





RESEARCH PAPER

 OPEN ACCESS 

DNA methylome-based validation of induced sputum as an effective protocol to study lung immunity: construction of a classifier of pulmonary cell types

Jyotirmoy Das ^a, Nina Idh^a, Liv Ingunn Bjoner Sikkeland^b, Jakob Paues^{a,c}, and Maria Lerm ^a

^aDivision of Inflammation and Infection, Department of Biomedical and Clinical Sciences, Faculty of Medicine and Health Sciences, Linköping University, Linköping, Sweden; ^bDepartment of Respiratory Medicine, Rikshospitalet, Oslo University Hospital and University of Oslo, Oslo, Norway; ^cDivision of Infectious Diseases, Department of Biomedical and Clinical Sciences, Faculty of Medicine and Health Sciences, Linköping University, Linköping, Sweden

ABSTRACT

Flow cytometry is a classical approach used to define cell types in peripheral blood. While DNA methylation signatures have been extensively employed in recent years as an alternative to flow cytometry to define cell populations in peripheral blood, this approach has not been tested in lung-derived samples. Here, we compared bronchoalveolar lavage with a more cost-effective and less invasive technique based on sputum induction and developed a DNA methylome-based algorithm that can be used to deconvolute the cell types in such samples. We analysed the DNA methylome profiles of alveolar macrophages and lymphocytes cells isolated from the pulmonary compartment. The cells were isolated using two different methods, sputum induction and bronchoalveolar lavage. A strong positive correlation between the DNA methylome profiles of cells obtained with the two isolation methods was found. We observed the best correlation of the DNA methylomes when both isolation methods captured cells from the lower parts of the lungs. We also identified unique patterns of CpG methylation in DNA obtained from the two cell populations, which can be used as a signature to discriminate between the alveolar macrophages and lymphocytes by means of open-source algorithms. We validated our findings with external data and obtained results consistent with the previous findings. Our analysis opens up a new possibility to identify different cell populations from lung samples and promotes sputum induction as a tool to study immune cell populations from the lung.

ARTICLE HISTORY

Received 30 March 2021
Revised 2 August 2021
Accepted 11 August 2021

KEYWORDS



Bronchoalveolar lavage; sputum induction; DNA methylation; linear regression analysis; lung cell deconvolution


Background


Much of our knowledge of the immune system is derived from studies on cells isolated from blood, although extrapolation of the performance of those lymphocytes and myeloid cells gives very limited information on immunological events in peripheral tissues. It is becoming increasingly evident that in order to understand tissue-specific immunity, samples from relevant tissues have to be collected and studied. To understand how different environmental factors contribute to airway inflammation, sputum induction (SI) has been employed as replacement for the invasive bronchoalveolar lavage (BAL) in several studies [1–4].

Epigenetic research has its roots in plant science, which emerged in the early 20th century. In human medicine, cancer biology has driven the

field forwards during the last two decades and in combination with modern, array-based techniques and next-generation sequencing now provides the scientific community with an easily accessible tool to study epigenetics at a whole-genome level. Presently, numerous clinical studies outside the cancer field are emerging, contributing an expanding portfolio of DNAm signatures for a variety of conditions, including psychiatric conditions (PTSD [5], ADHD [6]), exposure to cigarette smoke [7], ageing [8] and metabolic syndrome [9]. The DNAm signature analysis tools, the GrimAge algorithm [10] deserves a special mention since it has proven to be a more powerful age-classifier than those based on telomere length and importantly, it can provide information on both

CONTACT Maria Lerm  maria.lerm@liu.se  Professor in Medical Microbiology, Div. of Inflammation and Infection, Dept. of Biomedical and Clinical Sciences, Faculty of Medicine and Health Sciences, Linköping University, Lab 1, Floor 12, SE-58185 Linköping, Sweden

 Shared first authorship

 Supplemental data for this article can be accessed [here](#).

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

chronological and biological age from DNA methylation data. Several diseases have now been shown to involve age acceleration (higher biological age compared to chronological age), including severe Covid-19 [11], cardiovascular disease, cancer and diabetes [12]. A study based on DNA methylation data showed that cigarette smoking increased the relative age of the lungs as compared to the rest of the body [13].

In 2012, Houseman *et al* presented a cell-sorting algorithm based on DNA methylation data to identify the fraction of five different cell types in peripheral blood mononuclear cells (PBMCs) [14]. The classifier has had a fundamental impact on deconvolution of blood-derived DNA methylome data (nearly 2000 citations as per March 2021, including [15–20]) and has also been extrapolated to pulmonary cell DNA methylomes [21].

In this report, we isolated both macrophages (HLA-DR+, CD3- cells) and lymphocytes (CD3+ cells) from samples obtained through SI and BAL from the same donors in order to compare the cell populations obtained through the two protocols. Instead of using the marker-biased flow cytometry approach for cell characterization, we employed whole-genome DNA methylome analyses for the characterization. Our data demonstrate that the DNA methylation profile of the cell populations display strong overlap and conclude that the SI approach represents a valuable tool for non-invasive collection of samples representing the mucosal immunity of the airways. In addition, we provide a classifier algorithm that, based on DNA methylome data, can distinguish macrophages in samples derived from the pulmonary compartment.

Results

DNA methylomes from cells isolated through SI and BAL are highly correlated

To investigate whether the pulmonary cells collected using SI are similar to those obtained through BAL, we collected both specimens from the same subjects and isolated lymphocytes (CD3+) and macrophages (HLA-DR+/CD3-). DNA methylomes were captured from the isolated DNA using the Illumina 450 K protocol (Figure

1). The variation (β values after the data filtration) in the DNA methylome datasets derived from cells from SI and BAL were calculated and a Spearman's rank correlation test from each cell populations and each subject were performed. The analysis revealed a strong and highly significant positive correlation ($r_{TS|BAL\bar{x}} = 0.95$, p value $< 2.2e-16$) in each pair of data obtained through the two isolation protocols (Figure 2a, Supplementary Figure S1), suggesting that the cell identities and phenotypes were very similar. To analyse any possible common differences between sputum cells and BAL cells, we compared the mean global β values for the respective cell type and found that in both cell populations, the values differed slightly and significantly between the two protocols (SI_{mean} = 0.4504, BAL_{mean} = 0.4669 in HLA-DR+/CD3- cells and SI_{mean} = 0.4727, BAL_{mean} = 0.4666 in CD3+ cells, Figure 3(a,b)), suggesting that there is a relevant difference between the two protocols. Of note, a previous study demonstrated that the DNA methylomes of macrophages recovered from lower parts or apical parts of the lung display differences in the DNA methylomes, probably reflecting differences in the phenotypes [22]. Therefore, to investigate whether our HLA-DR data could be used to predict from which part of the lung the SI-derived cell populations came from, we used previously published data from Armstrong *et al* (GSE132547) that consists of Illumina EPIC array 850 K data obtained from 12 healthy subjects and grouped them separately as upper and lower lung for each subject. The linear regression was calculated and residuals were estimated ($\Delta y = \hat{y}_i - y_i \geq 0.3$, the distance of the β value of each CpG from the linear regression line, Figure 2(b,c)) to obtain the CpGs which are at the farthest distance from the regression line and strongly correlated to either SI or BAL samples. The percentage of cells derived from the upper and/or lower lungs was identified by using extracted CpGs from the HLA-DR datasets as test data with the Armstrong *et al* data [22] as training dataset in the EpiDISH package (Figure 4). The EpiDISH package provides tools to infer the fractions of a priori known cell subtypes present in a sample representing a mixture of

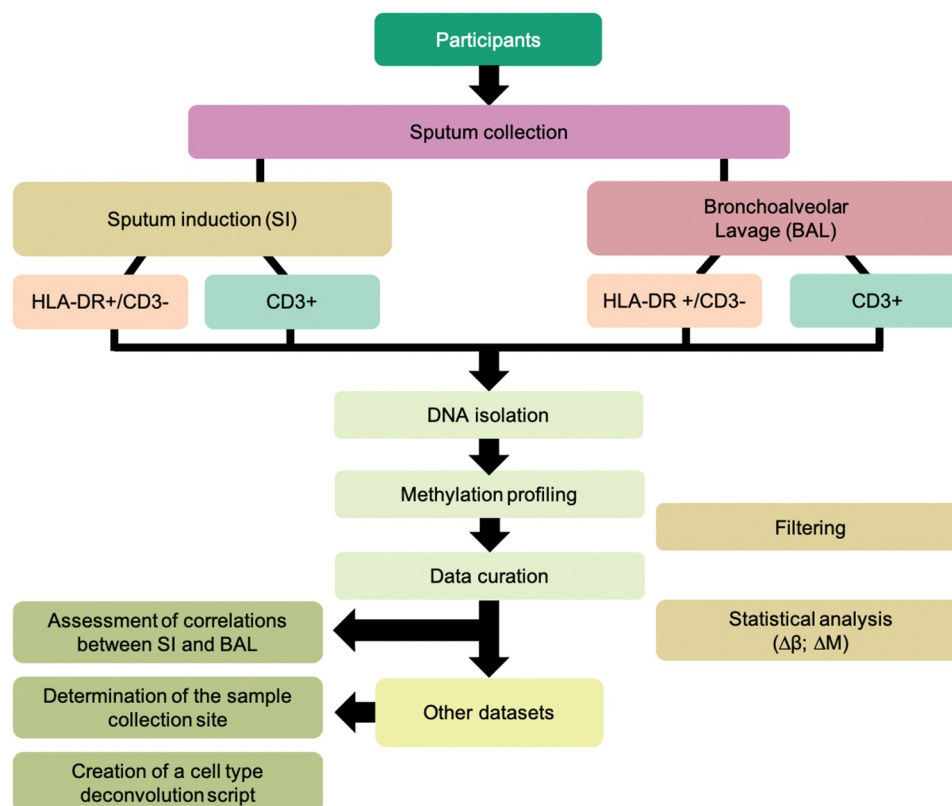


Figure 1. Flow diagram of the study.

such cell-types. For the subjects P4 and P5, the analysis predicted that BAL-recovered cells were from the lower lung and SI cells were from the upper lung, whereas for P2 and P3, SI and BAL cells were identified as cells from the lower lung. In P1, a fraction of the BAL cells (21%) were identified from the upper lung, while the rest of the BAL cells and the SI cells came from the lower lung.

A DNA methylome-based cell-sorting algorithm for pulmonary samples

The datasets compiled in our present study constitute a relevant source of data to create a classifier similar to the Houseman algorithm but instead tailored for the pulmonary compartment.

Based on our demonstration that the DNA methylomes from both isolation procedures (SI and BAL) were highly similar, we averaged the SI and BAL data for each cell population (HLA-DR and CD3) as a first step. To identify cell type-

specific CpGs, we employed three different reference-free EWAS algorithms, *RefFreeEWAS* [23], *SVA* [24] and *RefFreeCellMix* [23]. To estimate the variation only from the strongly correlated CpGs, we again set the residuals, $\Delta y = \hat{y}_i - y_i \leq 0.001$, and identify those CpGs close to the regression line to avoid SI or BAL-specific CpGs (Supplementary Figure S1b). Using the above-mentioned algorithms (with p_{BH} -value < 0.05), we determined the CpGs specific for HLA-DR+ and CD3+ cells and (Table 1) identified a total of 594 CpGs and 2,292 CpGs for HLA-DR cells and CD3 cells, respectively (Figure 5(a,b)).

Table 1. Cell specific CpGs to determine cell proportion estimation using different algorithms.

Algorithms	HLA-DR specific CpGs	CD3 specific CpGs
<i>RefFreeEWAS</i>	42,958	39,742
<i>SVA</i>	2,946	6,507
<i>RefFreeCellMix</i>	38,030	55,491

RefFreeEWAS: Epigenome-Wide Association Study (EWAS) using Reference-Free DNA Methylation Mixture Deconvolution. SVA: Surrogate Variable Analysis and RefFreeCellMix: Reference-Free Cell Mixture Projection

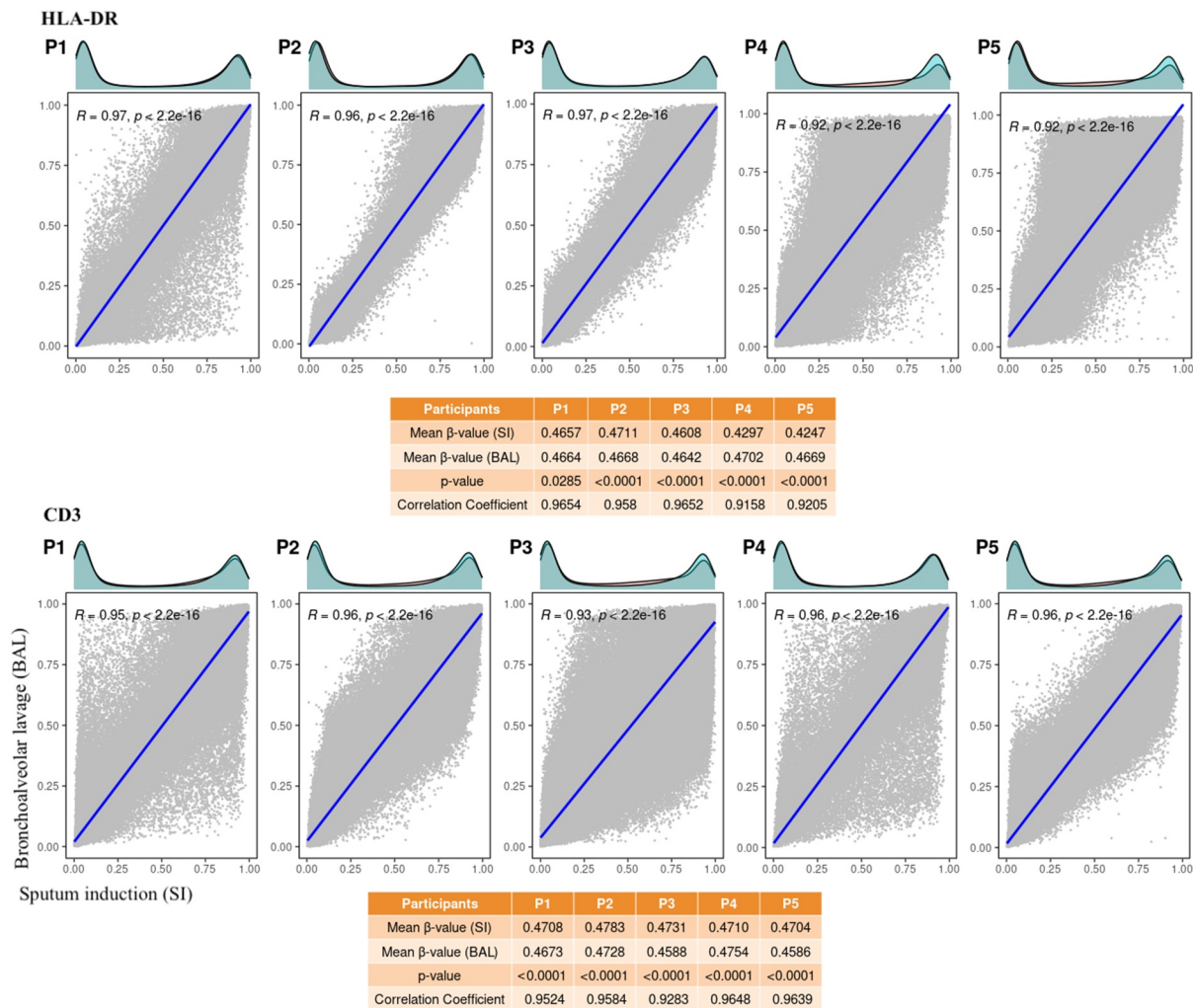


Figure 2. (a). Spearman's rank correlation analysis of β values obtained from DNA methylome analyses of cells (as indicated) from sputum induction (SI) and bronchoalveolar lavage (BAL). The density plots describe the β value distribution among SI (turquoise) and BAL (pink) methods. The adjacent tables represent statistical data from each sample. (b) Figure representing the calculation of residual value. The tangent distance was calculated using the equation mentioned in **methods**. The ellipses were drawn to show the CpGs closer to SI and BAL samples. The dashed line represents the cut-off applied to extract the CpGs above $\Delta y = 0.3$. Each dot represents one CpG. (c). Histogram representing the distribution of normalized β values of CD3 and HLA-DR cells. The blue vertical line representing the cut-off score ($\text{Mean} \pm 2\text{SD} = 0.3$) applied to select the CpGs for the analysis.

Finally, after applying an additional cut-off (≤ 0.1) to remove CpG with only minor differences in β values, we compiled a total of 2,763 CpGs as the reference dataset for the alveolar cell sorting (Supplementary Table S1).

Validation of the classifier against another dataset reveals a high accuracy

In order to validate the identified cell-specific CpGs, we compared the data with a recently published BAL sample DNA methylome dataset

(GSE133062) by Ringh *et al* [21]. In this study, the authors demonstrated by analysing Giemsa-stained BAL cytopins that more than 90% of the cells were alveolar macrophages. We performed a cell proportion analysis using our classifying CpG set in combination with the *EpiDISH* package in R with the whole 850 K DNA methylation data of Ringh *et al* [21]. The results revealed a high accuracy of the reference data in all samples with a high proportion of alveolar macrophages (85–100%) and low proportion of alveolar lymphocytes (0–15%) (Figure 6) in line with the published data

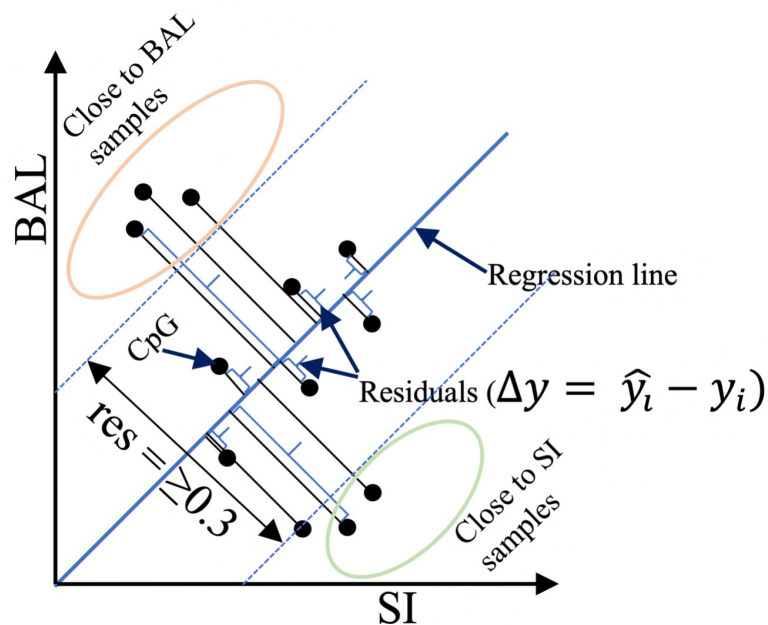


Figure 2.

[21]. To explain the performance of our identified cell-specific CpGs, we calculated the quantile relation between our cell-specific CpG data with Ringh *et al* data for BAL and SI separately (Supplementary Figure S3). The computational validation of the result showed that the identified CpGs can be a potential signature to deconvolute cell proportions in lung samples based on DNA methylation data.

Discussion

In this study, we demonstrated a strong correlation between the DNA methylome profiles of two cell populations isolated from SI and BAL. Both isolation methods are well known and have been used for decades for clinical diagnosis of pulmonary conditions. The sensitiveness of both methods has been widely compared [25–27] and one obvious advantage with BAL is the option to collect samples from defined sites in the lung. However, a major disadvantage with BAL is that it is an invasive procedure performed by physicians at larger medical centres that requires sedation and monitoring of the patient. It is also considerably more expensive than SI. On the other hand, SI is a safe, relatively cheap, non-invasive technique that can be performed by

nurses or physiotherapists at an outpatient clinic, which all contributes to the limitation of BAL samples for research [28].

To explore the comparativeness of cells obtained through SI and BAL, we collected samples from five patients using both methods and isolated HLA-DR+/CD3- and CD3+ cells. We performed a genome-wide DNA methylation analysis to compare the β values from ~400k CpG sites in the DNA from these cell types and found a strong and significant correlation between these two methods in both cell types. Although this analysis suggested that the cell populations obtained through the two protocols were very similar, comparison of the global mean β values significantly differed between SI and BAL. Through exploration of a previous study's finding [22] that designated unique CpG signatures to macrophages collected from the upper or lower parts of the lungs, we examined the possibility that this slight discrepancy was related to the local pulmonary site of sample collection. Indeed, the analysis allowed prediction of the site of collection (either upper or lower lung or in one case a mixture of both). Importantly, like BAL, which allows sample collection at defined sites, SI samples could result in cells obtained from the lower parts, which demonstrates that SI is an effective method for

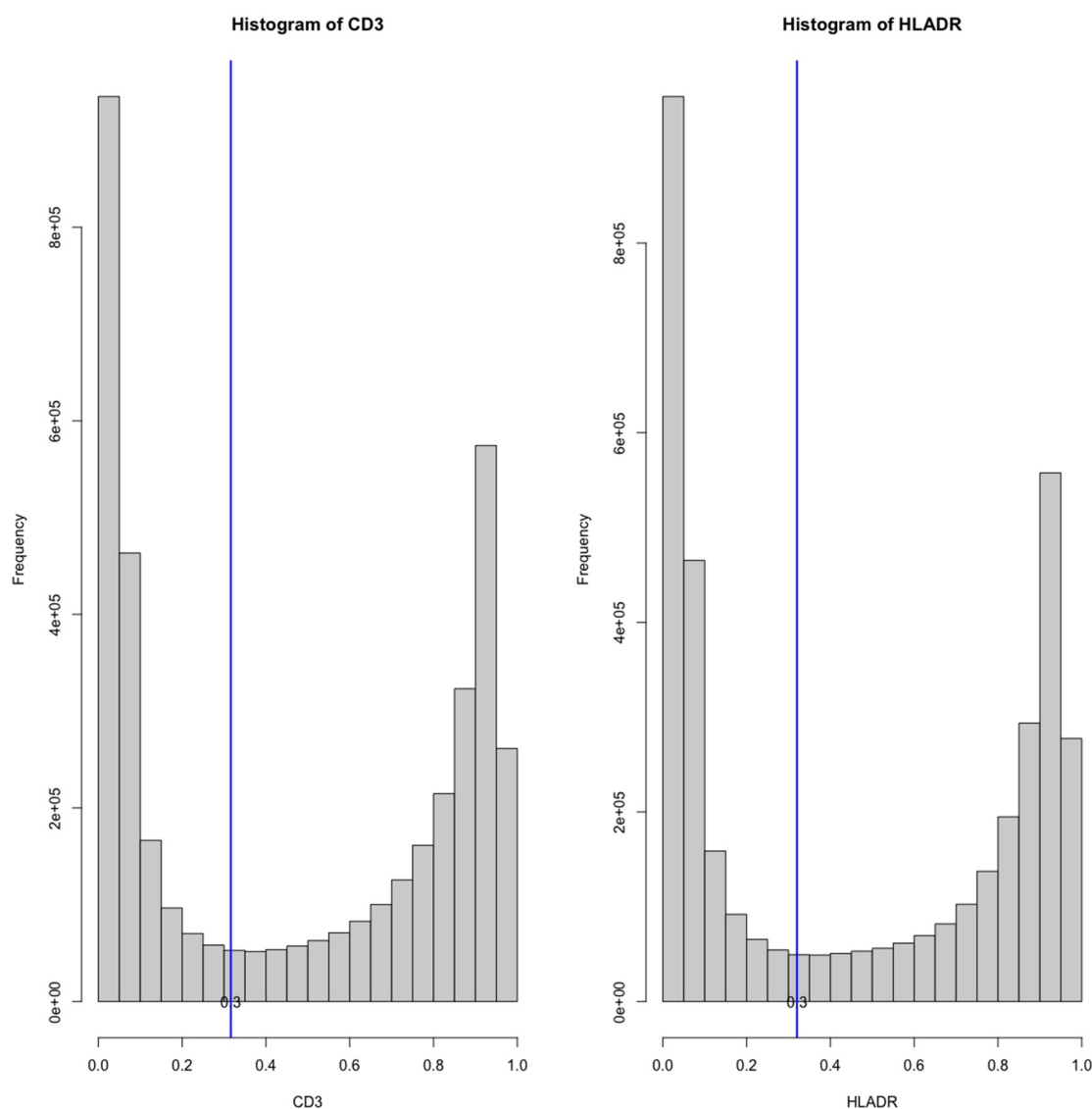


Figure 2.

studies of pulmonary immunity. A limitation of the present study was the absence of information on which part of the lungs was targeted in the BAL collection.

Using DNA methylation-based signatures to classify cell types is gaining increasing interest. For example, a recent study showed that tissue-specific DNA methylation signatures of free circulating DNA in plasma of patients with severe Covid-19 can reveal specific organ injury [29]. The most widely used cell-type deconvolution algorithm was described by Houseman *et al* [14], which uses cell-type-specific DNA methylation signatures to determine the frequencies of T cells,

B cells, monocytes, NK cells and neutrophils in PBMCs. Here, we present a comparable algorithm to assess the proportion of T cells and macrophages in pulmonary samples by combining three reference-free algorithms to derive the cell-specific CpGs. We used reference-free algorithms since there is no available reference data. We observed a large variation among these three algorithms and therefore intersected the CpG list and used the overlapping CpGs as cell type markers. In a validation step, our algorithm was applied to another publicly available dataset and it generated a similar prediction of cell proportions as the previously published, microscopy-validated data. At present,

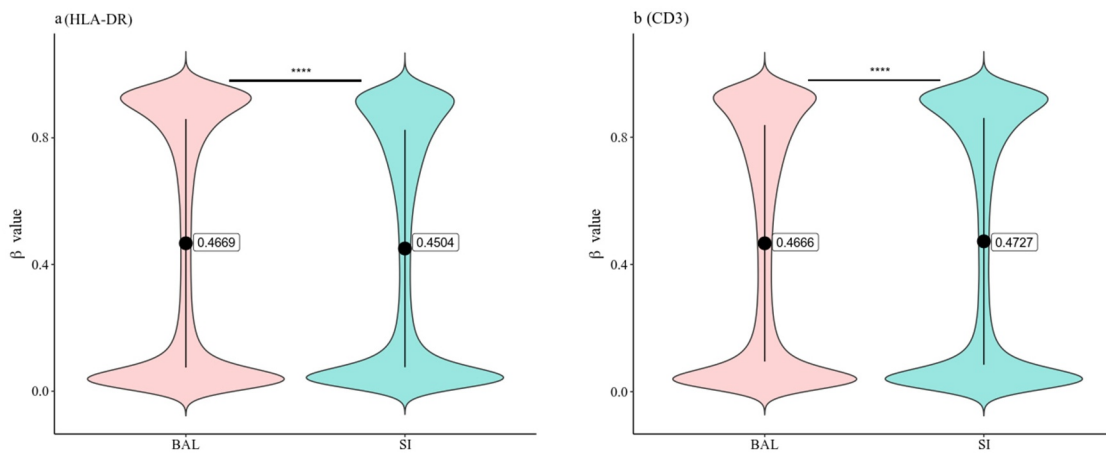


Figure 3. Distribution of DNA methylation β values in SI and BAL data. **a)** HLA-DR and **b)** CD3. The mean value is indicated by the black dot and the vertical line represents the interquartile range. Statistics: Mann-Whitney-Wilcoxon test, **** represents a p -value $<1e-4$. $N_{\text{HLA-DR}} = 382,591$, $N_{\text{CD3}} = 398,390$.

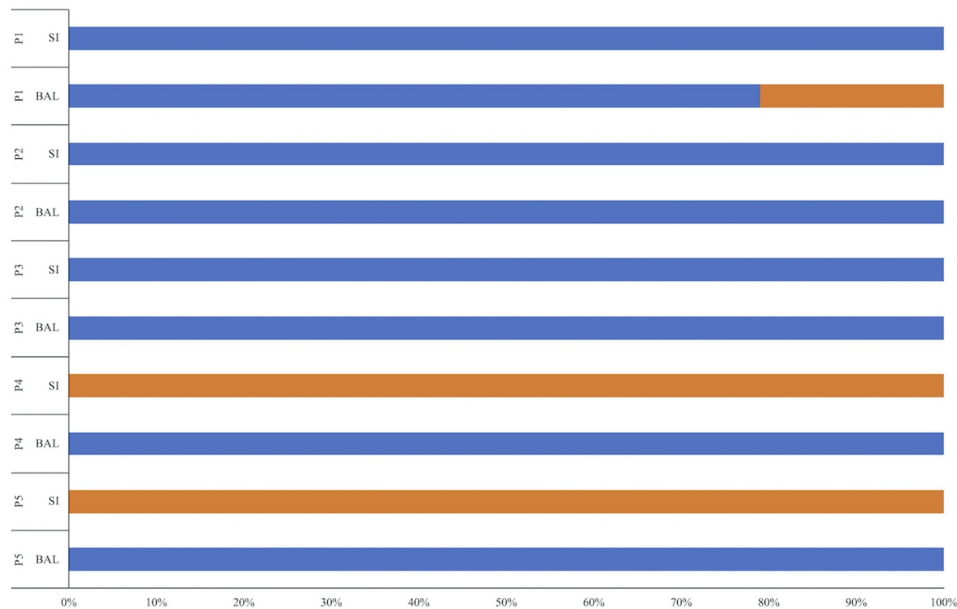


Figure 4. Bar plot displaying the distribution of lower- (blue) and upper- (orange) lung-specific CpGs as identified by the EpiDISH package for the included individuals' SI and BAL samples.

very few studies on lung-derived cells exist that include DNA methylation data along with cytological data; however, future studies will be able to comprehensively capture DNA methylation signatures for all pulmonary cell types. The results has the potential to facilitate interpretation of DNA methylation data originating from pulmonary samples, which is receiving increasing attention with a large number of studies defining pulmonary

CpG signatures for conditions including cystic fibrosis, lung cancer, chronic obstructive pulmonary diseases and smoking [21,30–33].

Conclusions

Our analyses demonstrate that SI is an attractive method for studies of pulmonary immunity that can replace BAL and that DNA methylomes

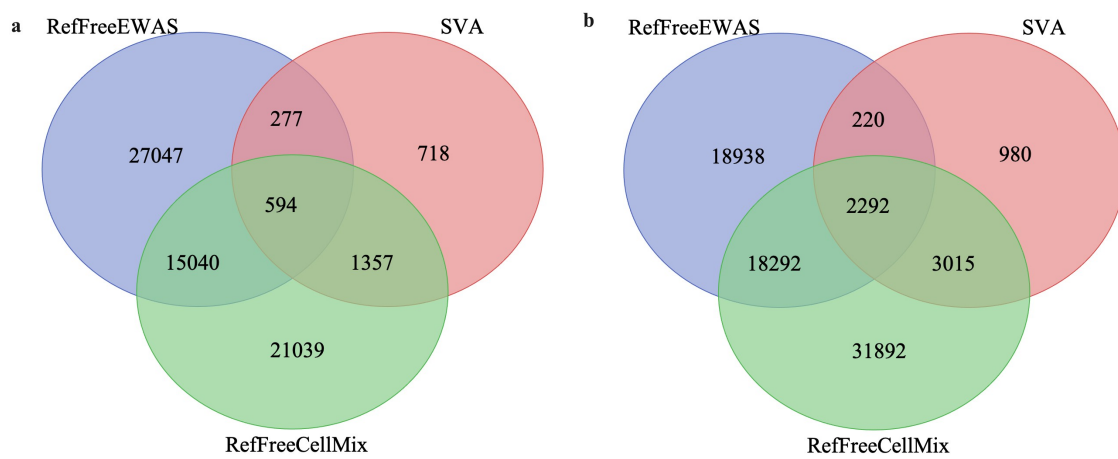


Figure 5. Venn diagrams representing HLA-DR-specific (a) and CD3-specific (b) CpGs based on three different reference-free EWAS deconvolution methods (RefFreeEWAS, SVA and RefFreeCellMix).

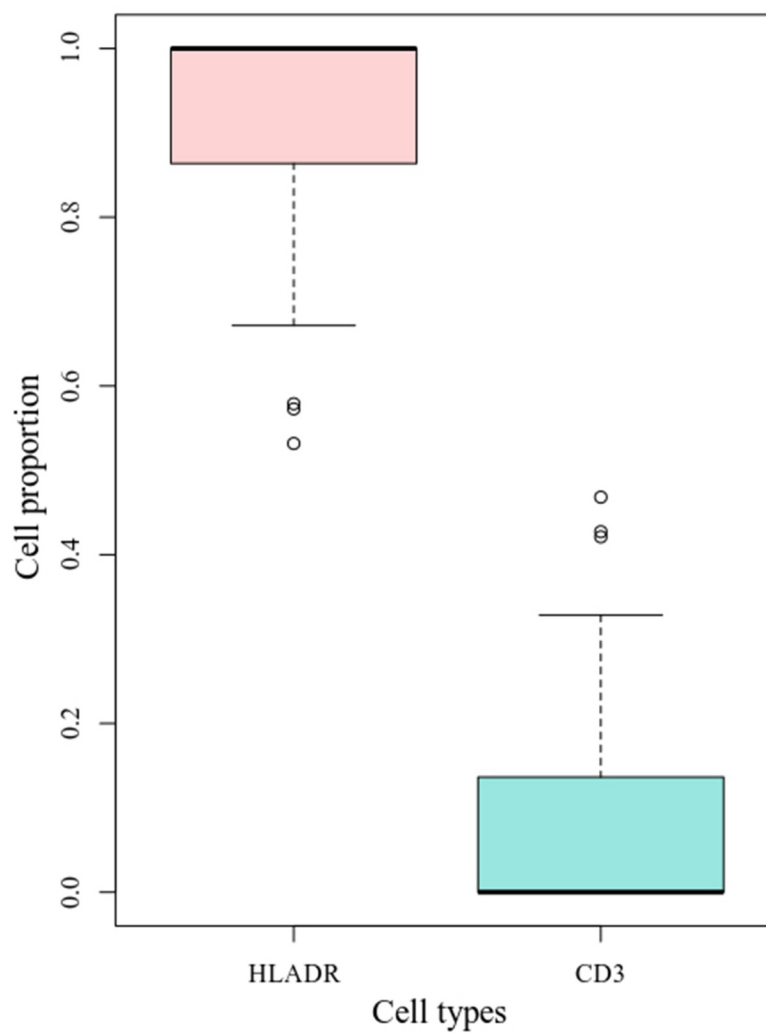


Figure 6. Boxplot illustrating a validation of the cell sorting algorithm against R Singh *et al.* data.

derived from the cells obtained through this protocol can distinguish between cells collected from the upper or lower parts of the lungs as well as predict the proportion of macrophages and T cells.

Methods

Study design, ethics and participants

Participants, five female patients scheduled for BAL as part of their clinical investigation, were recruited from Linköping University Hospital between 2017 and 2018 (Table 2). The participants donated induced sputum at a time point $-/+$ 2 weeks before/after collection of the BAL sample. After collection of BAL for diagnostic purposes, leftovers were used for the planned research.

Sputum induction and sputum processing

Sputum induction and processing was done as described by Sikkeland *et al* (2012) [34,35] with some modifications. Briefly, hypertonic saline at 3%, 4%, and 5% were inhaled for three 7-minute periods using a nebulizer (eFlow Rapid, PARI, Germany) after nebulizer-administered premedication with an adrenergic β 2-agonist, salbutamol 1 ml (1 mg/ml). After each inhalation period expectorates were collected, the subjects performed a three-step cleansing procedure to reduce contamination from nasopharynx (blow nose, rinse mouth and gargle three times with water), before coughing deeply involving the thorax and spitting the expectorate into a 50 ml sterile tube, (Falcon).

Sputum was processed within two hours after induction. From the sputum sample, mucus plugs, were manually selected, weighed and treated with four times its volume of 0.1% dithiothreitol (DTT,

Thermo Fisher). The sample was aspirated with a Pasteur pipet five times, vortexed 15 sec and rocked at 4°C at 40 rev/min for 15 min. The sample was then diluted four times with phosphate buffered saline (PBS), and rocked for 5 min before filtrated through a 50 μ m cell strainer (CellTrics® by Sysmex) and centrifuged for 5 min. Cell viability, total cell counts, and squamous cell counts were manually evaluated by counting cells in a Bürker chamber with trypan blue staining.

BAL was obtained according to standard clinical procedures at the Linköping University Hospital. 125 ml of sterile, physiological sodium chloride (NaCl) solution was instilled and aspirated to collect cells. The BAL fluid was kept on ice until processing. PBS was added and the samples were centrifuged at 400 g for 5 minutes at 4°C, then washed with PBS and followed by recentrifuged at 300 g for 13 minutes at 4°C before isolation of DNA as described below.

DNA extraction and data processing

The pulmonary HLA-DR and CD3-positive cells were isolated using superparamagnetic beads coupled with anti-human CD3 and Pan Mouse IgG antibodies (Invitrogen Dynabeads®, Life Technologies AS, Norway) and HLA-DR/human MHC class II antibodies (Invitrogen Dynabeads®, Life Technologies AS, Norway). An initial positive selection was done with CD3 beads followed by a positive HLA-DR selection. Bead coating and cell isolation was performed according to manufacturer's protocol. The DNA and RNA were extracted from the lung immune cells using the AllPrep® DNA/RNA Mini Kit (Qiagen, Germany), per the manufacturer's instructions.

The DNA methylome data of HLA-DR+/CD3- and CD3+ cells was analysed using the HumanMethylation450K (450 K) BeadChip (Illumina, USA) array as per the manufacturer's instructions. The raw IDAT files of the DNA methylation data was processed using the BMIQ normalization function of the *ChAMP* package [36] in R (v3.6.3) after using a robust filtration criteria to generate the β values from each CpG for each sample. The sex chromosome data was removed to get rid of the gender biasness, multi-hit probes were taken away from the data and also

Table 2. Demographic characteristics of participants.

Characteristics	Value
Age (year) [†]	55 \pm 10
Height (cm) [†]	166 \pm 1
Weight (kg) [†]	70 \pm 10
Body Mass Index (BMI) [†]	25 \pm 3
Sex (male/female)	0/5
BCG* (yes/no)	5/0
Smoking (current/previous/never)	0/4/1

[†]Values are mean \pm standard deviation. *BCG is the abbreviation of *Bacillus Calmette–Guérin*.

CpGs with lower p -values (< 0.01) was drawn out from the data ([37]; submitted manuscript). The filtered data was normalized using the beta-mixture quantile normalization (BMIQ) function using the ChAMP package [38]. To compare our 450 K data with other previously published dataset from Illumina 850 K array [GSE133062, GSE132547 [21,22]], we used *merge* function in R to extract the CpG sites present in the Illumina 450 K array analysis.

Statistical calculations

All statistical analysis was performed in R v3.6.3 and bioconductor packages (v3.10). Anderson-Darling (AD) normality test [39] was used to calculate the non-parametric distribution of the dataset using *nortest* package [40] in R (used for large dataset, >380 K, Table 3). As all of the data show non-parametric distribution, Spearman's rank correlation test was performed to calculate the correlation using *ggpubr* package [41] in R. The linear regression line with *geom smooth* function from *ggplot2* package was added to the plot to calculate the confidence interval (C.I. = 95%). The correlation matrix was calculated using the Spearman's correlation coefficient to evaluate the pairwise relationship among all samples. The value was considered significant if p -value < 0.05 throughout the current study.

Table 3. Normality test result from Anderson-Darling test.

Participants		Anderson-Darling A^2 value		p -value
		HLA-DR (N = 382,591)	CD3 (N = 398,390)	
P1	BAL	33934	29,223	$< 2.2e-16$
	SI	32022	34,212	
P2	BAL	34328	25,691	
	SI	33048	33,348	
P3	BAL	30508	22,776	
	SI	32047	31,399	
P4	BAL	33831	32,325	
	SI	24798	34,150	
P5	BAL	29871	31,828	
	SI	22223	24,985	

The table describes the normality test from each participant in both HLA-DR and CD3 cell types. Bronchoalveolar lavage (BAL) and sputum induction (SI). The calculation was done in 382,591 CpGs in HLA-DR and 398,390 CpGs in CD3 cell types.

Cell-specific CpGs identification and validation

To identify the cell type-specific CpGs, three reference-free algorithms were used, *RefFreeEWAS* [23], *SVA* [24] and *RefFreeCellMix* [23]. First, the linear regression model was applied on the β values for each sample combining both SI and BAL samples. The residual value was then calculated using the equation, $\Delta y = \hat{y}_i - y_i$, and $\Delta y < 0.001$ filter was set to extract the CpGs close to the regression line (Figure 2b). Data was prepared from the β values and the phenotypes (e.g., HLA-DR and CD3) per sample as per algorithm requirement and covariates were calculated using Bonferroni-Hochberg corrected p -value < 0.05 . Each algorithm was used separately to predict the cell-type-specific CpGs. Venn analysis was used to intersect the overlap CpGs identified using the three algorithms. To validate our results with another external dataset, a previously published dataset was used as the test data GSE133062 [21] only in the HLA-DR cell types using the *EpiDISH* package [42] in R. A Q-Q plot was also calculated using the *ggplot* function in R between the cell-specific CpGs identified in our dataset and the test data GSE133062 for BAL and SI for HLA-DR cells (Supplementary Figure S3). No publicly available dataset was identified that could be used to validate the CD3-specific cell population.

Acknowledgments

We direct our gratitude to the staff at Linköping University Hospital for assistance in sample collection and all the subjects for donating samples. The DNA methylome data were generated at the Bioinformatics and Expression Analysis (BEA) Core Facility at the Department of Biosciences and Nutrition, which is supported by the Board of Research at the Karolinska Institute, Stockholm. The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Linköping University campus partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

Authors' contributions

M.L. N.I. and J.P. wrote the ethical application, M.L., N.I., J.P. and J.D. conceived and designed the study, N.I. and J.P. collected the patient samples and compiled the demographics tables, L.I.B.S. designed the sputum induction method, J.D. and M.L. designed and performed the bioinformatic analysis

of the data, J.D. wrote the scripts and created the figures, M. L., N.I. and J.D. wrote the manuscript.

Availability of data and materials

Illumina BeadChip 450K array data will be made available after final publication and the R scripts are available in github.com/JD2112/SivsBALanalysis (upon request). The list of cell-type-specific CpGs are added in the Supplementary Table S1.

Consent for publication

All authors have approved to the final version of the manuscript.

Ethics approval and consent to participate

The study was approved by the ethics committee, EPN, in Linköping, Dnr 2016/237-31. Oral and written informed consent were obtained from all participants before inclusion.

Disclosure statement

No potential conflict of interest was reported by the author (s).

Funding

This study was funded through generous grants from Forskningsrådet Sydöstra Sverige (FORSS-932096), the Swedish Research Council (2015-02593 and 2018-02961) and the Swedish Heart Lung Foundation (20150709 and 20180613). J.D. is a postdoctoral fellow supported through the Medical Infection and Inflammation Center (MIIC) at Linköping University; Forskningsrådet i Sydöstra Sverige [FORSS-932096]; Hjärt-Lungfonden [20150709 and 20180613]; Vetenskapsrådet [2015-02593 and 2018-02961];

ORCID

Jyotirmoy Das  <http://orcid.org/0000-0002-5649-4658>
 Maria Lerm  <http://orcid.org/0000-0002-5092-9892>

References

- [1] Sikkeland LIB, Kongerud J, Stangeland AM, et al. Macrophage enrichment from induced sputum [3]. *Thorax*. 2007;62(6):558–559.
- [2] Kononova N, Sikkeland LIB, Mahmood F, et al. Annual decline in forced expiratory volume and airway inflammatory cells and mediators in a general population-based sample. *BMC Pulm Med*. 2019;19(1). DOI:10.1186/s12890-018-0765-7.
- [3] Belli AJ, Bose S, Aggarwal N, et al. Indoor particulate matter exposure is associated with increased black carbon content in airway macrophages of former smokers with COPD. *Environ Res*. 2016;150:398–402.
- [4] Li W, Gao R, Xin T, et al. Different expression levels of interleukin-35 in asthma phenotypes. *Respir Res*. 2020; 21:89, <https://doi.org/10.1186/s12931-020-01356-6>.
- [5] Sheerin CM, Lancaster EE, York TP, et al. Epigenome-Wide study of posttraumatic stress disorder symptom severity in a treatment-seeking adolescent sample. *J Trauma Stress*. 2021;34(3):607–615.
- [6] Carbon S, Douglass E, Dunn N, et al. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res*. 2019; Jan 8;47(D1):D330–D338. doi: 10.1093/nar/gky1055.
- [7] Bollepalli S, Korhonen T, Kaprio J, et al. EpiSmokEr: a robust classifier to determine smoking status from DNA methylation data. *Epigenomics*. 2019;11(13):1469–1486.
- [8] McCrory C, Fiorito G, Hernandez B, et al. GrimAge outperforms other epigenetic clocks in the prediction of age-related clinical phenotypes and all-cause mortality. *J Gerontol Ser A*. 2021;76(5):741–749.
- [9] Nuotio ML, Pervjakova N, Joensuu A, et al. An epigenome-wide association study of metabolic syndrome and its components. *Sci Rep*. 2020;10(1). DOI:10.1038/s41598-020-77506-z.
- [10] Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol*. 2013;14(10):R115.
- [11] Corley MJ, Pang APS, Dody K, et al. Genome-wide DNA methylation profiling of peripheral blood reveals an epigenetic signature associated with severe COVID-19. *J Leukoc Biol*. 2021;110(1):21–26.
- [12] Oblak L, van der Zaag J, Higgins-Chen AT, et al. A systematic review of biological, social and environmental factors associated with epigenetic clock acceleration. *Ageing Res Rev*. 2021;69:101348.
- [13] Wu X, Huang Q, Javed R, et al. Effect of tobacco smoking on the epigenetic age of human respiratory organs. *Clin Epigenetics*. 2019;11(1). DOI:10.1186/s13148-019-0777-z
- [14] Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012;13(1). DOI:10.1186/1471-2105-13-86.
- [15] Liu Y, Aryee MJ, Padyukov L, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol*. 2013;31(2):142–147.
- [16] Bick AG, Weinstock JS, Nandakumar SK, et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature*. 2020;586(7831):763–768.

- [17] Wahl S, Drong A, Lehne B, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature*. 2017;541(7635):81–86.
- [18] Bock C. Analysing and interpreting DNA methylation data. *Nat Rev Genet*. 2012;13(10):705–719.
- [19] Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol*. 2014;15(2):R31.
- [20] Verma D, Parasa VR, Raffetseder J, et al. Antimycobacterial activity correlates with altered DNA methylation pattern in immune cells from BCG-vaccinated subjects. *Sci Rep*. 2017;7(1). DOI:10.1038/s41598-017-12110-2.
- [21] Ringh MV, Hagemann-Jensen M, Needhamsen M, et al. Tobacco smoking induces changes in true DNA methylation, hydroxymethylation and gene expression in bronchoalveolar lavage cells. *EBioMedicine*. 2019;46:290–304.
- [22] Armstrong DA, Chen Y, Dessaint JA, et al. DNA methylation changes in regional lung macrophages are associated with metabolic differences immunohorizons. *Am Assoc Immunol*. 2019;3:274–281.
- [23] Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics*. 2014;30(10):1431–1439.
- [24] Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*. 2007;3(9):e161.
- [25] Conde MB, Soares SLM, Mello FCQ, et al. Comparison of sputum induction with fiberoptic bronchoscopy in the diagnosis of tuberculosis: experience at an acquired immune deficiency syndrome reference center in Rio de Janeiro, Brazil. *Am J Respir Crit Care Med*. 2000;162(6):2238–2240.
- [26] McWilliams T, Wells AU, Harrison AC, et al. Induced sputum and bronchoscopy in the diagnosis of pulmonary tuberculosis. *Thorax*. 2002;57(12):1010–1014.
- [27] Saglam L, Akgun M, Aktas E. Usefulness of induced sputum and fibreoptic bronchoscopy specimens in the diagnosis of pulmonary tuberculosis. *J Int Med Res*. 2005;33(2):260–265.
- [28] Anderson C, Inhaber N, Menzies D. Comparison of sputum induction with fiber-optic bronchoscopy in the diagnosis of tuberculosis. *Am J Respir Crit Care Med*. 1995;152(5):1570–1574.
- [29] Cheng AP, Cheng MP, Gu W, et al. Cell-free DNA in blood reveals significant cell, tissue and organ specific injury and predicts COVID-19 severity. *medRxiv*. 2020; Jul 29. doi: 10.1101/2020.07.27.20163188.
- [30] Sundar IK, Yin Q, Baier BS, et al. DNA methylation profiling in peripheral lung tissues of smokers and patients with COPD. *Clin Epigenetics*. 2017;9(1). DOI:10.1186/s13148-017-0335-5.
- [31] Magalhães M, Tost J, Pineau F, et al. Dynamic changes of DNA methylation and lung disease in cystic fibrosis: lessons from a monogenic disease. *Epigenomics*. 2018;10(8):1131–1145.
- [32] Tsou JA, Hagen JA, Carpenter CL, et al. DNA methylation analysis: a powerful new tool for lung cancer diagnosis. *Oncogene*. 2002;21(35):5450–5461.
- [33] Shen N, Du J, Zhou H, et al. A diagnostic panel of DNA methylation biomarkers for lung adenocarcinoma. *Front Oncol*. 2019;9. DOI:10.3389/fonc.2019.01281.
- [34] Sikkeland LIB, Eduard W, Skogstad M, et al. Recovery from workplace-induced airway inflammation 1 year after cessation of exposure. *Occup Environ Med*. 2012;69(10):721–726.
- [35] Sikkeland LIB, Dahl CP, Ueland T, et al. Increased levels of inflammatory cytokines and endothelin-1 in alveolar macrophages from patients with chronic heart failure. *PLoS One*. 2012;7(5):e36815.
- [36] Morris TJ, Butcher LM, Feber A, et al. ChAMP: 450k chip analysis methylation pipeline. *Bioinformatics*. 2014;30(3):428–430.
- [37] Das J, Verma D, Gustafsson M, et al. Identification of DNA methylation patterns predisposing for an efficient response to BCG vaccination in healthy BCG-naïve subjects. *Epigenetics [Internet]*. 2019;1–13. Available from: <https://www.tandfonline.com/doi/full/10.1080/15592294.2019.1603963>
- [38] Teschendorff AE, Marabita F, Lechner M, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*. 2013;29(2):189–196.
- [39] Razali NM, Wah YB. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J Stat Model Anal*. 2011; 2(1): 21–33.
- [40] Gross J, Ligges U. nortest: tests for normality [Internet]. 2015. Available from: <https://cran.r-project.org/package=nortest>
- [41] Kassambara A. ggpubr: “ggplot2” based publication ready plots [Internet]. 2020. Available from: <https://cran.r-project.org/package=ggpubr>
- [42] Zheng SC, Breeze CE, Beck S, et al. Identification of differentially methylated cell types in epigenome-wide association studies. *Nat Methods*. 2018;15(12):1059–1066.