nature methods

SUPPLEMENTARY INFORMATION

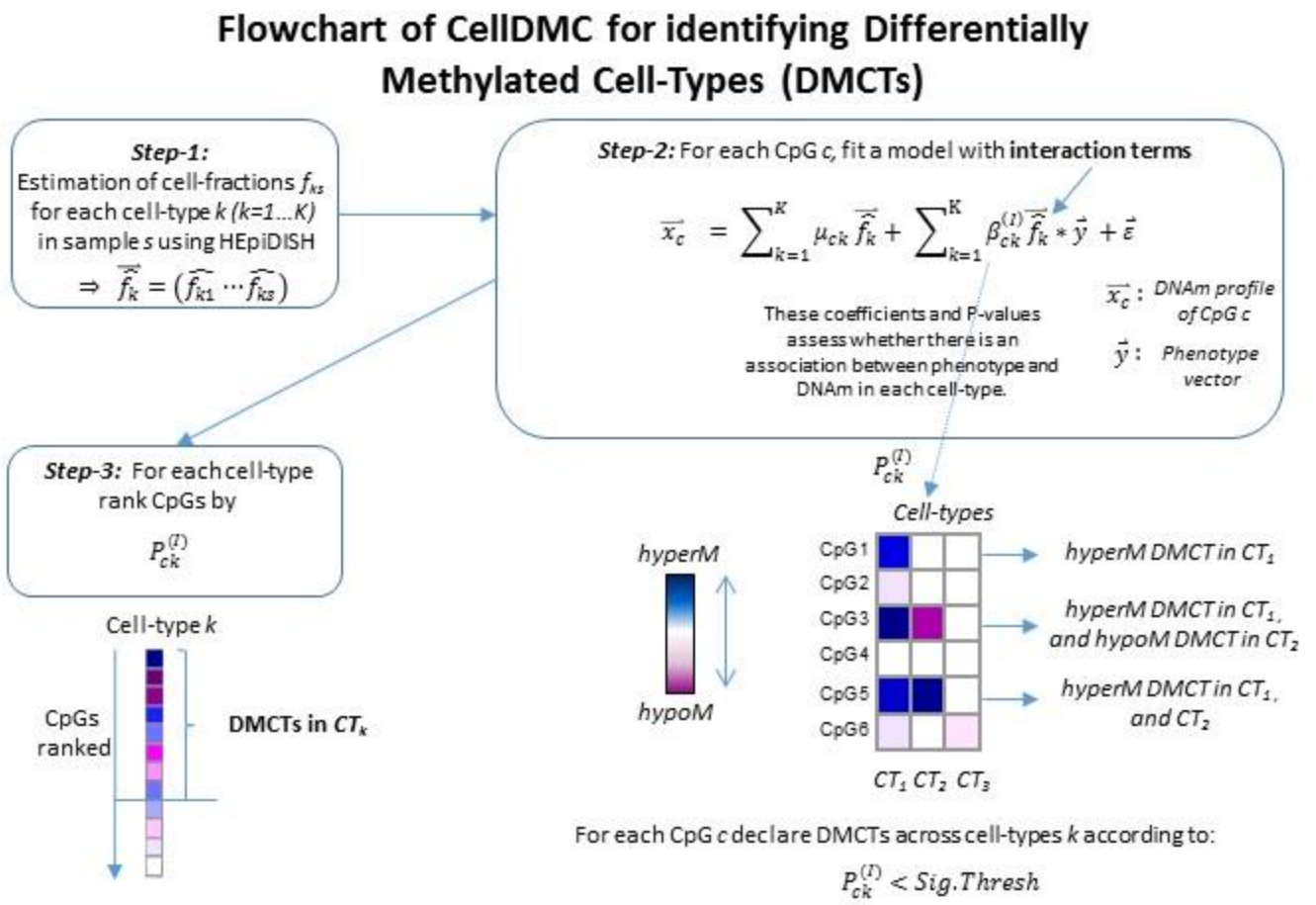
https://doi.org/10.1038/s41592-018-0213-x

In the format provided by the authors and unedited.

Identification of differentially methylated cell types in epigenome-wide association studies

Shijie C. Zheng[1], Charles E. Breeze[2], Stephan Beck[3] and Andrew E. Teschendorff[1,3,4]*

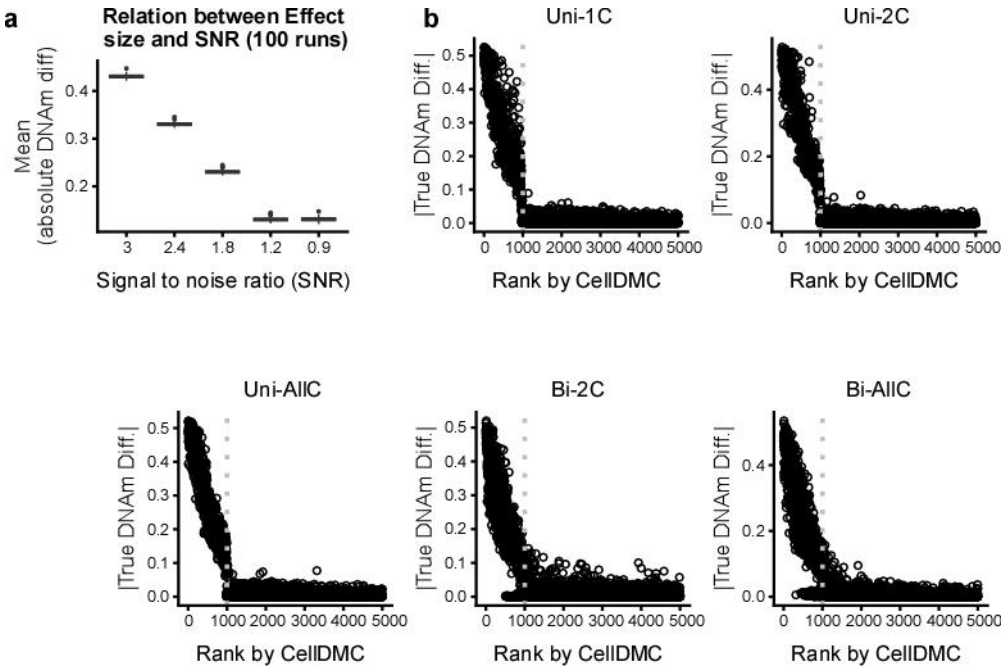
[1]CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China. [2]Altius Institute for Biomedical Sciences, Seattle, WA, USA. [3]UCL Cancer Institute, Paul O'Gorman Building, University College London, London, UK. [4]Department of Women's Cancer, University College London, London, UK. *e-mail: a.teschendorff@ucl.ac.uk



Supplementary Figure 1

CellDMC flowchart.

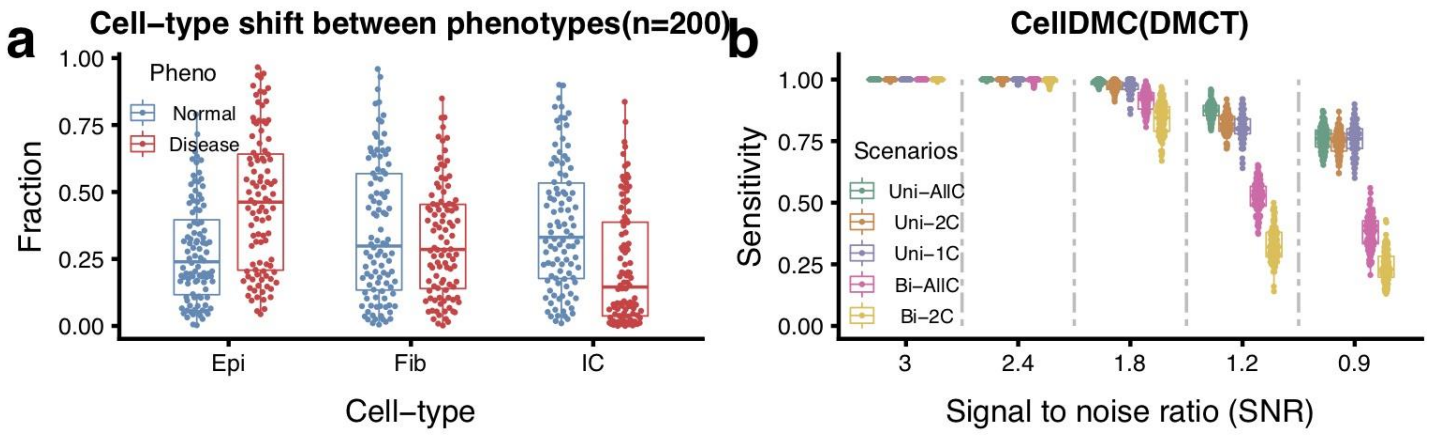
CellDMC consists of three main steps: (Step-1) Estimation of cell-type fractions in a given sample using a DNAm reference matrix for *K* cell-types and the DNAm profile of the sample *s*. The estimation procedure is based on the Hierarchical EpiDISH (HEpiDISH) algorithm and uses DNAm references for a generic epithelial cell, fibroblasts and for the seven main immune cell subtypes. (Step-2) For each CpG, a linear model is fit to the CpG's DNAm profile, which includes covariates for all cell-type fractions and interaction terms between cell-type fractions and phenotype (only K-1 terms are independent in each summation). The P-values for the interaction terms inform on the nature of the association (if any) between phenotype, cell-type and DNAm. For instance, as shown for CpG1, only the P-value of the interaction term for cell-type CT_1 is highly significant, indicating that CpG1 is a DMCT only in cell-type CT_1. The sign of the estimated regression coefficient of the interaction term is positive indicating that CpG1 is hypermethylated in cases vs controls. CpGs are declared significant DMCTs across cell-types according to the rule as indicated. (Step-3) Finally, for all significant DMCTs, these can then be ranked for each cell-type, according to the P-value of the interaction term.



Supplementary Figure 2

Choice of signal-to-noise ratio (SNR) and validation of CellDMC's ranking procedure

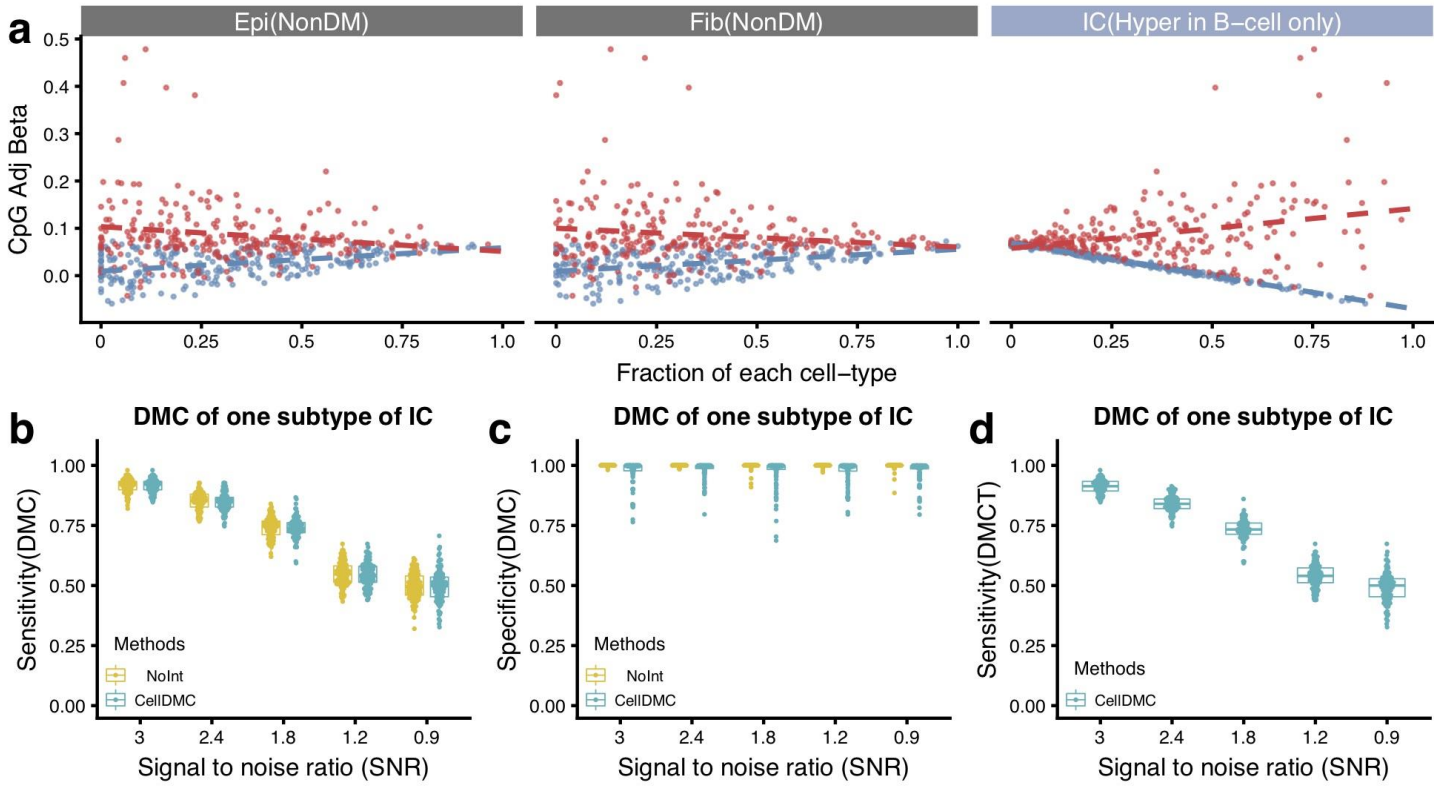
a) Box plot of the effect size (i.e. the average of absolute DNAm differences between cases and controls within an affected cell type, averaged over 100 DMCs) against the corresponding SNR value. Each box plot contains 100 values representing 100 different Monte Carlo runs. Thus, for SNR ~ 3, the mean difference in DNAm between case and control within an affected cell type is just over 0.4. The cases SNR ~ 1.2 and SNR ~ 0.9 do not differ in terms of the average difference in DNAm, but differ in terms of the variance, with a larger variance in the SNR ~ 0.9 case. Boxes indicate 25th and 75th centiles. Whiskers extend to the largest values no further than 1.5 times the IQR from these centiles, where IQR is the interquartile range. **b)** Scatterplots of the absolute true methylation difference of DMCTs between cases and controls (y-axis) against CellDMC's predicted rank position (x-axis), for a range of different simulation scenarios (Uni-1C, Uni-2C, Uni-AllC and Bi-2C,Bi-AllC) with 1,000 true DMCTs.



Supplementary Figure 3

CellDMC's sensitivity for detecting DMCTs in a scenario where there is a shift in the mean cell-type fractions between cases and controls.

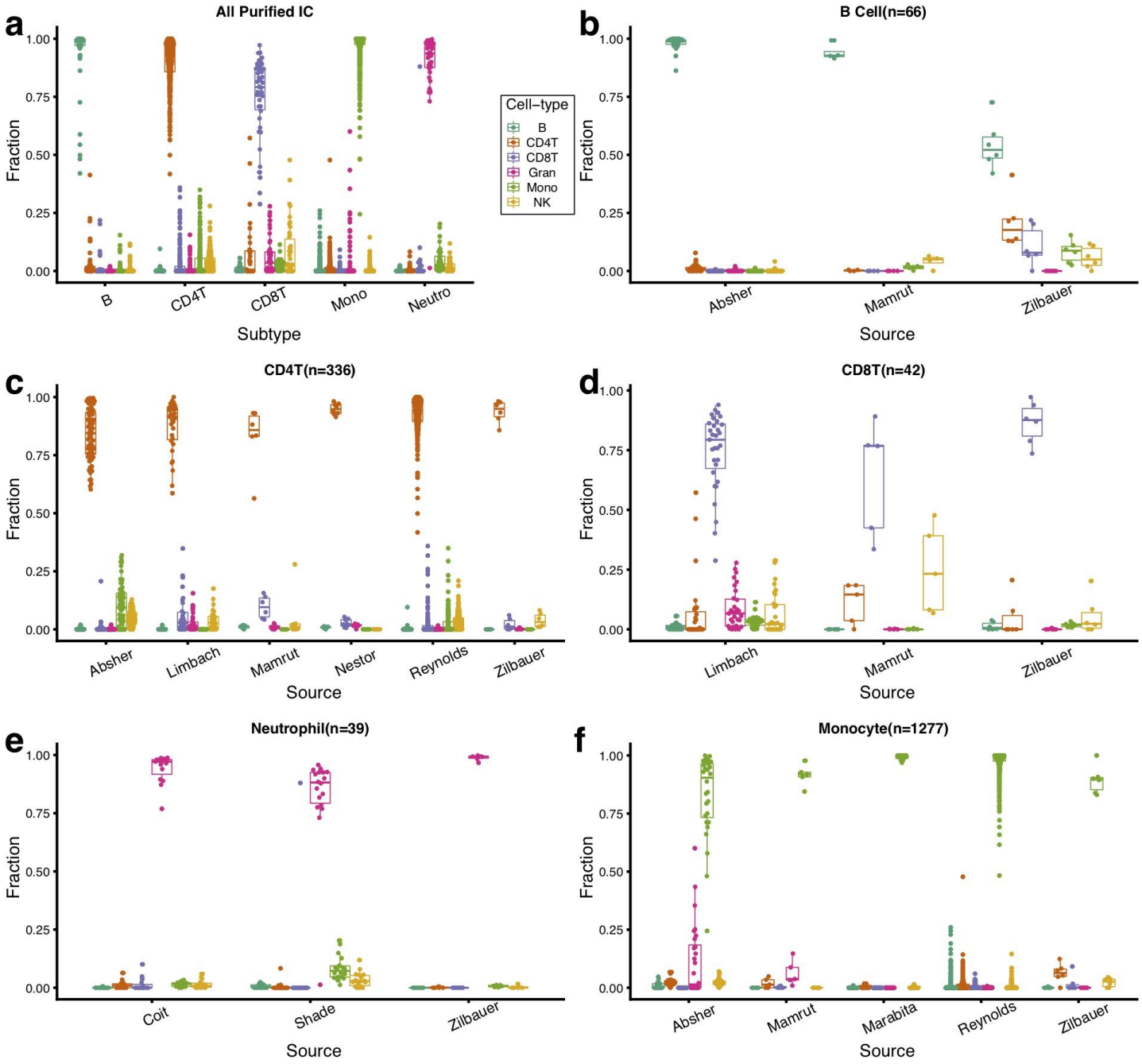
a) Box plot of cell-type fractions between cases and controls for one illustrative Monto Carlo run. **b)** Sensitivity measure to detect DMCTs for CellDMC. All results were based on 100 Monte Carlo runs, with 100 'control' in silico mixtures and 100 'disease' in silico mixtures. Boxes indicate 25th and 75th centiles. Whiskers extend to the largest values no further than 1.5 times the interquartile range from these centiles.



Supplementary Figure 4

Performance of CellDMC and non-interaction-based model (NoInt) when only one IC subtype is differentially methylated.

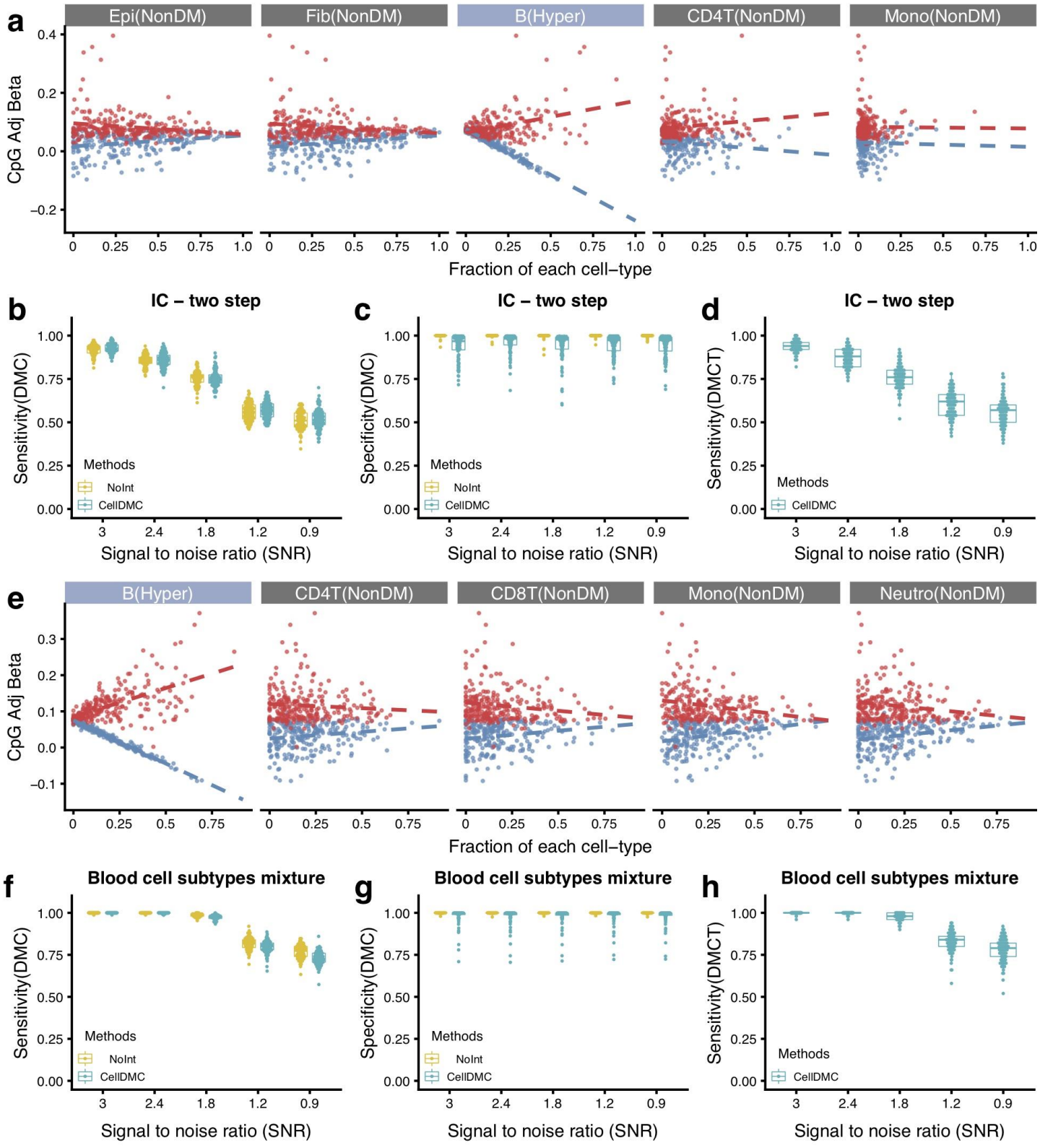
Unlike previous figures, where differential methylation in the immune cell (IC) compartment affected all immune cell subtypes equally, we are now considering only one immune cell subtype (here B cells) to be altered. **a)** Scatterplot of the adjusted beta-value (adjusted for all cell-type fractions) against cell-type fraction for a hypothetical example where differential methylation happens for only one IC cell subtype (e.g. B cells). **b, c)** Corresponding sensitivity and specificity measures of CellDMC and non-interaction-based model (NoInt) for detecting DMCs. **d)** Sensitivity measure of CellDMC to detect DMCTs. All results were based 100 Monte Carlo runs, with 100 'control' in silico mixtures and 100 'disease' in silico mixtures. Boxes indicate 25th and 75th centiles. Whiskers extend to the largest values no further than 1.5 times the interquartile range from these centiles.



Supplementary Figure 5

Validation of purified IC subtype datasets.

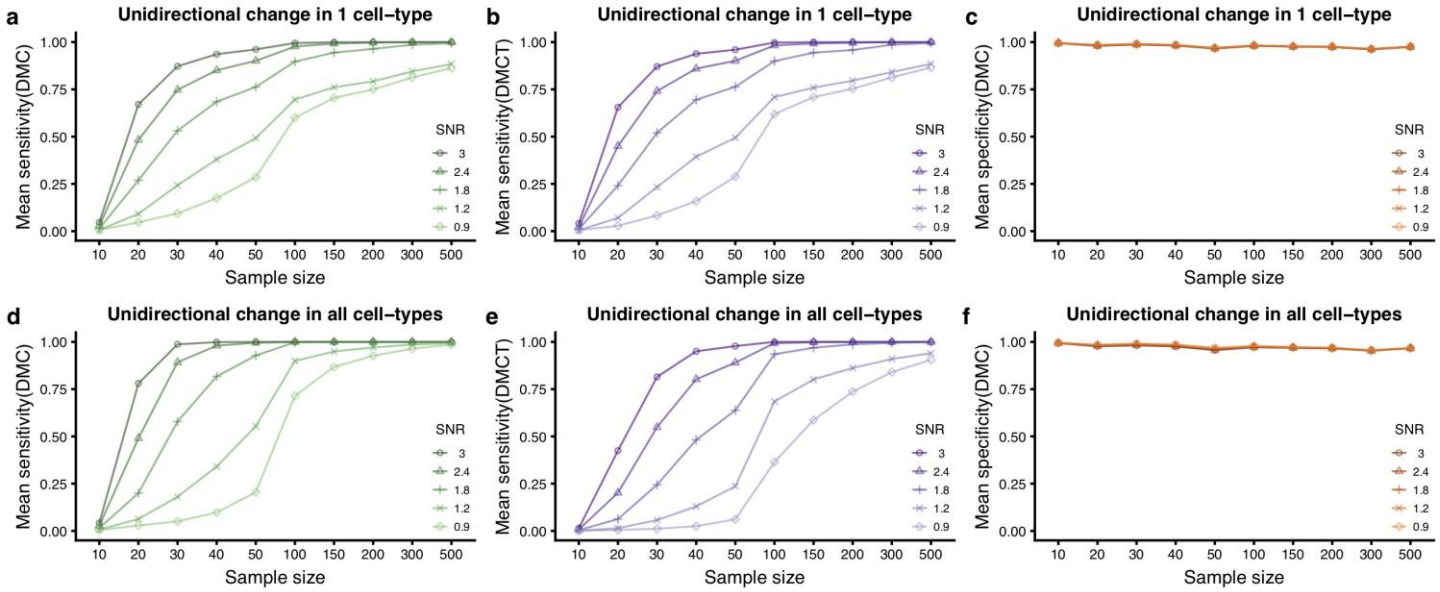
In this figure, we checked the purities of B cells, $CD4^+$ T cells, $CD8^+$ T cells, neutrophils and monocytes as profiled in independent studies. Cell-type fractions were estimated using EpiDISH (RPC mode) with our previously validated blood subtype DNAm reference, which consists of 333 cell-type-specific DMCs, to estimate fractions of six major blood subtypes (B cells, $CD4^+$ T cells, $CD8^+$ T cells, NK cells, monocytes and granulocytes/neutrophils). We display the estimated fractions for all purified blood datasets in **a)**, and for each cell subtype over the different source datasets in **b, c, d, e, f)**. Boxes indicate 25th and 75th centiles. Whiskers extend to the largest values no further than 1.5 times the inter-quartile range from these centiles.



Supplementary Figure 6

Performance comparison of CellDMC and the no-interaction model when only one IC subtype is altered including blood EWAS context.

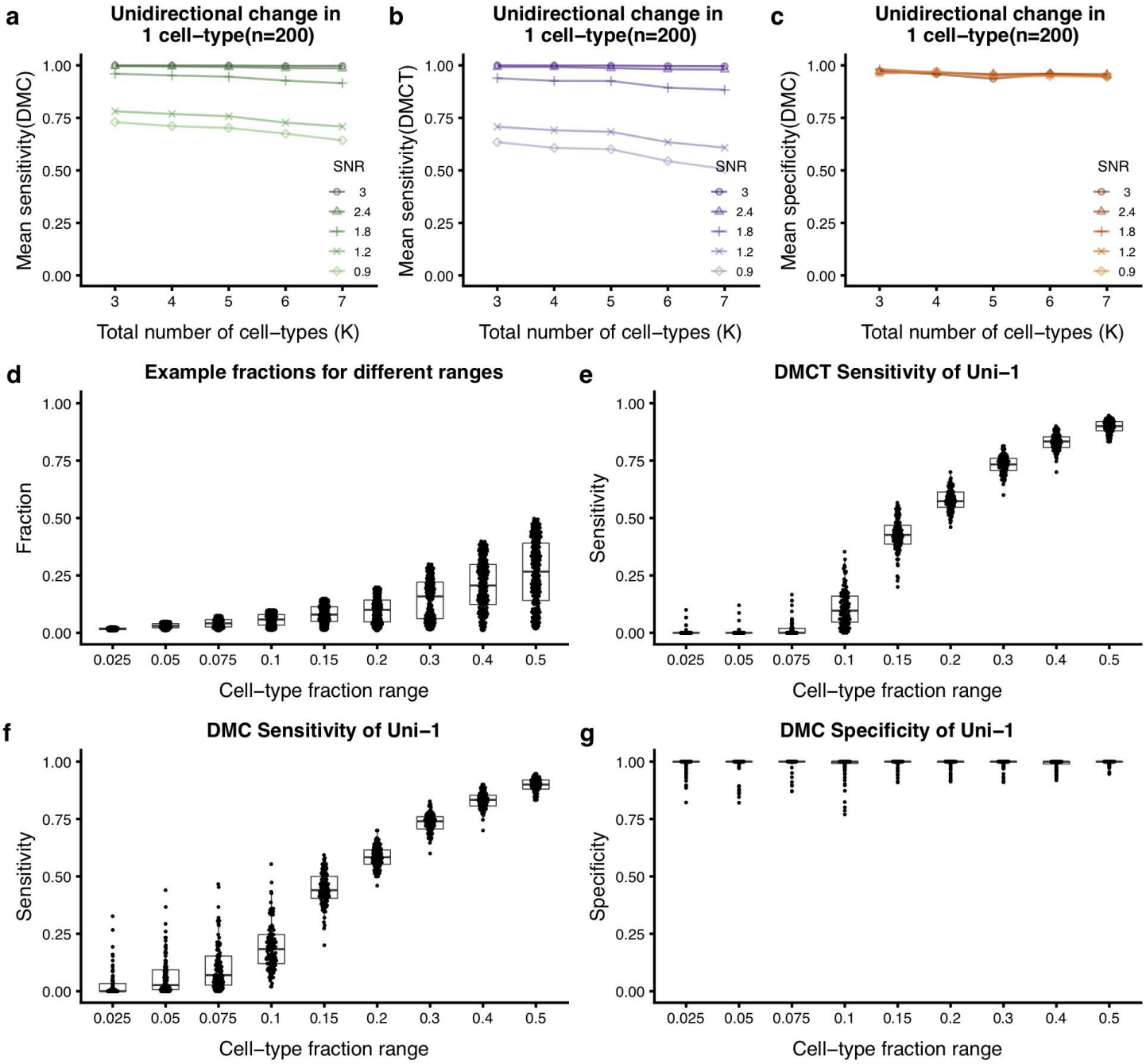
a-d) Similar to Supplementary Fig. 4, but now we used a two-step method (HEpiDISH) to estimate the fractions of individual IC subtypes. **a)** Scatterplot of an example where differential methylation happens only in B cells. HEpiDISH first estimates the fractions of total immune cells, epithelial cells, and fibroblasts. It then reapplies EpiDISH using the modified blood DNAm reference to estimate relative fractions of IC blood subtypes, which are then rescaled to obtain approximate absolute fractions for all cell types. Y-axis shows the cell-type fractions' adjusted beta value, and the x-axis gives estimated fractions of each cell type. **b, c)** Sensitivity and specificity measures of DMC calling for CellDMC and non-interaction (NoInt)-based model. **d)** Sensitivity measures of DMCT calling for CellDMC. All results were based on 100 Monte Carlo runs, with 100 'control' and 100 'disease' in silico mixtures. Boxes indicate 25th and 75th centiles. Whiskers extend to the largest values no further than 1.5 times the interquartile range from these centiles. **e-h)** As **a-d**), but now for the simulation scenario of a blood EWAS, where only blood cell subtypes are mixed together, with B cells the only altered cell type.



Supplementary Figure 7

Power analysis of CellDMC.

a, b) Sensitivity of CellDMC to detect **a**) DMCs and **b**) DMCTs as a function of total sample size (numbers of cases and controls are assumed equal and therefore one-half of this value) for five different SNR values in the scenario of a uni-directional change in one cell type. **c)** Corresponding specificity of CellDMC to detect DMCs/DMCTs. **d,e,f)** As in **a-c**), but now for the scenario of a uni-directional change affecting all cell types in the mixture. Each data point represents the mean over 100 Monte Carlo runs, where in each run 200 mixture samples were simulated: 100 'control' in silico mixtures and 100 'disease' in silico mixtures.

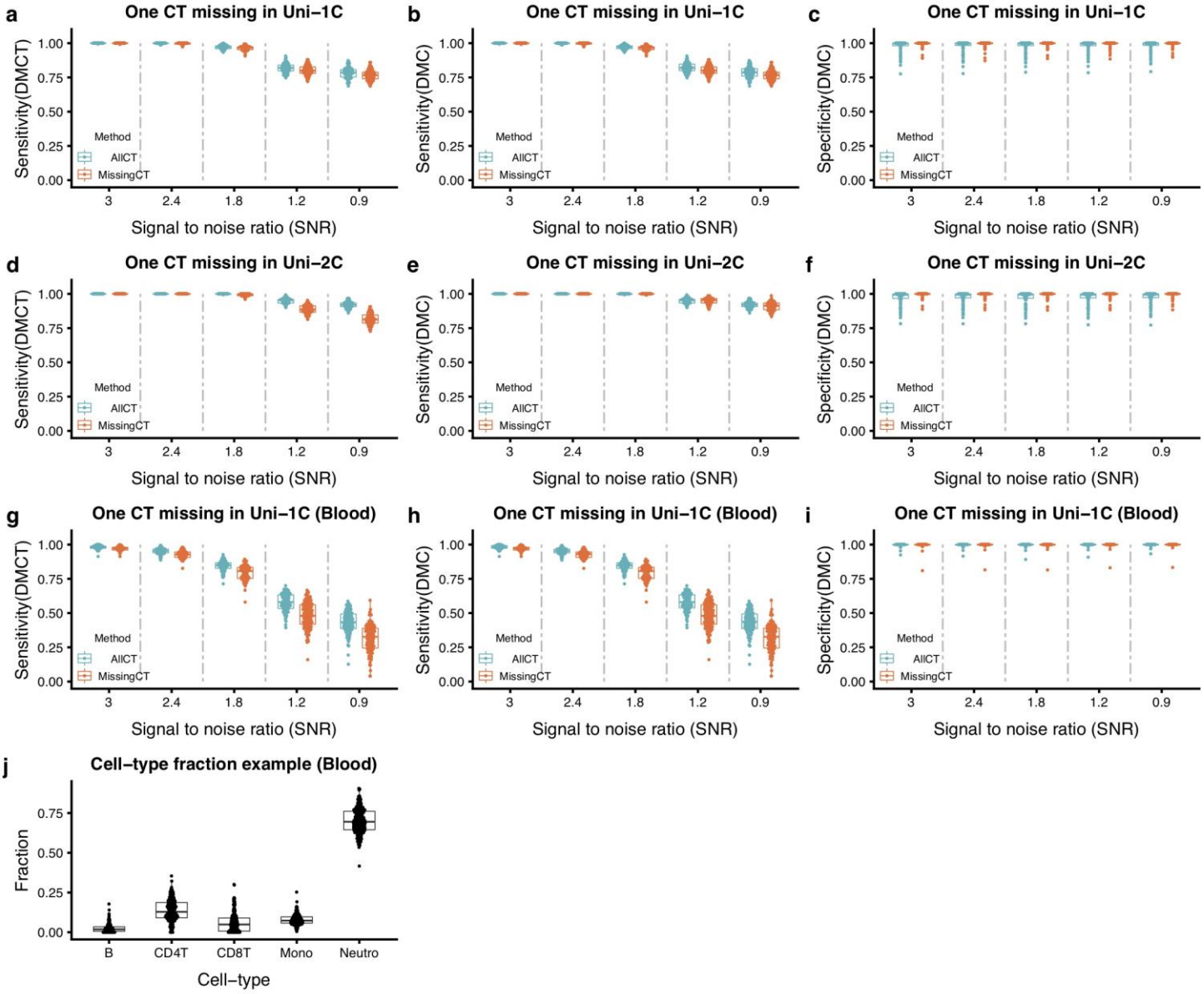


Supplementary Figure 8

Performance of CellDMC as a function of cell-type complexity and cell-type fraction range.

Assuming all cell types exhibit fractions of same underlying mean and variance, **a, b**) Sensitivity to detect DMCs and DMCTs as a function of the total number of cell types in the mixtures. Scenario considered is where only one cell type is differentially methylated. **c**) As **a**) but for the specificity. Each data point represents the mean over 100 Monte Carlo runs, where in each run 200 mixture samples were simulated: 100 'control' in silico mixtures and 100 'disease' in silico mixtures. In all cases, cell-type fractions were sampled from a Dirichlet distribution. For K=3, we mixed epithelial, fibroblast and monocytes and used EpiDISH with centEpiFibIC DNAm reference to estimate cell-type fractions. For larger K values, we iteratively added one more IC cell subtype to the mix in the order of neutrophils (K=4), $CD4^+$ T cells (K=5), B cells (K=6) and $CD8^+$ T cells (K=7). For K>3, we used HEpiDISH. For all K, in each Monte Carlo, there were 150 true DMCTs (75 hyper, and 75 hypo) occurring specifically in the monocytes. Considering the case of DMCs occurring only in

one cell type, **d)** box plots of cell-type fractions of the altered cell type (y-axis) against the cell-type fraction range. **e)** Box plots display sensitivity to detect DMCTs is larger than 50% as long as the range of variation in the cell-type fraction is 0.2 or higher. Even if the variation range in the cell-type fraction is only 0.15 (i.e. 15%), we still have a sensitivity close to 50%. **f-g)** The corresponding sensitivity and specificity to detect DMCs. All results were based on 100 Monte Carlo runs, with 100 'control' and 100 'disease' in silico mixtures. Boxes indicate 25th and 75th centiles. Whiskers extend to the largest values no further than 1.5 times the interquartile range from these centiles.

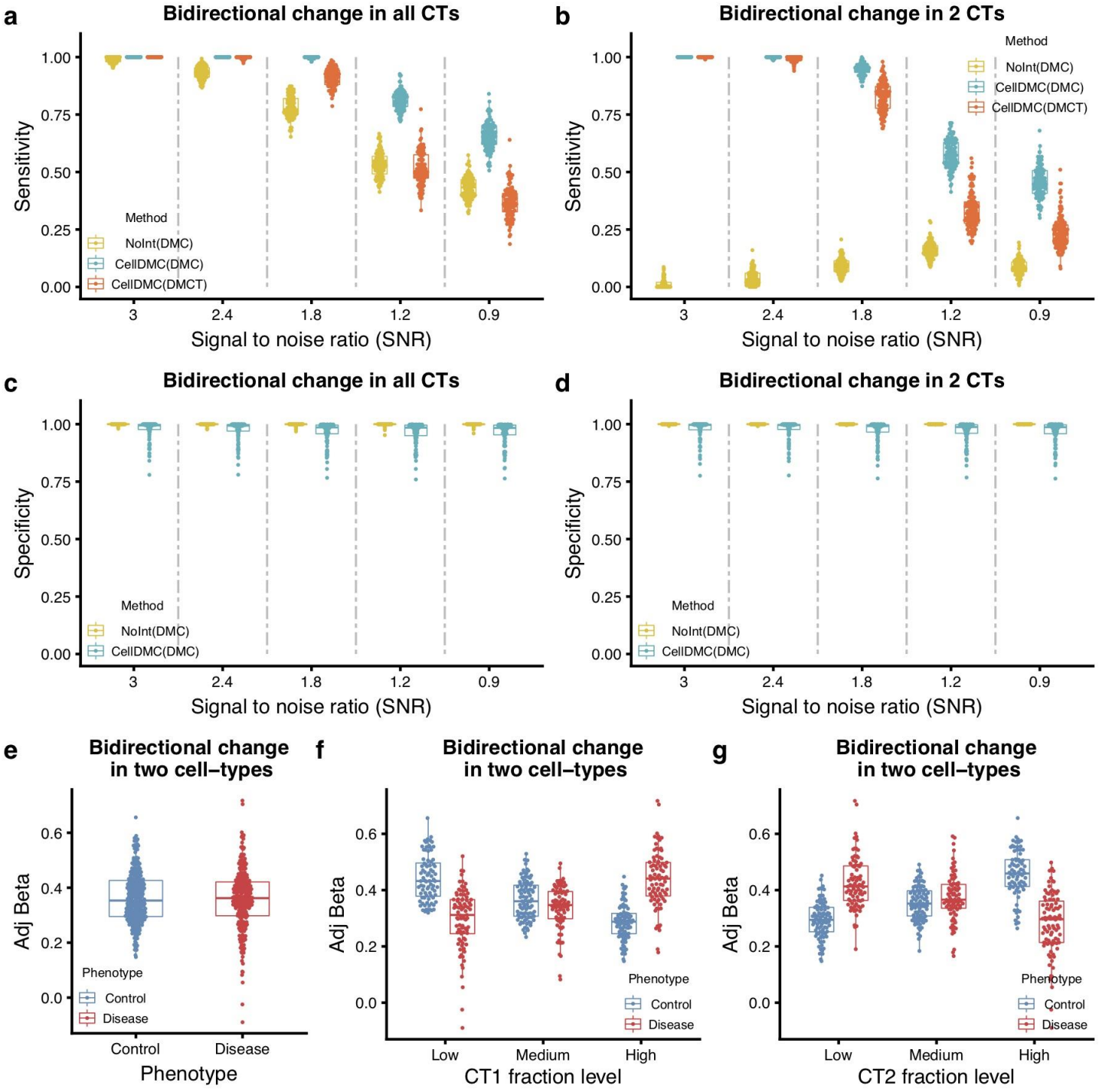


Supplementary Figure 9

Robustness of CellDMC to a missing cell type in DNAm reference matrix.

Assuming all cell types exhibit fractions of same underlying mean and variance, **a)** Comparison of the sensitivity to detect DMCTs as a function of SNR when all major cell types have representative DNAm profiles in the reference DNAm matrix (AllCT) versus the case when one major cell type is missing from the DNAm reference matrix (MissingCT). Scenario considered is where only one cell type out of three is differentially methylated (Uni-1C) and with the three cell-type fractions per sample drawn from a uniform Dirichlet distribution. **b**) As **a**) but for the sensitivity to detect DMCs. **c)** As **b**) but for the specificity. **d-f)** As **a-c**), but now for the scenario where the same DNAm alterations are happening in two cell types, with one of these missing from the DNAm reference matrix. **g-j**) Robustness of CellDMC to a missing cell type in the blood DNAm reference matrix. **g)** Sensitivity of CellDMC to detect DMCTs occurring in $CD4^+$ T cells at various SNRs and for the case where all depicted five blood cell subtypes are present in the DNAm reference matrix ("AllCT"-blue) versus the case where one cell type ($CD8^+$ T cells) was missing/removed from the DNAm reference matrix. **h-i)** As **g**), but for the sensitivity to detect DMCs (i.e. not caring which cell type the DMCT is called in) and specificity. In all cases, samples were in silico mixtures of the five underlying blood cell subtypes**. j)** Realistic distribution of cell-type fractions in blood from which the simulated fractions were modeled. Each box plot contains result from 100 Monte Carlo runs, where in each run 200 mixture samples were simulated: 100 'control' and 100 'disease' in silico mixture samples. Boxes indicate 25th and 75th centiles. Whiskers extend to the

largest values no further than 1.5 times the interquartile range from these centiles.

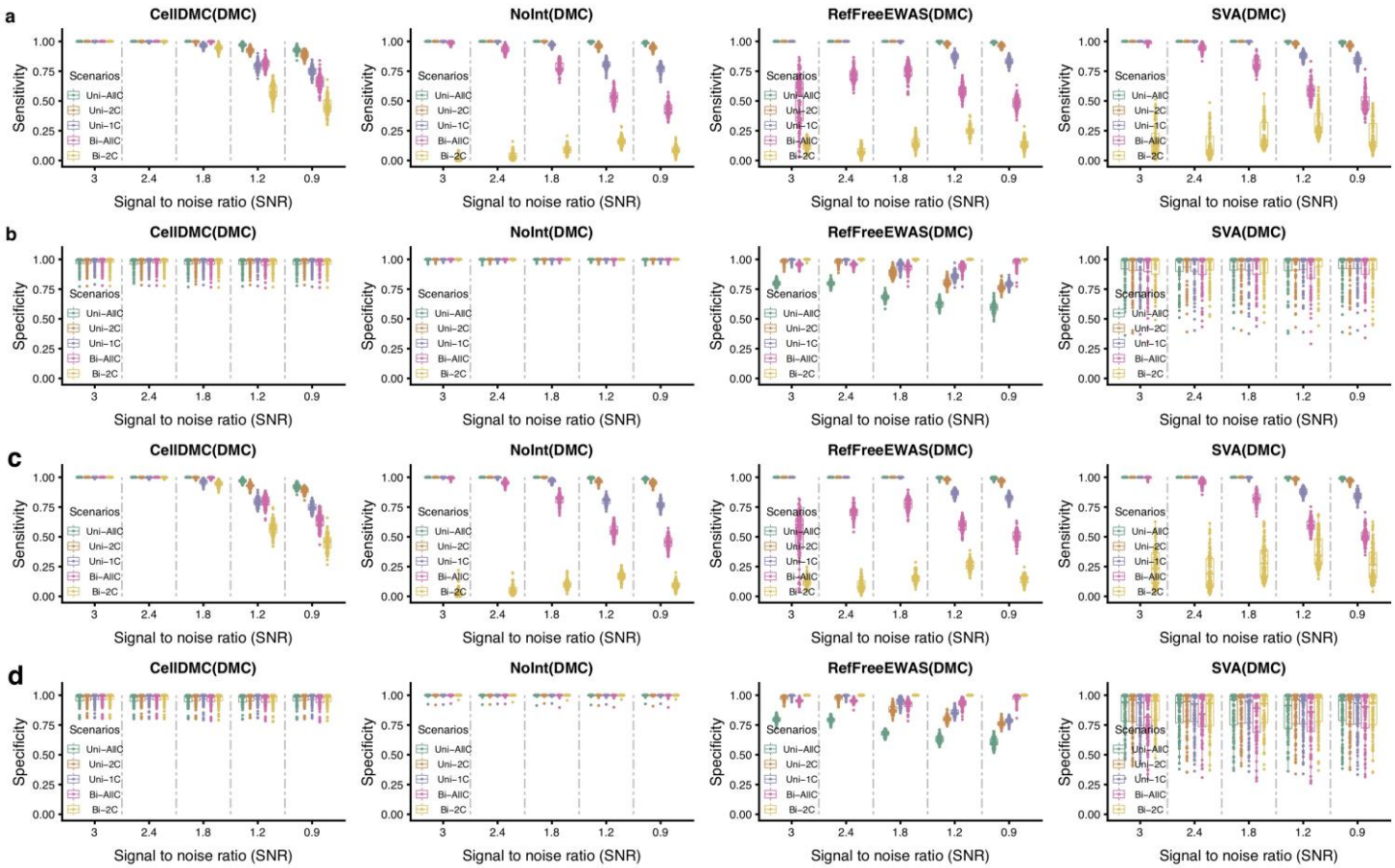


Supplementary Figure 10

Benchmarking of CellDMC to non-interaction-based DMC calling (NoInt) in bi-directional scenarios.

a-b) Sensitivity measures of the non-interaction-based model and CellDMC for detecting DMCs and DMCTs for the bi-directional DNAm change scenarios affecting all or only two cell-types, and for five signal-to-noise ratios (SNRs), as indicated. Note that for the bi-directional two-cell scenario depicted in **b**), a non-interaction-based analysis cannot detect DMCs because the opposite directional DNAm change in two of the underlying cell types (assumed here to be of equal magnitude) 'cancel out'. **c-d)** As **a-b)**, but now for the specificity. All results were based 100 Monte Carlo runs, with 100 'control' in silico mixtures and 100 'disease' in silico mixtures. **e)**

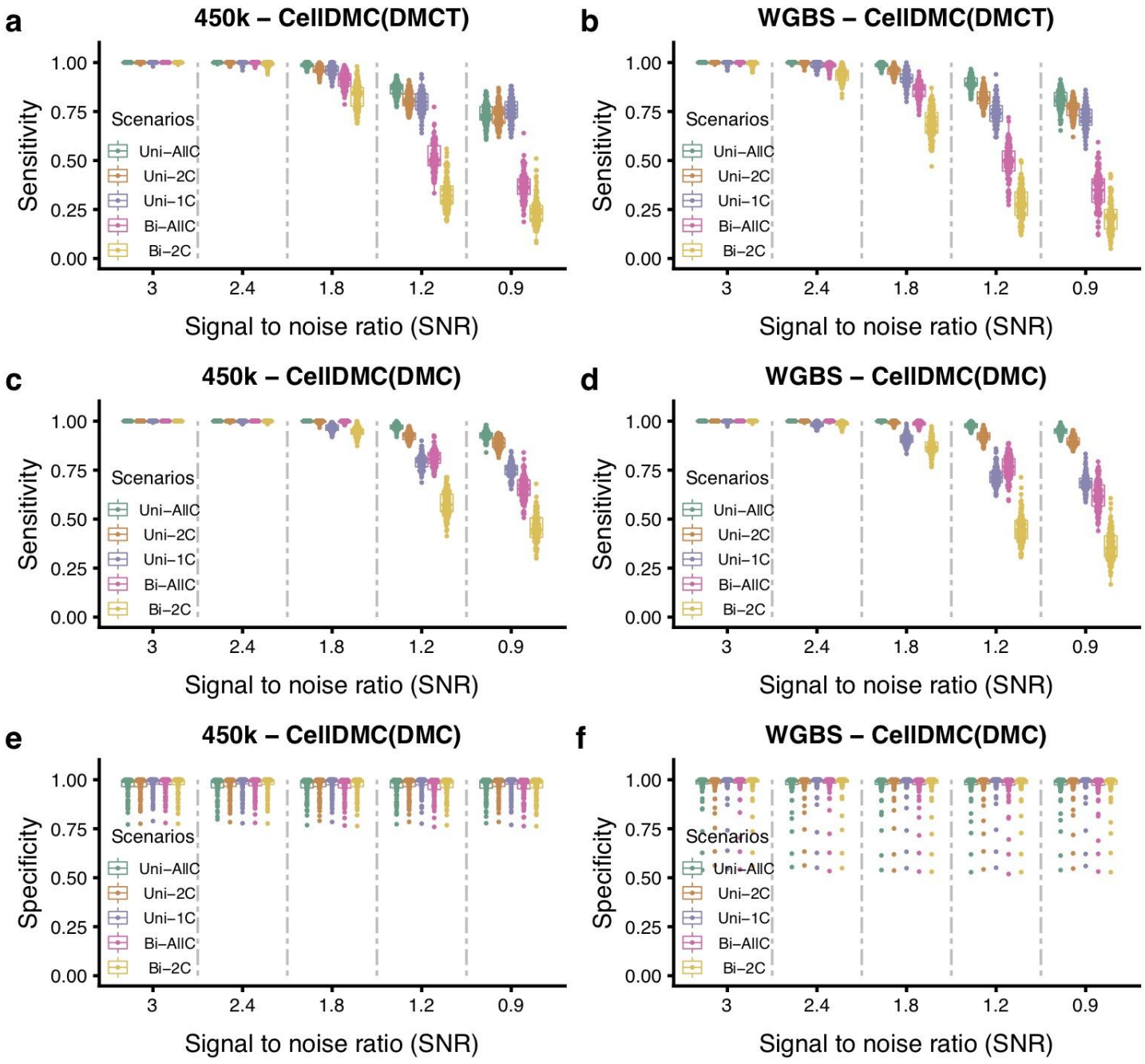
Illustrative example of a DMCT, which would be missed by a conventional DM analysis (the non-interaction-based model) because the bi-directional change in the two cell-types cancel out. **f-g)** Same CpG, but now with the adjusted beta-value plotted against cell-type fraction for the two underlying cell types, showing differences in DNAm between cases and controls, with the directionality switching depending on the cell-type fraction. Boxes indicate 25th and 75th centiles. Whiskers extend to the largest values no further than 1.5 times the interquartile range from these centiles.



Supplementary Figure 11

Benchmarking of CellDMC to NoInt model, RefFreeEWAS and SVA.

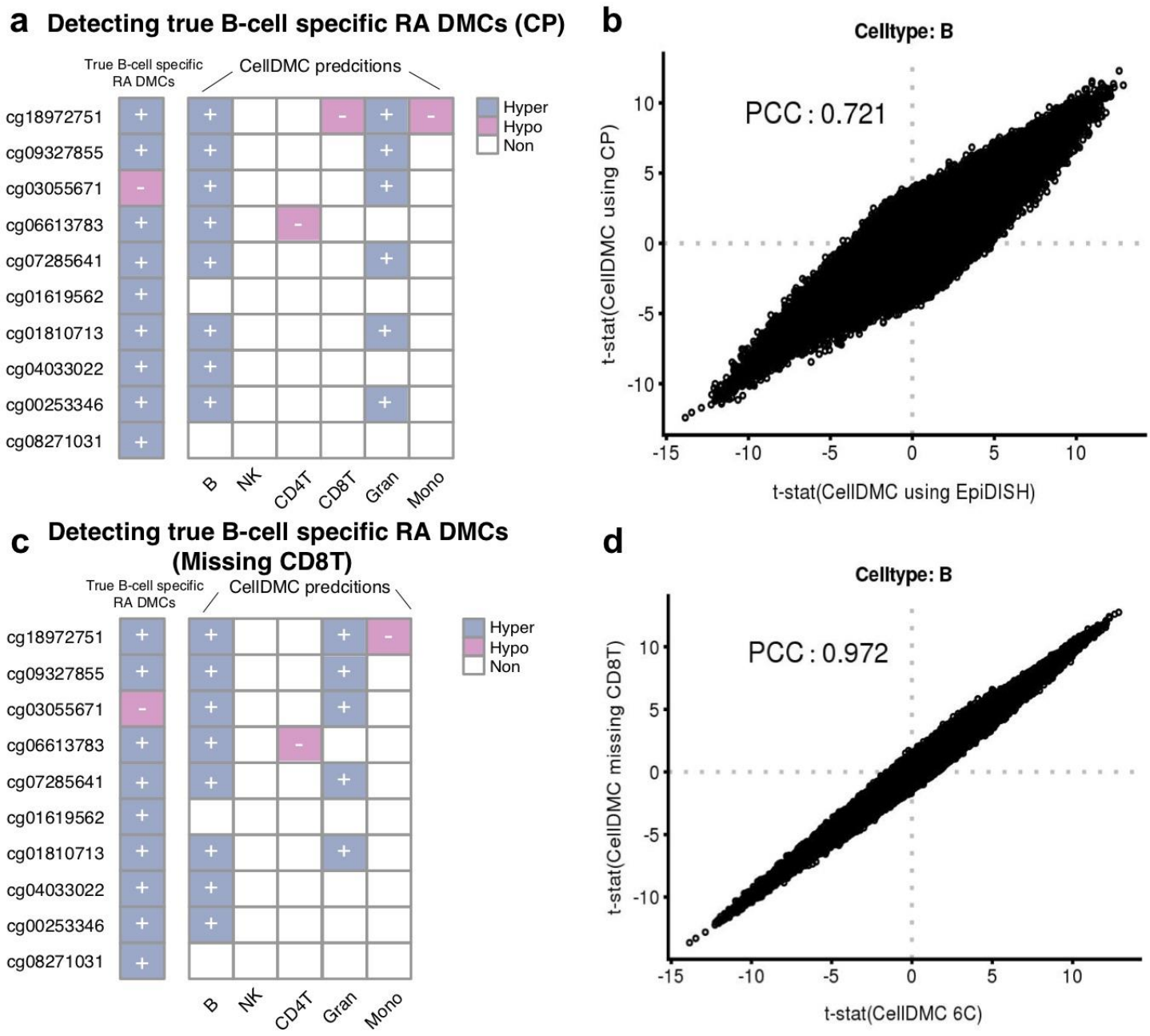
a) From left to right, plots of the sensitivity of CellDMC, of reference-based DMC calling (which includes cell-type fractions only as covariates without interaction terms, denoted “NoInt”), and of the reference-free methods RefFreeEWAS and SVA, against the SNR for five different differentially methylated cell-type scenarios where mean cell-type fractions are not different between cases and controls. Because the competing methods can only detect DMCs, and not DMCTs, we compare the methods in terms of their sensitivity to detect DMCs. Results are for 100 Monte Carlo runs, where in each run 200 in silico mixture samples were simulated (100 ‘control’ and 100 ‘disease’). **b)** As **a**), but now for the specificity. **c-d**) As **a-b**), but now there is a change in mean cell-type fractions between cases and controls. Boxes indicate 25th and 75th centiles. Whiskers extend to the largest values no further than 1.5 times the interquartile range from these centiles.



Supplementary Figure 12

Evaluation of CellDMC's performance on WGBS data and comparison to Illumina 450k.

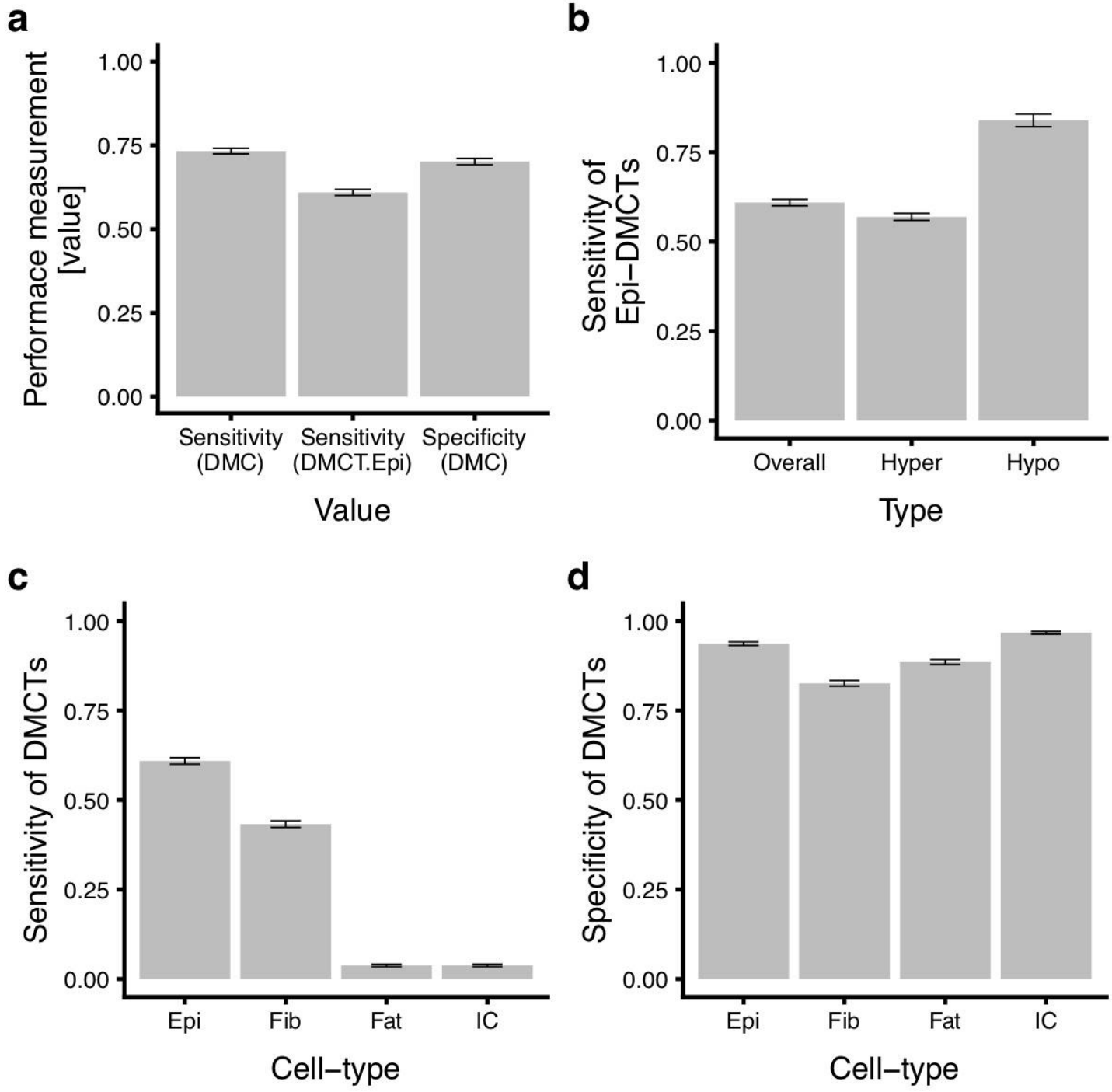
a,c,e) Sensitivity to detect DMCTs, DMCs and specificity for DMC/DMCT detection for CellDMC in five different simulation scenarios and five different SNRs, as indicated, on in silico mixtures of DNAm samples generated from Illumina 450k data. **b,d,f)** As **a,c,e**, but now displaying CellDMC's performance on in silico mixtures of DNAm profiles generated from WGBS data derived from IHEC. All results were based on 100 Monte Carlo runs, with 100 'control' and 100 'disease' in silico mixtures. Boxes indicate 25th and 75th centiles. Whiskers extend to the largest values no further than 1.5 times the interquartile range from these centiles.



Supplementary Figure 13

Robustness of CellDMC on the rheumatoid arthritis EWAS in whole blood.

a) Heat map shows the true B-cell-specific RA DMCs and predicted DMCTs by CellDMC (with cell-type fractions inferred with constrained projection-CP). **b)** Scatterplot of the interaction t-statistics for B cells as inferred using CellDMC with EpiDISH RPC (x-axis) vs. constrained projection (CP; y-axis). Statistic values larger than 10 or smaller than –10 are highly significant and are very robust to the choice of cell-type-fraction estimation algorithm. **c-d**) As **a-b**), but now cell-type fractions were estimated using EpiDISH RPC with $CD8^+$ T cells missing from the DNAm reference. Calling of DMCTs was based on a two-tailed *t*-test in CellDMC.

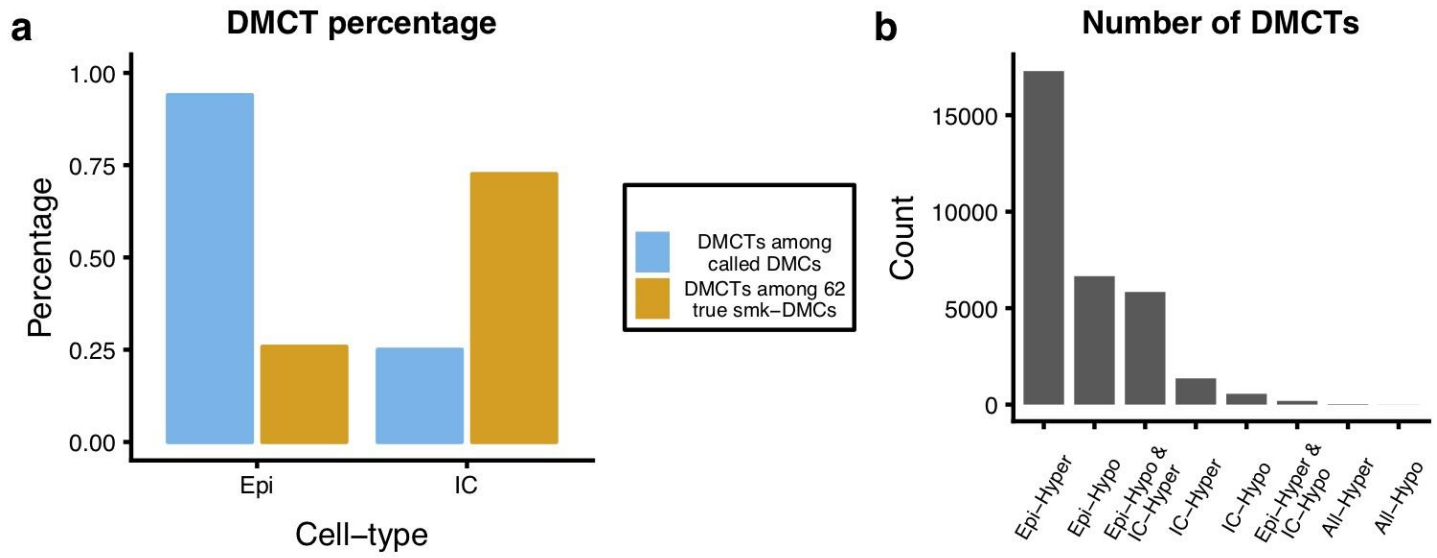


Supplementary Figure 14

Validation of CellDMC in detecting breast cancer epithelial DMCTs.

a) Bar plots display CellDMC's sensitivity to detect DMCs and Epi-DMCTs and specificity. **b)** Bar plot showing the percentage of breast cancer epithelial DMCTs, correctly predicted by CellDMC to occur in the epithelial compartment. There were a total of 19,379 true positive breast cancer epithelial DMCTs, of which 11,807 were correctly predicted to be a DMCT in epithelial cells. We also display the percentages for true hypermethylated and hypomethylated DMCTs, separately. **c)** Bar plot of percentage of breast cancer epithelial DMCs being predicted by CellDMC to be DMCTs in the epithelial fibroblast, fat, and immune cell compartments. **d)** Bar plot of

percentage of predicted nonDMCTs (nonDMCs) in breast cancer for a set of 15,212 true negatives, confirming that the great majority of these are not predicted to be epithelial DMCTs. We note that these CpGs are also not predicted to be DMCs in fibroblasts, fat or immune cells. All numbers on bar plots denote the number of CpGs. Error bars indicate 99% confidence interval, which was calculated as $p \pm 2.576 * \sqrt{[p * (1 – p)/n]}$.



Supplementary Figure 15

Identification of smoking DMCTs in buccal tissues

a) Bar plots comparing the percentage of DMCTs that occur in epithelial and IC compartments, measured relative to the set of 62 smoking DMCs in blood, or relative to all called DMCs/DMCTs. **b)** Bar plots displaying the number of different types of DMCT, as indicated. Epi, epithelial only; IC, immune cell only; Hyper, hypermethylated in smokers; Hypo, hypomethylated in smokers. “All” means a DMC that occurs in both cell types.

Scenario	Epi	Fib	IC	Number of DMCs
Uni-1 (Uni-1C)	Hyper			25
	Hypo			25
		Hyper		25
		Hypo		25
			Hyper	25
			Hypo	25
Total				150
Uni-2 (Uni-2C)	Hypo	Hypo		50
	Hyper	Hyper		50
		Hyper	Hyper	50
Total				150
Uni-3 (Uni-AllC)	Hyper	Hyper	Hyper	75
	Hypo	Hypo	Hypo	75
Total				150
Bi-2 (Bi-2C)	Hyper		Hypo	50
	Hypo	Hyper		50
		Hypo	Hyper	50
Total				150
Bi-3 (Bi-AllC)	Hyper	Hyper	Hypo	50
	Hypo	Hyper	Hypo	50
	Hypo	Hypo	Hyper	50
Total				150

Supplementary Table 1. Summary of the different in-silico mixture simulation scenarios. We considered a total of five different simulation scenarios. For the case where we restricted to K=3 cell-types (epithelial-Epi, Fibroblasts-Fib, and total immune cell (IC)), the scenarios are called Uni-1C (for unidirectional DNAm changes occurring in 1-cell type), Uni-2C (for unidirectional DNAm changes occurring in 2 cell-types), Uni-AllC (for unidirectional DNAm changes occurring in all 3 cell-types), Bi-2C (for DNAm changes occurring in 2 cell-types with opposite directionality) and Bi-AllC (for DNAm changes occurring in all 3 cell-types with varying directionality). The number of differentially methylated cytosines (DMCs) in each cell-type and their directionality in each scenario are indicated.

	CellDMC			NonAdj	EpiDISHAdj
Number of DMCs	16332			98582	40555
	All		Number of overlaps	5610	1375
Number of DMCTs	B	14929		4445	233
	NK	16		4	1
	CD4T	434		165	1
	CD8T	16		5	1
	Gran	1308		1104	1142
	Mono	7		0	0

Supplementary Table 2. Number of predicted DMCTs and DMCs using different methods. The table shows the number of DMCTs for each cell-type as predicted by CellDMC. And the number of DMCs predicted by CellDMC, the non-interaction based model (denoted EpiDISHAdj) and the non-adjusted method. The number of overlap between CellDMC predicted DMCTs and DMCs predicted by other two methods are also shown. These results are based on Bonferroni adjusted P-values as used in RA EWAS from Liu et al.