Cabús *et al. Biomarker Research* (2022) 10:62
https://doi.org/10.1186/s40364-022-00409-w

Biomarker Research

REVIEW

Open Access

Current challenges and best practices for cell-free long RNA biomarker discovery

Lluc Cabús[1,2], Julien Lagarde[2], Joao Curado[2], Esther Lizano[1] and Jennifer Pérez-Boza[2*]

Abstract

The analysis of biomarkers in biological fluids, also known as liquid biopsies, is seen with great potential to diagnose complex diseases such as cancer with a high sensitivity and minimal invasiveness. Although it can target any biomolecule, most liquid biopsy studies have focused on circulating nucleic acids. Historically, studies have aimed at the detection of specific mutations on cell-free DNA (cfDNA), but recently, the study of cell-free RNA (cfRNA) has gained traction. Since 2020, a handful of cfDNA tests have been approved for therapy selection by the FDA, however, no cfRNA tests are approved to date. One of the main drawbacks in the field of RNA-based liquid biopsies is the low reproducibility of the results, often caused by technical and biological variability, a lack of standardized protocols and insufficient cohorts. In this review, we will identify the main challenges and biases introduced during the different stages of biomarker discovery in liquid biopsies with cfRNA and propose solutions to minimize them.

Keywords: Liquid biopsies, Cell-free RNA, Long RNA, RNA sequencing, Technical bias, Early diagnosis

Background

In the past few years, there has been an increased interest in finding minimally invasive methods for disease-specific biomarker detection [1]. Following this trend, liquid biopsies are becoming promising alternatives to replace more invasive diagnostic methods such as tissue biopsies or image-based methods in the future. Although liquid biopsies can theoretically be applied to any biomolecule in any biological fluid, during the last decades there has been an increase in the studies that target circulating nucleic acids in blood [2–5]. Although the discovery and validation of these biomarkers require a considerable economic effort, their implementation in the clinical practice will be cheap, since the process only requires the obtention of a blood sample, but the true economical relief of this method is yet to be determined [6]. Liquid biopsies will constitute a great advantage for early detection, since adherence to the current screening tests are one of the major problems of the healthcare system.

Even though other biofluids are of special interest for certain diseases (urine for prostate or bladder cancer [7, 8], or cerebrospinal fluid for brain diseases, such as Parkinson's disease or some variants of brain cancer [9, 10]), for the majority of diseases, studies have focused only on blood and the fractions derived from it (plasma, serum and platelets). For this reason, in this review we have focused on circulating biomarkers in blood.

To date, most studies on circulating nucleic acids have centered on diagnosis, prognosis and response to treatment in oncology using cell-free tumor DNA (ctDNA). These are molecules of DNA shed by the tumor and present in the circulation [11]. CtDNA biomarkers provide information about specific mutations and are of special interest for targeted therapy: treatment with drugs directed to those specific mutations present only in the tumor cells [12]. Highlighting their potential, several ctDNA-based screening tests have been approved for the clinical practice during the last years, with many more currently undergoing clinical trials [13]. Although promising, there is one main limitation associated with the

*Correspondence: jennifer.perez@flomics.com

[2] Flomics Biotech, Barcelona, Spain
Full list of author information is available at the end of the article

translation of ctDNA-based screening tests to the clinic: the abundance of ctDNA in blood is directly linked to tumor burden. CtDNA derives exclusively from tumor cells, and it is secreted into the extracellular milieu either during the processes of cellular death or through active export [14]. These nucleic acids are biologically informative, but present important limitations for early detection in early cancer stages or in diseases with low tumor cells [15, 16]. Unlike ctDNA, cell-free RNA (cfRNA) is released from cancerous and non-cancerous cells. It can derive from non-transformed tissues such as stroma or from the immune system responding to the presence of tumors, both of which can be highly informative for the diagnosis [17].

Changes in RNA expression in cells are a dynamic process that can reflect tissue damage or disease [18]. Moreover, the study of cfRNA is not merely based on the differential abundance of a set of specific genes, but also on additional factors such as pathogenic alternative splicing [19] or A-to-I RNA editing [20], changes that are not detectable in the genome, only in the transcriptome. Due to these limitations, there has been a rising interest in the field of cfRNA over ctDNA in the last few years.

The field of cell-free RNA biomarkers has mostly focused on the study of microRNAs (miRNAs) as biomarkers of disease in the circulation due to their higher stability in blood [21]. However, there is a rising interest in the study of long RNAs (> 200 nt), including but not limited to messenger RNAs (mRNAs) and long non-coding RNAs (lncRNAs). As an example of this, lately some studies have suggested circulating biomarkers based on long RNAs for diseases such as fetal congenital heart defects [22], Alzheimer's disease [23] and lung cancer [24]. However, none of these have reached the level of validation necessary to reach the clinic yet.

Although more technically challenging, there is one main advantage associated with the study of long RNAs: the number of known long RNAs is much higher than that of known miRNAs (37,911 known long RNA genes between mRNAs and lncRNAs [25] vs 4571 between hairpin and mature miRNA sequences) [26]. This is also the case in biofluids, where the number of mRNAs detected is between 5 and 450 times the number of miRNAs detected [27]. As a result of this higher number of RNAs, the potential to obtain biomarkers that reliably assess the state of a specific disease using these biomolecules is much higher.

Despite its promising future, the field of liquid biopsies is young and still strongly biased by technical and biological limitations. In this review, we will identify the main challenges associated with the use of cell-free long RNAs for the discovery of diagnostic biomarkers. For that purpose, we have listed the different steps involved in long RNA biomarker discovery starting from blood collection to data analysis, highlighting the main limitations associated with each step (Figs. 1 and 2).

Plasma and serum as main biofluids

Blood carries oxygen to all organs in the body and is also the vessel of biological information shed into the circulation. This complex biofluid contains immune and blood cells, blood-clotting factors, proteins, lipoproteins, extracellular vesicles (EVs), cell fragments and nucleic acids, among other types of biomolecules [28]. In blood, cfRNA is either encapsulated inside microvesicles or forming ribonucleoprotein complexes [29]. To reduce RNA contamination from blood cells, the most common approach is to use serum (the fraction of blood remaining after the blood clotting) or plasma (the acellular fraction of the blood) instead of whole blood.

Although intrinsically different, there is limited information about which biofluid provides more biologically relevant information. Promising results have derived from both the study of serum and plasma. Even though not studied in depth, some groups reached opposite conclusions while attempting to quantify the differences associated with the analysis of serum vs plasma. For instance, Dufourd et al. suggested that circulating miRNA profiles of healthy subjects may not be affected by the type of biofluid studied [30] while others have proposed that miRNA profiles of serum and plasma are not comparable and can influence subsequent analyses [31]. An advantage of using plasma is that Klaas et al. reported higher amounts of cfRNA in plasma compared to serum [32]. Although the reason for this difference is unclear, the authors suggested that cfRNA could adhere to the blood clot during coagulation, resulting in a reduction in the cfRNA quantity.

Whether using plasma or serum, a common limitation of the study of these biofluids for biomarker discovery is the release of RNAs derived from red blood cells (RBC) and/or platelets during sample processing [33]. Regarding this, during blood clotting, a necessary step for serum isolation, RNAs are released from blood cells and platelets, affecting the circulating RNA spectrum [34, 35]. However, to separate plasma from other blood components, blood cells and platelets have to be removed with several centrifugation steps. At this point, an incorrect centrifugation or handling of the samples could result in cellular contamination [36].

In order to assess the compositional bias induced by platelet-derived EVs on the plasma transcriptome, Kim et al. [37] examined the variation of exosomes and cfRNA in human plasma due to blood processing and freeze-thaw effect. They discovered a significant reduction in the 1000-3000 nm EVs in platelet-free plasma samples,

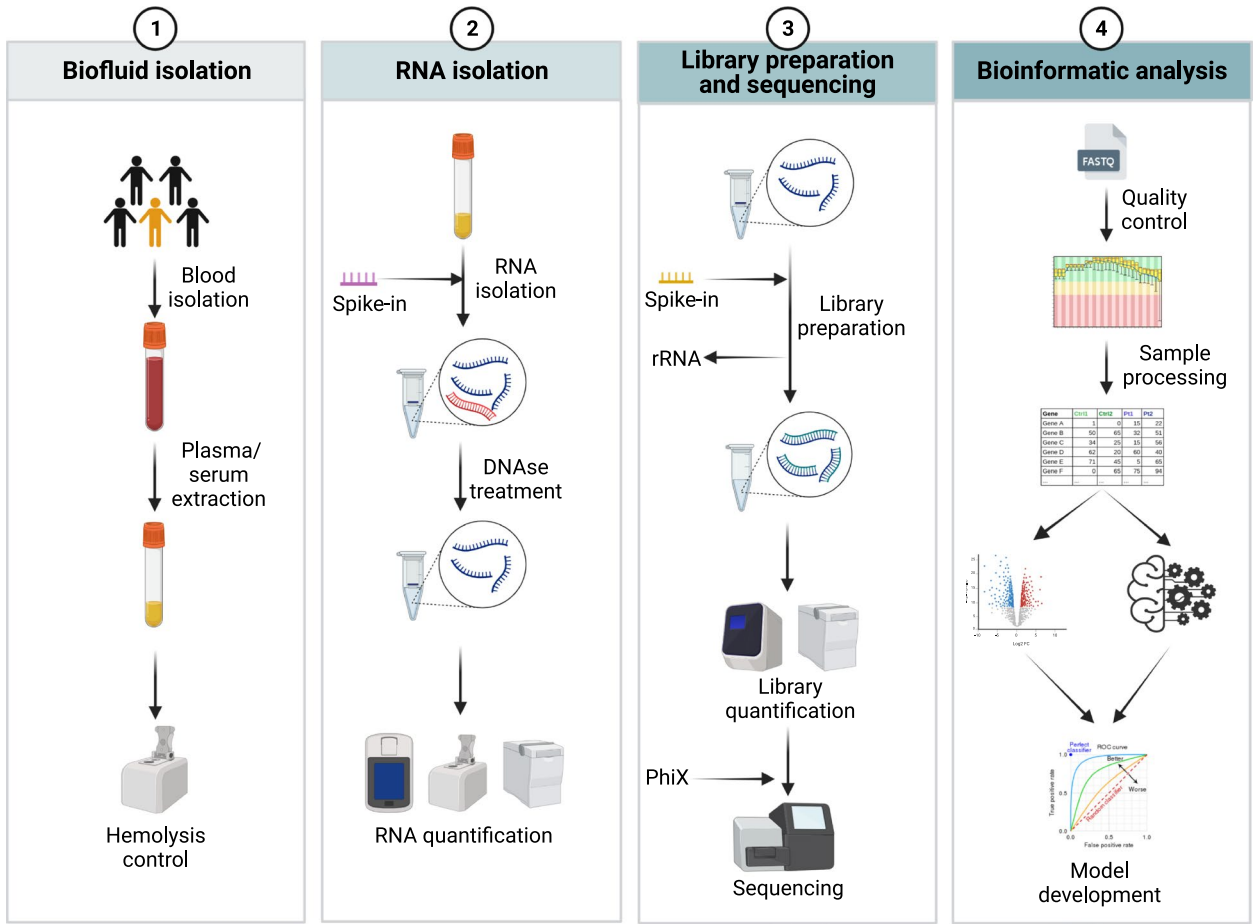

Fig. 1 Schematic timeline of all the steps involved in the development of cfRNA biomarkers. (1) Biofluid isolation: after obtaining blood from the patients and centrifugation to get plasma or serum, it is necessary to perform a hemolysis control to measure the contamination by cellular lysis. (2) RNA isolation: prior to the processing of plasma/serum, it is recommended to add external RNA molecules to act as proxies for correct RNA isolation (spike-ins). After isolation of the nucleic acids and before further processing, a step of DNAse digestion is required to limit the contamination of the sample with co-purified DNA. A final step of RNA quantification is required before moving forward to (3) library preparation. To improve reproducibility, a different set of spike-ins are added before starting the process of preparing the RNA for sequencing, and rRNA, lacking biological information, is depleted from the samples. Once the libraries are prepared and quantified, the next step is to sequence them. During this step, it is also possible to add an exogenous library (PhiX), to measure technical variability and ensure reproducibility. (4) Bioinformatic analysis: after the initial quality control of the sample, the data is processed and the expression of several genes linked to a specific phenotype through differential expression analysis or machine learning, to develop a robust biomarker model

showing an ex vivo platelet EV release. Moreover, they show that while post-thaw processing reduces the amount of platelet-derived EVs, it irreversibly affects the cfRNA profile. These results suggest that banked plasma samples with different degrees of platelet removal could be incomparable.

While hemolysis controls are more common than platelet assessment, there is still a lack of consensus about the optimal steps necessary to quantify the presence of RBC RNA. In an attempt to characterize the effect of hemolysis during the study of circulating RNA biomarkers, Kirschner et al. [38] suggested the addition of a pre-analytical step to quantify RBC lysis using a spectrophotometer. According to their results, blood samples with low absorbance at 414 nm, wavelength characteristic of oxyhemoglobin [39], had similar levels of miR16, a miRNA often present in RBC and found in abundance in hemolytic samples.

Despite the absence of studies performed on cfRNA, a study found that for cellular RNA freeze-thaw cycles have a detrimental effect on the quality of the RNA obtained downstream, resulting in significantly shorter fragments [40]. Additionally, other studies have also suggested that storing plasma samples at − 80 °C leads to degradation of

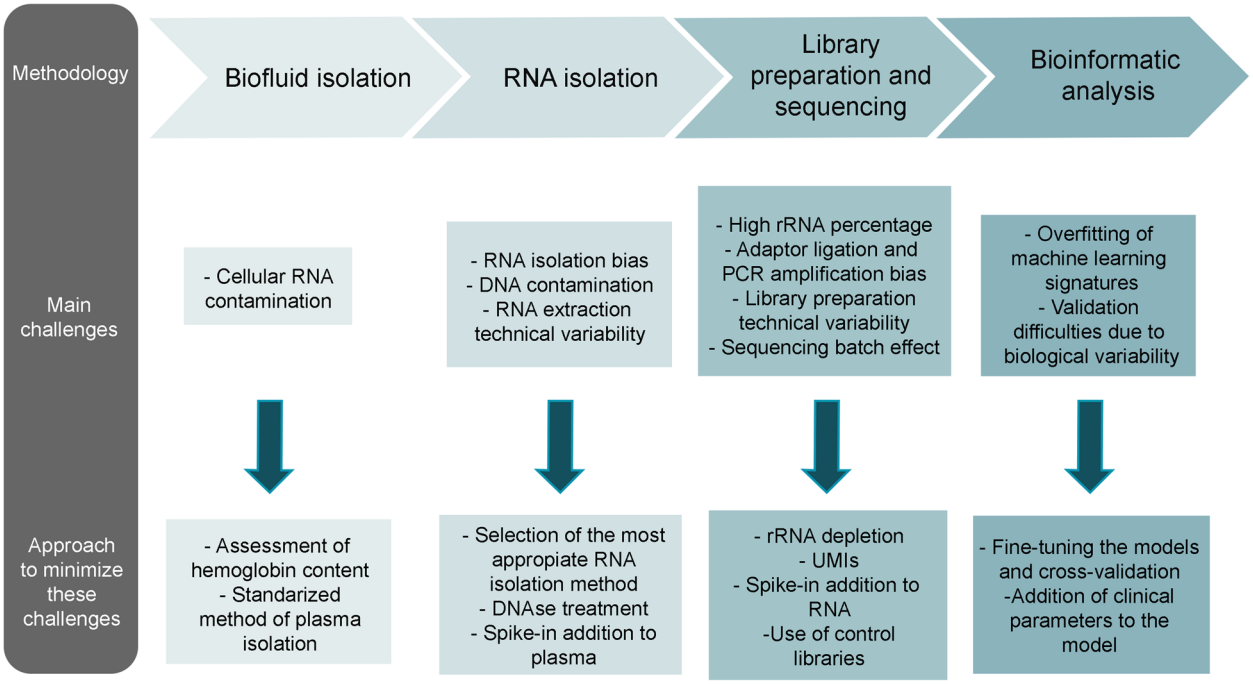

Fig. 2 Overview of the main challenges in the process of biomarker discovery and specific steps to minimize them

the nucleic acids over time [41]. This is of special interest for the study of long RNAs, which are remarkably more prone to degradation than miRNAs in plasma [21].

In summary, the obtention of plasma or serum from whole blood can result in contamination from different cellular sources, such as platelets or erythrocytes. Avoiding and assessing cellular contamination, with correct processing and quality controls, is a critical step in order to prevent and monitor the introduction of unwanted biases that could lead to wrong conclusions. In addition, avoiding freeze-thaw cycles and long-term storage improves the quality of the RNA extracted, leading to more robust and reproducible results.

RNA isolation

The cfRNA obtained from a blood sample represents an approximation to a snapshot of the transcriptome of the individual. However, the use of different methodologies for RNA isolation leads to biases that can mask any relevant biological information.

The most common strategy for RNA isolation from biofluids is based on RNA extraction kits designed and optimized for plasma and/or serum. Commercial column-based kits are more commonly used than traditional guanidium-thiocyanate or phenol-chloroform methods. Classic methods tend to favor the isolation of selective RNA populations and often lead to reduced quantities of RNA [42]. However, the widespread use of different kits has led to intrinsic technical differences associated with kit-dependent biases. Li et al. showed that different cfRNA isolation kits yield different RNA quantities. They also found kit-dependent biases linked with the recovery of long RNAs and issues with detecting some of the most common mRNAs. Their results highlight the importance of selecting the best approach to isolate RNA depending on the end goal of the study [43].

One major concern regarding RNA isolation from plasma is DNA contamination. The majority of cfRNA isolation kits recover a fraction of the cfDNA present in the biofluid [43]. During library preparation, DNA contamination is amplified along with RNA affecting the results [44]. To minimize this bias, the most common strategy is the incorporation of an additional step where the samples are treated with DNAse. This can be done before the extraction (on-column) [45] or after [27].

In order to limit the effect of other technical biases associated with RNA isolation, some groups have proposed the addition of exogenous RNAs as controls to compare the efficiency of the RNA extraction [46, 47]. Spike-ins are exogenous RNAs with similar GC content to endogenous RNAs with sequences not found in the human genome. They are added to the samples prior to RNA isolation in a known concentration [48] and their detection is useful to assess biases introduced during RNA isolation. Spike-ins have been successfully used to

compare RNA profiles across biofluids, allowing absolute quantification and revealing a 10,000 fold difference in concentration [27].

The quantity and quality of input RNA has a strong impact on downstream processes [49]. However, there is a lack of consensus on which is the best method to quantify blood-derived RNA. To date, the most commonly used methods to measure the quality and quantity of RNA are based on spectrophotometry such as Qubit [50] and Nanodrop [51]; or on capillarity, like the Bioanalyzer [52]. These methods were developed to assess the concentration of RNA in cellular samples and they are less accurate when evaluating the highly fragmented samples derived from blood. Additionally, other groups have proposed the incorporation of PCR-based methods in biomarker discovery pipelines. They use the abundance of specific genes in blood as internal controls to set a minimum concentration of RNA [53]. This approach accepts the lack of reproducible quantification methods when starting from very low input RNA and attempts only to confirm the abundance of certain genes above a threshold.

Different strategies for RNA isolation are linked to different recovery rates and to the enrichment of selective RNA types. However, several strategies are available to minimize these biases: consistency in the extraction kit used, comparison of the RNA extraction efficiency with spike-ins, assessment of samples with low RNA quality and removal of DNA contamination.

Library preparation and sequencing

During the last decade, RNA sequencing (RNA-seq) has become the gold standard methodology for biomarker discovery. It allows the detection of known and novel transcripts and their quantification in a sample, with higher sensitivity and accuracy than other methods, such as microarrays [54]. While years ago the costs associated with this methodology were high, the drop in prices for sequencing has opened the door for the application of NGS to biomarker discovery and as a screening tool. To adapt to this new field, multiple protocols for library preparation have been developed to perform RNA-seq starting from very low RNA inputs.

Approximately 80% of the cellular RNA is ribosomal RNA (rRNA) [55], and this number is even bigger in cfRNA, with more than 90% of the RNA found in the circulation belonging to this type of RNA [45]. In order to remove high concentrations of rRNA, there are two methods: polyA enrichment or rRNA depletion. However, only the rRNA depletion step is viable in this case, due to the high fragmentation of the RNA in plasma and lack of polyA tail in most of the cfRNA molecules [56].

During library preparation, another step leading to the introduction of biases is adaptor ligation and PCR amplification. Due to secondary structures and enzyme affinity, some sequences are more prone to ligate to adaptor sequences and amplify than others [57, 58]. In order to reduce this bias, many protocols have incorporated the use of Unique Molecular Identifiers (UMIs) in the early stages of library preparation. UMIs are random sequences of 4-10 nt that are ligated to the DNA/RNA molecules before PCR amplification [59]. The use of UMIs allows the identification of PCR-duplicated reads that originate from the same initial molecule, and the in silico correction of this clonal amplification. Additionally, if the UMIs are incorporated before adaptor ligation, they can also help minimize adaptor ligation biases. When compared to traditional in silico removal of technical duplicates [60], the presence of UMIs has shown to improve reproducibility in differential gene expression analysis, especially when starting from a very low input [58, 61], as in the case of cfRNA samples.

Following the same rationale used to evaluate biases during RNA isolation, spike-ins are also useful in measuring technical variability introduced during library preparation. Similar to the spike-ins used to assess RNA isolation, they are added at a known concentration to correct for the amplification bias during library preparation [48] and can be useful to quantify and normalize in silico the sequencing output.

Finally, another possible step of bias introduction is the sequencing of the libraries. Often, batch effects are observed when samples used in the same study are sequenced in different rounds [62]. Commercially available control libraries can be added at known concentrations to measure the reproducibility of sequencing and later used to mitigate batch effects. The most common example is the PhiX Control Libraries, generated from the PhiX virus [63]. These spike-in libraries have two functions. First, they act as positive controls of the sequencing run, ensuring clustering reaction and generating a number of clusters depending on the quantity of the spike-in added. And second, as a technical control for sequencing accuracy, aligning the sequences to the reference genome of the spike-in library.

In summary, the main problems associated with library preparation and sequencing are PCR amplification and batch effects caused by technical variability. However, the use of UMI sequences and spike-ins to minimize this bias has become very prevalent in RNA-seq studies during the last years and has shown to increase the reproducibility of the results [64].

Bioinformatics analysis

One of the most challenging steps in biomarker discovery using RNA-seq is the computational analysis of the generated data. This analysis allows to correctly interpret the molecular processes that occur in the patient. Bioinformatics analysis goes from data quality control to transcript quantification and various downstream analyses. However, since the results of the analysis are closely linked to the input quality of the data, quality assessment is a critical step in this workflow [65].

One of the most prevalent issues with the processing of raw RNA-seq data is normalization. It consists in correcting in silico for technical biases that could mask biological information [66]. RNA-seq is the most used tool to measure gene expression, although unlike other methods like RT-qPCR, it does not allow for absolute quantification. A normalization based on spike-in controls, added at the library preparation step, is one of the methods used to achieve absolute quantification of the transcriptome in cellular RNA [67]. This method is also shown to be highly reproducible in plasma samples [64]. However, some studies have found that spike-in normalization can translate to poor results due to the high variability in the spike-in amplification [68, 69]. Due to this, some groups are opting for a normalization of the samples according only to library size and gene length [45, 70]. UMI deduplication corrects for PCR biases and allows for the counting of RNA molecules in a sample, thereby improving reproducibility [61].

Once the raw data is processed, the identification of potential biomarkers can follow two different routes: comparative analysis of cfRNA profiles and machine learning (ML) methods. The former strategy is the simplest: it consists in finding genes whose expression is associated with a phenotype. This approach attempts to determine the presence of a disease or its prognosis only from the expression of a few genes. An example of this methodology is the recent study of Rasmussen et al., where they found a signature of 7 RNAs associated with preeclampsia, a condition marked by maternal hypertension that is a significant cause of maternal morbidity. This signature has a positive predictive value of 32.3%, higher than the current clinical state-of-the-art models [2]. The latter, ML-based approach, is more complex as it involves the application of ML algorithms to detect biomarker signatures predicting the likelihood of a phenotype. A ML signature combines multiple genes selected by ML algorithms and determines the presence of a disease or its prognosis. An ML-based signature normally has higher sensitivity and specificity than a signature coming from comparative analysis of cfRNA profiles [71], presumably because ML algorithms automatically assign a weight to every gene in order to maximize the accuracy of the classification. An example of this methodology is the study of Wang et al., where they found a panel of 57 RNA biomarkers that could detect COVID-19 infection with 98.1% accuracy [72].

In relatively small datasets, ML algorithms tend to generate models that fit artificially the initial cohort of samples, causing poor replication of the results in new cohorts of patients. This is especially prevalent in omics data, where the number of variables is very high [73]. This process is known as overfitting. To mitigate this, some of the most used methods are cross-validation and model simplification [74]. Cross-validation is a resampling method that uses a fraction of the data to evaluate the performance of the algorithm, whereas model simplification consists in reducing the number of genes included in the signature. This strategy of cross-validation and model simplification has already been used by multiple studies for diagnosis or prognosis of diseases.

Additionally, other research groups have proposed new approaches to enhance the reproducibility of the findings. For instance, Larson et al. [45] have focused on the study of "Dark channel" biomarkers, which are genes not expressed in non-cancer plasma, upregulated in cancer samples and detected in multiple samples to improve specificity and reduce drastically the number of false positives. Vorperian et al. [75] also proposed an interesting alternative method. They have identified the transcriptomic fingerprint of a certain number of cell types and used these profiles to deconvolute the cell types of origin of cfRNA. Using this approach it could be possible to narrow down the set of genes studied to focus only on those derived from the organ of interest, thus improving reproducibility and reducing variability.

On the one hand, the comparative analysis of cfRNA profiles is easier to implement in the clinical practice than more complex ML signatures due to its lower cost and improved practicality. On the other hand, ML signatures tend to have higher accuracy. Both methods are useful in obtaining signatures of high value for diagnosis and prognosis.

Limitations of cfRNA biomarker discovery

The main issues limiting the applicability of liquid biopsies as screening and diagnostic tools in the clinic are technical and biological reproducibility biases. This is often linked to a lack of gold standard methods for sample processing and data analysis [76]. Due to the high technical variability observed between research groups, different studies often find unrelated RNA signatures for the same disease and type of sample. This lack of reproducibility is one of the major problems that the field is facing, with many RNA signatures entering clinical trials but none reaching the clinic so far. To try and mitigate

this lack of standardization, the Global Biological Standards Institute (GBSI) published an article in 2014 to raise awareness about the lack of reproducibility and the urgent need for standards in cancer research, especially for high-throughput screening methods [77]. Since then, numerous efforts have focused on methodological standardization, such as the implementation of data repositories or the creation of reference RNA spike-in controls [76].

Besides the technical bias introduced during sample handling, external factors such as age or gender of the individual, have a strong effect on the cfRNA profile [78]. To control for these biases, every step of the study must be well controlled and documented, with balanced cohorts in all of these factors [79]. In cases where it is not possible to have balanced cohorts, these external factors should be accounted for in the statistical analysis.

A biological limitation of the study of cfRNA is that the interpersonal variability is very high [80], with some individuals showing a consistently higher expression of certain genes than others. Although their results are preliminary, they highlight the importance of a consistent normalization method to account for this biological variability.

Current challenges in the field of RNA-based liquid biopsies

According to a recent market research study, the liquid biopsy industry is expected to exceed 5.8 billion dollars by 2026 [81], although few cfDNA tests are already in use in the clinical practice [82]. In 2020, the FDA already approved three cfDNA-based tests and, of these, Guardant Health's Guardant360 CDx is the first one to use next generation sequencing for diagnosis [83]. Since the field of cfRNA is still young, there are no diagnostic tests based on RNA approved for the use in the clinical practice yet. There are many challenges to be addressed to translate an RNA-based liquid biopsy biomarker into the clinic, although several studies are currently undergoing clinical trials [84–88]. A robust and standardized methodology needs to be established, assessing all the possible biases that can alter the results. This will lead to more reproducible results and more robust statistical models.

One of the main challenges in the field of liquid biopsies is caused by a limitation on the number of donors in the training cohorts. Most of the studies comparing cases with controls use small retrospective cohorts to detect a disease after it is clinically reported [52, 89–91], which is suboptimal for biomarkers for early diagnosis. Only a handful of prospective studies have attempted population screening to find undiagnosed patients [92]. This approach, although optimal for diagnostic biomarker discovery and validation, requires the screening of a significant part of the population (depending on disease prevalence) and a great budgetary effort. The technical and economical requirements for such an attempt are beyond the grasp of most research centers and the biomarkers discovered in the first cohorts are not strong enough to pass an initial step of validation and attract big-pharma companies. Although multiple collaborative consortia are created to compile biological data for specific pathologies (such as the NCI Cohort Consortium for cancer), the lack of standardized methods often leads to results that highlight technical differences over biologically relevant biomarkers.

Another important aspect for the incorporation of liquid biopsies into the clinical practice is the ability of the physicians to be able to interpret the results of the biomarker model. This can be an arduous task for the medical staff that is not proficient in bioinformatics and statistical analysis. For this purpose, several research centers and companies have created cloud computing pipelines that take directly the sequencing data and generate comprehensive reports [93, 94].

Conclusions

The field of liquid biopsies, and more specifically cell-free long RNA liquid biopsies is promising, but still young. With a relatively reduced number of studies published, there are candidate biomarkers undergoing clinical trials, but none have been approved by the regulatory agencies at the moment. In the last few years, there has been an increasing interest in liquid biopsy-based biomarkers using RNAs. However, it has been only in the last 5 years that the focus has started to switch from miRNAs to long RNAs, leading to the discovery of new disease-associated RNAs. Although there is still much work left to do to translate long cfRNA into clinical practice, a number of recent promising results suggest that long cfRNA-based liquid biopsies could be one of the next big revolutions in the field of screening and diagnosis.

Abbreviations

ctDNA: Cell-free tumor DNA; cfRNA: Cell-free RNA; miRNAs: MicroRNAs; mRNAs: Messenger RNAs; lncRNAs: Long non-coding RNAs; EVs: Extracellular vesicles; RNA-seq: RNA sequencing; UMIs: Unique Molecular Identifiers; ML: Machine learning; GBSI: Global Biological Standards Institute.

Acknowledgements

Not applicable.

Authors' contributions

LC was the main contributor to the manuscript. JPB designed this manuscript. LC and JPB contributed to the final version of the manuscript. JPB, JL, JC and EL revised the manuscript. The authors read and approved the final manuscript.

Funding
This project has received funding from the Doctorats industrials grant (2019 DI 091) and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska - Curie grant agreement No 801342 (Tecniospring INDUSTRY)" and the Government of Catalonia's Agency for Business Competitiveness (ACCIÓ) (ACE003/20/000028).

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
LC, JL and JPB are employees, JC and EL are co-founders of Flomics, a biotechnology company providing human plasma cell-free RNA sequencing research and services.

Author details
[1]Institut de Biologia Evolutiva, Universitat Pompeu Fabra, Barcelona, Spain. [2]Flomics Biotech, Barcelona, Spain.

Received: 25 May 2022 Accepted: 4 August 2022
Published online: 18 August 2022

References

1. Ignatiadis M, Sledge GW, Jeffrey SS. Liquid biopsy enters the clinic — implementation issues and future challenges. Nat Rev Clin Oncol. 2021;18(5):297–312.
2. Rasmussen M, Reddy M, Nolan R, Camunas-Soler J, Khodursky A, Scheller NM, et al. RNA profiles reveal signatures of future health and disease in pregnancy. Nature. 2022;601(7893):422–7.
3. Moufarrej MN, Vorperian SK, Wong RJ, Campos AA, Quaintance CC, Sit RV, et al. Early prediction of preeclampsia in pregnancy with cell-free RNA. Nature. 2022;602(7898):689–94.
4. Mathios D, Johansen JS, Cristiano S, Medina JE, Phallen J, Larsen KR, et al. Detection and characterization of lung cancer using cell-free DNA fragmentomes. Nat Commun. 2021;12(1):5060.
5. Vandekerkhove G, Lavoie JM, Annala M, Murtha AJ, Sundahl N, Walz S, et al. Plasma ctDNA is a tumor tissue surrogate and enables clinical-genomic stratification of metastatic bladder cancer. Nat Commun. 2021;12(1):184.
6. Ladabaum U, Alvarez-Osorio L, Rösch T, Brueggenjuergen B. Cost-effectiveness of colorectal cancer screening in Germany: current endoscopic and fecal testing strategies versus plasma methylated Septin 9 DNA. Endosc Int Open. 2014;2(2):E96–104.
7. Lee B, Mahmud I, Marchica J, Dereziński P, Qi F, Wang F, et al. Integrated RNA and metabolite profiling of urine liquid biopsies for prostate cancer biomarker discovery. Sci Rep. 2020;10(1):3716.
8. Dudley JC, Schroers-Martin J, Lazzareschi DV, Shi WY, Chen SB, Esfahani MS, et al. Detection and surveillance of bladder cancer using urine tumor DNA. Cancer Discov. 2019;9(4):500–9.
9. Hossein-nezhad A, Fatemi RP, Ahmad R, Peskind ER, Zabetian CP, Hu SC, et al. Transcriptomic Profiling of Extracellular RNAs Present in Cerebrospinal Fluid Identifies Differentially Expressed Transcripts in Parkinson's Disease. J Parkinsons Dis. 2016;6(1):109–17.
10. Kopkova A, Sana J, Machackova T, Vecera M, Radova L, Trachtova K, et al. Cerebrospinal fluid MicroRNA signatures as diagnostic biomarkers in brain tumors. Cancers. 2019;11(10):1546.
11. Pessoa LS, Heringer M, Ferrer VP. ctDNA as a cancer biomarker: a broad overview. Crit Rev Oncol Hematol. 2020;155:103109.
12. Chattopadhyay I. Application of Circulating Cell-free DNA for Personalized Cancer Therapy. In: Precision Medicine in Oncology. Wiley; 2020. p. 83–97. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119432487.ch3 [cited 7 Mar 2022].
13. Cowling T, Loshak H. An Overview of Liquid Biopsy for Screening and Early Detection of Cancer. In: CADTH Issues in Emerging Health Technologies. Ottawa: Canadian Agency for Drugs and Technologies in Health; 2016. Available from: http://www.ncbi.nlm.nih.gov/books/NBK555478/ [cited 7 Mar 2022].
14. Kustanovich A, Schwartz R, Peretz T, Grinshpun A. Life and death of circulating cell-free DNA. Cancer Biol Ther. 2019;20(8):1057–67.
15. Zaporozhchenko IA, Ponomaryova AA, Rykova EY, Laktionov PP. The potential of circulating cell-free RNA as a cancer biomarker: challenges and opportunities. Expert Rev Mol Diagn. 2018;18(2):133–45.
16. Yi Z, Ma F, Rong G, Liu B, Guan Y, Li J, et al. The molecular tumor burden index as a response evaluation criterion in breast cancer. Signal Transduct Target Ther. 2021;6(1):1–8.
17. Nabet BY, Qiu Y, Shabason JE, Wu TJ, Yoon T, Kim BC, et al. Exosome RNA Unshielding couples stromal activation to pattern recognition receptor signaling in Cancer. Cell. 2017;170(2):352–66 e13.
18. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat Rev Genet. 2016;17(5):257–71.
19. El Marabti E, Younis I. The Cancer Spliceome: reprograming of alternative splicing in Cancer. Front Mol Biosci. 2018;5:80.
20. Ben-Aroya S, Levanon EY. A-to-I RNA editing: an overlooked source of Cancer mutations. Cancer Cell. 2018;33(5):789–90.
21. Glinge C, Clauss S, Boddum K, Jabbari R, Jabbari J, Risgaard B, et al. Stability of circulating Blood-based MicroRNAs – pre-analytic methodological considerations. PLoS One. 2017;12(2):e0167969.
22. Gu M, Zheng A, Tu W, Zhao J, Li L, Li M, et al. Circulating LncRNAs as novel, non-invasive biomarkers for prenatal detection of fetal congenital heart defects. Cell Physiol Biochem. 2016;38(4):1459–71.
23. Feng L, Liao YT, He JC, Xie CL, Chen SY, Fan HH, et al. Plasma long non-coding RNA BACE1 as a novel biomarker for diagnosis of Alzheimer disease. BMC Neurol. 2018;18:4.
24. Lin Y, Leng Q, Zhan M, Jiang F. A plasma long noncoding RNA signature for early detection of lung Cancer. Transl Oncol. 2018;11(5):1225–31.
25. Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, et al. GENCODE 2021. Nucleic Acids Res. 2021;49(D1):D916–23.
26. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2019;47(D1):D155–62.
27. Hulstaert E, Morlion A, Cobos FA, Verniers K, Nuytens J, Eynde EV, et al. Charting extracellular Transcriptomes in the human biofluid RNA atlas. Cell Rep. 2020;33(13) Available from: https://www.cell.com/cell-reports/abstract/S2211-1247(20)31541-2 [cited 7 Mar 2022].
28. Perakis S, Speicher MR. Emerging concepts in liquid biopsies. BMC Med. 2017;15:75.
29. Pös O, Biró O, Szemes T, Nagy B. Circulating cell-free nucleic acids: characteristics and applications. Eur J Hum Genet. 2018;26(7):937–45.
30. Dufourd T, Robil N, Mallet D, Carcenac C, Boulet S, Brishoual S, et al. Plasma or serum? A qualitative study on rodents and humans using high-throughput microRNA sequencing for circulating biomarkers. Biol Methods Protoc. 2019;4(1):bpz006.
31. Mompeón A, Ortega-Paz L, Vidal-Gómez X, Costa TJ, Pérez-Cremades D, Garcia-Blas S, et al. Disparate miRNA expression in serum and plasma of patients with acute myocardial infarction: a systematic and paired comparative analysis. Sci Rep. 2020;10(1):5373.
32. Max KEA, Bertram K, Akat KM, Bogardus KA, Li J, Morozov P, et al. Human plasma and serum extracellular small RNA reference profiles and their clinical utility. Proc Natl Acad Sci. 2018;115(23):E5334–43.
33. McDonald JS, Milosevic D, Reddi HV, Grebe SK, Algeciras-Schimnich A. Analysis of circulating MicroRNA: Preanalytical and analytical challenges. Clin Chem. 2011;57(6):833–40.
34. Wang K, Yuan Y, Cho JH, McClarty S, Baxter D, Galas DJ. Comparing the MicroRNA Spectrum between serum and plasma. PLoS One. 2012;7(7):e41561.
35. Witwer KW, Buzás EI, Bemis LT, Bora A, Lässer C, Lötvall J, et al. Standardization of sample collection, isolation and analysis methods in extracellular vesicle research. J Extracell Vesicles. 2013;2. https://doi.org/10.3402/jev.v2i0.20360.
36. Rieske RR, Kutcher ME, Audia JP, Carter KT, Lee YL, Tan YB, et al. Analysis of plasma products for cellular contaminants: comparing standard preparation methods. J Am Coll Surg. 2020;230(4):596–602.

37. Kim HJ, Rames MJ, Tassi Yunga S, Armstrong R, Morita M, Ngo ATP, et al. Irreversible alteration of extracellular vesicle and cell-free messenger RNA profiles in human plasma associated with blood processing and storage. Sci Rep. 2022;12:2099.
38. Kirschner MB, Kao SC, Edelman JJ, Armstrong NJ, Vallely MP, van Zandwijk N, et al. Haemolysis during sample preparation alters microRNA content of plasma. PLoS One. 2011;6(9):e24145.
39. Stroes JW, Van Rijn HJM. Quantitative measurement of Blood pigments in cerebrospinal fluid by derivative spectrophotometry. Ann Clin Biochem. 1987;24(2):189–97.
40. Yu K, Xing J, Zhang J, Zhao R, Zhang Y, Zhao L. Effect of multiple cycles of freeze–thawing on the RNA quality of lung cancer tissues. Cell Tissue Bank. 2017;18(3):433–40.
41. Sozzi G, Roz L, Conte D, Mariani L, Andriani F, Verderio P, et al. Effects of prolonged storage of whole plasma or isolated plasma DNA on the results of circulating DNA quantification assays. JNCI J Natl Cancer Inst. 2005;97(24):1848–50.
42. Wright K, de Silva K, Purdie AC, Plain KM. Comparison of methods for miRNA isolation and quantification from ovine plasma. Sci Rep. 2020;10(1):825.
43. Li X, Mauro M, Williams Z. Comparison of plasma extracellular RNA isolation kits reveals kit-dependent biases. BioTechniques. 2015;59(1):13–7.
44. Verwilt J, Trypsteen W, Van Paemel R, De Preter K, Giraldez MD, Mestdagh P, et al. When DNA gets in the way: a cautionary note for DNA contamination in extracellular RNA-seq studies. Proc Natl Acad Sci. 2020;117(32):18934–6.
45. Larson MH, Pan W, Kim HJ, Mauntz RE, Stuart SM, Pimentel M, et al. A comprehensive characterization of the cell-free transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection. Nat Commun. 2021;12(1):2357.
46. Zhang X, Nie X, Yuan S, Li H, Fan J, Li C, et al. Circulating long non-coding RNA ENST00000507296 is a prognostic Indicator in patients with dilated cardiomyopathy. Mol Ther Nucleic Acids. 2019;16:82–90.
47. Buschmann D, Haberberger A, Kirchner B, Spornraft M, Riedmaier I, Schelling G, et al. Toward reliable biomarker signatures in the age of liquid biopsies - how to standardize the small RNA-Seq workflow. Nucleic Acids Res. 2016;44(13):5995–6018.
48. Chen K, Hu Z, Xia Z, Zhao D, Li W, Tyler JK. The Overlooked Fact: Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses. Mol Cell Biol. 2015;36(5):662–7.
49. Schuierer S, Carbone W, Knehr J, Petitjean V, Fernandez A, Sultan M, et al. A comprehensive assessment of RNA-seq protocols for degraded and low-quantity samples. BMC Genomics. 2017;18(1):442.
50. Ward Gahlawat A, Lenhardt J, Witte T, Keitel D, Kaufhold A, Maass KK, et al. Evaluation of storage tubes for combined analysis of circulating nucleic acids in liquid biopsies. Int J Mol Sci. 2019;20(3):704.
51. Moret I, Sánchez-Izquierdo D, Iborra M, Tortosa L, Navarro-Puche A, Nos P, et al. Assessing an improved protocol for plasma microRNA extraction. PLoS One. 2013;8(12):e82753.
52. Yuan T, Huang X, Woodcock M, Du M, Dittmar R, Wang Y, et al. Plasma extracellular RNA profiles in healthy and cancer patients. Sci Rep. 2016;6:19413.
53. Androvic P, Romanyuk N, Urdzikova-Machova L, Rohlova E, Kubista M, Valihrach L. Two-tailed RT-qPCR panel for quality control of circulating microRNA studies. Sci Rep. 2019;9(1):4255.
54. Rao MS, Van Vleet TR, Ciurlionis R, Buck WR, Mittelstadt SW, Blomme EAG, et al. Comparison of RNA-Seq and microarray gene expression platforms for the Toxicogenomic evaluation of liver from short-term rat toxicity studies. Front Genet. 2019;9:636 Available from: https://www.frontiersin.org/article/10.3389/fgene.2018.00636 [cited 7 Mar 2022].
55. Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J. Processing of rRNA and tRNA. Freeman WH. Molecular Cell Biology 5th ed. W.H. Freeman and Company; 2004. p. 525.
56. Herbert ZT, Kershner JP, Butty VL, Thimmapuram J, Choudhari S, Alekseyev YO, et al. Cross-site comparison of ribosomal depletion kits for Illumina RNAseq library construction. BMC Genomics. 2018;19(1):199.
57. Fuchs RT, Sun Z, Zhuang F, Robb GB. Bias in ligation-based small RNA sequencing library construction is determined by adaptor and RNA structure. PLoS One. 2015;10(5):e0126049.
58. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. The impact of amplification on differential expression analyses by RNA-seq. Sci Rep. 2016;6(1):25533.
59. Sena JA, Galotto G, Devitt NP, Connick MC, Jacobi JL, Umale PE, et al. Unique molecular identifiers reveal a novel sequencing artefact with implications for RNA-Seq based gene expression analysis. Sci Rep. 2018;8(1):13121.
60. Klepikova AV, Kasianov AS, Chesnokov MS, Lazarevich NL, Penin AA, Logacheva M. Effect of method of deduplication on estimation of differential gene expression using RNA-seq. PeerJ. 2017;5:e3091.
61. Fu Y, Wu PH, Beane T, Zamore PD, Weng Z. Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. BMC Genomics. 2018;19(1):531.
62. Jaffe AE, Hyde T, Kleinman J, Weinbergern DR, Chenoweth JG, McKay RD, et al. Practical impacts of genomic data "cleaning" on biological discovery using surrogate variable analysis. BMC Bioinformatics. 2015;16:372.
63. How much PhiX spike-in is recommended when sequencing low diversity libraries on Illumina platforms? Available from: https://support.illumina.com/bulletins/2017/02/how-much-phix-spike-in-is-recommended-when-sequencing-low-divers.html. [cited 7 Mar 2022].
64. Everaert C, Helsmoortel H, Decock A, Hulstaert E, Van Paemel R, Verniers K, et al. Performance assessment of total RNA sequencing of human biofluids and extracellular vesicles. Sci Rep. 2019;9(1):17574.
65. Li X, Nair A, Wang S, Wang L. Quality control of RNA-Seq experiments. In: Picardi E, editor. RNA bioinformatics. New York: Springer; 2015. p. 137–46. https://doi.org/10.1007/978-1-4939-2291-8_8. [cited 7 Mar 2022]. (Methods in molecular biology).
66. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Brief Bioinform. 2017;19(5):776–92.
67. Bayega A, Oikonomopoulos S, Gregoriou ME, Tsoumani KT, Giakountis A, Wang YC, et al. Nanopore long-read RNA-seq and absolute quantification delineate transcription dynamics in early embryo development of an insect pest. Sci Rep. 2021;11(1):7878.
68. Qing T, Yu Y, Du T, Shi L. mRNA enrichment protocols determine the quantification characteristics of external RNA spike-in controls in RNA-Seq studies. Sci China Life Sci. 2013;56(2):134–42.
69. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. Nat Biotechnol. 2014;32(9):903–14.
70. Ibarra A, Zhuang J, Zhao Y, Salathia NS, Huang V, Acosta AD, et al. Non-invasive characterization of human bone marrow stimulation and reconstitution by cell-free messenger RNA sequencing. Nat Commun. 2020;11:400.
71. Zhang X, Jonassen I, Goksøyr A. Machine Learning Approaches for Biomarker Discovery Using Gene Expression Data. In: Helder IN, editor. Bioinformatics. Brisbane: Exon Publications; 2021. Available from: http://www.ncbi.nlm.nih.gov/books/NBK569564/ [cited 7 Mar 2022].
72. Wang Y, Li J, Zhang L, Sun HX, Zhang Z, Xu J, et al. Plasma cell-free RNA characteristics in COVID-19 patients. Genome Res. 2022;32(2):228–41.
73. Han H, Jiang X. Overcome support vector machine diagnosis Overfitting. Cancer Informat. 2014;13(Suppl 1):145–58.
74. Hawkins DM. The problem of Overfitting. J Chem Inf Comput Sci. 2004;44(1):1–12.
75. Vorperian SK, Moufarrej MN, Quake SR. Cell types of origin of the cell-free transcriptome. Nat Biotechnol. 2022;40(6):855–61.
76. Geeurickx E, Hendrix A. Targets, pitfalls and reference materials for liquid biopsy tests in cancer diagnostics. Mol Asp Med. 2020;72:100828.
77. Freedman LP, Inglese J. The increasing urgency for standards in basic biological research. Cancer Res. 2014;74(15):4024–9.
78. Rounge TB, Umu SU, Keller A, Meese E, Ursin G, Tretli S, et al. Circulating small non-coding RNAs associated with age, sex, smoking, body mass and physical activity. Sci Rep. 2018;8:17650.
79. Zheng Y. Study design considerations for Cancer biomarker discoveries. J Appl Lab Med. 2018;3(2):282–9.
80. Wagner JT, Kim HJ, Johnson-Camacho KC, Kelley T, Newell LF, Spellman PT, et al. Diurnal stability of cell-free DNA and cell-free RNA in human plasma samples. Sci Rep. 2020;10(1):16456.
81. Liquid Biopsy Market - Global Forecast to 2026 | MarketsandMarkets. Available from: https://www.marketsandmarkets.com/Market-Reports/liquid-biopsy-market-13966350.html. [cited 7 Mar 2022].
82. Rolfo C, Cardona AF, Cristofanilli M, Paz-Ares L, Diaz Mochon JJ, Duran I, et al. Challenges and opportunities of cfDNA analysis implementation in clinical practice: perspective of the International Society of Liquid Biopsy (ISLB). Crit Rev Oncol Hematol. 2020;151:102978.

83. FDA Approves Blood Tests That Can Help Guide Cancer Treatment - National Cancer Institute. 2020. Available from: https://www.cancer.gov/news-events/cancer-currents-blog/2020/fda-guardant-360-foundation-one-cancer-liquid-biopsy [cited 7 Mar 2022].
84. University of Southern California. Studies of Cell-Free DNA and RNA in Blood From Patients Being Treated for Prostate Cancer 2021. clinicaltrials.gov. Report No.: NCT02853097. Available from: https://clinicaltrials.gov/ct2/show/NCT02853097 [cited 3 Mar 2022].
85. Illumina, Inc. Prospective Collection of Whole Blood Specimens of Subjects Diagnosed With Preeclampsia With Severe Features and/or Fetal Growth Restriction in Support of a Molecular Assay Development 2017. clinicaltrials.gov; Report No.: NCT02808494. Available from: https://clinicaltrials.gov/ct2/show/NCT02808494 [cited 3 Mar 2022].
86. Das S. Circulating RNAs in Acute Heart Failure 2020. clinicaltrials.gov; Report No.: NCT03345446. Available from: https://clinicaltrials.gov/ct2/show/NCT03345446 [cited 3 Mar 2022].
87. Fondazione Policlinico Universitario Agostino Gemelli IRCCS. Translational Analysis In Longitudinal Series of Ovarian Cancer ORganoids 2020. clinicaltrials.gov; Report No.: NCT04555473. Available from: https://clinicaltrials.gov/ct2/show/NCT04555473 [cited 3 Mar 2022].
88. Vento A. Epitranscriptomic Biomarkers for Ischemic Heart Disease (IHD-EPITRAN) - A Prospective Cohort Study 2020. clinicaltrials.gov; Report No.: NCT04533282. Available from: https://clinicaltrials.gov/ct2/show/NCT04533282 [cited 3 Mar 2022].
89. Hu X, Bao J, Wang Z, Zhang Z, Gu P, Tao F, et al. The plasma lncRNA acting as fingerprint in non-small-cell lung cancer. Tumour Biol J Int Soc Oncodevelopmental Biol Med. 2016;37(3):3497–504.
90. Yao Y, Chen X, Lu S, Zhou C, Xu G, Yan Z, et al. Circulating long noncoding RNAs as biomarkers for predicting head and neck squamous cell carcinoma. Cell Physiol Biochem. 2018;50(4):1429–40.
91. Li Y, Zhao J, Yu S, Wang Z, He X, Su Y, et al. Extracellular vesicles long RNA sequencing reveals abundant mRNA, circRNA, and lncRNA in human Blood as potential biomarkers for Cancer diagnosis. Clin Chem. 2019;65(6):798–808.
92. Lennon AM, Buchanan AH, Kinde I, Warren A, Honushefsky A, Cohain AT, et al. Feasibility of blood testing combined with PET-CT to screen for cancer and guide intervention. Science. 2020;369(6499):eabb9601.
93. Sreedharan VT, Schultheiss SJ, Jean G, Kahles A, Bohnert R, Drewe P, et al. Oqtans: the RNA-seq workbench in the cloud for complete and reproducible quantitative transcriptome analysis. Bioinformatics. 2014;30(9):1300–1.
94. Flomics Stratus. Flomics Bioinformatics Cloud. Flomics. Available from: https://stratus.flomics.com/. [cited 7 Mar 2022].

Publisher's Note