Clinical Chemistry 64:3
566–575 (2018)

Leukocyte Counts Based on DNA Methylation at Individual Cytosines

Joana Frobel,[1,2] Tanja Božić,[1,2] Michael Lenz,[3,4,5] Peter Uciechowski,[6] Yang Han,[1,2] Reinhild Herwartz,[7] Klaus Strathmann,[8] Susanne Isfort,[7] Jens Panse,[7] André Esser,[9] Carina Birkhofer,[10] Uwe Gerstenmaier,[10] Thomas Kraus,[9] Lothar Rink,[6] Steffen Koschmieder,[7] and Wolfgang Wagner[1,2*]

BACKGROUND: White blood cell counts are routinely measured with automated hematology analyzers, by flow cytometry, or by manual counting. Here, we introduce an alternative approach based on DNA methylation (DNAm) at individual CG dinucleotides (CpGs).

METHODS: We identified candidate CpGs that were nonmethylated in specific leukocyte subsets. DNAm levels (ranging from 0% to 100%) were analyzed by pyrosequencing and implemented into deconvolution algorithms to determine the relative composition of leukocytes. For absolute quantification of cell numbers, samples were supplemented with a nonmethylated reference DNA.

RESULTS: Conventional blood counts correlated with DNAm at individual CpGs for granulocytes ($r = -0.91$), lymphocytes ($r = -0.91$), monocytes ($r = -0.74$), natural killer (NK) cells ($r = -0.30$), T cells ($r = -0.73$), CD4+ T cells ($r = -0.41$), CD8+ T cells ($r = -0.88$), and B cells ($r = -0.66$). Combination of these DNAm measurements into the "Epi-Blood-Count" provided similar precision as conventional methods in various independent validation sets. The method was also applicable to blood samples that were stored at 4 °C for 7 days or at −20 °C for 3 months. Furthermore, absolute cell numbers could be determined in frozen blood samples upon addition of a reference DNA, and the results correlated with measurements of automated analyzers in fresh aliquots ($r = 0.84$).

CONCLUSIONS: White blood cell counts can be reliably determined by site-specific DNAm analysis. This approach is applicable to very small blood volumes and frozen samples, and it allows for more standardized and cost-effective analysis in clinical application.

Analysis of the composition of white blood cells is among the most frequently requested laboratory tests in hematological diagnostics *(1)*. Leukocyte differential counts (LDCs)[11] can be determined by microscopic evaluation and manual counting. Since the advent of automated cell counters, LDCs are particularly analyzed by flow cytometric technologies *(2)*. Such automated analyzers sense electrical impedance, optical light-scattering properties, or fluorescence signal intensities *(3–5)*. Fluorescent staining of specific epitopes is the gold standard for definition of lymphocyte subsets. However, immunophenotypic analysis is costly, relatively labor-intensive, and not trivial to standardize. Furthermore, all of the aforementioned methods are applicable to only fresh blood samples; samples cannot be frozen for shipment or later analysis *(3, 6)*. Recently, genome-wide gene-expression profiles *(7–10)* and epigenetic profiles *(11–17)* have been used to deconvolute the cellular composition in whole blood. Such alternative approaches might overcome some of the limitations of the well-established state-of-the-art procedures for LDCs. While these deconvolution procedures may have several advantages, it must be taken into account that they are not applicable for analysis of erythrocytes and thrombocytes with inconsistent mRNA content and lack of DNA. Because erythrocyte and thrombocyte counts are of particular clinical relevance, conventional procedures cannot be completely replaced by gene expression or epigenetic parameters.

[1] Helmholtz-Institute for Biomedical Engineering, Stem Cell Biology and Cellular Engineering, RWTH Aachen University Medical School, Aachen, Germany; [2] Institute for Biomedical Engineering – Cell Biology, University Hospital of RWTH Aachen, Aachen, Germany; [3] Joint Research Center for Computational Biomedicine, RWTH Aachen University, Aachen, Germany; [4] Aachen Institute for Advanced Study in Computational Engineering Science (AICES), RWTH Aachen University, Aachen, Germany; [5] Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, Maastricht, the Netherlands; [6] Institute of Immunology, Faculty of Medicine, RWTH Aachen University, Aachen, Germany; [7] Department of Hematology, Oncology, Hemostaseology, and Stem Cell Transplantation, Faculty of Medicine, RWTH Aachen University, Aachen, Germany; [8] Institute for Transfusion Medicine, University Hospital Aachen, Aachen, Germany; [9] Institute for Occupational and Social Medicine, RWTH Aachen University, Aachen, Germany; [10] Varionostic GmbH, Ulm, Germany.

* Address correspondence to this author at: Helmholtz-Institute for Biomedical Engineering, Stem Cell Biology and Cellular Engineering, RWTH Aachen University Medical School, Pauwelsstraße 20, 52074 Aachen, Germany. E-mail wwagner@ukaachen.de.

Received July 21, 2017; accepted October 17, 2017.
Previously published online at DOI: 10.1373/clinchem.2017.279935

[11] Nonstandard abbreviations: LDC, leukocyte differential count; DNAm, DNA methylation; CpG, CG dinucleotide; NNLS, nonnegative least-squares; cfDNA, cell-free circulating DNA.

DNA methylation (DNAm) represents the best understood epigenetic modification. Methyl groups can be added to the fifth carbon atom of cytosines, predominantly in a cytosine-guanine dinucleotide context (CpG site). DNAm patterns have many advantages compared to immunophenotypic analyses: (*a*) DNAm is directly linked to cellular differentiation; (*b*) DNAm facilitates absolute quantification at single-base resolution (ranging from 0% to 100% DNAm); (*c*) every cell has only 2 copies of DNA and hence the results can be easily extrapolated to the cellular composition (in contrast to RNA, which can be highly overrepresented in small subsets); and (*d*) DNA is relatively stable, able to be isolated from lysed or frozen cells and shipped at room temperature for further analysis. So far, epigenetic estimations of LDCs are based on microarray data taking multiple CpGs into account; however, such profiling procedures are relatively costly and hardly applicable in daily clinical routine.

In this study, we hypothesized that site-specific analysis of DNAm at individual CpG sites could reflect the relative composition of leukocytes. Furthermore, we conceived a method, based on DNAm patterns, for absolute quantification of cell counts.

Methods

SELECTION OF CANDIDATE CG DINUCLEOTIDES

For selection of cell-type-specific CpG sites, we used DNAm profiles of purified leukocyte subsets that were generated on the Illumina Infinium HumanMethylation450 BeadChip platform (Gene Expression Omnibus ID: GSE35069) *(18)*. We used β-values, ranging from 0 to 1, which roughly correspond to percentages of DNAm. CpG sites on X and Y chromosomes were excluded. Initially, we considered various statistical approaches, but due to the relatively small number of available data sets and because a large difference in DNAm is of particular relevance, we finally preselected candidate CpGs on the basis of the following 2 simple parameters: (*a*) we sorted CpGs by the difference between the mean β-value of 1 purified leukocyte subset and the mean β-value of all other subsets to select CpGs with the highest Δ for each subset and (*b*) we sorted CpGs by the sums of variation of β-values within each subset and all other subsets to select CpGs with relatively little variation between the biological replicates of the data set GSE35069. The discriminatory power of these CpG sites was then tested on an independent data set of purified leukocyte subsets provided on the Array Express database (E-MTAB-2145) *(19)*. Furthermore, CpGs with systematic DNAm changes upon aging or between male and female samples were excluded. For the remaining candidate CpGs, different combinations were tested for precision of cell-type predictions on the GSE35069 data set. Combinations that showed the highest linear correlation with known "real" leukocyte counts were further pursued for pyrosequencing assays. This work flow for selection of cell-type-specific CpGs is also depicted in Fig. 1 in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol64/issue3).

For cellular quantification, we selected CpGs that were consistently methylated across all hematopoietic cell types. We used the following DNAm profiles (all generated with HumanMethylation450 BeadChips): (*a*) purified leukocyte subsets: GSE35069 *(18)* and E-MTAB-2145 *(19)*; (*b*) whole blood from healthy donors: GSE32148 *(20)* and GSE41169 *(21)*; and (*c*) DNAm profiles of blood disorders such as acute myeloid leukemia: The Cancer Genome Atlas *(22)*, GSE58477 *(23)*, GSE62298 *(24)*; myelodysplastic syndrome: GSE51758 *(25)*; B-cell lymphoma: GSE37362 *(26)*; acute lymphoblastic leukemia: GSE69954 *(27)*. Candidate CpGs were selected that are consistently highly methylated in each of these data sets (mean β-value >0.975).

BLOOD SAMPLES

Peripheral blood samples for the training set (n = 60) and for validation set I (n = 44) were obtained from the Health Effects in High-Level Exposure to PCB (HELPcB) program *(28)*. The study was approved by the local ethics committee of the RWTH Aachen University (EK 176/11). Peripheral blood samples for validation set II (n = 70, including patient samples without hematopoietic malignancy), validation set III (n = 41), and validation set IV (n = 38), as well as serum samples (n = 18) were obtained from the Department of Hematology, Oncology, Hemostaseology, and Stem Cell Transplantation and from the Department of Transfusion Medicine according to the guidelines specifically approved by the local ethics committee of the RWTH Aachen University (EK 099/14).

CONVENTIONAL ANALYSIS OF BLOOD COUNTS

Blood samples from the HELPcB program were analyzed with the Sysmex XN-9000 hematology analyzer (Sysmex Deutschland GmbH) and immunophenotypic analysis was performed as previously described *(29)*. In brief, EDTA anticoagulated whole blood was incubated for 20 min at room temperature with fluorescently labeled antibody pairs (CD3/CD4, CD3/CD8, CD3/CD19, CD3/CD16+CD56) and isotype-matched controls (IgG1 FITC/IgG2a PE, all from Becton Dickinson). Erythrocytes were lysed with BD FACS lysing solution and leukocytes were analyzed on a FACSCalibur with use of the BD Simulset software (Becton Dickinson). LDCs of validation set II and IV were determined either (*a*) with an automated hematology analyzer (Coulter AcT diff2, Beckman Coulter), (*b*) by microscopic analysis of

blood smears, and/or (*c*) by immunophenotyping and flow cytometric analysis on a Navios flow cytometer (Beckman Coulter). Blood samples of validation set III were analyzed with an Abbott Cell-Dyn Emerald hematology system (Abbott Laboratories).

ISOLATION OF DNA AND BISULFITE CONVERSION

Genomic DNA was isolated from blood with the QIAamp DNA Mini Kit (Qiagen). Genomic DNA from 1 mL of serum was isolated with the PME free-circulating DNA extraction kit (GS/VL system; Analytik Jena). Either 1 μg of DNA from peripheral blood or the complete DNA sample from serum was bisulfite-converted with the EZ DNA Methylation Kit (Zymo Research).

GENERATION OF NONMETHYLATED REFERENCE DNA FOR QUANTIFICATION

Target regions were PCR amplified (Eppendorf Mastercycler 5341; Eppendorf AG), cloned into the pBR322 vector (Thermo Fischer), expanded in DH5α *E.coli*, and isolated with the plasmid DNA purification kit (Macherey-Nagel). Mixtures of blood and reference DNA were subjected to DNA isolation and bisulfite conversion, as described above.

PYROSEQUENCING

Specific regions covering the CpG site of interest were amplified by PCR (Eppendorf Mastercycler 5341) with primers as indicated in Table 1 in the online Supplemental Data. Pyrosequencing was performed on a PyroMark Q96 ID System with use of region-specific sequencing primers and results were analyzed with the PyroMark Q CpG software (Qiagen).

MASSARRAY ANALYSIS

Converted DNA was amplified by PCR with use of the HotStart Plus PCR Master Mix (Qiagen; Table 2 in the online Supplemental Data). A 10-μL portion of PCR product was in vitro transcribed and cleaved in a base-specific (U-specific) manner with use of RNase A (T-Cleavage MassCleave Kit; Agena Bioscience). The cleaved products were then analyzed by the MALDI-TOF mass spectrometer (MassARRAY Analyzer 4 System; Agena Bioscience).

DECONVOLUTION OF LEUKOCYTE SUBSETS ON THE BASIS OF DNA METHYLATION MEASUREMENTS

DNAm measurements can be represented by a matrix W of size f × k [f: number of CpGs (features); k: number of cell types]. The methylation data of the blood samples are represented by a matrix V of size f × n (n: number of blood samples) and are modeled as a linear combination of the purified cell types W, with their mixture proportions H [k × n matrix—each of the n columns corresponds to the mixture proportion of the respective blood sample (same column in V)]: $V \cong WH$.

For estimation of H, a nonnegative least-squares (NNLS) approach is used to avoid negative mixture proportions. For implementation purposes, we use the multiplicative update rule of Lee et al. *(30)*:

$$H^{j+1}_{a\mu} = H^{j}_{a\mu}\frac{(W^T V)_{a\mu}}{(W^T W H^j)_{a\mu}}$$

Here, j is the iteration index, W^T indicates the transpose of matrix W, and a and μ are the row and column indices, respectively. Leukocyte proportions were then adjusted to a total sum of 100%.

To inversely predict the percentages of DNAm in individual leukocyte subsets, we used the respective iterative formula for estimating W *(30)*:

$$W^{j+1}_{ia} = W^{j}_{ia}\frac{(VH^T)_{ia}}{(W^j H H^T)_{ia}}$$

QUANTIFICATION OF CELL NUMBERS ON THE BASIS OF DNA METHYLATION

On mixture of genomic DNA with a nonmethylated reference DNA, the amount of DNAm can be mathematically described as the ratio of methylated to total DNA:

$$DNAm = \frac{a \times C_R + b \times C_G}{C_R + C_G}$$

Here, C_R and C_G resemble the copy number of the reference DNA and the genomic DNA, respectively; a and b are absolute DNAm levels in controls consisting of either pure reference DNA or blood DNA, respectively (e.g., 7% and 93% DNAm in our analysis). To determine the copy number of the reference plasmids (C_R), we used the following formula:

$$C_R = 1.5 \times \frac{m_R \times N_A}{MW}$$

where m_R is the added reference amount (e.g., 0.011 ng of *LSM14B*), N_A is Avogadro's constant, and MW is the molar weight of the reference DNA (calculated for the plasmid with *LSM14B* sequence: 2.85×10^6 gmol^{-1}). The correction factor 1.5 was empirically determined and relates to the fact that purified plasmids comprise fragments of the bacterial genome or other plasmids. With these parameters, the copy numbers of the genomic DNA (C_G) can be inversely calculated, and hence the cell numbers in a given blood sample:

$$cells/\mu L = \frac{C_R \times (DNAm - a)}{2 \times v \times (b - DNAm)}$$

The term "2" stems from the fact that each cell comprises 2 copies of genomic DNA; v is the volume of analyzed blood in μL.

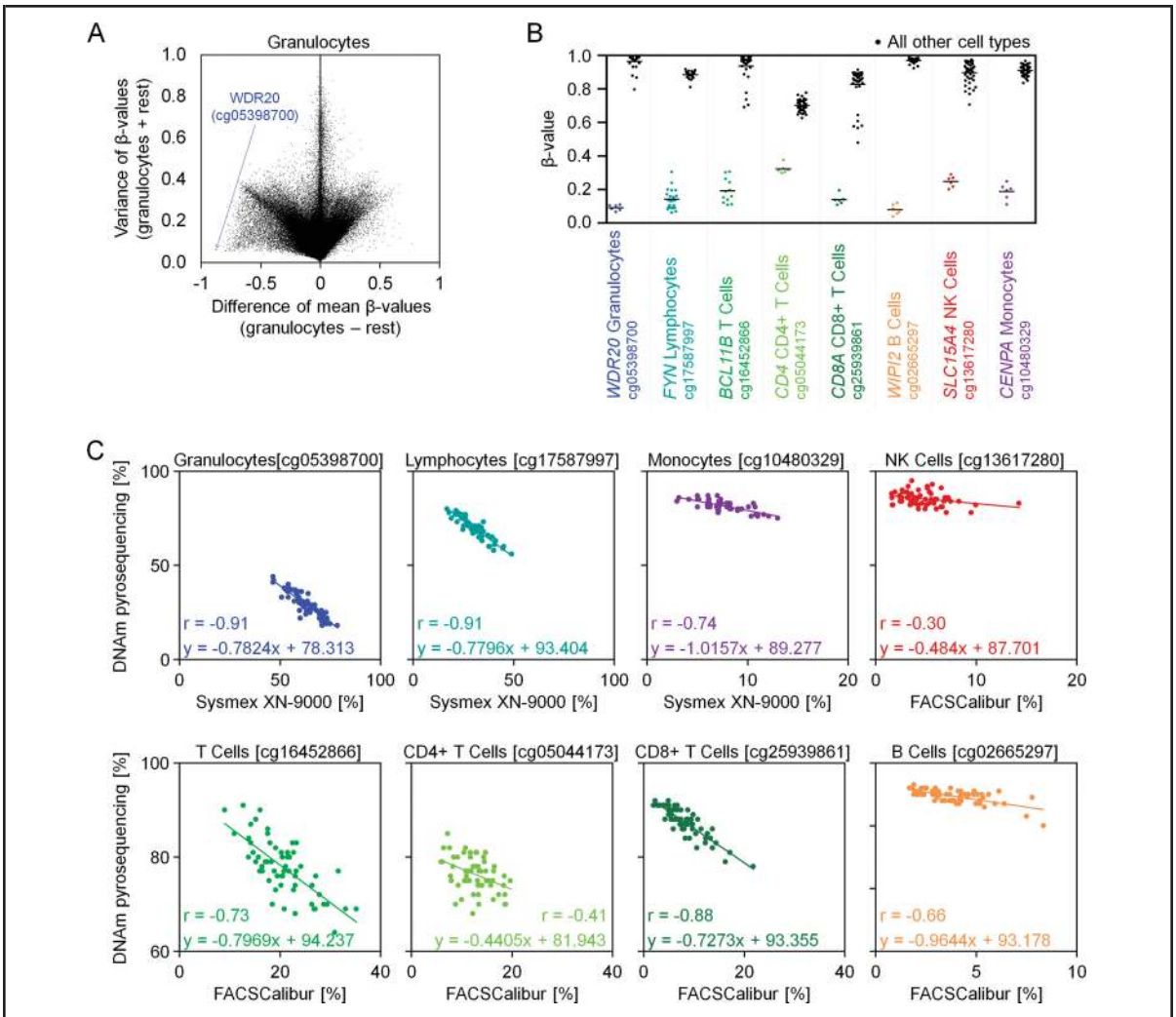

Fig. 1. Selection of cell-type-specific CpG sites for the Epi-Blood-Count.

Scatterplot to exemplarily depict selection criteria for the CpG site for granulocytes (cg05398700): (*i*) high difference between mean β-values in granulocytes and the mean β-values of all other cell types (rest) in DNAm profiles of purified cell types (GSE35069) *(18)*; and (*ii*) a low variance of β-values within the granulocytes and within the other hematopoietic subsets (A). For all selected candidate CpGs, the β-values of the reference data set (GSE35069) *(18)* are depicted for each leukocyte subset as compared to all other cell types (B). White blood counts in peripheral blood were analyzed with a Sysmex XN-9000 (for granulocytes, lymphocytes, and monocytes), and lymphocytes were further classified with a FACSCalibur (training set; n = 60) (C). DNAm levels at the cell-type-specific CpG sites were analyzed by pyrosequencing. Linear regression formulas are indicated for each cell type.

Results

IDENTIFICATION OF INDIVIDUAL CPG SITES TO DISCERN LEUKOCYTE SUBSETS

For selection of candidate CpGs, we used DNAm data of purified granulocytes, CD4+ T cells, CD8+ T cells, B cells, NK cells, and monocytes (GSE35069) *(18)*. For each of these cell types, we selected CpG sites that facilitated best discrimination based on the following 2 criteria: (*a*) highest difference in mean β-value of the subset and the mean β-value of all other hematopoietic cell types and (*b*) low variance of β-values within samples of the corresponding subset and within all other cell types. This analysis was initially performed for granulocytes (Fig. 1A) and then repeated for the other cell types (see Figs. 2 and 3 in the online Supplemental Data). Furthermore, DNAm profiles of T cells, B cells, and NK cells were combined to identify CpGs that reflected the entire lymphocyte population. Best performing CpG sites were validated on a second data set of purified leukocyte sub-

sets (E-MTAB-2145; see Fig. 4 in the online Supplemental Data) *(19)*.

For granulocytes, we selected a CpG site in the gene WD repeat domain 20 (*WDR20*;[12] cg05398700). Notably, CpGs with the highest discriminatory power for CD4+ T cells and CD8+ T cells are linked to the genes *CD4* (cg05044173) and *CD8A* (cg25939861), respectively. The selected CpG site for lymphocytes was in the gene FYN protooncogene (*FYN;* cg17587997); for T cells in B-cell CLL/lymphoma 11B (*BCL11B;* cg16452866); for B cells in WD repeat domain, phosphoinositide interacting 2 (*WIPI2;* cg02665297) that is involved in maturation of phagosomes; for NK cells in solute carrier family 15 member 4 (*SLC15A4;* cg13617280) that has been implicated in systemic lupus erythematosus; and for monocytes in centromere protein A (*CENPA*; cg10480329; Fig. 1B). Thus, our straightforward procedure identified CpGs that are associated with genes of relevant function in the corresponding cell types. To estimate whether DNAm at these CpGs is also reflected on gene expression level, we used microarray data sets of purified subsets (GSE28490) *(31)*. In fact, cell-type-specific hypomethylation of the relevant CpGs was often associated with higher gene expression, albeit this was not observed for *WDR20*, *WIPI2*, and *CENPA* (see Fig. 5 in the online Supplemental Data).

Subsequently, we analyzed if DNAm at our candidate CpGs correlated with the fractions of corresponding subsets. To this end, we established pyrosequencing assays for the selected CpG sites and analyzed 60 peripheral blood samples. Cell counts with a Sysmex XN-9000 hematology analyzer correlated well with DNAm at the respective CpG sites for granulocytes (Pearson correlation coefficient: $r = -0.91$), lymphocytes ($r = -0.91$), and monocytes ($r = -0.74$). Furthermore, immunophenotypic analysis correlated for T cells ($r = -0.73$), CD4+ T cells ($r = -0.41$), CD8+ T cells ($r = -0.88$), B cells ($r = -0.66$), and to a lesser extent for NK cells ($r = -0.30$; Fig. 1C). The candidate CpGs did not reveal a clear association with age or gender (see Fig. 6 in the online Supplemental Data). Taken together, DNAm measurements at our CpGs correlated with the frequency of corresponding leukocyte subsets in whole blood samples.

DECONVOLUTION OF GRANULOCYTES, MONOCYTES, AND LYMPHOCYTES

Subsequently, we analyzed if the fractions of granulocytes, monocytes, and lymphocytes can be recapitulated in 44 independent blood samples by pyrosequencing of DNAm at the 3 relevant CpGs. Initially, the percentages of cells were simply calculated on the basis of the linear regression formulas of the subsets in the training set (Fig. 1C). In comparison to measurements of the Sysmex XN-9000 analyzer, these linear regression models revealed a high correlation ($r = 0.99$ across all cell types). The mean absolute deviation was only 3.2% for granulocytes, 2.2% for lymphocytes, and 1.4% for monocytes (Fig. 2A).

Alternatively, we integrated the DNAm levels of the 3 CpGs into an NNLS linear regression model. This model was trained on 60 blood samples of the training set and subsequently termed "Epi-Blood-Count." The NNLS linear regression approach does not depend on an a priori database of cell-type-specific DNAm reference profiles for the selected CpG sites. In fact, DNAm estimates based on deconvolution were very similar to the β-values of DNAm profiles of purified subsets *(18)* (Fig. 2B). This approach gave similar accuracies as the linear regression formulas for individual CpGs (Fig. 2C). To simplify application for the users, an Excel calculator for the 3-CpG NNLS model is provided in Table 3 in the online Supplemental Data.

There are notorious differences between cell counting systems *(1)*. Therefore, we applied the Epi-Blood-Count on a second validation set (in total 70 blood samples) that were either measured with a Coulter counter (Coulter ACT diff2; n = 24), and/or by manual counting of blood smears by highly specialized laboratory staff (n = 66). Coulter counter results revealed high correlation with Epi-Blood-Count, albeit mean numbers of granulocytes and lymphocytes were underestimated by 4.4% and overestimated by 5.5%, respectively, indicating that there might be a systemic deviation between the 2 analyzers (Fig. 2D). The correlation between manual blood counts and Epi-Blood-Count was slightly lower (Fig. 2E), but direct comparison of Coulter counter results and manual counting revealed lower correlations, too (Fig. 7 in the online Supplemental Data). Furthermore, we have exemplarily analyzed if the Epi-Blood-Count was also applicable to MassARRAY measurements. The DNAm measurements by pyrosequencing and MassARRAY analysis correlated, but our pyrosequencing measurements appeared to be more reliable (Fig. 8 in the online Supplemental Data).

Subsequently, the accuracy of the Epi-Blood-Count was determined with blood samples that had been stored for 7 days at 4 °C (n = 10). The results correlated with manual counting at day 0, but mean numbers of granulocytes and lymphocytes were underestimated by 9.1% or overestimated by 7.2%, respectively (Fig. 2F). A similar shift has been described for automated analyzers upon storage of blood samples for only 72 h *(32)*. Furthermore, Epi-Blood-Count was compared in aliquots of

[12] Human Genes: *WDR20,* WD repeat domain 20; *CD4,* CD4 molecule; *CD8A,* CD8a molecule; *FYN,* FYN protooncogene, Src family tyrosine kinase; *BCL11B,* B-cell CLL/lymphoma 11B; *WIPI2,* WD repeat domain, phosphoinositide interacting 2; *SLC15A4,* solute carrier family 15 member 4; *CENPA,* centromere protein A; *LSM14B,* LSM family member 14B; *ZC3H3,* zinc finger CCCH-type containing 3.

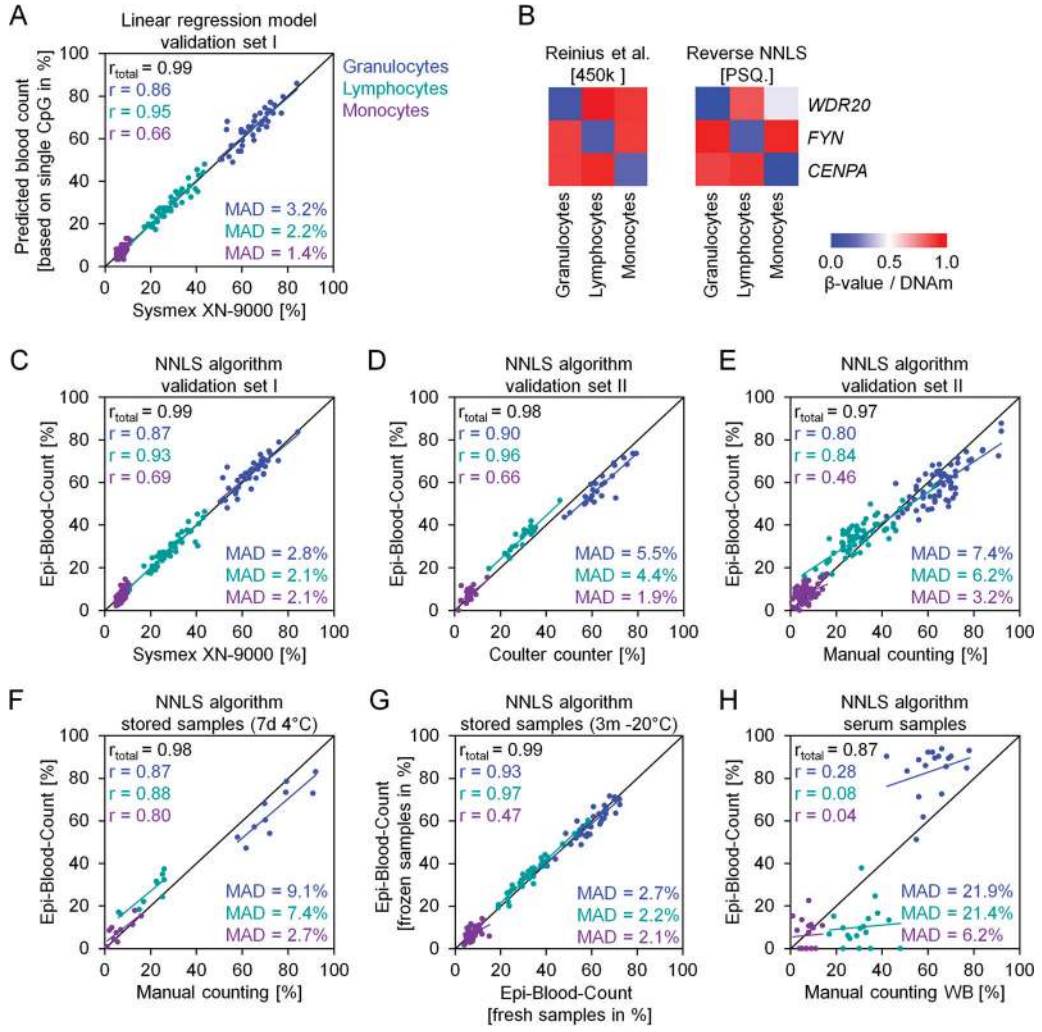

Fig. 2. Epi-Blood-Count of granulocytes, lymphocytes, and monocytes.

Blood samples of validation set I (n = 44) were analyzed by pyrosequencing at the 3 CpGs related to granulocytes, lymphocytes, and monocytes (A). The DNAm levels were implemented into the linear regression formulas of Fig. 1C to estimate cell fractions, and the results correlated with measurements on a Sysmex XN-9000 hematology analyzer. The heat maps compare the β-values for purified cell types in the reference data set (Reinius et al., 450k Bead Chip; GSE35069) and estimated DNAm levels for these cell types on the basis of deconvolution of pyrosequencing measurements (PSQ) (B). These estimations are based on DNAm levels at the 3 CpGs for granulocytes (*WDR20*), lymphocytes (*FYN*), and monocytes (*CENPA*) in whole blood of the training set (n = 60) that were then implemented into the reverse approach of the nonnegative least-squares (NNLS) linear model. Estimates for DNAm levels of individual subsets were in line with β-values of purified subsets in microarray data. These estimates for DNAm levels were then used for NNLS predictions in independent validation sets, and Epi-Blood-Count results correlated with measurements on a Sysmex XN-9000 hematology analyzer (C; n = 44), on a Coulter counter (D; n = 24), and microscopic analysis of blood smears and manual counting by a trained operator (E; n = 66). Epi-Blood-Count measurements were tested on blood samples after storage for 7 days at 4 °C (F; n = 10), or for 3 months at −20 °C (G; n = 41). Epi-Blood-Count was used to estimate the cell type of origin of cell-free circulating DNA (cfDNA) in serum samples (H). Measurements are compared to cell counts in whole blood (WB) by manual counting. These results indicate that cfDNA is particularly derived from granulocytes. r = Pearson correlation coefficient; MAD = mean absolute deviation.

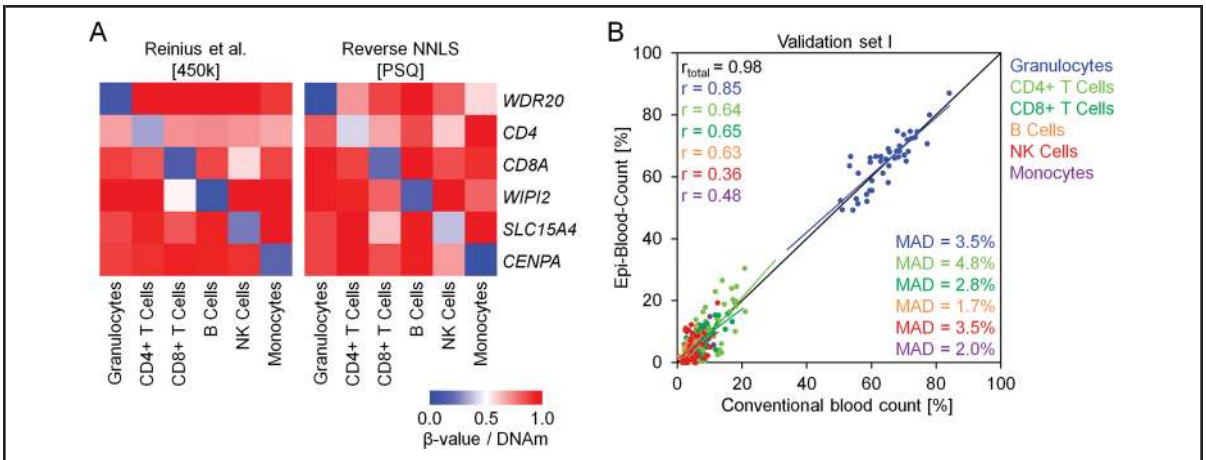

Fig. 3. Leukocyte differential counts with 6 CpG sites.

The heat maps compare the β-values of the relevant CpGs in the reference DNAm data sets from Reinius et al. (GSE35069) and estimated DNAm levels based on the reverse approach of NNLS with use of the pyrosequencing (PSQ) results of the training set (n = 60; in analogy to Fig. 2B) (A). These estimates of DNAm values were subsequently implemented into the NNLS matrix to estimate the proportions of cell types. Leukocyte differential counts were determined based on pyrosequencing of the 6 CpGs in validation set I (n = 44) (B). The results correlated with conventional measurements on a Sysmex XN-9000 hematology analyzer and immunophenotypic analysis with FACSCalibur. r = Pearson correlation coefficient; MAD = mean absolute deviation.

fresh blood vs storage for 3 months at −20 °C (validation set III; n = 41). The results revealed a high correlation ($r = 0.99$; Fig. 2G), indicating that the Epi-Blood-Count was applicable to frozen samples. Another advantage of the Epi-Blood-Count is that it requires only very small amounts of DNA. We have exemplarily used this approach to estimate the origin of cell-free circulating DNA (cfDNA) in serum. The results indicated that cfDNA in serum is particularly derived from granulocytes, which indeed have a very short half-life *(33)* (Fig. 2H).

ADDITIONAL CLASSIFICATION OF LYMPHOCYTE SUBSETS

The Epi-Blood-Count was further extended to classify lymphocytes subsets. To this end, DNAm at the candidate CpGs for B cells, NK cells, CD4+ T cells, and CD8+ T cells were analyzed by pyrosequencing in the 60 blood samples from the training set. To estimate DNAm in leukocyte subsets, we imputed immunophenotypic and DNAm measurements into the NNLS regression model. With this deconvolution approach, the estimated percentages of DNAm for each hematopoietic subset closely resembled the β-values of the purified subsets in DNAm profiles *(18)* (Fig. 3A). The 6-CpG Epi-Blood-Count model was tested on the training set (Fig. 9A in the online Supplemental Data) and on 2 independent validation sets (Fig. 3B; and see Fig. 9B in the online Supplemental Data). Immunophenotypic analysis and Epi-Blood-Count revealed a clear correlation: across all cell types the correlation coefficient was $r = 0.98$ with a mean of 3.1% for the mean absolute deviation. An Excel calculator for the 6-CpG NNLS model is provided in Table 4 in the online Supplemental Data. Furthermore, the measurements were also relatively stable after storage of blood samples at 4 °C for 7 days without fixation (see Fig. 9C in the online Supplemental Data).

QUANTIFICATION OF CELL NUMBERS ON THE BASIS OF DNA METHYLATION

We reasoned that quantification of cell numbers on the basis of DNAm would be feasible if samples were supplemented with a suitable reference DNA of known concentration (Fig. 4A). To this end, we identified 3 CpG sites that were consistently highly methylated (β-value >0.975) across DNAm profiles of leukocyte subsets and of whole blood of healthy individuals, patients with leukemia, or patients with lymphoma (see Fig. 10 in the online Supplemental Data). The selected CpG sites were within "like SM" domain (LSM) family member 14B (*LSM14B*; cg06096175), zinc finger CCCH-type containing 3 (*ZC3H3*; cg25834632), and a CpG site not associated with any gene (cg09414987). The corresponding sequences were cloned into plasmids to obtain nonmethylated reference DNAs.

Initially, we analyzed serial dilutions of reference DNA (*LSM14B*) in 2 independent peripheral blood samples (Fig. 4B). Notably, the results were in line with theoretical estimates by mathematical calculation, indicating that the method was robust for cellular quantification (Fig. 4C). The precision of this approach was particularly high for DNAm levels between 20% and 80%, if copy numbers of reference DNA and genomic DNA were sim-

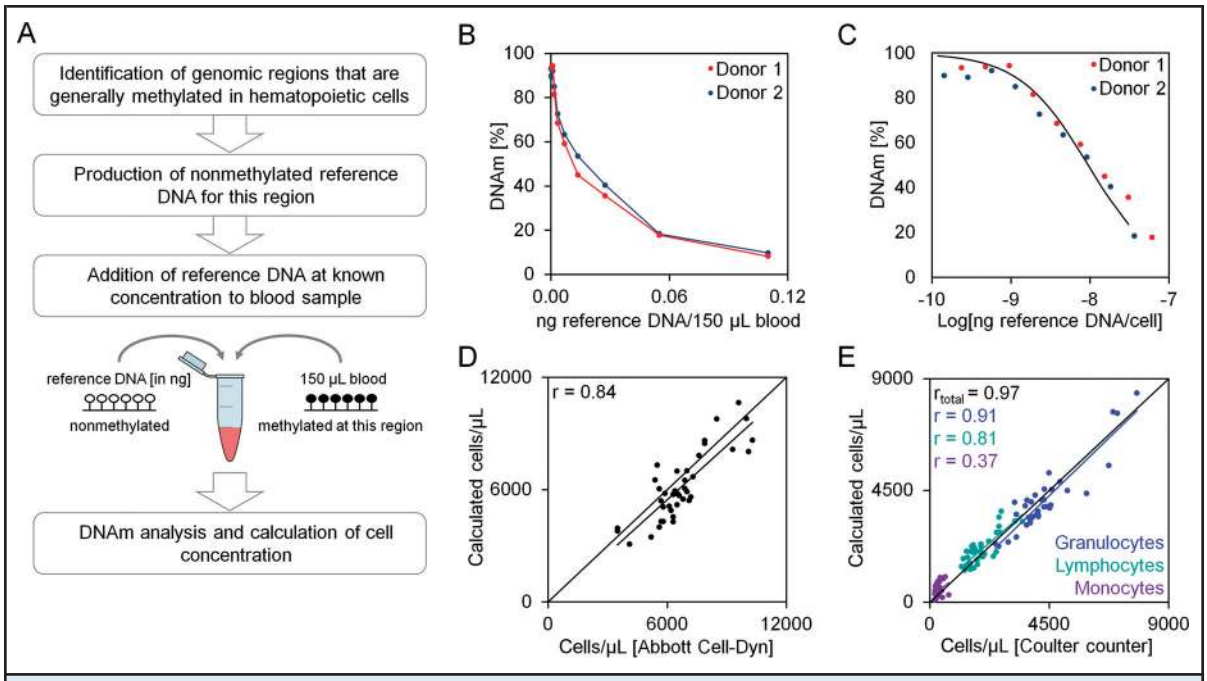

Fig. 4. Cell quantification based on DNA methylation.

Schematic presentation of cellular quantification based on DNAm levels with a nonmethylated reference DNA (A).Two blood samples (donor 1 and 2; 150 µL) were mixed with a serial dilution of the reference plasmid comprising the nonmethylated sequence of *LSM14B* (0.0002 ng to 0.1100 ng) (B). DNAm levels (analyzed by pyrosequencing) continuously declined with higher concentrations of reference DNA. If the results were plotted as a logarithmic ratio of reference DNA [ng] per cell (determined with the Abbott Cell-Dyn), there was an almost linear association in the DNAm range between 20% and 80%. Notably, the observed DNAm levels closely resembled the mathematically expected DNAm levels (black curve) (C). Calculated cell numbers based on the reference plasmid *LSM14B* clearly correlated with cell numbers determined with the Abbott Cell-Dyn analyzer (n = 41; validation set III) (D). Furthermore, epigenetic quantification could be combined with epigenetic LDCs: Cell numbers for granulocytes, lymphocytes, and monocytes correlated with cell numbers determined with a Coulter counter (n = 38; validation set IV) (E). r = Pearson correlation coefficient; MAD = mean absolute deviation.

ilar. To increase this range, we used the other 2 reference DNAs at higher and lower concentration, respectively (see Fig. 11 in the online Supplemental Data).

Subsequently, we mixed 150 µL of frozen blood (n = 41; validation set III) with our *LSM14B* reference DNA and analyzed DNAm at the relevant CpG site by pyrosequencing. The calculated cell numbers correlated well with cell counts that were automatically measured in fresh blood ($r = 0.84$; Fig. 4D). Furthermore, combined epigenetic analysis of relative LDCs with absolute cell numbers correlated with measurements of an automated hematology analyzer for individual leukocyte subsets (n = 38; validation set IV; $r = 0.97$; Fig. 4E).

Discussion

Analysis of DNAm patterns in blood holds enormous diagnostic potential. We demonstrate that site-specific analysis at individual CpG sites facilitates relative quantification of leukocyte subpopulations. In analogy, immunophenotypic analysis is based on individual cell-type-specific epitopes. Notably, several candidate CpGs of the Epi-Blood-Count are related to the same genes addressed in immunophenotypic analysis. Overall, the precision of the Epi-Blood-Count was comparable to the well-established conventional methods *(1, 34)*.

Other groups have previously described LDC algorithms based on genome-wide DNAm profiles of Illumina BeadChip microarrays *(11, 12, 35)*. This enables combination of a multitude of CpGs into bioinformatic predictors, which generally increases the precision of epigenetic signatures *(36)*. On the other hand, the precision of DNAm measurements at individual CpGs is higher in pyrosequencing data than β-values on Illumina BeadChips *(37)*. Microarray analysis is relatively time-consuming and expensive. This might be the reason why the number of available DNAm profiles with matched flow cytometric analysis is still relatively low. Reinius et al. provided flow cytometric analysis for 6 DNAm profiles *(18)*, and Absher and colleagues provided 44

DNAm profiles with conventional LDCs *(38)*. Notably, the precision of genome-wide algorithms on these data sets was similar to the performance of the Epi-Blood-Count in our cohorts (see Tables 5 and 6 in the online Supplemental Data) *(14)*. Furthermore, Koestler and coworkers compared their microarray-based predictions with complete blood counts, and the correlation for monocytes ($r = 0.60$) and lymphocytes ($r = 0.61$) was not better than our 3-CpG Epi-Blood-Count *(36)*. Either way, site-specific analysis of individual CpGs by pyrosequencing is better applicable to daily routine in clinical diagnostics than microarray analysis of DNAm profiles: analysis is feasible in 2 days and might be implemented into semiautomated procedures.

It remains to be demonstrated if the Epi-Blood-Count is also applicable to patient material. Particularly, hematopoietic malignancies have a major effect on the epigenetic makeup that needs to be taken into account. So far, the Epi-Blood-Count does not consider eosinophils, basophils, immature granulocytic precursors, or more specialized lymphocyte subsets such as naïve, memory, or regulatory T cells. Furthermore, we expect that it should be possible to integrate CpGs that are indicative for blasts, atypical lymphocytes, and hematopoietic progenitors for extended epigenetic differential counts. Alternative methods for DNAm analysis, such as barcoded bisulfite amplicon sequencing or digital PCR, may ultimately pave the way for more sensitive deconvolution of rare subsets. Furthermore, analysis of neighboring CpG sites of the same amplicon may increase robustness as described for detection of circulating tumor DNA *(39)*. It is, however, unlikely that epigenetic analysis of LDCs will completely replace the conventional cell counters, because it cannot address erythrocytes and thrombocytes, which hardly comprise DNA.

In this study, we describe an entirely new approach for cellular quantification based on DNAm that is based on addition of a nonmethylated reference sequence of known concentration. In analogy, quantification of cell numbers has been established in flow cytometry by addition of beads as quantification standards *(40)*. Our results indicate that the DNAm-based approach reaches a similar precision as manual, semiautomated, and automated cell counts *(41)*, but it is also applicable to cryopreserved samples.

In summary, our Epi-Blood-Count has various advantages over the well-established conventional methods: (*a*) blood can be frozen after sampling for long-term storage, shipment, and subsequent analysis; (*b*) it is applicable to small volumes of blood (few microliters, whereas at least 700 μL is required for immunophenotypic analysis); and (*c*) DNAm levels at individual CpGs provide an absolute measure that may facilitate better standardization between laboratories than immunophenotypic analysis by flow cytometry. Our proof-of-concept study therefore opens the door for epigenetic white blood cell counts in clinical application.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:*

Employment or Leadership: U. Gerstenmaier, varionostic GmbH.
Consultant or Advisory Role: None declared.
Stock Ownership: U. Gerstenmaier, varionostic GmbH.
Honoraria: None declared.
Research Funding: M. Lenz, the Ministry for Innovation, Science and Research of German Federal State of North Rhine-Westphalia, Germany, and the Dutch Province of Limburg, the Netherlands; W. Wagner, the Else Kröner-Fresenius-Stiftung (2014_A193), the German Research Foundation (WA 1706/8–1), the German Ministry of Education and Research (01KU1402B).
Expert Testimony: None declared.
Patents: RWTH Aachen University Medical School has applied for relevant patents for the Epi-Blood-Count and quantification of cell numbers based on DNAm. J. Frobel and W. Wagner, EP17163798.6 (patent application), T. Božić and W. Wagner, Az: 10 2017 004 108.3 (patent application).

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or final approval of manuscript.

References

1. Buttarello M, Plebani M. Automated blood cell counts: state of the art. Am J Clin Pathol 2008;130:104–16.
2. Roussel M, Benard C, Ly-Sunnaram B, Fest T. Refining the white blood cell differential: the first flow cytometry routine application. Cytometry A 2010;77:552–63.
3. Briggs C, Culp N, Davis B, d'Onofrio G, Zini G, Machin SJ, International Council for Standardization in Haematology (ICSH). ICSH guidelines for the evaluation of blood cell analysers including those used for differential leukocyte and reticulocyte counting. Int J Lab Hematol 2014;36:613–27.
4. Roussel M, Davis BH, Fest T, Wood BL, International Council for Standardization in Haematology (ICSH). Toward a reference method for leukocyte differential counts in blood: comparison of three flow cytometric candidate methods. Cytometry A 2012;81:973–82.
5. Cherian S, Levin G, Lo WY, Mauck M, Kuhn D, Lee C, Wood BL. Evaluation of an 8-color flow cytometric reference method for white blood cell differential enumeration. Cytometry B Clin Cytom 2010;78:319–28.
6. Zini G. Stability of complete blood count parameters with storage: toward defined specifications for different diagnostic applications. Int J Lab Hematol 2014;36:111–3.
7. 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;12:453–7.
8. Shen-Orr SS, Gaujoux R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. Curr Opin Immunol 2013;25:571–8.
9. Gong T, Hartmann N, Kohane IS, Brinkmann V, Staedtler F, Letzkus M, et al. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. PLoS One 2011;6:e27156.
10. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z,

Clark HF. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS One 2009;4:e6098.
11. Accomando WP, Wiencke JK, Houseman EA, Nelson HH, Kelsey KT. Quantitative reconstruction of leukocyte subsets using DNA methylation. Genome Biol 2014; 15:R50.
12. 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.
13. McGregor K, Bernatsky S, Colmegna I, Hudson M, Pastinen T, Labbe A, Greenwood CM. An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. Genome Biol 2016;17:84.
14. Waite LL, Weaver B, Day K, Li X, Roberts K, Gibson AW, et al. Estimation of cell-type composition including T and B cell subtypes for whole blood methylation microarray data. Front Genet 2016;7:23.
15. Houseman EA, Kim S, Kelsey KT, Wiencke JK. DNA methylation in whole blood: uses and challenges. Curr Environ Health Rep 2015;2:145–54.
16. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol 2014;15:R31.
17. Adalsteinsson BT, Gudnason H, Aspelund T, Harris TB, Launer LJ, Eiriksdottir G, et al. Heterogeneity in white blood cells has potential to confound DNA methylation measurements. PLoS One 2012;7:e46705.
18. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlen SE, Greco D, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS One 2012; 7:e41361.
19. 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:e52–60.
20. Harris RA, Nagy-Szakal D, Pedersen N, Opekun A, Bronsky J, Munkholm P, et al. Genome-wide peripheral blood leukocyte DNA methylation microarrays identified a single association with inflammatory bowel diseases. Inflamm Bowel Dis 2012;18:2334–41.
21. Horvath S, Levine AJ. HIV-1 infection accelerates age according to the epigenetic clock. J Infect Dis 2015; 212:1563–73.
22. The Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med 2013;368:2059–74.
23. Qu Y, Lennartsson A, Gaidzik VI, Deneberg S, Karimi M, Bengtzen S, et al. Differential methylation in CN-AML preferentially targets non-CGI regions and is dictated by DNMT3A mutational status and associated with predominant hypomethylation of HOX genes. Epigenetics 2014;9:1108–19.
24. Ferreira HJ, Heyn H, Vizoso M, Moutinho C, Vidal E, Gomez A, et al. DNMT3A mutations mediate the epigenetic reactivation of the leukemogenic factor MEIS1 in acute myeloid leukemia. Oncogene 2016; 35:3079–82.
25. Zhao X, Yang F, Li S, Liu M, Ying S, Jia X, Wang X. CpG island methylator phenotype of myelodysplastic syndrome identified through genome-wide profiling of DNA methylation and gene expression. Br J Haematol 2014;165:649–58.
26. Asmar F, Punj V, Christensen J, Pedersen MT, Pedersen A, Nielsen AB, et al. Genome-wide profiling identifies a DNA methylation signature that associates with TET2 mutations in diffuse large B-cell lymphoma. Haematologica 2013;98:1912–20.
27. Borssen M, Haider Z, Landfors M, Noren-Nystrom U, Schmiegelow K, Asberg AE, et al. DNA methylation adds prognostic value to minimal residual disease status in pediatric T-cell acute lymphoblastic leukemia. Pediatr Blood Cancer 2016;63:1185–92.
28. Schettgen T, Gube M, Esser A, Alt A, Kraus T. Plasma polychlorinated biphenyls (PCB) levels of workers in a transformer recycling company, their family members, and employees of surrounding companies. J Toxicol Environ Health A 2012;75:414–22.
29. Haase H, Fahlenkamp A, Schettgen T, Esser A, Gube M, Ziegler P, et al. Immunotoxicity monitoring in a population exposed to polychlorinated biphenyls. Int J Environ Res Public Health 2016;13.
30. Lee DD, Seung HS. Algorithms for non-negative matrix factorization. Adv Neural Inform Process Systems 2001; 13:556–62.
31. Allantaz F, Cheng DT, Bergauer T, Ravindran P, Rossier MF, Ebeling M, et al. Expression profiling of human immune cell subsets identifies miRNA-mRNA regulatory relationships correlated with cell type specific expression. PLoS One 2012;7:e29979.
32. Joshi A, McVicker W, Segalla R, Favaloro E, Luu V, Vaniasinkam T. Determining the stability of complete blood count parameters in stored blood samples using the SYSMEX XE-5000 automated haematology analyser. Int J Lab Hematol 2015;37:705–14.
33. Summers C, Rankin SM, Condliffe AM, Singh N, Peters AM, Chilvers ER. Neutrophil kinetics in health and disease. Trends Immunol 2010;31:318–24.
34. Estridge BH, Reynolds AP. Basic clinical laboratory techniques. 6th ed. Clifton Park (NY): Delmar Cengage Learning; 2011.
35. 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;18:105.
36. 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;8:816–26.
37. BLUEPRINT consortium. Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. Nat Biotechnol 2016;34: 726–37.
38. 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:e1003678.
39. Lehmann-Werman R, Neiman D, Zemmour H, Moss J, Magenheim J, Vaknin-Dembinsky A, et al. Identification of tissue-specific cell death using methylation patterns of circulating DNA. Proc Natl Acad Sci USA 2016; 113:E1826–34.
40. Montes M, Jaensson EA, Orozco AF, Lewis DE, Corry DB. A general method for bead-enhanced quantitation by flow cytometry. J Immunol Methods 2006;317:45–55.
41. Cadena-Herrera D, Esparza-De Lara JE, Ramírez-Ibañez ND, López-Morales CA, Perez NO, Flores-Ortiz LF, Medina-Rivero E. Validation of three viable-cell counting methods: Manual, semi-automated, and automated. Biotechnology Reports 2015;7:9–16.