ITM Workflow for handling sc-RNA sequencing isoform analysis

Saad Yousuf
saad.yousuf@metu.edu.tr
Middle East Technical University
Ankara, Çankaya

KEYWORDS

isoforms, single cell RNA sequencing, breast cancer

ACM Reference Format:
Saad Yousuf. 2021. ITM Workflow for handling sc-RNA sequencing isoform analysis. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

1 ABSTRACT

Single-cell RNA(sc-RNA) Sequencing is at the front and foremost go-to analysis for generating useful insights on the genomic nature of individual cells, most notably tumor cells, in order to shed light on possible activities. Yet even more important in this field are the challenges of identifying isoforms; these are different mRNA variants of the same gene which create further specificty in curing cancers as every isoform takes us deeper from the cellular level to individual gene's frequency of occurrence. In the study below, a sample of 551 cell samples across 15371 trasncript ids has been adapted from a common data set of primary breast cancer cells across multiple subtypes. We applied a common framework called the ISOP(ISoform Patterns Pathway) to yield isoform pairs clustered across 6 patterns in our ITM(Isoform TPM(transcripts-per-million) Matrix) Workflow [3], incorporating the analysis of the 3'UTR(untranslated region) as well, our region of interest, in order to find a possible relationship between the two factors [10].

2 INTRODUCTION

Isoforms refer to the varying mRNA transcripts, as in the case of Figure 1.'s pre-mRNA transcripts, that can correspond to varying products(polypeptides or proteins) of the same gene. A particular isoform's dominance in the expression of a gene dictates the frequency of that isoform's product. However, isoform detection is still under due scrutiny given how difficult they are to detect given limitations in the accuracy of sc-RNA analysis machines [5].We propose a new way to analyse sc-RNA data from multiple tumor samples in order to shed information with regards to isoform identification. We have chosen to focus on the results of a sc-RNA pipeline as implemented on cells of primary breast cancer. The genomic institute had captured 551 samples of 11 breast cancer patients across 15731 quality controlled transcripts. Such a bulkeous sample available on the common GEO(gene expression omnibus) database allows us to sufficiently calculate trasncript counts across many cells with different tumor subtypes. We propose adding of ISOform Patterns(ISOP) mixture modelling approach to resolve such preexisting sc-RNA data to allow for isoform analysis.

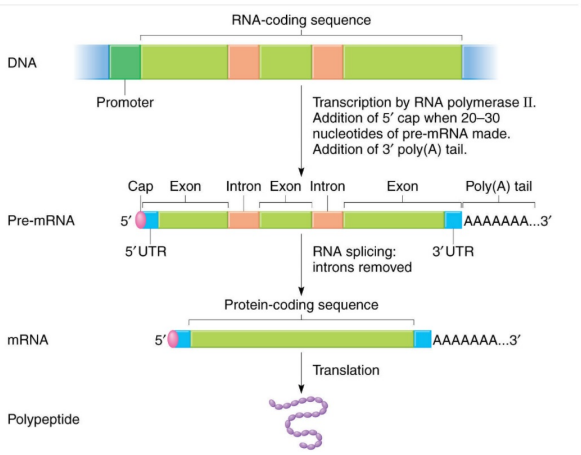

Figure 1: mRNA transcripts and the 3'UTR(untranslated region) on them [9]

The ISOP-R package, also available for use on Github, allows researchers to analyse transcriptional diversity via usage of transcript-cell TPM matrices. After using a common Salmon Workflow of version 1.2 on the CGC(Cancer Genomics Cloud) Server operated by SevenBridges genomic technology, we produced the required transcript-cell matrices as part of Differential Transcript Usage(DTU), defined as 'identifying changes in total activity of transcripts , relative to other transcripts of the same gene'.

Whereas traditional sc-RNA approaches, like those previously implemented on our data, use traditional dimension reduction on gene to cell TPM matrices to create PCA(principle component analysis) and t-SNE(t-distributed Stochastic Neighbor Embedding) clustered plots, our workflow extends this vision. As illustrated on Figure 2., our work can be considered to correspond to Downstream Analysis of a typical sc-RNA pipeline. In order to better show the relation of isoforms of the same gene across multiple cells , we use Gaussian mixture modelling to our advantage [8].
An isoform pair from the same gene is used as the basic unit in the ISOP method that we implement here. Using predicted model parameters, every pair can be fit into preset patterns, of which ISOP has found 6 major categories. In our isoform analysis, we have

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

chosen to focus on the 3'UTR(untranslated region) end of RNA transcripts, given the lack of methods applied on this particular region.

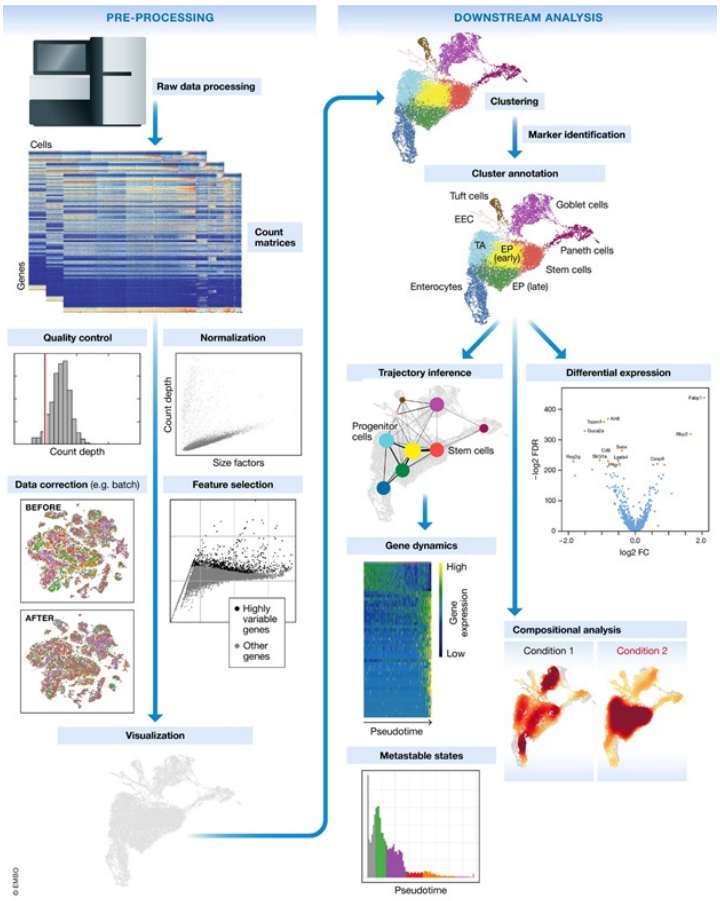

Figure 2: A typical sc-RNA pipeline [5]

3 METHODS

Patients and tumour specimens. A total of 10 of 11 patients diagnosed with invasive ductal carcinoma underwent breast-conserving surgery or total mastectomy without prior treatment. In all, 11 primary tumour specimens (BC01–BC11) were collected and processed for single-cell RNA sequencing. For BC09, two runs of single-cell RNA sequencing were performed and combined for downstream analysis. Molecular subtypes of tumours were predicted using the R package genefu and assigned accordingly as seen in Table.1.

Following the relevant pipeline, in total, 579 single-cell cDNAs were subjected to sc-RNA sequencing.

RNA sequencing and bioinformatics The RNA reads were aligned to the reference sequences using the 2-pass mode of STAR2.4.0b (default parameters) alignment algorithm , and relative gene expression was quantified as transcript per million (TPM) using RSEM v1.2.17 (default parameters) workflow . Isoform expression levels

Table 1: Patient Descriptions

Patient IDs	Cancer Subtype
BC01-2	Luminal A
BC03	Luminal B
BC04-6	HER2
BC07-11	TNBC

for each gene were summed to derive the TPM values. Quality control assessment of aligned single-cell RNA-seq reads was performed using RNA-SeQC57 standards, and the relevant high quality ones were selected.

Data Preprocessing To remove genes with low expression values, the following steps were applied. First, TPM values <1 were considered unreliable and substituted with zero. Second, TPM values were log2-transformed after adding a value of one. Third, genes expressed in less than 10 percent of all tumour groups were removed. In total, 551 samples and 17,779 genes passed the QC criteria. The initial matrix from the above procedures was obtained with the NCBI GEO(Gene Expression Omnibus) Database with Accession Code: GSE75688. The FastQ files were parsed into Seven Bridge's Cancer Genomics Cloud's Salmon Workflow Version 1.2.0 using gencode.v27 as the index file while also using the gencode.v27 GTF file to annotate our file. The resulting matrix file contained TPM values across 551 samples; inclusive of 515 cells' samples mapped onto across Ensemble Ids of 199615 known transcripts. These transcript ids were converted to UCSC ids via the UCSC Genome Browser and the matrix was filtered by removing transcripts expressed in less than 10 percent of the samples to eventually yield our trimmed data of 551 samples across 15731 transcripts(isoforms). The rest of remaining 36 samples are bulkeous samples.

ISOP Application the standard workflow is applied on the 551 by 15731 matrix. In the workflow, the UCSC hg 19 is used as reference annotation to map our UCSC transcript ids, while TPM read counts of less than 3 are filtered out. Gaussian Mixture Modelling is applied across the 4 main cancer subtypes available in the Supplementary Information's link. Every pair's isoform's(if the relevant data was available) was linked with their respective average 3'UTR lengths in order to create the relevant scatterplots colored as per their deciding patterns.

4 RESULTS

The ISOP workflow defines the 6 patterns clustering our isoform pairs as shown in Figure 3.'s table, with special consideration of the heterogeneity of the amount of expression that the isoforms show within that pair. After the relevant pie chart analysis in Figure 4., the 3'UTR lengths of every known isoform per pair were merged to create scatterplots comparing the Lengths as in Figure 5. These UTR Lengths, taken from the UCSC genome database, were merged accordingly to align them with their respective isoforms.

Every spot on figure 4 describes a pair of isoforms from the same gene. In case of a gene having more than 2 known isoforms, there will be multiple spots for a single gene. Initial observations reveal how almost all spots on the five scatterplots are the same, apart

Patterns	Definitions
I	There is no cell-to-cell heterogeneity in the isoform pair.
II	Extension of the I-pattern with an additional mixture component capturing the zero-inflation of cells where isoform expression is not detected, or where isoforms are expressed at close to equal amounts in both isoforms.
V	Unlike the I-pattern, which has a unimodal distribution in both isoforms, the V-pattern generally has a unimodal expression in one of the isoforms and a bimodal distribution in the other isoform.
VI	Extension of the V-pattern with an additional component in the mixture model with its mean close to zero
X	This pattern captures pairs of isoforms with mutually exclusive expression
X1	Extension to X considering zero-inflations

Figure 3: Our 6 Primary Patterns

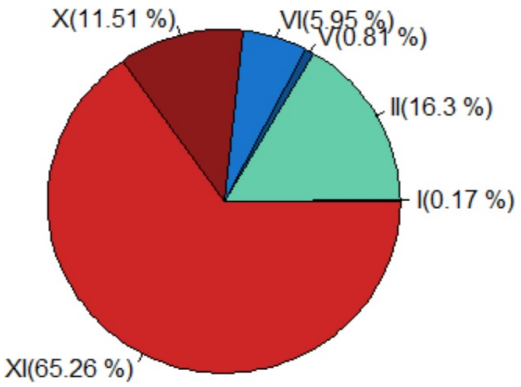

Figure 4: Percentages of every pattern across the main dataset. The remaining piecharts are provided in the Supplementary Information Section

from having a different pattern in a rare number of circumstances. Overall, the XI pattern appears to be the most common across all subtypes. Unlike the average 3' UTR length which lies around 800 nucleotide base pairs, our scatter plots reveal how the majority of these isoform pairs have lengths more than 800 in every subtype [6]. Interestingly, numerous isoforms also lie on the diagonal of

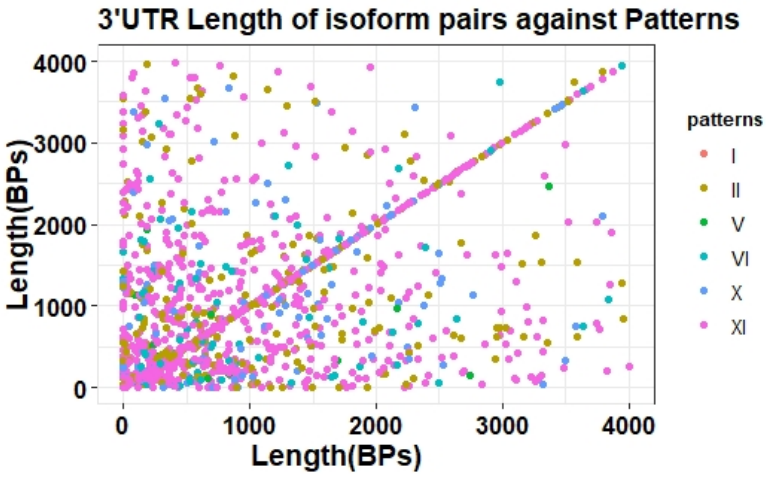

Figure 5: 3'UTR Lengths and our patterns for our main dataset. The rest of subplots are provided in the Supplementary Information Section

every figure for each subtype, suggesting similar lengths of both isoforms in a pair. In our first impressions, there hardly seems to be any co-relation between patterns and the relevant UTR Lengths for every pair, yet considering how more than half of the pairs in every plot is greater than the average length of the 3' UTR of a normal cell, we are hence indicated towards the importance of the 3'UTR.

5 DISCUSSION AND CONCLUSION

In perhaps more simple terms, the II, VI, and XI patterns can be considered to be advanced versions of their original I, V, and X patterns, considering additional zero-inflated data (isoforms in pairs having 0 expression throughout every cell in all samples) in their calculations respectively.

Assuming a sample isoform pair has isoforms A, B(in that particular order) the I pattern can be simplified to mean the expression of A being greater than that of B or vice versa. On the other hand, the V and VI patterns refers to each of the isoform A or B in the pair having a single or double levels of expression respectively in a pair. Lastly, the patterns X,XI refer to the literal case of mutual exclusivity, A OR B. Hence, cancer treatment could focus on targetting isoform products of patterns such as those of the X,XI pattern, where exclusivity would allow physicians to target a single isoform's products more efficiently than the actual gene's less-specific product.

3'UTR lengths have historically been a non-focal point for studies as it lies just after the translation termination codon on any given sample mRNA transcript. Lying just before the 3' Polyadenylation(A) tail at the end of a mRNA transcript, the average length of this region lies between 0 and 4000 nucleotide base pairs(bps), with a usual mean of 800 bps. Turning our attention towards the diagonals observed on Figure 4., these isoforms are those such that their CDS(coding regions) and 3'UTR match, yet this time it is their 5'UTR regions that are different but out of scope for this

study's purpose. These 5'UTR regions have been found to involved in mRNA export and transport which affect possible gene expression [2]. These 3'UTR lengths are significant in the fact that their regions have been thought to have sites for miRNA(micro-RNA, a gene regulator protein) interaction, along with having sites for independent regions which could influence mRNA translation, splicing, polyadenylation and eventually gene expression [1, 7]. Hence, longer isoforms have the tendency to have more binding sites for regulator proteins, leading to decreased gene expression from that particular transcript. From Figures 3 and 4., there is a lack of evidence to realize the above relation given how regulatory proteins in cancer cells could be targeting the 3'UTRs of tumor suppressing genes and hence inhibiting them. This could lead to the production of factors which support the typical abnormal and irregular cell division observed in tumor cells [4]. Conversely, oncogenes might be enhanced in previously normal cells leading to cancerous growth and proliferation as well.

Considering limitations, the number of patients could be increased across more cancer subtypes or even more cancer types in order to check for plausible 3'UTR phenomena. While we use expected(average) 3'UTR lengths from the UCSC Genome Browser's database, the actual cell's UTR length might not be properly considered by our cell gene collection as per the sequencing machine. More accurate and specific 3'UTR focused collection protocols could be advised to use given how cancer cells themselves already seem to cause the creation of a chaotic genetic makeup of a tumor cell.

In conclusion, cancer cells do show a highly chaotic and unstable genetic makeup, which in our case seems to be the main culprit behind the lack of a linear(or any co-relation for that matter) between the ISOP patterns and the 3'UTR lengths. Future studies could apply such a method at the individual patient's cellular levels, finding that patient's own specific isoforms and their eventual genetic products which could be used as susceptible targets by treatments.

6 SUPPLEMENTARY INFORMATION

Github: https://github.com/SYanon/ITM-Workflow

7 ACKNOWLEDGEMENTS

The author would like to acknowledge and is grateful for the expertise and supervision of Prof. Dr. Tolga Can, Department of Computer Engineering, Middle East Technical University, Ankara and Prof.Dr. Ayşe Elif Erson from the Department of Biological Sciences, Middle East Technical University, Ankara in this paper.

REFERENCES

[1] Lucy W Barrett, Sue Fletcher, and Steve D Wilton. 2012. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cellular and molecular life sciences* 69, 21 (2012), 3613–3634.

[2] Can Cenik, Hon Nian Chua, Hui Zhang, Stefan P Tarnawsky, Abdalla Akef, Adnan Derti, Murat Tasan, Melissa J Moore, Alexander F Palazzo, and Frederick P Roth. 2011. Genome analysis reveals interplay between 5 UTR introns and nuclear mRNA export for secretory and mitochondrial genes. *PLoS genetics* 7, 4 (2011), e1001366.

[3] Woosung Chung, Hye Hyeon Eum, Hae-Ock Lee, Kyung-Min Lee, Han-Byoel Lee, Kyu-Tae Kim, Han Suk Ryu, Sangmin Kim, Jeong Eon Lee, Yeon Hee Park, et al. 2017. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature communications* 8, 1 (2017), 1–12.

[4] Ayse Elif Erson-Bensan. 2020. RNA-biology ruling cancer progression? Focus on 3 UTRs and splicing. *Cancer and Metastasis Reviews* 39, 3 (2020), 887–901.

[5] Malte D Luecken and Fabian J Theis. 2019. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular systems biology* 15, 6 (2019), e8746.

[6] Flavio Mignone and Graziano Pesole. 2011. mRNA untranslated regions (UTRs). *eLS* (2011).

[7] Xavier Pichon, Lindsay A Wilson, Mark Stoneley, Amandine Bastide, Helen A King, Joanna Somers, and Anne E Willis. 2012. RNA binding protein/RNA element interactions and the control of translation. *Current Protein and Peptide Science* 13, 4 (2012), 294–304.

[8] Douglas A Reynolds. 2009. Gaussian mixture models. *Encyclopedia of biometrics* 741 (2009), 659–663.

[9] Peter J Russell and Keith Gordey. 2002. *IGenetics.* Number QH430 R87. Benjamin Cummings San Francisco.

[10] Trung Nghia Vu, Quin F Wills, Krishna R Kalari, Nifang Niu, Liewei Wang, Yudi Pawitan, and Mattias Rantalainen. 2018. Isoform-level gene expression patterns in single-cell RNA-sequencing data. *Bioinformatics* 34, 14 (2018), 2392–2400.