


METHODOLOGY ARTICLE

Open Access

# Deconvolution of cellular subsets in human tissue based on targeted DNA methylation analysis at individual CpG sites



Marco Schmidt<sup>1,2†</sup>, Tiago Maié<sup>3†</sup>, Edgar Dahl<sup>4</sup>, Ivan G. Costa<sup>3</sup> and Wolfgang Wagner<sup>1,2\*</sup> 

## Abstract

**Background:** The complex composition of different cell types within a tissue can be estimated by deconvolution of bulk gene expression profiles or with various single-cell sequencing approaches. Alternatively, DNA methylation (DNAm) profiles have been used to establish an atlas for multiple human tissues and cell types. DNAm is particularly suitable for deconvolution of cell types because each CG dinucleotide (CpG site) has only two states per DNA strand—methylated or non-methylated—and these epigenetic modifications are very consistent during cellular differentiation. So far, deconvolution of DNAm profiles implies complex signatures of many CpGs that are often measured by genome-wide analysis with Illumina BeadChip microarrays. In this study, we investigated if the characterization of cell types in tissue is also feasible with individual cell type-specific CpG sites, which can be addressed by targeted analysis, such as pyrosequencing.

**Results:** We compiled and curated 579 Illumina 450k BeadChip DNAm profiles of 14 different non-malignant human cell types. A training and validation strategy was applied to identify and test for cell type-specific CpGs. We initially focused on estimating the relative amount of fibroblasts using two CpGs that were either hypermethylated or hypomethylated in fibroblasts. The combination of these two DNAm levels into a “FibroScore” correlated with the state of fibrosis and was associated with overall survival in various types of cancer. Furthermore, we identified hypomethylated CpGs for leukocytes, endothelial cells, epithelial cells, hepatocytes, glia, neurons, fibroblasts, and induced pluripotent stem cells. The accuracy of this eight CpG signature was tested in additional BeadChip datasets of defined cell mixtures and the results were comparable to previously published signatures based on several thousand CpGs. Finally, we established and validated pyrosequencing assays for the relevant CpGs that can be utilized for classification and deconvolution of cell types.

**Conclusion:** This proof of concept study demonstrates that DNAm analysis at individual CpGs reflects the cellular composition of cellular mixtures and different tissues. Targeted analysis of these genomic regions facilitates robust methods for application in basic research and clinical settings.

**Keywords:** Cell types, Deconvolution, DNA methylation, Epigenetic, Human, Fibrosis, Cancer, CpG, Pyrosequencing, NNLS

\* Correspondence: [wwagner@ukaachen.de](mailto:wwagner@ukaachen.de)

†Marco Schmidt and Tiago Maié contributed equally to this work.

<sup>1</sup>Helmholtz-Institute for Biomedical Engineering, Stem Cell Biology and Cellular Engineering, RWTH Aachen University Medical School, 52074 Aachen, Germany

<sup>2</sup>Institute for Biomedical Engineering – Cell Biology, University Hospital of RWTH Aachen, 52074 Aachen, Germany

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

The human body comprises hundreds of different cell types, but a clear and commonly accepted classification is still elusive [1]. The cellular characterization is usually based on ontogenetic origin within a tissue, cellular morphology, and particularly on expression of cell type-specific surface markers. These markers can also be used to isolate and purify distinct cellular subsets, e.g., by flow cytometry upon labeling with specific antibodies. However, most cell types do not have a unique panel of surface markers and bulk analysis without physical sorting masks the contribution of rare cell types [2, 3]. In the advent of single-cell omics data, e.g., by transcriptomics, ATAC-seq, or even single-cell proteomics, it is possible to discern between cells by molecular means on a cell-by-cell basis [4]. However, these methods require fresh material, they are relatively expensive, and clear demarcation of cell types remains a challenge. Alternatively, it is possible to use transcriptomic or epigenetic bulk data to estimate the cellular composition in tissues based on deconvolution algorithms [2, 5–8]. Better insight into the composition of cell types may support pathological assessment, target identification, and staging of various diseases [6, 9]. To this end, a robust, simple, and cost-effective method to estimate the cellular composition in a given tissue sample would be advantageous.

DNA methylation (DNAm) at CG dinucleotides (CpGs) is a stable and heritable modification that is directly linked to cellular differentiation [9–11]. It can be analyzed quantitatively on single base resolution and—in contrast to gene expression—every cell has only two alleles, which makes DNAm ideally suited for deconvolution approaches [12]. Amongst the first applications was the estimation of leukocyte subsets in blood [5]. More recently, it has been shown that comprehensive human cell type DNAm profiles facilitate the estimation of the origin of circulating cell-free DNA [7]. Deconvolution may either be based on a reference dataset, or it can be trained reference-free [9, 13, 14]. So far, epigenetic deconvolution was mostly based on genome-wide DNAm profiles, generated by the Illumina BeadChip technology. This method is relatively cost-effective and provides a very broad insight into genome-wide DNAm patterns. However, targeted methods for DNAm analysis, such as pyrosequencing of specific CpGs, may facilitate faster and even more cost-efficient analysis with less starting material, while reducing batch to batch variation and other technical challenges [15]. We have recently developed targeted DNAm signatures for pyrosequencing of individual CpGs to achieve deconvolution of leukocyte subsets that correlate with conventional blood counts [16]. In this study, we followed the hypothesis that the relative proportion of fibroblasts or even the complex cellular composition of human tissues can be estimated by targeted analysis of DNAm at individual CpGs.

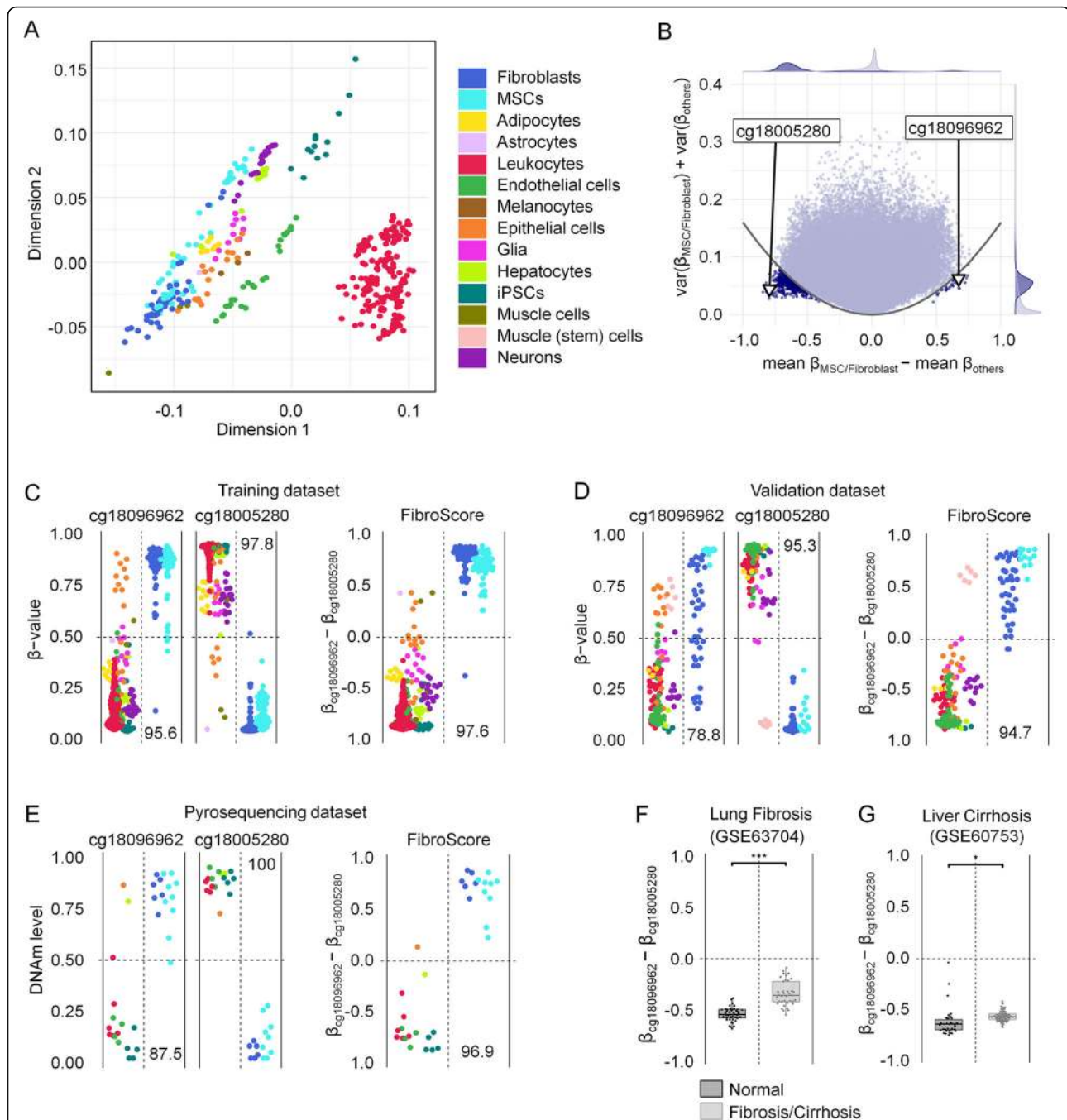
## Results

### Compilation of global DNAm profiles of different cell types

To identify cell type-specific CpGs for targeted methylation assays and tissue deconvolution, we curated and compiled 579 samples from 46 different studies, mostly generated with the Illumina 450K BeadChip technology. We only considered non-malignant samples and retrieved datasets of the following purified and characterized human cell types: fibroblasts, mesenchymal stromal cells (MSCs), adipocytes, astrocytes, leukocytes, endothelial cells, melanocytes, epithelial cells, glia, hepatocytes, muscle cells, muscle stem cells, neurons, and induced pluripotent stem cells (iPSCs). Four hundred nine samples were used as a training set and 170 samples from independent studies were used as a validation set (Additional file 1: Fig. S1A and Table S1) [7, 17–61]. Multidimensional scaling (MDS) of genome-wide DNAm profiles revealed that samples of the same cell type cluster together across different studies, supporting the notion that the cell type has major impact on DNAm patterns (Fig. 1a; Additional file 1: Fig. S1B).

### DNA methylation at fibroblast-associated CpGs can be indicative of fibrosis

Initially, we selected CpGs that might discern fibroblasts from other cell types. Such fibroblast-specific DNAm patterns could reflect the relative proportion of fibroblasts, for example for staging of fibrotic diseases. In our previous work, we addressed differences in DNAm profiles of fibroblasts versus MSCs, albeit classification of these cell types is hardly reflected by clear functional or molecular characteristics [63]. This is also reflected by their close relationship in the MDS plot. Therefore, we have decided to group both cell types together into the fibroblast category for subsequent analysis. To select fibroblast-specific CpGs that are either characteristically methylated or unmethylated in fibroblasts, we filtered for CpGs based on (1) the highest difference in mean DNAm in fibroblasts versus other cells and (2) small variance in DNAm levels within each of the two groups (Fig. 1b). CpG candidates were evaluated in terms of classification performance and ranked based on results from a 10-fold cross-validation setup. Based on this, we selected cg18096962 (associated with the lncRNA *RP11-60A8.1*) as hypermethylated and cg18005280 (associated with the gene leucine rich repeats and immunoglobulin like domains 1 [*LRIG1*]) as hypomethylated CpG site (Additional file 1: Fig. S1C,D). The difference in DNAm levels between these CpGs ( $[\beta$  value at cg18096962] –  $[\beta$  value at cg18005280]), referred to as FibroScore, could clearly distinguish fibroblasts from most other cell types (Fig. 1c, d). Only muscle stem cells, which have been differentiated for 24 h towards the myogenic lineage and



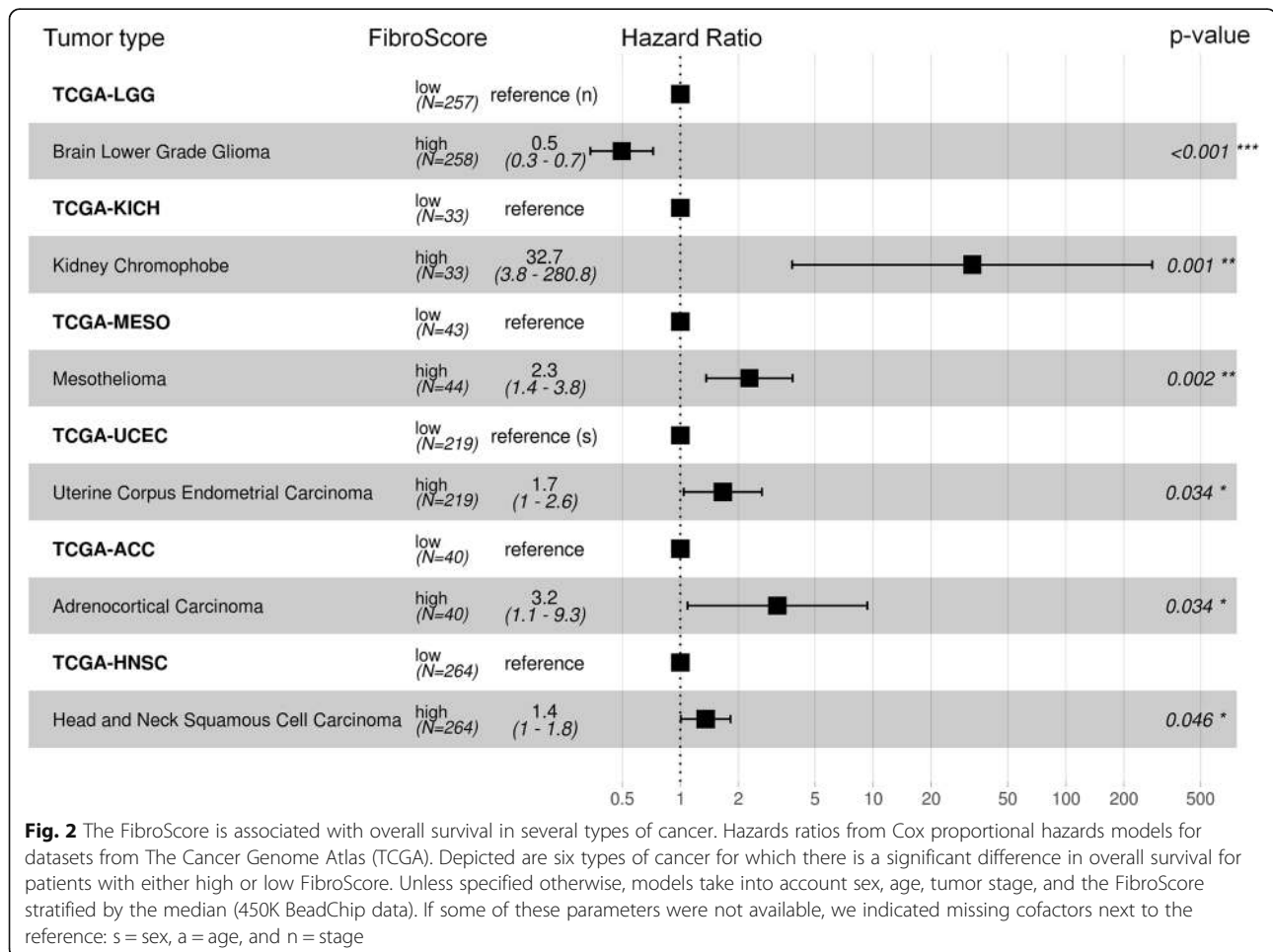
**Fig. 1** Selection of cell type-specific CpGs for fibroblasts. **a** Multidimensional scaling (MDS) plot of the training data set ( $n = 409$ ) demonstrates that samples cluster by cell type across different studies. All CpGs shared between the 450K and EPIC BeadChip were considered (except XY chromosomes). **b** Differential mean DNAm levels of fibroblasts/MSCs versus all other cell types were plotted against the sum of variances within both groups. The CpGs, which have been selected for the FibroScore, are indicated. **c** DNAm levels ( $\beta$  values) of the two selected CpGs of the FibroScore in the training set. Numbers correspond to classification accuracy in percentage values. **d** DNAm levels of the two selected CpGs and the FibroScore for the validation set. Only muscle stem cells, which might closely resemble MSCs, were classified with fibroblasts/MSCs. Numbers correspond to classification accuracy in percentage values. **e** DNAm levels of the two selected CpGs and the FibroScore as determined by pyrosequencing in samples of different cell types. Almost all cell preparations (with exception of the HaCat cell line) were classified correctly. **f** The FibroScore is significantly higher in lung fibrosis versus healthy control tissue (GSE63704; 450K data) [62]. \*\*\* $p < 0.001$ . **g** The FibroScore is significantly higher in liver cirrhosis versus healthy control tissue (GSE60753; 450K data) [29]. \* $p < 0.05$

might therefore closely resemble MSCs, were classified in the fibroblast category [50]. To further validate applicability of these CpG sites for targeted DNAm analysis, we analyzed DNA samples from cultured cells, frozen blood, and commonly used cell lines with pyrosequencing (Fig. 1e). Only one immortalized cell line was misclassified by the FibroScore: HaCat (spontaneously transformed keratinocytes for epithelial cells), which might be due to aberrant DNAm patterns by malignant transformation. Thus, targeted analysis of the two CpGs might be indicative of the fraction of fibroblasts/MSCs in tissue. In fact, when we applied the FibroScore to Illumina BeadChip datasets of lung fibrosis (GSE63704, Fig. 1f; Additional file 1: Fig. S1E) and liver cirrhosis (GSE60753, Fig. 1g; Additional file 1: Fig. S1F), we observed a significantly higher FibroScore in the fibrotic tissues as compared to healthy controls (two-sided  $t$  test:  $p = 2.51 \times 10^{-12}$ , and  $p = 0.0396$ , respectively) [29, 62].

#### FibroScore correlates with overall survival in various types of cancer

Cancer-associated fibroblasts (CAFs) determine the tumor microenvironment and play a crucial role for

progression of malignancies [64]. Therefore, we anticipated that the FibroScore might also be of prognostic relevance for various types of cancers. To address this question, we utilized 32 datasets from The Cancer Genome Atlas (TCGA) and determined the FibroScore based on the DNAm at the two relevant CpGs. For each cancer type, the patient data was then stratified by the median FibroScore. A higher FibroScore was indicative of a significantly worse overall survival in chromophobe renal cell carcinoma (TCGA-KICH,  $p = 0.001$ ), mesothelioma (TCGA-MESO,  $p = 0.002$ ), uterine corpus endometrial carcinoma (TCGA-UCEC,  $p = 0.034$ ), adrenocortical carcinoma (TCGA-ACC,  $p = 0.034$ ), and head and neck squamous cell carcinoma (TCGA-HNSC,  $p = 0.046$ ) (Fig. 2). Brain lower-grade glioma showed a significantly better survival outcome in patients with a higher FibroScore (TCGA-LGG,  $p < 0.001$ ). These results are reflected in the cox-proportional hazards adjusted survival curves for FibroScore (Additional file 1: Fig. S2). For all other cancer types, the stratification by the median FibroScore did not reveal a significant association with



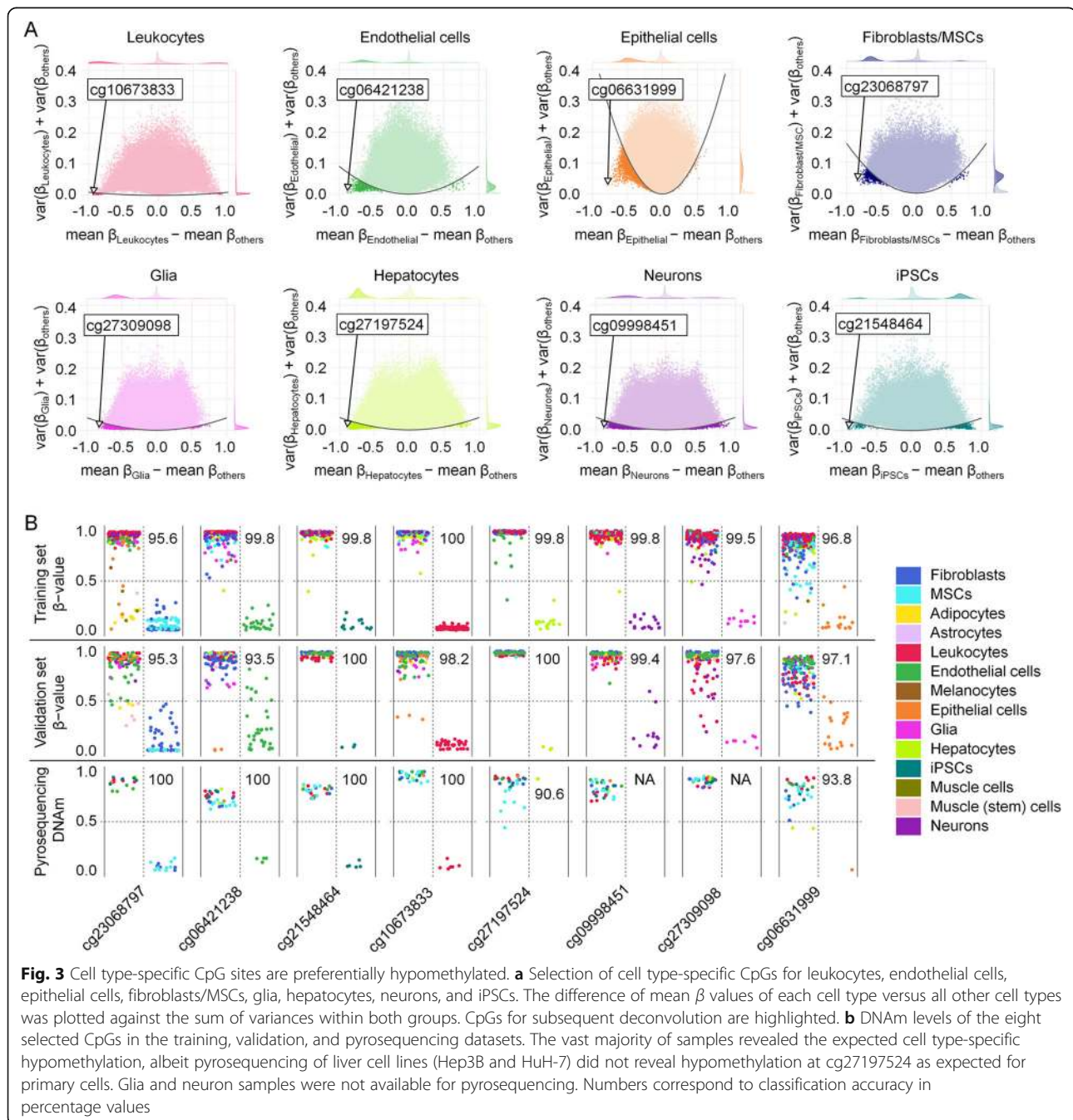


overall survival (Additional file 1: Fig. S3). These results support the notion that the DNAm at fibroblast-associated CpGs might be indicative of the fraction of CAFs, which is relevant for progression of various types of cancer.

### Deconvolution of cell types based on individual cell type-specific CpGs

Subsequently, we followed the question if targeted analysis of individual CpGs might also reflect the

composition of tissues. To this end, we have identified characteristic CpG sites for additional cell types using a similar procedure of CpG selection as mentioned above (difference in mean DNAm, variance in DNAm levels and classification performance). Notably, for all cell types—except for iPSCs, which resemble a ground state of non-differentiated cells—we identified more hypomethylated than hypermethylated CpGs in our feature selection (Fig. 3a). One hypomethylated CpG site was selected for every cell type, most of which were within



introns and exons of corresponding genes (Additional file 1: Fig. S4): cg23068797 (*DNM2*, dynamin-2) for fibroblasts, cg10673833 (*MYOIG*, myosin IG) for leukocytes, cg06631999 (*STMNI*, stathmin) for epithelial cells, cg27197524 (*POLE*, DNA polymerase epsilon catalytic subunit A) for hepatocytes, cg06421238 (*WSCDI*, WSC domain-containing protein 1) for endothelial cells, cg27309098 (*AGAPI*, Arf-GAP with GTPase, ANK repeat and PH domain-containing protein 1) for glia, cg09998451 (*RAB3A*, ras-related protein Rab-3A) for neurons, and cg21548464 (lncRNA *DLEUI*, deleted in lymphocytic leukemia 1) for iPSCs [65, 66]. Cell type-specific hypomethylation was validated with the Illumina BeadChip data from the validation set and by pyrosequencing of various cell types and tissues (Fig. 3b). To estimate if the cell type-specific differential DNAm might also be reflected in gene expression levels, we utilized the Primary Cell Atlas [67]. In fact, *MYOIG*, *WSCDI*, *RAB3A*, and *DLEUI* seemed to reveal cell type-specific upregulation, albeit only the CpG for *MYOIG* was located in the promoter region (Additional file 1: Fig. S5).

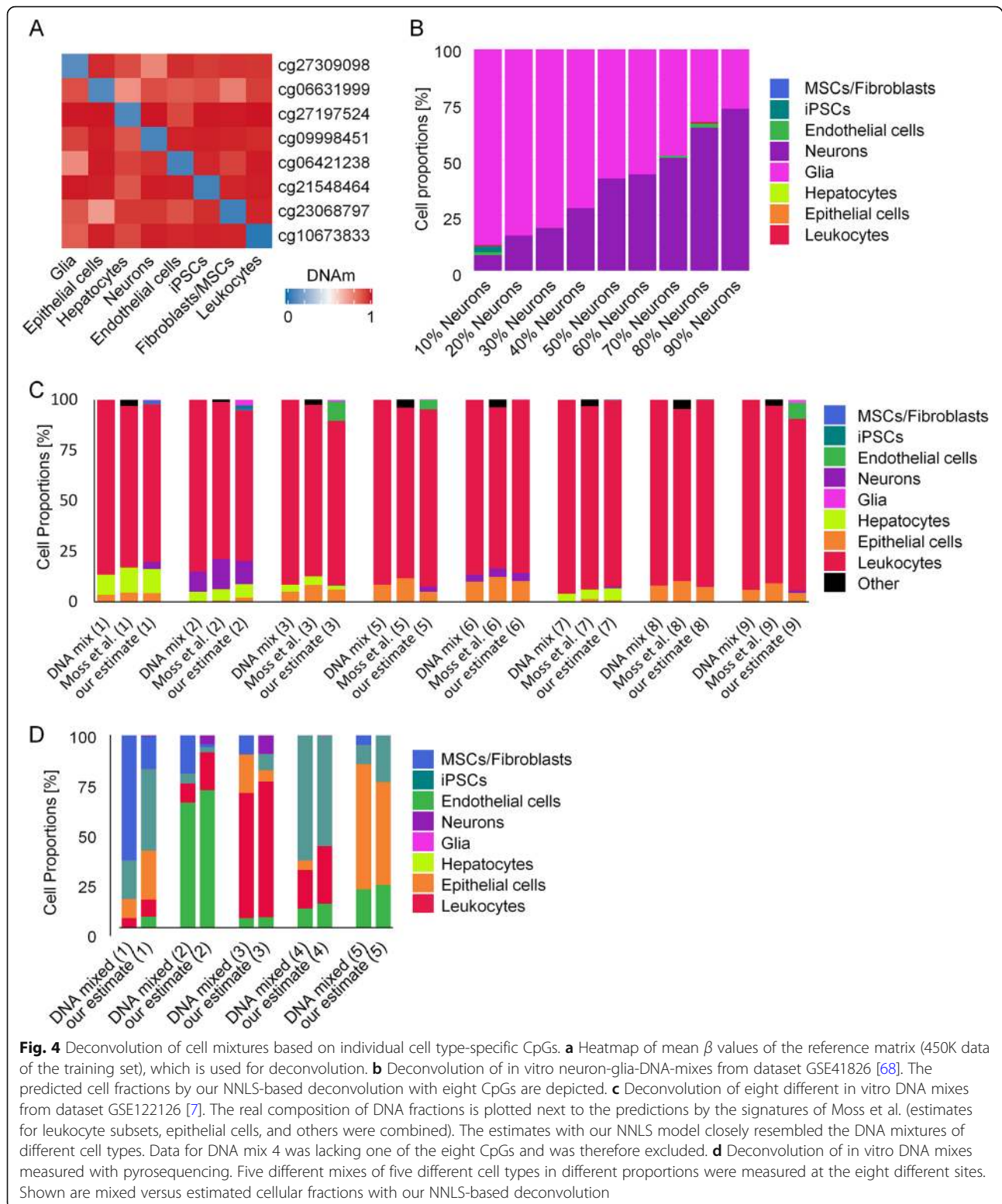
We used the mean DNAm levels for the selected CpGs in eight distinct cell types in the training dataset as our reference matrix when applying the non-negative least squares (NNLS) deconvolution algorithm (Fig. 4a; Additional file 2). The NNLS algorithm could then be used to generate estimates for the cellular composition of tissues and DNA mixes, based on the DNAm of eight CpGs. To assess the performance of our deconvolution model, we used a 450K Illumina BeadChip dataset of neuron-glia-DNA-mixes in incremental proportions [68]. Our predictions correlated very well with the neuronal/glia proportions, with only a small fraction of other cell types being predicted as present (Fig. 4b; Additional file 1: Fig. S6A). Alternatively, we tested non-negative matrix factorization (NMF) and EpiDISH, which performed not as well as the NNLS approach (Additional file 1: Fig. S6A,B) [69, 70]. Next, we tested the deconvolution performance on 450K data from in vitro DNA mixes of various different cell types [7]. Our training set did not include categories for lung and colon epithelial cells and therefore we assigned them to our epithelial cell category. Again, the predictions of our NNLS approach overall closely represented the real composition of cell types and the results were similar to the previously described deconvolution results that considered about 6000 CpGs [7] (Fig. 4c; Additional file 1: Fig. S6C). We then tested if deconvolution of different cell types would also be feasible by targeted methods. Therefore, we prepared five in vitro DNA mixes of five different cell types in varying proportions and analyzed all cell type-specific CpGs by pyrosequencing. The estimated composition closely resembled the previously mixed fractions (Fig. 4d; Additional file 1: Fig. S6D).

Validation of our deconvolution approach for complex tissue was hampered by the availability of DNAm profiles for samples with a defined composition of cell types. Therefore, we applied our deconvolution to various non-malignant tissue samples from TCGA (Additional file 1: Fig. S7A). Overall, the estimates for the different cell types are compatible with the assumed real cellular composition of the tissue. Furthermore, we have also applied our targeted pyrosequencing approach to various DNA samples from tissues, and the predictions indicated similar composition of different cell types as estimated for the Illumina BeadChip data (Additional file 1: Fig. S7B).

## Discussion

Epigenetic modifications govern cellular differentiation into specific lineages and therefore DNAm is ideally suited for cellular characterization [71–73]. Previous approaches for DNAm-based deconvolution of different cell types utilized larger signatures with multiple CpGs from Illumina BeadChip datasets [5, 7, 14]. Our proof of concept study demonstrates that estimates for the cellular composition are also feasible by targeted analysis of individual CpGs. Currently, classification of cell types is often based on antibody detection of individual epitopes—thus, estimates of the cellular composition by individual CpGs may be feasible, too.

There is always a trade-off between different methods: combining a multitude of CpGs into bioinformatic predictors generally increases the precision of epigenetic signatures [74]. On the other hand, the precision of DNAm measurements at individual CpGs is higher in pyrosequencing data as compared to  $\beta$  values on Illumina BeadChips [75]. Furthermore, the choice of regimen depends on various other aspects as cost, amount of DNA, and privacy regulations (summarized in Additional file 1: Table S2): The anticipated costs for consumables may vary considerably between different countries and institutions, but they are projected to be lower for pyrosequencing than for Illumina BeadChips. It is not trivial to compare working time as this is largely dependent on the number of samples that can be processed in parallel. However, the targeted analysis with pyrosequencing is feasible within 2 to 3 days, whereas processing and analysis of Illumina BeadChips takes longer in most core facilities. The recommended amount of genomic DNA is lower for pyrosequencing (about 10–20 ng per reaction) than for Illumina BeadChips (250–500 ng DNA, albeit also feasible with less [7]). The availability of instrumentation and of bioinformatics support needs to be considered, but again these requirements are overall lower for targeted sequencing. Furthermore, regulatory requirements, such as data protection, privacy regulations, and certification of the procedures,



may be easier met with targeted approaches. While we focused on pyrosequencing in this study, there are several alternative approaches for site-specific DNAm analysis, such as the Sequenom's EpiTYPER assay,

Single Base Primer Extension Assay (SNaPshot), droplet digital PCR (ddPCR), or bisulfite amplicon sequencing (BA-seq). In a recent study, we compared the accuracy of pyrosequencing, ddPCR and BA-seq for epigenetic age



predictions [76]: pyrosequencing provided very robust results, albeit the PCR bias might be smaller in ddPCR. The accuracy was slightly lower for BA-seq, but this method enables longer reads with more neighboring CpGs that may be considered for pattern analysis and it can facilitate multiplexing [77]. In principle, the cell type-specific genomic regions identified in this study could also be addressed by these alternative methods for site-specific DNAm analysis. Taken together, genome scale approaches with larger libraries of CpG sites as well as targeted methods have advantages and limitations, which need to be considered.

Albeit feature selection for cell type-specific CpGs did not take biological relevance into consideration, several of them were associated with potentially functionally relevant genes. DNAm and gene expression are often correlated, but not always in the expected direction of negative for promoter CpGs and positive for gene-body CpGs [22]. Particularly, the hypomethylation in cg10673833, which is located in the promoter region of *MYOIG*, may contribute to higher expression of this gene in leukocytes. It needs to be considered that regulation of gene expression is also dependent on many other regulatory mechanisms and epigenetic modifications, such as histone code, DNA accessibility, and higher order chromatin conformation. These features may also be cell type-specific, and it is conceivable they can also be utilized for cellular deconvolution in the future.

Fibroblasts are embedded into the extracellular matrix in native tissue, but there is no distinct cell marker that allows reliable quantification of this subset. Our FibroScore was significantly increased in lung fibrosis and liver cirrhosis. Targeted pyrosequencing of the two CpGs may therefore provide a simple estimate for relative changes of fibroblasts, e.g., for staging of fibrotic diseases. Furthermore, cancer-associated fibroblasts (CAFs) play a central role for tumorigenesis, progression, and metastasis in many cancers [78, 79]. It has been shown that the fraction of CAFs, which was estimated for example by the percentage of cells that stained positive for alpha smooth muscle actin, is associated with overall survival in several types of solid cancer [80, 81]. Our findings support the notion that an epigenetic fibroblast signature can support stratification of cancer samples. In the future, it will be important to better understand the epigenetic heterogeneity of CAFs and how these signatures are affected by epigenetic aberrations of the malignant clone. While the FibroScore may be indicative of the relative fraction of fibroblasts in tissue, it does not provide a quantitative measure for the percentage of fibroblasts. To this end, we have further developed our targeted approach for deconvolution of various cell types in tissue. It is difficult to access the accuracy of our NNLS-based deconvolution for tissue samples, since we

were lacking precise and validated information on their cellular composition. Nevertheless, the results of the in vitro mixes showed that deconvolution with individual cell type-specific CpGs is feasible.

A bottleneck of our analysis is the limited number of defined cellular subsets with available DNAm profiles. The lack of precise measures to distinguish between cell types is also reflected by the ongoing quest of the Human Cell Atlas Project, to define all human cell types in terms of distinctive molecular profiles and to connect this information with classical cellular descriptions (such as location and morphology) [1]. For example, fibroblasts and MSCs could possibly resemble the same type of cell [82]. On the other hand, fibroblasts are very heterogeneous and can differ greatly depending on their tissue of origin [83, 84]. For leukocyte subsets, it has been suggested that particularly cell subset-specific hypomethylation is permissive for gene expression and regulates corresponding cell functions [85]. Indeed, in our analysis, cell type-specific CpGs were predominantly hypomethylated—with the exception of iPSCs that resemble a rather non-differentiated ground state. In previous work, we have extensively studied characteristic DNAm patterns of hematopoietic subsets [16], but we have chosen to not over-represent the hematopoietic compartment in our deconvolution approach and to therefore combine all leukocytes into one category. Our hypomethylated CpGs for leukocytes, fibroblasts, endothelial cells, epithelial cells, hepatocytes, glia, neurons, and iPSCs cannot span the many facets of cellular classification but at least most cell types can be subordinated to at least one of these broad categories.

## Conclusions

Our results demonstrate that individual CpGs, which are particularly hypomethylated in specific cell types, can be used to estimate the fraction of fibroblasts or the composition of cellular mixes and tissues. In contrast to genome-wide DNAm profiles, targeted analysis, e.g., by pyrosequencing, provides new perspectives for small amounts of DNA and to derive robust procedures according to directives for in vitro diagnostic devices. Such analysis may be useful to gain insight into the composition of unknown tissue specimen or to correlate the percentage of specific cellular subsets with clinical parameters. Furthermore, it might provide estimates for the composition of cell-free DNA (cfDNA), which is increasingly relevant for liquid biopsy [7, 86].

## Methods

### Data acquisition and processing of DNAm profiles

We compiled a curated dataset of DNAm profiles (450K and EPIC Illumina BeadChip platforms) of well-characterized and non-malignant human cell types. All



analysis and data retrieval was performed with the R programming language v3.6.2 and functions from Bioconductor v3.9. The data was retrieved from Gene Expression Omnibus (through the GEOquery v2.52.0 R package (GEOquery, RRID:SCR\_000146), Additional file 1: Table S1), and data processing was performed using the minfi v1.30.0 R package (minfi, RRID:SCR\_012830) and in-house scripts. Features were limited to CpGs shared between the 450K and EPIC platforms, and we excluded probes related to sexual chromosomes and probes not shared across all samples (missing data), resulting in 415,366 CpGs for further selection. During the data acquisition process (at time of analysis), several samples and features were dropped due to conflicting or missing names, bad file formatting, and missing data. In the end, we had a total of 579 samples from 14 different cell types. For samples where raw data was available (IDAT files), ssNoob normalization method was applied [87]; otherwise, no additional normalization of beta values was performed. To avoid bias and overfitting, samples were divided into two independent datasets, a training ( $n = 409$ ) and validation set ( $n = 170$ ). Datasets from The Cancer Genome Atlas (TCGA) project (level 1 methylation array data) were downloaded and preprocessed with the TCGAbiolinks v2.12.6 (TCGAbiolinks, RRID:SCR\_017683) and SeSAME v1.2.0 packages in R (except for TCGA-STAD, which at the time of the analysis was unreachable) following their respective pipelines (Fig. 2; Additional file 1: Fig. S2) [88, 89].

Gene expression data (Affymetrix UG 133 Plus 2.0) from corresponding genes of the selected cell type-specific CpGs was extracted from [www.biogps.org](http://www.biogps.org) and the Primary Cell Atlas dataset [67]. Groups have been adjusted to fit the selected cell types. Redundant samples (e.g., time course experiments), experimentally treated samples (e.g., drugs, antibodies), tissue samples, and cells differentiated from iPSC or ES cells have been removed from the dataset (from a total of 754 samples, down to 383). If different probe sets were available for one gene, we selected the one that addressed the entire gene and correlated best with gene expression.

#### Feature selection and signatures for classification and deconvolution

In order to find the best CpGs to perform classification, we subjected the training data to a stratified  $k$ -fold cross-validation setup ( $k = 10$ ). We defined  $C_i$  as our cell of interest and  $C_{other}$  as the class that englobes all the other cell types. For a given fold, we calculate the difference in means between  $C_i$  and  $C_{other}$  (dMean) and the sum of variances within  $C_i$  and  $C_{other}$  (sVar) for each CpG. The relationship between mean and variances is exemplified in Fig. 1b, where we assume that CpGs with higher absolute dMean, and lower sVar were considered

more discriminative. To capture a set of discriminative CpGs, we define a parabola function and select all CpGs for which  $y < (ax)^2$  (Fig. 1b). Initially,  $a$  is set to 0.1. If less than 10 hypermethylated ( $x > 0$ ) or 10 hypomethylated ( $x < 0$ ) CpGs are selected,  $a$  is incremented by 0.1 until the previous criteria is reached. Next, we compute the area under the precision-recall curve (AUPR) on the remaining folds and scale it by the absolute dMean [90]. We consider here hypo- and hypermethylated CpGs separately for estimating dMean. The above procedure is repeated for each fold. A final score is obtained by the average scaled AUPR multiplied by the proportion of folds where a CpG was selected as a top scoring candidate. This is then used to obtain a final CpG ranking. This measure selects CpGs present in more folds, having higher AUPR and higher absolute dMean. The best iPSC CpG was not suitable for primer design for pyrosequencing and therefore the second best was selected for this cell type. For the FibroScore, we used the F1-score (without scaling) and selected from the best CpGs, one hypo- and one hypermethylated CpG, after initial screening with pyrosequencing.

#### Deconvolution of cell type proportions

Using the cell type-specific CpGs previously selected for classification and their mean methylation value (for each cell type) on the training dataset as our reference matrix, we applied a reference-based non-negative least-squares (NNLS) algorithm [16, 69]. An application for cell type deconvolution is provided as a separate Excel tool (Additional file 2) and as the DeconvolutionApp, <https://cost-alab.ukaachen.de/shiny/tmaie/deconapp/> (accessed 24 July 2020) [91].

#### Survival analysis

A multivariate survival analysis was performed on TCGA data using Cox proportional hazards models, taking into account (when available) sex, age, tumor stage, and the FibroScore stratified by median. Plots were created with the survival v3.1.12 and survminer v0.4.6 packages in R [92, 93].  $P$  values are based on the log-rank test.

#### Cell culture

Human mesenchymal stromal cells [94], dermal fibroblasts [84], human umbilical vein endothelial cells (HUVECs) [38], and iPSCs [94, 95] were isolated and thoroughly characterized as described in our previous work. Human cell lines HepG2, HuH-7, Hep3B, and HaCat were maintained at RWTH Aachen Medical School under standard culture for isolation of genomic DNA. For HepG2 and HaCat, DNA was directly isolated from cryopreserved vials.

### Isolation of genomic DNA and bisulfite conversion

Genomic DNA from cells and tissues was isolated with the NucleoSpin® Tissue Kit (Macherey-Nagel) and from blood (150 µl) with the QIAamp DNA Blood Mini Kit (Qiagen). DNA concentration was measured using the NanoDrop™ 2000 spectrophotometer (Thermo Scientific™) and bisulfite converted using the EZ DNA Methylation Kit (Zymo Research).

### Pyrosequencing

Bisulfite converted DNA (10–20 ng) was amplified with a region-specific biotinylated/unmodified DNA primer pair (Metabion; Additional file 1: Table S3) using the PyroMark PCR Kit (Qiagen) according to the manufacturer's instructions: Initial activation at 95 °C for 15 min, then 45 cycles of 30 s at 94 °C, 30 s at 56 °C, and 30 s at 72 °C followed by a final extension at 72 °C for 10 min. Pyrosequencing was performed on the PyroMark Q96 and the Q48 Autoprep platforms. Exemplary pyrograms are provided in Additional file 1: Fig. S8. The assay for the neuron-specific CpG site was designed for the complementary strand to stay within a more reasonable sequencing distance. The results were analyzed using the Pyro Q-CpG 1.0.9 or the PyroMark Q48 Advanced Software, respectively.

### Quantification and statistical analysis

In total, we used DNAm profiles of 579 samples from 46 different studies for training and validation sets. The pyrosequencing signatures were validated with four cell lines, 12 MSC samples, 6 fibroblast samples, 4 HUVEC preparations, 5 iPSC lines, 8 blood samples, and 14 different tissue samples. To estimate the significance of differential DNAm and FibroScore in the lung fibrosis and liver cirrhosis datasets, we utilized the two-sided *t* test: \*\*\* < 0.001, \*\* < 0.01, \* < 0.05. *P* values for overall survival in cancer are based on the log-rank test.

### Supplementary Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1186/s12915-020-00910-4>.

**Additional file 1: Figure S1.** Selection of cell-type specific CpGs for fibroblasts. **Figure S2.** Cox proportional hazards adjusted survival curves for FibroScore. **Figure S3.** Survival analysis in other types of cancer. **Figure S4.** Genomic context of cell-type-specific CpG sites. **Figure S5.** Gene expression of CpG related genes. **Figure S6.** Comparison of different deconvolution methods. **Figure S7.** Deconvolution of cell mixtures based on individual cell-type-specific CpGs. **Figure S8.** Representative pyrograms. **Table S1.** 450k/EPIC Illumina BeadChip datasets used in this study. **Table S2.** Advantages and limitations of pyrosequencing versus Illumina BeadChips. **Table S3.** Primer DNA sequences used for pyrosequencing.

**Additional file 2.** Excel application for cell type deconvolution. An application for cell type deconvolution is provided as separate Excel tool. The mean DNAm values for the cell-type-specific CpGs of the Illumina BeadChip training dataset are given as reference matrix. Furthermore, the

application allows NNLS predictions to estimate the cellular composition in independent datasets. This table was generated in analogy to the NNLS application for Epi-Blood-Count [16].

### Acknowledgements

Tissue samples used in this study were provided by the RWTH centralized biobank (RWTH cBMB) of the Medical Faculty of RWTH Aachen Medical School.

### Authors' contributions

Conception of the project: W.W. and I.C.; curation of datasets: M.S. and T.M.; analysis of DNAm profiles: T.M. and M.S.; pyrosequencing: M.S.; support for cell and tissue preparations: E.D.; the initial draft of this manuscript was written by M.S. and reviewed and edited by all authors. All authors read and approved the final manuscript.

### Funding

This work was supported by the Interdisciplinary Center for Clinical Research within the faculty of Medicine at the RWTH Aachen University (WW & IGC: IZKF O3–3), by the Deutsche Forschungsgemeinschaft (WW: WA 1706/8–1; WA1706/12–1, IC: KFO 344/1) and by the German Ministry of Education and Research (WW: VIP+ Epi-Blood-Count, 03VP06120, IC: Fibromap Consortia e:Med). Open Access funding enabled and organized by Projekt DEAL.

### Availability of data and materials

The datasets analyzed during the current study are available in the Gene Expression Omnibus (GEO): GSE34486, GSE40699, GSE41933, GSE43976, GSE50222, GSE52025, GSE52112, GSE58622, GSE59065, GSE59091, GSE59250, GSE59796, GSE60753, GSE63409, GSE65078, GSE68134, GSE71955, GSE74877, GSE77135, GSE79144, GSE79695, GSE82234, GSE85647, GSE87095, GSE87177, GSE88824, GSE92843, GSE95096, GSE98203, GSE99716, GSE103253, GSE107226, GSE11921, GSE53302, GSE68851, GSE71244, GSE74486, GSE85566, GSE86258, GSE86829, GSE87797, GSE104287, GSE106099, GSE109042, GSE111396, GSE122126, GSE41826, GSE60753, GSE63704, and The Cancer Genome Atlas (TCGA) repositories (see also Additional file 1: Table S1). A DeconvolutionApp is provided at <https://costalab.ukaachen.de/shiny/tmaie/deconvapp/> (accessed 24 July 2020) [91].

### Ethics approval and consent to participate

Whole blood and tissue samples were taken after informed consent according to the guideline of the local ethics committee (EK 206/09) and provided by the RWTH Centralized Biomaterial Bank (RWTH cBMB).

### Consent for publication

Not applicable.

### Competing interests

W.W. is cofounder of Cygenia GmbH that can provide service for analysis of epigenetic signatures ([www.cygenia.com](http://www.cygenia.com)). Apart from that, the authors declare that they have no competing interests.

### Author details

<sup>1</sup>Helmholtz-Institute for Biomedical Engineering, Stem Cell Biology and Cellular Engineering, RWTH Aachen University Medical School, 52074 Aachen, Germany. <sup>2</sup>Institute for Biomedical Engineering – Cell Biology, University Hospital of RWTH Aachen, 52074 Aachen, Germany. <sup>3</sup>Institute for Computational Genomics, Joint Research Center for Computational Biomedicine, RWTH Aachen University Medical School, 52074 Aachen, Germany. <sup>4</sup>RWTH centralized Biomaterial Bank (RWTH cBMB), Medical Faculty, RWTH Aachen University, Aachen, Germany.

Received: 28 July 2020 Accepted: 28 October 2020

Published online: 24 November 2020

### References

- Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The human cell atlas. *Elife*. 2017;6:e27041.
- Avila Cobos F, Vandesompele J, Mestdagh P, De Preter K. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics*. 2018;34(11):1969–79.

3. Kuhn A, Kumar A, Beilina A, Dillman A, Cookson MR, Singleton AB. Cell population-specific expression analysis of human cerebellum. *BMC Genomics*. 2012;13:610.
4. Roy AL, Conroy RS. Toward mapping the human body at a cellular resolution. *Mol Biol Cell*. 2018;29(15):1779–85.
5. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012;13:86.
6. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun*. 2019;10(1):380.
7. 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*. 2018;9(1):5068.
8. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*. 2015;12(5):453–7.
9. Titus AJ, Gallimore RM, Salas LA, Christensen BC. Cell-type deconvolution from DNA methylation: a review of recent applications. *Hum Mol Genet*. 2017;26(R2):R216–R24.
10. Khavari DA, Sen GL, Rinn JL. DNA methylation and epigenetic control of cellular differentiation. *Cell Cycle*. 2010;9(19):3880–3.
11. Zeng Y, Chen T. DNA methylation reprogramming during mammalian development. *Genes (Basel)*. 2019;10(4):257.
12. Houseman EA, Christensen BC, Karagas MR, Wrensch MR, Nelson HH, Wiemels JL, et al. Copy number variation has little impact on bead-array-based measures of DNA methylation. *Bioinformatics*. 2009;25(16):1999–2005.
13. Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics*. 2017;18(1):105.
14. Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics*. 2014;30(10):1431–9.
15. Delaney C, Garg SK, Yung R. Analysis of DNA methylation by pyrosequencing. *Methods Mol Biol*. 2015;1343:249–64.
16. Frobél J, Bozic T, Lenz M, Uciechowski P, Han Y, Herwartz R, et al. Leukocyte counts based on DNA methylation at individual cytosines. *Clin Chem*. 2018;64(3):566–75.
17. Bronneke S, Bruckner B, Peters N, Bosch TC, Stab F, Wenck H, et al. DNA methylation regulates lineage-specifying genes in primary lymphatic and blood endothelial cells. *Angiogenesis*. 2012;15(2):317–29.
18. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57–74.
19. Reinisch A, Etchart N, Thomas D, Hofmann NA, Fruehwirth M, Sinha S, et al. Epigenetic and in vivo comparison of diverse MSC sources reveals an endochondral signature for human hematopoietic niche formation. *Blood*. 2015;125(2):249–60.
20. Kular L, Liu Y, Ruhrmann S, Zheleznyakova G, Marabita F, Gomez-Cabrero D, et al. DNA methylation as a mediator of HLA-DRB1\*15:01 and a protective variant in multiple sclerosis. *Nat Commun*. 2018;9(1):2397.
21. Nestor CE, Barrenas F, Wang H, Lentini A, Zhang H, Bruhn S, et al. DNA methylation changes separate allergic patients from healthy controls and may reflect altered CD4+ T-cell population structure. *PLoS Genet*. 2014;10(1):e1004059.
22. Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol*. 2014;15(2):R37.
23. Fernandez AF, Bayon GF, Urdinguio RG, Torano EG, Garcia MG, Carella A, et al. H3K4me1 marks DNA regions hypomethylated during aging in human stem and differentiated cells. *Genome Res*. 2015;25(1):27–40.
24. Dahlman I, Sinha I, Gao H, Brodin D, Thorell A, Ryden M, et al. The fat cell epigenetic signature in post-obese women is characterized by global hypomethylation and differential DNA methylation of adipogenesis genes. *Int J Obes*. 2015;39(6):910–9.
25. Tserel L, Kolde R, Limbach M, Tretyakov K, Kasela S, Kisand K, et al. Age-related profiling of DNA methylation in CD8+ T cells reveals changes in immune response and transcriptional regulator genes. *Sci Rep*. 2015;5:13107.
26. Butcher LM, Ito M, Brimpari M, Morris TJ, Soares FAC, Ahrlund-Richter L, et al. Non-CG DNA methylation is a biomarker for assessing endodermal differentiation capacity in pluripotent stem cells. *Nat Commun*. 2016;7:10458.
27. Absher DM, Li X, Waite LL, Gibson A, Roberts K, Edberg J, et al. Genome-wide DNA methylation analysis of systemic lupus erythematosus reveals persistent hypomethylation of interferon genes and compositional changes to CD4+ T-cell populations. *PLoS Genet*. 2013;9(8):e1003678.
28. Zhang X, Ulm A, Sominen HK, Oh S, Weirauch MT, Zhang HX, et al. DNA methylation dynamics during ex vivo differentiation and maturation of human dendritic cells. *Epigenetics Chromatin*. 2014;7:21.
29. Hlady RA, Tiedemann RL, Puszyk W, Zendejas I, Roberts LR, Choi JH, et al. Epigenetic signatures of alcohol abuse and hepatitis infection during human hepatocarcinogenesis. *Oncotarget*. 2014;5(19):9425–43.
30. 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. *Nat Commun*. 2015;6:8489.
31. Burrows CK, Banovich NE, Pavlovic BJ, Patterson K, Gallego Romero I, Pritchard JK, et al. Genetic variation, not cell type of origin, underlies the majority of identifiable regulatory differences in iPSCs. *PLoS Genet*. 2016;12(1):e1005793.
32. Wang XM, Yik WY, Zhang P, Lu W, Huang N, Kim BR, et al. Induced pluripotent stem cell models of Zellweger spectrum disorder show impaired peroxisome assembly and cell type-specific lipid abnormalities. *Stem Cell Res Ther*. 2015;6:158.
33. Limbach M, Saare M, Tserel L, Kisand K, Eglit T, Sauer S, et al. Epigenetic profiling in CD4+ and CD8+ T cells from Graves' disease patients reveals changes in genes associated with T cell receptor signaling. *J Autoimmun*. 2016;67:46–56.
34. Holm K, Staaf J, Lauss M, Aine M, Lindgren D, Bendahl PO, et al. An integrated genomics analysis of epigenetic subtypes in human breast tumors links DNA methylation patterns to chromatin states in normal mammary cells. *Breast Cancer Res*. 2016;18(1):27.
35. Ivanov NA, Tao R, Chenoweth JG, Brandtjen A, Mighdoll MI, Genova JD, et al. Strong components of epigenetic memory in cultured human fibroblasts related to site of origin and donor age. *PLoS Genet*. 2016;12(2):e1005819.
36. Do C, Lang CF, Lin J, Darbary H, Krupska I, Gaba A, et al. Mechanisms and disease associations of haplotype-dependent allele-specific DNA methylation. *Am J Hum Genet*. 2016;98(5):934–55.
37. von der Heide EK, Neumann M, Vosberg S, James AR, Schroeder MP, Ortiz-Tanchez J, et al. Molecular alterations in bone marrow mesenchymal stromal cells derived from acute myeloid leukemia patients. *Leukemia*. 2017;31(5):1069–78.
38. Franzen J, Zirkel A, Blake J, Rath B, Benes V, Papanonis A, et al. Senescence-associated DNA methylation is stochastically acquired in subpopulations of mesenchymal stem cells. *Aging Cell*. 2017;16(1):183–91.
39. Mamrut S, Avidan N, Truffault F, Staun-Ram E, Sharshar T, Eymard B, et al. Methylome and transcriptome profiling in myasthenia gravis monozygotic twins. *J Autoimmun*. 2017;82:62–73.
40. Julia A, Absher D, Lopez-Lasanta M, Palau N, Pluma A, Waite Jones L, et al. Epigenome-wide association study of rheumatoid arthritis identifies differentially methylated loci in B cells. *Hum Mol Genet*. 2017;26(14):2803–11.
41. Uehiro N, Sato F, Pu F, Tanaka S, Kawashima M, Kawaguchi K, et al. Circulating cell-free DNA-based epigenetic assay can detect early breast cancer. *Breast Cancer Res*. 2016;18(1):129.
42. Kennedy DW, White NM, Benton MC, Fox A, Scott RJ, Griffiths LR, et al. Critical evaluation of linear regression models for cell-subtype specific methylation signal from mixed blood cell DNA. *PLoS One*. 2018;13(12):e0208915.
43. Kiehl S, Zimmermann T, Savai R, Pullamsetti SS, Seeger W, Bartkuhn M, et al. Epigenetic silencing of downstream genes mediated by tandem orientation in lung cancer. *Sci Rep*. 2017;7(1):3896.
44. Oleksiewicz U, Gladych M, Raman AT, Heyn H, Mereu E, Chlebanowska P, et al. TRIM28 and interacting KRAB-ZNFs control self-renewal of human pluripotent stem cells through epigenetic repression of pro-differentiation genes. *Stem Cell Rep*. 2017;9(6):2065–80.
45. Kozlenkov A, Jaffe AE, Timashpolsky A, Apones P, Rudchenko S, Barbu M, et al. DNA methylation profiling of human prefrontal cortex neurons in heroin users shows significant difference between genomic contexts of hyper- and hypomethylation and a younger epigenetic age. *Genes (Basel)*. 2017;8(6):152.

46. Takasawa K, Arai Y, Yamazaki-Inoue M, Toyoda M, Akutsu H, Umezawa A, et al. DNA hypermethylation enhanced telomerase reverse transcriptase expression in human-induced pluripotent stem cells. *Hum Cell*. 2018;31(1):78–86.
47. Herzog EM, Eggink AJ, Willemsen SP, Sliker RC, Wijnands KPJ, Felix JF, et al. Early- and late-onset preeclampsia and the tissue-specific epigenome of the placenta and newborn. *Placenta*. 2017;58:122–32.
48. Lee JU, Son JH, Shim EY, Cheong HS, Shin SW, Shin HD, et al. Global DNA methylation pattern of fibroblasts in idiopathic pulmonary fibrosis. *DNA Cell Biol*. 2019;38(9):905–14.
49. Fernandez-Santiago R, Carballo-Carbajal I, Castellano G, Torrent R, Richaud Y, Sanchez-Danes A, et al. Aberrant epigenome in iPSC-derived dopaminergic neurons from Parkinson's disease patients. *EMBO Mol Med*. 2015;7(12):1529–46.
50. Bigot A, Duddy WJ, Ouandaogo ZG, Negroni E, Mariot V, Ghimbovski S, et al. Age-associated methylation suppresses *SPRY1*, leading to a failure of re-quietence and loss of the reserve stem cell pool in elderly muscle. *Cell Rep*. 2015;13(6):1172–82.
51. Vizoso M, Puig M, Carmona FJ, Maqueda M, Velasquez A, Gomez A, et al. Aberrant DNA methylation in non-small cell lung cancer-associated fibroblasts. *Carcinogenesis*. 2015;36(12):1453–63.
52. Mamrut S, Avidan N, Staud-Ram E, Ginzburg E, Truffault F, Berrih-Aknin S, et al. Integrative analysis of methylome and transcriptome in human blood identifies extensive sex- and immune cell-specific differentially methylated regions. *Epigenetics*. 2015;10(10):943–57.
53. Mendioroz M, Do C, Jiang X, Liu C, Darbary HK, Lang CF, et al. Trans effects of chromosome aneuploidies on DNA methylation patterns in human Down syndrome and mouse models. *Genome Biol*. 2015;16:263.
54. Nicodemus-Johnson J, Myers RA, Sakabe NJ, Sobreira DR, Hogarth DK, Naureckas ET, et al. DNA methylation in lung cells is associated with asthma endotypes and genetic risk. *JCI Insight*. 2016;1(20):e90151.
55. Pidsley R, Lawrence MG, Zotenko E, Niranjani B, Statham A, Song J, et al. Enduring epigenetic landmarks define the cancer microenvironment. *Genome Res*. 2018;28(5):625–38.
56. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol*. 2016;17(1):208.
57. Fernandez-Rebollo E, Menstrup B, Ebert R, Franzen J, Abagnale G, Sieben T, et al. Human platelet lysate versus fetal calf serum: these supplements do not select for different mesenchymal stromal cells. *Sci Rep*. 2017;7(1):5132.
58. Verma D, Parasa VR, Raffetseder J, Martis M, Mehta RB, Netea M, et al. Antimicrobial activity correlates with altered DNA methylation pattern in immune cells from BCG-vaccinated subjects. *Sci Rep*. 2017;7(1):12305.
59. Cvitic S, Novakovic B, Gordon L, Ulz CM, Muhlberger M, Diaz-Perez FI, et al. Human fetoplacental arterial and venous endothelial cells are differentially programmed by gestational diabetes mellitus, resulting in cell-specific barrier function changes. *Diabetologia*. 2018;61(11):2398–411.
60. Lussier AA, Morin AM, MacIsaac JL, Salmon J, Weinberg J, Reynolds JN, et al. DNA methylation as a predictor of fetal alcohol spectrum disorder. *Clin Epigenetics*. 2018;10:5.
61. Clifford RL, Fishbane N, Patel J, MacIsaac JL, McEwen LM, Fisher AJ, et al. Altered DNA methylation is associated with aberrant gene expression in parenchymal but not airway fibroblasts isolated from individuals with COPD. *Clin Epigenetics*. 2018;10:32.
62. Wielscher M, Vierlinger K, Kegler U, Ziesche R, Gsur A, Weinhausel A. Diagnostic performance of plasma DNA methylation profiles in lung cancer, pulmonary fibrosis and COPD. *EBioMedicine*. 2015;2(8):929–36.
63. de Almeida DC, Ferreira MR, Franzen J, Weidner CI, Frobel J, Zenke M, et al. Epigenetic classification of human mesenchymal stromal cells. *Stem Cell Rep*. 2016;6(2):168–75.
64. Shiga K, Hara M, Nagasaki T, Sato T, Takahashi H, Takeyama H. Cancer-associated fibroblasts: their characteristics and their roles in tumor growth. *Cancers (Basel)*. 2015;7(4):2443–58.
65. Pierce RA, Field ED, Mutis T, Golovina TN, Von Kap-Herr C, Wilke M, et al. The HA-2 minor histocompatibility antigen is derived from a diallelic gene encoding a novel human class I myosin protein. *J Immunol*. 2001;167(6):3223–30.
66. Fagerberg L, Hallstrom BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics*. 2014;13(2):397–406.
67. Mabbott NA, Baillie JK, Brown H, Freeman TC, Hume DA. An expression atlas of human primary cells: inference of gene function from coexpression networks. *BMC Genomics*. 2013;14:632.
68. Guintivano J, Aryee MJ, Kaminsky ZA. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics*. 2013;8(3):290–302.
69. Lee DD, Seung HS. Algorithms for non-negative matrix factorization. *Adv Neur In*. 2001;13:556–62.
70. Zheng SC, Breeze CE, Beck S, Teschendorff AE. Identification of differentially methylated cell types in epigenome-wide association studies. *Nat Methods*. 2018;15(12):1059–66.
71. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. *Nat Rev Genet*. 2013;14(3):204–20.
72. Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, et al. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res*. 2013;23(3):555–67.
73. Tang B, Zhou Y, Wang CM, Huang TH, Jin VX. Integration of DNA methylation and gene transcription across nineteen cell types reveals cell type-specific and genomic region-dependent regulatory patterns. *Sci Rep*. 2017;7(1):3626.
74. Koestler DC, Christensen B, Karagas MR, Marsit CJ, Langevin SM, Kelsey KT, et al. Blood-based profiles of DNA methylation predict the underlying distribution of cell types: a validation analysis. *Epigenetics*. 2013;8(8):816–26.
75. BLUEPRINT consortium. Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. *Nat Biotechnol*. 2016;34(7):726–37.
76. Han Y, Franzen J, Stiehl T, Gobs M, Kuo CC, Nikolic M, et al. New targeted approaches for epigenetic age predictions. *BMC Biol*. 2020;18(1):71.
77. Lam D, Luu PL, Song JZ, Qu W, Risbridger GP, Lawrence MG, et al. Comprehensive evaluation of targeted multiplex bisulphite PCR sequencing for validation of DNA methylation biomarker panels. *Clin Epigenetics*. 2020;12(1):90.
78. Gieniec KA, Butler LM, Worthley DL, Woods SL. Cancer-associated fibroblasts—heroes or villains? *Br J Cancer*. 2019;121(4):293–302.
79. Monteran L, Erez N. The dark side of fibroblasts: cancer-associated fibroblasts as mediators of immunosuppression in the tumor microenvironment. *Front Immunol*. 2019;10:1835.
80. Liu L, Liu L, Yao HH, Zhu ZQ, Ning ZL, Huang Q. Stromal myofibroblasts are associated with poor prognosis in solid cancers: a meta-analysis of published studies. *PLoS One*. 2016;11(7):e0159947.
81. Dourado MR, Guerra ENS, Salo T, Lambert DW, Coletta RD. Prognostic value of the immunohistochemical detection of cancer-associated fibroblasts in oral cancer: a systematic review and meta-analysis. *J Oral Pathol Med*. 2018;47(5):443–53.
82. Soundararajan M, Kannan S. Fibroblasts and mesenchymal stem cells: two sides of the same coin? *J Cell Physiol*. 2018;233(12):9099–109.
83. Lynch MD, Watt FM. Fibroblast heterogeneity: implications for human disease. *J Clin Invest*. 2018;128(1):26–35.
84. Koch C, Suschek CV, Q L, S B, M G, S J, et al. Specific age-associated DNA methylation changes in human dermal fibroblasts. *PLoS One*. 2011;6(2):e16679.
85. Zilbauer M, Rayner TF, Clark C, Coffey AJ, Joyce CJ, Palta P, et al. Genome-wide methylation analyses of primary human leukocyte subsets identifies functionally important cell-type-specific hypomethylated regions. *Blood*. 2013;122(25):e52–60.
86. Bronkhorst AJ, Ungerer V, Holdenrieder S. The emerging role of cell-free DNA as a molecular marker for cancer management. *Biomol Detect Quantif*. 2019;17:100087.
87. Fortin JP, Triche TJ Jr, Hansen KD. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics*. 2017;33(4):558–60.
88. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. 2016;44(8):e71.
89. Zhou W, Triche TJ Jr, Laird PW, Shen H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Res*. 2018;46(20):e123.
90. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning. Pittsburgh: Association for Computing Machinery; 2006. p. 233–40.
91. Maié T, Schmidt M, Wagner W, Costa IG. DeconvolutionApp. 2020: <https://costalab.ukaachen.de/shiny/tmaie/deconvapp/>.



92. Therneau TM, editor. Extending the Cox model. New York: Springer US; 1997.
93. Kassambara A, Kosinski M, Biecek P. survminer: Drawing Survival Curves using 'ggplot2'. 2019; R package version 0.4.6. <https://CRAN.R-project.org/package=survminer>.
94. Fernandez-Rebollo E, Franzen J, Goetzke R, Hollmann J, Ostrowska A, Oliverio M, et al. Senescence-associated metabolomic phenotype in primary and iPSC-derived mesenchymal stromal cells. *Stem Cell Rep.* 2020;14(2):201-209.
95. Sontag S, Forster M, Qin J, Wanek P, Mitzka S, Schuler HM, et al. Modelling IRF8 deficient human hematopoiesis and dendritic cell development with engineered iPSC cells. *Stem Cells.* 2017;35(4):898-908.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

