bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Functional normalization of 450k methylation array data improves replication in large cancer studies

Jean-Philippe Fortin[1], Aurélie Labbe[2,3], Mathieu Lemire[4], Brent W. Zanke[5], Thomas J. Hudson[4,6], Elana J. Fertig[7], Celia M.T. Greenwood[2,8,9], and Kasper D. Hansen[1,10,*]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
[2]Department of Epidemiology, Biostatistics and Occupational Health, McGill University
[3]Douglas Mental Health University Institute, McGill University
[4]Ontario Institute of Cancer Research
[5]Clinical Epidemiology Program, Ottawa Hospital Research Institute
[6]Departments of Molecular Genetics and Medical Biophysics, University of Toronto
[7]Department of Oncology, Sidney Kimmel Cancer Center, Johns Hopkins School of Medicine
[8]Lady Davis Institute for Medical Research, Jewish General Hospital Montreal
[9]Department of Oncology, McGill University
[10]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine

Abstract

We propose an extension to quantile normalization which removes unwanted technical variation using control probes. We adapt our algorithm, functional normalization, to the Illumina 450k methylation array and address the open problem of normalizing methylation data with global epigenetic changes, such as human cancers. Using datasets from The Cancer Genome Atlas and a large case-control study, we show that our algorithm outperforms all existing normalization methods with respect to replication of results between experiments, and yields robust results even in the presence of batch effects. Functional normalization can be applied to any microarray platform, provided suitable control probes are available.

*To whom correspondence should be addressed. Email: khansen@jhsph.edu

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

In humans, DNA methylation is an important epigenetic mark occurring at CpG dinucleotides, implicated in gene silencing. In 2011, Illumina released the IlluminaHumanMethylation450 bead array [Bibikova et al., 2011], also known as the "450k array". This array has enabled population-level studies of DNA methylation by providing a cheap, high-throughput and comprehensive assay for DNA methylation. Applications of this array to population level data includes epigenome-wide association studies (EWAS) [Rakyan et al., 2011, Liu et al., 2013] and large-scale cancer studies such as the ones available through The Cancer Genome Atlas (TCGA). To date, around 9000 samples are available from NCBI GEO, and around 8000 samples from TCGA have been profiled on either the 450k array, the 27k array or both.

Studies of DNA methylation in cancer pose a challenging problem for array normalization. It is widely accepted that most cancers show massive changes in their methylome, making the marginal distribution of methylation across the genome different between cancer and normal samples. The global hypomethylation commonly observed in human cancers was recently shown to be organized into large, well-defined domains [Hansen et al., 2011, Berman et al., 2012]. It is worth noting that there are other situations where global methylation differences can be expected, such as between cell types and tissues.

Several methods have been proposed for normalization of the 450k array, including Quantile normalization [Touleimat and Tost, 2012, Aryee et al., 2014], SWAN [Maksimovic et al., 2012], BMIQ [Teschendorff et al., 2013], dasen [Pidsley et al., 2013], and noob [Triche et al., 2013]. A recent review examined the performance of many normalization methods in a setting with global methylation differences and concluded "there is to date no between-array normalization method suited to 450K data that can bring enough benefit to counterbalance the strong impairment of data quality they can cause on some data sets" [Dedeurwaerder et al., 2013]. The authors note that absence of normalization outperforms the methods they evaluate, highlighting the importance of benchmarking any method against raw data.

The difficulties in normalizing DNA methylation data across cancer samples have been recognized for a while. In earlier work on the CHARM platform [Irizarry et al., 2008], Aryee et al. [2011] proposed a variant of subset quantile normalization [Wu and Aryee, 2010] as a solution. The CHARM platform is based on a different assay than the 450k platform. For CHARM, input DNA is compared to DNA processed by a methylation dependent restriction enzyme. Aryee et al. [2011] used subset quantile normalization to normalize the input channels from different arrays to each other. The 450k assay does not involve an input channel; it is based on bisulfite conversion.

Any high-throughout assay suffers from unwanted variation [Leek et al., 2010]. This is commonly addressed by an unsupervised normalization step, followed by application of a batch effect removal tool such as SVA [Leek and Storey, 2007, 2008], ComBat [Johnson

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

et al., 2007] or RUV [Gagnon-Bartsch and Speed, 2012]. The term "batch effect" often refers to unwanted variation remaining after initial normalization, and batch effects have unsurprisingly been observed in studies using the 450K array [Harper et al., 2013]

In this work we propose a method we call functional normalization, which uses control probes to act as surrogates for unwanted variation and some batch effects. We apply this method to the analysis of 450k array data, and show that functional normalization outperforms all existing normalization methods for analysis of datasets with global methylation differences, including studies of human cancer. We also show that functional normalization outperforms the batch removal tool SVA [Leek and Storey, 2007, 2008] in this setting. Our evaluation metrics are focusing on assessing the degree of replication between large-scale studies, arguably the most important, biologically relevant end-point for such studies. Our method is available as the "preprocessFunnorm" function in the minfi package [Aryee et al., 2014] through the Bioconductor project [Gentleman et al., 2004].

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

Control probes may act as surrogates for batch effects

The 450k array contains 848 control probes. These probes can roughly be divided into negative control probes (613), probes intended for between array normalization (186) and the remainder (51) which are designed for quality control, including assessing the bisulfite conversion rate (see Methods and Supplementary Materials). Important for our proposed method, none of these probes are designed to measure biological signal.

Figure 1a shows a heatmap of a simple summary (see Methods) of these control probes, for 200 samples assayed on 8 plates (Ontario-EBV dataset). Columns are the control measure summaries and rows are samples. The samples have been processed on different plates, and we observe a clustering pattern correlated with plate. Figure 1b shows the first two principal components of the same summary data and there is evidence of clustering according to plate. Figure 1c shows how the marginal distributions of the methylated channel very across plates. This suggests that the summarized control probes can be used as surrogates for batch effects, as first proposed by Gagnon-Bartsch and Speed [2012] in the context of gene expression microarrays.

Functional normalization

We propose functional normalization (see Methods), a method which extends the idea of quantile normalization. Quantile normalization forces the empirical marginal distributions of the samples to be the same, which removes all variation in this statistic. In contrast, functional normalization only removes variation explained by a set of covariates, and is intended to be used when covariates associated with technical variation are available and are independent of biological variation. We adapted functional normalization to data from the 450k array (see Methods), using our observation that the control probe summary measures are associated with technical variability and batch effects. As covariates we recommend using the first $m = 2$ principal components of the control summary matrix, a choice with which we have obtained consistently good results; this is discussed in greater depth below. We have also examined the contributions of the different control summary measures in several different datasets, and we have noted that the control probe summaries given the most weight vary across different data sets.

Functional normalization, like most normalization methods, does not require the analyst to provide information about the experimental design. In contrast, batch removal methods such as SVA [Leek and Storey, 2007, 2008], ComBat [Johnson et al., 2007] and RUV [Gagnon-Bartsch and Speed, 2012] require the user to to provide either batch parameters or an outcome of interest. Like functional normalization, RUV also utilizes the control probes as surrogates for batch effects, but builds the removal of batch effects into a linear

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

model, which returns test statistics for association between probes and phenotype. This limits the use of RUV to a specific statistical model, and methods such as clustering, bumphunting [Aryee et al., 2014, Jaffe et al., 2012] or other regional approaches [Sofer et al., 2013] for identifying differentially methylated regions (DMRs) cannot readily be applied.

Funnorm improves the replication between experiments, even when a batch effect is present

As a first demonstration of the performance of our algorithm, we compare lymphocyte samples from the Ontario data set to EBV-transformed lymphocytes samples from the same collection (see Methods). We have recently studied this transformation [Hansen et al., 2014] and have shown that EBV transformation induces large blocks of hypomethylation encompassing more than half the genome, similar to what is observed between most cancers and normal tissue. This introduces a global shift in methylation as shown by the marginal densities in Supplementary Figure S1.

We divided the dataset into discovery and validation cohorts (see Methods), with 50 EBV transformed lymphocytes and 50 normal lymphocytes in each cohort. As illustrated in Supplementary Figure S2a, we introduced an *in silico* batch effect confounding EBV transformation status in the validation cohort (see Methods), to evaluate the performance of normalization methods in the presence of a known confounding batch effect. This has been previously done by others in the context of genomic prediction [Parker and Leek, 2012]. We normalized the discovery cohort, identified the top k differentially methylated positions (DMPs) and asked: "how many of these k DMPs can be replicated in the validation cohort". We normalized the validation cohort separately from the discovery cohort to mimic a replication attempt in a separate experiment. We identified DMPs in the validation cohort using the same method and the result is quantified using an ROC curve where the analysis result on the discovery cohort is taken as the gold standard.

To enable the comparison between normalization methods, we fix the number of DMPs across all methods. Because we know from previous work [Hansen et al., 2014] that EBV transformation induces large blocks of hypomethylation covering more than half of the genome, we expected to find a large number of DMPs, and we set $k = 100,000$. The resulting ROC curves are shown in Figure 2a. In this figure we show, for clarity, what we have found to be the most interesting alternatives to functional normalization in this setting: Raw data, Quantile normalization as suggested by Touleimat and Tost [2012] and implemented in minfi [Aryee et al., 2014] and the noob background correction [Triche et al., 2013]. Supplementary Figures S3a,b contains results for additional preprocessing methods: BMIQ [Teschendorff et al., 2013], SWAN [Maksimovic et al., 2012] and dasen [Pidsley et al., 2013]. Note that each preprocessing method will results in its own set of gold-standard DMPs and these ROC curves therefore measures the internal consistency of each preprocessing method. We note that functional normalization outperforms raw,

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

quantile and noob normalizations when the specificity is above 90% (which is the relevant range for practical use).

We also measured the agreement between the top k DMPs from the discovery cohort with the top k DMPs from the validation cohort by looking at the overlap percentage. The resulting concordance curves are shown in Figure 2b, with additional methods in Supplementary Figure S3c, and show functional normalization outperforming the other methods.

We can assess the quality of the DMPs replicated between the discovery and validation cohorts by comparing them to the previously identify methylation blocks and differentially methylated regions [Hansen et al., 2014]. In Figure 2c, we present the percentage of the initial $k = 100,000$ DMPs that are both replicated and present among the latter blocks and regions. We note that these previously reported methylation blocks represent large scale, regional changes in DNA methylation and not regions where every single CpG is differentially methylated. Nevertheless, such regions are enriched for DMPs. This comparison shows that functional normalization achieves a greater overlap with this external dataset, with an overlap of 66% compared to 58% for raw data, while other methods, but noob, perform worse than the raw data.

Replication between experiments in cancer study (TCGA-KIRC datasets)

We applied the same discovery-validation scheme to measure performance, used for the analysis of the Ontario-EBV study, on kidney renal clear cell carcinoma samples (KIRC) from TCGA. In total, TCGA has profiled 300 KIRC cancer and 160 normal samples on the 450K platform. Therefore we defined a discovery cohort containing 65 cancers and 65 normals and a validation cohort of 157 cancers and 95 normals (see Methods).

Our *in silico* batch confounding for this experiment succeeded in producing a validation cohort where the cancer samples have greater variation in background noise (Supplementary Figure S2b). This difference in variation is a less severe batch effect compared to the difference in mean background noise we achieved in the Ontario-EBV dataset (Figure S2a). As for the dataset containing EBV transformed samples, we expect large scale hypomethylation in the cancer samples and therefore we again consider $k = 100,000$ loci. The resulting ROC curves are shown in Figure 3a, with additional methods in Supplementary Figure S4a,b. Functional normalization and noob are best and do equally well. Again, the gold-standard set of probes that is used to measure performance in these ROC curves differs between normalization methods, and hence these ROC curves reflect the degree of consistency between experiments within each method.

To further compare the quality of the DMPs found by the different methods, we used an additional dataset from TCGA where the same cancer was assayed with the Illumina 27k platform (see Methods). We focused on the 25,978 CpG sites that were assayed on both

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

platforms and asked about the size of the overlap for the top k DMPs. For the validation cohort, with maximal batch effects, this is depicted in Figure 3b and Supplementary Figure S4c for additional methods; for the discovery cohort, with minimal batch effects, results are presented in Supplementary Figure S4d. Functional normalization shows the best concordance in the presence of a batch effect in the 450k data (the validation cohort) and is comparable to no normalization in the discovery cohort. Note that while noob and Funnorm show equivalent performance in the ROC curves, only Funnorm improves the validation with an external dataset compared to the raw data.

Funnorm preserves subtype heterogeneity in tumor samples (TCGA-AML datasets)

To measure how good our normalization method is at preserving biological variation among heterogeneous samples while removing technical biases, we use 192 acute myeloid leukemia samples from TCGA for which every sample has been assayed on both the 27K and the 450K platforms (see Methods). These two platforms assay 25,978 CpGs in common (but note the probe design changes between array types), and we can therefore assess the degree of agreement between measurements of the same sample on two different platforms, assayed at different time points. The 450k data appears to be affected by batch and dye bias, see Supplementary Figure S5a,b.

Each sample was classified by TCGA according to the French-American-British (FAB) classification scheme [Bennett et al., 1976], which proposes 8 tumor subtypes, and methylation differences can be expected between the subtypes [Figueroa et al., 2010, Akalin et al., 2012]. Using data from the 27k arrays, we identified the top k DMPs which distinguish the 8 subtypes. In this case, we are assessing the agreement of subtype variability, as opposed to cancer-normal differences. The analysis of the 27k data uses unnormalized data but adjusts for sample batch in the model (see Methods). Using data from the 450k arrays, we first processed the data using the relevant method, and next identified the top k DMPs between the 8 subtypes. The analysis of the 450k data does not include sample batch in the model, which allows us to see how well the different normalization methods remove technical artifacts introduced by batch differences. While both of the analyses are conducted on the full set of CpGs, we focus on the CpGs common between the two platforms and ask "what is the degree of agreement between the top k DMPs identified using the two different platforms". Figure 4a shows that functional normalization and noob outperforms both quantile normalization and raw normalization for all values of k, and functional normalization is marginally better than noob for some values of k. Supplementary Figure S6a shows the results for additional methods. We can also compare the two datasets using ROC curves, with the results from the 27k data as gold standard (Figure 4b and Supplementary Figure S6b). As DMPs for the 27k data we used the 5,451 CpGs that demonstrate an estimated false-discovery rate less than 5%. On the ROC curve functional normalization and noob outperform Quantile and Raw data for the full range of specificity.

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Replication between experiments with small changes

To measure the performance of functional normalization in a setting where there are no global changes in methylation, we used the Ontario-Blood dataset which assays lymphocytes from individuals with and without colon cancer. We expect a very small, if any, impact of colon cancer on the blood methylome. As above, we selected cases and controls to form discovery and validation cohorts, and we introduced an *in silico* batch effect that confounds case-control differences in the validation dataset only (see Methods). The discovery and validation datasets contain respectively 283 and 339 samples. For $k = 100$ loci, both functional and quantile normalization show good agreement between discovery and validation datasets, wheres noob and Raw data show an agreement which is not better than a random selection of probes (Figure 5a, Supplementary Figure S8a).

Funnorm improves X and Y chromosomes probes prediction in blood samples

As suggested previously [Pidsley et al., 2013], one can benchmark performance by identifying DMPs associated with sex. One copy of the X chromosome is inactivated and methylated in females, and the Y chromosome is absent. On the 450k array, 11,232 and 416 probes are annotated to be on the X and Y chromosomes, respectively. For this analysis it is sensible to remove regions of the autosomes which are similar to the sex chromosomes to avoid artificial false positives that are independent of the normalization step. We therefore remove a set of 30,969 probes which have been shown to cross-hybridize between genomic regions [Chen et al., 2013].

To evaluate our method in the presence of batch effects, we introduced an *in silico* batch effect by selecting 101 males and 105 females from different plates (see Methods), thereby confounding plate with sex. Results show that functional normalization performs well (Figure 5b, Supplementary Figure S8b).

Funnorm reduces technical variability

From the Ontario-Replicates lymphocyte dataset (see Methods), we have 19 individuals assayed in technical triplicates dispersed among 51 different chips. To test the performance of each method to remove technical variation, we calculated the probe-specific variance within each triplicate, and averaged the variances across the 19 triplicates. Figure 6 presents box plots of these averaged probe variances of the all methods. All normalization methods improve on Raw data, and functional normalization is in the top 3 of the normalization methods. Dasen in particular does well on this benchmark, which shows that improvements in reducing technical variation do not necessarily lead to similar improvements in the ability to replicate associations.

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Each 450k array is part of a slide of 12 arrays, arranged in 2 columns and 6 rows (see Figure 7). Figure 7a-c shows an effect of column and row position on quantiles of the beta value distribution, across several slides. This effect is not present in all quantiles of the beta distribution, and it depends on the dataset which quantiles are affected. Figure 7d-f shows that functional normalization corrects for this spatial artifact.

Number of principal components

As described above, we recommend using functional normalization with the number of principal components set to $m = 2$. Supplementary Figure S7 shows the impact of varying the number of principal components on various performance measures we have used throughout, and shows that $m = 2$ always is a good choice. It is outperformed by $m = 6$ in the analysis of the KIRC data and by $m = 3$ in the analysis of the AML data, but these choices perform worse in the analysis of the Ontario-EBV data. While $m = 2$ is a good choice across datasets, we leave m to be a user-settable parameter in the implementation of the algorithm. This analysis assumes we use the same m for the analysis of both the discovery and validation dataset. We do this to prevent overfitting and to construct an algorithm with no user input. It is possible to obtain better ROC curves by letting the choice of m vary between discovery and validation, because one dataset is confounded by batch and the other is not.

Comparison to batch effect removal tools

Batch effects are often considered to be unwanted variation remaining after an unsupervised normalization. In the previous assessments we have comprehensively compared functional normalization to existing normalization methods and have shown great performance even in the presence of a batch effect. While functional normalization is an unsupervised normalization procedure, we were interested in comparing its performance to a proper batch removal tool such as SVA [Leek and Storey, 2007, 2008].

We applied SVA to all dataset analyzed previously. We let SVA estimate the number of surrogate variables, and allow this estimation to be done separately on the discovery and the validation datasets, which allows for the best possible performance by the algorithm. Figure 8 compares SVA to functional normalization and Raw data for our evaluation datasets. In the analysis of the Ontario-EBV dataset and the TCGA KIRC dataset, SVA and functional normalization achieves similar performance based on the ROC curves, with functional normalization having slightly higher concordance with an external dataset based on the 27k array. For the TCGA AML dataset and the Ontario-Blood dataset functional normalization clearly outperforms SVA.

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

We have presented functional normalization, an extension of quantile normalization, and have adapted this method to Illumina 450k methylation microarrays. We have shown that this method is especially valuable for normalizing large-scale studies where we expect substantial global differences in methylation, such as in cancer studies or when comparing between tissues, and when the goal is to perform inference at the probe level. Although a normalization method, functional normalization is robust in the presence of a batch effect, and performs better than the batch removal tool SVA on our assessment datasets. This method fills a critical void in the analysis of DNA methylation arrays.

We have evaluated the performance of our method on a number of large scale cancer studies. Critically, we define a successful normalization strategy as one that enhances the ability to reliably detect associations between methylation levels and phenotypes of interest, across multiple experiments. Various other metrics for assessing the performance of normalization methods have been used in the literature on preprocessing methods for Illumina 450k arrays. These metrics include assessing variability between technical replicates [Triche et al., 2013, Teschendorff et al., 2013, Maksimovic et al., 2012, Wu et al., 2013, Dedeurwaerder et al., 2013], and comparing methylation levels to an external gold standard measurement such as bisulfite sequencing [Teschendorff et al., 2013, Dedeurwaerder et al., 2013, Touleimat and Tost, 2012]. We argue that a method which yields unbiased and precise estimates of methylation in a single sample does not necessarily lead to improvements in estimating the differences between samples, yet this is the relevant end-goal for any scientific investigation. For example, it is well known that the RMA method [Irizarry et al., 2003] for analysis of Affymetrix gene expression microarrays introduces bias into the estimation of fold-changes for differentially expressed genes; however, this bias is offset by a massive reduction in variance for non-differentially expressed genes, leading to the method's proven performance.

In our comparisons, we have separately normalized the discovery and the validation dataset, to mimic replication across different experiments. We have shown that functional normalization is always amongst the top performing methods, whereas other normalization methods tend to perform well on some, but not all, of our test datasets. As suggested by Dedeurwaerder et al. [2013], our benchmarks show the importance of comparing performance to Raw data, which outperforms (using our metrics) some of the existing normalization methods. In several datasets we have observed that the within-array normalization methods, SWAN and BMIQ, have very modest performance compared to Raw data and between-array normalization methods. This suggest that within-array normalization methods does not lead to improvements in the ability to replicate findings between experiments.

Our closest competitor is noob, [Triche et al., 2013], which includes both a background correction and a dye-bias equalization. We outperform noob on our concordance curves in all datasets, and our performance is notably better in the analysis the Ontario-EBV,

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Ontario-Blood and Ontario-Sex data sets. We find the results on the Ontario-EBV dataset to be particular compelling since the *in silico* batch effect that we introduced was substantial, and yet we were still able to obtain good agreement between discovery and validation analyses.

Our method relies on the fact that control probes carry information about technical batch effects. This idea was previously proposed by Gagnon-Bartsch and Speed [2012] and used to design the batch removal tool RUV. As discussed in the Results section, the RUV method is tightly integrated with a specific statistical model, requires the specification of the experimental design, and cannot readily accommodate regional methods [Aryee et al., 2014, Jaffe et al., 2012, Sofer et al., 2013] nor clustering. In contrast, functional normalization does not require specification of the experimental design and is applied as a preprocessing step. Batch effects are often considered to be unwanted variation remaining after an unsupervised normalization, and we conclude that functional normalization removes a greater amount of unwanted variation in the preprocessing step.

However, control probes cannot measure unwanted variation arising from factors such as cell type heterogeneity, which is known to be an important confounder in methylation studies of samples containing mixtures of cell types [Jaffe and Irizarry, 2014]. This is an example of unwanted variation from a biological, as opposed to technical, source. Cell type heterogeneity is a particular challenge in EWAS studies of whole blood, but this requires other tools and approaches to address.

Surprisingly, we show that functional normalization improves on the batch removal tool SVA applied to raw data, in the datasets we have assessed. It is a very strong result that an unsupervised normalization method improves on batch removal tool, which requires the specification of the comparison of interest.

While we have shown that functional normalization performs well in the presence of batch effects, we still recommend that any large-scale study consider the application of batch removal tools such as SVA [Leek and Storey, 2007, 2008], ComBat [Johnson et al., 2007] or RUV [Gagnon-Bartsch and Speed, 2012] after using functional normalization, due to their proven performance and their potential for removing unwanted variation which cannot be measured by control probes. As an example, Jaffe and Irizarry [2014] discusses the use of tools to control for cell type heterogeneity.

The analysis of the Ontario-Blood dataset suggests that functional normalization has potential to aid the analysis in a standard EWAS setting, where only a small number of differentially methylated loci are expected. However, if only very few probes are expected to demonstrate changes, and if those changes are small, it becomes difficult to evaluate the performance of our normalization method using our criteria of successful replication.

The main ideas of functional normalization can readily be applied to other microarray platforms, including gene expression and miRNA arrays, provided that the platform of interest contains a suitable set of control probes. We expect the method to be particular

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

useful when applied to data with large anticipated differences between samples.

Acknowledgements

Funding: Funding for the methylation genotyping was obtained from a GL2 grant from the Ontario Research Fund to BWZ and TJH. CCFR is supported by the National Cancer Institute, National Institutes of Health under RFA #CA-95-011. The OFCCR is supported by grant U01 CA074783. JPF was partially supported by les Fonds de Recherche en Santé du Québec, awarded to AL, the Natural Sciences and Engineering Research Council of Canada and by les Fonds de recherche Nature et technologies du Québec, and a pilot project from the Johns Hopkins Head and Neck Cancer SPORE awarded to EJF. Research reported in this publication was supported by National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM083084. TJH and BWZ are recipients of Senior Investigator Awards from the Ontario Institute for Cancer Research, through generous support from the Ontario Ministry of Research and Innovation. EJF is also supported by the National Cancer Institute (CA141053).

The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the CFRs, nor does mention of trade names, commercial products or organizations imply endorsement by the U.S. Government or the CFR. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: None declared.

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Methods

Infinium HumanMethylation450 BeadChip

We use the following terminology, consistent with the minfi package [Aryee et al., 2014]: the 450k array is available as slides consisting of 12 arrays. These arrays are arranged in a 6 row by 2 columns layout. The scanner can process up to 8 slides in a single plate.

We use the standard formula $\beta = M/(M + U + 100)$ for estimating percent methylation given (un)methylation intensities U and M.

Functional normalization uses information from the 848 control probes on the 450k array, as well as the "out-of-band" probes discussed in Triche et al. [2013]. These control probes are not part of the standard output from GenomeStudio, the default Illumina software. Instead we use the IDAT files from the scanner together with the open source illuminaio [Smith et al., 2013] package to access the full data from the IDAT files. This step is implemented in minfi [Aryee et al., 2014].

Control probe summaries

We transform the 848 control probes, as well as the out-of-band probes [Triche et al., 2013] into 42 summary measures. The control probes contributes 38 of these 42 measures and the out-of-band contributes 4. An example of a control probe summary is the mapping of 61 "C" normalization probes to a single summary values, their mean. The out-of-band probes are the intensities of the Type I probes measured in the opposite color channel from the probe design. For the 450k platform, this means 92,596 green intensities, and 178,406 red intensities that can be used to estimate background intensity, and we summaries these values into 4 summary measures. A full description of how the control probes and the out-of-band probes are transformed into the summary control measures are listed in the Supplementary Material.

Functional normalization: the general framework

Functional normalization extends the idea of quantile normalization, by adjusting for known covariates measuring unwanted variation. In this section we present a general model that is not specific to methylation data. The adaptation of this general model to the 450k data is discussed in the next section. The general model is as follows. Consider $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ independent high dimensional vectors each associated with a set of scalar covariates $Z_{i,j}$ with $i = 1, \ldots, n$ indexing samples and $j = 1, \ldots, m$ indexing covariates. Ideally these known covariates are associated with unwanted variation and unassociated with biological variation; functional normalization attempts to remove their influence.

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

For each high-dimensional observation $\mathbf{Y}_i$, we form the empirical quantile function for its marginal distribution, and denote it by q_i^{emp}. Quantile functions are defined on the unit interval and we use the variable $r \in [0,1]$ to evaluate them pointwise, like $q_i^{\text{emp}}(r)$. We assume the following model in pointwise form

$$q_i^{\text{emp}}(r) = \alpha(r) + \sum_{j=1}^{m} Z_{i,j}\beta_j(r) + \epsilon_i(r) \tag{1}$$

which has the functional form

$$q_i^{\text{emp}} = \alpha + \sum_{j=1}^{m} Z_{i,j}\beta_j + \epsilon_i$$

The parameter function α is the mean of the quantile functions across all samples, β_j are the coefficient functions associated with the covariates and ϵ_i are the error functions which are assumed to be independent and centered around 0.

In this model, the term

$$\sum_{j=1}^{m} Z_{i,j}\beta_j$$

represents variation in the quantile functions explained by the covariates. By specifying known covariates that measure unwanted variation and that are not associated with biological signal, functional normalization removes unwanted variation by regressing out the latter term. An example of a known covariate could be processing batch. In a good experimental design, processing batch will not be associated with biological signal.

In particular, assuming we have obtained estimates $\hat{\beta}_j$ for $j = 1, \ldots, m$, we form the functional normalized quantiles by

$$q_i^{\text{Funnorm}}(r) = q_i^{\text{emp}}(r) - \sum_{j=1}^{m} Z_{i,j}\hat{\beta}_j(r)$$

We then transform $\mathbf{Y}_i$ into the functional normalized quantity $\tilde{\mathbf{Y}}_i$ using the formula

$$\tilde{\mathbf{Y}}_i = q_i^{\text{Funnorm}}\left((q_i^{\text{emp}})^{-1}(\mathbf{Y}_i)\right)$$

This ensures that the marginal distribution of $\tilde{\mathbf{Y}}_i$ has q_i^{Funnorm} as its quantile function.

We now describe how to obtain estimates $\hat{\beta}_j$ for $j = 1, \ldots, m$. Our model 1 is an example of function-on-scalar regression, described in [Reiss et al., 2010]. The literature on function-on-scalar regression makes assumptions about the smoothness of the coefficient functions and uses a penalized framework because the observations appear noisy and non-smooth. In contrast, because our observations $\mathbf{Y}_i$ are high-dimensional, the observed empirical quantile functions have a high degree of smoothness. This allows us to circumvent the

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

smoothing approach used in traditional function-on-scalar regression. We use a dense grid of H equidistant points between 0 and 1, and we assume that H is much smaller than the dimension of $\mathbf{Y}_i$. On this grid, model 1 reduces pointwise to a standard linear model. Because the empirical quantile functions $q^{\text{emp}}(r)$ are smooth in r, the parameter estimates are smooth across r as well. This allows us to use H standard linear model fits to compute estimates $\hat{\alpha}(h), \hat{\beta}_j(h), j = 1, \ldots, m$ with h being on the dense grid $\{h \in d/H : d = 0, 1, \ldots, H\}$. We next form estimates $\hat{\alpha}(r), \hat{\beta}_j(r), j = 1, \ldots, m$ for any $r \in [0, 1]$ by linear interpolation. This is much faster than traditional approaches in the function-on-scalar regression literature.

Importantly, in this framework, using a saturated model in which all the variation (but the mean) is explained by the covariates results in removing all variation and is equivalent to quantile normalization. In our notation, quantile normalized quantile functions are

$$q_i^{\text{quantile}}(r) = \hat{\alpha}(r)$$

where $\hat{\alpha}$ is the mean of the empirical quantile functions. This corresponds to the maximum variation that can be removed in our model. In contrast, including no covariates makes the model comparable to no normalization at all. By choosing covariates which only measure unwanted technical variation, functional normalization will only remove the variation explained by these technical measurements and will leave biological variation intact. Functional normalization allows a sensible tradeoff between not removing any technical variation at all (no normalization) and removing too much variation, including global biological variation, as it can occur in quantile normalization.

Functional normalization for 450k arrays

We apply the functional normalization model to the methylated (M) and unmethylated (U) channels separately. Since we expect the relationship between the methylation values and the control probes to differ between Type I and Type II probes, functional normalization is also applied separately by probe type to obtain more representative quantile distributions. We address probes mapping to the sex chromosomes separately, see below. This results in 4 separate applications of functional normalization, using the exact same covariate matrix, with more than 100,000 probes in each normalization fit. For functional normalization, we pick $H = 500$ equidistant points (see notation in previous section). As covariates we use the first $m = 2$ principal components of the summary control measures as described above. We do this because the control probes are not intended to measure biological signal since they are not designed to hybridize to genomic DNA. Our choice of $m = 2$ is based on empirical observations on several datasets.

Following the ideas from quantile normalization for 450k arrays Touleimat and Tost [2012], Aryee et al. [2014], we normalize the probes mapping to the sex chromosomes (11,232

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and 416 probes for the X and Y chromosomes respectively) separately from the autosomal probes. For each of the two sex chromosomes, we normalize males and females separately. For the X chromosome we use functional normalization and for the Y chromosome we use quantile normalization since the small number of probes on this chromosome violates the assumptions of functional normalization which results in instability.

Functional normalization only removes variation in the marginal distributions of the two methylation channels associated with control probes. This preserves any biological global methylation difference between samples.

Data

The Ontario study.

The Ontario study consists of samples from 2200 individuals from the Ontario Familial Colon Cancer Registry (OFCCR) [Cotterchio et al., 2000] who had previously been genotyped in a case-control study of colorectal cancer in Ontario [Zanke et al., 2007]. The majority of these samples are lymphocytes derived from whole blood. We use various subsets of this dataset for different purposes.

The Ontario-EBV dataset. Lymphocyte samples from 100 individuals from the Ontario study were transformed into immortal lymphoblastoid cell lines (LCL) using EBV transformation. We divided the 100 EBV-transformed samples into two equal equally sized datasets (discovery and validation). For the discovery dataset, we matched the 50 EBV-transformed samples to 50 other lymphocyte samples assayed on the same plates. For the validation dataset, we matched the 50 EBV-transformed samples to 50 other lymphocyte samples assayed on different plates.

The Ontario-Blood dataset. From the Ontario study, we first created a discovery-validation design where we expect only a small number of loci to be differentially methylated. For the discovery dataset, we selected all cases and controls on 3 plates showing little evidence of plate effects among the control probes, which yielded a total of 52 cases and 231 controls. For the validation dataset, we selected 4 plates where the control probes did show evidence of a plate effect and then selected cases and controls from separate plates, to maximize a confounding effect of plate. This yielded a total of 175 cases and 163 controls.

The Ontario-Sex dataset. Among 10 plates for which the control probes demonstrated differences in distribution depending on plate, we selected 101 males from a set of 5 plates and 105 females from another set of 5 plates, attempting to maximize the confounding effect of batch on sex.

The Ontario-Replicates dataset. Amongst the lymphocyte samples from the Ontario study, 19 samples have been assayed 3 times each. One replicate is a hybridization repli-

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

cate and the other replicate is a bisulfite conversion replicate. The 57 samples have been assayed on 51 different slides across 11 different plates.

The TCGA-KIRC datasets. From The Cancer Genome Atlas (TCGA) we have access to kidney renal clear cell carcinoma and normal samples, assayed on two different methylation platforms. We use the level 1 data which contains the IDAT files. For the 450k platform, TCGA has assayed 300 tumor samples and 160 normal samples. For the discovery set, we select 65 tumor samples and 65 matched normals from slides showing little variation in the control probes. These 130 samples were assayed on 3 plates. For the validation dataset we select the remaining 95 normal samples together with all 157 cancer samples part of the same TCGA batches as the 95 normals. These samples were spread over all 9 plates, therefore maximizing potential batch effects. For the 27k platform, TCGA has assayed 219 tumor samples and 199 normals. There is no overlap between the individuals assayed on the 450k platform and the individuals assayed on the 27k platform.

The TCGA-AML datasets. Also from TCGA, we used data from 194 acute myeloid leukemia (AML) samples, where each sample was assayed twice: first on the 27K Illumina array and subsequently on the 450K array. Every sample but 2 have been classified according to the French-American-British (FAB) subtype classification scheme [Bennett et al., 1976] which classifies the tumor into one of 8 subtypes. The 2 unclassified samples were removed post-normalization. We use the data known as level 1, which contains the IDAT files.

WGBS EBV data. Hypomethylated blocks and small differentially methylated regions (DMRs) between transformed and quiescent cells were obtained from a previous study [Hansen et al., 2014]. Only blocks and DMRs with a family-wise error rate equal to 0 was retained (see the reference). A total of 228,696 probes on the 450K array overlap with the blocks and DMRs.

Comparison to normalization methods

We have compared our functional normalization, "Funnorm", to the most popular normalization methods used for the 450k array. This includes the following between-array normalization methods: (1) "Quantile": stratified quantile normalization as proposed by Touleimat and Tost [2012] and implemented in minfi [Aryee et al., 2014], (2) "dasen": background adjustment and between-sample quantile normalization of M and U separately as implemented in Pidsley et al. [2013] and (3) "noob": a background adjustment model using the out-of-band control probes followed by a dye bias correction [Triche et al., 2013], implemented in the methylumi package. We also consider two within -array normalization methods: (4) "SWAN": Subset-quantile within-array normalization [Maksimovic et al., 2012] and (5) "BMIQ": Beta-mixture quantile normalization [Teschendorff et al., 2013]. Finally, we consider (6) "Raw" data: no normalization.

In its current implementation, noob yields missing values for at most a couple of thou-

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

sand loci per array. This is based on excluding loci below an array-specific detection limit. We have discarded those loci from our performance measures, but only for the noob performance measures. In its current implementation, BMIQ produced missing values for all type II probes in 5 samples for the TCGA AML dataset. We have excluded these samples for our performance measures, but only for our BMIQ performance measures.

For clarity, in figures we focus on the top-performing methods which are Quantile, Raw and noob. The assessments of the other methods, dasen, BMIQ and SWAN are available in Supplementary Materials.

Comparison to batch removal tools

We used the reference implementation of SVA in the "sva" package [Leek et al., 2012]. We applied SVA to the M-values obtained from the raw data. Surrogate variables were estimated using the iteratively re-weighted surrogate variable analysis algorithm [Leek and Storey, 2008], and were estimated separately for the discovery and validation cohorts. In the analysis of the Ontario-EBV dataset, SVA found 21 and 24 surrogate variables respectively for the discovery and the validation cohorts. In the analysis of the Ontario-Blood dataset, SVA found 18 and 21 surrogate variables respectively for the discovery and the validation cohorts. In the analysis of the TCGA KIRC dataset, SVA found 29 and 32 surrogate variables respectively for the discovery and the validation cohorts. In the analysis of the TCGA LAML dataset, SVA found 24 surrogate variables.

Identification of differentially methylated positions

To identify differentially methylated positions (DMPs), we used F-statistics from a linear model on the beta values from the array. The linear model was applied on a probe-by-probe basis. In most cases, the model includes case/control status as a factor. In the 27K data, we adjust for batch by including a plate indicator (given by TCGA), as a factor in the model.

Discovery-Validation comparisons

To measure the consistency of each normalization method at finding true DMPs, we compare results obtained on a discovery-validation split of a large dataset. Comparing results between two different subsets of a large dataset is an established idea and have been applied to the context of 450k normalization [Wu et al., 2013, Teschendorff et al., 2013]. We extend this basic idea in a novel way by introducing an *in silico* confounding of treatment (case/control status) by batch effects as follows. In a first step, we selected a set of samples to be the discovery cohort, by choosing samples where the treatment variable is not visibly confounded by plate effects. Then the validation step is achieved by selecting

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

samples demonstrating strong potential for treatment confounding by batch, for example by choose sample from different plates (see Data descriptions). The extent to which it is possible to introduce such a confounding is dataset dependent. In contrast to earlier work Wu et al. [2013], we normalize the discovery and the validation cohort separate, to more realistic mimic an independent replication experiment. The idea of creating *in silico* confounding between batch and treatment has been previously explored in the context of genomic prediction [Parker and Leek, 2012].

We quantify the agreement between validation and discovery in two ways: by an ROC curve and a concordance curve.

For the ROC curve we use the discovery cohort as the gold-standard. Because the validation cohort is affected by a batch effect, a normalization method that is robust to batch effects will show better performance in the ROC curve. Making this ROC curve requires us to choose a set of DMPs for the discovery cohort. The advantage of the ROC curve is that the plot displays immediately interpretable quantities such as specificity and sensitivity.

For the concordance curve we compare the top k DMPs from the discovery and the validation set each, and display the percentage of the overlap for each k. These curves do not require us to select a set of DMPs for the discovery cohort. Note that these curves have been previously used in the context of multiple-laboratory comparison of microarray data [Irizarry et al., 2005].

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

Altuna Akalin, Francine E Garrett-Bakelman, Matthias Kormaksson, Jennifer Busuttil, Lu Zhang, Irina Khrebtukova, Thomas A Milne, Yongsheng Huang, Debabrata Biswas, Jay L Hess, C David Allis, Robert G Roeder, Peter J M Valk, Bob Löwenberg, Ruud Delwel, Hugo F Fernandez, Elisabeth Paietta, Martin S Tallman, Gary P Schroth, Christopher E Mason, Ari Melnick, and Maria E Figueroa. Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. *PLoS Genetics*, 8(6):e1002781, 2012. doi:10.1371/journal.pgen.1002781.

Martin J Aryee, Zhijin Wu, Christine Ladd-Acosta, Brian Herb, Andrew P Feinberg, Srinivasan Yegnasubramanian, and Rafael A Irizarry. Accurate genome-scale percentage DNA methylation estimates from microarray data. *Biostatistics*, 12(2):197–210, 2011. doi:10.1093/biostatistics/kxq055.

Martin J. Aryee, Andrew E. Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P. Feinberg, Kasper D. Hansen, and Rafael A. Irizarry. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays. *Bioinformatics*, 2014. doi:10.1093/bioinformatics/btu049. In press.

J M Bennett, D Catovsky, M T Daniel, G Flandrin, D A Galton, H R Gralnick, and C Sultan. Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. *British Journal of Haematology*, 33(4):451–8, 1976. doi:10.1111/j.1365-2141.1976.tb03563.x.

Benjamin P Berman, Daniel J Weisenberger, Joseph F Aman, Toshinori Hinoue, Zachary Ramjan, Yaping Liu, Houtan Noushmehr, Christopher P E Lange, Cornelis M van Dijk, Rob A E M Tollenaar, David Van Den Berg, and Peter W Laird. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nature Genetics*, 44(1):40–46, 2012. doi:10.1038/ng.969.

Marina Bibikova, Bret Barnes, Chan Tsan, Vincent Ho, Brandy Klotzle, Jennie M Le, David Delano, Lu Zhang, Gary P Schroth, Kevin L Gunderson, Jian-Bing Fan, and Richard Shen. High density dna methylation array with single cpg site resolution. *Genomics*, 98 (4):288–95, 2011. doi:10.1016/j.ygeno.2011.07.007.

Yi-an Chen, Mathieu Lemire, Sanaa Choufani, Darci T Butcher, Daria Grafodatskaya, Brent W Zanke, Steven Gallinger, Thomas J Hudson, and Rosanna Weksberg. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*, 8(2):203–209, 2013. doi:10.4161/epi.23470.

M Cotterchio, G McKeown-Eyssen, H Sutherland, G Buchan, M Aronson, A M Easson, J Macey, E Holowaty, and S Gallinger. Ontario familial colon cancer registry: methods and first-year response rates. *Chronic diseases in Canada*, 21(2):81–86, 2000.

Sarah Dedeurwaerder, Matthieu Defrance, Martin Bizet, Emilie Calonne, Gianluca Bon-

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

tempi, and François Fuks. A comprehensive overview of Infinium HumanMethylation450 data processing. *Briefings in Bioinformatics*, 2013. doi:10.1093/bib/bbt054.

Maria E Figueroa, Sanne Lugthart, Yushan Li, Claudia Erpelinck-Verschueren, Xutao Deng, Paul J Christos, Elizabeth Schifano, James Booth, Wim van Putten, Lucy Skrabanek, Fabien Campagne, Madhu Mazumdar, John M Greally, Peter J M Valk, Bob Löwenberg, Ruud Delwel, and Ari Melnick. DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer Cell*, 17(1):13–27, 2010. doi:10.1016/j.ccr.2009.11.020.

J A Gagnon-Bartsch and T P Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012. doi:10.1093/biostatistics/kxr034.

Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004. doi:10.1186/gb-2004-5-10-r80.

Kasper D Hansen, Sarven Sabunciyan, Ben Langmead, Noemi Nagy, Rebecca Curley, Georg Klein, Eva Klein, Daniel Salamon, and Andrew P Feinberg. Large-scale hypomethylated blocks associated with Epstein-Barr virus-induced B-cell immortalization. *Genome Research*, 24(2):177–184, 2014. doi:10.1101/gr.157743.113.

Kasper Daniel Hansen, Winston Timp, Héctor Corrada Bravo, Sarven Sabunciyan, Benjamin Langmead, Oliver G McDonald, Bo Wen, Hao Wu, Yun Liu, Dinh Diep, Eirikur Briem, Kun Zhang, Rafael A Irizarry, and Andrew P Feinberg. Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, 43(8):768–775, 2011. doi:10.1038/ng.865.

Kristin N Harper, Brandilyn A Peters, and Mary V Gamble. Batch effects and pathway analysis: two potential perils in cancer studies involving dna methylation array analysis. *Cancer Epidemiol Biomarkers Prev*, 22(6):1052–60, 2013. doi:10.1158/1055-9965.EPI-13-0114.

GenomeStudio Methylation Module v1.8 User Guide. Illumina, 2010.

Rafael A Irizarry, Benjamin M Bolstad, Francois Collin, Leslie M Cope, Bridget Hobbs, and Terence P Speed. Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31(4):e15, 2003. doi:10.1093/nar/gng015.

Rafael A Irizarry, Daniel Warren, Forrest Spencer, Irene F Kim, Shyam Biswal, Bryan C Frank, Edward Gabrielson, Joe G N Garcia, Joel Geoghegan, Gregory Germino, Constance Griffin, Sara C Hilmer, Eric Hoffman, Anne E Jedlicka, Ernest Kawasaki, Francisco Martínez-Murillo, Laura Morsberger, Hannah Lee, David Petersen, John Quack-

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

enbush, Alan Scott, Michael Wilson, Yanqin Yang, Shui Qing Ye, and Wayne Yu. Multiple-laboratory comparison of microarray platforms. *Nature Methods*, 2(5):345–50, 2005. doi:10.1038/nmeth756.

Rafael A Irizarry, Christine Ladd-Acosta, Benilton Carvalho, Hao Wu, Sheri A Brandenburg, Jeffrey A Jeddeloh, Bo Wen, and Andrew P Feinberg. Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Research*, 18(5):780–790, 2008. doi:10.1101/gr.7301508.

Andrew E Jaffe and Rafael A Irizarry. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*, 15(2):R31, 2014. doi:10.1186/gb-2014-15-2-r31.

Andrew E Jaffe, Peter Murakami, Hwajin Lee, Jeffrey T Leek, M Daniele Fallin, Andrew P Feinberg, and Rafael A Irizarry. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology*, 41(1): 200–209, 2012. doi:10.1093/ije/dyr238.

W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007. doi:10.1093/biostatistics/kxj037.

Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):1724–1735, 2007. doi:10.1371/journal.pgen.0030161.

Jeffrey T Leek and John D Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723, 2008. doi:10.1073/pnas.0808709105.

Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010. doi:10.1038/nrg2825.

Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 2012. doi:10.1093/bioinformatics/bts034.

Yun Liu, Martin J Aryee, Leonid Padyukov, M Daniele Fallin, Espen Hesselberg, Arni Runarsson, Lovisa Reinius, Nathalie Acevedo, Margaret Taub, Marcus Ronninger, Klementy Shchetynsky, Annika Scheynius, Juha Kere, Lars Alfredsson, Lars Klareskog, Tomas J Ekström, and Andrew P Feinberg. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology*, 31(2):142–147, 2013. doi:10.1038/nbt.2487.

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Jovana Maksimovic, Lavinia Gordon, and Alicia Oshlack. SWAN: Subset quantile Within-Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biology*, 13(6):R44, 2012. doi:10.1186/gb-2012-13-6-r44.

Hilary S Parker and Jeffrey T Leek. The practical effect of batch on genomic prediction. *Statistical Applications in Genetics and Molecular Biology*, 11(3):Article 10, 2012. doi:10.1515/1544-6115.1766.

Ruth Pidsley, Chloe C Y Wong, Manuela Volta, Katie Lunnon, Jonathan Mill, and Leonard C Schalkwyk. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*, 14(1):293, 2013. doi:10.1186/1471-2164-14-293.

Vardhman K Rakyan, Thomas A Down, David J Balding, and Stephan Beck. Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics*, 12(8): 529–541, 2011. doi:10.1038/nrg3000.

Philip T Reiss, Lei Huang, and Maarten Mennes. Fast function-on-scalar regression with penalized basis expansions. *Int J Biostat*, 6(1):Article 28, 2010.

Mike L Smith, Keith A Baggerly, Henrik Bengtsson, Matthew E Ritchie, and Kasper D Hansen. illuminaio: An open source IDAT parsing tool for Illumina microarrays. *F1000Research*, 2(264), 2013. doi:10.12688/f1000research.2-264.v1.

Tamar Sofer, Elizabeth D Schifano, Jane A Hoppin, Lifang Hou, and Andrea A Baccarelli. A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure. *Bioinformatics*, 29(22):2884–2891, 2013. doi:10.1093/bioinformatics/btt498.

A E Teschendorff, F Marabita, M Lechner, T Bartlett, J Tegner, D Gomez-Cabrero, and S Beck. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, 29(2):189–196, 2013. doi:10.1093/bioinformatics/bts680.

Nizar Touleimat and Jörg Tost. Complete pipeline for Infinium Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics*, 4(3):325–341, 2012. doi:10.2217/epi.12.21.

Timothy J Triche, Daniel J Weisenberger, David Van Den Berg, Peter W Laird, and Kimberly D Siegmund. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Research*, 41(7):e90, 2013. doi:10.1093/nar/gkt090.

Michael C Wu, Bonnie R Joubert, Pei-Fen Kuan, Siri E Håberg, Wenche Nystad, Shyamal D Peddada, and Stephanie J London. A systematic assessment of normalization approaches for the infinium 450k methylation platform. *Epigenetics*, 9(2), 2013. doi:10.4161/epi.27119.

Zhijin Wu and Martin J Aryee. Subset quantile normalization using nega-

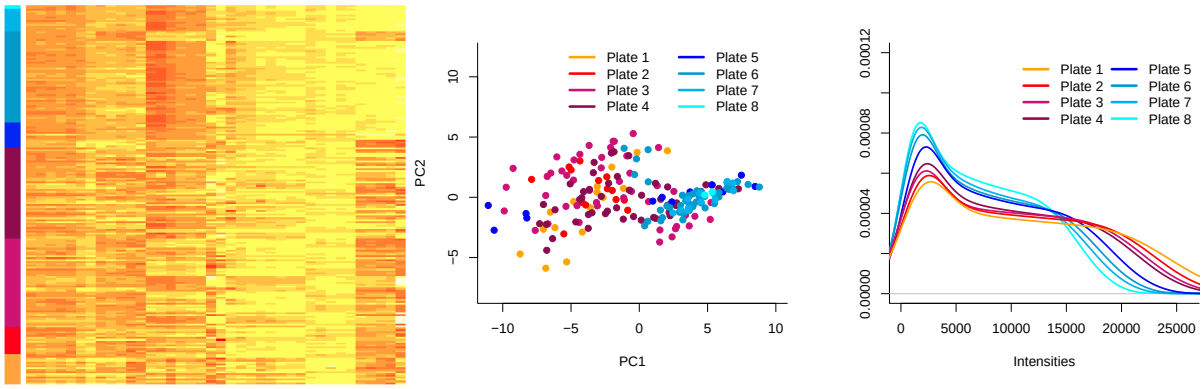
bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

tive control features. *Journal of Computational Biology*, 17(10):1385–1395, 2010. doi:10.1089/cmb.2010.0049.

Brent W Zanke, Celia M T Greenwood, Jagadish Rangrej, Rafal Kustra, Albert Tenesa, Susan M Farrington, James Prendergast, Sylviane Olschwang, Theodore Chiang, Edgar Crowdy, Vincent Ferretti, Philippe Laflamme, Saravanan Sundararajan, Stéphanie Roumy, Jean-François Olivier, Frédérick Robidoux, Robert Sladek, Alexandre Montpetit, Peter Campbell, Stephane Bezieau, Anne Marie O'Shea, George Zogopoulos, Michelle Cotterchio, Polly Newcomb, John McLaughlin, Ban Younghusband, Roger Green, Jane Green, Mary E M Porteous, Harry Campbell, Helene Blanche, Mourad Sahbatou, Emmanuel Tubacher, Catherine Bonaiti-Pellié, Bruno Buecher, Elio Riboli, Sebastien Kury, Stephen J Chanock, John Potter, Gilles Thomas, Steven Gallinger, Thomas J Hudson, and Malcolm G Dunlop. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nature Genetics*, 39(8): 989–994, 2007. doi:10.1038/ng2089.

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figures



(a) **(b)** **(c)**

Figure 1. Control probes acts as surrogates for batch effects. (a) A heatmap of a summary (see Methods) of the control probes, with samples on the y-axis and control summaries on the x-axis. Samples were processed on a number of different plates indicated by the color label. Only columns have been clustered. (b) The first two principal components of the matrix depicted in (a). Samples partially cluster according to batch, with some batches showing tight clusters and other being more diffuse. (c) The distribution of methylated intensities averaged by plate. These three panels suggests that the control probe summaries partially measure batch effects.

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

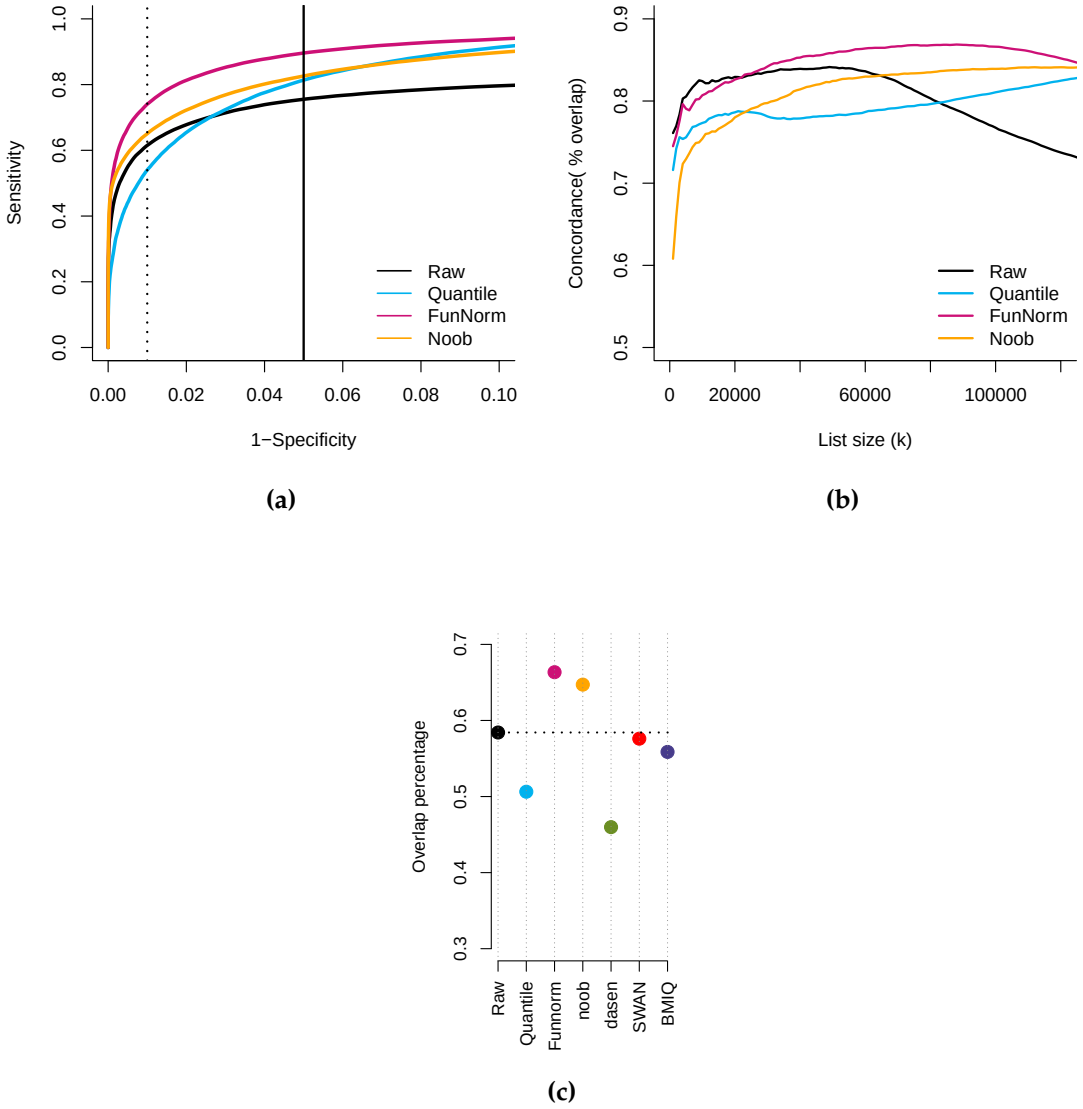
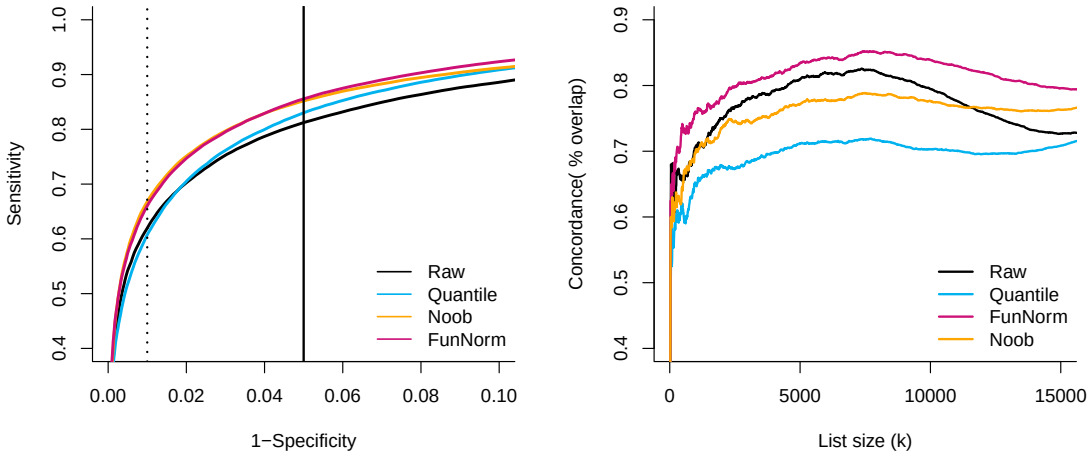


Figure 2. Improvements in replication for the EBV dataset. (a) ROC curves for replication between a discovery and a validation dataset. The validation dataset was constructed to show *in silico* batch effects. The dotted and solid lines represent respectively the commonly used false discovery rate cutoffs of 0.01 and 0.05. (b) Concordance curves showing the percent overlap between the top k DMPs in the discovery and validation cohort. Additional normalization methods assessed in Supplementary Figure S3. Functional normalization shows a high degree of concordance between datasets. (c) The percentage of top 100,000 DMPs which are replicated between the discovery and validation cohort and also inside a differentially methylated block or region from Hansen et al. [2014].

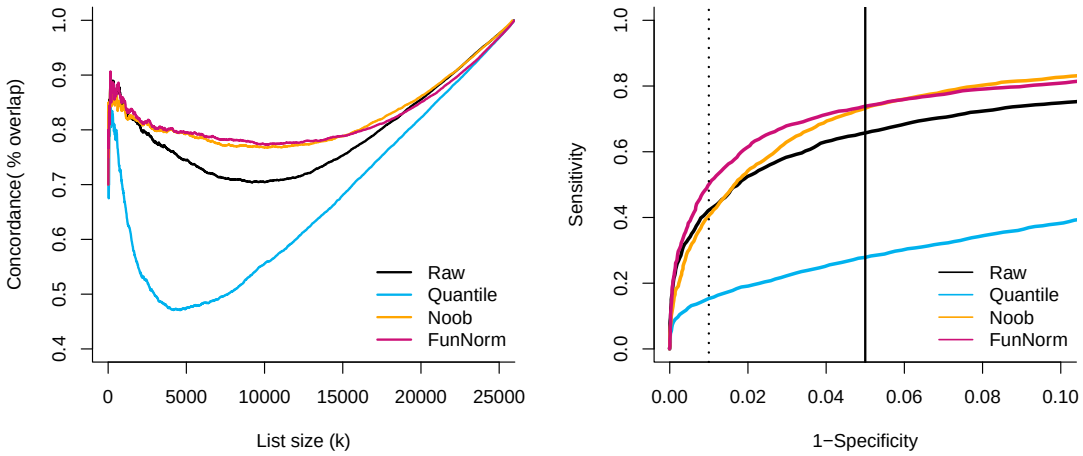
bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a) **(b)**

Figure 3. Improvements in replication for the TCGA KIRC dataset. (a) ROC curves for replication between a discovery and a validation dataset. The validation dataset was constructed to show *in silico* batch effects. (b) Concordance plots between an additional cohort assayed on the 27k array and the validation dataset. Additional normalization methods assessed in Supplementary Figure S4. Functional normalization shows a high degree of concordance between datasets.

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a) **(b)**

Figure 4. Improvements in replication of tumor subtype heterogeneity. In the AML dataset from TCGA, the same samples have been assayed on 450k and 27k arrays. (a) Concordance plots between results from the 450k array and the 27k array. (b) ROC curves for the 450k data, using the results from the 27k data as gold standard.

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

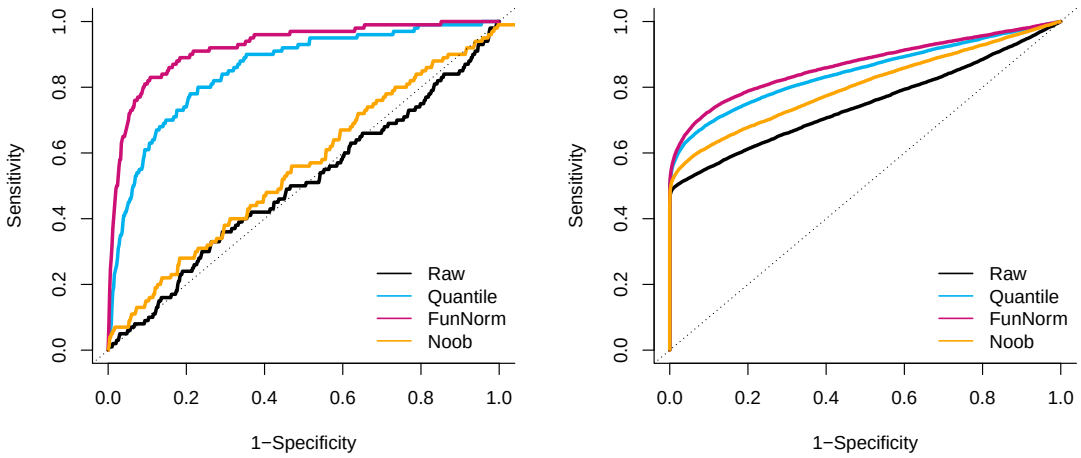


Figure 5. Performance Improvements on blood samples dataset. (a) An ROC curve for replication of case-control differences between blood samples from colon cancer patients and blood samples from normal individuals, the Ontario-Blood dataset. The validation dataset was constructed to show an *in silico* batch effect. (b) An ROC curve for identification of probes on the sex chromosomes for the Ontario-Sex dataset. Sex is confounded by an *in silico* batch effect. Both evaluations show good performance of functional normalization.

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

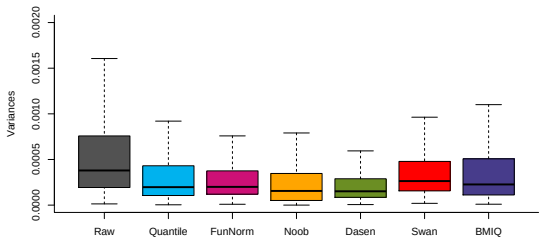


Figure 6. Variance across technical triplicates. Boxplots of the probe-specific variances estimated across 19 individuals assayed in technical triplicates. All normalization methods improve upon Raw data, and Funnorm performs well.

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

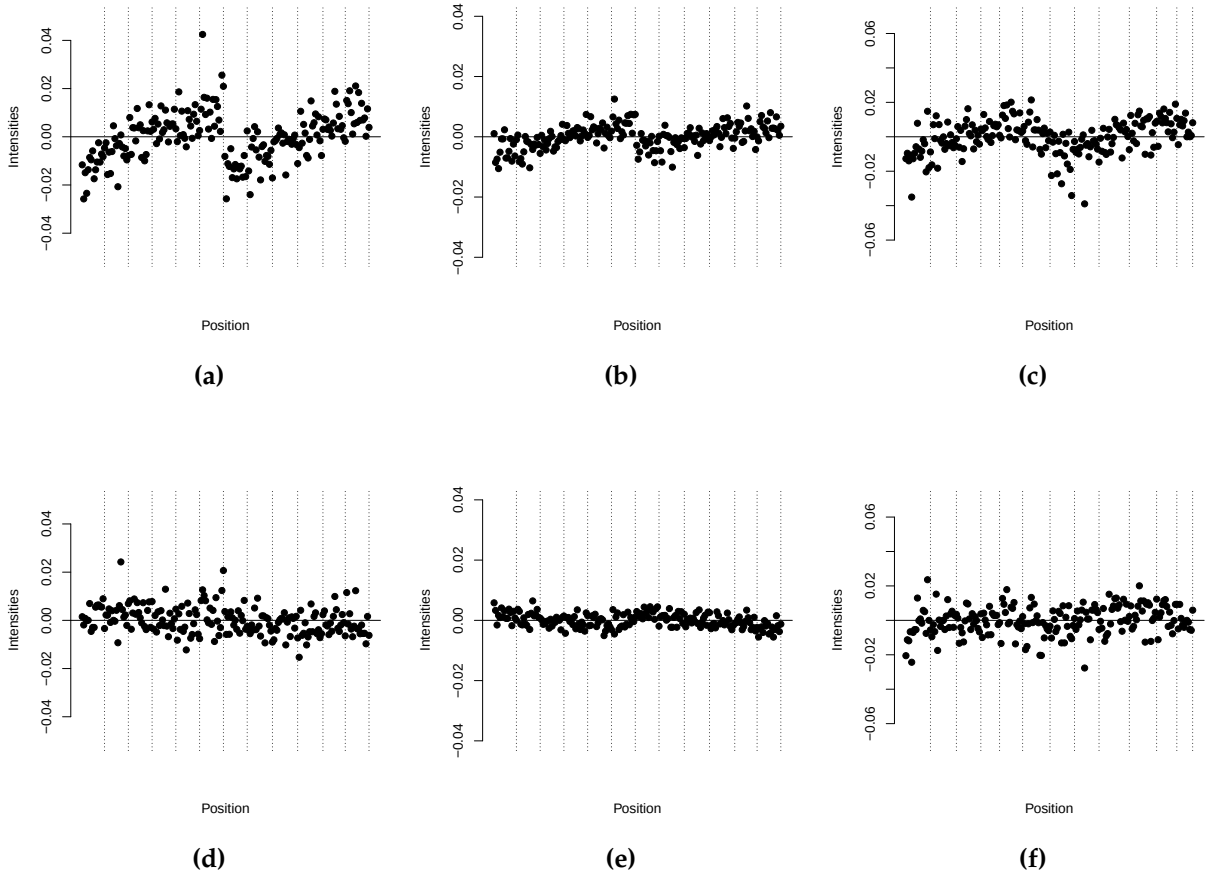


Figure 7. Spatial location affects overall methylation. Quantiles of the beta distributions adjusted for a slide effect. The 12 vertical stripes are ordered as rows 1-6 in column 1 followed by rows 1-6 in column 2. (a) 10th percentile for Type II probes for the unnormalized AML dataset. (b) 15th percentile for Type I probes for the unnormalized AML dataset. (c) 85th percentile for Type II probes for the unnormalized Ontario-Lympho dataset. Panel (a-c) show that the top of the slide has a different beta distribution from the bottom. (d-f) Like (a-c) but after functional normalization, which corrects this spatial artifact.

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

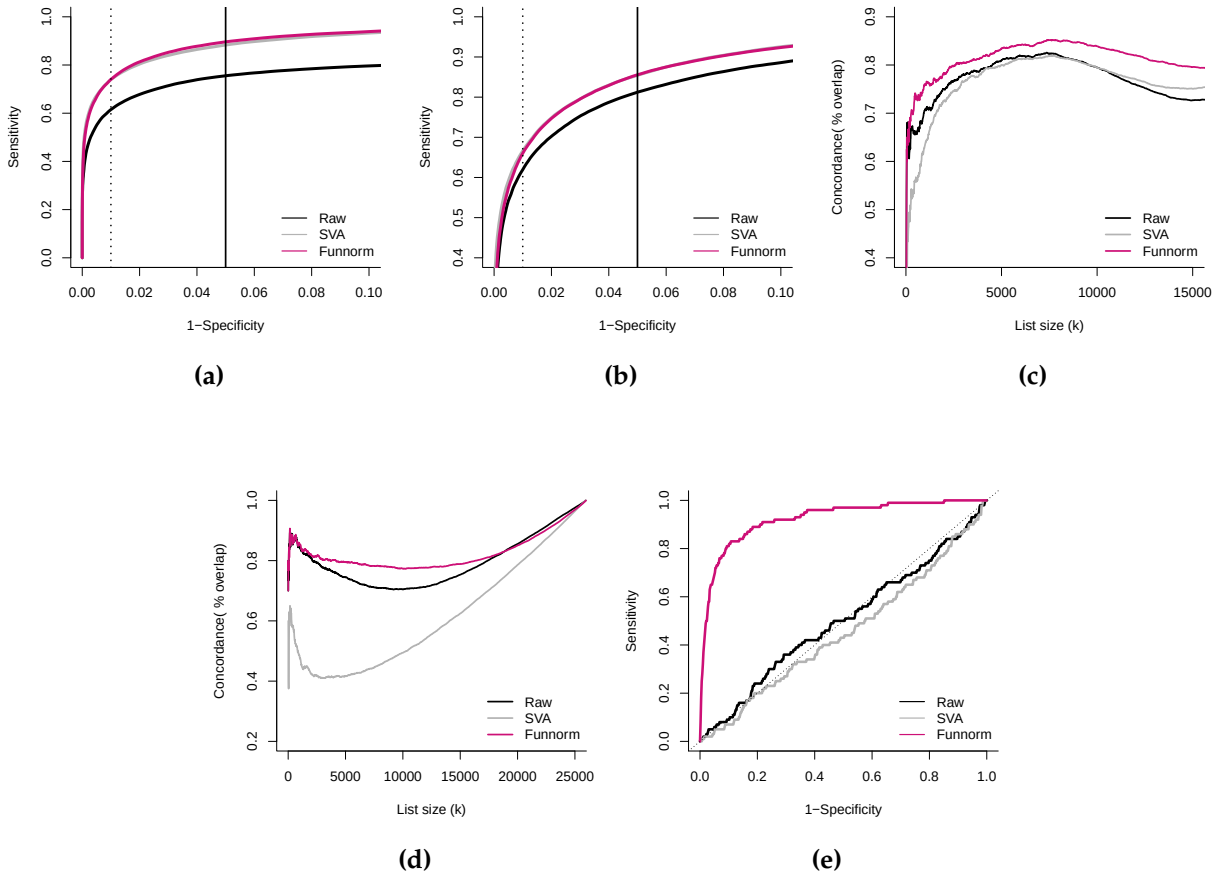
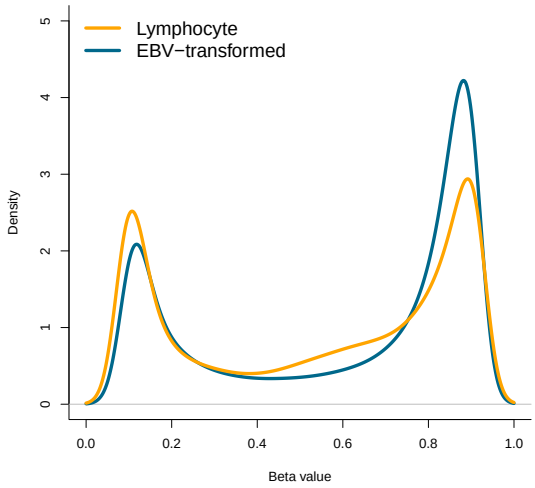


Figure 8. Comparison to batch effect removal tool SVA (a) Like Figure 2a; an ROC curve for the Ontario-EBV dataset. (b) Like Figure 3a, an ROC curve for the TCGA KIRC dataset. (c) Like Figure 3b, a concordance curve between the validation cohort from 450k data and the 27k data for TCGA KIRC dataset. (d) Like Figure 4a, a concordance plots between results from the 450k array and the 27k array for TCGA AML dataset. (e) Like Figure 5a, an ROC curve for Ontario-Blood dataset.

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

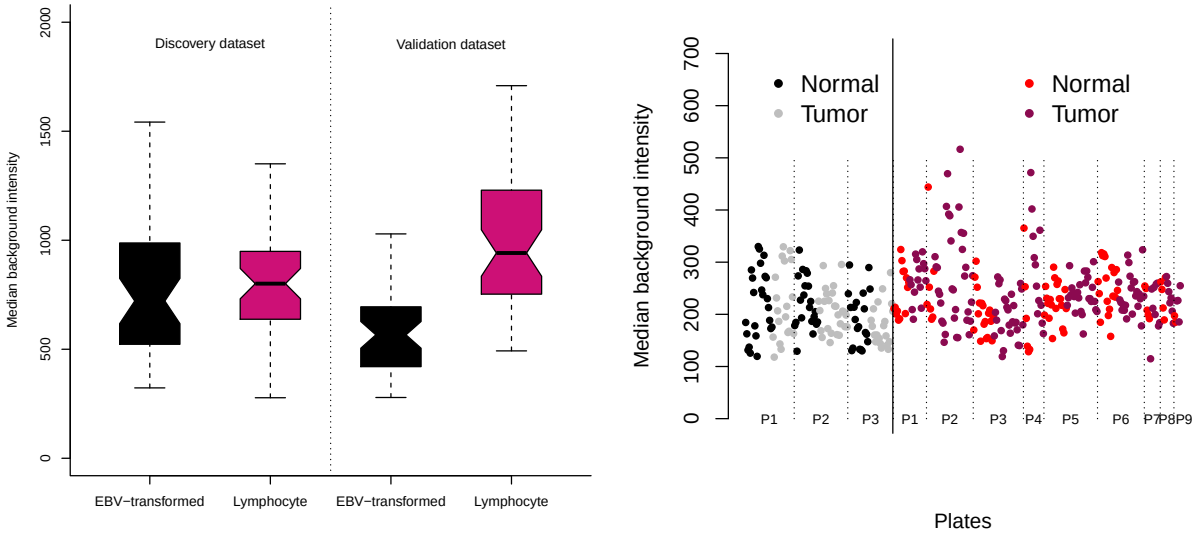
Supplementary Figures



(a)

Supplementary Figure S1. Global changes in DNA methylation for the EBV dataset. (a) The average density of unnormalized beta values across both EBV transformed lymphocytes and normal lymphocytes, showing global hypomethylation caused by EBV transformation.

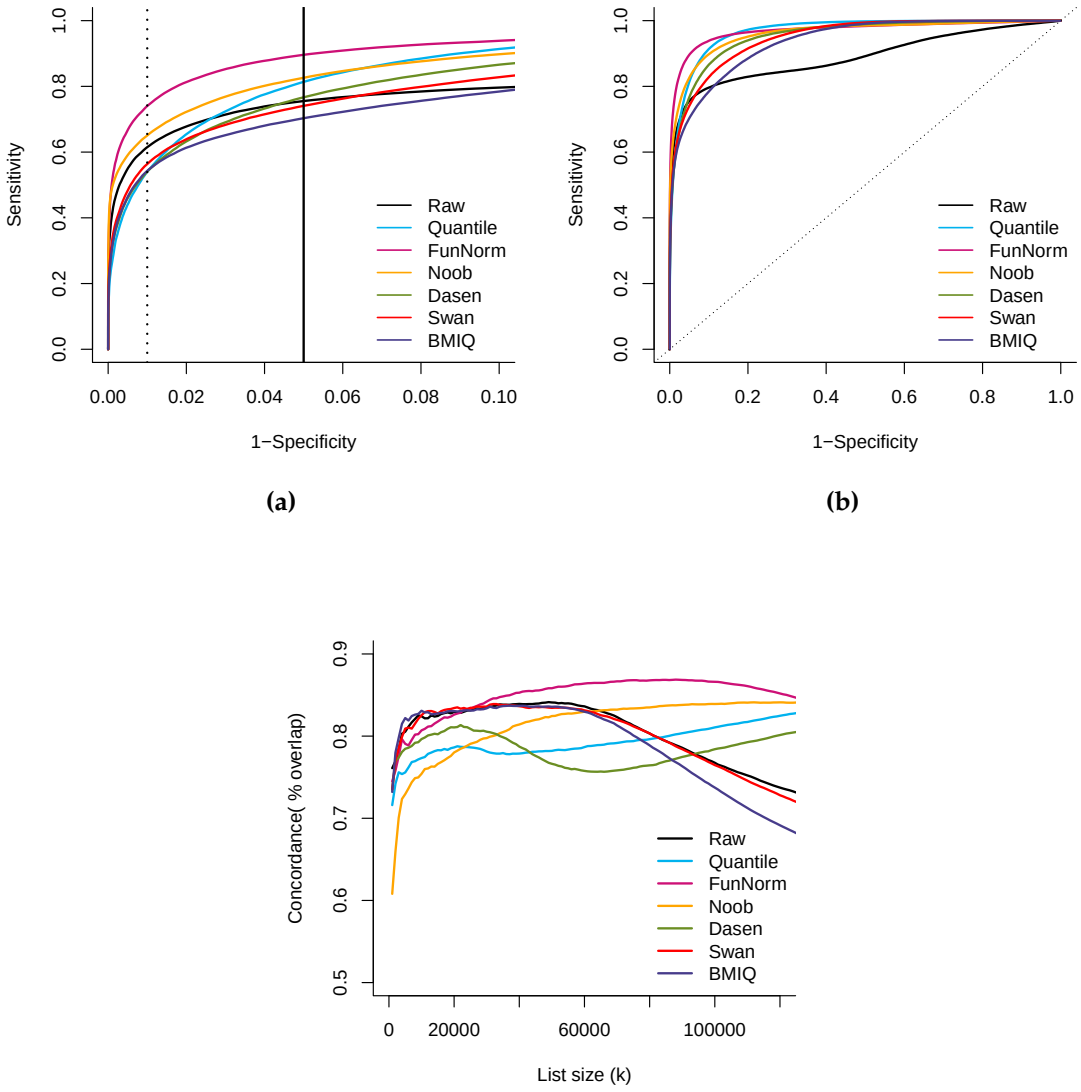
bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a) **(b)**

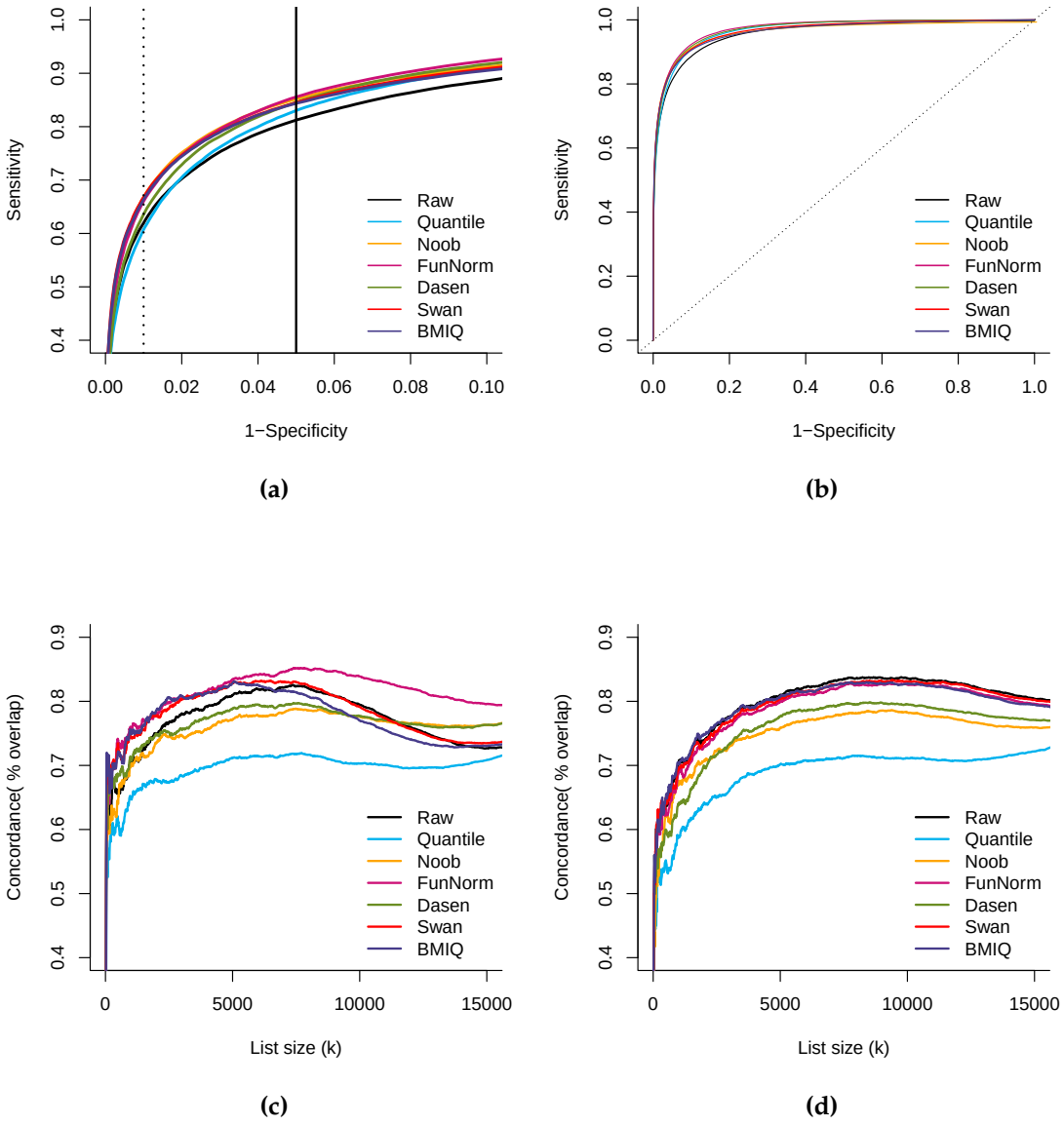
Supplementary Figure S2. Illustration of *in silico* batch effects. (a) Distribution of background intensity for the Ontario-EBV dataset, showing an *in silico* introduced difference in the validation dataset. (b) Distribution of background intensity for the TCGA KIRC dataset, showing an *in silico* introduced difference in variation in the validation dataset.

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



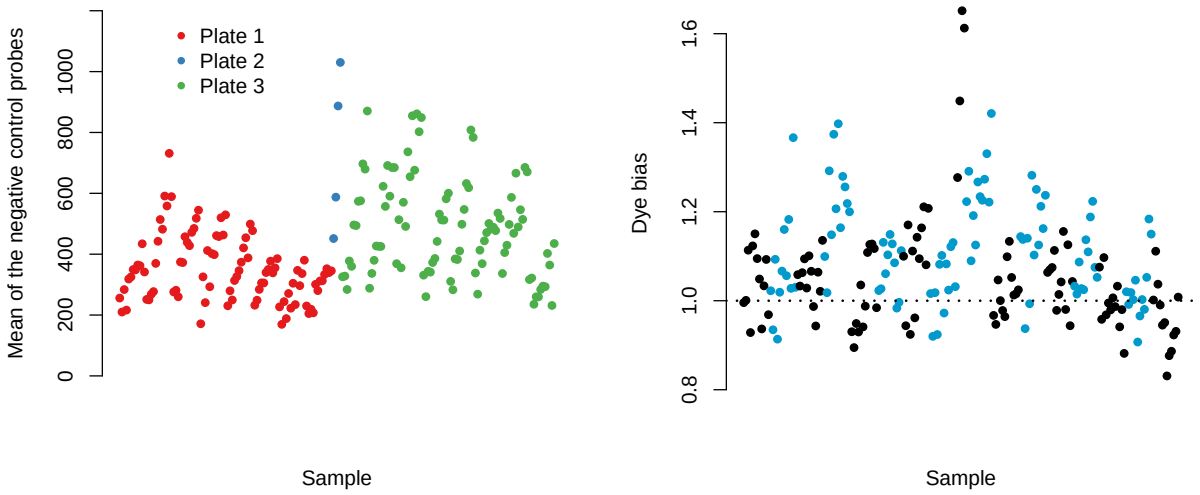
Supplementary Figure S3. Improvements in replication for the EBV dataset, all methods. (a) Like Figure 2b, but for all examined normalization methods. (b) Like (a) but for the full range of specificity. (c) Like Figure 2c but for all examined normalization methods.

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Supplementary Figure S4. Improvements in replication for the TCGA KIRC dataset, all methods. (a) Like Figure 3a but for all normalization methods we assess. (b) Like (a) but for the full range of specificity. (c) Like Figure 3b but for all normalization methods we assess. (d) Like (c) but comparing to the discovery dataset instead of the validation dataset; the discovery dataset is less affected by unwanted variation.

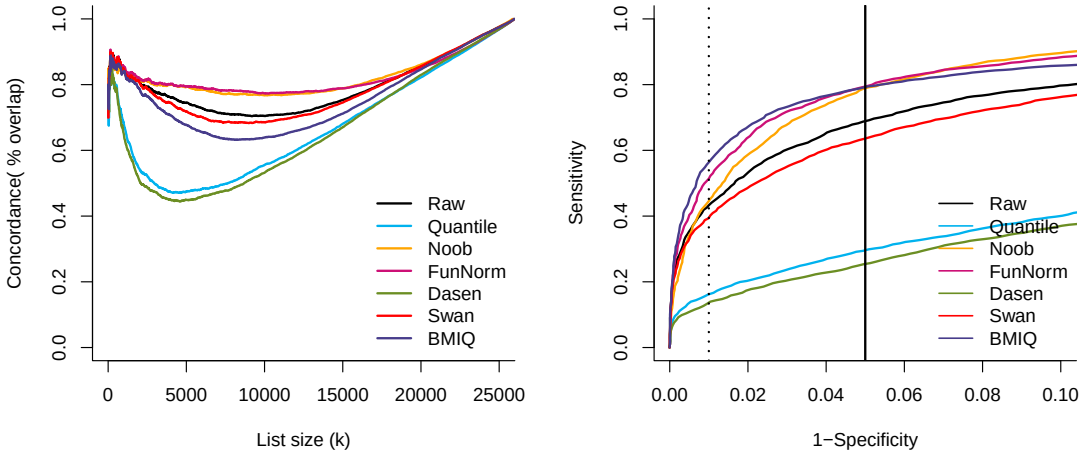
bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a) **(b)**

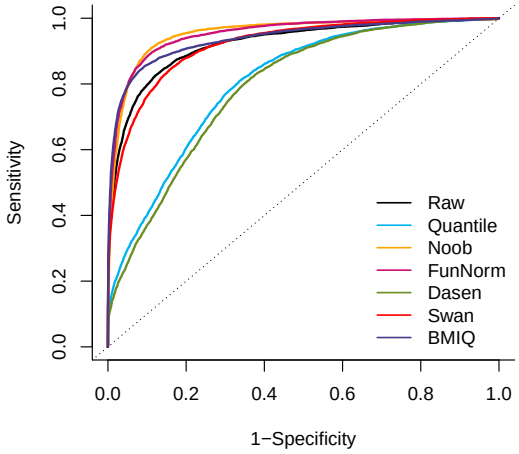
Supplementary Figure S5. Plate effects and dye bias for the AML dataset (a) The means of the negative control probes is correlated with the processing plate (96 samples) indicating that background intensity is affected by batch. (b) We measure dye bias by taking the ratio of the negative control probes in the green channel and the negative control probes in the red channel. A value of 1 means that there is no dye-bias. We plot the dye bias (y-axis) for samples ordered by plate and then by slide (x-axis). The plate order is the same as (a). We use two different alternating colors to differentiate the slides. We observe that dye bias is orthogonal to plate effect and highly slide dependent.

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a)

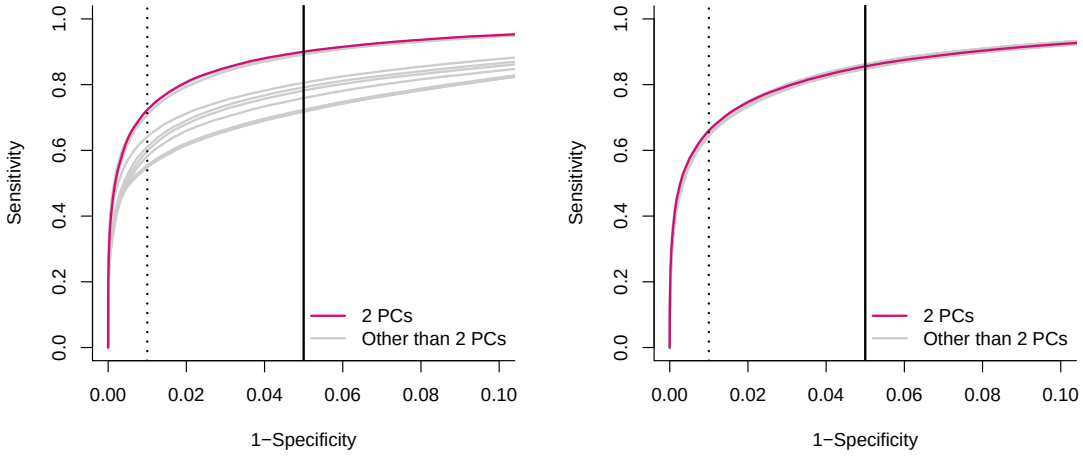
(b)



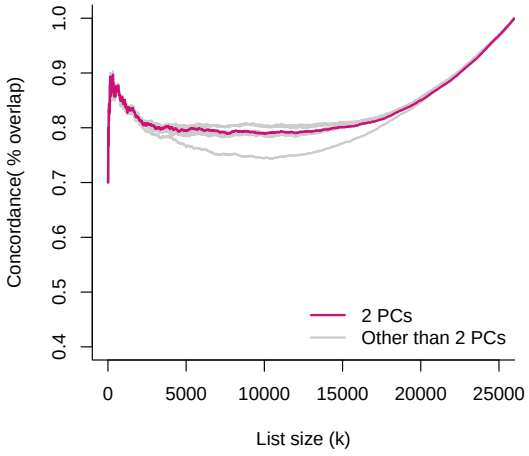
(c)

Supplementary Figure S6. Improvements in replication of tumor subtype heterogeneity. (a) Like Figure 4a (b) Like Figure 4b but for all normalization methods. (c) Like (b) but for the full range of specificity

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



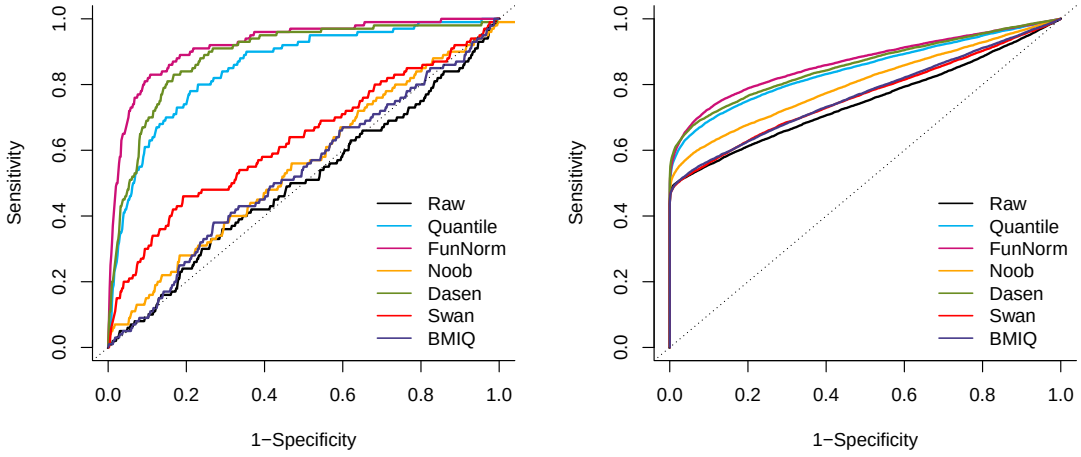
(a) **(b)**



(c)

Supplementary Figure S7. The impact of the number of principal components. (a) Like Figure 2a, (b) Like Figure 3a (c) Like Figure 4a, but showing the difference between using $m = 2$ components and other choices of $m \leq 10$.

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a) **(b)**

Supplementary Figure S8. Improvements in blood samples, all methods Like Figure 5, but for all normalization methods we assess.

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary material

In the following text we describe exhaustively the transformations that we applied to the control probes to create the control probe summaries that were used throughout the paper as covariates in the functional normalization. Note that we chose the transformations by considering the recommendations made in the GenomeStudio Methylation manual [Ill, 2010].

- For "Bisulfite Conversion I" probes, 3 probes (C1,C2,C3) are expected to have high signal in the green channel in case the bisulfite conversion reaction was successful, and similarly 3 additional probes (C4,C5,C6) are expected to have high signal in the red channel. We therefore consider these 6 intensities and take the mean as a single summary value.
- For "Bisulfite Conversion II" probes, 4 probes are expected to have high intensities in the red channel in case the bisulfite conversion reaction was successful. Therefore we consider the mean of these 4 intensities as a single summary value.
- For the "Extension" control probes, 2 probes must be monitored in the red channel (A,T) and 2 probes must be monitored in the green channel (C,G). We consider the 4 raw intensities as output values for a total of 4 summary values.
- For the "Hybridization" probes, the 3 probes have to be monitored in the green channel. We consider the raw intensities as output values, corresponding to low, medium and high hybridization signals, for a total of 3 summary values.
- For the "Staining" probes, we select the green intensity of the probe that is expected to have high intensity in the green channel, and similarly for the probe that is expected to have high intensity in the red channel. This results in 2 summary values.
- For "Non-polymorphic" controls, we consider the 2 probes that are expected to be high in the green channel (C and G) and the two probes that are expected to be high in the red channel (A and T). We consider the 4 raw intensities as output for a total of 4 summary values.
- "Target removal" probes have to be monitored only in the green channel. We use the raw intensities for the 2 probes as output values for a total of 2 summary values.
- For "Specificity II" probes, we monitor the 3 probes in both the green and red channels, for a total of 6 output values. The green channel is expected to be low and the red channel to be high, so we also form the ration of the mean of the green channel to the mean of the red channel as output value. This results in 7 summary values.
- For "Specificity I" probes, 3 probes are expected to have high signal in the green channel, and 3 different probes are expected to have high signal in the red channel. For each trio, we take the raw intensities in the corresponding "good" channel, giving us 6 output values. For each trio, we also consider the signal-to-noise ratio

bioRxiv preprint doi: https://doi.org/10.1101/002956; this version posted February 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

by taking the ratio of the mean of the 3 intensities measured in the good channel (high signal expected) and the mean of the three intensities in the opposite channel (low signal expected). This gives us 2 ratio summaries. We also consider the mean of these two ratios, giving us 1 additional output value. This results in a total of 9 summary values.

- "Normalization" probes: probes targeting A bases (32) and T bases (61) have to be monitored in the red channel and probes targeting G (32) and C (61) have to be monitored in the green channel. For each type (A,C,T,G), we consider the mean of the intensities in their corresponding channel, that we denote by $normA$, $normC$, $normT$ and $normG$ (4 output values). Moreover, we consider the ratio $(normC + normG)/(normA + normT)$ as a surrogate for dye bias computed with positive controls (1 additional output value), for a total of 5 summary values.
- For the Out-of-band probes (Oob), we first take the 1st, 50th and 99th percentiles of the 92,596 green intensities (3 output values). Because the variation seen in the green Oob probes is similar to the that of the red Oob probes, we omit the latter. Nevertheless, we consider the ratio of the median of the 92,596 green intensities and the median of the 178,406 red intensities, as a surrogate for dye bias. This results in a total of 4 summary values.