



# 'Longing' for the Next Generation of Liquid Biopsy: The Diagnostic Potential of Long Cell-Free DNA in Oncology and Prenatal Testing

Stephanie C. Y. Yu<sup>1,2,3</sup> · L. Y. Lois Choy<sup>1,2,3,4</sup> · Y. M. Dennis Lo<sup>1,2,3,4</sup>

Accepted: 15 June 2023 / Published online: 20 July 2023  
© The Author(s) 2023

## Abstract

Liquid biopsy using cell-free DNA (cfDNA) has gained global interest as a molecular diagnostic tool. However, the analysis of cfDNA in cancer patients and pregnant women has been focused on short DNA molecules (e.g.,  $\leq 600$  bp). With the detection of long cfDNA in the plasma of pregnant women and cancer patients in two recent studies, a new avenue of long cfDNA-based liquid biopsy has been opened. In this review, we summarize our current knowledge in this nascent field of long cfDNA analysis, focusing on the fragmentomic and epigenetic features of long cfDNA. In particular, long-read sequencing enabled single-molecule methylation analysis and subsequent determination of the tissue-of-origin of long cfDNA, which has promising clinical potential in prenatal and cancer testing. We also examine some of the limitations that may hinder the immediate clinical applications of long cfDNA analysis and the current efforts involved in addressing them. With concerted efforts in this area, it is hoped that long cfDNA analysis will add to the expanding armamentarium of liquid biopsy.

## 1 Introduction

The analysis of cell-free DNA (cfDNA) in bodily fluids for disease diagnosis and monitoring has been gaining importance. The detection of fetal chromosomal aneuploidies based on maternal plasma cfDNA analysis has revolutionized the field of prenatal testing [1–4]. The detection of circulating tumor DNA has been used to guide targeted therapy for lung cancer [5]. However, most of the existing research efforts have been focused on short cfDNA molecules (e.g.,  $\leq 600$  bp). This can be explained by the fact that most of the studies utilized massively parallel short-read sequencing

### Key Points

Long cfDNA exists in the plasma of healthy subjects, pregnant women, and cancer patients.

Efforts have been made to understand the fragmentomic, epigenetic, and tissue-of-origin information of long cfDNA, with demonstrated potential clinical utilities in prenatal testing and oncology.

The high cost and low throughput of long-read sequencing are limitations that need to be addressed before the clinical potential of long cfDNA analysis can be fully realized.

Stephanie C. Y. Yu and L. Y. Lois Choy contributed equally to this work.

✉ Y. M. Dennis Lo  
loym@cuhk.edu.hk

- <sup>1</sup> Centre for Novostics, Hong Kong Science Park, Pak Shek Kok, Hong Kong SAR, China
- <sup>2</sup> Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China
- <sup>3</sup> Department of Chemical Pathology, The Chinese University of Hong Kong, Prince of Wales Hospital, New Territories, Shatin, Hong Kong SAR, China
- <sup>4</sup> State Key Laboratory of Translational Oncology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, Hong Kong SAR, China

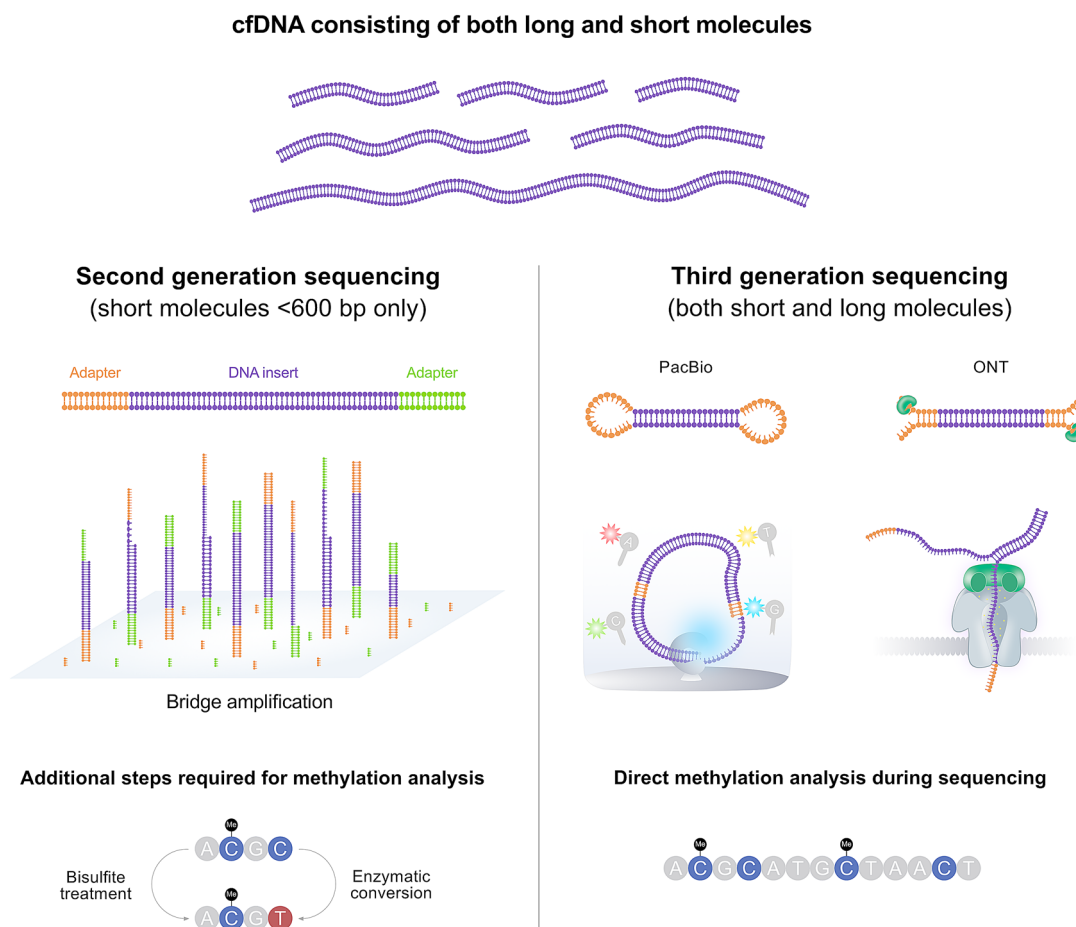
technologies, for example, the Illumina sequencing platform, which have limitations in detecting long DNA molecules (Fig. 1). For instance, the bridge amplification in the Illumina sequencing technology favors the amplification of short DNA molecules over that of longer DNA molecules [6]. Also, long DNA molecules would generate more diffuse clusters during bridge amplification, thereby lowering signal intensities [6, 7].

The development of third-generation sequencing platforms has sparked interest in long-read sequencing. Single-molecule real-time (SMRT) sequencing by Pacific Biosciences (PacBio) and nanopore sequencing by Oxford

Nanopore Technologies (ONT), which can generate significantly longer reads than next-generation sequencing [8], are key players in the field (Fig. 1). The Telomere-to-Telomere Consortium (T2T) utilized both SMRT sequencing and nanopore sequencing to fill in the gaps of the human reference genome [9]. Because of the potential of resolving long repeats and complex structural variants, long-read sequencing was named the “Method of the Year” by Nature Methods in 2022 [10].

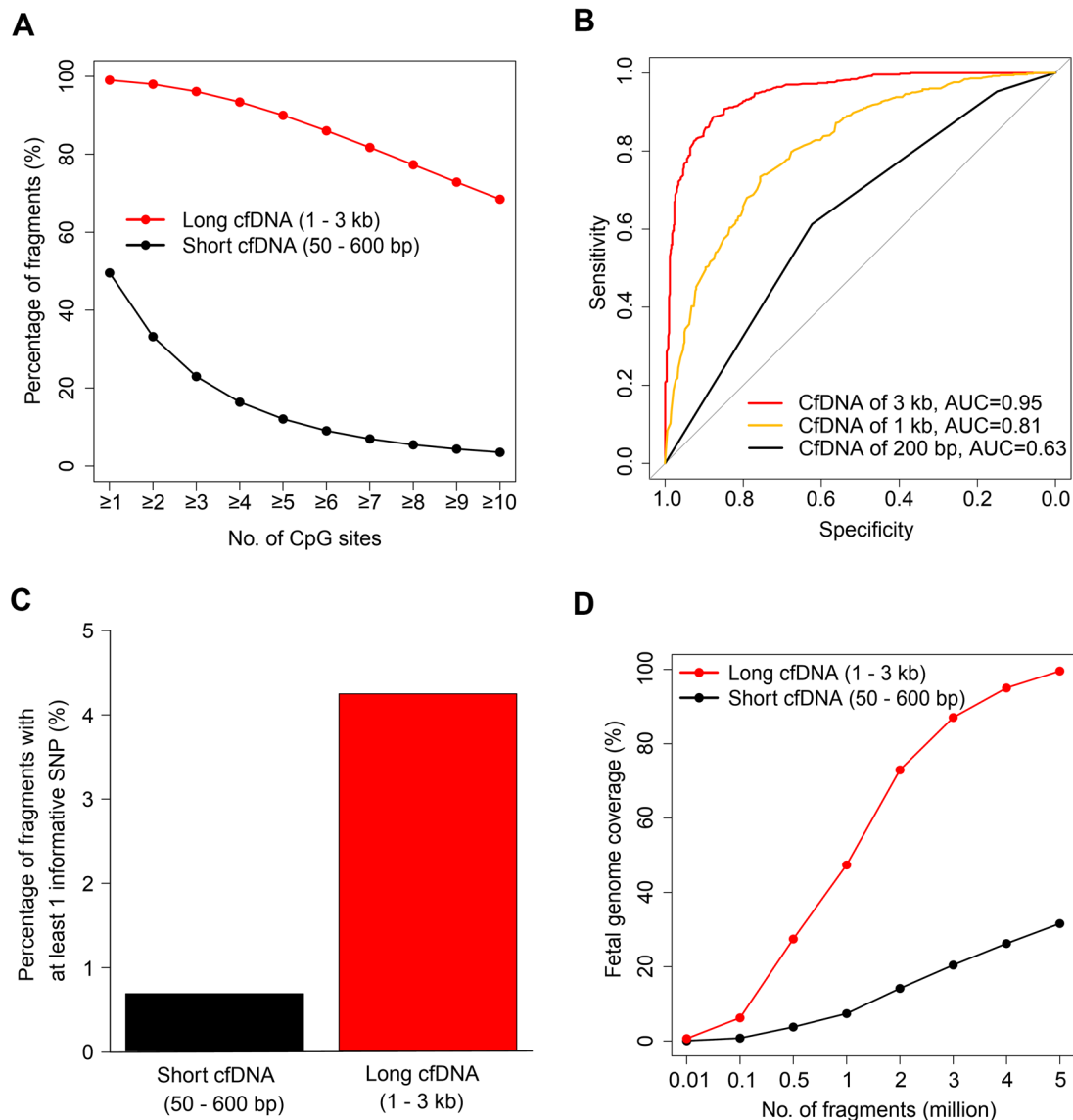
Recently, using single-molecule sequencing, two studies have demonstrated the presence of long cfDNA in the plasma of pregnant women [11] and cancer patients [12], respectively. This has opened a new avenue of long cfDNA-based liquid biopsy. Coupled with the development of novel approaches, such as the holistic kinetic model [13] that allows direct DNA methylation analysis without a prior chemical or enzymatic conversion step, we are able to obtain genetic, fragmentomic features, as well as epigenetic

characteristics [14] of long cfDNA. Compared to short cfDNA, long cfDNA generally contains more CpG and SNP sites (Fig. 2). As previously reported, among long cfDNA molecules in the size range of 1–3 kb, more than 80% carried at least seven CpG sites, whereas among shorter cfDNA molecules in the size range of 200–600 bp and below 200 bp, less than 20% and 5% carried at least seven CpG sites, respectively [12] (Fig. 2A). The increased number of CpG sites associated with long cfDNA molecules was found to enhance the performance of an approach for deducing the tissue-of-origin of individual cfDNA molecules based on an analysis of their methylation patterns at the single-molecule level (Fig. 2B). Moreover, based on computer simulations, long cfDNA molecules in the size range of 1–3 kb had a sixfold higher percentage of molecules with at least one informative SNP site for differentiating fetal- and maternal-derived cfDNA than short cfDNA molecules in the size range of 50–600 bp (Fig. 2C). As a result, a smaller number



**Fig. 1** Cell-free DNA molecules consist of both short and long molecules. In second-generation sequencing, e.g., Illumina sequencing, only the short DNA molecules  $\leq 600$  bp are analyzed. For epigenetic analysis using second-generation sequencing platforms, additional steps such as bisulfite treatment or enzymatic conversion are neces-

sary before sequencing. In contrast, for third-generation sequencing platforms, e.g., single-molecule real-time (SMRT) sequencing by Pacific Biosciences (PacBio) and nanopore sequencing by Oxford Nanopore Technologies (ONT), they can sequence long DNA molecules and perform direct methylation analysis



**Fig. 2** Advantages of the analysis of long cfDNA. **A** Long cfDNA molecules contain a larger number of CpG sites than short cfDNA molecules. **B** The performance of a classifier for differentiating the tissue-of-origin (e.g., placenta vs. buffy coat) of individual cfDNA molecules based on the analysis of single-molecule methylation pattern of cfDNA improves when cfDNA molecules of longer sizes are

used. **C** There is a higher percentage of molecules containing an informative SNP site for differentiating fetal- and maternal-derived cfDNA molecules among long cfDNA compared to short cfDNA. **D** A smaller number of long cfDNA molecules will need to be sequenced to attain the same coverage of the fetal genome compared to short cfDNA

of cfDNA molecules would need to be sequenced to achieve the same coverage of the fetal genome (Fig. 2D). These have tremendous potential in clinical implementation in oncology and noninvasive prenatal testing (NIPT). Table 1 compares the characteristics, sequencing methods and limitations of short cfDNA and long cfDNA. In this article, we review the current efforts that have been made in prenatal and cancer testing and probe into the future of long cfDNA analysis.

## 2 Long Cell-Free DNA Fragmentomics

Recent studies have shown that, with the use of SMRT sequencing, about 10–20% of long cfDNA molecules of > 1 kb were analyzable in the plasma [11, 12]. When looking at high-resolution size profiles of cfDNA obtained from SMRT sequencing, besides the well-described mononucleosomal

**Table 1** Comparison between short cell-free (cf) DNA and long cfDNA

	Short cfDNA	Long cfDNA
Length <sup>a</sup>	≤ 600 bp	> 600 bp
Characteristics	Mononucleosomal and dinucleosomal peaks Predominantly ended with C	Peaks at multiples of nucleosomal units extending to 2 kb Increasing proportion of A end
Sequencing methods	Next-generation sequencing	Third-generation sequencing (SMRT or nanopore sequencing)
Limitations	Carry smaller number of CpG sites, less information provided for methylation pattern analysis Difficult to detect structural variants and repeat expansions	Relatively low throughput of third-generation sequencing platforms: The number of DNA molecules available for analysis is less than that of short cfDNA Relatively high cost of third-generation sequencing platforms

<sup>a</sup>600 bp is the detection limit of the Illumina sequencing platform. Definition of long and short cfDNA can vary according to different features and applications

dinucleosomal DNA peaks observable in previous studies using short-read sequencing, an additional series of peaks at multiples of nucleosomal units extending to molecules of 2 kb in size could be observed. Such ladder patterns suggest that apoptosis might be an important mechanism for the generation of long cfDNA.

It is well known that both the cell-free fetal DNA in maternal plasma and the tumor-derived DNA in plasma of cancer patients showed a shorter modal size of 143 bp compared to their background DNA (maternal-derived DNA and non-tumoral DNA) [15, 16]. With the analysis of long cfDNA, such a conclusion could now be generalized to a broader size spectrum of up to 3 kb [11, 12]. In other words, both fetal fraction and tumor fraction were negatively correlated with cfDNA size. Nevertheless, long cfDNA from fetal and tumoral origin could be detected in the respective plasma samples. Thus far, the longest detected fetal- and tumor-derived DNA molecules are 24 kb and 14 kb, respectively [11, 12]. Thus, the use of long-read sequencing for cfDNA analysis has extended the analyzable size spectrum of cfDNA.

The use of long-read sequencing also allows the DNA sequences at fragment ends to be deduced [11, 12]. While short cfDNA fragments of < 500 bp are predominantly ended with C, an increasing proportion of A end has been observed in longer cfDNA fragments of > 500 bp. Previously, biological links between fragment end characteristics and activities of various nucleases have been established [17]. The enrichment of 5' A end motifs in long cfDNA fragments suggests that their generation might be related to the nuclease DNA fragmentation factor subunit β (DFFB).

### 3 Single-Molecule Methylation Pattern Analysis of Long cfDNA

Based on distinct methylation patterns among different tissues, numerous methods have been developed for tracing the tissue origins of cfDNA [18–21]. Most studies have

employed an approach that uses combined methylation signals from bulk collections of cfDNA molecules that have been mapped to tissue-specific marker regions to deduce the corresponding tissue contribution to cfDNA in a sample [18, 19]. These studies have provided data demonstrating the relative contributions of different tissues to cfDNA in plasma in various physiologic or pathologic conditions.

Another approach is to use methylation signals on individual cfDNA molecules to deduce their respective tissues of origin [11, 12, 21]. In fact, both short-read sequencing [21] and long-read sequencing [11, 12] allow the methylation status of each CpG site on a cfDNA molecule to be determined. However, due to the size detection limit of short-read sequencing and the fact that there is on average only one CpG site in every 100 nucleotides, most cfDNA fragments that can be analyzed by a short-read sequencing platform (i.e., mostly in the size range of 150–400 bp) would probably contain only one to four CpG sites on average. On the other hand, for the reference tissue methylomes which methylation patterns of individual cfDNA molecules have been compared to, they can either be in the form of population-based average methylation levels at individual CpG sites or single molecule-based methylation patterns. While the former could be readily obtained from methylation data generated by short-read bisulfite sequencing or DNA methylation microarray analysis, the latter would require data generated from high-depth long-read sequencing of genomic DNA from different tissues.

Taking advantage of the increased number of CpG sites associated with long cfDNA molecules, as well as the ability of a long-read sequencing platform to simultaneously analyze long cfDNA molecules and their associated methylation, Yu et al. have developed an approach for the classification of individual cfDNA molecules in maternal plasma samples as being derived from the placenta or buffy coat based on the comparison of the single-molecule methylation pattern of cfDNA with the reference methylation profiles of placenta and buffy coat obtained from high-depth bisulfite

sequencing [11]. This approach achieved an area under the receiver operating characteristic curve (AUC) of 0.88. A similar approach was used by Choy et al. to deduce the tissue origin of individual cfDNA molecules and to quantify liver-derived cfDNA in plasma samples from healthy individuals, patients with chronic hepatitis B virus (HBV) infection, and patients with hepatocellular carcinoma (HCC) [12].

Based on a computer simulation analysis, it was found that the performance of such tissue-of-origin analysis would improve with an increasing number of informative CpG sites on individual cfDNA molecules [11]. Hence, single-molecule methylation pattern analysis enhances the resolution of the tissue-of-origin analysis, with the use of long cfDNA allowing a more accurate tissue-of-origin determination than the use of short cfDNA.

## 4 Analytical Platforms for Long cfDNA Analysis

There are two predominant long-read sequencing platforms, namely SMRT sequencing from PacBio and nanopore sequencing from ONT, which can be used for long cfDNA analysis [22]. Based on the analysis of artificial mixtures of sonicated human and mouse genomic DNA

of different sizes (1500 bp vs. 200 bp) at a molar ratio of 1:1, researchers found that both the PacBio and the ONT showed bias towards sequencing of longer DNA fragments, with the extent of bias from PacBio stronger than that from ONT (a fivefold vs. a twofold over-representation of long fragments) [22].

Table 2 provides a summary of the comparison between PacBio and ONT for long cfDNA analysis [22].

## 5 Potential Clinical Utilities of Long cfDNA Analysis

### 5.1 Size-Based Detection of Pre-eclampsia

Pre-eclampsia is a pregnancy complication associated with increased maternal and neonatal morbidity and mortality. Both the absolute amounts of total cfDNA and fetal cfDNA have been reported to be elevated in pregnancies with pre-eclampsia [23, 24]. In addition to such quantitative differences, a recent study by Yu et al. revealed that there was a significant reduction in the proportion of long cfDNA molecules in pregnancies with pre-eclampsia [11]. Therefore, a classifier that was based on the percentage of long cfDNA in a maternal plasma sample for differentiating pregnancies with and without pre-eclampsia has been built. Interestingly,

**Table 2** Comparison between PacBio and ONT for long cell-free (cf) DNA analysis

	Comparison between PacBio and ONT <sup>a</sup>	Clinical or analytical implication
Sequencing data output and sequencing cost per flow cell <sup>b</sup>	ONT gives 10 times more mapped reads per flow cell Sequencing cost per flow cell is comparable between the two platforms (~ \$1000 USD)	It may be more cost-effective to use ONT for clinical applications that require higher data output
Sequencing error	PacBio shows lower substitution error rates (0.8% vs. 3.4%)	PacBio may be more suitable for clinical applications that require accurate variant calling
Size and end motif analysis	From ONT sequencing data: 35% of base-called reads contain no adapters or adapters at only one end; > 50% of aligned reads contain soft clips Data from PacBio show 30 times higher percentages of long cfDNA > 1 kb ONT data provides 4 times more fragments that are eligible for size analysis PacBio allows the analysis of the 5' fragment ends of both the Watson and Crick strands of a double-stranded cfDNA molecule while ONT only allowed the analysis of the 5' fragment end of one of the two strands (the sequenced strand) Platform-dependent end motif profiles could be observed	ONT data requires additional processing steps to identify and remove reads with incomplete adapter sequences before adapter trimming and to remove reads with soft clips. Consequently, not all mapped reads can be used for size and end motif analysis End motif profiles obtained from different platforms may not be comparable
Tissue-of-origin analysis	ONT data provide 2–4 times higher number of fragments that are eligible for tissue-of-origin analysis Performance (in terms of AUC) is comparable between the two platforms	

<sup>a</sup>The comparison is based on the analysis of a cfDNA sample with one SMRT Cell 8M on a Sequel II sequencer and the analysis of the same cfDNA sample with one flow cell on a PromethION 24 sequencer [22]

<sup>b</sup>Sequencing cost per flow cell includes the cost for library preparation, flow cell, and sequencing reagents



such a cfDNA size-based classifier was found to perform much better when a long-read sequencing platform was used (AUC: 1 vs. 0.7 when a short-read sequencing platform was used), possibly related to an extended spectrum of cfDNA size that was analyzable by a long-read sequencing platform. Nevertheless, future studies are needed to validate these findings and to investigate the predictive power of such a cfDNA size-based biomarker for pre-eclampsia before the onset of clinical symptoms.

## 5.2 Single-Molecule Methylation-Based Analysis for the Noninvasive Prenatal Testing of Monogenic Diseases

A major technical challenge for the noninvasive prenatal testing (NIPT) of monogenic diseases had been the assessment of maternally inherited mutations of the fetus because of the mother's own contribution of the mutant alleles to the maternal plasma. To overcome this challenge, strategies including relative mutation dosage analysis (RMD) [25] and relative haplotype dosage analysis (RHDO) [15], which determine the maternal inheritance of the fetus by detecting the slight allelic and haplotype dosage imbalance, respectively, in the maternal plasma, have been developed.

As an alternative to these quantitative approaches, Yu et al. recently described the principle of a qualitative approach for determining both maternal and paternal inheritance of the fetus based on the aforementioned single-molecule tissue-of-origin analysis of long cfDNA [11]. The principle is that if a cfDNA molecule that carried a paternal- or maternal-specific genetic alteration, such as an SNP or a disease-causing mutation, was identified as being derived from the placenta based on the methylation pattern analysis, such genetic alteration would be considered as being inherited by the fetus.

In a proof-of-concept study, Yu et al. demonstrated the feasibility of deducing the maternal inheritance of the fetus on a chromosome-arm level based on the single-molecule tissue-of-origin analysis of long cfDNA [11]. Of note, as the accuracy of the current version of the tissue-of-origin analysis and the sequencing output of the current long-read sequencing platforms are far from sufficient for the realization of such a qualitative approach, a quantitative approach comparing the number of cfDNA molecules being determined to be of placental origin between the two maternal haplotypes has been employed. Furthermore, this analysis was performed with cfDNA molecules that were pooled from 28 maternal plasma samples due to the limited number of cfDNA molecules that were associated with informative SNPs in each sample. Nevertheless, the approach was shown to achieve a classification rate of 90% and an accuracy of

100%. Finally, Yu et al. demonstrated the feasibility of NIPT for fragile X syndrome and the detection of a recombination event on chromosome X in a pregnancy involving a male fetus who was at risk of fragile X syndrome [11].

## 5.3 Single-Molecule Methylation-Based Analysis for Cancer Detection

In another proof-of-concept study, Choy et al. applied the single-molecule tissue-of-origin analysis to cancer detection [12]. A scoring system, which was named the HCC methylation score, was developed to assess the likelihood for a patient to have HCC, with a higher score indicating a higher likelihood. It was shown that HCC patients have significantly higher HCC methylation scores when compared to HBV carriers and healthy individuals. Choy et al. showed that the use of long cfDNA molecules with at least seven CpG sites further improved the discriminating power of the HCC methylation score when compared to the use of shorter cfDNA molecules with one to six CpG sites, with the AUC improved from 0.75 to 0.91.

In addition to the use of single-molecule methylation-based analysis alone for cancer detection, one can combine this methylation-based analysis of long cfDNA with the detection of cancer-associated somatic mutation to improve the diagnostic specificity for cancer detection. In the detection of cancer-associated somatic mutations from cfDNA, clonal hematopoiesis, which is an age-related phenomenon due to the clonal expansion of blood cells carrying somatic mutations, can lead to false-positive identification of cancer-associated mutations [26]. The analysis of long cfDNA in this scenario will be potentially useful. Due to its longer length, it is more likely for a long cfDNA molecule to carry a mutation together with multiple CpG sites. Combining the mutation information with the methylation pattern information offered by multiple CpG sites in long cfDNA can potentially reduce false-positive results. For example, if a cfDNA molecule contains both the mutation and the methylation pattern that is suggestive of tumoral origin instead of hematopoietic origin, then one can be more confident that the mutation identified is from the tumor, rather than the result of clonal hematopoiesis.

## 5.4 Detection of Structural Variants and Repeat Expansions

Structural variants (SVs), which are defined as DNA rearrangements of at least 50 bp [27], and repeat expansions are known causes of monogenic diseases [28]. Also, they are associated with complex disorders such as cancers [29, 30]. The detection of SVs and repeat expansions are valuable for disease diagnosis, monitoring, or treatment stratification [29, 30]. However, the detection of SVs and

the accurate determination of the number of repeats from cfDNA remain technically challenging with limited sensitivity [31], which can be partly attributed to the analysis of only the short cfDNA using the short-read sequencing technology. Such short cfDNA molecules of  $\leq 600$  bp in length rarely span the entire length of the SV or the repeat regions, resulting in some ambiguity during sequence alignment [27]. We believe that the analysis of long cfDNA with the use of long-read sequencing technologies would likely reduce alignment ambiguity and improve the coverage of repetitive regions, thereby improving the detection of SVs and repeat expansions.

## 6 Conclusion and Future Perspectives

Currently, the Achilles’ heel of long cfDNA analysis is low throughput. The long-read sequencing platforms have lower throughput than second-generation sequencing platforms. Depending on the sequencing platform used, long-read sequencing platforms can generate up to 300 gigabytes of data per flow cell, whereas the second-generation sequencing platforms can generate up to 3 terabytes of data per flow cell. Such a relatively low throughput for current long-read sequencing platforms hinders the immediate clinical application of long cell-free DNA analysis. For instance, in the study by Yu et al., due to the inadequate number of plasma DNA molecules with informative SNPs in individual samples, a chromosome-arm-level analysis with pooled data from multiple maternal plasma samples had to be employed for the deduction of maternal inheritance of the fetus [11]. In addition to low throughput, the high cost of long-read sequencing also hinders its widespread use in resource-limited settings.

Despite the current limitations, there have been ongoing efforts to improve the hardware of long-read sequencing. Recently, PacBio launched a new sequencer called Revio, with the number of zeromode waveguides in each SMRT cell increased from 8 million to 25 million [32], which may potentially increase the throughput. Also, ONT has released a new sequencing kit (Kit 14), with raw-read accuracy at 99.6% [33].

In addition to hardware, efforts have been made in refining the software of long-read sequencing. Recently, partnering with Google, PacBio has introduced DeepConsensus, which utilized an alignment-based loss to train a gap-aware transformer-encoder for sequence correction [34]. DeepConsensus has been claimed to reduce read errors by 42% when compared to the standard approach based on a hidden Markov model [34].

In addition to single-molecule sequencing technologies (e.g., SMRT sequencing or nanopore sequencing), which sequence long DNA molecules directly without the need

for PCR amplification, new players using the synthetic long-read sequencing approach are also emerging. Examples include the LoopSeq synthetic long read by Element Biosciences and the Complete Long-Read technology by Illumina. In general, the synthetic long-read sequencing approach involves labelling short fragments from individual long DNA molecules with unique barcode sequences, amplifying and then sequencing the barcoded short fragments with a short-read sequencing platform, and finally computationally reconstructing long reads from short sequencing reads. While the use of a short-read sequencing platform may potentially reduce the cost, long cfDNA analysis using the synthetic long-read sequencing approach may suffer from the potential loss of accurate fragmentomic information, such as the size and end motif signatures, as a result of the reassembling of long reads from barcoded short sequencing reads. Also, current synthetic long-read sequencing technologies do not support direct DNA methylation analysis, which may impede the tissue-of-origin analysis of cfDNA. Thus, these limitations may prevent the immediate application of the synthetic long-read sequencing approach for the analysis of long cfDNA. Nevertheless, the long-read sequencing field is actively working towards improving sequencing throughput with lower cost and higher accuracy, which may ultimately benefit cfDNA analysis in the future.

As we gradually learn more about long cfDNA, further studies into the preanalytical factors affecting long cfDNA analysis would be valuable. Certain preanalytical factors, such as specimen collection tubes, blood sample storage conditions, blood-processing protocols, and DNA extraction methods, might affect long and short DNA molecules to different extents. A detailed and thorough understanding of the preanalytical aspects would be necessary before clinical implementation of long cell-free DNA-based liquid biopsy to ensure consistent clinical result interpretation. With continuous efforts, it is hoped that long cfDNA analysis can enhance the spectrum of diagnostic applications of liquid biopsy.

**Acknowledgements** We thank Dr. Mary-Jane L. Ma and Mr. Wenlei Peng for their assistance in the preparation of this review.

## Declarations

**Funding** Centre for Novostics is supported by the Innovation and Technology Commission under the InnoHK initiative. Y.M.D.L. is supported by an endowed chair from the Li Ka Shing Foundation.

**Conflicts of interest** S.C.Y.Y. has filed patents on cell-free DNA analysis, received royalties from Illumina, Xcelom, Take2, and DRA, and received financial support from Oxford Nanopore Technologies for attending a meeting. L.Y.L.C. has filed patents on cell-free DNA analysis. Y.M.D.L. has filed patents on circulating nucleic acid analyses and has received royalties from Illumina, Grail, Xcelom, DRA, Take2, Se-

quenom and Insighta. Y.M.D.L. has equity interest in Grail/Illumina, DRA, Take2 and Insighta.

**Availability of Data and Materials** Not applicable.

**Compliance with Ethical Standards** Not applicable.

**Ethical Approval** Not applicable.

**Consent** Not applicable.

**Code Availability** Not applicable.

**Author Contributions** All authors have made substantial contributions to the intellectual content of the article.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc/4.0/>.

## References

- Lo YMD, Corbetta N, Chamberlain PF, Rai V, Sargent IL, Redman CW, et al. Presence of fetal DNA in maternal plasma and serum. *Lancet*. 1997;350(9076):485–7. [https://doi.org/10.1016/S0140-6736\(97\)02174-0](https://doi.org/10.1016/S0140-6736(97)02174-0).
- Bianchi DW, Chiu RWK. Sequencing of circulating cell-free DNA during pregnancy. *N Engl J Med*. 2018;379(5):464–73. <https://doi.org/10.1056/NEJMr1705345>.
- van der Meij KRM, Sistermans EA, Macville MVE, Stevens SJC, Bax CJ, Bekker MN, et al. TRIDENT-2: national implementation of genome-wide non-invasive prenatal testing as a first-tier screening test in the Netherlands. *Am J Hum Genet*. 2019;105(6):1091–101. <https://doi.org/10.1016/j.ajhg.2019.10.005>.
- American College of Obstetricians and Gynecologists' Committee on Practice Bulletins—Obstetrics; Committee on Genetics; Society for Maternal-Fetal Medicine. Screening for fetal chromosomal abnormalities: ACOG practice bulletin, number 226. *Obstet Gynecol*. 2020;136(4):e48–69. <https://doi.org/10.1097/AOG.0000000000004084>.
- Kwapisz D. The first liquid biopsy test approved. Is it a new era of mutation testing for non-small cell lung cancer? *Ann Transl Med*. 2017;5(3):46. <https://doi.org/10.21037/atm.2017.01.32>.
- Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, et al. Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*. 2014;56(2):61–4. <https://doi.org/10.2144/000114133>. (66, 68, passim).
- Buermans HP, den Dunnen JT. Next generation sequencing technology: advances and applications. *Biochim Biophys Acta*. 2014;1842(10):1932–41. <https://doi.org/10.1016/j.bbdis.2014.06.015>.
- Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol*. 2020;21(1):30. <https://doi.org/10.1186/s13059-020-1935-5>.
- Nurk S, Koren S, Rhie A, Rautiainen M, Bizikadze AV, Mikheenko A, et al. The complete sequence of a human genome. *Science*. 2022;376(6588):44–53. <https://doi.org/10.1126/science.abj6987>.
- Marx V. Method of the year: long-read sequencing. *Nat Methods*. 2023;20(1):6–11. <https://doi.org/10.1038/s41592-022-01730-w>.
- Yu SCY, Jiang P, Peng W, Cheng SH, Cheung YTT, Tse OYO, et al. Single-molecule sequencing reveals a large population of long cell-free DNA molecules in maternal plasma. *Proc Natl Acad Sci U S A*. 2021;118(50):e2114937118. <https://doi.org/10.1073/pnas.2114937118>.
- Choy LYL, Peng W, Jiang P, Cheng SH, Yu SCY, Shang H, et al. Single-molecule sequencing enables long cell-free DNA detection and direct methylation analysis for cancer patients. *Clin Chem*. 2022;68(9):1151–63. <https://doi.org/10.1093/clinchem/hvac086>.
- Tse OYO, Jiang P, Cheng SH, Peng W, Shang H, Wong J, et al. Genome-wide detection of cytosine methylation by single molecule real-time sequencing. *Proc Natl Acad Sci U S A*. 2021;118(5):e2019768118. <https://doi.org/10.1073/pnas.2019768118>.
- Lo YMD, Han DSC, Jiang P, Chiu RWK. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science*. 2021;372(6538):eaaw3616. <https://doi.org/10.1126/science.aaw3616>.
- Lo YMD, Chan KCA, Sun H, Chen EZ, Jiang P, Lun FM, et al. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci Transl Med*. 2010;2(61):61ra91. <https://doi.org/10.1126/scitranslmed.3001720>.
- Jiang P, Chan CW, Chan KCA, Cheng SH, Wong J, Wong VW, et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc Natl Acad Sci U S A*. 2015;112(11):E1317–25. <https://doi.org/10.1073/pnas.1500076112>.
- Han DSC, Ni M, Chan RWY, Chan VWH, Lui KO, Chiu RWK, et al. The biology of cell-free DNA fragmentation and the roles of DNASE1, DNASE1L3, and DFFB. *Am J Hum Genet*. 2020;106(2):202–14. <https://doi.org/10.1016/j.ajhg.2020.01.008>.
- Sun K, Jiang P, Chan KCA, Wong J, Cheng YK, Liang RH, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc Natl Acad Sci U S A*. 2015;112(40):E5503–12. <https://doi.org/10.1073/pnas.1508736112>.
- Lehmann-Werman R, Neiman D, Zemmour H, Moss J, Magenheimer J, Vaknin-Dembinsky A, et al. Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc Natl Acad Sci U S A*. 2016;113(13):E1826–34. <https://doi.org/10.1073/pnas.1519286113>.
- Guo S, Diep D, Plongthongkum N, Fung HL, Zhang K, Zhang K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat Genet*. 2017;49(4):635–42. <https://doi.org/10.1038/ng.3805>.
- Li W, Li Q, Kang S, Same M, Zhou Y, Sun C, et al. Cancer-Detector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. *Nucleic Acids Res*. 2018;46(15):e89. <https://doi.org/10.1093/nar/gky423>.
- Yu SCY, Deng J, Qiao R, Cheng SH, Peng W, Lau SL, et al. Comparison of single molecule, real-time sequencing and nanopore sequencing for analysis of the size, end-motif, and tissue-of-origin of long cell-free DNA in plasma. *Clin Chem*. 2023;69(2):168–79. <https://doi.org/10.1093/clinchem/hvac180>.



23. Lo YMD, Leung TN, Tein MS, Sargent IL, Zhang J, Lau TK, et al. Quantitative abnormalities of fetal DNA in maternal serum in preeclampsia. *Clin Chem*. 1999;45(2):184–8. <https://doi.org/10.1093/clinchem/45.2.184>.
24. Carbone IF, Conforti A, Picarelli S, Morano D, Alviggi C, Farina A. Circulating nucleic acids in maternal plasma and serum in pregnancy complications: are they really useful in clinical practice? A systematic review. *Mol Diagn Ther*. 2020;24(4):409–31. <https://doi.org/10.1007/s40291-020-00468-5>.
25. Lun FMF, Tsui NBY, Chan KCA, Leung TY, Lau TK, Charoenkwan P, et al. Noninvasive prenatal diagnosis of monogenic diseases by digital size selection and relative mutation dosage on DNA in maternal plasma. *Proc Natl Acad Sci U S A*. 2008;105(50):19920–5. <https://doi.org/10.1073/pnas.0810373105>.
26. Yu SCY, Chan KCA. What is the importance of analyzing tumor tissues and blood cells in the study of circulating tumor-derived DNA? *Clin Chem*. 2022;68(12):1481–3. <https://doi.org/10.1093/clinchem/hvac168>.
27. Liu Z, Roberts R, Mercer TR, Xu J, Sedlazeck FJ, Tong W. Towards accurate and reliable resolution of structural variants for clinical diagnosis. *Genome Biol*. 2022;23(1):68. <https://doi.org/10.1186/s13059-022-02636-8>.
28. Malik I, Kelley CP, Wang ET, Todd PK. Molecular mechanisms underlying nucleotide repeat expansion disorders. *Nat Rev Mol Cell Biol*. 2021;22(9):589–607. <https://doi.org/10.1038/s41580-021-00382-6>.
29. van Belzen IAEM, Schönhuth A, Kemmeren P, Hehir-Kwa JY. Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology. *NPJ Precis Oncol*. 2021;5(1):15. <https://doi.org/10.1038/s41698-021-00155-6>.
30. Erwin GS, Gürsoy G, Al-Abri R, Suriyaprakash A, Dolzhenko E, Zhu K, et al. Recurrent repeat expansions in human cancer genomes. *Nature*. 2023;613(7942):96–102. <https://doi.org/10.1038/s41586-022-05515-1>.
31. Mc Connell L, Gazdova J, Beck K, Srivastava S, Harewood L, Stewart JP, et al. Detection of structural variants in circulating cell-free DNA from sarcoma patients using next generation sequencing. *Cancers (Basel)*. 2020;12(12):3627. <https://doi.org/10.3390/cancers12123627>.
32. Revio System Specification Sheet. Pacific Biosciences of California, Inc. 2023. <https://www.pacb.com/wp-content/uploads/Revio-specification-sheet.pdf>. Accessed 23 Mar 2023.
33. The power of Q20+ chemistry. Oxford Nanopore Technologies plc. 2023. <https://nanoporetech.com/q20plus-chemistry> Accessed 23 Mar 2023.
34. Baid G, Cook DE, Shafin K, Yun T, Llinares-López F, Berthet Q, et al. DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nat Biotechnol*. 2023;41(2):232–8. <https://doi.org/10.1038/s41587-022-01435-7>.