC CUTOFF on the DE results would violate the FDR,

A rigorous statistical testing strategy is required for differential expression relative to a FC threshold. In edgeR, this can be done with the TREAT method

TREAT used for detecting biologically significant DE gene when there are already numerous genes

EdgeR has many other function such as gene set testing, GO term analysis, KEGG pathways analysis.
EdgeR can also perform transcript level differential expression analysis.

Application to single cell analysis
Important to find maker genes for cell clusters
This is done with DE analysis at the single cell level
Applying bulk RNA-seq to scRNA is not recommended because cells are treated as replicates and not individual samples

Pseudo bulk DE analysis
Create pseudo bulk expression profiles then apply standard bulk RNA seq pipeline

The scRNA seq using edgeR can be used to characterise cell gene expression in a time course analysis

Can also be used for DE methylation of bisulphate-seq

Normalisation

Ha Tran (a1743091@adelaide.edu.au) - Nov 12, 2021, 7:31 AM AEDT

Read counts are normalised for:

1. The sequencing depth (the 'Million' part)
    a. Sequences ran with greater depth will subsequently produce greater number of reads, thus, requiring normalisation prior to analysis
2. The read length (the 'Kilobase' part)
    a. Longer genes will have more reads mapping to them

For the demonstration of the differences among the following normalisation method, imagine the following RNA-seq data:

GeneID	Rep1	Rep2	Rep3
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

- Rep 3 have greater read sequencing depth as there are more count for all genes
- Gene B is twice the size of gene A, thus, is on average twice the read counts of gene A

RPKM (Reads Per Kilobase Million)

1. Normalise for read depth
    a. Calculate total read counts for all replicate
    b. Scale all the total read counts

	Rep1	Rep2	Rep3
Total	35	45	106
Scaled	3.5	4.5	10.6

c. Re-calculate the matrix by dividing the read counts by the scaled reads to obtained the Reads Per Million matrix

GeneID	Rep1 RPM	Rep2 RPM	Rep3 RPM
A (2kb)	2.86	2.67	2.83
B (4kb)	5.71	5.56	5.66
C (1kb)	1.43	1.78	1.43
D (10kb)	0	0	0.09

2. Normalise for gene length
    a. Re-calculate the RPM matrix by dividing the RPM by the length of the gene to obtain the Reads per Kilobase Million

GeneID	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009

FPKM (Fragments Per Kilobase Million)

FPKM and RPKM are similar with the crucial distinction being that RPKM isfor single end RNA-seq while FPKM is for PE-RNA-seq. FPKM keeps track of the two fragments so that they are not counted twice.

TPM (Transcripts per Million)

TPM is also similar to RPKM and FPKM with the crucial difference being that the calculation are inversed (i.e. normalise for gene length first, followed by normalisation for sequence depth)

1. Normalise for gene length
    a. Re-calculate the matrix by dividing the read counts by the length of the gene to obtain the Reads per Kilobase

GeneID	Rep1 RPK	Rep2 RPK	Rep3 RPK
A (2kb)	5	6	15
B (4kb)	5	6.35	15
C (1kb)	5	8	15
D (10kb)	0	0	0.1

2. Normalise for sequence depth
    a. Calculate total read counts for all RPK replicate
    b. Scale all the total read counts

	Rep1	Rep2	Rep3
Total	15	20.25	45.1
Scaled	1.5	2.025	4.51

c. Re-calculate the RPK matrix by dividing the RPK read counts by the scaled reads to obtained the Transcript per Million matrix

GeneID	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02

RPKM vs TMP

The major advantage of TMP over RPKM is that the proportion of read counts are more easily comparible between rep since they all have the same total. RPKM have different total for each Rep and thus it is harder to compare the proportion of gene between samples.

GeneID	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02
Total	**10**	**10**	**10**

GeneID	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42

D (10kb)	0	0	0.009
Total	**4.29**	**4.5**	**4.25**

- It is generally acceptable to use RPKM and TPM for 'within sample' transcript expression comparison
- Both RPKM and TPM are not recommended for cross-sample transcript expression comparison

03/08/2021 Dave Adelson

Ha Tran (a1743091@adelaide.edu.au) - Nov 12, 2021, 7:25 AM AEDT

Introductory meeting

Questions and answer

1. Topics of lab placement
   Compound kushen injection, differential gene expression analysis of the CKI-1 deletion generated last year
2. Duration of lab placement
   Whole semester
3. Will the placement be during the semester or in the break
4. Key literature
   2019 paper (see annotated bibliography)
5. Activities conducted during the placement
   DGE analysis
   Presentation in a lab meeting
6. Recommendation for lab book
   Lab archives
7. Big picture of the project
   Potential discovery of therapeutic origin of anti-cancer properties of CKI
8. Specific project objectives
   DGE analysis
   Gene enrichment analysis
   Pathway enrichment analysis
   Co-expression analysis
9. Resource required and approximate budget
   Not many resources are required for the following project, however, a lab server will be provided additional computational and memory power
10. Type of data that will be generated
    Plots that illustrate DE gene expression between control and CKI-1
11. Is the data confidential and subjected to privacy concerns
    NO
12. Suggestions on organisation of data
13. Storage of physical and digital data
    Can be stored on the lab server or locally
14. Safeguard measures around physical and digital data
    Password protected account on lab server and local computer
15. Data collaborators
    Dave
16. Data dissemination
    To be further discussed
17. Upon completion, what will happen with data
    To be further discussed
18. Digital data storage volume
    Maximum a couple of Gb, Rstudio and related packages excluded
19. Data backup
    Can be backup to the lab server or cloud storage such as google drive or the university's box
20. Specific metadata format
    None, up to personal preference
21. Bioinformatic related seminars
    TBA

Introduction to the lab setting and people in the lab
Set up with a desk and granted key assess to the building
Completed induction forms
Invitation to the lab server and lab discord server
Suggestions to use putty ssh to connect to the lab server
Invitation to fortnightly lab meetings
Invitation to present at a lab meeting after the project is done

06/08/21 Zhipeng Qu

Ha Tran (a1743091@adelaide.edu.au) - Nov 12, 2021, 7:26 AM AEDT

Discussion of the primary Aung et al., 2019 paper,

> Since Aung et al., 2019 performed DGE on multiple treatment groups, the only overlapping pathway was cytokine-cytokine receptor signalling. Therefore, this project will take a step back and assess only the DGE between the CKI-1 deletion group. Furthermore, Aung et al., 2019 illustrated that CKI-1 significantly perturbed many pathways without demonstrating any phenotypical effects in the bioassays, therefore, warrant re-investigation of the DGE in CKI-1 samples.

Further discussion of the specific results related to Aung et al., 2019 paper. These results are clearly annotated on the pdf file of the paper. See annotated bibliography.

Introduction to related paper, namely (Qu, Z et al., 2016) and (Aung, T. N et al., 2017) papers. A brief discussion of the topic and results of the paper.

20/08/21 Zhipeng Qu

Ha Tran (a1743091@adelaide.edu.au) - Nov 12, 2021, 7:26 AM AEDT

Brief discussion of the main transcriptomic techniques and tools that will be used for the differential gene expression analysis.

For DGE analysis these included
edgeR, DESeq2, limma for the actually DGE analysis process.

For gene enrichment analysis two main tools are used,
Gene ontology (GO) term enrichment analysis, and
KEGG pathway analysis

For co expression analysis
WGCNA weighted correlation network analysis

Further discussion about the potential figures that will be produced during the analysis

Further discussion about the full breath of the research project
If time permits, raw reads can be used to perform transcriptome assembly (i.e. start from the beginning)
Or, co-expression analysis can also be performed, however, this is only of time permits and not need for the project

Overview of the initial preliminary DGE pipeline and review of the results.

03/09/21 Zhipeng Qu

Ha Tran (a1743091@adelaide.edu.au) - Nov 12, 2021, 7:28 AM AEDT

Project update, progress on the research proposal,
More details about the DMP and time management of the project

Project Description

Project Plan

DMP