SUPPLEMENTARY MATERIALS

TITLE

A novel approach for *in vivo* DNA footprinting using short double-stranded cell-free DNA from plasma

AUTHORS

Jan Müller[1,2,3,4], Christina Hartwig[1,5], Mirko Sonntag[1,6], Lisa Bitzer[1], Christopher Adelmann[1], Yevhen Vainshtein[1], Karolina Glanz[1], Sebastian O. Decker[7], Thorsten Brenner[8], Georg F. Weber[9,10], Arndt von Haeseler[11,12] and Kai Sohn[1,*]

[1] Innovation Field In-vitro Diagnostics, Fraunhofer Institute for Interfacial Engineering and Biotechnology IGB, Stuttgart, Germany

[2] Max Perutz Labs, Vienna Biocenter Campus (VBC), Vienna, Austria

[3] University of Vienna, Max Perutz Labs, Department of Structural and Computational Biology, CIBIV, Vienna, Austria

[4] Vienna BioCenter PhD Program, Doctoral School of the University of Vienna and Medical University of Vienna, Vienna, Austria

[5] Institute for Interfacial Engineering and Plasma Technology (IGVP), University of Stuttgart, Stuttgart, Germany

[6] Interfaculty Graduate School of Infection Biology and Microbiology (IGIM), Eberhard Karls University Tübingen, Tübingen, Germany

[7] Heidelberg University, Medical Faculty Heidelberg, Department of Anesthesiology, Heidelberg, Germany

[8] Department of Anesthesiology and Intensive Care Medicine, University Hospital Essen, University Duisburg-Essen, Essen, Germany

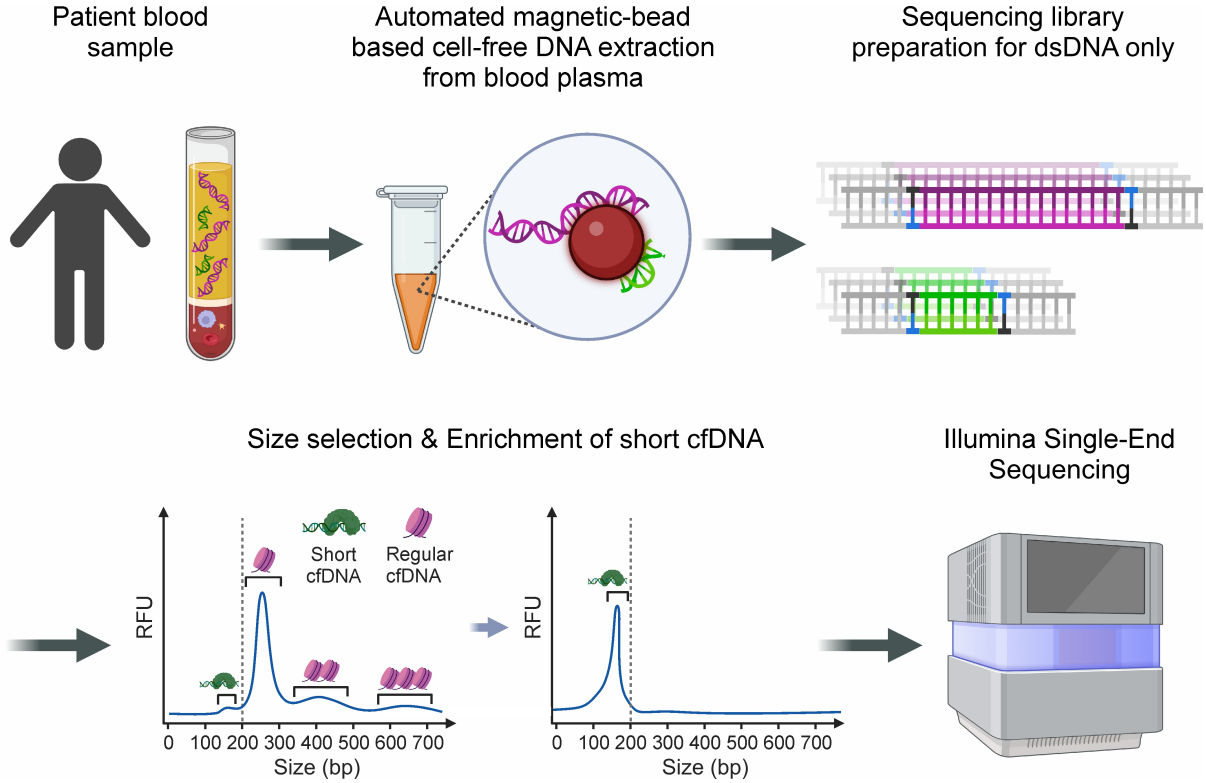
[9] Department of Surgery, Friedrich-Alexander University (FAU) Erlangen-Nürnberg and Universitätsklinikum Erlangen, Erlangen, Germany

[10] Comprehensive Cancer Center (CCC) Erlangen-EMN, Friedrich-Alexander University (FAU) Erlangen-Nürnberg and Universitätsklinikum Erlangen, Erlangen, Germany

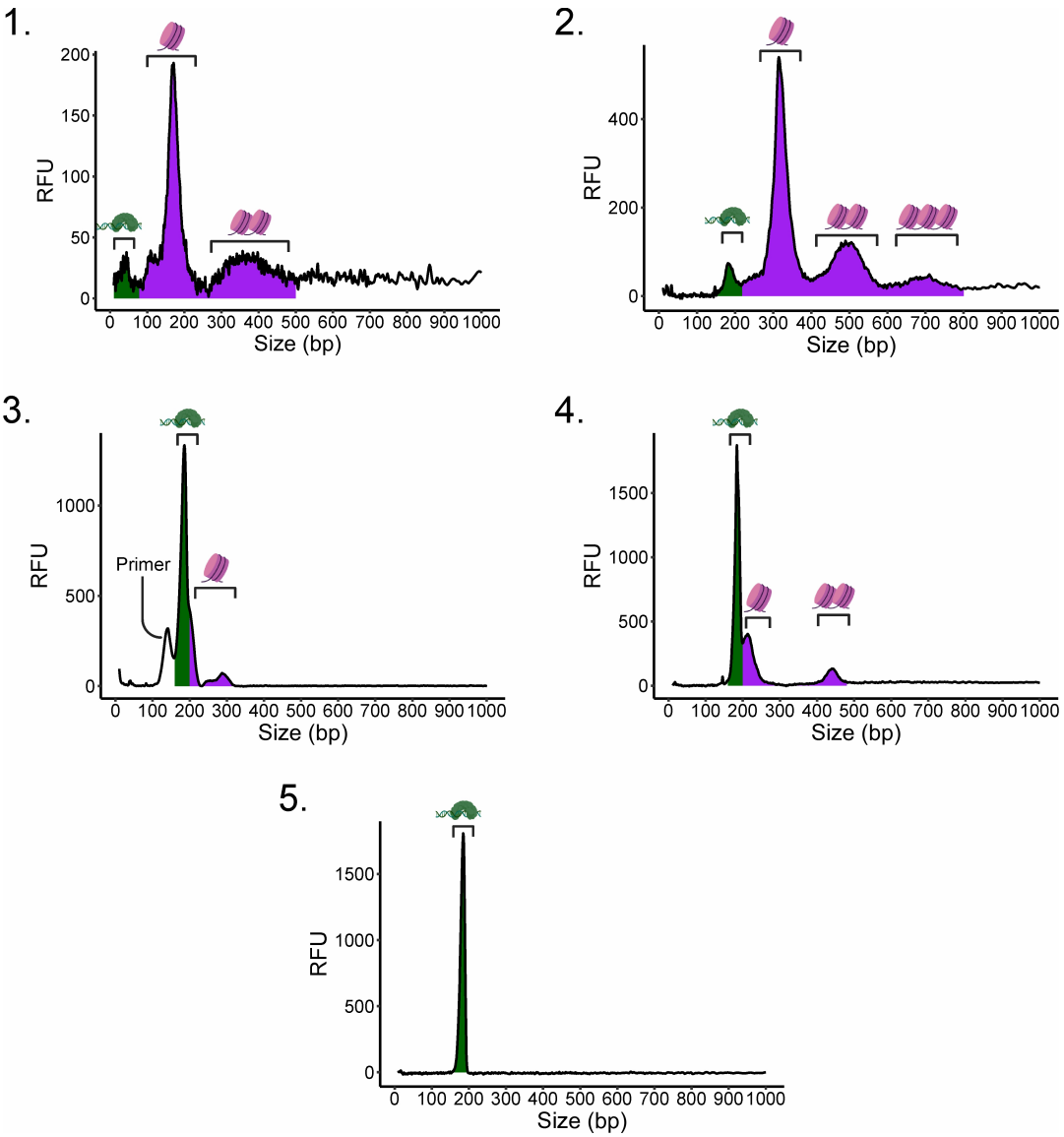
[11] Center of Integrative Bioinformatics Vienna (CIBIV), Max Perutz Labs, University of Vienna and Medical University of Vienna, Vienna BioCenter (VBC), Vienna, Austria

[12] University of Vienna, Faculty of Computer Science Bioinformatics and Computational Biology, Vienna, Austria

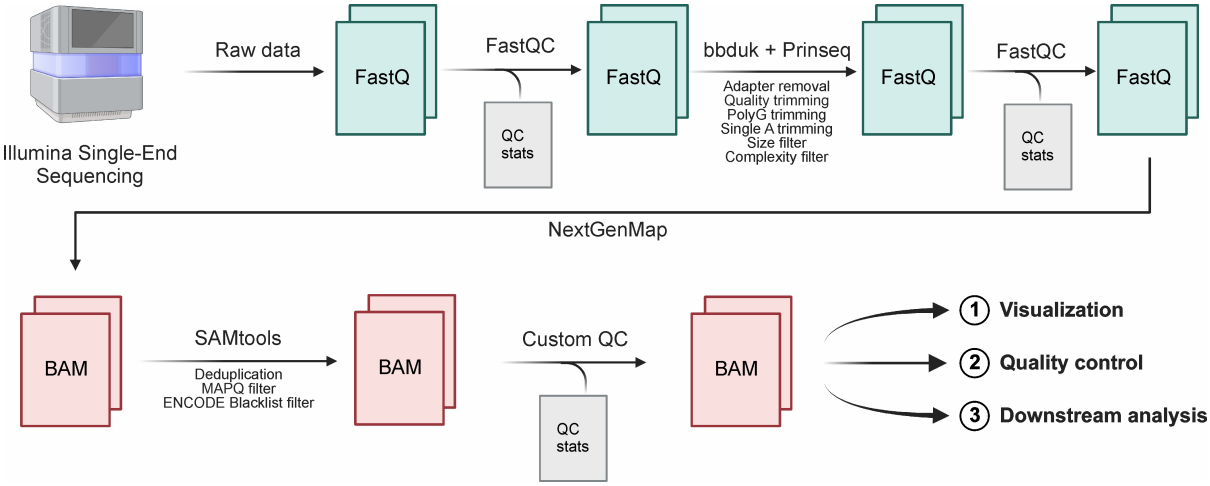
* To whom correspondence should be addressed. Tel: +49 711 970-4055; Email: kai.sohn@igb.fraunhofer.de



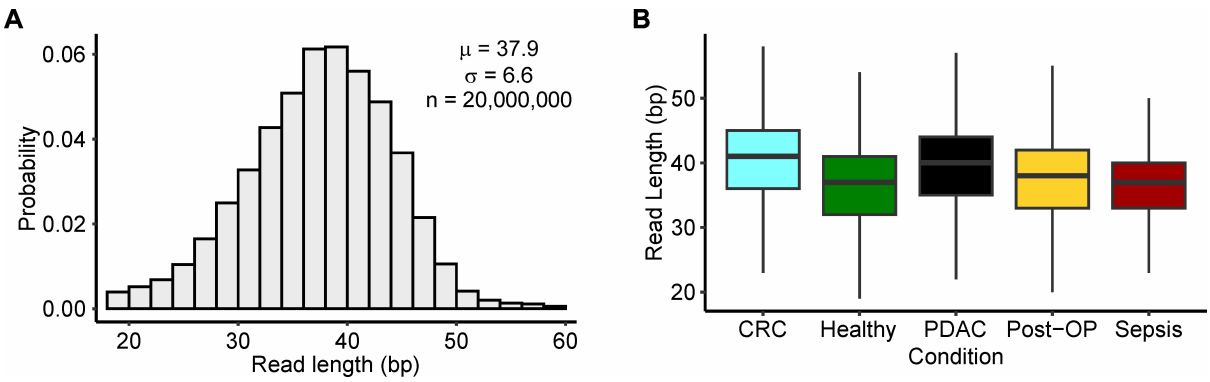
Supplementary Figure S1. Visual summary of short cfDNA extraction and sequencing from blood plasma. Cell-free DNA was extracted from the blood plasma of patients using an automated magnetic bead-based kit. Sequencing library preparation was performed by fragment end-repair and adapter ligation. Only intact double-stranded DNA fragments are enriched in the final sequencing library because of the PCR amplification. Short cfDNA is enriched by size selection from sequencing libraries. Short cfDNA sequencing libraries are sequenced on an Illumina platform in single-end mode. This figure was created with biorender.



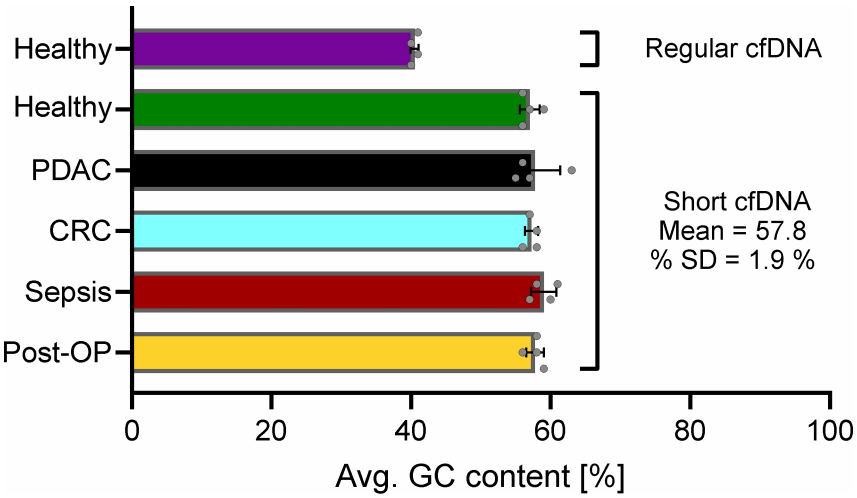
Supplementary Figure S2. Depiction of the full short cfDNA enrichment process from cell-free DNA for high-throughput sequencing. The steps shown correspond to: 1. Isolated cell-free DNA. 2. Double-stranded DNA sequencing library. 3. Size-selected sequencing library. 4. PCR-amplified sequencing library. 5. Sequencing library after second size selection. Size selection was performed using an automated preparative gel electrophoresis instrument. For each step, the Fragment Analyzer profiles of S19 are shown. The library preparation adds around 100 bps to the DNA fragments resulting in the fragment size shift seen from 2. onwards. The purple color indicates DNA fragments that can be assigned to nucleosomes or regular cfDNA, while the green color indicates DNA fragments that can be assigned to short cfDNA. The icons originally created with biorender for the supplementary figure S1 were reused in this figure.



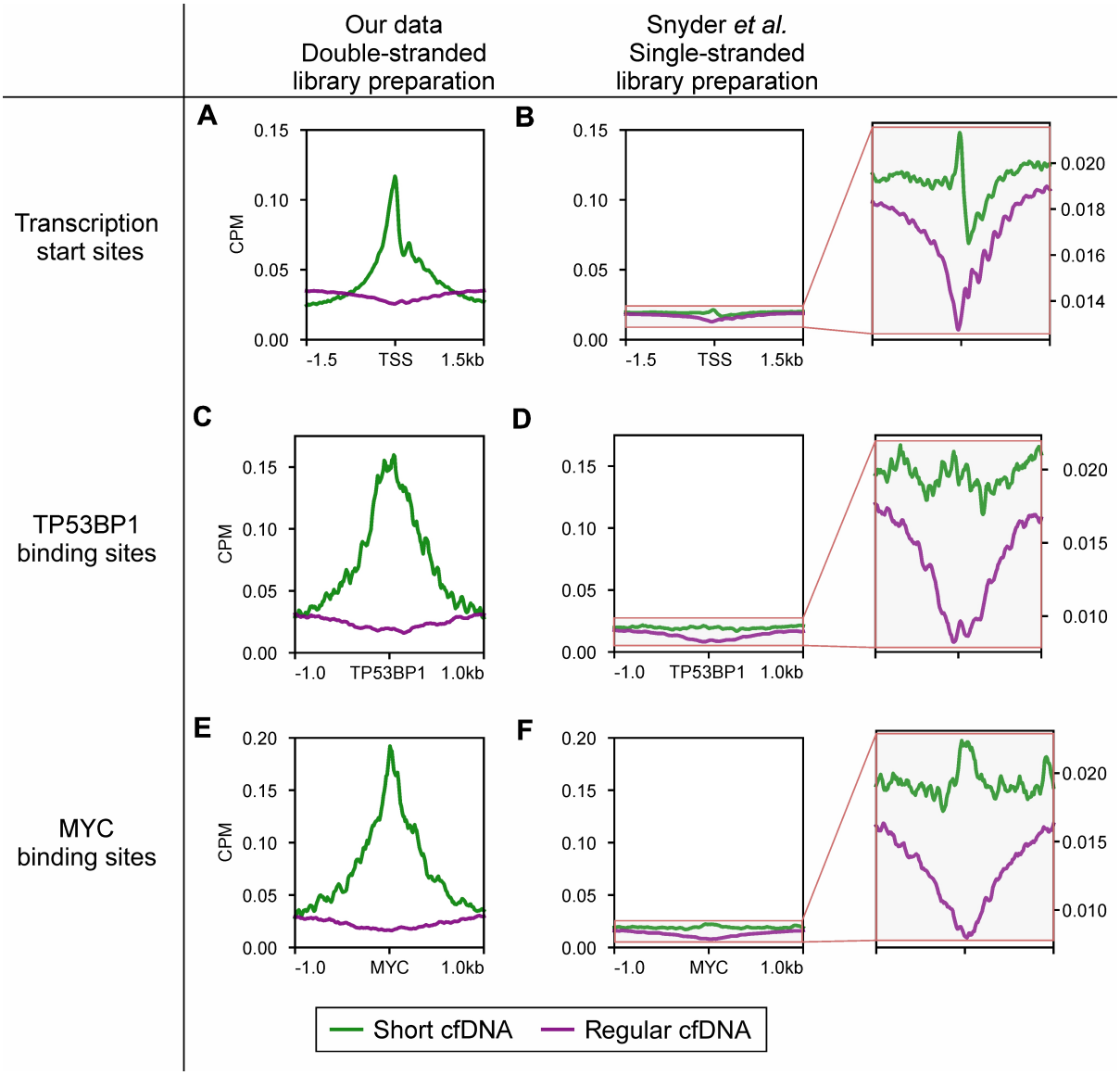
Supplementary Figure S3. Visual summary of the short cfDNA data processing pipeline. In short, the raw short cfDNA sequencing data is cleaned and filtered, mapped to a human reference genome, and filtered again before further downstream analysis. The individual steps are described in detail in the methods section. This figure was created with biorender.



Supplementary Figure S4. Analysis of read length distributions. (a) Histogram depicting the observed length of fully processed short cfDNA sequencing reads. One million random reads were taken from all twenty sequenced short cfDNA samples and their read lengths were plotted in a histogram (total n = 20 million). The observed distribution has a mean read length of = 37.9 bp (μ) and a standard deviation = 6.6 bps (σ). (b) Boxplot with data from (a) split by patient conditions.

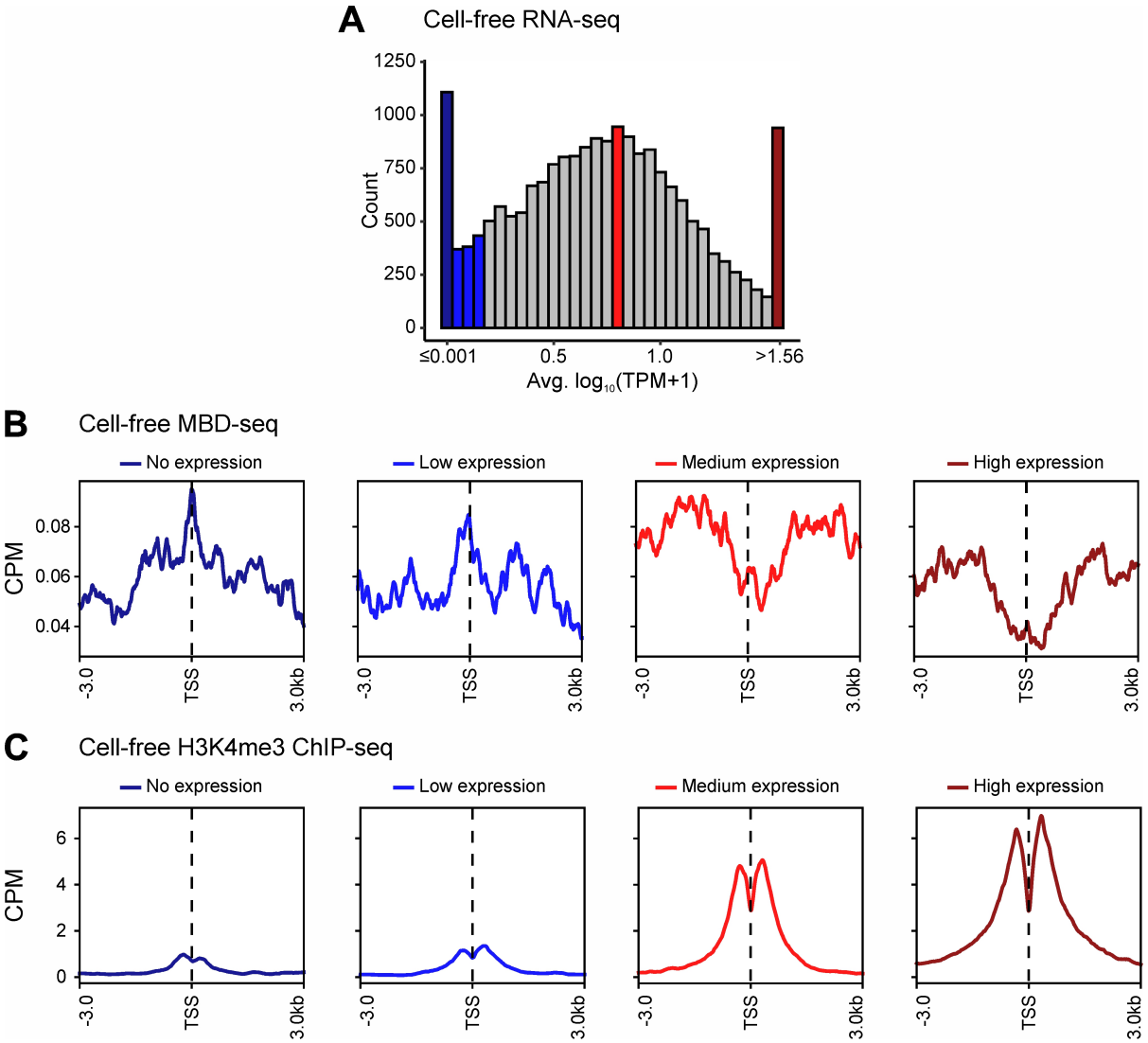


Supplementary Figure S5. Average GC content of processed sequencing reads for regular cfDNA and short cfDNA sequencing. Data from the samples S01 - S24 were used for this analysis (see Supplementary Table S1). The bar lengths represent the mean value, while error bars indicate standard deviations.

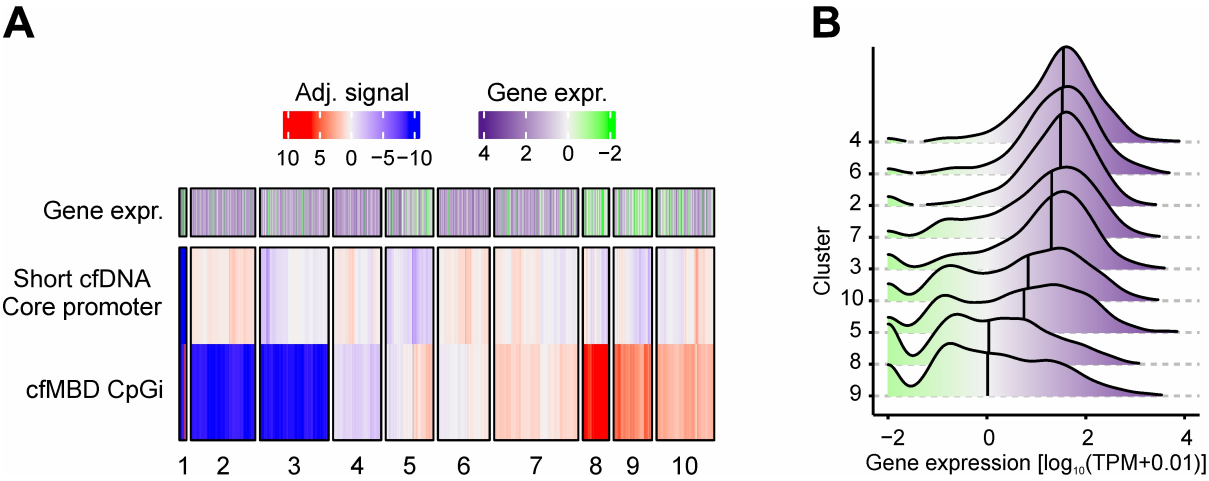


Supplementary Figure S6. Comparison of our short cfDNA sequencing data to single-stranded cfDNA sequencing data from Snyder *et al.* (Snyder *et al.* 2016). (a, c, e) Average coverage profiles based on our short cfDNA sequencing data (Sequencing depths: Short cfDNA (S03) = 8.29×10^6, Regular cfDNA (S06) = 2.60×10^7). (b, d, f) Average coverage profiles based on sequencing data from Snyder *et al.* generated with a single-stranded library preparation method (Sequencing depths: Short cfDNA = 1.33×10^8, Regular cfDNA = 3.84×10^8). These average coverage profiles are also plotted on a smaller scale to better visualize the dynamics of the data. (a and b) Average coverage profiles of annotated transcription start sites for short cfDNA and regular cfDNA. (c and d) Average coverage profiles of one thousand ChIP-seq validated tumor

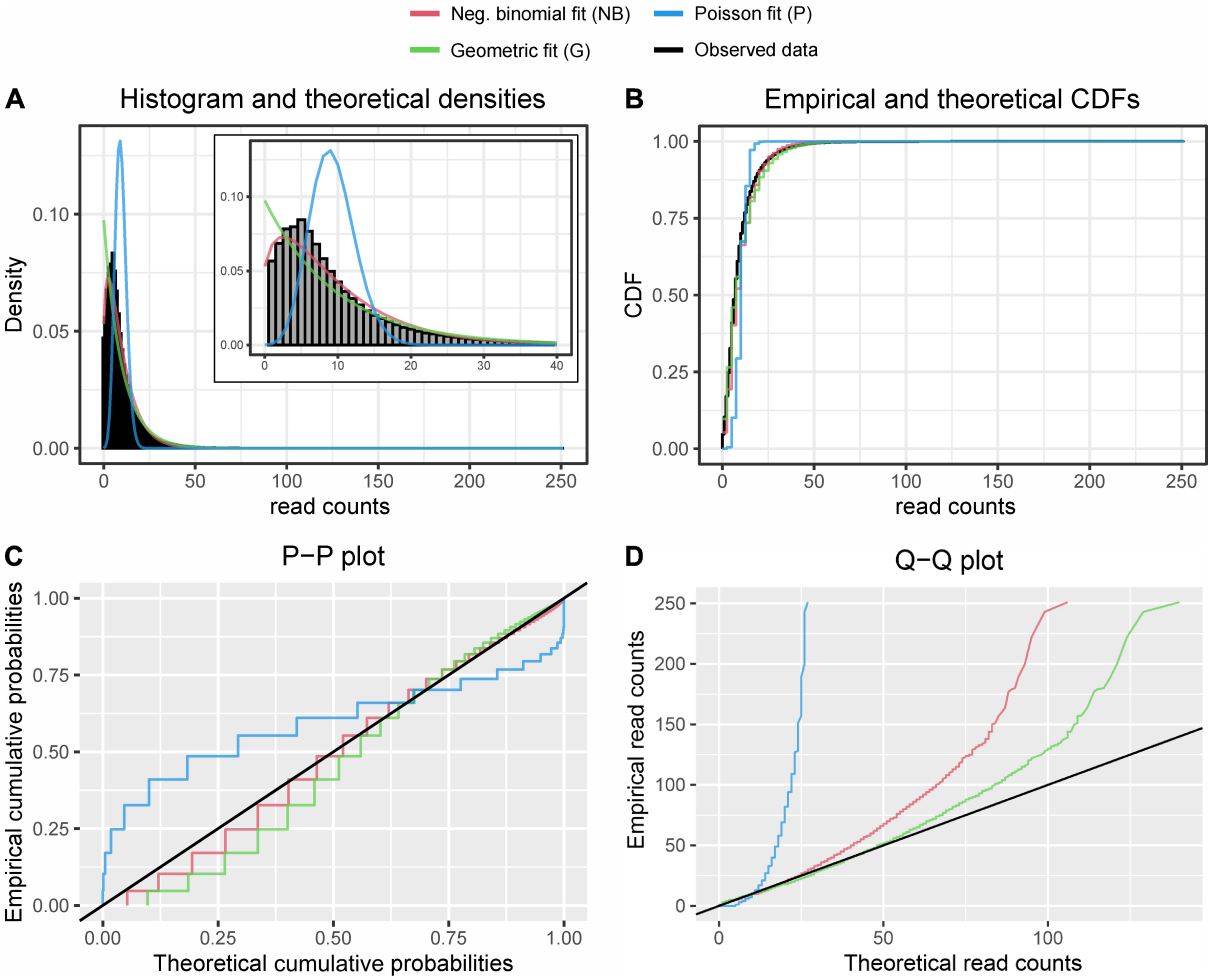
protein p53 binding protein 1 (TP53BP1) binding sites for short cfDNA and regular cfDNA. (e - f) Average coverage profiles of one thousand ChIP-seq validated MYC proto-oncogene, bHLH transcription factor (MYC) binding sites for short cfDNA and regular cfDNA. The purple color indicates data from regular cfDNA, while the green color indicates data from short cfDNA. Ultra-deep sequencing data by Snyder *et al*. was produced from the blood plasma of a healthy individual with a single-stranded library preparation method (Snyder *et al.* 2016). Raw sequencing data were retrieved from SRA (file ID = SRR2130051) and split into short cfDNA (35-80 nt) and regular cfDNA (120-180 nt) *in silico*.



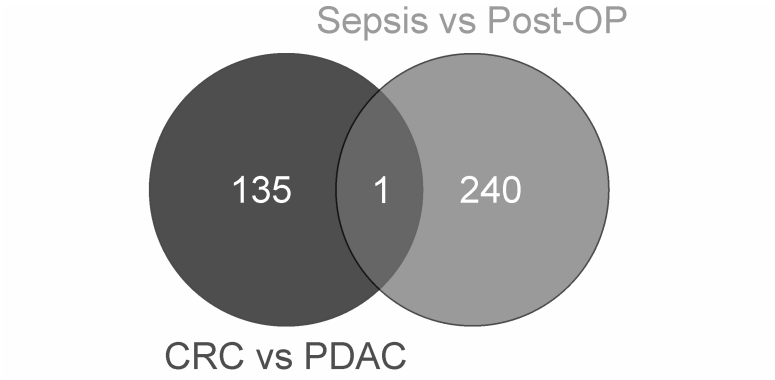
Supplementary Figure S7. Relationship of cell-free MBD-seq and cell-free H3K4me3 ChIP-seq with expression levels of genes. (a) Histogram showing average expression levels of protein-coding genes in publicly available cell-free RNA sequencing data. For each category 938 genes were selected (5 % of all analyzed genes): no expression (dark blue), low expression (blue), medium expression (red), and high expression (dark red). (b) Average coverage profiles for cell-free MBD-seq reads at selected transcription start sites. (c) Average coverage profiles for H3K4me3 ChIP-seq reads at selected transcription start sites.



Supplementary Figure S8. Analysis of joint influence of short cfDNA abundance and DNA methylation on gene expression levels. (a) Clustered heatmap with average short cfDNA signals at core promoters and average cfMBD signals at CpG islands of protein coding genes. Average gene expression levels from whole blood are annotated. (b) Sorted ridgeplot of the gene expression annotation per cluster from (a) shows the composite influence of short cfDNA abundance and DNA methylation on gene expression level distributions. Data were created from samples S26-S37.



Supplementary Figure S9. Distribution fitting and evaluation for observed read count data from short cfDNA. (A) Histogram of observed read count data (black), with three different fitted distributions as line plots (negative binomial (NB) = red, geometric (G) = green, poisson (P) = blue). An inset plot shows the read count range from 0 to 40. (B) Cumulative distribution function (CDF) plots for observed and fitted data. (C) Probability-probability plot (P-P plot) for observed and fitted data. (D) Quantile-quantile plot (Q-Q plot) for observed and fitted data.



Comparison	PDAC vs CRC		Sepsis vs Post-OP	
High in	PDAC	CRC	Sepsis	Post-OP
# DERs	105	30	155	85

Supplementary Figure S10. Comparison of loci with differential enrichment of short cfDNA. The table shows the number of differentially enriched regions (DERs) derived from short cfDNA sequencing data for the comparisons of pancreatic ductal adenocarcinoma (PDAC) versus colorectal carcinoma (CRC) and post-operative controls (Post-OP) versus sepsis. The Venn diagram shows the overlap of the two sets of DERs.

Supplementary Table S1. Metadata for the generated patient sequencing datasets included in this study. All sequenced datasets in this table are available at the SRA (PRJNA1033613).

Sequencing type	Sample	Condition	Age	Gender	Staging	Diagnosis
Short	S01	Healthy	50	male	-	-
Short	S02	Healthy	50	male	-	-
Short	S03	Healthy	45	female	-	-
Short	S04	Healthy	45	male	-	-
Regular cfDNA	S05	Healthy	47	female	-	-
Regular cfDNA	S06	Healthy	61	female	-	-
Regular cfDNA	S07	Healthy	50	male	-	-
Regular cfDNA	S08	Healthy	52	male	-	-
Short cfDNA	S09	Post-OP	20	female	-	Sigmoid colon perforation
Short	S10	Post-OP	38	female	-	Colitis ulcerosa
Short	S11	Post-OP	29	female	-	Gastroparesis
Short cfDNA	S12	Post-OP	56	female	-	Cholelithiasis with chronic cholecystitis
Short	S13	Sepsis	50	male	-	Sepsis
Short	S14	Sepsis	65	female	-	Sepsis
Short	S15	Sepsis	74	male	-	Sepsis
Short	S16	Sepsis	71	male	-	Sepsis
Short cfDNA	S17	PDAC	68	male	III	Pancreatic head ductal adenocarcinoma
Short	S18	PDAC	75	male	II	Pancreatic tail
Short cfDNA	S19	PDAC	60	male	I	Pancreatic ductal adenocarcinoma
Short cfDNA	S20	PDAC	63	female	III	Pancreatic corpus carcinoma

Short cfDNA	S21	CRC	84	male	II	Adenocarcinoma of colon ascendens
Short	S22	CRC	65	male	II	Coecal carcinoma
Short cfDNA	S23	CRC	72	female	III	Deep-seated rectal carcinoma
Short cfDNA	S24	CRC	86	male	0 (regression)	Deep-seated rectal carcinoma
cfMBD-seq	S25	Healthy	45	female	-	-
Short	S26	Sepsis	75	male	-	Sepsis
Short	S27	Sepsis	57	male	-	Sepsis
Short	S28	Sepsis	45	male	-	Sepsis
Short	S29	Sepsis	59	male	-	Sepsis
cfMBD-seq	S30	Sepsis	75	male	-	Sepsis
cfMBD-seq	S31	Sepsis	57	male	-	Sepsis
cfMBD-seq	S32	Sepsis	45	male	-	Sepsis
cfMBD-seq	S33	Sepsis	59	male	-	Sepsis
RNA-seq	S34	Sepsis	75	male	-	Sepsis
RNA-seq	S35	Sepsis	57	male	-	Sepsis
RNA-seq	S36	Sepsis	45	male	-	Sepsis
RNA-seq	S37	Sepsis	59	male	-	Sepsis

Supplementary Table S2. Detailed metadata of sepsis patients. Sepsis samples S30-S37 correspond to the same four individuals as S26-S29. Abbreviations: Erythrocytes = Erythro, Thrombocytes = Thrombo, Leukocytes = Leuko. Units: Erythrocytes = 10^{12} cells/L, Thrombocytes = 10^{9} cells/L, Leukocytes = 10^{9} cells/L.

Sample	**Age**	**Gender**	**Pathogens**	**Erythro**	**Thrombo**	**Leuko**
S13	50	male	*Bacteroides fragilis; Escherichia coli*	2.4	1127	13.67
S14	65	female	*Bordetella petrii; Human herpesvirus 5; Klebsiella pneumoniae; Pseudomonas mendocina; Salmonella enterica*	NA	NA	NA

S15	74	male	*Enterococcus faecium*	2.7	273	10.06
S16	71	male	*Enterococcus faecium*	3.3	338	24.13
S26	75	male	*Bifidobacterium animalis; Enterococcus faecium; Escherichia coli; Klebsiella pneumoniae; Lactobacillus fermentum; Staphylococcus warneri*	4	156	2.12
S27	57	male	*Haemophilus influenzae; Haemophilus parainfluenzae*	3.5	221	18.15
S28	45	male	NA	4	457	17.33
S29	59	male	*Propionibacterium spp.*	2.9	83	4.39

Supplementary Table S3. Metadata for the utilized public datasets in this study. Cell-free H3K4me3 ChIP-seq data are available from Zenodo: https://zenodo.org/records/4277001. ATAC-seq data from ENCODE were downloaded with the GRCh38 assembly and converted locally to the GRCh37 assembly with liftOver.

Data type	Sample type	Source	Project ID	File ID
ATAC-seq	Cell line (GM12878)	ENCODE	ENCSR095QNB	ENCFF646NWY
Cell-free RNA-seq	Plasma of healthy individuals	SRA	PRJNA598835	SRR10822588, SRR10822583, SRR10822579, SRR10822594, SRR10822591
Cell-free H3K4me3 ChIP-seq	Plasma of a healthy individual	Zenodo	-	H013.1
DNase-seq	Cell line (GM12878)	ENCODE	ENCSR000EMT	ENCFF783ZLL
DNase hypersensitive sites	Cell line (GM12878)	ENCODE	ENCSR000EMT	ENCFF273MVV
Double-stranded cell-free DNA	Plasma of a healthy individual	SRA	PRJNA291063	SRR2130050
Single-stranded cell-free DNA	Plasma of a healthy individual	SRA	PRJNA291063	SRR2130051

REFERENCES

Snyder M, Kircher M, Hill A, Daza R, Shendure J. 2016. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164:** 57–68. doi:10.1016/j.cell.2015.11.050.