Which Aligner Software is the Best for Our Study?
Abolfazl Bahrami *

Department of Animal Science, University College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran

*Corresponding author: A Bahrami, Department of Animal Science, Tehran University, Karaj, I.R. Iran, Tel/Fax: +98-9199300665

Abstract
Aligners are the most important software used in the field of Transcriptomics studies and related fields. In the recent study, almost all the aligners could be configured to give good results, but still, researchers and scientists who use such software face challenges in choosing accurate, sensitive, requiring fewer hardware facilities and ultimately appropriate with their research goals. We try to clarify the various challenges and misunderstandings, below.

Introduction
Alignment is the first step in most RNA-seq analysis pipelines, and the accuracy of downstream analyses depends heavily on it. Many algorithms have been developed for this alignment step. Due to the increasing growth in the use of aligning and mapping software, this software has become particularly important. This seemingly worthless issue but it is confusing and difficult to compare results from different approaches. We performed a comprehensive benchmarking of 4 popular and common aligners and compared default with optimized parameters. Another thing that should be considered is how robust the results are to different parameters. In the previous studies, almost all the aligners could be configured to give good results, but they differed in the performance of the default options [1], with HISAT2 (hierarchical indexing for spliced alignment of transcripts 2) looking pretty good in those terms [2]. We have to say though, we use HISAT2 a lot just because of how easy it is and how few resources it requires. Therefore, in this research, we have done a statistical analysis using SPSS software on simulated (The simulation engine BEERS [3] was used to generate simulated data. Data were generated for human. Each data set consists of 15 million 100-base paired-end strand-specific reads.

The genomes used were Homo sapiens hg19. For human data, 30,000 transcript models were chosen at random) and real data in the GEO (Gene Expression Omnibus) database [4] (Supplementary Table 1 shows the accession number, number of samples and related study title of data used in this study that obtained from different experiments (~116 billion reads)). Some studies have shown related software in comparison with other software in terms of sensitivity, precision, run time and memory usage and shown HISAT2 is more acceptable (Supplementary Table 2) but it’s not about the number of mapped reads and the power of other software. In this way, the most important of these software include HISAT2, TopHat2 [5] and STAR [6], and the other hand because HISAT2 uses the Bowtie2 [7] implementation, so we compared these four software in terms of mapping percentage averages (Supplementary Table 3).

The simplified results of the comparisons are presented in Table 1. Table 2 shows the percentages of mapping on the human reference genome using Trim
The precision of the HISAT2 is higher when considering the mapped percentage parameter on a particular location. On the other hand, TopHat2 has the power to detect introns from exons and map more reads to more than one specific location. STAR maps a greater percentage of reads as incorrectly mapped. Finally, Bowtie2, which is more specific to DNA-Seq data, is not practical for using in RNA-Seq mapping studies.

For data science, the software must be provided with automatic software [8] output files and then using HISAT2, TopHat2, STAR and Bowtie2 aligner (existed aligner in the galaxy server). Reads declared ‘aligned’ can be summarized in three main groups: Correctly mapped, correctly multimapped, and incorrectly mapped reads. Hopefully, an effective tool will report the majority of reads aligned correctly, with a few reads aligned ambiguously and very few reads aligned incorrectly (Figure 1).

Results of analysis as follows:

- TopHat2 maps a greater percentage of reads on the reference genome. As well as, correctly multimapped percentage is higher than other software; this can be useful in capturing non-coding regions such as miRNAs and other non-coding RNAs (Supplementary Figure 1).
- The precision of the HISAT2 is higher when considering the mapped percentage parameter on a particular location.
- On the other hand, TopHat2 has the power to detect introns from exons and map more reads to more than one specific location.
- STAR maps a greater percentage of reads as incorrectly mapped.
- Finally, Bowtie2, which is more specific to DNA-Seq data, is not practical for using in RNA-Seq mapping studies.

For data science, the software must be provided
via an easy to use, unified interface, such that they can be easily deployed and sustainably managed. With an understanding of its ability to analyze data set, the researchers will have a better interpretation of their results. Eventually, the results of the statistical analysis of this research can be a good guide for researchers using this software.

**Competing Interests**

The authors have no financial conflicts of interest.

**Additional Information**

The authors declare that data supporting the findings of this study are available within the article and its Additional files.

**Acknowledgements**

We would like to thank Prof. Reza Miraie-Ashtiani and Prof. Mostafa Sadeghi for providing critical feedback and reagents, Ghorban Elyasi Zarringhabaie for technical assistance and the University of Tehran, Department of Animal Science. We also thank the anonymous reviewers for their constructive feedback which significantly improved the quality of our analysis.

**References**

1. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37: 907-915.
2. Kim D, Langmead B, Salzberg SL (2015) HISAT: A fast spliced aligner with low memory requirements. Nat Methods 12: 357-360.
3. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, et al. (2011) Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics 27: 2518-2528.
4. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. (2013) NCBI GEO: Archive for functional genomics data sets--update. Nucleic Acids Res 41: 991-995.
5. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, et al. (2013) TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14: 36.
6. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. (2013) STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29: 21-211.
7. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359.
8. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120.
### Supplementary Table 1: The accession number, number of samples and related study title of data used in this study that obtained from different experiments (~116 billion reads).

| Accession number | Samples | Title                                                                                                                                 |
|------------------|---------|----------------------------------------------------------------------------------------------------------------------------------------|
| GSE130401        | 21      | The hippo pathway effector protein YAP modulated resistance to trametinib neuroblastomas with hyperactivated RAS pathway signalling       |
| GSE137290        | 21      | Distinct mechanisms of acquired resistance to oncogenic kinase inhibition in cancer cells revealed using a single-step, high-dose selection scheme |
| GSE135902        | 40      | The Transcriptome Of Cml Monocytes Is Highly Inflammatory And Reflects Leukemia-Specific And Age-Related Alterations                      |
| GSE124326        | 480     | Whole blood transcriptome analysis in bipolar disorder reveals strong lithium effect                                                 |
| GSE120597        | 50      | Genetic Abnormalities in Large to Giant Congenital Nevi: Beyond NRAS mutations                                                        |
| GSE139250        | 51      | Exploring the impact of chronic hypoxia on the expression of DNA repair gene in Glioblastoma and Medulloblastoma cells.             |
| GSE129705        | 128     | RNA-sequencing of whole blood samples from biologic naïve rheumatoid arthritis patients initiating anti-TNF treatment              |
| GSE115046        | 110     | Massively parallel characterization of regulatory dynamics during neural induction                                                   |
| GSE139181        | 33      | Transcriptomics analysis of trimester-specific full-term placentas from three Zika virus-infected women                               |
| GSE130289        | 139     | Dynamics of Trophoblast Differentiation in Peri-implantation Stage Human Embryos                                                      |
| GSE118912        | 32      | Activity-by-Contact model of enhancer specificity from thousands of CRISPR perturbations                                             |
| GSE105160        | 197     | RNASeq of mouse, human, and non-human primate primary dermal fibroblasts to poly(I:C) transfection                                      |
| GSE138988        | 24      | Transcriptome-wide comparison of stress granules and P-bodies reveals that translation plays a major role in RNA partitioning        |
| GSE138853        | 30      | Impact of transcriptional mutagenesis on p53 transactivation                                                                            |
| GSE137392        | 60      | MITF regulates SCD and fatty acid saturation to control melanoma phenotypic state.                                                    |
| GSE137391        | 24      | Transcriptomics profiling of some commonly used cell lines at the base-line culture condition                                         |
| GSE137390        | 36      | Lineage-restricted regulation of SCD and fatty acid saturation by MITF controls melanoma phenotypic plasticity                       |
| GSE116698        | 76      | Co-Stimulation–Induced AP-1 Activity is Required for Chromatin Opening During T Cell Activation.                                     |
| GSE112855        | 45      | Next generation sequencing profiling experimental circulating tumor cells-derived metastatic variants [RNA-seq]                     |
| GSE138730        | 32      | Altered m6A Modification of Specific Cellular Transcripts Affects Flaviviridae Infection                                            |
| GSE124685        | 84      | mRNA Sequencing to identify transcriptional changes in early and late stages of lung in human Idiopathic Pulmonary Fibrosis       |
| GSE94690         | 40      | eIF4A2 drives repression of translation at initiation by Ccr4-Not through purine-rich motifs in the 5'UTR                           |
| GSE138485        | 46      | Retrospective gene expression analysis of human RNA samples from Hepatocellular Carcinoma in relation with survival                |
| GSE130751        | 63      | Non-oncogene addiction to SIRT3 plays a critical role in lymphomagenesis                                                             |
| GSE127696        | 78      | Transcriptomic profile of cystic fibrosis airway epithelial cells undergoing repair                                                   |
| GSE133151        | 74      | Clonal selection confers distinct evolutionary trajectories in BRAF-driven cancers                                                     |
| GSE125873        | 31      | RNA-Seq of blood in preterm infants with Bronchopulmonary dysplasia.                                                                  |

### Supplementary Table 2: Comparison of studies.

| Un Mapped (AVERAGE) | Total Mapped (AVERAGE) | 1 Paired (AVERAGE) | > 1 Paired (AVERAGE) | Un Mapped (AVERAGE) | Total Mapped (AVERAGE) | 1 Paired (AVERAGE) | > 1 Paired (AVERAGE) |
|---------------------|------------------------|--------------------|----------------------|---------------------|------------------------|--------------------|----------------------|
| 1417277.889         | 17455782.67            | 16826840.05        | 628942.1667          | 7.558871667         | 92.52623056           | 89.20201278        | 3.324218333          |
| 1084243.778         | 17772150.39            | 16565979.83        | 1206170.778          | 5.803278333         | 94.19672167           | 87.81203444        | 6.384687222          |
| 5175759.481         | 13693971.39            | 10933493.13        | 2760478.275          | 27.45770389         | 72.54202401           | 57.91073651        | 14.63628698          |
| Software | Sample | Total Sequence (AVERAGE) | Un Mapped (AVERAGE) | Total Mapped (AVERAGE) | 1 Paired (AVERAGE) | >1 Paired (AVERAGE) | Un Mapped (AVERAGE) (%) | Total Mapped (AVERAGE) (%) | 1 Paired (AVERAGE) (%) | >1 Paired (AVERAGE) (%) |
|----------|--------|--------------------------|---------------------|------------------------|-------------------|-------------------|------------------------|------------------------|------------------------|------------------------|
| HISAT2   | 2045   | 18856393.89              | 1745782.67          | 13693971.39            | 2760475.89        | 27.45770389       | 1093343.13             | 72.54702401            | 72.45770389            | 27.60478.275           |
| Tophat2  | 2045   | 18856393.89              | 1084243.778         | 16932744.44            | 1206170.778       | 5.8032783344      | 1093343.13             | 72.54702401            | 72.45770389            | 27.60478.275           |
| Bowtie2  | 2045   | 18866804.93              | 92.53223056         | 17772150.44            | 1206170.778       | 12.06170.778      | 1093343.13             | 72.54702401            | 72.45770389            | 27.60478.275           |
| HISAT2   | 2045   | 18856393.89              | 1417277.889         | 1745782.67            | 1206170.778       | 5.8032783344      | 1093343.13             | 72.54702401            | 72.45770389            | 27.60478.275           |