Minh Khue Nguyen 12.02.2024

Benchmarking long-read RNA-seq technologies for short-read tasks

Studienprojekt Exposé

1 Introduction

1.1 RNA sequencing

RNA sequencing (RNA-seq) has revolutionized the way scientists explore the transcriptome, the complete set of transcripts in a cell, and its quantitative expression levels. This technology enables the identification of novel genes, splice variants, and fusion transcripts, and allows for the comprehensive analysis of gene expression patterns across different conditions, tissues, and developmental stages (Kapranov et al., 2007, Birney et al., 2007).

Initially, short-read sequencing technologies, prominently represented by Illumina's sequencing platforms, dominated the field due to their high-throughput capabilities, cost-effectiveness, and high accuracy in quantifying gene expression levels (Van Dijk et al., 2014). These technologies generate millions of short nucleotide sequences, typically ranging from 50 to 600 base pairs (bp) in length, which are then aligned to a reference genome or transcriptome to infer the expression of genes or transcripts.

A key challenge with short-read sequencing technology is its limited read length, currently up to 600 bases. This limitation hinders precise quantification of different transcript variants and the discovery of new ones. Short-reads often do not span entire gene transcripts or large exon regions, making it difficult to map these reads unambiguously back to the correct locations in the genome. In contrast, long-read sequencing technologies produces much longer reads, typically ranging from 1 to 100 kilobases (kb). This range comfortably covers the entire length of human spliced genes. As a result, long-read sequencing can capture entire transcript variants in a single read. However, the trade-offs of this long-read sequencing technology are lower throughput and somewhat less accuracy when compared to short-read sequencing (Dong et al., 2021). The long-read sequencing landscape is primarily dominated by two technologies: Pacific Biosciences' (PacBio) (Rhoads and Au, 2015) Single-Molecule Real-Time (SMRT) sequencing and Oxford Nanopore Technologies' (ONT) (Bowden et al., 2019) nanopore sequencing.

1.2 Applications of RNA sequencing data

- **Differential gene expression (DGE) analysis**

  The information encoded in the selected genes is transcribed into RNA molecules, which in turn can be translated into proteins or can be directly used to finely control gene expression. As a result, the collection of RNAs produced under specific circumstances and at a given moment represents the present condition of a cell and can disclose the pathogenic mechanisms driving illnesses (Finotello and Di Camillo, 2015).

  A fundamental use of RNA-seq data is differential gene expression (DGE) analysis, which identifies genes whose expression levels differ significantly between various biological conditions, such as healthy versus diseased tissues, treatment versus control groups, or different stages. This analysis can be extended to the transcript level, enabling the exploration of differential transcript expression (DTE), which provides insights into changes at a more granular level of gene regulation.

  While differential transcript usage (DTU) refers to differences in transcript proportions within a gene, implying DTE, the reverse is not necessarily true; for instance, if total gene expression doubles and each transcript type proportionally doubles, this results in DTE (and thus DGE) without DTU (Liang and Pardee, 2003, Singh et al., 2018). Thus, DGE analysis not only identifies upregulated or downregulated genes in response to particular stimuli or environmental changes but also serves as a foundational tool for more detailed functional investigations and hypothesis development, paving the way for the discovery of disease biomarkers and understanding the molecular bases of phenotypic variations.

- **Differential transcript usage (DTU) analysis**

  While DGE focuses on changes in overall gene expression, DTU analysis dives deeper into the transcriptome to explore variations in the usage of transcript isoforms between conditions.

  To fully comprehend the intricacy of gene regulation, DTU analysis is essential since numerous genes can generate different isoforms by alternative splicing, alternative promoter use, or alternative polyadenylation. It pinpoints particular isoforms that exhibit differential utilisation, which may have unique functional ramifications like regulating discrete components or encoding various protein variations. In the setting of complicated disorders like cancer, where aberrant splicing patterns might contribute to carcinogenesis and progression, DTU analysis is especially crucial (Marques-Coelho et al., 2021).

1.3 SIRV Spike-ins - Benchmarking RNA-sequencing methodologies

Synthetic Internal RNA Variants, or SIRV Spike-ins, are carefully designed RNA molecules that are employed as internal standards in RNA sequencing (RNA-seq) investigations. These synthetic RNA variations are made to resemble natural RNA transcripts in terms of complexity, size, and structure. They are added to RNA samples prior to the sequencing process and have a number of uses, the main one being to offer a controlled baseline for RNA-seq techniques that may be used to calibrate, validate, and increase their accuracy.

2 Scope of project

Despite the trade-offs between short-read and long-read sequencing technologies, there has been no work done on directly comparing the applications of RNA sequencing data generated from these technologies, especially between PacBio, ONT and Illumina. To my knowledge, the closest work by (Dong et al., 2023) benchmarked different analysis tool using deep sequencing data from Illumina and Oxford Nanopore Technologies. This study however only uses one method for DGE and 2 methods for DTU.

The primary objective of this study project is to assist the pipeline construction for benchmarking long-read RNA-seq datasets from Illumina, PacBio and ONT for tasks typically performed with short-read data. These tasks include quantification, replicability, differential gene/transcript expression analysis, and differential transcript usage. The concrete tasks of this project involves constructing a pipeline for running differential gene/transcript expression analysis using multiple different tools and evaluating the results of DGE/DTE. Since this study project is intended for a limited time-frame, only a part of the intended tools will be implemented, as well as only a part of the intended analysis will be done (See 4. Approach for the specific tools and analysis). This pipeline will then be integrated with an existing pipeline that performs quantifications on the raw RNA-seq data.

The novelty of this project lies in the direct comparison of high-depth long-read datasets from both PacBio and ONT, which has been rare in previous studies. By focusing on these comparisons and considering pipeline factors such as quantification strategies (on transcriptome vs. genome), inclusion of novel transcripts, and downsampling.

3 Relevant work

Very few studies have explored the capabilities and limitations of long-read RNA-seq technologies.

For instance, work by Byrne et al. has demonstrated the potential of long-read sequencing for improving transcriptome annotation and identifying novel transcripts. However, there has been limited research directly comparing the performance of long-read technologies with short-read sequencing for specific RNA-seq applications.

Dong et al. presents both gene- and isoform-level analyses of long-read nanopore transcriptome datasets, using largely conventional methods developed for short-read data. Their study also highlights the limitations of existing methods for isoform identification from long-read data and introduces a new method, FLAMES, to improve isoform-level analysis. However, the datasets contain only a few million reads each, in comparison to the high-depth long-read datasets that we have from PacBio and will be receiving from ONT.

In a recent study assessing RNA sequencing tools, (Dong et al., 2023) compared various analysis tools using deep sequencing data from Oxford Nanopore Technologies and Illumina. However, like mentioned above, this study only employs two DTU methods and one DGE approach.

4 Approach

The first step to this project is researching libraries in R and Python that perform DGE/DTE and DTU and and write scripts to perform the analysis for each method. We aim to implement 3 methods for DGE/DTE (`DESeq2` (Love et al., 2014), `edgeR` (Robinson et al., 2010), and `limma` (Ritchie et al., 2015)) and 2 methods for DTU (`DRIMSeq` (Nowicka and Robinson, 2016) and `satuRn` (Gilis et al., 2021)).

After all methods have been wrapped in convenient scripts, they will be integrated into a `snakemake` workflow, including the existing pipeline to perform quantifications on the raw RNA-seq data, forming a complete workflow for DGE and DTU analysis. This pipeline will then be run on the entire

RNA-seq dataset, sequenced with PacBio and Illumina from the widely used WTC11 iPSC cell line (Kreitzer et al., 2013). Since we are still waiting for nanopore data from ONT, the evaluation of nanopore data from ONT does not belong to the scope of this project.

Finally, an early comparison of one Illumina quantification and three PacBio quantifications on the following three key aspects will be delivered:

a) Quantification accuracy (relative-only, i.e., how well are log-fold changes between conditions captured) on SIRV spike-ins.

b) DTE accuracy on SIRV spike-ins and DGE accuracy on appropriate EC markers reflecting our differentiation.

c) DTU accuracy on SIRV spike-ins.

References

E Birney, JA Stamatoyannopoulos, A Dutta, R Guigó, TR Gingeras, EH Margulies, Z Weng, M Snyder, ET Dermitzakis, and RE Thurman. Baylor college of medicine human genome sequencing center; washington university genome sequencing center; broad institute; children's hospital oakland research institute.(2007). identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447(7146):799–816, 2007.

Rory Bowden, Robert W Davies, Andreas Heger, Alistair T Pagnamenta, Mariateresa de Cesare, Laura E Oikkonen, Duncan Parkes, Colin Freeman, Fatima Dhalla, Smita Y Patel, et al. Sequencing of human genomes with nanopore technology. *Nature communications*, 10(1):1869, 2019.

Ashley Byrne, Anna E Beaudin, Hugh E Olsen, Miten Jain, Charles Cole, Theron Palmer, Rebecca M DuBois, E Camilla Forsberg, Mark Akeson, and Christopher Vollmers. Nanopore long-read rnaseq reveals widespread transcriptional variation among the surface receptors of individual b cells. *Nature communications*, 8(1):16027, 2017.

Xueyi Dong, Luyi Tian, Quentin Gouil, Hasaru Kariyawasam, Shian Su, Ricardo De Paoli-Iseppi, Yair David Joseph Prawer, Michael B Clark, Kelsey Breslin, Megan Iminitoff, et al. The long and the short of it: unlocking nanopore long-read rna sequencing data with short-read differential expression analysis tools. *NAR genomics and bioinformatics*, 3(2):lqab028, 2021.

Xueyi Dong, Mei RM Du, Quentin Gouil, Luyi Tian, Jafar S Jabbari, Rory Bowden, Pedro L Baldoni, Yunshun Chen, Gordon K Smyth, Shanika L Amarasinghe, et al. Benchmarking long-read rna-sequencing analysis tools using in silico mixtures. *Nature Methods*, 20(11):1810–1821, 2023.

Francesca Finotello and Barbara Di Camillo. Measuring differential gene expression with rna-seq: challenges and strategies for data analysis. *Briefings in functional genomics*, 14(2):130–142, 2015.

Jeroen Gilis, Kristoffer Vitting-Seerup, Koen Van den Berge, and Lieven Clement. saturn: Scalable analysis of differential transcript usage for bulk and single-cell rna-sequencing applications. *F1000Research*, 10, 2021.

Philipp Kapranov, Aarron T Willingham, and Thomas R Gingeras. Genome-wide transcription and the implications for genomic organization. *Nature Reviews Genetics*, 8(6):413–423, 2007.

Faith R Kreitzer, Nathan Salomonis, Alice Sheehan, Miller Huang, Jason S Park, Matthew J Spindler, Paweena Lizarraga, William A Weiss, Po-Lin So, and Bruce R Conklin. A robust method to derive functional neural crest cells from human pluripotent stem cells. *American journal of stem cells*, 2(2):119, 2013.

Peng Liang and Arthur B Pardee. Analysing differential gene expression in cancer. *Nature Reviews Cancer*, 3(11):869–876, 2003.

Michael Love, Simon Anders, and Wolfgang Huber. Differential analysis of count data–the deseq2 package. *Genome Biol*, 15(550):10–1186, 2014.

Diego Marques-Coelho, Lukas da Cruz Carvalho Iohan, Ana Raquel Melo de Farias, Amandine Flaig, Jean-Charles Lambert, and Marcos Romualdo Costa. Differential transcript usage unravels gene expression alterations in alzheimer's disease human brains. *npj Aging and Mechanisms of Disease*, 7(1):2, 2021.

Malgorzata Nowicka and Mark D Robinson. Drimseq: a dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research*, 5, 2016.

Anthony Rhoads and Kin Fai Au. Pacbio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 13(5):278–289, 2015.

Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, 26(1):139–140, 2010.

Arun J Singh, Stephen A Ramsey, Theresa M Filtz, and Chrissa Kioussi. Differential gene regulatory networks in development and disease. *Cellular and Molecular Life Sciences*, 75:1013–1025, 2018.

Erwin L Van Dijk, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. Ten years of next-generation sequencing technology. *Trends in genetics*, 30(9):418–426, 2014.