Identification of tissue-specific cell death using methylation patterns of circulating DNA

Roni Lehmann-Werman1,3,1, Daniel Neiman1,3, Hai Zemmour1, Joshua Moss1,3, Judith Magenheim1, Adi Vaknin-Dembinsky1,2, Sten Rubertsson2, Bengt Nellgård2,3, Kai Blennow2, Henrik Zetterberg2,3, Kirsty Spalding2,3, Michael J. Haller2,3, Clive H. Wasserfall2,3, Desmond A. Schatz4, Carla J. Greenbaum5, Craig Dorrell5,6, Markus Grompe5,6, as providing a better understanding of normal tissue dynamics.

Circulating DNA (cfDNA) released from dying cells is emerging as a diagnostic tool for monitoring cancer dynamics and graft failure. However, existing methods rely on differences in DNA sequences in source tissues, so that cell death cannot be identified in tissues with a normal genome. We developed a method of detecting tissue-specific cell death in humans based on tissue-specific methylation patterns in cfDNA. We interrogated tissue-specific methylene databases to identify cell type-specific DNA methylation signatures and developed a method for detecting these signatures in mixed DNA samples. We isolated cfDNA from plasma or serum of donors, treated the cfDNA with bisulfite, PCR-amplified the cfDNA, and sequenced it to quantify cfDNA carrying methylation markers of the cell type of interest. Pancreatic β-cell DNA was identified in the circulation of patients with recently diagnosed type-1 diabetes and islet-graft recipients; oligodendrocyte DNA was identified in patients with relapsing multiple sclerosis; neuronal/glial DNA was identified in patients after traumatic brain injury or cardiac arrest; and exocrine pancreas DNA was identified in patients with pancreatic cancer or pancreatitis. This proof-of-concept study demonstrates that the tissue origins of cfDNA and thus the rate of death of specific cell types can be determined in humans. The approach can be adapted to identify cfDNA derived from any cell type in the body, offering a minimally invasive window for diagnosing and monitoring a broad spectrum of human pathologies as well as providing a better understanding of normal tissue dynamics.

Significance

We describe a blood test for detection of cell death in specific tissues based on two principles: (i) dying cells release fragmented DNA to the circulation, and (ii) each cell type has a unique DNA methylation pattern. We have identified tissue-specific DNA methylation markers and developed a method for sensitive detection of these markers in plasma or serum. We demonstrate the utility of the method for identification of pancreatic β-cell death in type 1 diabetes, oligodendrocyte death in relapsing multiple sclerosis, brain cell death in patients after traumatic or ischemic brain damage, and exocrine pancreas cell death in pancreatic cancer or pancreatitis. The approach allows minimally invasive monitoring of tissue dynamics in humans in multiple physiological and pathological conditions.


Conflict of interest statement: C.D. and M.G. are the inventors of antibodies directed against human pancreatic cells, HPx1/HC0-3B3 and HPd3/DHIC5-4D9. Oregon Health & Science University (OHSU) has commercially licensed this technology. This potential conflict of interest has been reviewed and managed by OHSU.

This article is a PNAS Direct Submission.

1R.L.-W., D.N., H. Zemmour, and J. Moss contributed equally to this work.
2To whom correspondence may be addressed. Email: beng@cc.huji.ac.il, shemer.rut@mail.huji.ac.il, or yuvalld@ekmd.huji.ac.il.

This article contains information available online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1519286113/-/DCSupplemental.
kidney, liver, or heart transplantation, based on SNPs that distinguish the DNA of the donor from that of the recipient (9–11). In all these cases, genetic differences exist between the DNA sequence of the tissue of interest (fetus, tumor, or graft) and that of the host, providing the basis for highly specific assays.

Blood levels of cfDNA increase in many conditions, such as TBI (12), cardiovascular disease (13), sepsis (14), and intensive exercise (15). However, in these cases, the cfDNA sequence is identical to that of genomic DNA, making it impossible to use sequence variation to identify the source of the elevated cfDNA and thus greatly compromising the utility of cfDNA as a diagnostic or prognostic tool. For example, the cfDNA could originate from parenchymal cells of the injured tissue but also from dying inflammatory cells.

Despite having an identical nucleotide sequence, the DNA of each cell type in the body carries unique methylation marks correlating with its gene-expression profile, and these marks represent a fundamental aspect of tissue identity. Methylation patterns are unique to each cell type, are conserved among cells of the same type in the same individual and among individuals, and are highly stable under physiologic or pathologic conditions (16). Therefore, it is possible to use the DNA methylation pattern of cfDNA to determine its tissue of origin and hence to infer cell death in the source organ.

DNA methylation in cfDNA has been studied extensively but usually in the context of epigenetic aberrations unique to the specific pathology, such as abnormal methylation in promoters of tumor-suppressor genes, which may lead to insights into cancer biology (17–20). By contrast, our approach seeks to identify, in plasma or serum, the normal, stable methylation signature of a specific tissue as a sensitive biomarker of cell death. Therefore the approach can be applied to pathologies that retain a stable genome and epigenome.

A classic example of tissue-specific DNA methylation is provided by the insulin gene (INS) promoter, which is unmethylated in insulin-producing pancreatic β cells and is methylated elsewhere. Recent studies have identified unmethylated INS promoter DNA in the circulation of patients with recently diagnosed T1D and in islet-graft recipients, likely reflecting both autoimmune and alloimmune destruction of β cells (21–25). However, published data suggest that the analytic approaches used were not sufficiently specific to differentiate robustly between β-cell- and non-β-cell-derived DNA. Here, we present a method and an analysis that demonstrate the approach toward proof of the concept that tissue-specific methylation patterns in cfDNA can be used to detect tissue cell death with a high level of specificity and sensitivity in multiple human pathologies. Here we demonstrate the detection of cell death in different pathologies of the pancreas and the brain.

Results

Identification of Tissue-Specific Methylation Markers. We started by identifying tissue-specific DNA methylation markers distinguishing individual tissues or cell types from other tissues. Particular attention was given to markers that differ between the tissue of interest and hematopoietic cells, which contribute the majority of cfDNA in healthy individuals. We analyzed publically available (The Cancer Genome Atlas and Gene Expression Omnibus) and locally generated methylomes to identify individual CpG dinucleotides with differential methylation patterns, i.e., unmethylated in the tissue of interest but methylated elsewhere (SI Materials and Methods and the schematic of the procedure in Fig. S1).

The Illumina Infinium HumanMethylation450 BeadChip array provides information on the methylation status of individual CpG dinucleotides. The discriminatory power of any single CpG is limited, because it can be randomly methylated or unmethylated in a small fraction of molecules from tissues where it typically is unmethylated or methylated, respectively. To increase the signal-to-noise ratio of the assay, we exploited the regional nature of DNA methylation. We defined an “expanded window” of four to nine CpG sites adjacent to the original CpG marker site, reasoning that accidental methylation or demethylation of multiple adjacent cytosines in the same molecule is unlikely. We obtained DNA from different human tissues and treated it with bisulfite to convert unmethylated cytosines to uracils. We then PCR-amplified short fragments containing the signature CpG site and multiple adjacent CpGs. Amplicons were designed to maximize the number of potentially informative CpG sites (to improve specificity in the detection of tissue-specific patterns) while minimizing overall fragment size (to improve the sensitivity of the assay, given that the average size of cfDNA fragments is 165 bp). We sequenced multiple molecules from the PCR product using Illumina MiSeq to assess the fraction of molecules with tissue-specific methylation patterns. In some cases, as an alternative approach to the comparisons between Illumina methylene arrays, we selected and validated tissue-specific markers based on promoters of known tissue-specific genes (Fig. S1). As shown in the examples below, scoring for DNA molecules in which multiple adjacent CpG sites share the same tissue-specific methylation pattern gave a much greater discriminatory power between the tissue of interest and other tissues compared with the information content of individual CpG sites.

Thus, we have defined short sequences of DNA containing four to nine CpG sites whose combined methylation status constitutes an epigenetic signature unique to a tissue of interest relative to blood cells and other tissues.

Presence of Unmethylated INS Promoter in the Circulation of T1D Patients. To detect cfDNA derived from β cells, we used the INS promoter as a β-cell–specific methylation marker. Previous studies seeking to identify DNA derived from β cells in peripheral blood samples have used methylation-specific PCR based on the methylation status of two or three CpG dinucleotides in the INS promoter (22). However, the INS promoter contains additional CpG sites in close proximity, which can be used to improve the distinction between DNA of β cells and other tissues (Fig. 1A). To test this concept, we amplified a 160-bp fragment of the INS promoter from bisulfite-treated DNA obtained from multiple tissues and sequenced the product to determine the methylation status of each CpG in each tissue. As shown in Fig. 1B, each individual CpG was unmethylated in 90–95% of the DNA molecules from human β cells and in 5–15% of the DNA molecules from other tissues. However, when we assessed six sites in combination, we found that six fully unmethylated sites were present in roughly 80% of the DNA molecules from β cells but in less than 0.01% of the molecules from any other tissue, thus giving an extremely high specificity for β cells.

To determine the sensitivity and linearity of the assay, we spiked human β-cell DNA into human lymphocyte DNA in different proportions and determined the frequency of unmethylated INS promoter DNA. The measured methylation signal was in excellent correlation with the input material, and β-cell DNA could be detected even when diluted 1:1,000 in lymphocyte DNA (Fig. 1C).

We then used this information to look for β-cell–derived cfDNA in the circulation. Plasma DNA from healthy volunteers and T1D patients was treated with bisulfite, PCR amplified, and sequenced to determine the fraction of molecules containing fully unmethylated INS promoter DNA. The fraction obtained was multiplied by the concentration of cfDNA measured in each sample to obtain the concentration of β-cell–derived DNA circulating in the blood of each patient (Fig. S1).

The cfDNA of healthy volunteers (n = 31) had an extremely low frequency of fully unmethylated INS promoter molecules (i.e., with all six CpGs unmethylated); less than 0.12% of circulating fragments had this sequence. When multiplied by the total amount of cfDNA in each individual, we found that less than 0.06 ng cfDNA/mL plasma was derived from β cells (equivalent to 10 genomes/mL), consistent with a very low rate of β-cell turnover in healthy adults (Fig. 1D). Plasma from all T1D patients
sampled 2–16 wk after diagnosis (n = 11) showed a clear signal of unmethylated INS promoter DNA in cfDNA, (350–2,900 copies of unmethylated INS promoter DNA/mL of plasma, equivalent to 175–1,450 β-cell genomes/mL), indicating ongoing β-cell death (Fig. 1E). The fraction of cfDNA derived from β-cells ranged from 0 to 0.1% in controls and from 1.9–5.5% in patients (Table S1).

To confirm that the combined methylation pattern of multiple CpG sites at the INS promoter was necessary to detect β-cell-derived DNA in the circulation, we examined the methylation status of each individual CpG in the plasma of healthy individuals and of persons with recently diagnosed T1D. Each individual CpG did not have a different pattern in the plasma of healthy controls or of T1D patients (unmethylated in ∼15% of cfDNA molecules), but collectively the six CpG sites yielded a clear signal in the plasma of T1D patients that was absent in healthy controls (Fig. S2).

We also studied plasma samples taken from patients with long-standing T1D who had been transplanted with cadaveric allogeneic islets and treated with immune suppressants (26). As shown in Fig. 1F, the plasma of all patients (n = 10) had a high signal (unmethylated INS promoter DNA) 1–2 h after transplantation, which declined dramatically in the hours and days that followed.
The extensive loss of grafted β cells immediately after transplantation is consistent with a previous imaging study of a transplanted patient (27). The levels of β-cell cfDNA shortly after transplantation were in good correlation with the amount of transplanted islets, presumably because transplanted islets contained a fraction of dead or dying cells (Fig. 1G). In most patients, signals above background were clearly detected at 7 d and even 1 mo after transplantation. Given the short half-life of cfDNA (2), the observed signals (up to 200 copies/mL plasma) may reflect significant ongoing loss of β cells despite immune suppression. More work will be required to determine if cfDNA levels predict long-term transplant outcome.

The current standard assay to assess β-cell damage in the transplant setting is plasma c-peptide, which reflects both physiologic insulin secretion and insulin released from damaged or dead β cells. In the hours immediately after transplantation, when patients are maintained in normoglycemic status by exogenous insulin, circulating c-peptide reflects β-cell damage. Indeed, c-peptide levels and unmethylated INS promoter cfDNA were in excellent correlation in the plasma of patients 1–2 h after transplantation (Fig. 1H), supporting the validity of β-cell cfDNA as a marker of acute β-cell death. These results demonstrate that our NGS-based method represents a highly sensitive and specific assay for the detection of cfDNA derived from specific tissues. With respect to T1D, signal analysis achieved complete separation between healthy controls and recently diagnosed patients; this result contrasts favorably with previous reports, which demonstrated a significant signal overlap between healthy controls and diabetic patients (21–24).

![Cell culture](image_url)

**Fig. 2.** Identification of oligodendrocyte-derived cfDNA in MS. (A) Methylation status of MBP3 and WM1 in DNA from multiple tissues and from sorted human neurons and oligodendrocytes (see also Figs. S3 and S4). (B) Spike-in experiments. Brain DNA was mixed with lymphocyte DNA, and lack of methylation of oligodendrocyte markers was used to estimate the fraction of oligodendrocyte DNA in the mixtures. Note that the measured frequency is lower than the input frequency, likely because input (brain DNA in this case) is a mixture of DNA from glial and other cell types. (C) Oligodendrocyte-derived DNA in the serum or plasma of healthy individuals, derived from the fraction of oligodendrocyte DNA (Table S1) and the total amount of cfDNA. (D) Oligodendrocyte-derived DNA in the serum of remitting and relapsing MS/NMO patients. The graph shows the cumulative values of unmethylated MBP3 and WM1 in each sample. Controls vs. stable disease, $P = 0.6$; controls vs. relapsing disease, $P < 0.0001$; stable vs. relapsing disease, $P < 0.0001$; controls vs. all patients, $P = 0.021$. 

Lehmann-Werman et al. PNAS | Published online March 14, 2016 | E1829
well as in additional settings of interest, such as monitoring the efficacy of immune suppression used to prevent rejection of transplanted β cells.

Identification of Oligodendrocyte-Derived cfDNA in Multiple Sclerosis. Noninvasive detection of brain-cell death is particularly challenging. To test the hypothesis that brain-specific methylation patterns can be used to identify brain-derived cfDNA, we looked for evidence of oligodendrocyte DNA circulating in the blood of patients with multiple sclerosis (MS) and neuromyelitis optica (NMO). Both are autoimmune diseases in which oligodendrocyte and astrocyte cell death occurs and in which the blood-brain barrier is disrupted (28, 29). We analyzed the published methylation of normal human white matter (30) and identified clusters of adjacent CpG sites in the UTR of myelin basic protein (termed here MBP3) and around an unannotated locus (CG10809560 in the Illumina array, herein WM1) for white matter 1 that were unmethylated selectively in sorted human oligodendrocytes (Fig. 2A). As with the INS promoter, individual CpGs in these clusters had a moderate signal-to-noise ratio, but combining all CpGs at the MBP3 or WM1 loci greatly increased the discrimination between oligodendrocyte DNA and DNA from other sources, including lymphocytes (Figs. S3 and S4). Thus, DNA fragments from the MBP3 or WM1 loci, when unmethylated in all CpG sites, can serve as an exclusive marker of oligodendrocytes.

We then spiked human brain DNA into human lymphocyte DNA in different proportions and found excellent correlation between measured methylation signal and the input material, so that oligodendrocyte DNA was detected even when diluted 1:1,000 in lymphocyte DNA (Fig. 2B). We then tested if these oligodendrocyte-derived fragments can be found in plasma or serum. Healthy individuals (n = 47) had negligible levels of unmethylated MBP3 or WM1 in their circulation, suggesting minimal basal turnover of oligodendrocytes (Fig. 2C). Strikingly, during disease exacerbation (a relapse documented both clinically and using brain MRI close to the time of sampling), most MS patients (14/19) had unmethylated cfDNA of MBP3, WM1, or both in their serum (Fig. 2D). Stable MS patients (n = 30) had a minimal or absent signal in serum (Fig. 2D). This observation is consistent with the notion that short-lived unmethylated MBP3 or WM1 cfDNA reflects acute oligodendrocyte cell death. Initial analysis did not reveal clinical correlates to the lack of signal in some relapsing patients. No correlation was observed between the signal and age, sex, Expanded Disability Status Scale (EDSS), or disease duration.

These results indicate that acute oligodendrocyte death can be detected by increased levels of fully unmethylated cfDNA fragments from the MBP3 or WM1 loci, presenting an opportunity for diagnosis and monitoring of demyelinating diseases.

Identification of Brain-Derived cfDNA After Acute Brain Damage. To obtain a more generally applicable marker of brain injury, we scanned the Illumina arrays for loci whose methylation status distinguished brain DNA from other tissues. A cluster of nine CpG sites around locus CG09787504 (here termed "Brain1") was fully unmethylated in 70% of DNA from various sources of brain tissue (enriched for either neurons or glia) and in <5% of DNA molecules from other tissues (likely reflecting DNA from peripheral neurons present in these tissues). Importantly, <0.03% of molecules in lymphocytes were unmethylated, giving a >2,000-fold difference in methylation of this locus between brain and lymphocytes (Fig. 3d and Fig. S5). Spike-in experiments showed that human cortex DNA can be detected even when diluted 1:1,000 into human lymphocyte DNA and that the signal recorded correlates perfectly with the level of input material (Fig. 3B).

Healthy individuals (n = 47) had extremely low levels of fully unmethylated Brain1 in serum or plasma (Fig. 3C). This low baseline signal may reflect minimal physiological neuronal turnover or an alternative mechanism for the clearance of DNA from dying brain cells. We then examined serum samples from patients in two scenarios, cardiac arrest and TBI; both situations involve neuronal injury in combination with disruption of the blood-brain barrier.

Following cardiac arrest, patients with documented ischemic brain damage sampled at multiple time points (n = 10 patients, 60 samples) (31) all had high levels of unmethylated Brain1 in serum (Fig. 3D). Similarly, patients hospitalized in an intensive care unit after severe TBI (n = 15 patients, 102 samples) had elevated unmethylated Brain1 in serum (Fig. 3E). Both sets of results are consistent with cfDNA fragments derived from dead brain cells (neurons or glia) in these patients. The amount and temporal patterns of brain-derived cfDNA varied between patients with cardiac arrest and those with TBI. In the group of patients with cardiac arrest, the strongest signals were observed at the first time point, shortly after resuscitation, with a decline in subsequent days in most patients. In the group of patients with TBI, a more delayed rise in brain-derived cfDNA was observed.

Identification of Exocrine Pancreas-Derived cfDNA in Pancreatic Cancer and Pancreatitis. Finally, we tested whether the approach can be used to detect cfDNA in the context of cancer. Although tumors have extensive methylation changes compared with normal tissue, the majority of tissue-specific methylation sites remain intact in tumors (32). Thus, cell death in tumors should release cfDNA carrying the normal methylation pattern of the source tissue. Pancreatic ductal adenocarcinoma is thought to originate from eitheracinor duct cells in the exocrine pancreas. We used antibodies to FACS-purify duct and acinar cells from cadaveric human material and obtained their methylomes using Illumina 450k arrays. Analysis of these data revealed multiple CpGs that were unmethylated in the exocrine pancreas and methylated in most other tissues, including the endocrine pancreas. We selected two sites for further analysis and identified clusters of adjacent CpGs that could be used as markers for the exocrine pancreas, distinguishing acinar and ductal cells from other cell types (Fig. 4A and Figs. S6 and S7). Spike-in experiments showed that human pancreas DNA can be detected even when diluted 1:1,000 into human lymphocyte DNA and that the signal recorded correlates perfectly with the level of input material (Fig. 4B). Healthy subjects (n = 47) had very low levels of unmethylated exocrine pancreas markers in their cfDNA (both plasma and serum), consistent with a low turnover of this tissue (Fig. 4C). Nearly half the patients with pancreatic cancer (20/42) displayed exocrine pancreas-derived cfDNA above background level in plasma or serum (Fig. 4D). There was a trend toward a stronger signal in patients with advanced disease, and these patients were more likely to show a signal above background level. Nevertheless, some patients (11/29) with stage 1 and 2 (localized) disease had a clear signal, suggesting this method has the potential to identify cell death in pancreatic cancer at a resectable stage.

The majority of pancreatic cancers carry somatic mutations in the KRAS (Kirsten rat sarcoma) gene, and these mutations can be detected in blood to monitor tumor cell death (although the presence of mutant KRAS per se does not reveal the tissue origin of mutant DNA) (33, 34). To compare the performance of our methylation-based assay with the detection of mutant KRAS, we
Fig. 3. Identification of brain-derived cfDNA after brain damage. (A) Methylation status of CpG sites at the CG079787504 locus (Brain1) in multiple tissues, as determined by deep sequencing. Bars represent the percentage of molecules in which all nine CpGs of the locus are unmethylated. (B) Spike-in experiment. Cortex DNA was mixed with lymphocyte DNA, and lack of methylation of Brain1 was used to estimate the fraction of brain DNA in the mixtures. (C) Brain-derived DNA in the serum or plasma of 47 healthy volunteers, derived from the fraction of fully unmethylated Brain1 molecules (Table S1) and the amount of cfDNA in each individual. (D) Brain-derived DNA in the serum of 10 patients after cardiac arrest. Each patient was sampled immediately after resuscitation (“acute”) and at subsequent time points. Healthy controls vs. patients (all time points), P < 0.0001. (E) Brain-derived DNA in the serum of 15 patients after TBI, sampled at different days after admission to a neurotrauma unit. Healthy controls vs. patients (all time points), P < 0.005.
used sequencing to quantify codon 12/13 KRAS mutations in cfDNA of patients diagnosed with pancreatic cancer. We detected mutant KRAS in 22% of the samples examined (n = 27), all among patients with stage 4 disease; by contrast, 48% of samples showed a methylation signal. Overall, there was 59% agreement between the two tests (Fig. 4D). More work will be needed to compare the sensitivity of KRAS mutation detection in cfDNA and pancreas-derived cfDNA by methylation.

To test further the hypothesis that cell death leads to increases in tissue-specific cfDNA irrespective of etiology, we examined the serum of patients with chronic pancreatitis. Indeed, 7 of 10 patients had elevated pancreas-derived cfDNA levels (Fig. 4D). No mutant KRAS was found in the serum of patients with pancreatitis (Fig. 4D). We note that patients with pancreatitis had a clearer signal with a marker that was unmethylated in both acinar and ductal cells (cut-like homeobox 2; CUX2), perhaps reflecting the different epigenetic identities of dying cells in the two pathologies. In summary, cfDNA carrying the methylation patterns of the exocrine pancreas is present in the blood of patients with pancreatic cancer and pancreatitis, reflecting death of exocrine cells in these conditions.
Discussion

The dynamics of cell death in vivo remain largely inaccessible to noninvasive investigation and diagnosis. The assays presented herein rely on two well-established principles in biology, namely, that dying cells release cfDNA and that each tissue has a unique DNA methylation pattern. Combining these principles allowed us to identify the tissue origins of cfDNA and hence to assess the rate of cell death in tissues of interest. A unique feature of the method is the ability to detect cfDNA derived from tissues with normal genomes that are not accessible to mutation-based or genomic variant-based cfDNA analysis. As a proof of principle, we assessed six tissue-specific methylation markers in >600 plasma/serum samples from >200 individuals and demonstrated the presence of tissue-specific cfDNA in patients with T1D (β-cell DNA), MS (oligodendrocyte DNA), brain damage caused by TBI or cardiac arrest (neuron or glia DNA), and pancreatic cancer or pancreatitis (exocrine pancreas DNA). In all cases, a considerable proportion of patients had tissue-specific cfDNA at levels far above baseline. This proof-of-principle study did not allow a full assessment of the diagnostic performance of our assay, but there is a reason to believe that measurements of cfDNA methylation-based cell death can contribute to clinical practice. In islet-graft recipients, cfDNA methylation-based cell death can contribute to clinical our assay, but there is a reason to believe that measurements of cancer or pancreatitis (exocrine pancreas DNA). In all cases, a considerable proportion of patients had tissue-specific cfDNA at levels far above baseline. This proof-of-principle study did not allow a full assessment of the diagnostic performance of our assay, but there is a reason to believe that measurements of cfDNA methylation-based cell death can contribute to clinical practice. In islet-graft recipients, cfDNA methylation-based cell death can contribute to clinical

For β-cells the known uniqueness of insulin expression directed us and others to study the insulin-regulatory regions, but the selection of target loci for other cell types was not trivial. We took a genome-wide approach to identify marker genomic loci, and these provided excellent discrimination between the methylation patterns in white matter, brain cells, or exocrine pancreas as opposed to all other tissues tested. In general, the resolution of the assay should increase with the use of more methylation markers for each tissue tested.

Our current protocol searches for cfDNA derived from a specific tissue of interest based on one or more unique methylation marks. We present data on six markers in four different cell types, but the method can be used to identify cell-type–specific makers in any tissue of the body. While this paper was being prepared for submission, Sun et al. (35) reported a different approach for mapping the tissue origins of plasma DNA using genome-wide bisulfite sequencing. Although both approaches rely on tissue-specific methylation patterns, they differ substantially. The Sun study requires deep and expensive sequencing (currently >$1,000 per sample), which may not be practical for routine diagnostics; our approach is based on targeted sequencing of specific markers (~$10 per sample). In addition, the Sun study demonstrated the tissue origins of cfDNA in conditions in which the source tissue differs genetically from the host (pregnancy, transplantation, and cancer), whereas we show the primary role of cfDNA—and infer cell death—in pathologies in which no alternative to methylation exists, e.g., β-cell death in diabetes and brain cell death in MS and head trauma.

In summary, we present a method for the detection of cell death in specific tissues, based on tissue-specific methylation patterns in circulating DNA. The approach may have multiple applications, including assessment of tissue damage after injury, evaluation of both targeted and off-target (toxicity) cell death in response to therapy, and early diagnosis of diseases such as T1D, neurodegenerative disease, and cancer. However, its use will require a deeper understanding of the rules that govern cfDNA dynamics and of the baseline distribution of tissue-specific cfDNA in different individuals across a variety of physiological conditions. In the long run, we envision a new type of blood test aimed at the sensitive detection of tissue damage without a priori suspicion of disease in a specific organ. We believe that such a tool will have broad utility in diagnostic medicine and in the study of human biology.

Materials and Methods

Patients. All clinical studies were approved by the relevant local ethics committees. The study in cardiac arrest patients was approved by the Ethics Committee at Uppsala University. The study in patients with TBI was approved by the Ethics Committee at University of Gothenburg, Sweden. Studies in patients with pancreatic cancer and pancreatitis were approved by the Ethics Committees of the Hebrew University-Hadassah Medical Center of Jerusalem, Sheba Medical Center, Israel; and the Department of Surgery, Philippus University of Marburg, Germany. The study in patients with T1D was approved by the Ethics Committee at the University of Florida. The study with islet-graft recipients was approved by the Ethics Committee at the University of Alberta, Canada. The study with MS and NMO patients was approved by the Ethics Committee at the Hadassah Medical Center, Jerusalem. Informed consent was obtained from all subjects or from their legal guardians before blood sampling. Fully de-identified samples were shipped to the Hebrew University Medical School for analysis. See SI Materials and Methods for detailed information about patient selection.

Biomarkers. Tissue-specific methylation biomarkers were selected after a comparison of extensive genome-wide DNA methylation datasets generated using Illumina Infinium HumanMethylation450 BeadChip array. See SI Materials and Methods for a detailed explanation of the datasets used and the method for selecting tissue-specific methylation markers. To detect mutant KRAS in plasma, bisulfite-treated cfDNA was PCR amplified using primers flanking codons 12 and 13 of the KRAS gene and specific for bisulfite-treated DNA. PCR products were sequenced on an Illumina MiSeq machine.

Lehmann-Werman et al.
PNAS | Published online March 14, 2016 | E1833
Sample Preparation and DNA Processing. Blood samples were collected in plasma-preparation tubes or serum-separator tubes and were centrifuged for 10 min at 3,320 x g within 2 h after collection. The experiments with T1D patients were performed using plasma. The MS, TBI, cardiac arrest, and pancreatitis studies were performed using serum because of the limited availability of archived material. Pancreatic ductal adenoma carcinoma samples were both serum and plasma, with no consistent differences detected between the sources. Control samples were matched to patient samples in both serum and plasma allowed effective detection of methylation patterns.

We used primers (single-strand molecular probes (Invitrogen). Bisulfite-treated DNA was PCR amplified using denaturants (SI Materials and Methods) specific for bisulfite-treated DNA but independent of methylation status at monitored CpG sites. Primers were bar-coded, allowing the mixing of samples from different individuals when sequencing products using MiSeq (Illumina). Sequencing was performed on PCR products using MiSeq Reagent Kit v2 (MiSeq, Illumina). The bisulfite treated DNA but independent of methylation status at monitored CpG sites. We used a two-tailed Mann–Whitney test based on values of unmethylated tissue-specific DNA in each patient.

Statistical Analysis. To assess the significance of differences between groups, we considered methylated if “CG” was read and were considered unmethylated if “TG” was read. Efficiency of bisulfite conversion was assessed by analyzing the methylation of non-CpG cytosines.

Acknowledgments. We thank the T1D Exchange Biobank and investigators and staff of T1D Exchange Protocols for subject recruitment, Tomer Nir and Adi Nir for discussions, and Marc Gotkine for critical reading of the manuscript. This work was supported by the T1D Exchange, a program of T1D First; in part by a grant from the American Schools and Hospitals Abroad Program of the US Agency for International Development for upgrading the Hebrew University Medical School Flow Cytometry Laboratory; and by grants from the Juvenile Diabetes Research Foundation (3-SDA-2013-39-C-R); the Beta Cell Biology Consortium and the Human Islet Research Network of the National Institutes of Health (DK104216); the Sir Zalman Coven Universities Fund; a Trilateral German-Israel-Palestine program of the Deutsche Forschungsgemeinschaft; the Soyka Pancreatic Cancer Fund; and the Israeli Centers for Research Excellence Program of The Israel Science Foundation 41.11 (to Y.D.). R.L.-W. was supported by a training grant from Teva Pharmaceutical Industries Ltd. as part of the Israeli National Network of Excellence in Neuroscience. K.B. holds the Torsten Söderberg Professorship in Medicine at the Royal Swedish Academy of Sciences.

Supporting Information

Lehmann-Werman et al. 10.1073/pnas.1519286113

SI Materials and Methods

Patient Selection.

Recently diagnosed T1D patients. Plasma was prepared from 11 patients (age 4–20 y, mean age, 9.5 y) who had been diagnosed with T1D 1–4 mo earlier.

Islet-graft recipients. Patients were 44–57 y old, with a T1D duration of 10–36 y and poorly controlled blood glucose levels (HbA1C 6.4–10). Anti-rejection therapy included alemtuzumab, etanercept, and anakinra. Maintenance therapy included tacrolimus and mycophenolate mofetil.

MS and NMO patients. MS and NMO patients were diagnosed according to the 2010 McDonald criteria (37) and the NMO diagnostic criteria, respectively (38). Patients’ characteristics were as follows: relapsing–remitting MS patients, n = 49; 74% females; average age = 36 ± 12.5 y, range 18–68 y; disease duration 4 ± 4.5 y, range 0–14 y; EDSS score 2.8 ± 1.8, range 1–7.5. None of the patients had received steroid treatment for 2 mo before testing. In patients in relapse, blood was drawn before i.v. steroid therapy. At the time of sampling one patient was treated with Copaxone, and four were treated with Imuran.

Cardiac arrest patients. Samples from cardiac arrest patients were collected at the intensive care unit at Upssala University Hospital, Uppsala. Ten unconscious patients with cardiac arrest were resuscitated with restoration of spontaneous circulation. Hypothermia treatment to a body temperature of 32–34 °C for 24 h, ventilation, and pharmacologic support were administered immediately after resuscitation as described (39). Patients were defined as comatose if they were (i) not awake, (ii) not following any commands, and (iii) not responding to any stimuli. All patients received an arterial line in the radial or femoral artery for blood sampling. Serial blood samples were collected, starting as soon as possible in the emergency phase and continuing at 24, 48, 72, 96, and 108 h after cardiac arrest. Serum aliquots were frozen at −70°C until analysis. The study was approved by the Ethics Committee at Uppsala University; informed consent was obtained from the patient’s closest relative.

TBI. Eleven patients (all of European ancestry; average age 39 y) with isolated severe TBI were enrolled from a clinical study at the Neurointensive Care Unit (NICU) at the Sahlgrenska University Hospital, Gothenburg, Sweden. All patients had a mixture of focal contusions and general edema. All had severe TBI according to the following criteria: (i) had a score of 4 on the reaction level scale, corresponding to a score of 8 on the Glasgow coma scale (3 = no response, 15 = awake) (40); (ii) were in need on ventilator treatment; (iii) were monitored for intracranial pressure (ICP). Venous blood samples were taken on serial days after trauma. After clinical and radiologic evaluations, the patients underwent neurosurgical intervention within hours after admission to receive an indwelling ventricular catheter for ICP monitoring/therapeutic CSF drainage. When appropriate, space-occupying lesions such as hemorhages and contusions were removed surgically.

Patients then were treated in accordance with a standardized protocol, the Lund concept, with the aim of maintaining cerebral perfusion pressure >60 mm Hg and the ICP <20 mm Hg (41). Data collected included demographic and clinical variables such as age, sex, and time of injury. Physiologic and laboratory variables were recorded continuously throughout the study period and were adjusted concomitantly to be kept within the following limits: hemoglobin >120 g/L; serum sodium >135 to <150 mmol/L; serum potassium >4.0 to <5.0 mmol/L; serum albumin >35 to <50 g/L; core temperature 37 ± 0.5 °C; mean arterial blood pressure (MABP) between 70 and 100 mm Hg; ICP <20 mm Hg; cerebral perfusion pressure (MABP – ICP) >60 mm Hg; PO2 >12 to <18 kPa; PCO2 around 4.5 kPa; and normalized pH. Blood glucose was kept between 4 and 6 mmol/L according to the NICU routine. No patient received steroids.

The study was approved by the Ethics Committee at the University of Gothenburg; informed consent was obtained from the patient’s closest relative.

Pancreatic cancer and chronic pancreatitis. Plasma or serum was obtained from 42 patients with pathologically confirmed pancreatic adenocarcinoma (28 males and 14 females; average age 68 y, range 41–87 y) and 10 patients with chronic pancreatitis. The 2010 American Joint Committee on Cancer TNM staging of pancreatic cancer (42) was used. At the time of blood collection 29 patients had localized disease (4 at stage 1 and 25 at stage 2, all preoperative), and 13 patients had metastatic stage 4 disease. Samples were obtained from patients treated in three centers: the Hebrew University-Hadassah Medical Center, Jerusalem; Sheba Medical Center, Israel; and the Department of Surgery, Philipps University of Marburg, Germany. Studies were approved by the Ethics Committee of each institution, and informed written consent was obtained from all patients.

Healthy controls. A total of 46 healthy volunteers (50% females; age 22–60 y) participated in the study as unpaid healthy controls. All denied having any signs of symptoms related to the disease state studied.

Locally Generated Methylome Data. To obtain pure pancreatic α and β cells, we dissociated cadaveric human islets by Trypsin-EDTA solution B (EDTA 0.05%, Trypsin 0.25%, with Phenol Red) into a single-cell suspension. Cell pellets were fixed in 4% paraformaldehyde/PBS (Thermo Scientific) for 15 min at room temperature; then a 1% saponin (Sigma)/PBS solution was added to a final saponin concentration of 0.1% for 30 min at room temperature. After incubation, cells were washed once in Detergent Wash Buffer (WBD) (1% BSA, 0.1% saponin in PBS×1). For staining, guinea pig anti-insulin (1:500; Abcam) and mouse anti-glucagon (1:800; Abcam) were added for 90 min at room temperature; then cells were washed once in WBD, and CY5 anti-guinea pig (1:500) or CY2 anti-mouse (1:200) secondary antibodies were added to the cell pellet for 30 min. Cells were washed twice in WBD and once in wash buffer without saponin (1% BSA in PBS×1). Then the cells were filtered using a 5-mL polystyrene round-bottomed test tube with a cell-strainer cap (40-μm nylon mesh) designed for flow cytometry applications. We used BD FACS Aria III for sorting.

Duct and acinar cells were isolated from dissociated cadaveric human pancreas as described previously (43). Live cells were stained with cell-surface markers and sorted.

Genomic DNA was isolated using phenol/chloroform and was processed for Illumina 450k arrays according to the manufacturer’s instructions.

Selection of Methylation Biomarkers. Tissue-specific DNA methylation markers were selected by comparing genome-wide methylation data from 35 human tissues generated using the Illumina Infinium HumanMethylation450k BeadChip kit. Publicly available datasets used for comparison included The Cancer Genome Atlas (bladder, brain, breast, cervix, colon, head and neck, kidney, liver, lung, mesenchyme, pancreas, prostate, rectum, skin, stomach, thyroid, uterus), and Gene Expression Omnibus repositories GSE48472 (blood, liver, muscle, omentum, pancreas, spleen, s.c. fat, buccal swab, hair, saliva), GSE40360 (frontal lobe white matter), GSE53162 (prefrontal cortex, cerebellum), GSE5277 (heart), and GSE40279 (whole blood). In
addition, we locally generated methylome data for human pancreatic α cells (two replicates), pancreatic β cells (three replicates), pancreatic duct cells, and pancreatic acinar cells. The mean beta values for each tissue underwent quantile normalization. Candidate marker loci were defined as CpG sites with an average methylation beta value <0.5 in the tissue of interest, >0.9 in leukocytes, and >0.8 in >90% of tissues. The specific CpG sites that were selected for further validation by PCR and sequencing of adjacent CpG sites were as follows:

**INS.** The INS locus was found previously to be unmethylated in β cells and methylated in a number of other tissues (21, 44, 45). **Myelin basic protein.** Myelin basic protein has been characterized previously as important for oligodendrocyte function (46).

An individual CpG site in the 3′ end of the gene (Chr18:74692045–74692213) was found to have an average methylation beta value <0.2 in white matter, >0.9 in leukocytes, and >0.8 in >90% of tissues. WM1. The WM1 locus (Chr10:79936919–79937042) was selected for having an average methylation beta value <0.5 in white matter and >0.9 in leukocytes and in >90% of all tissues.

**CG0978 (Brain).** The CG0978 locus (chr10:3283832–3283996) had a methylation beta value <0.3 in cerebellum, >0.9 in heart and leukocytes, and >0.75 in all tissues other than β cells.

**REG1A.** The REG1A gene is highly expressed in the pancreas and digestive tract (47). The locus used as a marker (Chr2:79347448–79347588) was found to have an average methylation beta value <0.05 in pancreatic acinar cells, >0.9 in leukocytes, and >0.8 in all tissues not part of the digestive tract (0.581 stomach, 0.524 rectum, 0.49 colon).

**CUX2.** The CUX2 locus (Chr12:111664570–111664700) had an average methylation beta value of 0.11 in pancreatic duct cells, >0.9 in leukocytes and all nonpancreatic tissues, 0.79 in pancreatic acinar cells, 0.4 in α and β cells, and 0.7 in the whole pancreas.

Based on the identification of individual CpG sites with a tissue-specific methylation pattern, we developed a PCR-sequencing assay to assess the methylation status of multiple cytosines in the vicinity of these sites. Primers for differentially methylated areas were designed to complement bisulfite-treated DNA. The primers were selected to amplify a sequence shorter than 150 bp (to allow amplification of fragmented cfDNA) that contained four adjacent CpGs as a way to increase specificity for the tissue of interest. Primers were designed to include standard Illumina MiSeq adaptors plus barcodes. Genomic sequences of primers (from 5′ to 3′) were as follows:

**INS L:** TTTTGGGAGTTATGTTTAGT
**INS R:** ACTCACCCTCAAATCTCTTAC
**MBP3 L:** GTGTATTTTGGTTTTTGAAGATTT
**MBP3 R:** ACCAAACATTAAATAAAAACATTA
**WM1 L:** GTTGGATGTGGTTGGATTTATTAA
**WM1 R:** AAAAACCAATTAAACCCATTAAA
**Brain1 L:** TTATTTTAAAGATTTGGGTAGTTT
**Brain1 R:** AAATCCTTCTACAATCTCTACCT
**REG1A L:** AGTTGTTGTTATTTGTTGTTG
**REG1A R:** ATATCAAAATCTCTAAATAAAAACCA
**CUX2 L:** TTATTTGAAGATTTGGTTG
**CUX2 R:** TCCACACAAACTAATTATTAAC.

**DNA Methylation Analysis.** Sequenced reads were separated by barcode, aligned to the target sequence, and analyzed using custom scripts written and implemented in Matlab. Reads were quality filtered based on Illumina quality scores. Reads were identified by having at least 80% similarity to the target sequences after alignment and containing all the expected CpGs in the sequence. CpGs were considered methylated if “C” was read and were considered unmethylated if “T” was read. Successful bisulfite conversion was assessed by analyzing the methylation of non-CpG cytosines. The Matlab script used for data processing of ampalonic bisulfite sequencing reads is provided in Dataset S1.

**Defining the Sensitivity of the Assay.** To determine the linearity and sensitivity limits of our assay, we diluted DNA from specific tissues into lymphocyte DNA (representing the major source of DNA in plasma). We treated the mixed DNA with bisulfite and PCR amplified markers of the specific tissue. After sequencing the PCR products on MiSeq, we determined the fraction of reads with tissue-specific methylation patterns and used this information to assess the frequency of DNA from the specific tissue in the mixture. The results of the analysis for different tissue-specific methylation markers used in this study are shown in Figs. 1C, 2B, 3B, and 4B. The data suggest that tissue-specific signals can be detected at dilutions as low as 1:1,000 and that the observed frequency matches well the input frequency of tissue-specific DNA.

**Demonstrating the Advantage of Using Multiple Adherent CpGs.** As discussed above, the methylation pattern of a single CpG usually discriminates poorly between tissues; therefore we defined multiple adjacent unmethylated CpG sites as markers. The rationale for the use of multiple CpGs is that each cytosine might undergo accidental demethylation in an irrelevant tissue, but it is unlikely that multiple adjacent cytosines will all be demethylated accidentally. Thus, the more adjacent unmethylated cytosines present in a molecule, the less likely the molecule is to derive from irrelevant tissues. To test this assumption experimentally, we examined all possible methylation patterns of the INS promoter in the circulation of healthy individuals described in Fig. 1. As shown in detail in Fig. S8, the more unmethylated CpG sites in the INS promoter are taken into account, the lower is the background in plasma of healthy people. A similar phenomenon was observed in all other markers tested in the paper, namely, that longer stretches of unmethylated CpG sites are less likely to exist in the plasma of healthy individuals (Figs. S3–S7). These findings support the use of the methylation status of multiple adjacent CpGs as a way to increase specificity for the tissue of interest.

**Tissue-Specific cfDNA Versus Absolute Levels of cfDNA.** To calculate the concentration of cfDNA derived from a specific tissue, we multiplied the fraction of cfDNA derived from a specific tissue (as determined from the frequency of molecules carrying a tissue-specific methylation pattern) by the concentration of cfDNA measured in each particular patient. The collective data on the concentration of cfDNA are shown in Fig. S9. Although the total levels of cfDNA are elevated in some patients, with a trend toward elevation in particular pathologies, this information alone is not sufficient to make conclusions about cell death in specific tissues. Moreover, in some cases (e.g., in many T1D patients), the general level of cfDNA is not changed significantly. Furthermore, the concentration of total cfDNA is influenced by technical issues related to sample preparation, particularly white cell lysis before plasma/serum separation. Together, these issues underscore the necessity of using tissue-specific methylation markers to detect cell death in particular tissues.

Lehmann-Werman et al. www.pnas.org/cgi/content/short/1519286113 2 of 15
Fig. S1. Flowchart of the method of detecting circulating DNA derived from a specific tissue. (A) Procedure for identifying tissue-specific methylation markers. (B) Procedure for determining levels of tissue-specific cfDNA.

Fig. S2. Methylation of the *INS* promoter in the plasma of healthy volunteers and patients with recently diagnosed T1D. (A) Methylation status of individual CpG sites at the *INS* promoter. (B) Methylation status of an expanded window of four to six CpGs expressed as the percent of unmethylated DNA in the patients in A.
Fig. S3. (Continued)
Fig. S3. Methylation of the 3′ UTR of MBP3. (A) Structure of the MBP3 3′ UTR fragment used as marker. Lollipops represent CpGs. The empty lollipop represents the CpG detected in the Illumina 450k array. Arrows mark positions of PCR primers. (B) Methylation status of the individual CpG site at the MBP3 locus captured in the Illumina 450k array. Data are from publicly available 450k arrays (see Selection of Methylation Biomarkers). (C) Methylation status of individual CpG sites and an expanded window of multiple CpGs from the MBP3 locus, in multiple tissues, as determined by deep sequencing. (D) Methylation of individual CpG sites from the MBP3 locus in the serum of healthy controls and relapsing MS/NMO patients. (E) Fraction of unmethylated MBP3 locus fragments in the serum of healthy volunteers and in the MS/NMO patients in D. Total unmethylated MBP3 locus DNA expressed in nanograms per milliliter of serum is shown in Fig. 2C.
Fig. S4. (Continued)
Fig. S4. Methylation of CG10809560 and adjacent CpG sites (the WM1 locus). (A) Structure of the WM1 locus fragment used as marker. Lollipops represent CpGs. The empty lollipop represents the CpG detected in the Illumina 450k array. Arrows mark positions of PCR primers. (B) Methylation status of WM1 in multiple tissues as recorded in publicly available Illumina 450k arrays. (C) Methylation status of individual CpG sites and expanded window of multiple CpGs from the WM1 locus, in multiple tissues, as determined by deep sequencing. (D) Methylation of individual CpG sites from the WM1 locus in the serum of healthy controls and relapsing MS/NMO patients. (E) Fraction of unmethylated WM1 DNA fragments in the serum of healthy volunteers and in the MS/NMO patients in D.
Fig. S5. (Continued)
Fig. S5. Methylation of brain marker CG09787504 (Brain1) and adjacent CpG sites. (A) Structure of Brain1 locus fragment used as marker. Lollipops represent CpGs. The empty lollipop represents the CpG detected in the Illumina 450k array. Arrows mark positions of PCR primers. (B) Methylation status of Brain1 in multiple tissues as recorded in publicly available Illumina 450k arrays. (C) Methylation status of individual CpG sites and expanded window of multiple CpGs from the Brain1 locus in multiple tissues, as determined by deep sequencing. (D) Methylation of individual CpG sites from the Brain1 locus in the serum of healthy controls and patients after cardiac arrest. (E) Fraction of unmethylated Brain1 DNA fragments in the serum of healthy volunteers and the patients in D.
Fig. S6. (Continued)
Fig. S6. Methylation of the CpG cluster near the REG1A gene. (A) Structure of the REG1A fragment used as marker. Lollipops represent CpGs. The empty lollipop represents the CpG detected in the Illumina 450k array. Arrows mark positions of PCR primers. (B) Methylation status of the individual CpG site in the REG1A locus that is captured in the Illumina 450k array. Data are from publicly available 450k arrays (see Selection of Methylation Biomarkers). (C) Methylation status of individual CpG sites and expanded window of multiple CpGs from the REG1A locus in multiple tissues, as determined by deep sequencing. (D) Methylation of individual CpG sites from the REG1A locus in the serum of healthy controls and patients with pancreatic cancer. (E) Fraction of unmethylated REG1A fragments in the serum of healthy volunteers and patients in D.
Fig. S7. (Continued)
Fig. S7. Methylation of the CpG cluster near the CUX2 gene. (A) Structure of the CUX2 fragment used as marker. Lollipops represent CpGs. The empty lollipop represents the CpG detected in the Illumina 450K array. Arrows mark positions of PCR primers. (B) Methylation status of the individual CpG site at the CUX2 locus that is captured in the Illumina 450K array. Data are from publicly available 450K arrays (see Selection of Methylation Biomarkers). (C) Methylation status of individual CpG sites and the expanded window of multiple CpGs from the CUX2 locus in multiple tissues, as determined by deep sequencing. (D) Methylation of individual CpG sites from the CUX2 locus in the serum of healthy controls and patients with pancreatic cancer. (E) Fraction of unmethylated CUX2 fragments in the serum of healthy volunteers and patients in D.
Fig. S8. Frequency of unmethylated INS promoter molecules in the plasma of healthy individuals, determined by analyzing all possible combinations of CpGs within the amplified segment. The fraction of unmethylated molecules is shown for the INS promoter of 22 nondiabetic plasma samples, with all possible combinations of unmethylated CpGs. Fractions of unmethylated molecules are defined as (molecules with all analyzed CpGs unmethylated)/(all sequenced reads). The mean and SEM are shown for each CpG combination. The samples are colored according to the number of CpGs analyzed: one CpG, green; two CpGs, cyan; three CpGs, orange; four CpGs, purple; five CpGs, red; six CpGs, brown.

Fig. S9. Concentration of cfDNA in healthy individuals and patients. Each dot represents one patient. We measured the concentration after isolation of cfDNA and bisulfite treatment (before use as template for PCR). CP, patients with chronic pancreatitis; MS, patients with multiple sclerosis; Panc, patients with pancreatic cancer; T1D, recently diagnosed type 1 diabetes; TBI, patients with traumatic brain injury; transplants, islet-graft recipients.
Table S1. Percentage of cfDNA molecules derived from specific tissues

Dataset S1. Matlab script used to analyze MiSeq data
function methyl_plasma_pipeline()

%% Pipeline for plasma methylation analysis
%%
%% The user will be prompted for three files: a sequence fastq file, a barcode
%% fastq file, and an input tab-delimited text file.
%% The input text file should have the following format: target_name
%% sample_name barcode pre-bis_amplicon
%% Output: fastq files per barcode, raw,histogram,summary file for
%% each target-sample, summary of all results to Summary_results.txt
%% For questions please contact joshua.moss@mail.huji.ac.il

matlabpool open local
%% Files enter

[indexFile,folder,~] = uigetfile('../*.fastq','Choose Index file');
[readsFile,~,~] = uigetfile([folder '/*.fastq'],'Choose Reads
file');
[seqFile,~,~] = uigetfile([folder '/*.txt'],'Choose Sequence file');

cd(folder);

%% import info from sequence file

fid = fopen(seqFile);
info = textscan(fid,'%s %s %s %s','Delimiter',
{'	'},'CollectOutput',1,'Whitespace','''');
fclose(fid);
info = info{1};
sample_name1 = info(:,1);
sample_name2 = info(:,2);
sample_barcodes = info(:,3);
sample_seq = info(:,4);
[barcodes,~,sample_barcodes_num] = unique(sample_barcodes);
sampleNum = length(sample_barcodes_num);

%% write input summary to file

fid = fopen('input_summary.txt','w');
fprintf(fid, '%s	', 'Sample #');
fprintf(fid, '%s	', 'Name');
fprintf(fid, '%s	', 'Info');
fprintf(fid, '%s	', 'Barcode');
fprintf(fid, '%s	', 'Sequence');
for i = 1:sampleNum
    fprintf(fid,'
');
    fprintf(fid,'%f	',i);
    for j = 1:4
        fprintf(fid,'%s	',info{i,j});
    end
end
fclose(fid);

%% write fasta files
for i = 1:sampleNum
    fastawrite(['sample' int2str(i) '.fasta'], [sample_name1{i} ' ' sample_name2{i} i], sample_seq{i});
end

%% write barcode summary
fid = fopen('barcode_order.txt','w');
fprintf(fid,'%s	', 'Barcode #');
fprintf(fid, '%s	', 'Barcode');
for i = 1:length(barcodes)
    fprintf(fid,'
');
    fprintf(fid,'%f	', i);
    fprintf(fid,'%s	',barcodes{i});
end
fclose(fid);

%% quality filter
fprintf('%s','Fasta files written. Quality filter and barcode division will now begin. Please be patient!');
filter_and_seperate(indexFile,readsFile,barcodes);
fprintf('%s','Barcode division is done! Analysis will now be performed');
fprintf('
');

%% Methanalysis run

% Assign Params structs for each sample
Params_array = cell(sampleNum,1);
for i = 1:sampleNum
    Params.RefFile = ['sample' int2str(i) '.fasta']
    Params.ReadsFile = ['barcode' int2str(sample_barcodes_num(i)) '.fastq']
    Params_array{i} = Params;
end

% RUN
parfor i = 1:sampleNum
    RunMethAnalysisFile(Params_array{i});
end

matlabpool close

%% Summary file
num_CpGs = zeros(sampleNum,1);
for i = 1:sampleNum
    num_CpGs(i) = length(strfind(sample_seq{i},'CG'));
end
max_num_CpGs = max(num_CpGs);
numT = zeros(sampleNum,max_num_CpGs+3);
umT(:,:) = NaN;
for i = 1:sampleNum
    numT_sample = numToutput(num_CpGs(i),['sample' int2str(i)]);
    numT(i,1:length(numT_sample)) = numT_sample;
end

%print
fid = fopen('Summary_results.txt','w');
fprintf(fid,'%s	', 'Sample #');
fprintf(fid,'%s	', 'Gene');
fprintf(fid,'%s	', 'Sample');
fprintf(fid,'%s	', 'CpGs');
fprintf(fid,'%s	', 'All reads');
fprintf(fid,'%s	', 'All T');
for i = 1:max_num_CpGs
    fprintf(fid,'%s	', ['All T - ' int2str(i)]);
end
for i = 1:sampleNum
    fprintf(fid,'
');
    fprintf(fid,'%f	', i);
    fprintf(fid,'%s	', sample_name1{i});
    fprintf(fid,'%s	', sample_name2{i});
    for j = 1:length(numT(1,:))
        if isnan(numT(i,j))
            break;
        end
        fprintf(fid, '%f	', numT(i,j));
    end
end
fclose(fid);
msgbox('Congratulations - All done!');
end

function filter_and_seperate(indexFile,readsFile,barcodes)
    % Filters reads by the quality of their barcode read and separates
    % reads by barcode

    % set quality cutoff (can be adjusted)
    qt = 32;

    % read in relevant barcodes
    for i = 1:length(barcodes)
        barcodes{i} = seqrcomplement(barcodes{i});
    end

    % get general sequence information
    info_I = fastqinfo(indexFile);
    len_I = double(info_I.NumberOfEntries);

    fid_barcodes = cell(length(barcodes),1);
    for i = 1:length(barcodes)
        fid_barcodes{i} = fopen([['barcode' int2str(i) '.fastq'],'a']);
    end
fidI = fopen(indexFile);
 fidR = fopen(readsFile);

 tmp = fgetl(fidI);
i=0;
tic
while tmp ~= -1
 i = i+1;
 if (mod(i,1000)==0)
  t = toc;
  rem = t*(len_I-i)/i;
  disp(['Seperating barcodes ' num2str(i) '/' num2str(len_I)
  ' (' num2str(round(100*i/len_I)) '%) Remaining: ' datestr(datenum(0,0,0,0,0,rem),'HH:MM:SS')]);
 end

% read barcode info
 seq_I = fgetl(fidI);
 tmp = fgetl(fidI);
 qual_I = fgetl(fidI);
 tmp = fgetl(fidI);
 bc_compare = find(strcmp(seq_I,barcodes));

% read seq info
 header_R = fgetl(fidR);
 seq_R = fgetl(fidR);
 header2_R = fgetl(fidR);
 qual_R = fgetl(fidR);

 if ~isempty(bc_compare)
  % filter quality barcode reads
  qual_I_num = mean(double(qual_I)-33);
  if qual_I_num > qt
   % write to barcode file
   fprintf(fid_barcodes{bc_compare}, '%s
', header_R);
   fprintf(fid_barcodes{bc_compare}, '%s
', seq_R);
   fprintf(fid_barcodes{bc_compare}, '%s
', header2_R);
   fprintf(fid_barcodes{bc_compare}, '%s
', qual_R);
  end
 end
end

fclose(fidI);
fclose(fidR);

for i = 1:length(barcodes)
  fclose(fid_barcodes{i});
end

fprintf('%s','Quality filter is done.');
fprintf('
');

function RunMethAnalysisFile(Params)
    %% Performs local alignment of all reads in a fastq file to target amplicon in fasta file
    %% Outputs three files for each sample:
    %% .meth_raw - shows raw alignments and CpG information
    %% .meth_CpG_hist - shows counts of each different combinations of CpGs in aligned reads
    %% .meth_summary - shows the methylation information for each CpG from target amplicon
    Params.OutputFileName = strrep(Params.RefFile,'.fasta','.meth_raw');
    name_core = strrep(Params.RefFile,'.fasta','');
    
    Params.Thresholds.PercentCorrectAlignment = 0.8; % sets minimum percent of target sequence aligned to read; can be adjusted
    Params.Thresholds.MinReadLength = 50; % sets minimum read length; can be adjusted
    Params.Thresholds.MinAlignmentLength = 50; % sets minimum length of aligned region in read; can be adjusted

    if (exist(Params.RefFile) == 0)
        disp(['Reference file does not exist: ' Params.RefFile]);
        return;
    end
    if (exist(Params.ReadsFile) == 0)
        error(['Reads file does not exist: ' Params.ReadsFile]);
    end
    disp(['Counting Lines for file: ' Params.OutputFileName]);
    info = fastqinfo(Params.ReadsFile);
    len = double(info.NumberOfEntries);
    Refs = fastaread(Params.RefFile);
    fidOut = fopen(Params.OutputFileName,'w');
    if (fidOut == -1)
        error(['Cannot create output file: ' Params.OutputFileName]);
    end
    % Convert all C which is not CG to T
    RefsConv = ConvertNonCGtoT(Refs);
    Refs.Sequence = upper(Refs.Sequence);
    CGinds = findstr(Refs.Sequence,'CG');
    Cinds = find(Refs.Sequence=='C');
    NonCGinds = setdiff(Cinds,CGinds);
    [~, CGloc] = intersect(Cinds,CGinds);
    NonCGloc = setdiff(1:length(Cinds),CGloc);
    Nhist = zeros(5,length(Cinds));
Nhist_CG_Reads = struct;

ReadsCount = 0;

ConvTable('A') = 1;
ConvTable('a') = 1;
ConvTable('C') = 2;
ConvTable('c') = 2;
ConvTable('G') = 3;
ConvTable('g') = 3;
ConvTable('T') = 4;
ConvTable('t') = 4;
ConvTable('-') = 5; % No Match
ConvTable('N') = 5; % No Match

fidReads = fopen(Params.ReadsFile,'r');
if (fidReads == -1)
   error(['Cannot open reads file: ' Params.ReadsFile]);
end

i=0;
line=1;
tic;
while (line ~= -1)
i=i+1;
   if (mod(i,1000)==0)
      t = toc;
      rem = t*(len-i)/i;
      disp([name_core ': Finished ' num2str(i) '/' num2str(len)
            '(%) Remaining: '
datestr(datenum(0,0,0,0,0,rem),'HH:MM:SS')]);
   end

   lineName = fgetl(fidReads);
   line = fgetl(fidReads);
   lineTmp = fgetl(fidReads);
   lineTmp = fgetl(fidReads);
   if (lineTmp == -1)
      break;
   end

   % Print read name
   fprintf(fidOut,'%s\t', lineName);

   % Ignore short reads
   if (length(line) < Params.Thresholds.MinReadLength)
      fprintf(fidOut,'Read_Length<%d
\t',Params.Thresholds.MinReadLength);
   end

   % Print original line
   fprintf(fidOut,'%s\n', line);
continue;
end
foundRef = 0;

 [~, Al, St] = swalign(RefsConv.Sequence,line);
% Ignore short alignments
if (length(Al(2,:)) < Params.Thresholds.MinAlignmentLength)
    fprintf(fidOut,'Align_Length<%.d
\t',Params.Thresholds.MinAlignmentLength);
% Print original line
    fprintf(fidOut,'%s\n', line);
    continue;
end
RefScore = sum(Al(2,:)=='|')/length(Al(2,:));
% Check correct target read
if (RefScore >= Params.Thresholds.PercentCorrectAlignment)
    ReadCheck = Al(3,find(Al(1,:)=='-'));
    CindsCheck = Cinds(find(Cinds>=St(1)))-St(1)+1;
    iShift = length(Cinds) - length(CindsCheck)+1;
    CindsCheck = CindsCheck(find(CindsCheck <= length(ReadCheck)));
    foundRef = 1;
    inds = sub2ind(size(Nhist),
    ConvTable(ReadCheck(CindsCheck)),
    iShift:length(ReadCheck(CindsCheck))+iShift-1);
    alignStr=repmat('-',1,length(RefsConv.Sequence));
    alignStr(St(1):St(1)+length(ReadCheck)-1) = ReadCheck;
    ReadsCount = ReadsCount+1;
    Nhist(inds) = Nhist(inds)+1;
    histf = alignStr(CGinds);
    histf(histf == '-') = 'X';
    if (isfield(Nhist_CG_Reads,histf) == 0)
        Nhist_CG_Reads.(histf) = 1;
    else
        Nhist_CG_Reads.(histf) = Nhist_CG_Reads.(histf)+1;
    end
    if (foundRef == 0)
        fprintf(fidOut,'No_Ref_Found\t');
% Print original line
        fprintf(fidOut,'%s\n', line);
    else
% Print Reference name
        fprintf(fidOut,'%s\t', Refs.Header);
% Print original line
        fprintf(fidOut,'%s\t', line);
% print alignment sring
        fprintf(fidOut,'%s	', alignStr);
% print CG indices

fprintf(fidOut,'CpGs:	');
for k=1:length(CGinds)
    fprintf(fidOut,'%s ', alignStr(CGinds(k)));
end
fprintf(fidOut,'
');
end
close(fidOut);
close(fidReads);

% Write summary data
Params.OutputSummary =
strrep(Params.RefFile,'.fasta','.meth_summary');

fid = fopen(Params.OutputSummary,'w');
if (fid == -1)
    error(["Cannot open output summary file: 
        Params.OutputSummary"]) end
fprintf(fid,'Reads file name:	%s
',Params.ReadsFile);
fprintf(fid,'Total Reads:	%d
',i);
fprintf(fid,'Sample name:	%s
',Refs.Header);
fprintf(fid,'# of reads aligned:	%d
',ReadsCount);
fprintf(fid,'CpG sites
index	#A	#C	#G	#T	#-	%%Meth
');
for j=1:length(CGloc)
    fprintf(fid,'%d	%d	%d	%d	%d	%d	%.2f
',CGinds(j),Nhist(1,CGloc(j)),Nhist(2,CGloc(j)),Nhist(3,CGloc(j)),
            Nhist(4,CGloc(j)),Nhist(5,CGloc(j)),Nhist(2,CGloc(j))/
            (Nhist(2,CGloc(j))+Nhist(4,CGloc(j))));
end
fprintf(fid,'Non CpG sites
index	#A	#C	#G	#T	#-	%%Meth
');
for j=1:length(NonCGloc)
    fprintf(fid,'%d	%d	%d	%d	%d	%d	%.2f
',NonCGinds(j),Nhist(1,NonCGloc(j)),Nhist(2,NonCGloc(j)),Nhist(3,NonCGloc(j)),
            Nhist(4,NonCGloc(j)),Nhist(5,NonCGloc(j)),Nhist(2,NonCGloc(j))/
            (Nhist(2,NonCGloc(j))+Nhist(4,NonCGloc(j))));
end
close(fid);

% Write histogram data

% CpG hist data
Params.OutputHist =
strrep(Params.RefFile,'.fasta','.meth_CpG_hist');

fid = fopen(Params.OutputHist,'w');
if (fid == -1)
    error(["Cannot open output summary file: 
        Params.OutputHist"]) end
fprintf(fid,'%d\t%d\t%d\t%d\t%d\t%d\t%.2f
',CGinds(j),Nhist(1,CGloc(j)),Nhist(2,CGloc(j)),Nhist(3,CGloc(j)),
            Nhist(4,CGloc(j)),Nhist(5,CGloc(j)),Nhist(2,CGloc(j))/
            (Nhist(2,CGloc(j))+Nhist(4,CGloc(j))));
end
close(fid);
for i=1:length(Refs)
    fprintf(fid,'Sample name:	%s
',Refs.Header);
    fprintf(fid,'Count	');
    fprintf(fid,'# of A	# of C	# of G	# of T	# of Sites	');
    for j=1:length(CGinds)
        fprintf(fid,'CpG Index %d	',CGinds(j));
    end
    fprintf(fid,'
');
    [a, b] = sort(struct2array(Nhist_CG_Reads),'descend');
    names = fieldnames(Nhist_CG_Reads);
    for j=1:length(a)
        name = names{b(j)};
        name(name == 'X') = '-';
        fprintf(fid,'%d	',a(j));
        fprintf(fid,'%d	%d	%d	%d	',sum(name=='A'),sum(name=='C'),sum(name=='G'),sum(name=='T'),length(name)-sum(name=='-'));
        for k=1:length(name)
            fprintf(fid,'%c	',name(k));
        end
        fprintf(fid,'
');
    end
end
fclose(fid);
end

function [RefsConv] = ConvertNonCGtoT(Refs)
%% Converts CH nucleotides in sequence to TH
CGinds = findstr(Refs.Sequence,'CG');
Cinds = findstr(Refs.Sequence,'C');
RefsConv.Sequence = Refs.Sequence;
RefsConv.Sequence(setdiff(Cinds,CGinds)) = 'T';
end

function results=numToutput(cpgs,sample_name1)
%% Creates final summary table of all samples
%% Counts # of molecules with all CpGs unmethylated, all-1 CpGs
%% unmethylated, etc.
results=zeros(1,cpgs+3);
file= [sample_name1 '.meth_CpG_hist'];
fid=fopen(file,'r');
data=textscan(fid,'%f %f %f %f %f %f %f %*[^\n]',
    'delimiter','	','HeaderLines', 2, 'CollectOutput',1);
fclose(fid);

% remove irrelevant lines (not all cpgs/contains A or G)
data{1}(data{1}(:,6)<cpgs,:) = [];
data{1}(data{1}(:,2)>0,:) = [];
data{1}(data{1}(:,4)>0,:) = [];

% output 2
results(1) = cpgs;

% Total aligned
results(2) = sum(data{1}(:,1));

for i = 0:cpgs
    idx = find(data{1}(:,5) == (cpgs - i));
    if ~isempty(idx)
        results(3+i) = sum(data{1}(idx,1));
    end
end

end