Biol-4315 Project: Comparing two different aligners and see if and how they may impact alignment

Gavin Ieong

2022-12-05

Table of Contents

Abstract

Aligners are alignment programs used for mapping Ribonucleic acid (RNA) and deoxyribonucleic acid (DNA) reads to a genome. Currently, there are many aligners available for use; such examples include Hisat2, Subread, Bowtie2, Tophat2, Kallisto, and many more.

Although similar in function, many of the programs' algorithms have a different approach with different focus areas in its implementation of its use. Such factors include usage of short reads vs long reads, computational speed, better memory management, and some may even consider different methods of indexation. Furthermore, some aligners have different capabilities regarding recognition of alignment at exon splice sites and splice junctions.

For example, Bowtie2, is an aligner that is memory efficient and is best used for alignment short reads between 50 to 100bp, or longer reads that are 1000bp for bigger genomes, but is a splice unaware aligner. However, looking at Hisat2, it is best suited for reads under 500bp, but conversely is splice aware aligner.

If the aligners were to be compared, will the different approaches used between Hisat2 and Bowtie2 result in different alignment of short mRNA reads to a reference genome, and thus impacting downstream differential gene expression analysis?

Using Hisat2 and Bowtie2 as the chosen aligners, they are run through the same exact project pipeline in analyzing a dataset of *Arabidposis thaliana* and the results were contrasted to each other and was evaluated to check if identical results were produced. The resulting graphs and data, such as hierarchal clustering trees, bar graphs, and heatmaps, were set side by side, and both the differences and general trends were noted.

Between Hisat2 and Bowtie2, the former produced a higher number and percent of aligned reads to the genome, and better grouping within the hierarchal clustering trees. However, similar trends of DEG were seen in the gene ontology bar plots, and similar trends were further seen in the heat map produced.

Due to the similarities of the trend between the aligners, the aligners cannot be said to produce different alignments, however further research with bigger datasets and more aligners can help confirm this result.

General Background

The science and study of “omics” can be seen as the studies of processes in totality. Emergence of technologies allows us to study many “omics”, and two of those major branches encompasses genomics and transcriptomics.[19] Furthermore, there are other branches such as metaomics which covers metagenomics and metatranscriptomics.[2]

Genomics is a branch of genetics which comprehensively studies the structure and function of DNA within a gene and genome.[19] Genomics makes use of Genome-Wide

Association Studies (GWAS) and Whole-Genome Sequencing (WGS) as tools to fully investigate the complexity surrounding genotypes and phenotypes. Simply, this includes integration of capturing patterns within genes population-wide, but further invokes complex topics with examples such as using variants as markers for risk prediction and can further lead to personalized genomic medicine due to big data.[3] Studies have shown that personalized genomic sequencing, particularly for medicine, has been on the rise in the USA and Canada, and has been used as marker tools as a lookout for cancer susceptibility and prenatal anomalies.[15] However due to increased use of personalized genomics, this also has created issues in regards to ethics.[15]

Transcriptomics is the complete study of RNA, which includes a transcriptome (a set of all RNA transcripts).[6] A comprehensive analysis of the transcriptome can provide complex information about gene structure, expression, and even differential expressed genes. In fact, DEGs are one of the earliest goals of transcriptomics, and due to those studies, technologies such as microarrays have been further developed and can be used to quickly analysis DEGs. However, it is not without its caveats when working with RNA, and there can be problems with quantifying and defining transcripts as DEG occur under different environmental conditions, and RNA having a post-transcriptional process.[6]

Metagenomics is the study of a whole collection of microbes within an environment. Currently studies are done on microbial communities with environmental pollutants to better study the microbe community composition and activity in an area that is contaminated. With the use of next-generation sequencing (NGS), there are hopes that vital information about key enzymes and genes that are involved with the degradation and detoxification of the pollution in the area can be obtained and further researched.[2]

Meta transcriptomics informs us of the genes that are expressed by the community as a whole as it deals with the analysis of meta genomic mRNA (known as metatranscriptome).[2] It can provide a snapshot of the conditions the microbe communities in a given sample by looking at the mRNA. However, currently it is a difficult process due to the need to extract high quality mRNA samples from the environment.[2]

Project background

Two aligners (Hisat2 and Bowtie2) will be used to align a dataset from SRA SRP010938, which contains 18 paired end (PE) read sets from *Arabidposis thaliana*, to see if they will produce identical results.

The first aligner used, Hisat2, is an external software package to R. Its uniqueness stems from the fact that it uses large numbers of indexes and an indexing scheme based along two mathematically operations (Burrows-Wheeler transform and Ferragina-Manzini) for alignment.[8] Instead of a singular index that represents the genome, it has both a global index, and numerous smaller indexes that cover the genome. Those smaller indexes further help align reads containing exons, and it boasts of high accuracy in detection of slice sites and alignment of those reads to the region.[8]

On the other hand, Bowtie2 is an aligner that is also external to R. It boasts of an effective method of combining speed, sensitivity, and even accuracy, across various lengths of reads and thus associated technologies producing these reads.[10] In order to achieve this, Bowtie2 utilizes a technique called full-text minute indexing, and in combination, further utilizes hardware accelerating algorithms to produce its alignments.[10]

These differences could contribute to a different alignment of reads to a genome if the algorithm used varies within its implementation. Will the different approaches used between Hisat2 and Bowtie2 result in different alignment of short mRNA reads to a reference genome, and thus impacting downstream differential gene expression analysis? Despite the differences between the two aligners, the basis of the aligners is to correctly align reads to a genome which should not change between aligners. Therefore, regardless of the differences between chosen aligners, the alignment should not change and will lead to the same alignment to the genome.

Project Methodology

Program ran using:

- R version 4.2.1 (2022-06-23)
- Platform: x86_64-apple-darwin17.0 (64-bit)
- Running under: macOS Monterey 12.6.1

Libraries:

- systemPipeRdata(2.0.1) [7]

- systemPipeR (2.2.2)[1]

Packages:

- BiocManager (1.30.19)[14]
- docopt (0.7.1)[5]
- DT (0.26)[18]
- pheatmap (1.0.12) [9]
- Rbowtie (1.36.0) [11]
- GenomicFeatures (1.48.4)[12]
- DESeq2 (1.36.0)[13]
- edgeR (3.38.4)[17]
- BiocStyle (2.24.0)[16]
- Go.db (3.15.0) [4]
- Hisat2 (2-2.2.1) [8]

Dataset:

Using SRA: SRP010938 which contains 18 paired-end (PE) read sets from *Arabidposis thaliana*. There are a mix of three treatments within, which are separated into Virulent (V), some infected with buffer (B), and those infected with A-Virulent (A). The A and B treatment are negative controls.

Project Step Summary

The project pipeline consists of a total of 30 steps.

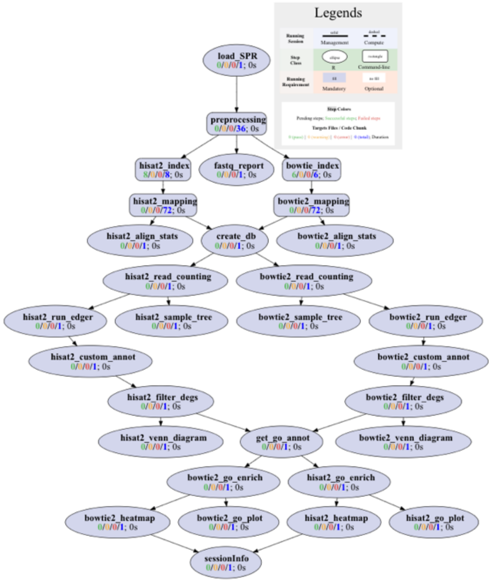


Figure 1. Full Project Pipeline with 30 steps in total

Main Steps

Load SPR

- Load required libraries for project pipeline (Loads SystemPipeR)

Pre-Processing

- Clean and trim reads by removing adapter sequences, primers, PCR chimeras and low-quality reads using ShortReads
- Print out Fastq Report to see quality, length, and other information after trimming the reads involved

Index and Mapping

- Will call Bowtie2 and Hisat2 to both index and map those reads to genome
- This is a two-step process and involves Samtools
- Alignment Stat and Database can be created after mapping those reads

Count table and Annotations

- Requires library called “GenomicFeatures”

- Can get annotations from a file called tair10.gff, using arguments:
    - format = "gff", dataSource = "TAIR", organism = "Arabidopsis thaliana")
- Will create count table and put in file called countDFeByg.xls and RPKM in file called rpkmDFeByg.xls

Sample Wise Correlation and Clustering

- Sample Tree of Hierarchical Clustering Based on Gene Expression
    - Requires libraries “Deseq2” and “ape”
    - Made with DeSeq2 and method = spearman
- EdgeR will be used for DEG analysis
    - Requires library “edgeR”
    - Will read and use counts from previously made countDFeByg.xls

Adding Gene Descriptions and Custom Annotations

- Create annotations from online host
    - Requires library “biomaRt”
    - With arguments:
    - dataset = "athaliana_eg_gene”, host = "https://plants.ensembl.org")
    - Saved into file called edgeRglm_allcomp.xls
- DEG Count Analysis
    - Loads previous file edgeRglm_allcomp.xls
    - Filter DEG using arguments Fold=2, FDR = 20
    - Save to file called DEGcounts.xls
    - Create bar graph of DEG counts

  - Venn Diagram is also created

Gene Ontology (GO) Analysis

  - Create annotations from online host
    - Requires library “biomaRt”
    - With arguments:
    - dataset = "athaliana_eg_gene”, host = "https://plants.ensembl.org")
  - Create Enrichment Analysis Consisting Of:
    - Molecular Function
    - Biological Processes
    - Cellular Components
  - Heat map is generated
    - Requires library “pheatmap”

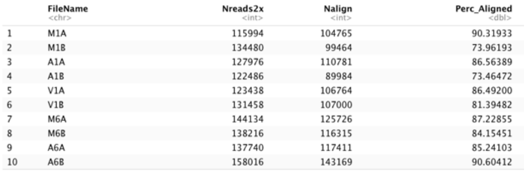
Report System

  - Render and Print Logs
  - Render Final Report

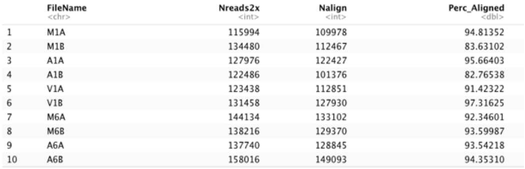
Project Results

To examine the results, a comparison of 5 different outputs between Hisat2 and Bowtie2 will be made.

Comparison Of Alignment Stats



	FileName	Nreads2x	Nalign	Perc_Aligned
	<chr>	<int>	<int>	<dbl>
1	M1A	115994	104765	90.31933
2	M1B	134480	99464	73.96193
3	A1A	127976	110781	86.56389
4	A1B	122486	89984	73.46472
5	V1A	123438	106764	86.49200
6	V1B	131458	107000	81.39482
7	M6A	144134	125726	87.22855
8	M6B	138216	116315	84.15451
9	A6A	137740	117411	85.24103
10	A6B	158016	143169	90.60412



	FileName	Nreads2x	Nalign	Perc_Aligned
	<chr>	<int>	<int>	<dbl>
1	M1A	115994	109978	94.81352
2	M1B	134480	112467	83.63102
3	A1A	127976	122427	95.66403
4	A1B	122486	101376	82.76538
5	V1A	123438	112851	91.42322
6	V1B	131458	127930	97.31625
7	M6A	144134	133102	92.34601
8	M6B	138216	129370	93.59987
9	A6A	137740	128845	93.54218
10	A6B	158016	149093	94.35310

Figure 2. Bowtie2 (left) and Hisat2 (Right) Alignment Stats showing raw number along with percentage of reads aligned.

Bowtie2 is seen to display fewer number of alignments and small percentage of overall alignment for all ten samples compared to Hisat2. For example, for file M1A, 90.3% was aligned for Bowtie2 while 94.8% was aligned for Hisat2. However, pure alignment numbers are not conclusive nor informative enough, and further examination into trees and DEG will follow.

Comparison of Hierarchal Clustering Trees

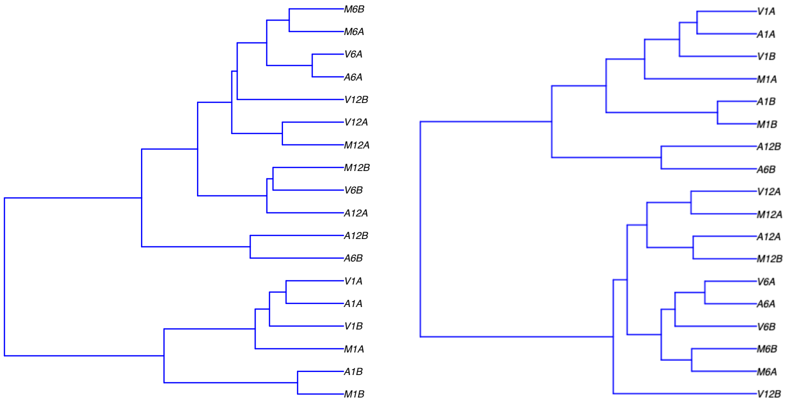


Figure 3. Hierarchal Clustering Trees for Bowtie2 (left) and Hisat2(Right)

At first glance, the trees seem similar as we can see the same grouping with 1A and 1B, in the lower portion of Bowtie2 and top portions of Hisat2. However, with Bowtie2, we can see more mixing between the time points as many of the 12 and 6 time points are grouped together in the middle of the tree. Unlike Hisat2, where the time groups are well defined together. This showcases that Hisat2 is better at grouping time points together and furthermore the tree is more balances as roughly half of the groups belong to each branch.

Comparison of DEGs Counts

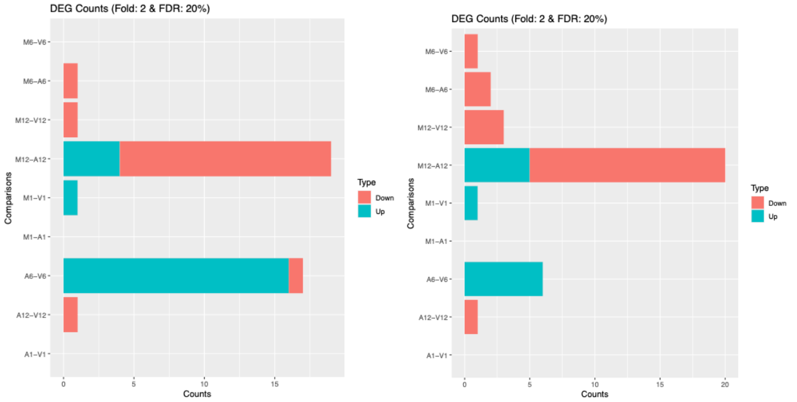


Figure 4. Comparision of DEG count for Bowtie2(left) and Hisat2(right)

Looking at the DEG count for both aligners where the FDR is at 20%, and they show a similar trend. Exceptions can be seen where Bowtie2 showcases more up regulated DEG expression in A6-V6, and Hisat2 showcases more down regulated DEG expression through the other comparisons.

Comparison Of Gene Ontology Bar Graphs

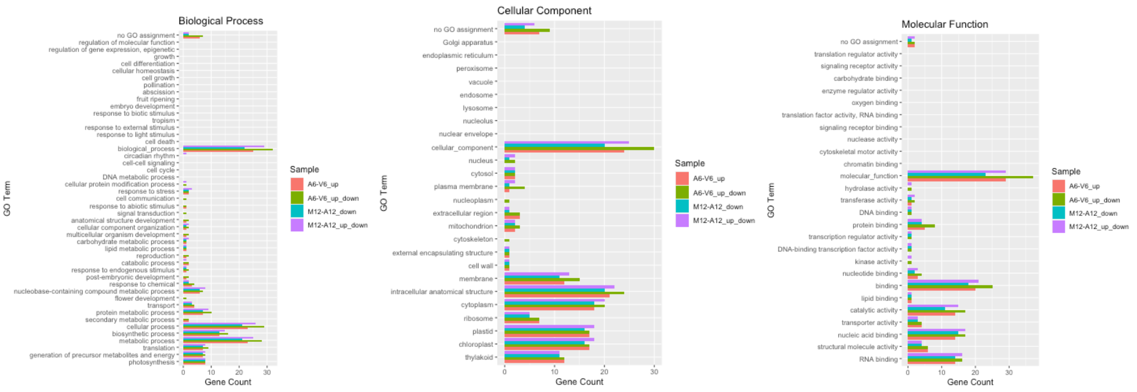


Figure 5. Gene Ontology Bar Graphs for Biological Process, Cellular Component, and Molecular Function for Bowtie2

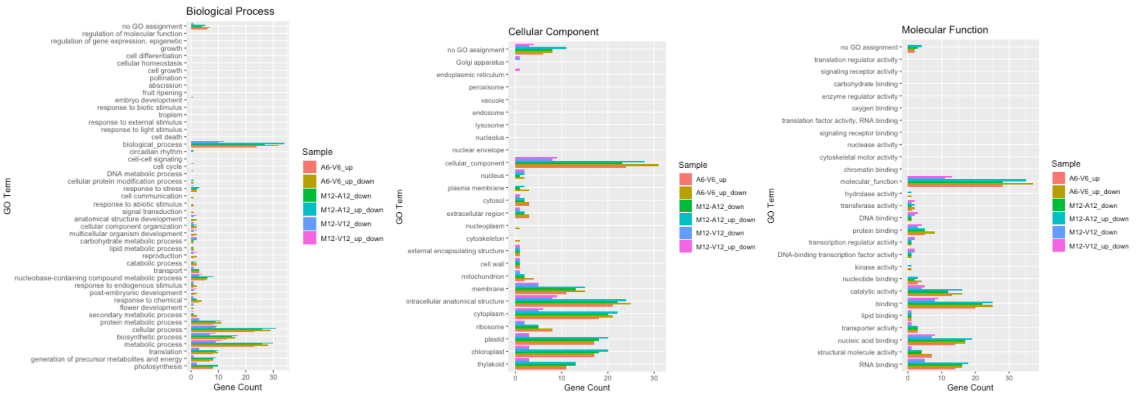


Figure 6. Gene Ontology Bar Graphs for Biological Processes, Cellular Components, and Molecular Function for Hisat2

Once again, looking at the general trend, there is not much difference between Bowtie2 and Hisat2. However, with Hisat2, there are two more pairwise comparisons, which provide more information in analysis.

Comparison of Heat Maps

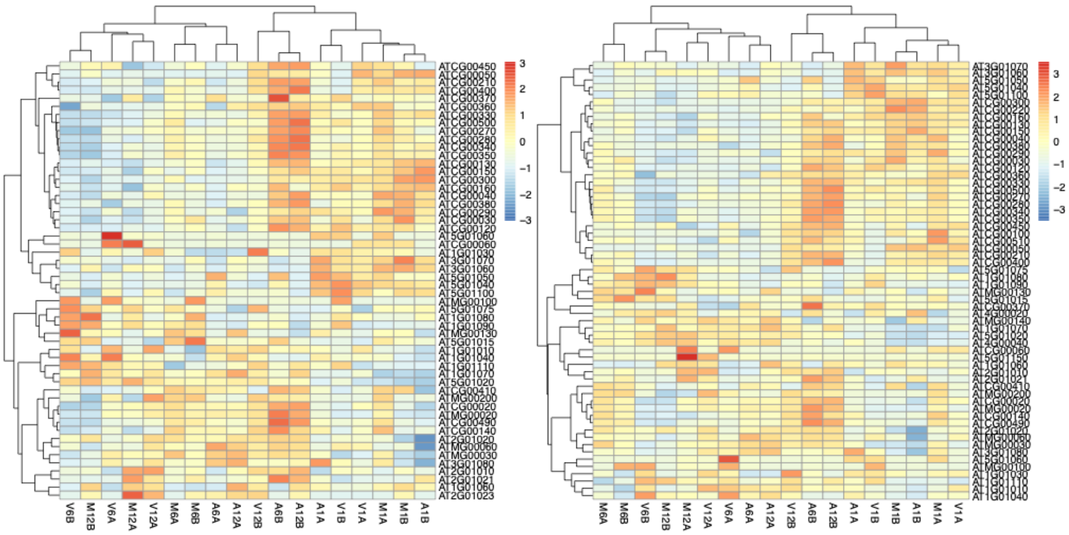


Figure 7 Heatmap generated with data from Bowtie2(left) and Hisat2(right)

The heat maps are harder to compare due to the complexity surrounding it; however, once again we can see the general trends and patterns stay the same despite differences seen. The top twenty genes are different as noted on the right y-axis, and the tree topology is also different as noted on the x-axis. But on further note, looking at the general trend in both heat maps, the upper right corner and bottom left corner, there is more up regulated DEG as seen with the increase in red. Conversely, the top left corner in both figures has more down regulated DEG as seen with the increase in blue.

Project discussion

As noted from our results, using the same project pipeline, but substituting different aligners used within, led to similar but different results. First comes the difference in percentage of reads aligned, with Bowtie2 having a smaller percentage aligned. One possible reason for this could be that it is splice unaware. As we are working with plant reads which are eukaryotic, and RNA-seq data, there would be splice sites within. As mentioned previously, Hisat2 has algorithms within to help detect these sites and thus could potentially be the reason for more reads aligned to the genome.

Secondly, the visualized diagrams (Hierarchal Clustering Trees, DEG counts, Gene Ontology Plots, and Heatmaps), showcase similar results with small differences. Although there are minor differences, the results still had the same general trend. There were no big outlier differences to be seen between the aligners. To reiterate, the goal of an aligner is to produce an alignment of reads to a reference genome. If the aligners have the means to produce similar outputs independently, then this goal was accomplished. Thus, the conclusion to the question posed is that the alignment with different aligners do produce the same outputted results.

However, there are caveats to this conclusion which must be kept in mind while examining these two aligners. Firstly, the data is subsetted to a smaller size and this will result in a large portion data to be unnoticed and unused and may carry differences not seen in the smaller sample size. Furthermore, as the general trend and pattern was used as a criterion during judgement of same alignment, a full dataset is needed to account for the complexity of the data because differences would change the trend more drastically than noted.

Additionally, both aligners chosen coincidently used Samtools when mapping and indexing, and thus contributing to the similarities. Finally, from research, Hisat2 uses Bowtie2 as inspiration for implementation of many low-level operation, and once again contributes to the similarities seen.[8]

In regard to the importance of these findings, the emphasis of choosing specific aligners might not be a crucial step in a project pipeline if the general trends staying the same despite the aligner used. This may allow more flexibility in overall design of a project but due to those minor differences we cannot say it is good to use same aligners for every case; however, this could mean there are aligners fitted for more specialized cases.

In future trials to further study and confirm this conclusion, more aligners with bigger dissimilarities should be further researched alongside a bigger dataset to confirm if these findings hold true.

Citations

1. Backman TWH, Girke T (2016) systemPipeR: NGS workflow and report generation environment. BMC Bioinformatics 17 (1), doi:10.1186/s12859-016-1241-0.

2. Bharagava, R. N., Purchase, D., Saxena, G. & Mulla, S. I. Applications of metagenomics in microbial bioremediation of pollutants. *Microbial Diversity in the Genomic Era* 459–477 (2019). doi:10.1016/b978-0-12-814849-5.00026-5

3. Capps, B. New Technologies: Ethics of Genomics. *International Encyclopedia of Public Health* 240–247 (2017). doi:10.1016/b978-0-12-803678-5.00300-3

4. Carlson M (2022). _GO.db: A set of annotation maps describing the entire Gene Ontology_. R package version 3.15.0.

5. de Jonge E (2020). _docopt: Command-Line Interface Specification Language_. R package version 0.7.1, <https://CRAN.R-project.org/package=docopt>.

6. Dong, Z. C. & Chen, Y. Transcriptomics: Advances and approaches. *Science China Life Sciences* **56,** 960–967 (2013).

7. Girke T (2022). _systemPipeRdata: systemPipeRdata: Workflow templates and sample data_. https://github.com/tgirke/systemPipeRdata, https://systempipe.org/.

8. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A fast spliced aligner with low memory requirements. *Nature Methods* **12,** 357–360 (2015).

9. Kolde R (2019). _pheatmap: Pretty Heatmaps_. R package version 1.0.12, <https://CRAN.R-project.org/package=pheatmap>.

10. Langmead, B., Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012). https://doi.org/10.1038/nmeth.1923

11. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10(3):R25 (2009).

12. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, et al. (2013) Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol 9(8):

e1003118. doi:10.1371/journal.pcbi.1003118

13. Love, M.I., Huber, W., Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 Genome Biology 15(12):550 (2014)

14. Morgan M (2022). _BiocManager: Access the Bioconductor Project Package Repository_. R package version 1.30.19, <https://CRAN.R-project.org/package=BiocManager>.

15. Shickh, S. *et al.* The role of digital tools in the delivery of genomic medicine: Enhancing patient-centered care. *Genetics in Medicine* **23,** 1086–1094 (2021).

16. Oleś A (2022). _BiocStyle: Standard styles for vignettes and other Bioconductor documents_. R package version 2.24.0, <https://github.com/Bioconductor/BiocStyle>.

17. Robinson MD, McCarthy DJ and Smyth GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140

18. Xie Y, Cheng J, Tan X (2022). _DT: A Wrapper of the JavaScript Library 'DataTables'_. R package version 0.26, <https://CRAN.R-project.org/package=DT>.

19. Yadav, D., Tanveer, A., Malviya, N. & Yadav, S. Overview and principles of bioengineering. *Omics Technologies and Bio-Engineering* 3–23 (2018). doi:10.1016/b978-0-12-804659-3.00001-4