Received: 30 June 2021 | Revised: 13 July 2021 | Accepted: 14 July 2021

DOI: 10.14814/phy2.14993

ORIGINAL ARTICLE

Physiological Reports

Whole genome sequencing of low input circulating cell-free DNA obtained from normal human subjects

Julie F. Foley[1] | **Brian Elgart**[3] | **B. Alex Merrick**[1] | **Dhiral P. Phadke**[3] | **Molly E. Cook**[2] | **Jason A. Malphurs**[2] | **Gregory G. Solomon**[2] | **Ruchir R. Shah**[3] | **Michael B. Fessler**[2] | **Frederick W. Miller**[2] | **Kevin E. Gerrish**[2]

[1]Division of National Toxicology Program, NIEHS, Durham, North Carolina, USA

[2]Division of Intramural Research, NIEHS, Durham, North Carolina, USA

[3]Sciome, LLC, Durham, North Carolina, USA

Correspondence
Julie F. Foley, NIEHS, 111. T.W. Alexander Dr., Research Triangle Park, NC 27709, USA.
Email: foley1@nih.gov

Funding information
NIH; National Institute of Environmental Health Sciences, Grant/Award Number: ES-103318-05

Abstract

Cell-free DNA circulates in plasma at low levels as a normal by-product of cellular apoptosis. Multiple clinical pathologies, as well as environmental stressors can lead to increased circulating cell-free DNA (ccfDNA) levels. Plasma DNA studies frequently employ targeted amplicon deep sequencing platforms due to limited concentrations (ng/ml) of ccfDNA in the blood. Here, we report whole genome sequencing (WGS) and read distribution across chromosomes of ccfDNA extracted from two human plasma samples from normal, healthy subjects, representative of limited clinical samples at <1 ml. Amplification was sufficiently robust with ~90% of the reference genome (GRCh38.p2) exhibiting 10X coverage. Chromosome read coverage was uniform and directly proportional to the number of reads for each chromosome across both samples. Almost 99% of the identified genomic sequence variants were known annotated dbSNP variants in the hg38 reference genome. A high prevalence of C>T and T>C mutations was present along with a strong concordance of variants shared between the germline genome databases; gnomAD (81.1%) and the 1000 Genome Project (93.6%). This study demonstrates isolation and amplification procedures from low input ccfDNA samples that can detect sequence variants across the whole genome from amplified human plasma ccfDNA that can translate to multiple clinical research disciplines.

KEYWORDS

circulating cell-free DNA, genomic sequencing, plasma DNA, variants

1 | INTRODUCTION

Short extracellular DNA fragments (~170 bp) normally circulate in blood at low concentrations (ng/ml). Programmed cell death in hematopoietic cells and tissues (normal turnover) is the primary source of extracellular nuclear and mitochondrial DNA (mtDNA) in whole blood. Circulating cell-free DNA was first described in the late 1940's (Mandel & Metais, 1948). However, in the past decade, a growing number of investigators have recognized quantitation and sequence analysis of ccfDNA as an emerging,

blood-based biomarker for disease diagnosis, staging, and therapeutic decisions in oncology (Crowley et al., 2013; Otandault et al., 2019; Suraj et al., 2016), fetal aneuploidy disorders (Breveglieri et al., 2019; Zhang et al., 2019), organ transplantation (Hayward & Chitty, 2018; Oellerich et al., 2015), autoimmune disorders (Duvvuri & Lood, 2019; Tug et al., 2014), and other pathologies (Cerne & Bajalo, 2014; Haghiac et al., 2012; Hamaguchi et al., 2015). Since ccfDNA is limited in quantity and heterogeneous in size (multiples of ~170 bp), concentrations can be highly variable, and even with the use of successful downstream methods, detecting sequence variants at a whole genome level can be challenging.

Targeted amplicon sequencing has been the preferred DNA sequencing application because when compared to tissue genomic DNA, extracted ccfDNA concentrations are low and not sufficient for WGS (Plagnol et al., 2018; Zakrzewski et al., 2019; Zhang et al., 2019). However, a targeted approach precludes gene discovery and constrains variant detection to a select few candidate genes for tissue-specific diseases, primarily cancer. Here, we report procedures for isolation, amplification, and sequencing of plasma ccfDNA from two normal human plasma donors followed by the analysis of ccfDNA yield and quality performance metrics. These procedures should not be limited to WGS and allow for exome sequencing and candidate gene enrichment. Population databases such as COSMIC, Catalogs of Mutations in Cancer (COSMIC, Catalogue of Somatic Mutations in Cancer http://cancer.sanger.ac.uk/cosmic, Accessed 22 Aug 2017) enable comparisons of ccfDNA sequencing data against curated mutations for disease-specific molecular signatures. Here, we describe procedures for the extraction of ccfDNA from human plasma, sequencing of the whole genome, and analysis for genomic alterations. Development of these isolation and amplification procedures from low input ccfDNA is critical when clinical plasma volume samples are limiting to 1 ml or less.

2 | MATERIALS AND METHODS

2.1 | Sample collection

Circulating cfDNA was extracted from healthy, de-identified volunteers recruited from the NIEHS Clinical Research Unit (Research Triangle Park). The Institutional Review Board of NIEHS approved the protocol for this study and methodologies conformed to the standards set by the Declaration of Helsinki.

Peripheral whole blood was collected from two healthy male volunteers in Streck Cell-Free DNA BCT® tubes (La Vista), and the plasma was separated within 2 h of collection by a double centrifugation protocol. The first spin at 1600× g was performed at room temperature for 10 min in a mega-centrifuge (Sorvall RT7, Thermo Fisher Scientific). Following careful plasma separation, centrifugation was repeated at 16,000× g for 10 min in a benchtop microcentrifuge (Eppendorf 5415D, Hauppauge), and the clean plasma was transferred to a cryovial for storage at −80°C until ccfDNA isolation.

2.2 | DNA extraction, quantification, and fragment analyses

Circulating cfDNA was extracted from 1 ml of plasma using a magnetic bead-based kit (Maxwell® RSC ccfDNA Plasma Kit, Promega) and eluted with nuclease-free water (50 µl volume) on a Promega Maxwell® RSC instrument. Sample yield was determined by a fluorometric method (Quantus™ Fluorometer, Promega). Fragmentation pattern analysis (FA) was used to assess the plasma DNA quality to verify the presence of characteristic ccfDNA short bp fragments of ~170 bp and the absence of genomic contamination (5200 Fragment Analyzer System, Agilent Inc.). Purified samples were stored at −80°C.

2.3 | Library preparation and sequencing

Amplification was required to generate sufficient amounts of DNA for WGS. The initial step of library preparation typically begins with DNA shearing. Since ccfDNA is comprised of short base pair fragments (<200), it was not necessary to perform this step. Amplified sequencing libraries were prepared using a three-step library preparation protocol for sequencing plasma ccfDNA (Table 1). Briefly, 10 µl of eluted ccfDNA at 0.1 ng/µl was used for template preparation and library synthesis. Amplification was performed in a single tube with the SMARTer ThruPLEX® Plasma Seq Library Preparation protocol according to

TABLE 1 Plasma ccfDNA sample summary

Sample	Pre-amplification concentration (ng/ul)	ccfDNA input	PCR cycles	Post-amplification clean up concentration (ng/uL)	Library size	Fold enrichment
1	0.1	1 ng	14	13	340 bp	650
2	0.1	1 ng	14	11	340 bp	550

the manufacturer's instructions (Takara Bio USA). NGS libraries were purified with Agencourt AMPure XP beads (Beckman Coulter Genomics), then amplified by a round of PCR (14 cycles) and tagged with Illumina-compatible indexes for multiplex sequencing (SMARTer® DNA HT Dual Index Kit, Takara Bio USA). Fragment analysis of the libraries for each sample showed a median sample size of 310 bp (170 bp ccfDNA plus 140 bp adapters) containing library products of the mononucleosomal DNA fragments. Libraries were sequenced on a NovaSeq™ 6000 (Illumina) S1 platform by the NIEHS Epigenomics Sequencing Core (Research Triangle Park) at an average of 20X coverage from 2 × 150 bp paired-end reads. Sequences were output in FASTQ format. Sequencing reads for all samples have been deposited in NCBI SRA (SUB9071852).

2.4 | Bioinformatics analysis

Sequence data quality was evaluated by the FASTQC tool v0.11.8 (www.bioinformatics.babraham.ac.uk/projects/fastqc/). Genome Analysis Toolkit (GATK) workflow describes the data pre-processing procedures and identification of germline short variants (single nucleotide polymorphisms [SNPs] and indels). Read pairs were mapped to the human reference genome (GRCh38.p2) using the BWA alignment tool (v0.7.12) (Li & Durbin, 2009). Duplicate reads were removed using the MarkDuplicates program from Picard tools v1.99 (http://broadinstitute/github/io/picard). Aligned read numbers were obtained after duplicate reads and reads aligning on two different chromosomes were removed. Utilities from the BEDTools package and custom summarization scripts were used to obtain the coverage at each hg38 genome base (McKenna et al., 2010). Genomic coverage was summarized at the following different levels: 1X, 10X, 20X, 30X, 50X, and 100X. To identify the presence of bias during the sequencing workflow, percent aligned reads were normalized by chromosome length. In addition, using an in silico approach, coverage was also evaluated specifically at all coding regions as annotated for a commercial human whole exome sequencing library (Agilent SureSelect® Human All Exon V7 probe design). Exon probes targeted approximately 214,000 coding exons included in RefSeq (99.3%), GENCODE v24 (99.6%), CCDS (99.6%), and UCSC databases for known genes (99.6%).

Tools from GATK version 4.1.2 were used for data pre-processing, variant calling, and filtering of variants. Base quality scores were recalibrated using the GATK tools, BaseRecalibrator, and ApplyBQSR. Identification of germline variants was performed using GATK's HaplotypeCaller (Poplin et al., 2017). Raw variants were filtered using GATK tools VariantRecalibrator and ApplyVQSR (Depristo et al., 2011; Van der et al. 2013). Variants in hg38 repeat regions or blacklisted regions were removed. Functional annotation of the variants was carried out using Snpeff version 4.3t (Cingolani et al., 2012).

Additional analyses were conducted to further characterize the sample variants. We excluded the low impact variants and then generated a mutation frequency spectrum. The analysis was performed with the R package, SomaticSignatures (v3.6; Bioconductor) (Gehring et al., 2015). Finally, we compared the identified variants with two human germline variant databases aligned against the reference genome GRCh38.p2: genome Aggregation Database, gnomAD, (v3.1.1, Broad), (Karczewski et al., 2020), and The International Genome Sample Resource, 1000 Genomes Project (Clarke et al., 2017).

3 | RESULTS

The human genome was sequenced after the amplification of ccfDNA (1ng) using a sequencing library preparation procedure specifically designed for circulating short ccfDNA bp fragments extracted from plasma. Amplified fragments were sequenced on an Illumina NovaSeq™ 6000, S1 Sequencing System (San Diego, CA) from two unique human blood-based samples.

3.1 | ccfDNA yield

Amplification was performed in a single tube to reduce the possibility of contamination and loss of ccfDNA. Following the manufacturer's recommendation (Takara Bio) of 14 cycles for low input ccfDNA concentrations (ng/μl), we observed a ccfDNA enrichment of greater than 500-fold (Table 1). Post-amplification fragment analysis demonstrated 340 bp library size without contamination for WGS.

3.2 | Performance metrics

Paired-end sequencing was performed and aligned across the hg38 genome build. The mean number of sequencing reads was 577.7 million. Aligned mapped reads for samples 1 and 2 were 53% and 73% to the GRCh38.p2 reference genome (Table 2). Decreased alignment observed in sample 1 was attributed to artifact of unknown origin that compromised the sample. A PCR bubble present in the FA of sample 1 indicated slight over amplification that likely accounts for the increased read duplicate numbers for that sample. A reasonable duplication rate (25%) was observed in the other sample. After removal of the duplicate reads,

TABLE 2 Summary performance statistics for WGS from amplified ccfDNA

Sample ID	Read length	Total reads (Million)	Aligned reads	Alignment rate	Duplicate reads
1	PE−150[a]	857.9	456.7	53%	45%
2	PE−150[a]	955.4	699.3	73%	25%

[a]PE-150: Paired-end, 150 base pairs

and reads aligning on two different chromosomes, a reliable number of reads was achieved for alignment to the reference genome for both samples.

Genomic coverage was assessed for each sample to determine the capture sensitivity. We analyzed genomic coverage at 1X, 10X, 20X, 30X, 40X, 50X, 75X, and 100X to examine what percentage of the genomic positions is adequately included at varying coverage thresholds (Figure 1). Accurate base sequencing calls are typically based on a minimum 10- to 20X-fold depth of coverage (Lelieveld et al., 2015; Parla et al., 2011). Our results indicate that ~90% of the hg38 bases were covered with 10X or higher coverage in each sample. Competent base pair read coverage was demonstrated at 20X. Read coverage for sample 1 dropped to 53% but was acceptable with ~250 million deduplicated reads. Sample 2 retained confident coverage with 84% read coverage at 20X. Regions of the genome not covered by any reads were minimal with 6% of the bases not covered.

The Reads Per Million (RPM) normalized signal at various genes were reviewed to assess coverage consensus between the two samples (Figure 2). Coverage patterns for both samples paralleled each other. Figure 2 demonstrates raw base read coverage patterns along with normalized read scores for the housekeeping genes, albumin, (*ALB*) and beta-2-microglobulin, (*B2M*). Sample normalized read values were similar and consistent for each gene. *ALB* displayed high coverage with a normalized read score of ~0.01 RPM. In contrast to *ALB*, low base pair read coverage was observed for *B2M* and reflected in the similar, decreased normalized read scores (0.04 RPM).

We examined the distribution of coverage across each chromosome to determine the presence of sequencing bias by looking at the percentage of aligned reads normalized by chromosome length. For each sample, proportional to the number of aligned reads, there is uniform coverage of each chromosome across the genome (Figure 3). Of the five acrocentric chromosomes (Cingolani et al., 2012; Gehring et al., 2015; Hamaguchi et al., 2015; Plagnol et al., 2018; Zakrzewski et al., 2019), coverage was decreased for chromosomes 13–15. Annotated clone assembly problems for GRCh38.p2 have been described for the short arm regions of these chromosomes since they are heterochromatic and contain families of repeated sequences including ribosomal RNA gene arrays. The long

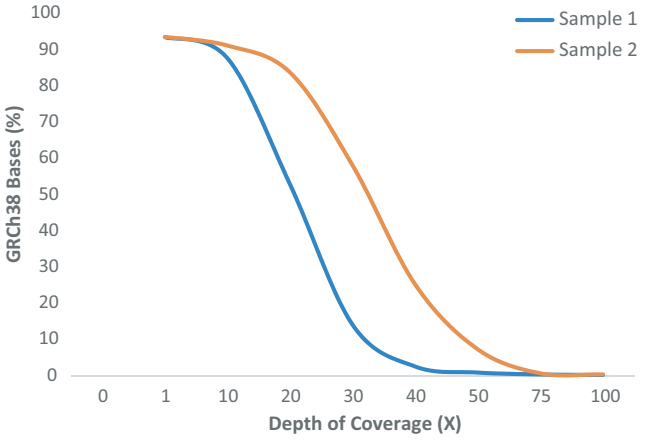

FIGURE 1 Coverage on the hg38 reference genome. The percentage of bases covered in the human GRCh38.p2 reference genome is shown at 1X, 10X, 20X, 30X, 50X, and 100X depth of coverage. Testing at 10X and 20X demonstrated sufficient coverage of the reference genome for accurate future downstream variant calls

arm is euchromatic and contains the protein-coding genes of the chromosome (Dunham et al., 2004; Shepelev et al., 2015). Coverage for the Y-sex chromosome is uniform, although chromosome coverage is substantially lower due to a sequencing gap in the reference genome. The majority (41 Mb) of the Y chromosome (63 Mb) is comprised of three blocks of highly reiterated satellites as well as other repeat sequences which complicates short read alignment (Kirsch et al., 2005).

Variants identified in repeat or blacklisted regions of the human genome were excluded from the analysis. Known variant calls (2,106,417 SNPs and indels) based on dbSNP build 151 (Sherry et al., 2001) accounted for 98.8% of the variants. The remaining 1.2% (25,277 SNPs and indels) were classified as "novel" with 0.002% identified as synonymous or non-synonymous, missense mutations. Indels and non-synonymous frameshift mutations accounted for the remaining novel variants.

WGS of amplified ccfDNA effectively accounted for all protein coding regions of the genome. We matched the coding exon sequences from this study to the well annotated Agilent SureSelect® Human All Exon V7 probe design. Exon probes target coding regions from RefSeq (99.3%), GENCODE v24 (99.6%), CCDS (99.6%), and UCSC known genes (99.6%) (https://www.agilent.com/

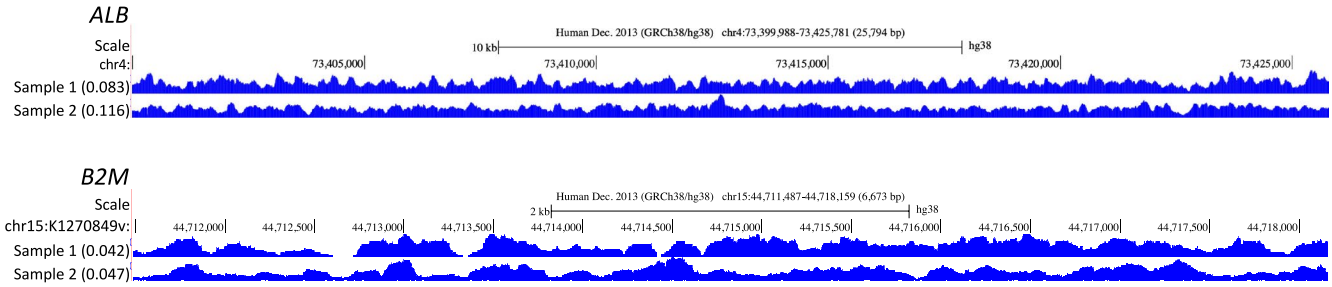

FIGURE 2 Parallel sample coverage of RPM normalized signal for the housekeeping genes, *ALB* and *B2M*

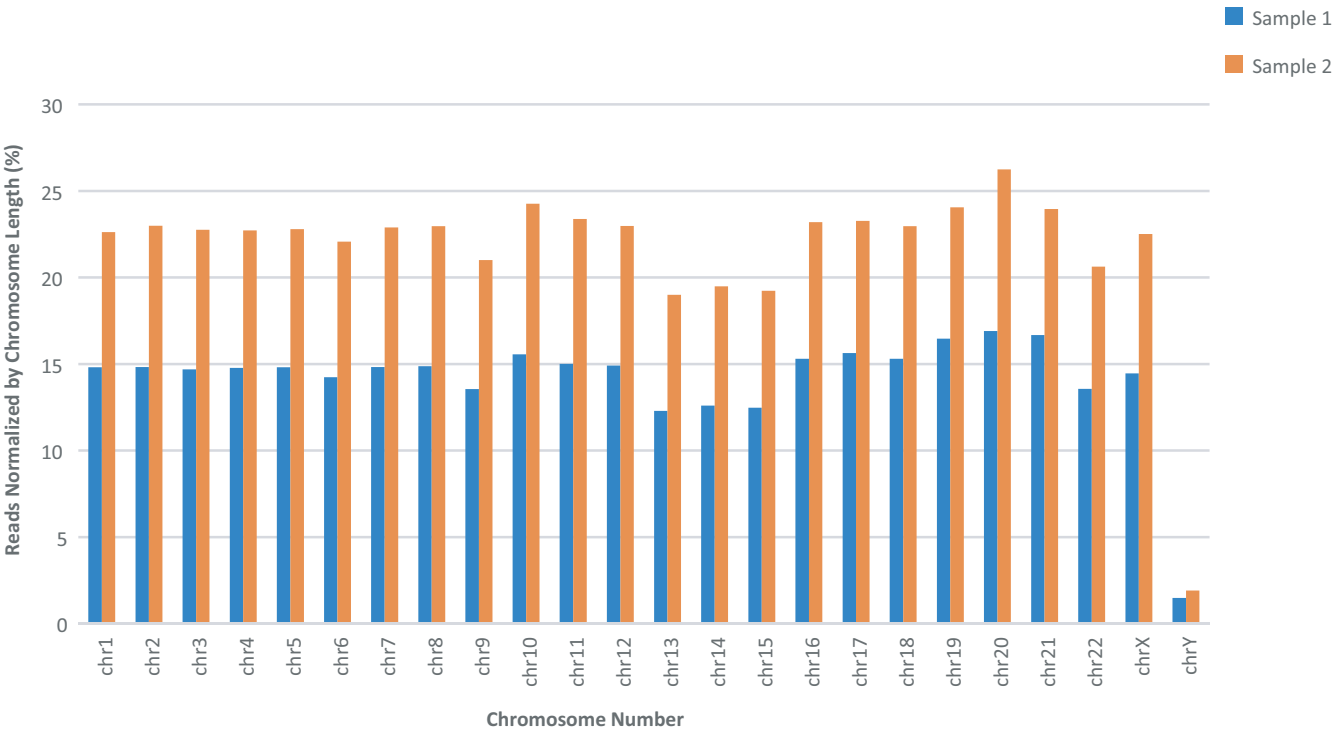

FIGURE 3 Distribution of coverage across each chromosome to determine the evidence of sequence bias. The percentage of aligned reads was normalized by chromosome length. For each sample, there is uniform coverage of each chromosome across the genome. Variant calls were performed using the GATK Pipeline (Broad)

cs/library/datasheets/public/5991-9040EN%20SureSelect%20V7%20Datasheet.pdf). Capture sensitivity was consistent across both samples (Figure 4). At a minimum coverage depth of 10X, >98% of the protein coding exon bases were sequenced for all samples. With increasing stringency at 20X, coverage decreased for both samples; however, sample 2 retained sufficient sensitivity (65%) to obtain valuable sequencing information. Protein coding exon bases not covered by any reads were negligible with less than 1% non-covered bases.

We further characterized the final genomic variants where low impact variants were excluded with respect to the reference genome, GRCh38.p2, and generated a mutational frequency plot. A high percentage of C>T (39.9%) and T>C mutations (38.4%) was observed, Figure 5.

Spontaneous, 5-methylcytosine deamination in resting cells creates point mutations from purine mismatching and leads to the frequently observed C>T transition mutations (Mustjoki & Young, 2021). We computed the frequency of each mutation in other known germline databases such as dbSNP (ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606/VCF/GATK/All_20180418.vcf), HapMap and 1000 Genomes (https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0/), and found that the percentage of C>T and T>C mutations in these resources matched with our study findings (Clarke et al., 2017; Gibbs et al., 2003).

In our final analysis, the final variant set was compared to two additional germline variant resource databases, gnomAD and 1000 Genomes Project (Clarke et al., 2017;

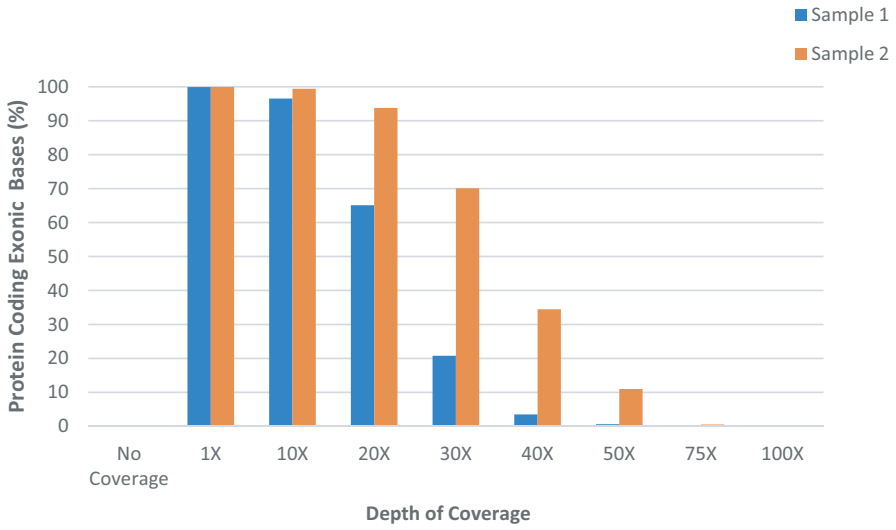

FIGURE 4 Coverage in protein coding exon bases from Agilent SureSelect® Human All Exon V7 probe design. The percentage of protein coding exon bases is shown at 1X, 10X, 20X, 30X, 50X, and 100X depth of coverage. Depth of coverage at 20X demonstrated sufficient coverage of the exonic bases for accurate, future downstream variant calls

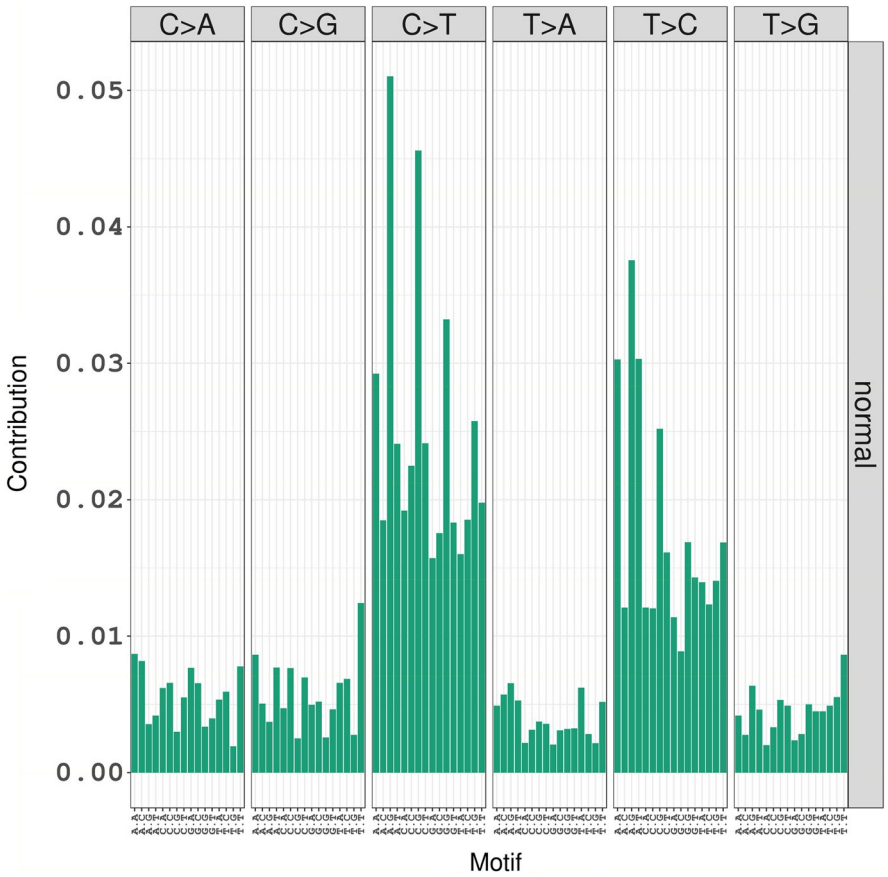

FIGURE 5 Mutational spectrum of the final variant set after excluding the low impact variants. Approximately two thirds of the variants were C>T and T>C mutations

Karczewski et al., 2020). When we compared our final variant set, concordance of 81.1% and 93.6% was found between our data set and that of the respective germline database. The high degree of overlap of shared variants between two germline genome resources demonstrates the potential capability of identifying variants from low input ccfDNA by WGS.

4 | DISCUSSION

Clinical oncology research has recently been transformed by technological advancements to isolate, quantify, and sequence released cell-free nucleic acids (DNA, RNA, and mtDNA) from a blood "liquid biopsy" (e.g., non-small cell lung carcinoma) (Chen & Zhao, 2019; Malapelle et al.,

2017; Plagnol et al., 2018). Since the half-life of ccfDNA is between 16 min and 2.5 h, ccfDNA can mirror disease status in real time (Diehl et al., 2008). This carries the promise for future individualized diagnostic, prognostic, and therapeutic cancer applications, but technical challenges remain. Circulating cell-free DNA concentrations are often inadequate in the absence of active disease or other factors that typically increase ccfDNA levels, and this can be further complicated by small volume plasma collections (≤ 4 ml) in clinical samples (Alborelli et al., 2019). As ccfDNA concentrations correlate with tumor size and stage, typically being lowest in patients with small tumors (Chen & Zhao, 2019), technical advances will be required for ccfDNA analysis to be useful in early stage cancer and other preclinical conditions. As a biomarker, ccfDNA concentrations have been disease informative, but downstream sequencing analysis could expand the utility of this molecule as a biomarker. In this study, we successfully isolated, amplified, and performed whole genome sequencing from low input amounts of plasma ccfDNA, thereby optimizing the methodology that may extend the range of ccfDNA analysis, and, thereby, its potential clinical and research utility.

Next-generation sequencing in the form of targeted amplicon sequencing has been the primary approach to identify cancer-related mutations from genomic DNA and ccfDNA samples. In the field of oncology, targeted amplicon sequencing has been extensively used for hereditary cancer screening, disease recurrence detection, or determining a therapeutic response (Marrugo-Ramírez et al., 2018; Russano et al., 2020). Due to insufficient starting amounts of ccfDNA from vascular liquid biopsies, a limited number of studies have attempted whole genome or whole exome sequencing on these sample types. However, improvements with library preparation protocols for highly fragmented DNA samples have made WGS possible. We applied a library preparation method with reformulated repair and ligation reactions specific to plasma ccfDNA to successfully amplify and sequence the whole genome. Generation of blunt end fragments, followed by ligation and then extension, cleavage, and amplification were performed in a single tube to minimize the ccfDNA loss. We found base pair coverage was uniform across all chromosomes with enough reads for reliable variant calling, even though duplication rates were increased in one sample. Not unexpected, duplication rates do vary among samples as was found in this study. We observed a relatively high duplication rate (45%) for sample 1. Either low ccfDNA sample input (0.1 ng/μl in 10 μl), low library complexity, or the possibility of variance in ccfDNA fragment size with over representation of smaller fragments following PCR amplification were possible factors explaining the observed increased duplication rate. Samples were handled identically (duplication rate ≤30%) and fragment analysis indicated no differences in either sample quality until post-amplification. By expanding the sequencing capabilities from gene targeting to include whole genome and whole exome sequencing, there is an acceleration of opportunities for disease discovery.

Few WGS studies have been conducted on clinical ccfDNA samples. Coverage of the genome at 10X and 20X was comparable to or better than other reports for WGS of ccfDNA (Wang et al., 2017). In a tumor ccfDNA matched study, Ma et al. (2017) validated that WGS of ccfDNA with low average coverage depth (~10X) was sufficient to detect variants from late stage lung and colon cancer samples with confidence (Ma et al., 2017). Our data at 20X coverage supports the reliability of sequenced calls from normal subjects with ~99% of the called variants present in dbSNP. The remaining variants absent in dbSNP are of unknown/uncertain significance (VUS, variants of unknown certainty). These variants could be attributed to clonal hematopoiesis and aging (Zink et al., 2017). Hematopoietic clone-derived mutations including driver and passenger mutations are prevalent in ccfDNA of healthy individuals. VUS may be pathogenic or protective, an area that needs more investigation (Oulas et al., 2019). Those identified in this study will be followed up in subsequent work.

Many of the same single nucleotide variants are shared between germline and somatic mutation databases (Meyerson et al., 2020). Meyerson et al. (2020) found after strict filtering to exclude common germline polymorphisms and poor coverage or mapability sites, 336,987 common variants between the gnomAD germline database and the TCGA cancer genome database. They concluded that shared variants depict true biological occurrences of the same variant in the germline and somatic setting and arise primarily because DNA has some of the same basic chemical vulnerabilities in either setting. This helps to explain the mutation frequency pattern we observed from our variant data. A higher frequency of C>T and T>C transition mutations was present in the ccfDNA from the two normal liquid biopsy samples. This frequency has been reported frequently in the literature and resembles the COSMIC database clock-like mutation signature, SBS5. Fiala and Diamandis (2020) reviewed mutations present in normal tissues and venous liquid biopsies in the context of developing specific ctDNA test for cancer, evaluating clonal hematopoiesis, cardiovascular pathology, and neural mosaicism in Alzheimer's disease (Fiala & Diamandis, 2020). In this review paper, work by Lodato et al. (2018), delineated three mutational signatures from single cell neurons isolated from the prefrontal cortex and hippocampus brain regions from healthy subjects and those diagnosed with early onset, hereditary neurodegenerative disorders, Cockayne syndrome, and Xeroderma pigmentosum

(Lodato et al., 2018). The first signature, characterized by C>T and T>C mutations, increased with age, in a clock-like fashion (Fiala & Diamandis, 2020; Lodato et al., 2018). The cancer genome database, COSMIC also reports this pattern as signature, SBS5 where approximately two thirds of the mutations are C>T and T>C. Even though COSMIC is a cancer genome database, the signature is described as clock-like in that the number of mutations in most cancers and normal cells correlates with the age of the individual (https://cancer.sanger.ac.uk/signatures/sbs/sbs5/?genome=GRCh38). Our data from low input, plasma ccfDNA samples, directly correlate with germline resources and the well-described COSMIC SBS5 signature for normal and cancer-derived tissue.

Library preparation is a critical step in NGS sequencing and has a direct impact on the quality of sequencing results. Utilizing library preparation procedures specific to the amplification of highly fragmented plasma DNA, we successfully amplified ccfDNA and sequenced the whole human genome with uniform base pair distribution across each chromosome. Future work will continue to improve and refine the technique for clinical studies, and to translate the technique to exposure science, including both human and preclinical experimental animal model systems.

In summary, our results demonstrate that reliable data from WGS of low input clinical, ccfDNA samples can be used for biomarker and sequence variant discovery that can be applied to oncology, inflammation, transplantation, maternal/fetal disease, and other pathologies.

ACKNOWLEDGMENTS

The authors express their gratitude to Dr. Stephanie Smith-Roe and Dr. Erik Tokar for review of the manuscript. This research was supported by the NIH, National Institute of Environmental Health Sciences through Intramural Research Program Project ES-103318-05.

CONFLICT OF INTERESTS

JFF, BE, BAM, KEG, MEC, JAM, GGS, MBF, and FWM are employees of the National Institute of Environmental Health Sciences (NIEHS) and National Institutes of Health (NIH), of the United States Government. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of NIEHS, NIH, or the United States Government. DPP and RRS are Bioinformaticians at Sciome, LLC and performed WGS analysis and annotation under NIEHS contract HHSN273201700001C under NIEHS supervision. The NIEHS Epigenomics core facility (GGS) performed the genome sequencing. Sciome did not contribute to financial funding for publication of this work. The authors adhere to policies on public sharing of data and materials.

AUTHORS' CONTRIBUTIONS

JFF, BE, BAM, and KEG made substantial contributions to the conception, design, and acquisition of data for the study. DPP and RSS conducted the bioinformatic analysis. MEC, JAM, and GGS provided NGS services through the NIEHS Epigenomics Core Laboratory. DPP, JFF, BE, BAM, and KEG contributed to the WGS data analysis and interpretation. MBF and FWM provided scientific study design input and review of the manuscript. All the co-authors gave final approval for publication of the final version of the manuscript.

ORCID

Julie F. Foley https://orcid.org/0000-0001-9726-2821

REFERENCES

Alborelli, I., Generali, D., Jermann, P., Cappelletti, M. R., Ferrero, G., Scaggiante, B., Bortul, M., Zanconati, F., Nicolet, S., Haegele, J., Bubendorf, L., Aceto, N., Scaltriti, M., Mucci, G., Quagliata, L., & Novelli, G. (2019). Cell-free DNA analysis in healthy individuals by next-generation sequencing: A proof of concept and technical validation study. *Cell Death & Disease*, *10*, https://doi.org/10.1038/s41419-019-1770-3

Breveglieri, G., D'Aversa, E., Finotti, A., & Borgatti, M. (2019). Non-invasive prenatal testing using fetal DNA. *Molecular Diagnosis & Therapy*, *23*, 291–299. https://doi.org/10.1007/s40291-019-00385-2

Cerne, D., & Bajalo, J. (2014). Cell-free nucleic acids as a non-invasive route for investigating atherosclerosis. *Current Pharmaceutical Design*, *20*, 5004–5009. https://doi.org/10.2174/1381612819666131206110317

Chen, M., & Zhao, H. (2019). Next-generation sequencing in liquid biopsy: Cancer screening and early detection. *Human Genomics*, *13*, 34. https://doi.org/10.1186/s40246-019-0220-8

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, *6*, 80–92. https://doi.org/10.4161/fly.19695

Clarke, L., Fairley, S., Zheng-Bradley, X., Streeter, I., Perry, E., Lowy, E., Tassé, A.-M., & Flicek, P. (2017). The International Genome Sample Resource (IGSR): A worldwide collection of genome variation incorporating the 1000 genomes project data. *Nucleic Acids Research*, *45*(D1), D854–D859. https://doi.org/10.1093/nar/gkw829

Crowley, E., Di Nicolantonio, F., Loupakis, F., & Bardelli, A. (2013). Liquid biopsy: Monitoring cancer-genetics in the blood. *Nature Reviews Clinical Oncology*, *10*, 472–484. https://doi.org/10.1038/nrclinonc.2013.110

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*, 491–501. https://doi.org/10.1038/ng.806

Diehl, F., Schmidt, K., Choti, M. A., Romans, K., Goodman, S., Li, M., Thornton, K., Agrawal, N., Sokoll, L., Szabo, S. A., Kinzler, K. W., Vogelstein, B., & Diaz Jr, L. A. (2008). Circulating mutant DNA to assess tumor dynamics. *Nature Medicine*, https://doi.org/10.1038/nm.1789

Dunham, A., Matthews, L. H., Burton, J., Ashurst, J. L., Howe, K. L., Ashcroft, K. J., Beare, D. M., Burford, D. C., Hunt, S. E., Griffiths-Jones, S., Jones, M. C., Keenan, S. J., Oliver, K., Scott, C. E., Ainscough, R., Almeida, J. P., Ambrose, K. D., Andrews, D. T., Ashwell, R. I., ... Ross, M. T. (2004). The DNA sequence and analysis of human chromosome 13. *Nature*, *428*, 522–528. https://doi.org/10.1038/nature02379

Duvvuri, B., & Lood, C. (2019). Cell-free DNA as a biomarker in autoimmune rheumatic diseases. *Frontiers in Immunology*, *10*, 1–21. https://doi.org/10.3389/fimmu.2019.00502

Fiala, C., & Diamandis, E. P. (2020). Mutations in normal tissues—some diagnostic and clinical implications. *BMC Medicine*, *18*, 1–9. https://doi.org/10.1186/s12916-020-01763-y

Gehring, J. S., Fischer, B., Lawrence, M., & Huber, W. (2015). SomaticSignatures: Inferring mutational signatures from single-nucleotide variants. *Bioinformatics*, *31*, 3673–3675. https://doi.org/10.1093/bioinformatics/btv408

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F. L., Yang, H. M., Ch'ang, L. Y., Huang, W., Liu, B., Shen, Y., Tam, P. K. H., Tsui, L. C., Waye, M. M. Y., Wong, J. T. F., Zeng, C. Q., Zhang, Q. R., Chee, M. S., Galver, L. M., Kruglyak, S., ... Tanaka, T. (2003). The international HapMap project. *Nature*, *426*, 789–796. https://doi.org/10.1038/nature02168

Haghiac, M., Vora, N. L., Basu, S., Johnson, K. L., Presley, L., Bianchi, D. W., & De Mouzon, S. H. (2012). Increased death of adipose cells, a path to release cell-free DNA into systemic circulation of obese women. *Obesity*, *20*, 2213–2219. https://doi.org/10.1038/oby.2012.138

Hamaguchi, S., Akeda, Y., Yamamoto, N., Seki, M., Yamamoto, K., Oishi, K. & Tomono, K. (2015). Origin of circulating free DNA in sepsis: Analysis of the CLP mouse model. *Mediators of Inflammation*, *2015*, 1–9. https://doi.org/10.1155/2015/614518

Hayward, J., & Chitty, L. S. (2018). Beyond screening for chromosomal abnormalities: Advances in non-invasive diagnosis of single gene disorders and fetal exome sequencing. *Seminars in Fetal & Neonatal Medicine*, *23*, 94–101. https://doi.org/10.1016/j.siny.2017.12.002

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A. ... MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*, 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kirsch, S., Weiss, B., Miner, T. L., Waterston, R. H., Clark, R. A., Eichler, E. E., Münch, C., Schempp, W., & Rappold, G. (2005). Interchromosomal segmental duplications of the pericentromeric region on the human Y chromosome. *Genome Research*, *15*, 195–204. https://doi.org/10.1101/gr.3302705

Lelieveld, S. H., Spielmann, M., Mundlos, S., Veltman, J. A., & Gilissen, C. (2015). Comparison of exome and genome sequencing technologies for the complete capture of protein-coding regions. *Human Mutation*, *36*, 815–822. https://doi.org/10.1002/humu.22813

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Lodato, M. A., Rodin, R. E., Bohrson, C. L., Coulter, M. E., Barton, A. R., Kwon, M., Sherman, M. A., Vitzthum, C. M., Luquette, L. J., Yandava, C. N., Yang, P., Chittenden, T. W., Hatem, N. E., Ryu, S. C., Woodworth, M. B., Park, P. J., & Walsh, C. A. (2018). Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science*, *359*, 555–559. https://doi.org/10.1126/science.aao4426

Ma, X., Zhu, L., Wu, X., Bao, H., Wang, X., Chang, Z., Shao, Y. W., & Wang, Z. (2017). Cell-free DNA provides a good representation of the tumor genome despite its biased fragmentation patterns. *PLoS One*, *12*, 1–18. https://doi.org/10.1371/journal.pone.0169231

Malapelle, U., Mayo De-Las-Casas, C., Rocco, D., Garzon, M., Pisapia, P., Jordana-Ariza, N., Russo, M., Sgariglia, R., De Luca, C., Pepe, F., Martinez-Bueno, A., Morales-Espinosa, D., González-Cao, M., Karachaliou, N., Viteri Ramirez, S., Bellevicine, C., Molina-Vila, M. A., Rosell, R., & Troncone, G. (2017). Development of a gene panel for next-generation sequencing of clinically relevant mutations in cell-free DNA from cancer patients. *British Journal of Cancer*, *116*, 802–810. https://doi.org/10.1038/bjc.2017.8

Mandel, P., & Metais, P. (1948). Nuclear acids in human blood plasma. *Comptes Rendus Des Seances De La Societe De Biologie Et De Ses Filiales*, *142*, 241–243.

Marrugo-Ramírez, J., Mir, M., & Samitier, J. (2018). Blood-based cancer biomarkers in liquid biopsy: A promising non-invasive alternative to tissue biopsy. *International Journal of Molecular Sciences*, *19*, https://doi.org/10.3390/ijms19102877

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110

Meyerson, W., Leisman, J., Navarro, F. C. P., & Gerstein, M. (2020). Origins and characterization of variants shared between databases of somatic and germline human mutations. *BMC Bioinformatics*, *21*, 1–22. https://doi.org/10.1186/s12859-020-3508-8

Mustjoki, S., & Young, N. S. (2021). Somatic mutations in "Benign" disease. *New England Journal of Medicine*, *384*, 2039–2052. https://doi.org/10.1056/nejmra2101920

Oellerich, M., Walson, P., Beck, J., Schmitz, J., Kollmar, O., & Schütz, E. (2015). Cell free DNA as a marker of transplant graft injury (liquid biopsy). *Therapeutic Drug Monitoring*, *38*(Suppl 1), 1. https://doi.org/10.1097/FTD.0000000000000239

Otandault, A., Anker, P., Al Amir Dache, Z., Guillaumon, V., Meddeb, R., Pastor, B., Pisareva, E., Sanchez, C., Tanos, R., Tousch, G., Schwarzenbach, H., & Thierry, A. (2019). Recent advances in circulating nucleic acids in oncology. *Annals of Oncology*, *30*, 374–384. https://doi.org/10.1093/annonc/mdz031

Oulas, A., Minadakis, G., Zachariou, M., & Spyrou, G. M. (2019). Selecting variants of unknown significance through network-based gene-association significantly improves risk prediction for disease-control cohorts. *Scientific Reports*, *9*, 1–15. https://doi.org/10.1038/s41598-019-39796-w

Parla, J. S., Iossifov, I., Grabill, I., Spector, M. S., Kramer, M., & McCombie, W. R. (2011). A comparative analysis of exome

capture. *Genome Biology*, *12*, R97. https://doi.org/10.1186/gb-2011-12-9-r97

Plagnol, V., Woodhouse, S., Howarth, K., Lensing, S., Smith, M., Epstein, M., Madi, M., Smalley, S., Leroy, C., Hinton, J., de Kievit, F., Musgrave-Brown, E., Herd, C., Baker-Neblett, K., Brennan, W., Dimitrov, P., Campbell, N., Morris, C., Rosenfeld, N., ... Forshew, T. (2018). Analytical validation of a next generation sequencing liquid biopsy assay for high sensitivity broad molecular profiling. *PLoS One*, *13*, 1–18. https://doi.org/10.1371/journal.pone.0193802

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., & Banks, E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, https://doi.org/10.1101/201178

Russano, M., Napolitano, A., Ribelli, G., Iuliani, M., Simonetti, S., Citarella, F., Pantano, F., Dell'Aquila, E., Anesi, C., Silvestris, N., Argentiero, A., Solimando, A. G., Vincenzi, B., Tonini, G., & Santini, D. (2020). Liquid biopsy and tumor heterogeneity in metastatic solid tumors: The potentiality of blood samples. *Journal of Experimental & Clinical Cancer Research*, *39*, 95. https://doi.org/10.1186/s13046-020-01601-2

Shepelev, V. A., Uralsky, L. I., Alexandrov, A. A., Yurov, Y. B., Rogaev, E. I., & Alexandrov, I. A. (2015). Annotation of suprachromosomal families reveals uncommon types of alpha satellite organization in pericentromeric regions of hg38 human genome assembly. *Genomics Data*, *5*, 139–146. https://doi.org/10.1016/j.gdata.2015.05.035

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). DbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311. https://doi.org/10.1093/nar/29.1.308

Suraj, S., Dhar, C., & Srivastava, S. (2016). Circulating nucleic acids: An analysis of their occurrence in malignancies. *Biomed Reports*, *6*, 8–14. https://doi.org/10.3892/br.2016.812

Tug, S., Helmig, S., Menke, J., Zahn, D., Kubiak, T., Schwarting, A., & Simon, P. (2014). Correlation between cell free DNA levels and medical evaluation of disease progression in systemic lupus erythematosus patients. *Cellular Immunology*, *292*, 32–39. https://doi.org/10.1016/j.cellimm.2014.08.002

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*(1), https://doi.org/10.1002/0471250953.bi1110s43

Wang, W., Kong, P., Ma, G., Li, L., Zhu, J., Xia, T., Xie, H., Zhou, W., & Wang, S. (2017). Characterization of the release and biological significance of cell-free DNA from breast cancer cell lines. *Oncotarget*, *8*, 43180–43191. https://doi.org/10.18632/oncotarget.17858

Zakrzewski, F., Gieldon, L., Rump, A., Seifert, M., Grützmann, K., Krüger, A., Loos, S., Zeugner, S., Hackmann, K., Porrmann, J., Wagner, J., Kast, K., Wimberger, P., Baretton, G., Schröck, E., Aust, D., & Klink, B. (2019). Targeted capture-based NGS is superior to multiplex PCR-based NGS for hereditary BRCA1 and BRCA2 gene analysis in FFPE tumor samples. *BMC Cancer*, *19*, 396–411. https://doi.org/10.1186/s12885-019-5584-6

Zhang, J., Li, J., Saucier, J. B., Feng, Y., Jiang, Y., Sinson, J., McCombs, A. K., Schmitt, E. S., Peacock, S., Chen, S., Dai, H., Ge, X., Wang, G., Shaw, C. A., Mei, H., Breman, A., Xia, F., Yang, Y., Purgason, A., ... Eng, C. M. (2019). Non-invasive prenatal sequencing for multiple Mendelian monogenic disorders using circulating cell-free fetal DNA. *Nature Medicine*, *25*, 439–447. https://doi.org/10.1038/s41591-018-0334-x

Zink, F., Stacey, S. N., Norddahl, G. L., Frigge, M. L., Magnusson, O. T., Jonsdottir, I., Thorgeirsson, T. E., Sigurdsson, A., Gudjonsson, S. A., Gudmundsson, J., Jonasson, J. G., Tryggvadottir, L., Jonsson, T., Helgason, A., Gylfason, A., Sulem, P., Rafnar, T., Thorsteinsdottir, U., Gudbjartsson, D. F., ... Stefansson, K. (2017). Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood*, *130*, 742–752. https://doi.org/10.1182/blood-2017-02-769869

How to cite this article: Foley, J. F., Elgart, B., Alex Merrick, B., Phadke, D. P., Cook, M. E., Malphurs, J. A., Solomon, G. G., Shah, R. R., Fessler, M. B., Miller, F. W., & Gerrish, K. E. (2021). Whole genome sequencing of low input circulating cell-free DNA obtained from normal human subjects. *Physiological Reports*, 9, e14993. https://doi.org/10.14814/phy2.14993