Normalization benchmark of ATAC-seq datasets shows the importance of accounting for GC-content effects

Koen Van den Berge[1,2,3], Hsin-Jung Chou[4,+], Hector Roux de Bézieux[5,6], Kelly Street[7,8], Davide Risso[9], John Ngai[4,10,†], Sandrine Dudoit[1,5,6,*]

[1] Department of Statistics, University of California, Berkeley, CA, USA

[2] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

[3] Bioinformatics Institute Ghent, Ghent University, Ghent, Belgium

[4] Department of Molecular and Cell Biology, University of California, Berkeley, CA, USA

[5] Division of Biostatistics, School of Public Health, University of California, Berkeley, CA, USA

[6] Center for Computational Biology, University of California, Berkeley, CA, USA

[7] Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA

[8] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[9] Department of Statistical Sciences, University of Padova, Italy

[10] Helen Wills Neuroscience Institute, University of California, Berkeley, CA, USA

+ Current address: Audentes Therapeutics Inc, San Francisco, CA, USA

† Current address: National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA

* To whom correspondence should be addressed: sandrine@stat.berkeley.edu

January 26, 2021

Abstract

Modern assays have enabled high-throughput studies of epigenetic regulation of gene expression using DNA sequencing. In particular, the assay for transposase-accessible chromatin using sequencing (ATAC-seq) allows the study of chromatin configuration for an entire genome. Despite the gain in popularity of the assay, there have been limited studies investigating the analytical challenges related to ATAC-seq data, and most studies leverage tools developed for bulk transcriptome sequencing (RNA-seq). Here, we show that GC-content effects are omnipresent in ATAC-seq datasets. Since the GC-content effects are sample-specific, they can bias downstream analyses such as clustering and differential accessibility analysis. We evaluate twelve different normalization procedures on eight public ATAC-seq datasets and show that no method uniformly outperforms all others. However, our work clearly shows that accounting for GC-content effects in the normalization is crucial for common downstream ATAC-seq data analyses, such as clustering and differential accessibility analysis, leading to improved accuracy and interpretability of the results. Using two case studies, we show that exploratory data analysis is essential to guide the choice of an appropriate normalization method for a given dataset.

1 Introduction

Genomic DNA is packaged into chromatin in the eukaryotic nucleus via a highly-deliberate and dynamic process. The study of chromatin configuration aids in unraveling the complex epigenetic regulation of gene expression. Chromatin accessibility, which reflects the relatively open or closed chromatin conformation, affects the ability of nuclear proteins to physically interact with chromatin DNA and hence regulate gene expression [1]. Genome-wide mapping of chromatin accessibility delineates the functional chromatin landscape corresponding to transcription start sites (TSS), transcription factor (TF) binding sites, and all classes of cis-regulatory elements (e.g., promoters and enhancers) [2,3]. ATAC-seq, a robust assay for transposase-accessible chromatin using sequencing, has been used to provide insight into chromatin accessibility with a relatively simple and time-saving protocol [1,4].

ATAC-seq relies on a hyperactive Tn5 transposase that can simultaneously cut accessible DNA fragments and ligate sequencing adapters to both strands [4]. The tagged DNA fragments are amplified, sequenced, and mapped back to the genome, upon which accessible regions are identified by the enrichment of mapped read ends, traditionally using peak-calling algorithms [5,6]. The data used for downstream analysis thus typically consist of a count matrix,

where each row corresponds to a genomic region (or "peak") and each column corresponds to a sample. The count in each cell of the matrix represents the number of read ends mapped to a particular peak for a given sample, and is a proxy for the accessibility of the genomic region.

High-throughput sequencing studies are often influenced by several factors of technical variation, e.g., sample preparation, library preparation, and sequencing batch[7,8]. Notably, GC-content, the fraction of guanine and cytosine nucleotides in a particular genomic region or gene, has previously been identified as a sample-specific technical bias factor, e.g., in peak-calling for chromatin immunoprecipitation sequencing (ChIP-seq) data[9] or normalization and differential expression for RNA-seq data[10,11]. Similarly, ATAC-seq data have been shown to be affected by technical variation due to, for example, enzymatic cleavage effects, PCR bias, and duplicate reads[12]. Importantly, the effect of GC-content in ATAC-seq data has been noted previously[13,14]. Indeed, since a large fraction of accessible regions are around gene promoters, which often have a high GC-content and are enriched in CpG islands[15], we naturally expect an association between accessibility and GC-content. However, the impact of GC-content effects in ATAC-seq data on downstream analyses has not been studied in depth.

Notwithstanding the recent surge in popularity for the assay, surprisingly few studies investigate the analytical challenges of ATAC-seq data, e.g., normalization and differential accessibility (DA) analysis. Indeed, most data analysis workflows rely on statistical methods originally developed for ChIP-seq or bulk RNA-seq data to analyze bulk ATAC-seq datasets[16–18]. Recently, Reske et al.[18] compared pipelines for DA analysis and showed that normalization has a large influence on the results. While the authors advised comparing multiple normalization methods for a particular dataset at hand, they did not elaborate on GC-content effects nor normalization methods that take this into account. In particular, while most research papers analyzing bulk ATAC-seq data adopt standard bulk RNA-seq global-scaling normalization procedures (e.g., Rizzardi et al.[16], Philip et al.[17]), such as total-count normalization, edgeR's trimmed mean of M-values (TMM)[19], or DESeq2's median-of-ratios (MOR)[20], some account for GC-content effects[14], typically through conditional-quantile normalization (cqn)[21]. Given this dichotomy in normalization choices, we investigate the influence of accounting for possible GC-content effects on downstream analyses for ATAC-seq data.

In this manuscript, we show that GC-content effects in ATAC-seq data can be sample-specific, which indicates that they can bias downstream analyses such as clustering and differential accessibility analysis. We evaluate several GC-aware as well as GC-unaware normalization methods using a principled framework[22], and also study the impact of GC-content effects on the accuracy and interpretation of differential accessibility analysis results. While no normalization method uniformly performs best across all datasets, GC-aware normalization methods typically perform better than GC-unaware methods, emphasizing the need to correct for GC-content effects. We recommend that researchers use exploratory data analysis methods to guide the choice of normalization method.

2 Results

2.1 GC-content effects are sample-specific and confound downstream analyses

In ATAC-seq, accessibility counts are often positively associated with GC-content. We explore this using data from Calderon et al.[23], who generated an ATAC-seq atlas of human immune cell types. Here, we focus on six replicate samples of memory natural killer (MNK) cells in a control condition.

Figure 1a shows that the accessibility count of a particular genomic region is associated with its GC-content. However, the slope and shape of the curves may differ between samples, which indicates that GC-content effects are sample-specific and can therefore bias between-sample comparisons. This can be similarly observed for other cell types in this dataset (Supplementary Figure 1). Note that, while the width of each peak is also associated with its accessibility, this effect tends to be more uniform across different samples and therefore have lower impact on between-sample comparisons (Supplementary Figure 2). This could be considered analogous to the effect of gene length on read counts in RNA-seq data.

One might initially think that, because DA analyses involve comparing read counts between samples for a given genomic region with a fixed GC-content, GC-content effects would cancel out. However, because of their sample-specificity, GC-content effects also impact log-fold-changes (LFC) comparing accessibility between samples for a given region. A 3 vs. 3 mock null comparison of the MNK cells (i.e., a comparison of the same type of cells which should not exhibit DA), using both edgeR[24] and DESeq2[20], reveals a bias in the LFC with respect to GC-content (Figure 1b-c). That is, the LFC are not centered around zero, as expected for a null comparison of normalized data, and also vary with GC-content. Furthermore, both TMM and MOR normalizations, which are frequently used for DA analysis in ATAC-seq data, fail to remove GC-content effects. Full-quantile (FQ) normalization[25], another popular normalization method, also fails at removing GC-content effects (Figure 1d). Similar effects can

be observed for other cell types for which six replicates are available in the same condition (Supplementary Figures 3-4).

If not accounted for, GC-content effects can have a significant impact on a downstream differential accessibility analysis, masking biological signal and also leading to false positives, as was similarly previously observed in RNA-seq data[10,11,21].

2.2 GC-aware normalization

GC-content effects have been observed and accounted for in other work on ATAC-seq data[14,26], typically using conditional-quantile normalization (cqn)[21] (see Methods). While this method removes GC-content effects in many datasets, Risso et al.[10] observed that, in the context of RNA-seq, cqn's regression approach may be "too weak" for some datasets and "more aggressive normalization procedures" may be required. They proposed a method, implemented in the Bioconductor R package EDASeq, based on two rounds of full-quantile normalization: First, FQ normalization within a sample across bins of features (e.g., peaks or genes) with similar GC-content and, subsequently, FQ normalization between samples. We will denote this method as FQ-FQ normalization.

However, read count distributions may in some datasets be more comparable across samples within a GC-content bin than across GC-content bins within a sample (see Figure 2a). This motivates a variant of FQ-FQ normalization, which we call GC-FQ, that applies full-quantile normalization across samples for each GC-content bin separately. Note that this attempts to equalize the GC-content effect across samples and therefore performs between-sample normalization simultaneously.

For the ATAC-seq dataset of Calderon et al.[23], all three GC-aware methods (cqn, GC-FQ, and FQ-FQ) indeed effectively remove GC-content effects (Figure 2b-d) on the fold-changes.

All three of these GC-aware normalization methods rely on full-quantile between-sample normalization, which is an aggressive normalization method that does not come without assumptions. Since FQ normalization forces distributions to be equal across all samples, the underlying assumption is that global differences between distributions are the result of technical effects. In other words, if there were no technical nor sampling variability in the data, the distributions of all samples should be identical, hence comparable, and this is what FQ normalization is trying to achieve.

This assumption is restrictive and is not guaranteed to hold for all datasets. Hicks et al.[27] recently developed a generalization of FQ normalization, smooth-quantile normalization, or qsmooth, that can account for global differences between distributions due to biological effects of interest. The method is based on the assumption that the read count distribution of each sample should be equal within biological groups or conditions, but could vary between groups. Essentially, the method is a weighted combination of FQ normalization between samples for each biological group separately and FQ normalization across all samples of all biological groups.

We therefore also implement a variant of GC-FQ, which we call smooth GC-FQ, that applies smooth-quantile normalization across samples within each GC-content bin separately and is therefore capable of dealing with biological groups that have global distributional differences between them.

2.3 Benchmarking ATAC-seq normalization using scone

We evaluate normalization methods using eight public bulk ATAC-seq datasets[14,17,23,28–32], including ATAC-seq atlases of mouse tissues[30], human blood cells[23], and human brain cells[29], thus spanning a multitude of biological systems. For each dataset, we use the publicly-available raw (i.e., unnormalized) accessibility count matrix; the different datasets hence also span a realistic range of preprocessing and peak-calling pipelines.

First, we determine an appropriate number of GC-content bins to use in GC-FQ normalization based on relative log expression (RLE) measures[33] (see Methods). This leads us to select 50 bins for GC-FQ normalization in all downstream analyses (Supplementary Figure 5).

The scone framework is used to evaluate and rank normalization procedures in a principled way through investigation of the normalized data in terms of clustering properties, association with factors of wanted and unwanted variation, and global distributional characteristics[22]. Note that while scone relies on a range of relevant summaries to evaluate normalization, it may be more favorable to methods that generally make samples more similar to each other, due to its use of RLE and clustering characteristics. We use peaks that overlap with housekeeping genes as negative control features and incorporate two new evaluation measures that specifically assess GC-content effects based on RLE[33] (see Methods). Where available, we use known subject (e.g., patient) effects, batch effects, and quality control (QC) measures as known factors of unwanted variation to guide the evaluation. As factors of wanted variation, we use the biological groups relevant to the experiment (e.g., brain region, tissue, treatment). We compare GC-aware normalization methods (smooth) GC-FQ, cqn, and FQ-FQ, with GC-unaware normalization methods

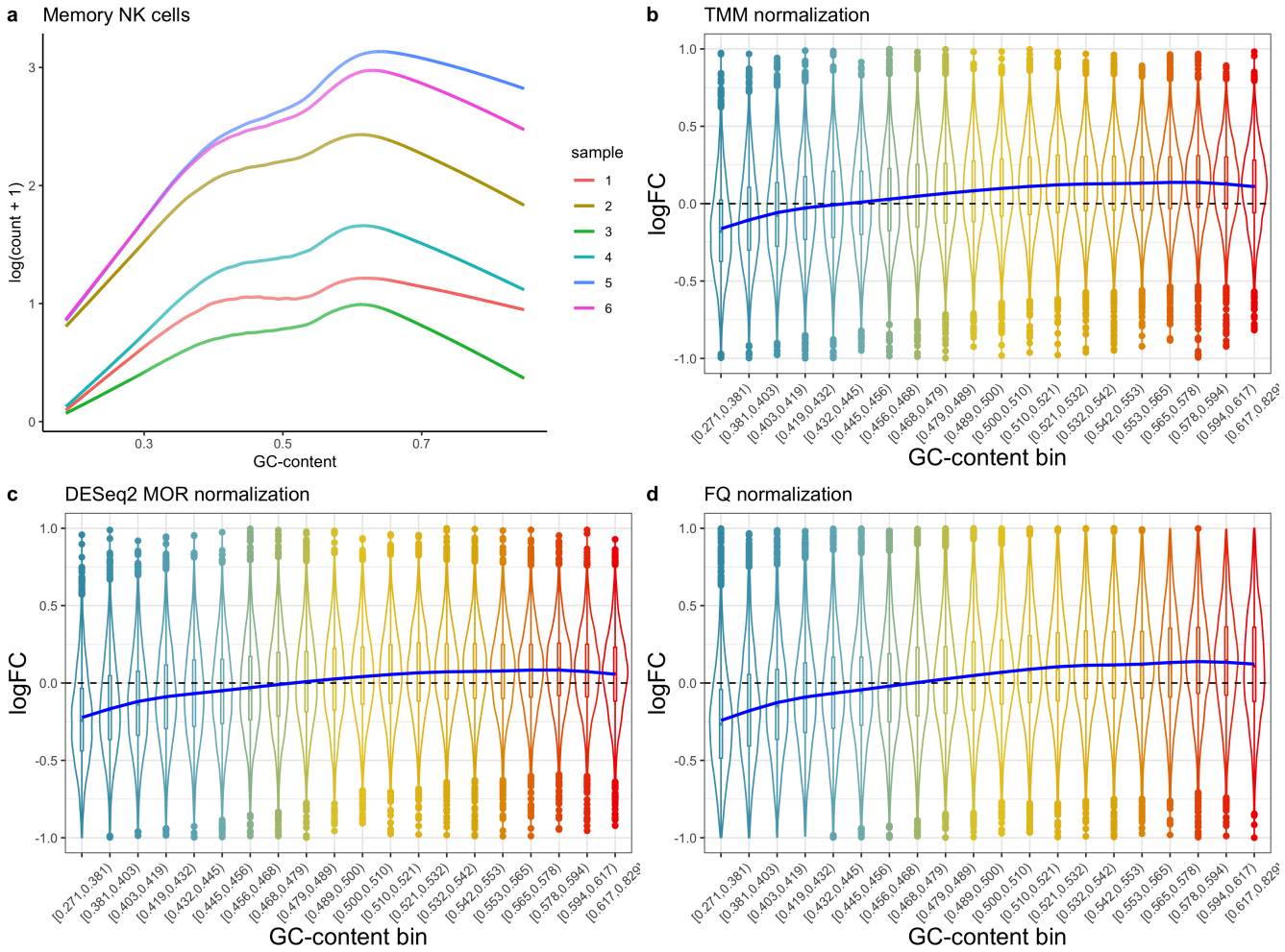


Figure 1: *GC-content effects are sample-specific and confound differential accessibility analysis.* (a) Fitted lowess curves of log-count as a function of GC-content for the six MNK cell control samples in Calderon et al. [23]. The shape and slope of the curves can be different for different samples, especially Sample 1 in comparison to other samples. This is also reflected in the data for other cell types (Supplementary Figure 1). (b) Differential accessibility log-fold-changes for a 3 vs. 3 mock null comparison based on normalization and differential accessibility analysis using edgeR show a bias for peaks with low and high GC-content (in a null setting, LFC should be centered around zero). The blue curve represents a generalized additive model (GAM) fit. (c) Similar to (b), but using DESeq2 for normalization and differential accessibility analysis. (d) Similar to (b), but using full-quantile normalization and edgeR differential accessibility analysis.

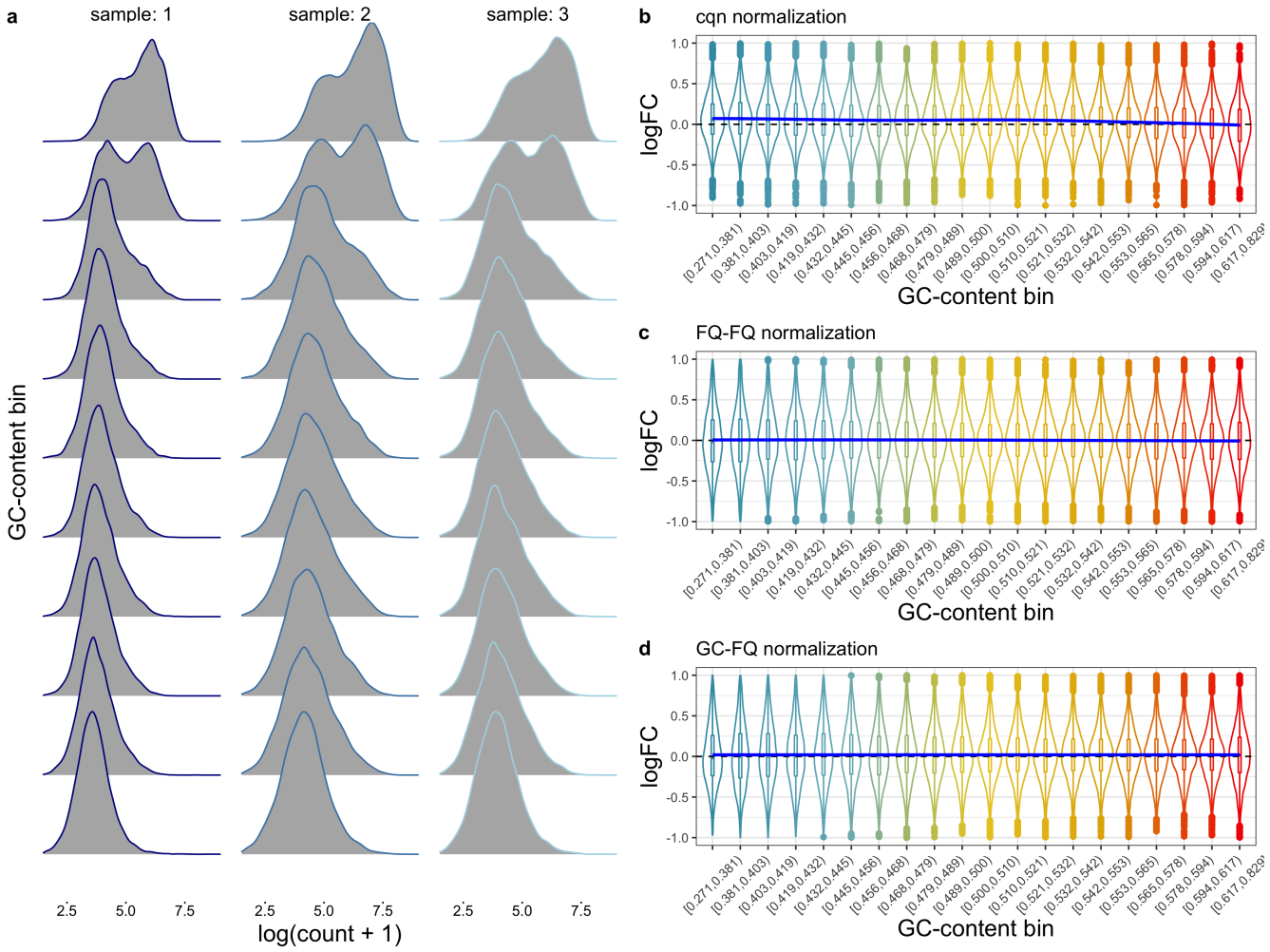


Figure 2: *GC-aware normalization methods* cqn, *FQ-FQ, and GC-FQ are successful in eliminating GC-content effects on the differential accessibility fold-change estimates.* (a) Accessibility distributions for three replicates from the dataset of Philip et al. [17]. The peaks are grouped into 10 equally-sized bins according to their GC-content (rows) and the accessibility distribution (kernel density estimate) is plotted for each bin. The distributional shapes are more comparable across samples for a particular GC-content bin, than they are across GC-content bins for a particular replicate. (b-d) There is no visible GC-content effect on fold-changes estimated using edgeR following normalization with GC-aware methods cqn, EDASeq, and GC-FQ, in the mock comparisons for the dataset from Calderon et al. [23]. The blue curve represents a GAM fit.

TMM (edgeR normalization), MOR (DESeq2 normalization), full-quantile, total-count, upper-quartile, and no normalization; see Methods for a description of each normalization procedure. We also assess RUVg normalization, as implemented in RUVSeq[34], to assess the effect of additionally removing up to five unknown factors of unwanted variation.

None of the normalization methods uniformly performs best across all datasets (Figure 3), suggesting that the optimal normalization is dataset-specific. FQ-FQ and GC-FQ normalizations perform well on average, followed by cqn (Figure 3a). In general, accounting for the width of the peaks in cqn normalization (i.e., method cqn_length) can be advantageous. Notably, the performance of smooth GC-FQ is variable; whereas it performs substantially better than most other methods for some datasets, e.g., from Bryois et al. [28] and Fullard et al. [29], it performs worse than FQ normalization for others, possibly reflecting global distributional differences between biological groups in some datasets. These results are broadly recapitulated when using each normalization method in conjunction with RUVSeq (Supplementary Figure 6). In summary, while no normalization method is uniformly best, GC-unaware methods perform worse than GC-aware methods, strongly suggesting the need to account for GC-content effects in ATAC-seq data.

For each dataset, we can use the scone evaluation measures to explore the different normalizations using principal component analysis (PCA). A clear clustering of normalization methods can often be observed, that is driven by whether or not the methods adopt FQ normalization and account for GC-content (see Supplementary Figures 7 and 8), suggesting that both features have a large, positive impact on normalization. A similar structure is identified when additionally using RUVSeq (Supplementary Figures 10 and 11). In general, accounting for factors of unwanted variation using peaks overlapping with housekeeping genes as negative control features in RUVSeq benefits the normalization (Supplementary Figure 9), but choosing an appropriate normalization method still remains crucial.

Taken together, our evaluation shows that accounting for GC-content effect is critical for normalization of ATAC-seq datasets, and methods FQ-FQ (implemented in the EDASeq package) and GC-FQ provide good results across most datasets.

2.4 Robustness of the evaluation benchmark

The evaluation of normalization methods using scone depends on an appropriate set of evaluation measures. Here, we assess the robustness of our results to the choice of measures.

While we have observed GC-content effects in all eight public datasets we have investigated, it is still informative to assess the performance of each normalization method if, in fact, there is no GC-content effect. We achieve this by randomly permuting the GC-content values across peaks and re-running the benchmarks using scone. In this evaluation, we no longer use the RLE-based measures that assess GC-content effects, since these effects were removed by the permutation and these measures are therefore no longer relevant. For the permuted datasets, FQ-FQ and GC-FQ surprisingly still perform best (Figure 3b and Supplementary Figure 12), followed by FQ normalization. The observation that two rounds of FQ normalization, using FQ-FQ and GC-FQ normalization, still performs better than standard FQ normalization when there are in fact no GC-content effects, likely results from the scone evaluation measures favoring normalization methods that tend to make samples more similar to each other. When additionally accounting for factors of unwanted variation, these results are generally confirmed.

The scenario of permuted GC-content could be unrealistic for ATAC-seq data, however, and we therefore also evaluate all normalization procedures without using the RLE-based measures that assess removal of GC-content effects, on the original datasets without permuting GC-content. In this scenario, FQ-FQ and GC-FQ, but also cqn, still perform well (Figure 3c and Supplementary Figure 13).

Since RLE-based measures favor normalization methods that force distributions to be similar, which is what explicitly happens in full-quantile normalization, one could argue that the use of the RLE-based measures provides an a priori competitive advantage to methods based on full-quantile normalization. To assess the robustness of the benchmarking results, we therefore also provide a ranking of the normalization procedures without using any RLE-based measures. This evaluation thus only focuses on correlation with factors of wanted and unwanted variation, clustering of the samples according to biology and batch, and unsupervised clustering. Notably, smooth GC-FQ normalization achieves the highest ranking, followed by cqn_length (Figure 3d and Supplementary Figure 14). The good performance of smooth GC-FQ in this setting shows the sensitivity of the results to the chosen evaluation measures; a separate, independent evaluation will be provided in the next section. Indeed, since smooth GC-FQ may not necessarily align distributions across conditions, which would be penalized by the RLE-based measures, its increased performance in this setting is to be expected.

Across all evaluations in this section, simple normalization procedures such as total-count normalization or no normalization performed reasonably well as soon as at least 4 factors of unwanted variation were accounted for in RUVSeq. This suggests that RUV may adjust for GC-content effects in an unsupervised manner, and additionally

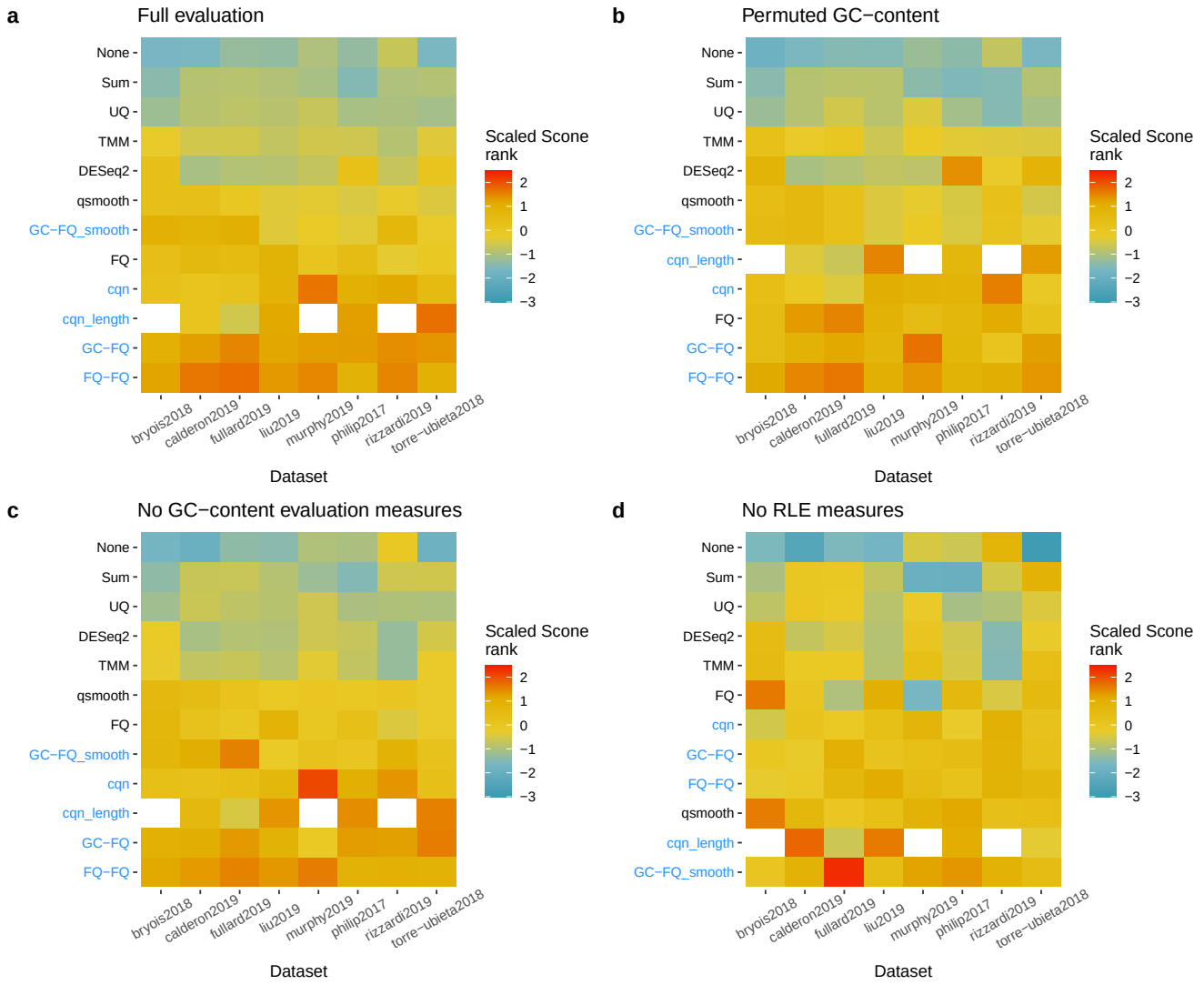


Figure 3: *Benchmark of twelve normalization methods across eight public ATAC-seq datasets using* scone. Twelve different normalization procedures are evaluated on eight public ATAC-seq datasets using scone. The pseudo-color images display matrices of scaled scone ranks (see Methods), with rows corresponding to normalization procedures and columns to datasets, and where a high score corresponds to successful normalization. Methods are ordered according to their average scaled score across all datasets, and their names colored based on whether they explicitly account for GC-content (blue) or not (black). A white cell indicates that the corresponding normalization method was not applied to the corresponding dataset. (a) Benchmark using all evaluation measures. (b) As (a), but removing GC-content effects by permuting GC-content across genomic features. (c) Benchmark without considering evaluation measures that assess GC-content bias. (d) Benchmark without considering any RLE-based evaluation measures.

suggests that using ATAC-seq peaks which overlap with known housekeeping genes may provide appropriate negative control features in ATAC-seq datasets.

2.5 Impact of normalization on differential accessibility analysis

While the previous section provided robust evidence that accounting for GC-content effects is crucial for the normalization of ATAC-seq data, we have also seen that some performance assessment measures may unfairly penalize some normalization methods. Therefore, we provide a second, independent evaluation in terms of differential accessibility (DA) results. In DA analysis, the accessibility of genomic regions is compared between biological groups in order to interpret their epigenetic profiles using, e.g., transcription factor motif enrichment. It is therefore of interest to assess the impact of different normalization approaches on DA analysis.

Parametric simulation approaches, where a specific distribution is fitted for each feature, are often adopted to investigate method performance, e.g., in bulk RNA-seq differential expression analysis[35]. However, typical parametric simulation procedures do not allow for the incorporation of sample-specific effects, such as GC-content effects, therefore motivating a different approach. Additionally, the models may not provide good fits for the actual distribution of read counts. Here, we develop a simple framework that introduces DA signal in real datasets while retaining sample-specific (e.g., GC-content) effects (see Methods).

First, we use a null dataset for the mouse olfactory epithelium (OE), which consists of 10 samples and where we expect no/limited differential accessibility between any grouping of samples. We perform DA analyses for different sample sizes, creating a 4-vs.-4 and a 5-vs.-5 setting by randomly splitting the 10 samples into two groups, each consisting of either 4 or 5 samples, and introducing DA signal for either 10% or 25% of the features in each setting (see Methods). This results in a total of 4 simulation scenarios, which are each evaluated 14 times (a greater number of simulations was not feasible due to local memory limitations). We find consistent good performance for FQ, GC-FQ, qsmooth, and smooth GC-FQ, while, surprisingly, FQ-FQ normalization has low performance (Supplementary Figure 15). cqn performs well in the 5-vs.-5, but less so in the 4-vs.-4 comparisons, while RUVg performs well in datasets with 10% DA features, but not in datasets with 25% DA features. In general, global-scaling methods TMM and DESeq2 perform poorly.

We confirm these results using the dataset from Calderon et al.[23], where we use stratified sampling by cell type, to create random groups of samples, each consisting of comparisons between a total of 10, 20, 30, and 40 samples and, again, introducing signal in either 10% or 25% of the features (see Methods). In these evaluations, global-scaling methods TMM and DESeq2, as well as RUVg, generally perform worse than all other methods (Supplementary Figure 16). Top performing normalization methods are GC-FQ, smooth GC-FQ, and FQ-FQ, again underscoring the effectiveness of combining full-quantile normalization with accounting for GC-content effects. Other good performing methods include cqn and FQ normalization.

In summary, using a separate evaluation framework, we confirm that accounting for GC-content effects is critical for ATAC-seq data normalization and that FQ-based methods generally perform well.

2.6 Case studies

In what follows, we consider the normalization of ATAC-seq datasets in greater depth using two case studies. These serve as demonstrations of how one can evaluate normalization procedures in practice using exploratory data analysis techniques. We assess each of the normalization methods that were benchmarked above, except for the basic total-count and upper-quartile normalization methods. Additionally, to keep the message clearer, we do not apply RUVSeq after normalization by each of the other methods.

2.6.1 Mouse Tissue Atlas

Liu et al.[30] presented an ATAC-seq atlas of 20 tissues in adult mice, consisting of 296,416 peaks across 66 samples. Hierarchical clustering based on the Euclidean distance of the log-transformed normalized counts shows that normalization is essential to derive a biologically sensible clustering of the samples (Supplementary Figure 17). Without normalization, several tissues do not cluster together. The clustering is improved by using FQ normalization or global-scaling normalization methods TMM (edgeR) and MOR (DESeq2), but these still fail to cluster the ovary or adrenal gland tissue samples properly. By contrast, GC-aware methods cqn, FQ-FQ, and smooth GC-FQ successfully group the samples of each tissue type together, while GC-FQ misclusters one adrenal gland sample.

Next, we perform a differential accessibility analysis using the normalized counts from each normalization method as input (see Methods on how normalized counts were obtained). We model the accessibility counts using a negative binomial distribution as implemented in edgeR and assess DA between heart and liver tissues. Assuming that either

a small fraction of peaks are DA, or that there is symmetry in the direction of DA between the groups under comparison, log-fold-changes should be centered at zero and similarly distributed across different GC-content bins. However, log-fold-changes are biased for peaks with both low GC-content and high GC-content values for all GC-unaware normalization methods (Supplementary Figure 18). While this technical artefact is successfully removed by FQ-FQ and GC-FQ normalizations, cqn and smooth GC-FQ still suffer from substantial bias (Supplementary Figure 18). Since a high GC-content is also associated with a high accessibility count, which is in turn associated with high statistical power, we naturally expect the top DA peaks to be skewed in terms of GC-content, i.e., we expect a dominance of high GC-content values for the DA peaks. This is indeed the case for all normalization methods (Supplementary Figure 19), while the top peaks for TMM normalization are remarkably balanced. If we focus on the significant peaks at a nominal false discovery rate (FDR) threshold of 5%[36], most methods discover a comparable number of around 130×10^3 peaks. However, cqn flags a substantially higher number of peaks, $\sim 153 \times 10^3$. The peaks discovered by cqn are also more balanced with respect to GC-content, as compared to other methods (Supplementary Figure 20).

2.6.2 Brain Open Chromatin Atlas

Fullard et al.[29] published the Brain Open Chromatin Atlas (BOCA), where chromatin accessibility is measured in five human postmortem brain samples. The dataset consists of a total of 14 brain regions and two cell types (neuronal and non-neuronal). These 14 brain regions can be classified into six broader regions, namely, the neocortex (NCX), primary visual cortex (PVC), amygdala (AMY), hippocampus (HIP), mediodorsal thalamus (MDT), and striatum (STR). After normalizing the counts using each normalization method (see Methods), we first investigate how the samples cluster according to hierarchical clustering based on the Euclidean distance of the log-transformed normalized counts. In particular, we check which normalizations succeed in separating neuronal from glial (i.e., non-neuronal) cells, which is a major source of variability in this dataset (Figure 4a and Supplementary Figure 21). All methods except for no normalization and cqn correctly cluster each cell type in partitions that contain only a single cell type (Supplementary Figure 22). While most methods result in two large partitions that directly separate the two cell types, GC-FQ and smooth GC-FQ separately cluster a single glial and five neuronal samples, respectively.

We also assess how well each normalization method is able to recover the cell types and the major brain regions within each of these cell types by clustering the datasets using partitioning around medoids (PAM) based on the first 2 to 10 principal components. We consider PAM clustering at two resolution levels: First, we search for two clusters and check how well these correspond to the known cell types (i.e., glial and neuronal); next, we search for 12 clusters and check how well these correspond to the known regions within each cell type. We evaluate the clusterings using the adjusted Rand index (ARI)[37,38], comparing the derived partitions with the ground truth. Interestingly, all methods typically cluster the majority or all samples correctly according to cell type, except for cqn and no normalization (Supplementary Figure 23). However, when checking how well the different brain regions within each cell type are recovered by clustering the normalized datasets into 12 clusters (Figure 4b), for each selected number of PCs, GC-FQ performs best, while cqn and no normalization perform worst. The good performance of GC-FQ, and also FQ-FQ, was already noticeable in the PCA plots in Supplementary Figure 21, since these are the only methods where the STR and MDT regions are clearly separated from other brain regions for the neuronal cells in 2 dimensions.

Asides from clustering, researchers often focus on discovering peaks that are differentially accessible between biological groups. Here, we use edgeR to fit a negative binomial generalized linear model for each peak in each normalized dataset. For each peak, we test for differences in average accessibility between neuronal and non-neuronal cells. Since euchromatin is typically more GC-rich, there is a natural association of accessibility with GC-content. However, mean-difference plots (MD-plots)[39] also show a prominent GC-content bias in the fold-changes for all GC-unaware normalization methods (Figure 4c and Supplementary Figure 24). Likewise, stratified violin plots of the fold-changes by GC-content show substantial bias of the fold-changes as a function of GC-content (Supplementary Figure 25). Only FQ-FQ and GC-FQ successfully remove GC-content effects on the fold-changes. smooth GC-FQ removes the bias partially, but not completely. Interestingly, the log-fold-changes following cqn normalization tend to be biased. However, we note that the log-fold-change bias in cqn can be alleviated by allowing more peaks to be included in the normalization, since default settings only used $\sim 5\%$ of all peaks, and that other DA methods such as glmGamPoi[40] do not show such pronounced fold-change bias (Supplementary Figure 26). The peculiar results for cqn normalization are also reflected in the number of DA peaks from edgeR, which is at least $\sim 40\%$ higher as compared to all other methods (Supplementary Figure 27). The scone evaluations of Figure 3 showed that cqn has a lower score for the dataset of Fullard et al.[29], as compared to other datasets. These results emphasize the need for exploratory data analysis in order to choose an appropriate normalization method.

To assess the relevance of the discovered DA peaks, we check the genomic features and enriched gene sets associated with them, where we assign a gene to a peak if its promoter is within a $5,000$bp distance of the peak. The intersection of $66,091$ DA peaks discovered across all methods is enriched in genomic features such as exons, promoters, and $5'$ UTRs, while depleted in intergenic regions, as compared to the background of all peaks (Supplementary Figure 28). The enriched biological process (BP) gene sets are highly relevant, including *neurogenesis* and *nervous system development*, among others (Supplementary Table 1). The DA results also allow us to assess whether accounting for GC-content effects can aid biological interpretation. We therefore interpret the set of peaks discovered by FQ-FQ and GC-FQ and not by their GC-unaware counterpart, FQ normalization. These peaks are enriched in genomic features such as promoters, $5'$ UTRs, and exons (Supplementary Figure 28), and are significantly enriched in 28 biological processes, including *neuron projection morphogenesis* and *synapse organization* (Supplementary Table 2), reinforcing their biological relevance. We also further investigate the peaks uniquely discovered by cqn and cqn_noLength. These peaks are enriched in intergenic regions (Supplementary Figure 28), and no enriched gene sets are found at a nominal 5% FDR level, which encourages our intuition that these are likely false positive peaks. Indeed, while we were able to annotate 14% of the peaks discovered by FQ-FQ and GC-FQ but not FQ, only 4% of the peaks uniquely discovered by cqn could be annotated.

Taken together, these results again suggest that while GC-content normalization is crucial, the most appropriate normalization method depends on the dataset being analyzed. While cqn generally performs well across datasets, other methods should be preferred on the BOCA dataset. Exploratory data analysis is essential for evaluating and guiding the choice of effective normalization and removal of technical GC-content effects.

3 Discussion

The evaluations in this manuscript highlight the importance of accounting for GC-content effects in ATAC-seq datasets. Because of the sample-specificity of GC-content effects, failing to adjust for GC-content using an appropriate normalization method can bias downstream analyses such as clustering and differential accessibility analysis. We have proposed GC-aware normalization procedures and benchmarked these against state-of-the-art procedures using eight public ATAC-seq datasets. While GC-aware procedures perform better than GC-unaware procedures, none uniformly outperforms all others. The choice of an appropriate normalization procedure is dataset-specific and exploratory data analysis is essential to guide this choice.

Similar GC-content effects have also been noted in DNA-seq[41], RNA-seq[11,21], and ChIP-seq[9], amongst others. For ChIP-seq datasets, Teng and Irizarry[9] recently developed a negative binomial mixture model to correct for GC-content effects in both background and binding signal at the peak-calling stage, by accounting for GC-content in the abundance estimation for a particular genomic region. While their method has been evaluated using ChIP-seq data, it may also be useful for ATAC-seq data. However, none of the publicly-available and processed datasets we have looked into account for GC-content effects during the peak-calling stage.

While in this manuscript we have focused on correcting GC-content bias at the level of called peaks, other approaches are possible. For example, Benjamini and Speed[41] argue that it is the GC-content of the full DNA fragment (vs. only the sequenced read) that most influences read counts. A comparison of the peak-level normalization approaches discussed here with fragment-level approaches would be an interesting avenue for further research on how to best correct for GC-content effects.

Our work has focused on normalization of bulk ATAC-seq datasets. While FQ-based normalization procedures were found to perform favorably in this setting, it remains to be seen whether they perform equally well on single-cell ATAC-seq (scATAC-seq) datasets. The sparsity associated with scATAC-seq data suggests that their application could be limited and alternative normalization procedures may be needed.

4 Methods

4.1 Datasets

Philip et al.[17] study CD8 T-cell dysfunction in acutely infected and chronic tumoral tissue, over several time points. We only focus on the mouse samples and consider time and treatment as the biological variables of interest. We did not find metadata on quality control (QC) or batch variables, so we do not use any in the scone evaluation. The count matrix corresponds to $75,689$ peaks for 41 samples and was downloaded from the Gene Expression Omnibus (GEO) with accession number GSE89308.

Bryois et al.[28] study the adult human prefrontal cortex brain. We remove samples that are not schizophrenic or control samples, leaving a total of 272 samples, consisting of 135 individuals with schizophrenia and 137 controls. We

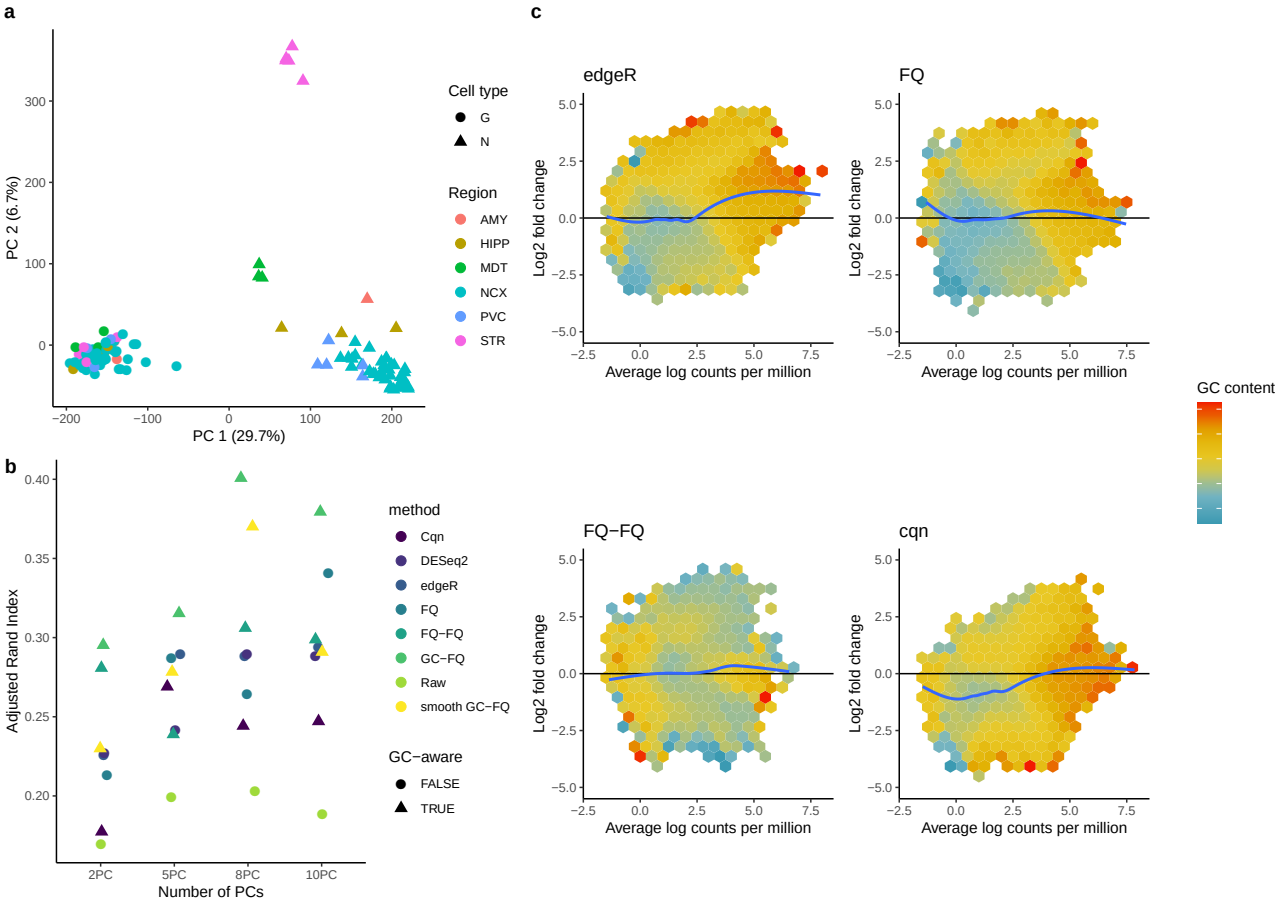


Figure 4: *Analysis of the Brain Open Chromatin Atlas dataset.* (a) PCA plot of the dataset after FQ-FQ normalization. The plotting symbols denote cell type, neuronal (N) and glial (G); the colors represent the six broad brain regions. (b) The samples were clustered using PAM based on a variable number of PCs (x-axis), after normalization with each of nine methods. The y-axis corresponds to the adjusted Rand index comparing the PAM clusters with the true partitioning according to brain region and cell type (12 clusters in total). Different normalizations are represented by different colors and GC-aware normalization methods are represented with triangles. GC-aware methods generally perform better, on average. (c) MD-plots for differential accessibility analysis comparing neuronal vs. non-neuronal cells. The peaks are grouped into hexagons, where the color of each hexagon denotes the average GC-content of its corresponding peaks. There is substantial GC-content bias for GC-unaware normalization methods edgeR and FQ, where low GC-content is associated with low fold-changes and vice versa. This is not the case for FQ-FQ normalization. The log-fold-change distribution for cqn is skewed towards negative values.

consider the disease status as the biological variable of interest. In the evaluation, we use the 32 QC variables that were available in the metadata, along with the top 10 principal components derived from the patients' genotypes. The sequencing index in the metadata is used as batch variable. The count matrix corresponds to $118,152$ peaks and was obtained through personal communication with the authors. It was not relevant to correct for the width of the peaks using `cqn` in this dataset, since all peaks have a length of 301bp.

de la Torre-Ubieta et al.[14] study human cortical neurogenesis in the germinal zone and cortical plate of the developing cerebral cortex. Samples were derived from three individual donors and each donor was handled and processed separately, so we treat each donor as a batch and the brain region as the biological variable of interest. The count matrix corresponds of $62,005$ peaks across 19 samples and was downloaded from the GEO with accession number GSE95023. Note that the replication in this dataset is technical, i.e., consists of samples from the same human donor.

Calderon et al.[23] study a repertoire of 32 immune cell types under resting and activated conditions in humans. The metadata include three QC variables (number of peaks called, number of sequenced reads, and transcription start site enrichment for each sample), which we use in the `scone` evaluation. Most donors are processed and sequenced separately, and therefore each donor represents a different batch. However, some samples underwent a second round of sequencing and the accessibility counts from these two sequencing rounds were summed, so we also treat these additionally sequenced samples as a separate batch. The biological variables of interest are cell type and treatment. The count matrix corresponds to $829,942$ peaks across 175 samples and was downloaded from the GEO with accession number GSE118189. This dataset was filtered to retain peaks with at least 2 counts per million in at least 10 samples, reducing the dataset to $203,448$ peaks.

Murphy et al.[31] study photoreceptors and bipolar cells in the mouse retina. The dataset combines two experiments and we define each experiment as a batch. We consider the cell type as the biological variable of interest. The count matrix corresponds to $110,715$ peaks across 12 samples and was downloaded from the GEO with accession number GSE131625. It was not relevant to correct for the width of the peaks using `cqn` in this dataset, since all peaks have a length of 201bp.

Rizzardi et al.[16] study neuronal and non-neuronal cell populations in the prefrontal cortex and nucleus accumbens in humans. We consider the combination of brain region and cell type as the biological variable of interest. We define a batch variable as the combination of the donor, flow cytometry run, and sequencing date variables. The count matrix corresponds to $961,916$ peaks across 22 samples and was downloaded from the GEO with accession number GSE96614.

Brain Open Chromatin Atlas (case study). Fullard et al.[29] developed a human brain atlas of neuronal and non-neuronal cells across 14 distinct brain regions from 5 human donors. We define a batch variable as the flow cytometry date. Note that while the sequencing date is nested within the flow cytometry date, there are other variables in the metadata that might also be considered to define batches, e.g., PCR date. A total of 49 variables corresponding to potential technical effects were included as QC measures. The biological variable of interest is defined as the combination of cell type and brain region. The count matrix corresponds to $300,444$ peaks across 115 samples and was downloaded from the Brain Open Chromatin Atlas (BOCA) website, at `https://bendlj01.u.hpc.mssm.edu/multireg/`.

Mouse Tissue Atlas (case study). Liu et al.[30] created an ATAC-seq atlas of mouse tissues, spanning a total of 20 tissues for both male and female mice. We use the "Slide lane of sequencer" variable recorded in the metadata as batch variable. Four variables (mitochondrial reads, usable reads, transcription start site (TSS) enrichment, and number of reproducible peaks, as identified using the irreproducible discovery rate (IDR)[42]) are used as QC measures. The combination of gender and tissue type is used as the biological variable of interest. The count matrix corresponds to $296,574$ peaks across 79 samples and was downloaded from Figshare at `https://doi.org/10.6084/m9.figshare.c.4436264.v1`.

Mouse olfactory epithelium null dataset. This ATAC-seq dataset was generated by us for the present study and consists of 10 samples of either FACS-sorted ICAM1+ or ICAM1- cells that were dissected from the mouse olfactory epithelium (OE). Samples have variation in input cell number, Tn5 transposase concentration, ATAC-seq method modification, processing and sequencing batch, and treatment that induced olfactory neuron death. The dataset consists of $96,081$ peaks and 10 samples.

4.2 GC-content retrieval

For each dataset, we use the Bioconductor R package `Biostrings` to retrieve the GC-content of every peak region, using the reference genome of the relevant organism. Table 1 provides the genome version used for each dataset.

Table 1: *Genome version used for each dataset.*

Dataset	Organism	Genome
Philip et al. [17]	Mouse	GRCm38
Bryois et al. [28]	Human	GRCh37.75
Calderon et al. [23]	Human	GRCh37.75
de la Torre-Ubieta et al. [14]	Human	GRCh37.75
Murphy et al. [31]	Mouse	GRCm38
Rizzardi et al. [16]	Human	GRCh37.75
Fullard et al. [29]	Human	GRCh37.75
Liu et al. [30]	Mouse	GRCm37.67

4.3 Normalization procedures

Let Y_{ji} denote the accessibility count for peak $j = 1, \ldots, J$ in sample $i = 1, \ldots, n$. The evaluated normalization procedures can be summarized as follows.

No normalization. The raw counts are used for analysis.

Total-count (TC) normalization. Each count is divided by the total library size, $N_i = \sum_j Y_{ji}$, for its corresponding sample.

Upper-quartile (UQ) normalization. Each count is divided by the upper-quartile (i.e., 75th percentile) of the counts for its corresponding sample. UQ can be beneficial over TC normalization, as the latter can be affected by a few very high counts that dominate the total library size N_i.

Trimmed mean of M-values (TMM) normalization. TMM is a global-scaling normalization procedure that was originally proposed by Robinson and Oshlack [19]. As the name suggests, it is based on a trimmed mean of fold-changes (M-values) as the scaling factor. A trimmed mean is an average after removing a set of "extreme" values. Specifically, TMM calculates a normalization factor $F_i^{(r)}$ for each sample i as compared to a reference sample r,

$$\log_2(F_i^{(r)}) = \frac{\sum_{j \in \mathcal{J}^*} w_{ji}^r M_{ji}^r}{\sum_{j \in \mathcal{J}^*} w_{ji}^r}, \tag{1}$$

where M_{ji}^r represents the $\log_2$-fold-change of the accessibility fraction as compared to a reference sample r, i.e.,

$$M_{ji}^r = \log_2 \left(\frac{Y_{ji}/N_i}{Y_{jr}/N_r} \right),$$

w_{ji}^r represents a weight calculated as

$$w_{ji}^r = \frac{N_i - Y_{ji}}{N_i Y_{ji}} + \frac{N_r - Y_{jr}}{N_r Y_{jr}}, \tag{2}$$

and $\mathcal{J}^*$ represents the set of peaks after trimming those with the most extreme values.

The procedure only takes peaks into account where both $Y_{ji} > 0$ and $Y_{jr} > 0$. By default, TMM trims peaks with the 30% most extreme M-values and 5% most extreme average accessibility, and chooses as reference r the sample whose upper-quartile is closest to the across-sample average upper-quartile. The normalized counts are then given by $\tilde{Y}_{ji} = Y_{ji}/N_i^s$, where

$$N_i^s = \frac{N_i F_i^{(r)}}{\frac{1}{n} \sum_{i=1}^n N_i F_i^{(r)}}.$$

Median-of-ratios (MOR) normalization. The median-of-ratios method is used in DESeq2 [20]. It assumes that the expected value $\mu_{ji} = E(Y_{ji})$ is proportional to the true accessibility of the peak, q_{ji}, scaled by a normalization factor s_i for each sample,

$$\mu_{ji} = s_i q_{ji}.$$

The normalization factor s_i is then estimated using the median-of-ratios method compared to a synthetic reference sample r defined based on geometric means of counts across samples

$$s_i = \text{median}_{\{j : Y_{jr}^* \neq 0\}} \frac{Y_{ji}}{Y_{jr}^*}, \tag{3}$$

with

$$Y^*_{jr} = \left(\prod_{i=1}^{n} Y_{ji} \right)^{1/n}.$$

From this, we calculate the normalized count as $\tilde{Y}_{ji} = Y_{ji}/s_i$.

Full-quantile (FQ) normalization. In full-quantile normalization[43], the samples are forced to each have a distribution identical to the distribution of the median/average of the quantiles across samples. In practice, we implement full-quantile normalization using the following procedure

1. given a data matrix $\mathbf{Y}_{J\times n}$ for J peaks (rows) and n samples (columns),
2. sort each column to get $\mathbf{Y}^S$,
3. replace all elements of each row by the median (or average) for that row,
4. obtain the normalized counts $\tilde{\mathbf{Y}}$ by re-arranging (i.e., unsorting) each column.

Smooth-quantile (SQ) normalization (qsmooth). Full-quantile normalization assumes that the read count distribution has a similar shape for each sample and that the observed variability in global distributional properties corresponds to technical effects. However, this may not always be the case in practice. To tackle this, Hicks et al.[44] developed smooth-quantile normalization, a variant of full-quantile normalization that is able to deal with datasets where there are large global differences between biological conditions of interest. It provides a balance between (a) full-quantile normalization between samples of each condition separately, and (b) full-quantile normalization on the full dataset. This balance is struck by calculating data-driven weights for each quantile, that specify which of the two normalization options is more appropriate. The weights are estimated in a smooth way across the quantiles, by contrasting the within-condition with the between-condition variability for each quantile. If the within-condition variability is significantly smaller than the between-condition variability, then the weights will favor normalization for each condition separately.

Within-and-between-sample full-quantile (FQ-FQ) normalization. The FQ-FQ method, implemented in the EDASeq package[10], accounts for GC-content effects by performing two rounds of full-quantile normalization. First, the features of each sample are grouped into (by default, 10) GC-content bins and full-quantile normalization is performed across bins within each sample (referred to as 'within-lane normalization'). Next, the data are normalized using full-quantile normalization across all samples.

Conditional-quantile normalization (cqn). The cqn method[45] starts by assuming a Poisson model for the accessibility counts Y_{ji}. Median regression is used to model, for each sample, the log-transformed accessibility count as a smooth function of GC-content, focusing on peaks with high average count (above 50). Next, subset quantile normalization[46] is performed on the residuals of that model (i.e., on the counts adjusted for GC-content) for between-sample normalization. The method could intuitively be thought of as full-quantile normalization after removing a smoothed sample-specific GC-content effect. Normalized counts are calculated as recommended in the cqn vignette, i.e.,

$$\tilde{Y}_{ji} = \left(\frac{(Y_{ji}+1)10^6}{\sum_j Y_{ji}} \right) 2^{O_{ji}}, \tag{4}$$

with O_{ji} the normalization offset estimated by cqn, which is on the $\log_2$ scale.

GC-full-quantile (GC-FQ) normalization. GC-FQ is similar to FQ-FQ, but relies on the observation that, in ATAC-seq, read count distributions are often more comparable between samples within a GC-content bin, than between GC-content bins within a sample (Figure 2). It therefore applies between-sample FQ normalization for each GC-content bin separately, with 50 bins by default.

Smooth GC-FQ normalization. smooth GC-FQ is a variant of GC-FQ that applies smooth-quantile normalization across samples within each GC-content bin. Like GC-FQ, it uses 50 bins by default.

4.4 scone benchmark

We use the Bioconductor R package scone to implement and evaluate different normalization procedures. The first step in the scone workflow is to normalize the data using all normalization methods of interest, and then perform regression-based normalization to adjust for both known and unknown factors of unwanted variation. The unknown factors of unwanted variation are inferred using RUVSeq[34], based on negative control features which we here set to be the peaks that fall within known housekeeping genes. We use all available evaluation criteria implemented in scone and add two criteria that specifically assess GC-content effects and are described below.

Each normalization procedure is assigned a score for each evaluation criterion, constructed such that a high score corresponds to a good performance of the normalization procedure with respect to that evaluation criterion. Since scores are not directly comparable between criteria, we first rank the scores for each evaluation criterion separately, where a high rank reflects good normalization. As a summary for each normalization procedure, we average the ranked scores across evaluation criteria. Finally, since not all normalization procedures could be applied to each dataset (e.g., in some datasets, all peaks have the same size and therefore one cannot apply cqn_length), we scale the average ranks for a given dataset to have zero mean and unit standard deviation across methods; this scaling makes scores more comparable between datasets.

4.4.1 Assessing GC-content bias

To assess GC-content bias after normalization, we define two measures based on relative log expression (RLE) values[33] across GC-content bins, which are inspired by the RLE_MED and RLE_IQR measures already implemented in scone.

Let Y_{ji} denote the accessibility measure for peak j in sample i and $L_{ji} = \log(Y_{ji} + 1)$. Then, the RLE is defined as

$$r_{ji} = \text{Ł}_{ji} - \bar{L}_{j.},$$

where $\bar{L}_{j.}$ denotes the median of L_{ji} across all samples i. For a set of peaks within a GC-content bin b, we can define a measure of GC-content bias as the mean squared median RLE[22], i.e., as the average squared deviation of the median RLE from zero,

$$d_b = \frac{1}{n} \sum_{i=1}^{n} \bar{r}_{.ib}^2,$$

where $\bar{r}_{.ib}$ is the median RLE value for peaks in bin b for sample i. A small value of d_b generally corresponds to a good normalization of the data, since the median RLE is close to zero. However, if sample-specific GC-content effects exist, then the normalization may only be working for certain ranges of GC-content. We therefore assess whether the mean squared median RLE varies with GC-content by computing its variance across GC-content bins,

$$RLE_MED_{GC} = \frac{1}{B-1} \sum_{b=1}^{B} (d_b - \bar{d})^2, \tag{5}$$

where $\bar{d} = \sum_b d_b / B$ is the average of d_b across GC-content bins. In the evaluation, we let the number of bins B depend on the total number of peaks in a given dataset by constructing bins containing around $4,000$ peaks each.

A similar measure is calculated based on the variance of the interquartile range (IQR) of the RLE measures for each bin,

$$v_b = \frac{1}{n-1} \sum_{i=1}^{n} (q_{ib} - \bar{q}_b)^2,$$

where q_{ib} is the interquartile range of the RLE values for peaks in bin b for sample i and $\bar{q}_b$ its average across all samples. Using a similar reasoning as above, we then evaluate the variance of v_b across different GC-content bins

$$RLE_IQR_{GC} = \frac{1}{B-1} \sum_{b=1}^{B} (v_b - \bar{v})^2, \tag{6}$$

where $\bar{v}$ is the average of v_b across all bins.

4.5 Simulation study

In order to evaluate the impact of normalization on DA analysis, we simulate data based on two real datasets, an unpublished mouse olfactory epithelium null dataset from our group and the dataset from Calderon et al.[23] (see Methods, Datasets for a descriptions of each dataset). For the null dataset, samples are randomly split into two equally-sized groups. For the Calderon et al.[23] dataset, we focus on unstimulated T-cells (42 samples over 12 cell types, with 3 or 4 samples per cell type) and perform stratified random sampling, where the samples for each cell type are randomly split into two approximately equally-sized groups (i.e., for cell types with 4 samples, there are 2 samples per group, and for cell types with 3 samples, one group comprises 1 sample and the other 2 samples). For each dataset, following the assignment of samples to groups, we evaluate the performance of normalization and DA analysis methods for a range of total sample sizes (8 and 10 for the null dataset, by subsampling from each group;

10, 20, 30, and 40 for the Calderon et al.[23] dataset, by subsampling from each group within each cell type) and differential accessibility signal strengths (10% and 25% DA peaks).

Specifically, for each selected sample i, we calculate its accessibility fraction for each peak j as

$$F_{ji} = \frac{Y_{ji}}{\sum_{j=1}^{J} Y_{ji}}. \tag{7}$$

A proportion of peaks (10% or 25%) is simulated to be differentially accessible, with equal up-/down-regulation between groups, via independent binary random variables S_j, equal to either -1 or 1, each with 1/2 probability. The S_j's define the group $g \in \{1, 2\}$ of samples for which the accessibility fractions will be altered as follows

$$\gamma_{ji} = \begin{cases} F_{ji} \exp^{I(S_j=-1)Z_j}, & \text{if } g(i) = 1 \\ F_{ji} \exp^{I(S_j=1)Z_j}, & \text{if } g(i) = 2 \end{cases}, \tag{8}$$

where $g(i) \in \{1, 2\}$ denotes the group to which sample i belongs and Z_j are independent Gaussian random variables with mean 0.8 and standard deviation 0.1. That is, the log-fold-change in accessibility between groups 2 and 1 is $S_j Z_j$. Sequencing depths N_i^* are simulated from a uniform distribution

$$N_i^* \sim U(N_{min}, N_{max}),$$

where $N_{min} = \min_i \sum_j Y_{ji}$ and $N_{max} = \max_i \sum_j Y_{ji}$ denote, respectively, the minimum and maximum library sizes across all samples in the dataset. Given γ_{ji} and N_i^*, accessibility counts Y_{ji}^* are then simulated using a Multinomial distribution

$$Y_{ji}^* | \gamma_{ji}, N_i^* \sim Mult(\gamma_{ji}, N_i^*). \tag{9}$$

For each combination of sample size and DA signal strength, we evaluate 14 simulated datasets (a greater number of simulations was not feasible due to local memory limitations). Differential accessibility analysis is performed using edgeR for all normalization methods, except for MOR (DESeq2) normalization where we rely on the native DESeq2 pipeline. For RUVg normalization, we use peaks overlapping with housekeeping genes as negative control features, as in the scone evaluation, excluding features that were simulated to contain signal between the groups.

For each simulated dataset, we calculate the true positive rate (TPR) and false discovery rate (FDR), defined as

$$\begin{aligned} TPR &= \frac{TP}{TP + FN} \\ FDR &= \frac{FP}{TP + FP}, \end{aligned} \tag{10}$$

where FN, FP, and TP denote, respectively, the numbers of false negatives, false positives, and true positives.

Method performance is visualized using FDR-TPR curves, constructed by calculating, for each of the 14 simulated datasets, FDR and TPR ratios by sequentially moving from the most to the least significant DA peak and then averaging the FDR and TPR over the 14 simulations.

4.6 Case studies

4.6.1 Mouse Tissue Atlas

The raw count matrix from Liu et al.[30] is obtained as described in the Datasets section and is normalized using each of the twelve evaluated normalization procedures, with no peaks filtered out. Euclidean distances between samples are calculated on the log-normalized counts, adding an offset of 1 to avoid taking the log of zero. Hierarchical clustering trees are derived using complete linkage. Differential accessibility analysis is performed using edgeR for all normalization methods, except for MOR normalization where we rely on the native DESeq2 pipeline. Normalized counts are used directly as input for FQ, FQ-FQ, and (smooth) GC-FQ normalization, while normalization offsets are used for TMM and cqn. For RUVSeq normalization, we incorporate 4 inferred factors of unwanted variation in the design matrix.

4.6.2 Brain Open Chromatin Atlas

The raw count matrix from Fullard et al.[29] is obtained as described in the Datasets section. We do not filter any peaks. PCA and hierarchical trees are based on the log-normalized counts, adding an offset of 1 to avoid taking the log of zero. Hierarchical trees use the Euclidean distance between samples and are constructed using complete linkage. DA analysis is performed as described in the Mouse Tissue Atlas case study. The contrast matrix is defined for comparing the average expression of neuronal vs. non-neuronal samples.

5 Data and code availability

A GitHub repository containing all code for the analyses can be found at `https://github.com/koenvandenberge/bulkATACGC`, which also contains a link to Zenodo to download the data used for these analyses. All datasets used in this study are publicly-available, either online or by contacting the original authors, as described in the Methods section.

6 Acknowledgements

The authors thank Michael Love for his input on scaling normalization, and Ameek Bindra, who contributed to this project through the Undergraduate Research Apprenticeship Program (URAP) of the Department of Statistics of the University of California, Berkeley.
KVdB is a postdoctoral fellow of the Belgian American Educational Foundation (BAEF) and is supported by the Research Foundation Flanders (FWO) grant 1246220N. H-J C was supported by a postodoctoral research grant from the UC Berkeley Siebel Stem Cell Center. SD and JN are supported by the National Institutes of Health, grant R01 DC007235.

References

[1] Sandy L. Klemm, Zohar Shipony, and William J. Greenleaf. Chromatin accessibility and the regulatory epigenome, 4 2019. ISSN 14710064. URL https://www.nature.com/articles/s41576-018-0089-8.

[2] Alan P. Boyle, Sean Davis, Hennady P. Shulha, Paul Meltzer, Elliott H. Margulies, Zhiping Weng, Terrence S. Furey, and Gregory E. Crawford. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell*, 132(2):311–322, 1 2008. ISSN 00928674. doi: 10.1016/j.cell.2007.12.014. URL http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg17.

[3] Robert E. Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T. Maurano, Eric Haugen, Nathan C. Sheffield, Andrew B. Stergachis, Hao Wang, Benjamin Vernot, Kavita Garg, Sam John, Richard Sandstrom, Daniel Bates, Lisa Boatman, Theresa K. Canfield, Morgan Diegel, Douglas Dunn, Abigail K. Ebersol, Tristan Frum, Erika Giste, Audra K. Johnson, Ericka M. Johnson, Tanya Kutyavin, Bryan Lajoie, Bum Kyu Lee, Kristen Lee, Darin London, Dimitra Lotakis, Shane Neph, Fidencio Neri, Eric D. Nguyen, Hongzhu Qu, Alex P. Reynolds, Vaughn Roach, Alexias Safi, Minerva E. Sanchez, Amartya Sanyal, Anthony Shafer, Jeremy M. Simon, Lingyun Song, Shinny Vong, Molly Weaver, Yongqi Yan, Zhancheng Zhang, Zhuzhu Zhang, Boris Lenhard, Muneesh Tewari, Michael O. Dorschner, R. Scott Hansen, Patrick A. Navas, George Stamatoyannopoulos, Vishwanath R. Iyer, Jason D. Lieb, Shamil R. Sunyaev, Joshua M. Akey, Peter J. Sabo, Rajinder Kaul, Terrence S. Furey, Job Dekker, Gregory E. Crawford, and John A. Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 9 2012. ISSN 00280836. doi: 10.1038/nature11232. URL https://www.nature.com/articles/nature11232.

[4] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218, 12 2013. ISSN 1548-7091. doi: 10.1038/nmeth.2688. URL http://www.nature.com/articles/nmeth.2688.

[5] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nussbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, 9 2008. ISSN 1465-6906. doi: 10.1186/gb-2008-9-9-r137. URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2008-9-9-r137.

[6] Peter V Kharchenko, Michael Y Tolstorukov, and Peter J Park. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology*, 26(12):1351–1359, 12 2008. ISSN 1087-0156. doi: 10.1038/nbt.1508. URL http://www.nature.com/articles/nbt.1508.

[7] Sheng Li, Pawel P. Labaj, Paul Zumbo, Peter Sykacek, Wei Shi, Leming Shi, John Phan, Po Yen Wu, May Wang, Charles Wang, Danielle Thierry-Mieg, Jean Thierry-Mieg, David P. Kreil, and Christopher E. Mason. Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nature Biotechnology*, 32(9): 888–895, 9 2014. ISSN 15461696. doi: 10.1038/nbt.3000. URL https://www.nature.com/articles/nbt.3000.

[8] Zhenqiang Su, Paweł P. Łabaj, Sheng Li, Jean Thierry-Mieg, Danielle Thierry-Mieg, Wei Shi, Charles Wang, Gary P. Schroth, Robert A. Setterquist, John F. Thompson, Wendell D. Jones, Wenzhong Xiao, Weihong Xu, Roderick V. Jensen, Reagan Kelly, Joshua Xu, Ana Conesa, Cesare Furlanello, Hanlin Gao, Huixiao Hong, Nadereh Jafari, Stan Letovsky, Yang Liao, Fei Lu, Edward J. Oakeley, Zhiyu Peng, Craig A. Praul, Javier Santoyo-Lopez, Andreas Scherer, Tieliu Shi, Gordon K. Smyth, Frank Staedtler, Peter Sykacek, Xin Xing Tan, E. Aubrey Thompson, Jo Vandesompele, May D. Wang, Jian Wang, Russell D. Wolfinger, Jiri Zavadil, Scott S. Auerbach, Wenjun Bao, Hans Binder, Thomas Blomquist, Murray H. Brilliant, Pierre R. Bushel, Weimin Cai, Jennifer G. Catalano, Ching Wei Chang, Tao Chen, Geng Chen, Rong Chen, Marco Chierici, Tzu Ming Chu, Djork Arné Clevert, Youping Deng, Adnan Derti, Viswanath Devanarayan, Zirui Dong, Joaquin Dopazo, Tingting Du, Hong Fang, Yongxiang Fang, Mario Fasold, Anita Fernandez, Matthias Fischer, Pedro Furió-Tari, James C. Fuscoe, Florian Caimet, Stan Gaj, Jorge Gandara, Huan Gao, Weigong Ge, Yoichi Gondo, Binsheng Gong, Meihua Gong, Zhuolin Gong, Bridgett Green, Chao Guo, Lei Guo, Li Wu Guo, James Hadfield, Jan Hellemans, Sepp Hochreiter, Meiwen Jia, Min Jian, Charles D. Johnson, Suzanne Kay, Jos Kleinjans, Samir Lababidi, Shawn Levy, Quan Zhen Li, Li Li, Li Li, Peng Li, Yan Li, Haiqing Li, Jianying Li, Shiyong Li, Simon M. Lin, Francisco J. López, Xin Lu, Heng Luo, Xiwen Ma, Joseph Meehan, Dalila B. Megherbi, Nan Mei, Bing Mu, Baitang Ning, Akhilesh Pandey, Javier Pérez-Florido, Roger G. Perkins, Ryan Peters, John H. Phan, Mehdi Pirooznia, Feng Qian, Tao Qing, Lucille Rainbow, Philippe Rocca-Serra, Laure

Sambourg, Susanna Assunta Sansone, Scott Schwartz, Ruchir Shah, Jie Shen, Todd M. Smith, Oliver Stegle, Nancy Stralis-Pavese, Elia Stupka, Yutaka Suzuki, Lee T. Szkotnicki, Matthew Tinning, Bimeng Tu, Joost Van Delft, Alicia Vela-Boza, Elisa Venturini, Stephen J. Walker, Liqing Wan, Wei Wang, Jinhui Wang, Jun Wang, Eric D. Wieben, James C. Willey, Po Yen Wu, Jiekun Xuan, Yong Yang, Zhan Ye, Ye Yin, Ying Yu, Yate Ching Yuan, John Zhang, Ke K. Zhang, Wenqian Zhang, Wenwei Zhang, Yanyan Zhang, Chen Zhao, Yuanting Zheng, Yiming Zhou, Paul Zumbo, Weida Tong, David P. Kreil, Christopher E. Mason, and Leming Shi. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology*, 32(9):903–914, 9 2014. ISSN 15461696. doi: 10.1038/nbt.2957. URL `https://www.nature.com/articles/nbt.2957`.

[9] Mingxiang Teng and Rafael A Irizarry. Accounting for GC-content bias reduces systematic errors and batch effects in ChIP-seq data. *Genome research*, 27(11):1930–1938, 11 2017. ISSN 1549-5469. doi: 10.1101/gr.220673.117. URL `http://www.ncbi.nlm.nih.gov/pubmed/29025895http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5668949`.

[10] Davide Risso, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics*, 12(1):480, 12 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-480. URL `http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-480`.

[11] Michael I Love, John B Hogenesch, and Rafael A Irizarry. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nature Biotechnology*, 9 2016. ISSN 1087-0156. doi: 10.1038/nbt.3682. URL `http://www.nature.com/doifinder/10.1038/nbt.3682`.

[12] Clifford A. Meyer and X. Shirley Liu. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics*, 15(11):709–721, 11 2014. ISSN 1471-0056. doi: 10.1038/nrg3788. URL `http://www.ncbi.nlm.nih.gov/pubmed/25223782http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4473780http://www.nature.com/articles/nrg3788`.

[13] M. Ryan Corces, Jason D. Buenrostro, Beijing Wu, Peyton G. Greenside, Steven M. Chan, Julie L. Koenig, Michael P. Snyder, Jonathan K. Pritchard, Anshul Kundaje, William J. Greenleaf, Ravindra Majeti, and Howard Y. Chang. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics*, 48(10):1193–1203, 10 2016. ISSN 15461718. doi: 10.1038/ng.3646. URL `https://www.nature.com/articles/ng.3646`.

[14] Luis de la Torre-Ubieta, Jason L. Stein, Hyejung Won, Carli K. Opland, Dan Liang, Daning Lu, and Daniel H. Geschwind. The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. *Cell*, 172(1-2):289–304, 1 2018. ISSN 0092-8674. doi: 10.1016/J.CELL.2017.12.014. URL `https://www.sciencedirect.com/science/article/pii/S0092867417314940`.

[15] R. Fenouil, P. Cauchy, F. Koch, N. Descostes, J. Z. Cabeza, C. Innocenti, P. Ferrier, S. Spicuglia, M. Gut, I. Gut, and J.-C. Andrau. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Research*, 22(12):2399–2408, 12 2012. ISSN 1088-9051. doi: 10.1101/gr.138776.112. URL `http://www.ncbi.nlm.nih.gov/pubmed/23100115http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3514669http://genome.cshlp.org/cgi/doi/10.1101/gr.138776.112`.

[16] Lindsay F. Rizzardi, Peter F. Hickey, Varenka Rodriguez DiBlasi, Rakel Tryggvadóttir, Colin M. Callahan, Adrian Idrizi, Kasper D. Hansen, and Andrew P. Feinberg. Neuronal brain-region-specific DNA methylation and chromatin accessibility are associated with neuropsychiatric trait heritability. *Nature Neuroscience*, 22(2):307–316, 2 2019. ISSN 1097-6256. doi: 10.1038/s41593-018-0297-8. URL `http://www.nature.com/articles/s41593-018-0297-8`.

[17] Mary Philip, Lauren Fairchild, Liping Sun, Ellen L. Horste, Steven Camara, Mojdeh Shakiba, Andrew C. Scott, Agnes Viale, Peter Lauer, Taha Merghoub, Matthew D. Hellmann, Jedd D. Wolchok, Christina S. Leslie, and Andrea Schietinger. Chromatin states define tumour-specific T cell dysfunction and reprogramming. *Nature*, 545(7655):452–456, 5 2017. ISSN 0028-0836. doi: 10.1038/nature22367. URL `http://www.nature.com/doifinder/10.1038/nature22367`.

[18] Jake J. Reske, Mike R. Wilson, and Ronald L. Chandler. ATAC-seq normalization method can significantly affect differential accessibility analysis and interpretation. *Epigenetics & Chromatin*, 13(1):22, 12 2020. ISSN

1756-8935. doi: 10.1186/s13072-020-00342-y. URL https://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/s13072-020-00342-y.

[19] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, 2010. doi: 10.1186/gb-2010-11-3-r25. URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-3-r25.

[20] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 12 2014. ISSN 1465-6906. doi: 10.1186/s13059-014-0550-8. URL http://genomebiology.com/2014/15/12/550.

[21] K. D. Hansen, R. A. Irizarry, and Z. WU. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, 13(2):204–216, 4 2012. ISSN 1465-4644. doi: 10.1093/biostatistics/kxr054. URL http://www.ncbi.nlm.nih.gov/pubmed/22285995http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3297825https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/kxr054.

[22] Michael B. Cole, Davide Risso, Allon Wagner, David DeTomaso, John Ngai, Elizabeth Purdom, Sandrine Dudoit, and Nir Yosef. Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq. *Cell Systems*, 8(4):315–328, 4 2019. ISSN 2405-4712. doi: 10.1016/J.CELS.2019.03.010. URL https://www.sciencedirect.com/science/article/abs/pii/S2405471219300808.

[23] Diego Calderon, Michelle L. T. Nguyen, Anja Mezger, Arwa Kathiria, Fabian Müller, Vinh Nguyen, Ninnia Lescano, Beijing Wu, John Trombetta, Jessica V. Ribado, David A. Knowles, Ziyue Gao, Franziska Blaeschke, Audrey V. Parent, Trevor D. Burt, Mark S. Anderson, Lindsey A. Criswell, William J. Greenleaf, Alexander Marson, and Jonathan K. Pritchard. Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nature Genetics*, 51(10):1494–1505, 10 2019. ISSN 1061-4036. doi: 10.1038/s41588-019-0505-9. URL http://www.nature.com/articles/s41588-019-0505-9.

[24] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–40, 1 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp616. URL http://www.ncbi.nlm.nih.gov/pubmed/19910308http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2796818.

[25] James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in {mRNA-Seq} experiments. *BMC Bioinformatics*, 11:94, 2 2010.

[26] Dan Liang, Angela L. Elwell, Nil Aygün, Michael J. Lafferty, Oleh Krupa, Kerry E. Cheek, Kenan P. Courtney, Marianna Yusupova, Melanie E. Garrett, Allison Ashley-Koch, Gregory E. Crawford, Michael I. Love, Luis de la Torre-Ubieta, Daniel H. Geschwind, and Jason L. Stein. Cell-type specific effects of genetic variation on chromatin accessibility during human neuronal differentiation. *bioRxiv*, page 2020.01.13.904862, 1 2020. doi: 10.1101/2020.01.13.904862. URL https://www.biorxiv.org/content/10.1101/2020.01.13.904862v1.

[27] Stephanie C Hicks, Kwame Okrah, Joseph N Paulson, John Quackenbush, Rafael A Irizarry, and Héctor Corrada Bravo. Smooth quantile normalization. *Biostatistics*, 19(2):185–198, 4 2018.

[28] Julien Bryois, Melanie E. Garrett, Lingyun Song, Alexias Safi, Paola Giusti-Rodriguez, Graham D. Johnson, Annie W. Shieh, Alfonso Buil, John F. Fullard, Panos Roussos, Pamela Sklar, Schahram Akbarian, Vahram Haroutunian, Craig A. Stockmeier, Gregory A. Wray, Kevin P. White, Chunyu Liu, Timothy E. Reddy, Allison Ashley-Koch, Patrick F. Sullivan, and Gregory E. Crawford. Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. *Nature Communications*, 9(1):3121, 12 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-05379-y. URL http://www.nature.com/articles/s41467-018-05379-y.

[29] John F Fullard, Mads E Hauberg, Jaroslav Bendl, Gabor Egervari, Maria-Daniela Cirnaru, Sarah M Reach, Jan Motl, Michelle E Ehrlich, Yasmin L Hurd, and Panos Roussos. An atlas of chromatin accessibility in the adult human brain. *Genome research*, 28(8):1243–1252, 8 2018. ISSN 1549-5469. doi: 10.1101/gr.232488.117. URL http://www.ncbi.nlm.nih.gov/pubmed/29945882http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6071637.

[30] Chuanyu Liu, Mingyue Wang, Xiaoyu Wei, Liang Wu, Jiangshan Xu, Xi Dai, Jun Xia, Mengnan Cheng, Yue Yuan, Pengfan Zhang, Jiguang Li, Taiqing Feng, Ao Chen, Wenwei Zhang, Fang Chen, Zhouchun Shang, Xiuqing Zhang, Brock A. Peters, and Longqi Liu. An ATAC-seq atlas of chromatin accessibility in mouse tissues. *Scientific Data*, 6(1):65, 12 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0071-0. URL `http://www.nature.com/articles/s41597-019-0071-0`.

[31] Daniel P Murphy, Andrew EO Hughes, Karen A Lawrence, Connie A Myers, and Joseph C Corbo. Cis-regulatory basis of sister cell type divergence in the vertebrate retina. *eLife*, 8, 10 2019. ISSN 2050-084X. doi: 10.7554/eLife.48216. URL `https://elifesciences.org/articles/48216`.

[32] Lindsay Rizzardi, Peter Hickey, Varenka Rodriguez, Rakel Tryggvadottir, Colin Callahan, Adrian Idrizi, Kasper Hansen, and Andrew P Feinberg. Neuronal brain region-specific DNA methylation and chromatin accessibility are associated with neuropsychiatric disease heritability. *bioRxiv*, page 120386, 3 2017. doi: 10.1101/120386. URL `https://www.biorxiv.org/content/early/2017/03/24/120386`.

[33] Luke C. Gandolfo and Terence P. Speed. RLE plots: Visualizing unwanted variation in high dimensional data. *PLOS ONE*, 13(2):e0191629, 2 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0191629. URL `https://dx.plos.org/10.1371/journal.pone.0191629`.

[34] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.*, 32:896, 8 2014.

[35] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of RNA-seq data. *BMC bioinformatics*, 14:91, 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-91. URL `http://www.ncbi.nlm.nih.gov/pubmed/23497356http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3608160`.

[36] Yaov Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. URL `https://www.jstor.org/stable/2346101?seq=1#page_scan_tab_contents`.

[37] William M. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850, 12 1971. ISSN 0162-1459. doi: 10.1080/01621459.1971.10482356. URL `http://www.tandfonline.com/doi/abs/10.1080/01621459.1971.10482356`.

[38] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 12 1985. ISSN 0176-4268. doi: 10.1007/BF01908075. URL `http://link.springer.com/10.1007/BF01908075`.

[39] Sandrine Dudoit, Yee Hwa Yang, Matthew J. Callow, and Terence P. Speed. No Title. *Statistica Sinica*, 12 (1):111–139, 2002.

[40] Constantin Ahlmann-Eltze and Wolfgang Huber. glmGamPoi: Fitting Gamma-Poisson Generalized Linear Models on Single Cell Count Data. *bioRxiv*, page 2020.08.13.249623, 9 2020. doi: 10.1101/2020.08.13.249623. URL `https://www.biorxiv.org/content/10.1101/2020.08.13.249623v2`.

[41] Yuval Benjamini and Terence P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10):e72–e72, 5 2012. ISSN 1362-4962. doi: 10.1093/nar/gks001. URL `https://academic.oup.com/nar/article/40/10/e72/2411059`.

[42] Qunhua Li, James B. Brown, Haiyan Huang, and Peter J. Bickel. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5(3):1752–1779, 9 2011. ISSN 1932-6157. doi: 10.1214/11-AOAS466. URL `http://projecteuclid.org/euclid.aoas/1318514284`.

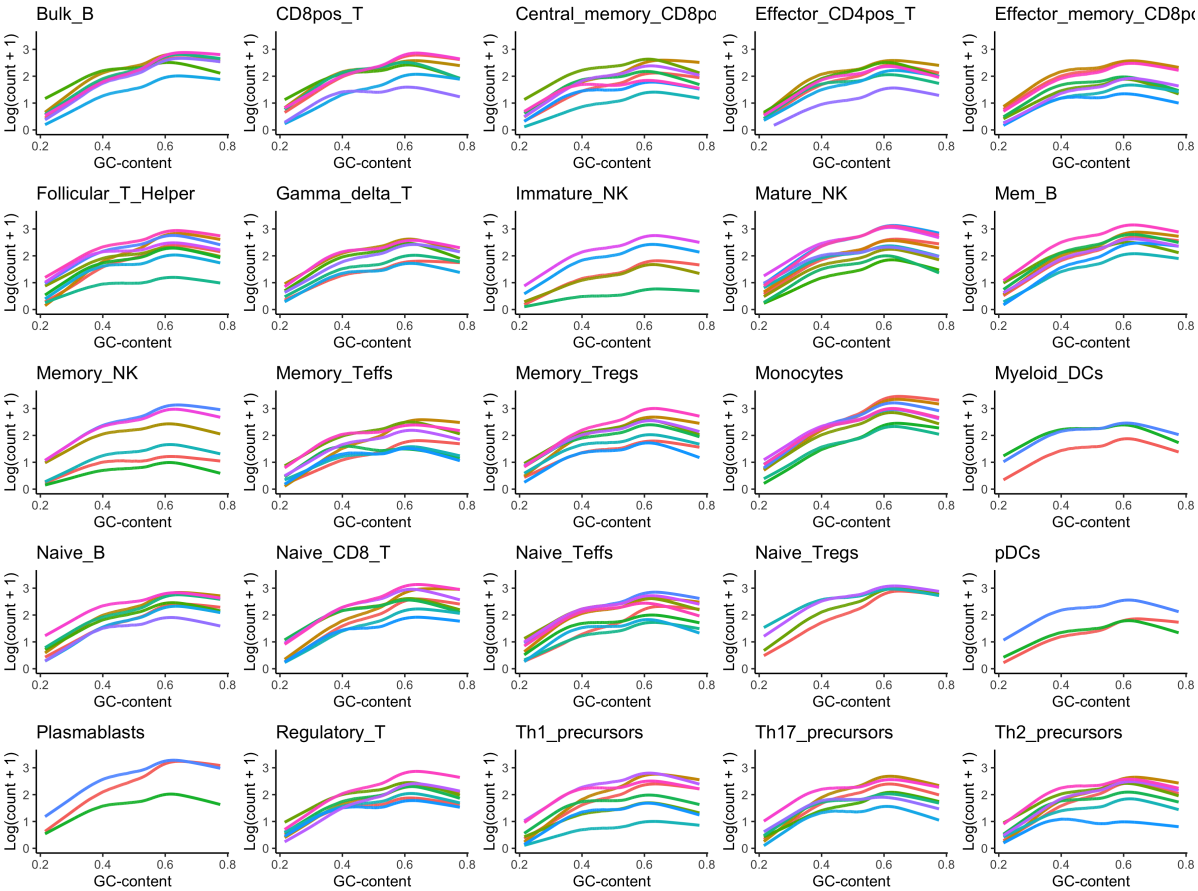
[43] B M Bolstad, R A Irizarry, M Astrand, and T P Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)*, 19(2):185–93, 1 2003. ISSN 1367-4803. URL `http://www.ncbi.nlm.nih.gov/pubmed/12538238`.

[44] Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 19(4):562–578, 10 2018.

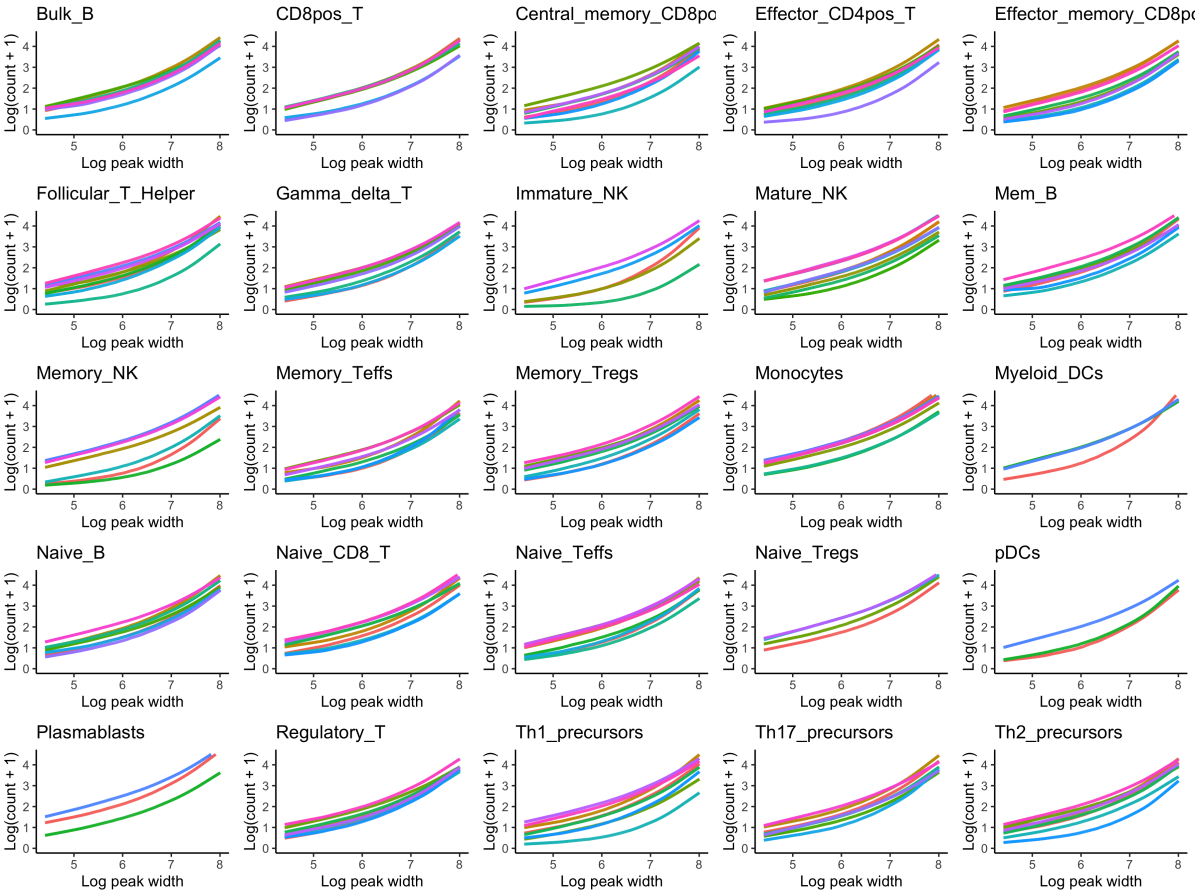
[45] Kasper D Hansen, Benjamin Langmead, and Rafael A Irizarry. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome biology*, 13(10):R83, 1 2012. ISSN 1465-6914. doi: 10.1186/gb-2012-13-10-r83. URL `http://genomebiology.com/2012/13/10/R83`.

[46] Zhijin Wu and Martin J Aryee. Subset quantile normalization using negative control features. *Journal of computational biology : a journal of computational molecular cell biology*, 17(10):1385–95, 10 2010. ISSN 1557-8666. doi: 10.1089/cmb.2010.0049. URL http://www.ncbi.nlm.nih.gov/pubmed/20976876http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3122888.

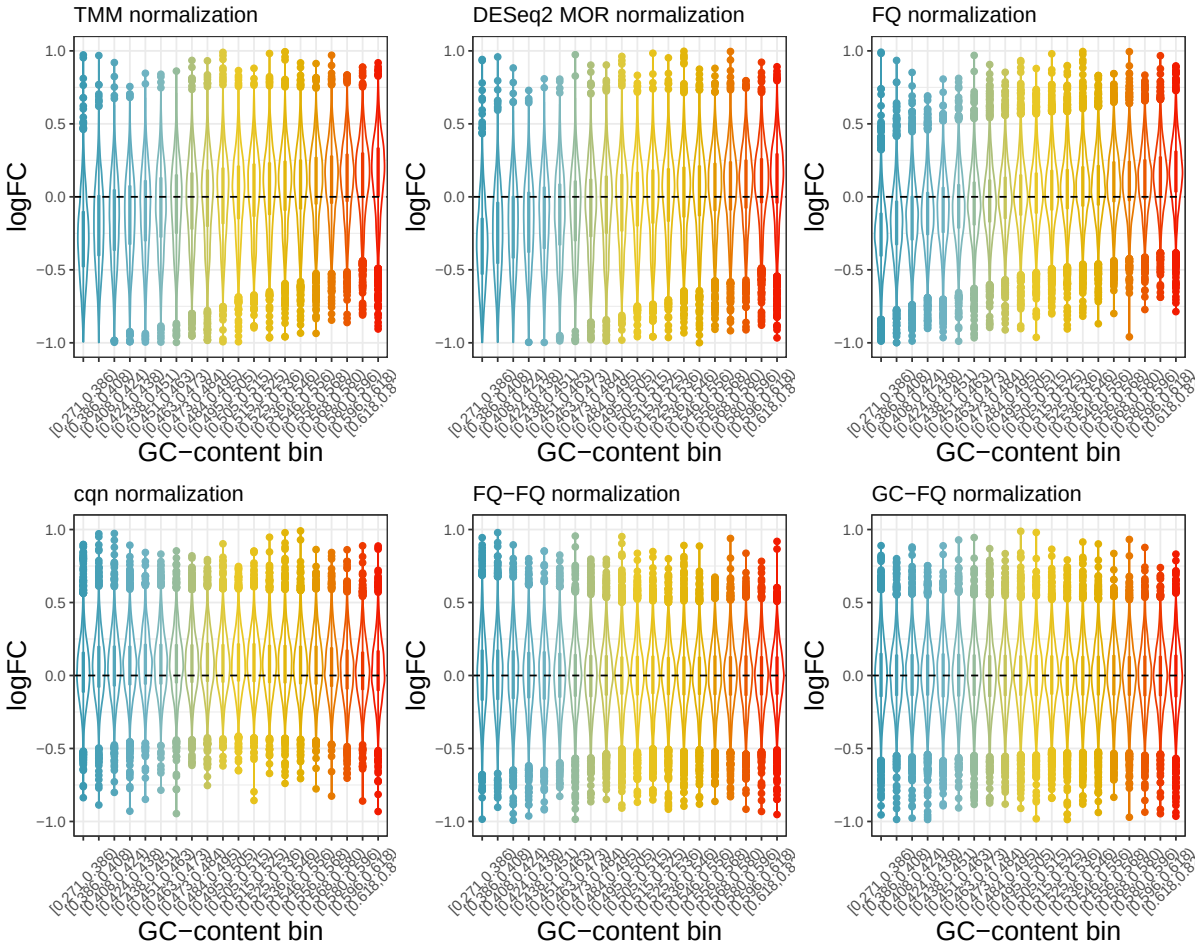
Supplementary Figures



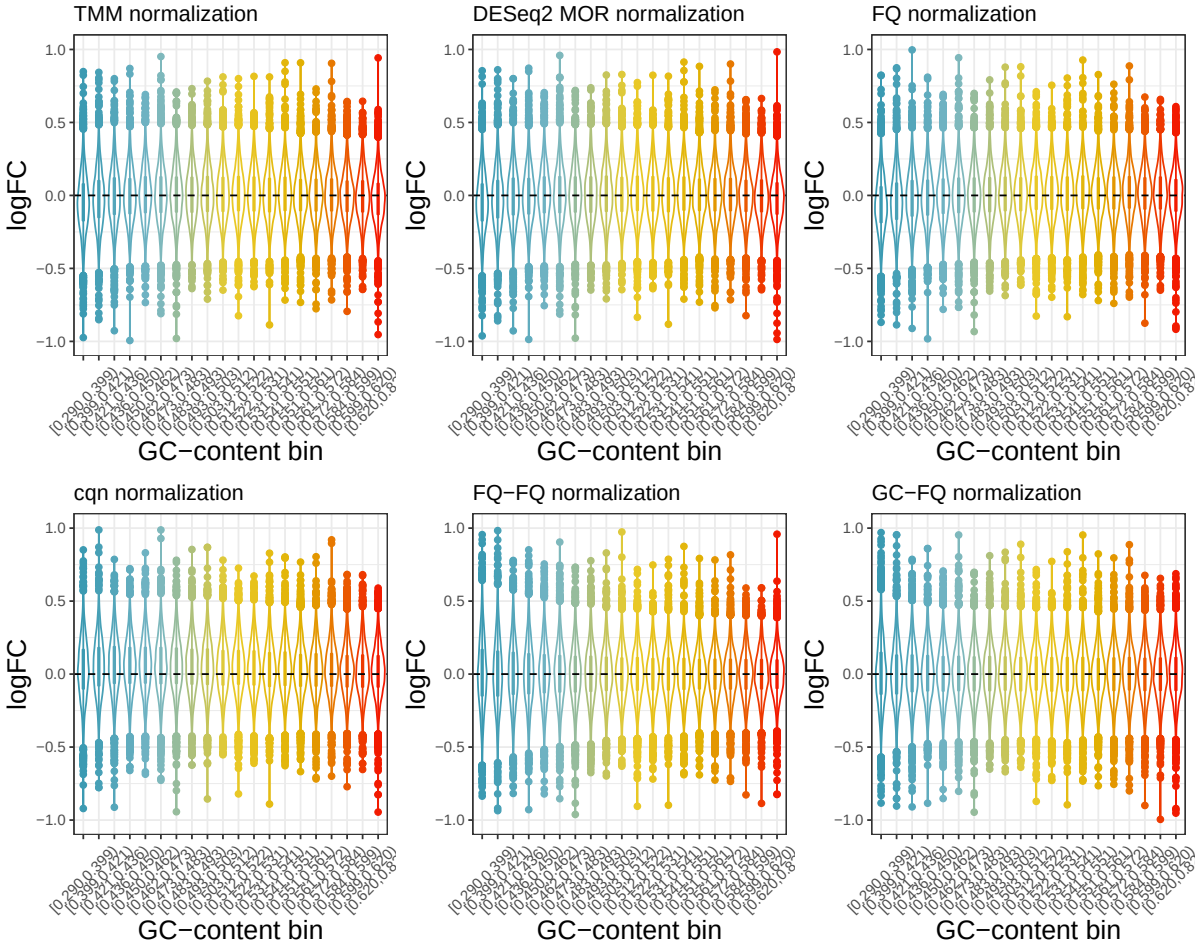
Supplementary Figure 1: *Lowess fits for the log accessibility count as a function of GC-content for the dataset from Calderon et al.*[23].



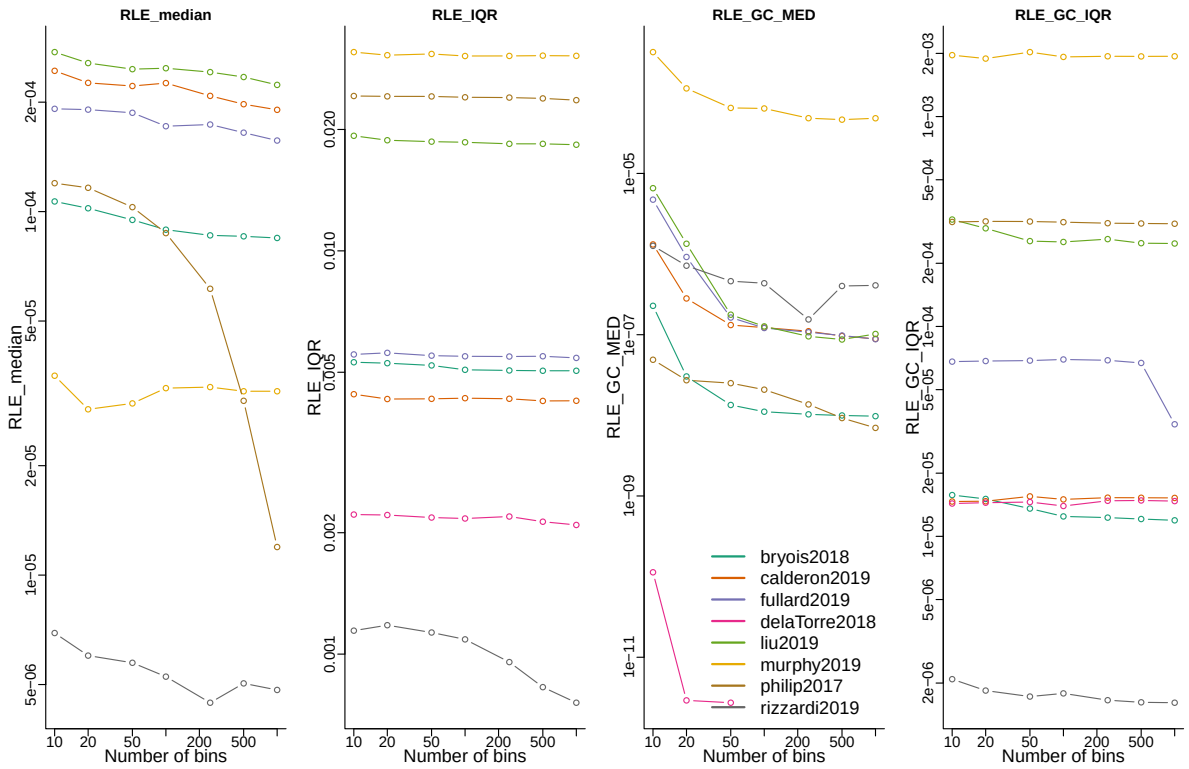
Supplementary Figure 2: *Lowess fits for the log accessibility count as a function of peak width for the dataset from Calderon et al.*[23].



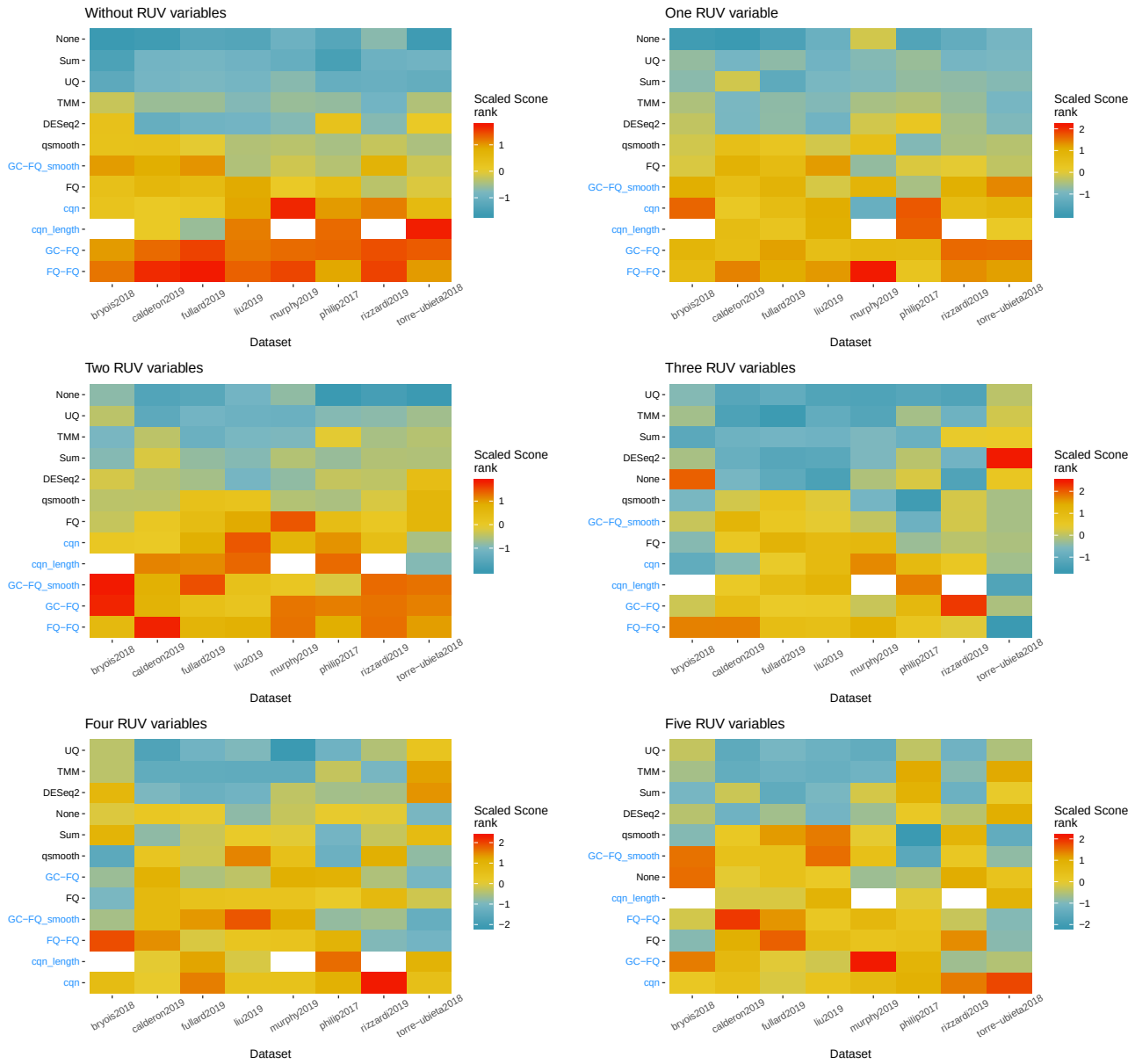
Supplementary Figure 3: *Fold-change bias for a mock comparison of stimulated mature natural killer cells in the dataset from Calderon et al.*[23].



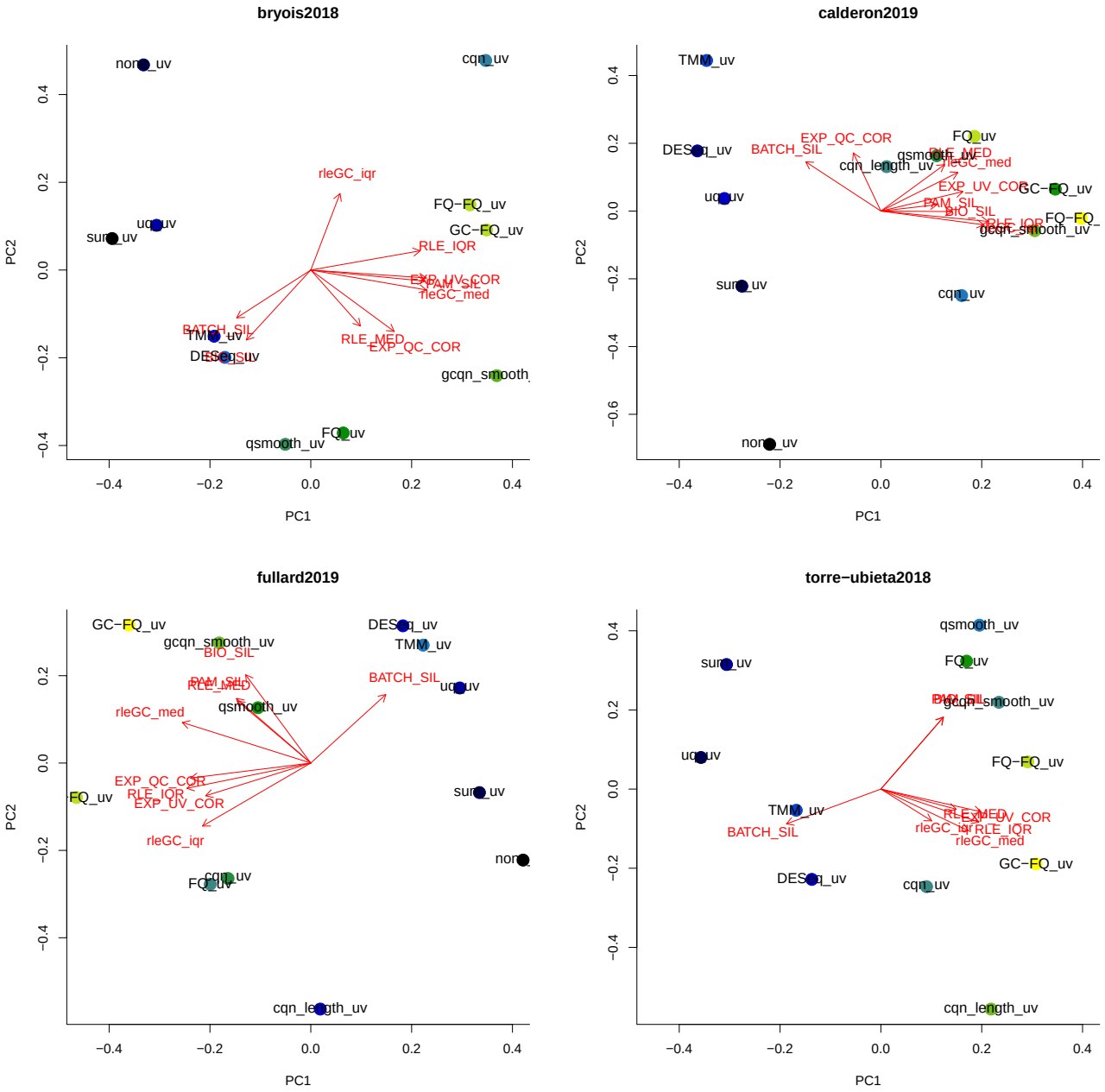
Supplementary Figure 4: *Fold-change bias for a mock comparison of stimulated nonocyte cells in the dataset from Calderon et al.*[23].



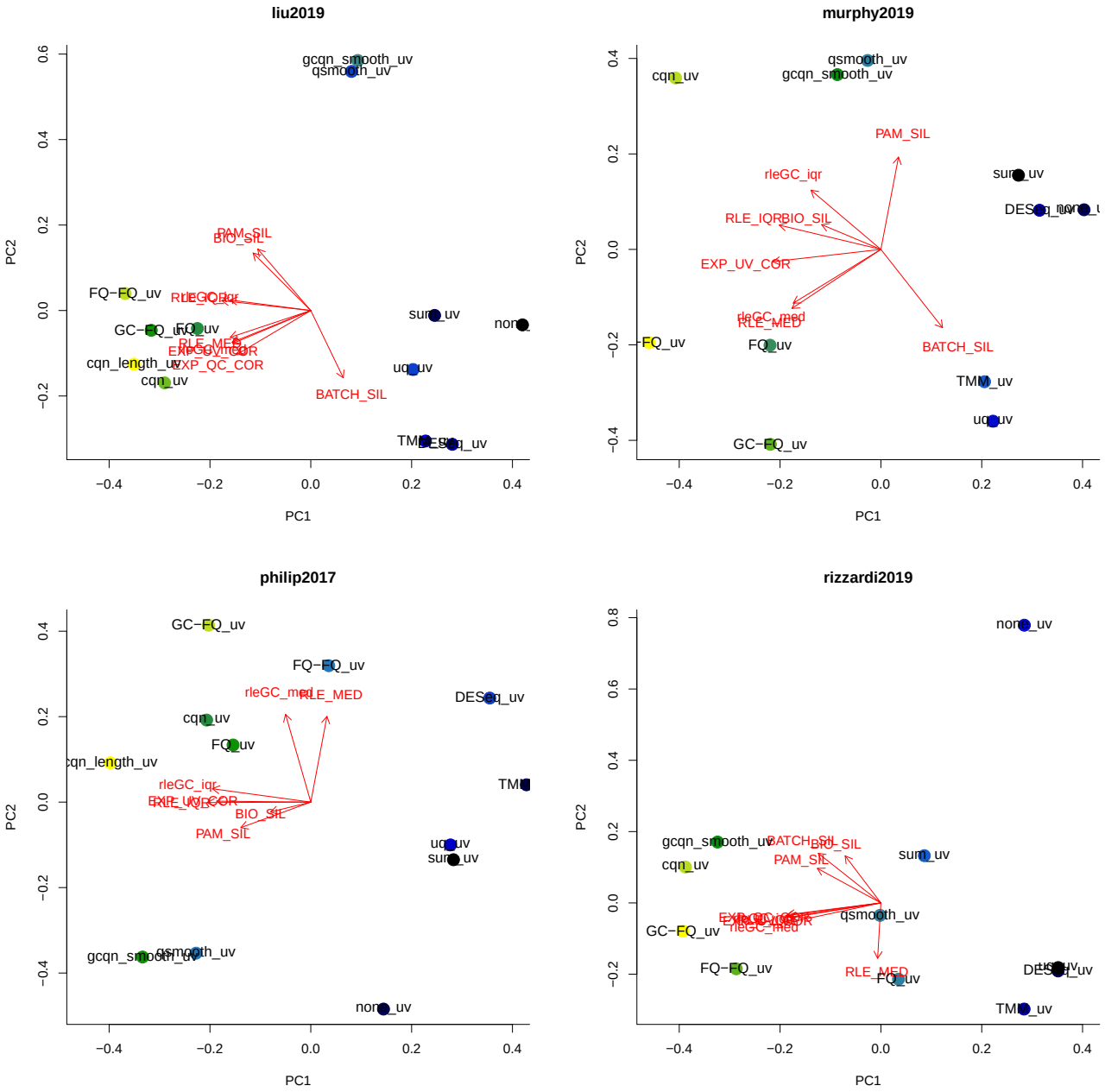
Supplementary Figure 5: *Evaluating the number of bins for GC-FQ normalization.* Normalization performance measures based on relative log expression (RLE) values (as defined in Methods) are shown as a function of the number of bins used for GC-FQ normalization, across all eight datasets. Each dataset was normalized using 10, 20, 50, 100, 250, 500, and 1,000 bins. The number of peaks across datasets ranges from $\sim 62,000$ to $\sim 300,000$. Using 50 bins results in a good trade-off between having a reasonable amount of data in each bin and removing global between-sample differences and GC-content bias. Note that the dataset from de la Torre-Ubieta et al. [14] is not displayed in the RLE_median panel as well as for 100+ bins in the RLE_GC_MED panel, since its RLE values are equal to zero. Interestingly, this is also the only dataset that corresponds to technical, rather than biological, replication.



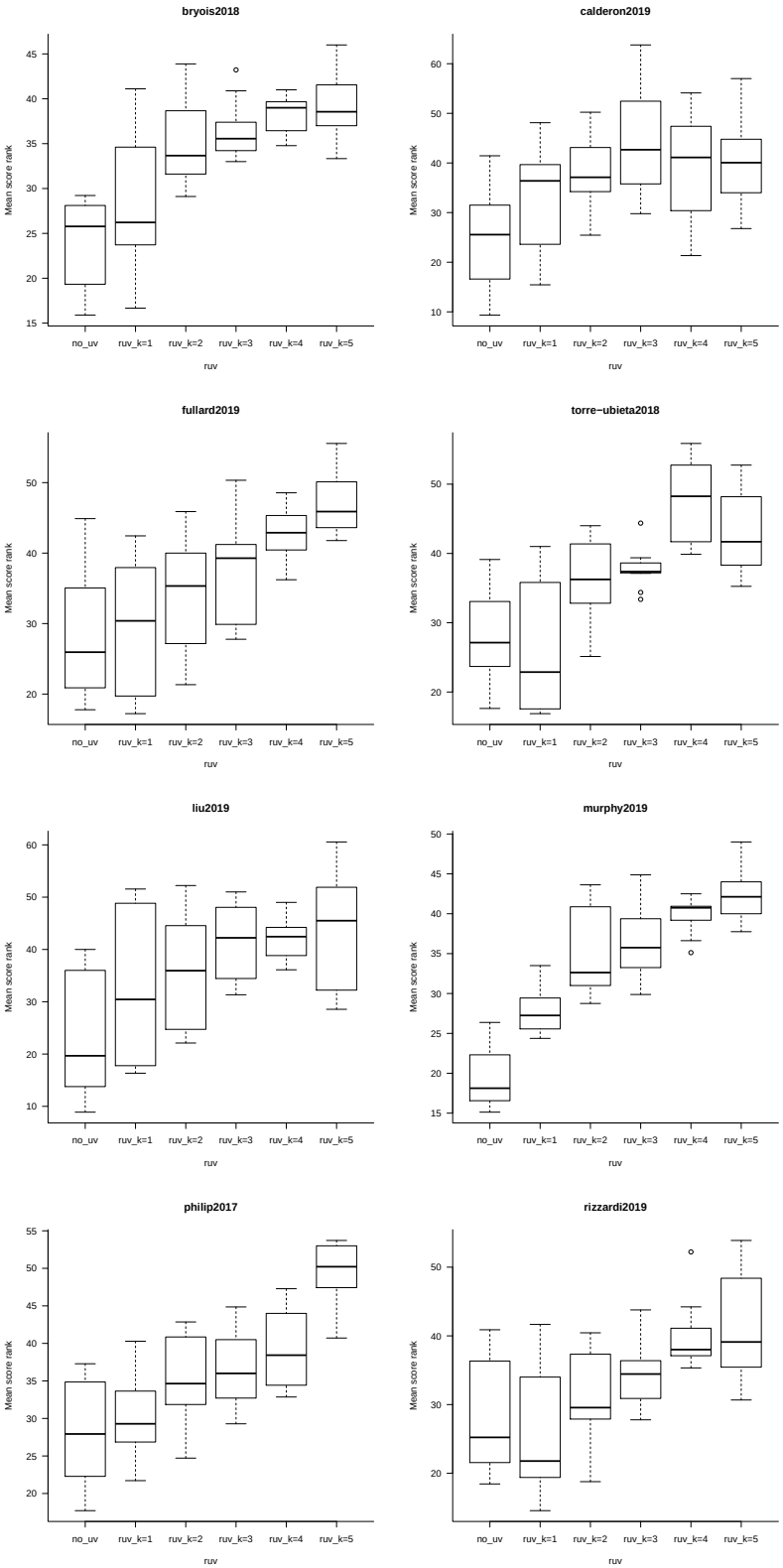
Supplementary Figure 6: *Benchmark of normalization methods across eight public ATAC-seq datasets using* scone. Twelve different normalization methods are applied to eight public ATAC-seq datasets. We evaluate the appropriateness of each normalization in each dataset using scone. The color of each cell denotes its scaled average scone rank (see Methods), where a high score corresponds to successful normalization. Methods are ordered according to their average score across all datasets. The different panels correspond to removing additional factors of unwanted variation using RUVSeq after normalization, as noted in the title.



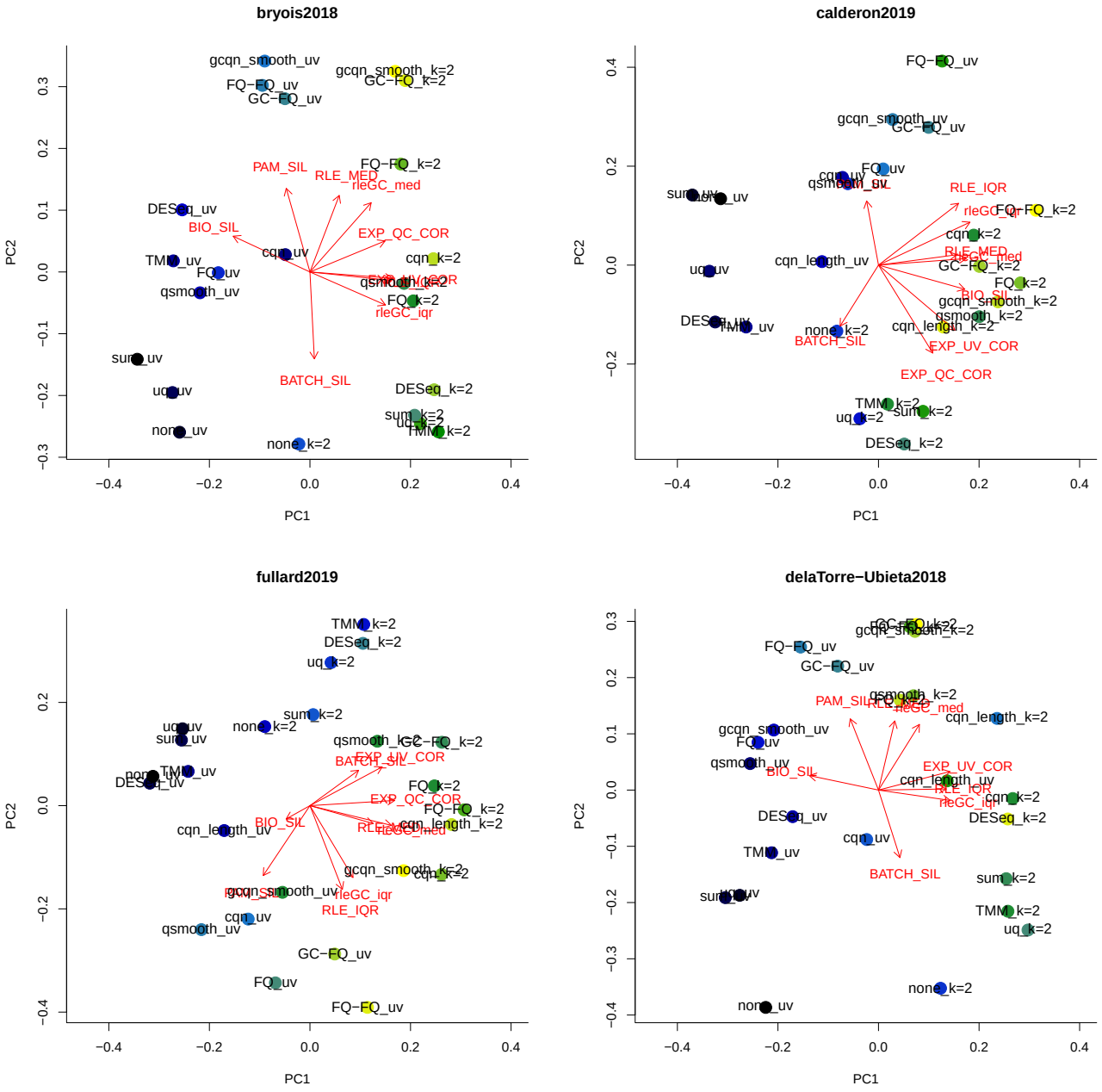
Supplementary Figure 7: *Biplots of* scone *normalization performance measures, without removing unwanted variation with* RUVSeq. Each normalization method's point in the biplot is colored according to its performance measure, from dark blue (bad normalization) to bright yellow (good normalization).



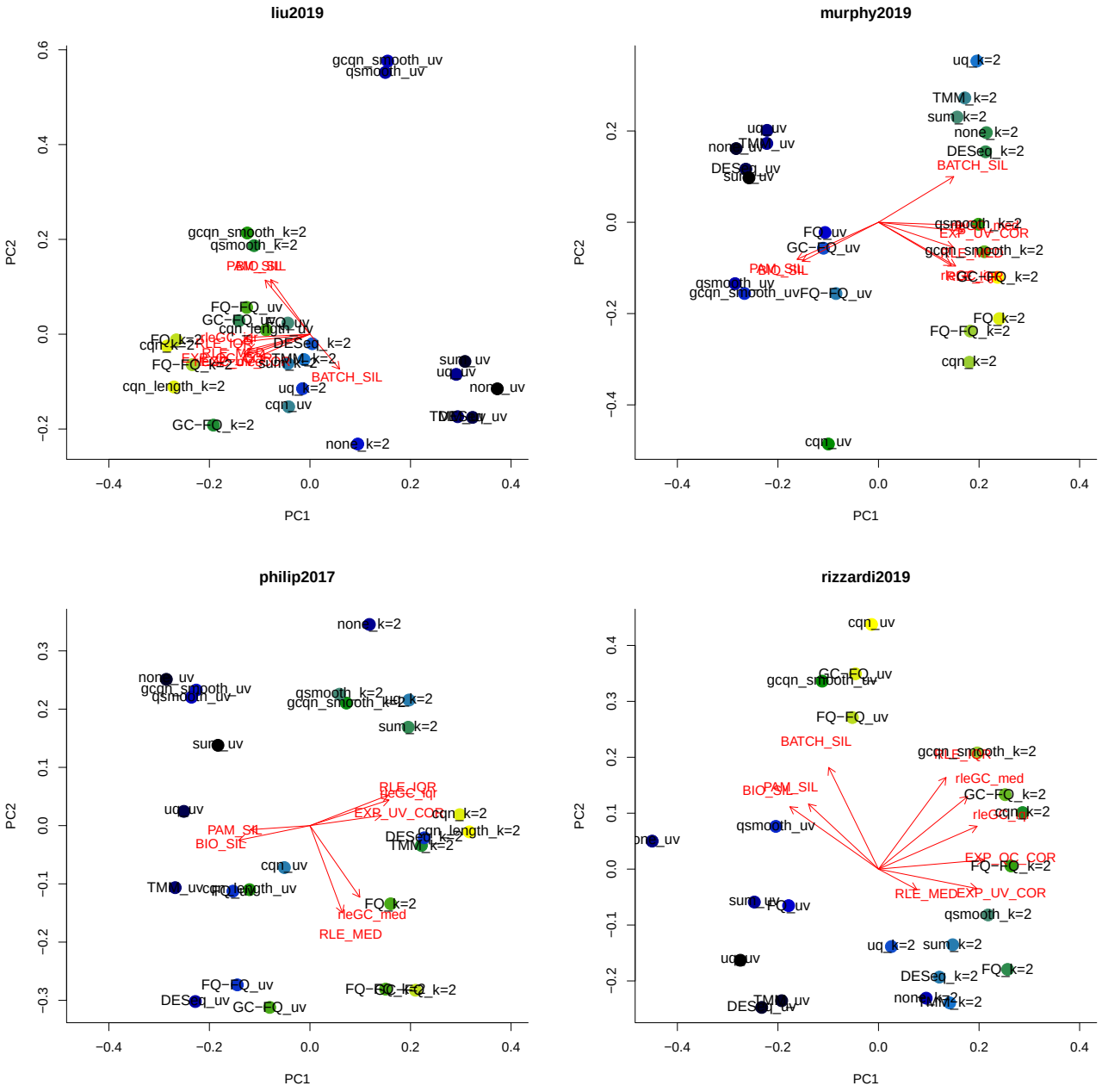
Supplementary Figure 8: *Biplots of* scone *normalization performance measures, without removing unwanted variation with* RUVSeq*.* Each normalization method's point in the biplot is colored according to its performance measure, from dark blue (bad normalization) to bright yellow (good normalization).



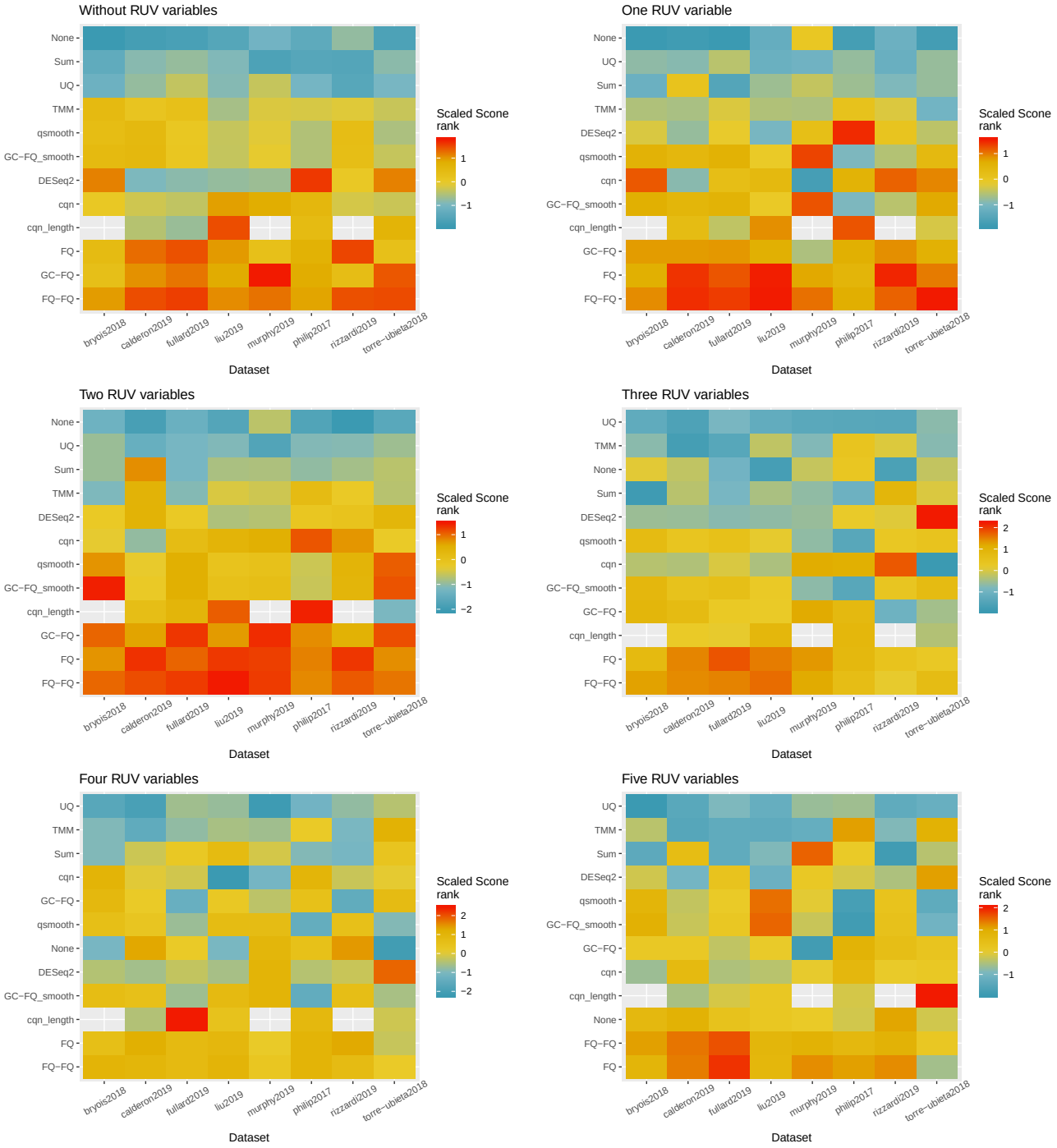
Supplementary Figure 9: *Boxplots of average* scone *ranks for all* 12 *normalization methods, ordered according to the number of factors of unwanted variation removed with* RUVSeq. A high score denotes good normalization.



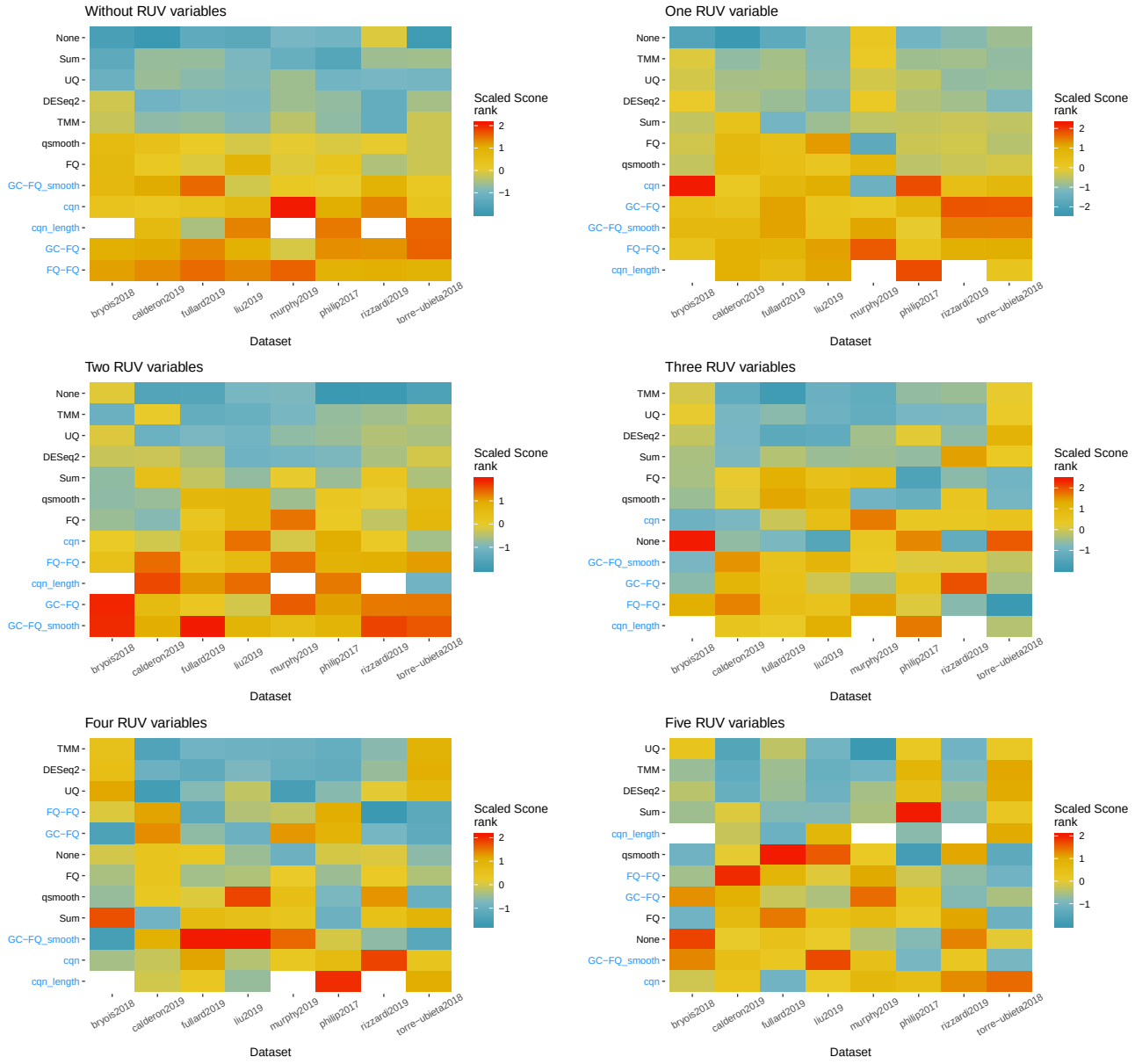
Supplementary Figure 10: *Biplots of* scone *normalization performance measures, with and without removing unwanted variation with* RUVSeq*.* Each normalization method's point in the biplot is colored according to its performance measure, from dark blue (bad normalization) to bright yellow (good normalization). The `_uv` suffix denotes not removing unwanted variation, while the `_k=2` suffix denotes removing 2 factors of unwanted variation using RUVSeq[34].



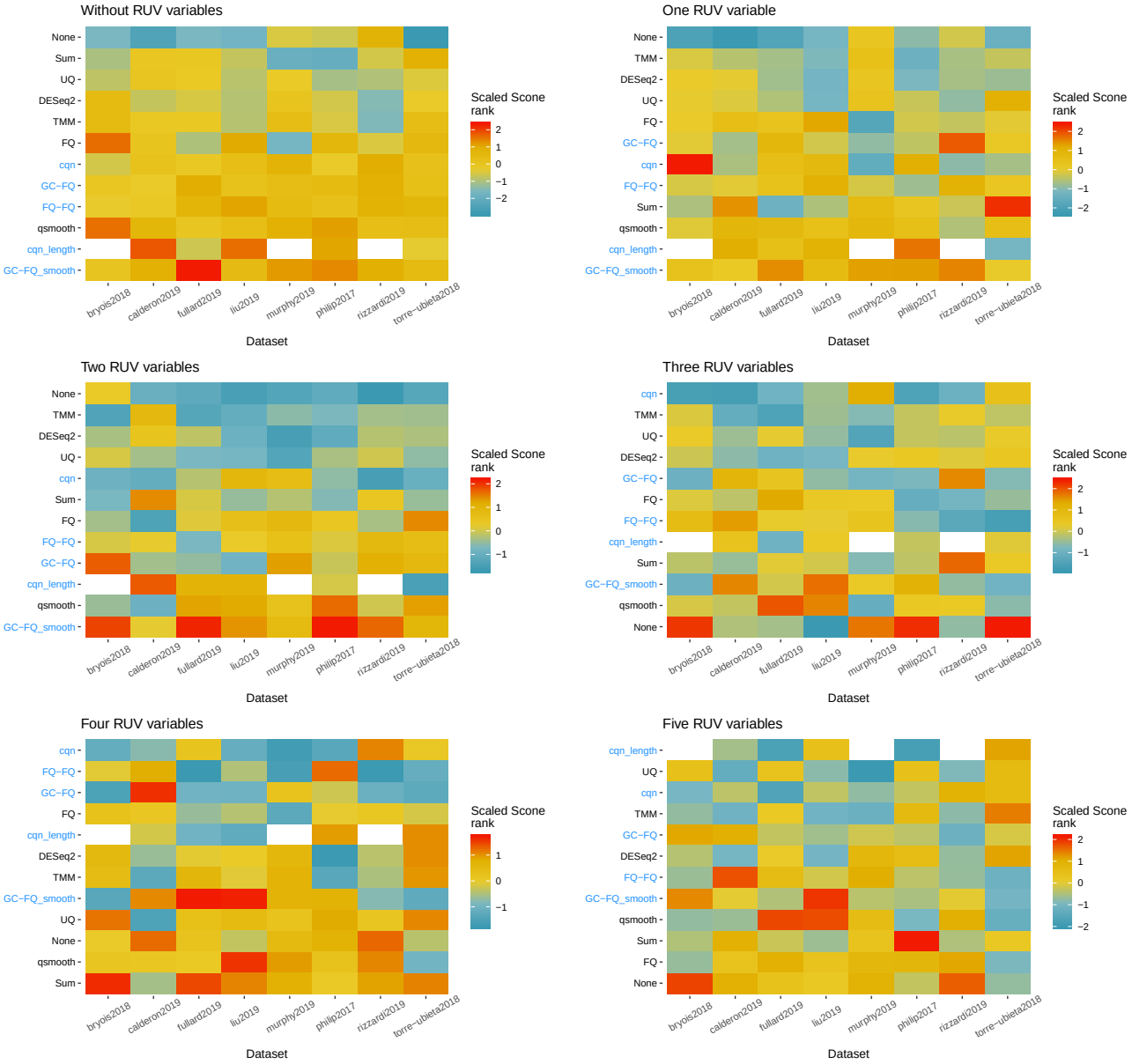
Supplementary Figure 11: *Biplots of* scone *normalization performance measures, with and without removing unwanted variation with* RUVSeq*.* Each normalization method's point in the biplot is colored according to its performance measure, from dark blue (bad normalization) to bright yellow (good normalization). The `_uv` suffix denotes not removing unwanted variation, while the `_k=2` suffix denotes removing 2 factors of unwanted variation using RUVSeq[34].



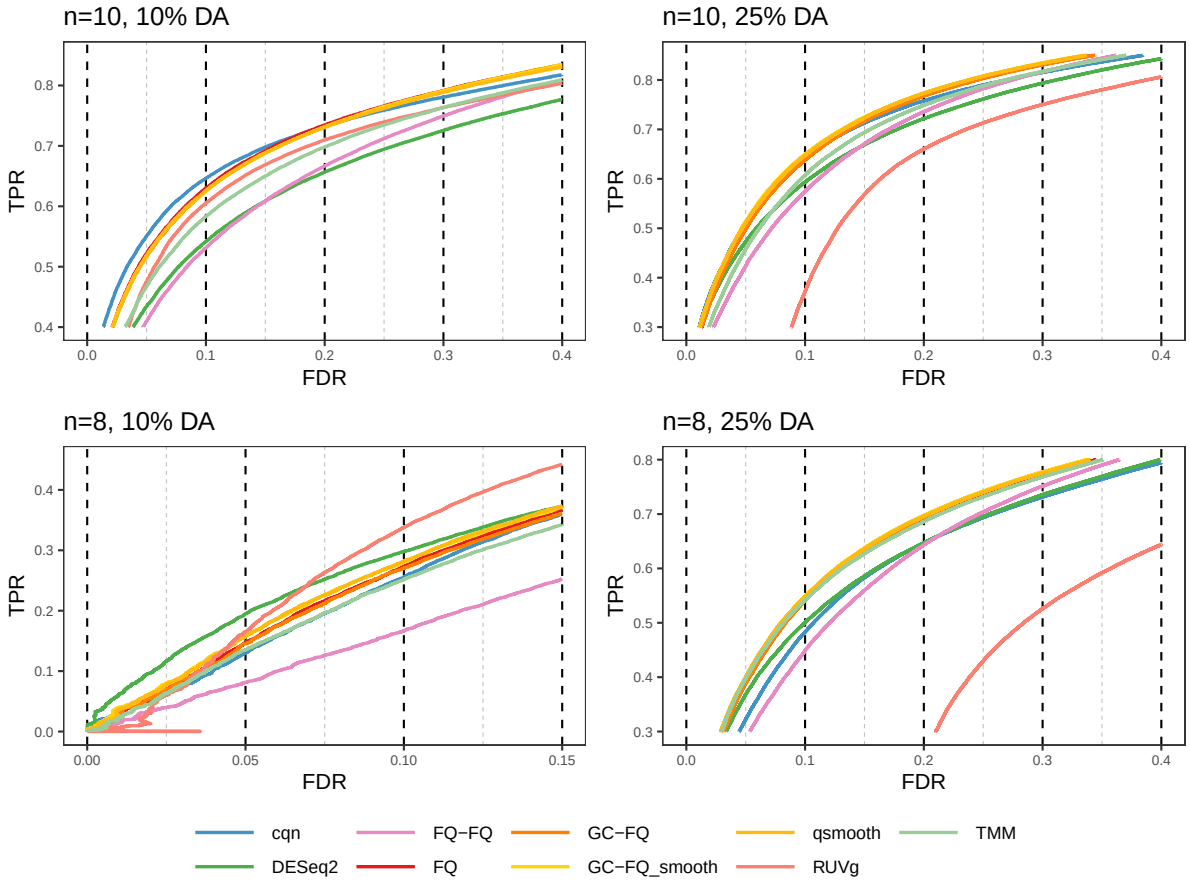
Supplementary Figure 12: *Benchmark of normalization methods across eight public ATAC-seq datasets using* scone*, after removing GC-content effects by permuting GC-content values accross peaks.*



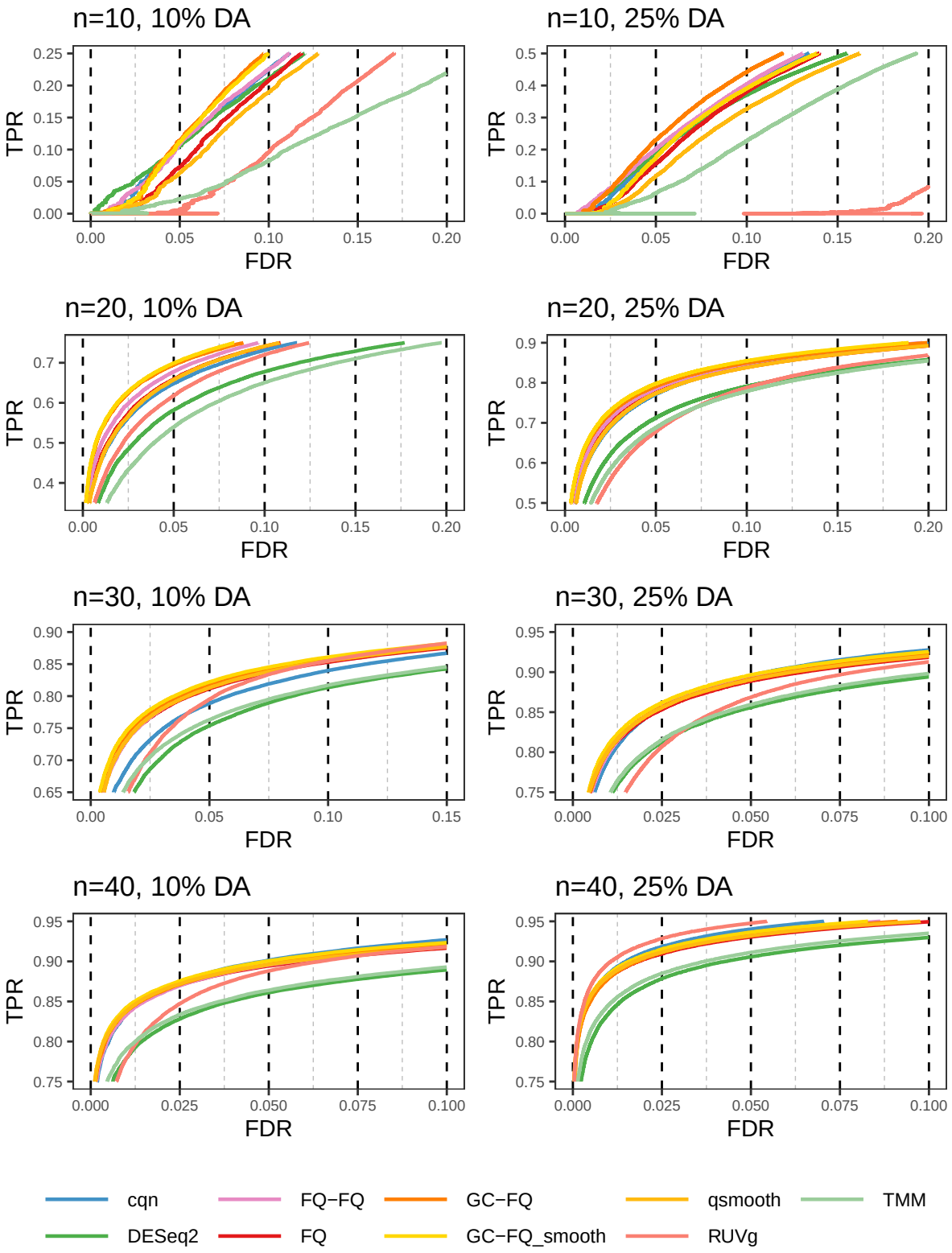
Supplementary Figure 13: *Benchmark of normalization methods across eight public ATAC-seq datasets using* scone*, without the two criteria that assess GC-content bias.*



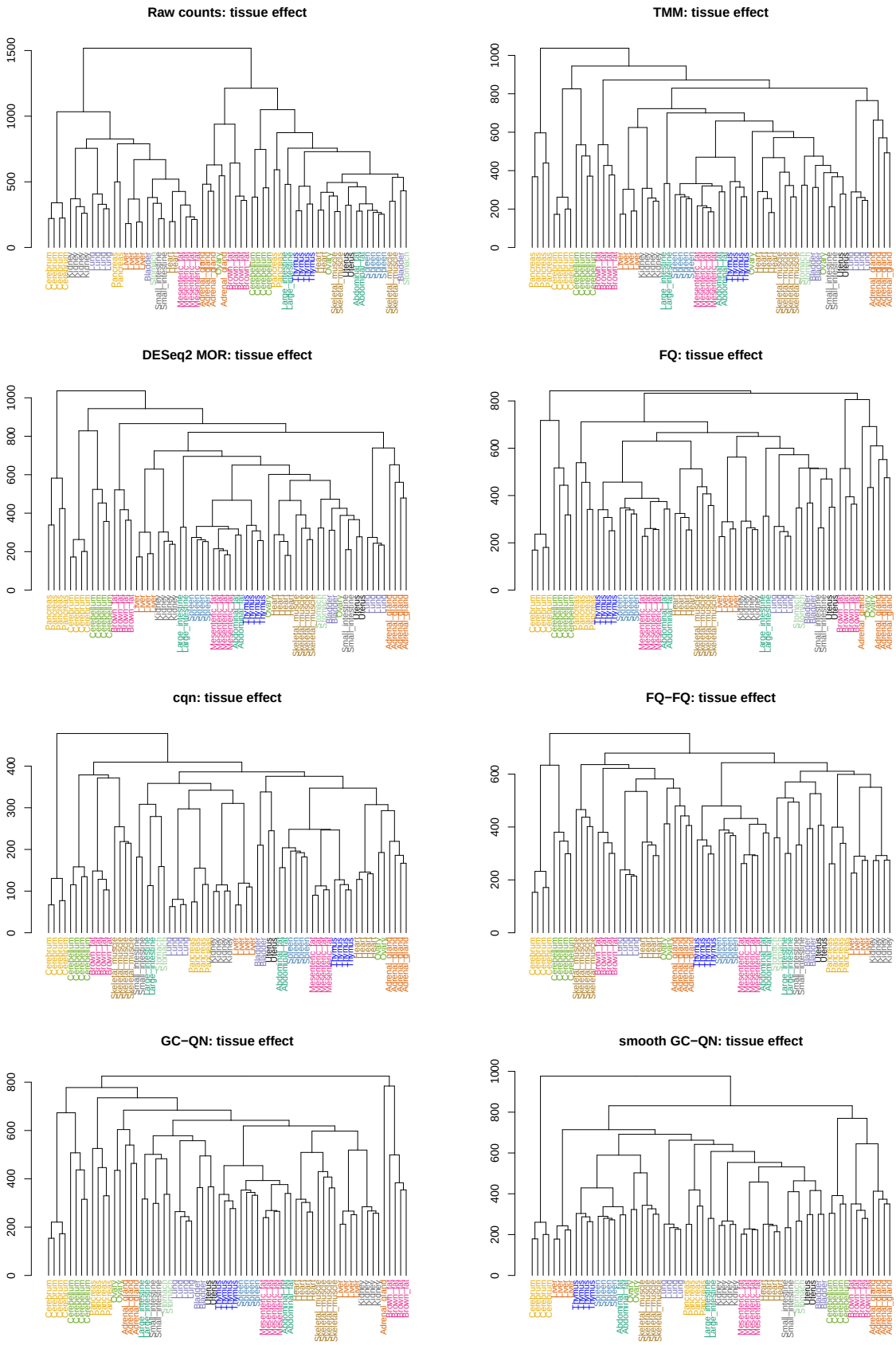
Supplementary Figure 14: *Benchmark of normalization methods across eight public ATAC-seq datasets using* scone*, without the four RLE-based evaluation criteria.*



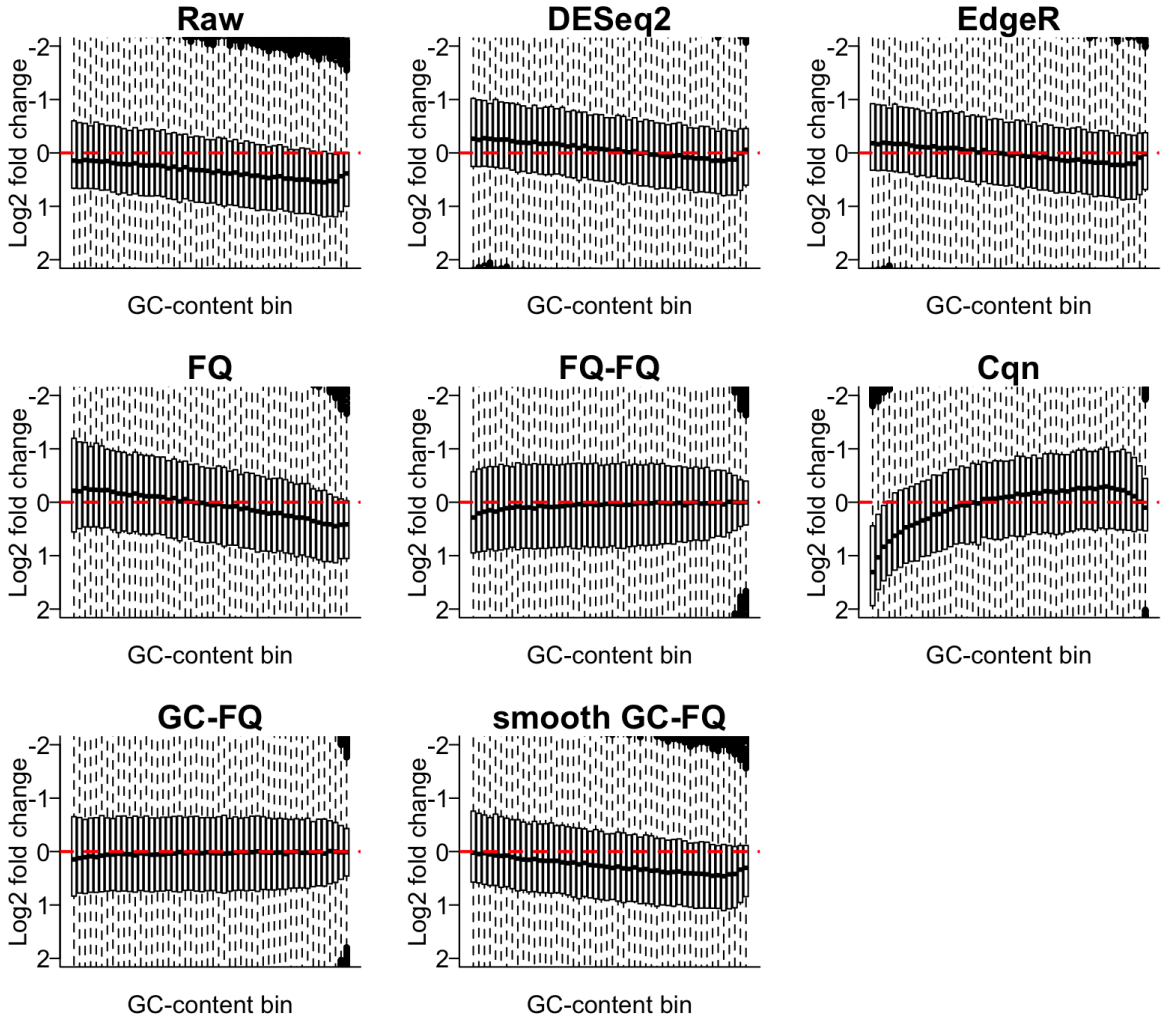
Supplementary Figure 15: *FDR-TPR performance curves for simulated datasets based on the olfactory epithelium null dataset.* Each panel represents an average FDR-TPR performance curve for each method over 14 simulated datasets based on the olfactory epithelium null dataset described in the Methods section. The sample size and proportion of DA features are noted in the title of each panel. Note that for the 5-vs.-5 comparisons there is no variability contributed by sample assignments to mock groups across the iterations, since the dataset contains 10 samples and the sample assignment was fixed prior to the simulation. The axis limits are cropped to focus on the relevant regions that discriminate between methods, separately for each evaluation.



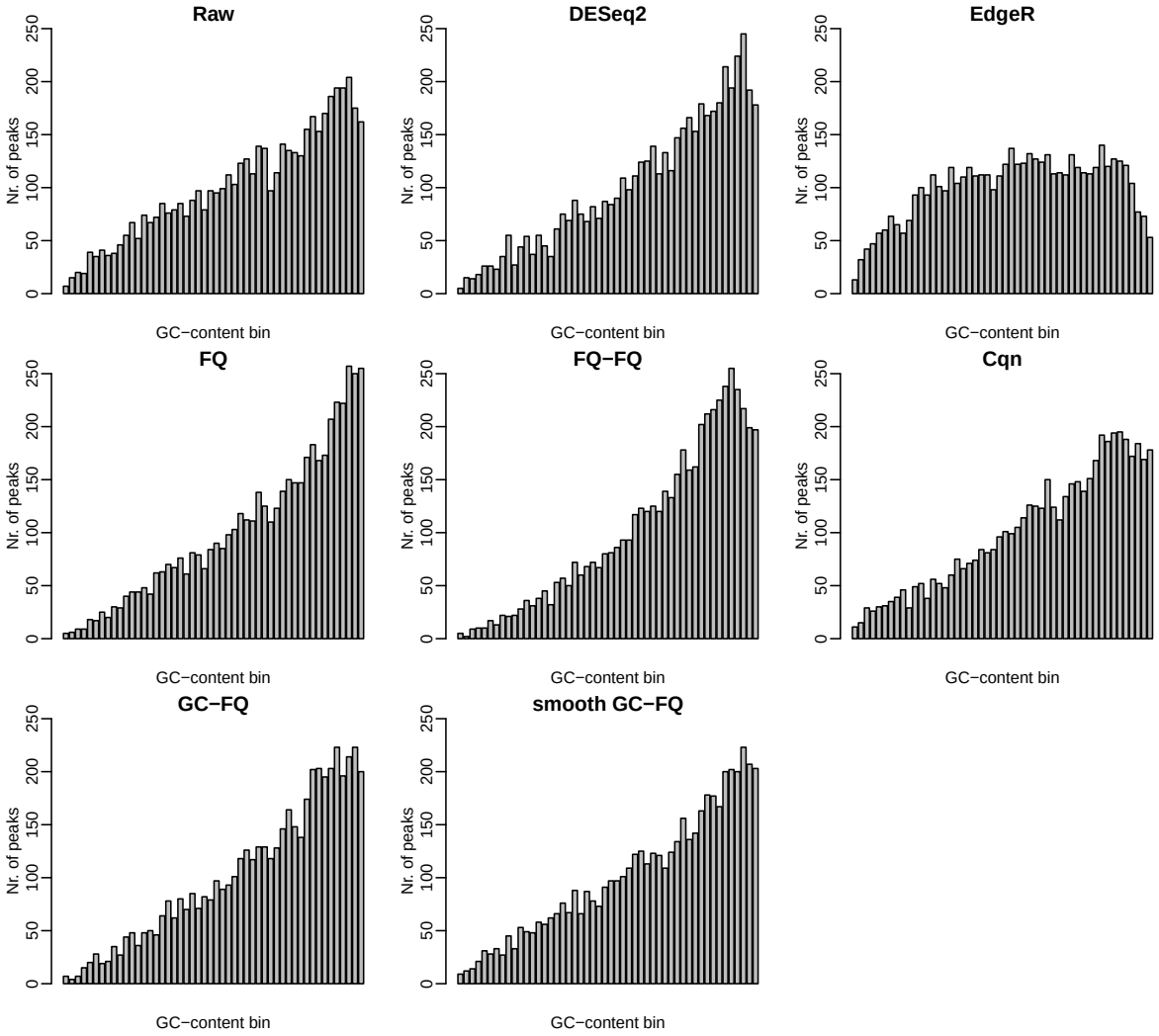
Supplementary Figure 16: *FDR-TPR performance curves for simulated datasets based on the dataset from Calderon et al.*[23]. Each panel represents an average FDR-TPR performance curve over 14 simulated datasets based on the dataset from Calderon et al.[23]. The total sample size and proportion of DA features are noted in the title of each panel. The axis limits are cropped to focus on the relevant regions that discriminate between methods, separately for each evaluation.



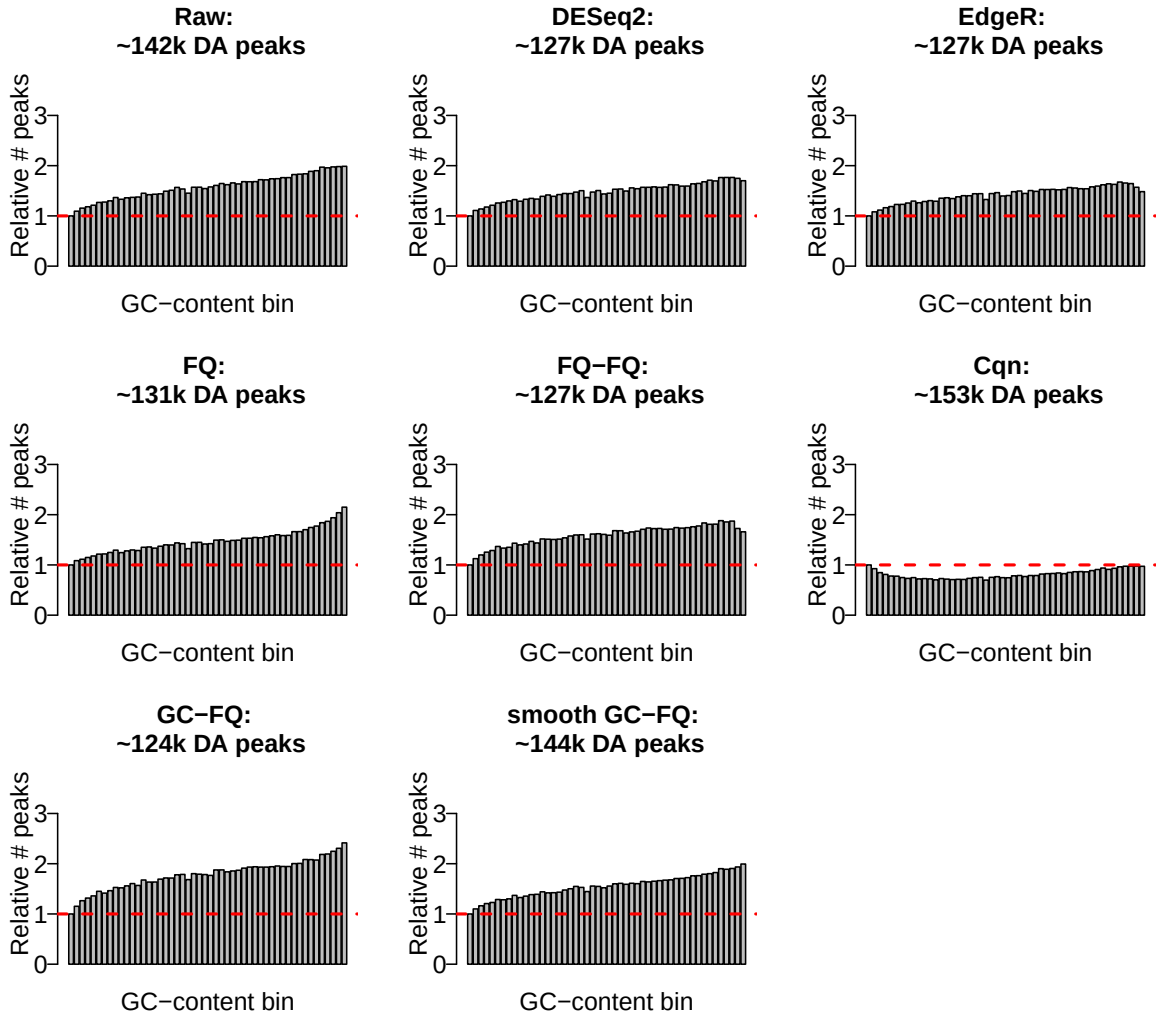
Supplementary Figure 17: *Hierarchical clustering of the dataset from Liu et al.*[30]. For each method, the normalized counts are log-transformed and pairwise Euclidean distances between samples are calculated. Agglomerative hierarchical trees are constructed using complete linkage. Samples are labeled and colored according to tissue type. The y-axis represents the Euclidean distance at which two (groups of) samples are agglomerated.



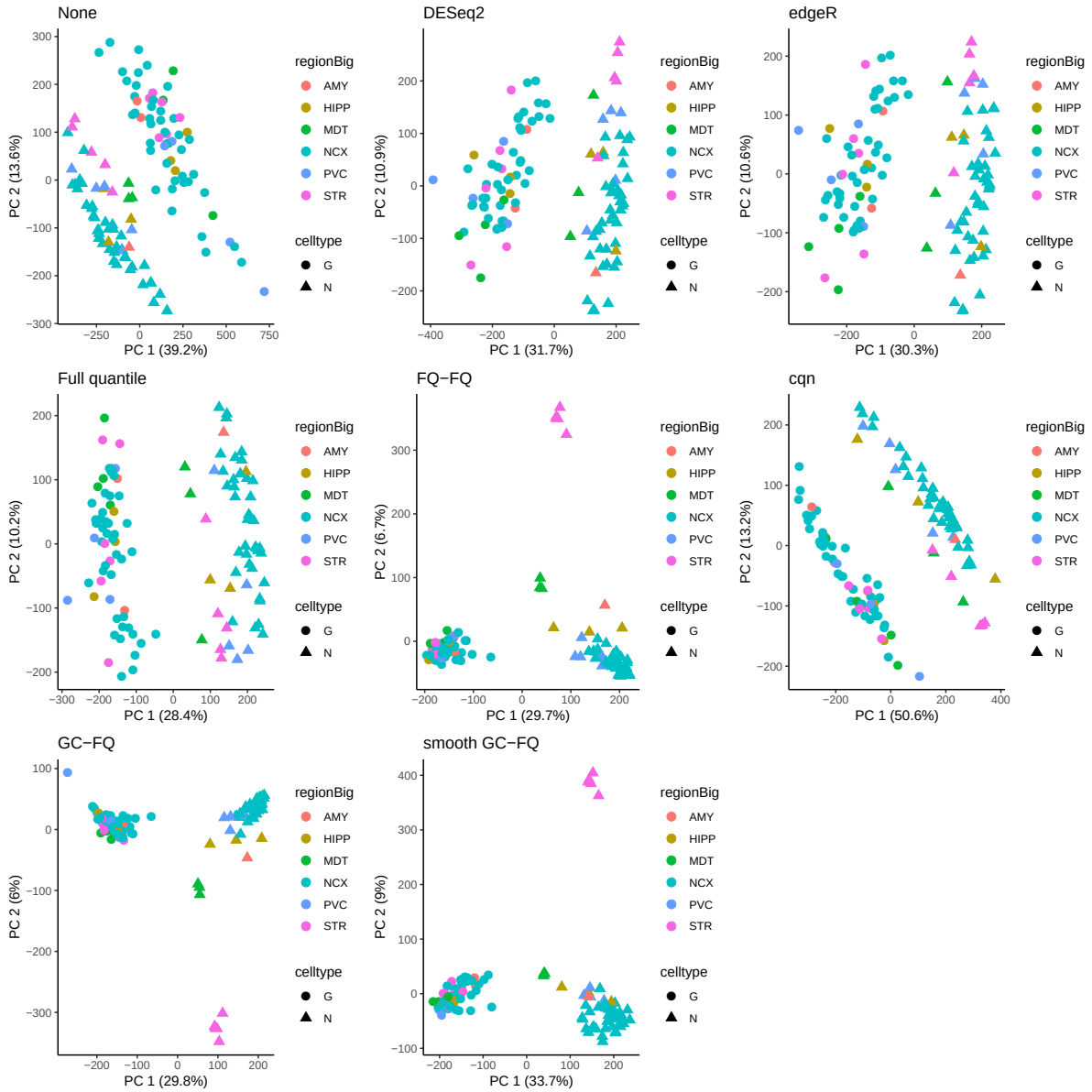
Supplementary Figure 18: *Estimated log-fold-changes for a differential accessibility analysis comparing liver with heart tissues for the dataset from Liu et al.*[30]. Each panel represents a normalization method, which was applied in its native pipeline (see Methods).



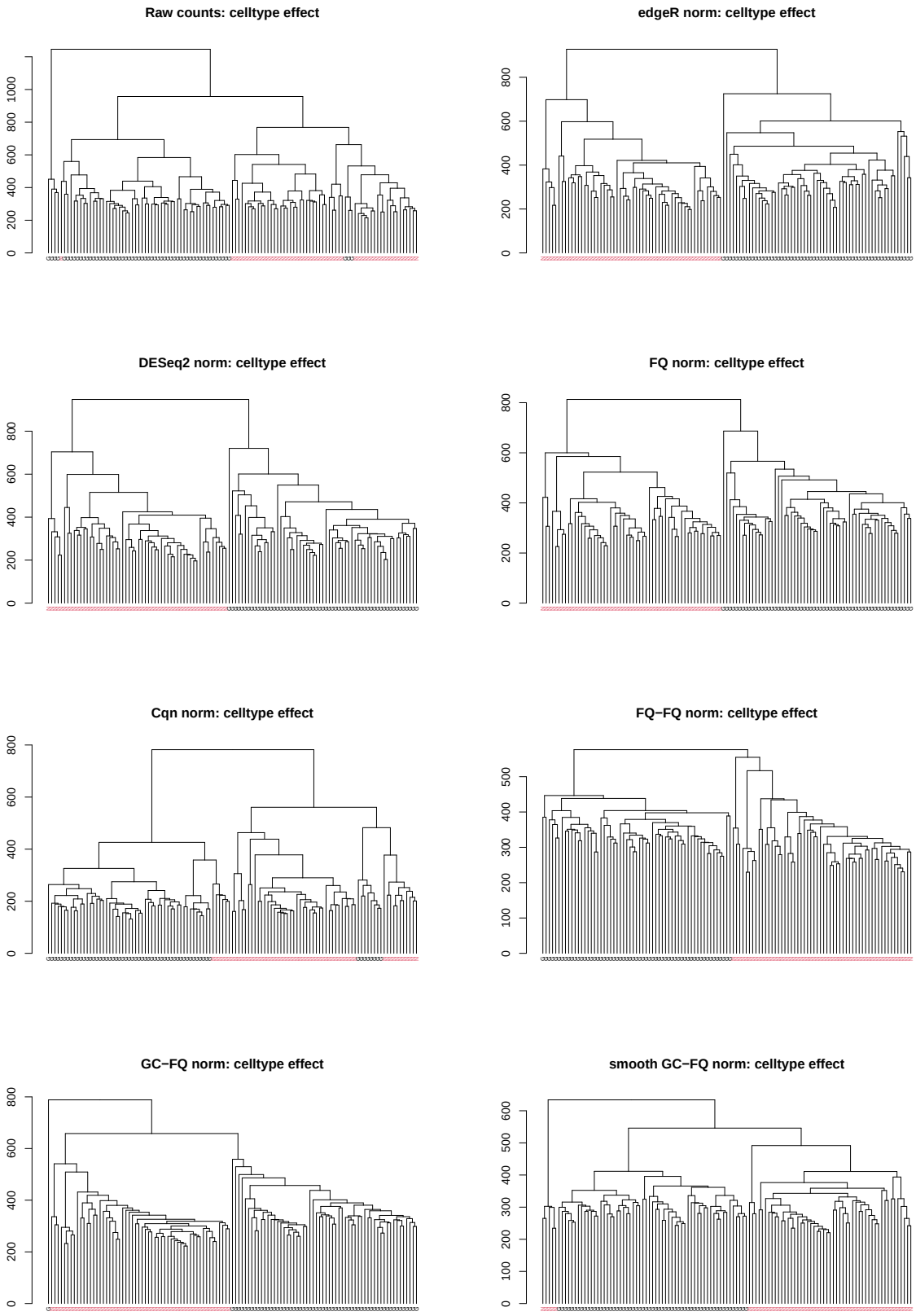
Supplementary Figure 19: *Distribution by GC-content of the top* $5,000$ *peaks for a DA analysis comparing liver with heart tissues for the dataset from Liu et al.*[30]. Each panel represents a normalization method, which was applied in its native pipeline (see Methods). For each normalization method, the barplots display the number of peaks (among the top $5,000$) by GC-content bin. A strong bias is observed for all methods; the most uniform distribution is seen for TMM (edgeR) normalization.



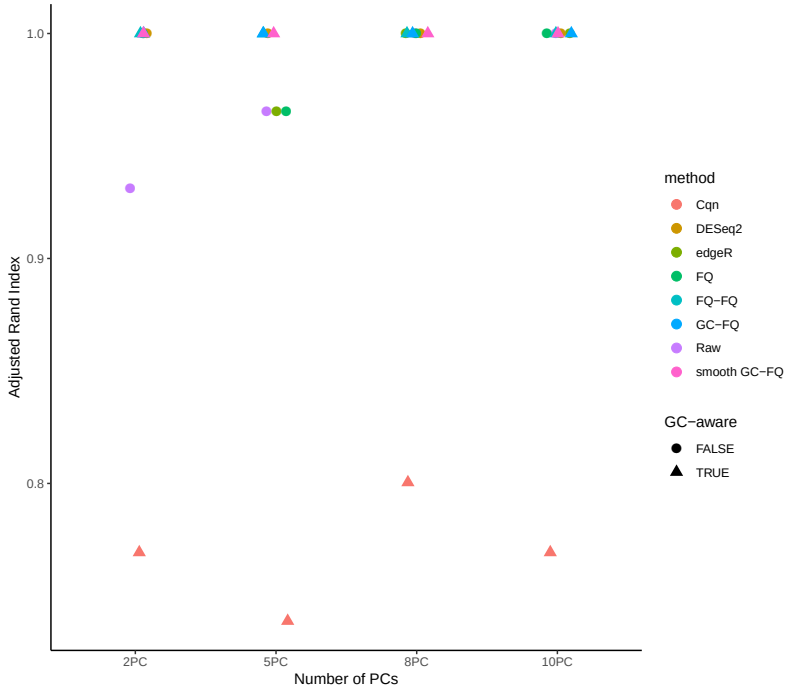
Supplementary Figure 20: *Distribution by GC-content of the significantly DA peaks for a DA analysis comparing liver with heart tissues for the dataset from Liu et al.*[30]. Each panel represents a normalization method, which was applied in its native pipeline (see Methods). For each normalization method, the barplots display the number of DA peaks at nominal FDR level of 5% relative to the corresponding number of DA peaks for the lowest GC-content bin. The total number of DA peaks found at the nominal 5% FDR level is indicated in the title of each panel.



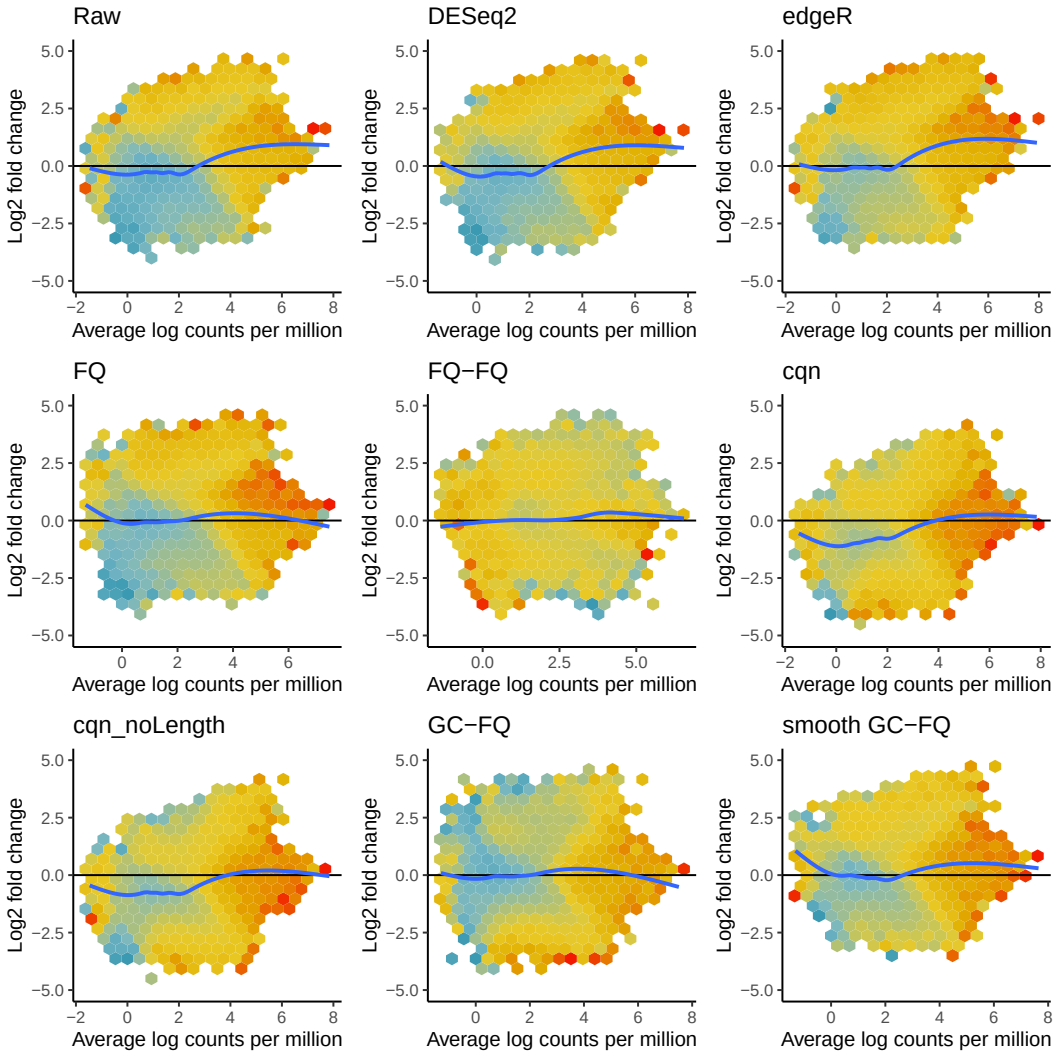
Supplementary Figure 21: *PCA plot for the dataset from Fullard et al.*[29] *after normalization.* The points are colored according to the brain region and the plotting symbol denotes the cell type.



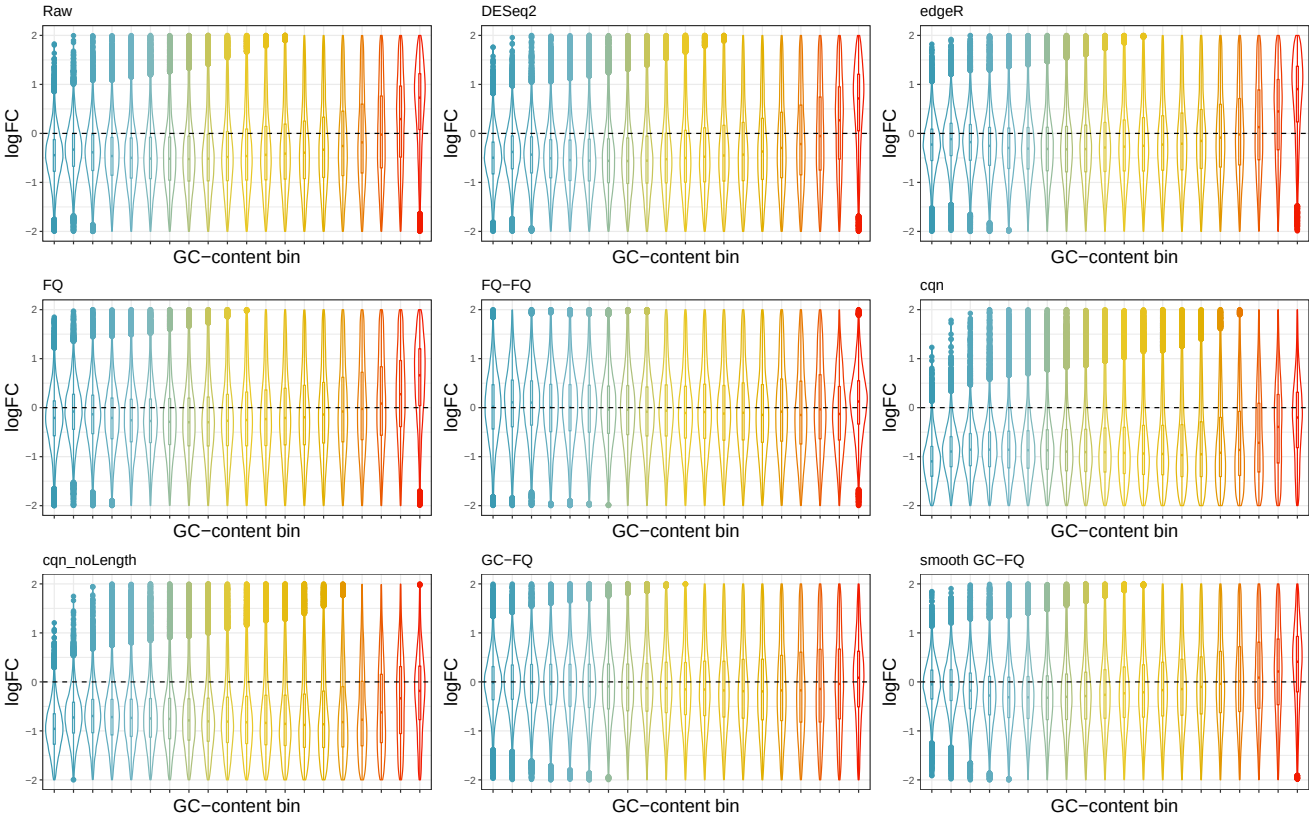
Supplementary Figure 22: *Hierarchical clustering of the dataset from Fullard et al.*[29]. For each method, the normalized counts are log-transformed and pairwise Euclidean distances between all samples are calculated. Agglomerative hierarchical trees are constructed using complete linkage. Samples are labeled and colored according to cell type: neuronal (black) and glial (red) cells. The y-axis represents the Euclidean distance at which two (groups of) samples are agglomerated.



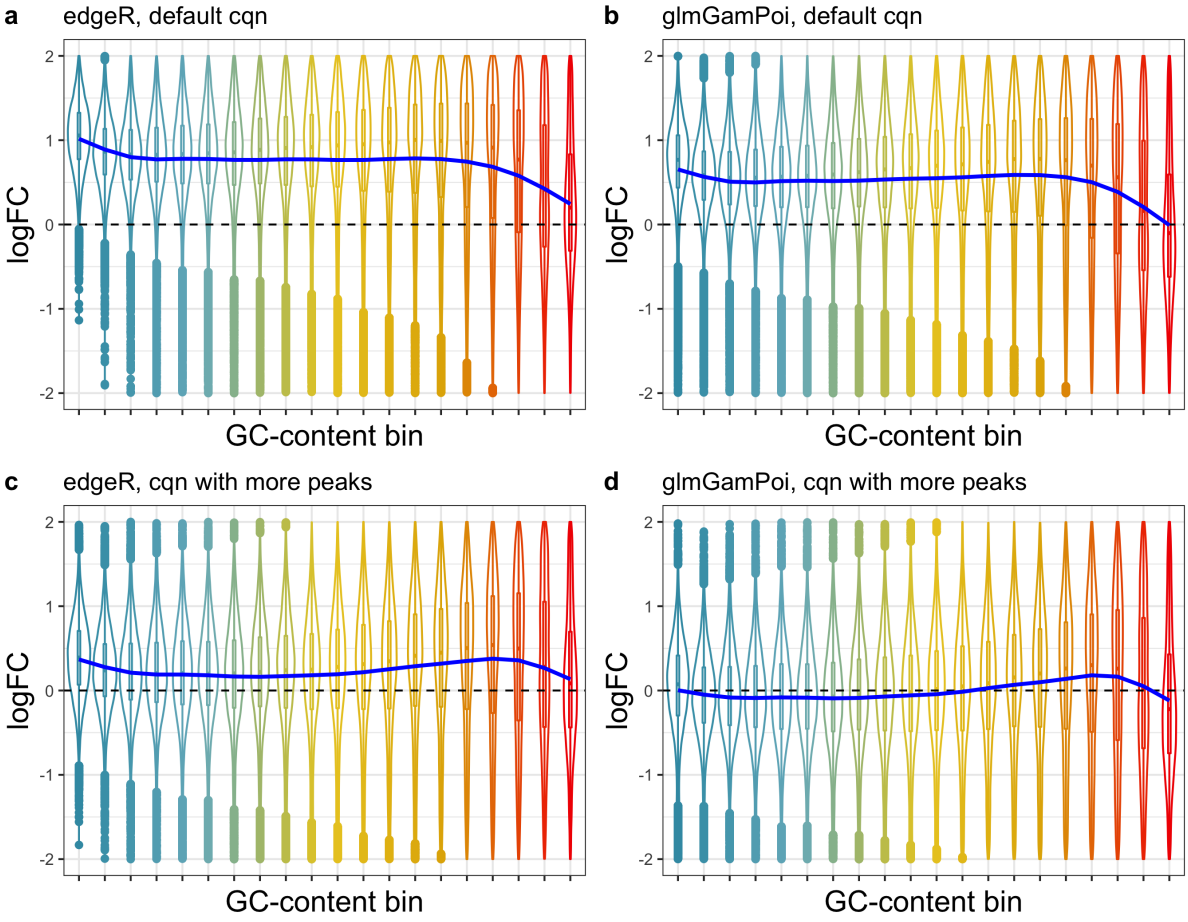
Supplementary Figure 23: *Clustering using the top principal components (PCs) for the dataset from Fullard et al.*[29]. Clustering is performed using partitioning around medoids (PAM) based on different numbers of PCs and the adjusted Rand index (ARI) is calculated with respect to the true cell type labels. All methods except `cqn` correctly recover the two cell types, resulting in an ARI value of 1.



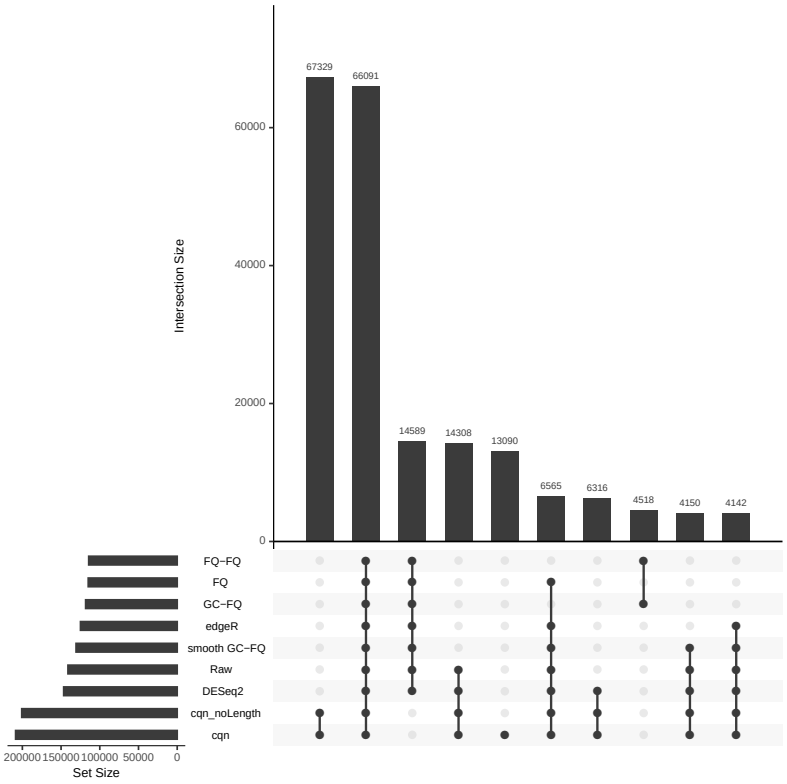
Supplementary Figure 24: *MD-plots with hexagonal binning comparing neuronal to non-neuronal cells for the dataset from Fullard et al.*[29]. The x-axis represents average log-count per million and the y-axis represents log-fold-change as estimated based on a negative binomial model fitted using edgeR. Each panel corresponds to a normalization method. The hexagons are colored according to the average GC-content of the corresponding peaks, where red denotes high GC-content and blue denotes low GC-content.



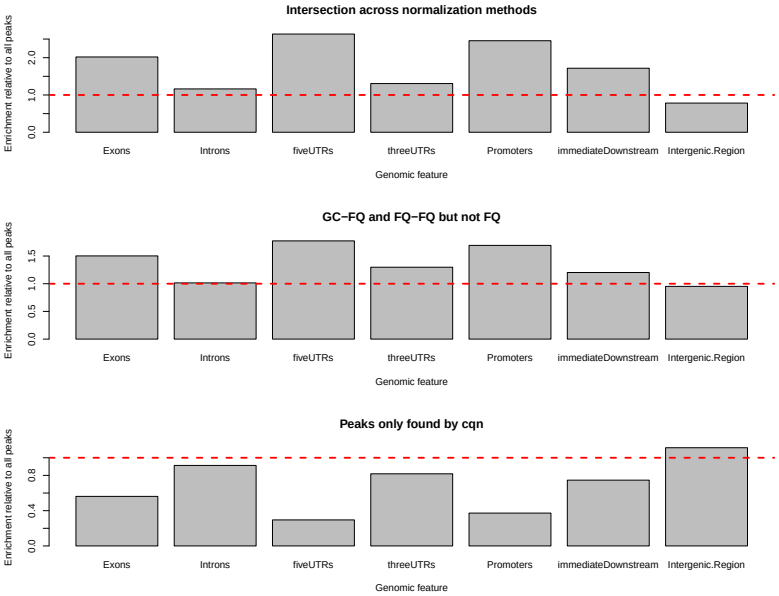
Supplementary Figure 25: *Stratified violin plots of log-fold-changes by GC-content for the dataset from Fullard et al.*[29]. The x-axis represents equally-sized GC-content bins and the y-axis represents log-fold-change as estimated based on a negative binomial model fitted using edgeR. Each panel corresponds to a normalization method.



Supplementary Figure 26: *Log-fold-change bias for different* cqn *settings and DA methods, for the dataset from Fullard et al.*[29]. Log-fold-changes comparing neuronal vs. non-neuronal cells are grouped according to the GC-content of their corresponding peak into 20 equally-sized groups. **(a)-(b)**: Log-fold-changes obtained using a default cqn analysis, using only peaks with mean accessibility larger than 50, in an edgeR (a) and glmGamPoi (b) analysis. Default settings only use $\sim 5\%$ of all peaks in the dataset for normalization. **(c)-(d)**: Log-fold-changes obtained using an adjusted cqn analysis, where we allow allow peaks with mean accessibility larger than 15 to be used for normalization. Log-fold-change bias is less apparent in an edgeR (c) and glmGamPoi (d) analysis. These adapted settings use $\sim 34\%$ of all peaks in the dataset for normalization.



Supplementary Figure 27: *UpSet plot comparing overlap in DA peaks for different normalization procedures for the dataset from Fullard et al.*[29]. Left panel: Barplot of the number of DA peaks for each normalization method (5% nominal FDR level). Top panel: Barplot of the number of DA peaks in common for every combination of methods. Center panel: Combination of methods under consideration for the top panel. For example, the first column shows that $67,329$ DA peaks are only found by sdRPackcqn and `cqn_noLength`.



Supplementary Figure 28: *Enrichment of genomic features relative to background for the dataset from Fullard et al.*[29]. Upper panel: Barplot of the enrichment of each genomic feature relative to all peaks in the dataset, for the intersection of peaks discovered by all normalization methods, except for RUVSeq. An enrichment of relevant genomic features, such as promoters, $5'$ UTRs, and exons is observed. Middle panel: Similar, for peaks discovered by FQ-FQ and GC-FQ but not FQ, where substantial enrichment of promoters, $5'$ UTRs, and exons is observed. Bottom panel: Similar, for peaks only discovered by cqn normalization. No enrichment of relevant features is observed.

Supplementary Tables

Supplementary Table 1: *Gene Ontology analysis of DA peaks.* Top 20 enriched Biological Process (BP) Gene Ontology terms for the DA peaks identified by all normalization procedures, when comparing neuronal vs. non-neuronal cells for the BOCA dataset from Fullard et al.[29]. A total of 26 GO terms were significantly enriched at a 5% nominal FDR level. Abbreviations: oligmeriz. = oligomerization, pwy = pathway.

	GO ID	GO term	BH-adjusted p-value
1	GO:0010243 on	response to organonitrogen compound	0.0365
2	GO:0000122	negative regulation of transcription by RNA polymerase II	0.0365
3	GO:0035601	protein deacylation	0.0365
4	GO:0031329	regulation of cellular catabolic process	0.0365
5	GO:0070431	nucleotide-binding oligomeriz. domain containing 2 signaling pwy	0.0365
6	GO:0098732	macromolecule deacylation	0.0365
7	GO:0050768	negative regulation of neurogenesis	0.0365
8	GO:0051961	negative regulation of nervous system development	0.0365
9	GO:0032990	cell part morphogenesis	0.0365
10	GO:0022008	neurogenesis	0.0398
11	GO:0007399	nervous system development	0.0398
12	GO:0045665	negative regulation of neuron differentiation	0.0398
13	GO:0120039	plasma membrane bounded cell projection morphogenesis	0.0398
14	GO:0006476	protein deacetylation	0.0398
15	GO:0048667	cell morphogenesis involved in neuron differentiation	0.0398
16	GO:0048709	oligodendrocyte differentiation	0.0398
17	GO:0048666	neuron development	0.0398
18	GO:0048858	cell projection morphogenesis	0.0459
19	GO:0007417	central nervous system development	0.0473
20	GO:1901698	response to nitrogen compound	0.0473

Supplementary Table 2: *Gene Ontology analysis of DA peaks.* Top 20 enriched Biological Process Gene Ontology terms for DA peaks discovered by FQ-FQ and GC-FQ but not by FQ normalization, when comparing neuronal vs. non-neuronal cells for the BOCA dataset from Fullard et al. [29]. A total of 28 GO terms were significantly enriched at a 5% nominal FDR level.

	GO ID	GO term	BH-adjusted p-value
1	GO:0006813	potassium ion transport	0.0270
2	GO:0032990	cell part morphogenesis	0.0270
3	GO:0120039	plasma membrane bounded cell projection morphogenesis	0.0270
4	GO:0048858	cell projection morphogenesis	0.0270
5	GO:0048812	neuron projection morphogenesis	0.0270
6	GO:0043266	regulation of potassium ion transport	0.0270
7	GO:0098655	cation transmembrane transport	0.0270
8	GO:0031175	neuron projection development	0.0270
9	GO:0048666	neuron development	0.0270
10	GO:0030182	neuron differentiation	0.0270
11	GO:0120036	plasma membrane bounded cell projection organization	0.0320
12	GO:0050808	synapse organization	0.0320
13	GO:0099536	synaptic signaling	0.0320
14	GO:0099537	trans-synaptic signaling	0.0331
15	GO:0071804	cellular potassium ion transport	0.0352
16	GO:0071805	potassium ion transmembrane transport	0.0352
17	GO:0055085	transmembrane transport	0.0408
18	GO:0050804	modulation of chemical synaptic transmission	0.0467
19	GO:0030030	cell projection organization	0.0467
20	GO:0051641	cellular localization	0.0467