

Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease

Joshua Moss¹, Judith Magenheim¹, Daniel Neiman¹, Hai Zemmour¹, Netanel Loyfer², Amit Korach³, Yaacov Samet⁴, Myriam Maoz⁵, Henrik Druid^{6,7}, Peter Arner⁸, Keng-Yeh Fu⁹, Endre Kiss⁹, Kirsty L. Spalding^{8,9}, Giora Landesberg¹⁰, Aviad Zick⁵, Albert Grinshpun⁵, AM James Shapiro¹¹, Markus Grompe¹², Avigail Dreazan Wittenberg¹, Benjamin Glaser¹³, Ruth Shemer^{1,*}, Tommy Kaplan^{2,*}, and Yuval Dor^{1,*}

¹Dept. of Developmental Biology and Cancer Research, Institute for Medical Research Israel-Canada, The Hebrew University-Hadassah Medical School, Jerusalem 9112001, Israel

²School of Computer Science and Engineering, The Hebrew University, Jerusalem 9190401, Israel

³Dept of Cardio-Thoracic Surgery, ⁴Vascular Surgery, ⁵Oncology, ¹⁰Anesthesiology and Critical Care Medicine, ¹³Endocrinology and Metabolism Service, Hadassah-Hebrew University Medical Center, Jerusalem 9112001, Israel

⁶Dept. of Oncology-Pathology, ⁹Cell and Molecular Biology, Karolinska Institutet, Stockholm SE 17177, Sweden

⁷Dept of Forensic Medicine, The National Board of Forensic Medicine, Stockholm SE 11120, Sweden

⁸Department of Medicine, Karolinska University Hospital, Karolinska Institutet, Stockholm SE 17176, Sweden

¹¹Department of Surgery and the Clinical Islet Transplant Program, University of Alberta, Edmonton, AB T6G 2R3, Canada

¹²Papé Family Pediatric Research Institute, Oregon Health & Science University, Portland OR 97239

*Corresponding authors: shemer.ru@mail.huji.ac.il, tommy@cs.huji.ac.il, and yuvald@ekmd.huji.ac.il

1 **Abstract**

2 Methylation patterns of circulating cell-free DNA (cfDNA) contain rich information about recent
3 cell death events in the body. Here, we present an approach for unbiased determination of the
4 tissue origins of cfDNA, using a reference methylation atlas of 25 human tissues and cell types.
5 The method is validated using *in silico* simulations as well as *in vitro* mixes of DNA from
6 different tissue sources at known proportions. We show that plasma cfDNA of healthy donors
7 originates from white blood cells (55%), erythrocyte progenitors (30%), vascular endothelial
8 cells (10%) and hepatocytes (1%). Deconvolution of cfDNA from patients reveals tissue
9 contributions that agree with clinical findings in sepsis, islet transplantation, cancer of the
10 colon, lung, breast and prostate, and cancer of unknown primary. We propose a procedure
11 which can be easily adapted to study the cellular contributors to cfDNA in many settings,
12 opening a broad window into healthy and pathologic human tissue dynamics.

13

14 **Introduction**

15 Small fragments of DNA circulate freely in the peripheral blood of healthy and diseased
16 individuals. These cell-free DNA (cfDNA) molecules are thought to originate from dying cells and
17 thus reflect ongoing cell death taking place in the body¹. In recent years, this understanding
18 has led to the emergence of diagnostic tools, which are impacting multiple areas of medicine.
19 Specifically, next generation sequencing of fetal DNA circulating in maternal blood has allowed
20 non-invasive prenatal testing (NIPT) of fetal chromosomal abnormalities^{2,3}; detection of donor-
21 derived DNA in the circulation of organ transplant recipients can be used for early identification
22 of graft rejection^{4,5}; and the evaluation of mutated DNA in circulation can be used to detect,
23 genotype and monitor cancer^{1,6}. These technologies are powerful at identifying genetic
24 anomalies in circulating DNA, yet are not informative when cfDNA does not carry mutations.

25 A key limitation is that sequencing does not reveal the tissue origins of cfDNA, precluding the
26 identification of tissue-specific cell death. The latter is critical in many settings such as
27 neurodegenerative, inflammatory or ischemic diseases, not involving DNA mutations. Even in
28 oncology, it is often important to determine the tissue origin of the tumor in addition to
29 determining its mutational profile, for example in cancers of unknown primary (CUP) and in the
30 setting of early cancer diagnosis⁷. Identification of the tissue origins of cfDNA may also provide
31 insights into collateral tissue damage (e.g. toxicity of drugs in genetically normal tissues), a key
32 element in drug development and monitoring of treatment response.

33 Several approaches have been proposed for tracing the tissue sources of cfDNA, based on
34 tissue-specific epigenetic signatures. Snyder et al. have used information on nucleosome
35 positioning in various tissues to infer the origins of cfDNA, based on the idea that nucleosome-

36 free DNA is more likely to be degraded upon cell death and hence will be under-represented in
37 cfDNA⁸. Ulz et al. used this concept to infer gene expression in the cells contributing to cfDNA⁹.
38 The latter can theoretically indicate not only the tissue origins of cfDNA, but also cellular states
39 at the time of cell death, for example whether cells died and released cfDNA while engaged in
40 the cell division cycle or during quiescence.

41 An alternative approach is based on DNA methylation patterns. Methylation of cytosine
42 adjacent to guanine (CpG sites) is an essential component of cell type-specific gene regulation,
43 and hence is a fundamental mark of cell identity¹⁰. We and others have recently shown that
44 cfDNA molecules from loci carrying tissue-specific methylation can be used to identify cell
45 death in a specific tissue^{11, 12, 13, 14, 15, 16, 17, 18}. Others have taken a genome-wide approach to the
46 problem, and used the plasma methylome to assess the origins of cfDNA. Sun et al inferred the
47 relative contributions of four different tissues, using deconvolution of cfDNA methylation
48 profiles from low-depth whole genome bisulfite sequencing (WGBS)¹⁹. Guo et al demonstrated
49 the potential of cfDNA methylation for detecting cancer as well as identifying its tissue of origin
50 in two cancer types, using a reduced representation bisulfite sequencing (RRBS) approach²⁰.
51 Kang et al and Li et al described CancerLocator²¹ and CancerDetector²², probabilistic
52 approaches for cancer detection based on cfDNA methylation sequencing.

53 While these studies show the potential of DNA methylation in identifying the cellular
54 contributions to cfDNA, it remains to be seen whether cfDNA methylation can be analyzed in an
55 unbiased and comprehensive manner, in settings where it is unclear which cell types contribute
56 to cfDNA and which underlying diseases a patient may have. To address this challenge, we took
57 advantage of the Illumina Infinium methylation array, which allows the simultaneous analysis of
58 the methylation status of >450,000 CpG sites throughout the human genome. Illumina
59 methylation arrays have been previously used in the deconvolution of whole blood methylation
60 profiles to determine the relative proportions of white blood cells in a sample, a crucial step in
61 Epigenome-Wide Association Studies (EWAS)^{23, 24, 25}. However, to date, array deconvolution
62 has been applied only to whole blood samples, where all contributing cells are well-studied
63 types of white blood cells²³.

64 Here we demonstrate that plasma methylation patterns can be used to accurately identify cell
65 type-specific cfDNA in healthy and pathological conditions. We have generated an extensive
66 reference atlas of 25 human tissues and cell types, covering major organs and cells involved in
67 common diseases. As we show, our approach allows for a robust and accurate deconvolution of
68 plasma methylation from as little as 20ml of blood, and using only a small number (4039) of
69 selected genomic loci. We quantify the major cell types contributing to cfDNA in healthy
70 individuals, and demonstrate the origins of cfDNA in islet transplantation, sepsis and cancer.
71 We propose principles for effective plasma methylome deconvolution, including the key
72 importance of a reference atlas consisting of cell type, rather than whole-tissue methylomes,

73 and discuss the potential of global cfDNA methylation analysis as a diagnostic modality for early
74 detection and monitoring of disease.

75

76 **Results**

77 **Development of a DNA methylation atlas**

78 To obtain a comprehensive DNA methylation database of human cell types, we took advantage
79 of datasets which were previously published, either as part of The Cancer Genome Atlas (TCGA)
80 ²⁶ or by individual groups that deposited data in the Gene Expression Omnibus (GEO). In
81 selecting datasets to be included in the database, we used the following criteria: 1) we only
82 used primary tissue sources, which have not been passaged in culture – reasoning that culture
83 may change methylation patterns or alter the cellular composition of a mixed tissue, e.g.
84 enriched for fibroblasts; 2) used the methylomes of healthy human tissues, which are expected
85 to be universally conserved (that is, be nearly identical among cells of the same type, among
86 individuals, throughout life, and be largely retained even in pathologies) ²⁷; 3) excluded tissue
87 methylomes that contained a high proportion of blood-derived DNA, as previously described²⁸;
88 4) merged the methylomes of highly similar tissues (e.g. rectum and colon, stomach and
89 esophagus, cervix and uterus); and 5) preferred the methylomes of specific cell types, rather
90 than whole tissues. We reasoned that since whole tissues are a composite of multiple
91 heterogeneous cell types (e.g. different types of epithelial cells, blood, vasculature and
92 fibroblasts), methylation signatures of minority populations might be difficult to identify, and
93 unique tissue signatures might be masked by the methylome of stroma. Unfortunately, other
94 than isolated blood cell types, the vast majority of publically available methylomes comes from
95 bulk tissues. We therefore generated methylation profiles of key human cell types, which have
96 not been previously published. We have isolated primary human adipocytes, cortical neurons,
97 hepatocytes, lung alveolar cells, pancreatic beta cells, pancreatic acinar cells, pancreatic duct
98 cells, and vascular endothelial cells. As detailed in the Materials and Methods and
99 Supplementary File 1, surgical samples from each tissue were enzymatically dissociated, stained
100 with antibodies against a cell type of interest, and isolated using either flow cytometry (FACS)
101 or magnetic beads (MACS). We then prepared DNA from sorted cells, and obtained the
102 genome-wide methylome using Illumina 450K or EPIC BeadChip array platforms. The result of
103 this effort was a comprehensive human methylome reference atlas, composed of 25 tissues or
104 cell types (Figure 1a).

105

106 **Deconvolution algorithm using cell type-specific CpGs**

107 To analyze novel DNA methylation samples, composed of admixed methylomes from various
108 cell types, we devised a computational deconvolution algorithm. We approximate the plasma

109 cfDNA methylation profile as a linear combination of the methylation profiles of cell types in
110 the reference atlas. According to this model, the relative contributions of different cell types to
111 plasma cfDNA can be determined using non-negative least squares linear regression (NNLS)^{23,}
112 ^{29, 30}. In addition, the relative contributions of cfDNA can be multiplied by the total
113 concentration of cfDNA in plasma to obtain the absolute concentrations of cfDNA originating
114 from each cell type (genome equivalents/ml) (Figure 1b).

115 For accurate inference, we first selected a subset of CpG sites in the genome that are
116 differentially methylated among the cell types and tissues in our atlas. We chose to use only a
117 subset of the methylome for deconvolution based on several considerations. First, almost half
118 of the CpG sites represented in the Illumina arrays show similar methylation patterns across all
119 cells and are therefore uninformative. Second, we found that using a limited subset of CpGs,
120 that are uniquely methylated or unmethylated in a cell type, allows one to detect rare cell types
121 contributing only small amounts of cfDNA and reduces false detection of contributors
122 (Supplementary Figures 1-2). Third, a smaller subset of genomic regions can be the basis of a
123 simpler, capture-based method, increasing the feasibility of routine use.

124 After removing CpG sites with little variance across cell types, we selected, for each tissue or
125 cell type in the atlas, 100 CpG sites uniquely hypermethylated and 100 sites uniquely
126 hypomethylated when compared to other tissues, as well as CpGs located adjacently (within 50
127 bp) to the originally selected set (Methods, Supplementary File 1). This process resulted in
128 ~7,390 CpGs, to which we added 500 CpGs, by iteratively identifying the two most similar cell
129 types in the atlas, and adding the CpG site upon which these two cell types differ the most
130 (Methods, Supplementary File 1). In total, our selection includes ~7,890 CpGs, covering ~4,039
131 genomic regions. We found this set of CpGs to perform favorably on simulated datasets when
132 compared to other selection criteria, including the full set of CpGs (Supplementary Figure 1, 2).

133

134 ***In silico* mix-in simulations**

135 We initially performed *in silico* experiments to assess the performance of the deconvolution
136 approach in determining the relative contributions of various cell types to a methylation profile
137 of DNA from a heterogeneous mixture of cell types. For an exhaustive and realistic assessment,
138 we used whole-blood samples from 18 individuals measured using EPIC Illumina arrays³¹. We
139 then computationally mixed-in methylation profiles of individual samples of cell types and
140 tissues at varying admixtures, reapplied the feature selection and deconvolution algorithms
141 using an atlas from which the individual mixed-in sample was removed. We then compared the
142 actual percentage with the predicted one. We simulated such data for every cell type in the
143 reference methylation atlas, except for white blood cells, at mixing levels varying from 0% to
144 10% (in 1% intervals) across 36-180 replicates (18 independent leukocyte samples, times 2-10
145 replicates for each cell type). As shown in Figure 2a, the deconvolution algorithm performed

146 well for almost all cell types. Most cell types were accurately detected when composing >1% of
147 the mixture, with many cell types detected even below 1% (Supplementary Figure 1).

148 Importantly, almost no non-leukocyte cells (<0.25%) were detected at mixing level of 0%
149 (namely, analysis of pure leukocytes) (Figure 2a, leftmost side of each plot; Supplementary
150 Figure 1). In preliminary analysis we noticed that some confusion might occur between cell
151 types of similar developmental origin (e.g. cervix/uterus, stomach/esophagus, colon/rectum),
152 and therefore have merged these samples in the reference atlas (Methods). Overall the
153 confusion between cell types was minimal, as shown using confusion matrices (Supplementary
154 Figure 3, 4).

155

156 **Cell-type vs whole-tissue reference methylomes**

157 We then tested the importance of using cell type-specific versus tissue-specific or cell-line
158 derived methylomes. A reference methylation atlas containing the methylome of purified
159 hepatocytes outperformed atlases containing either whole liver or HepG2 hepatoma cell line
160 methylomes, with the former leading to overestimation of hepatocyte in the mixture, and the
161 latter leading to a gross underestimation (Figure 2b). Similarly, an atlas containing the
162 methylomes of purified pancreatic cells (acinar, duct and beta cells) was superior in detecting
163 pancreatic DNA within blood, compared to a reference atlas containing the methylome of the
164 whole pancreas, with the latter being ineffective in detecting small contributions (<2%) of
165 pancreatic DNA (Figure 2c). These findings support the feasibility of highly sensitive
166 deconvolution of the plasma methylome, and highlight the importance of using a
167 comprehensive, cell type-specific DNA methylation atlas for sensitive detection of rare
168 contributors to mixed methylomes.

169 ***In vitro* DNA mixing**

170 We then mixed DNA samples from four specific tissues (Liver, Lung, Neurons and Colon, each
171 from a single donor), into leukocytes from a healthy donor, at different proportions varying
172 from 0% to 10%, and reapplied the computational deconvolution analysis (Figure 3,
173 Supplementary File 1). For all samples, our algorithm identified the correct cell type in a specific
174 and sensitive manner (Pearson's r 0.88-0.99, p -value<1e-3 for all mixes). These findings lend
175 further support to the feasibility of deconvolution, but they do not fully address real life issues
176 such as inter-individual variation in methylation.

177 **Tissue origins of Healthy cfDNA**

178 To determine the main contributors to cfDNA in healthy individuals, we collected plasma from
179 multiple healthy donors (n=105). The samples were classified by sex and age (young: 19-30 or

180 old: 67-97; see Supplemental File 1), and cfDNA was pooled accordingly to obtain 250ng cfDNA
181 in each pool.

182 We then obtained methylation profiles of each sample (n=8) using Illumina arrays and
183 performed a deconvolution analysis to estimate the relative contribution of each tissue/cell-
184 type to the cfDNA. The predicted distribution of contributing tissues/cell types was similar
185 among all pools (Figure 4a,b). Additionally, cfDNA from four additional healthy individuals was
186 analyzed and found to be consistent with the findings in the pooled samples (Supplementary
187 File 1). As previously reported³², we found that the main contributors to cfDNA were of
188 hematopoietic origin. On average, 32.0% ($\pm 1.1\%$ mean SD) of cfDNA came from granulocytes,
189 29.7% ($\pm 0.8\%$) from erythrocyte progenitors, 10.5% ($\pm 1.1\%$) from monocytes, and 12.1%
190 ($\pm 0.7\%$) from lymphocytes. The main solid tissue sources of cfDNA were vascular endothelial
191 cells (8.6% $\pm 0.9\%$) and hepatocytes (1.2% $\pm 0.4\%$). The signal from erythrocyte progenitors,
192 endothelial cells and hepatocytes is expected to be present in cfDNA but not in DNA isolated
193 from leukocytes. Indeed, deconvolution of blood cell (leukocyte) methylomes predicted signals
194 from these tissues at much lower levels than in plasma, supporting validity of the algorithm
195 ($p < 1e-10$, Figure 4c).

196 Furthermore, the predicted proportions of monocytes, neutrophils and lymphocytes in whole
197 blood methylomes were in excellent agreement with the actual proportions of these cell types
198 in each individual blood sample, as obtained from a Complete Blood Count (CBC) (Figure 4d).

199 Unexpectedly, deconvolution of the healthy plasma methylome revealed also a signal from
200 neurons, accounting for as much as 2% of cfDNA (Figure 4a,b). The significance of this finding
201 remains to be determined, as it is not consistent with findings using PCR-sequencing of specific
202 brain markers¹¹; we favor the idea that the neuronal signal is an artifact of the assay, perhaps
203 reflecting contribution from a tissue not included in our atlas (see Discussion).

204 While the young and old samples showed similar relative contributions of the different cell
205 types, the plasma of older people showed a significantly higher levels of total cfDNA, as
206 measured in genome equivalents per ml of plasma (Supplementary Figure 5). The similar
207 proportions of cfDNA origins may suggest a slower clearance rate of circulating DNA in older
208 individuals (Figure 4b), rather than an increased rate of cell death in all tissues. Further work is
209 required to define the determinants of cfDNA clearance in difference physiologic and
210 pathologic conditions. In summary, these findings provide the first detailed description of the
211 composition of cfDNA in healthy people.

212

213 **Deconvolution of cfDNA in islet transplant recipients**

214 We analyzed the plasma methylome of patients with long standing type 1 diabetes, 1 hour after
215 receiving a cadaveric pancreatic islet transplant (pool of n=5 recipients). The total concentration

216 of cfDNA in these samples was ~20-fold higher than healthy control levels, suggesting a massive
217 process of cell death shortly after islet transplantation. The deconvolution algorithm identified
218 a large proportion (~20%) of cfDNA as derived from pancreatic origin (from beta, acinar and
219 duct cells, Figure 5a-b), in stark contrast to cfDNA from healthy plasma. These findings strongly
220 support the validity of our deconvolution procedure. Strikingly, we observed that most of the
221 increase in cfDNA levels in islet transplant recipients was of an immune cell origin
222 (granulocytes, monocytes and lymphocytes). This finding suggests an acute immune response
223 to the infusion of islets into recipient blood, or alternatively a response to the procedure itself
224 and/or pre-transplant immune suppression treatment, resulting in massive immune cell death
225 (Figure 5b). Follow up studies will attempt to distinguish between these possibilities.

226 To examine the dynamics of cfDNA of pancreatic origin, we determined the plasma methylome
227 of 3 individual recipients before (<1 day), 1 hour after, and 2 hours after transplantation. As
228 expected, the algorithm identified no pancreas cfDNA before islet transplantation, a large
229 increase immediately after transplantation, and a subsequent decrease in levels of pancreatic
230 cfDNA (Figure 5c). Interestingly, cfDNA originating from immune cells as inferred by
231 deconvolution showed a different dynamics, likely reflecting the response of the innate
232 immune system to the transplantation (Supplementary Figure 6). In addition, we used a
233 previously described targeted bisulfite-sequencing approach to quantify the amount of
234 unmethylated CpGs at a haplotype block located over the insulin promoter¹¹. We observed a
235 high correlation ($r=0.995$, $p \leq 2.6e-8$) between the amount of beta cell cfDNA estimated by
236 deconvolution and by targeted PCR-based method, further supporting validity of the
237 deconvolution algorithm (Figure 5d). Finally, we tested the deconvolution algorithm using a
238 reference matrix containing either whole-tissue or cell type-specific methylomes. Consistently
239 with results from deconvolution of *in silico* mixes (Figure 2b-c), a reference matrix containing
240 cell type-specific methylomes showed higher sensitivity compared with an atlas containing a
241 whole-tissue methylome, which failed to identify pancreatic cfDNA in one of the three
242 recipients (Figure 5e).

243 **The origin of cfDNA in sepsis**

244 An increase in total cfDNA levels in septic patients has been previously documented, and even
245 shown to have a prognostic value^{33,34}. However, it is unclear which cell types are contributing
246 to the elevated cfDNA. We analyzed the cfDNA methylation profile of 14 samples from patients
247 with sepsis. In most patients (13/14) the main contributors to the increase in cfDNA were
248 leukocytes (mainly granulocytes), elevated > 20-fold relative to healthy levels (Figure 6a,b). In
249 some cases, varying amounts of hepatocyte cfDNA were detected (patients SEP-026, SEP-017,
250 SEP-016). Importantly, the levels of hepatocyte cfDNA were strongly correlated (Pearson's
251 $r=0.931$, $p < 5e-7$) with levels of Alanine Aminotransferase (ALT) in circulation, a marker of
252 hepatocyte damage (Figure 6c).

253 **Identifying tumor origin by cfDNA methylation**

254 We deconvoluted the cfDNA methylation profiles of patients with metastatic colon cancer
255 (n=4), lung cancer (n=4) and breast cancer (n=3) (Supplementary File 1). All had elevated
256 concentration of cfDNA compared to healthy individuals (>20 fold increase). The tissue of origin
257 was the strongest signal (most genome equivalents/ml) in the majority of cases (8/11 total, 3/4
258 colon, 2/4 lung, 3/3 breast, Figure 7a-c). These findings indicate the ability of the deconvolution
259 algorithm to correctly detect cfDNA from advanced cancer, despite potential changes to the
260 epigenome of cancer cells.

261 To assess the accuracy of cancer detection using deconvolution, we performed a mixing
262 experiment, where plasma from a patient with colon cancer was mixed with plasma of healthy
263 donors at different proportions (Supplementary File 1), and the methylome of the resulting
264 mixture was deconvoluted. The algorithm correctly identified the presence of colon DNA in the
265 mixes, in the correct proportion, down to 3% (33-fold dilution of the original cancer plasma
266 sample, $r=0.92, p<1.2e-3$) (Figure 7d).

267 To further assess the performance of the deconvolution algorithm, we applied it to recently
268 published dataset where plasma samples of prostate cancer patients were assessed using
269 Illumina 450K arrays, before and after treatment with Abiraterone Acetate, including patients
270 that were responsive or not responsive to therapy³⁵. As shown in Figure 7e, the algorithm
271 detected prostate DNA in most patients (as compared to a lack of signal in all healthy controls).
272 Strikingly, the deconvolution algorithm also detected a sharp decline in the levels of prostate
273 cfDNA in treatment-sensitive patients ($p<0.019$, paired t-test) but not in treatment-resistant
274 patients ($p<0.909$, paired t-test), further supporting validity of the method.

275 Finally, we tested whether an unbiased deconvolution approach could be useful in identifying a
276 cancer tissue of origin, even in the absence of an identifiable primary tumor. To this end, we
277 analyzed the plasma cfDNA of four patients with Cancer of Unknown Primary (CUP). All patients
278 had metastatic disease with no clear pathological identification of the primary source of cancer
279 (detailed in Supplementary File 1). In each case the suspected origin of the tumor, based on
280 clinical history and pathology reports, showed a strong signal in the deconvolution analysis
281 (Figure 7f). Patient 3, for example, presented with metastases in bones and lungs without
282 identifiable histopathology. Six years earlier, the patient had a local bladder carcinoma that was
283 treated and removed. Deconvolution analysis of plasma cfDNA identified a significant
284 contribution by bladder cells (>5,000 genome equiv./ml), suggesting that the current disease
285 originated from previously disseminated bladder cancer cells (Figure 7f).

286 These findings indicate that cfDNA methylation deconvolution can be the basis of a non-
287 invasive approach to identify the origin of cancer, similar to what has been described using
288 biopsy material³⁶.

289 Discussion

290 In many diseases, DNA from dying cells is released into the bloodstream. Tools that can identify
291 the source tissue of this DNA could be instrumental in identifying and locating disease. DNA
292 methylation reflects cell identity, and is therefore an ideal marker of the origin of DNA in
293 circulation. In this study, we present a method to decipher the cellular origins of cfDNA by
294 deconvoluting genome-wide methylation profiles, and use it to determine which cells release
295 DNA into blood in several clinically relevant situations.

296 When assessing the tissues that contribute to human cfDNA, we first made an effort to define
297 the healthy baseline. Previous studies used plasma from female patients who had received
298 bone marrow transplants from male donors, and concluded that most cfDNA is derived from
299 cells of hematopoietic origin³²; however, the contribution of individual blood cell types was not
300 assessed, nor was the contribution of non-blood cells. More recently, Guo et al analyzed the
301 plasma methylome of healthy and cancer patients using WGBS, and reported the contribution
302 of white blood cells (without subtypes) as well as nine solid tissues and two tumor types²⁰. Our
303 deconvolution assay revealed the specific contributors to healthy plasma, namely granulocytes,
304 monocytes, lymphocytes and erythrocyte progenitors. The latter is consistent with a recent
305 report that used specific erythroid lineage methylation markers to identify erythroid lineage-
306 derived cfDNA¹⁵. Note that unlike the other sources of cfDNA, in this case the process reflected
307 by cfDNA might be cell birth (the generation of enucleated red blood cells) rather than cell
308 death. Refinement of the methylome atlas will likely result in further refinement of cfDNA
309 interpretation, even retrospectively on the samples reported here. For example, it should be
310 possible to determine the relative contribution of neutrophils and other cell types to the
311 granulocyte cfDNA pool, and of circulating monocytes and tissue resident macrophages to the
312 monocyte cfDNA pool.

313 Beyond blood cells, we found that ~10% of cfDNA in healthy individuals is derived from vascular
314 endothelial cells (a finding made possible by the generation of a vascular endothelial cell
315 methylome reference), and that ~1% of cfDNA is derived from hepatocytes, which is consistent
316 with our recent observation of hepatocyte cfDNA in healthy plasma using 3 targeted
317 hepatocyte markers¹⁸. The cfDNA signal from the vasculature and the liver reflects the sum of
318 multiple parameters: total cell number in these organs, the degree of baseline turnover, and
319 the fact that cfDNA from these tissues is apparently cleared via blood. The absence of a cfDNA
320 signal from other tissues in the body, known to have a high turnover rate, likely reflects
321 alternative clearance routes: for example, dying intestinal epithelial cells under healthy
322 conditions likely shed cfDNA into the lumen of the intestine, rather than to blood. Similar
323 considerations apply to the lung, kidney and skin. The algorithm also detected a neuronal-
324 derived signal comprising as much as ~2% of the healthy plasma methylome. While this finding
325 may reflect a baseline turnover of central or peripheral neurons³⁷, we cannot rule out the

326 possibility that it is an artifact of the deconvolution algorithm, due to a partial and imperfect
327 reference atlas. One argument in favor of the latter interpretation is that our directed PCR-
328 sequencing assays using brain-specific methylation markers show only a negligible neuronal
329 signal in healthy individuals (~0.1%), while positive controls with brain damage do show a clear
330 signal (manuscript in preparation and ¹¹). More experiments are needed to determine the
331 actual contribution of neuronal DNA to the healthy cfDNA.

332 We also performed a preliminary analysis of cfDNA composition as a function of age, using
333 pools of samples from healthy individuals aged 75 and above and between the ages of 19 and
334 30. Two striking findings emerge from the analysis of these samples: first, the total
335 concentration of cfDNA in aged individuals is about twice that of people in their 3rd decade of
336 life; second, deconvolution revealed a distribution of sources that is highly similar between
337 aged and young individuals. We propose that this similarity reflects a decrease in the rate of
338 cfDNA clearance in old age, rather than a concordant increase in cell death within all tissues.
339 Additional studies are required to definitively interpret the biology of the circulating
340 methylome in old age.

341 The application of cfDNA deconvolution to selected pathologies provided further support as to
342 the validity of the approach. This included the identification of pancreas cfDNA in islet
343 transplant recipients (but not in healthy controls) and the identification of elevated hepatocyte
344 cfDNA in patients with sepsis, which correlated with an independent circulating liver marker. In
345 both transplantation and sepsis we found that elevated cfDNA was mostly derived from
346 immune cells. Both scenarios likely involve strong immune reactions and the increase in
347 leukocyte-derived cfDNA may be derived from cells that died during cell division or as part of an
348 immune response. We also demonstrated that deconvolution can identify cfDNA from a
349 cancer's tissue of origin, even in advanced tumors presumably presenting with epigenomic
350 instability. While more studies with plasma samples from cancer patients are needed, in
351 particular from early stage diseases, our findings from multiple type of cancer (colon, lung,
352 breast and prostate) are highly encouraging in this respect. Lastly, using plasma samples from
353 patients with cancer of unknown primary, we showed that the tissue source of metastases can
354 be identified by analysis of cfDNA methylation, even in cases where the primary tissue of the
355 cancer is missing and unclear. Whilst most current approaches aim to monitor cancer via
356 identification of mutations in cfDNA, we propose that combining such an analysis with cfDNA
357 methylation deconvolution may eventually allow for early and unbiased diagnosis of cancer and
358 its location ⁷.

359 Our work provides a proof of concept for the utility of plasma methylome deconvolution in
360 studying human tissue dynamics in health and disease, adding insights beyond those of recent
361 reports in this emerging field ^{19, 20, 21, 22}. Furthermore, our approach can easily be adapted to
362 determine the cellular contributors to cfDNA in virtually any setting in which there is a question

363 regarding the composition of cfDNA. We selected to work with Illumina arrays as a platform for
364 both the tissue reference atlas and the plasma methylome assay. This platform has multiple
365 advantages, perhaps most importantly the vast amount of public data available that can be
366 used to construct a tissue methylome atlas. Additionally, it is the most affordable method
367 available for obtaining high-resolution genome-wide methylation profiles and is simple to
368 perform and analyze as well as scalable. However, arrays have also important limitations: they
369 cover only a small fraction of the genome-wide methylome; they report on the methylation
370 status of individual CpG sites, missing the information embedded in the status of methylation
371 haplotype blocks^{11, 20}; they suffer from batch effects; they require a relatively large amount of
372 DNA (100ng cfDNA, shown here to be sufficient for deconvolution, requires about 40ml of
373 blood); and their sensitivity (ability to detect a small fraction of molecules with a different
374 methylation status in a mixture) is limited compared with sequencing of individual molecules.
375 We believe that in the long run, for applications requiring maximal sensitivity and affordability
376 (such as for early detection of cancer in asymptomatic individuals), a cfDNA methylation
377 deconvolution approach based on deep sequencing of a collection of informative CpG blocks,
378 possibly following capture of key loci from plasma, using a sequencing-based comprehensive
379 atlas, will likely be the preferred approach.

380 Nonetheless, our study does provide some important insights into design principles of effective
381 plasma methylome technology, which are general and would hold for other platforms including
382 massively parallel bisulfite sequencing or nanopore sequencing. These include: 1) The key
383 importance of generating a comprehensive methylation atlas composed of individual cell types
384 (purified from fresh tissue), rather than whole tissues. The inclusion of cell-type specific
385 methylomes allows the identification of important tissue contributions to cfDNA, including cell
386 types that comprise a small minority of their host tissue (e.g. beta cells in the pancreas), and
387 cell types that are present within multiple organs and hence might be masked (e.g. vascular
388 endothelial cells). 2) Not all CpG sites contribute to accurate deconvolution; in fact,
389 deconvolution based on a defined subset of informative sites performs better than an approach
390 taking into account all sites, including those that are not differentially methylated between
391 tissues and hence contribute mostly noise; 3) A specific subset of ~4000 CpG sites that is
392 informative enough for accurate estimation of cfDNA contributors. We propose that a capture-
393 based approach, applying deep bisulfite sequencing to probe multiple neighboring CpGs from
394 the same molecule around selected loci, would offer deconvolution at a much greater
395 resolution, and potentially using lower amount of DNA.

396 In summary, we report a method for interpreting the circulating methylome using a reference
397 methylome atlas, allowing inference of tissue origins of cfDNA in a specific and sensitive
398 manner. We propose that deconvolution of the plasma methylome is a powerful tool for
399 studying healthy human tissue dynamics and for identifying and monitoring a wide range of
400 pathologies.

401

402 **Methods**

403 **Reference matrix**

404 All DNA methylation profiles were determined either on the Illumina Infinium Human
405 Methylation 450K or EPIC BeadChip arrays. DNA methylation data for white blood cells
406 (neutrophils, monocytes, B-cells, CD4+ T-cells, CD8+ T-cells, NK-cells, n=6 each) were
407 downloaded from GSE110555 (EPIC)
408 [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE110555>] ³⁸. Data for erythrocyte
409 progenitors (n=5) were downloaded from GSE63409
410 [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63409>] (450K) ³⁹, and data for left
411 atrium (n=4) were downloaded from GSE62727
412 [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62727>] (450K) ⁴⁰. Data for bladder
413 (n=19), breast (n=98), cervix (n=3), colon (n=38), esophagus (n=16), oral cavity (n=34), kidney
414 (n=160), prostate (n=50), rectum (n=7), stomach (n=2), thyroid (n=56) and uterus (n=34) were
415 downloaded from TCGA ²⁶. DNA methylation data for adipocytes (n=3, 450K), hepatocytes (n=3,
416 450K and EPIC), alveolar lung cells (n=3, EPIC), neurons (n=3, 450K and EPIC), vascular
417 endothelial cells (n=2, EPIC) pancreatic acinar cells (n=3, 450K and EPIC), duct cells (n=3, 450K
418 and EPIC), beta cells (n=4, 450K and EPIC), colon epithelial cells (n=3, EPIC) were generated in
419 house and are available from the corresponding authors upon reasonable request. Detailed
420 sample information is available in Supplementary File 1.

421 **Cell isolation**

422 Cancer-free primary human tissue was obtained from consenting donors, dissociated to single
423 cells, sorted using cell type-specific antibodies and lysed to obtain genomic DNA, from which
424 250ng were applied to an Illumina methylation array. Adipocytes (n=3) were isolated from fat
425 tissue according to the collagenase procedure of Rodbell ⁴¹. In brief, tissue was cut into ≈20 mg
426 pieces and incubated (10 g tissue/25ml buffer) in Krebs-Ringer phosphate (KRP buffer, pH 7.4)
427 containing 4% bovine serum albumin (BSA) and 0.5 mg/ml of collagenase type 1 for 45 min at
428 37°C in a shaking water bath. The isolated adipocytes were collected through a 250µm nylon
429 mesh filter and were washed 3-4 times with 1% KRP-BSA washing buffer. The stromal vascular
430 fraction (SVF) in the washing buffer was collected by 500 x g centrifuge at 4°C for 10 min. Cells
431 were then homogenized in lysis buffer (0.32 M sucrose, 25 mM KCl, 5 mM MgCl₂, 0.1 mM
432 EDTA, 10 mM Tris-HCl pH 7.5, 0.005 % NP-40, 1 mM DTT) transferred to ultracentrifuge tubes,
433 layered onto a sucrose cushion solution (1.8 M sucrose, 25 mM KCl, 5 mM MgCl₂, 0.1 mM
434 EDTA, 10 mM Tris-HCl pH 7.5, 1 mM DTT) and centrifuged at 106,750 x g for 1hr at 4°C to
435 isolate nuclei. Cortical neurons (n=1) were isolated from human occipital cortex by sucrose-
436 gradient centrifugation and labeled with Alexa Fluor 647 conjugate of neuron-specific

437 monoclonal anti-NeuN antibody (A-60) (Millipore, 1:1,000). NeuN-positive and negative nuclei
438 were sorted by FACS and DNA was extracted^{42, 43}. Hepatocytes (n=2) were isolated as
439 previously described⁴⁴. Pancreatic acinar cells and duct cells (n=3) were obtained from
440 cadaveric donors as described⁴⁵. Pancreatic beta cells (n=4) were isolated from cadaveric islets
441 as previously described⁴⁶. Vascular endothelial cells were isolated from the saphenous vein,
442 surgically excised due to chronic insufficiency. Dissociated endothelial cells were captured using
443 mouse anti-human CD105 magnetic beads (cat #130-051-201, Miltenyi, 1:5) (n= 3 donors,
444 pooled to 2 samples, one containing material from two donors and one containing material
445 from one sample). Distal lung tissue (n= 3 donors, 3 samples) was dissociated using an
446 adaptation of previous protocols^{47, 48, 49, 50}. Briefly, alveolar epithelial cells were enriched
447 using mouse anti-human CD105 magnetic beads for depletion of endothelial cells (cat #130-
448 051-201, Miltenyi, 1:5) and subsequently mouse anti-human Epcam (CD326) magnetic beads to
449 capture epithelial cells (cat #130-061-101, Miltenyi, 1:4) or by FACS sorting using the following
450 antibodies: CD45 eFluor 450 (cat #48-9459-41), CD31 eFluor 450 (cat #48-0319-42) and CD235a
451 eFluor 450 (cat #48-9987-42) (all from eBioscience, 1:20) and CD326-APC (cat #130-113-260,
452 Miltenyi, 1:50). Colon epithelial cells were dissociated using an adaptation of a published
453 protocol⁵¹ and were sorted by FACS using CD45 eFluor 450 (cat #48-9459-41), CD31 eFluor 450
454 (cat #48-0319-42) and CD235a eFluor 450 (cat #48-9987-42, eBioscience, 1:20) (for blood and
455 endothelial cell lineage depletion), and CD326-APC (Miltenyi, 1:50, cat #130-113-260)
456 antibodies. FACS gating strategies are shown in Supplementary Figure 8.

457 **Blood samples**

458 Donors were consented and whole blood (usually 20 ml) was drawn, collected into an EDTA
459 tube, and spun quickly to separate plasma, which was stored at -20°C until isolation of cfDNA.

460 **Human research participants**

461 Tissue and plasma samples were obtained in accordance with the principles endorsed by the
462 Declaration of Helsinki and written informed consent was obtained from all subjects. Protocols
463 were approved by the Institutional review boards of Hadassah-Hebrew University Medical
464 Center, The University of Alberta, Karolinska Institute and Oregon Health & Science University.

465 **Sample pooling**

466 Pooled DNA samples were obtained by mixing DNA from several individuals. DNA was extracted
467 from 8ml of plasma and samples were added until 250ng reached (7-19 samples per pool). No
468 individual contributed more than 2 times as much DNA to a pool than another individual.

469 **DNA extraction**

470 250 ng was collected from each sample, except where otherwise specified. DNA concentration
471 was measured with Qubit. cfDNA extraction from plasma was performed with the

472 QIA Symphony liquid handling robot. cfDNA was treated with the Illumina Infinium FFPE
473 restoration kit and hybridized to the Illumina 450K or EPIC arrays.

474 For adipocytes, we used a modified protocol from Miller et al.⁵². 500 μ l DNA lysis buffer (200
475 mM NaCl, 5 mM EDTA, 100 mM Tris-HCl pH 8, 1 % SDS) and 6 μ l Proteinase K (20 mg/ml) were
476 added to the collected nuclei and incubated at 55°C overnight. RNase cocktail (Ambion) was
477 then added and incubated at 55°C for 1hr. Half of the existing volume of 5 M NaCl solution was
478 added and the mixture agitated for 15 s. The solution was spun down at 16,000 x g for 3 min.
479 The supernatant containing DNA was transferred to a new Eppendorf tube. 3 times of the
480 existing volume of 95 % ethanol was added and the tube was inverted several times to
481 precipitate adipocytes or SVF DNA. The DNA precipitate was washed three times in 75 %
482 ethanol and air-dried at 55°C for 2 hr. 500 μ l DNase/RNase-free water was used to suspend
483 the dried DNA. All DNA samples were quantified and purity-checked by UV spectroscopy
484 (Nanodrop).

485 Neuronal DNA was extracted by adding 500 μ l DNA lysis buffer (100 mM Tris-HCl [pH 8.0], 200
486 mM NaCl, 1% SDS, and 5 mM EDTA) and 6 μ l Proteinase K (20 mg/ml, Invitrogen) to the sorted
487 nuclei and incubated overnight at 65°C. Following overnight incubation, an RNase cocktail was
488 added (3 μ l, Ambion) and incubated at 65°C for 45 min. Half of the existing volume of 5 M NaCl
489 solution was added and the mixture agitated for 15 s and centrifuged at 16000 x g for 3 min.
490 The supernatant containing the DNA was transferred to a 12 ml glass vial. Three times the
491 volume of 95% ethanol was added to the glass vial and inverted several times to precipitate the
492 DNA. The DNA precipitate was washed in DNA-washing solution (70% [v/v] ethanol and 0.5 M
493 NaCl) for 15 min for three times and transferred to 200 μ l DNase-/ RNase-free water (Gibco/Life
494 Technologies) and air-dried at 65°C overnight. Finally, the DNA was dissolved in 500 μ l TE
495 buffer (pH 8.0) (10 mM Tris-HCl [pH 8.0] and 1 mM EDTA). The DNA was quantified and its
496 purity was verified using a NanoDrop 2000 spectrophotometer (ThermoScientific).

497 **Data processing**

498 Methylation array data were processed with the minfi package in R. For each sample analyzed
499 on the Illumina Methylation array, CpG sites were filtered out if they were represented by less
500 than 3 beads on the array, if the detection p-value (representing total fluorescence of the
501 relevant probes) was greater than 0.01, or if they mapped to a sex chromosome. Background
502 correction and normalization were performed with the preprocessIllumina function, which
503 removes background calculated based on internal control probes and normalizes all samples to
504 a predetermined control sample.

505 **Comparison of EPIC and 450K platforms**

506 As the reference database included samples analyzed with two highly similar yet not identical
507 platforms, the Illumina 450K array and the Illumina EPIC array, we looked to identify and

508 remove sites with low reproducibility between the platforms. To this end, we collected data
509 from samples analyzed on both platforms: 15 samples from GSE86833
510 [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE86833>] ⁵³, 12 samples from
511 GSE92580 [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92580>] ⁵⁴, and one
512 sample from our generated dataset (hepatocytes). For each overlapping CpG, we then
513 calculated the median absolute error (MAE) between the 450K samples and the corresponding
514 EPIC samples, and removed 37,747 CpGs with an MAE>0.05.

515 **CpG feature selection**

516 First, CpGs whose variance across the entire methylation atlas was below 0.1%, or CpGs with
517 missing values were excluded. We then selected the $K=100$ most specific hyper-methylated
518 CpGs for each cell type. Let us denote the methylation matrix \mathbf{X} , composed of N rows (CpGs) by
519 d columns (cell types). We then divided each row (the methylation pattern of one CpG over all
520 cell types) by its sum $X'_i = \frac{X_i}{\sum_j^d X_{i,j}}$. For each cell type j , we identified the top K hyper-methylated
521 CpGs with the highest $X'_{i,j}$ values. To identify uniquely hypo-methylated CpGs, we performed a
522 similar process for the reversed methylation matrix ($\mathbf{1-X}$). Finally, for each cell type we included
523 both the top K hypermethylated and the top K unmethylated CpGs in the reference matrix
524 (Supplementary File 1). To this set of CpGs, we added neighboring CpGs, up to 150bp.

525 Pairwise-specific CpGs were iteratively selected as follows: Given the current set S of CpGs, we
526 projected the reference atlas on those coordinates, and calculated the Euclidean distances
527 between pairs of cell types. Once the closest pair of cell types was identified, we selected the
528 CpG site where they differ the most, and added it into the set S . This process was iteratively
529 repeated, focusing on the most confusing pair of cell types in each iteration.

530 **Deconvolution**

531 To calculate the relative contribution of each cell type to a given sample, we performed non-
532 negative least squares, as implemented in the nnls package in MATLAB (an efficient alternative
533 to lsqnonneg). Given a matrix \mathbf{X} of reference methylation values with N CpGs and d cell types,
534 and a vector \mathbf{Y} of methylation values of length N , we identified non-negative coefficients β , by
535 solving $\operatorname{argmin}_{\beta} \|\mathbf{X}\beta - \mathbf{Y}\|_2$, subject to $\beta \geq 0$. We then adjusted the resulting β to have a sum
536 of 1, where for each β_j we defined $\beta'_j = \frac{\beta_j}{\sum_j^d \beta_j}$. To obtain absolute levels of cfDNA (genome
537 equivalent/ml) per cell type, we multiplied the resulting β'_j by the total concentration of cfDNA
538 present in the sample, as measured by Qubit. It was assumed that the mass of a haploid
539 genome is 3.3 pg and as such, the concentration of cfDNA could be converted from units of
540 ng/ml to haploid genome equivalents/ml by multiplying by a factor of 303. To estimate
541 deconvolution error rates, we used a bootstrap approach, where we also analyzed the
542 observation vector (\mathbf{Y}) using $N=100$ different instances of the methylation atlas. Following

543 Houseman et al ³⁰, and due to the limited number of replicate per cell type, we used a
544 parametric approach, where the original replicates for each tissue were used to estimate the
545 mean CpG methylation and its standard deviation. We then generated $N=100$ new methylation
546 atlases (\mathbf{X}') by sampling from Normal distributions centered at these values for each CpG/tissue.
547 Finally, we deconvoluted the observation vector (\mathbf{Y}) using each atlas, and estimated the
548 empirical standard deviation of the admixture parameters across atlases (\mathbf{X}'). The same
549 approach was used to estimate the variation for contribution of specific cell types, including
550 DNA mixes (Fig 3a-d), pancreas (Fig 5c-e), hepatocytes (Fig 6c), and plasma mixes (Fig 7d).

551 **Simulations**

552 We analyzed 18 leukocyte samples (whole-blood) with Illumina methylation EPIC arrays. For
553 each cell type, we mixed in every available replicate with each leukocyte sample in ratios of 0 to
554 100, 0.1 to 99.9, 1 to 99, 2 to 98, etc. up to 10 to 90. For every combination of leukocytes and
555 cell type replicate, we updated the reference atlas by excluding the mixed-in sample and then
556 re-computing the average methylome for that cell type using all other replicates. We then re-
557 applied the feature selection process (using the new atlas), applied the deconvolution
558 algorithm, and estimated the admixture coefficients for all cell types. This procedure ensures
559 that the training set is completely separated from the test set. Finally, we calculated for each
560 cell type, at each admixture ratio, the average predicted proportion over all replicates, its
561 median, and the range between the 1st and 3rd quartiles.

562 **Reproducibility**

563 We assayed three cfDNA samples in duplicate (Supplementary Figure 7a-c). The predicted
564 proportions of cell types contributing to the samples were highly correlated ($r > 0.99$).
565 Furthermore, as the amount of cfDNA available is often limited, we also evaluated the
566 possibility of using less than the 250 ng cfDNA (as recommended by Illumina for analysis with
567 methylation array). The results were reproducible with as little as 50 ng of cfDNA ($r > 0.9$)
568 (Supplementary Figure 7a-d).

569 **Code Availability**

570 A standalone program for deconvolution of array methylome is available at
571 https://github.com/nloyfer/meth_atlas or from the corresponding authors.

572 **Data Availability**

573 The datasets generated and analyzed during this study are summarized in Supplementary File 1
574 and available from the corresponding authors on reasonable request. A reporting summary for
575 this Article is available as a Supplementary Information file.

576 **Acknowledgements**

577 This research was performed using grants from The Ernest and Bonnie Beutler Research
578 Program of Excellence in Genomic Medicine, the Juvenile Diabetes Research Foundation, The
579 Alex U Soyka pancreatic cancer fund (to YD), the NIDDK-supported Human Islet Research
580 Network (HIRN) (RRID:SCR_014393; <https://hirnetwork.org>; UC4 DK104216-01), the Kahn
581 Foundation (to B.G, R.S and Y.D) and GRAIL (to B.G, R.S, T.K. and Y.D). T.K. was supported by the
582 Israeli Centers of Excellence (I-CORE) for Gene Regulation in Complex Human Disease (no.
583 41/11) and Chromatin and RNA in Gene Regulation (no. 1796/12), and by Israel Science
584 Foundation grant (no. 913/15). N.L. was supported by a fellowship from the Leibnitz Center for
585 Research in Computer Science. K.L.S. and P.A. acknowledge the Swedish Research Foundation,
586 Novo Nordisk Foundation and the Diabetes Research Program at Karolinska Institutet. K.L.S.
587 also acknowledges the KI-AZ ICMC and the Vallee Foundation. Illumina array experiments were
588 performed with the help of the Roswell Park Cancer Institute.

589

590 **Author Contributions**

591 Conceived and designed the methods: J. Moss, B.G., R.S., T.K., and Y.D.; Data collection and
592 contribution: J. Moss, J. Magenheimer, D.N., H.Z., A.K., Y.S., M.M., H.D., P.A., K.-Y. F., E.K., K.L.S.,
593 G.L., A.Z., A.G., AM J.S., M.G., A.D.W., B.G., and R.S.; Analyzed the data: J. Moss, N.L., and T.K.;
594 Wrote the paper: J. Moss, B.G., R.S., T.K. and Y.D.

595

596 **Competing interests**

597 The Authors declare no Competing Non-Financial Interests but the following Competing
598 Financial Interests: JM, RS, BG, TK and YD are inventors on a patent entitled “CELL FREE DNA
599 DECONVOLUTION AND USE THEREOF” (US provisional application No. 62/661,179).

References

1. Wan JC, *et al.* Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat Rev Cancer* **17**, 223-238 (2017).
2. Fan HC, Gu W, Wang J, Blumenfeld YJ, El-Sayed YY, Quake SR. Non-invasive prenatal measurement of the fetal genome. *Nature* **487**, 320-324 (2012).
3. Lo YM, *et al.* Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci Transl Med* **2**, 61ra91 (2010).
4. De Vlaminck I, *et al.* Circulating cell-free DNA enables noninvasive diagnosis of heart transplant rejection. *Sci Transl Med* **6**, 241ra277 (2014).
5. De Vlaminck I, *et al.* Noninvasive monitoring of infection and rejection after lung transplantation. *Proc Natl Acad Sci U S A*, (2015).
6. Abbosh C, *et al.* Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* **545**, 446-451 (2017).

7. Aravanis AM, Lee M, Klausner RD. Next-Generation Sequencing of Circulating Tumor DNA for Early Cancer Detection. *Cell* **168**, 571-574 (2017).
8. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164**, 57-68 (2016).
9. Ulz P, *et al.* Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nat Genet* **48**, 1273-1278 (2016).
10. Dor Y, Cedar H. Principles of DNA methylation and their implications for biology and medicine. *Lancet*, (2018).
11. Lehmann-Werman R, *et al.* Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc Natl Acad Sci U S A* **113**, E1826-1834 (2016).
12. Gala-Lopez BL, *et al.* Beta Cell Death by Cell-Free DNA and Outcome after Clinical Islet Transplantation. *Transplantation*, (2018).
13. Akirav EM, *et al.* Detection of beta cell death in diabetes using differentially methylated circulating DNA. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 19018-19023 (2011).
14. Lebastchi J, *et al.* Immune Therapy and beta-Cell Death in Type 1 Diabetes. *Diabetes* **62**, 1676-1680 (2013).
15. Lam WKJ, *et al.* DNA of Erythroid Origin Is Present in Human Plasma and Informs the Types of Anemia. *Clinical chemistry* **63**, 1614-1623 (2017).
16. Gai W, *et al.* Liver- and Colon-Specific DNA Methylation Markers in Plasma for Investigation of Colorectal Cancers With or Without Liver Metastases. *Clinical chemistry*, (2018).
17. Zemmour H, *et al.* Non-invasive detection of human cardiomyocyte death using methylation patterns of circulating DNA. *Nature communications* **9**, 1443 (2018).
18. Lehmann-Werman R, *et al.* Monitoring liver damage using hepatocyte-specific methylation markers in cell-free circulating DNA. *JCI Insight* **3**, (2018).
19. Sun K, *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc Natl Acad Sci U S A* **112**, E5503-5512 (2015).
20. Guo S, Diep D, Plongthongkum N, Fung HL, Zhang K, Zhang K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat Genet* **49**, 635-642 (2017).
21. Kang S, *et al.* CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome Biol* **18**, 53 (2017).
22. Li W, *et al.* CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. *Nucleic acids research*, (2018).
23. Accomando WP, Wiencke JK, Houseman EA, Nelson HH, Kelsey KT. Quantitative reconstruction of leukocyte subsets using DNA methylation. *Genome Biol* **15**, R50 (2014).
24. Titus AJ, Gallimore RM, Salas LA, Christensen BC. Cell-type deconvolution from DNA methylation: a review of recent applications. *Human molecular genetics* **26**, R216-R224 (2017).

25. Kaushal A, *et al.* Comparison of different cell type correction methods for genome-scale epigenetics studies. *BMC Bioinformatics* **18**, 216 (2017).
26. Weisenberger DJ. Characterizing DNA methylation alterations from The Cancer Genome Atlas. *The Journal of clinical investigation* **124**, 17-23 (2014).
27. Bergman Y, Cedar H. DNA methylation dynamics in health and disease. *Nat Struct Mol Biol* **20**, 274-281 (2013).
28. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. *Nature communications* **6**, 8971 (2015).
29. Houseman EA, *et al.* Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics* **9**, 365 (2008).
30. Houseman EA, *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
31. Salas LA, *et al.* An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. *Genome Biol* **19**, 64 (2018).
32. Lui YY, Chik KW, Chiu RW, Ho CY, Lam CW, Lo YM. Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. *Clinical chemistry* **48**, 421-427 (2002).
33. Rhodes A, Wort SJ, Thomas H, Collinson P, Bennett ED. Plasma DNA concentration as a predictor of mortality and sepsis in critically ill patients. *Critical care* **10**, R60 (2006).
34. Dwivedi DJ, *et al.* Prognostic utility and characterization of cell-free DNA in patients with severe sepsis. *Critical care* **16**, R151 (2012).
35. Gordevicius J, *et al.* Cell-Free DNA Modification Dynamics in Abiraterone Acetate-Treated Prostate Cancer Patients. *Clin Cancer Res* **24**, 3317-3324 (2018).
36. Moran S, *et al.* Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *Lancet Oncol* **17**, 1386-1395 (2016).
37. Rao M, Gershon MD. Neurogastroenterology: The dynamic cycle of life in the enteric nervous system. *Nat Rev Gastroenterol Hepatol* **14**, 453-454 (2017).
38. Reinius LE, *et al.* Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS One* **7**, e41361 (2012).
39. Jung N, Dai B, Gentles AJ, Majeti R, Feinberg AP. An LSC epigenetic signature is largely mutation independent and implicates the HOXA cluster in AML pathogenesis. *Nature communications* **6**, 8489 (2015).
40. Zhou J, *et al.* Human atrium transcript analysis of permanent atrial fibrillation. *Int Heart J* **55**, 71-77 (2014).
41. Rodbell M. Metabolism of Isolated Fat Cells. I. Effects of Hormones on Glucose Metabolism and Lipolysis. *J Biol Chem* **239**, 375-380 (1964).
42. Spalding KL, Bhardwaj RD, Buchholz BA, Druid H, Frisen J. Retrospective birth dating of cells in humans. *Cell* **122**, 133-143 (2005).
43. Spalding KL, *et al.* Dynamics of hippocampal neurogenesis in adult humans. *Cell* **153**, 1219-1227 (2013).

44. Duncan AW, *et al.* Frequent aneuploidy among normal human hepatocytes. *Gastroenterology* **142**, 25-28 (2012).
45. Dorrell C, *et al.* Transcriptomes of the major human pancreatic cell types. *Diabetologia* **54**, 2832-2844 (2011).
46. Neiman D, *et al.* Islet cells share promoter hypomethylation independently of expression, but exhibit cell-type-specific methylation in enhancers. *Proc Natl Acad Sci U S A* **114**, 13525-13530 (2017).
47. Yu W, *et al.* Formation of cysts by alveolar type II cells in three-dimensional culture reveals a novel mechanism for epithelial morphogenesis. *Mol Biol Cell* **18**, 1693-1700 (2007).
48. Ehrhardt C, Kim KJ, Lehr CM. Isolation and culture of human alveolar epithelial cells. *Methods Mol Med* **107**, 207-216 (2005).
49. Bove PF, *et al.* Human alveolar type II cells secrete and absorb liquid in response to local nucleotide signaling. *J Biol Chem* **285**, 34939-34949 (2010).
50. Mao P, *et al.* Human alveolar epithelial type II cells in primary culture. *Physiol Rep* **3**, (2015).
51. Roche JK. Isolation of a purified epithelial cell population from human colon. *Methods Mol Med* **50**, 15-20 (2001).
52. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic acids research* **16**, 1215 (1988).
53. Pidsley R, *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol* **17**, 208 (2016).
54. Kling T, Wenger A, Beck S, Caren H. Validation of the MethylationEPIC BeadChip for fresh-frozen and formalin-fixed paraffin-embedded tumours. *Clin Epigenetics* **9**, 33 (2017).

Figure Legends

Figure 1: Identification of tissue-of-origin of cfDNA using deconvolution of the plasma methylome aided by a comprehensive methylation atlas. (a) Methylation atlas composed of 25 tissues and cell types (columns) across ~8000 CpGs (rows). For each cell type, we selected the top 100 uniquely hypermethylated (top) and 100 most hypomethylated (bottom) CpG sites, giving a total of 5,000 tissue-discriminating individual CpGs. We then added neighboring (up to 50bp) CpGs, as well as 500 CpGs that are differentially methylated across pairs of otherwise similar tissues. Overall, we used 7,890 CpGs that are located in 4,039 500bp genomic blocks. **(b)** Deconvolution of plasma DNA. Cell-free DNA (cfDNA) is extracted from plasma and analyzed with a methylation array. It is then deconvoluted using a reference methylation atlas to quantify the contribution of each cell type to the cfDNA sample.

Figure 2: DNA methylation patterns allow for accurate deconvolution of simulated admixed samples. (a) The methylome of each cell type was mixed *in silico* with the methylome of leukocytes such that it contributed between 0% to 10% of DNA, in 1% intervals (x-axis of each plot) and compared to the prediction of deconvolution using our reference methylation atlas (y-axis). Red horizontal bars represent the median predicted contribution for each mixed-in level, across 36-180 replicates for each cell type (2-10 replicates of measured cell type methylomes, each mixed within any of 18 leukocyte replicates). The

blue area represents a box plot spanning the 25th to 75th percentiles for each mixing ratio, with black vertical lines marking the 9th to 91st percentiles. **(b)** Primary tissue methylome allows a more accurate deconvolution than whole-tissue or a cell line. Hepatocyte methylome was mixed *in silico* with blood methylome as in (a). The level of inferred admixture (y-axis) was calculated using a reference tissue methylome atlas that included other hepatocyte samples (green), whole liver methylomes (blue) or the methylome of the HepG2 cell line (red). Dotted red line marks accurate prediction. **(c)** Cell type-specific methylomes allow a more accurate deconvolution than whole tissue methylomes. The methylome of pancreatic acinar, duct, or beta cells was diluted *in silico* into leukocyte methylomes (left, middle and right, respectively); the level of admixture was calculated using a comprehensive reference atlas that contained either independent samples of the spiked-in pancreas cell types (green lines), or a whole pancreas methylome (blue lines). Note assay linearity, but reduced sensitivity, when using a whole pancreas methylome.

Figure 3: *in vitro* mixing experiments. Genomic DNA derived from liver **(a)**, lung **(b)**, neurons **(c)** and colon **(d)** (each from a single donor) was mixed in 9 different combinations (detailed in Supplementary File 1) with genomic DNA extracted from the blood of a single healthy donor, in the proportions indicated in the X axis. A total of 250ng DNA from each mixture was subjected to an Illumina EPIC array, and the resulting methylome was deconvoluted to predict the contribution of each mixed-in tissue (Y axis). Each dilution point represents one mixing experiments.

Figure 4: Cellular contributors to cfDNA in healthy individuals **(a)** Predicted distributions of contributors to circulating cfDNA, averaged across eight sample pools of healthy donors. Contributions smaller than 1% were included in “Other”. **(b)** Deconvolution results for eight sets of pooled DNA samples, expressed as absolute levels of DNA (genome equivalents/ml plasma, derived by multiplying the fraction contribution of each tissue by the total amount of cfDNA in 1ml plasma). Shown are contributions larger than 1%. Young, 19-30 years old; Old, 67-97 years old (pool average > 75yr). **(c)** Comparison of estimated proportion of various cell types in healthy plasma samples (blue) vs. leukocytes (orange), as predicted by deconvolution. Shown, from left, are the contributions of erythrocyte progenitor cells, vascular endothelial cells and hepatocytes, all of which are not expected in leukocyte samples. Also shown are the predicted contributions of lymphocytes, that represent a large fraction of leukocyte cell population. Shaded boxes mark 95% confidence interval of the sample mean. **(d)** Deconvolution of whole blood methylomes (not plasma), showing excellent correlation (Pearson’s $r=0.985$, $p<2e-16$) between the estimated proportions of monocytes, neutrophils and lymphocytes and the actual proportions of these cells obtained via standard Complete Blood Count (CBC) for each sample.

Figure 5: Cellular contributors to cfDNA in islet transplant recipients. **(a)** Deconvolution results for pooled sample of cfDNA from five patients, 1 hour after islet transplantation. The patients present a noticeable amount of pancreas-derived cfDNA (typically absent in healthy donors). Cell types contributing <1% were included in “Other”. **(b)** Same as (a), expressed as absolute levels of cfDNA (genome equivalents per ml plasma). Also shown is the prediction for a healthy individual. **(c)** Inferred amount of cfDNA from all three pancreas cell types for three individuals prior to, 1 hour after, and 2 hours after islet transplantation. Error bars: SD, estimated using Bootstrapping. **(d)** Comparison of pancreatic cfDNA estimations using deconvolution (y-axis) to results of targeted insulin promoter

methylation assay (x-axis). Pearson's $r=0.996$, $p\text{-value}=1.6e-8$. **(e)** Same as (c), using a reference atlas with whole pancreas methylome, instead of purified pancreas cell types. Here, deconvolution fails to identify pancreatic cfDNA in recipient 1.

Figure 6: Cellular contributors to cfDNA in sepsis. **(a)** Predicted cellular contributions are shown for 14 samples of cfDNA from patients with sepsis. Cell types present at $<1\%$ were included in "Other". **(b)** Pie charts representing predicted distribution of cell types contributing to cfDNA of two of the sepsis patients. **(c)** Predicted levels of hepatocyte cfDNA compared to serum levels of Alanine Aminotransferase (ALT), a standard biomarker for hepatocyte damage (Pearson's $r=0.93$, $p\text{-value}\leq 4e-7$). Error bars: SD, as estimated using Bootstrapping.

Figure 7: Cellular contributors to cfDNA in cancer. **(a-c)** Predicted contributions of breast, colon and lung DNA to the plasma methylome of 4 patients with colon cancer (CC), 4 patients with lung cancer (LC), 3 patients with breast cancer (BRC) and 4 healthy donors (H). All patients were at advanced stages of disease. **(d)** A mix-in experiment. The plasma of a patient with advanced colon cancer was mixed with 3 healthy plasma samples in varying proportions (detailed in Supplementary File 1), and the fraction of colon-derived cfDNA was assessed using deconvolution of the methylome. **(e)** Identification of prostate-derived cfDNA in published plasma methylomes of patients with prostate cancer³⁵ before and after treatment. Patients classified as abiraterone acetate (AA) treatment responsive (blue) show a dramatic drop in prostate-derived cfDNA, compared with the AA-resistant patients (red). **(f)** Deconvolution of cfDNA methylation predicts cfDNA origin for CUP cancer patients. Shown are the predicted cellular contributors for cfDNA samples from 4 patients diagnosed with a Cancer of Unknown Primary (CUP). Blood cell types and cells contributing $<1\%$ are not shown. For each patient, the location of metastases and the presumed tissue source of cancer according to clinical history are listed. Deconvolution results agreeing with clinical predictions are shown as orange bars.

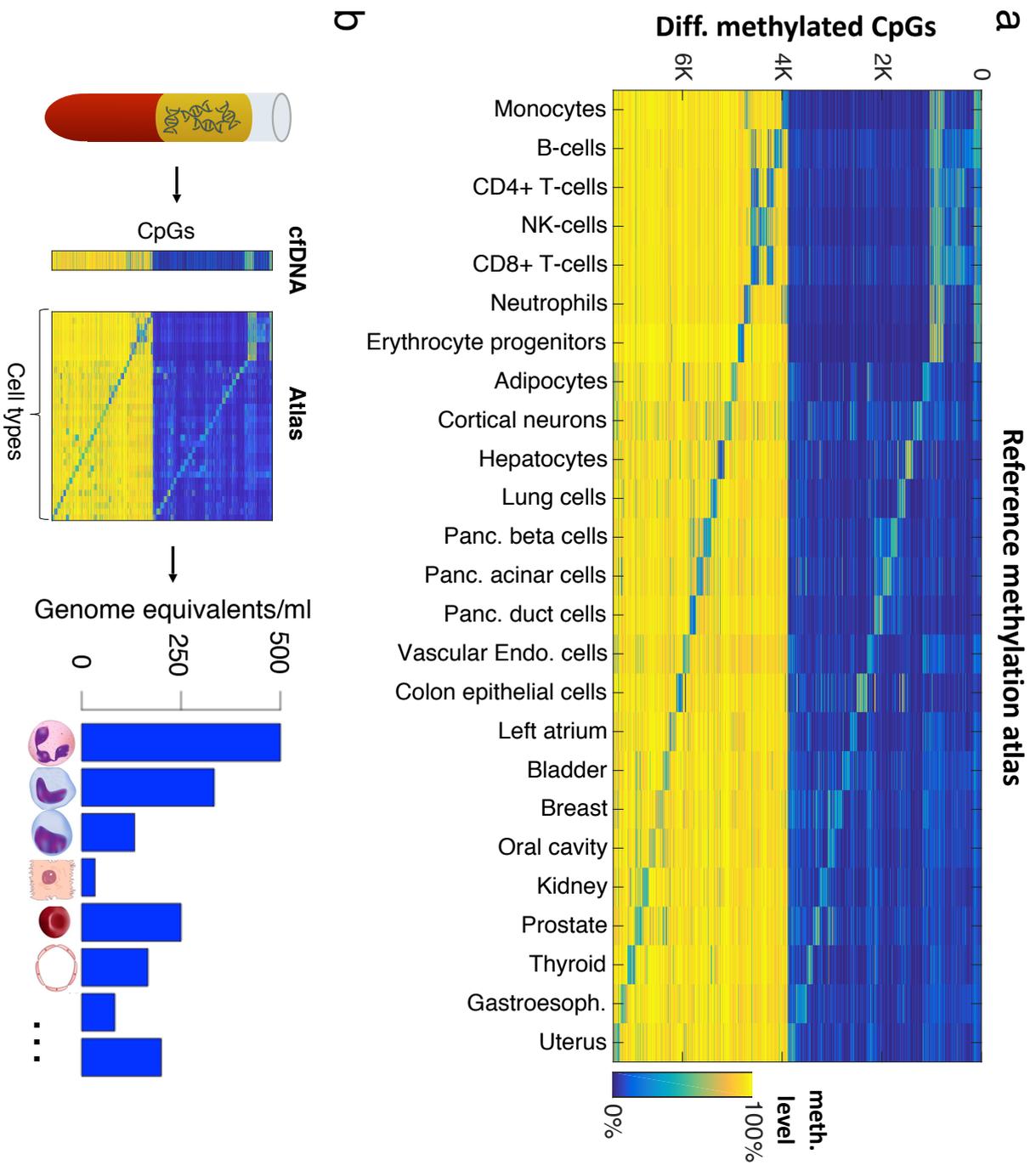
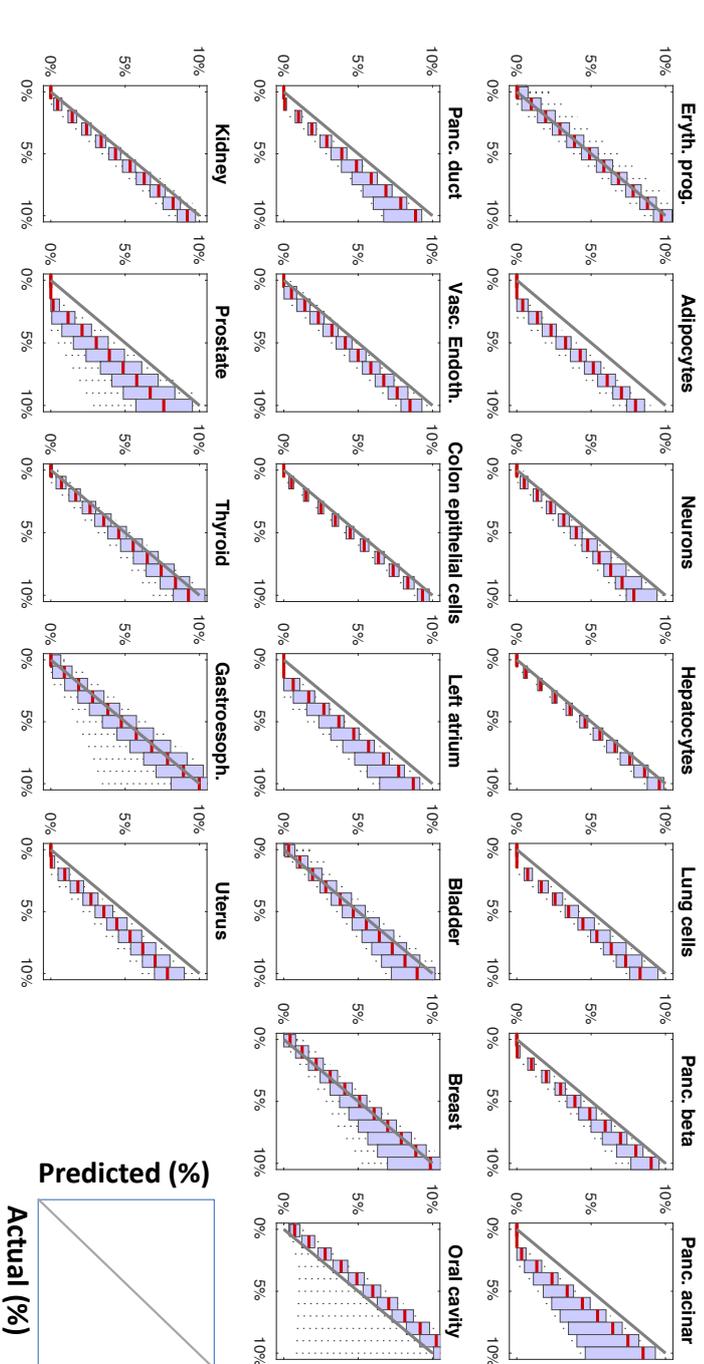
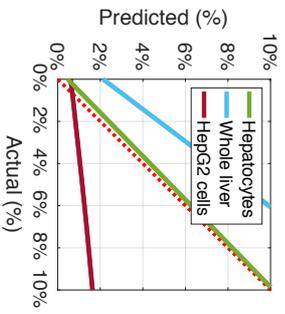


Figure 2

a



b



c

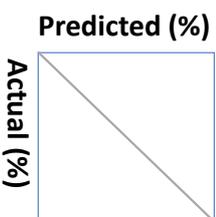
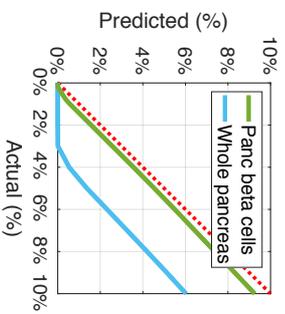
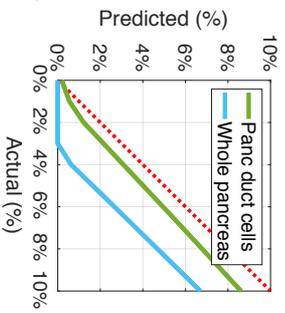
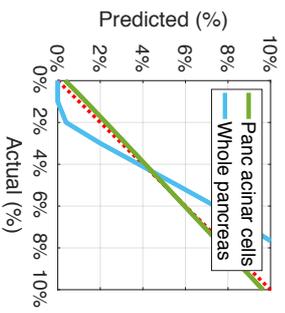


Figure 3

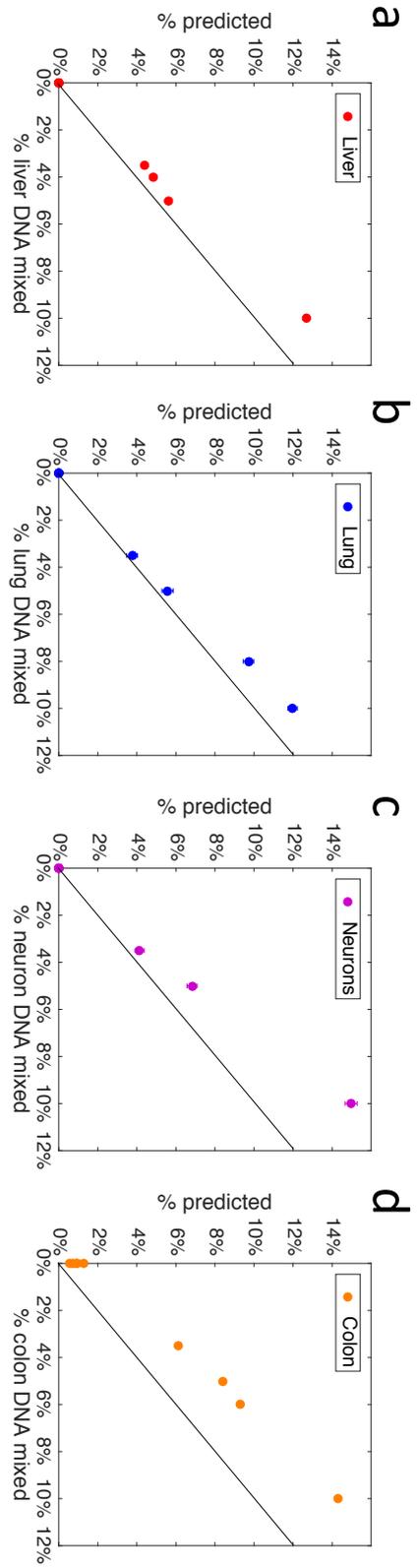


Figure 4

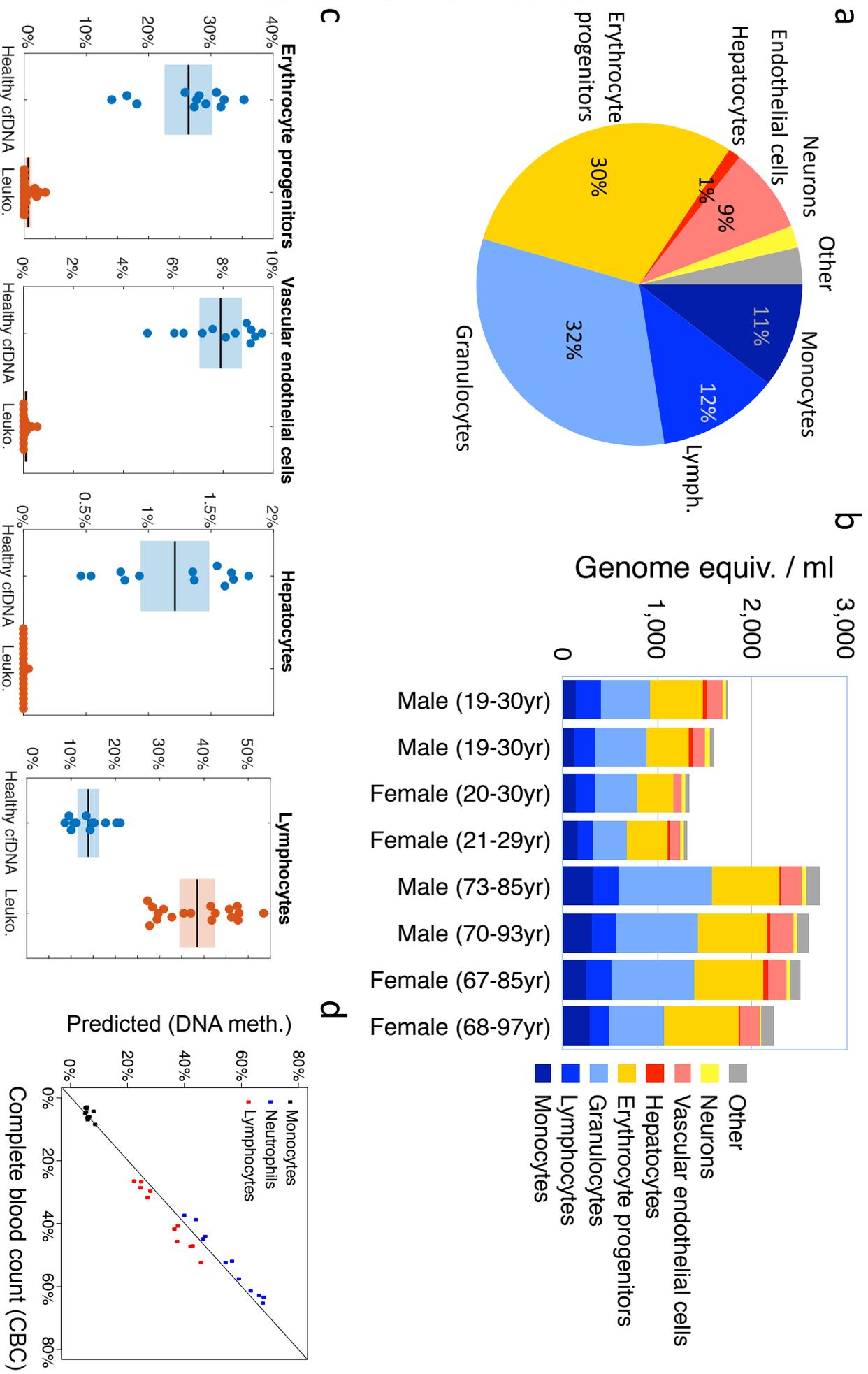


Figure 5

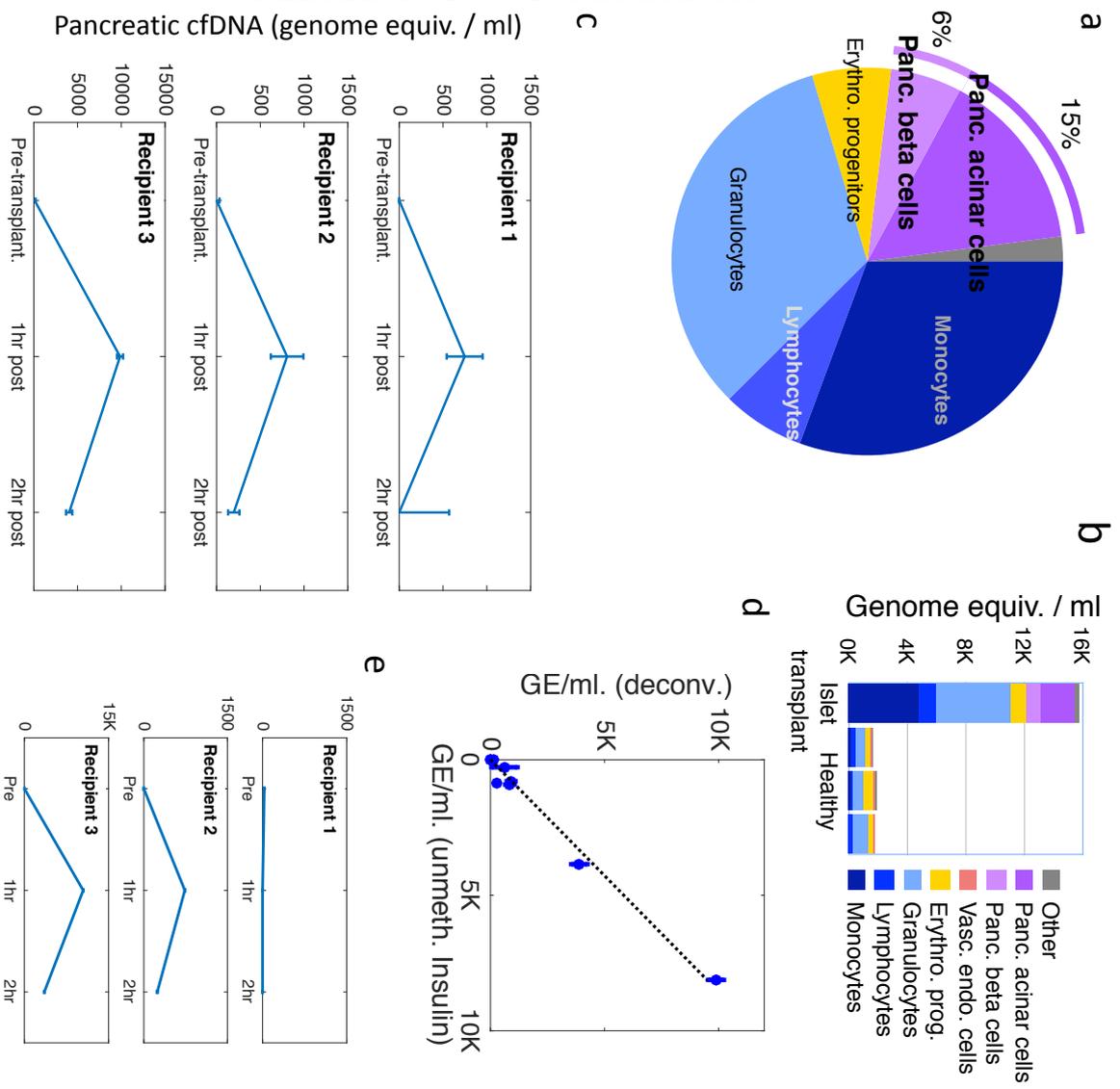


Figure 6

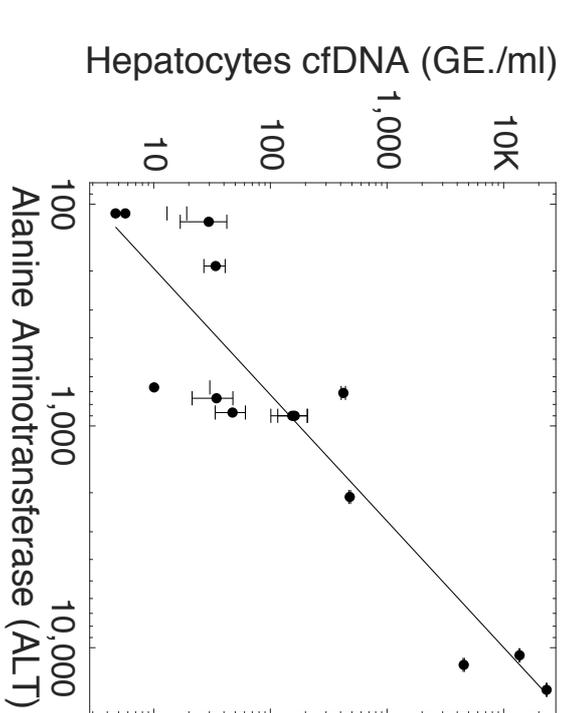
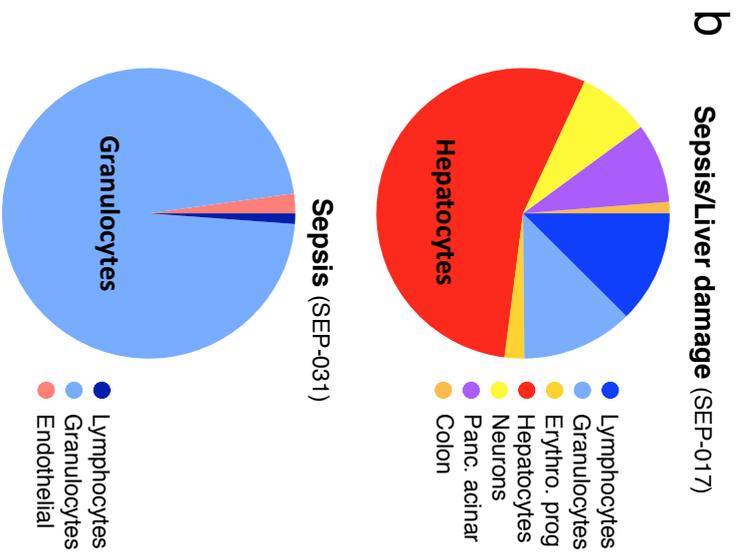
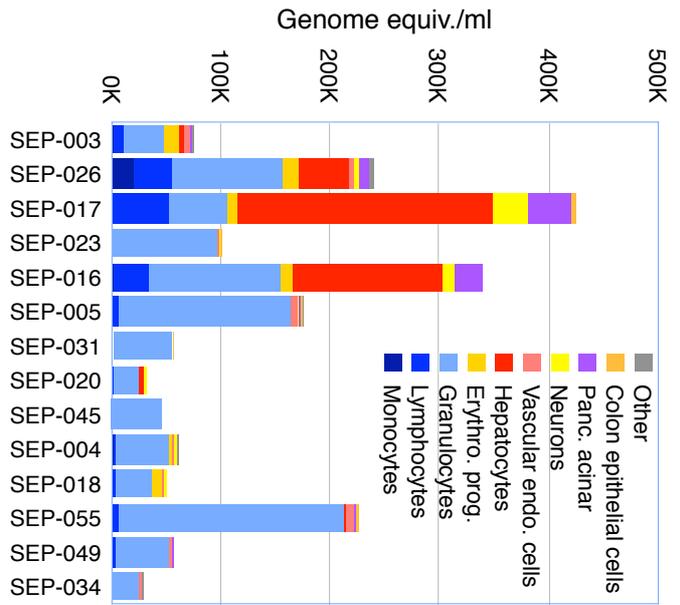
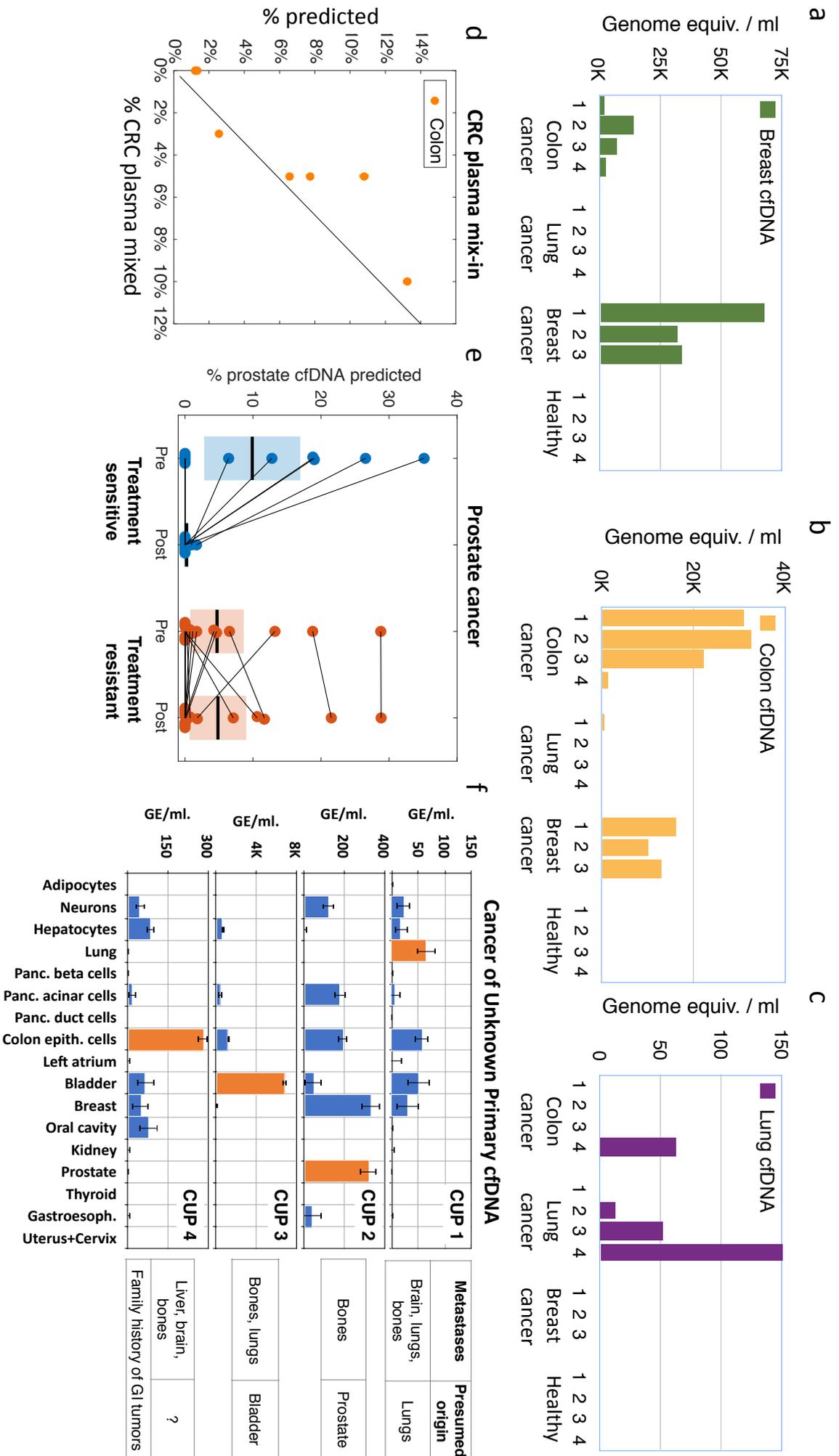
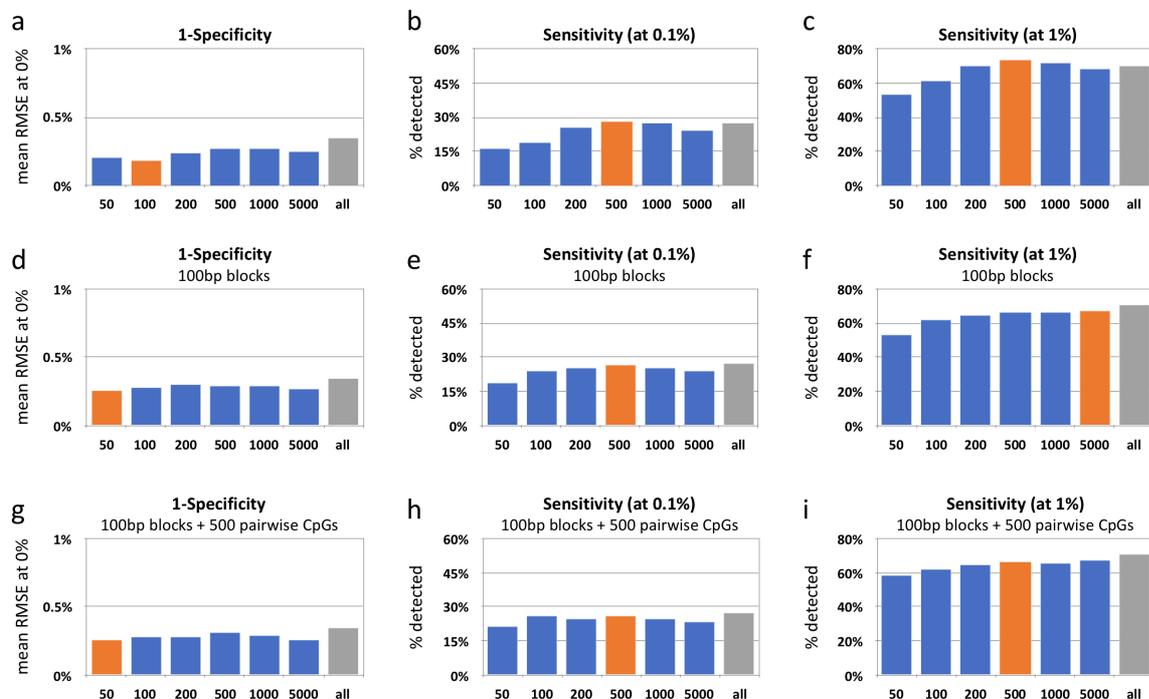


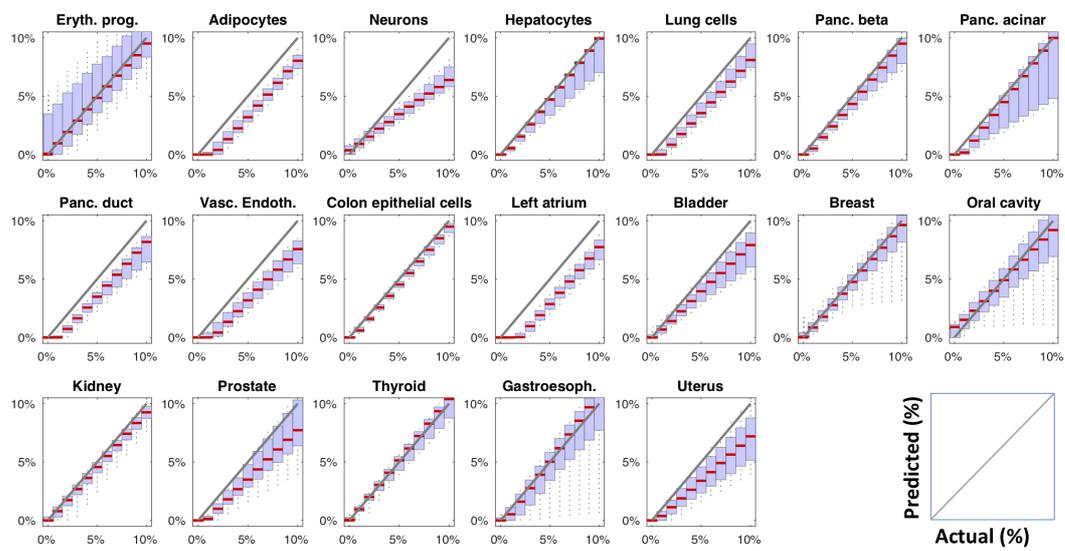
Figure 7



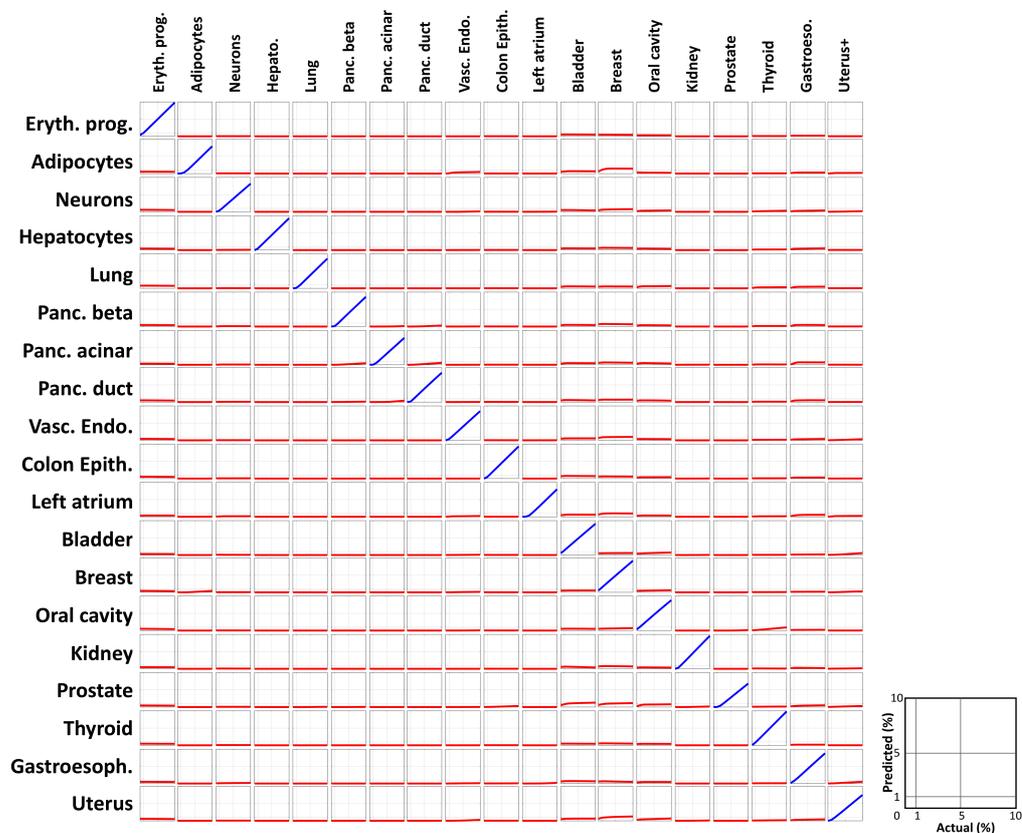
Supplementary Figures



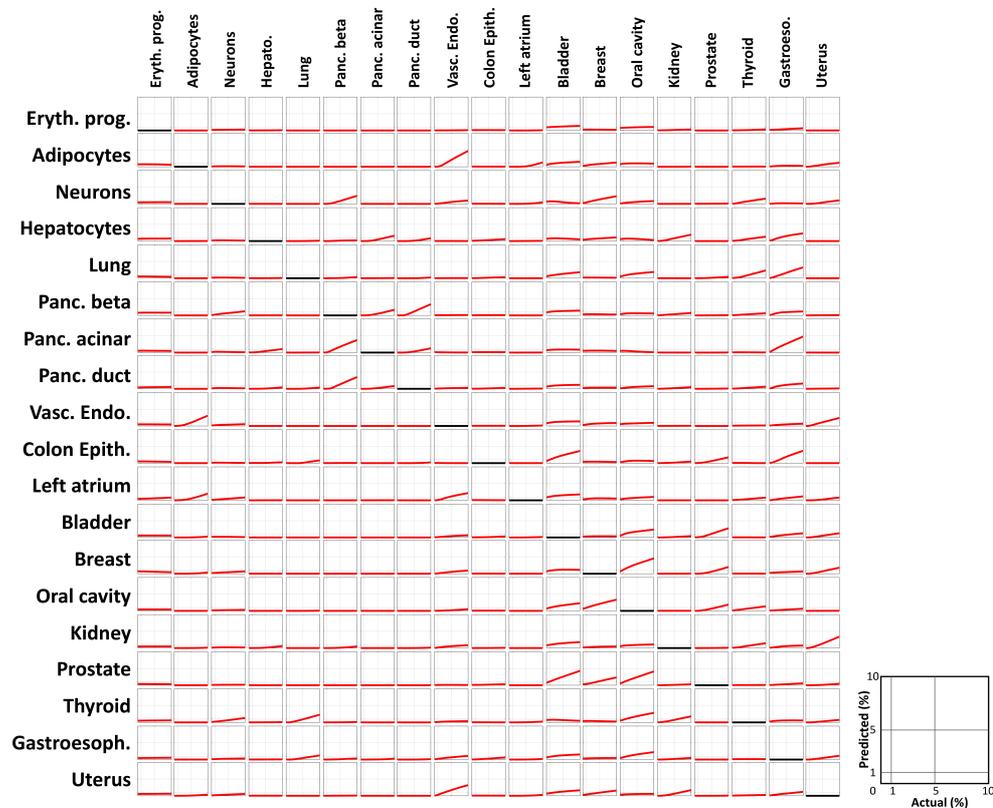
Supplementary Figure 1. Estimations of specificity (false positive rate) and sensitivity (detection rate) for various CpG selection strategies. (a) Specificity values, estimated as the mean error (RMSE) for simulations with 0% mixing (namely, pure leukocytes), for convolution with different numbers of selected CpGs per cell types. These include the top K differentially hypermethylated CpGs and top K hypomethylated CpGs per cell types, with K varying from 50 CpGs (50x2x25=2500 total), through 100, 200, 500, 1000, 5000, or all methylation array CpGs (right-most column). Note that deconvolution with all CpGs is less accurate (and less efficient) compared to models with fewer, selected, CpGs. **(b)** Sensitivity values, estimated as percent of in silico mixes (at 0.1%) correctly detected. Orange bar marks optimal selection (500 hyper + 500 hypomethylated CpGs per cell type, for total of 25,000 CpGs). **(c)** Same as (b), but at a 1% mixing-in level. **(d-f)** Same as (a-c), with deconvolution also based on all neighboring CpGs (up to 50bp away) from previously selected ones. The addition of neighboring CpGs allows for accurate deconvolution with few CpGs, e.g. 2x100 CpG blocks per cell type (total of 7,390 CpGs in 4,039 CpG “haplotype blocks”). **(g-i)** Specificity using previous CpGs with additional pairwise-specific 500 CpGs that are specifically selected to distinguish between similar cell types (e.g. different T cells, adipocytes vs. vascular endothelial cells, Bladder vs Prostate, etc), allowing for a further improvement in sensitivity with few CpGs.



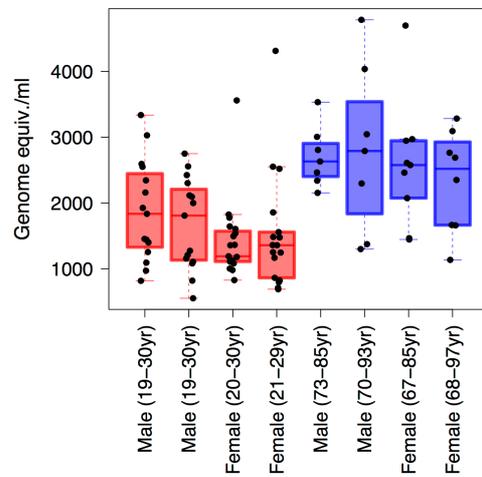
Supplementary Figure 2. Same as Fig 2, without feature selection (250,777 CpGs in total). Included are all CpGs, except for those with missing values or those with variance < 0.1% across the methylation atlas.



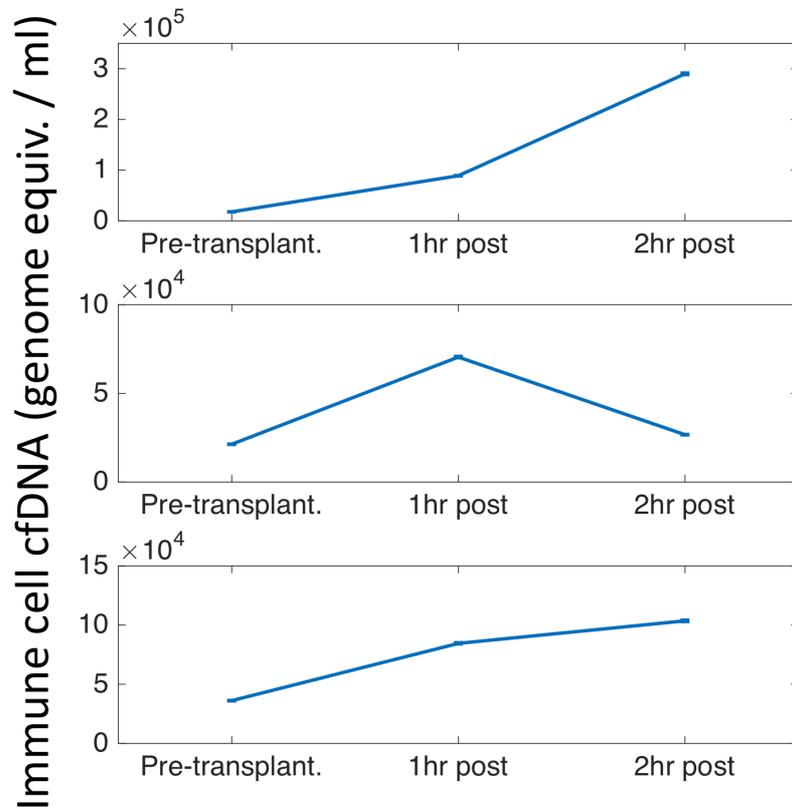
Supplementary Figure 3. Confusion matrix for deconvolution of plasma cfDNA methylomes. Each row corresponds to one cell type, in silico admixed with leukocytes at various mixing ratios from 0% to 10% (x-axis) and depicts the inferred proportion of the mixed (in blue) and all other cell types (in red). Most cell types are completely invariant of each other.



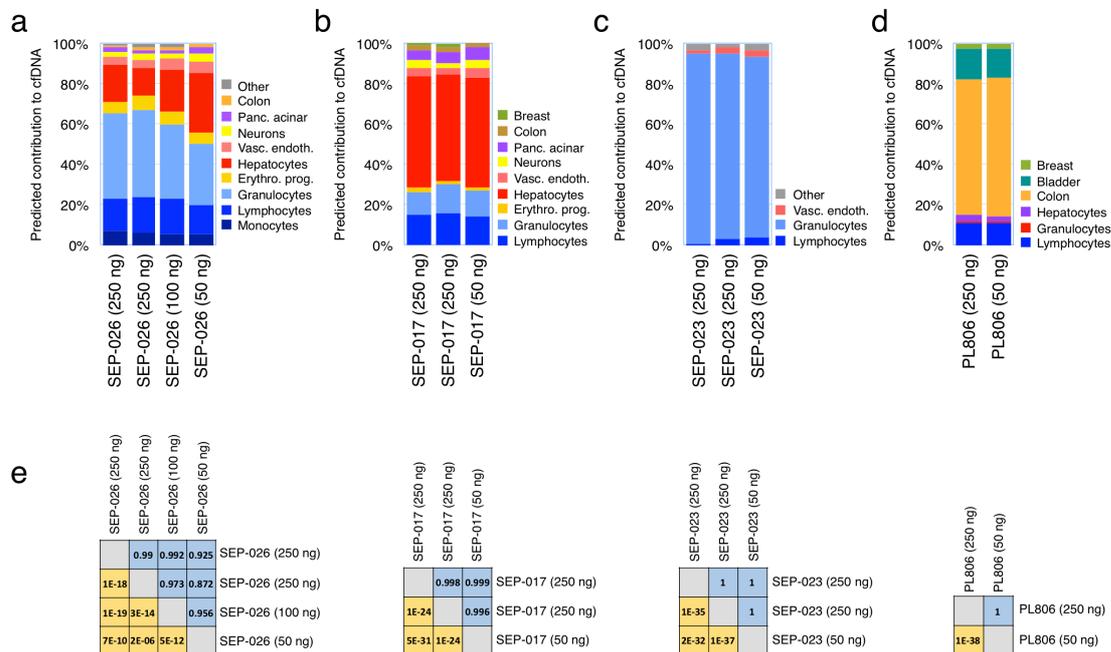
Supplementary Figure 4. Confusion matrix for deconvolution of plasma cfDNA methylomes. Unlike Supplementary Figure 3, here the admixed cell type was completely removed from the methylation atlas prior to deconvolution (black lines), resulting with some confusion between functionally or biologically related cell types (adipocytes and endothelial cells, pancreatic cells, etc).



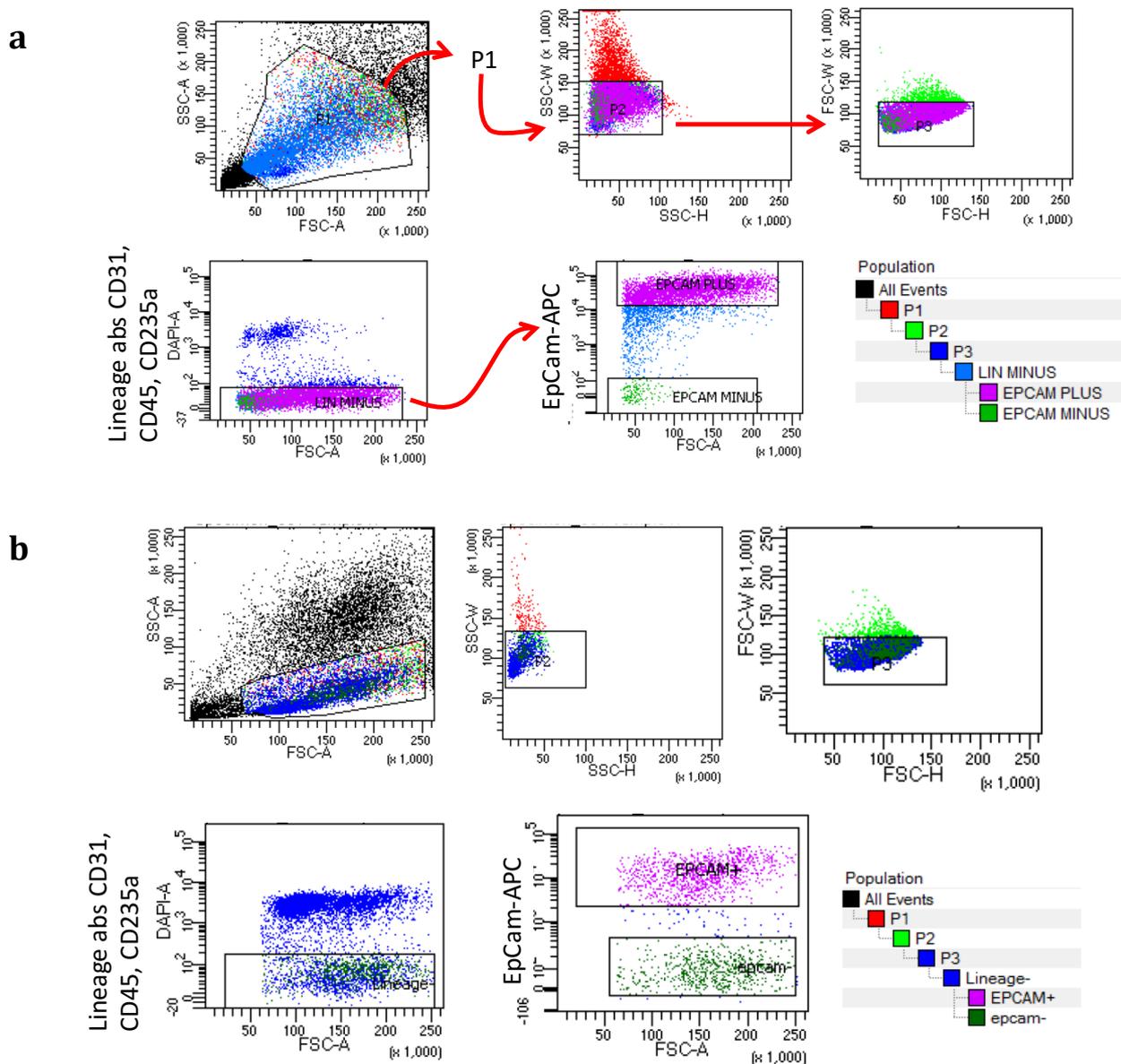
Supplementary Figure 5. cfDNA concentration of healthy individuals. Total concentration of cfDNA (in haploid genome equivalents/ml plasma) from all samples used to generate healthy pools are shown, grouped by the pool they were used in, as in Fig 4b. The cfDNA concentrations of the older individuals (blue boxplots) were significantly greater than those of the younger individuals (red boxplots) (p -value $< 3.96e-7$, Mann-Whitney test).



Supplementary Figure 6. Immune cell cfDNA in pancreatic islet transplantation. Same as Figure 5c, plotting the inferred amount of cfDNA (in haploid genome equivalents/ml plasma) from all immune cell types for three individuals prior to, 1 hour after, and 2 hours after islet transplantation. Error bars: SD, estimated using Bootstrapping.



Supplementary Figure 7. Reproducibility of deconvolution results. (a-d) Predicted distribution of cellular contributors shown for four samples using different amounts of DNA (50 ng, 100 ng or 250 ng). **(e)** Pearson correlation coefficients (in blue) and p-values (in yellow) shown for different pairs of analyzed plasma cfDNA methylomes (each set from the same individual).



Supplementary Figure 8. Sorting and gating strategies for colon epithelial cells (a) and lung alveolar epithelial cells (b). We chose a P1 population according to size and granularity cell distribution, followed by gate P2 and then gate P3 to avoid doublets. From the P3 gate we chose the Lineage negative cells, and then selected the EpCam positive cells.

Legends for Supplementary Data file 1:

Supplementary Data File 1 contains nine individual tables, covering:

Table 1: Cell type-specific CpGs selected for deconvolution.

Table 2: Pairwise-differential CpGs selected for deconvolution.

Table 3: Ages of samples used in healthy pools.

Table 4: Inferred composition of healthy plasma cfDNA.

Table 5: Reference sample donor data.

Table 6: In vitro mixes.

Table 7: Cancer patient data.

Table 8: Healthy and cancer cfDNA mixes.

Table 9: Cancer of unknown primary site (CUP) data.