Towards a gold-standard workflow for the profiling of cell-free RNA in blood plasma

Cristina Tuñí i Domínguez[1], Lluc Cabús[1], Sonia Belmonte[1], Phil Sanders[1], Marc Weber[1], Julien Lagarde[1]

[1] *Flomics Biotech, Barcelona, Spain. E-mail: cristina.tuni@flomics.com*

Summary

Liquid biopsies, particularly those using plasma cell-free RNA (cfRNA), are emerging as tools for early cancer detection due to their non-invasive nature and high sensitivity. However, effectively using cfRNA as a biomarker involves several technical challenges both at the wet and dry lab levels, including acquiring quality cfRNA libraries, reducing genomic DNA contamination, and refining bioinformatics analysis pipelines. NGS of cfRNA has gained interest recently and lacks a gold standard, making the evaluation of methodologies more difficult. Our study addresses these issues by conducting a comparative analysis of cfRNA-Seq datasets, both from public sources and our in-house samples. We focus on comparing quality control metrics to identify key differences in the data sources. Despite consistently high exon mapping rates, our study reveals significant variability in metrics such as mapped read percentage and library complexity among different studies.

By analyzing successive iterations of our experimental protocol, we have assessed the quality improvement in cfRNA-Seq data. This emphasizes the need for continuous quality control and protocol refinement in the field for more reliable results.
This research enhances cfRNA-Seq data processing and analysis, supporting the development of precise diagnostic methods. We aim to establish benchmarks for sample quality and bioinformatics analysis in cfRNA-Seq, advancing non-invasive cancer diagnostics. Our findings provide a foundation for future validation studies across diverse populations and cancer types, potentially improving early detection and patient outcomes.

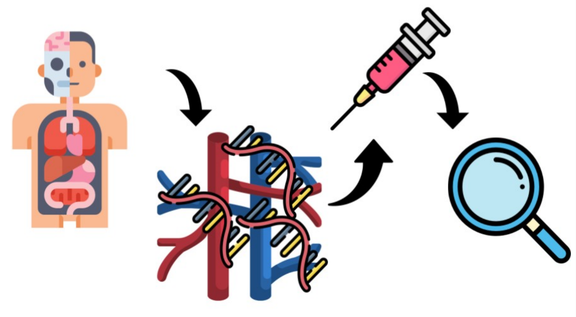

Figure 1. *Schematic representation of cell free RNA liquid biopsy*

Selection of publicly available datasets

We selected **cfRNA** samples from 8 studies that had the raw FASTQ files available on SRA. A 100 or less samples were selected at random, downloaded, and sub-sampled to have 1 million read pairs. Additionally, we included a study that used ***cfDNA*** and another one that used ***bulk tissue RNA***.

Study	Source	RNA extraction	DNAse	Library prep kit	Nº samples
Flomics 1st generation	cfRNA	Plasma/Serum Purification kit (Norgen)	No	CORALL Total	63
Flomics 2nd generation	cfRNA	Maxwell RSC	Yes	SMARTer pico	8
Flomics 3rd generation	cfRNA	Maxwell RSC	Yes (x2)	SMARTer pico	16
Block 2022 [1]	cfRNA	Qiagen RNeasy	No	SMARTer pico	41
Lu 2021 [2]	cfRNA	Plasma/Serum Purification kit (Norgen)	Yes	SMARTer pico	65
Lu 2022 [3]	cfRNA	Plasma/Serum Purification kit (Norgen)	Yes	SMARTer pico	100
Roskams 2022 [4]	cfRNA	Plasma/Serum Purification kit (Norgen)	Yes	SMARTer pico	90
Ngo 2018 [5]	cfRNA	Plasma/Serum Purification kit (Norgen)	Yes	SMARTer pico	15
Ibarra 2020 [6]	cfRNA, exome capture	QIAamp Circulating Nucleic Acid Kit (Qiagen)	Yes	Unspecified	100
Toden 2020 [7]	cfRNA, exome capture	QIAamp Circulating Nucleic Acid Kit (Qiagen)	No	Swift 2S kit	100
Chalasani 2021 [8]	cfRNA, exome capture	QIAamp Circulating Nucleic Acid Kit (Qiagen)	Yes	Unspecified	100
ENCODE 2023 (bulk RNA) [9]	Tissue RNA	RNeasy Fibrous Tissue Mini Kit (Qiagen)	Yes	DNA sample kit (Illumina)	82
Wei 2020 (cfDNA) [10]	cfDNA	NA	No	Truseq Nano DNA	10

Table 1. *Characteristics of the datasets used in the analysis*

Bioinformatic analysis

The same version of the in-house **Flomics/rnaseq** analysis pipeline (***Figure 2*****),** made with Nextflow and based on **nf-core/rnaseq** [11], was run for each dataset, with adequate parameters dependent on the sample processing. **Kraken2** was added to the pipeline to detect bacterial contamination, essential for cfRNA experiments due to the low amount of RNA material. The **Flomics QC** module aggregates several quality control metrics into a single TSV file. This was used as input to plot results metrics, useful to compare the quality of the samples between datasets, and overall performance of the dataset.

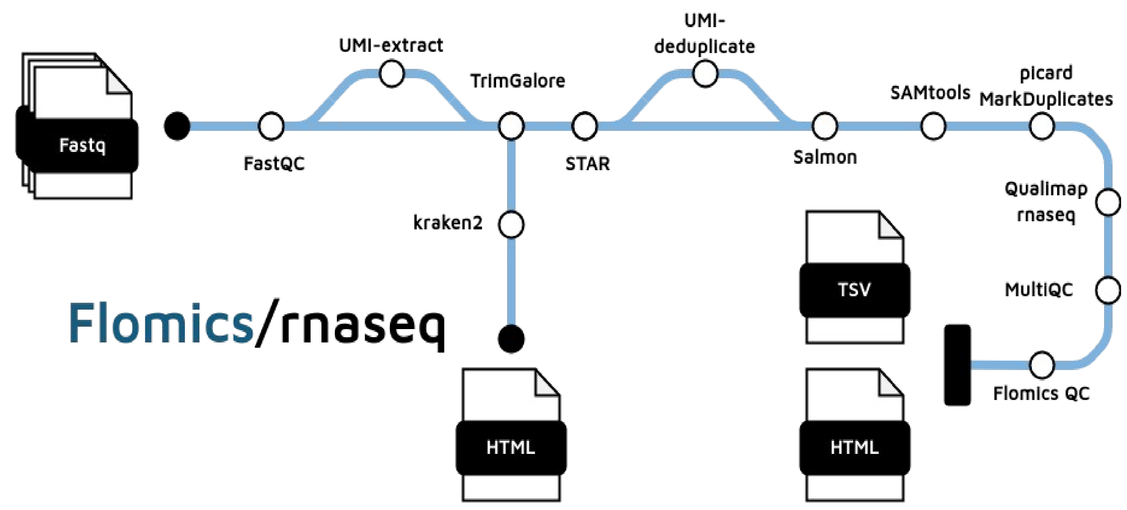

Figure 2. *Schematic representation of the analysis pipeline*

Results

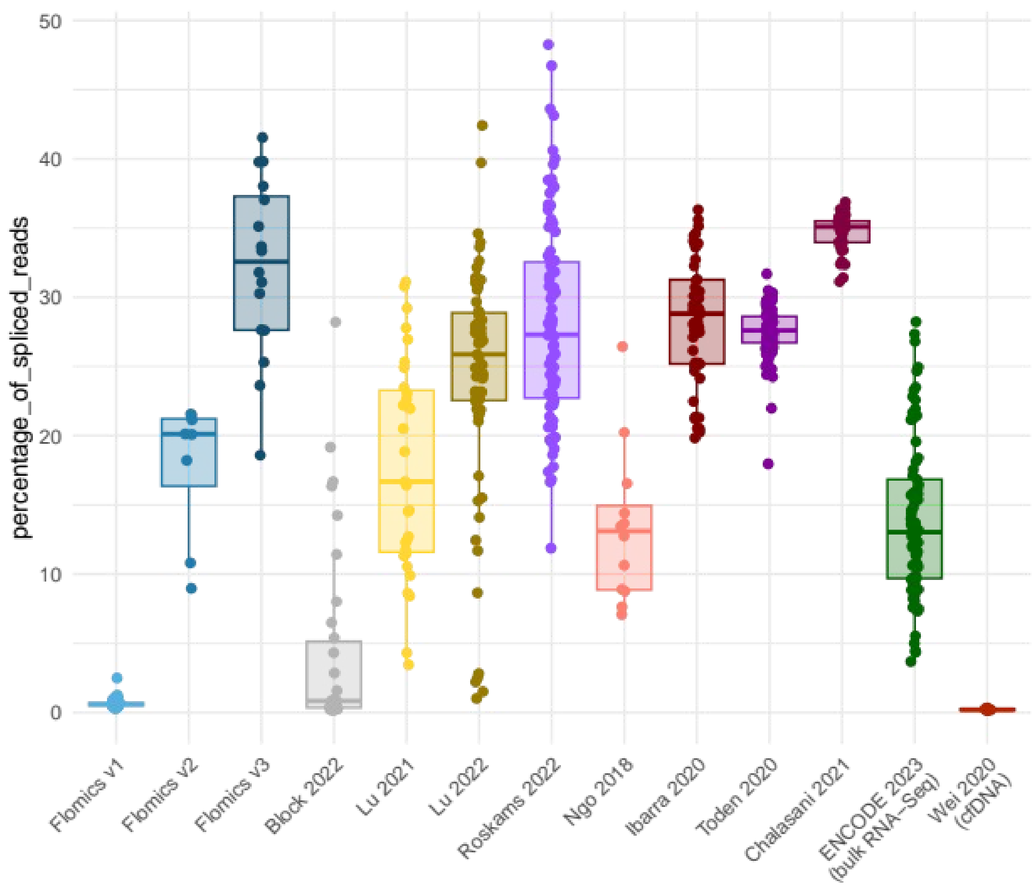

Figure 3. *Boxplot representation of the percentage of spliced reads.*

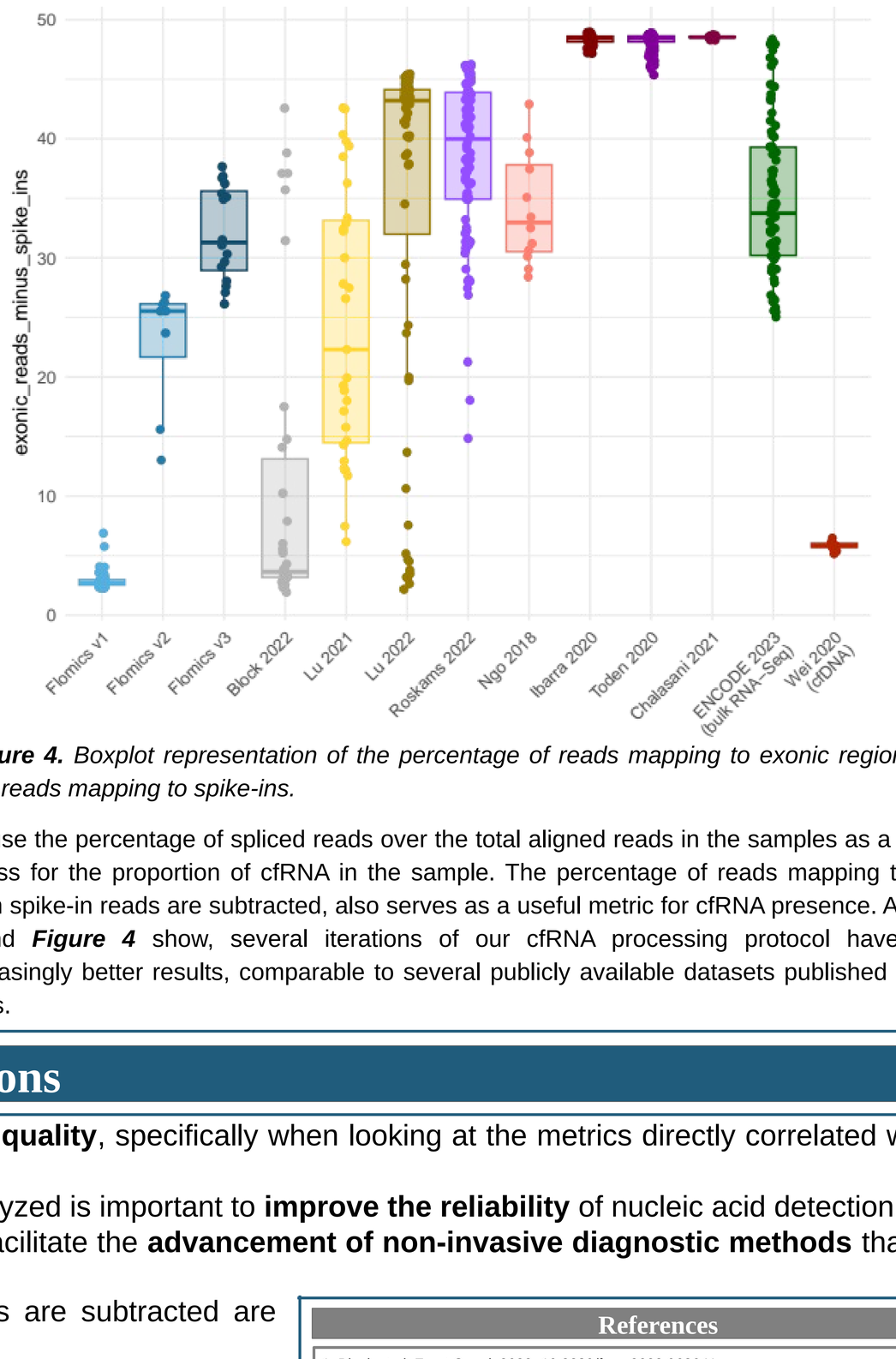

Figure 4. *Boxplot representation of the percentage of reads mapping to exonic regions minus the reads mapping to spike-ins.*

We use the percentage of spliced reads over the total aligned reads in the samples as a metric to assess for the proportion of cfRNA in the sample. The percentage of reads mapping to exons, when spike-in reads are subtracted, also serves as a useful metric for cfRNA presence. As ***Figure 3*** and ***Figure 4*** show, several iterations of our cfRNA processing protocol have yielded increasingly better results, comparable to several publicly available datasets published in recent years.

Conclusions

- **Iterative updates** to our own protocol consistently yielded **improvements in data quality**, specifically when looking at the metrics directly correlated with the higher abundance of cfRNA.
- Continuous and extensive **quality control of several metrics** for each sample analyzed is important to **improve the reliability** of nucleic acid detection.
- By establishing **benchmarks for sample quality** and bioinformatics analysis, we facilitate the **advancement of non-invasive diagnostic methods** that could improve early cancer detection and patient monitoring.
- **Percentage of spliced reads** and **percentage of exonic reads** when spike-ins are subtracted are **good quality control metrics** to assess for the proportion of cfRNA in the sample.

References

1. Block et al. Front Oncol. 2022. 10.3389/fonc.2022.963641
2. Zhu et al. Theranostics. 2021. 10.7150/thno.48206
3. Chen et al. Elife. 2022. 10.7554/eLife.75181
4. Roskams-Hieter et al. NPJ Precis Oncol. 2022. 10.1038/s41698-022-00270-y
5. Ngo et al. Science. 2018. 10.1126/science.aar3819
6. Ibarra et al. Nat Commun. 2020. 10.1038/s41467-019-14253-4
7. Toden et al. Sci Adv. 2020. 10.1126/sciadv.abb1654
8. Chalasani et al. Am J Physiol Gastrointest Liver Physiol. 2021. 10.1152/ajpgi.00397.2020
9. Zhang et al. Genome Research. 2019. 10.1101/gr.249789.119
10. Tao Wei et al. Molecular Oncology. 2020. 10.1002/1878-0261.12757
11. Ewels et al. Nat Biotechnol. 2020. 10.1038/s41587-020-0439-x.

Contact us: