Computational and Structural Biotechnology Journal 23 (2024) 3163–3174

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

LBFextract: Unveiling transcription factor dynamics from liquid biopsy data

Isaac Lazzeri [a,*], Benjamin Gernot Spiegl [a], Samantha O. Hasenleithner [b,*], Michael R. Speicher [a,c,1,2], Martin Kircher [d,e,**,1]

[a] Institute of Human Genetics, Diagnostic and Research Center for Molecular BioMedicine, Medical University of Graz, Neue Stiftingtalstr. 6, Graz 8010, Austria
[b] Division of Oncology, Department of Internal Medicine, Medical University of Graz, 8010 Graz, Austria
[c] BioTechMed-Graz, Graz, Austria
[d] Berlin Institute of Health (BIH) at Charité – Universitätsmedizin Berlin, Berlin 10178, Germany
[e] Institute of Human Genetics, University Medical Center Schleswig-Holstein, University of Lübeck, Lübeck, Germany

ARTICLE INFO

Keywords:
Cell-free DNA
Bioinformatics
Whole-genome sequencing
Transcription factors
Fragmentomics

ABSTRACT

Motivation: The analysis of circulating cell-free DNA (cfDNA) holds immense promise as a non-invasive diagnostic tool across various human conditions. However, extracting biological insights from cfDNA fragments entails navigating complex and diverse bioinformatics methods, encompassing not only DNA sequence variation, but also epigenetic characteristics like nucleosome footprints, fragment length, and methylation patterns.

Results: We introduce Liquid Biopsy Feature extract (LBFextract), a comprehensive package designed to streamline feature extraction from cfDNA sequencing data, with the aim of enhancing the reproducibility and comparability of liquid biopsy studies. LBFextract facilitates the integration of preprocessing and postprocessing steps through alignment fragment tags and a hook mechanism. It incorporates various methods, including coverage-based and fragment length-based approaches, alongside two novel feature extraction methods: an entropy-based method to infer TF activity from fragmentomics data and a technique to amplify signals from nucleosome dyads. Additionally, it implements a method to extract condition-specific differentially active TFs based on these features for biomarker discovery. We demonstrate the use of LBFextract for the subtype classification of advanced prostate cancer patients using coverage signals at transcription factor binding sites from cfDNA. We show that LBFextract can generate robust and interpretable features that can discriminate between different clinical groups. LBFextract is a versatile and user-friendly package that can facilitate the analysis and interpretation of liquid biopsy data.

Data and Code Availability and Implementation: LBFextract is freely accessible at https://github.com/Isy89/LBF. It is implemented in Python and compatible with Linux and Mac operating systems. Code and data to reproduce these analyses have been uploaded to 10.5281/zenodo.10964406.

1. Introduction

Analyses of circulating cell-free DNA (cfDNA), i.e. the analysis of naturally occurring short DNA fragments in bodily fluids like blood and urine, are increasingly being adopted for the identification, diagnostic assessment and surveillance of various pathological and physiological states in humans [1–5]. This gain in traction can be attributed to many factors that are fueling the growth of the cfDNA field, such as the increasing prevalence of cancer [6], rising preference for non-invasive procedures, various advantages of liquid biopsies, i.e. diagnostic approaches using samples of bodily fluids, over standard tissue biopsies, favorable government initiatives, and growing public and private interest. However, the bioinformatics approaches to harvesting the inherent biological information from cfDNA fragments are becoming more sophisticated and complex. The cfDNA field has begun to extend beyond the analysis of observed DNA sequence variation, such as single

* Corresponding authors.
** Corresponding author at: Berlin Institute of Health (BIH) at Charité – Universitätsmedizin Berlin, Berlin 10178, Germany.
E-mail addresses: isaac.lazzeri@medunigraz.at (I. Lazzeri), samantha.hasenleithner@medunigraz.at (S.O. Hasenleithner), martin.kircher@bih-charite.de (M. Kircher).
[1] Contributed equally
[2] Posthumous authorship

https://doi.org/10.1016/j.csbj.2024.08.007
Received 6 June 2024; Received in revised form 9 August 2024; Accepted 9 August 2024
Available online 11 August 2024

nucleotide variants (SNVs) and somatic copy number alterations (SCNAs). For example, several studies employing nucleosome position mapping have now provided evidence that cfDNA reflects nucleosome footprints and coverage profiles at regulatory regions such as transcription start sites (TSS) and transcription factor binding sites (TFBS) have been employed for the inference of gene expression, cancer detection and tissue deconvolution [7–10]. Various works have shown that the cellular nucleosomal architecture significantly influences DNA fragmentation, creating distinct patterns in not only the length of the fragments, but also the frequency and type of specific motifs, which have also been used for cancer detection and classification [11–17]. The extraction of epigenetic alterations, such as cfDNA fragment length, diverse cfDNA fragment patterns, methylation markers, and signals at open chromatin regions, will play an important role in the development of more advanced liquid biopsy technologies [18,19]. Although this development of cfDNA-derived features is rapidly evolving and individual methods of feature extraction involve similar steps, a package that offers a straightforward and extendable collection of feature extraction methods that enables experiment reproducibility and comparability is currently lacking. A variety of bioinformatics tools are available to calculate diverse types of genomic coverage measures, spanning from bin-wise genomic read coverage like BEDtools [40,41] to region-specific fragment coverage like DeepTools [36]. In the liquid biopsy field, new methods aimed at enhancing the signal derived from nucleosome dyads have been proposed [9,14,20–22]. However, these specialized methods are mostly provided as part of workflows, intertwined with preprocessing and GC bias correction steps, which hinders their reusability. In Liquid Biopsy Feature extract (LBFextract), we provide diverse feature extraction methods based on whole-genome sequencing (WGS) coverage. Further, we provide an easy way to integrate GC bias correction methods in the form of alignment fragment tags, compatible with software like GCparagon [23], while remainingtool-agnostic, thus making this step uncoupled from the LBFextract feature extraction methods. To this end, in LBFextract, we implement commonly used coverage strategies like midpoint coverage, used in the Griffin package [21], and read coverage, proposed in [9], but generalizing them with a user-defined number of bases to retain from each fragment. We further implement a new method, which we call coverage around dyads (CAD), with the aim of enhancing the signal derived from nucleosome dyads through a better modelling of nucleosome position on DNA fragments. We also included several feature extraction methods to extract different types of fragment length distributions (FLD) and fragment length ratios (FLR), which provide orthogonal information to coverage-based features as well as new entropy-based fragmentomics features. Through a hook mechanism, we provide entry points to integrate extra pre- and postprocessing steps, i.e., plugins, allowing to use third-party software to customize the way reads are collected, how they are transformed, the process of signal extraction as well as the way genomic ranges are normalized. Finally, LBFextract provides a way to identify condition-specific statistically significant transcription factor (TF) signals and their enrichment analysis, enabling condition-specific biomarker discovery from liquid biopsy WGS data. In the first part of this article, we present the package structure, followed by a description of the feature extraction methods. In the final part, we provide a demonstration of a clinical use case of this package for the extraction of coverage signals at TFBSs for subtype classification in the context of advanced prostate cancer (PC) (Fig. 1).

2. Materials and methods

2.1. Package structure

LBFextract has been developed as a plugin system (Supp. Fig. 1) in which several hooks define entry points that a user can use to customize the workflow without having to reimplement code or functionality that the package or other plugins already implement. To achieve this, it uses Pluggy [24], a python package that allows the user to change the behavior of the host program. In this context, the modification of the behavior is defined as python functions called hooks, which are loaded and registered at runtime and change or exchange certain parts of the host program. We developed two types of hooks: Command line interface (CLI) hooks and workflow-specific hooks. Using the CLI hooks, a user can implement CLI-plugin commands that are registered at installation time and are made available through the CLI as LBFextract subcommands. The workflow-specific hooks allow the customization of different steps of the default workflow. Specifically, we implemented the following hooks: fetch_reads, save_reads, load_reads, transform_reads, transform_single_intervals, transform_all_intervals, save_signal, plot_signal and save_plot. The read fetching hook handles binary sequence alignment map (BAM) files, generally done by retrieving specific regions of interest defined in one or multiple browser extensible data (BED) files. As most feature extraction methods rely both on the start and end information of DNA fragments, the current version of LBFextract supports only paired-end WGS data, provided as BAM files. LBFextract is sequencing platform-agnostic and supports any BAM file generated by aligners that adhere to the SAM format specification and implement the observed template length as TLEN#1. The way reads are saved and loaded is defined by the save_reads and the load_reads hooks. Signal extraction methods are implemented as transform_single_intervals hooks, which handles the extraction of the signal in each region defined in the BED file. The transform_all_intervals hook is available for transformations requiring all genomic intervals as an input. The save_signal hook defines the way extracted signals are stored. Finally, signal-specific plots can be defined in the plot_signal hook and saved with the save_plot hook.

2.2. Feature extraction methods

LBFextract defines a set of feature extraction methods, which can be divided into coverage and fragmentomics-based methods. Recent work has described diverse types of coverage signals that can be derived from cfDNA. Here, we implemented coverage (fragment coverage, midpoint coverage, middle-n points coverage, coverage around the dyad, sliding coverage, central 60 bp-coverage) as well as fragmentomics signals (Windowed Protection Score (WPS), FLD and FLR). A central aspect of LBFextract is the introduction of a novel feature extraction method defined as coverage around dyads (CAD) and novel entropy-derived features such as entropy and relative fragment entropy (RFE).

To better describe the feature extraction methods, we introduce a mathematical notation for a BED file, DNA fragments and DNA fragments relative to a genomic interval.

We define a BED file as a multiset of genomic intervals g_i:

$$G = \{g : g = [g_s, g_e) : g_s, g_e \in Z^+ \cup 0, g_s < g_e\} \quad (1)$$

where g_s, g_e represent the start and end of a genomic interval g.

and the set of all fragments present in a sample as F:

$$F = \{f : f = [f_s, f_e) : f_s, f_e \in Z^+ \cup 0, f_s < f_e\} \quad (2)$$

where f_s and f_e represent the start and end of fragments respectively. Further, let g be a target region, we define f^g as the set of all positions of fragment f overlapping interval g expressed relative to interval g, which can be defined as:

$$f^g = \{i : \quad i - g_s \quad \forall \mathrm{i} \in \mathrm{f} \cap \mathrm{g}\} \quad (3)$$

and F^g as the multiset of all f^g overlapping an interval g

$$F^g = \{f^g : \forall f \in F\} \quad (4)$$

Hereafter, all intervals g will be considered to have the same length defined as w, which is by default set to 4000, which corresponds to $\pm$ 2000 bp around the TFBS center.

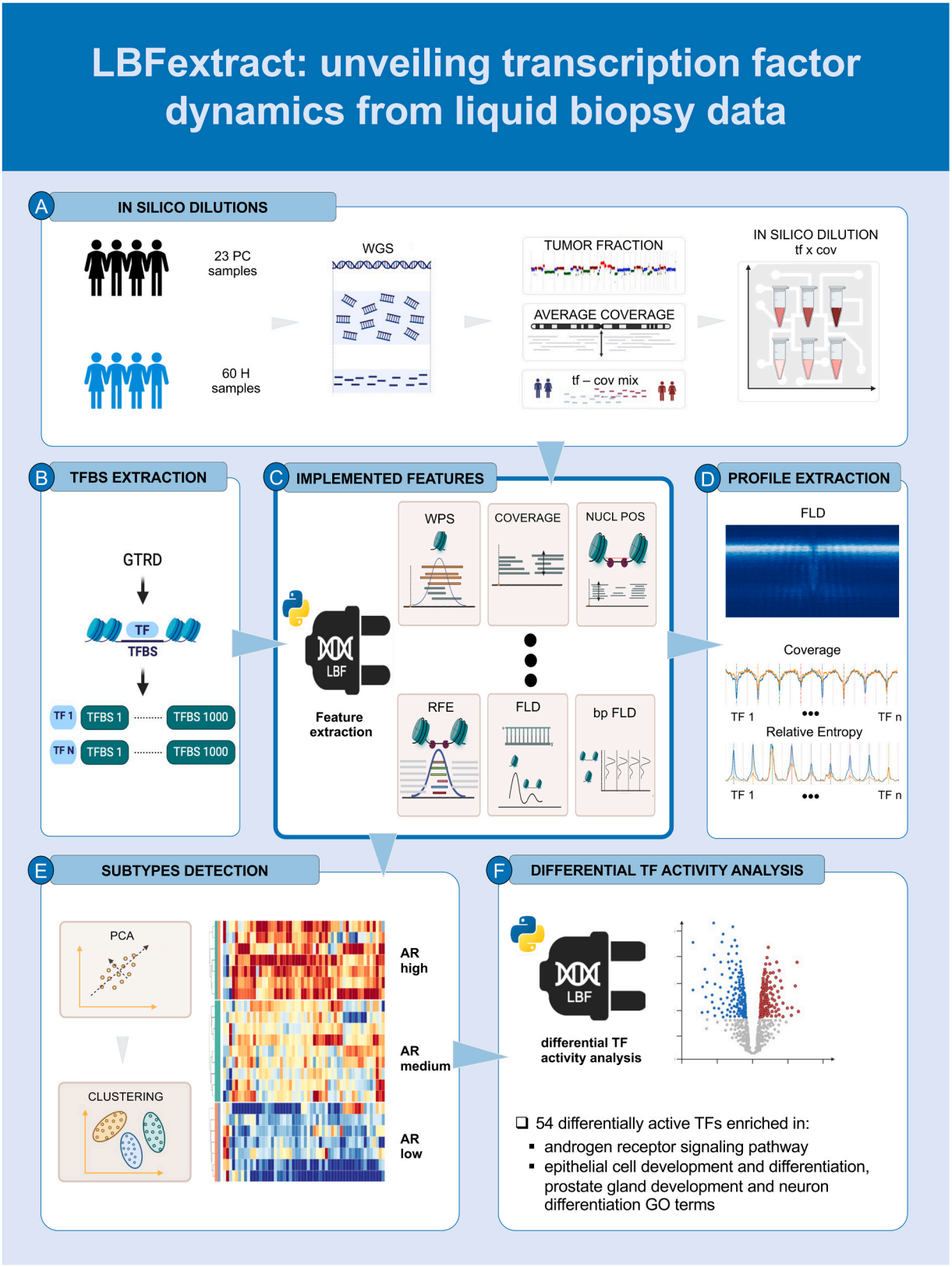

Fig. 1. Workflow showing the application of LBFextract for prostate cancer subtype biomarker discovery. A) Prostate cancer (PC) samples are diluted to 20x coverage and 20 % tumor fraction (Supplementary Info). B) TFBSs per TF are retrieved by the GTRD database. C) Coverage- and fragmentomics-based features are extracted from cfDNA WGS data using LBFextract. D) Example of features extracted with LBFextract. E) Pre-defined set of features per TF are used and down projected to a lower dimensional space using PCA. This is followed by clustering to extract PC subtypes. F) Differential TF binding activity analyses are performed to extract condition-specific TFs. Further enrichment analyses are performed to place the detected differentially active TFs into context.

2.2.1. Coverage

In fragment coverage, information concerning paired reads is used to infer the length of a fragment and the coverage at a specific position is defined as the number of fragments overlapping a given position. Here, positions covered by a fragment are those spanning from the left-most start to the right-most end of two mates in a read pair. This is different from read coverage, which considers positions of sequenced nucleotides, without taking overlaps or missing segments between the paired-end reads into account, thus possibly introducing an artifact. For any $0 \leq l \leq w$, define c_l^g to be the coverage at position l in genomic interval g relative to interval g defined as:

$$c_l^g = \sum_{fg \in Fg} 1_{l \in fg} \tag{5}$$

and $\boldsymbol{c}^g \in R^w$ to be the coverage of interval g defined as:

$$\boldsymbol{c}^g = \left(c_l^g\right)_{l=0}^{w-1} \tag{6}$$

the coverage profile for all regions in a BED file can be described as:

$$\boldsymbol{c} = \left(\frac{1}{|G|}\sum_{g \in G} c_l^g\right)_{l=0}^{w-1} \tag{7}$$

in which $\boldsymbol{c} \in R^w$.

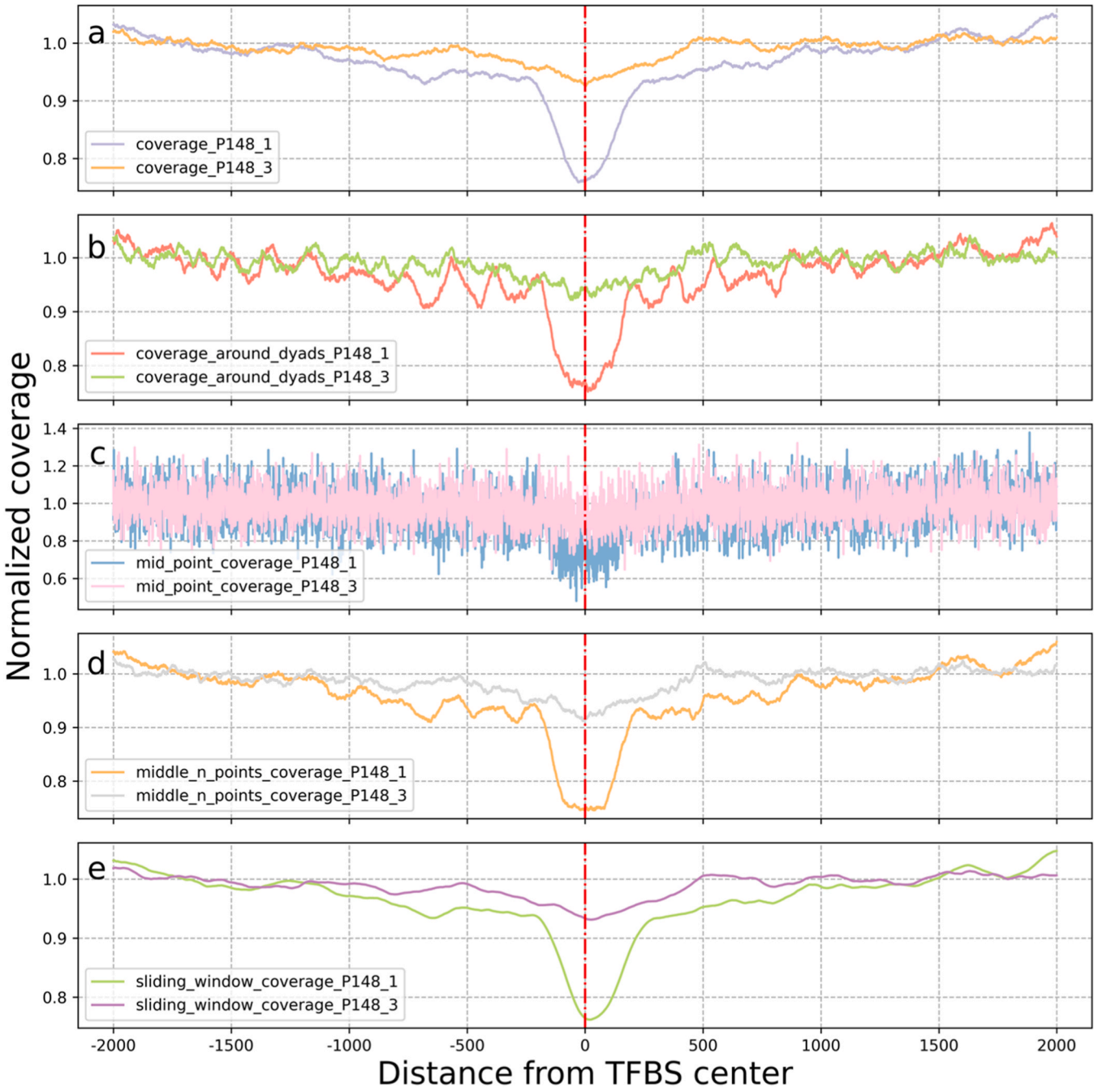

Fig. 2. Comparison of coverage signals of cfDNA samples obtained for the same individual at two time points (P148_1 and P148_3). a) Normal coverage b) Coverage around dyads c) Midpoint coverage d) Middle-n points coverage e) Sliding window coverage. The figure shows the respective overlays of 1000 sites for the AR TF and highlights a difference between these samples around the central location of the considered windows, suggesting regions of open chromatin in P148_1 that are absent in P148_3.

Midpoint coverage is calculated as the number of fragments having their central position located at a certain genomic position. If $\lfloor \frac{(f_e - f_s)}{2} \rfloor - g_s \in [0, w)$ then define the midpoint to be:

$$m_f^g = \lfloor \frac{(f_e - f_s)}{2} \rfloor - g_s \tag{8}$$

Then, formula 5 changes to:

$$c_l^g = \sum_{f \in F \; l = m_f^g} 1 \tag{9}$$

This measure might be problematic, as coverage at each position is dramatically decreased by what could be considered an extreme *in silico* trimming of reads. Instead of midpoint coverage, *n* positions around the middle of each fragment might be used, which in LBFextract is called middle-n points coverage. To define this, let n be the number of positions around m and we can then describe $\widehat{f}^g$ as the subset of f representing the positions around the midpoint of f and $\widehat{F}^g$ the multiset of all $\widehat{f}^g$ mapping to a genomic interval g:

$$\widehat{f}_n^g = \left[m_f^g - n, \quad m_f^g + n \right) \tag{10}$$

$$\widehat{F}_n^g = \{ \widehat{f}_n^g : \quad \forall f \in F \quad \} \tag{11}$$

Then, the middle-n points coverage at position l with respect to genomic interval g can be defined as:

$$c_l^g = \sum_{\widehat{f}_n^g \in \widehat{F}_n^g} 1_{l \in \widehat{f}_n^g} \tag{12}$$

Midpoint coverage and middle-n points coverage have been widely used to extract information concerning nucleosome positioning [21]. The middle points are assumed to be the location of the nucleosome dyads, which represent the most protected part of the DNA wrapped around nucleosomes. While this might generally be true for fragments shorter than 250 base pairs (bp), it is not true for longer fragments, i.e. > 250 bp. Indeed, the fragment midpoint of poly-nucleosomal structures like di-nucleosomes falls within the unprotected region between two nucleosomes, interfering with proper nucleosome dyad localization (Supplementary Methods). To avoid this problem, we implemented "coverage around dyads" (Supp. Fig. 2). This method takes into consideration the presence of poly-nucleosomal structures and models the probability of each read coming from a n or n + 1 poly-nucleosomal structure (Supplementary Information, Algorithm 1 lines 7–8). This is used to reconstruct the size of each fragment before degradation (Supplementary Information, Algorithm 1 lines 13–17), which in turn is used to better place the position of the dyad and obtain a stronger nucleosome-derived signal (Supplementary Information, Algorithm 1 lines 18–22).

The fragment size of mono-nucleosomal-derived cfDNA is determined from the FLD in the pre-defined region chr12:34300000–34500000. This region, located at the centromeric portion of chromosome 12, is characterized by highly phased nucleosomes (Supp. Figure 8), the absence of TSS, as per the GRCh38.p13 annotation file (release 38) provided by the GENCODE project, and the presence of only 16 TFBS (Supp. Figure 8). Consequently, the influence of open chromatin regions on the FLD shape is expected to be minimal. Additionally, the presence of phased nucleosomes in this region has been previously shown [7,8,25,26].

Sliding coverage is beneficial when the average depth of coverage is low, as it helps to smooth the signal and mitigate the impact of artifacts, such as drops in coverage at individual positions. This calculation involves applying a moving average over each genomic interval to determine the coverage at each position. Special consideration is given to the edges of the coverage vector, where the size of the sliding window must be adjusted accordingly to prevent exceeding the vector boundaries (i.e. when $l = w - 1$, n needs to be adjusted to 1). Finally, the sliding coverage for all genomic intervals in a BED file is averaged at each position. This process can be mathematically described as follows:

$$\mathbf{c}_{sliding}^{n} = \left(\frac{1}{|G|} \sum_{g \in G} \frac{1}{\min(w-l, n) + 1} \sum_{i=0}^{\min(w-l,\ n)} c_{i+l}^g \right)_{l=0}^{w-1} \tag{13}$$

where n is the window parameter with a default of 4 bp.

Central 60 bp-coverage (Supp. Fig. 2) introduced in [9], trims 53 bases from both fragment sides and uses 60 bp from each side for coverage calculation. We generalized this, introducing two variables (default [53,113]) describing the range of bases that should be retained.

Profiles of each interval defined in the BED files are further normalized to the mean coverage of the flanking regions. Let r be the length of the flanking region, then the normalization step can be defined as follows:

$$\mathbf{c}_{norm}^{g} = \left(\frac{\mathbf{c}_l * r * 2}{\sum_{j=0}^{r-1} \mathbf{c}_j + \sum_{j=w-1\ -r}^{w-1} \mathbf{c}_j} \right)_{l=0}^{w-1} \tag{14}$$

2.2.2. Windowed Protection Score (WPS)

Further, we implemented the WPS introduced in [7], which quantifies the protective effect of nucleosomes in a genomic region by assigning a score to each position. This score is determined by the count of fragments that entirely cover the span of a window centered on a genomic position, subtracted by the number of fragments that begin/-terminate within that window.

This can be represented as follows:

$$\boldsymbol{wps} = \left(\frac{1}{|G|} \sum_{g \in G} \boldsymbol{wps}_l^g \right)_{l=0}^{w} \tag{15}$$

$$\boldsymbol{wps}_l^g = \quad (\boldsymbol{c}_l^g - \boldsymbol{s}_l^g) - \widetilde{\boldsymbol{c}_l^g} \tag{16}$$

in which $\boldsymbol{wps} \in R^w$, $\boldsymbol{c} \in R^w$ is the coverage vector for genomic interval g calculated using a multiset of trimmed fragments corresponding to: $\{ f : (f_s + v,\ f_e - v)\ \forall f \in F \}$, w is the length of the genomic interval by default set to 4000, $\widetilde{\boldsymbol{c}_i^g}$ is the running median coverage vector, and $\boldsymbol{s}^g$ defined as:

$$\boldsymbol{s}^g = \left(\sum_{\substack{x \in T_v^g \\ l \in x}} \mathbb{1} \right)_{l=0}^{w-1} \tag{17}$$

is the vector with the coverage of the genomic intervals corresponding to a window v surrounding the start and the end of each fragment, which are represented by the multiset T:

$$T_v^g = \{ \quad t_v^g : ((f_s - v, \quad f_s + v) \cup (f_e - v, \quad f_e + v)) \quad \cap g \forall \quad f \in F \} \tag{18}$$

Profiles of each interval defined in the BED files are further normalized as done for the coverage-based signals (Formula 14).

2.2.3. Fragment length distribution

Recently, there has been a growing utilization of fragmentomics features [26,27]. For example, the ratio between long and short fragments has been used for cancer prediction [11,13] or for determining the proportion of placental cfDNA [28]. To improve upon these methods, we provide the possibility to extract the full distribution of fragment lengths per position given the genomic intervals in one or multiple BED files. To

calculate this, for all fragments F in BED file G, the fragment length distribution at position i $\boldsymbol{d}_i$ can be defined as:

$$\boldsymbol{d}_i = \left(\frac{1}{|F|} \sum_{\substack{f \in F \\ length(f)=p \\ i \in f}} \mathbb{1} \right)_{p=p_s}^{p_e} \tag{19}$$

where p_s is the minimum fragment length to be considered and p_e the maximum. For each BED file containing the regions of interest, a matrix $D = [\boldsymbol{d}_0, \boldsymbol{d}_1, \ldots, \boldsymbol{d}_{w-1}]$ in which w represents the length of the regions, is generated. By describing fragments as a set of positions, we can define different types of fragment length distributions. For example, in "coverage around dyads", we described how inferred dyad locations are used in the set of positions rather than the fragment itself. The same principle can be applied to FLD feature extraction methods. To this end, we implemented the FLD around dyads, the FLD around the midpoint, the middle-n points FLD, and the central 60 bp FLD.

2.2.4. Entropy and relative fragment entropy (RFE)

Previously, it was shown that fragmentation patterns at active TSSs change, resulting in higher diversity of fragment lengths with respect to DNA-protected regions. Prior work described a peak in the fragment length distribution around 160 bp as well as a correlation with RNA expression levels of individual genes [29]. For this approach, termed epigenetic expression inference from cfDNA-sequencing (EPIC-seq), coverages of about 500x to 2000x are required, which are reached through hybrid capture-based targeted deep sequencing.

We hypothesized that higher diversity of fragment lengths may not

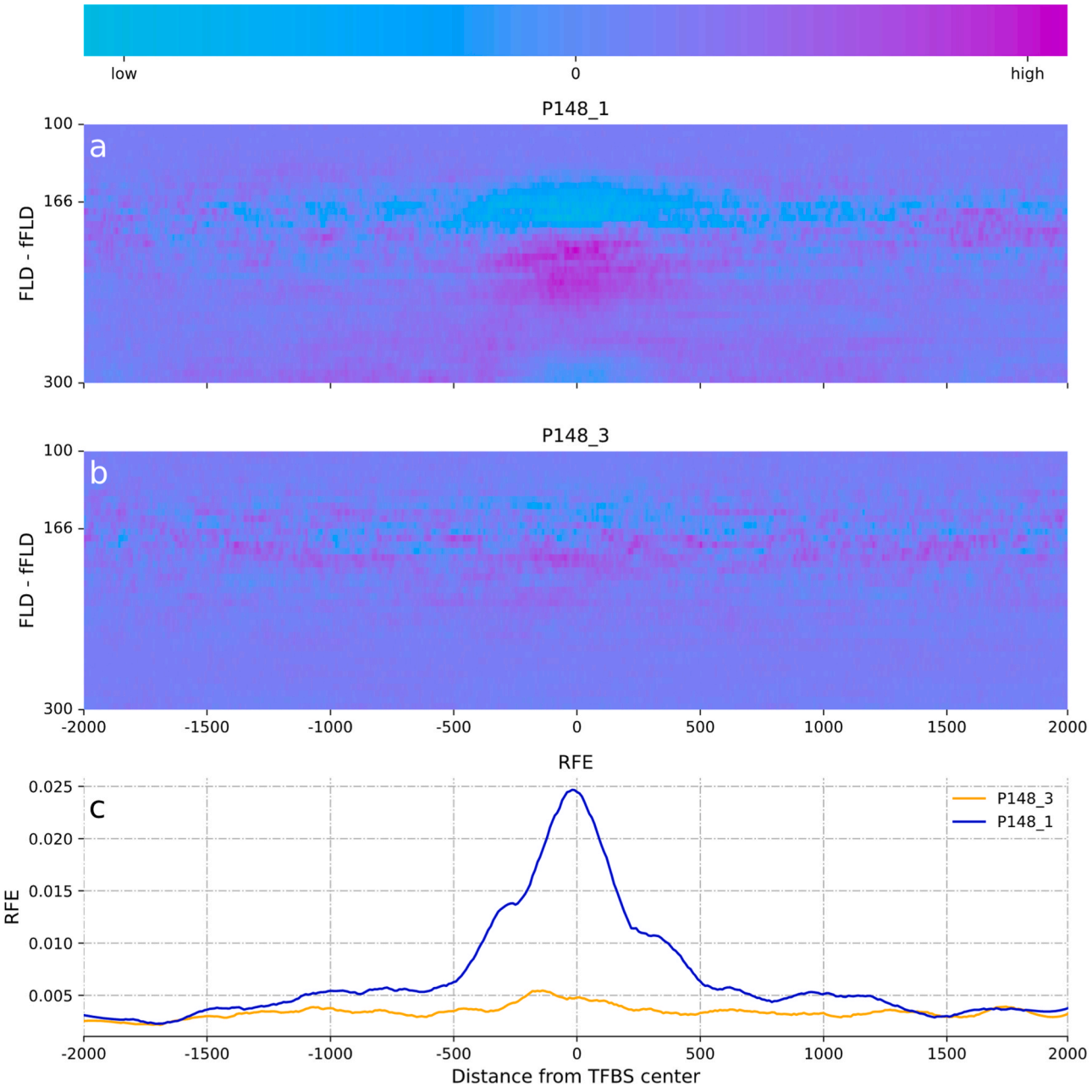

Fig. 3. Fragment length distributions (FLD) per position minus the FLD in the flanking regions (fFLD) at androgen receptor (AR)-specific TFBSs. a) FLD - fFLD P148_1. b) FLD - fFLD P148_3. c) RFE per position at AR-specific TFBSs smoothed with Savitzky-Golay filter.

only apply to TSSs, but also to TFBSs, where nucleosome displacement and depletion occurs. Therefore, we defined a way to extract an entropy signal from genomic shotgun data as a measure of TF activity. To overcome the problem of low depth of coverage at each position for typical sequencing data, we use multiple regions, such as multiple TFBSs representing the same TF. At a coverage of 20x, this results in an average number of twenty thousand reads per position for TFs with 1000 TFBSs along the genome. We implemented an entropy and a normalized entropy signal.

The former calculates the entropy of the fragment lengths at each position. Given a random variable E which takes values defined in the set $O = \{i : i \in [p_s, p_e)\}$ distributed according to $P_l : O \rightarrow \boldsymbol{d}_l$, entropy at position l can be defined as:

$$E_l = -\sum_{x \in O} P_l(x) \cdot \log P_l(x) \tag{20}$$

The latter, RFE, calculates the divergence in the fragment length distribution at each position over genomic intervals and the fragment length distribution in the flanking regions.

This can be defined as:

$$RFE_l = D_{\mathrm{KL}}(Q_l|F) = \sum_{x \in O} Q_l(x) \cdot \log\left(\frac{Q_l(x)}{F(x)}\right) \tag{21}$$

Where l represents the genomic position, Q_l represents the fragment length distribution at position l and F is the fragment length distribution in the flanking region.

At genuine TFBSs, we expect higher fragmentation than in the flanking regions where DNA is expected to be protected by nucleosomes. The FLD in the central part of the TFBS should differ from the FLD in the flanking region and therefore the RFE should show a peak in the center for TFs possessing higher activity (Supp. Fig. 3).

To avoid effects due to differences in coverage between positions, the same number of reads is used to calculate the entropy and the RFE signals at each position.

2.3. Differentially active genomic intervals

In LBFextract, we implemented a way to calculate differentially active genomic regions for which, in the case of activity, a peak or a dip is expected in the central part of the region. The general procedure is summarized in Algorithm 2 (Supplementary Information). In the first part, the algorithm extracts the features f for each TF t from a sample's BAM file (line 7–9). Subsequently, it calculates the accessibility of each feature and applies an appropriate statistical test (line 10–23). It uses the accessibility values of each TF grouped according to a label vector l, which assigns each sample to a specific group. Correction for multiple testing is applied to adjust for the higher probability of observing statistically significant results when testing multiple groups and multiple TFs. Finally, for all TFs that were found to be differentially active, an enrichment analysis is retrieved through the STRING API.

3. Results

3.1. Difference in androgen receptor signaling in prostate cancer and castration resistant prostate cancer

Our previous work utilized tissue and cancer type-specific chromatin accessibility datasets to identify tissue-dependent TFBS accessibility patterns and we found evidence that nucleosome footprints in cfDNA are informative of TFBSs [9]. In addition, we demonstrated that TFs are amenable to molecular PC subtyping, which is an important issue in the management of PC [30]. More specifically, we focused on the phenomenon of transdifferentiation of prostate adenocarcinoma (PRAD) to a treatment-emergent small-cell neuroendocrine prostate cancer (t-SCNC), which is a frequent mechanism in the development of treatment resistance against androgen deprivation therapy (ADT), and constitutes a subtype that is no longer dependent on androgen receptor (AR) signaling [31]. The ability to detect this critical transition in longitudinal sampling has clinical implications, i.e. as it indicates a change in therapy is needed [30]. The involvement of TFs in this transdifferentiation process to neuroendocrine prostate cancer (NEPC) has been extensively studied [31–33]. Also, we have previously leveraged this information to confirm transdifferentiation events in our PC cohort (Supplementary Methods) [9]. Herein, we use LBFextract to reproduce our previous findings and expand on them by extracting not only coverage, but also fragmentomic features at diverse TFBSs of TFs involved in this transdifferentiation. We apply the differential activity analysis provided by LBFextract to shed light on subtype-specific TFs. To demonstrate the validity of these signals and their potential, we showcase a previously described patient P148 [8]. Within 12 months, the time interval between collection of the samples P148_1 and P148_3, the PRAD transdifferentiated to a t-SCNC, which was accompanied by a clinical observation of a decrease of prostate-specific antigen (PSA) and an increase of neuron-specific enolase (NSE). We look at AR chromatin accessibility via coverage and FLDs at TFBSs of AR for P148. To reduce confounding effects like tumor fraction and coverage, which may bias the analysis, we additionally performed *in silico* dilutions for both samples to a 20x coverage and 20 % tumor fraction (Supplementary Information). We retrieved TFBSs from the Gene Transcription Regulation Database (GTRD v21.12), sorted them based on the number of peaks supporting each meta-cluster, removed TFBSs belonging to sex-related chromosomes and retained the top 1000 TFBSs. In doing so, we obtained 1058 TFs with 1000 TFBSs each (Fig. 1).

By analyzing general coverage at AR-specific TFBSs with the LBFextract extract_coverage method, we observed the expected central increase in the normalized coverage in P148_3 relative to P148_1, showing the reduced AR activity (Fig. 2a-e). By analyzing the coverage signal with extract_coverage_around_dyads, we were also able to observe the peaks flanking the central positions of the TFBSs in the case of P148_1, which indicates the presence of nucleosome phasing, which was not observed using normal fragment coverage. Furthermore, these peaks were found to be absent for P148_3, suggesting reduced AR binding to the TFBSs and therefore a decreased phasing of the neighboring nucleosomes (Fig. 2b).

The same could be observed from fragmentomics features. When looking at the FLD 2000 bp around the TFBS of AR (Fig. 3), from which we subtracted the signal of the FLD in the flanking regions (fFLD), an increased diversity in fragmentation patterns towards the center of the TFBS at fragment lengths above 180 bp can be observed in P148_1 (Fig. 3a), which is almost absent in P148_3 (Fig. 3b). Furthermore, a decreased representation of the fragment lengths around 166 bp toward the center of the TFBS is visible in the case of P148_1 (light blue region in Fig. 3a), while less pronounced in P148_3. Similarly, this information is captured by the RFE (Fig. 4, Fig. 3c), which exploits the difference between the distribution in the flanking regions and the one at each position. A higher RFE value at AR for sample P148_1 is evidence for a higher diversity in fragment lengths at the center of the TFBSs. Because peaks or valleys are not dependent on coverage here, but only on the fragment length, the information contained in the FLD or entropy-derived signals offers a unique perspective on TF-specific TFBS chromatin states.

3.2. Coverage around dyads and relative fragment entropy: analysis of signal while varying tumor fraction

In this work, we investigate metastatic prostate cancer (mPC) samples and showcased our newly implemented measures on samples that had been previously in-silico diluted to reach a tumor fraction of 20 % and an average coverage of 20x, which represent an average case for our mPC cohort. Here, we analyzed the coverage around dyads (CAD) and relative fragment entropy (RFE) signals while varying the tumor fraction from 1 % to 30 %. Interestingly, we observed that the efficacy of our

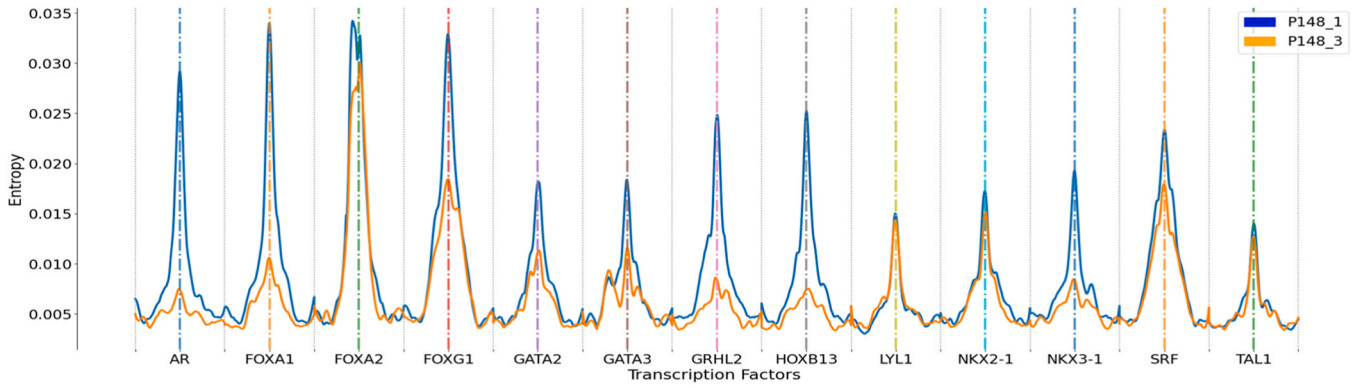

Fig. 4. Relative fragment entropy for TFs involved in the transdifferentiation process (i.e., AR, FOXA1/2, FOXG1, GATA2/3, GRHL2, HOXB13, NKX2–1, NKX3–1, SRF) as well as TFs involved in hematopoiesis (LYL1, TAL1).

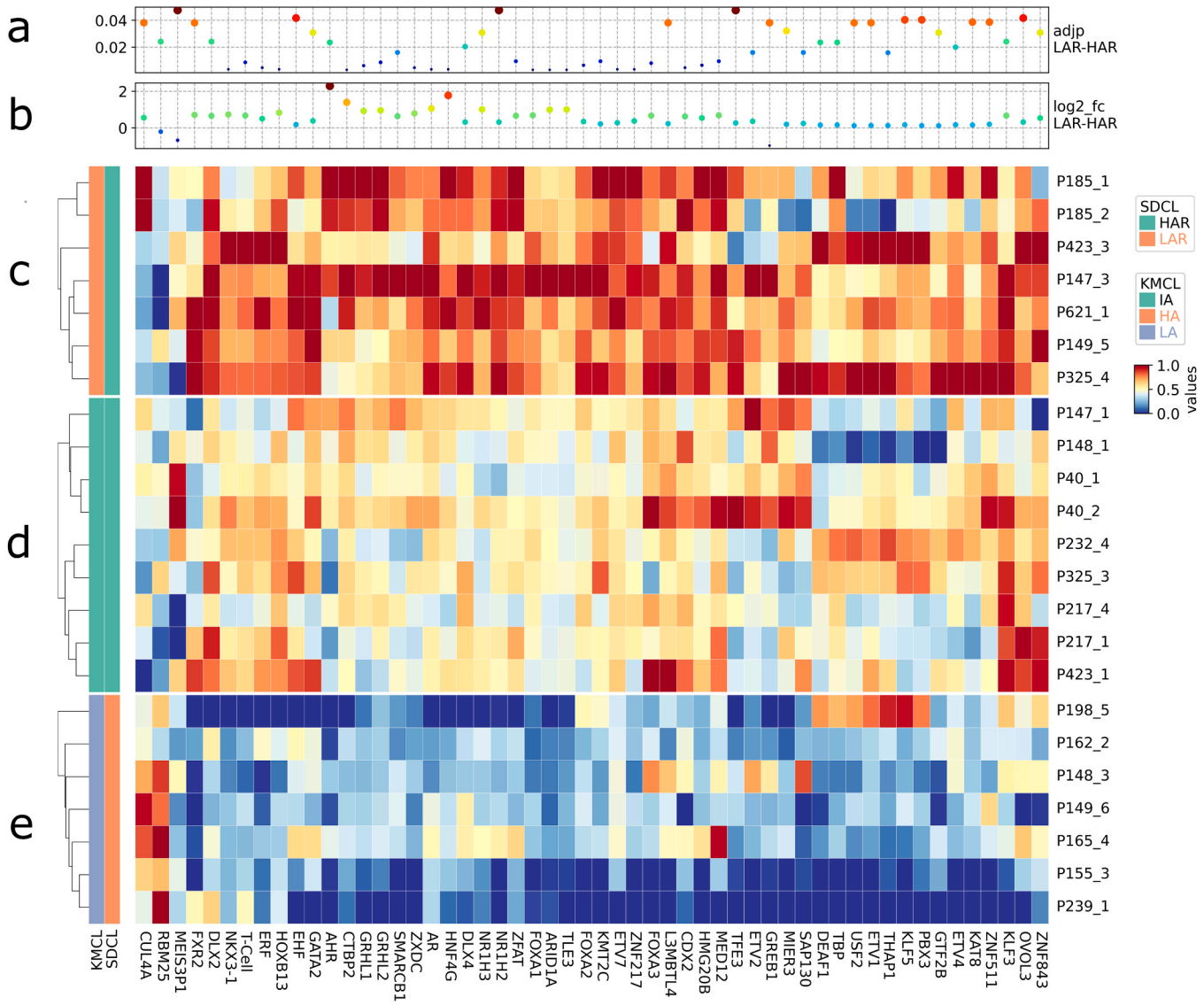

Fig. 5. Heatmap of the accessibilities of differentially active TFs found in the HAR and LAR clusters. a) Dot-plot of the adjusted p-values. b) Dot-plot of the pseudo log2 fold changes. c-e) Heat maps of the signals in the HAR (c, d) and LAR clusters (e). Bars on the left represent the clustering according to semi-supervised clustering (SDCL) and k-means (KMCL), respectively.

metrics—RFE and CAD—varies with different transcription factors (TFs). We observed distinct relationships between signal strength and tumor fraction, with the signal's intensity or attenuation exhibiting unique patterns. Specifically, hematopoietic TFs demonstrate an inverse relationship, where signal strength diminishes as the tumor fraction increases. In contrast, tumor-specific TFs exhibit a direct relationship, with signal strength enhancing as the tumor fraction rises. Additionally, certain TFs show a consistent signal regardless of the tumor fraction.

The accompanying figures (Supp. Figure 10, Supp. Figure 11) elucidate these relationships by plotting our two metrics against varying tumor fractions for sample P148_1. For instance, LYL-1, a hematopoietic TF, exhibits a decline in coverage and entropy signals with increasing tumor fraction. Conversely, FOXA1 and AR, both tumor-specific TFs, display robust signals at a 0.3 tumor fraction, which attenuate as the tumor fraction decreases, with AR's signal vanishing at a tumor fraction of 0.05. Interestingly, the signal for NKX3–1 remains detectable at a tumor fraction of 0.03, while the signal for NKX2–1 remains relatively stable across varying tumor fractions.

This suggests that binding activity plays a major role in the limit of detection when using different metrics and, for some TFs, signal attenuation happens faster for the RFE metric than for CAD.

3.3. Characterization of androgen receptor high and low signal-specific TFs in advanced prostate cancer

Despite the presence of samples with high and low AR signals, it is essential to emphasize that the signal patterns are not uniform and sampled from at least two distributions. Further, we postulate the existence of PC subtypes characterized by high and low AR signals within our cohort (n = 23). To further explore this theory, we applied semi-supervised and k-means clustering methods to detect different subgroups, detecting 2 and 3 clusters respectively. With this aim, in the semi-supervised approach, we selected samples P148_1 (PRAD) and P148_3 (NEPC), which display high and low AR coverage signals, respectively, as prototypes for high AR (HAR) and low AR (LAR) clusters. To make use of prior knowledge, we expanded the sets of TFs with those found to be differentially active in NEPC and PRAD in [20] and [8]. We obtained a final set of TFs for the neuroendocrine subtype composed of: AR, FOXA1, NKX3–1, HOXB13, GRHL2, ASCL1, GATA2 and HNF4G. Each PC sample was assigned to its nearest cluster prototype (lowest Euclidean distance). This generated an HAR cluster (n = 15), characterized by the presence of a valley around the TFBSs centers of AR (Supp. Fig. 5a), FOXA1 (Supp. Fig. 5b), NKX3–1 (Supp. Fig. 5c), HOXB13 (Supp. Fig. 5d), GRHL2 (Supp. Fig. 5f) and HNF4G (Supp. Fig. 5i) and an LAR cluster (n = 7), characterized by flatter profiles for the same TFs. We extracted coverage signals for all 1058 TFs and assessed differences in TF activities between clusters using the Mann-Whitney U test. After correction for multiple testing using the Benjamini-Hochberg method (multipletests function in python statsmodel package [34]), we rejected all null hypotheses with an adjusted-p value lower than 0.05. From the first analysis between HAR and LAR groups, we found 53 TFs with significantly increased accessibility (Table 1). Most of the TFs nominated for the initial clustering dataset—with the exception of ASCL1—were also found to be present within the list of differentially active TFs, which supports the validity of the initial set of TFs. Concurrently, we observed genes which were previously shown to be linked to the AR CRPC subtype, such as AR and FOXA1, at the top of the list of differentially active TFs. AR, which has a statistically relevant 1.06 log2 fold downregulation in the LAR cluster, belongs to the steroid hormone group of nuclear receptors and was shown to have a central role in PC development and progression. FOXA1, which was identified as downregulated in our analysis, is a pioneer TF that plays a pivotal role in partnering with AR to promote its attachment to chromatin. Remarkably, prior research revealed the role of FOXA1 as a suppressor of neuroendocrine differentiation and links its downregulation to the promotion of NEPC progression [35] and reprogrammed activity in NEPC [36]. We further detected HNF4G as being significantly upregulated (adjusted p-value 0.0035, 1.77 log2 fold change) in the HAR cluster. This TF is generally involved in gastrointestinal NEPC (GI-NEPC) and was shown to be expressed in 5 % of primary PCs and 30 % of CRPCs. It is responsible for the activation of an AR-independent resistance mechanism involving the activation of gastrointestinal transcription and chromatin patterns [37]. We also observed the presence of ARID1A and SMARCB1 among the top 5 differentially active TFs. Both of these TFs were shown to cooperate in the SWI/SNF chromatin remodeling complex and to be involved in PC lineage plasticity [38]. Further, HOXB13, NKx3–1, FOXA2 and GATA2 are other TFs that are linked to NEPC [20,39,40] and were found to be differentially active. Lastly, our analysis highlighted TLE3 as a key player among the differentially active TFs whose loss was previously linked to the development of a glucocorticoid receptor (GR)-mediated resistance mechanism under androgen receptor inhibitors [41].

Table 1
Results showing the top 20 differentially active TFs from analysis performed on the HAR and LAR groups (full table available in Supplementary Information). To include the direction of change, we calculated a pseudo log2 fold change, which retains the sign information.

	TF	μ_{HAR}	μ_{LAR}	p_value	adj_p-val	log2_fc
1	CTBP2	0.0041	0.0016	< 0.0001	0.0031	1.3885
2	TLE3	0.0056	0.0028	< 0.0001	0.0031	1.0007
3	ARID1A	0.0064	0.0032	< 0.0001	0.0031	0.9881
4	FOXA1	0.0087	0.0054	< 0.0001	0.0031	0.6823
5	HNF4G	0.0036	0.0011	< 0.0001	0.0035	1.7712
6	AR	0.0071	0.0034	< 0.0001	0.0035	1.0557
7	HOXB13	0.0082	0.0046	< 0.0001	0.0035	0.8275
8	NKX3−1	0.0066	0.0040	< 0.0001	0.0035	0.7305
9	ZNF217	0.0090	0.0069	< 0.0001	0.0035	0.3758
10	ETV7	0.0076	0.0063	< 0.0001	0.0035	0.2774
11	ZXDC	0.0039	0.0023	0.0001	0.0046	0.7897
12	CDX2	0.0036	0.0023	0.0001	0.0046	0.6228
13	ERF	0.0086	0.0061	0.0001	0.0046	0.5016
14	GRHL1	0.0065	0.0034	0.0001	0.0062	0.9228
15	HMG20B	0.0035	0.0024	0.0001	0.0065	0.5378
16	FOXA2	0.0090	0.0071	0.0001	0.0065	0.3426
17	FOXA3	0.0028	0.0018	0.0001	0.0080	0.6653
18	GRHL2	0.0065	0.0034	0.0001	0.0086	0.9550
19	T-Cell	0.0070	0.0044	0.0002	0.0086	0.6698
20	MED12	0.0030	0.0019	0.0002	0.0095	0.6824

3.4. Enrichment analysis

After the analysis of differential activity, LBFextract performs an enrichment analysis step, which is carried out through the STRING API using all identified differentially active TFs. For the enrichment step, one can use either the default parameters of the STRING API, the provided list of TFs, or a specific background. In this analysis, the default STRING API settings were used to include all potential ontology terms and to avoid the exclusion of potentially relevant biological pathways. Through this step, the enrichment of the differentially active genes in different databases, which span from gene ontology to KEGG, REACTOME and WikiPathways, is retrieved. Here, we focused on the LAR and HAR clusters and found a significant enrichment in the "Androgen receptor signaling pathway", alongside various processes associated with epithelial cell development and differentiation, prostate gland development, and neuron differentiation in the Gene Ontology process category. Moreover, the results of the DISEASES enrichment analysis contained many significant terms that are closely related to PC such as "prostate cancer", "prostate carcinoma", "prostate adenocarcinoma", "adenocarcinoma", and "reproductive organ disease". The enrichment analysis in the TISSUES category indicated statistically significant enrichment in "whole blood" and "prostate epithelium cell line" as well as "prostate epithelium cell". Furthermore, pathways including "Androgen receptor network in prostate cancer", "Nuclear receptors" and

"Endoderm differentiation" were found in the WikiPathways category. Additionally, pathways such as "Signaling by Nuclear Receptors," "Estrogen-dependent gene expression", "NR1H2 & NR1H3 regulate gene expression to limit cholesterol uptake", and "NR1H2 & NR1H3 regulate gene expression linked to triglyceride lipolysis in adipose" in the REACTOME database were significantly enriched.

4. Discussion

We showcased the features and applications of our LBFextract package in a biomarker discovery setting for PC, highlighting its unique capabilities in feature extraction, differential TF activity analysis, and its plugin structure, which was designed to keep up with the continuous growth in feature extraction methods in liquid biopsy research.

In our package, we provide a default set of liquid biopsy feature extraction methods based on fragment coverage, with major differences to coverage features extraction methods offered by other packages like SAMTools [42], Picard [43], DeepTools [44] Pysam [42,45,46], Mosdepth [47] and BEDTools [48,49].

While most of these general tools are based on read coverage or coverage in ranges, the coverage methods implemented in LBFextract are based on fragment coverage, in which the positions between read pairs are filled, and different strategies to increase the strength of the dyads-derived signals are applied, making nucleosome phasing around active TFBSs more visible. (Fig. 2b-d, Supp. Fig. 2). For example, we implemented midpoint coverage and middle-n points coverage (Fig. 2c-d), previously used in [21], to increase the strength of the phasing signal. We also re-implemented the central 60 bp-coverage [8] with the following modifications:removal of dependencies from other tools like fastx_trimmer; generalized to a user-defined central region; prevented double counting of overlapping reads; and made it compatible with reads < 150 bp, thus reducing sequencing requirements and costs.

When investigating the FLD signals derived after diverse *in silico* trimming strategies at CTCF TFBSs, an issue using midpoint coverage with fragments > 220 bp became visible. Indeed, the center of dinucleosome-derived fragments is positioned between nucleosome-derived dyads. Unphased dinucleosomal and mononucleosomal fragments produce a weaker signal with shifted nucleosome-derived peaks, which can be improved by better modelling the position of the dyad on polynucleosomal fragments. Therefore, we implemented the coverage around dyads signal, which provides a stronger dyad-derived signal. We modelled information about dyads coming from polynucleosomal structures, i.e. fragments having a length between multiples of the mono-nucleosomal length, thus increasing the amount of information used, which resulted in a stronger dyad-derived signal for phased nucleosomes (Supp Fig. 2). With this strategy, we further improved upon the coverage analysis performed in our previous work [8]. Indeed, for dinucleosome-derived fragments, an increasing bias concerning the location of the dyad is introduced the more a fragment becomes digested if only the central part of the read is considered. In the extreme case of fragments < 150 bp, the positions are counted twice.

In our investigation, we examined the AR coverage signal within sample P148, which transdifferentiated into cancer NEPC. Our results reveal a flat coverage profile at AR TFBSs, along with the disappearance of the recurrent peaks induced by nucleosome phasing in P148_3. These observations align with and support the conclusions drawn in prior studies [9,20], in which a similar behavior was described. This shows the importance of cfDNA-specific features. Indeed, with general coverage strategies, this information was not visible.

Finally, we expanded with the analysis capacity of LBFextract further into the fragmentomics space by implementing entropy-derived features, which efficiently summarize variation in FLD signals. We used these features to infer the TF activity in PC and NEPC samples, showing that RFE is powerful for recapitulating findings of our previous work [9,20]. The validity of this signal is also supported by the fact that hematopoietic TFs, such as TAL1 and LYL1, show similar profiles between samples. In contrast, condition-specific TFs, such as AR, HOXB13 and FOXA1, provide different signals in different conditions.

To identify potentially new TFs associated with these cancer subtypes, we extracted the coverage signal of 1058 TFs retrieved from the GTRD database for PC samples in the cohort. We performed cluster analysis and looked for differentially active TFs between the groups obtained with this approach, highlighting LBFextract's capability of discovering both subtypes and subtype-specific differentially active TFs from cfDNA. Indeed, through this analysis, we found a higher expression of the NKX3–1 TF in the HAR cluster. This aligns with the finding that NKX3–1 is generally expressed in the luminal cell of the prostate, where it is intertwined in a regulative feed-forward loop with AR in both normal prostate and AR-dependent PC [50], but it may be downregulated or lost in NEPC. We also observed a loss in the LAR group for HOXB13, which is generally active in AR-dependent PC and CRPC, but downregulated in NEPC [51]. This agrees with the low AR signal, low PSA values and high NSE values found in P148_3 and P198_5, which are part of the LAR cluster. Interestingly, two subunits of the mSWI/SNF complex, ARID1A and SMARCB1, were found to be downregulated in the LAR cluster, suggesting a possible impairment of the BAF complex known to be involved in damage response. As suggested by Park Y. et al. [52], this can be exploited to challenge the tumor with PARP inhibitors combined with ionizing radiation.

Although herein we focused on the PC use case, the analysis of differential TF activity enabled by LBFextract can be extrapolated to other tumor types and may provide valuable insights into various biological and pathological conditions by identifying changes in gene regulation. For example, in breast cancer, which is a heterogeneous disease with multiple subtypes, differential TF activity may help classify these subtypes [53,54] and reveal distinct regulatory networks. Detection of other lineage-specific TFs may offer new avenues of early detection [9] or monitoring. Furthermore, as cfDNA can also provide information about physiological processes and conditions such as aging or pregnancy, LBFextract may help address unmet needs in these areas.

While the aforementioned applications of LBFextract underscore its potential in liquid biopsy research, it is important to acknowledge its limitations and challenges. The package is specifically designed for feature extraction from BAM files, ensuring reproducibility solely for the feature extraction process. Therefore, variations in results may still occur due to the use of different aligners, modifications to aligner parameters, or alternative preprocessing steps, such as trimming. Additionally, LBFextract currently does not include feature extraction methods for fragment end motifs or genome-wide analyses like DELFI [11]. Nevertheless, these functionalities could potentially be incorporated through the plugin mechanism. Finally, as LBFextract relies on template length to infer DNA fragment length, it is currently restricted to paired-end sequencing data.

5. Conclusion

In this article, we introduced LBFextract, a Python package designed for the extraction and analysis of features from liquid biopsy WGS data, with a specific emphasis on transcription factor-specific coverage and entropy features. A notable strength of LBFextract lies in its flexibility, allowing seamless integration of new feature extraction methods in the form of plugins, enabling adaptability to research approaches as needed. This should streamline the generation of multiview datasets that can be used in multiview machine learning models, which have been shown to improve cancer classification [55]. Our study demonstrated the capabilities of LBFextract in suggesting TFs for follow-up research and showcasing its ability to recapitulate signals observed in previous work. This validation was performed from both a coverage and fragmentomics perspective. We outlined the ability of identifying condition-specific TFs, demonstrating the tool's utility in uncovering potential biomarkers in diverse biological contexts. LBFextract's open architecture and compatibility with plugins seek to not only make it a standalone tool

for reproducible feature extraction and biomarker identification, but also to contribute to the dynamic and collaborative nature of broader scientific and Python communities.

Funding

This work was supported by the Austrian Science Fund (FWF) [KLI 764 and TCS 101]. S.O.H. and B.S. were funded by a fellowship program of The Austrian Research Promotion Agency (FFG) [Cancer and Aging Liquid Chromatin Diagnostics; FFG-Nr. 899093].

CRediT authorship contribution statement

Benjamin Spiegl: Writing – review & editing, Data curation. **Isaac Lazzeri:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Michael R. Speicher:** Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization. **Samantha O. Hasenleithner:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Investigation, Funding acquisition. **Martin Kircher:** Writing – review & editing, Writing – original draft, Supervision, Resources.

Declaration of Competing Interest

A patent application has been filed for aspects of the paper (inventors: I.L.; S.O.H.; M.R.S.). S.O.H. is a co-founder of Vessel FlexCo and is on the advisory board of CureMatch. The remaining authors declare no competing interests.

Acknowledgements

We are especially grateful for the help and guidance of Prof. Michael Speicher, who passed away unexpectedly in September 2023. He was a great mentor, colleague, and friend. During his scientific career, he made many important contributions to the circulating cfDNA field, pioneering the analysis of WGS data of cfDNA and developing some of the earliest applications of these methods. His scientific passion will live on in our work. Special thanks to Hossein Hajiabolhassan, the other dedicated members of the BPDP group and Kircher lab, whose insights and collaboration have enriched the scientific depth of this work. We are very grateful to Prof. Ellen Heitzer for the critical review of this manuscript. We extend our sincere appreciation to Thomas Bauernhofer for generously providing the samples and clinical information used in this study. Fig. 1 was created with BioRender.com.

Supplementary Information

For additional details see Supplementary Information. For usage of the package, refer to https://lbf.readthedocs.io/.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.08.007.

References

[1] Hasenleithner SO, Speicher MR. A clinician's handbook for using ctDNA throughout the patient journey. Mol Cancer 2022 Mar;21(21):81.
[2] Heitzer E, Haque IS, Roberts CES, Speicher MR. Current and future perspectives of liquid biopsies in genomics-driven oncology. Nat Rev Genet 2019 Feb;20(2):71–88.
[3] Wan JCM, Mughal TI, Razavi P, Dawson SJ, Moss EL, Govindan R, et al. Liquid biopsies for residual disease and recurrence. Med 2021 Dec 10;2(12):1292–313.
[4] Ignatiadis M, Sledge GW, Jeffrey SS. Liquid biopsy enters the clinic — implementation issues and future challenges. Nat Rev Clin Oncol 2021 May;18(5):297–312.
[5] Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies [Internet]. [cited 2023 Dec 19]. Available from: https://www.science.org/doi/10.1126/science.aaw3616.
[6] Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA: A Cancer J Clin 2023;73(1):17–48.
[7] Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA Comprises an in vivo nucleosome footprint that informs its tissues-of-Origin. Cell 2016 Jan 14;164(1–2):57–68.
[8] Ulz P, Thallinger GG, Auer M, Graf R, Kashofer K, Jahn SW, et al. Inferring expressed genes by whole-genome sequencing of plasma DNA. Nat Genet 2016;48(10):1273–8.
[9] Ulz P, Perakis S, Zhou Q, Moser T, Belic J, Lazzeri I, et al. Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nat Commun 2019 Oct 11;10(1):4666.
[10] Sun K, Jiang P, Cheng SH, Cheng THT, Wong J, Wong VWS, et al. Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. Genome Res 2019 Jan 3;29(3):418–27.
[11] Mathios D, Johansen JS, Cristiano S, Medina JE, Phallen J, Larsen KR, et al. Detection and characterization of lung cancer using cell-free DNA fragmentomes. Nat Commun 2021 Aug 20;12(1):5060.
[12] Underhill HR, Kitzman JO, Hellwig S, Welker NC, Daza R, Baker DN, et al. Fragment length of circulating tumor DNA. PLoS Genet 2016 Jul 18;12(7):e1006162.
[13] Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Sci Transl Med 2018 Nov 7;10(466):eaat4921.
[14] Genome-wide cell-free DNA fragmentation in patients with cancer | Nature [Internet]. [cited 2023 Dec 19]. Available from: ⟨https://www.nature.com/articles/s41586–019-1272–6⟩.
[15] Mouliere F, Smith CG, Heider K, Su J, van der Pol Y, Thompson M, et al. Fragmentation patterns and personalized sequencing of cell-free DNA in urine and plasma of glioma patients. EMBO Mol Med 2021 Aug 9;13(8):e12881.
[16] Jiang P., Xie T., Ding S.C., Zhou Z., Cheng S.H., Chan R.W.Y., et al. Detection and characterization of jagged ends of double-stranded DNA in plasma. Genome Res [Internet]. 2020 Aug 14 [cited 2023 Dec 19]; Available from: ⟨https://genome.cshlp.org/content/early/2020/08/14/gr.261396.120⟩.
[17] Jiang P, Sun K, Peng W, Cheng SH, Ni M, Yeung PC, et al. Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. Cancer Discov 2020 May;10(5):664–73.
[18] Dor Y, Cedar H. Principles of DNA methylation and their implications for biology and medicine. Lancet 2018 Sep 1;392(10149):777–86.
[19] Leypold NA, Speicher MR. Evolutionary conservation in noncoding genomic regions. Trends Genet 2021 Oct 1;37(10):903–18.
[20] De Sarkar N, Patton RD, Doebley AL, Hanratty B, Adil M, Kreitzman AJ, et al. Nucleosome patterns in circulating tumor DNA reveal transcriptional regulation of advanced prostate cancer phenotypes. Cancer Discov 2023;13(3):632–53.
[21] Doebley A.L., Ko M., Liao H., Cruikshank A.E., Kikawa C., Santos K., et al. Griffin: Framework for clinical cancer subtyping from nucleosome profiling of cell-free DNA [Internet]. medRxiv; 2021 [cited 2023 Dec 29]. p. 2021.08.31.21262867. Available from: ⟨https://www.medrxiv.org/content/10.1101/2021.08.31.21262867v1⟩.
[22] Peneder P, Stütz AM, Surdez D, Krumbholz M, Semper S, Chicard M, et al. Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden. Nat Commun 2021 May 28;12(1):3230.
[23] Spiegl B, Kapidzic F, Röner S, Kircher M, Speicher MR. GCparagon: evaluating and correcting GC biases in cell-free DNA at the fragment level. NAR Genom Bioinforma 2023 Oct 11;5(4):lqad102.
[24] pluggy — pluggy 1.3.1.dev20+g4b5b2d4 documentation [Internet]. [cited 2024 Jan 18]. Available from: ⟨https://pluggy.readthedocs.io/en/latest/⟩.
[25] Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A. Determinants of nucleosome organization in primary human cells. Nature 2011 May 22;474(7352):516–20.
[26] Liu Y. At the dawn: cell-free DNA fragmentomics and gene regulation. Br J Cancer 2022 Feb;126(3):379–90.
[27] Ding SC, Lo YMD. Cell-Free DNA fragmentomics in Liquid Biopsy. Diagn (Basel) 2022 Apr 13;12(4):978.
[28] Yu SCY, Chan KCA, Zheng YWL, Jiang P, Liao GJW, Sun H, et al. Size-based molecular diagnostics using plasma DNA for noninvasive prenatal testing. Proc Natl Acad Sci USA 2014 Jun 10;111(23):8583–8.
[29] Esfahani MS, Hamilton EG, Mehrmohamadi M, Nabet BY, Alig SK, King DA, et al. Inferring gene expression from cell-free DNA fragmentation profiles. Nat Biotechnol 2022 Apr;40(4):585–97.
[30] Aggarwal R, Huang J, Alumkal JJ, Zhang L, Feng FY, Thomas GV, et al. Clinical and genomic characterization of treatment-emergent small-cell neuroendocrine prostate cancer: a multi-institutional prospective study. J Clin Oncol 2018 Aug 20;36(24):2492–503.
[31] Puca L, Vlachostergios PJ, Beltran H. Neuroendocrine differentiation in prostate cancer: emerging biology, models, and therapies. Cold Spring Harb Perspect Med 2019 Jan 2;9(2):a030593.
[32] Puca L, Bareja R, Prandi D, Shaw R, Benelli M, Karthaus WR, et al. Patient derived organoids to model rare prostate cancer phenotypes. Nat Commun 2018 Jun 19;9(1):2404.
[33] Beltran H, Prandi D, Mosquera JM, Benelli M, Puca L, Cyrta J, et al. Divergent clonal evolution of castration resistant neuroendocrine prostate cancer. Nat Med 2016 Mar;22(3):298–305.

[34] Seabold S., Perktold J. statsmodels: Econometric and statistical modeling with python. In: 9th Python in Science Conference. 2010.
[35] Kim J, Jin H, Zhao JC, Yang YA, Li Y, Yang X, et al. FOXA1 inhibits prostate cancer neuroendocrine differentiation. Oncogene 2017 Jul 13;36(28):4072–80.
[36] Baca SC, Takeda DY, Seo JH, Hwang J, Ku SY, Arafeh R, et al. Reprogramming of the FOXA1 cistrome in treatment-emergent neuroendocrine prostate cancer. Nat Commun 2021 Mar 30;12(1):1979.
[37] Shukla S, Cyrta J, Murphy DA, Walczak EG, Ran L, Agrawal P, et al. Aberrant activation of a gastrointestinal transcriptional circuit in prostate cancer mediates castration resistance. Cancer Cell 2017;32(6):792–806.
[38] Cyrta J, Augspach A, De Filippo MR, Prandi D, Thienger P, Benelli M, et al. Role of specialized composition of SWI/SNF complexes in prostate cancer lineage plasticity. Nat Commun 2020;11(1):5549.
[39] Sandhu HS, Portman KL, Zhou X, Zhao J, Rialdi A, Sfakianos JP, et al. Dynamic plasticity of prostate cancer intermediate cells during androgen receptor-targeted therapy. Cell Rep 2022;40(4).
[40] Park JW, Lee JK, Witte ON, Huang J. FOXA2 is a sensitive and specific marker for small cell neuroendocrine carcinoma of the prostate. Mod Pathol 2017;30(9):1262–72.
[41] Palit SA, Vis D, Stelloo S, Lieftink C, Prekovic S, Bekers E, et al. TLE3 loss confers AR inhibitor resistance by facilitating GR-mediated human prostate cancer cell growth. Elife 2019;8:e47430.
[42] Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience 2021 Feb 16;10(2). giab008.
[43] Picard toolkit [Internet]. Broad Institute, GitHub repository. Broad Institute; 2019. Available from: ⟨https://broadinstitute.github.io/picard/⟩.
[44] Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res 2016 Jul 8;44(W1):W160–5.
[45] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009 Aug 15;25(16):2078–9.
[46] Bonfield JK, Marshall J, Danecek P, Li H, Ohan V, Whitwham A, et al. HTSlib: C library for reading/writing high-throughput sequencing data. Gigascience 2021 Feb 16;10(2):giab007.
[47] Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics 2018 Mar 1;34(5):867–8.
[48] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010 Mar 15;26(6):841–2.
[49] Dale RK, Pedersen BS, Quinlan AR. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics 2011 Dec 15;27(24):3423–4.
[50] Tan PY, Chang CW, Chng KR, Wansa KSA, Sung WK, Cheung E. Integration of regulatory networks by NKX3-1 promotes androgen-dependent prostate cancer survival. Mol Cell Biol 2012;32(2):399–414.
[51] Cheng S, Yang S, Shi Y, Shi R, Yeh Y, Yu X. Neuroendocrine prostate cancer has distinctive, non-prostatic HOX code that is represented by the loss of HOXB13 expression. Sci Rep 2021;11(1):2778.
[52] Park Y, Chui MH, Suryo Rahmanto Y, Yu ZC, Shamanna RA, Bellani MA, et al. Loss of ARID1A in tumor cells renders selective vulnerability to combined ionizing radiation and PARP inhibitor therapy. Clin Cancer Res 2019 Sep 13;25(18):5584–94.
[53] Rao S, Han AL, Zukowski A, Kopin E, Sartorius CA, Kabos P, et al. Transcription factor-nucleosome dynamics from plasma cfDNA identifies ER-driven states in breast cancer. Sci Adv 2022 Aug 26;8(34). eabm4358.
[54] Doebley AL, Ko M, Liao H, Cruikshank AE, Santos K, Kikawa C, et al. A framework for clinical cancer subtyping from nucleosome profiling of cell-free DNA. Nat Commun 2022 Dec 3;13(1):7475.
[55] Moldovan N, van der Pol Y, van den Ende T, Boers D, Verkuijlen S, Creemers A, et al. Multi-modal cell-free DNA genomic and fragmentomic patterns enhance cancer survival and recurrence analysis. Cell Rep Med 2024 Jan 16;5(1):101349.