Inference of tumor cell-specific transcription factor binding from cell-free DNA

Peter Ulz[1], Samantha Perakis[1], Qing Zhou[1], Tina Moser[1], Jelena Belic[1], Albert Wölfler[2], Armin Zebisch[2], Armin Gerger[3], Gunda Pristauz[5], Edgar Petru[5], Michael G. Schimek[4], Jochen B. Geigl[1], Thomas Bauernhofer[3], Heinz Sill[2], Christoph Bock[6,7,8], Ellen Heitzer[1,9,10], Michael R. Speicher[1,9]

[1]Institute of Human Genetics, Diagnostic and Research Center for Molecular BioMedicine, Medical University of Graz, Graz, Austria

[2]Department of Hematology, Medical University of Graz, Graz, Austria.

[3]Department of Internal Medicine, Division of Oncology, Medical University of Graz, Graz, Austria

[4]Institute of Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria

[5]Department of Obstetrics and Gynecology, Medical University of Graz, Graz, Austria.

[6]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria

[7]Department of Laboratory Medicine, Medical University of Vienna, Vienna, Austria

[8]Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany

[9]BioTechMed-Graz, Graz, Austria

[10]Christian Doppler Laboratory for Liquid Biopsies for Early Detection of Cancer, Graz, Austria

Correspondence should be addressed to M.R.S. (email: michael.speicher@medunigraz.at)

Abstract

Alterations in transcription factors are important drivers of tumorigenesis, but non-invasive assays for assessing transcription factor activity are lacking. Here, we evaluated the feasibility of inferring transcription factor binding in solid tumors from their nucleosome footprint in circulating cell-free DNA. We developed a novel analysis pipeline to determine accessibility of transcription factor binding sites and applied it to 244 cell-free DNA samples from patients with prostate, breast and colon cancer. We observed patient-specific as well as tumor-specific patterns. Specifically, inferred binding patterns for the transcription factors AR, HOXB13, and NKX3-1 allowed us to classify patients by tumor type, including subtypes of prostate cancer, which has clinical implications for the management of patients. Our approach for mapping tumor-specific transcription factor binding *in vivo* based on blood samples makes a key part of the noncoding genome amenable for clinical analysis.

Introduction

Transcription factors (TFs) modulate the expression of their target genes and often play a key role in development and differentiation [1]. In order to bind regulatory DNA, TFs generally have to interact with nucleosomes [2] and next-generation sequencing-based assays have recently provided unprecedented insight into nucleosome occupancy patterns at TF binding sites (TFBSs) [1, 3, 4, 5]. However, although TF activity is frequently disturbed in cancer [1, 6, 7], there is a lack of non-invasive means from blood to measure TF activity or their modulations under therapies.

Cell-free DNA (cfDNA) from plasma is released after enzymatic digestion from apoptotic cells [8] and therefore circulates mostly as mono-nucleosomal DNA [9, 10]. Hence, whole-genome sequencing of cfDNA fragments enabled the generation of nucleosome maps where dyads, i.e. the midpoint of a canonical nucleosome, of sites with high nucleosome preferences resulted in a strong peak of reads whereas dyads of less preferentially positioned nucleosomes showed reduced peaks or none at all (Fig. 1a) [9, 10]. Previously, we generated *in vivo* maps of nucleosome occupancy at transcription start sites (TSSs) to infer expressed genes from cells releasing their DNA into the circulation [10], while another study used a similar approach to detect TF footprints [9].

As the inference of TF binding from cfDNA has tremendous diagnostic potential in cancer and beyond, we developed a new and optimized bioinformatics pipeline, capable of resolving those constituents involved in nucleosome signatures at TFBSs to objectively assess and to compare TFBS accessibility in different plasma samples. To validate this pipeline for clinical purposes, we produced deep whole-genome sequencing (WGS) data from 24 plasma samples from healthy donors and from 15 plasma samples of patients with metastasized prostate, colon or breast cancer, where cfDNA also comprises circulating tumor DNA (ctDNA) [11, 12, 13, 14]. Furthermore, we generated shallow WGS data for 229 plasma samples from patients with the aforementioned tumor entities (>18.5 billion mapped plasma sequence reads in total). Importantly, our approach profiles individual TFs instead of establishing general tissue-specific patterns using mixtures of cfDNA signals resulting from multiple cell types and analyses by Fourier transformation as previously suggested [9]. Hence, our approach offers a more nuanced view of both tissue contributions and biological processes and allowed us therefore to identify lineage-specific TFs suitable for both tissue-of-origin and tumor-of-origin analyses. Furthermore, for the first time we show TFBS plasticity in cfDNA from patients with cancer and we demonstrate the potential of TFs in classifying prostate cancer subtypes, which has relevant clinical implications.

Results

Nucleosome occupancy inferred from cfDNA shows characteristic TF binding footprints

We started with establishment of nucleosome occupancy maps at TFBSs and tested for similarities and differences among healthy individuals and cancer patients. To this end, we obtained and analyzed high-coverage cfDNA samples from 24 healthy controls (males and females, 12 each) where the vast majority (>90%) of cfDNA is derived from apoptosis of white blood cells with minimal contribution from other tissues [15, 16] and 11 plasma samples derived from 7 patients with 3 common tumor entities (Supp. Table 1), i.e. four cases with prostate cancer (P40, P147, P148, P190), one colorectal cancer (CRC; C2), and two breast cancers (B7 and B13) with ctDNA fractions ranging from 18-78% (Supp. Fig. 1; Supp. Table 2).

We focused our analysis on the 676 TFs contained in the Gene Transcription Regulation Database (GTRD; Version 18.01; http://gtrd.biouml.org) [17], which provides detailed TFBS information based on ChIP-seq data for a variety of tissue samples. We annotated these TFs with an up-to-date curated list of 1,639 known or likely human TFs (http://humantfs.ccbr.utoronto.ca/; version 1.01) [1] (Fig. 1b; Supp. Table 3). Because of the potentially high number of TFBSs per TF, we defined three different stringency criteria (Fig. 1b): first, all TFBSs for all tissue samples in the GTRD; second, those peaks supported by >50% of the maximum number of samples (subsequently referred to as “>50%-TFBSs”; in these two analyses all 676 GTRD TFs were included); third, the 1,000 TFBSs per TF that were supported by the majority of samples (“1,000-msTFBSs”; 505 TFs fulfilled this criterion).

CTCF binding sites, which are known to be surrounded by arrays of strongly positioned nucleosomes [18, 19], yielded the expected oscillating coverage pattern, which remained similar throughout all analyzed samples, regardless whether the cfDNA was derived from healthy controls or from patients with cancer (Fig. 1c). This was consistent with DNase hypersensitivity assays from the Encyclopedia of DNA Elements (ENCODE) database for cell lines GM12878 (B-lymphocyte cell line from a female donor with European ancestry), LNCaP (androgen-sensitive human prostate adenocarcinoma cell line) and HCT116 (human colon cancer cell line) (Fig. 1c).

The ctDNA in the plasma from patients with cancer altered the balance between DNA from hematopoietic versus epithelial cells compared to the healthy controls, resulting for the cancer-derived samples in decreased amplitudes for the lineage-restricted hematopoietic TFs purine-rich box1 (PU.1) [20], LYL1 (lymphoblastic leukemia 1) [21], and the lymphocyte lineage-restricted transcription factor SPIB [22] and an increased amplitude for TF GRHL2, a pioneer TF for epithelial cells [23] (Fig. 1d). We confirmed the lineage-specificity of these TFs with data of publicly available DNase hypersensitivity assays (Fig. 1d). As another example for a well-established TF, we analyzed FOXA1, which cooperates with nuclear hormone receptors in endocrine-driven tumors of the breast and prostate [24]. Indeed, consistent with DNase hypersensitivity assays, we observed preferentially increased accessibility of FOXA1 in the plasma samples of prostate and breast cancer patients (Fig. 1e).

For further confirmation, we also conducted comparisons with ENCODE data, where mono-nucleosome-bound DNA fragments were generated by micrococcal nuclease (MNase) digestion (Supp. Fig. 2a, Supp. Fig. 3). We performed coverage-independent analyses [25] (Supp. Fig. 2b; Supp. Info.) and computed the spatial density of cfDNA fragments related to the single recognition sequences (Supp. Fig. 2c; Supp. Info.). We also generated catalogues of TFBSs which may be affected by co-binding of more than one TF for all 676 TFs and the 505 TFs from the 1,000-msTFBSs (Supp. Fig. 2d; Supp. Table 4). Furthermore, using purified, high molecular weight DNA as a negative control, we observed only an even coverage over TFBSs (Supp. Fig. 2e). In summary, these results showed that the corresponding TFBS coverage profiles closely resembled each other, suggesting a high accuracy of our approach and that the obtained patterns for any given TF are reproducible throughout all samples.

The "accessibility score" enables accurate inference of TF binding from cfDNA

The aforementioned results suggested that certain lineage-specific TFs are suitable for determining the tissue-of-origin of plasma DNA. However, currently no means of assessing the accessibility at their binding sites in cfDNA as proxy for their activity exist. To implement such an approach, we first investigated TF-specific nucleosome coverage profiles (Supp. Info.), which led us to conduct calculations separately for TFBSs within and outside of TSSs (Supp. Fig. 4a), and furthermore, for all GTRD tissues versus the >50%-TFBSs (Supp. Fig. 4b). These analyses suggested that average TFBS patterns comprise two signals: a TSS-proximal (within 2 kb of TSS, resulting in a "low frequency pattern") and a TSS-distal (>2 kb away from TSS peak, generating a "high-frequency pattern"), corresponding to the more evenly spaced peak signal. To suppress effects on the coverage not contributed by preferential nucleosome positioning and to remove local biases from the nucleosome data, we used Savitzky-Golay filters for detrending (Material & Methods) (Fig. 2a). Subsequently, we recorded the data range (maximum minus the minimum of the data values, corresponds to the amplitude) of the high-frequency signals, corrected them by LOESS smoothing as they depend on the number of TFBSs (Fig. 2b) (with the exception of the 1,000-msTFBSs), and then used calculated ranks (Material & Methods) as a measure for the accessibility of each TFBS. As TF binding only opens or "primes" its target enhancers without necessarily activating them per se [23], we refer to the rank values as "accessibility score".

However, we also wanted to test potential alternatives for TF accessibility assessment, and to this end we reconstructed an unbiased, detrended signal at a period between 135 and 235 bp by wavelet analysis and summed up the powers of the signal across the 2,000bp flanking TFBSs (Fig. 2c-d). To benchmark the performance of Savitzky-Golay filtering and wavelet analysis, we used cfRNA data [26] and observed significantly reduced accessibility for unexpressed TFs (i.e. <0.01 FPKM [Fragments Per Kilobase exon per Million reads]) as compared to the accessibility of expressed (i.e. >10 FPKM) TFs (>50%-TFBSs; Savitzky-Golay filtering: $p=1.75\text{x}10^{-13}$; the sum of powers (wavelet analysis): p=0.0004049; 1,000-msTFBSs; Savitzky-Golay filtering: $p=1.254\text{x}10^{-11}$;

Mann-Whitney-U test each) (Fig. 2e). These differences were also significant when we compared the adjusted ranges to mean DNase coverage (>50%-TFBSs; Savitzky-Golay filtering: $p<2.2x10^{-16}$; the sum of powers (wavelet analysis): $p<2.2x10^{-16}$; 1,000-msTFBSs; Savitzky-Golay filtering: $p<2.2x10^{-16}$; Mann-Whitney-U test each) (Fig. 2f). As Savitzky-Golay filtering performed slightly better, we favored this approach and then defined detection thresholds for TFBS accessibilities deviating from the normal samples as ±3 mean of the standard deviation (as a z-score of 3). For assessments based on all or >50%-TFBSs, the detection thresholds for our normalized accessibility score were ±253 and ±88 for the 1,000-msTFBSs, which have a lesser number of analyzable TFs (Supp. Fig. 4c, Supp. Table 6, Supp. Table 7).

Hence, we established robust means to assess TFBS accessibility and we reasoned that this should offer novel options to use cfDNA in clinical diagnostics.

Inference of TF binding from cfDNA supports molecular subtyping in prostate cancer

We next assessed to what extent tissue-specific TFs are suitable for the identification of tumor-of-origin and molecular subtyping. To this end, prostate cancer is a particularly interesting tumor entity because a frequent (~20%) mechanism in the development of treatment-resistance to novel agents targeting the AR pathway, such as abiraterone or enzalutamide, is the transdifferentiation of an adenocarcinoma to a treatment-emergent small-cell neuroendocrine prostate cancer (t-SCNC) [27]. This transdifferentiation has enormous clinical implication because it requires a change in therapy [27] and the involvement of several TFs in such a transdifferentiation process has been extensively studied [28, 29, 30] (Fig. 3a).

We first screened data for expression of human TFs across tissues and cell types provided by [1] and the human protein atlas (www.proteinatlas.org) [31] and confirmed the well-established prostate lineage specificity of TFs AR, HOXB13, and NKX3-1 [32, 33, 34, 35], which was also reflected in the DNase hypersensitivity assays of the prostate cancer cell line LNCaP (Fig. 3b-d). Accordingly, these TFs displayed increased accessibility at their binding sites only in the cfDNA of patients with prostate cancer. Because of the extraordinary relevance of AR in prostate cancer, we used not only the AR binding sites as defined by the GTRD, but additionally employed those reported by Pomerantz and colleagues [36], who, by analyzing the AR cistrome, identified 9,179 tumor AR binding sites with higher binding intensity in tumors and 2,690 normal AR binding sites with high binding intensity in normal samples. Indeed, whereas normal AR binding sites were not accessible in cfDNA of both controls and patients, the tumor AR binding sites demonstrated increased accessibility in the plasma samples from patients (Fig. 3d).

As a further confirmation for the robustness and reproducibility of lineage-specific TFs in cfDNA, we reasoned that if we analyze pools of multiple cfDNA samples generated by shallow-coverage (<0.2x) [37], that those TFs with increased accessibility in the majority or all samples, i.e. lineage-specific TFs, should have an increased accessibility score, whereas others will be averaged out. To this end, we pooled cfDNA samples separately for prostate (*n*=69), for colon (*n*=100) and for breast (*n*=60) cancer cases and repeated the analyses. The epithelial TF GRHL2 and the hematopoietic TFs reiterated their increased and decreased, respectively, accessibility patterns (Supp. Fig. 5). Within the prostate cancer cfDNA pool, the lineage-specific TFs AR, HOXB13, and NKX3-1 again showed increased accessibilities (Supp. Fig. 5), suggesting that these features are universally present in prostate cancer and may be suitable for the identification of tumor-of-origin from cfDNA.

For tumor subclassification, we started with an index case, P148, where we had the opportunity to analyze two plasma samples (P148_1, P148_3) taken 12 months apart, during which the prostate adenocarcinoma transdifferentiated to a t-SCNC (Supp. Table 1; Supp. Info.) [38]. These two samples showed significant TFBS accessibility changes (Kendall's Tau: 0.7573, Supp. Table 8), specifically reflected in several TFs. The t-SCNC is no longer an androgen-dependent stage of prostate cancer [29] and, consequently, accessibility of AR binding sites was no longer observed in sample P148_3 (Fig. 3e). Due to its close cooperation with nuclear hormone receptors [24], accessibility to FOXA1 was correspondingly reduced (Fig. 3e). Furthermore, the change in the cell type identity became apparent as observed by the reduced accessibility to the binding sites of the prostate-specific lineage TFs HOXB13 and NKX3-1 (Fig. 3e) and the epithelial TF GRHL2 (Supp. Fig. 6a). TF changes associated with neuronal development included augmented accessibility of GLI-similar 1 (GLIS1) (Supp. Fig. 6b), a TF whose expression is dramatically increased under hypoxic conditions [39]. Hypoxia has been discussed to facilitate the development of prostate adenocarcinoma to an androgen-independent state [40] and furthermore to downregulate repressor element-1 (RE-1) silencing transcription factor (REST), which induces neuroendocrine reprogramming [41] and we indeed observed a significantly decreased accessibility of REST (Fig. 3e). Furthermore, N-MYC is involved in AR signaling suppression and neuroendocrine program regulation [29, 42], which was mirrored in an increased accessibility (Fig. 3e).

These observations suggested that in certain cancer disease stages, TFBSs may have a high plasticity affecting pathways.

In order to test whether prostate cancer subtype classification based on TFBSs from cfDNA is possible, we added plasma samples as a proof-of-principle from 4 further t-SCNCs cases (P170_2, P179_4, P198_5, and P240_1) (Supp. Table 1). For these cases, we tested in addition whether our approach is also applicable to cfDNA sequenced with a lesser coverage by down-sampling plasma samples P148_1 (819,607,690 reads) and P148_3 (768,763,081 reads) to ~50 million reads. The reduction of reads resulted in an increase of noise levels, which was dependent on the number of TFBSs but neglectable for TFs with more than 1,000 TFBSs (Supp. Fig. 7) so that analyses for the aforementioned highly relevant TFs were not affected. We then repeated the analyses for the aforementioned 4 samples, each sequenced with ~50 million reads, and observed again the decreased accessibilities for TFs AR, FOXA1, HOX-B13, and NKX3-1, or the increased accessibility of N-MYC (Fig. 3f). Interestingly, we noted decreased accessibility of REST only in two of these four cases (P170_2 and P198_5; Fig. 3f), which is consistent with reports that REST downregulation is usually observed in 50% of neuroendocrine prostate cancer cases [29]. Only in these two cases did GLIS1 again have an increased accessibility (z-scores: P170_2: 4.3; P198_5: 4.4), suggesting that this hypoxia-associated TF may be linked to REST downregulation.

Discussion

We developed a method and bioinformatics software pipeline for inferring tumor cell-specific TF binding from cell-free DNA in the blood, with relevance for clinical diagnostics and non-invasive tumor classification. While most studies have adopted a gene-centric focus when evaluating somatically acquired alterations, we evaluated an important part of the noncoding genome, focusing on TFBSs. As many TFs bind preferentially within open chromatin and have to therefore interact with nucleosomes [1, 7], we utilized the largely mono-nucleosomal cfDNA, which allows the mapping of nucleosome positions [9, 10]. A unique feature of our approach is that we generated *in vivo* data on TFBSs from an endogenous physiological process in contrast to *in vitro* assays, which may be affected by technical variations. Our data correlated strongly with DNase I hypersensitivity data for cell lines GM12878, LNCaP or HCT116 suggesting the reliability of our approach.

Compared to a previous publication [9], our study has several distinct differences. First, Snyder and colleagues generated data only for a few TFs and were more focused on general DNase hypersensitive regions. Consequently, they suggested the use of general tissue-specific patterns using mixtures of cfDNA signals resulting from multiple cell types and analyses by Fourier transformation [9]. In contrast, we profiled individual TFs and thereby established lineage-specific TFs for clinical applications. Second, our analysis pipeline with the novel metric, the accessibility score, enables the objective comparison of TF binding events in various plasma samples, which paves the way for entirely new diagnostic procedures. Third, due to the improved resolution of TFBS analyses, we were for the first time able to use cfDNA to show TFBS plasticity during a disease course, such as reprogramming to a different cell lineage. Such a dynamic TF view instead of a static view obtained from tissue [7] is a unique feature of cfDNA analyses. Fourth, we demonstrate that our cfDNA TFBS bioinformatics pipeline allows subclassification of tumor entities and hence fills an important diagnostic gap in the managing of patients with prostate cancer [27]. Finally, whereas Snyder and colleagues required 1.5 billion reads per plasma sample [9], which is prohibitive for routine clinical use from a cost perspective, we were able to conduct in-depth TF analysis with ~50 million reads, making our approach amenable for clinical applications.

Our study has limitations, as our TF nucleosome interaction maps are inevitably heterogeneous, comprising signals of all cell types that give rise to cfDNA. Furthermore, as in other studies [9], we used plasma samples from individuals who appeared to have large burdens of ctDNA and therefore we cannot yet provide data on sensitivity.

Nevertheless, advanced prostate cancer, the main tumor entity analyzed here, is a classic example of the intractability and consequent lethality that characterizes metastatic carcinomas. Tumor studies lack dynamic models, and, in particular, dynamic profiling of clinical samples, for exploring transitions and interplays between pathways. Because of the potential of TFs to regulate gene transcription throughout the genome and their often exquisitely lineage-specific manner, their detailed non-invasive analyses based on cfDNA offer a unique opportunity to improve clinical diagnostics. Our data also provides the foundation for further dissection of the non-coding genome through novel means of transcription regulation profiling.

Methods

Patients

The study was approved by the Ethics Committee of the Medical University of Graz (approval numbers 21-227 ex 09/10 [breast cancer], 21-228 ex 09/10 [prostate cancer], 21-229 ex 09/10 [colorectal cancer], and 29-272 ex 16/17 [High resolution analysis of plasma DNA]), conducted according to the Declaration of Helsinki and written informed consent was obtained from all patients and healthy probands, respectively.

Detailed information on the patients is provided in the Supplementary Information.

Blood sampling and library preparation

Peripheral blood was collected from patients with metastatic prostate, breast and colon cancer at the Department of Oncology and from anonymous healthy donors without known chronic or malignant disease at the Department of Hematology at the Medical University of Graz. CfDNA was isolated from plasma using the QIAamp Circulating Nucleic Acids kit (QIAGEN, Hilden, Germany) in accordance with the manufacturer's protocol. Library preparation for WGS was performed as described previously [37].

Sequencing

Control and high-coverage tumor samples were sequenced on the Illumina NovaSeq S4 flowcell at 2x150bp by the Biomedical Sequencing Facility at CeMM, Vienna, Austria. For the control samples, an average of 435,135,450 (range: 352,904,231-556,303,420) paired-end reads were obtained. For the tumor samples (P40_1, P40_2, P147_1, P147_3, P148_1, P148_3, C2_6, C2_7), an average of 688,482,253 reads (range: 541,216,395-870,285,698) were sequenced. Additional samples were sequenced using the Illumina NextSeq platform (B7_1, B13_1, P190_3, P170_2, P179_4, P198_5, P240_1; average sequencing yield: 195,425,394 reads; range: 115,802,787-379,733,061).

Low-coverage tumor samples which were used to create single-entity pools were sequenced on either the Illumina Next-Seq or MiSeq platform. This resulted in 382,306,130 reads from 69 prostate cancer samples, 254,490,128 reads from 60 breast cancer samples and 604,080,473 reads from 100 colon cancer samples.

Characterization of plasma samples

Some plasma samples, i.e. of patients B7 and B13 [10] and P40, P147, and P148 [38] were previously analyzed within other studies. From these analyses, we had information regarding mutations, specific SCNAs, and tumor content of the plasma samples based on the algorithm ichorCNA [43].

P40: Mutations: *BRCA1*: NM_007294: Q975R; specific SCNAs: *TMPRSS2-ERG* fusion; *AR* amplification in sample 2; chr12 amplification (containing *ARID2*, *HDAC7*); tumor content: P40_1: 30%; P40_2: 24%;

P147: Mutations: *BRCA2*: T298fs; *TP53*: F338I; specific SCNAs: *RET* amplification in sample 3; *AR* amplification; *BRAF* amplification (7q34); *PTEN* loss; tumor content: P147_1: 52%; P147_3: 73%;

P148: Mutations: *TP53*: R213X; specific SCNAs: *MYC* amplification; *PTEN* loss; *FOXP1*, *RYBP*, *SHQ1* loss; *TMPRSS2-ERG* fusion; *AR* amplification (lost in P148_3); tumor content: P148_1: 38%; 148_3: 49%

C2: specific SCNAs: high level amplification on chromosome 12 (*KRAS*) in C2_6, not visible in C2_7; tumor content: C2_6: 18%; C2_7: 28%

Transcription factor binding site definitions

Data from the GTRD database were downloaded (http://gtrd.biouml.org/downloads/18.01/human_meta_clusters.interval.gz) and individual BED files per TF were extracted. The position was recalculated by focusing on the reported point where the meta-cluster has the highest ChIP-seq signal. An additional BED file was created which only included peaks that are supported by >50% of the maximum number of samples (>50%-TFBSs) analyzed for this specific transcription factor. All BED files were then converted to hg19 (from original hg38) using the liftOver tool provided by UCSC.

Transcription factor binding site overlaps

In order to check whether binding sites of transcription factors overlap, regions of the binding sites from GTRD (of the sites supported by > 50% of the samples) were increased by 25, 50 and 100bp, respectively, on either side using bedtools slop. Subsequently, the number of overlap was calculated by using bedtools intersect via pybedtools for every transcription factor with every other transcription factor.

Single-end sequencing data preparation

In order to enhance the nucleosome signal, reads were trimmed to remove parts of the sequencing read that are associated with the linker region. Hence, forward reads were trimmed to only contain base 53-113 (this would correspond to the central 60bp of a 166bp fragment). Reads were then aligned to the human hg19 genome using bwa and PCR duplicates were removed using samtools rmdup. Average coverage is calculated by bedtools genomecov.

Paired-end sequencing data preparation

Paired-end reads were aligned to the human hg19 genome using bwa mem and PCR duplicates were marked with picard MarkDuplicates

MNase-seq data preparation

BAM files of MNase-seq experiments of GM12878 were downloaded from the ENCODE portal. Reads in BAM files were trimmed directly from the BAM file using pysam. In brief, left-most alignment positions in the BAM file were shifted 53bp in the respective direction and the sequence length was adjusted to 60bp. The coverage patterns were then calculated in the same way as the trimmed cell-free DNA sequencing data.

Coverage patterns at transcription factor binding sites

For every transcription factor in the GTRD, coverage patterns were calculated. To this end, coverage data was extracted for every region using pysam count_coverage in a region +/- 1000 bp around the defined binding sites. Coverage data at every site were normalized by regional copy number variation and by mean coverage. For every position around the TFBS, coverage was averaged and 95% confidence intervals were calculated. If >100,000 positions were defined for a transcription factor, 100,000 sites were randomly chosen to be analyzed.

Insert sizes around transcription factor binding sites

To see whether fragment sizes around transcription factor binding sites were biased, insert size data from paired-end analyses were used. Every position from -1000 until 1000bp from the binding site was traversed and (single-end) reads where the central 3 bp around the midpoint are located at this position were fetched using pysam. Also, paired-end alignments from the same sample were fetched and the insert size information was designated to the respective reads. All insert sizes at specific positions relative to the TFBS were then summarized and 1000 data points were sampled and plotted for each position in the range of -1000 to 1000bp from the TFBS.

Measuring transcription factor binding site size

In order to measure the size of the transcription factor binding site, the respective coverage pattern was smoothed using a third order Savitzky-Golay filter (window-size: 31). Peaks were identified by searching for data points that were larger than the neighboring 20 data points on either side. Peaks were removed if they resided within 50bp of the center of the supposed binding site. The distance between the closest peaks next to the binding site peak was specified as the transcription factor binding site size.

Since binding site estimates are only reasonable if nucleosome synchronization is detectable, we filtered the signals by various criteria:

- High-frequency signal amplitude > 0.1
- Mean normalized coverage of the central 100bp < 1
- Amount of peaks is less than 15
- Median distance between peaks is >150bp
- The binding site sets comprises over 500 sites

228 binding site sets passed these filters and were used for binding site estimation.

Measurement of transcription factor accessibility using Savitzky-Golay filters

As we hypothesize that two distinct signals make up the coverage pattern, two signals of different frequencies were extracted. The lower range frequency data was extracted by a Savitzky-Golay filter (3rd order polynomial and window size of 1001). A high-frequency signal was extracted by a different Savitzky-Golay filter (3rd order polynomial and window size of 51). The high-frequency signal was then normalized by division by the results of the low-frequency signal. Subsequently, the data range of the high-frequency signal was recorded. Since coverage profiles from transcription factors with few described binding sites are inherently noisier, a LOESS smoothing was performed over the signal range and the amount of described binding sites. The range values were corrected by the smoothed LOESS and ranks of the adjusted range were calculated.

Measurement of transcription factor accessibility using wavelet transformation

As an additional method to measure accessibility of transcription factors, we applied wavelet transformation by using the R-package “WaveletComp”. For every signal, we recorded peaks in the power spectrum along the periods between 2 and 512 bp. The highest peak in the range between 135 and 235 (185bp +/- 50bp) was used to reconstruct a de-noised higher frequency nucleosome signal at that specific period. Moreover, any residual baseline was removed using detrending of the original data series. Three parameters of the reconstructed signal were analyzed: The maximum amplitude of the signal, the sum of the signal powers (amplitudes squared) and the sum of the absolute amplitudes along the 2000bp surrounding the transcription factor binding site.

For comparing tumor to normal samples, mean value and standard deviation for the respective parameters were recorded in normal samples for every transcription factor and z-scores were calculated by taking the respective parameter in the cancer sample, subtracting the mean value of the normal and dividing by the standard deviation.

Comparing tumor and control samples

In order to compare tumor and control samples, the ranks of the respective transcription factors in the adjusted range values were compared. Rank differences were calculated between a tumor sample and every control sample and mean rank differences were recorded. Moreover, z-scores were calculated for every transcription factor from the accessibility ranks, by taking the respective rank, subtracting the mean rank of the control samples and dividing by the standard deviation of this transcription factor ranks of the control samples.

DNase hypersensitivity data analysis

BAM_files from DNase hypersensitivity experiments were downloaded from the ENCODE database for GM12878, LNCaP and HCT116 cell lines. Binding site regions of a transcription factor were increased by 25bp on either side using bedtools slop. Coverage at the respective binding sites was extracted using mosdepth and normalized by million mapped reads per sample.

Analysis of somatic copy-number alterations (SCNAs)

For control data, paired-end alignments were subsampled using samtools view to only include 2% of the initial alignments and converted to FastQ using samtools fastq. For the cancer samples, separate low-coverage whole-genome sequencing was performed. Plasma-Seq [37] was applied to the subsampled FastQ files and the low-coverage data of the cancer samples, respectively. In brief, reads were aligned to the human hg19 genome and reads were counted within pre-specified bins. The bin size was determined by the amount of theoretically mappable positions to account for differences in mappability throughout the genome. Read counts were normalized for total amount of reads and GC content of bins were corrected for by LOESS smoothing over the GC spectrum. Moreover, corrected read counts were normalized by the mean read counts of non-cancer controls per bin to control for additional positional variation.

Data and code availability

Sequencing raw data are available in EBI-EGA under the accession EGAS00001003206. Code is available in GitHub at https://github.com/PeterUlz/TranscriptionFactorProfiling.

Supplementary Information

Coverage-independent and spatial cfDNA fragment analyses

As a coverage independent confirmation of our TFBS signals, we plotted the length of each cfDNA fragment as a function of the distance of the fragment midpoint to the CTCF binding site [25]. The resulting heatmap confirmed the signal periodicity consistent with the coverage-based oscillating pattern (Supp. Fig. 2b). In addition, in order to more closely analyze the landscape of fragments related to the single recognition sequences, we computed the spatial density of cfDNA fragments within a 2kb region centered on the TFBSs and ranked the sites according to the coverage of the central 40bp. The resulting heatmap showed that nucleosome phasing in most sites analyzed is even (Supp. Fig. 2c), which is again consistent with the coverage profiles.

TF-specific nucleosome coverage profiles

To establish the shape of TFBSs, we investigated TF-specific nucleosome coverage profiles, as we observed that some TFs showed evenly spaced nucleosome peaks, including their binding sites (e.g. PU.1 and GRHL2 in Fig. 1d), whereas other TFs had wider troughs at their binding sites (e.g. CREM in Supp. Fig. 2a), resembling those which we have previously described for TSSs [10]. Altogether, we identified 55 TFBSs where the TFBS exceeded 300 bp and from these, 26 had binding sites close to di-nucleosomal sizes (312-352 bps; Supp. Fig. 4d; Supp. Table 5). For these patterns, we found highly significant increases of overlap for both CpG islands ($p=4.2x10^{-11}$; Mann-Whitney U test) and TSSs ($p=8.5x10^{-12}$; Mann-Whitney U test) for TFBSs with sizes >300 bp (Supp. Fig. 4e).

Pairwise comparison of plasma samples

To address the question whether TF accessibility remains stable over time, we also analyzed two samples each from patients P40, P147, and C2). However, with our very stringent criteria, i.e. by confining the analyses to 1,000-msTFBSs, we did not observe significant differences in these plasma sample pairs (Controls: Median: 0.8404 ±0.0196 (IQR); P40: 0.8620; P147: 0.8370; C2: 0.8719; each Kendall's Tau) (Supp. Fig. 8, Supp. Table 8).

Between P147_1 and P147_3, a novel, high-amplitude amplification including the *RET* gene evolved, whereas C2_7 had lost an amplification including *KRAS*, which we had observed in the previous sample C2_6. *RET* in prostate cancer and *KRAS* in CRC both may affect the PI3K/AKT/mTOR pathway [44] (Pek et al. Oncogene 36:4975-4986) and we therefore investigated downstream targets such as the TF CREB; however, the accessibility was not different from the control plasma samples and furthermore remained unchanged. Between P40_1 and P40_2, resistance against androgen deprivation therapy (ADT) had evolved, which was reflected in a high-level amplification of the AR gene [45]. However, if AR expanded its repertoire of transcriptional targets, it did not become apparent at the aforementioned T-ARBSs and N-ARBSs [36] (Supp. Fig. 9). Our very conservative approach limiting the analyses to 1,000-msTFBSs may explain why we may not have observed differences between these samples.

Acknowledgements

This work was supported by CANCER-ID, a project funded by the Innovative Medicines Joint Undertaking (IMI JU; #115749-1), by the BioTechMed-Graz flagship project

“EPIAge”, and by the Christian Doppler Research Fund for Liquid Biopsies for Early Detection of Cancer.

Author contributions

PU and MRS designed the study, SP, TM, QZ and JB performed the experiments. PU, JBG, EH and MRS analyzed the data. AW, AZ, AG, GP, EP, TB and HS provided clinical samples and clinical information. PU, EH, CB and MRS supervised the study and wrote the manuscript. All authors revised the manuscript.

Competing financial interests

The authors declare the following competing financial interests: A patent application has been filed for aspects of the paper (inventors: PU; EH, MRS). PU is a technical consultant to Freenome Inc. EH and MRS have an unrelated sponsored research agreement with Servier within CANCER-ID, a project funded by the Innovative Medicines Joint Undertaking (IMI JU), the salary of JB was paid through this arrangement. The other authors have no competing interests to declare.

Figures

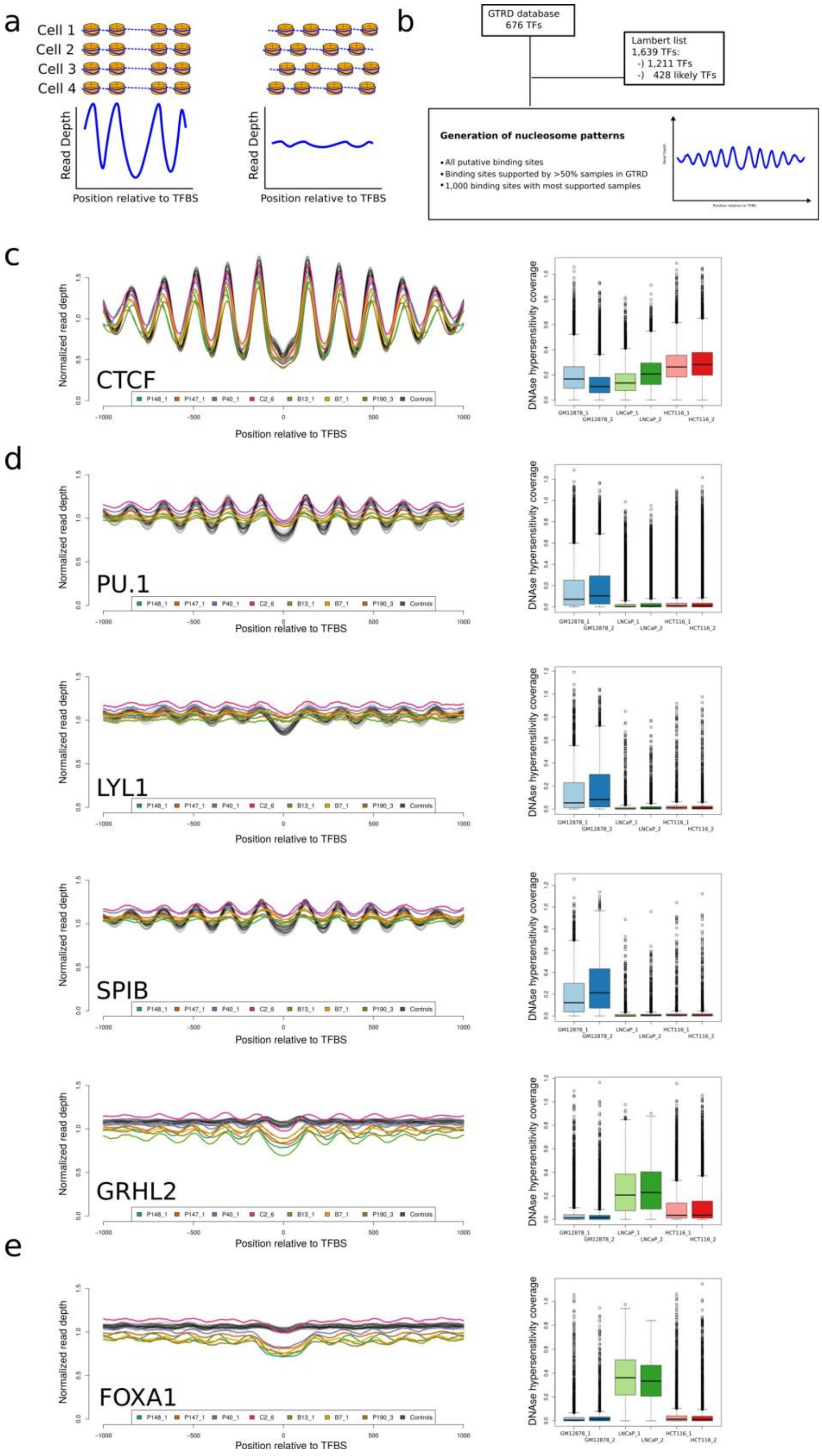


Figure 1

Establishment of TF-nucleosome interactions from cfDNA

a) Regions with highly organized, i.e. phased, nucleosomes result in an oscillating read depth pattern where a peak of reads indicate the positions of dyads, i.e. the midpoint of a canonical nucleosome. A less-defined positioning of nucleosomes yields a rather flat coverage profile.

b) TFBS data for 676 TFs were retrieved from the GTRD [17] and aligned with a curated list of known or likely human TFs [1]. Three different calculations, each with increased stringency, were conducted (for details see text).

c) The coverage pattern of CTCF is similar across all analyzed cfDNAs, which is consistent with DNase hypersensitivity data showing approximately equal accessibility in blood (GM12878) and epithelial tissues, i.e. prostate (LNCaP) and colon (HCT116). In this panel and in the respective subsequent panels, the profiles calculated from healthy controls are shown in gray whereas the patient-derived profiles are displayed in the indicated colors.

d) The hematopoietic lineage-specificity of TFs (PU.1, LYL1, SPIB) was confirmed by DNA hypersensitivity assays and their amplitude is reduced in plasma from cancer patients compared to healthy controls. In contrast, the amplitudes for the epithelial TF GRHL2 increase in cfDNA from patients with cancer.

e) Accessibility plots and DNase hypersensitivity for TF FOXA1 illustrating the preferential amplitude change in patients with hormone-dependent cancers, i.e. prostate and breast cancer.

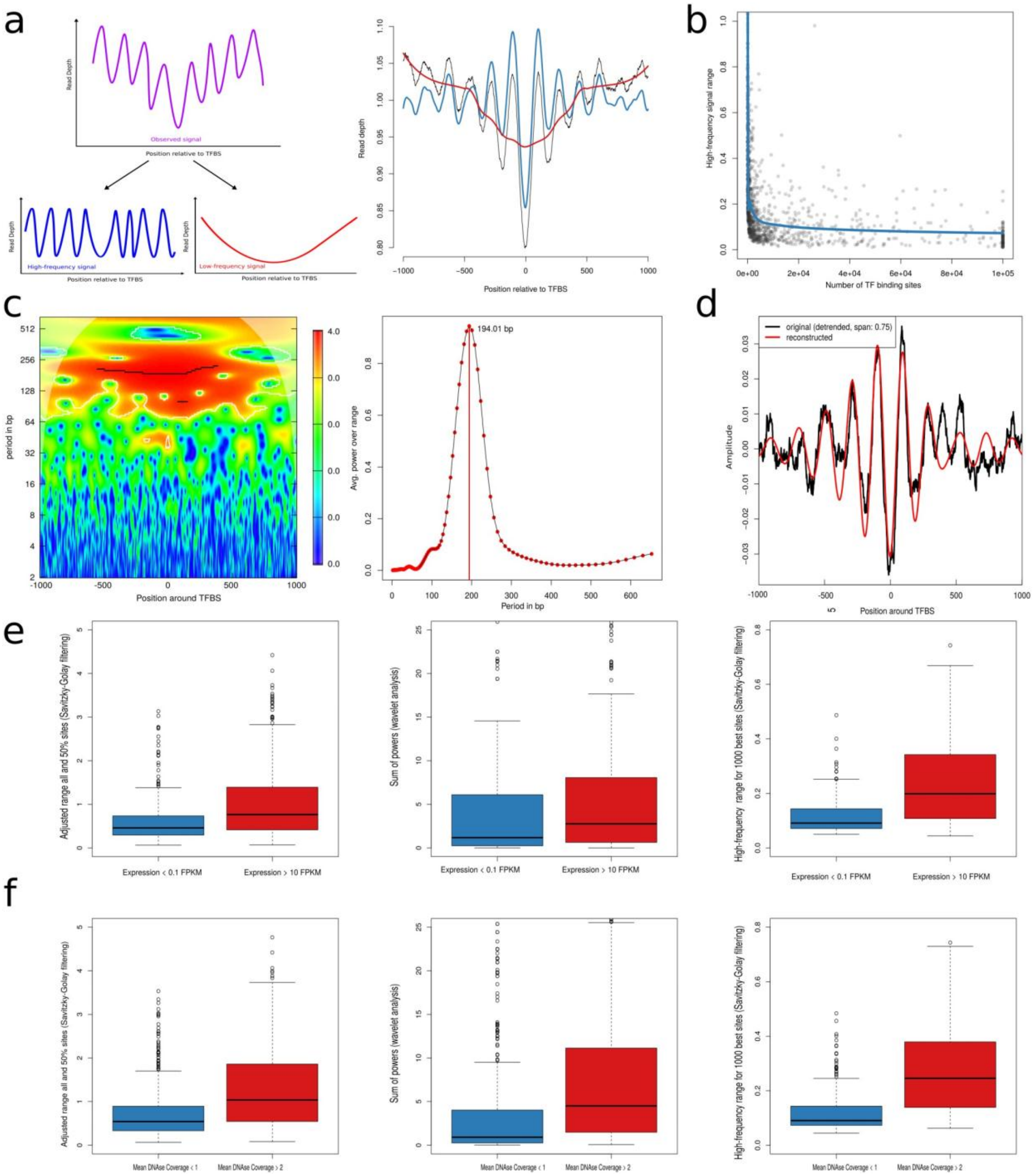


Figure 2

Accessibility score for the characterization of TFBSs

a) To measure TF accessibility, the observed raw coverage signal (purple in left and black in right panel) was split by Savitzky-Golay filtering into a low-frequency signal (red) and a high-frequency signal (blue) using different window sizes. The right panel illustrates an overlay of these three signals. The high-frequency signal is used as a measure for accessibility.

b) The range of the high-frequency signal (Y-axis) critically depends on the number of TFBSs (X-axis), as TFs with few binding sites have more noise due to lesser averaging. A LOESS model is fitted (blue) in order to correct for this bias.

c) Wavelet analysis of GRHL2: Heatmap of periods along the region surrounding the TFBSs of GRHL2 (left panel). Color code represents quantiles of the signal power distribution. Average power of periods of transcription factor GRHL2 (right panel).

d) Detrended original (black) and reconstructed (red) nucleosome coverage profile of transcription factor GRHL2 resulting from wavelet analysis.

e) All tested procedures (left: >50%-TFBSs, Savitzky-Golay filtering; center: the sum of powers, wavelet analysis; right: 1,000-msTFBSs, Savitzky-Golay filtering), showed increased values as a measure of accessibility for transcription factors that are expressed in blood (>10 FPKM), but not in genes that show no or low signs of expression (<0.1 FPKM).

f) Transcription factors with a mean DNase hypersensitivity coverage of >2 in GM12878 DNase data from the ENCODE project have higher adjusted ranges and higher sum of powers than factors that have a mean coverage of <1 in all three analyses conducted (left: >50%-TFBSs, Savitzky-Golay filtering; center: the sum of powers, wavelet analysis; right: 1,000-msTFBSs, Savitzky-Golay filtering).

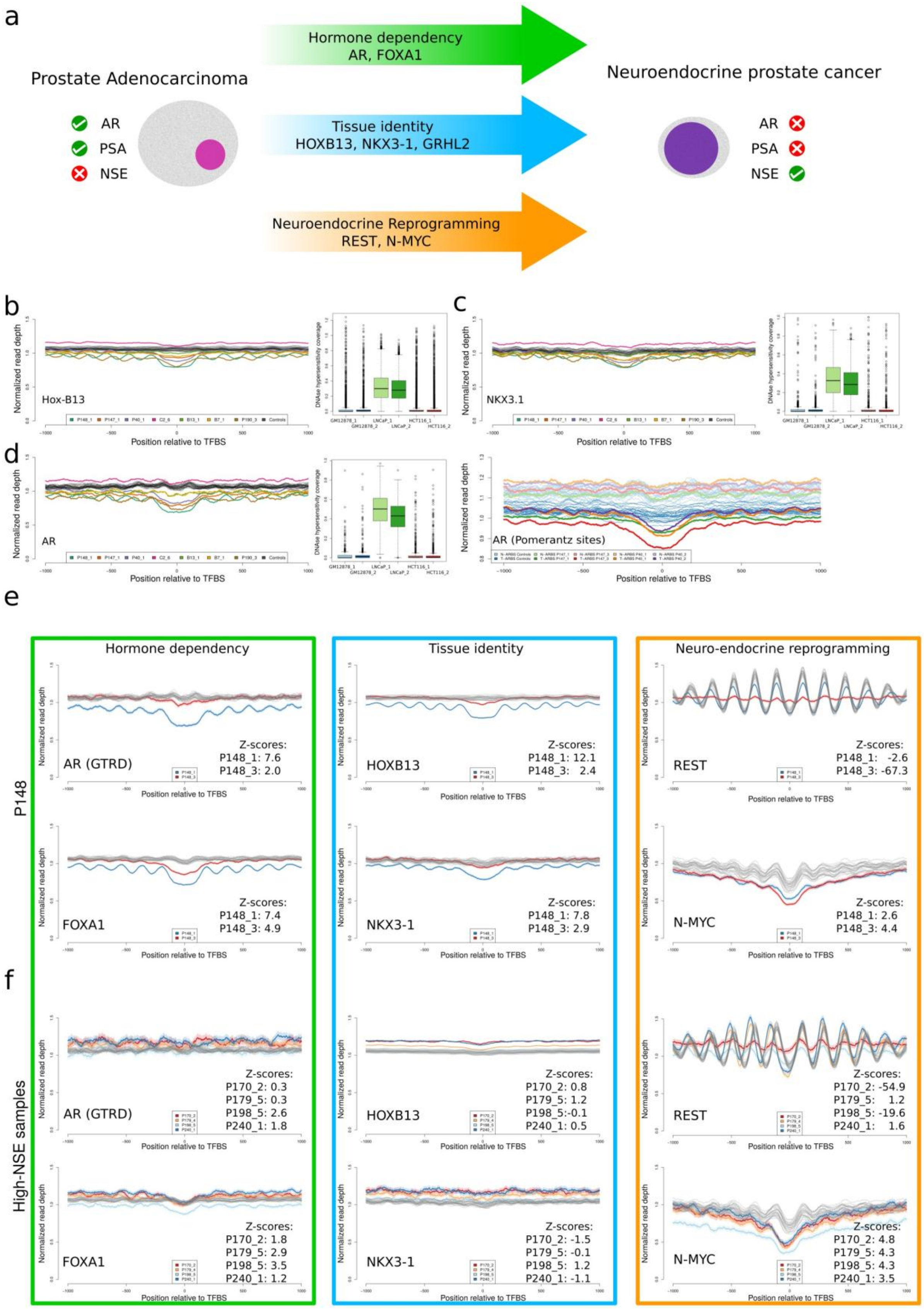


Figure 3

Prostate lineage-specific TFs, their plasticity and suitability for tumor classification

a) Prostate adenocarcinomas are AR-dependent and accordingly have frequently increased PSA (prostate-specific antigen) levels and normal NSE (neuron-specific enolase) values. In contrast, t-SCNC are no longer dependent on AR and usually have low PSA and increased NSE levels. Several TFs involved in the transdifferentiation process from an adenocarcinoma to a t-SCNC have been identified and are indicated in the arrows.

b) Accessibility profile of the prostate lineage-specific homeobox TF HOXB13 and the respective DNase hypersensitivity assays of prostate cancer cell line LNCaP. In this and the subsequent panels, the profiles calculated from healthy controls are shown in gray whereas the patient-derived profiles are displayed in the indicated colors.

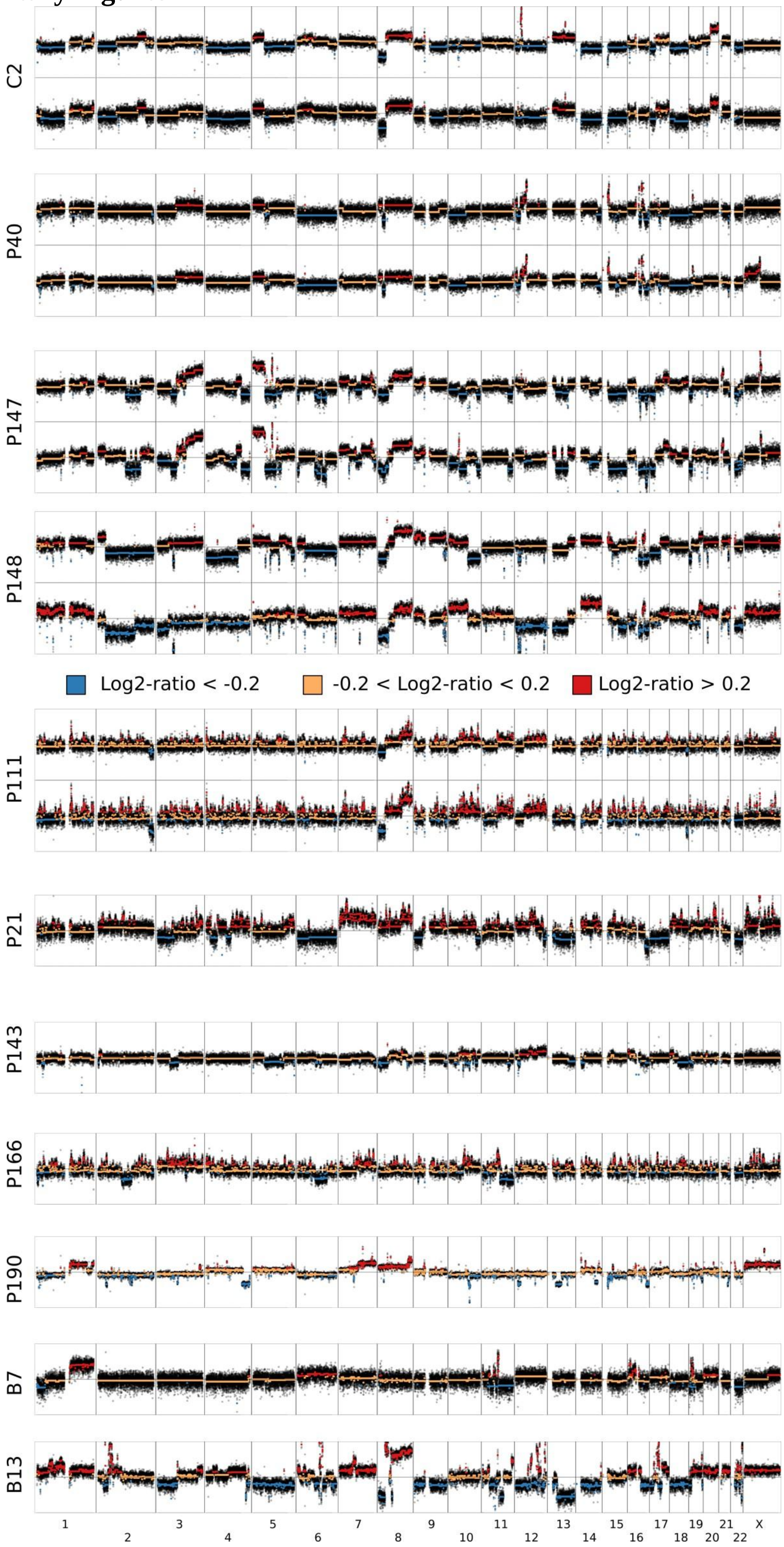
c) Accessibility pattern and DNA hypersensitivity assay of NKX3-1, one of the earliest genes expressed during prostatic epithelium maturation.

d) AR accessibility was analyzed for all AR binding sites in the GTRD and in addition for AR binding sites with higher binding intensity in tumors (T-ARBSs), and for sites with high binding intensity in normal samples (normal AR binding sites, N-ARBSs) [36]. The well-established lineage specificity of AR was confirmed by DNA hypersensitivity assays.

e) Coverage pattern changes during transdifferentiation from an adenocarcinoma to a neuroendocrine carcinoma established from two plasma samples from patient P148 for hormone-dependent (AR, FOXA1), tissue identity-specific (HOXB13, NKX3-1), and neuroendocrine reprogramming (REST, N-MYC) TFs.

f) Analysis of the same TFs as in a) from 4 plasma samples from patients with neuroendocrine prostate cancers.

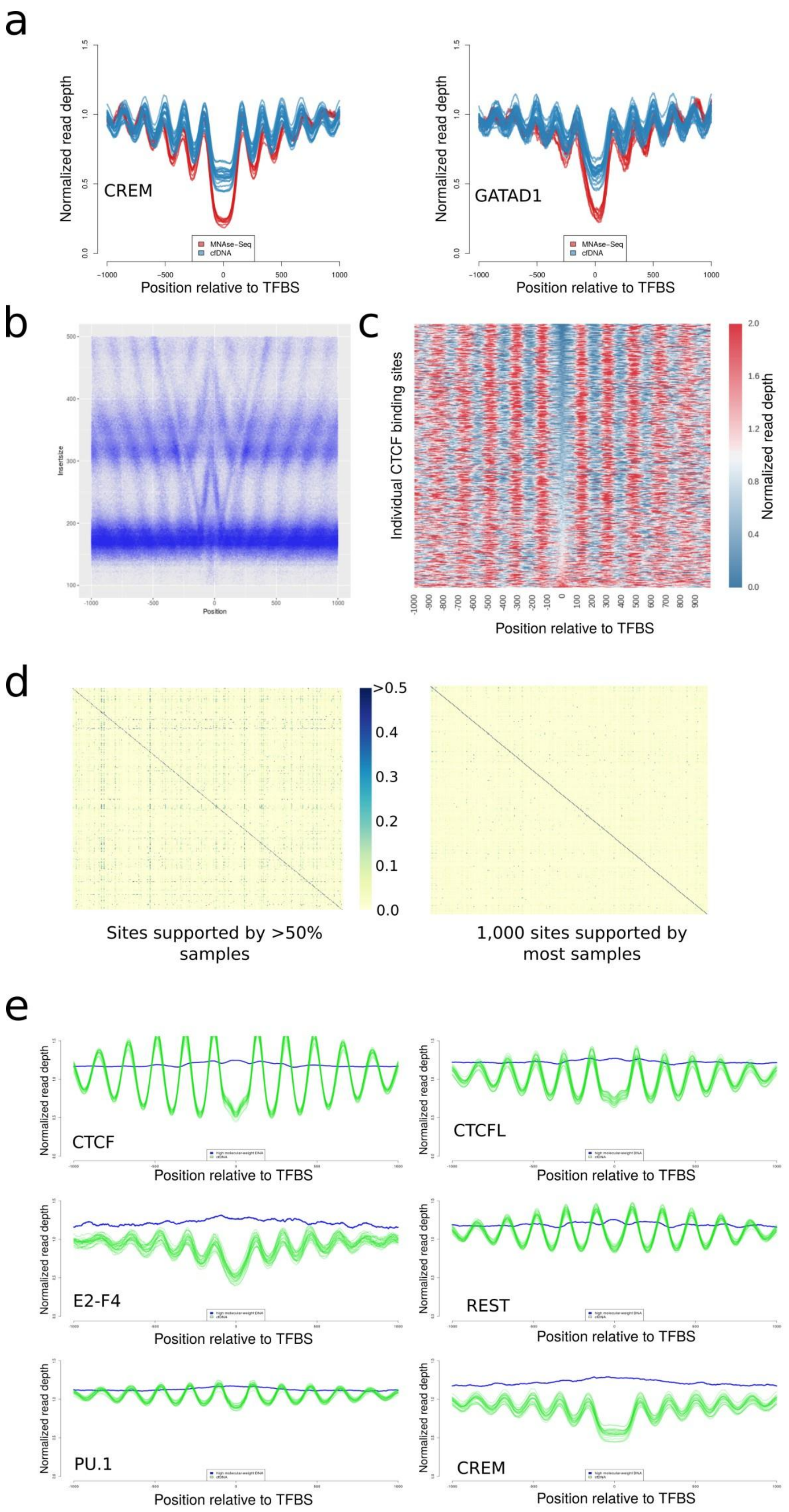
Supplementary Figures



Supplementary Figure 1

Somatic copy number alterations (SCNAs) in plasma samples from patients with cancer

SCNAs identified after whole-genome sequencing of 8 plasma samples from four patients (C2, P40, P147, P148).



Supplementary Figure 2

TF-nucleosome interaction map for 676 high-confidence TFs with reliable binding site information

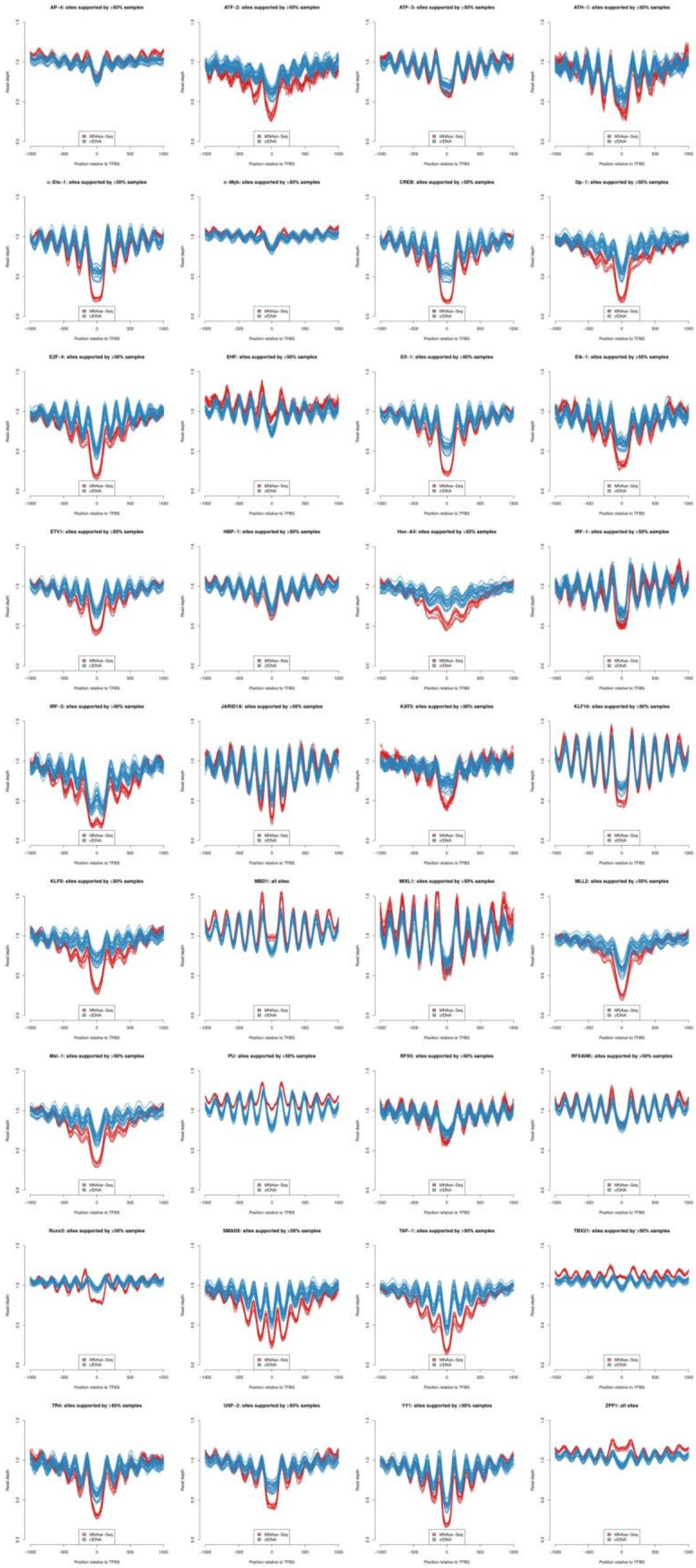
a) TFBS-nucleosome coverage profiles for two representative TFs, CREM and GATAD1, established from 24 cfDNA samples from healthy controls, each shown with an individual blue line. The MNase-seq coverage patterns from the lymphoblastoid cell line GM12878 obtained from ENCODE are illustrated in red. Additional MNase plots are illustrated in Supp. Fig. 9.

b) Heatmap of fragment sizes around CTCF binding sites displayed as a plot of the length of each sequencing read (Y-axis) as a function of the distance from the fragment midpoint to the center of the site for each annotated feature (X-axis) (details in Supp. Info.).

c) Heatmap of individual CTCF binding sites and surrounding regions. Regions are ordered by the coverage within the central 50bp around the TFBS. The spatial density of cfDNA fragments within a 1kb region centered on the TFBSs were computed and ranked.

d) Matrices of overlaps between TFBSs (left panel: all 676 GRTD TFs; right panel: 505 TFs with the 1,000-msTFBSs). Each point represents the percentage of overlaps (within 50bp) in binding site definitions (complete list of TFs in Supp. Table 3).

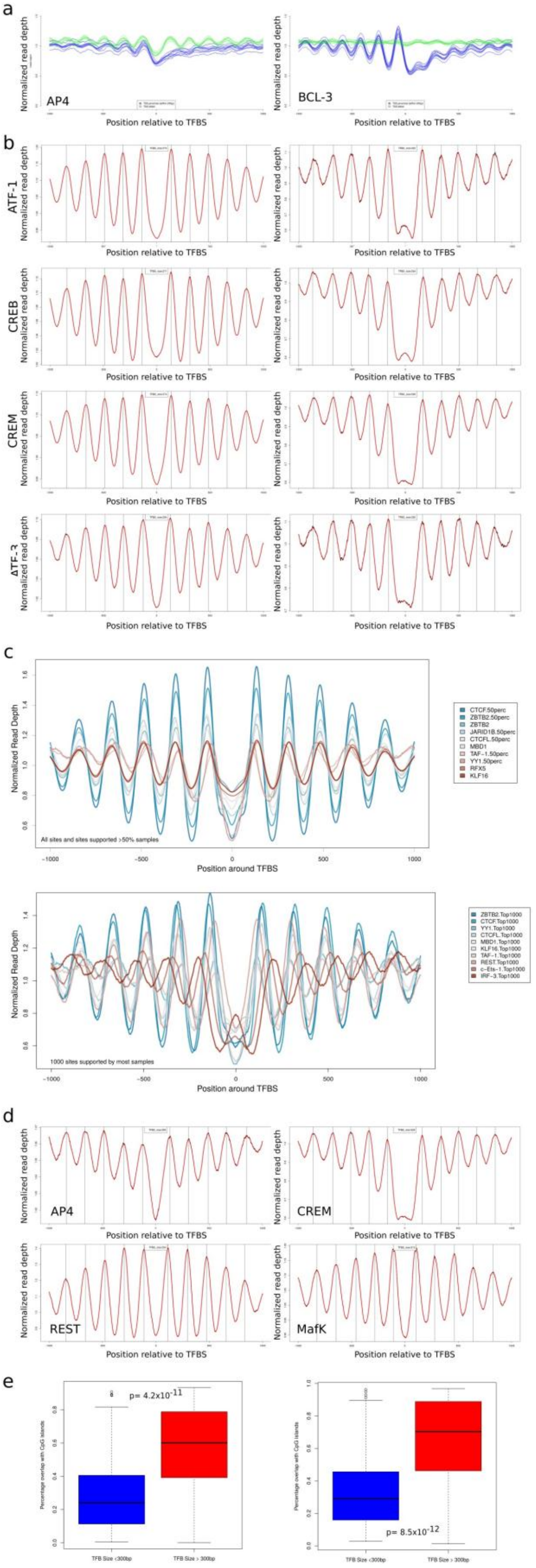
e) TFBS analyses with high molecular weight DNA, which is not mono-nucleosomal DNA, yields a uniform, non-oscillating pattern (blue) in contrast to plasma DNA (green).



Supplementary Figure 3

Further TF-nucleosome interaction maps

Additional comparisons between coverage profiles of cfDNA and MNAse-seq around TFBSs.



Supplementary Figure 4

The shape of TFBSs

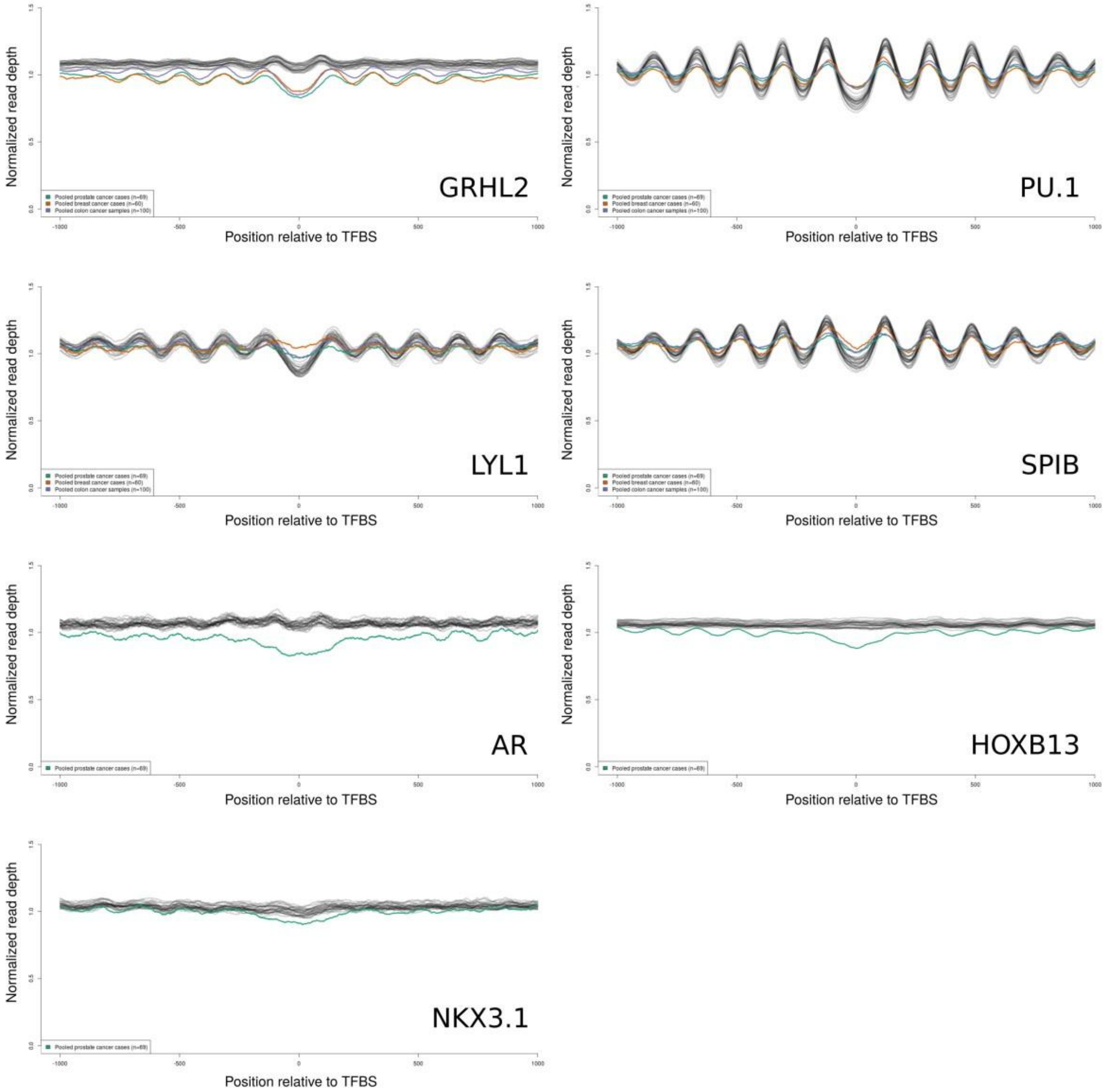
a) Coverage profiles for TFs AP-4 and BCL-3 after calculations conducted separately for TFBS within and outside of TSSs.

b) Analyses of TFBSs for TFs ATF1, CREB, CREM and ATF-3 may result in evenly spaced or in TSS-like coverage patterns, dependent on whether all tissues in the GTRD or whether, more strictly, only those peaks that are supported by >50% of the maximum number of samples (>50%-TFBSs) were included.

c) Exemplary TF-nucleosome profiles calculated for all and >50%-TFBS (upper panel) and for 1,000-msTFBSs (lower panel), illustrating the variable nucleosome patterns of different TFs in cfDNA.

d) Measurements of TFBS widths revealed substantial differences among various TFBSs.

e) Boxplot illustrating the percentage of overlap for CpG islands (left panel) and TSSs (right panel).

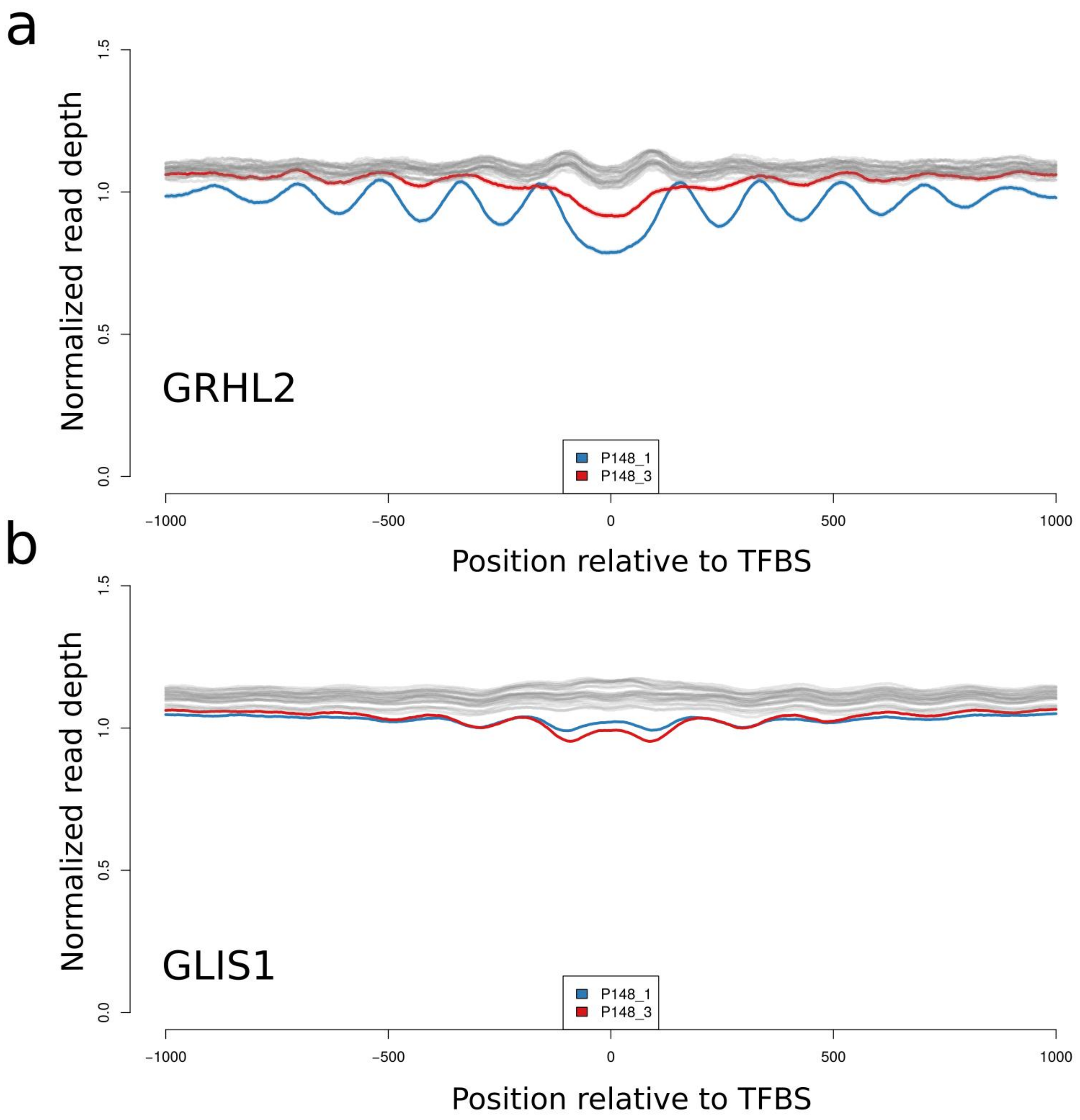


Supplementary Figure 5

Analyses of pooled shallow-coverage cfDNA

Accessibility in pooled cfDNA samples from prostate (*n*=69), colon (*n*=100) and breast (*n*=60) cancer cases of the epithelial TF GRHL2 and of hematopoietic TFs (PU.1, LYL1, and SPIB).

Accessibility within the prostate cancer cfDNA pool of the lineage-specific TFs AR, HOXB13, and NKX3-1.

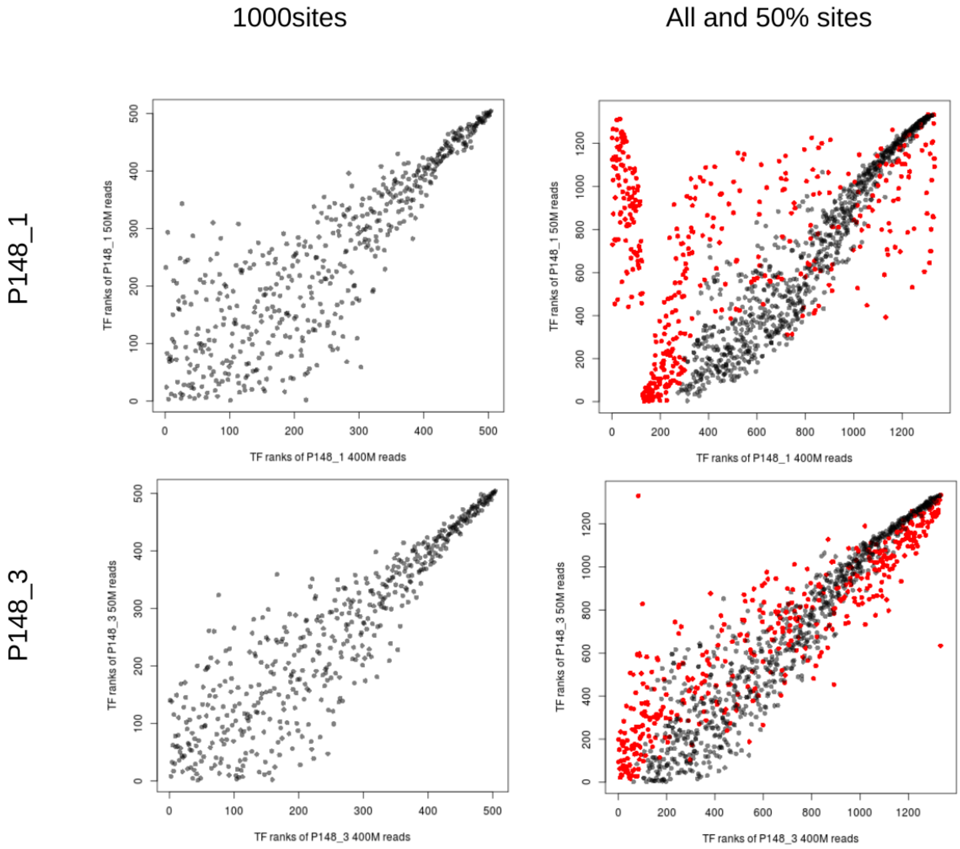


Supplementary Figure 6

TFs involved in transdifferentiation from an adenocarcinoma to a t-SCNC

a) GRHL2 accessibility in plasma samples P148_1 and P148_3.

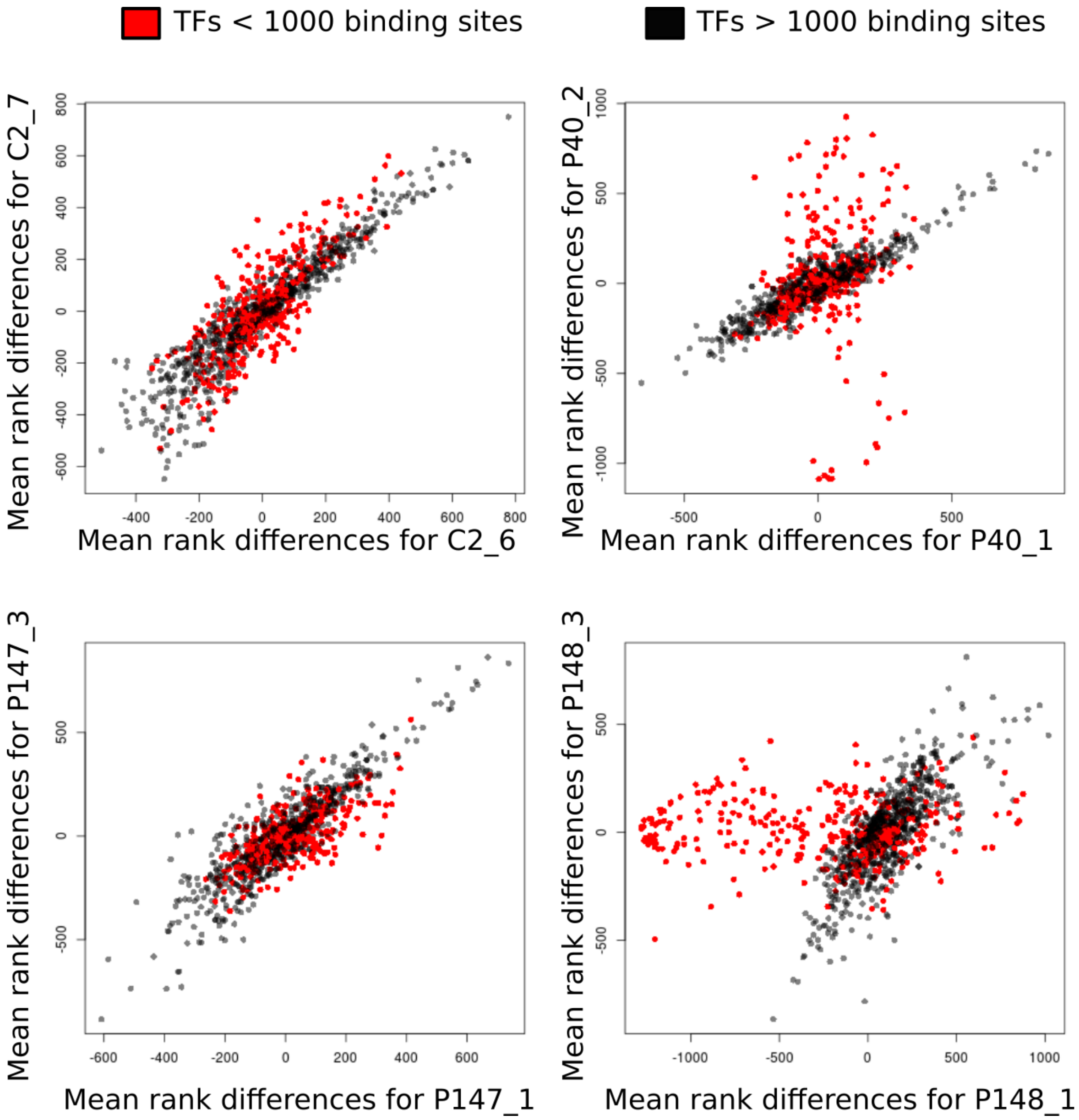
b) Analysis of GLIS1 in the two plasma samples from patient P148.



Supplementary Figure 7

Down-sampling of plasma samples P148_1 and P148_3

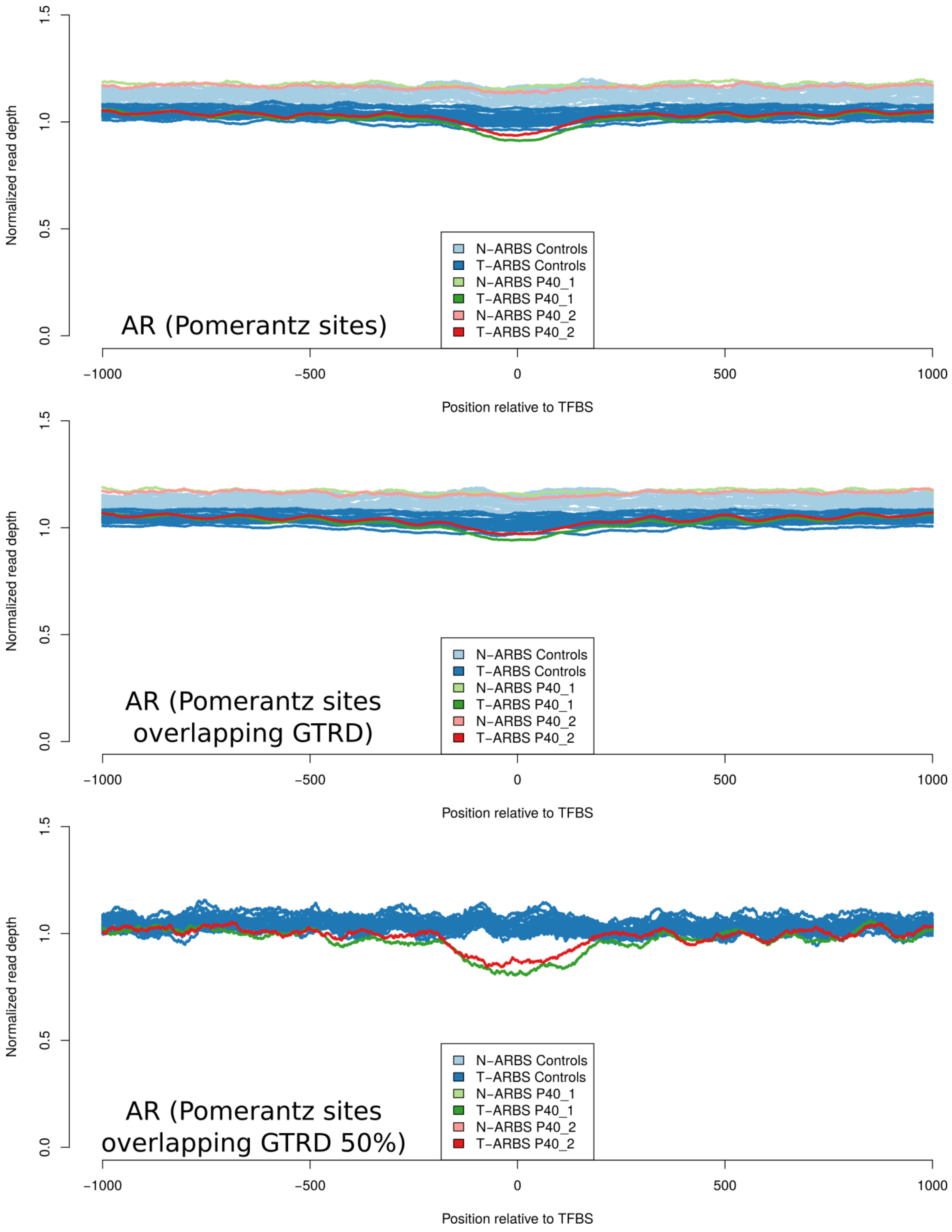
Plasma samples P148_1 (819,607,690 reads) and P148_3 (768,763,081 reads) were down-sampled to ~50 million reads and analyzed for 1,000-msTFBSs (left column) and all and >50%-TFBSs (right column). The analysis suggests that preferentially TFs with a low number of TFBSs are affected by increased noise.



Supplementary Figure 8

Comparison of TFBS accessibility in serial analysis

Plots of correlation between serial samples from patients C2, P147, P40, and P148. The X-axis represents the first, the Y-axis the second plasma sample.



Supplementary Figure 9

Analyses of AR binding sites for plasma samples from P40

This patient with prostate cancer received ADT treatment and developed a high-level AR amplification between samples P40_1 and P40_2.

Supplementary Tables

Supplementary Table 1

Clinical data on prostate cancer patients P40, P147, P148, P190, P170, P179, P198, and P240, on breast cancer patients B7 and B13, and patient C2 with colorectal cancer. Some plasma samples, i.e. of patients B7 and B13 [10] and P40, P147, and P148 [38] were previously analyzed within other studies.

Supplementary Table 2

Overview of samples and sequencing statistics with accompanying information about tumor fractions from ichorCNA.

Supplementary Table 3

List of TFs from the GTRD used in this study and accompanying information.

Supplementary Table 4

Pairwise analysis of overlaps within 50bp as a fraction of the first TF for TFs with > 1000 defined binding sites.

Supplementary Table 5

List of filtered 228 TFBS sets with information about estimated binding site size and details about nucleosome peak calls.

Supplementary Table 6

Comparison between TF accessibility of tumor samples and control samples by mean rank differences between a tumor sample and all control samples and Z-scores for every TF for all sites and 50%-sites sets.

Supplementary Table 7

Comparison between TF accessibility of tumor samples and control samples by mean rank differences between a tumor sample and all control samples and Z-scores for every TF for the 1,000-msTFBSs sets.

Supplementary Table 8

Comparison between TF accessibility of paired samples by rank differences.

References

1. Lambert SA, *et al.* The Human Transcription Factors. *Cell* **172**, 650-665 (2018).

2. Lai WKM, Pugh BF. Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nat Rev Mol Cell Biol* **18**, 548-562 (2017).

3. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213-1218 (2013).

4. Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

5. Lai B, *et al.* Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. *Nature* **562**, 281-285 (2018).

6. Rodriguez-Bravo V, Carceles-Cordon M, Hoshida Y, Cordon-Cardo C, Galsky MD, Domingo-Domenech J. The role of GATA2 in lethal prostate cancer aggressiveness. *Nat Rev Urol* **14**, 38-48 (2017).

7. Corces MR, *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, (2018).

8. Lo YM, *et al.* Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Science translational medicine* **2**, 61ra91 (2010).

9. 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* **164**, 57-68 (2016).

10. Ulz P, *et al.* Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nature genetics* **48**, 1273-1278 (2016).

11. Siravegna G, Marsoni S, Siena S, Bardelli A. Integrating liquid biopsies into the management of cancer. *Nature reviews Clinical oncology* **14**, 531-548 (2017).

12. Wan JC, *et al.* Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nature reviews Cancer* **17**, 223-238 (2017).

13. Alix-Panabieres C, Pantel K. Clinical Applications of Circulating Tumor Cells and Circulating Tumor DNA as Liquid Biopsy. *Cancer discovery* **6**, 479-491 (2016).

14. Heitzer E, Haque IS, Roberts CES, Speicher MR. Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nature Reviews Genetics*, (2018).

15. Lui YY, Chik KW, Chiu RW, Ho CY, Lam CW, Lo YM. Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. *Clinical chemistry* **48**, 421-427 (2002).

16. Sun K, *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proceedings of the National Academy of Sciences of the United States of America* **112**, E5503-5512 (2015).

17. Yevshin I, Sharipov R, Valeev T, Kel A, Kolpakov F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic acids research* **45**, D61-D67 (2017).

18. Ong CT, Corces VG. CTCF: an architectural protein bridging genome topology and function. *Nature reviews Genetics* **15**, 234-246 (2014).

19. Fu Y, Sinha M, Peterson CL, Weng Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS genetics* **4**, e1000138 (2008).

20. Sizemore GM, Pitarresi JR, Balakrishnan S, Ostrowski MC. The ETS family of oncogenic transcription factors in solid tumours. *Nature reviews Cancer* **17**, 337-351 (2017).

21. Zohren F, *et al.* The transcription factor Lyl-1 regulates lymphoid specification and the maintenance of early T lineage progenitors. *Nat Immunol* **13**, 761-769 (2012).

22. Solomon LA, Li SK, Piskorz J, Xu LS, DeKoter RP. Genome-wide comparison of PU.1 and Spi-B binding sites in a mouse B lymphoma cell line. *BMC Genomics* **16**, 76 (2015).

23. Jacobs J*, et al.* The transcription factor Grainy head primes epithelial enhancers for spatiotemporal activation by displacing nucleosomes. *Nature genetics* **50**, 1011-1020 (2018).

24. Augello MA, Hickey TE, Knudsen KE. FOXA1: master of steroid receptor function in cancer. *The EMBO journal* **30**, 3885-3894 (2011).

25. Henikoff JG, Belsky JA, Krassovsky K, MacAlpine DM, Henikoff S. Epigenome characterization at single base-pair resolution. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 18318-18323 (2011).

26. Koh W*, et al.* Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 7361-7366 (2014).

27. Aggarwal R*, et al.* Clinical and Genomic Characterization of Treatment-Emergent Small-Cell Neuroendocrine Prostate Cancer: A Multi-institutional Prospective Study. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **36**, 2492-2503 (2018).

28. Puca L*, et al.* Patient derived organoids to model rare prostate cancer phenotypes. *Nature communications* **9**, 2404 (2018).

29. Puca L, Vlachostergios PJ, Beltran H. Neuroendocrine Differentiation in Prostate Cancer: Emerging Biology, Models, and Therapies. *Cold Spring Harbor perspectives in medicine*, (2018).

30. Beltran H*, et al.* Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nature medicine* **22**, 298-305 (2016).

31. Uhlen M*, et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

32. Labbe DP, Brown M. Transcriptional Regulation in Prostate Cancer. *Cold Spring Harbor perspectives in medicine*, (2018).

33. Baca SC*, et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666-677 (2013).

34. Ewing CM*, et al.* Germline mutations in HOXB13 and prostate-cancer risk. *The New England journal of medicine* **366**, 141-149 (2012).

35. Tan PY, Chang CW, Chng KR, Wansa KD, Sung WK, Cheung E. Integration of regulatory networks by NKX3-1 promotes androgen-dependent prostate cancer survival. *Molecular and cellular biology* **32**, 399-414 (2012).

36. Pomerantz MM*, et al.* The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. *Nature genetics* **47**, 1346-1351 (2015).

37. Heitzer E*, et al.* Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing. *Genome Med* **5**, 30 (2013).

38. Ulz P*, et al.* Whole-genome plasma sequencing reveals focal amplifications as a driving force in metastatic prostate cancer. *Nature communications* **7**, 12008 (2016).

39. Khalesi E*, et al.* The Kruppel-like zinc finger transcription factor, GLI-similar 1, is regulated by hypoxia-inducible factors via non-canonical mechanisms. *Biochemical and biophysical research communications* **441**, 499-506 (2013).

40. Yamasaki M, Nomura T, Sato F, Mimata H. Chronic hypoxia induces androgen-independent and invasive behavior in LNCaP human prostate cancer cells. *Urologic oncology* **31**, 1124-1131 (2013).

41. Svensson C*, et al.* REST mediates androgen receptor actions on gene repression and predicts early recurrence of prostate cancer. *Nucleic acids research* **42**, 999-1015 (2014).

42. Beltran H*, et al.* Molecular characterization of neuroendocrine prostate cancer and identification of new drug targets. *Cancer discovery* **1**, 487-495 (2011).

43. Adalsteinsson VA, *et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nature communications* **8**, 1324 (2017).

44. Ban K, Feng S, Shao L, Ittmann M. RET Signaling in Prostate Cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **23**, 4885-4896 (2017).

45. Visakorpi T, *et al.* In vivo amplification of the androgen receptor gene and progression of human prostate cancer. *Nature genetics* **9**, 401-406 (1995).