BMC Research Notes

RESEARCH NOTE

Open Access

The importance of data transformation in RNA-Seq preprocessing for bladder cancer subtyping

Ariadna Acedo-Terrades[1], Júlia Perera-Bel[1*] and Lara Nonell[2*]

Abstract

Objective RNA-Seq provides an accurate quantification of gene expression levels and it is widely used for molecular subtype classification in cancer, with special importance in prognosis. However, the reliability and validity of these analyses can significantly be influenced by how data are processed. In this study we evaluate how RNA-Seq preprocessing methods influence molecular subtype classification in bladder cancer. By benchmarking various aligners, quantifiers and methods of normalization and transformation, we stress the importance of preprocessing choices for accurate and consistent subtype classification.

Results Our findings highlight that log-transformation plays a crucial role in centroid-based classifiers such as consensusMIBC and TCGAclas, while distribution-free algorithms like LundTax offer robustness to preprocessing variations. Non log-transformed data resulted in low classification rates and poor agreement with reference classifications in consensusMIBC and TCGAclas classifiers. Additionally, LundTax consistently demonstrated better separation among subtypes, compared to consensusMIBC and TCGAclas, regardless of preprocessing methods. Nonetheless, the study is limited by the lack of a true reference for objective assessment of the accuracy of the assigned subtypes. Hence, future work will be necessary to determine the robustness and scalability of the obtained results.

Keywords Molecular subtypes, RNA sequencing, Preprocessing, Bladder cancer

Introduction

While recent advancements in technologies, such as single-cell RNA sequencing and multi-omics approaches, have facilitated novel cancer classifications [1–3], RNA sequencing (RNA-Seq) remains a powerful and cost-effective tool that enables comprehensive transcriptome analysis by providing an accurate quantification of gene expression. RNA-Seq data are largely used for molecular subtype classification of diseases, crucial for improving prognostic accuracy and enhancing understanding of cancer biology [4–9]. In bladder cancer (BC), subtypes have shown to correlate with prognosis and treatment response [10–12]. Pioneering efforts by Robertson et al. defined five molecular subtypes and developed the TCGA classification (TCGAclas) [10], followed by the LundTax classifier developed by Sjöhdal et al. which focused on tumor differentiation and immune response [11]. Building on these efforts, Kamoun et al. unified previously reported molecular taxonomies and developed ConsensusMIBC, a robust classification for BC [12]. However, RNA-Seq preprocessing choices significantly impact the reliability of downstream analyses, including molecular subtype classification [13].

*Correspondence:
Júlia Perera-Bel
jperera@researchmar.net
Lara Nonell
laranonell@vhio.net
[1] Hospital del Mar Research Institute (HMRI), Barcelona, Spain
[2] Bioinformatics Unit, Vall d'Hebron Institute of Oncology, Barcelona, Spain

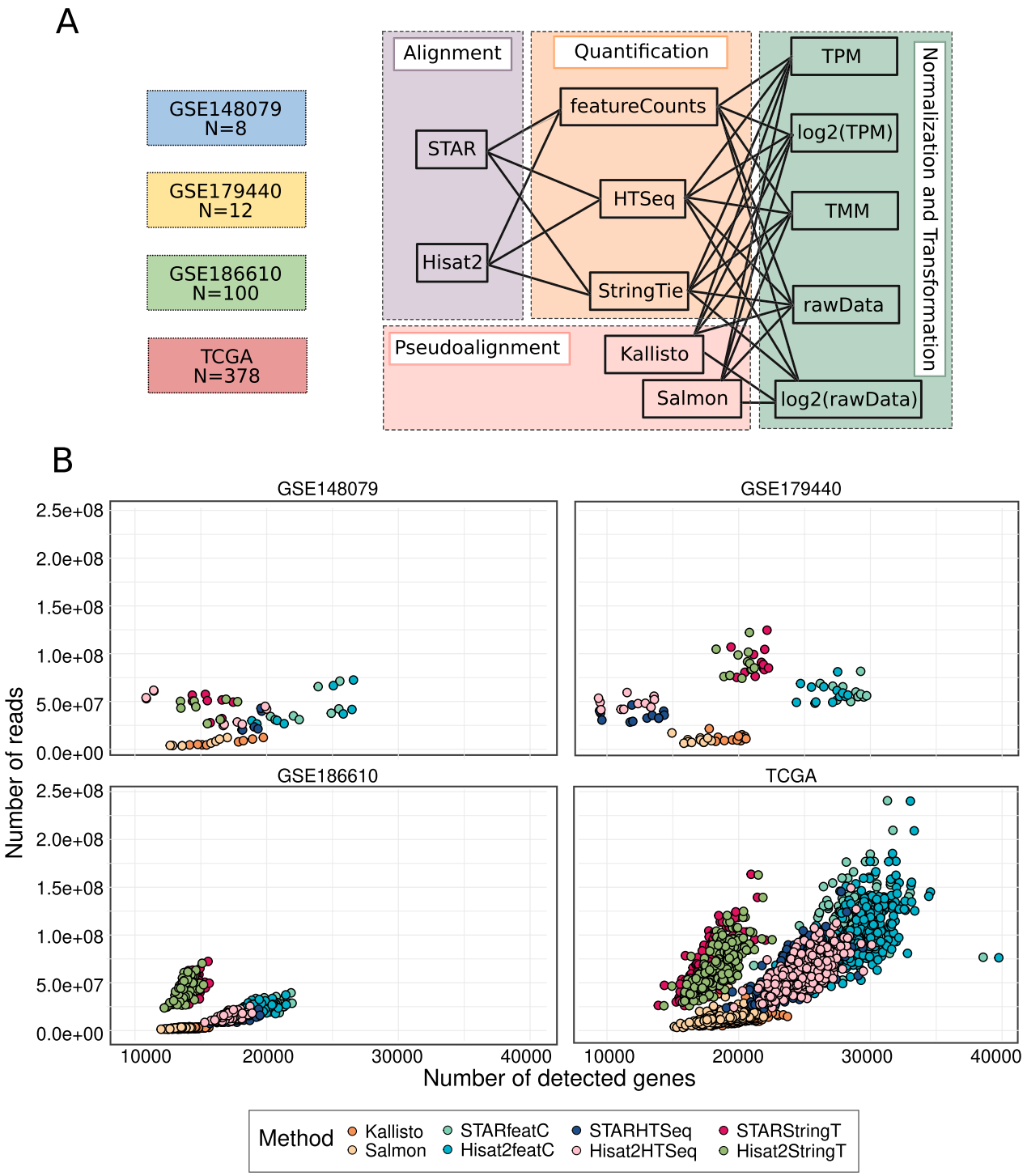

Fig. 1 Overview of methodology and count summaries of benchmarked tools. **A** Combination of datasets and preprocessing methods benchmarked. **B** Total counts for each dataset. Colors indicate distinct preprocessing methods for alignment and quantification. X axis shows the number of detected genes and Y axis shows the number of total counts

RNA-Seq data preprocessing workflow encompassess several steps to transform raw reads into a data matrix with the gene expression levels. The first step is the alignment of the reads to a reference genome or transcriptome using tools like STAR [14] or Hisat2 [15] among others. STAR typically offers higher accuracy in

aligning complex and repetitive regions of the genome, while Hisat2 tends to be more precise in detecting single nucleotide polymorphisms. The next step is gene expression quantification, where the number of reads aligned are quantified to determine the gene expression levels. Widely used methods such as featureCounts [16] and HTSeq [17] perform straightforward read-counting methods, whereas StringTie [18] employs an EM algorithm to estimate transcript abundances, which helps in resolving ambiguities when the reads map to multiple transcripts. Although alignment and quantification are typically performed as separate steps, they can be performed together using pseudoaligners like Kallisto [19] and Salmon [20].

Gene expression levels are affected by biological and technical variability (e.g. sample quality, sequencing depth, batch effect). Normalization and transformation, such as transcript per million (TPM) or Trimmed Mean of M-values (TMM) [21], adjust raw counts to make expression values comparable. Log transformation is also essential as it balances skewed data and stabilizes the variance reducing the impact of outliers.

The main objective of this study is to evaluate the impact of different RNA-Seq preprocessing tools methods on molecular subtype classification in BC. We foresee that by ensuring the best choice for every preprocessing step, we can maximize the reliability of subtype classification [22].

Methods

We evaluated twelve combinations of preprocessing methods on three molecular subtype classifiers using four bladder cancer datasets: GSE148079 (n = 8) [23], GSE179440 (n = 12) [24], GSE186610 (n = 100) [25] and TCGA (n = 378) [10] (Fig. 1A). The preprocessing workflow included quality control (Supplementary Fig. 1), alignment (STAR, Hisat2), quantification (FeatureCounts, StringTie, HTSeq), and pseudoalignment (Salmon, Kallisto), using hg38 annotations. Different normalization and/or transformation methods (TMM, TPM and log2TPM), together with rawData and log2rawData were studied. ConsensusMIBC, LundTax, and TCGAclas classifiers were assessed using separation and coincidence metrics. (Supplementary Methods).

Results

The number of total counts and detected genes were consistent among the different alignment and quantification methods across datasets. STAR and Hisat2 outperformed the two pseudoaligners (Kallisto and Salmon) in the number of counts. Even though both pseudoaligners yielded the lowest number of counts, they detected an equivalent number of genes as the StringTie quantifier, regardless of the aligner. In contrast, featureCounts, followed by HTSeq, consistently detected the highest number of genes, except for the GSE179440 dataset, in which StringTie detected more genes than HTSeq. The number of counts was proportional to the number of sequenced reads, being the highest for the TCGA (average of 62 million reads) (Fig. 1B, Supplementary Table 1 and Supplementary Methods). Overall, the best performing methods across datasets were STAR or Hisat2 combined with featureCounts since they retrieve the highest number of genes (Fig. 1B, Supplementary Table 1).

We next evaluated if the differences observed in the number of counts and number of detected genes have an impact on BC subtyping. Our results on the consensusMIBC classifier [12] showed that, across all methods and datasets, using non log-transformed data (rawData or TPM) resulted in low correlation values and many unclassified samples (up to 87.5–100% in the two smallest datasets and 34.4%-64% in the largest). Even when few or no samples were unclassified (eg. featureCounts), correlation values were consistently higher for log transformed data, being log2TPM and TMM the highest. Of note, HTSeq and StringTie quantifiers were more affected by this issue, regardless of the aligner and specially in combination with TMM normalization (e.g. 0% vs 1.06–34.4% of unclassified samples, using log-transformed vs non log-transformed data in TCGA dataset). Similar results were observed with the pseudoaligners (0% vs 0–16.7%) (Fig. 2, Supplementary Table 2).

Additionally, we assessed two other BC classifiers: TCGAclas [10] and LundTax [26]. The results for the TCGAclas classifier showed a high variability across

(See figure on next page.)

Fig. 2 Heatmap of consensusMIBC classification using benchmarked tools. Each column shows the results for the different RNA-Seq combinations of aligner, quantifier, of normalization and transformation method. Each row represents a sample from several public datasets (GSE148079, GSE179440, GSE186610 and TCGA). Colors of the heatmap indicate the predicted molecular subtype. Column on the left represents the reference for each sample (i.e. the most commonly predicted subtype). Luminal papillary (LumP), luminal unstable (LumU), luminal non-specified (LumNS), stroma rich (Stroma-rich), basal squamous (Ba/Sq) and neuroendocrine-like (NE-like). Kallisto, Salmon, STAR + featureCounts (STAR + featC), STAR + HTSeq, STAR + StringTie (STAR + StringT), Hisat2 + featureCounts (Hisat2 + featC), Hisat2 + HTSeq, Hisat2 + StringTie (Hisat2 + StringT) and Reference. Barplots show the mean and standard deviation of the correlation values for each methodology

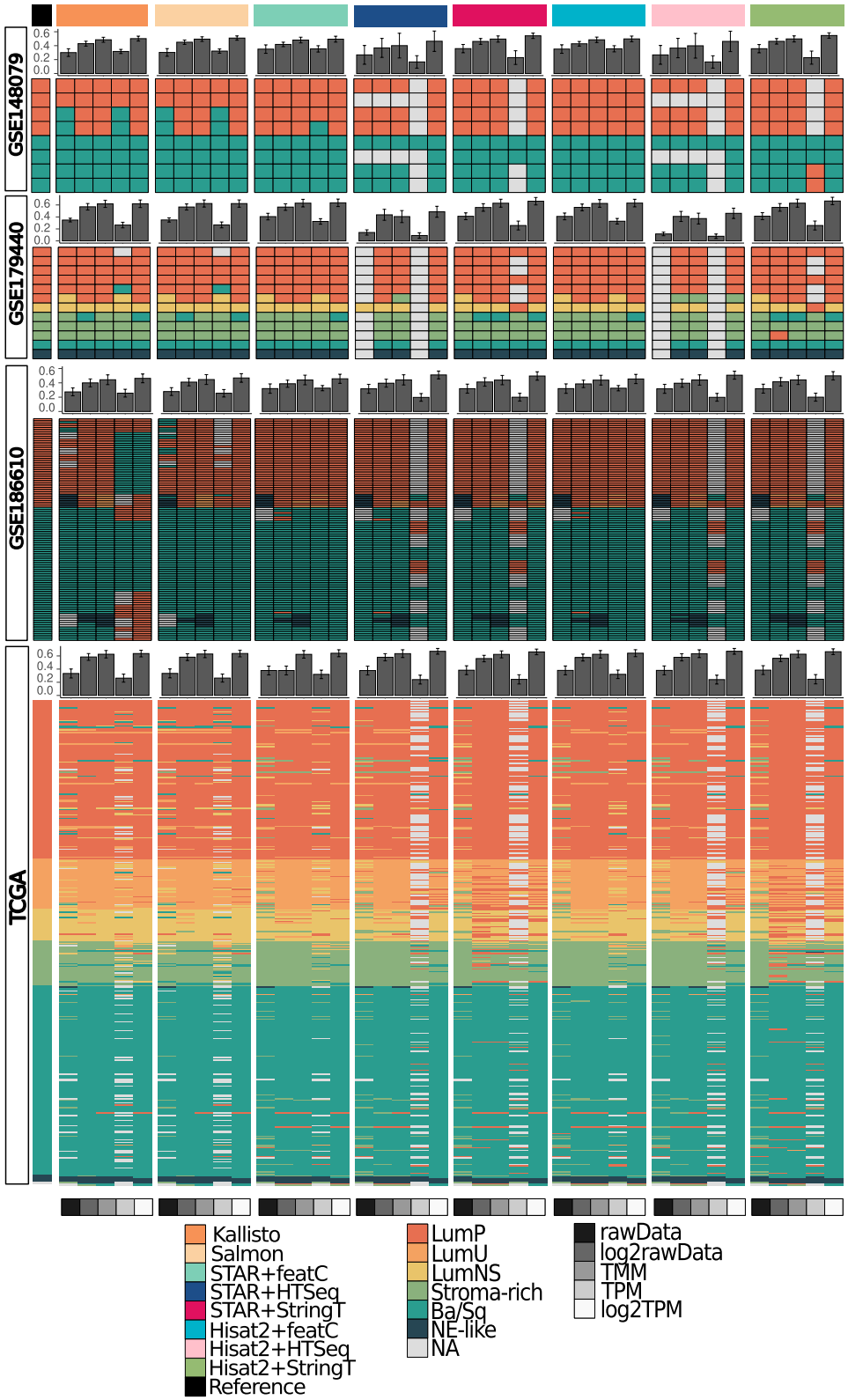

Fig. 2 (See legend on previous page.)

methods, especially when comparing log-transformed and non log-transformed data (Supplementary Fig. 2). In contrast, LundTax appeared very stable, even across log and non-log data (Supplementary Fig. 3). For instance, the TCGAclas on TPM compared to log2TPM predicted 53% vs 18% of basal squamous, 37% vs 68% of luminal papillary (GSE186610, STAR + featureCounts). In contrast, the subtype classification of the same dataset using consensusMIBC and LundTax classifiers were more consistent across both pipelines, with variations of at most 3%. These results were observed in most of the different preprocessing pipelines that were applied, such as Hisat2 + featureCounts, Kallisto or STAR + HTSeq (Supplementary Figs. 2, 3 and Supplementary Tables 3, 4).

In an effort to fairly compare across the three classifiers, we evaluated the distribution of their scores (Supplementary Methods). The distribution of the non log-transformed data was distinctly separated from that of the log-transformed for consensusMIBC and TCGAclas classifiers. Importantly, all scores below the minimum confidence threshold belonged to non log data (Fig. 3A). In contrast, LundTax scores distribution was not influenced by log transformation (Supplementary Tables 5 and 6).

To assess the performance, two metrics were evaluated: separation and coincidence. Separation shows how a sample is representative of its subtype whereas coincidence is the percentage of samples corresponding to the most frequent subtype (Supplementary Methods, Supplementary Fig. 4). The most important differences were influenced by the classifier and the normalization and transformation methods. Specifically, TPM normalization combined with HTSeq or StringTie exhibited the lowest coincidence scores across both the consensusMIBC (0.22–0.40) and TCGAclas (0.46–0.66) classifiers. LundTax was not influenced by log transformation and showed more stable metrics. Finally, pseudoaligners, when combined with log transformed data, performed similarly to other quantifiers, regardless of the aligner (Supplementary Fig. 4, Supplementary Table 7).

Of note, consensusMIBC and TCGAclas classifiers demonstrated low separation values regardless of the preprocessing methods used (0.1–0.32), indicating that the samples were less representative and less distinctly separated from other molecular subtypes. Conversely, the LundTax classifier achieved consistently the highest separation values (0.45–0.63) across different methods (Supplementary Fig. 4, Supplementary Table 7).

As observed in Fig. 2, the stability of the molecular subtype classification was highly sample-dependent. We studied the changes in molecular subtype classification across datasets and methods to analyze their exchangeability. The most frequent subtypes, Ba/Sq and LumP, were the most stable ones (93.8%; 90.3%), followed by the less frequent, NE-like (89.7%). As expected, we observed high exchangeability among luminal subtypes (LumP, LumU, and LumNS). Around 11.4% of LumNS and 13.2% of LumU were also classified as LumP, the most abundant luminal subtype. In contrast, the likelihood of a luminal subtype being classified as a non-luminal subtype was low or zero, as in the case of NE-like (Fig. 3B, Supplementary Table 8). The likelihood of interchange among non-luminal subtypes (Stroma-rich, Ba/Sq, and NE-Like) was not as high as among luminal subtypes (0–7.22%), but it was still higher than being reclassified as any of the luminal subtypes. Stroma-rich showed similar reclassification proportions with most subtypes (Fig. 3B, Supplementary Table 8).

Conclusions

Evaluating the impact of RNA-Seq preprocessing tools is essential for standardizing RNA-Seq pipelines, and thus maximizing the reliability of molecular subtype classification in bladder cancer. Our findings suggest that preprocessing steps including read alignment, gene quantification, normalization and data transformation, have an impact on the downstream analysis of molecular subtypes in bladder cancer. Despite some preprocessing methodologies being more affected by normalization and transformation, results showed that the critical factor when using consensusMIBC [12] and TCGAclas [10] classifiers is the log transformation of the data. Both classifiers are centroid-based and therefore highly sensitive to the distribution of the data. Indeed, Kamoun et al. recommend using log-transformed data to reduce common issues such as outliers or skewness [9]. The LundTax classifier [26], in contrast, is less affected by

(See figure on next page.)

Fig. 3 Score distribution across classifiers and molecular subtypes interconnections of the consensusMIBC classifier. **A** Density plot of the score values distribution for each classifier, splitted by log-transformed and non log-transformed values. Shadows in each density plot represent the 95% confidence intervals. **B** Directed graph with reclassification percentages. Each node represents a consensusMIBC molecular subtype. The size shows the number of samples classified as that specific molecular subtype across methods. The distance between nodes and the width of the links represent the proportion of reclassified samples across methods between the subtypes. Edge colors are inherited by the original node. Edges below 1% were filtered out

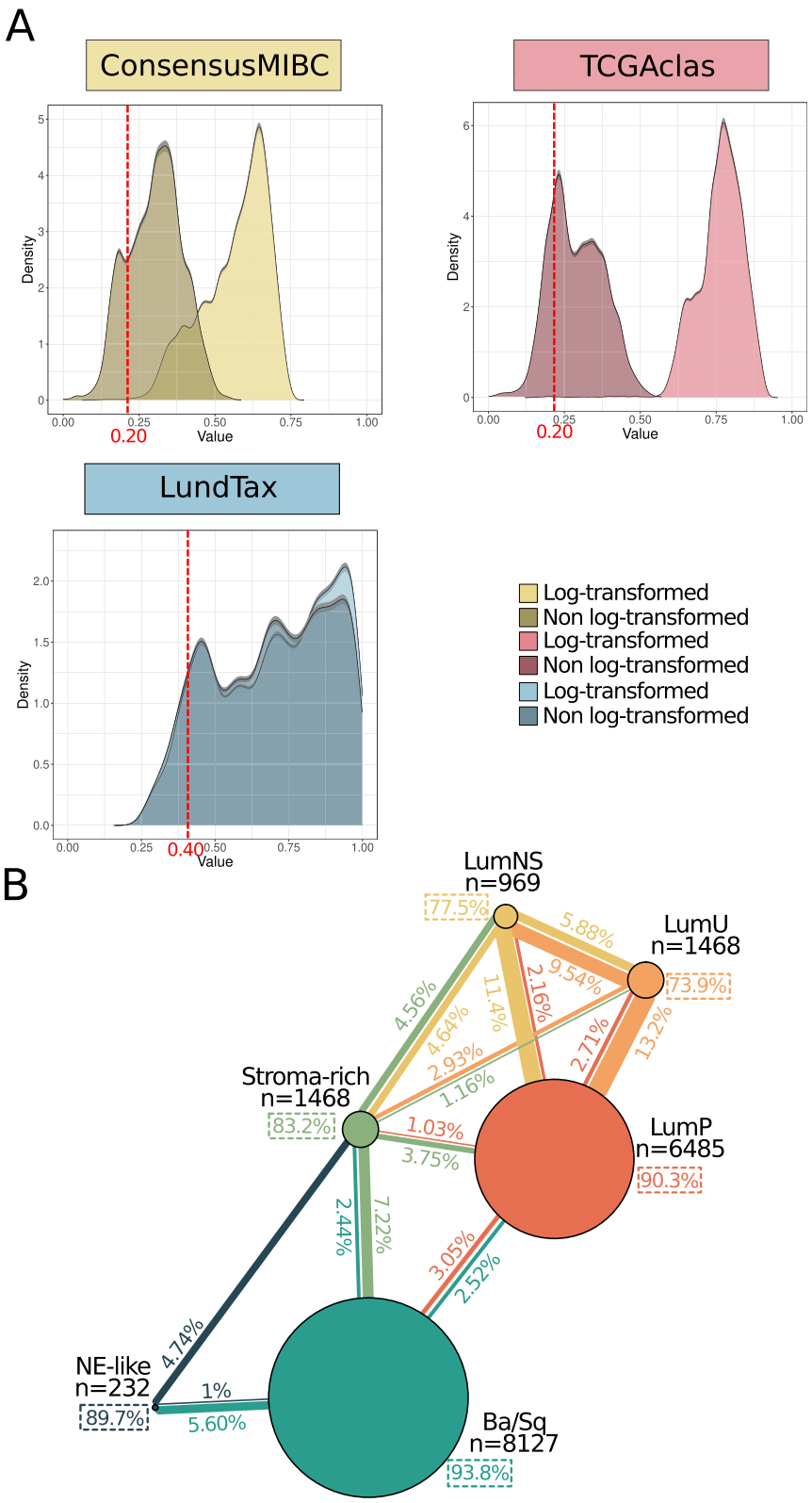

Fig. 3 (See legend on previous page.)

log-transformation. This may be attributed to the use of a Random Forest algorithm, which relies on relative comparisons rather than absolute values [27–29].

We examined the interconnections among molecular subtypes within the consensusMIBC classification to assess their exchangeability. As expected and previously reported, luminal subtypes showed a higher likelihood of being reclassified within the luminal group compared to non-luminal subtypes [28]. However, reclassification between luminal and non-luminal subtypes is not likely to happen, suggesting a distinct separation between these groups greater than differences across preprocessing methods [27, 30].

Based on our comprehensive evaluation and previous studies, both pseudoaligners (Kallisto [19] and Salmon [20]) and the featureCounts [16] quantifier combined with any aligner (STAR [14] or Hisat2 [15]) are effective for molecular subtype classification. While featureCounts provides better sensitivity in detecting lowly expressed genes, Kallisto and Salmon offer notable advantages in speed and memory efficiency [31]. Our findings suggest that, although Kallisto and Salmon identify a smaller number of genes, this does not compromise the accuracy of subtype classification, likely because the classifiers rely only on a subset of genes.

Moreover, we evaluated the classification rates to compare among the results obtained by the three classifiers. Our results showed low classification rates for HTSeq [17] and StringTie [18] quantifiers, regardless of the aligner, across all datasets and classifiers, indicating a low-accuracy in subtype assignment. The highest classification rates were achieved by featureCounts combined with STAR or Hisat2 aligners, along with Kallisto and Salmon.

We have tested gene-based quantifiers (featureCounts and HTSeq), as well as transcript quantifiers (StringTie), with the objective of assessing subtyping classification for which most methods are based on gene estimations. Despite gene-level results being often more accurate, powerful and interpretable than transcript-level results, difference between transcript-based and exon-based quantifiers is expected to be relatively minor when performing molecular subtype classification compared to other analyses, such as differential expression analysis, where gene quantification plays a crucial role [32–34].

In conclusion, our results show that log-transformation is a required step for centroid-based classifiers such as consensusMIBC and TCGAclas, that were trained on log-transformed data. In addition, we recommend distribution-free algorithms such as LundTax, which show less sensitivity to preprocessing steps. Future research should be focused on the validation of the robustness and scalability from the findings of our study in MIBC as well as on other cancer type molecular classifiers. Having standardized workflows could improve the application of molecular subtype classification across cancers, enabling a more thorough study of their association with treatment response and promoting the development of personalized therapeutic strategies in clinical practice.

Limitations

The main limitation of this study is the lack of a true reference to compute objective evaluation metrics, which makes it challenging to assess the accuracy and reliability of assigned subtypes. Coincidence can also be misleading as it measures the agreement of an assigned subtype with the most frequent subtype, without considering subtype designation accuracy.

Additionally, relying only on limited methods and datasets might introduce bias and miss important insights, compromising the reliability of the findings. Nevertheless, our findings reveal parallel constraints in predictive models development, which are heavily reliant on the datasets and methods used, particularly when external validation is not feasible.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13104-025-07138-x.

Supplementary material 1.

Supplementary material 2.

Supplementary material 3.

Acknowledgements

Not applicable.

Author contributions

A.A.-T. compiled the samples and performed all computational analyses. L.N. contributed to the conceptualization of the study. J.P.-B., and L.N. supervised the study. All authors wrote the manuscript.

Funding

The work was supported by the following grants and agencies: Project PI19/00004 and PI22/00171, funded by Instituto de Salud Carlos III (ISCIII) and co-funded by the European Union; a grant from FIS-ISCIII (FI20/00095), 2021SGR00042 by Generalitat de Catalunya.

Availability of data and materials

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The data analyzed in this study can be accessed in the following databases: GSE148079 [23], GSE179440 [24], GSE186610 [25] and TCGA [10]. Additionally, scripts used for the different preprocessing steps are also available in the following github repository: https://github.com/MARData-BU/Methods_for_preprocessing_steps_RNA-Seq.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests
The authors declare no competing interests.

Received: 29 October 2024 Accepted: 4 February 2025
Published online: 10 February 2025

References

1. Yu Lu, et al. Prognostic significance of lineage diversity in bladder cancer revealed by single-cell sequencing. Front Genet. 2022;13:862634. https://doi.org/10.3389/fgene.2022.862634.
2. Al-Ghafer IA, et al. NMF-guided feature selection and genetic algorithm-driven framework for tumor mutational burden classification in bladder cancer using multi-omics data. Netw Model Anal Health Inf Bioinform. 2024. https://doi.org/10.1007/s13721-024-00460-7.
3. Sheng L, et al. Integrated analysis of bulk and single-cell RNA -seq data reveals cell differentiation-related subtypes and a scoring system in bladder cancer. J Cell Mol Med. 2024;28(19): e70111. https://doi.org/10.1111/jcmm.70111.
4. Perou CM, Sørlie T, Eisen MB, Van De Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge Ø, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale A-L, Brown PO, Botstein D. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747–52. https://doi.org/10.1038/35021093.
5. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511(7511):543–50. https://doi.org/10.1038/nature13385.
6. Abeshouse A, Ahn J, Akbani R, Ally A, Amin S, Andry CD, Annala M, Aprikian A, Armenia J, Arora A, Auman JT, Balasundaram M, Balu S, Barbieri CE, Bauer T, Benz CC, Bergeron A, Beroukhim R, Berrios M, Zmuda E. The molecular taxonomy of primary prostate cancer. Cell. 2015;163(4):1011–25. https://doi.org/10.1016/j.cell.2015.10.025.
7. Guinney J, Dienstmann R, Wang X, De Reyniès A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, Bot BM, Morris JS, Simon IM, Gerster S, Fessler E, De Sousa E, Melo F, Missiaglia E, Ramay H, Barras D, Tejpar S. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015;21(11):1350–6. https://doi.org/10.1038/nm.3967.
8. Bailey P, Chang DK, Nones K, Johns AL, Patch A-M, Gingras M-C, Miller DK, Christ AN, Bruxner TJC, Quinn MC, Nourse C, Murtaugh LC, Harliwong I, Idrisoglu S, Manning S, Nourbakhsh E, Wani S, Fink L, Holmes O, Grimmond SM. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature. 2016;531(7592):47–52. https://doi.org/10.1038/nature16965.
9. Lindskrog SV, et al. An integrated multi-omics analysis identifies prognostic molecular subtypes of non-muscle-invasive bladder cancer. Nat Commun. 2021;12(1):2301. https://doi.org/10.1038/s41467-021-22465-w.
10. Robertson AG, Kim J, Al-Ahmadie H, Bellmunt J, Guo G, Cherniack AD, Hinoue T, Laird PW, Hoadley KA, Akbani R, Castro MAA, Gibb EA, Kanchi RS, Gordenin DA, Shukla SA, Sanchez-Vega F, Hansel DE, Czerniak BA, Reuter VE, Lerner SP. Comprehensive molecular characterization of muscle-invasive bladder cancer. Cell. 2018;174(4):1033. https://doi.org/10.1016/j.cell.2018.07.036.
11. Sjödahl G. Molecular subtype profiling of urothelial carcinoma using a subtype-specific immunohistochemistry panel. In: Schulz WA, Hoffmann MJ, Niegisch G, editors. Urothelial Carcinoma, vol. 1655. New York: Springer; 2018. p. 53–64. https://doi.org/10.1007/978-1-4939-7234-0_5.
12. Kamoun A, De Reyniès A, Allory Y, Sjödahl G, Robertson AG, Seiler R, Hoadley KA, Groeneveld CS, Al-Ahmadie H, Choi W, Castro MAA, Fontugne J, Eriksson P, Mo Q, Kardos J, Zlotta A, Hartmann A, Dinney CP, Bellmunt J, Zlotta A. A consensus molecular classification of muscle-invasive bladder cancer. Eur Urol. 2020;77(4):420–33. https://doi.org/10.1016/j.eururo.2019.09.006.
13. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17(1):13. https://doi.org/10.1186/s13059-016-0881-8.
14. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635.
15. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–15. https://doi.org/10.1038/s41587-019-0201-4.
16. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30. https://doi.org/10.1093/bioinformatics/btt656.
17. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–9. https://doi.org/10.1093/bioinformatics/btu638.
18. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5. https://doi.org/10.1038/nbt.3122.
19. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525–7. https://doi.org/10.1038/nbt.3519.
20. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417–9. https://doi.org/10.1038/nmeth.4197.
21. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25. https://doi.org/10.1186/gb-2010-11-3-r25.
22. Van R, Alvarez D, Mize T, Gannavarapu S, Chintham Reddy L, Nasoz F, Han MV. A comparison of RNA-Seq data preprocessing pipelines for transcriptomic predictions across independent studies. BMC Bioinformatics. 2024;25(1):181. https://doi.org/10.1186/s12859-024-05801-x.
23. Iyyanki T, et al. Subtype-associated epigenomic landscape and 3D genome structure in bladder cancer. Genome Biol. 2021;22(1):105. https://doi.org/10.1186/s13059-021-02325-y.
24. Feng C, et al. Integrative transcriptomic, lipidomic, and metabolomic analysis reveals potential biomarkers of basal and luminal muscle invasive bladder cancer subtypes. Front Genet. 2021;12:695662. https://doi.org/10.3389/fgene.2021.695662.
25. Green JL, et al. Molecular characterization of type I IFN-induced cytotoxicity in bladder cancer cells reveals biomarkers of resistance. Mol Therapy Oncolytics. 2021;23:547–59. https://doi.org/10.1016/j.omto.2021.11.006.
26. Cotillas EA, Bernardo C, Veerla S, Liedberg F, Sjödahl G, Eriksson P. A versatile and upgraded version of the LundTax classification algorithm applied to independent Cohorts. J Mol Diagn. 2024. https://doi.org/10.1016/j.jmoldx.2024.08.005.
27. Sjödahl G, Lauss M, Lövgren K, Chebil G, Gudjonsson S, Veerla S, Patschan O, Aine M, Fernö M, Ringnér M, Månsson W, Liedberg F, Lindgren D, Höglund M. A molecular taxonomy for urothelial carcinoma. Clin Cancer Res. 2012;18(12):3377–86. https://doi.org/10.1158/1078-0432.CCR-12-0077-T.
28. Eriksson P, Marzouka N-A-D, Sjödahl G, Bernardo C, Liedberg F, Höglund M. A comparison of rule-based and centroid single-sample multiclass predictors for transcriptomic classification. Bioinformatics. 2022;38(4):1022–9. https://doi.org/10.1093/bioinformatics/btab763.
29. De Jong JJ, Zwarthoff EC. Molecular and clinical heterogeneity within the luminal subtype. Nat Rev Urol. 2020;17(2):69–70. https://doi.org/10.1038/s41585-019-0262-7.
30. Choi W, Czerniak B, Ochoa A, Su X, Siefker-Radtke A, Dinney C, McConkey DJ. Intrinsic basal and luminal subtypes of muscle-invasive bladder cancer. Nat Rev Urol. 2014;11(7):400–10. https://doi.org/10.1038/nrurol.2014.129.
31. Zhang C, Zhang B, Lin L-L, Zhao S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. BMC Genom. 2017;18(1):583. https://doi.org/10.1186/s12864-017-4002-1.
32. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research. 2016. https://doi.org/10.12688/f1000research.7563.2.
33. Sarantopoulou D, Brooks TG, Nayak S, Mrčela A, Lahens NF, Grant GR. Comparative evaluation of full-length isoform quantification from RNA-Seq. BMC Bioinform. 2021;22(1):266. https://doi.org/10.1186/s12859-021-04198-1.
34. Zhao S, Xi L, Zhang B. Union exon based approach for RNA-Seq gene quantification: to be or not to be? PLoS ONE. 2015;10(11): e0141910. https://doi.org/10.1371/journal.pone.0141910.

Publisher's Note