MetDecode: methylation-based deconvolution of cell-free DNA for non-invasive multi-cancer typing

Dhanya Sudhakaran[1†], Stefania Tuveri[1†], Antoine Passemiers[2†], Tatjana Jatsenko[1], Tina Laga[3,9], Kevin Punie[4,5,6], Sabine Tejpar[7], An Coosemans[8], Els Van Nieuwenhuysen[3,9], Dirk Timmerman[9], Giuseppe Floris[10], Anne-Sophie Van Rompuy[10], Xavier Sagaert[10], Antonia Testa[11], Daniela Ficherova[12], Daniele Raimondi[2], Frederic Amant[3,9,13], Liesbeth Lenaerts[3], Yves Moreau[2] and Joris R. Vermeesch[1*]

*To whom correspondence should be addressed. Center for Human Genetics, University Hospitals Leuven, KU Leuven, Herestraat 49, box 602, Leuven 3000, Belgium.
Telephone: +32 16 34 5941
Email: joris.vermeesch@kuleuven.be
† Joint Authors

1. Laboratory for Cytogenetics and Genome Research, Department of Human Genetics, KU Leuven, Leuven, Belgium.

2. Dynamical Systems, Signal Processing and Data Analytics (STADIUS), KU Leuven, Leuven, Belgium.

3. Gynaecological Oncology, Department of Oncology, KU Leuven, Leuven, Belgium.

4. Multidisciplinary Breast Centre, University Hospitals Leuven, Leuven, Belgium.

5. Laboratory of Experimental Oncology, Department of General Medical Oncology, University Hospitals Leuven, KU Leuven, Leuven, Belgium.

6. Department of Oncology, GZA Ziekenhuis, Antwerp, Belgium

7. Digestive Oncology Unit, University Hospital Gasthuisberg, Leuven, Belgium.

8. Laboratory of Tumour Immunology and Immunotherapy, Department of Oncology, Leuven Cancer Institute, KU Leuven, Leuven, Belgium

9. Gynaecology and Obstetrics, University Hospitals KU Leuven, Leuven, Belgium

10. Translational Cell & Tissue Research, Department of Pathology, KU Leuven, Leuven, Belgium

11. Department of Woman, Child and Public Health, Fondazione Policlinico Universitario Agostino Gemelli, IRCCS, Rome, Italy

12. Obstetrics and Gynaecology, First Faculty of Medicine, Charles University and General University Hospital in Prague, Prague, Czech Republic

13. Gynaecologic Oncology, Netherlands Cancer Institute, Amsterdam, the Netherlands

Abstract

Cell-free DNA (cfDNA) mediated early cancer detection is based on detecting alterations in the cfDNA components. However, the underlying pathology can usually not be readily identified. We built a reference atlas based on the methylome of multiple cancer and blood-cell types and developed MetDecode, an epigenetic signature-based deconvolution algorithm. MetDecode accurately estimates the tumour proportion in *in-silico* mixtures and identifies the tissue of origin in 81.25% cfDNA samples from cancer patients. This method will complement cancer screening programs and guide clinical follow-up.

Keywords

cell-free DNA, liquid biopsy, tissue-of-origin identification, epigenomic profiling, methylome deconvolution, reference-based deconvolution

Background

Cell-free DNA (cfDNA) is present as DNA fragments floating in the blood. The fragments, mainly derived from dying cells, contain the genomic and epigenetic signatures of the cells of origin [1]. When cancer is present, a fraction of the cfDNA can be derived from tumour cells, defined as circulating-tumour DNA (ctDNA) [2]. ctDNA is now being widely explored as a non-invasive biomarker for cancer screening and diagnosis [3]. A major focus is on the detection of cancer-specific single nucleotide mutations and copy number alterations (CNAs). Whereas somatic mutations are usually identified by targeted sequencing, the detection of CNAs is done by genome-wide sequencing of the cfDNA [4,5]. Though somatic mutations can be tumour-specific, their application for cancer detection and screening is hampered by the low variant allele frequency at the early stages of the disease [5,6]. Alternatively, genome-wide detection of tumour-associated CNAs in cfDNA has been shown to allow cancer detection, also in population screening settings with low-pass sequencing [7–10].

CfDNA-based screening for cancers can often detect the presence of abnormal signals indicative of a cancer, but not its origin or cancer type. Especially for metastatic disease of unknown primary, profiles do not readily allow the identification of the tissue of origin (TOO) albeit this would be of clinical value [11]. When performing non-invasive prenatal screening for foetal chromosomal aneuploidies, incidental occult maternal malignancies can be detected without insights into the origin [10,12,13]. If the TOO or cancer type could be deduced from cfDNA analysis, this would tremendously speed up the diagnosis and start of treatment, hence streamlining subsequent clinical follow-up, reducing costs and minimising the need for extensive radiologic imaging [14]. For patients, this might reduce the anxiety associated with a positive screening test outcome.

Methods to deduce the origin of cfDNA fragments have been based on epigenetic markers such as nucleosome positioning, fragmentation and methylation profiles [15–17]. These profiles are tissue and cell-type-specific [18], offering the possibility to identify the different components of the cfDNA pool, alongside an estimation of the relative proportion of each of them. Tumour-associated methylation changes have been described during cancer initiation and progression. Hence, they are promising markers for early cancer TOO identification [19].

Recently, several algorithms have been developed to deconvolute the plasma cfDNA composition based on methylation profiles. Typically, reference atlases consisting of either normal tissues or cell-type-specific methylation markers are used to identify tissue-specific methylation signals [14,19–26]. Although each method has its specific merits, it also has its limitations. For instance, none of the methods deconvolutes multiple cancer tissues. Also, most methods do not consider missing variables due to the incompleteness of the atlas [22,28] or operate in a reference-free fashion [26]. However, cfDNA mixtures are more complex and could carry fragments from tissues not represented in the atlas.

To address these limitations, we developed an alternative reference-based deconvolution method, named MetDecode. The method builds on gradient-based optimization and extends existing methods by simultaneously modelling the presence of noise and the lack of comprehensiveness of the reference atlas in a deterministic and lowly-parametrized fashion. We used in-house sequenced or

publicly available tumour ssamples to build a reference atlas of tissue-specific methylation markers for four different cancer tissues, namely breast, ovarian, cervical and colorectal cancer and combined it with white blood cell (WBC)-derived entities. The reference atlas is subsequently exploited by MetDecode to estimate the contribution of each atlas entity. Additionally, the reference atlas is extended with unknown methylation patterns learnt on-the-fly from cfDNA methylation profiles to account for missing data. This method could complement cancer screening programs to direct clinical follow-up to the right cancer type and will expedite treatment.

Methods

Plasma cfDNA and genomic DNA collection and extraction

Peripheral blood was collected in Roche cell-free DNA blood collection tubes® (Roche, Switzerland) or a Streck Cell-Free BCT® (Streck, USA) and extracted as described previously [7]. Archived [7] and prospectively collected plasma cfDNA samples of healthy individuals were included as control samples (18-90 years old). We included only individuals without cancer and no known autoimmune condition to exclude the introduction of confounding factors to the analysis, as both pathological conditions can influence the shedding and the composition of cfDNA [2,29]. Archived plasma cfDNA was also obtained from patients with a known diagnosis of breast, colorectal or ovarian cancer (mean age: 61.88 years old). Treatment-naïve formalin-fixed paraffin-embedded (FFPE) tumour biopsies were collected. Genomic DNA (gDNA) was extracted from the FFPE tumour biopsies as well as from WBC from healthy subjects or patients with a diagnosis of breast, colorectal, cervical or ovarian cancer using the QIAamp DNA FFPE Tissue Kit or the DNeasy Blood & Tissue Kits (Qiagen, Hilden, Germany), respectively. The extracted gDNA was fragmented using Covaris M220 before library preparation (Covaris Inc., Woburn, MA, USA). The study was approved by the ethical committee of the University Hospitals Leuven (study protocols S62285, S62795, S63983, S66450, S59207 and S51375).

Complete blood count

Advia 2120 hemacytometer was used to perform the complete blood count (CBC) analysis on whole blood following manufacturer's instructions.

Whole-genome methylation sequencing and data analysis

cfDNA and gDNA extracted from FFPE tumour biopsies or WBC was subjected to whole-genome DNA methylation sequencing using the NEBNext Enzymatic Methyl-seq kit (New England Biolabs, Ipswich, MA, USA) following manufacturer's instructions. Enzymatic conversion was preferred over bisulfite conversion for methylation analysis, avoiding fragmentation and loss of DNA in the process [28,30]. In addition, for gDNA from cervical and ovarian FFPE tumour biopsies that were used to build the reference atlas, bisulfite conversion was performed, to be consistent with the method used for the remaining samples in the atlas. Libraries were prepared with the same kit, thereby replacing the enzymatic conversion reactions with the bisulfite treatment using EZ-96 DNA Methylation-Direct MagPrep (Zymo Research, Irvine, CA, USA). The conversion efficiency was evaluated by spiking unmethylated Lambda DNA in one sample per batch, irrespective of the conversion method used. Libraries were quantified using Qubit dsDNA high-sensitivity assay kit and Qubit 3.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA). Libraries were sequenced on NovaSeq 6000 S4 flowcell (Illumina, San Diego, CA, USA) generating PE150bp reads at an average depth of 15X. The data after demultiplexing was quality checked and trimmed using fastp (v0.20) and then aligned to human genome hg38 using bwa-meth (v0.2.2). Deduplication was done using Picard (v2.20.3) and methylation calling via MethylDackel (v0.5.1). The tumour fraction in the cfDNA samples was calculated using ichorCNA [31]

Generation of a DNA methylation marker atlas for multiple blood cell types and tumour tissues

A DNA methylation marker atlas, covering markers for 6 tumour tissues and 7 blood cell types was generated solely using whole-genome bisulfite sequencing (WGBS) data. From public repositories, we downloaded genome-wide CpG site methylation ratios for B cells, CD4+ T cells, CD8+ T cells, natural killer cells, monocytes, neutrophils and erythroblasts (BLUEPRINT [32], GSE186458), and for breast invasive carcinoma, colon adenocarcinoma and rectal adenocarcinoma tissues (The Cancer Genome Atlas [33] (Supplementary Table 1, Additional File 2). WGBS data for high-grade serous ovarian carcinoma, cervical adenocarcinoma and cervical squamocellular carcinoma were generated in-house from FFPE samples. Available

samples (n=2-7) per tissue/cell type were merged after removing highly variant CpG site which resulted in a combined atlas entity for every tissue as explained in the Additional File 1. Using these combined atlas entities, the sites which were uniquely methylated in one tissue type, with at least 30% difference between the absolute methylation value in that tissue versus the rest, were extracted using R scripts and extended to cover a region with a minimum of 4 CpGs (if the sites were within 500 bp). Once the start and end coordinates of the marker regions were obtained, the total number of reads and the number of methylated reads in these regions for every tissue/cell-type in the atlas, namely $D^{(atlas)}$ and $M^{(atlas)}$, were obtained using custom scripts. The same was also extracted for the samples to be deconvoluted ($D^{(cfdna)}$, $M^{(cfdna)}$) and then used as input for the deconvolution algorithm.

Deconvolution algorithm

We created a methylation-based reference atlas, composed of two matrices $D^{(atlas)}$ and $M^{(atlas)}$, where $D_{jk}^{(atlas)}$ is the total CpG count for atlas entity j and marker region k, and $M_{jk}^{(atlas)}$ the corresponding methylated CpG count. Because each CpG site can be spanned by multiple reads, it may contribute multiple times to the same count. Therefore, these values must not be confused with read counts. We also provided the algorithm with two other input matrices, $D^{(cfdna)}$ and $M^{(cfdna)}$, representing the cfDNA mixtures. Our algorithm has been designed to infer a matrix A of cell type contributions, where A_{ij} is the estimated proportion of cell type j to cfDNA profile i. A was found by minimizing a weighted mean squared error between the methylation ratios of cfDNA samples and the ratios of convoluted atlas entities. Marker region k in sample i was re-weighted by $\frac{\left(D_{ik}^{(cfdna)}\right)^{\beta}}{\sum_{a=1}^{n}\sum_{l=1}^{p}\left(D_{al}^{(cfdna)}\right)^{\beta}}$ to better reflect the confidence in the estimation of the methylation ratio $R_{ik}^{(cfdna)}$. β is a hyper-parameter (default=1) controlling the importance given by the end user to the coverage. To account for the presence of unknown cell types missing from the reference atlas, we extended the atlas with estimates of missing cell types. When the number of cfDNA samples is largely greater than the atlas size, the methylation patterns of these unknown contributors can be learned from the data directly. The assumed number of unknown cell types was defined as a hyper-parameter (default = 1). We accounted for the unknown contributors in the cfDNA mixture by appending extra rows to the

$D^{(cfdna)}$ and $M^{(cfdna)}$ matrices. Methylation ratio matrices $R^{(cfdna)}$ and $R^{(atlas)}$ were computed from the corresponding read count matrices.

$R^{(cfdna)}$ was next deconvoluted using non-negative least squares (NNLS) algorithm and reference matrix $R^{(atlas)}$, the residuals were used to define the missing contributor and extend $D^{(atlas)}$, $R^{(atlas)}$ and $M^{(atlas)}$ by one row each. This procedure was repeated h times. A more technical description of the algorithm is provided in Additional File 1.

Evaluation metric

Pearson Correlation Coefficient and mean squared error (MSE) were used to evaluate the reliability of MetDecode estimations. We evaluated the accuracy of multiclass cancer TOO prediction as $\frac{\#(correctly\ assigned\ samples)}{\#(total\ samples)}$. P-values were considered significant when <0.001.

Results

Creation of a reference atlas and tissue-specific epigenetic marker selection

To enable the deconvolution of a methylome into its potential contributors by assigning the relative proportion to a specific tissue type, a methylation reference atlas with 13 entities was created. We included methylome data from seven cells of hematopoietic origin which are the most represented in plasma cfDNA [22,34] as well as methylome data from six different tumours. The tumour tissues included breast cancer, ovarian cancer, colon adenocarcinoma, rectum adenocarcinoma, cervical adenocarcinoma and cervical squamous cell carcinoma. These cancers were selected to serve as a proof-of-concept. The seven cell types from the haematopoietic lineage included neutrophils, monocytes, erythroblasts, natural killer, B cells, CD4+ T cells and CD8+ T cells. Tumour methylome data was downloaded from TCGA for five breast invasive carcinomas from different subtypes (n=1 luminal A, n=1 luminal B, n=1 basal-like; n=2 HER2), two rectum adenocarcinoma and two colon adenocarcinoma (Supplementary Table 1, Additional File 2). Since publicly available data was lacking for cervical and ovarian tumours, we generated genome-wide methylome data for three high-grade serous ovarian carcinomas (HGSOC), two cervical adenocarcinomas and three cervical squamocellular carcinoma in-house and included it in the atlas. For interpretation purposes, we combined deconvoluted

percentages for the two cervical cancer subtypes and from the colon and the rectal adenocarcinoma to identify the TOO.

Differentially methylated sites were selected by comparing the CpG site methylation ratio of one tissue against the rest of the entities in the reference (one-versus-all strategy) and then extended to regions. To ensure that the methylation marker regions were unique, a methylated or unmethylated region should have a distinct methylation pattern in one tissue versus the other entities in the reference atlas (Figure 1A). We identified 17874 differentially methylated regions across the genome for the 13 reference entities. The number of marker regions per reference entity ranged from 23 to 7058 with a median count of 5 CpGs and median length of 512bp (Figure 1B).

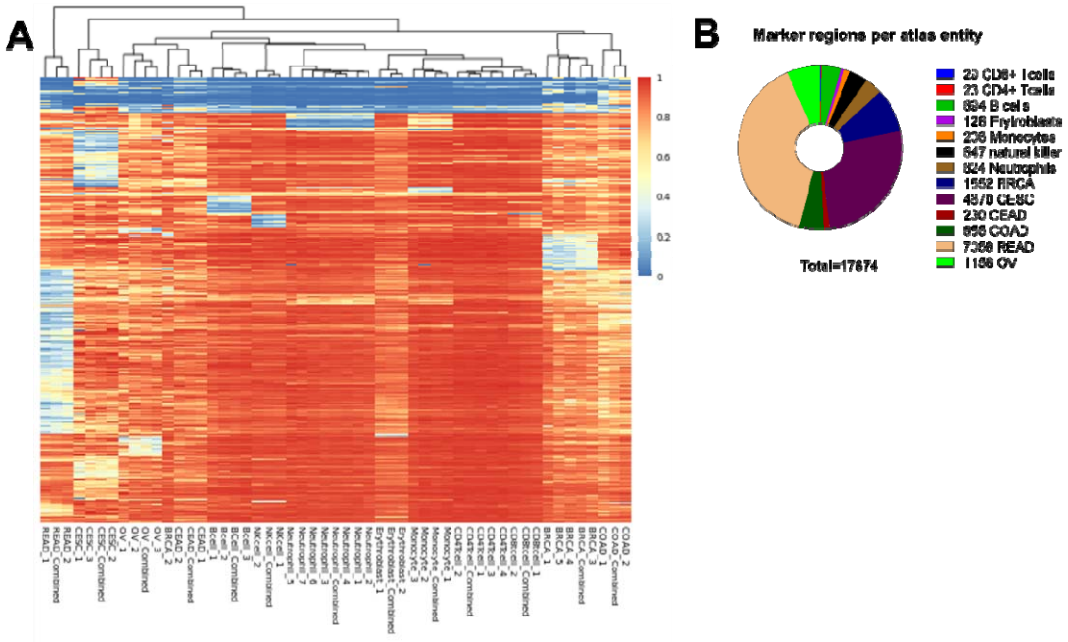


Figure 1 (A) Heatmap displaying the methylation ratio of the selected marker regions across the atlas entities. The methylation ratio is represented on the colour scale with red indicating a value close to one, meaning hypermethylation, and blue to zero, meaning hypomethylation. The individual samples per tissue/cell-type and the combined entity (named Combined, as described Materials and Methods) considered for the marker selection are shown on the X-axis. (B) Number of marker regions per atlas entity. BRCA, breast carcinoma; CEAD, cervical adenocarcinoma; CESC, cervical squamocellular carcinoma, COAD, colorectal adenocarcinoma; OVCA, ovarian carcinoma; READ, rectal adenocarcinoma.

MetDecode is robust against noise and the presence of unknown contributors

Along with cancerous cells, cfDNA samples from cancer patients also contain other cell types, such as immune cells. A reference atlas of tumours and immune cells will often be incomplete and hence the analysis methods run the risk of being blinded for

unknown contributions. While the primary goal of our method is to predict the type of cancer, standard supervised classifiers will fall short because (i) the number of training samples per class is too low (2 to 7 per class in this study) for standard classifiers, (ii) samples contain a variable proportion from cells from some classes, and (iii) samples might also contain cells of unknown types/profiles. Instead, we develop an algorithm to estimate the proportions from a mixture of reference samples also accounting for unknown cell-types. The method then assigns the sample to a particular tissue type by considering the atlas entity with the highest proportion falling outside the established normal range.

Contrary to methylation arrays or targeted sequencing which focus on specific loci, whole-genome methylation sequencing produces generally lower coverage due to the reads being spread over the entire genome. Therefore, the coverage of each CpG site is reduced and noise is exacerbated [23]. Not only does the genome coverage vary from sample to sample due to differences in sequencing depth, but it also varies along the genome itself (e.g., due to differences in mappability). Accordingly, we devised MetDecode to account for the reliability of the methylation ratio estimates (e.g., variability at methylation loci both in the atlas and cfDNA samples [35]) in the presence of noise. Since higher coverage of a marker region enables a more accurate estimate of its methylation ratio (under the assumption of the absence of biases, which we implicitly assumed), we re-weighted our objective function (see Methods) to lower the contribution of lower-coverage marker regions to the objective function.

The importance attached to the coverage is controlled by a hyper-parameter β, which determines the rate at which the weight of a marker region increases with its coverage. To assess the relevance of this new feature, we compared our method in default settings ($\beta = 1$) to our method without considering the coverage ($\beta = 0$). For this purpose, we designed simulations based on real data (see Supplementary Figure 1, Additional File 1), with random noise injection based on binomial distributions and deconvoluted these random mixtures. In Figure 2A, we reported the distribution of Pearson correlation coefficients between estimated and expected cell type proportions across 30 runs. When averaging the correlation coefficients across cell types (bottom right violin plot in Figure 2A), Pearson coefficient appears to be significantly higher in the $\beta = 1$ setting ($p<0.001$; T-test; one-sided), highlighting the

gain in deconvolution performance obtained when increasing the attention of our deconvolution algorithms on high-coverage regions, both in the atlas and cfDNA samples.

To create the atlas, only a limited number of cell and tissue types were selected. However, the cfDNA is made up of DNA derived from many different cell and tissue types, albeit usually in lower amounts. To account for this incompleteness of the atlas, we included the possibility to model unknown cell types. We opted for a data-driven approach that infers the unknowns using the cfDNA samples as well as the (incomplete) atlas, based on the residuals obtained after deconvolution (difference between the original and reconstructed/convoluted cfDNA samples). To demonstrate the relevance of this novel feature, we performed experiments analogous to those described above to quantify the performance characteristics of MetDecode when modelling one unknown cell type (unk=1), compared to the situation where the atlas is assumed to be complete (unk=0). In our simulations, cfDNA mixtures were defined as random linear combinations of the atlas entities (with proportions sampled from a Dirichlet distribution, see Additional File 1), plus an unknown entity with random binary methylation pattern. We observed a significant improvement of Pearson correlation coefficients across 30 runs ($p<0.001$; T-test; one-sided Figure 2B) for 5 out of the 6 cancer tissues included in our atlas. Overall, unknown modelling enhanced deconvolution accuracy for 9 out of the 13 cell types, and decreased performance for the 4 remaining cell types. When averaging the Pearson correlation coefficients across all the cell types (bottom right violin plot in Figure 2B), we observed a p-value of 0.001. These results highlight the relevance of unknown modelling when unknown cell types in the sample of interest have their methylation patterns uncorrelated with the atlas entities.

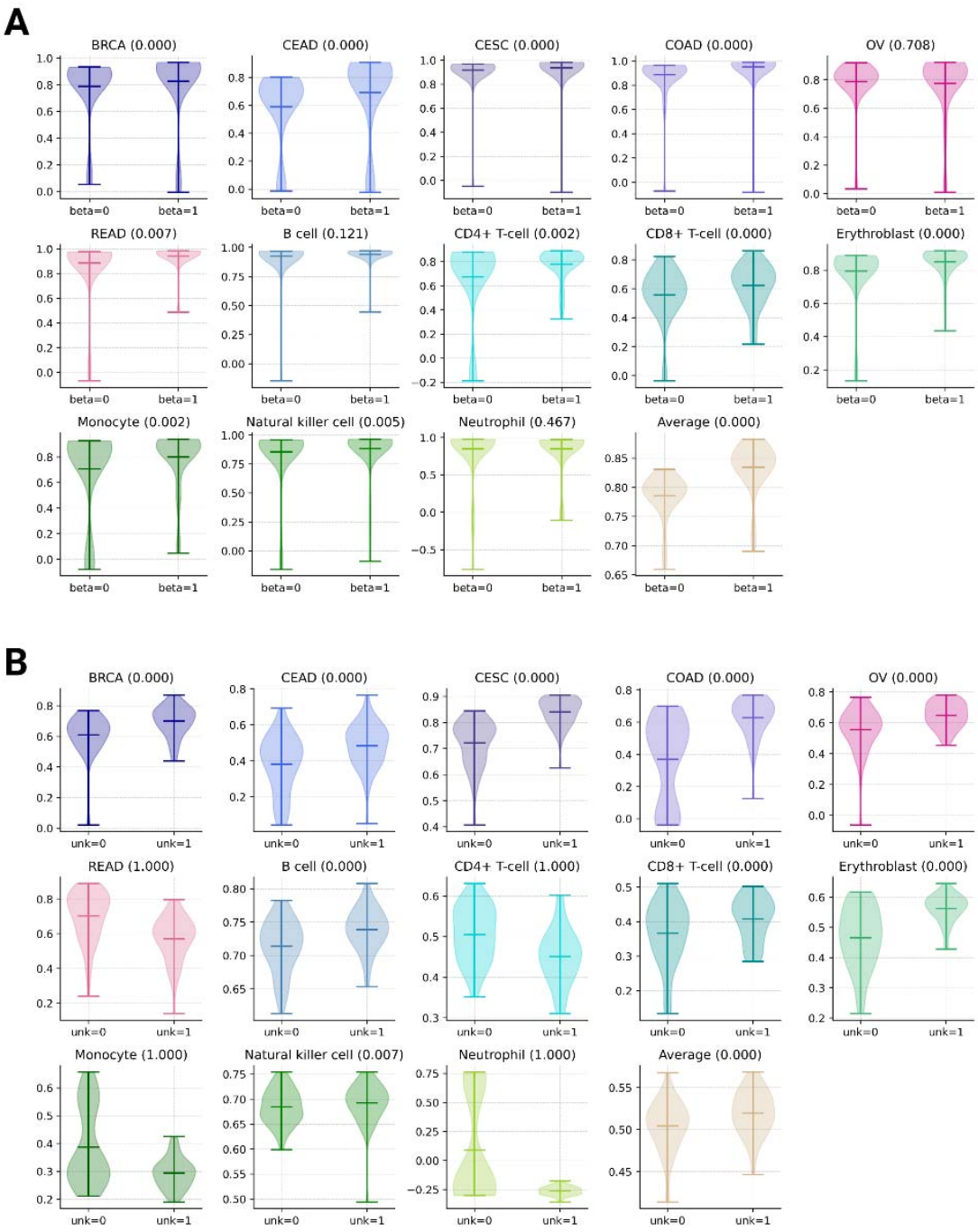


Figure 2. (A) Pearson correlation coefficients of MetDecode without ( ) and with ( ) consideration to the coverage across 30 simulation runs. For each cell type, a one-sided T-test has been performed to assess the difference in the distributions of Pearson coefficients, and the corresponding p-value reported between brackets. (B) Pearson correlation coefficients without (unk=0) and with exactly one (unk=1) unknown modelled by MetDecode. BRCA, breast carcinoma; CEAD, cervical adenocarcinoma; CESC, cervical squamocellular carcinoma, COAD, colorectal adenocarcinoma; OVCA, ovarian carcinoma; READ, rectal adenocarcinoma.

MetDecode identifies the correct TOO in genomic DNA from leukocytes and tumour tissue

To evaluate the accuracy of MetDecode for deconvoluting and assigning the correct tissue type, we applied the method on 12 WBC-derived gDNA methylomes from healthy controls (mean age: 48.08 years; range: 22-77; M/F:5/7) and 20 gDNA methylomes from tumour tissue biopsy of breast (n=5), colorectal (n=6), cervical (n=6) and ovarian (n=3) cancer (Figure 3).

When deconvoluting, the methylomes are distributed amongst different atlas entities. When the major contributor amongst all the atlas entities was the expected tissue, the assignment was considered correct. The healthy controls were considered as correctly assigned when neutrophils were deconvoluted as the main contributor [36]. MetDecode assigned the correct tissue in 29 out of 32 samples (Overall Accuracy: 90.63%). All WBC-derived samples showed neutrophils as the main contributor. All 5 breast tumour samples and 6 colorectal samples were assigned to their respective cancer. In addition, 5/6 and 1/3 cervical and ovarian tumours were classified correctly. One of the 6 cervical samples was classified as an ovarian tumour. In addition, two out of 3 ovarian tumour samples (n=2 clear cell carcinoma) were misclassified as colorectal cancers.

To assess the accuracy of mixture deconvolution, we compared the results of the WBC deconvolution to Complete Blood Counting (CBC) using matched blood samples. We observed a high correlation for the neutrophils fraction ($r=0.879$, p-value<0.001, Figure 3B), when comparing CBC and MetDecode deconvolution estimates. Lower correlation was found for lymphocytes and monocytes ($r=0.60$; $r=0.48$, respectively). However, this is similar to other reports [34,37]. In conclusion, these results demonstrate that MetDecode can identify major contributors in samples containing mixture of cell-types.

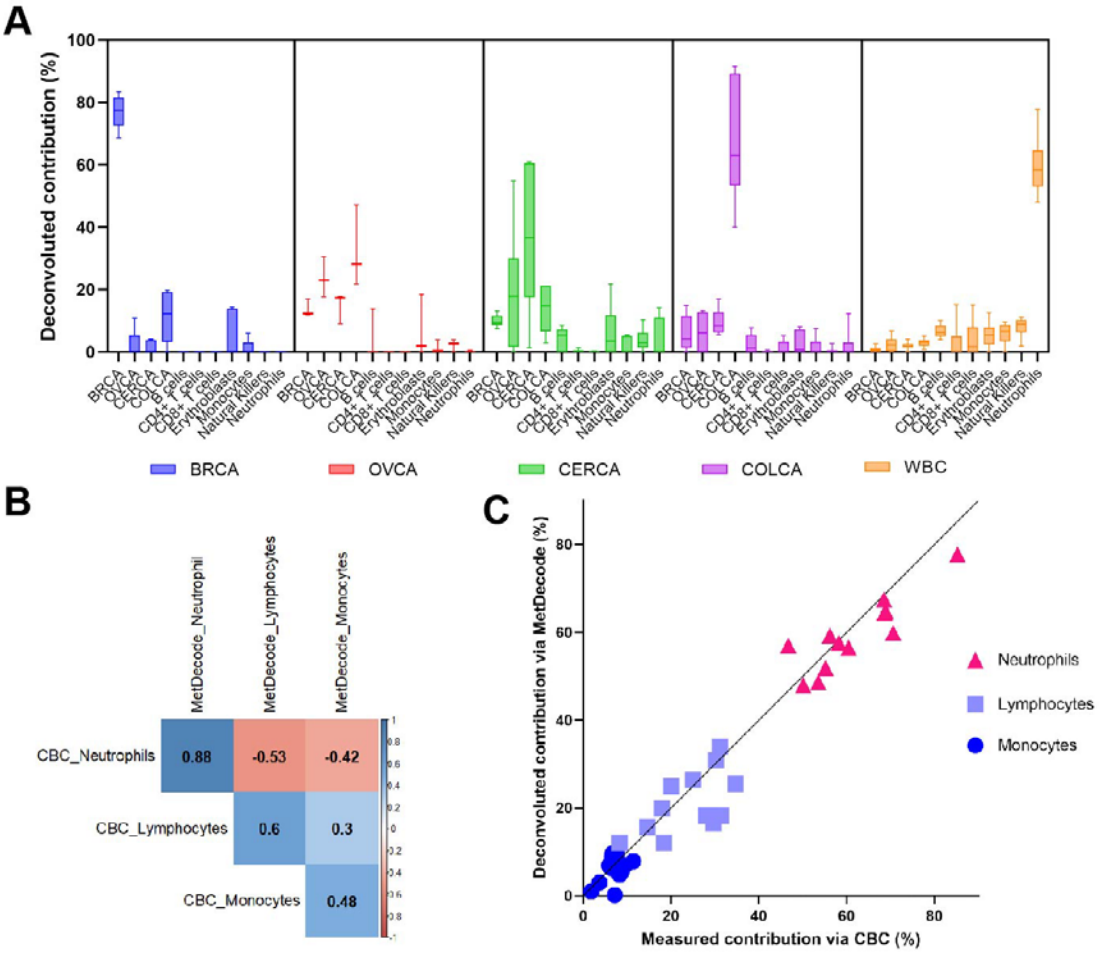


Figure 3. Deconvolution of genomic DNA methylation profiles using MetDecode. (A) Deconvolution of gDNA derived from WBC and tumour FFPE biopsies. For each sample group, the deconvoluted contribution from every atlas entity is shown. The box represents the interquartile range, the extremity represents the minimum and the maximum value. The median is marked with a horizontal bar inside the box. WBC, white blood cells; BRCA, breast carcinoma; CERCA, cervical carcinoma; COLCA, colorectal carcinoma; OVCA, ovarian carcinoma. (B) Pearson correlation between deconvoluted contributions measured via complete blood count (CBC) and MetDecode for neutrophils, monocytes and lymphocytes. (C) Relative contribution measured via CBC (X-axis) vs the deconvoluted proportion estimated using MetDecode (Y-axis) for the neutrophils, lymphocytes and monocytes in the gDNA derived from WBC. The diagonal line represents the identity line.

Metdecode allows accurate deconvolution of in-silico mixes

We next evaluated the limit of detection and assessed the concordance in the relative contribution estimation of MetDecode using *in-silico* mixtures of tumour DNA. Breast, colorectal, cervical and ovarian cancer data was combined with cfDNA from healthy donors at ratios varying from 50-0.1% to reflect an average depth of 6X (see Additional File 1). Deconvolution of the *in-silico* mixes showed that MetDecode detects tumour DNA proportions down to 4.3%. The mean correlation between the expected and the deconvoluted percentage was r=0.8603 ($p<0.001$). The highest

correlation was obtained for the breast cancer *in-silico* mixes, followed by ovarian, cervical and colorectal (0.998, p<0.001; 0.982, p<0.001; 0.976, p<0.001 and, 0.866 p<0.001, respectively, Figure 4). The expected and estimated percentages of the spiked-in component show a strong correlation, indicating that MetDecode's relative proportion estimations are indeed reliable.

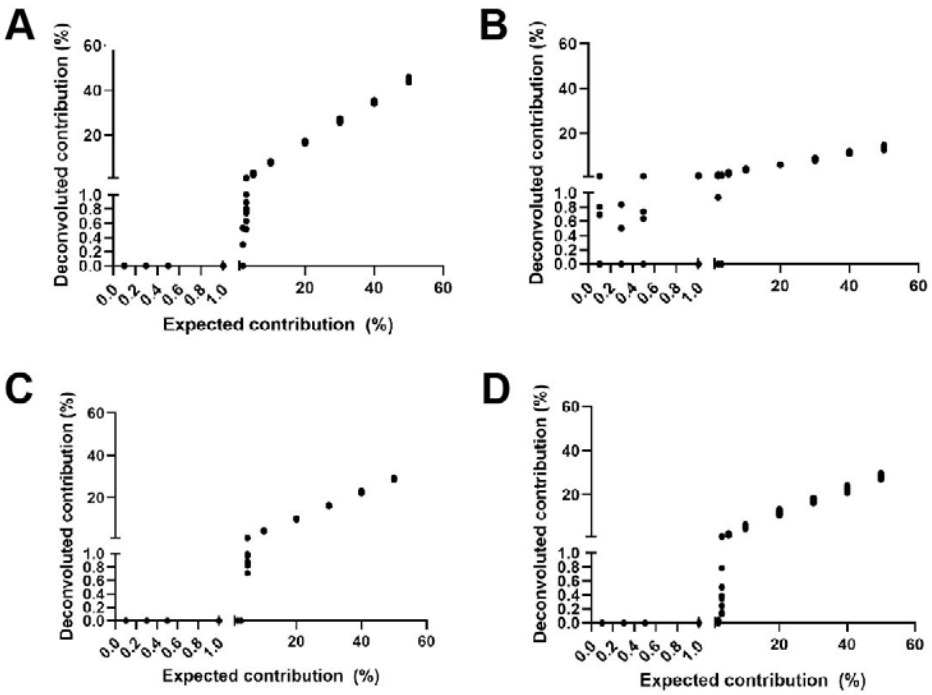


Figure 4. Correlation plots between deconvoluted and expected contribution in percentage of the tumour tissue spiked-in in the *in-silico* mixes. Random reads were combined from a healthy control BAM file and a tumour gDNA BAM file to create an *in-silico* mix and repeated 10 times to obtain replicates (A), (B), (C), (D) represent the *in-silico* mixes for breast, ovarian, colorectal and cervical cancer, respectively. Each dot represents the value for a replicate with a deconvoluted % (y-axis) vs expected % (x-axis) of tumour tissue DNA.

MetDecode correctly identified the tumour origin in cfDNA from cancer patients

MetDecode was subsequently applied to whole-genome cfDNA methylation sequencing data from healthy controls (n=93; mean age: 66.8 years; range:18-90) and treatment-naive patients with a confirmed cancer diagnosis (n=16; n=5 breast, n=4 colorectal, and n=7 ovarian cancers; Table 1). We selected samples with a minimum tumour fraction (TF) of 3% (measurement based on ichorCNA [31]).

Table 1. Clinical information of the cancer cohort and deconvolution outcome

Sample	Sex	Age	Cancer type	Cancer subtype	Stage	TOO assigned	MetDecode (%)*	TF (%)
Case 1	F	80-89	Breast cancer	Triple Negative	IIIA	BRCA	3.398	3.96
Case 2	F	50-59	Breast cancer	Luminal B	IIIA	BRCA	5.02	5.86
Case 3	F	40-49	Breast cancer	Triple Negative	IA	**NA**	**NA**	4.44
Case 4	F	50-59	Breast cancer	Her2 positive	IV	BRCA	14.014	6.27
Case 5	F	50-59	Breast cancer	Triple Negative	IV	**COLCA**	12.439	13.66
Case 6	F	80-89	Colorectal cancer	Unknown	IIIC	COLCA	33.909	10.89
Case 7	M	70-79	Colorectal cancer	Unknown	IVA	COLCA	40.852	15.14
Case 8	M	80-89	Colorectal cancer	MSS	II	COLCA	13.693	3.36
Case 9	M	60-69	Colorectal cancer	MSS	IV	COLCA	21.473	11.38
Case 10	F	50-59	Ovarian cancer	Mucinous carcinoma	IVB	**CERCA**	5.793	12.25
Case 11	F	70-79	Ovarian cancer	LGSOC	IIIC	OVCA	20.682	19.07
Case 12	F	60-69	Ovarian cancer	Endometrioid carcinoma	IA	OVCA	13.388	7.91
Case 13	F	40-49	Ovarian cancer	LGSOC	IIIC	OVCA	11.039	5.50
Case 14	F	60-69	Ovarian cancer	HGSOC	IIIC	OVCA	22.998	11.14
Case 15	F	60-69	Ovarian cancer	HGSOC	IIIA	OVCA	9.439	4.47
Case 16	F	50-59	Ovarian cancer	HGSOC	IV	OVCA	49.481	27.12

* The number reported refers to the deconvoluted percentage of the putative TOO.
In bold are indicated the misassigned TOO.
TF, Tumour Fraction; TOO, Tissue Of Origin; MSS, Microsatellite Stable; LGSOC, Low-Grade Serous Carcinoma; HGSOC, High-Grade Serous Carcinoma, BRCA, breast carcinoma; CERCA, cervical carcinoma; COLCA, colorectal carcinoma; OVCA, ovarian carcinoma; NA, not assigned.

cfDNA data from the healthy individuals was used as control to establish the reference range of each atlas entity (Supplementary Table 2, Additional File 2). As

expected, [22,34,38] neutrophils were the major contributors to plasma cfDNA (37.51%±9.418), followed by erythroblasts (19.34%±4.625), and monocytes (18.16%±3.908) (Supplementary Figure 3, Additional File 1).

We then investigated the deconvoluted composition of the cancer patient-derived plasma cfDNA and compared it to the normal ranges in healthy controls. Overall, the tissue in which the primary tumour resides was increased in all three cancer types. The colorectal cancers showed an average 35.67-fold increase in COLCA contribution compared to the healthy controls, ovarian cancers an average 8.06-fold increase in OVCA contribution and breast cancers an average 7.8-fold increase in BRCA contribution (Figure 5A). This shows that tissue-specific signals are well classified.

We next assessed the ability of MetDecode to assign the correct TOO in these cfDNA samples from cancer patients. Among the deconvoluted values falling outside the normal range established, the highest contributor across the cancer components of the reference atlas was regarded as the putative TOO of the malignancy. Overall, MetDecode assigned the correct TOO in 13 out of 16 cancer cases (accuracy 81.25%). 100% of the colorectal and 85.71% of the ovarian cancer cfDNA samples were correctly classified. One mucinous ovarian carcinoma (stage IV) was predicted to have cervical cancer tissue as a major cancer contributor. For breast cancer, 60% of the samples were assigned to the correct tissue. One triple-negative breast cancer sample (case 5, stage IV) was misclassified as colorectal. For this sample, the deconvoluted contribution from several atlas entities, namely breast, ovary and the unknown component were also higher than normal (Supplementary Table 3, Additional File 2). This result might be caused by metastasis in multiple organs, such as liver, lung and lymph nodes. Additionally, one triple-negative breast cancer cfDNA sample (case 3, stage I) did not show any alteration compared to the controls.

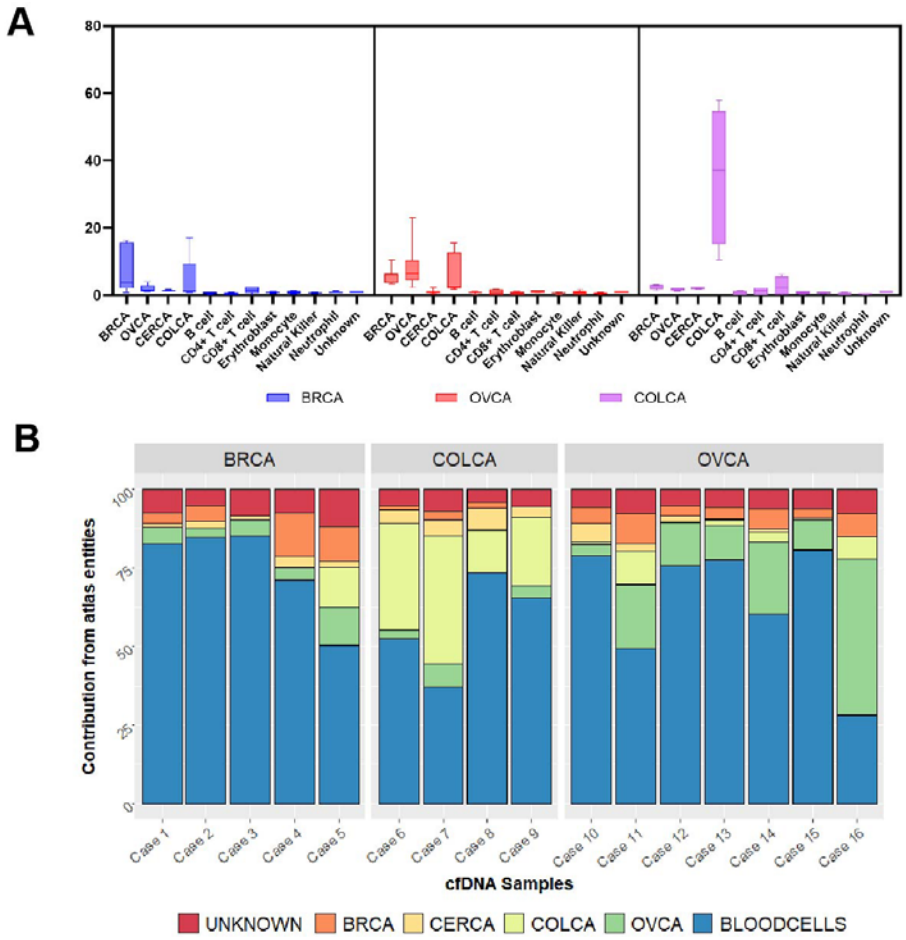


Figure 5. Deconvolution of the cfDNA derived from patients with a confirmed cancer diagnosis. (A) Fold enrichment of the deconvoluted percentages in the cfDNA from each cancer cohort was determined using healthy controls as the baseline. For each group of the cohort, namely breast, ovary and colorectal, the deconvoluted contribution of each atlas entity is represented in the box plot. The box represents the interquartile range, the extremity represents the minimum and the maximum value. The median is marked by a horizontal bar inside the box. (B) Distribution of the deconvoluted percentage in the 16 cfDNA samples from cancer patients. The contribution from the different blood cell types is summed up and shown in blue. Expected range for each atlas entity was established using mean±2SD of the contribution detected in healthy controls. An atlas entity is referred to as the tissue of origin (TOO), when the relative contribution in that specific tissue is higher than this expected range. If multiple entities fall outside the range, the highest is considered the putative TOO. BRCA, breast carcinoma; CERCA, cervical carcinoma; COLCA, colorectal carcinoma; OVCA, ovarian carcinoma

Discussion

While there are huge efforts to enable cfDNA-based early multi-cancer detection, the utility of such tests is being explored [39]. Here, we present a novel method for methylation-based cfDNA deconvolution to aid the identification of cancer types. MetDecode combines a methylome reference with a novel algorithm which can

model for unknown contributors absent in the reference and is mindful of the coverage of each marker in the reference. In addition to accurately estimating the tumour proportion in in-silico mixtures ($r=0.8603$, $p<0.001$), the method also assigned the correct TOO in 81.25% of the cfDNA samples from cancer patients.

Other deconvolution methods using epigenetic markers have been developed. MethAtlas [22] and cfNOMe [28], built on non-negative least squares and constrained programming respectively, modelled cfDNA mixtures as perfect linear combinations of the reference cell types. cfNOMe performed a multimodal analysis by complementing methylation with nucleosome occupancy profiles. To account for incomplete reference atlases, CelFiE [23] extended previous approaches by inferring the methylation patterns of unknown cell types from the data directly using a probabilistic model. MethylCIBERSORT [40] built on support vector regression (SVR) to perform robust deconvolution and discard the effect of markers with low reconstruction error whereas ARIC [24], also based on SVR, introduced a feature selection step to remove redundant markers, using condition numbers as a measure of collinearity. MethylResolver [25], on the other hand, alleviated the effect of outliers by using a least trimmed squares approach. CancerLocator [20] used a probabilistic model to estimate the tumour burden and identify the correct cancer type. CancerDetector [21] achieved higher sensitivity than CancerLocator by performing cancer classification at the level of individual sequencing reads. However, neither of these two methods allows full deconvolution of white blood cells but rather estimates the cancer proportions alone, therefore limiting the interpretability of the results. Finally, MeDeCom [26] is a reference-free approach based on regularized matrix factorization.

Among all these methods, CelFiE is the first and only full reference-based technique proposing to tackle issues related to both the non-completeness of the atlas and the noisy nature of sequencing data. However, the number of parameters in CelFiE's underlying model scales to the number of markers and (unknown) cell types, therefore exposing the method to overfitting risks. Full tissue–type deconvolution with modelling of unknowns has, to our knowledge, only been proposed in CelFiE and has not been properly addressed in the literature, as described above. Here we propose a novel computational method, coined MetDecode, to disentangle the methylation patterns of different cell types contributing to cfDNA mixtures, while

accounting for the potential incompleteness of the reference atlas and the inaccurate estimation of methylation ratios in low-coverage regions.

Incorporating these algorithms into our pipeline was not practically feasible, since the peculiarities of our data make them mostly unsuitable for most of these deconvolution approaches. For example, MeDeCom [26] was not designed to handle a reference atlas, as the method is unsupervised. MethAtlas [22], like many other methods, does not account for the coverage, as the method expects methylation array data as input instead of sequencing data. Finally, CfNOMe [28] requires nucleosome occupancy profiles which we did not compute, as such profiles are not handled by some of the other tools (e.g. CelFIE [23]).

The deconvolution of in-silico mixes affirms the performance of our deconvolution method. With respect to the gDNA samples, 2 ovarian carcinomas were misclassified as colorectal cancer. We hypothesize that the misassignment results from not including different ovarian carcinoma subtypes to build the reference atlas. In fact, the subtype used to build the reference atlas was high-grade ovarian carcinoma, while the three gDNA test samples were classified as clear cell ovarium carcinoma (n=2) and mucinous carcinoma (n=1). Similar to what was observed in the deconvolution of gDNA, one cfDNA sample from an ovarian cancer patient was not correctly classified. We hypothesize that the misassignment is a consequence of the absence of these ovarian carcinoma subtypes in the reference atlas and hence remains unrecognised [41,42].

A way to overcome this limitation might reside in using cell-type specific methylation data for the reference atlas creation [38]. The development of cell-type-based methylome atlases might provide an opportunity to dissect the different contributing cell types and may well outperform the methylome markers based on bulk tissue-specific entities. Cell-type specific methylation would ensure more precise deconvolution and offer clearer insight into the origin of the tumour. However, cell-type-based methylation data is not yet available for the cancer tissues of interest [22]. Similarly, including different subtypes for the marker selection could potentially allow subtype identification and improve overall cancer diagnosis. Given the small number of samples available, no clear correlation between (mis)classification and cancer stage could be drawn.

Unique to our approach of selecting methylation markers is that we select regions with a methylation pattern distinct to one cell or tissue type and depending on the aims of the end user, it can be applied to different tissues or cell types. Existing marker selection approaches seek to maximise the difference between methylation ratio and the median of the ratio across all tissues. This does not necessarily ensure that only one tissue or cell type is differentially methylated compared to the rest of the entities in the reference atlas [23]. The limitation of our approach is that the number of markers may reduce with an increment in the number of atlas entities. Additionally, our methylome atlas was built with a limited number of samples per atlas entity. We envision that increasing the number of samples per atlas entity may improve specificity of the selected methylation markers.

Conclusions

Deconvolution of the cfDNA epigenetic signatures is an elegant approach to deduce the TOO or cancer-type. To estimate the contributions and the type of cancer and white blood cell types in a cfDNA sample, we developed MetDecode, a methylome reference-based deconvolution algorithm. MetDecode can model the unknown contributors unavailable in the reference and account for the coverage of each marker region to alleviate the potential sources of noise.

Despite the limited sample size, the results presented here are encouraging and important for the future of liquid biopsy clinical application. In fact, a tool able to pinpoint the TOO of a malignancy can be used to streamline the diagnostic process in cancer patients. Emblematic cases in which the TOO detection via cfDNA can be of clinical utility are in the detection of cancer-like signals in maternal blood during routine non-invasive prenatal screening and in case of metastatic tumours of unknown primary. Deconvoluting and defining the TOO will aid the oncologists in identifying the tumour and direct treatment, streamlining the diagnostic process, especially in cases in which invasive examinations and radiological investigation are not ideal. Furthermore, if specific immune characteristics of the malignancy could be detected thanks to the blood-derived entities residing in the atlas, an important dowel for the treatment decision of the patient can be added simultaneously to the TOO identification. To conclude, we developed a method for deconvoluting the

components of plasma enabling detection of cancer origin using tissue-specific methylome information.

List of abbreviations

cfDNA - cell-free DNA

TOO - tissue of origin

CNA - copy number alterations

FFPE - formalin-fixed paraffin-embedded

WGBS - whole genome bisulfite sequencing

NNLS - non-negative least squares

MSE - mean squared error

HGSOC - high-grade serous ovarian carcinoma

WBC - white blood cells

CBC - complete blood counting

TF - tumour fraction

MSS - microsatellite stable

LGSOC - low-grade serous carcinoma

BRCA - breast carcinoma

CERCA - cervical carcinoma

COLCA - colorectal carcinoma

OVCA - ovarian carcinoma

NA - not assigned

SVR - support vector regression

Declarations

Ethics approval and consent to participate

The study was approved by the ethical committee of the University Hospitals Leuven (study protocols S62285, S62795, S63983, S66450, S59207 and S51375).

Consent for publication

Not applicable

Availability of data and materials

The data that support the findings of this study is in controlled access data storage in EGA under EGAS00001007493 and is available upon reasonable request. MetDecode is available on GitHub as a Python package: https://github.com/AntoinePassemiers/MetDecode

Competing interests

The authors declare that they have no competing interests

Funding

This study was supported by the Research Foundation-Flanders (FWO-Vlaanderen; 1S74420N to ST; 1SB2721N to AP, 12Y5623N to DR), Agentschap Innoveren en Ondernemen (VLAIO; Flanders Innovation & Entrepreneurship grant HBC.2018.2108 to TJ), Stichting tegen Kanker (STK grant 2018-134 to JRV and FA), Kom op tegen Kanker (KotK grant 2018/11468 to JRV and FA, KotK grant 2016/10728/2603 to AC). DT is Senior Clinical Investigator FWO - Fund for Scientific Research Flanders. GF is recipient of a post-doctoral mandate sponsored by KOOR of the University Hospitals Leuven. European Union's Horizon 2020 research and innovation program under grant agreement No 824110 - EASI-Genomics (JRV) and Institutional support from the KU Leuven, C1- C14/18/092, C14/22/125 to JRV and C3/20/100 to JRV and YM.

Authors contribution

D. Sudhakaran, S. Tuveri, T. Jatsenko, L. Lenaerts, and J.R. Vermeesch conceptualized and designed the study. S. Tuveri, T. Jatsenko, L. Lenaerts, Laga T., Punie K., Tejpar S., Coosemans A., Testa A., Ficherova D., Nieuwenhuysen E, Timmerman D. and Amant F. carried out clinical sample collection. G. Floris, A.S.

Van Rompuy and Sagaert X. performed diagnosis on the tissue and selected the paraffin material to be used in the study. S. Tuveri performed the wet-lab tasks and coordinated sequencing of cfDNA and gDNA. D. Sudhakaran designed the marker selection and performed bioinformatics analysis. A. Passemiers, D. Raimondi and Y. Moreau conceptualized the algorithm. A. Passemiers designed and implemented the algorithm. S. Tuveri, T. Jatsenko, L. Lenaerts, and D. Sudhakaran contributed to the interpretation of results. S. Tuveri, D. Sudhakaran, and A. Passemiers wrote the manuscript; all co-authors reviewed the manuscript.

Acknowledgements

We would like to thank the patients and healthy blood donors for their availability and Dr Leen Vancoillie and Dr Ilse Parijs for their help in extracting cfDNA from plasma samples.

References

1. Aucamp J, Bronkhorst AJ, Badenhorst CPS, Pretorius PJ. The diverse origins of circulating cell-free DNA in the human body: a critical re-evaluation of the literature. Biological Reviews [Internet]. 2018;93:1649–83. Available from: http://doi.wiley.com/10.1111/brv.12413

2. Schwarzenbach H, Hoon DSB, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients [Internet]. Nat Rev Cancer. 2011. p. 426–37. Available from: www.nature.com/reviews/cancer

3. Chen D, Xu T, Wang S, Chang H, Yu T, Zhu Y, et al. Liquid Biopsy Applications in the Clinic. Mol Diagn Ther [Internet]. 2020;24:125–32. Available from: https://pubmed.ncbi.nlm.nih.gov/31919754/

4. Chan KCA, Jiang P, Zheng YWL, Liao GJW, Sun H, Wong J, et al. Cancer genome scanning in plasma: detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. Clin Chem [Internet]. 2013 [cited 2023 Aug 28];59:211–24. Available from: https://pubmed.ncbi.nlm.nih.gov/23065472/

5. Song P, Wu LR, Yan YH, Zhang JX, Chu T, Kwong LN, et al. Limitations and opportunities of technologies for the analysis of cell-free DNA in cancer diagnostics. Nat Biomed Eng [Internet]. 2022 [cited 2023 Aug 28];6:232–45. Available from: https://pubmed.ncbi.nlm.nih.gov/35102279/

6. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science [Internet]. 2018;359:926–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29348365

7. Lenaerts L, Vandenberghe P, Brison N, Che H, Neofytou M, Verheecke M, et al. Genomewide copy number alteration screening of circulating plasma DNA: Potential for the detection of incipient tumors. Annals of Oncology. 2019;30:85–95.

8. Meng Z, Ren Q, Zhong G, Li S, Chen Y, Wu W, et al. Noninvasive Detection of Hepatocellular Carcinoma with Circulating Tumor DNA Features and α-Fetoprotein. J Mol Diagn [Internet]. 2021;23:1174–84. Available from: https://pubmed.ncbi.nlm.nih.gov/34182124/

9. Boons G, Vandamme T, Mariën L, Lybaert W, Roeyen G, Rondou T, et al. Longitudinal Copy-Number Alteration Analysis in Plasma Cell-Free DNA of Neuroendocrine Neoplasms is a Novel Specific Biomarker for Diagnosis, Prognosis, and Follow-up. Clinical Cancer Research [Internet]. 2022;28:338. Available from: /pmc/articles/PMC9401546/

10. Gao Q, Zeng Q, Wang Z, Li C, Xu Y, Cui P, et al. Circulating cell-free DNA for cancer early detection. The Innovation. Cell Press; 2022.

11. Qaseem A, Usman N, Jayaraj JS, Janapala RN, Kashif T. Cancer of Unknown Primary: A Review on Clinical Guidelines in the Development and Targeted Management of Patients with the Unknown Primary Site. Cureus. 2019;

12. Bianchi DW, Chudova D, Sehnert AJ, Bhatt S, Murray K, Prosen TL, et al. Noninvasive Prenatal Testing and Incidental Detection of Occult Maternal Malignancies. JAMA [Internet]. 2015;314:162. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2015.7120

13. Lenaerts L, Brison N, Maggen C, Vancoillie L, Che H, Vandenberghe P, et al. Comprehensive genome-wide analysis of routine non-invasive test data allows cancer prediction: A single-center retrospective analysis of over 85,000 pregnancies. EClinicalMedicine. 2021;35.

14. Etzioni R, Urban N, Ramsey S, McIntosh M, Schwartz S, Reid B, et al. The case for early detection. Nat Rev Cancer [Internet]. 2003;3:243–52. Available from: https://pubmed.ncbi.nlm.nih.gov/12671663/

15. Sun K, Jiang P, Chan KCA, Wong J, Cheng YKY, Liang RHS, 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 [Internet]. 2015;112:E5503–12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26392541

16. 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 [Internet]. 2016;164:57–68. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26771485

17. Lehmann-Werman R, Neiman D, Zemmour H, Moss J, Magenheim J, Vaknin-Dembinsky A, et al. Identification of tissue-specific cell death using methylation patterns of circulating DNA. Proc Natl Acad Sci U S A [Internet]. 2016;113:E1826-34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26976580

18. Dor Y, Cedar H. Principles of DNA methylation and their implications for biology and medicine. Lancet [Internet]. 2018;392:777–86. Available from: https://pubmed.ncbi.nlm.nih.gov/30100054/

19. McMahon KW, Karunasena E, Ahuja N. The Roles of DNA Methylation in the Stages of Cancer [Internet]. Cancer Journal (United States). Lippincott Williams and Wilkins; 2017. p. 257–61. Available from: https://journals.lww.com/journalppo/Fulltext/2017/09000/The_Roles_of_DNA_Methylation_in_the_Stages_of.2.aspx

20. Kang S, Li Q, Chen Q, Zhou Y, Park S, Lee G, et al. CancerLocator: Non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. Genome Biol [Internet]. 2017;18. Available from: https://pubmed.ncbi.nlm.nih.gov/28335812/

21. Same M, Liu C-C, Sher L, Wong WH, Sun C, Kang S, et al. CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. Nucleic Acids Res. 2018;46:e89–e89.

22. Moss J, Magenheim J, Neiman D, Zemmour H, Loyfer N, Korach A, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat Commun [Internet]. 2018;9:5068. Available from: http://www.nature.com/articles/s41467-018-07466-6

23. Caggiano C, Celona B, Garton F, Mefford J, Black BL, Henderson R, et al. Comprehensive cell type decomposition of circulating cell-free DNA with CelFiE. Nature Communications 2021 12:1 [Internet]. 2021;12:1–13. Available from: https://www.nature.com/articles/s41467-021-22901-x

24. Zhang W, Xu H, Qiao R, Zhong B, Zhang X, Gu J, et al. ARIC: accurate and robust inference of cell type proportions from bulk gene expression or DNA methylation data. Brief Bioinform [Internet]. 2022;23. Available from: https://pubmed.ncbi.nlm.nih.gov/34472588/

25. Arneson D, Yang X, Wang K. MethylResolver-a method for deconvoluting bulk DNA methylation profiles into known and unknown cell contents. Commun Biol [Internet]. 2020;3. Available from: https://pubmed.ncbi.nlm.nih.gov/32747663/

26. Lutsik P, Slawski M, Gasparoni G, Vedeneev N, Hein M, Walter J. MeDeCom: discovery and quantification of latent components of heterogeneous methylomes. Genome Biol [Internet]. 2017;18. Available from: https://pubmed.ncbi.nlm.nih.gov/28340624/

27. Keukeleire P, Makrodimitris S, Reinders M. Cell type deconvolution of methylated cell-free DNA at the resolution of individual reads. NAR Genom Bioinform [Internet]. 2023 [cited 2023 Aug 28];5. Available from: /pmc/articles/PMC10236360/

28. Erger F, Nörling D, Borchert D, Leenen E, Habbig S, Wiesener MS, et al. CfNOMe - A single assay for comprehensive epigenetic analyses of cell-free DNA. Genome Med [Internet]. 2020;12. Available from: /pmc/articles/PMC7315486/

29. Mondelo-Macía P, Castro-Santos P, Castillo-García A, Muinelo-Romay L, Diaz-Peña R. Circulating Free DNA and Its Emerging Role in Autoimmune Diseases. J Pers Med [Internet]. 2021;11:1–14. Available from: /pmc/articles/PMC7924199/

30. Vaisvila R, Ponnaluri VKC, Sun Z, Langhorst BW, Saleh L, Guan S, et al. Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. Genome Res [Internet]. 2021;31:1280–9. Available from: /pmc/articles/PMC8256858/

31. Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat Commun [Internet]. 2017;8:1324. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29109393

32. Fernández JM, de la Torre V, Richardson D, Royo R, Puiggròs M, Moncunill V, et al. The BLUEPRINT Data Analysis Portal. Cell Syst [Internet]. 2016 [cited 2023 Aug 25];3:491-495.e5. Available from: https://pubmed.ncbi.nlm.nih.gov/27863955/

33. The Cancer Genome Atlas. https://www.cancer.gov/tcga . Accessed May 2020.

34. Fox-Fisher I, Piyanzin S, Ochana BL, Klochendler A, Magenheim J, Peretz A, et al. Remote immune processes revealed by immune-derived circulating cell-free dna. Elife [Internet]. 2021;10. Available from: /pmc/articles/PMC8651286/

35. Margalit S, Abramson Y, Sharim H, Manber Z, Bhattacharya S, Chen YW, et al. Long reads capture simultaneous enhancer-promoter methylation status for cell-type deconvolution. Bioinformatics [Internet]. 2021 [cited 2023 Aug 28];37:I327–33. Available from: https://pubmed.ncbi.nlm.nih.gov/34252972/

36. Rosales C. Neutrophil: A Cell with Many Roles in Inflammation or Several Cell Types? Front Physiol [Internet]. 2018;9:113. Available from: /pmc/articles/PMC5826082/

37. Baron U, Werner J, Schildknecht K, Schulze JJ, Mulu A, Liebert UG, et al. Epigenetic immune cell counting in human blood samples for immunodiagnostics. Sci Transl Med [Internet]. 2018;10. Available from: https://pubmed.ncbi.nlm.nih.gov/30068569/

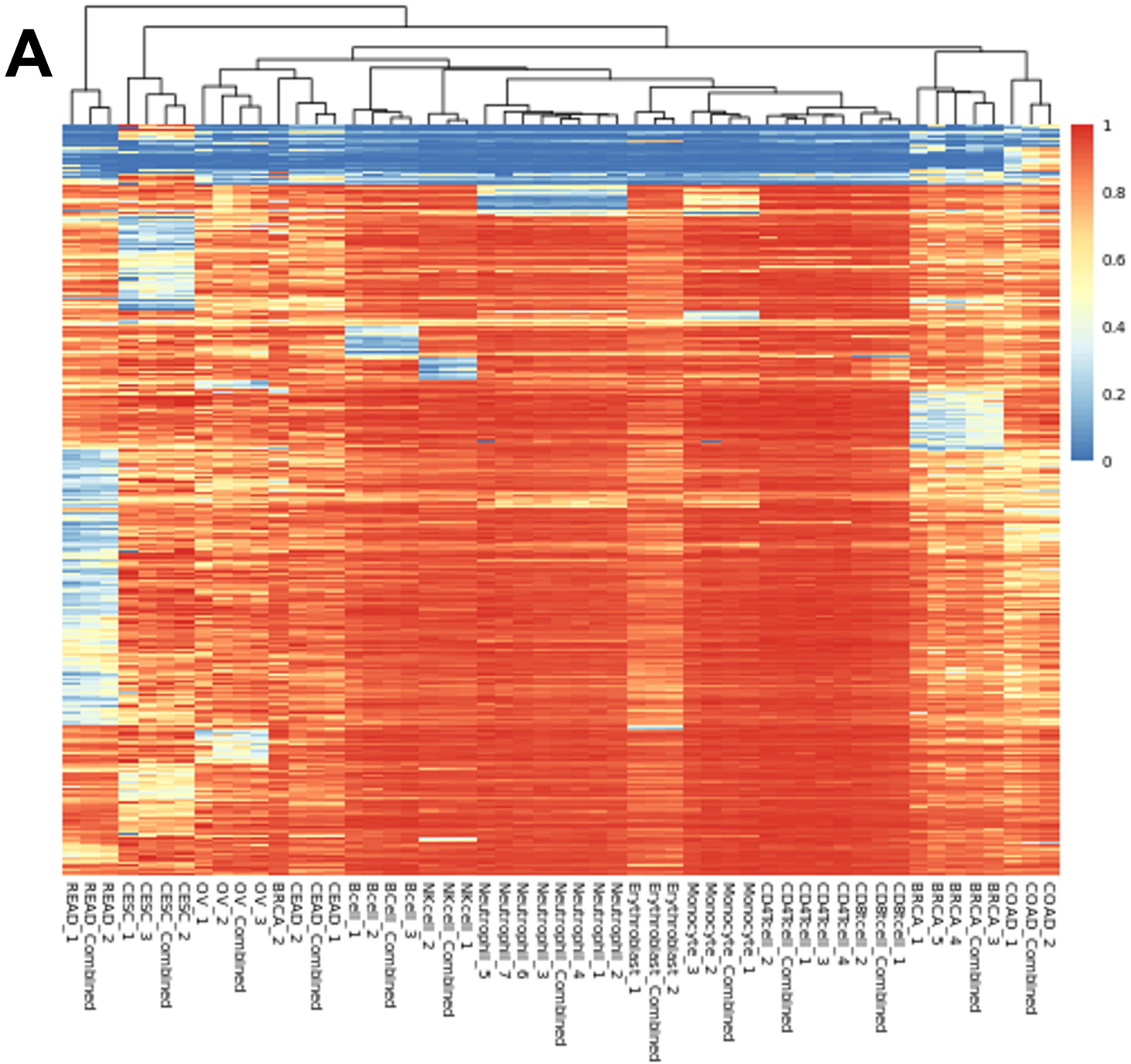
38. Loyfer N, Magenheim J, Peretz A, Cann G, Bredno J, Klochendler A, et al. A DNA methylation atlas of normal human cell types. Nature [Internet]. 2023;613:355–64. Available from: https://pubmed.ncbi.nlm.nih.gov/36599988/

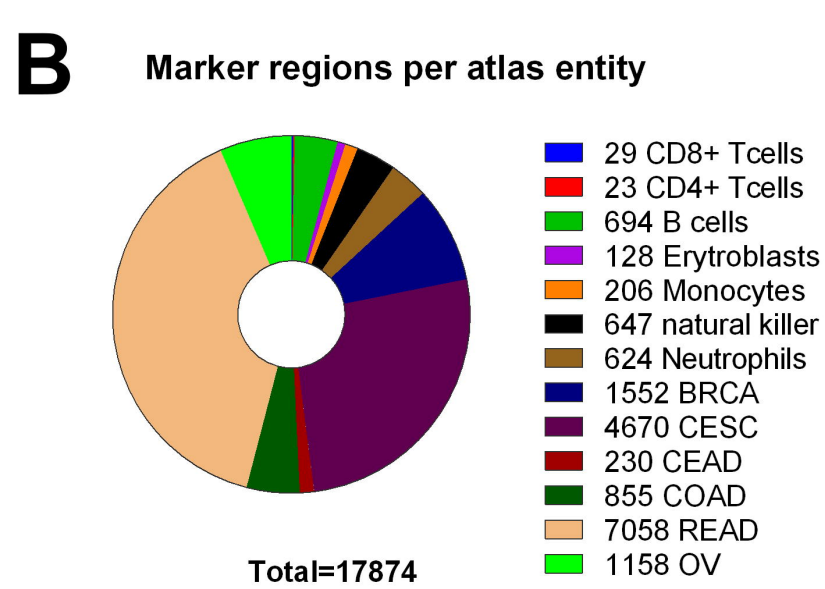
39. Bryce AH, Thiel DD, Seiden M V., Richards D, Luan Y, Coignet M, et al. Performance of a Cell-Free DNA-Based Multi-cancer Detection Test in Individuals Presenting With Symptoms Suspicious for Cancers. JCO Precis Oncol. 2023;

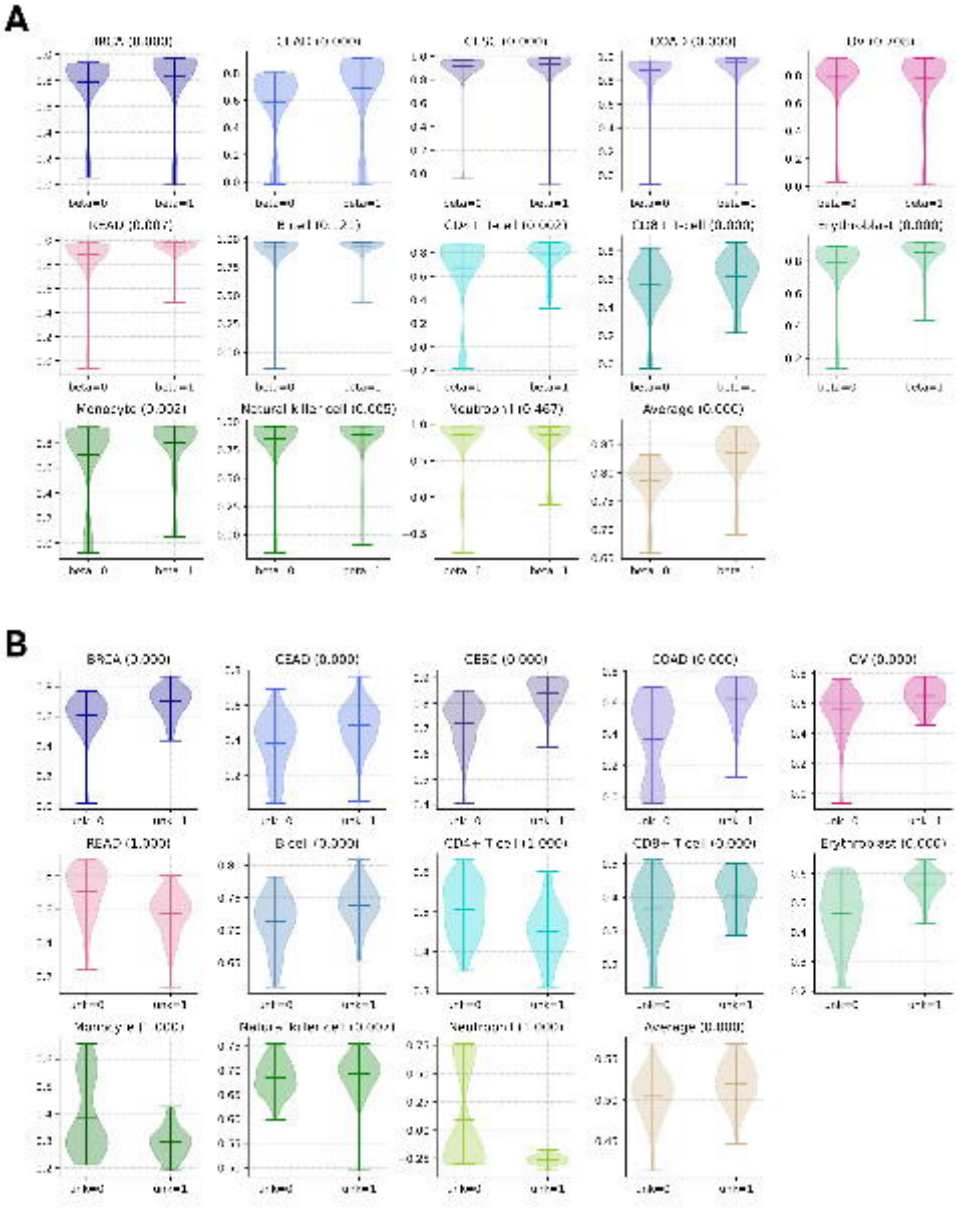
40. Filipski K, Scherer M, Zeiner KN, Bucher A, Kleemann J, Jurmeister P, et al. DNA methylation-based prediction of response to immune checkpoint inhibition in metastatic melanoma. J Immunother Cancer [Internet]. 2021;9. Available from: https://pubmed.ncbi.nlm.nih.gov/34281986/

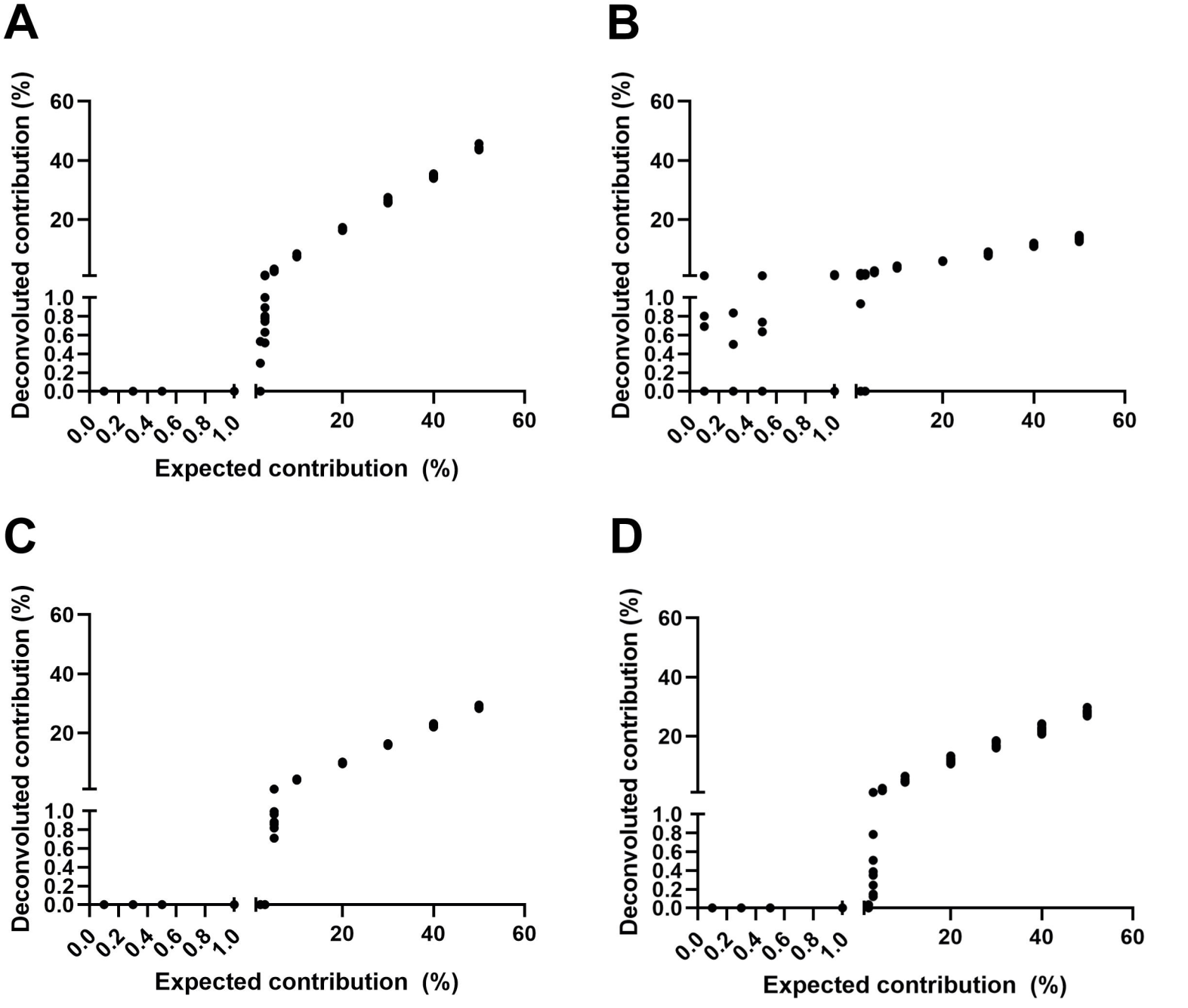
41. Köbel M, Kalloger SE, Boyd N, McKinney S, Mehl E, Palmer C, et al. Ovarian Carcinoma Subtypes Are Different Diseases: Implications for Biomarker Studies. PLoS Med [Internet]. 2008;5:1749–60. Available from: /pmc/articles/PMC2592352/

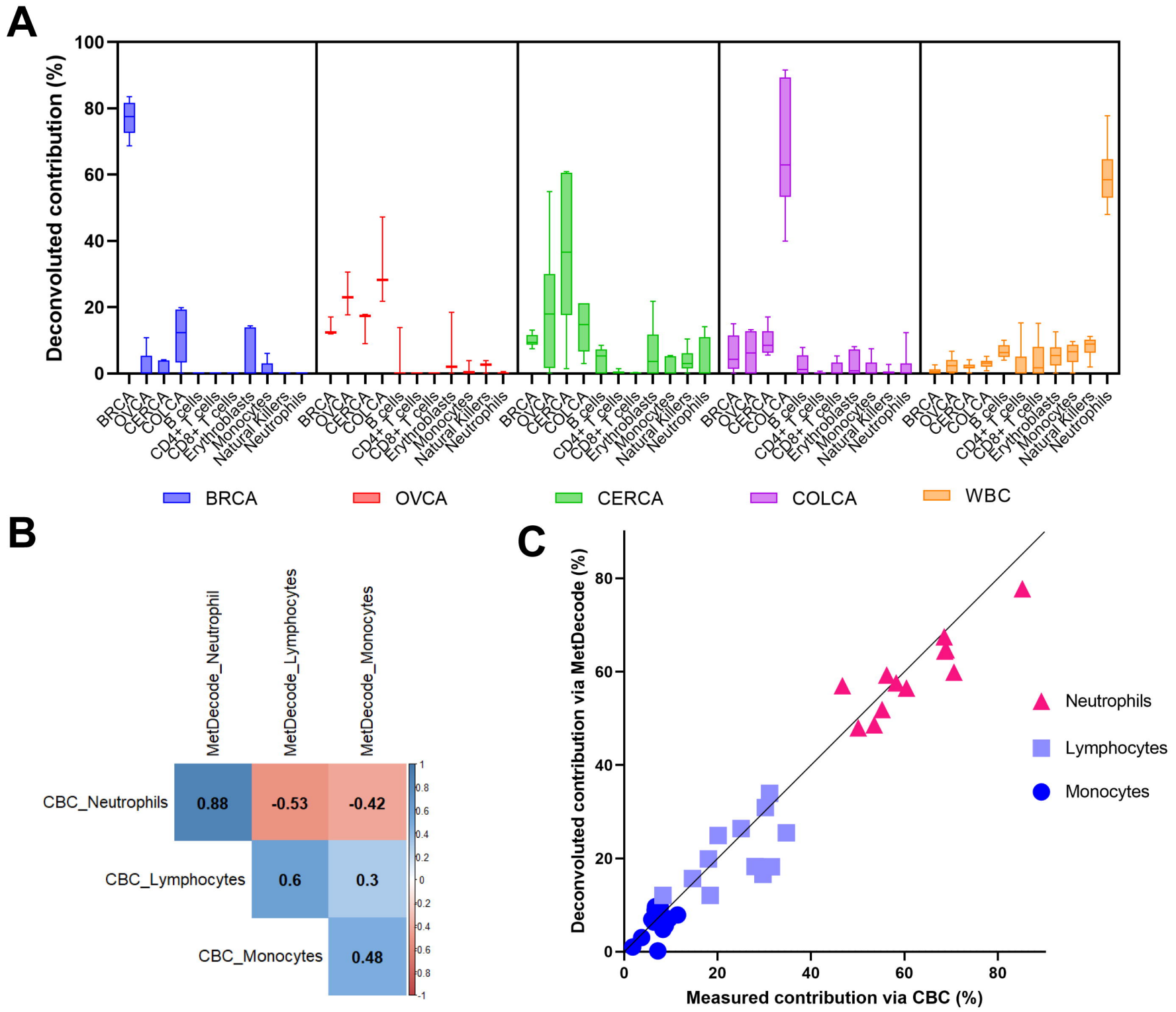
42. Doherty JA, Peres LC, Wang C, Way GP, Greene CS, Schildkraut JM. Challenges and Opportunities in Studying the Epidemiology of Ovarian Cancer Subtypes. Curr Epidemiol Rep [Internet]. 2017;4:211. Available from: /pmc/articles/PMC5718213/

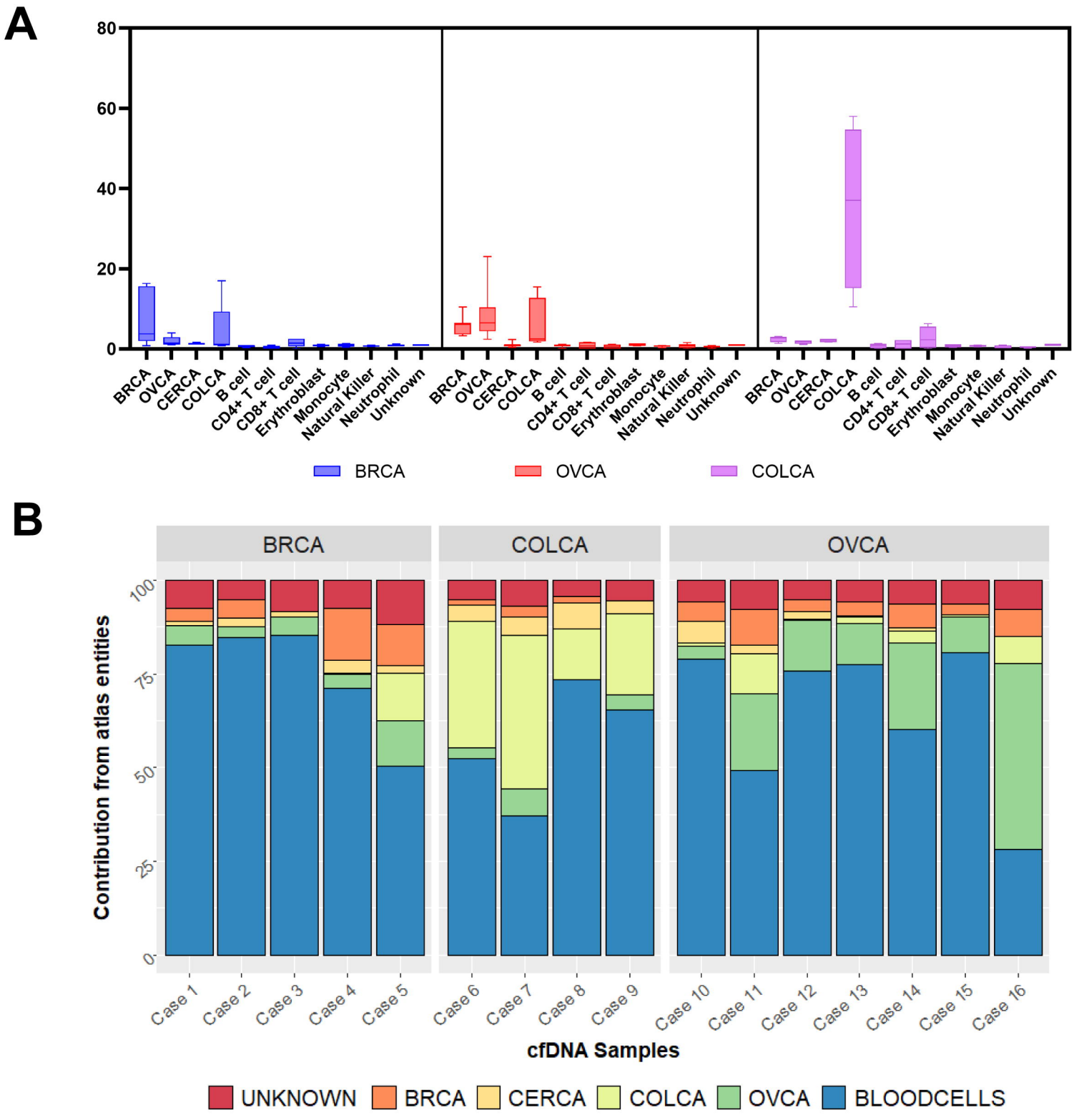

A
1
0.8
0.6
0.4
0.2
0
READ_1
READ_Combined
READ_2
CESC_1
CESC_3
CESC_Combined
CESC_2
OV_1
OV_2
OV_Combined
OV_3
BRCA_2
CEAD_2
CEAD_Combined
CEAD_1
Bcell_1
Bcell_2
Bcell_Combined
Bcell_3
NKcell_2
NKcell_Combined
NKcell_1
Neutrophil_5
Neutrophil_7
Neutrophil_6
Neutrophil_3
Neutrophil_Combined
Neutrophil_4
Neutrophil_1
Neutrophil_2
Erythroblast_1
Erythroblast_Combined
Erythroblast_2
Monocyte_3
Monocyte_2
Monocyte_Combined
Monocyte_1
CD4Tcell_2
CD4Tcell_Combined
CD4Tcell_1
CD4Tcell_3
CD4Tcell_4
CD8Tcell_2
CD8Tcell_Combined
CD8Tcell_1
BRCA_1
BRCA_5
BRCA_4
BRCA_Combined
BRCA_3
COAD_1
COAD_Combined
COAD_2


B
Marker regions per atlas entity
29 CD8+ Tcells
23 CD4+ Tcells
694 B cells
128 Erytroblasts
206 Monocytes
647 natural killer
624 Neutrophils
1552 BRCA
4670 CESC
230 CEAD
855 COAD
7058 READ
1158 OV
Total=17874


A
B


A
B
C
D
Deconvoluted contribution (%)
Expected contribution (%)
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60


A
Deconvoluted contribution (%)
100
80
60
40
20
0
BRCA
OVCA
CERCA
COLCA
B cells
CD4+ T cells
CD8+ T cells
Erythroblasts
Monocytes
Natural Killers
Neutrophils
BRCA
OVCA
CERCA
COLCA
WBC
B
MetDecode_Neutrophil
MetDecode_Lymphocytes
MetDecode_Monocytes
CBC_Neutrophils
CBC_Lymphocytes
CBC_Monocytes
0.88
-0.53
-0.42
0.6
0.3
0.48
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
C
Deconvoluted contribution via MetDecode (%)
Measured contribution via CBC (%)
0
20
40
60
80
Neutrophils
Lymphocytes
Monocytes


A
80
60
40
20
0
BRCA
OVCA
CERCA
COLCA
B cell
CD4+ T cell
CD8+ T cell
Erythroblast
Monocyte
Natural Killer
Neutrophil
Unknown
BRCA
OVCA
COLCA
B
BRCA
COLCA
OVCA
Contribution from atlas entities
100
75
50
25
0
Case 1
Case 2
Case 3
Case 4
Case 5
Case 6
Case 7
Case 8
Case 9
Case 10
Case 11
Case 12
Case 13
Case 14
Case 15
Case 16
cfDNA Samples
UNKNOWN
BRCA
CERCA
COLCA
OVCA
BLOODCELLS