NAR Genomics and Bioinformatics, 2024, **6**, lqae146
https://doi.org/10.1093/nargab/lqae146
Advance access publication date: 4 November 2024
High Throughput Sequencing Methods

Tumor purity estimated from bulk DNA methylation can be used for adjusting beta values of individual samples to better reflect tumor biology

Iñaki Sasiain[1,†], Deborah F. Nacer[1,2,*,†], Mattias Aine[2], Srinivas Veerla[1,2] and Johan Staaf[1,2,*]

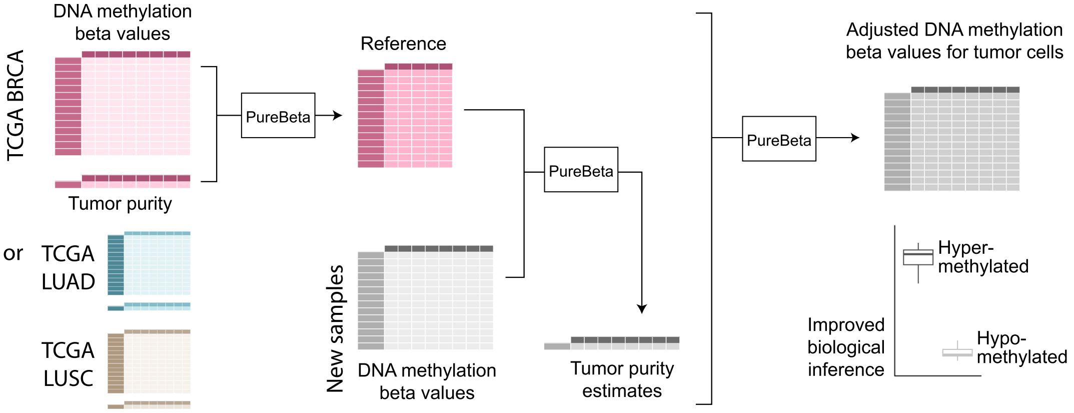
[1]Division of Translational Cancer Research, Department of Laboratory Medicine, Lund University, Lund 22381, Sweden
[2]Division of Oncology, Department of Clinical Sciences Lund, Lund University, Lund 22381, Sweden

[*]To whom correspondence should be addressed. Tel: +46 46 2221444; Fax: +46 46 2221444; Email: deborah.figueiredo_nacer_de_oliveira@med.lu.se
Correspondence may also be addressed to Johan Staaf. Email: johan.staaf@med.lu.se
[†]The first two authors should be regarded as Joint First Authors.

Abstract

Epigenetic deregulation through altered DNA methylation is a fundamental feature of tumorigenesis, but tumor data from bulk tissue samples contain different proportions of malignant and non-malignant cells that may confound the interpretation of DNA methylation values. The adjustment of DNA methylation data based on tumor purity has been proposed to render both genome-wide and gene-specific analyses more precise, but it requires sample purity estimates. Here we present PureBeta, a single-sample statistical framework that uses genome-wide DNA methylation data to first estimate sample purity and then adjust methylation values of individual CpGs to correct for sample impurity. Purity values estimated with the algorithm have high correlation (>0.8) to reference values obtained from DNA sequencing when applied to samples from breast carcinoma, lung adenocarcinoma, and lung squamous cell carcinoma. Methylation beta values adjusted based on purity estimates have a more binary distribution that better reflects theoretical methylation states, thus facilitating improved biological inference as shown for *BRCA1* in breast cancer. PureBeta is a versatile tool that can be used for different Illumina DNA methylation arrays and can be applied to individual samples of different cancer types to enhance biological interpretability of methylation data.

Graphical abstract



Introduction

Epigenetic modifications affect gene and genome regulation without altering DNA sequences and are considered a fundamental and enabling characteristic in tumor formation (1). Epigenetic changes including DNA methylation and histone modifications have been extensively studied in the context of cancer (e.g. (2–4)). Changes in DNA methylation, the most studied epigenetic mechanism to date, are directly involved in gene regulation and have been shown to be important for tumor development (5–9). Different methodologies have been developed to interrogate genomic DNA methylation patterns of cells in tumor samples (10). Among the most used methods are genome-wide Illumina Infinium DNA methylation arrays, the method of choice for methylation analysis of thousands of cancer samples publicly available as part of The Cancer Genome Atlas (TCGA) initiative. Such arrays survey individual CpG sites based on a well-proven protocol involving bisulfite conversion, hybridization, single-base extension, and

Received: January 30, 2024. **Revised:** October 1, 2024. **Editorial Decision:** October 16, 2024. **Accepted:** October 18, 2024

fluorescence scanning (11,12). CpGs are DNA dinucleotides (cytosine followed by guanine) that are overrepresented in gene promoters and that when methylated can be involved in gene silencing through e.g. interfering with the binding of proteins needed for transcription (13). Irrespective of the chosen methodology to study methylation patterns, a common challenge in tumor profiling is that malignant cells are intermixed with non-malignant cells (e.g. cells from the immune system) in the so-called tumor microenvironment. Malignant and non-malignant cells often differ in DNA methylation states, therefore data generated from bulk tumor tissue will be influenced by the composition of cell types present in the sample.

To address this compositional effect in DNA methylation studies, several approaches and methods have been proposed. These may encompass deconvolving data to provide a size estimate of the malignant and non-malignant compartments by cell type (e.g. (14–16)) or more generally estimating purity, usually on a sample level (e.g. (17–19)). Few methods have been reported that attempt to adjust individual CpGs for, e.g. estimated normal cell contamination. An example of the latter is the recent work by Staaf and Aine (20) demonstrating that adjustment of DNA methylation data at a single CpG level based on linear regression modelling can be performed in larger cohorts with tumor purity estimates available for each sample, improving both genome-wide and gene-specific analyses. To our knowledge, no other method has been proposed and is currently used within cancer research adjusting beta values for tumor purity on an individual CpG level. However, not all research data sets are comprised of hundreds of samples and accurate measurements of tumor purity often do not accompany publicly available DNA methylation data sets, rendering the proposed Staaf and Aine (20) methodology less widely applicable. Indeed, for purity information to be available, high-throughput DNA sequencing data of the same set of samples would typically be needed as it is currently the gold standard input for methods that calculate tumor purity, though alternative methods using e.g. DNA methylation as input exist (see (21) for a review).

Here, we present PureBeta, a framework that substantially extends and enhances the work by Staaf and Aine (20) by allowing purity estimation and individual CpG adjustment of samples from multiple cancer types using only the DNA methylation data itself as input. PureBeta builds on published information (e.g. from the TCGA consortium) and uses genome-wide DNA methylation data to first estimate sample purities and then adjust methylation values of individual CpGs with respect to tumor purity with a single sample approach. We show a strong concordance between purity estimates obtained from PureBeta and from sequencing-based technologies in different cancer types such as breast and lung cancer—the types with the highest incidence and mortality rate today respectively. Importantly, we demonstrate that when purity estimates calculated with this algorithm are used to adjust DNA methylation data, the distribution of beta values becomes more binary as would be expected biologically, thus facilitating the inference of unconfounded gene methylation states in tumor cells. PureBeta is a tool that can be used for both the Illumina HumanMethylation450 and MethylationEPIC arrays and can be applied to individual samples of different cancer types to enhance biological interpretability of methylation data with respect to somatic changes in DNA methylation states of malignant cells.

Materials and methods

Cohorts

Data sets were retrieved from the TCGA initiative following the workflow described in (20) for three malignancies: breast cancer (BRCA, $n = 630$), lung adenocarcinoma (LUAD, $n = 418$), and lung squamous cell carcinoma (LUSC, $n = 333$). These included clinical information and pre-processed DNA methylation beta values for 421 368 CpGs obtained with the Illumina Infinium HumanMethylation450 BeadChip array. Custom annotations for CpGs were derived by mapping their coordinates to other genomic information such as gene coordinates as described in (20). For brevity, relative to genes CpGs could be of three categories: in gene promoters, when within a 500-base pair (bp)-window upstream or downstream from the transcription start site of any gene; proximal to gene promoters, when within 5000 bp up- or downstream from the gene promoter window; or distal to gene promoters, when >5000 bp away from the gene promoter window. Expression of 60 483 transcripts for TCGA BRCA samples in fragments per kilobase million (FPKM) was also retrieved following (20). Tumor purity calculated with ABSOLUTE (22) from whole exome sequencing (WES) data was obtained from (23) and used as reference values. TCGA identifiers and other characteristics such as tumor purities estimated with different methods are available in Supplementary Table S1.

An additional cohort from the Sweden Cancerome Analysis Network – Breast (SCAN-B) initiative (24,25) (see original publications for the ethics statement) containing DNA methylation information for 82 triple-negative breast cancer (TNBC) samples reported by Glodzik et al. was used as a validation cohort (5). The triple-negative status is given to a breast cancer sample with negative status of both estrogen receptor (ER) and progesterone receptor, as well as a lack of amplification of the *HER2/ERBB2* (human epidermal receptor growth factor 2/Erb-B2 receptor tyrosine kinase 2) gene. DNA methylation data for this cohort was generated with the Illumina Infinium MethylationEPIC platform and retrieved from Gene Expression Omnibus (accession number GSE148748) as processed beta values. In addition to data from around 850 000 CpGs, SCAN-B TNBC samples had tumor purities calculated from whole genome sequencing (WGS) and had also been previously classified as *BRCA1*-hypermethylated ($n = 57$ tumors) based on pyrosequencing or *BRCA1*-null (gene inactivation by pathogenic germline or somatic variants, $n = 25$ tumors) as reported by Staaf et al. (26). A total of 29 CpGs available in the data sets were used to assess *BRCA1* methylation status based on their genomic position (within 1500 bp upstream and 500 bp downstream of the transcription start site of *BRCA1*). Notably, clustering beta values of these CpGs completely recapitulated *BRCA1* promoter hypermethylation status as assessed by pyrosequencing (5).

The PureBeta framework

Our approach to calculating tumor purities from bulk DNA methylation stems from the work of Staaf and Aine (20). Briefly, the authors observed that beta values of individual CpGs could be separated into populations that correlated differently with sample purities and that these patterns could be captured using simple linear regressions. They then devised a strategy to use these regressions to adjust methylation values to improve the interpretation of DNA methylation data from bulk samples. However, in the original publication, there was a need for precalculated purity estimates for each sample of interest, as well as for a higher number of samples to be analyzed at once for successful calculation of regressions. To solve these issues, PureBeta was developed using the concept delineated by Staaf and Aine (20) to create and/or use reference data to estimate tumor purity of individual samples or full cohorts and subsequently adjust methylation values based on the obtained estimates. For clarity, in this work we refer to purity or tumor purity as the fraction of malignant cells in a sample and to the complementary 1-purity as the fraction of non-malignant cells. PureBeta was written in R v4.3.0 and can be downloaded as an R package from GitHub at https://github.com/StaafLab/PureBeta and reference data are available in FigShare at https://doi.org/10.6084/m9.figshare.26272864. Pseudocode explaining the steps taken in the main functions of the package is available as Supplementary Methods. PureBeta can be summarized into three main modules: (i) creation of reference data from a cohort through the reference_regressions_generator() function, (ii) estimation of tumor purities for individual samples through the purity_estimation() function and (iii) subsequent adjustment of beta values per CpG per sample through the reference_based_beta_correction() function (Figure 1).

Creation of reference data

This is performed on a cohort level (e.g. TCGA BRCA) and it is required for calculating the reference linear regressions that will be used in the next modules. Detailed information on this module such as packages used and optimization criterion have been thoroughly described in the work by Staaf and Aine (20). DNA methylation beta values and accompanying tumor purities of samples from e.g. sequencing data like WES or WGS are needed as input. Briefly, for each CpG, input variables are contrasted, and flexible mixture modeling is used to divide samples from the reference cohort into one to three populations based on the different linear relationships between the variables, which are then summarized per population per CpG by linear regression (Figure 1A). These are referred to as reference regressions. Regressions can be calculated from any cohort given that the sample size is large enough to robustly detect populations. For this module, PureBeta generates an output file with properties such as the slope, intercept, degrees of freedom, and residual standard error of each regression, as well as the beta variance per CpG. PureBeta provides pre-calculated reference data for TCGA BRCA, TCGA LUAD and TCGA LUSC to be used by users when input information needed in this module of the algorithm is not available.

Purity estimation

This newly developed module is performed on a single sample level, i.e. one sample at a time is analyzed to have its purity estimated. DNA methylation beta values from a sample of interest go through a series of steps before a final purity estimate is made (Figure 1B-F). First, CpGs with reference regressions are filtered based on beta variance to retain only those with variance above 0.05. This is performed to remove low-varying, non-informative CpGs expected to have only one population with less clear links to tumor purity. Variance cutoff value was chosen taking into consideration all three TCGA cohorts using a 6-fold cross-validation scheme described at the end of this module (Supplementary Figure S1). Then, one remaining CpG at a time, a sample's beta value is assigned to one of the reference regressions calculated at cohort level in the previous module (Figure 1B). Not all CpGs with beta variance above the threshold contribute to the estimate: when a beta value can be assigned to more than one regression or when the assigned regression's slope is too small to be informative with regards to purity change, the information from that CpG is discarded. Slope cutoff was determined to be 0.2 after optimization using the 30 000 CpGs with highest beta variance in the TCGA BRCA set, corresponding to a variance of 0.05 and above, also using a 6-fold cross-validation scheme (Supplementary Figure S1). CpGs that are not discarded are used to calculate a 1-purity confidence interval for the sample being analyzed through bootstrapping performed individually per CpG (Figure 1C). To achieve this, first a 1-purity value is calculated based on the sample's original beta value for that CpG and the linear regression for the population it has been assigned to following a linear model with 1-purity as the independent variable (x), the original beta value (β_{orig}) as the dependent variable (y), and the intercept (a) and slope (b) for that population of that CpG as calculated in the previous module (Equation 1).

$$y = a + b \times x \;\rightarrow\; 1 - purity = \frac{\beta_{orig} - a}{b} \tag{1}$$

Next, data points (beta and 1-purity pairs) from the population that generated the linear regression are resampled with replacement to compose a new population of same size to which a new linear regression is fitted. Resampling and linear modeling are performed 500 times (default). A different resampling number can be chosen by the user but increasing it does not significantly impact purity estimation (Supplementary Figure S1). For each resampling (i), a beta value is predicted (β_{pred}) using the 1-purity estimate obtained in Equation 1 and an added randomly sampled residual (ε) to account for the intrinsic variability of the dependent variable according to the regression parameters (Equation 2).

$$\beta_{pred_i} = a_i + b_i \times (1 - purity) + \varepsilon_i \tag{2}$$

The described bootstrapping method thus generates a distribution of β_{pred} values that is used to obtain an interval of the most common beta values. This interval can be made narrower or wider by changing the confidence level alpha (α), set to a default value of 0.7 after a 6-fold cross-validation scheme similar to the one performed for the slope (Supplementary Figure S1). At 0.7, the interval would include 70% of the β_{pred} values ranging from the 15th to the 85th percentiles considering their distribution. Finally, a 1-purity interval is calculated using the predicted beta value interval from Equation 2 and the original regression's intercept and slope from Equation 1 (Equation 3).

$$(1 - purity_{min}, 1 - purity_{max}) = \frac{\left(\beta_{\frac{1-\alpha}{2}}, \beta_{1-\frac{1-\alpha}{2}}\right) - a}{b} \tag{3}$$

After all CpGs have been processed, 1-purity intervals are combined determining a 1-purity coverage, i.e. the number of CpGs and estimated intervals that contain a given value in the 1-purity scale (Figure 1D). Coverages showed a systematic overestimation of low purity sections seen as a secondary peak at 1-purity intervals close to 0.8 that is corrected by fitting a new regression to the raw coverage values and using the residuals as the new corrected coverage, an approach that does not distort the original meaning of the data (Figure 1E). As the final step, the corrected coverage is smoothed with a spline regression (27) and its maximum value is recorded as

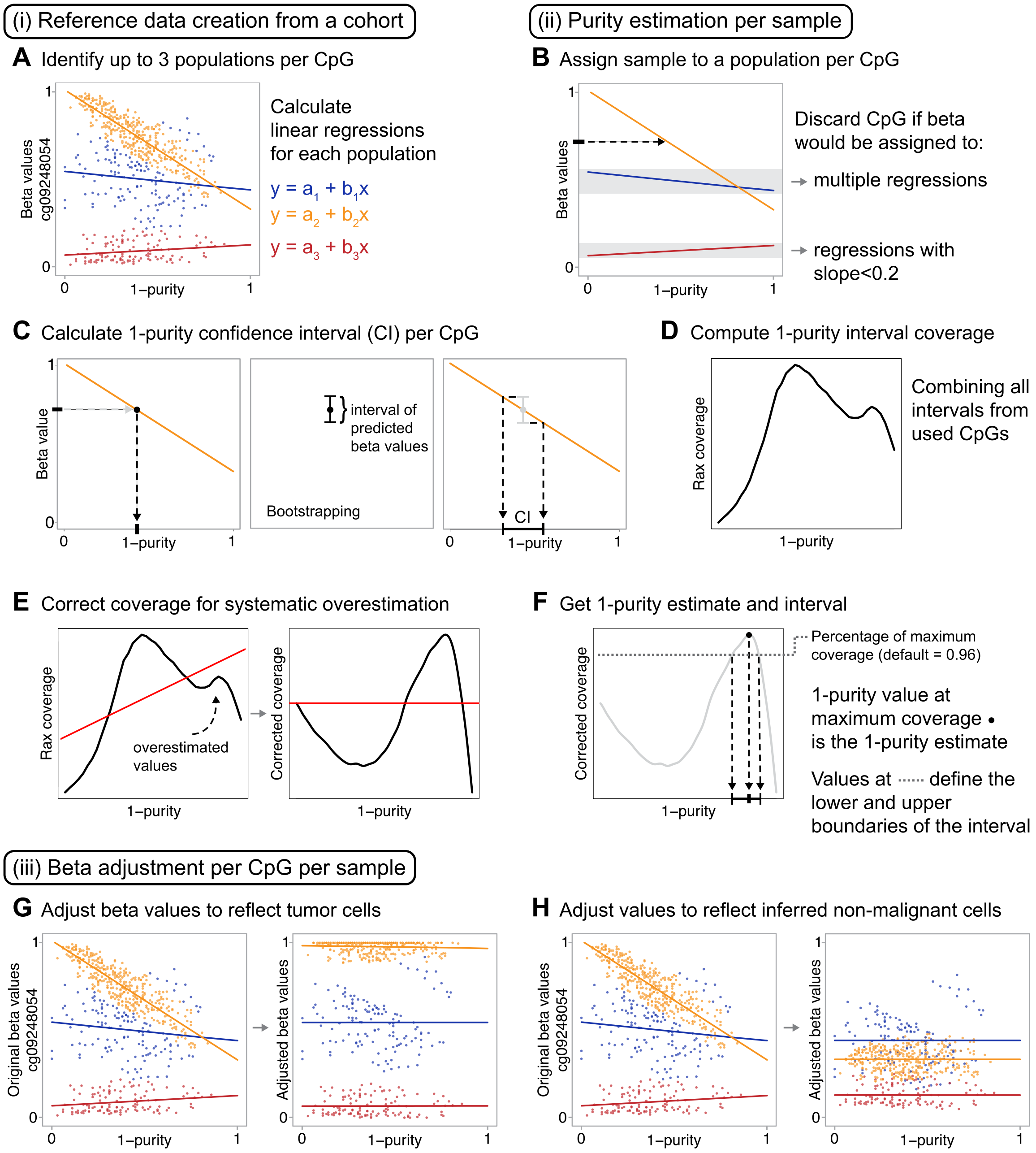

(i) Reference data creation from a cohort
A Identify up to 3 populations per CpG
Beta values cg09248054
1
0
0
1−purity
1
Calculate linear regressions for each population
y = a1 + b1x
y = a2 + b2x
y = a3 + b3x
(ii) Purity estimation per sample
B Assign sample to a population per CpG
Beta values
1
0
0
1−purity
1
Discard CpG if beta would be assigned to:
→ multiple regressions
→ regressions with slope<0.2
C Calculate 1-purity confidence interval (CI) per CpG
Beta value
1
0
0
1−purity
1
interval of predicted beta values
Bootstrapping
CI
0
1−purity
1
D Compute 1-purity interval coverage
Rax coverage
1−purity
Combining all intervals from used CpGs
E Correct coverage for systematic overestimation
Rax coverage
overestimated values
1−purity
Corrected coverage
1−purity
F Get 1-purity estimate and interval
Corrected coverage
1−purity
Percentage of maximum coverage (default = 0.96)
1-purity value at maximum coverage • is the 1-purity estimate
Values at ······ define the lower and upper boundaries of the interval
(iii) Beta adjustment per CpG per sample
G Adjust beta values to reflect tumor cells
Original beta values cg09248054
1
0
0
1−purity
1
Adjusted beta values
1
0
0
1−purity
1
H Adjust values to reflect inferred non-malignant cells
Original beta values cg09248054
1
0
0
1−purity
1
Adjusted beta values
1
0
0
1−purity
1

Figure 1. PureBeta's framework. Overview of the algorithm's three main modules (i–iii). (**A**) As the first module, samples are divided into up to three populations for each CpG and linear regressions are calculated between beta values and known sample purities as exemplified by one CpG (cg09248054) and the TCGA BRCA cohort. (**B**) To estimate the purity of a new sample in the second module of PureBeta, the sample is first assigned to a population based on its beta value. This is performed per CpG. If the beta value would be assigned to multiple regressions or to regressions with small slopes, the CpG is discarded from the estimation calculation. (**C**) A purity interval is calculated based on the original beta values and the assigned linear regression for each kept CpG using a bootstrapping approach. (**D**) All purity intervals are combined into a purity coverage. (**E**) The purity coverage is corrected for a detected systematic overrepresentation of low purity values. (**F**) The maximum coverage value is assigned as the sample's purity estimate together with a purity interval. (G, H) As the final module of the algorithm, original beta values are adjusted per CpG to reflect values of samples comprised of only tumor (**G**) or only normal (**H**) cells according to the calculated sample purities and linear regressions. Beta values shown in these panels are from the same CpG and samples in panel A. See Methods for more details on each step.

the 1-purity estimate for that sample (Figure 1F). Estimates are accompanied by a 1-purity interval to incorporate a degree of uncertainty that includes all 1-purity values above a user-specified percentage of the maximum coverage in the algorithm (Figure 1F). The default for this parameter is 0.96 returning all 1-purity values that have a higher coverage than 96% of the maximum obtained for that sample. Reducing this parameter will generate wider intervals. The 1-purity estimate and interval are the final output of this module of PureBeta. For optimization of parameters, estimates from PureBeta were compared to the reference values from WES looking to minimize the absolute difference between them. Optimization for variance, slope, and alpha parameters was performed using a 6-fold cross-validation scheme as mentioned above. For this, the entire TCGA cohort for one cancer type was randomly divided into six non-overlapping subsets of similar but not equal sizes and one subset at a time was used for purity estimation based on regressions calculated from the samples in the other five subsets combined (Supplementary Figure S1).

Adjustment of beta values

After acquiring tumor purity estimates for all samples in the cohort of interest, reference regressions calculated in the first module and purity estimates produced in the second module are used to adjust beta values of individual CpGs as described in the work from Staaf and Aine (20). Briefly, for each CpG the calculated regressions and residuals per sample are used to obtain the adjusted beta values considering purity values equal to 0 or to 1. This generates new beta values better reflecting pure tumor cell methylation (Figure 1G) when the 1-purity estimate is used as the independent variable in the linear regressions. The framework also generates new beta values for inferred non-malignant background methylation (Figure 1H) when linear regressions are calculated using purity values and these values are used as independent variable instead in the adjustment. The data resulting from this module is further investigated here using the hypermethylation status of the *BRCA1* gene as a case example in the SCAN-B TNBC cohort. This module can be performed on a single sample level using the provided set of reference regressions directly or for multiple samples from an entire new cohort of interest. If applied to a novel cohort, a refitting approach where regressions are recalculated from pooling together reference samples and novel samples first and then the refitted regressions are used for beta adjustment might be recommended. See the last section of the Results for more information on the refitting approach.

Statistical and computational analyses

All statistical analyses were performed with R v4.2.0. All *P*-values reported are two-sided and compared to a level of significance of 0.05. Correlations were calculated with the Pearson method using the *cor.test()* function. Correlation between DNA methylation and gene expression was calculated using beta values and FPKM from the full TCGA BRCA cohort. Two other estimates of tumor purity were calculated from DNA methylation data for test samples of the three TCGA cohorts. The first estimate was calculated with the InfiniumPurify (28) R package v1.3.1 and the *getPurity()* function without normal data and with cancer type set as BRCA, LUAD, or LUSC as appropriate. The second estimate was calculated through cell fraction imputation using the online version of CIBERSORTx (29) (quantile normalization disabled, 1000 permutations) with appropriate signatures extracted from the MethylCIBERSORT (15) R package v0.2.0 (breast_v2, lung_NSCLC_adenocarcinoma_v2, lung_NSCLC_squamous_cell_carcinoma_v2). Purity values from PureBeta were further compared to previously published estimates calculated with different methods (30). The gene set overrepresentation (GSO) analysis was done with the clusterProfiler (31) R package v4.4.4 and the *enricher()* function using the hallmark gene sets from the msigdbr (32) R package v7.5.1. Genes of interest were compared to a universe of 26 809 genes that contained any CpG available to this study located in their promoter regions based on transcription start sites as mentioned before.

Results

PureBeta estimates from DNA methylation data are concordant with those from WES in breast and lung cancer

PureBeta can be divided into three main modules: (i) creation of reference data through calculation of reference regressions from a cohort, (ii) estimation of tumor purity for each sample and (iii) subsequent adjustment of beta values given the calculated purities (Figure 1). Each module can be run independently if the required input data are available. To test the framework, the TCGA BRCA cohort was divided into a random 80–20 split, i.e. 80% of the samples ($n = 504$) were used for calculating the regressions per CpG that were then used to estimate the tumor purity of the remaining 20% ($n = 126$). After optimization of parameters (Supplementary Figure S1), tumor purities estimated with PureBeta showed good agreement (Pearson correlation $r = 0.84$) with the standard reference purity values calculated from WES data (Figure 2A). In addition, accompanying 1-purity intervals were generally narrow (mean: 0.052, min–max: 0.023–0.090) and included the WES value for almost half of the samples in the validation set ($n = 58$). As an error metric, we calculated the absolute difference between WES and PureBeta purities showing that there were no systematic miscalculations connected to lower or higher estimated tumor purity values (Figure 2B). Two samples stood out by having very large errors in their estimates, both purportedly composed entirely of tumor cells in the validation set (purity = 1 as estimated by TCGA). However, when compared to independent *in silico* tools used for calculating tumor purities (InfiniumPurify and MethylCIBERSORT) developed specifically for the Illumina arrays, PureBeta's performance showed good agreement (Pearson correlation $r > 0.85$, Supplementary Figure S2), making the WES estimate used as reference a comparative outlier. Indeed, upon further investigation of copy number profiles of the two outlier samples, they seemed to contain a substantial fraction of normal cells, rendering the TCGA estimate inaccurate (Supplementary Figure S2).

To test PureBeta's suitability on other types of cancer, we applied our framework to the two main histological types of lung cancer, lung adenocarcinoma (using the TCGA LUAD cohort) and lung squamous cell carcinoma (using the TCGA LUSC cohort). Analyses were performed in the same way as for TCGA BRCA: 80% of samples were randomly selected to generate the regressions (LUAD $n = 334$, LUSC $n = 266$) and the remaining 20% had their tumor purities estimated (LUAD $n = 84$, LUSC $n = 67$). Again, PureBeta showed good performance and the Pearson correlation between the estimates and the standard reference values obtained from WES was 0.90 and 0.83 for LUAD and LUSC, respectively (Figure 2A). However, data points were more widely spread around a 1:1 relationship than in the TCGA BRCA cohort as also reflected by an increase in mean distance between estimated purities and reference values (Figure 2B). Similarly, purity intervals were slightly wider in the lung cancer cohorts than in the breast cohort despite being calculated with the same optimized, default parameters, suggesting that larger cohorts could correlate with more precise estimates (Figure 2C). Additional comparisons were made between the estimates obtained with PureBeta and purity values obtained with other programs for the three TCGA cohorts evidencing our ap-

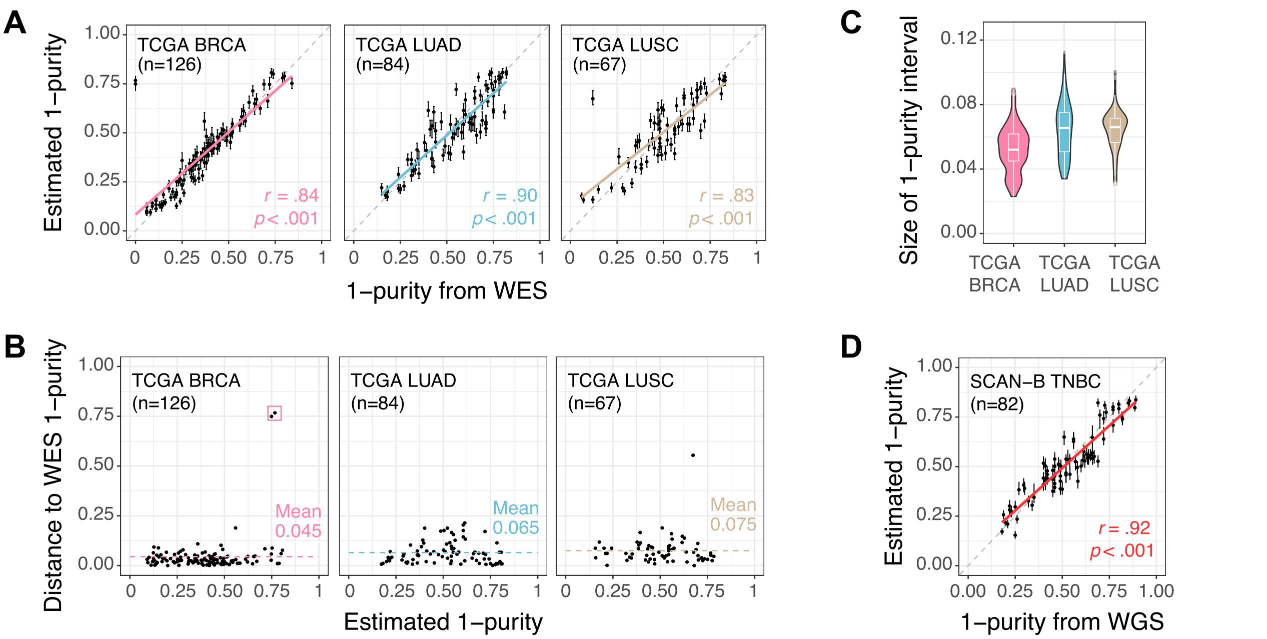


Figure 2. PureBeta's performance in the different cohorts. (**A**) Correlation between 1-purity as estimated with PureBeta and reference values calculated from WES on 20% of samples after using the other 80% to calculate regressions per TCGA cohort. Vertical bars correspond to the 1-purity interval for a sample. Dashed line corresponds to a 1:1 relationship. (**B**) Error between 1-purity as estimated with PureBeta and from WES calculated as absolute distance between estimates per sample for the three TCGA cohorts. Dashed line corresponds to the mean distance. (**C**) Violin plot of purity interval size as obtained from PureBeta per TCGA cohort. Violin width reflects cohort size. (**D**) Correlation between 1-purity as estimated with PureBeta and reference values calculated from WGS in SCAN-B TNBC. Vertical bars correspond to the 1-purity interval for a sample. Dashed line corresponds to a 1:1 relationship.

proach is comparable to that of other more established software (Supplementary Figure S2).

After establishing the framework and successfully applying it to three cancer types, reference regressions for hundreds of thousands of CpGs were calculated using all samples in the TCGA cohorts separately. To support independent analyses in breast and lung cancer, these reference regressions are available for public use with the PureBeta software. Importantly, these reference regressions allow the user to apply PureBeta down to a single sample of interest. Currently, the Illumina HumanMethylation450 platform has been superseded by the Illumina MethylationEPIC, interrogating >900 000 CpGs in its second version. To test PureBeta's applicability to MethylationEPIC data, we proceeded to estimate purities of the independent SCAN-B TNBC cohort ($n = 82$, with reference purity values determined from WGS) using reference regressions calculated from all 630 TCGA BRCA samples. Notably, we obtained a correlation of 0.92 between the estimates from PureBeta and the values calculated from WGS (Figure 2D). This shows that reference regressions based on data deposited in TCGA from the widely used but now discontinued HumanMethylation450 array can be used effectively to estimate tumor purity for samples profiled with the more current MethylationEPIC array given its backwards compatibility and both methylation arrays having a substantial number of overlapping probes.

CpG selection by PureBeta per cohort and per sample are not constant

In theory, PureBeta can utilize the entire DNA methylation data set but whether a given CpG contributes to the purity estimate for a specific sample is decided individually based on a set of requirements (Figure 1). Thus, CpG usage can vary between samples and likely also between cancer types. To investigate this in more detail we first studied the TCGA BRCA cohort. In this cohort, around 38 500 CpGs with regressions remained after variance filtering corresponding to 9.1% of the total number of CpGs. Of these CpGs, ∼23 000 CpGs were used per sample on average to make an estimate, but numbers varied from ∼17 000 to ∼28 000 CpGs. Samples with lower or higher tumor purities tended to need less CpGs for an estimate to be made than samples with intermediate purity values (Figure 3A). However, there was no relationship between the number of CpGs used and the distance metric calculated between the PureBeta estimate and the reference purity (Figure 3B).

Conversely, from the CpG perspective, the remaining ∼38 500 CpGs varied from not being used by PureBeta for any sample ($n = 270$ CpGs, 0.7%) to being used in any number up to all 126 samples when estimating purities (Figure 3C). Several CpGs have been previously associated with specific leukocytes and their beta values can be used to estimate purity (30) or to deconvolute DNA methylation data to estimate quantities of different immune cells in a sample (15,33). A comparison between these CpGs and CpGs used by our pipeline revealed that PureBeta's ability to estimate tumor purity is not directly associated to beta values of immune cells as very few CpGs used by PureBeta are among leukocyte-specific CpGs (Supplementary Figure S3). The CpG usage pattern allowed us to also investigate deeper the 3266 CpGs that were used in purity calculations of >97% of TCGA BRCA samples (Supplementary Table S2). When it comes to genomic distribution, no chromosomes seemed to be particularly favored by PureBeta after the variance filtering nor in CpGs in the top 97% that couldn't be related to different

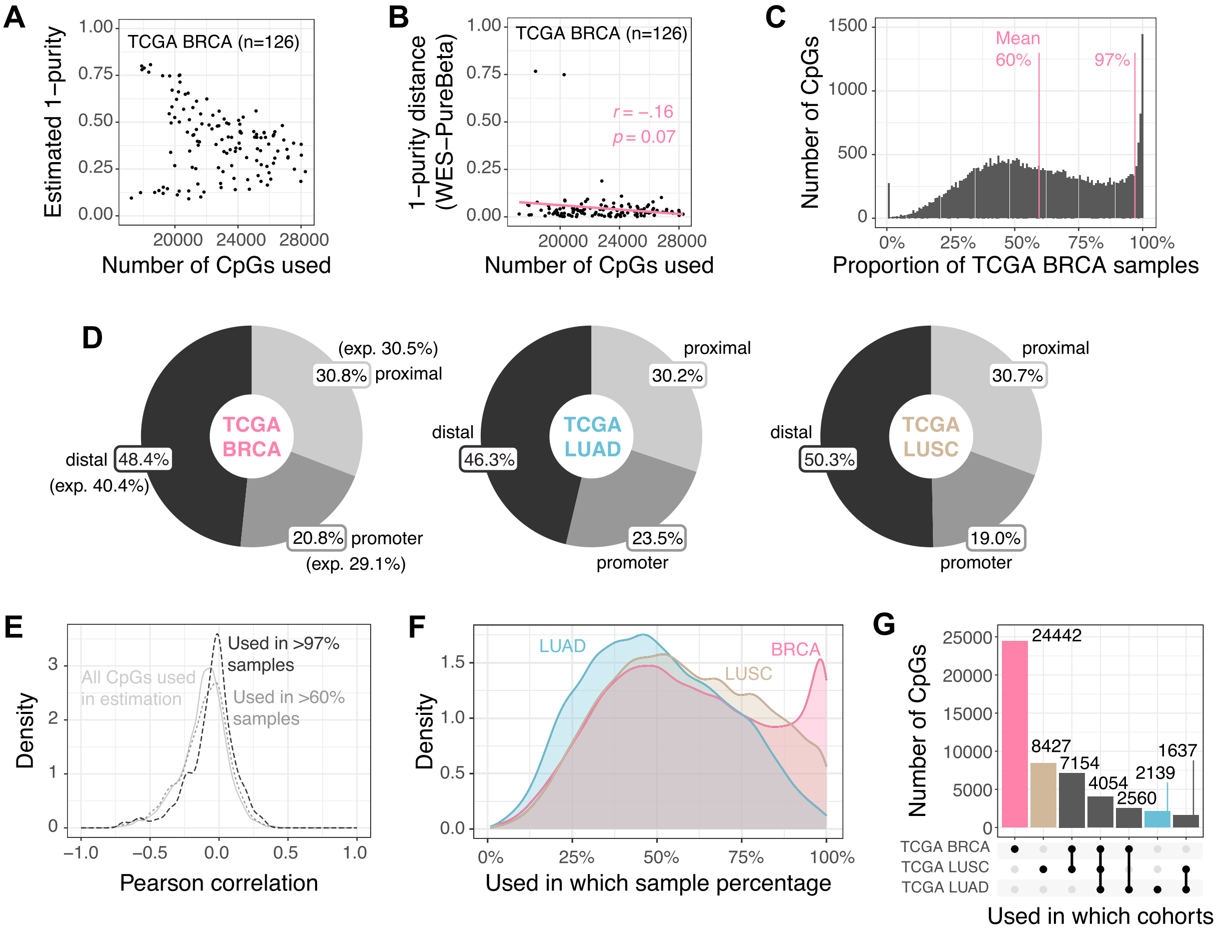


Figure 3. CpG usage by PureBeta in the TCGA cohorts. (**A**) 1-purity estimated with PureBeta in TCGA BRCA compared to the number of CpGs used for making the estimate. (**B**) Absolute distance between estimated and reference 1-purities in TCGA BRCA compared to the number of CpGs used during estimation. (**C**) Percentage of TCGA BRCA samples that had a same CpG used during purity estimation. (**D**) Distribution relative to genes of all CpGs used for purity estimation in the three cohorts compared to expected (exp.) values considering all ∼421 000 CpGs available. Expected values are the same across cohorts. (**E**) Distribution of Pearson correlation values between DNA methylation beta values and gene expression data after CpG to gene mapping through genomic coordinates. Only CpGs categorized as in promoters were kept. (**F**) Percentage of samples in a cohort using a same CpG during purity estimation. CpGs that were not used in any sample were excluded. (**G**) Upset plot of CpGs used during purity estimation and how many were in common across cohorts.

proportions of CpGs relative to genes and their beta variance (Supplementary Figure S3). Within chromosomes, most used CpGs were considered distal relative to gene promoters as expected based on the assayed positions by the Illumina platform, but used CpGs were also in greater proportion distal to and less often in promoters than expected (Figure 3D). This was not surprising given that we applied a variance cutoff in PureBeta and that distal CpGs tend to have greater beta variance when compared to those in gene promoters (Supplementary Figure S3). Among the ∼3000 CpGs used in more than 97% of TCGA BRCA samples, even more were considered distal (62.4%) complicating their connection to the regulation of specific genes.

It is more straightforward to associate CpGs to specific genes when the dinucleotides overlap directly with promoter regions. Among the ∼38 000 CpGs used during estimation, ∼8000 CpGs (21%) were in gene promoters and could be linked to 4472 genes in a many-to-many relationship, although most connections were unique (Supplementary Figure S3, Supplementary Table S2). A GSO analysis of these 4472 genes resulted in tumor related terms such as epithelial to mesenchymal transition and *KRAS* signaling among overrepresented hallmarks (Supplementary Table S3). When focusing only on the CpGs used in > 97% of samples, 370 of ∼3000 CpGs (11.3%) were located within the promoters of 381 genes. The GSO analysis did not return any hallmark as significantly overrepresented among these 381 genes. We also calculated the correlation between beta values of CpGs and the expression of the genes in which they were located. As expected, most correlation values were negative. Interestingly, the peak of correlation values moved gradually towards zero as CpGs were restricted to those used for estimating purities in more than 60% (mean) and 97% of samples suggesting an enrichment for e.g. cell type-specific distal elements (Figure 3E).

Unlike in breast cancer, CpGs in the lung cohorts had lower variance of beta values and only 2.7% (∼11 500) and 5.2% (∼22 000) were kept for calculating regressions in TCGA LUAD and TCGA LUSC, respectively, after filtering. It is important to note that variance cutoff was determined by taking into consideration all three TCGA cohorts (Supplementary Figure S1) and that increasing the number of CpGs up to a similar number to what was used in TCGA BRCA did not improve purity estimates considerably in the lung cancer cohorts (Supplementary Figure S4). How many CpGs were used per sample during purity estimation also varied and it was similar to TCGA BRCA in that there was a peak of CpG probes used in close to 50% of samples but diverged from the breast cohort by having very few CpGs used in all or almost all lung cancer samples (Figure 3F). Though fewer CpGs remained after variance filtering in TCGA LUAD, 8.5% of CpGs with regressions were not used for estimation in any sample, a higher percentage than the unused 2.1% in TCGA LUSC, which is in turn higher than the unused percentage in TCGA BRCA (0.7%).

Regarding the distribution of CpGs with regressions and CpGs used to estimate purity in chromosomes, the patterns in TCGA LUAD and TCGA LUSC were similar to those in TCGA BRCA except for CpGs in sex chromosomes (Supplementary Figure S5). This was expected as the proportion of women differed between cohorts (TCGA BRCA: 100%, TCGA LUAD: 53%, TCGA LUSC: 27%). Interestingly, purity estimates calculated with and without CpGs located in sex chromosomes were virtually the same (Supplementary Figure S5). Relative to genes, CpGs followed the distribution patterns seen in TCGA BRCA by favoring those distal to gene promoters in contrast to those directly in promoters (Figure 3D). A GSO analysis of the 1794 and 2539 genes that had 2445 and 4045 CpGs mapped to their promoters in TCGA LUAD and TCGA LUSC respectively matched TCGA BRCA and resulted in overrepresented hallmarks such as *KRAS* signaling in LUSC (Supplementary Table S3). This is consistent with the fact that, while most CpGs were used for estimating purities of samples in only one cohort, a considerable number was used across two or three of the cohorts included in this study (Figure 3G).

Beta adjustment using PureBeta's estimates improves biological interpretability of DNA methylation data

The final step in PureBeta is beta adjustment per CpG given the calculated regressions and tumor purity estimates. This step outputs one object with two different beta tables: one representing values of tumor cells and another representing the inferred values of normal cells as originally described by Staaf and Aine (20). To illustrate that beta adjustment through PureBeta is beneficial to biological inference, we applied the entire PureBeta framework to the 82 SCAN-B TNBC cases with dysregulated *BRCA1* gene function (by promoter hypermethylation, $n = 57$, or pathogenic germline/somatic variants referred to as null, $n = 25$) profiled by the MethylationEPIC platform. However, since clinical subgroups and molecular subtypes of breast cancer are important for disease progression and management (34), we posed the question whether cohort subtype composition would influence the beta adjustment. A first look of the proportions of clinical subgroups as defined by ER and HER2 status in TCGA BRCA showed that these were balanced between the regression and the estimation sample sets after the 80/20 split (Figure 4A) and that there was no difference between subgroups in terms of purity estimate performance (Figure 4B), implying that the beta adjustment step would not be affected.

To see if the proportion of TNBC samples would affect the adjustment outcome on SCAN-B TNBC, we devised a strategy where four different approaches were compared (Figure 4C). The approaches differed with respect to which data set was chosen for calculating the regressions per CpG (the entire TCGA BRCA with 630 samples or only TCGA BRCA TNBC with 86 samples) and whether reference regressions were recalculated by adding beta values and purity estimates of SCAN-B TNBC with those of the reference cohort before proceeding to beta adjustment (referred to as refitting the regressions). As predicted, the adjustment performed with any approach pushed tumor beta values closer to 0 or 1, reducing values around 0.5 (Figure 4D). This implies that adjusted beta values better reflect expected biological patterns of a binary methylation state as evidenced by the increasing separation between beta values of samples considered hypermethylated and null for *BRCA1* with any of the approaches (Figure 4E). Of interest, inferred beta values of normal cells showed the reversed pattern and became more similar between *BRCA1* subgroups also with any of the approaches as exemplified by approaches using TCGA BRCA TNBC (Figure 4F). Beta values per CpG for both tumor and non-malignant cells can be seen in Supplementary Figure S6.

While all approaches showed an improvement from the original data, it was evident that subgroup composition matters as the approaches using regressions from TCGA BRCA TNBC (approaches 3 and 4) outperformed those with all clinical subgroups included (approaches 1 and 2). While refitting regressions calculated from TCGA BRCA by adding SCAN-B TNBC information (approach 2) did improve beta adjustment when compared to without refitting (approach 1), this approach still did not achieve the improvement seen with the methods using only TNBC samples from the start (for instance the true single sample approach 3). Additionally, the approaches using TCGA BRCA TNBC (approaches 3 and 4) showed that even regressions calculated from smaller cohorts (86 samples in this case) can perform well during beta value adjustment when (likely) a similar molecular subtype is analyzed.

Discussion

Epigenetic modifications are a fundamental and enabling characteristic in tumor formation. In the tumor microenvironment, malignant and non-malignant cells are intermixed, and different proportions of these cells can skew DNA methylation data generated from bulk tumor tissue when cells differ regarding methylation states. To address this issue, the recent work by Staaf and Aine (20) proposed the adjustment of beta values of individual CpGs according to the tumor purity of a sample based on modeling the linear relationship between these two variables. In the original publication, beta adjustment was shown to e.g. clearly improve the separation between the well-established gene expression PAM50 Basal and Luminal subtypes of breast cancer using only DNA methylation data. However, important limitations that apply to their methodology include the need for tumor purity estimates of all samples involved and the need for a reasonably large sample cohort for the chosen flexible mixture modeling algorithm to work. Together, these limitations render their method less applicable when one is interested in analyzing few samples. To circumvent the limitations for the Staaf and Aine (20) methodology we developed PureBeta – a complete pipeline that allows for purity estimation and individual CpG adjustment of any number of samples from breast or lung cancer using only DNA methylation data as input through provided reference data. Importantly, PureBeta substantially enhances the work

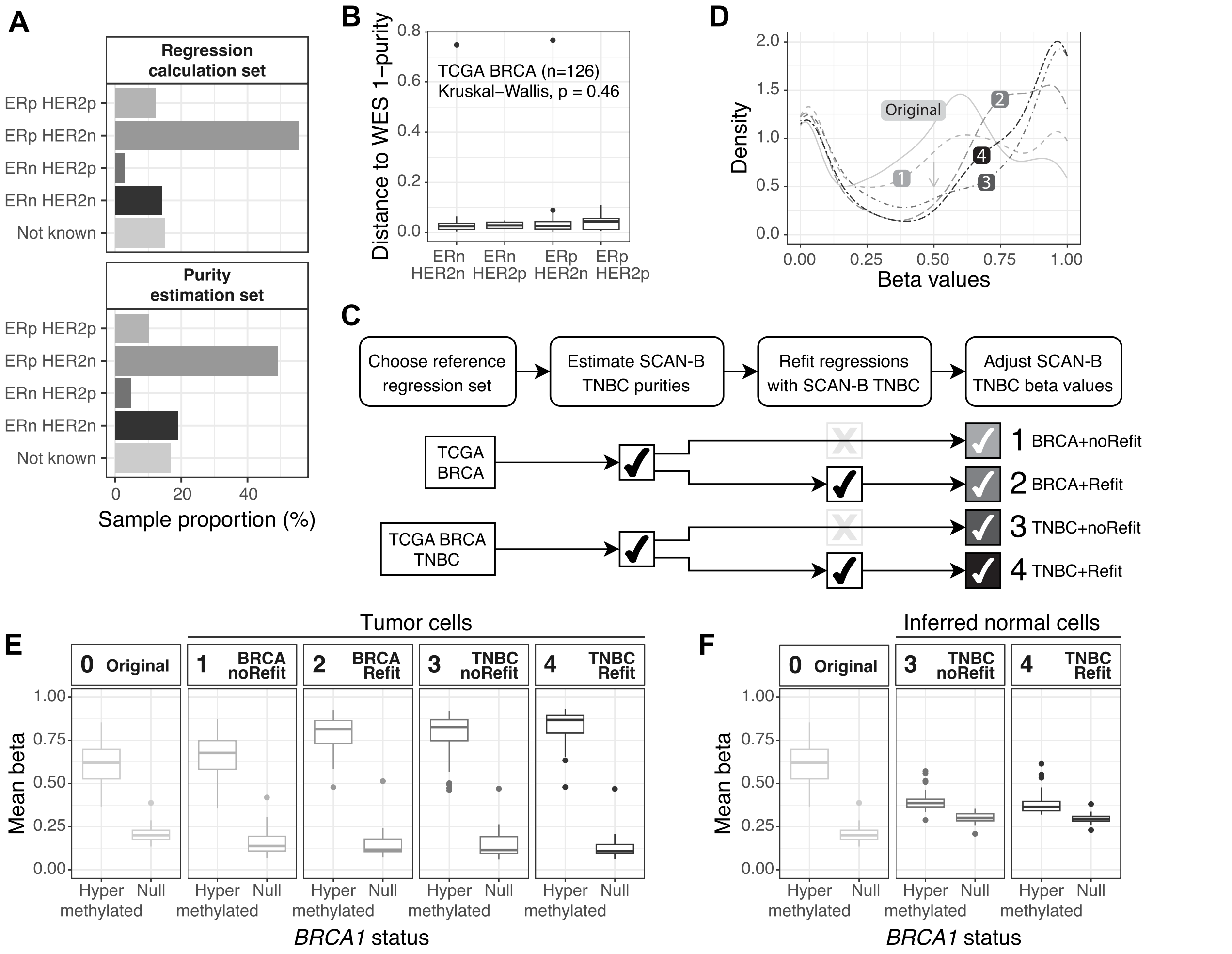


Figure 4. Beta adjustment and cohort influence. (**A**) Proportion of breast cancer clinical subgroups based on ER and HER2 status in the TCGA BRCA cohort when split into two sets. (**B**) Distance between estimated and reference 1-purity values distributed by clinical subgroup of samples. (**C**) Overview of strategy for investigating the influence of cohort composition on beta adjustment. (**D**) Density of beta values from the original cohort and from tumor cells after adjustment for purity with the four methods in (C) showing a decrease in values around 0.5 (arrow). (E, F) Mean value per sample of original beta values, (**E**) adjusted beta values for tumor cells, and (**F**) adjusted beta values for inferred non-malignant background cells as calculated with approaches presented in (C).

of Staaf and Aine (20) by solving the need for external tumor purity values and large sample sizes, which together considerably extends the applicability of their original approach.

PureBeta is a pipeline developed to estimate tumor purity of samples and subsequently adjust their beta values, but it is composed of three modules that can be run independently rendering it more flexible. By providing reference regressions calculated in the first module for cancer types that are widely studied, our approach allows the purity estimation and beta adjustment modules to be applied to a single sample of interest. The provided reference regressions also make this algorithm less computationally demanding for users. In addition, estimating purity directly from the DNA methylation data to be adjusted removes the need for accompanying sequencing data for samples of interest, data that is not always available. Most importantly, PureBeta showed high concordance with standard reference purities calculated from both WES and WGS.

The three modules of PureBeta are intended to be run sequentially with chained outputs and inputs, but modules can be integrated with equivalent estimates from other data sources. For instance, the final beta adjustment step can be performed based on purities obtained from other software instead of estimates from our algorithm. Several methods exist for estimating tumor purity from different biological data (e.g. (17–19)), all of which with specific assumptions, model fitting, and model selection through a comparison metric. Here we used sequencing data to obtain reference tumor purities, the current gold standard practice, despite methodological benchmarking studies having reported poor concordance between different sequencing-based purity and ploidy estimation models and pathologic estimations (35). As shown for two discrepant TCGA BRCA samples, purity estimates calculated with PureBeta can be more reasonable than those calculated based on WES illustrating that the used reference data type for purity estimation is not error free. In fact, for data sets with both DNA methylation and sequencing data available, PureBeta can be used for challenging cases to provide orthogonal support on the suitable sequencing-based purity solution to choose. Although PureBeta is more than just a tool for estimating purities from DNA methylation data, this particular step was also benchmarked against estimates from other software that use DNA methylation or other types of data as input. Our method showed good concordance with the other approaches, but differences between them existed as evidenced by differences in the slopes of the linear models showing that e.g. InfiniumPurify overestimates values at lower purities when compared to PureBeta.

PureBeta is not free from limitations. If instead of using the provided reference set users would like to generate reference regressions from their own cohorts, the algorithm will require more samples to robustly delineate populations per CpG. Still, as shown for TCGA BRCA TNBC, references can be calculated from data sets with <100 samples and used for purity estimation and beta adjustment successfully. PureBeta's performance is, however, influenced by cohort composition of reference and target samples, a question that had not been explored in the original work by Staaf and Aine (20). To partly address this issue, we added the refitting option of recalculating regressions including beta values and purity estimates of samples to be adjusted before performing the final beta adjustment. Original DNA methylation data and purity estimates for the three TCGA cohorts are available in FigShare to be used by PureBeta in refitting mode. As shown through comparing the four approaches of different reference data and refitting option combinations, even if a subgroup of interest is represented in the reference cohort, the methylation signal needed to derive encompassing linear models and adjust beta values might be diluted by samples of other subgroups. Therefore, users of PureBeta should aim to use data similar to the samples of interest for calculating the reference regressions. In addition, users can and should resort to regression refitting in at least two situations: (i) if subgroups of interest are knowingly under sampled or (ii) if the subgroup proportion in the reference data is unknown but samples of interest are available in a large enough number to potentially alter the linear regression modeling (illustrated by approach 2). Finally, PureBeta was not designed for, and may thus not be appropriate for, samples at the edges of the purity distribution such as cancer cell lines composed only of tumor cells.

We showed that CpG availability and use by PureBeta vary depending on cohort and sample analyzed. Analyses of DNA methylation data are often performed on the most varying CpGs in terms of beta values, and PureBeta follows this established practice. The beta variance cutoff was optimized using the three cohorts and adding more CpGs did not improve purity estimates significantly. Still, PureBeta offers the user the option of setting different cutoffs if that would suit their own data sets better. Another common procedure in DNA methylation studies is to discard CpGs located on the X and Y sex chromosomes as they are differentially methylated in males and females. However, excluding such CpGs did not change the pipeline outcome in LUAD and LUSC. This is presumably because beta values of male and female samples are likely assigned to different linear regressions given the inherent differences in chromosome copy and gene silencing between sexes. Consequently, CpGs in the X and Y chromosomes were kept in the reference regression sets provided with PureBeta. In general, CpGs used by PureBeta to estimate purities seemed to be connected to genes involved in tumor processes such as the epithelial to mesenchymal transition as evidenced by the GSO analyses. However, only focusing on CpGs located in gene promoters as explored in this study is an oversimplification of the complexity and importance of DNA methylation in tumorigenesis. Much of the variation observed in beta values is likely not directly related to the regulation of specific tumor suppressor or oncogenes but involved instead in genome regulation and integrity through affecting e.g. transposable elements and enhancers (36,37).

Irrespective of the scope chosen, be it on a gene or whole genome level, the ability to define purer (i.e. more binary) methylation states as done with PureBeta could help investigations of biology and epigenomic changes between different subgroups of any cancer type. PureBeta can also aid in contrasting malignant and non-malignant cells/tissue by providing inferred beta values of normal cells as a proxy for the normal background state of samples (see (20) for further elaboration), data that would have to be otherwise collected for each patient. PureBeta's ability of enhancing methylation levels was exemplified through the biological representation of somatic DNA methylation events in *BRCA1* in the SCAN-B TNBC cohort. The same cohort was used to showcase that PureBeta performs well also on DNA methylation data obtained from the newer MethylationEPIC platform despite reference regressions being calculated from the older HumanMethylation450, a testament to the continued usability of our tool in the future. Finally, as these widely used Illumina arrays are released with backwards compatibility, the amount of DNA methylation data generated with them and made publicly available through initiatives such as TCGA indicates the potential for extending PureBeta to any cancer type, even rarer ones with lower number of samples collected. Taken together, PureBeta provides researchers with a tool that can assist in furthering our understanding of pure methylation phenotypes in cancer, as well as of the contribution of the tumor microenvironment to observed (bulk) methylation profiles in cancer.

Data availability

Most data underlying this article are available from The Cancer Genome Atlas initiative and can be retrieved through the Genomic Data Commons Data Portal at https://portal.gdc.cancer.gov. Other data are available from Gene Expression Omnibus at https://www.ncbi.nlm.nih.gov and can be accessed with the accession number GSE148748. PureBeta is freely available under the GPL-3.0 license as an R package in GitHub at https://github.com/StaafLab/PureBeta and reference data is available in FigShare at https://doi.org/10.6084/m9.figshare.26272864.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

The authors would like to acknowledge patients and clinicians participating in the SCAN-B study, personnel at the central SCAN-B laboratory at the Division of Oncology and Pathology, Lund University, the Swedish national breast cancer quality registry (NKBC), Regional Cancer Center South (RCC Syd), and the South Sweden Breast Cancer Group (SSBCG).

Author contributions: Iñaki Sasiain: Conceptualization, Formal analysis, Methodology, Validation. Deborah F. Nacer: Conceptualization, Data curation, Formal analysis, Validation, Visualization, Writing – original draft (lead), Writing – review & editing. Mattias Aine: Resources, Data curation. Srinivas Veerla: Data curation, Supervision. Johan Staaf:

Conceptualization, Resources, Data curation, Supervision, Funding acquisition, Project administration.

Funding

Swedish Cancer Society [CAN 2021/1407]; Mrs Berta Kamprad Foundation [FBKS-2020-5]; Swedish Research Council [2021-01800]; and Swedish governmental funding [ALF, 2022/0021]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the paper. Funding for open access charge: Lund University.

Conflict of interest statement

None declared.

References

1. Hanahan,D. (2022) Hallmarks of cancer: new dimensions. *Cancer Discov.*, **12**, 31–46.
2. Garcia-Martinez,L., Zhang,Y., Nakata,Y., Chan,H.L. and Morey,L. (2021) Epigenetic mechanisms in breast cancer therapy and resistance. *Nat. Commun.*, **12**, 1786.
3. Chaligne,R., Gaiti,F., Silverbush,D., Schiffman,J.S., Weisman,H.R., Kluegel,L., Gritsch,S., Deochand,S.D., Gonzalez Castro,L.N., Richman,A.R., *et al.* (2021) Epigenetic encoding, heritability and plasticity of glioma transcriptional cell states. *Nat. Genet.*, **53**, 1469–1479.
4. Lianidou,E. (2021) Detection and relevance of epigenetic markers on ctDNA: recent advances and future outlook. *Mol Oncol*, **15**, 1683–1700.
5. Glodzik,D., Bosch,A., Hartman,J., Aine,M., Vallon-Christersson,J., Reutersward,C., Karlsson,A., Mitra,S., Nimeus,E., Holm,K., *et al.* (2020) Comprehensive molecular comparison of BRCA1 hypermethylated and BRCA1 mutated triple negative breast cancers. *Nat. Commun.*, **11**, 3747.
6. Mansouri,A., Hachem,L.D., Mansouri,S., Nassiri,F., Laperriere,N.J., Xia,D., Lindeman,N.I., Wen,P.Y., Chakravarti,A., Mehta,M.P., *et al.* (2019) MGMT promoter methylation status testing to guide therapy for glioblastoma: refining the approach based on emerging evidence and current challenges. *Neuro. Oncol.*, **21**, 167–178.
7. Shigeyasu,K., Nagasaka,T., Mori,Y., Yokomichi,N., Kawai,T., Fuji,T., Kimura,K., Umeda,Y., Kagawa,S., Goel,A., *et al.* (2015) Clinical Significance of MLH1 Methylation and CpG Island Methylator Phenotype as Prognostic Markers in Patients with Gastric Cancer. *PLoS One*, **10**, e0130409.
8. Rauluseviciute,I., Drablos,F. and Rye,M.B. (2020) DNA hypermethylation associated with upregulated gene expression in prostate cancer demonstrates the diversity of epigenetic regulation. *BMC Med Genomics*, **13**, 6.
9. Nishida,J., Momoi,Y., Miyakuni,K., Tamura,Y., Takahashi,K., Koinuma,D., Miyazono,K. and Ehata,S. (2020) Epigenetic remodelling shapes inflammatory renal cancer and neutrophil-dependent metastasis. *Nat. Cell Biol.*, **22**, 465–475.
10. Martisova,A., Holcakova,J., Izadi,N., Sebuyoya,R., Hrstka,R. and Bartosik,M. (2021) DNA methylation in solid tumors: functions and methods of detection. *Int. J. Mol. Sci.*, **22**, 4247.
11. Sandoval,J., Heyn,H., Moran,S., Serra-Musach,J., Pujana,M.A., Bibikova,M. and Esteller,M. (2011) Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*, **6**, 692–702.
12. Moran,S., Arribas,C. and Esteller,M. (2016) Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*, **8**, 389–399.
13. Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.
14. Teschendorff,A.E., Breeze,C.E., Zheng,S.C. and Beck,S. (2017) A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinf.*, **18**, 105.
15. Chakravarthy,A., Furness,A., Joshi,K., Ghorani,E., Ford,K., Ward,M.J., King,E.V., Lechner,M., Marafioti,T., Quezada,S.A., *et al.* (2018) Pan-cancer deconvolution of tumour composition using DNA methylation. *Nat. Commun.*, **9**, 3220.
16. Arneson,D., Yang,X. and Wang,K. (2020) MethylResolver-a method for deconvoluting bulk DNA methylation profiles into known and unknown cell contents. *Commun. Biol.*, **3**, 422.
17. Zheng,X., Zhao,Q., Wu,H.J., Li,W., Wang,H., Meyer,C.A., Qin,Q.A., Xu,H., Zang,C., Jiang,P., *et al.* (2014) MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes. *Genome Biol.*, **15**, 419.
18. Zheng,X., Zhang,N., Wu,H.J. and Wu,H. (2017) Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. *Genome Biol.*, **18**, 17.
19. Benelli,M., Romagnoli,D. and Demichelis,F. (2018) Tumor purity quantification by clonal DNA methylation signatures. *Bioinformatics*, **34**, 1642–1649.
20. Staaf,J. and Aine,M. (2022) Tumor purity adjusted beta values improve biological interpretability of high-dimensional DNA methylation data. *PLoS One*, **17**, e0265557.
21. Wang,F., Zhang,N., Wang,J., Wu,H. and Zheng,X. (2016) Tumor purity and differential methylation in cancer epigenomics. *Brief Funct Genomics*, **15**, 408–419.
22. Carter,S.L., Cibulskis,K., Helman,E., McKenna,A., Shen,H., Zack,T., Laird,P.W., Onofrio,R.C., Winckler,W., Weir,B.A., *et al.* (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.*, **30**, 413–421.
23. Hoadley,K.A., Yau,C., Hinoue,T., Wolf,D.M., Lazar,A.J., Drill,E., Shen,R., Taylor,A.M., Cherniack,A.D., Thorsson,V., *et al.* (2018) Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, **173**, 291–304.
24. Saal,L.H., Vallon-Christersson,J., Hakkinen,J., Hegardt,C., Grabau,D., Winter,C., Brueffer,C., Tang,M.H., Reutersward,C., Schulz,R., *et al.* (2015) The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. *Genome Med*, **7**, 20.
25. Ryden,L., Loman,N., Larsson,C., Hegardt,C., Vallon-Christersson,J., Malmberg,M., Lindman,H., Ehinger,A., Saal,L.H. and Borg,A. (2018) Minimizing inequality in access to precision medicine in breast cancer by real-time population-based molecular analysis in the SCAN-B initiative. *Br. J. Surg.*, **105**, e158–e168.
26. Staaf,J., Glodzik,D., Bosch,A., Vallon-Christersson,J., Reutersward,C., Hakkinen,J., Degasperi,A., Amarante,T.D., Saal,L.H., Hegardt,C., *et al.* (2019) Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nat. Med.*, **25**, 1526–1533.
27. Perperoglou,A., Sauerbrei,W., Abrahamowicz,M. and Schmid,M. (2019) A review of spline function procedures in R. *BMC Med. Res. Methodol.*, **19**, 46.
28. Qin,Y., Feng,H., Chen,M., Wu,H. and Zheng,X. (2018) InfiniumPurify: an R package for estimating and accounting for tumor purity in cancer methylation research. *Genes Dis*, **5**, 43–45.
29. Newman,A.M., Steen,C.B., Liu,C.L., Gentles,A.J., Chaudhuri,A.A., Scherer,F., Khodadoust,M.S., Esfahani,M.S., Luca,B.A., Steiner,D., *et al.* (2019) Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.*, **37**, 773–782.
30. Aran,D., Sirota,M. and Butte,A.J. (2015) Systematic pan-cancer analysis of tumour purity. *Nat. Commun.*, **6**, 8971.
31. Wu,T., Hu,E., Xu,S., Chen,M., Guo,P., Dai,Z., Feng,T., Zhou,L., Tang,W., Zhan,L., *et al.* (2021) clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (Camb)*, **2**, 100141.

32. Liberzon,A., Birger,C., Thorvaldsdottir,H., Ghandi,M., Mesirov,J.P. and Tamayo,P. (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.*, **1**, 417–425.
33. Salas,L.A., Koestler,D.C., Butler,R.A., Hansen,H.M., Wiencke,J.K., Kelsey,K.T. and Christensen,B.C. (2018) An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. *Genome Biol.*, **19**, 64.
34. Nolan,E., Lindeman,G.J. and Visvader,J.E. (2023) Deciphering breast cancer: from biology to the clinic. *Cell*, **186**, 1708–1728.
35. Haider,S., Tyekucheva,S., Prandi,D., Fox,N.S., Ahn,J., Xu,A.W., Pantazi,A., Park,P.J., Laird,P.W., Sander,C., *et al.* (2020) Systematic assessment of tumor purity and its clinical implications. *JCO Precis. Oncol.*, **4**, PO.20.00016.
36. Nishiyama,A. and Nakanishi,M. (2021) Navigating the DNA methylation landscape of cancer. *Trends Genet.*, **37**, 1012–1027.
37. Jones,P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, **13**, 484–492.

Received: January 30, 2024. **Revised:** October 1, 2024. **Editorial Decision:** October 16, 2024. **Accepted:** October 18, 2024