Coverage landscape of the human genome in nucleus DNA and cell-free DNA

Jiaqi Luo[1] **and Shuai Cheng Li**[1,2,✉]

[1] Department of Computer Science, City University of Hong Kong, 83 Tat Chee Ave, Kowloon Tong, Hong Kong
[2] Department of Biomedical Engineering, City University of Hong Kong, 83 Tat Chee Ave, Kowloon Tong, Hong Kong

For long, genome-wide coverage has been used as a measure of sequencing quality and quantity, but the biology hidden beneath has not been fully exploited. Here we performed comparative analyses on genome-wide coverage profiles between nucleus genome DNA (gDNA) samples from the 1000 Genomes Project (n=3,202) and cell-free DNA (cfDNA) samples from healthy controls (n=113) or cancer patients (n=362). Regardless of sample type, we observed an overall conserved landscape with coverage segmentation, where similar levels of coverage were shared among adjacent windows of genome positions. Besides GC-content, we identified protein-coding gene density and nucleosome density as major factors affecting the coverage of gDNA and cfDNA, respectively. Differential coverage of cfDNA vs gDNA was found in immune-receptor loci, intergenic regions and non-coding genes, reflecting distinct genome activities in different cell types. A further rise in coverage at non-coding genes/intergenic regions and a further drop of coverage at protein-coding genes/genic regions within cancer cfDNA samples suggested a relative loss of contribution by normal cells. Importantly, we observed the distinctive convergence of coverage in cancer-derived cfDNA, with the extent of convergence positively correlated to stages. Based on the findings we developed and validated an outlier-detection approach for cfDNA-based cancer screening without the need of cancer samples for training. The method achieved 97% sensitivity on pediatric sarcomas (n=241) and 44% sensitivity on early-stage lung cancers (n=36) with >90% specificity for condition-matched tasks, 100% sensitivity on late-stage cancers (n=85) for condition-unmatched tasks, outperforming current benchmarks.

Correspondence: ***sc.li@cityu.edu.hk***

Introduction

Cell-free DNA (cfDNA) is one of the most highly anticipated biomarkers for the detection and monitor of cancers or other diseases. A single tube of peripheral blood can extract both the peripheral blood mononuclear cells (PBMCs) carrying the reference nucleus genome DNA (gDNA), and the plasma cfDNA that potentially contains circulating tumour DNA (ctDNA) with cancer-related mutations (Figure 1). This paired mode of somatic mutation calling has become a mature strategy widely used in clinical settings for the identification of targetable mutations or the monitoring of cancer progression (1–3). However, only a part of patients (more likely in late-stages) carry known driver mutations related to cancers, and not all mutations are detectable in cfDNA. The phenomenon of clonal hematopoiesis (CH) where somatic mutations that could be shed into cfDNA accumulate in hematopoietic stem cells, may introduce 'false positive' in known oncogenes such as TP53, KRAS, GNAS, NRAS and PIK3CA (4) as it also happens in healthy individuals. To expand the scope of cfDNA-based liquid biopsy, alternatives are required to circumvent the limitations of somatic mutation calling. Intrinsic connections were discovered between nucleus gDNA and cfDNA, that cfDNA fragments of variable lengths are the products of nuclease activities on gDNA released by cell death (5, 6). The parts of the gDNA under protection of protein-binding (e.g. nucleosomes, transcription factors) remain intact, while the rest being naked are susceptible to nuclease cleavage, especially at DNase I hypersensitive sites (DHSs). Genomic regions enriched with long fragments (120-180bp) generally map the positions of nucleosomes while those with dense short fragments (35-80bp) correlate to DHSs and/or transcription factor binding sites. A hallmark of actively transcribed genes is the nucleosome-depleted region at the promoter and reduced nucleosome occupancy extending into the gene body that causes a drop of read coverage. The nucleosome and DHS positioning differs between tissues or cell types, thus providing possibility to infer the (tumour) tissue-of-origin. Besides, significant differences in fragment sizes, end points and end motifs have been found between ctDNA and control cfDNA. These fragmentomic features coupled with machine learning models yield good performances on single datasets where train and test samples come from the same study, despite the generalizability to independent external datasets were not fully evaluated (7–9).

Although being used locally as an auxiliary variable to infer mutations, protein-binding, transcriptions and tissue-of-origin in previous studies (1, 3, 5, 6), coverage itself has not been studied as an outcome variable at a genome-wide level. Coverage measures the number of sequencing reads that 'covers' a specified genomic position or region. Coverage information has been widely used as a quality control metric in sequencing experiments and mutation calling algorithms (10, 11). Lindner et al (12) introduced the use of genome coverage profiling to solve complex mixtures in metagenomics. The genome-wide coverage is read-based, regardless of fragment sizes. Inspired by previous work, we viewed the genome-wide coverage as an outcome of genomic, transcriptomic and epigenomic activities from multiple cell types, which may

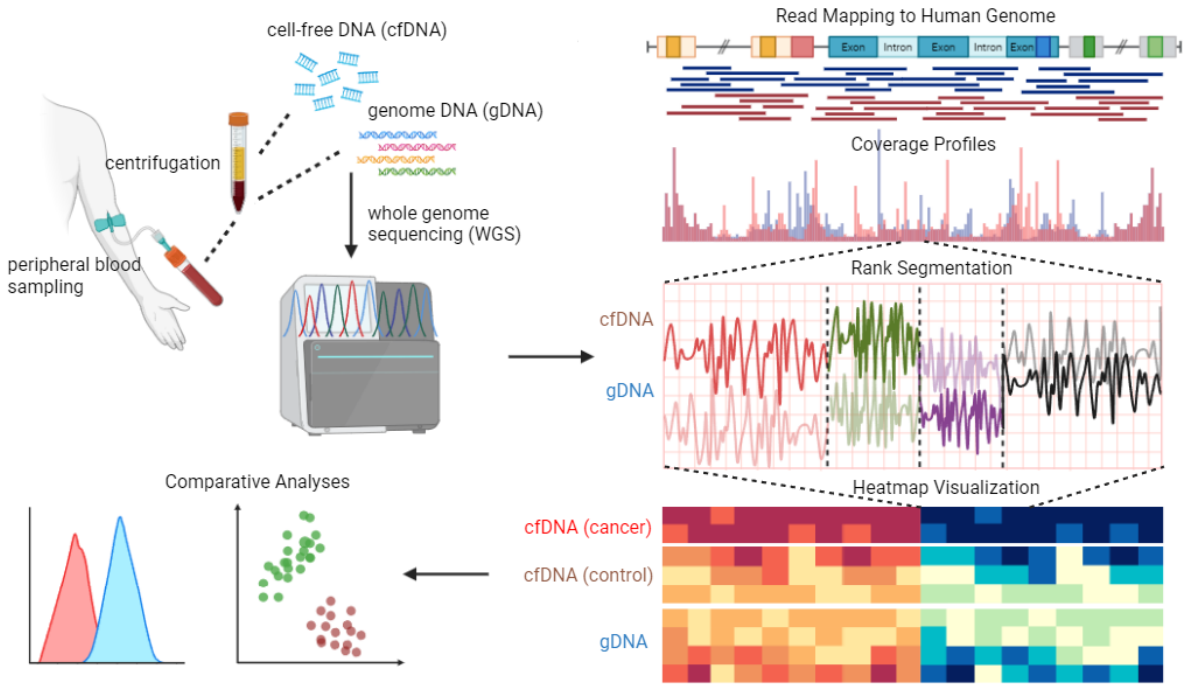

Figure 1. A graphical abstract of the workflow and methodology. Whole genome sequencing data of nucleus genome DNA (gDNA) derived from peripheral blood mononuclear cells (PBMCs) and cell-free DNA (cfDNA) derived from healthy controls or cancer patients were obtained from 1000 Genomes and European Genome Achieve. Raw data was mapped to the human reference genome, and genome-wide coverage profiles were calculated. Rank transformation was performed to transform coverage into rank to allow integration of data from different studies. Borders of the coverage rank segmentation were computed and mapped with the positioning of nucleosomes and genes (protein-coding, non-coding). The genome-wide coverage landscapes of gDNA vs cfDNA, and cfDNA from healthy control vs cancer patients were systemically compared and visualized with heatmaps. A proof-of-concept method based on outlier-detection was developed and validated based on the distinctive feature identified in cancer samples.

be biologically meaningful and distinguishable between health and diseases. We systemically compared and contrasted genome-wide coverage between gDNA versus cfDNA, and between cfDNA of healthy controls versus various cancers (Figure 1). We looked for coverage patterns that are conserved or distinct among sample types, and tried to explain the patterns with factors such as cell types, GC-content, positioning of genic/intergenic regions, protein-coding/non-coding genes and nucleosomes. Finally, we examined the feasibility of cfDNA-based cancer screening by genome-wide coverage profiling.

Conventionally, ctDNA refers to cfDNA released by tumour cells, as opposed to cfDNA released by non-cancerous cells such as white blood cells. Therefore, a cfDNA sample from a cancer patient should contains both ctDNA and normal cfDNA. Hereafter for the ease of expression, we use ctDNA to represent the **entire cfDNA sample drawn from cancer patients**, while cfDNA (if without specific clarification) represents the **entire cfDNA sample from healthy controls**.

Methods

Datasets

Data used in this study include the 3,202 gDNA samples of high-coverage, whole-genome sequencing data from the 1000 Genomes Project (1kGP) (13), and the cfDNA samples drawn from multiple studies (9, 14–17) of cancer samples with or without condition-matched controls (Table 1).

Data processing

The raw WGS data (fastq, bam) was pre-processed following the GATK best-practice pipeline (18). In brief, each sample was mapped to the hg38 human reference genome, with sequencing duplicates removed and base quality score recalibrated.

Genome-wide coverage profiling

10kb-window coverage fraction. Coverage at each genomic position was calculated with Samtools using the command 'samtools depth -J -s' (19). Coverage of all positions within each 10kb-window was sum together as the window's coverage. Each window coverage was then

divided by the total sum of all window coverage to get the coverage fraction. Note that repeated sequences located in pericentromeric regions can vary greatly between samples and affect the coverage fractions of the rest of the genome. Therefore, the three cytobands closest to the centromere from both arms of each chromosome were excluded from calculations. Non-autosomal chromosomes (chrX, chrY, chrM) were also excluded for coverage calculations to remove bias caused by gender. Finally, windows showing zero coverages in all samples were filtered, resulting in a total of 236,239 windows.

Rank transformation. A normalization/integration approach is required for the correct comparisons of samples within and between studies. Rank transformation was performed on the coverage fraction profile of each sample. The genome-wide 10kb-windows of fraction coverage were ranked, the bigger the coverage fraction value, the bigger the rank; tied elements were assigned the possible minimum rank. The rank transformation was compared with the raw coverage fraction, Z-score standardization and min-max normalization on dealing with batch effects. Among the methods compared, the rank transformation had been shown to minimize the differences between studies of control cfDNA samples (Supplementary Figure 1).

Calculations of gene and nucleosome density

Gene annotations were obtained from gencode gff3 version 46. Stable nucleosome positions from healthy individuals were obtained from the GSE81314 study (20). For each 10kb window, the count of all genes/protein-coding genes/non-coding genes/nucleosomes covered by the window was calculated as density.

Computing the borders of coverage segmentation

We observed the phenomenon of coverage segmentation, that similar levels of coverage ranks were shared among adjacent windows to form segments. To locate the segments we need to find the borders of each segment. The python package ruptures (21) was used to find the borders of the coverage segments for each chromosome, using the median coverage rank of the 3,202 1kGP gDNA samples at each 10kb window. Specifically, we chose binary segmentation as the search method, with minimum segment size set to 100 windows (i.e., $\geq$ 1Mbp). To allow automatic determination of number of segments (breakpoints), penalty was set to $\log(n) \cdot \text{sigma}$, where n is the total number of windows for the chromosome, sigma is the standard deviation of the window ranks of the chromosome.

Identification of up/downregulated segments

We compared the median segment ranks of all ctDNA samples (n=362) vs all cfDNA samples (n=113) to get top20 upregulated or downregulated segments and performed gene ontology enrichment analysis on genes underlying these segments using the R package clusterProfiler (22). While significant (FDR<0.05) enrichment was found in top20 downregulated segments, no significant enrichment was found in top20 upregulated segments (Supplementary Table 1 and Supplementary Table 2).

Then we examined each cancer dataset to identify ctDNA frequently up/downregulated segments. The coverage ratio of each segment was calculated between the median of each cancer dataset versus the median of the condition-matched controls (if existed) or the total control (n=113), except for the four prostate cancer samples from EGAD00001005343 due to too small sample size. Segments with at least 10% upregulation or downregulation were listed for each cancer dataset. The intersection of all lists was defined as frequently up/downregulated segments. Under this criteria, no intersection was found for downregulated segments, while three segments were found frequently upregulated (Supplementary Table 3).

Differentially covered genes

For each gene, coverage is defined as the mean rank of the windows covering the gene. Similar to low-expression genes, we discard genes with the lowest 20% mean coverage in ctDNA and cfDNA. Then for each gene the mean coverage ratio between ctDNA and cfDNA is calculated. Mann–Whitney–Wilcoxon tests with Bonferroni corrections are performed on the gene ratios. Genes with FDR<0.01 are identified as differentially covered genes.

Rank histogram and outlier detection

After the rank transformation each sample has 236,239 10kb-windows whose coverage ranked from 1 to 236,239. An eight-bin histogram (each bin sized 30,000) was computed for each chromosome describing the distribution of ranks for the chromosome. In the four control datasets, the maximum median standard deviation of segment ranks was 14766.6 (EGAD00001006237, n=38), hence the choice of the 30,000 bin size is roughly two standard deviations.

The outlier detection method only requires sufficient number of control samples ($n \geq 20$) for training. We used the Elliptic Envelope algorithm (23) implemented in scikit-learn (24) to learn the control histogram distribution of each chromosome, with the expected contamination rate of control set to near zero (1e-32), indicating no outlier is expected in the training data. Some chromosomes are more vulnerable to batch effects, as demonstrated in models trained and tested by ten unselected batches of 1kGP gDNA samples which predicted a high contamination rate of the training data itself (Supplementary Table 4). We filtered out chromosomes with a contamination rate larger than 0.05 (chr8, 9, 10, 11, 14) and chromosomes with $\leq$ 30 segments (chr16, 17, 18, 19 ,20, 21, 22), resulting in ten chromosomes (chr1, 2, 3, 4, 5, 6, 7, 12, 13, 15) with a total of 679 segments. Each model is specifically trained for a chromosome. Each model for each chromosome gives a prediction score of whether the given histogram is abnormal (0 for abnormal, 1 for normal). The prediction score for

each chromosome is weighted according to the number of segments in the chromosome and sum together. For example, chr1 has 98 segments, chr2 has 100 segments, then the weights for chr1 and chr2 are 98/679=0.1443, 100/679=0.1473, respectively. The weighted sum is the proportion of segments that support an inlier prediction; the proportion of inlier segments and outlier segments sums to 1. We defined the outlier proportion cutoff as α, samples with weighted sum less than 1-α is predicted as outlier that differs from the controls.

The outlier proportion cutoff α is determined by an iterative process that tests a range of α values. For each α, we calculate the leave-one-out specificity, by iterating N (number of control samples) times, each time training N-1 control samples, leaving one control sample for test, then calculate the specificity with the N test results. The smallest α value that satisfies a wanted level of leave-one-out specificity (0.9 as default) on the control samples is selected as cutoff for test. Test samples with outlier portion exceeding the cutoff are deemed outliers.

Benchmarks

For benchmarking, we chose ichorCNA (1) as an representative of mutation-based methods, and LIQUORICE (9) as an representative of fragmentomics-based methods. LIQUORICE was developed based on samples from EGAD00001007080, one of the major dataset for our performance evaluation; ichorCNA is a popular tool for tumour content prediction and has been used in the EGAD00001007080 study.

For ichorCNA, we obtained the detailed predictions for EGAD00001007080 with minimum, maximum, and top-ranked tumour content values from the original study (9), and we used the default cutoff value 0.03 as cutoff for the maximum tumour content, i.e., classifying as cancerous if a sample with predicted maximum tumour content $\geq$0.03. Since the samples in EGAD00001006237 are early-stage lung adenocarcinoma (LUAD) samples, parameters recommended for low tumour content was used with the default 1Mb bins and 0.03 tumour content cutoff.

For LIQUORICE, EwS-specific DHSs and Hematopoietic-specific DHSs were obtained from the original paper and used as region sets for EGAD00001007080. LUAD-specific DHSs was obtained from ENCODE DNase Clusters specific to A549 LUAD cell line as instructed; LUAD-specific DHSs and Hematopoietic-specific DHSs were used as region sets for EGAD00001006237. Both datasets were then run with default parameters.

Results

Coverage landscape is conserved between gDNA and cfDNA

We first looked at the overall similarity of genome-wide coverage landscape between the two sample types (gDNA, cfDNA). To integrate samples from different studies, we performed a rank transformation of the coverage profiles over 10kb windows. For each sample (row), windows were ranked based on their coverage, the higher the coverage the larger the rank (see Methods). To preserve the original coverage landscape and to investigate the contributions of multiple factors that could be confounded by GC-content, no GC-corrections were performed prior to the integration of multiple datasets. The rank-transformed coverage profiles were visualized with heatmaps (Figure 2a; Supplementary Figure 1). From the heatmaps one can observe apparent aggregation of windows into segments (adjacent windows share similar colours), meaning adjacent windows tend to present similar coverage (rank). We computed the borders of these segments throughout the genome using the 3,202 gDNA samples from 1kGP (see Methods), resulting in 1,090 segments of variable length ($\geq$ 1Mb) across the genome (Figure 2a, dashed vertical lines). The borders of segmentation seemed to be largely conserved even between gDNA and cfDNA. The conserved segmentation patterns allow the use of the 1,090 segments as a representation of the whole genome for noise and dimensionality reduction.

We identified major contributing factors to coverage (Figure 2c). Consistent with previous findings (25), GC-content showed a strong positive correlation with the 10kb-window coverage (R=0.816). We also confirmed the confounding effect of GC-content (Supplementary Figure 2) as it showed positive correlation with gene density (R=0.537) and weak negative correlation with nucleosome density (R=-0.197). Interestingly, significant partial correlation was found between gene density and gDNA coverage (ρ=0.404, GC-controled) but not cfDNA coverage (ρ=0.13). This contribution by gene density was mainly from protein-coding genes (ρ=0.395) - segments enriched with protein-coding genes seemed to present higher coverage, while segments with sparse protein-coding genes are less covered (Figure 2a-b). However in cfDNA, nuclosome density (ρ=0.364) instead was the contributor of coverage second to GC-content. Our results indicated that apart from GC-content, the positioning of protein-coding genes affects gDNA coverage, whereas nucleosome positioning affects cfDNA coverage.

cfDNA vs gDNA differential coverage at immune-receptor loci, intergenic regions and non-coding genes

Next, we compared the segment coverage profile of cfDNA (n=113) and gDNA (n=3,202). For each sample, the coverage of each segment is defined as the mean of window ranks within the segment. We identified top20 cfDNA (versus gDNA) downregulated or upregulated regions, respectively (Figure 3a). We observed a 3-fold upregulation of cfDNA at the segment mapping the IGH locus located 14q32.33 (Figure 3a-b; Supplementary Figure 3). We then examined whether differential coverage also occur in the TRB locus located at 7q34. Although not as strong as that of the IGH locus, a 32% increase was found in cfDNA on the segment mapping the TRB locus (Figure 3a-b). The

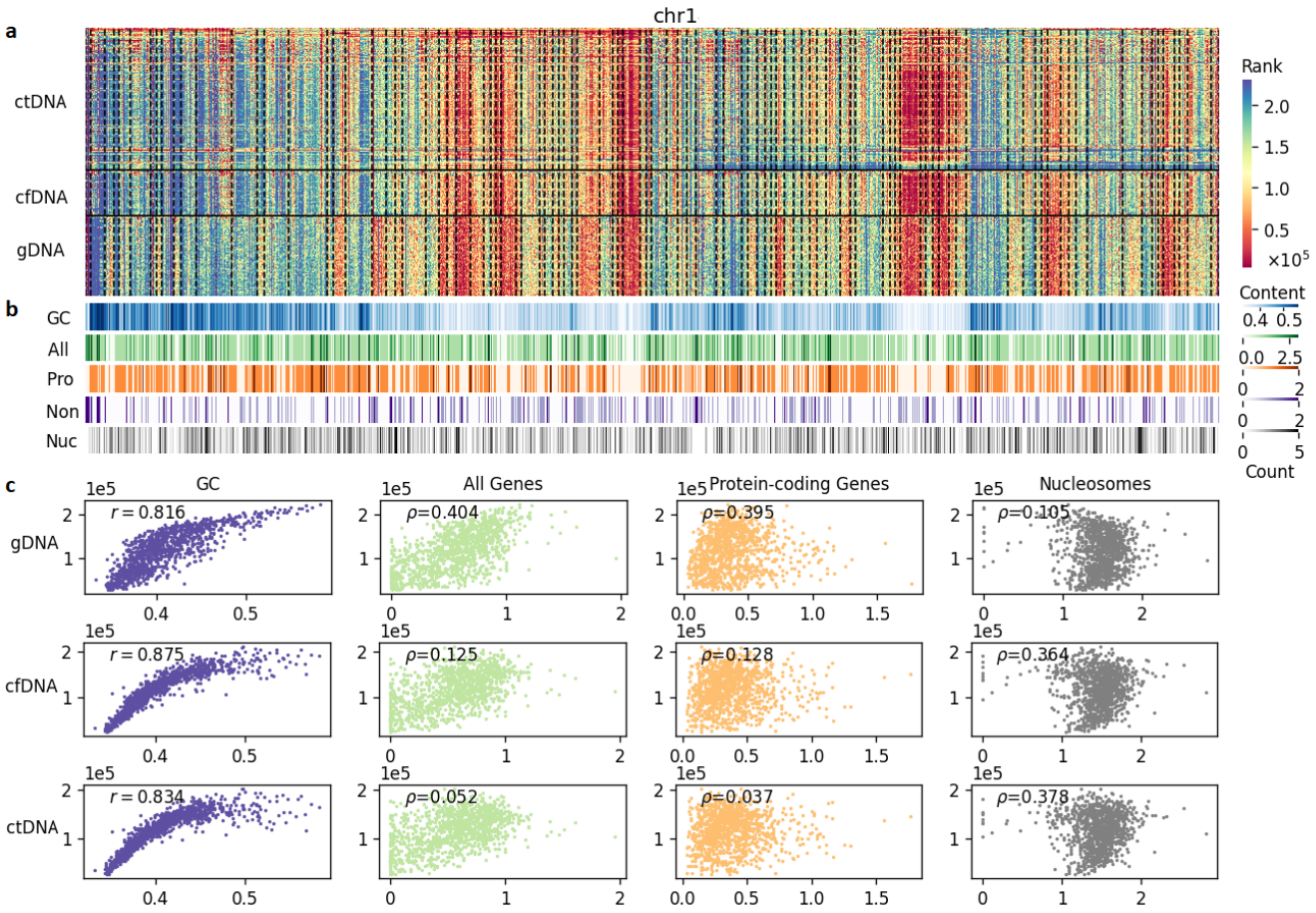

Figure 2. The coverage landscapes of gDNA, cfDNA and ctDNA. **(a)** Heatmap visualizing the rank-transformed coverage profiles of ctDNA, cfDNA and gDNA over 10kb windows at chr1. colour segments are evident, demonstrating the conserved patterns of coverage segmentation regardless of sample types. Dash lines indicate segment borders. Since there are too many gDNA (total=3,202) samples, here we only plot 150 unselected gDNA samples to fit the numbers of other samples. **(b)** Heatmap showing the GC-content (GC), gene density (All), protein-coding gene density (Pro), non-coding gene density (Non) and nucleosome density (Nuc) within each 10kb window at chr1. **(c)** Scatter plots of genome-wide coverage rank vs GC-content, (all) gene density, protein-coding gene density and nucleosome density in gDNA, cfDNA and ctDNA.

IGH locus is rearranged in B cells and partially rearranged in T cells (26), meaning a large part of the locus is lost in lymphocytes, which account for about 80% of total PBMCs. This explains the extremely low coverage at the IGH locus in the gDNA samples, because the cell sources of 1kGP gDNA samples are PBMCs. Similarly, from the dramatic IGH upregulation in cfDNA we can infer that lymphocytes are not the major cell sources of cfDNA. Indeed, neutrophils are the most abundant leukocytes, and their short lifespan will even add to their contribution to dead cells in the blood - the source of cfDNA. Compared to the broad IGH partial rearrangement in T cells, the TRB rearrangement in B cells from healthy individuals is less prevalent and restricted to T-cell receptor (TCR) δ genes (27). However, a subset of circulating neutrophils (5-8%) is known to express TCRs (28) and thus having rearranged TRB locus. As a result, greater coverage difference was observed in the IGH locus, compared to the TRB locus.

Besides immune-receptor loci, cfDNA was found significantly upregulated in intergenic regions and non-coding genes (Figure 3c), compared to genic regions and protein-coding genes, respectively. The cfDNA/gDNA ratio was not affected by GC-content (R=-0.067; Supplementary Figure 4). Gene density negatively correlated with cfDNA/gDNA ratio in most chromosomes (Figure 3a; Supplementary Figure 5). Upregulation of cfDNA was typically observed in segments without protein-coding genes, which seemed to be uniformly covered in gDNA but with apparent ups and downs in cfDNA that seemed to correspond with density of non-coding genes (Figure 3d, two cfDNA upregulated segments highlighted with green triangles). Meanwhile, cfDNA downregulation was typically observed in protein-coding regions (Figure 3e, two cfDNA downregulated segments highlighted with red triangles). The differential coverage between cfDNA and gDNA reflects the distinct activities (e.g. rearrangement, apoptosis) happen within different cell sources.

Despite significant differential coverage at local regions, the distribution of coverage is largely conserved between cfDNA and gDNA at chromosome level (Figure 3f). The mean and variance of each chromosome distinguish from one another; meanwhile, the mean and variance of each

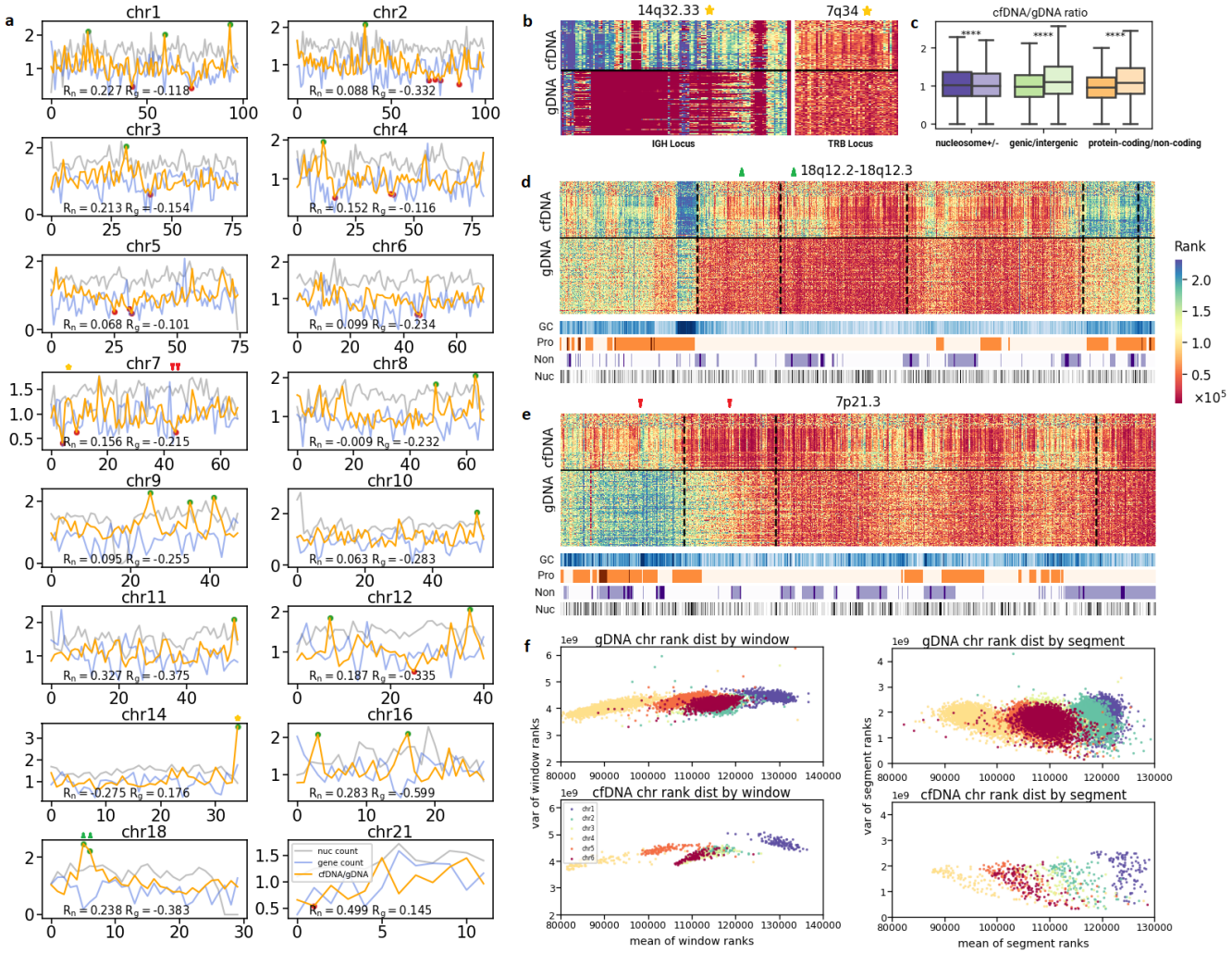

Figure 3. Differential coverage of cfDNA vs gDNA. **(a)** Line plots of segment coverage ratio of cfDNA vs gDNA in different chromosomes, plotted together with nucleosome density (count) and gene density (count). The Pearson correlation coefficient between cfDNA/gDNA ratio and nucleosome or gene density (Rn, Rg) are shown for each chromosome. Top20 upregulated and top20 downregulated segments are highlighted in green and red circles, respectively. Only chromosomes with top20 up/downregulated segments are present. A 3-fold coverage difference was observed in the last segment of chr14 (highlighted with star) mapping the IGH locus. The TRB locus at chr7 is also highlighted with star. Two typical cfDNA upregulated segments are highlited with green up-triangles, while two typical downregulated segments are highlighted with red down-triangles. **(b)** Heatmap visualization of the differential coverage at the IGH locus and the TRB locus. **(c)** The gDNA/cfDNA ratio at windows mapping nucleosome positive or negative regions, genic or intergenic regions, and protein-coding genes or non-coding genes. Mann–Whitney–Wilcoxon tests were performed to determine whether the difference in ratio was significant. ****: p<1e-4. **(d)** Heatmap showing typical examples of cfDNA upregulated segments (highlighted with green up-triangles), plotted together with GC-content (GC), protein-coding gene density (Pro), non-coding gene density (Non), nucleosome density (Nuc). Dash lines indicate segment borders. While gDNA is uniformly covered (evenly coloured with red), cfDNA is unevenly covered (blue, yellow, red intervals). **(e)** Heatmap showing typical examples of cfDNA downregulated segments (highlighted with red triangle), plotted together with GC-content (GC), protein-coding gene density (Pro), non-coding gene density (Non), nucleosome density (Nuc). Dash lines indicate segment borders. **(f)** Scatter plots showing the coverage mean and variance for chromosomes 1-6, calculated using the genome-wide 10kb-windows or segment representation.

chromosome remain relatively constant between cfDNA and gDNA. While levels of variances differ significantly between chromosomes with the genome-wide windows, the 1,090 segments well adjusts them to similar levels.

Distinctive convergence of segment coverage in ctDNA

We then compared segment coverage of ctDNA vs cfDNA to identify potential tumour-induced changes. Similarly, we identified top20 ctDNA (versus cfDNA) upregulated or downregulated segments (Figure 4a; Supplementary Table 1 and Supplementary Table 2). Significant differences in protein-coding gene density were found between the top20 upregulated and the top20 downregulated segments - the downregulated segments were enriched with protein-coding genes while the upregulated ones were sparse (Figure 4b). Moreover, we observed a further drop of coverages at genic regions or protein-coding genes and a further rise in coverages at intergenic regions or non-coding genes (Figure 4c), compared to the cfDNA vs gDNA differential coverage shown above. This suggested a decreased contribution by normal cells due to the replacement by other cell

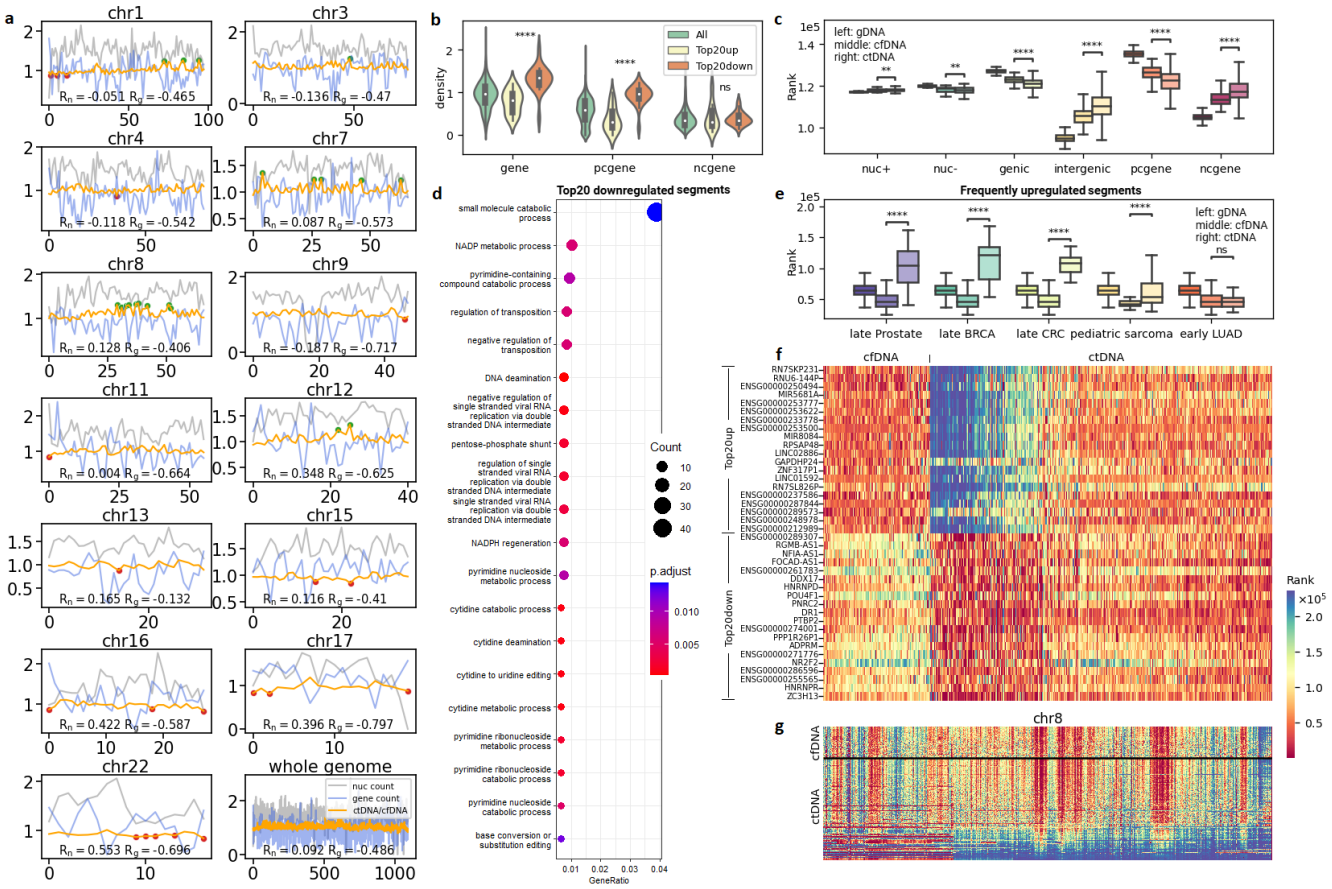

Figure 4. Differential coverage of ctDNA vs cfDNA. **(a)** Line plots of coverage rank ratio of ctDNA vs cfDNA on segments within different chromosomes, plotted together with nucleosome density (count) and gene density (count). The Pearson correlation coefficient between ctDNA/cfDNA ratio and nucleosome or gene density (Rn, Rg) are shown for each chromosome. Top20 upregulated and top20 downregulated segments are highlighted in green and red circles, respectively. Only chromosomes with top20 up/downregulated segments are present. Upregulated segments are significantly enriched at chr8. **(b)** Violin plot of gene density, protein-coding gene density and non-coding gene density in all segments, top20 upregulated segments and top20 downregulated segments. **(c)** A comparison of gDNA (n=3,202), cfDNA (n=113) and ctDNA (n=362) mean coverage at windows mapping nucleosome positive (short-named as nuc+) regions, nucleosome negative (nuc-) regions, genic regions, intergenic regions, protein-coding genes (pcgene) or non-coding genes (ncgene). Mann–Whitney–Wilcoxon tests were performed between ctDNA and cfDNA to determine whether differential coverage was significant. **: p<1e-2, ****: p<1e-4. **(d)** Significantly enriched gene ontology terms of biological process from genes within the top20 downregulated segments, FDR<0.05. **(e)** A comparison of gDNA, cfDNA and ctDNA mean coverage at three ctDNA frequently upregulated segments on multiple cancer datasets. **(f)** Differentially covered genes (FDR<0.01), with top20 upregulated and top20 downregulated shown. **(g)** Heatmap visualization on coverage landscape of ctDNA and cfDNA at chr8. The distinctive convergence of segment coverage is evident in ctDNA samples at the bottom.

sources with distinct activities, such as the tumour. We also observed a stronger negative correlation (R=-0.486, whole genome; Figure 4a; Supplementary Figure 6) between ctDNA/cfDNA ratio and gene density, compared to the negative correlation between cfDNA/gDNA ratio and gene density (R=-0.233, whole genome; Supplementary Figure 5). The top20 downregulated segments were significantly (FDR<0.05) enriched with genes associated with biological processes involving DNA/RNA mutation, nucleotide metabolism and NADPH metabolism (Figure 4d), which are closely related to cancers (29–31). Due to low gene density, no significant enrichment was found in the top20 upregulated segments. However, we identified three frequently upregulated segments with at least 10% increase in median coverage in all cancer datasets compared to (condition-matched) controls (Figure 4e; Supplementary Table 3). The frequently upregulated segments significantly (p<1e-4) distinguished ctDNA (prostate, breast, colorectal cancers and pediatric sarcoma) from cfDNA, except for early-stage lung cancer, which showed a non-significant trend (p=0.146). Similar to the concept of differentially expressed genes, here we defined differentially covered genes (see Methods). We identified 4,564 differentially covered genes with FDR<0.01 (Figure 4f; Supplementary Figure), in which 1,502 were upregulated and 3,062 were downregulated (Supplementary Table 5).

Of note, we observed a distinctive feature of ctDNA – the convergence of segment coverage during progression of cancer. A typical example is seen in the chromosome 8 (chr8), where ctDNA upregulated segments aggregated in the long arm (Figure 4a, 4g), consistent with the fact that chr8 gain is common in cancers (32, 33). The extent of segment convergence seemed to be positively associated with cancer stages (Supplementary Figure 7), where early-stage cancers may not show visible convergence while late-stage cancers often converged to only a few segments.

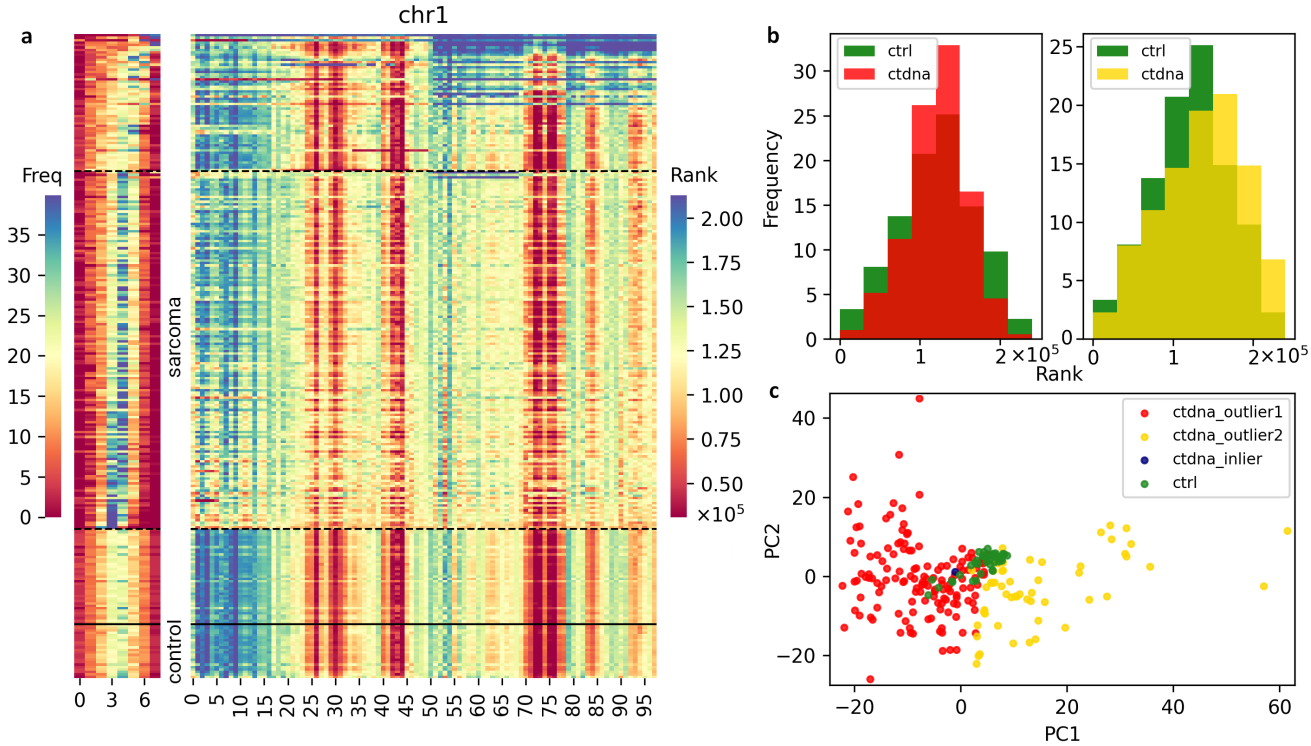

Figure 5. Outlier detection on coverage profiles, exampled with chr1. **(a)** Heatmap visualization of an 8-bin histogram (left panel) of segment coverage rank and the original segment coverage rank (right panel) at chr1 of ctDNA from pediatric sarcoma samples (n=241) versus condition-matched control cfDNA (n=22). Solid lines separate control and cancer samples. Dashed lines separate cancer samples exhibiting different directions of coverage convergence into three parts: convergence towards end-rank bins (top, correspond to the right panel of Figure 5c), convergence towards middle-rank bins (middle, correspond to the left panel of Figure 5c), no apparent convergence (bottom). **(b)** Differential rank histogram of sarcoma-derived vs control cfDNA. Convergence of segment coverage was observed, either towards the middle-bins (left panel) or towards the end-bins (right panel). **(c)** PCA plots for the visualization of ctDNA outliers (highlighted in red and yellow, colours correspond with Figure 5b) detected by Elliptic Envelope.

Measuring segment convergence for the detection of cancer

Finally, we examined the possibility of coverage-based cancer detection by calculating the segment convergence in chromosomes. From a chromosomal point of view, convergences can happen in any segments towards either directions – converging to higher or lower coverage ranks. Since the ranks are calculated genome-wide, changes of segment rank distribution in one chromosome would affect the segments in the other chromosomes. The segment rank distribution can be represented as a histogram, which partitions the ranks into bins. Figure 5 shows the comparison of an 8-bin segment rank histogram (alongside the segment heatmap) between ctDNA (pediatric sarcoma, n=241) and condition-matched control cfDNA (n=22) on chr1. From the comparative histogram, we observed a diverse spectrum of changes that could happen in ctDNA. In the most extreme cases, the whole chromosome or an arm converged to the highest or lowest bin; in most cases, segments belonging to the edge bins converged to the middle bins (Figure 5a-c).

Considering the position and direction of convergence is random, instead of looking for characteristics of the cancer samples, we transformed the question into a task of outlier detection. We assumed that the segment rank histograms of control cfDNA samples follow a multivariate Gaussian distribution. For each chromosome, we computed the 8-bin histogram, resulting in an 8-dim vector for each sample. The shape of an 8-dimensional Gaussian distribution is an ellipsoid, and we used the Elliptic Envelope (23) algorithm to first learn a model of the control elliptic distribution. Any points outside the ellipsoid were deemed outliers (Figure 5d). The model then can be used to predict whether a test 8-dim vector is outlier. For each test sample, we weighted and summarized the prediction results for all chromosomes (see Methods). When the percentage of segments flagged as outliers exceed a certain threshold (5% as default), the test sample is predicted abnormal.

The performances were assessed for two scenarios: prediction with or without condition-matched controls. Condition-matched controls are known control cfDNA samples sequenced together or under the same condition with the ctDNA samples. In this case we trained on the known controls and assessed the sensitivity and leave-one-out specificity of the model. When condition-matched controls were not available, condition-unmatched controls (i.e., controls from other studies) were used for

training; sensitivity and specificity were assessed (see Methods). Table 1 shows the performances of both scenarios (highlighted in red), with benchmarks from state-of-the-art fragmentomic-based (9) or mutation-based (1) methods. For the condition-matched tasks, our method achieved 97% sensitivity on pediatric sarcomas (n=241) with >90% leave-one-out specificity, 44% sensitivity on early lung cancers (n=36) with >90% leave-one-out specificity, outperforming current approaches. For the non-ideal, condition-unmatched tasks, our method achieved 100% sensitivity on late-stage cancers (n=85) with >95% specificity. Note that the control distribution still vary between studies, likely due to a variety of factors such as sequencing platform, experiment protocols and sequencing depth, although not as strong as the differences between control and most cancer cases (Supplementary Figures 8-11). The combined use of unmatched-control samples from several studies increased the tolerance to batch-effects, while sacrificing some sensitivity.

Discussion

In this study, we systemically compared genome-wide coverage profiles between gDNA samples from 1kGP (n=3,202) and cfDNA samples of healthy individuals (n=113) and cancer patients (n=362). We observed the phenomenon of coverage segmentation which was largely conserved among gDNA and cfDNA, where adjacent genomic regions share similar levels of coverage, forming segments that correspond to GC-content and protein-coding gene density or nucleosome density. We identified differential cfDNA vs gDNA coverage at immune-receptor locus, intergenic regions and non-coding genes, which were associated with differential genome activities from different cell sources. We also discovered an intensified coverage drop in genic regions or protein-coding genes and rise in intergenic regions or non-coding genes in ctDNA, indicating the increased proportion of tumour-derived DNA in the plasma. Moreover, we found the convergence of segment coverage as a distinctive characteristic during cancer development and progression. We demonstrated as a proof-of-concept an outlier-detection method to measure the segment convergence for cancer screening without using any cancer data for training, i.e., only trains on control cfDNA data. Our method achieved versatile performances and good generalizability to multiple independent datasets on both condition-matched and condition-unmatched tasks.

We uncovered biological insights hidden beneath coverage. Despite well known effects of GC-content to sequencing coverage, differential coverage caused by differences in protein-coding/non-coding genes or genic/intergenic regions were rarely discussed. Like expression profiles, our results demonstrated that the genome-wide coverage profiles are cell-type specific. Lymphocyte is the major cell type of PBMC, whereas neutrophil is the most abundant leukocyte in blood. Non-lymphocyte cells have intact IGH locus while lymphocytes have partial or full rearrangements. As a result, coverage at the IGH locus has a 3-fold difference between PBMC-derived gDNA and plasma cfDNA. Since the lymphocyte count is strongly associated with health status, the IGH coverage which is proportional to lymphocyte-derived cfDNA, could be further utilized as an indicator of diseases.

Several processing steps are key to the coverage profiling analyses: 1) the rank transformation of the coverage fraction, which allows integration of data from different studies. Although not commonly seen in DNA analysis, the rank transformation has been used in handling single-cell expression counts as a non-parametric method that circumvents the need of normalization (34). 2) The computation of coverage segment map to represent the genome, which extracts important features and de-noises the data. 3) Histogram binning of the coverage segment, allowing further dimensionality reduction. 4) Outlier detection, to detect outliers based on a known control distribution. This removes the dependence on cancer samples for training, which is considered more generalizable than conventional binary classifiers that requires cancer samples for training, because cancers are too heterogeneous that could not be fully explained by even thousands of samples.

Finally, to demonstrate the feasibility of using extracted features from genome-wide coverage profiles to detect early cancers, we chose the Elliptic Envelope algorithm (23) to perform ourlier detection. Other outlier/novelty detection approaches or algorithms could also fit the task, and may further improve the performances.

Data and Software Availability

All the data used within this study can be retrieved from public databases. Source code freely available at `https://github.com/deepomicslab/Genome-wide-coverage-profiling`.

Acknowledgments

Not applicable.

Conflict of interest statement. None declared.

References

1. Viktor A. Adalsteinsson, Gavin Ha, Samuel S. Freeman, Atish D. Choudhury, Daniel G. Stover, Heather A. Parsons, Gregory Gydush, Sarah C. Reed, Denisse Rotem, Justin Rhoades, Denis Loginov, Dimitri Livitz, Daniel Rosebrock, Ignaty Leshchiner, Jaegil Kim, Chip Stewart, Mara Rosenberg, Joshua M. Francis, Cheng Zhong Zhang, Ofir Cohen, Coyin Oh, Huiming Ding, Paz Polak, Max Lloyd, Sairah Mahmud, Karla Helvie, Margaret S. Merrill, Rebecca A. Santiago, Edward P. O'Connor, Seong H. Jeong, Rachel Leeson, Rachel M. Barry, Joseph F. Kramkowski, Zhenwei Zhang, Laura Polacek, Jens G. Lohr, Molly Schleicher, Emily Lipscomb, Andrea Saltzman, Nelly M. Oliver, Lori Marini, Adrienne G. Waks, Lauren C. Harshman, Sara M. Tolaney, Eliezer M. Van Allen, Eric P. Winer, Nancy U. Lin, Mari Nakabayashi, Mary Ellen Taplin, Cory M. Johannessen, Levi A. Garraway, Todd R. Golub, Jesse S. Boehm, Nikhil Wagle, Gad Getz, J. Christopher Love, and Matthew Meyerson. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nature Communications*, 8(1), 2017.
2. Joshua D Cohen, Lu Li, Yuxuan Wang, Christopher Thoburn, Ludmila Danilova, Christopher Douville, Ammar A Javed, Fay Wong, Austin Mattox, Ralph H Hruban, Christopher L Wolfgang, G Michael, Marco Dal Molin, Tian-li Wang, Richard Roden, and

Table 1. Performances of the proposed convergence-based outlier-detection method (highlighted in red) on various cancer-screening tasks.

Accession	Description	Condition-matched	Sensitivity	Specificity*
EGAD00001007080	Pediatric sarcoma (n=241) with condition-matched controls (n=22)	Yes	0.97	0.91
		No	0.68	0.95
	LIQUORICE (fragmentomics-based benchmark)	Yes	0.81	N/A
		No	0.53	0.36
	IchorCNA (mutation-based benchmark)	No	0.53	0.91
EGAD00001006237	Early lung cancer (n=36) with condition-matched controls (n=38)	Yes	0.44	0.92
		No	0.25	0.71
	LIQUORICE (fragmentomics-based benchmark)	Yes	0.11	N/A
		No	0.06	0.92
	IchorCNA (mutation-based benchmark)	No	0.08	1.0
EGAD00001005343	Late-stage cancers (n=4) with condition-matched controls (n=24)	Yes	1.0	0.96
		No	1.0	1.0
EGAD00001006421	Healthy controls (n=29)	No	N/A	1.0
EGAD00001006420	Late-stage breast cancer (n=9)	No	1.0	N/A
EGAD00001006419	Late-stage colon cancer (n=12)	No	1.0	N/A
EGAD00001008460	Late-stage prostate cancer (n=60)	No	1.0	N/A

*For condition-matched tasks, leave-one-out specificity was calculated.

P Alison. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, 359(6378):926–930, 2018.

3. Laura Keller, Yassine Belloum, Harriet Wikman, and Klaus Pantel. Clinical relevance of blood-based ctDNA analysis: mutation detection and beyond. *British Journal of Cancer*, 124(2):345–358, 2021.
4. Siddhartha Jaiswal and Benjamin L. Ebert. Clonal hematopoiesis in human aging and disease. *Science*, 366(6465):1–20, 2019.
5. Matthew W. Snyder, Martin Kircher, Andrew J. Hill, Riza M. Daza, and Jay Shendure. Cell-free DNA Comprises an in Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell*, 164(1-2):57–68, 2016.
6. Peter Ulz, Gerhard G. Thallinger, Martina Auer, Ricarda Graf, Karl Kashofer, Stephan W. Jahn, Luca Abete, Gunda Pristauz, Edgar Petru, Jochen B. Geigl, Ellen Heitzer, and Michael R. Speicher. Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nature Genetics*, 48(10):1273–1278, 2016.
7. Stephen Cristiano, Alessandro Leal, Jillian Phallen, Jacob Fiksel, Vilmos Adleff, Daniel C. Bruhm, Sarah Østrup Jensen, Jamie E. Medina, Carolyn Hruban, James R. White, Doreen N. Palsgrove, Noushin Niknafs, Valsamo Anagnostou, Patrick Forde, Jarushka Naidoo, Kristen Marrone, Julie Brahmer, Brian D. Woodward, Hatim Husain, Karlijn L. van Rooijen, Mai Britt Worm Ørntoft, Anders Husted Madsen, Cornelis J.H. van de Velde, Marcel Verheij, Annemieke Cats, Cornelis J.A. Punt, Geraldine R. Vink, Nicole C.T. van Grieken, Miriam Koopman, Remond J.A. Fijneman, Julia S. Johansen, Hans Jørgen Nielsen, Gerrit A. Meijer, Claus Lindbjerg Andersen, Robert B. Scharpf, and Victor E. Velculescu. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*, 570(7761):385–389, 2019.
8. Peiyong Jiang, Kun Sun, Wenlei Peng, Suk Hang Cheng, Meng Ni, Philip C. Yeung, Macy M.S. Heung, Tingting Xie, Huimin Shang, Ze Zhou, Rebecca W.Y. Chan, John Wong, Vincent W.S. Wong, Liona C. Poon, Tak Yeung Leung, W. K.Jacky Lam, Jason Y.K. Chan, Henry L.Y. Chan, K. C.Allen Chan, Rossa W.K. Chiu, and Y. M.Dennis Lo. Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. *Cancer Discovery*, 10(5):664–673, 2020.
9. Peter Peneder, Adrian M. Stütz, Didier Surdez, Manuela Krumbholz, Sabine Semper, Mathieu Chicard, Nathan C. Sheffield, Gaelle Pierron, Eve Lapouble, Marcus Tötzl, Bekir Ergüner, Daniele Barreca, André F. Rendeiro, Abbas Agaimy, Heidrun Boztug, Gernot Engstler, Michael Dworzak, Marie Bernkopf, Sabine Taschner-Mandl, Inge M. Ambros, Ola Myklebost, Perrine Marec-Bérard, Susan Ann Burchill, Bernadette Brennan, Sandra J. Strauss, Jeremy Whelan, Gudrun Schleiermacher, Christiane Schaefer, Uta Dirksen, Caroline Hutter, Kjetil Boye, Peter F. Ambros, Olivier Delattre, Markus Metzler, Christoph Bock, and Eleni M. Tomazou. Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden. *Nature Communications*, 12(1):1–16, 2021.
10. David Sims, Ian Sudbery, Nicholas E. Ilott, Andreas Heger, and Chris P. Ponting. Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121–132, 2014.
11. Daniel C. Koboldt. Best practices for variant calling in clinical sequencing. *Genome Medicine*, 12(1):1–13, 2020.
12. Martin S. Lindner, Maximilian Kollock, Franziska Zickmann, and Bernhard Y. Renard. Analyzing genome coverage profiles with applications to quality control in metagenomics. *Bioinformatics*, 29(10):1260–1267, 2013.
13. Marta Byrska-Bishop, Uday S. Evani, Xuefang Zhao, Anna O. Basile, Haley J. Abel, Allison A. Regier, André Corvelo, Wayne E. Clarke, Rajeeva Musunuri, Kshithija Nagulapalli, Susan Fairley, Alexi Runnels, Lara Winterkorn, Ernesto Lowy, Evan E. Eichler, Jan O. Korbel, Charles Lee, Tobias Marschall, Scott E. Devine, William T. Harvey, Weichen Zhou, Ryan E. Mills, Tobias Rausch, Sushant Kumar, Can Alkan, Fereydoun Hormozdiari, Zechen Chong, Yu Chen, Xiaofei Yang, Jiadong Lin, Mark B. Gerstein, Ye Kai, Qihui Zhu, Feyza Yilmaz, Chunlin Xiao, Paul Flicek, Soren Germer, Harrison Brand, Ira M. Hall, Michael E. Talkowski, Giuseppe Narzisi, and Michael C. Zody. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, 185(18):3426–3440.e19, 2022.
14. Peter Ulz, Samantha Perakis, Qing Zhou, Tina Moser, Jelena Belic, Isaac Lazzeri, Albert Wölfler, Armin Zebisch, Armin Gerger, Gunda Pristauz, Edgar Petru, Brandon White, Charles E.S. Roberts, John St John, Michael G. Schimek, Jochen B. Geigl, Thomas Bauernhofer, Heinz Sill, Christoph Bock, Ellen Heitzer, and Michael R. Speicher. Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. *Nature Communications*, 10(1), 2019.
15. Asaf Zviran, Rafael C. Schulman, Minita Shah, Steven T.K. Hill, Sunil Deochand, Cole C. Khamnei, Dillon Maloney, Kristofer Patel, Will Liao, Adam J. Widman, Phillip Wong, Margaret K. Callahan, Gavin Ha, Sarah Reed, Denisse Rotem, Dennie Frederick, Tatyana Sharova, Benchun Miao, Tommy Kim, Greg Gydush, Justin Rhoades, Kevin Y. Huang, Nathaniel D. Omans, Patrick O. Bolan, Andrew H. Lipsky, Chelston Ang, Murtaza Malbari, Catherine F. Spinelli, Selena Kazancioglu, Alexi M. Runnels, Samantha Fennessey, Christian Stolte, Federico Gaiti, Giorgio G. Inghirami, Viktor Adalsteinsson, Brian Houck-Loomis, Jennifer Ishii, Jedd D. Wolchok, Genevieve Boland, Nicolas Robine, Nasser K. Altorki, and Dan A. Landau. Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nature Medicine*, 26(7):1114–1124, 2020.
16. Guanhua Zhu, Yu A. Guo, Danliang Ho, Polly Poon, Zhong Wee Poh, Pui Mun Wong, Anna Gan, Mei Mei Chang, Dimitrios Kleftogiannis, Yi Ting Lau, Brenda Tay, Wan Jun Lim, Clarinda Chua, Tira J. Tan, Si Lin Koo, Dawn Q. Chong, Yoon Sim Yap, Iain Tan, Sarah Ng, and Anders J. Skanderup. Tissue-specific cell-free DNA degradation quantifies circulating tumor DNA burden. *Nature Communications*, 12(1):1–11, 2021.
17. Cameron Herberts, Matti Annala, Joonatan Sipola, Sarah W.S. Ng, Xinyi E. Chen, Anssi Nurminen, Olga V. Korhonen, Aslı D. Munzur, Kevin Beja, Elena Schönlau, Cecily Q. Bernales, Elie Ritch, Jack V.W. Bacon, Nathan A. Lack, Matti Nykter, Rahul Aggarwal, Eric J. Small, Martin E. Gleave, David A. Quigley, Felix Y. Feng, Kim N. Chi, and Alexander W. Wyatt. Deep whole-genome ctDNA chronology of treatment-resistant prostate cancer. *Nature*, 608(7921):199–208, 2022.
18. Christian Freksa and Ralf Rohrig. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 11 (1110):483–492, 2014.
19. Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, Genome Project, and Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *BIOINFORMATICS APPLICATIONS NOTE*, 25(16):2078–2079, 2009.
20. Chun Xiao Song, Senlin Yin, Li Ma, Amanda Wheeler, Yu Chen, Yan Zhang, Bin Liu, Junjie Xiong, Weihan Zhang, Jiankun Hu, Zongguang Zhou, Biao Dong, Zhiqi Tian, Stefanie S. Jeffrey, Mei Sze Chua, Samuel So, Weimin Li, Yuquan Wei, Jiajie Diao, Dan Xie, and Stephen R. Quake. 5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages. *Cell Research*, 27(10):1231–1242, 2017.
21. Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167, 2020.
22. Guangchuang Yu, Li Gen Wang, Yanyan Han, and Qing Yu He. ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology*, 16(5):284–287, 2012.
23. Stephen Howard. The Elliptical Envelope. pages 1–10, 2007.
24. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand

Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

25. Yuval Benjamini and Terence P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10):1–14, 2012.
26. Gita Kumari, Tatiana Gerasimova, Hansen Du, Supriyo De, William H. Wood, Kevin G. Becker, and Ranjan Sen. Misregulation of the IgHLocus in thymocytes. *Frontiers in Immunology*, 9(NOV):1–12, 2018.
27. Ondrej Krejci, Zuzana Prouzova, Ondrej Horvath, Jan Trka, and Ondrej Hrusak. Cutting Edge: TCR δ Gene Is Frequently Rearranged in Adult B Lymphocytes. *The Journal of Immunology*, 171(2):524–527, 2003.
28. Kerstin Puellmann, Wolfgang E. Kaminski, Mandy Vogel, C. Thomas Nebe, Josef Schroeder, Hans Wolf, and Alexander W. Beham. A variable immunoreceptor in a subpopulation of human neutrophils. *Proceedings of the National Academy of Sciences of the United States of America*, 103(39):14441–14446, 2006.
29. Nicholas J. Mullen and Pankaj K. Singh. Nucleotide metabolism: a pan-cancer metabolic dependency. *Nature Reviews Cancer*, 23(5):275–294, 2023.
30. Huai Qiang Ju, Jin Fei Lin, Tian Tian, Dan Xie, and Rui Hua Xu. NADPH homeostasis in cancer: functions, mechanisms and therapeutic implications. *Signal Transduction and Targeted Therapy*, 5(1):1–12, 2020.
31. Tracey W. Chan, Ting Fu, Jae Hoon Bahn, Hyun Ik Jun, Jae Hyung Lee, Giovanni Quinones-Valdez, Chonghui Cheng, and Xinshu Xiao. RNA editing in cancer impacts mRNA abundance in immune response pathways. *Genome Biology*, 21(1):1–25, 2020.
32. Ronald R. De Krijger, Sandra M.H. Claessen, Frieda Van Der Ham, Ad J.M. Van Unnik, Christina A. Hulsbergen-Van De Kaa, Leen Van Leuven, Max Van Noesel, and Ernst J.M. Speel. Gain of chromosome 8q is a frequent finding in pleuropulmonary blastoma. *Modern Pathology*, 20(11):1191–1199, 2007.
33. Mohammed Alshalalfa, Tiffany T. Nguyen, Konrad H. Stopsack, Anwar Khan, Idalid Franco, Crystal Seldon, Nishwant Swami, William Jin, Karthik Meiyappan, Minh Ton, Jeffrey M. Venstrom, Edward Christopher Dee, and Brandon A. Mahal. Chromosome 8q arm overexpression is associated with worse prostate cancer prognosis. *Urologic Oncology: Seminars and Original Investigations*, 41(2):106.e17–106.e23, 2023.
34. Alexander H.S. Vargo and Anna C. Gilbert. A rank-based marker selection method for high throughput scRNA-seq data. *BMC Bioinformatics*, 21(1):1–51, 2020.