nature genetics

Inferring expressed genes by whole-genome sequencing of plasma DNA

Peter Ulz[1], Gerhard G Thallinger[2,3], Martina Auer[1], Ricarda Graf[1], Karl Kashofer[4], Stephan W Jahn[4], Luca Abete[4], Gunda Pristauz[5], Edgar Petru[5], Jochen B Geigl[1], Ellen Heitzer[1] & Michael R Speicher[1]

The analysis of cell-free DNA (cfDNA) in plasma represents a rapidly advancing field in medicine. cfDNA consists predominantly of nucleosome-protected DNA shed into the bloodstream by cells undergoing apoptosis. We performed whole-genome sequencing of plasma DNA and identified two discrete regions at transcription start sites (TSSs) where nucleosome occupancy results in different read depth coverage patterns for expressed and silent genes. By employing machine learning for gene classification, we found that the plasma DNA read depth patterns from healthy donors reflected the expression signature of hematopoietic cells. In patients with cancer having metastatic disease, we were able to classify expressed cancer driver genes in regions with somatic copy number gains with high accuracy. We were able to determine the expressed isoform of genes with several TSSs, as confirmed by RNA-seq analysis of the matching primary tumor. Our analyses provide functional information about cells releasing their DNA into the circulation.

cfDNA from plasma is an intensively investigated biomarker. Numerous oncology publications have demonstrated that analyses of cancer-cell-derived DNA in the circulation, referred to as circulating tumor DNA (ctDNA), can be used to track tumor dynamics in real time[1–5]. cfDNA fragments have been associated with the release of DNA from apoptotic cells after enzymatic processing, as the distribution of their lengths has a mode near 166 bp, a size that corresponds approximately to the length of DNA wrapped around a nucleosome (~147 bp) plus a linker fragment (~20 bp)[6–8]. Indeed, evidence that cfDNA reflects nucleosome footprints was recently reported[9].

Importantly, micrococcal nuclease (MNase) assays, in which MNase digestion is used to produce mononucleosome-bound DNA fragments, have identified specific nucleosome patterns at promoters, which profoundly influence gene regulation[10–13]. In actively transcribed genes, the promoter region—the region of about 150 bp upstream of the TSS—is a nucleosome-depleted region (NDR) that facilitates access to the bulky transcriptional machinery and is flanked by arrays of well-positioned nucleosomes[11–13]. Furthermore, a reduction in nucleosome occupancy was found that extended up to 1 kb into the gene body, resulting in reduced frequencies of mapped reads[11,13]. Inactive promoters, by contrast, exhibited neither pronounced depletion nor strong positioning and phasing of nucleosomes[13].

Our investigation was hence threefold: to determine whether plasma DNA is able to reflect such expression-specific nucleosome occupancy at promoters; to assess whether plasma DNA possesses the sensitivity and accuracy to predict whether genes are expressed; and to determine whether blood samples from patients with cancer are informative for expressed cancer driver genes. To these ends, we conducted paired-end sequencing of 179 plasma samples and whole-genome sequencing of plasma DNA from 50 male and 54 female donors and 2 patients with breast cancer. We then analyzed 426 additional plasma samples from patients with cancer for their suitability for nucleosome promoter analysis.

Analyses of 179 paired-end-sequenced plasma DNA samples confirmed the expected unimodal size distribution of plasma nuclear DNA fragments, with a narrow range and mode at 166 bp, differing from plasma mitochondrial DNA, in which higher-order nucleosome packaging is absent[14] (**Fig. 1a**).

Sequencing of DNA fragments after MNase digestion has generated nucleosome maps where dyads (regions occupied by the center of a nucleosome) of 'perfectly positioned' nucleosomes—sites with high nucleosome preferences—corresponded to strong read peaks, reflecting the phasing of nucleosomes, whereas dyads of less preferentially positioned nucleosomes showed reduced peaks or none at all[13] (**Fig. 1b**). Near chromosome 12p11.1 is a ~76-kb region that contains over 400 consistently positioned nucleosomes independent of tissue type[10], and in this region we compared the plasma DNA read counts from female and male donors with those derived from sequencing of MNase-digested chromatin of the cell line GM12878 (a lymphoblastoid cell line (LCL) from a female donor) taken from the Encyclopedia of DNA Elements (ENCODE) Project (**Fig. 1c**). The plasma DNA read depth maps demonstrated wave-like patterns with peaks whose position showed high correlation with those found in the MNase maps (**Fig. 1c**).

Next, we compared read depth patterns at the TSSs of 3,804 housekeeping genes[15] with those at 670 genes unexpressed in all tissues (according to FANTOM5) in 104 control samples. The patterns

[1]Institute of Human Genetics, Medical University of Graz, Graz, Austria. [2]Institute of Molecular Biotechnology, Graz University of Technology, Graz, Austria. [3]BioTechMed OMICS Center Graz, Graz, Austria. [4]Institute of Pathology, Medical University of Graz, Graz, Austria. [5]Department of Obstetrics and Gynecology, Medical University of Graz, Graz, Austria. Correspondence should be addressed to M.R.S. (michael.speicher@medunigraz.at).

Received 31 March; accepted 22 July; published online 29 August 2016; doi:10.1038/ng.3648

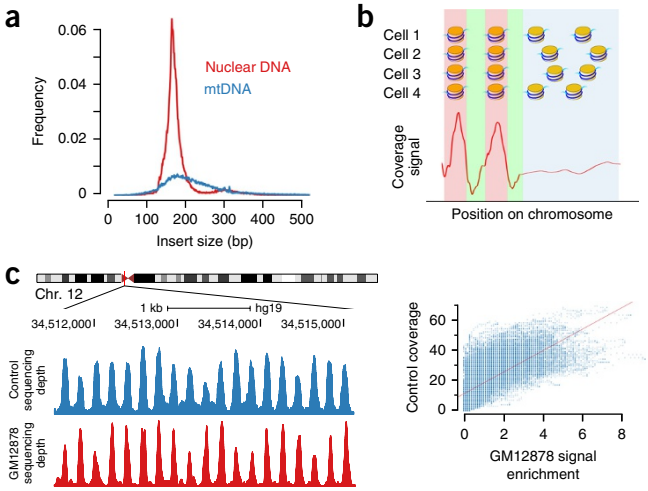

Figure 1 Plasma DNA fragment size and patterns of nucleosome positioning. (**a**) Size distribution calculated from paired-end sequencing data of nuclear (red; based on ~110,000 reads) and mitochondrial (blue; based on ~53,000 reads) plasma DNA fragments. (**b**) Schematic of the coverage patterns caused by nucleosomal protection of cfDNA. Nuclear chromatin digested by MNase is enriched for DNA protected by nucleosomes (dark blue) and depleted for connecting linker regions (light blue). In MNase assays, regions with perfectly positioned nucleosomes have strong coverage peaks reflecting the phasing of nucleosomes (left), whereas the pattern of read depth is different for regions with less preferentially positioned nucleosomes (right) (adapted from ref. 13). (**c**) Ideogram of chromosome 12 with enlargement of 12p11.1, which contains an extreme example of sequence-directed nucleosome positioning[10]. Read depth coverage for plasma DNA fragments from female and male donors ($n = 154$; merged data) is shown in blue, and the MNase midpoint density map from cell line GM12878 is shown in red. To the right is a comparison between plasma DNA read depth and the MNase midpoint density map, demonstrating a strong correlation (Pearson: 0.709, $P < 2.2 \times 10^{-16}$; Spearman: 0.708, $P < 2.2 \times 10^{-16}$).

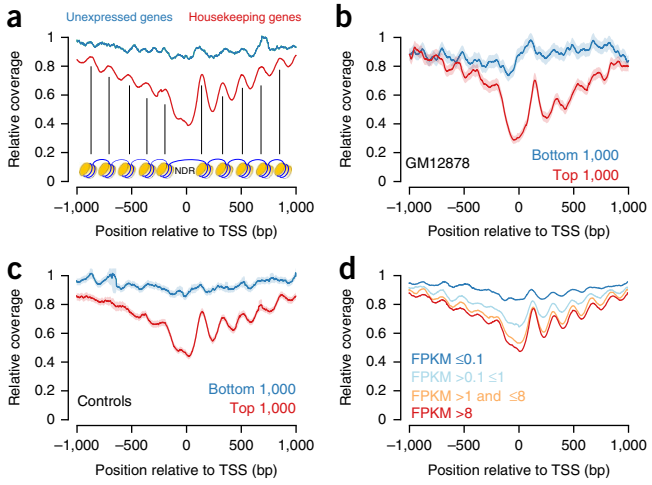

Figure 2 Nucleosome positioning at transcription start sites. (**a**) Sequencing coverage at promoter sites for housekeeping (red) and unexpressed (blue) genes generated with plasma samples from 104 donors. The coverage pattern reflects nucleosome organization. At the start of transcription, nucleosomes are removed to create an NDR over the promoter, allowing transcription factors to bind[11–13]. The reduction in nucleosome occupancy for expressed housekeeping genes corresponded to decreased coverage (*x* axis, distance from the TSS; *y* axis, relative coverage reflecting nucleosome dyads). (**b**) MNase midpoint density maps in the GM12878 cell line for the 1,000 (representing 1,334 TSSs) most highly expressed genes (red) and the 1,000 (representing 1,109 TSSs) least expressed genes (blue) in blood, identified on the basis of published plasma RNA-seq data[18]. Shaded areas represent 95% confidence intervals. (**c**) Plasma DNA read depth maps for the promoter regions of the genes in **b** (red, 1,000 most highly expressed genes; blue, 1,000 least expressed genes). (**d**) Plasma DNA read depth patterns at the promoters of differently expressed genes (red, FPKM >8; orange, FPKM >1 and ≤8; light blue, FPKM >0.1 and ≤1; dark blue, FPKM ≤0.1).

for housekeeping genes corresponded to those established by MNase assays[11–13], with depleted coverage at the TSS and oscillating periodicity upstream and downstream of the TSS (**Fig. 2a**). At promoters of inactive genes, by contrast, coverage increased, reflecting the denser nucleosome packaging of repressed genes[13] (**Fig. 2a**). We then wanted to test genes expressed in blood. Because the spacing of nucleosomes differs among cell types[13], we used MNase data from the GM12878 cell line for comparison with plasma read depth coverage from healthy donor samples, for which the vast majority (>90%) of DNA fragments are derived from white blood cells[16,17]. For the 1,000 (representing 1,334 TSSs) most highly and the 1,000 (1,109 TSSs) least expressed genes in blood[18], we observed similar coverage patterns for the publicly available GM12878 MNase data sets[19] and our own plasma DNA fragments (**Fig. 2b,c**). These plasma TSS sequence read depth maps differed depending on gene expression level (**Fig. 2d**).

To distinguish between expressed and silent genes on the basis of plasma read coverage characteristics, we conducted multiple tests that resulted in the identification of two discrete regions. The first region was based on the aforementioned reduction in nucleosome occupancy in the 2,000-bp region centered on the TSS[13] (2K-TSS coverage), which we had confirmed (**Fig. 2**). The second region was the most frequent position of the NDR, which we mapped to the region from −150 bp to +50 bp with respect to the TSS (NDR coverage) (**Supplementary Fig. 1**). Read depth coverage for both regions was normalized by relative copy number so that copy number alterations (CNAs), which are frequently observed in plasma samples from patients with cancer[20], did not affect the evaluation.

Kernel density estimation of the read depth coverage of these two regions for the 1,000 most highly and least expressed genes resulted in two separate clusters (**Fig. 3a**). To test whether these two clusters correspond to differently expressed gene sets, we classified the genes by employing support vector machines (SVMs), which allowed us to predict the expression status of the 100 most highly and least expressed genes[18] with a sensitivity and an accuracy of 0.91 each (**Fig. 3b**). Even for the 1,000 and 5,000 most highly and least expressed genes, the sensitivity was still 0.81 and 0.78 and the accuracy was 0.83 and 0.76, respectively (**Fig. 3b**, **Supplementary Fig. 2**, **Supplementary Tables 1** and **2**, and **Supplementary Note**). Accordingly, the genes in the two clusters were statistically significantly differentially expressed (Mann–Whitney *U* test, $P < 2.2 \times 10^{-16}$) (**Fig. 3c**). In contrast, when we performed training runs with random assignments of genes as expressed or unexpressed, the accuracy and sensitivity of the classifier both dropped to ~0.5 (**Supplementary Note**). Furthermore, reliable prediction of expressed genes was still possible at 5% of the original sequencing depth, as determined by subsampling data from the 104 merged controls (**Supplementary Table 3** and **Supplementary Note**). As examples, we show the different promoter coverage patterns for the genes *NCL* and *GABRR3* (**Fig. 3d**). Additionally, we conducted tenfold cross-validation, which yielded a total accuracy of 85.8%. We ranked genes on the basis of their FPKM values relative to their average 2K-TSS and NDR coverage (**Fig. 3e** and **Supplementary Fig. 3**) and confirmed the quantitative relationship already suggested

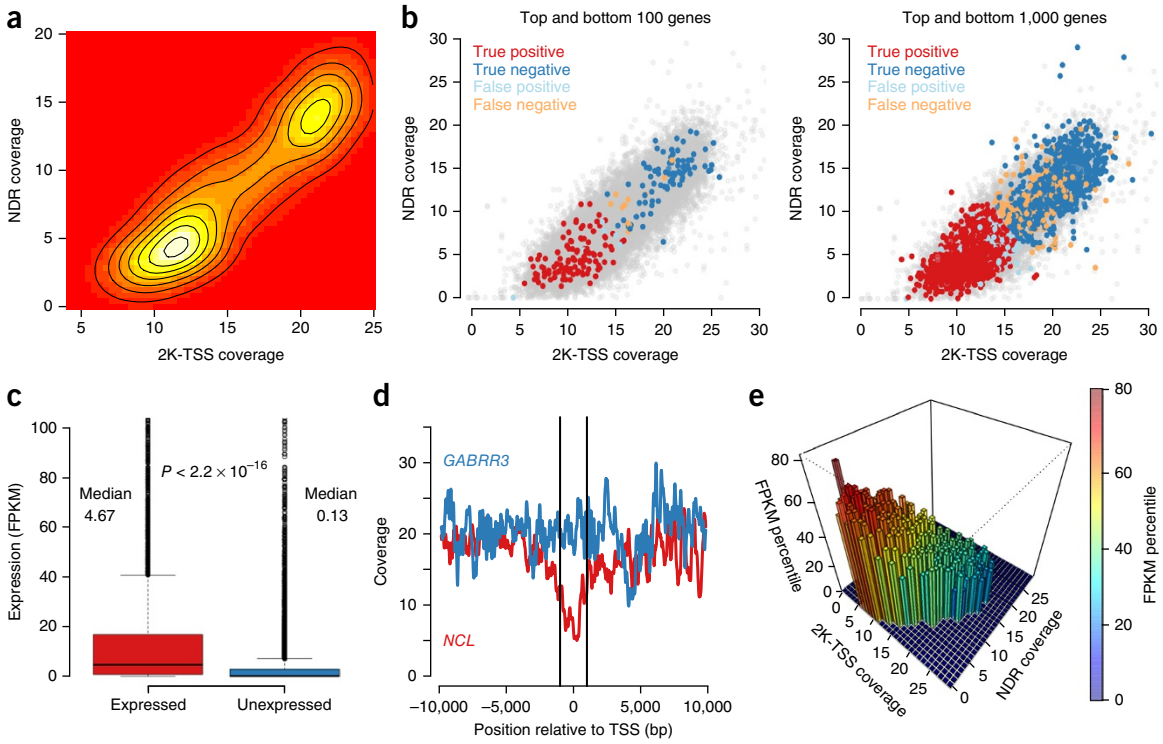

Figure 3 Classification of expressed and silent genes by plasma DNA read depth analyses. (**a**) Kernel density estimation identified two separate gene clusters based on normalized coverage patterns at 2K-TSS and NDR regions. (**b**) SVM classification based on normalized 2K-TSS and NDR coverage for the 100 (left) and 1,000 (right) most highly and least expressed genes. Red and dark blue circles represent genes correctly predicted to be expressed and unexpressed, respectively, whereas light blue and orange circles represent incorrectly predicted genes. (**c**) Box plots showing that the difference in FPKM values between genes predicted to be expressed (n = 11,345) and unexpressed (n = 9,156) is statistically highly significant (expressed: median = 4.67, s.d. = 675.3; unexpressed: median = 0.13, s.d. = 97.0; Mann–Whitney U test, two-sided, $P < 2.2 \times 10^{-16}$). Each box comprises data from the first to the third quartile (interquartile range, IQR) and the median. Whiskers extend to 1.5 × IQR from the box. (**d**) Example promoter coverage patterns for the *NCL* (red) and *GABRR3* (blue) genes, which are expressed with mean FPKM values of 2,000 and <0.5, respectively. The vertical bars delimit the regions from TSS – 1,000 bp to TSS + 1,000 bp. (**e**) Averaging FPKM percentiles within integer bins of 2K-TSS and NDR coverage showed a quantitative relationship between these two coverage parameters and gene expression (represented by percentile of FPKM values; for details, see the **Supplementary Note**).

by the TSS sequence read depth maps for different gene expression levels (**Fig. 2d** and **Supplementary Note**).

We then investigated whether plasma DNA from patients with cancer would allow us to draw conclusions regarding the expression of genes in their primary tumor. Because of the inevitable heterogeneity of these plasma samples (corresponding to mixtures of DNA released from tumor and hematopoietic cells in various proportions), we conducted *in silico* dilution simulations to establish resolution limits, which showed that, for this application, ≥75% of all DNA fragments for a given TSS must be released by tumor cells to infer expression status (**Fig. 4a**). For our proof-of-concept studies, we analyzed matching and synchronously obtained primary tumors from two metastasized breast cancer cases (B7 and B13) in addition to performing whole-genome sequencing of plasma DNA (**Supplementary Fig. 4**) and RNA-seq analysis (**Fig. 4b**). We sequenced the plasma DNA with high coverage (B7, ~8.2×; B13, ~9.1×) and calculated CNAs[21–23] (**Fig. 4c**). Nucleosome array (**Supplementary Fig. 5**) and promoter read depth (**Supplementary Fig. 6**) differences between unexpressed and housekeeping genes could again be established for the chromosome 12p11.1 region in these samples. We estimated tumor purity directly from observed relative copy number profiles[24] and found overall ctDNA allele frequencies of ~45% and ~72% in B7 and in B13, respectively (**Fig. 4d**), which are ctDNA allele frequencies common in metastatic breast cancer[25]. However, the actual ctDNA allele frequency for a specific region additionally depends on its copy number, as amplified regions are relatively enriched in ctDNA. Therefore, we calculated regional ctDNA allele frequencies on the basis of overall ctDNA allele frequency and the $\log_2$-transformed copy number ratio of the respective region (**Fig. 4e**). This analysis suggested that accurate gene expression predictions should at least be possible for chromosome 1q and the amplified regions on chromosomes 11q, 16p, and 19p in B7, whereas all over-represented regions should be suitable in B13.

We identified focal amplifications frequent in breast cancer as defined previously[26], such as amplifications of 11q13.3 (15 genes, including *CCND1*) in B7 or of 8p11 (31 genes, including *FGFR1*) and 17q12 (46 genes, including *ERBB2*) in B13 (**Fig. 4c**). We compared the FPKM values of each gene predicted to be expressed with those of genes predicted to be not expressed in these amplicons for B7 and B13 and observed statistically highly significant differences (**Fig. 5a**). When we extended these analyses to all amplicons, FPKM values were significantly different for expressed and unexpressed genes (Mann–Whitney U test, two-sided: B7, 3.79×10^{-6}; B13, 1.53×10^{-6}; **Supplementary Tables 4** and **5**, and **Supplementary Note**). We then analyzed the 100 most highly expressed genes, as determined by RNA-seq analysis of the primary tumor, on chromosome 1q in B7 and on chromosome 8p11-qter in B13 and found that 86.1% and 88.1%, respectively, were correctly classified in the expressed cluster (**Fig. 5b**). When we extended these analyses to the 100 most highly expressed genes for all gained regions in B13 (regions with

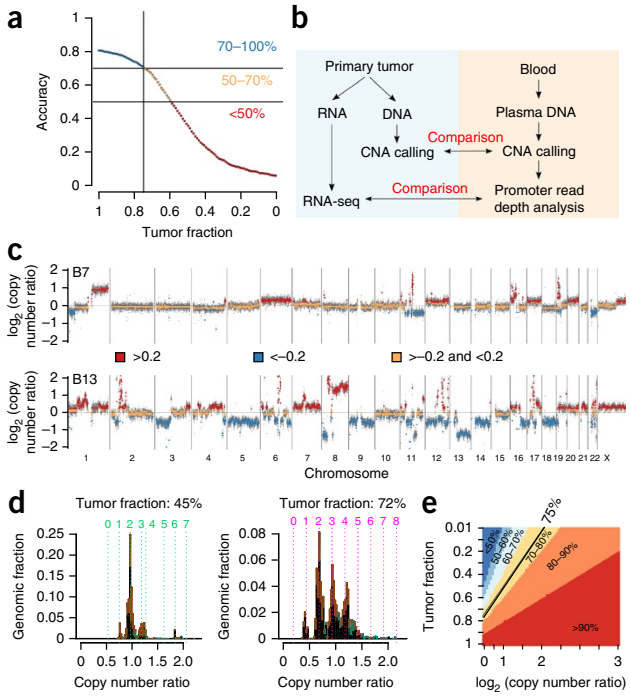

Figure 4 Procedure for predicting expressed genes in cancer from blood. (**a**) Simulation of resolution limits with *in silico* dilution employing 2K-TSS and NDR coverage mixing the 1,000 most highly expressed genes with random parameters from the distribution of the 1,000 least expressed genes in plasma (blue, accuracy of >70%; orange, accuracy of 50–70%; red, accuracy of <50%). (**b**) Workflow for the identification of expressed cancer driver genes in peripheral blood. Matching primary tumor tissue was synchronously obtained with blood samples. CNAs from both the primary tumor and plasma DNA were established for comparison. Expression patterns in the primary tumor were analyzed by RNA-seq and correlated with plasma DNA promoter coverage in relation to the respective copy number status. (**c**) Copy number profiles of two patients with breast cancer (B7 and B13) from plasma DNA. The *x* axis shows the chromosomes; the *y* axis shows $\log_2$-transformed copy number ratios. (**d**) Estimation of tumor purity and ploidy by the quantitative ABSOLUTE method, which estimates tumor purity and ploidy directly from observed relative copy number profiles[24], for B7 (left) and B13 (right). (**e**) Heat map illustrating how regional ctDNA allele frequencies are established in relation to overall ctDNA allele frequency (*y* axis) and copy number ($\log_2$-transformed ratio) (*x* axis). The black line represents an allele frequency of 75%, which was deemed suitable for gene expression prediction. Regions colored in different shades of red and yellow show the varying levels of ctDNA allele frequency.

Table 1 Identification of the most highly expressed isoforms of eight genes with more than one TSS in amplicons at 11q13.3 (B7) and 8p11 and 17q12 (B13)

Sample	Gene	TSS position	Isoforms	FPKM	Distance
B13	*DDHD2*	chr8:38,089,008	2	4.85	0.21
		chr8:38,089,470	1	0.01	0.09
	GRB7	chr17:37,894,161	1	2.95	1.26
		chr17:37,894,575	1	1.59	0.91
		chr17:37,895,023	1	1.33	0.84
		chr17:37,896,219	1	0.14	0.63
	PPP1R1B	chr17:37,784,750	2	9.24	0.69
		chr17:37,783,176	1	3.77	0.89
	GSDMB	chr17:38,074,903	3	4.27	0.58
		chr17:38,073,793	1	<0.01	0.30
	ANK1	chr8:41,522,804	3	4.93	0.44
		chr8:41,655,140	4	0.35	0.11
		chr8:41,754,280	1	0.01	0.17
	ERBB2	chr17:37,856,230	1	11.37	1.07
		chr17:37,844,336	1	4.43	0.58
	FGFR1	chr8:38,325,363	3	6.53	0.62
		chr8:38,326,352	6	3.02	0.36
B7	*ANO1*	chr11:69,924,407	1	2.06	0.24
		chr11:69,931,515	1	0.06	0.06

$\log_2$ (copy number ratio) >0.2; corresponding to ~1 Gb), 78.0% were correctly classified.

To provide more detailed examples for single genes, we sought to determine which isoforms were expressed in B13 for two highly relevant cancer-related genes—*ERBB2*, which is an important biomarker for treatment decisions involving the monoclonal antibody trastuzumab[27], and *FGFR1*, a potential target for fibroblast growth factor receptor (FGFR) inhibitors, which are currently in development[28]. *ERBB2* had promoter coverage corresponding to an expressed gene (**Fig. 5c**). For its two isoforms (NM_004448 and NM_001005862), we calculated the differences in the distance between 2K-TSS and NDR coverage in blood from patients with cancer and healthy controls. This calculation predicted that NM_004448 was the highly expressed isoform in the primary tumor (**Fig. 5d**,**e**), which was indeed confirmed by RNA-seq (NM_004448, 11.4 FPKM; NM_001005862, 4.4 FPKM). Using the same approach, we analyzed *FGFR1*, which has two TSSs but nine isoforms, and demonstrated that expression from TSS2 should be higher than that from TSS1, which again correlated with the RNA-seq data (TSS1, sum FPKM of 3.0; TSS2, sum FPKM of 6.5) (**Table 1**). This prompted us to analyze every gene in focal amplifications with several TSSs. Of these 93 genes, 8 had more than one TSS that gave rise to isoforms with FPKM differences of at least 2 (including *ERBB2* and *FGFR1*). We were able to correctly identify the most highly expressed isoform for seven of these (**Table 1**).

As our evaluations depend on ctDNA allele frequency and the $\log_2$-transformed copy number ratios of respective regions, we wanted to test whether this approach is broadly applicable. To this end, we analyzed 426 plasma samples from patients with metastasized cancer (colon, 128; prostate, 139; breast, 125; lung, 31; other tumor entities, 3) and calculated allele frequencies, which were within the range reported in metastatic disease[29]. We found that 220 (51.6%) of these samples had at least 100 genomic bins (>5.6 Mb) suitable for promoter read depth analysis. Certain regions such as high-level amplifications will almost always be amenable to our analyses, which is important as these regions frequently contain cancer driver genes[26,30–32]. However, our approach will likely not be applicable in minimal residual disease. Another hampering factor is that nucleosome-deprived states may also occur in paused genes, as genes with elongating or poised RNA polymerase II exhibited a similar pattern of nucleosome phasing as expressed genes[11,33]. Moreover, transcription is not always associated with chromatin reorganization[12].

Recently, nucleosome spacing was used to determine tissues of origin for cfDNA[9]. Using plasma samples from five individuals with cancer, the authors of this study found correlations to the correct non-hematopoietic cell sources in three of the five tested cases on the basis of nucleosome footprints. Another study used plasma DNA whole-exome sequencing data to investigate associations between gene expression and nucleosome fragmentation patterns[34]. However, as whole-exome sequencing data do not cover the region upstream of the TSS, including the NDR, predictions were only possible for a small number of genes. In contrast, we leveraged whole-genome sequencing data to include entire promoter regions for the establishment of gene

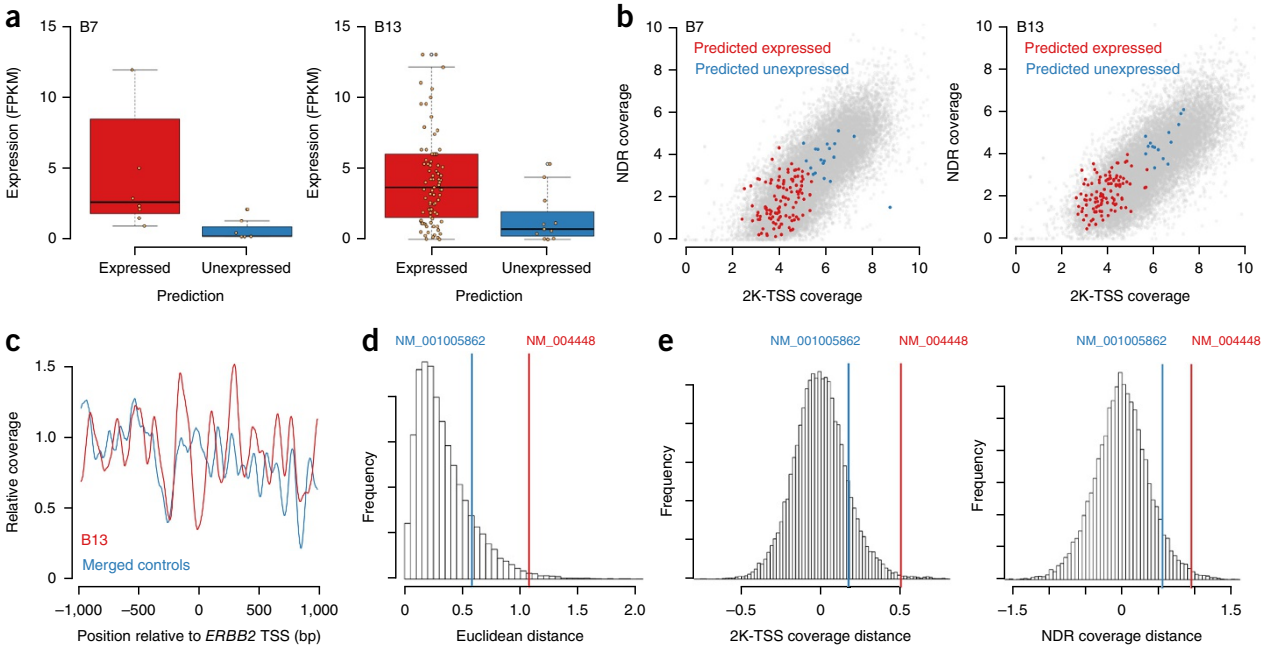

Figure 5 Identification of expressed genes in cancer from the peripheral blood. (**a**) Box plots showing FPKM values for genes predicted to be expressed or not expressed in focal amplifications of 11q13.3 (15 TSSs in 15 genes including *CCND1*; $n_{expressed} = 8$ and $n_{unexpressed} = 7$) in B7 (left) and in both 8p11 (39 TSSs in 31 genes including *FGFR1*) and 17q12 (59 TSSs in 46 genes including *ERBB2*) ($n_{expressed} = 87$ and $n_{unexpressed} = 11$) in B13 (right). Blue dots represent genes located in amplicons. Outliers, including *CCND1* (FPKM of 50 in B7) and *ERBB2* (FPKM of 15 in B13), are not shown because of scaling. The differences were statistically highly significant (B7: expressed: mean = 9.7, s.d. = 17.0; unexpressed: mean = 0.7, s.d. = 0.8; Mann–Whitney *U* test, $P = 0.003$; B13: expressed: mean = 5.7, s.d. = 9.7; unexpressed: mean = 1.5, s.d. = 1.8; $P = 0.001$). (**b**) Classification accuracy for the 100 most highly expressed genes on chromosomes 1q in B7 and 8p11-qter in B13 as assessed by RNA-seq of the respective primary tumor tissue. (**c**) Different coverage of *ERBB2* (mean for both isoforms) in B13 and control samples around the TSS. (**d**) *ERBB2* has two isoforms (NM_001005862 and NM_004448). Calculation of the differences in Euclidean distance for 2K-TSS and NDR coverage between blood from patients with cancer and healthy controls established that isoform NM_004448 was highly expressed in the tumor from B13. (**e**) The distances for 2K-TSS (left) and NDR (right) coverage separately confirm that isoform NM_004448 was highly expressed in the tumor from B13.

expression status. Our study provides a new view on the genomes of cells that release their DNA into the circulation and hence expands upon currently existing options for cfDNA analyses.

URLs. FANTOM5 (Functional Annotation of the Mammalian Genome), http://fantom.gsc.riken.jp/5/; European Genome-phenome Archive (EGA), http://www.ebi.ac.uk/ega/; Picard tools, http://broadinstitute.github.io/picard.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Raw sequencing reads, aligned and trimmed reads from high-coverage sequencing of plasma DNA (merged healthy controls and tumor samples) as well as raw reads from RNA-seq experiments have been deposited in the European Genome-phenome Archive (EGA) under accession EGAS00001001754.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to S. Perakis for critical reading and editing of this manuscript. This work was supported by CANCER-ID, a project funded by the Innovative Medicines Joint Undertaking (IMI JU).

AUTHOR CONTRIBUTIONS

P.U. and M.R.S. designed the study. M.A. and R.G. performed the experiments. P.U., G.G.T., J.B.G., E.H., and M.R.S. analyzed data. E.P. and G.P. provided clinical samples and clinical information. S.W.J. and L.A. performed pathology analyses. K.K. conducted RNA-seq. PU., E.H., and M.R.S. supervised the study. P.U., J.B.G., E.H., and M.R.S. wrote the manuscript. All authors revised the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Schwarzenbach, H., Hoon, D.S. & Pantel, K. Cell-free nucleic acids as biomarkers in cancer patients. *Nat. Rev. Cancer* **11**, 426–437 (2011).
2. Heitzer, E., Auer, M., Ulz, P., Geigl, J.B. & Speicher, M.R. Circulating tumor cells and DNA as liquid biopsies. *Genome Med.* **5**, 73 (2013).
3. Crowley, E., Di Nicolantonio, F., Loupakis, F. & Bardelli, A. Liquid biopsy: monitoring cancer-genetics in the blood. *Nat. Rev. Clin. Oncol.* **10**, 472–484 (2013).
4. Diaz, L.A. Jr. & Bardelli, A. Liquid biopsies: genotyping circulating tumor DNA. *J. Clin. Oncol.* **32**, 579–586 (2014).
5. Heitzer, E., Ulz, P. & Geigl, J.B. Circulating tumor DNA as a liquid biopsy for cancer. *Clin. Chem.* **61**, 112–123 (2015).
6. Diehl, F. *et al.* Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc. Natl. Acad. Sci. USA* **102**, 16368–16373 (2005).
7. Lo, Y.M. *et al.* Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci. Transl. Med.* **2**, 61ra91 (2010).
8. Ramachandran, S. & Henikoff, S. Replicating nucleosomes. *Sci. Adv.* **1**, e1500587 (2015).
9. Snyder, M.W., Kircher, M., Hill, A.J., Daza, R.M. & Shendure, J. Cell-free DNA comprises an *in vivo* nucleosome footprint that informs its tissues-of-origin. *Cell* **164**, 57–68 (2016).
10. Gaffney, D.J. *et al.* Controls of nucleosome positioning in the human genome. *PLoS Genet.* **8**, e1003036 (2012).
11. Schones, D.E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887–898 (2008).
12. Venkatesh, S. & Workman, J.L. Histone exchange, chromatin structure and the regulation of transcription. *Nat. Rev. Mol. Cell Biol.* **16**, 178–189 (2015).

13. Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474**, 516–520 (2011).
14. Chandrananda, D., Thorne, N.P. & Bahlo, M. High-resolution characterization of sequence signatures due to non-random cleavage of cell-free DNA. *BMC Med. Genomics* **8**, 29 (2015).
15. Eisenberg, E. & Levanon, E.Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).
16. Lui, Y.Y. *et al.* Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. *Clin. Chem.* **48**, 421–427 (2002).
17. Sun, K. *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc. Natl. Acad. Sci. USA* **112**, E5503–E5512 (2015).
18. Koh, W. *et al.* Noninvasive *in vivo* monitoring of tissue-specific global gene expression in humans. *Proc. Natl. Acad. Sci. USA* **111**, 7361–7366 (2014).
19. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
20. Heitzer, E., Ulz, P., Geigl, J.B. & Speicher, M.R. Non-invasive detection of genome-wide somatic copy number alterations by liquid biopsies. *Mol. Oncol.* **10**, 494–502 (2016).
21. Heidary, M. *et al.* The dynamic range of circulating tumor DNA in metastatic breast cancer. *Breast Cancer Res.* **16**, 421 (2014).
22. 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).
23. Mohan, S. *et al.* Changes in colorectal carcinoma genomes under anti-EGFR therapy identified by whole-genome plasma DNA sequencing. *PLoS Genet.* **10**, e1004271 (2014).
24. Carter, S.L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
25. Murtaza, M. *et al.* Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* **497**, 108–112 (2013).
26. Ulz, P., Heitzer, E. & Speicher, M.R. Co-occurrence of *MYC* amplification and *TP53* mutations in human cancer. *Nat. Genet.* **48**, 104–106 (2016).
27. Giordano, S.H. *et al.* Systemic therapy for patients with advanced human epidermal growth factor receptor 2–positive breast cancer: American Society of Clinical Oncology clinical practice guideline. *J. Clin. Oncol.* **32**, 2078–2099 (2014).
28. Helsten, T. *et al.* The FGFR landscape in cancer: analysis of 4,853 tumors by next-generation sequencing. *Clin. Cancer Res.* **22**, 259–267 (2016).
29. Bettegowda, C. *et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.* **6**, 224ra24 (2014).
30. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
31. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
32. Ulz, P. *et al.* Whole-genome plasma sequencing reveals focal amplifications as a driving force in metastatic prostate cancer. *Nat. Commun.* **7**, 12008 (2016).
33. Adelman, K. & Lis, J.T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.* **13**, 720–731 (2012).
34. Ivanov, M., Baranova, A., Butler, T., Spellman, P. & Mileyko, V. Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC Genomics* **16** (Suppl. 13), S1 (2015).

ONLINE METHODS

Patients. The study was approved by the Ethics Committee of the Medical University of Graz (approval numbers 21-227 ex 09/10 and 21-228 ex 09/10) and was conducted according to the Declaration of Helsinki; written informed consent was obtained from all patients.

Samples. Single-end sequencing data (trimmed to 60 bp) for 104 plasma samples from donors without tumors (male, $n = 50$; female, $n = 54$) were merged and used to generate coverage profiles around the TSS and to establish a gene expression prediction algorithm. Single-end sequencing data for two breast cancer cases (B7 and B13; also trimmed to 60 bp) were used to explore the applicability of expression prediction for cancer-related genes from peripheral blood. To assess differences in fragment size between nuclear and mitochondrial DNA, paired-end sequencing from 179 plasma DNA samples, including an independent set of 20 controls without cancer and 159 patients with cancer, was used. Finally, single-end sequencing data from an additional 426 cancer samples (colon, $n = 128$; prostate, $n = 139$; breast, $n = 125$; lung, $n = 31$) from ongoing studies were used to explore the general applicability of the method by testing tumor allele frequencies.

Plasma DNA preparation. Plasma DNA was prepared using the QIAamp DNA Blood Mini kit (Qiagen) as previously described[22]. Samples selected for sequencing library construction were analyzed on a Bioanalyzer instrument (Agilent Technologies) to observe the plasma DNA size distribution.

Sequencing. Shotgun libraries of plasma DNA and tumor DNA were prepared using the TruSeq DNA Nano library preparation kit by Illumina with a starting amount of 5–10 ng according to the protocol. However, because of the low DNA input, we increased the number of PCR cycles to 25. Furthermore, the fragmentation step was omitted because of the degradation of plasma DNA. Libraries were sequenced on Illumina MiSeq and NextSeq sequencers. All raw sequencing data were deposited in the European Genome-phenome Archive (EGA), which is hosted by the European Bioinformatics Institute.

Insert size calculations from paired-end sequencing. Paired-end reads from 179 plasma DNA samples were aligned with bwa[35] backtrack to the hg19 human reference genome. Resulting BAM files were merged using SAMtools[36], and alignments to the mitochondrial genome were extracted. Alignments to the nuclear genome were downsampled, and insert sizes for both BAM files were analyzed using the Picard InsertSizeMetrics function.

Single-end sequencing data preparation. Raw reads (150 bp) from the 104 control samples and the 2 breast cancer samples were trimmed from both ends to contain bases 53 to 113. These 60 bp should constitute the central 60 bp of a typical 166-bp cfDNA fragment and should thus be exclusively associated with a nucleosome. Reads were then aligned to the hg19 human genome using bwa-mem (version 0.7.4)[35], and PCR duplicates were removed using the SAMtools rmdup[36] function (version 0.1.18). Aligned BAM files for controls were merged using the SAMtools merge function.

CNA analysis. Raw reads for the two breast cancer samples and controls were aligned to the hg19 human reference genome using bwa[35] with the pseudoautosomal region of the Y chromosome masked. PCR duplicates were removed, and reads were counted in 50,000 genomic bins, each containing the same amount of mappable positions (approximately 56 kb). Raw read counts were normalized by median bin count, and correction for GC content was performed using Lowess smoothing. Furthermore, corrected read counts were normalized by mean bin counts for non-cancer controls and segmented using both the CBS and GLAD tools provided by the CGHweb framework[37].

Tumor fraction estimation. Tumor fraction for the two breast cancer samples was estimated by applying ABSOLUTE[24] to the segmented $\log_2$-transformed copy number ratios obtained in the CNA analysis. We used the most plausible karyotype model and extracted purity values.

Estimation of relative tumor fraction. To estimate the relative tumor fraction of a region depending on the copy number state, we derived the following. The undiluted copy number (cp_i) of a certain region i depends on the $\log_2$-transformed copy number ratio (lr_i) as measured in the CNA analysis step as well as the tumor fraction (tf)

$$cp_i = \frac{2 \times 2^{lr_i} - 2(1 - tf)}{tf}$$

The relative tumor fraction (rtf_i) for this region can then be computed using the (pure) copy number and the tumor fraction again.

$$rtf_i = \frac{tf \times cp_i}{tf \times cp_i + (1 - tf) \times 2}$$

Plasma RNA analysis and RNA-seq. Gene expression values from plasma RNA analyses with microarrays were provided by Koh *et al.*[18]. RMA values from four healthy (non-pregnant) subjects were averaged, and the 1,000 most highly expressed genes (Top1000) and the 1,000 least expressed genes (Bottom1000) were extracted. The fastq files with raw data from the RNA-seq step were downloaded for the four non-pregnant samples (Sequence Read Archive (SRA) accessions SRR1296080, SRR1296081, SRR1296082, and SRR1296083).

RNA-seq expression values were computed from tumor samples B7 and B13 and the aforementioned plasma RNA-seq data[18]. Briefly, we aligned RNA-seq reads to the hg19 human reference genome using TopHat2 (v2.0.7) and calculated the gene-wide FPKM value using Cufflinks 2 (ref. 38). Subsequently, FPKM values were averaged for each gene.

TSS profiles. Coverage values around TSSs were extracted from aligned BAM files using the SAMtools depth[36] function, and every value was normalized by the mean value of the combined regions: TSS – 3,000 to TSS – 1,000 and TSS + 1,000 and TSS + 3,000.

Copy number–normalized parameter extraction. Two parameters were used for the identification and prediction of genes as expressed or unexpressed:

1) The coverage from −1,000 bp to +1,000 bp with respect to the TSS (2K-TSS coverage)

2) The coverage from −150 bp to +50 bp with respect to the TSS (NDR coverage)

For every TSS in RefSeq, parameters were extracted and divided by the relative copy number of that region as determined in the CNA analysis step.

Prediction by SVMs. To predict the expression status of individual genes, we used SVMs. As a training set for expressed genes, we used a random subset of 300 housekeeping genes out of 3,804 housekeeping genes that are expressed uniformly in multiple tissues[15], and for unexpressed genes we used a random subset of 300 genes out of 670 reported to be unexpressed in most tissues by the FANTOM5 project. Random subset selection and prediction were repeated 1,000 times, and the prediction status for each TSS that was not part of the training set was recorded. Tenfold cross-validation indicated a mean accuracy of 85.79%. To maximize performance, we considered a gene to be expressed when the prediction consent of all the iterations was higher than 75%, which resulted in a modest improvement (**Supplementary Fig. 7** and **Supplementary Note**). As an independent test set, we then used the Top1000 and Bottom1000 gene sets.

Quantitative analysis. To test whether the 2K-TSS and NDR coverage parameters contained quantitative information about gene expression, we annotated every TSS from the merged controls with the FPKM values of the respective genes from the aforementioned plasma RNA-seq experiments[18]. FPKM values were ranked, and percentiles were calculated. Subsequently, data from the two parameters were binned, and the average percentile of every (integer) bin was calculated. Bins containing ten or fewer TSSs were discarded. For more details, see the **Supplementary Note**.

***In silico* dilution simulations.** We performed dilution simulations to test the reliability of prediction at varying tumor fractions. To this end, we modeled the distribution of the 2K-TSS and NDR coverage parameters for the 1,000 least expressed genes in plasma and added random numbers from

these distributions to the parameters for the Top1000 expressed genes at varying proportions.

Isoform discrimination. To check whether more abundant isoforms (transcripts) can be distinguished from less abundant isoforms of the same gene, we compared coverage data for tumors and merged controls. We first normalized both parameters (2K-TSS and NDR coverage) for the merged cfDNA controls to account for varying read depth between controls and tumor sequencing data. Subsequently, we normalized both parameters (2K-TSS and NDR coverage) for the tumor samples, and for every TSS we calculated a distance for both coverage parameters from the respective TSS in the normalized control data and the tumor data. The distances from the 2K-TSS and NDR coverage parameters were then combined using Euclidean distance metrics. Isoforms with higher expression should have decreased 2K-TSS and NDR coverage, as DNA should be less protected by nucleosomes for more abundant transcripts.

Code availability. Relevant code is available at https://github.com/PeterUlz/Nucleosome_ctDNA.

35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
36. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
37. Lai, W., Choudhary, V. & Park, P.J. CGHweb: a tool for comparing DNA copy number segmentations from multiple algorithms. *Bioinformatics* **24**, 1014–1015 (2008).
38. Trapnell, C. *et al*. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).

doi:10.1038/ng.3648