SINGLE-STRANDED APPROACHES FOR cfDNA FRAGMENTOMICS

Cell-free DNA (cfDNA), found circulating in blood plasma, contains a wealth of clinically relevant biological information which can be recovered by minimally-invasive procedures[1]. Next Generation Sequencing (NGS) data obtained from cfDNA, can be used to monitor prenatal health, organ transplant reception or rejection, cancer and other diseases[2].

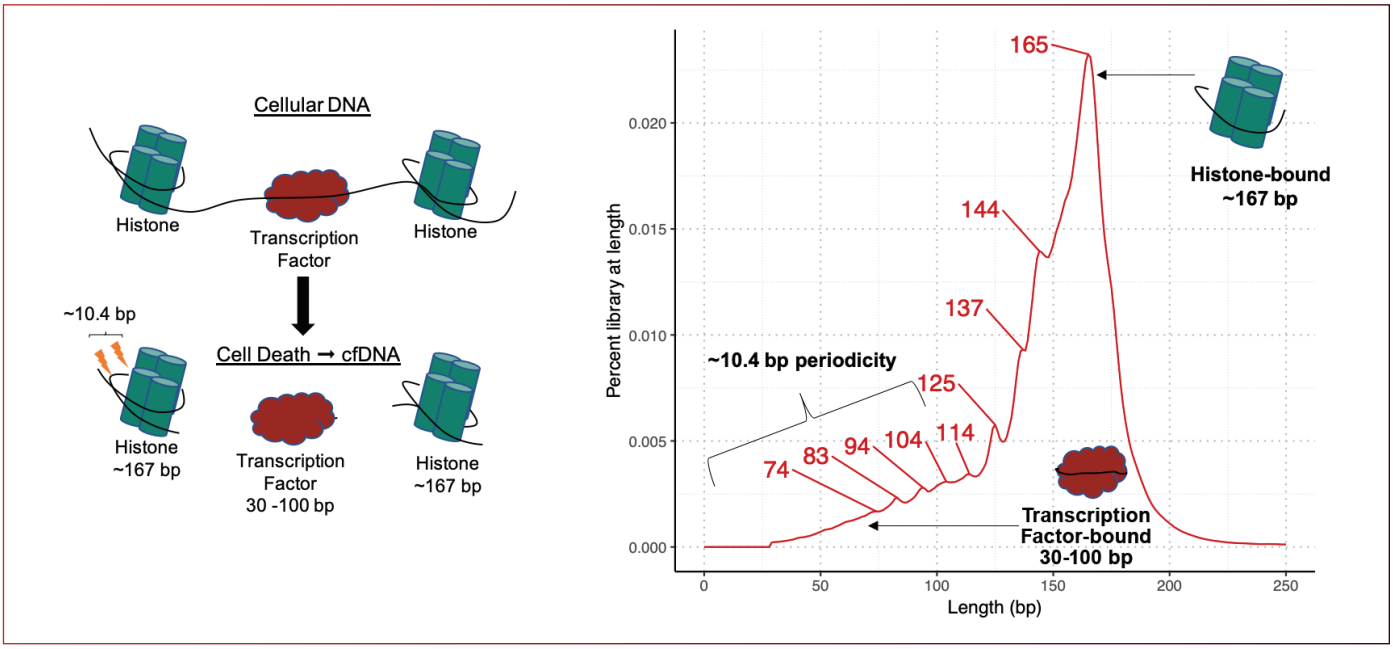

Figure 1. Left: Model for generation of cfDNA fragments. Nucleases preferentially cleave DNA at regions not associated with proteins, Thus, cfDNA fragments are often DNA segments associated with histones or other DNA binding proteins like transcription factors. Right: Typical cfDNA fragment size distribution. The main peak are DNA fragments associated with mono-nucleosomes.

The cfDNA pool arises from non-random fragmentation of cellular DNA during cell-death and subsequent nuclease activity within the blood, or through active secretion of DNA in the form of exosomal vesicles[3]. The majority of DNA fragments extracted from blood plasma are around 167 base-pairs (bp) in length (Figure 1)[4]. These fragments arise from DNA bound to the histone monomer, which is protected from nuclease degradation. However, the cfDNA pool also contains a valuable population of short length DNA fragments (30-100 bp). This subnucleosomal fraction contains DNA fragments protected by DNA binding proteins, mitochondrial DNA, and microbe-derived DNA, which have a smaller footprint than the nucleosome, all of which add a valuable layer of detail to cfDNA sequence data[4,5]. Additionally, some of these smaller fragments are products of sequential degradation of unbound DNA in the blood. This is reflected in the ~10.4 bp periodicity of the peak sizes within this region (Figure 1) suggesting that the DNA helical structure may also determine cleavage pattern of cfDNA.

WHY IS cfDNA FRAGMENTATION PROFILING IMPORTANT?

The fragmentation patterns of DNA, as revealed by sequencing cfDNA can be used to determine the position of nucleosomes and DNA-binding proteins and open-chromatin regions on the genome, at the time of cleavage. In this way, analysis of cfDNA fragments can reveal the information about the cell types and their biological state[6,7]. Plasma cfDNA contains a composite signal from all the tissues shedding DNA. Deconvolution of this signal can in turn throw light on underlying changes in chromatin organization, transcription factor positioning during disease progression, and can potentially reveal the identity of the damaged or diseased cell-type. Distinguishing fetal, tumor-derived, transplant-derived

fragments from cfDNA fragments originating from healthy tissues has clear diagnostic value[1,2].

Several statistical and machine-learning approaches have been used to better understand this complex signature. The Window Protection Score (WPS) approach developed by Snyder et al (2016), evaluates the differential DNA protection conferred by nucleosome and transcription factor binding in cancer and healthy states. This nucleosome spacing profile can be used to identify tumor-specific signals from cfDNA[4]. This study highlighted the benefit of a single-stranded library preparation in more accurate generation of WPS.

An alternative genome-wide approach called DNA EvaLuation of Fragments for early Interception (DELFI), incorporated cfDNA fragmentation sites in non-overlapping windows across the genome. This study not only observed larger median fragment size for healthy individuals, but also observed significant variation in fragment size of cancer-derived cfDNA molecules, which changed during the course of treatment[8]. While this study relied on a double-stranded library preparation method, the authors indicated that a single-stranded approach would improve the recovery of small fragments which harbor information that is more relevant to cancer diagnostics.

The Orientation-aware cfDNA Fragmentation (OCF) approach designed by Sun K et al (2019) evaluated the coverage imbalance at open-chromatin regions in cfDNA-derived data, based on fragmentation points.

Using existing databases for tissue-specific open-chromatin regions, this tool was used to elucidate tissue-of-origin information from cfDNA of healthy vs diseased individuals. The current opinion within the liquid biopsy field is that precise capture of cfDNA cut-sites will improve accurate determination of nucleosome positioning, particularly when evaluating changes to this profile during disease[9]. These studies stand to benefit from single-stranded library preparation methods that retain the exact fragmentation site[10].

COMMERCIAL NGS LIBRARY PREPARATION METHODS FOR cfDNA FRAGMENT PROFILING

Double-stranded Library preparation methods that convert cfDNA molecules into sequencing libraries are ineffective in capturing the complete picture of cfDNA fragmentomics (Figure 2). Most of these methods require end-repair of the double stranded input molecule, which alters the native ends of cfDNA fragments. These library protocols have several disadvantages for cfDNA analysis: (1) nicked and single-stranded cfDNA fragments do not ligate to the adapters and are eliminated from the final library (2) native DNA termini of the fragment molecules are altered by end-polishing in the final library (3) shorter cfDNA fragments are inefficiently captured.

While single-stranded methods such as those developed by Gansauge et al and Wu D.C et al, demonstrate effective capture of native ends, they suffer from low throughput and low library conversion rates[11,12]. The

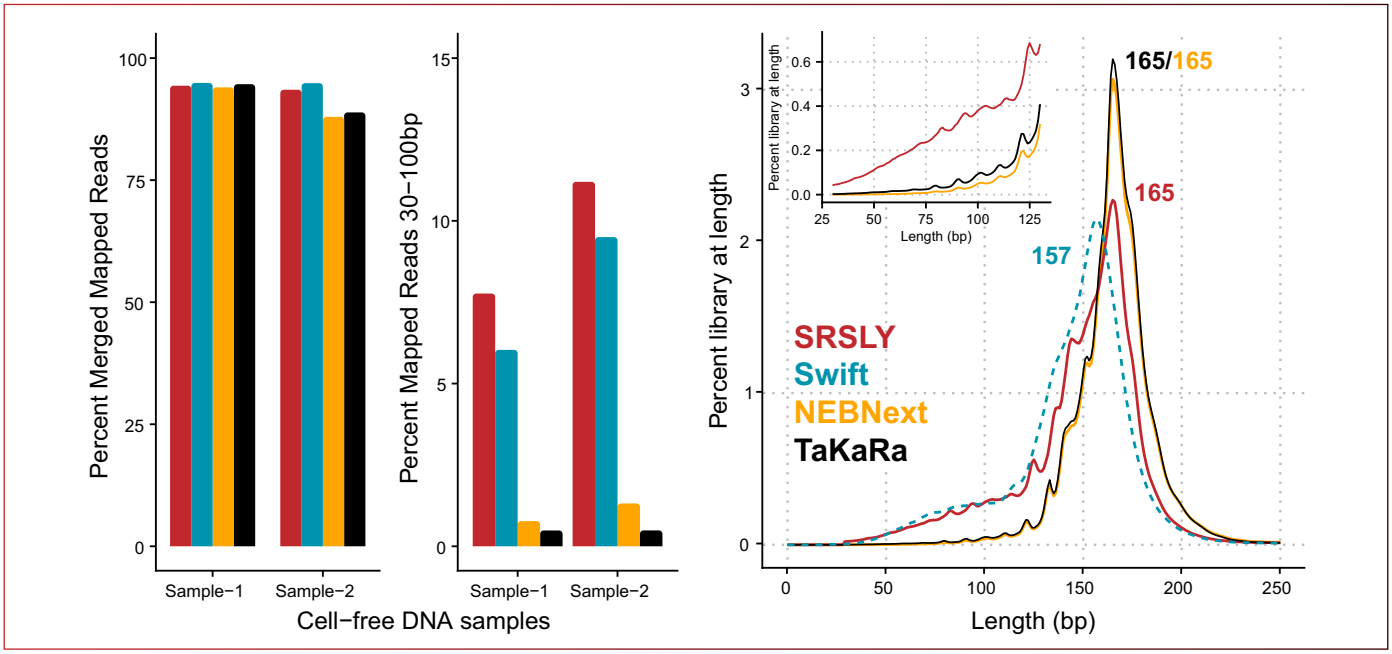

Figure 2. Left: Mapping rates of various library preparation methods. Middle and Right – cfDNA fragment size distribution showing retention of short fragment by single stranded methods (SRSLY and Swift). The size distribution for Swift is shifted to the left and loses the helical periodicity due to bioinformatic removal of terminal bases.

commercially available single-stranded kit – Accel-NGS 1S Plus kit (Swift Biosciences™) has demonstrated an improved capability in capturing short fragments. However, this method requires downstream bioinformatic processing of sequencing data that obfuscates true cfDNA fragment size and sequence information (Figure 2).

Claret Bioscience has developed a cfDNA library method, SRSLY that outperforms commercially available kits in capturing small fragments while retaining true length and sequence of all fragments.

IN VIVO FRAGMENTATION GENERATES UNIQUE cfDNA TERMINI

Nuclease degradation of gDNA within the cell and in the bloodstream can manifest as cfDNA fragments that harbor three types of DNA termini – 3' or 5' single-stranded overhangs (which range from 1 to several nucleotides in length) and blunt ends. cfDNA fragment overhang features may contain crucial information about the nature of cell-death mechanisms contributing to cfDNA generation; different DNA termini have been observed in apoptosis and necrosis[13]. Fragment ends can also reveal differences in underlying nuclease identity, activity or expression. The end polishing step prerequisite in traditional library methods converts all cfDNA fragments to blunt ended molecules by filling 5' ends and degrading 3' ends. The resulting sequencing reads are not representative of the original molecules. These artefactual blunt ends result in reads that are reverse complementary to each other. A method such as SRSLY, that retains all cfDNA molecules yet also captures the variation in overhang length and composition retains the original DNA fragment, from the first base to the last. Because of this feature, it is possible to analyze the 5' and 3' ends separately. Traditional double-stranded library preparation methods with end-polishing compromise the ability to discriminate between the base composition specific to the 5' and 3' ends as shown in Figure 3.

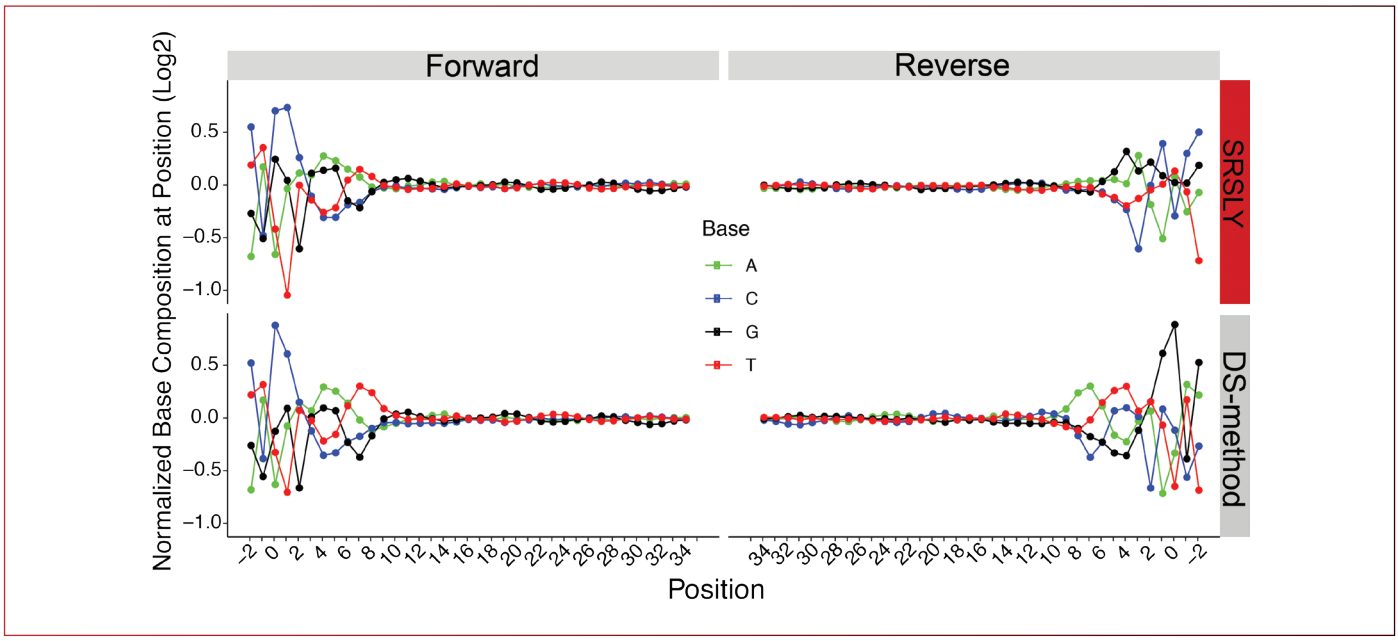

Figure 3. Base composition of ends of forward and reverse reads obtained from libraries generated with a double-stranded approach and SRSLY.

SINGLE-STRANDED APPROACHES REVEAL ACCURATE cfDNA DINUCLEOTIDE COMPOSITION

An oscillating pattern of A/T-rich and G/C-depleted regions followed by a G/C-rich and A/T-depleted region is expected near the labile regions of nucleosome-protected DNA[4,14]. The dinucleotide composition of cfDNA sequences that center around ~167 bp, i.e. the most common nucleosome-protected size, shows differences in the profiles obtained by double-stranded and single-stranded library preparation approaches. For double-stranded methods, end-polishing causes both fragment ends to be mirror images of each other. However, presumably due to the presence of diverse single-stranded overhangs at the 3' termini, distinct dinucleotide frequency patterns are obtained for 5' versus 3' termini using SRSLY. This loss of signal with double-stranded methods compromises the true sequence information of the fragment and consequently the downstream analyses. (Figure 4).

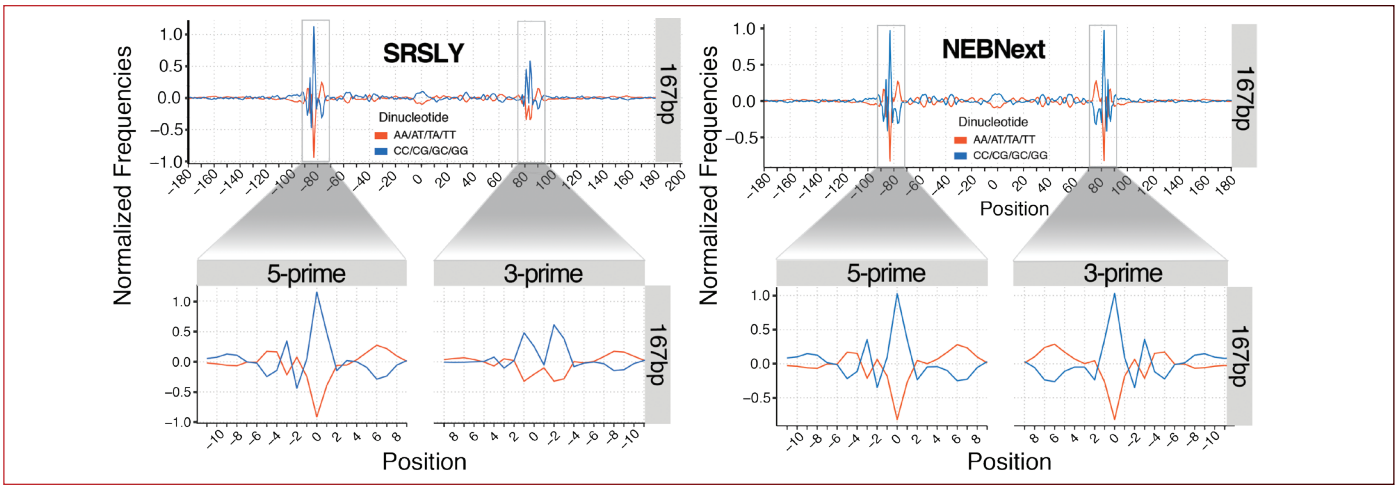

Figure 4. Dinucleotide frequency for a 100 or 11bp window centered around 5-prime and 3-prime fragmentation point for the fragment size with maximum abundance (~167bp), including 100 and 11 bp of genomic context.

CONCLUSION

Single-stranded library preparation methods like SRSLY are ideal for retention of sequence information along the length of input DNA fragments, including the termini. The exact 5' and 3' end points of each input DNA fragment are retained in an efficient and simple library preparation. In liquid biopsy, the information harbored in cfDNA ends can be used as an additional feature in machine learning approaches designed for the identification of tumor-derived molecules or patterns specific to cancer and consequently improve the accuracy of such tools.

Deconvolution of the signal obtained from cfDNA relies on publicly available databases of tissue- or tumor-specific nucleosome position profiles, transcription binding profiles and open chromatin regions. It must be noted that these databases were often generated using NGS methods which themselves required end-polishing and therefore inadvertently lost positioning accuracy. Apart from its utility in cfDNA fragmentomics, SRSLY and other methods that retain full sequence information can also be used in generating more accurate data to improve the performance of machine-learning based prediction tools.

REFERENCES

1. Schwarzenbach H, Hoon DS, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients. Nat Rev Cancer. 2011;11(6):426-437.
2. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proceedings of the National Academy of Sciences of the United States of America. 2008;105(42):16266-16271.
3. Mouliere F, Robert B, Arnau Peyrotte E, et al. High fragmentation characterizes tumour-derived circulating DNA. PloS one. 2011;6(9):e23418.
4. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell. 2016;164(1-2):57-68.
5. Burnham P, Kim MS, Agbor-Enoh S, et al. Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. Sci Rep. 2016;6:27859.
6. Frenkel ZM, Trifonov EN, Volkovich Z, Bettecken T. Nucleosome positioning patterns derived from human apoptotic nucleosomes. J Biomol Struct Dyn. 2011;29(3):577-583.
7. Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U. Predicting cell-type-specific gene expression from regions of open chromatin. Genome Res. 2012;22(9):1711-1722.
8. Cristiano S, Leal A, Phallen J, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. 2019;570(7761):385-389.
9. Sun K, Jiang P, Cheng SH, et al. Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. Genome Res. 2019;29(3):418-427.
10. Sanchez C, Snyder MW, Tanos R, Shendure J, Thierry AR. New insights into structural features and optimal detection of circulating tumor DNA determined by single-strand DNA analysis. NPJ Genom Med. 2018;3:31.
11. Gansauge MT, Gerber T, Glocke I, et al. Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. Nucleic acids research. 2017;45(10):e79.
12. Wu DC, Lambowitz AM. Facile single-stranded DNA sequencing of human plasma DNA via thermostable group II intron reverse transcriptase template switching. Sci Rep. 2017;7(1):8421.
13. Didenko VV, Ngo H, Baskin DS. Early necrotic DNA degradation: presence of blunt-ended DNA breaks, 3' and 5' overhangs in apoptosis, but only 5' overhangs in early necrosis. Am J Pathol. 2003;162(5):1571-1578.
14. Gaffney DJ, McVicker G, Pai AA, et al. Controls of nucleosome positioning in the human genome. PLoS genetics. 2012;8(11):e1003036.

P.O. Box 3052 | Santa Cruz, CA 95063 | +1.833.425.2738 | www.ClaretBio.com | info@ClaretBio.com