OXFORD

Briefings in Bioinformatics, 2024, **25(2)**, 1–15
https://doi.org/10.1093/bib/bbad538
Problem Solving Protocol

The CUT&RUN greenlist: genomic regions of consistent noise are effective normalizing factors for quantitative epigenome mapping

Fabio N. de Mello, Ana C. Tahira, Maria Gabriela Berzoti-Coelho and Sergio Verjovski-Almeida

Corresponding author. Sergio Verjovski-Almeida, Laboratório de Ciclo Celular, Instituto Butantan, Av. Vital Brasil 1500, 05503-900, São Paulo, SP, Brazil. Tel.: +55-11-2627-3855; E-mail: verjo@iq.usp.br

Abstract

Cleavage Under Targets and Release Using Nuclease (CUT&RUN) is a recent development for epigenome mapping, but its unique methodology can hamper proper quantitative analyses. As traditional normalization approaches have been shown to be inaccurate, we sought to determine endogenous normalization factors based on the human genome regions of constant nonspecific signal. This constancy was determined by applying Shannon's information entropy, and the set of normalizer regions, which we named the 'Greenlist', was extensively validated using publicly available datasets. We demonstrate here that the greenlist normalization outperforms the current top standards, and remains consistent across different experimental setups, cell lines and antibodies; the approach can even be applied to different species or to CUT&Tag. Requiring no additional experimental steps and no added cost, this approach can be universally applied to CUT&RUN experiments to greatly minimize the interference of technical variation over the biological epigenome changes of interest.

Keywords: CUT&RUN; CUT&Tag; normalization; epigenomics

INTRODUCTION

With the advent and popularization of high-throughput sequencing technologies, the field of genome biology experienced drastic changes as several genomic assays were adapted to high-throughput approaches. Chromatin immunoprecipitation and sequencing (ChIP-seq) is perhaps the most popular among them [1–3], allowing for target-specific isolation of DNA–protein complexes, and thus the genome-wide mapping of epigenetic modifications, chromatin-modifying enzymes and transcription factors. Similarly, Cleavage Under Targets and Release Using Nuclease (CUT&RUN) is a recent development by Skene and Henikoff [4] for genome-wide mapping of DNA–protein interactions, an optimization of chromatin immunocleavage [5] for high-throughput sequencing, which seeks to specifically address some of ChIP-seq's main drawbacks. Rather than relying on the random shearing and immunoprecipitation of chromatin, CUT&RUN relies on a micrococcal nuclease protein fusion guided by a protein-A/G-conjugated antibody, directing the cleavage activity to the genomic loci of interest. This approach greatly minimizes the generation of antibody-nonspecific fragments, i.e. noise [4, 6]. In turn, this significantly higher signal-to-noise ratio results in much lower requirements for both starting sample volumes and required read depths [4, 6, 7], greatly reducing the sequencing cost compared to ChIP-seq. In addition, this target-directed enzymatic cleavage enhances the accuracy [4, 7], as it is no longer necessary to rely on random shearing from sonication or undirected enzymatic cleavage.

The advantages of CUT&RUN have led to its increasing adoption [8–13], gaining popularity as a simpler, cheaper alternative to ChIP-seq. A derived technique has been developed, Cleavage Under Targets and Tagmentation (CUT&Tag) [14–17], by integrating Tn5 transposase tagmentation, resulting in an even higher signal-to-noise ratio. In addition, both protocols have been adapted and enhanced by multiple research groups, fine-tuning them for low sample inputs [18–20], minimization of off-target fragments [21, 22] and simplifications of the workflow [23, 24].

Early CUT&RUN mainly relied on the numerous tools and analysis concepts previously developed for ChIP-seq [4]. However, the methodological differences of CUT&RUN have since prompted the development of specialized tools and approaches, such as

Fabio N. de Mello is an MSc student in Bioinformatics at the University of São Paulo. His research interests include molecular biology, cell biology, bioinformatics and systems biology.
Ana C. Tahira is a research specialist at Instituto Butantan. Her main field of activity is in bioinformatics related to gene expression and next-generation sequencing analyses focused on cancer, neurodevelopmental disorder, psychiatry and schistosomiasis.
Maria Gabriela Berzoti-Coelho is a postdoctoral researcher at Instituto Butantan. Her research focus and interests include oncology, molecular biology, cell biology, epigenomics and non-coding RNAs.
Sergio Verjovski-Almeida is a scientific leader at Instituto Butantan and a retired Full Professor and Senior Collaborator at Universidade de São Paulo. His laboratory's research focus includes long non-coding RNAs as modulators of transcriptional programs in cancer cells as well as genome-wide long non-coding RNAs identification and characterization in the parasite *Schistosoma mansoni*.
Received: November 7, 2023. **Revised:** December 14, 2023. **Accepted:** December 25, 2023
© The Author(s) 2024. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/bib/article/25/2/bbad538/7590321 by Universidade de São Paulo user on 22 April 2024

specialized peak calling algorithms [7, 25] and quality control analysis [26, 27].

A recurrent challenge has been proper signal normalization to make samples comparable; for ChIP-seq analyses, straightforward library size-based approaches (such as fragments per kilobase million or fraction reads in peaks) have been extensively found to be inappropriate [28–30]. Normalizing by nonspecific noise can be an option; however, CUT&RUN's low noise generation typically results in an inconsistent background [4, 7], inappropriate for this kind of approach. So far, the accepted golden standard has been spike-in normalization [4, 6, 29, 31]; however, this approach fails to account for the cleavage efficiency and patterns, as the spike material is separately fragmented in advance [30, 32, 33].

In the course of our work with CUT&RUN (unpublished), we began to search for better alternatives, building upon concepts presented in the ENCODE blacklist of problematic regions for ChIP-seq [34]. This work identified that certain regions of the genome present very high background signals, regardless of other experimental factors, being reliably used in ChIP-seq analyses [34–36].

In the present work, we aimed to expand upon this concept by identifying a set of genomic regions that exhibit consistent background signals in CUT&RUN experiments. We have coined the term 'greenlist' to refer to this list of regions, following the naming scheme used for ChIP-seq blacklists [34], greylists [37] and sequencing barcoding whitelists. We demonstrate that these greenlist signals remain consistent across various CUT&RUN experiments involving different antibodies and cell types, thus making them suitable for use as control signals for normalization purposes. We have made available a human CUT&RUN greenlist and blacklist, a mouse CUT&RUN greenlist and blacklist, as well as a human CUT&Tag greenlist and blacklist as Supplementary Table S1. This provides a robust, intrinsic solution for quantitative CUT&RUN analysis, surpassing current methods without any additional cost or experimental steps.

RESULTS

Generating the greenlist

Similar to the ENCODE blacklist [34], we sought to develop a systematic pipeline to identify genomic regions with constant signals across all publicly available negative control CUT&RUN samples. After filtering, we obtained 463 human samples from 102 experiments, derived from 73 established cell lines and 112 patient biopsy samples, and using 30 commercial anti-IgG antibodies (plus 56 samples with no antibody); this variety is essential in ensuring that our results are not biased to specific cell types or experimental setups. As expected, we observed variable correlations between samples (Figure 1A), especially across different experiments. To better understand these differences, we performed a principal component analysis (PCA) (Figure 1B) to test the correlation of the components to known meta-factors, which revealed that this variation was mainly related to differences in protocol/commercial kit used, cell line and sequencing depth (Figure 1C, Supplementary Table S2). We also saw significant variation across experiments not attributed to other tested meta-factors (Figure 1C), indicating that there are still experiment-specific sources of noise of unknown origin. In fact, we expect a fair portion of noise generation to be stochastic in nature and not all variation to be fully explainable.

Of note, the processing of these datasets confirmed the ubiquity of CUT&RUN's varying yields, with several datasets presenting extreme variations in library size (up to 48× difference), even across supposedly identical replicates, as well as samples presenting very low alignment percentages (1.34–99.78%). These differences can have a major impact on quantitative analyses, highlighting the need for better quantitative tools.

To identify how constant the signal of each region is across all the samples, we relied on Shannon entropy [38, 39] (Figure 2A). Briefly, we expect bins with inconsistent high-count outliers to have low information entropy and bins with a more homogenous distribution across the dataset to have higher entropy (further detailed in the Methods). This approach is preferable to simply using the standard deviation or standard error, as those would require the assumption that all bins follow the same general distribution function, which cannot be assumed in advance. As expected, progressively filtering out the lowest entropy bins greatly increased the correlation between samples (Figure 2B), indicative of how consistent the signals from the high-entropy regions were.

We computed the entropy distributions of our test set (equivalent to 10% of bins) after normalization using candidate regions with progressively smaller percentages of highest entropy bins and compared them with a standard DESeq normalization based on library size (Figure 2C). We also compared with the entropy computed with non-normalized counts, which was very low (Figure 2C), as technical variation and different sequencing depths compound with the biological variation and lead to extreme differences between the samples. A nearly ideal threshold was approached by using normalization regions with progressively smaller percentages of highest entropy bins (Figure 2C); a perfect normalization would maximize entropy, nullifying technical variation and leaving only the biological. Based on these results, we selected the top 0.1% most entropic bins as the threshold for generating the greenlist (Figure 2C, highlighted). This threshold maximized entropy and avoided overfitting, whereas the normalization with very small fractions of highly entropic bins (<0.02%) decreased entropy, as more bias was introduced by enhancing the effect of minor stochastic variations, being less representative of the overall data.

We subsequently filtered out from the candidate greenlist any region overlapping or close to any known genes (<5 kb away from known genes), to avoid possible overlapping of true signals, i.e. fragments generated in an antibody-specific manner within gene bodies and neighborhoods. Finally, we extended and merged a selected region if other regions in its vicinity were in the top 1% highest entropy bins (similarly to the original blacklist [34]) in order to avoid short, scattered regions, thus obtaining our CUT&RUN greenlist (available as Supplementary Table S1). The pipeline for greenlist generation is summarized in Figure 2D.

Evaluating this final list by the same previous parameters, we observed results comparable to the initial threshold tests (Figure 3A and B). Some loss of efficiency is to be expected since many high-entropy regions are lost when filtering out gene neighborhoods, but this extra cautious approach did not seem to significantly affect the performance of the final list. To ensure this entropy maximization behavior was not simply due to the size of our greenlist, we performed a Monte Carlo simulation randomly selecting 0.1% of bins (Figure 3C) and saw that our greenlist normalization showed a significantly ($P < 10^{-6}$, 100 000 trials) higher entropy median than a random selection, both before and after filtering out gene neighborhoods. We also confirmed that the signal from these regions did not seem to be significantly biased to experimental differences such as different antibodies or cell types (Figure 3D, Supplementary Table S2), with each principal component only explaining small fractions of the total variance

Downloaded from https://academic.oup.com/bib/article/25/2/bbad538/7590321 by Universidade de São Paulo user on 22 April 2024

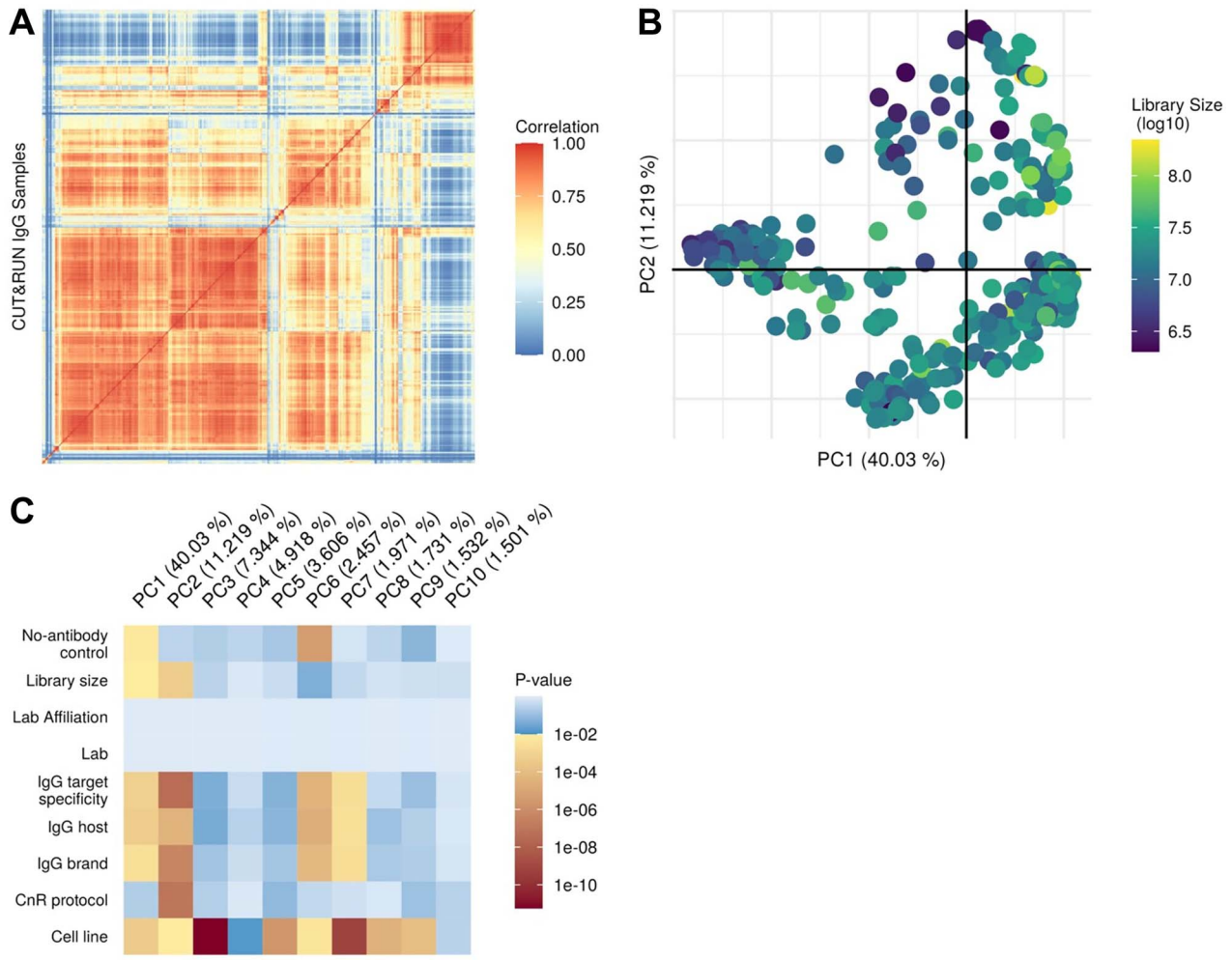

Figure 1. Overview of publicly available negative control CUT&RUN samples. Comparison of the 463 CUT&RUN samples selected for analysis. (**A**) Heatmap of Pearson correlation between the samples, considering all genomic bins (1 kb). (**B**) First two components of the PCA, colored by aligned library size as indicated on the scale at right. (**C**) *P*-value of association between each of the first 10 principal components (PC1–10, columns) to known meta-factors (lines), tested by ANCOVA; *P*-values considered significant ($P \leq 0.01$) are indicated on the scale at right.

(Figure 3E). We did observe a significant association ($P < 10^{-17}$) between library size and the first principal component (even after normalization), but that was expected: higher sequencing depths will highlight subtle variations along the regions, while lower depths will mask them. Regardless, this accounted for only 4% of the total variation of the dataset. Some of the remaining principal components were affected by cell type, suggesting that greenlist regions can still be slightly impacted by differences in chromatin accessibility. Importantly, differences in experimental setups, such as different antibody brands/isotypes, protocol optimizations or test variables of interest, showed limited impact on the overall observed variability (<25% counting the 20 PCs) (Figure 3D). As an additional estimate of the greenlist robustness, we assessed the stability of greenlist generation by random sampling different proportions of the CUT&RUN dataset from 98% down to 2% and found that the similarity between the original greenlist and the greenlists obtained from subsets of the data remains high (correlation >0.8) down to ~110 to 60 datasets (Figure 3F).

To further assess the possible biases in the generation of our greenlist, we next characterized the genomic properties of the high-entropy regions found. We observe no significant bias of greenlist regions toward gene bodies, euchromatic regions or heterochromatic regions when compared to their overall distribution across the human genome (Table S4), keeping in mind that we purposefully discarded candidates overlapping gene bodies. This indicates that the high entropy values observed are not simply a consequence of chromatin accessibility. There is, however, a bias toward centromere regions (Table S4), especially after filtering; this is in line with previous literature findings of centromere regions generating considerable nonspecific noise [34], but we show that our pipeline was able to separate which centromere regions had constant noise and which had not. Following, we observe no distinct pattern of chromosomal distribution of either greenlist regions (Figure S1) or pre-filtering candidate regions (Figure S2). The base composition of greenlist regions (Figure S3) also followed a random distribution, with a similar guanine-cytosine (GC)% as the total human genome. Analyzing the average normalized counts per greenlist region (Figure S4), we observe a few (40 regions) high-count outliers, overall fitting a negative binomial distribution. We observed a trend for greenlist regions to be slightly more repetitive than the overall genomic profile (Figure S5), but we believe this to be a consequence rather than a cause, as the overall repetitiveness of a bin shows no correlation to its calculated entropy (Figure S6), and thus should not be expected to significantly bias our pipeline.

Downloaded from https://academic.oup.com/bib/article/25/2/bbad538/7590321 by Universidade de São Paulo user on 22 April 2024

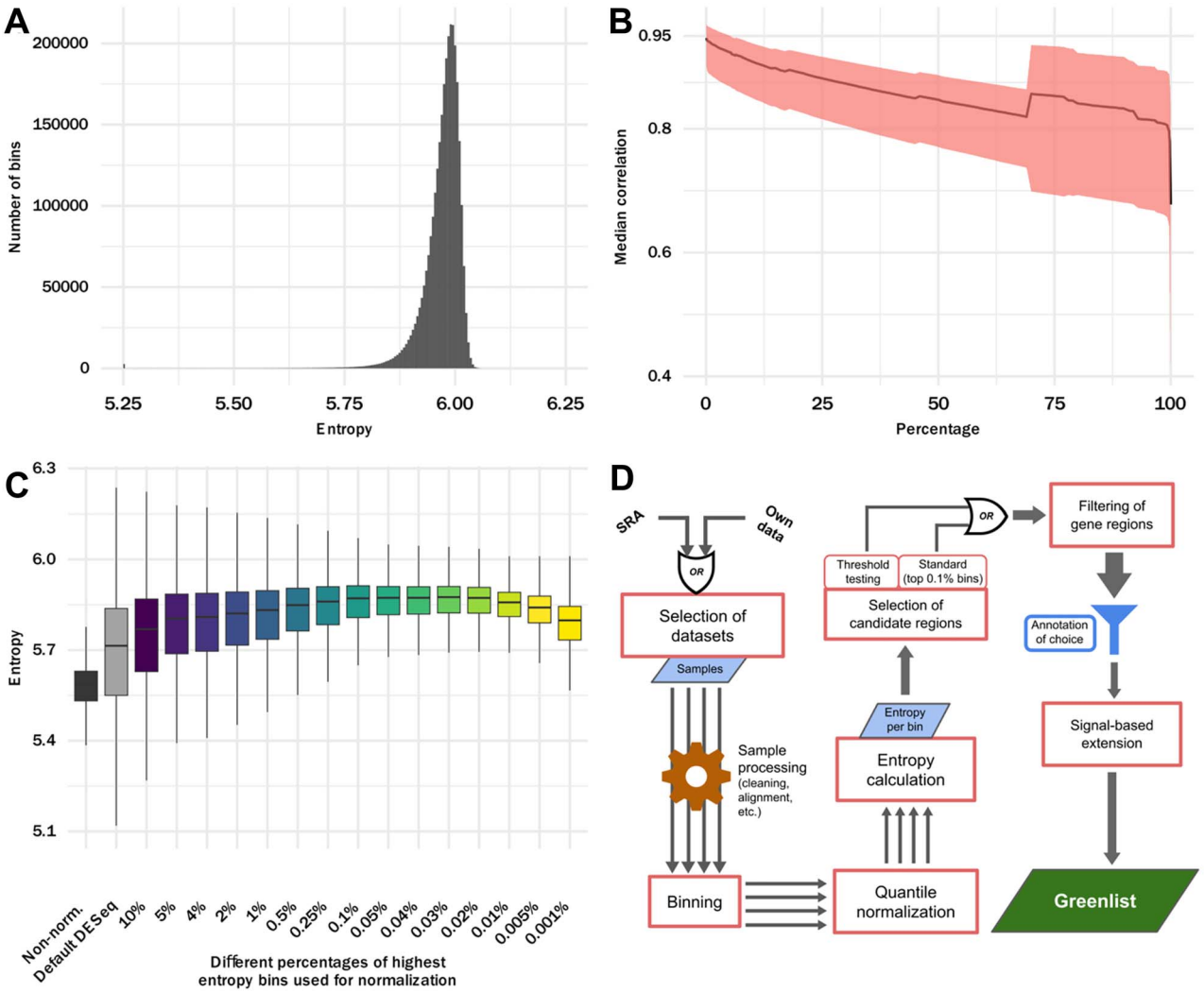

Figure 2. High-entropy regions are effective normalizers. Shannon entropy calculations over all genomic bins (excluding those with no counts on any samples). (**A**) Frequency distribution of entropy for all bins. (**B**) Median Pearson correlation between samples (solid line), considering different percentage fractions of the highest entropy bins, with Q25–Q75 range shown in the shaded area (red). (**C**) Entropy distributions for our test set (the 10% of bins with the lowest entropies) after normalization using different percentages of highest entropy bins, as indicated in the x-axis; in the y-axis, boxplot represents median entropy, first and third quartiles, whiskers extend to 5 and 95% quantiles. (**D**) Summary of the processing pipeline used for greenlist generation, generalized for applicability in new contexts.

A robust approach for different organisms, for CUT&RUN or CUT&Tag

Next, we sought to expand the applicability of our greenlist by defining the greenlist for the mouse genome. Application of the pipeline to 611 mouse negative control samples showed similar quality metrics as seen before (Figure 4A–C), indicative of the consistency of the method. Next, we asked whether a viable greenlist could be generated from a single experiment with a large enough sample size; this should cover cases where a greenlist is needed for an organism with none or few previously available public samples. For this, we used the GSE151326 GEO dataset [40], an antibody characterization study with 50 samples of 43 different antibodies on human cells, as a demonstration. We see that despite how different the samples were initially (Figure 4D), filtering the highest entropy regions yielded very consistent signals (Figure 4E), which did not appear to be biased by library size or antibody target.

Comparing this *de novo* greenlist with our previous hg38 greenlist, we observed a middling correlation of the entropy metrics (Figure 5A) but minimal overlap of the final lists (Figure 5B), as measured by precision-recall F1 scores. This is a direct consequence of the variety of cell types used for the original, which are lacking in a single experiment-based list; nevertheless, the consistency of the found *de novo* greenlist regions shows that this would be a suitable normalization option for the experiment at hand, just not as widely applicable as one built from several experiments.

In addition, we also expanded our efforts toward CUT&Tag. Despite intrinsically producing even less background noise than CUT&RUN, the analysis of 217 human CUT&Tag negative control samples still revealed a greenlist of consistent signal regions (Figure 5C and D), featuring similar entropy maximization results as observed for CUT&RUN (Figure 5E). Comparing the two lists, we again saw little overlap (Figure 5B), suggesting that each list is accurately optimized to their respective technique.

In parallel, we also took the opportunity to create blacklists for human CUT&RUN, mouse CUT&RUN and human CUT&Tag, following the methods of the original ChIP-seq blacklist [34].

Downloaded from https://academic.oup.com/bib/article/25/2/bbad538/7590321 by Universidade de São Paulo user on 22 April 2024

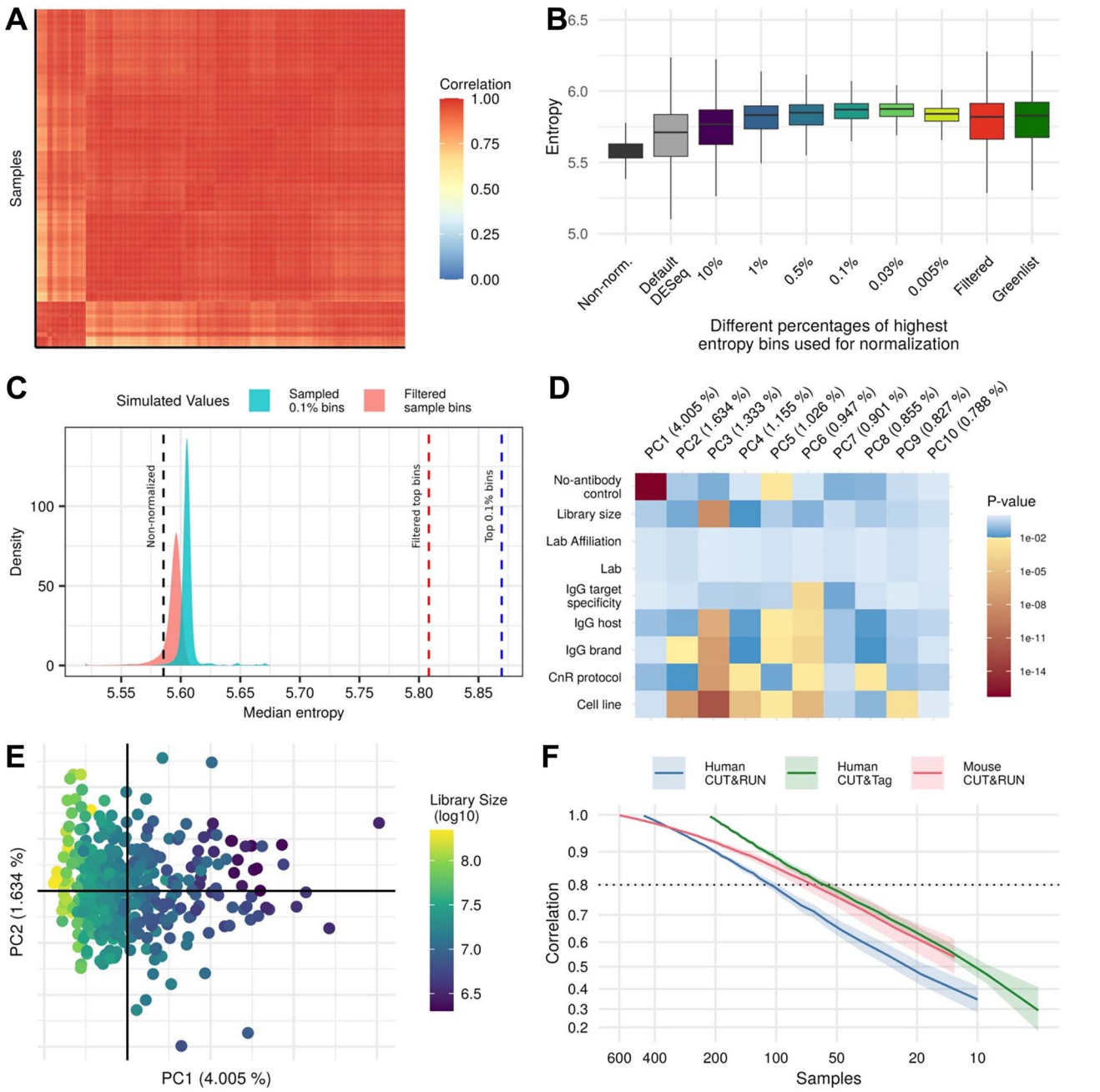

Figure 3. Greenlist regions show consistent signal across variate experimental conditions. Evaluation of the performance of the final greenlist for human CUT&RUN, after filtering and extension. (**A**) Heatmap of Pearson correlation across all samples, considering only greenlist regions signal. (**B**) Performance of final greenlist (last point on the right) and greenlist without extension (labeled as "Filtered", second to last point on the right) as normalizers of our test set (previous results similar to Figure 2C are shown). (**C**) Density distribution of median entropies of Monte Carlo simulations of normalizations with randomly selected regions, before (blue, rightmost curve) and after (red, leftmost curve) filtering (medians obtained in Figure 3B shown as dashed lines). (**D**) *P*-value of association between each of the first 10 principal components (PC1–10, columns) to known meta-factors (lines), tested by ANCOVA (*P*-values considered significant ($P \leq 0.01$) are indicated on the scale at right). (**E**) First two components of PCA of greenlist regions across all samples, color coded according to library size as indicated on the scale at right. Library size was the only significant ($P \leq 0.01$) cofactor found. (**F**) Spearman correlation of entropy between the original greenlists and greenlists generated after random subsamplings of our datasets. Between 4 and 196 trials per interval (see Methods), median correlation shown as a solid line, with standard deviation shown as shaded areas. Acceptable cutoff considered (0.8) shown as a horizontal black dotted line.

Despite having since been validated for other techniques, the ENCODE ChIP-seq blacklist has not been validated for CUT&RUN or CUT&Tag, and the vastly different methodological steps (i.e. the lack of random fragmentation and immunoprecipitation) may affect the generation of high-signal background regions. We observed a fair overlap between the CUT&RUN and CUT&Tag blacklists (Figure 5B), indicative of the similarity of these techniques, but as expected both feature little overlap with the ChIP-seq blacklist, confirming their drastically different noise profile. Previous efforts have been made to create a CUT&RUN-specific blacklist [41] but with limited samples ($N = 20$), which would present limited accuracy.

Importantly, we observe minimal overlap between the greenlists and blacklists specific to each method (0.010 for CUT&RUN and ~0.001 for CUT&Tag, Figure 5B), despite moderate correlations between the two parameters (0.472 and 0.730, respectively;

Downloaded from https://academic.oup.com/bib/article/25/2/bbad538/7590321 by Universidade de São Paulo user on 22 April 2024

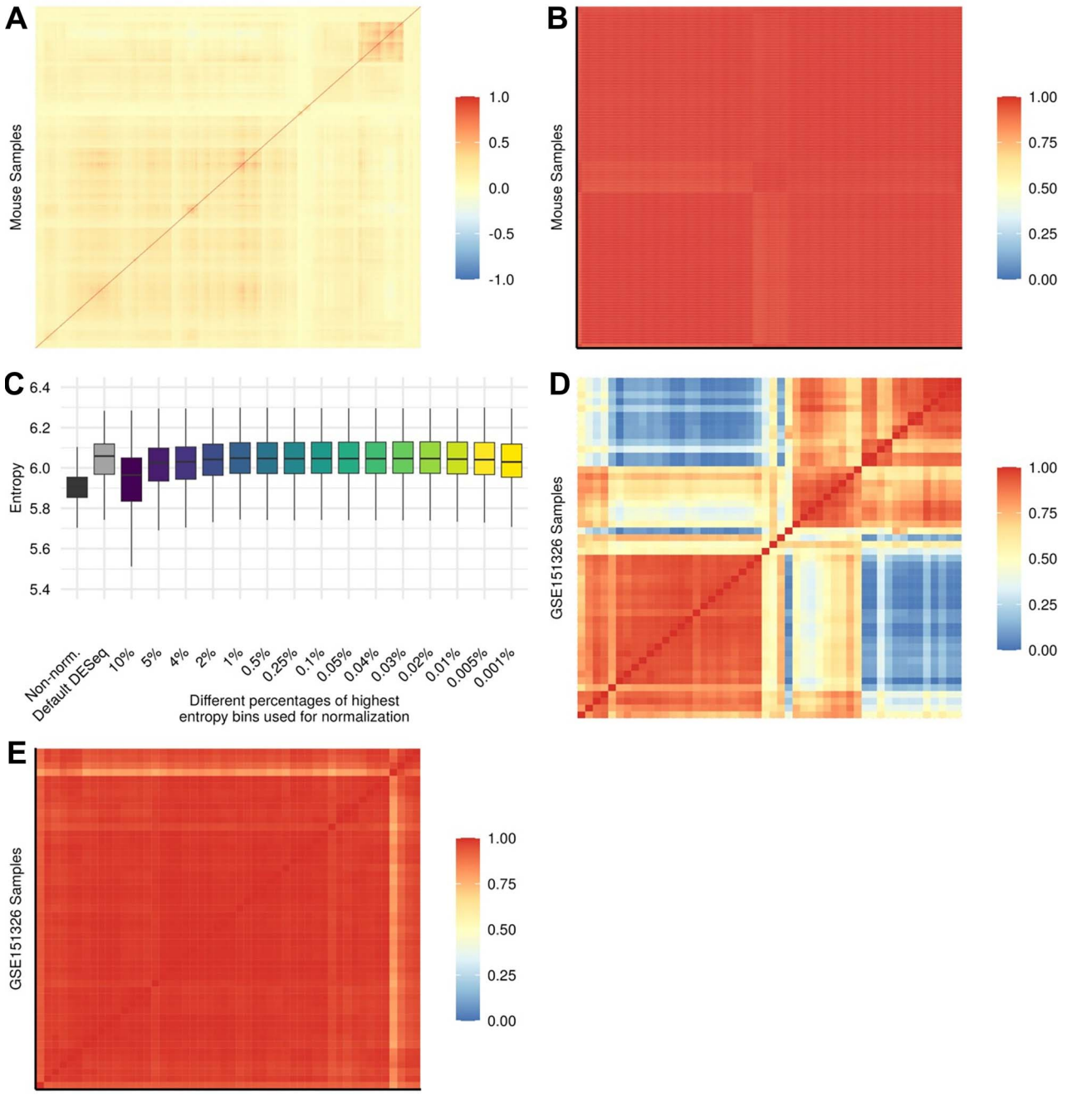

Figure 4. Application of our pipeline to mouse datasets or to a dataset from a single experiment show similar performances. (**A**) Heatmap of Pearson correlation of 611 mouse CUT&RUN samples, considering signal from all 1-kb genomic bins. (**B**) Heatmap of Pearson correlation of 611 mouse CUT&RUN samples, considering only the signal from greenlist regions. (**C**) Entropy distributions for our mouse CUT&RUN test set after normalization using different percentages of highest entropy bins, as indicated in the *x*-axis; in the *y*-axis, boxplot represents median entropy, first and third quartiles, whiskers extend to 5% and 95% quantiles. (**D**) Heatmap of Pearson correlation of the 50 samples from dataset GSE151326, considering signal from all 1-kb genomic bins. (**E**) Heatmap of Pearson correlation of the 50 samples from dataset GSE151326, considering only signal from the *de novo* greenlist constructed off this dataset alone.

Figure 5A). This shows that, although there is a tendency for high-signal regions to also display high entropy, i.e. blacklist regions tend to be reasonably consistent across experiments, the entropy values of the vast majority of them were not high enough to be selected in our pipeline. The exact reason for these observed moderate correlations remains elusive—noise generation is a heavily stochastic process, and the dynamics of how each technique generates specific patterns of noise remains to be characterized in the literature. Nevertheless, the low overlap scores observed here between greenlists and blacklists (Figure 5B) show that our pipeline was able to distinguish between the two and to identify that blacklist regions are not consistent enough to be effectively used as basis for normalization.

Greenlist normalization outperforms current standards

We have observed that the greenlist regions act as suitable normalizing factors in our tests, maximizing the entropy of our negative control datasets, so we next sought to compare them to consolidated normalization approaches. We reanalyzed dataset

Downloaded from https://academic.oup.com/bib/article/25/2/bbad538/7590321 by Universidade de São Paulo user on 22 April 2024

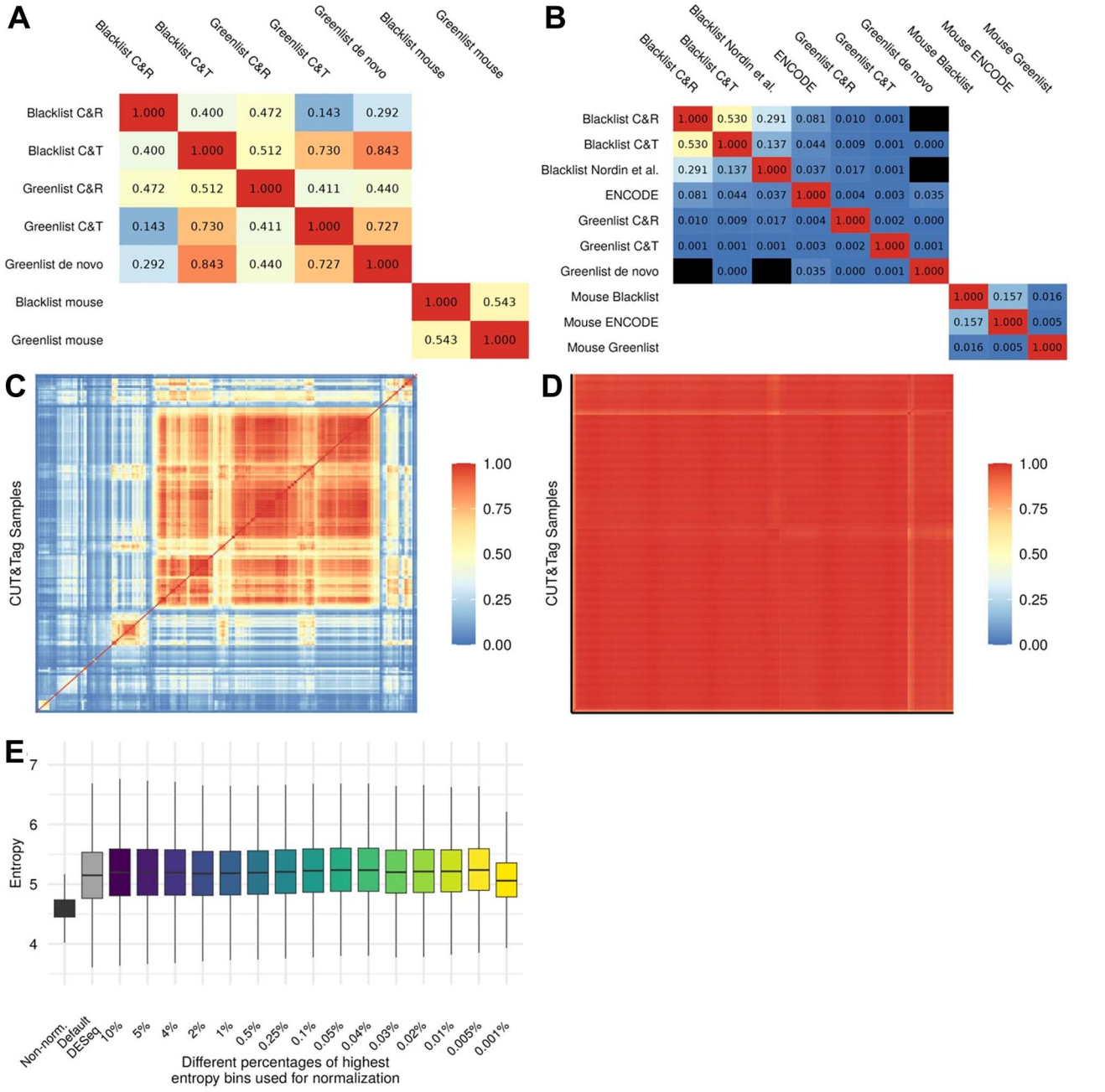

Figure 5. Greenlists are methodology specific and different from blacklists. (**A**) Spearman correlation between the quantifications used for creating each list, either entropy (for greenlists) or median signal (for blacklists), considering all genomic 1-kb bins. (**B**) Overlap between the final lists created and some from the literature, considering F1 scores of precision-recall. (**C**) Heatmap of Pearson correlation of 217 human CUT&Tag samples, considering signal from all 1-kb genomic bins. (**D**) Heatmap of Pearson correlation of 217 human CUT&Tag samples, considering only signal from the CUT&Tag greenlist regions. (**E**) Entropy distributions for our human CUT&Tag test set after normalization using different percentages of highest entropy bins, as indicated in the x-axis; in the y-axis, boxplot represents median entropy, first and third quartiles, whiskers extend to 5 and 95% quantiles.

GSE104550 [42], generated by the Henikoff laboratory and used by Meers *et al.* [6] to establish the comparability of normalizations by *Escherichia coli* carryover spike-in and external spike-in (such as added *Saccharomyces cerevisiae* DNA), the two normalization methods considered as the golden standards for CUT&RUN. The design of this experiment [42] proved very advantageous for testing normalization options, as the samples are expected to remain biologically consistent within each of the two antibodies tested, and the main sources of variation should be the starting number of cells (tested here at several values between 100 and 1 000 000 cells) and the technical variations expected from DNA extraction, library preparation and sequencing. Thus, we can consider the ratio of library size to starting cell count as the *de facto* technical variation, as we expect around the same number of fragments generated per cell in each sample, and we can evaluate each normalization method as to how well it can account for this variation.

Here, we again observe that the CUT&RUN signal across the entire genome shows drastic differences between the antibodies used (Figure 6A); nevertheless, greenlist signal remained constant and was adequate to be used as normalizer among these samples (Figure 6B). We saw that spike-in normalization (either by *E. coli* carryover or added *S. cerevisiae* material) exhibited low correlations to our known *de facto* technical variations ($R^2 = 0.513$)

Downloaded from https://academic.oup.com/bib/article/25/2/bbad538/7590321 by Universidade de São Paulo user on 22 April 2024

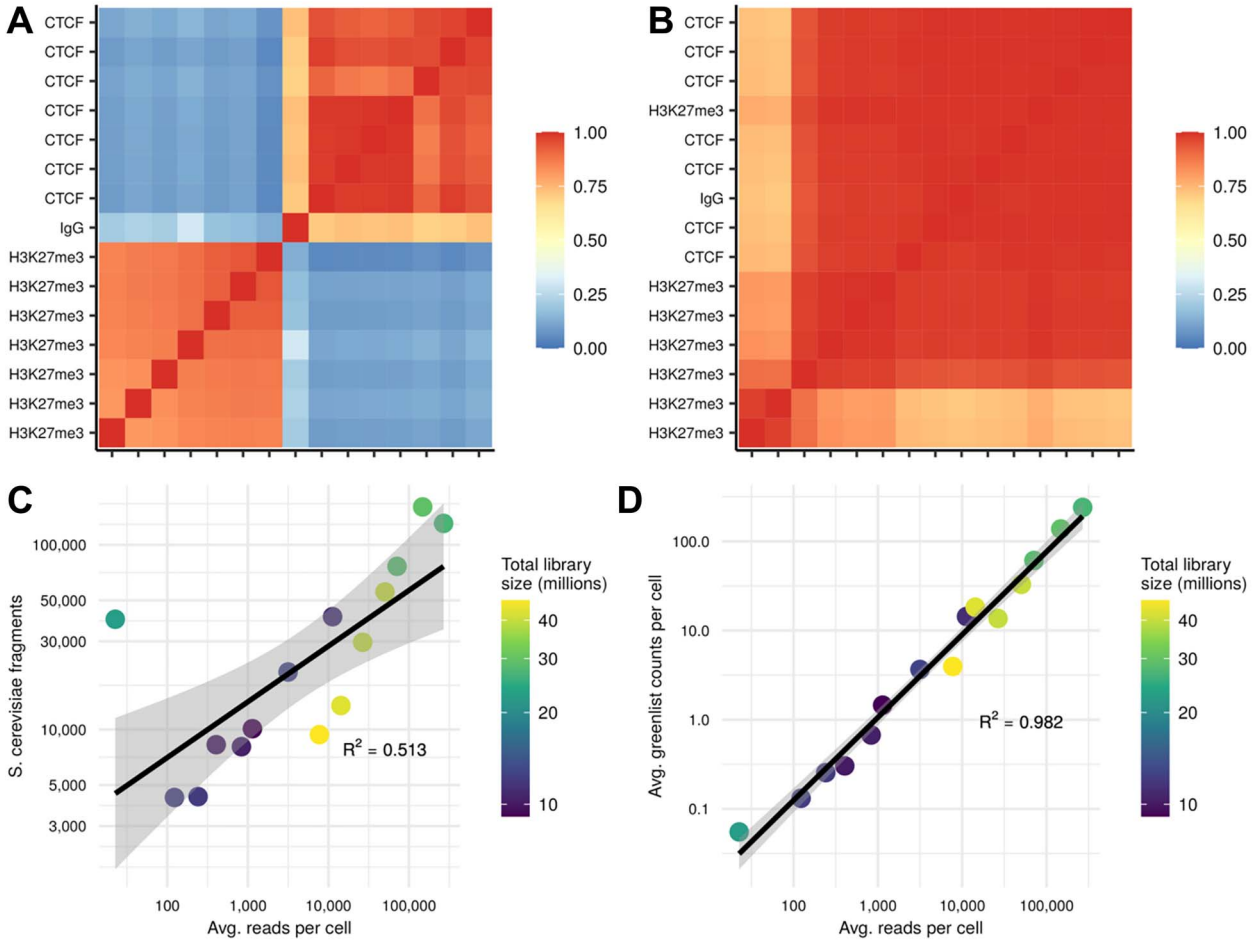

Figure 6. Greenlist normalization outperforms spike-in on dataset GSE104550 from the Henikoff laboratory. (**A**) Heatmap of Pearson correlation between all samples, considering all 1-kb genomic bins. Samples labeled by antibody used. (**B**) Heatmap of Pearson correlation between all samples, considering only greenlist regions. Samples labeled by antibody used. (**C**) Correlation between spike amplification (inferred as *Saccharomyces cerevisiae* fragments reads divided by original fragments, where original spike-in fragments remain constant for all samples and was thus omitted) and *de facto* library amplification (inferred as total library size reads divided by starting number of cells). (**D**) Correlation between greenlist amplification (inferred as total greenlist counts divided by starting number of cells) and *de facto* library amplification (as in **C** above). Linear regression and 95% confidence interval shown for graphs (**C**) and (**D**), R^2 calculated as fit to the linear regression.

(Figure 6C), failing to arrive at consistently normalized libraries. On the other hand, greenlist counts appropriately encapsulate this variation ($R^2 = 0.982$) (Figure 6D).

Importantly, while the exact cell count is known in this experiment [42], as the variation was an intended design, it does not need to be known in advance for other experimental applications; thus, greenlist normalization can more accurately account for unknown variations in cell counts that may arise from experimental conditions.

Overall, these results demonstrate an intrinsic advantage of endogenous normalization factors such as our greenlist. As they are generated concurrently with the fragments of interest, they are directly affected by the same technical and experimental variations, such as different starting cell counts or variations in yield as discussed earlier.

A detailed documentation on applying greenlist normalization to a typical CUT&RUN workflow is provided via a GitHub Jupyter notebook (https://github.com/fndemello/CUT-RUN_greenlist/blob/main/CUTandRUNAnalysis.ipynb). We chose dataset GSE157095 by Singh *et al.* [43] as an example and provided documentation for a greenlist normalization approach to call PAX3-FOXO1-binding peaks over normalized inputs, as well as a differential binding analysis using DESeq2 [44].

Greenlist normalization uncovers relevant variation from biased datasets

Lastly, we aimed to validate our greenlist's applicability as a normalization factor in real experimental scenarios. For this, we specifically looked for CUT&RUN studies that featured extreme differences in library sizes, as they stand to benefit the most from a novel, robust normalization approach. We first selected study GSE221701 [45], which featured 48 samples and used 11 antibodies, with sample library sizes ranging from 11 to 134 million read pairs. Secondly, study GSE194217 [46], with 26 samples and 6 antibodies, with sample library sizes ranging from 3 to 64 million reads. Notably, both works evaluated the CUT&RUN results in only a qualitative manner, instead seeking quantitative validations from other techniques such as RNA-seq, ATAC-seq and ChIP-seq.

To evaluate the impact of technical variation in the above datasets and evaluate the potential of proper normalization to mitigate it, we performed a PCA analysis of the genomic data under different normalization scalings. Should the samples be improperly normalized, we expected the variations of large library size samples to contribute to the composition of PCs more strongly, thus biasing the graph and eclipsing variation patterns of interest (such as those related to biological sources). Applying

Downloaded from https://academic.oup.com/bib/article/25/2/bbad538/7590321 by Universidade de São Paulo user on 22 April 2024

standard normalization, based only on total library size, we saw that the PCAs of each dataset seemed to form clusters divided by antibodies (Figure 7A and B), but there was a strong correlation between PC1 and library size (Figure 7C and D). Thus, the variation observed was strongly impacted by technical variation, even after normalization, and attempting to interpret the data under these conditions could easily lead to inaccurate biological conclusions. Zou *et al.* [46], authors of dataset GSE194217, did perform spike-in normalization with added *S. cerevisiae* DNA; performing the PCA analysis with this normalized data, we saw that it did remove the correlation with library size (Figure 7E)—but only to replace it with a correlation to raw spike-in counts (Figure 7F), thus still arriving at biased results.

On the other hand, performing the analysis on greenlist-normalized counts greatly minimized the bias of library size (Figure 8A and B, without creating a bias to the greenlist itself (Figure 8C and D). As such, the analysis can now be appropriately interpreted in a biological context as a typical PCA would be (Figure 8E and F). For the work of Zou *et al.* in particular, this new analysis now highlights two histone marks, H3K27ac and H3K27me3 (Figure 8F), which is in line with some of the major findings of their work [46].

This validation approach was repeated with three additional datasets for confirmation. Dataset GSE223997 by Xu *et al.* [47] featured 50 CUT&Tag samples, with total library sizes ranging from ∼180 000 to 59 million reads. As before, library size normalization (Figure S7a) and spike-in normalization (Figure S7b) still lead to biased PCA results, with technical variation masking biological findings; however, normalization with the CUT&Tag-specific greenlist greatly minimizes this bias (Figure S7c and d). Following, dataset GSE171327 by Weigel et al. [48] featured 24 mouse CUT&RUN samples, varying from 15 to 77 million reads. Although the technical bias after spike-in normalization was milder than other examples (Figure S8a and b), greenlist normalization still showed an improvement (Figure S8c and d), with a lesser contribution of technical variation to the PCA. Finally, dataset GSE166221 by Vinjamur *et al.* [49] featured both human and mouse CUT&RUN samples, with the mouse samples displaying a considerably smaller library size range than the human samples (6–9 million and 4–33 million reads, respectively). Consequently, PCA of the library size-normalized human samples was more influenced by technical variation (Figure S9a and b), which is again minimized by greenlist normalization (Figure S9c and d). Importantly, these additional validations not only corroborate our previous demonstrations (Figures 7 and 8), but also confirm that the mouse CUT&RUN and human CUT&Tag greenlists prove just as effective.

To further evaluate the performance of our greenlist as normalizing factors, we next considered its effects over the experimental replicates in the previously selected datasets. Regardless of each study's topic, experimental replicates should have minimal biological variation, and should thus arrive at similar results if properly normalized. We quantified this similarity with F1 scores and observed that greenlist normalization consistently outperforms spike-in approaches for human CUT&RUN (Figure 9A), CUT&Tag (Figure 9B) and mouse CUT&RUN (Figure 9C), with replicate pairs presenting on average greater similarity when peaks are called over greenlist-normalized inputs. As a contrast to the datasets chosen so far, dataset GSE157095 by Singh *et al.* [43] features very little library size variation (7.3 to 9 million reads), which results in both normalization approaches showing comparable results (Figure 9D). This highlights that while spike-in normalization can be an accurate approach in best-case experimental scenarios, its effectiveness falls behind for more challenging datasets; meanwhile, greenlist normalization matches spike-in's effectiveness at its best, and consistently surpasses it in more dire situations. In this sense, the GSE104550 dataset by Meers *et al.* [6] (analyzed in Figure 6) once again serves as a great test for normalization approaches, as the samples feature minimal experimental variation outside of the different number of cells used; as expected, greenlist normalization can better deal with the simulated technical variation introduced (Figure 9E), achieving more similar peaksets than those of spike-in normalization. And in contexts where spike-in was originally not performed, such as in the work of Vinjamur *et al.* [49], the option of greenlist normalization remains accessible, a clear improvement over the default library size normalization (Figure 9F).

DISCUSSION

High-throughput genomic techniques such as CUT&RUN can be especially sensitive to poor normalization, as the vastly different signal profiles of different targets and experimental designs can interfere with typical statistical assumptions. Consequently, this poor normalization risks jeopardizing the technique's quantitative power. And as new experimental protocols are developed, so too do specialized *in silico* tools become necessary for their proper analysis. In that sense, we here introduce a new normalization technique, highly specialized for either CUT&RUN or CUT&Tag, through a novel application of information theory approaches to the underlying concept of the ENCODE ChIP-seq blacklist [34]. Our approach leverages the extensive collection of datasets published thus far, and through rigorous testing and validation we observed our greenlist's applicability across a variety of experimental scenarios.

Naturally, this reliance on previously published datasets imposes limitations to future broadening of this method beyond the characterization work performed here. We have rigorously documented the capabilities and sensitivities of our pipeline in order to make it applicable to different contexts (such as different organisms and variant methodologies), including considerations to mitigate small sample sizes, but these future applications will inevitably depend on a robust availability of samples. This issue compounds with the inherent randomness of noise generation, as we are unable to deterministically define the exact causes of the observed constancy of greenlist regions. Although this uncertainty should always be kept in mind, we have sought to minimize sample bias as much as possible throughout our pipeline, and thus believe that the empirical observations at the core of greenlist normalization are reliably robust. Based on our results, we observe that roughly 60–120 samples are needed to generate a robust, widely applicable greenlist (Figure 3F). For organisms with fewer samples than that, the greenlist generated can still be a reliable normalizer (Figure 4D and E, 50 samples), but will likely be overfit to the datasets used, and thus may not be applicable for future studies with the same organism.

Furthermore, we showed that greenlist normalization outperforms the current standards. This offers a robust analysis option, which should apply to any CUT&RUN or CUT&Tag experiment, with no added experimental step, and incurring no additional costs. Even for organisms featuring few published datasets, we showed that this entropy-based greenlist construction should remain consistent and still offer a robust normalization. We expect that this methodology will enhance the quantitative potential of CUT&RUN and CUT&Tag in the scientific literature, fostering more comprehensive and reliable analyses.

Downloaded from https://academic.oup.com/bib/article/25/2/bbad538/7590321 by Universidade de São Paulo user on 22 April 2024

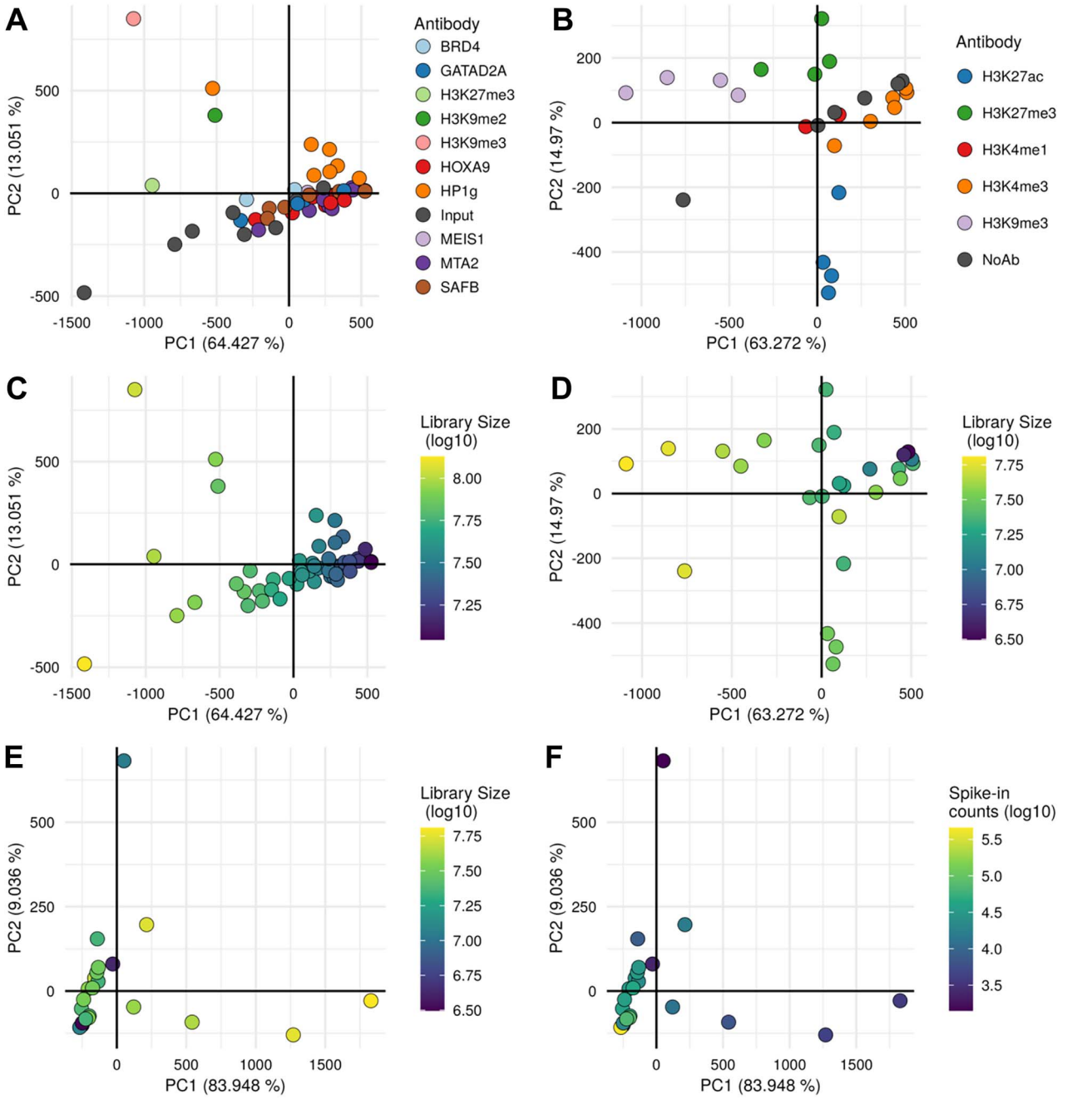

Figure 7. Established normalization approaches are unable to remove technical bias in extreme cases. Validation of greenlist normalization performance for two public datasets, GSE221701 (graphs **A**, **C**) and GSE194217 (graphs **B**, **D**, **E**, **F**). (**A**, **B**) First two components for PCA of each dataset, using default library size-based normalization; samples colored by antibodies used, as indicated on the legends at right. (**C**–**D**) First two components for PCA of each dataset, using default library size-based normalization, with samples colored by library sizes, as indicated on the scales at right. (**E**–**F**) First two components for PCA of dataset GSE194217, using spike-in normalization; samples colored by library sizes (**E**) or spike-in counts (**F**), as indicated on the scales at right.

MATERIALS AND METHODS

Sample selection and processing

Samples were selected from publicly available studies submitted to the Sequence Read Archive repository (SRA, https://www.ncbi.nlm.nih.gov/sra), and manually curated to ensure the consistency and accuracy of the metadata. Experimental meta-factors considered included cell line, cell type/characteristics, antibody brand and product, antibody target (e.g. anti-mouse IgG versus anti-human IgG), antibody host, CUT&RUN protocol/kit used, authors and authors' primary affiliations. Only negative control samples from studies with ≥6 samples total were selected. Samples with too few aligned reads were discarded (≤1.5 M for human CUT&RUN, ≤1 M for mouse CUT&RUN, ≤500 k for human CUT&Tag), based on recommendations of established protocols [6, 50–52]. In total, 463 samples were used for human CUT&RUN, 611 for mouse CUT&RUN and 217 for human CUT&Tag; a full list of datasets used is available in Supplementary Table S3.

Downloaded from https://academic.oup.com/bib/article/25/2/bbad538/7590321 by Universidade de São Paulo user on 22 April 2024

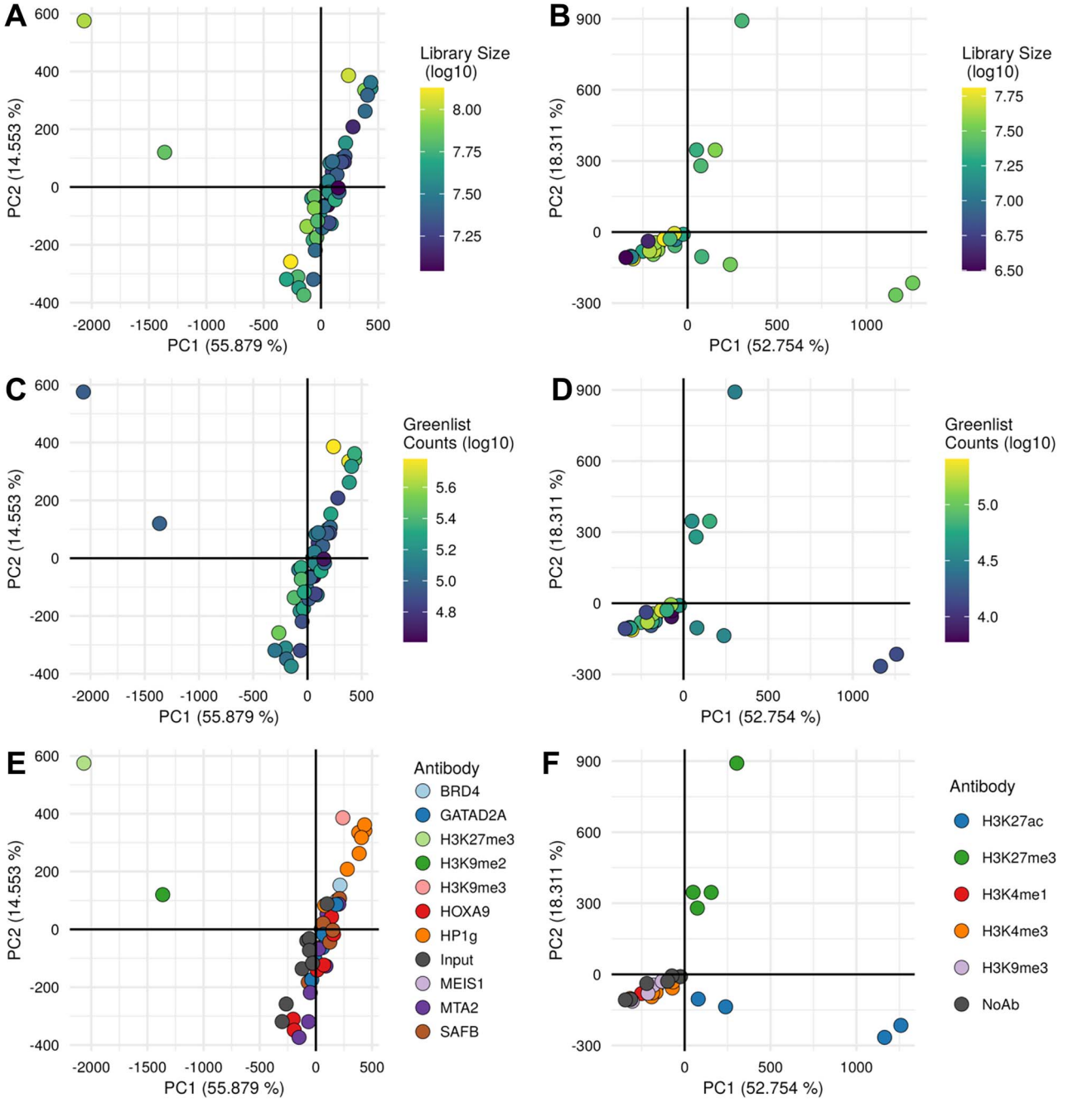

Figure 8. Greenlist normalization greatly minimizes technical variation, uncovering relevant biological data. Validation of greenlist normalization performance for two public datasets, GSE221701 (graphs **A**, **C**, **E**) and GSE194217 (graphs **B**, **D**, **F**). (**A**–**B**) First two components for PCA of each dataset, calculated after performing greenlist normalization; samples colored by library sizes, as indicated on the scales at right. (**C**–**D**) First two components for PCA of each dataset, calculated after performing greenlist normalization, with samples colored by total counts in greenlist regions, as indicated on the scales at right. (**E**–**F**) First two components for PCA of each dataset, calculated after performing greenlist normalization; samples colored by antibodies used, as indicated on the legends at right.

Samples were downloaded with SRA-toolkit (SRA Toolkit Development Team, v3.0.2 https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software), cleaned with fastp (v0.20.0) [53] with parameters '–detect_adapter_for_pe -W 1 -3 -5', and aligned with bowtie2 (v2.3.5.1) [54] with parameters '-X 1000 –no-mixed –dovetail –no-unal –very-sensitive-local -N 1'. Genome builds used were hg38/GRCh38 with gene annotation GENCODE Release 40 [55] for human, mm39/GRCm39 with gene annotation GENCODE Release M33 [55] for mouse, R64-1-1 (Saccharomyces Genome Database, https://www.yeastgenome.org/) for yeast and DH5alpha strain for *E. coli* (GenBank entry CP026085.1). Files were processed as needed with samtools (v1.10) [56] and bedtools (v2.26.0) [57].

For the validation analyses of specific experiments (GSE104550, GSE221701, GSE194217, GSE223997, GSE171327, GSE157095 and GSE166221), we followed their original published methodology for sample processing and peak calling [42, 43, 45–49].

Construction of the greenlist

Binning was performed with 1-kb windows while blacklisting only the 'Low-Mappability' regions of the ENCODE blacklist [34].

Downloaded from https://academic.oup.com/bib/article/25/2/bbad538/7590321 by Universidade de São Paulo user on 22 April 2024

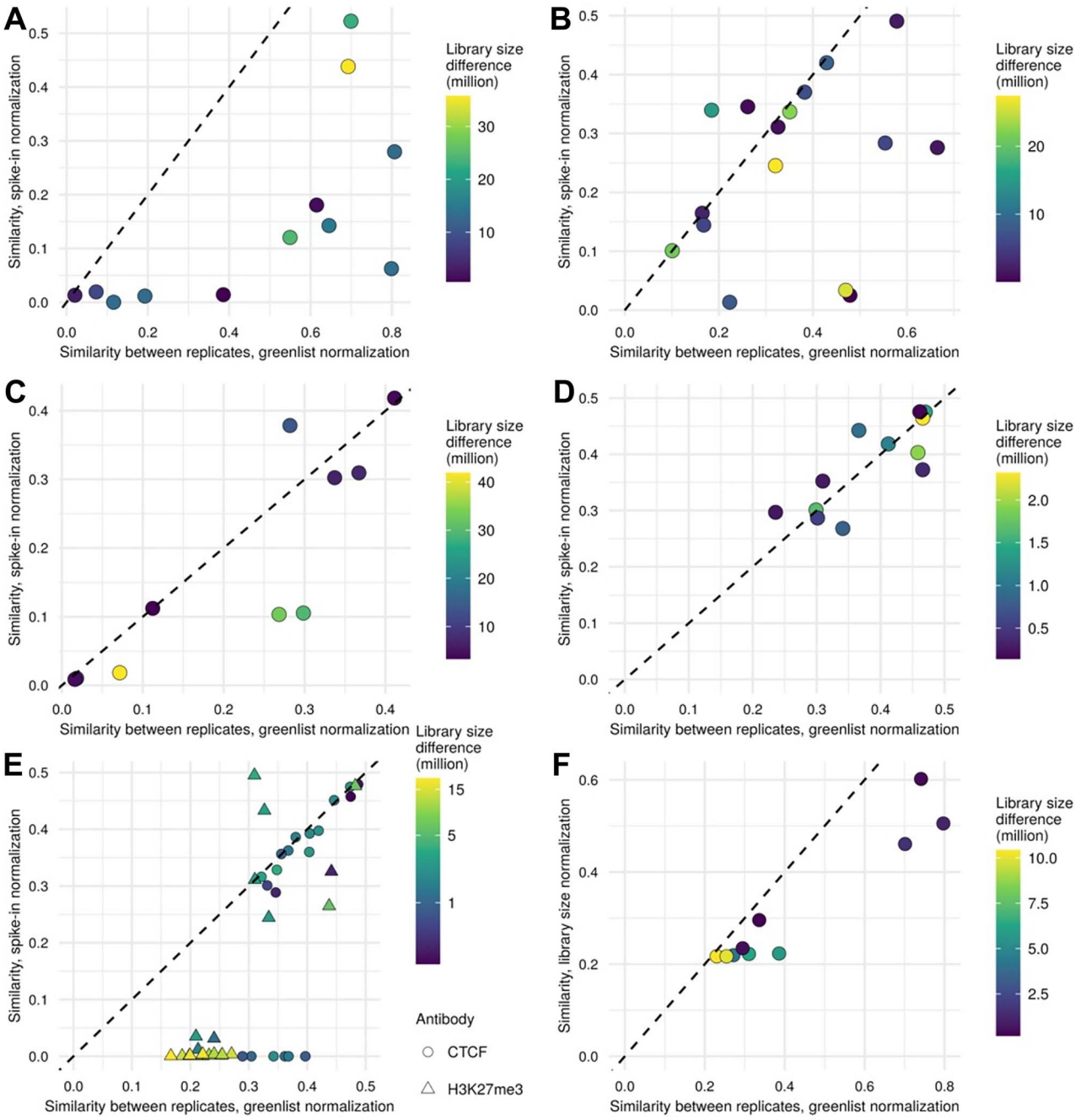

Figure 9. Greenlist normalization increases similarity between biological replicates. Comparison of normalization approaches for increasing similarity between replicates, as measured by F1 scores, with each point representing one replicate pair. Slope 'y = 1x' highlighted as a dashed line for visual clarity; points to the right of this line indicate a higher similarity between replicates normalized by greenlist compared with spike-in. The color scale on the right of each panel indicates the difference in the total number of reads (in millions) between each pair of replicates. (**A**) Similarity of replicate pairs of dataset GSE194217, of human CUT&RUN, when normalized by greenlist (*x*-axis) or spike-in (*y*-axis). (**B**) Similarity of replicate pairs of dataset GSE223997, of human CUT&Tag, when normalized by greenlist or spike-in (as above). (**C**) Similarity of replicate pairs of dataset GSE171327, of mouse CUT&RUN, when normalized by greenlist or spike-in (as above). (**D**) Similarity of replicate pairs of dataset GSE157095, a human CUT&RUN experiment with low experimental variation, when normalized by greenlist or spike-in (as above). (**E**) Similarity of replicate pairs of dataset GSE104550 by the Henikoff laboratory, testing different amounts of starting cells, when normalized by greenlist or spike-in (as above). (**F**) Similarity of replicate pairs of dataset GSE166221, of human and mouse CUT&RUN, when normalized by greenlist (*x*-axis) or default MACS2 parameters (*y*-axis).

Quantifications were done with deeptools (v3.5.1) [58] multiBamSummary, either in bins mode for bins or bed-file mode for quantifying finished greenlists. Quantified bins were normalized by quantile normalization with the broman (v0.80) R package [59], and Shannon entropy was calculated with the entropy (v1.3.1) R package [39], using the maximum-likelihood approximation. Briefly, this metric quantifies the expected value of the information content of a variable; in this case, for each bin X, we consider the proportion of counts per sample over the total counts of X as the probability to calculate X's information content, so that bins with high-count outliers have low information entropies and bins with a more homogenous distribution have high entropies.

Downloaded from https://academic.oup.com/bib/article/25/2/bbad538/7590321 by Universidade de São Paulo user on 22 April 2024

For calculating normalization factors, we tested both a manual approach (simply dividing (library size)/sum(greenlist counts)) or using DESeq2 (v1.32.0) [44] to calculate size factors; both approaches showed similar results, so DESeq2 was used for convenience.

After testing, final constructions of the greenlists were done by selecting bins in the top 0.1% of highest entropies, filtering out bins <5 kb away from known genes, and extending them if the bins in their proximity were in the top 1% of highest entropies. Blacklist constructions were done in a similar fashion but selecting bins with median normalized counts in the top 0.1 and 1% of highest signal instead.

In the interest of reproducibility, the R scripts used for greenlist construction and validation are provided at https://github.com/fndemello/CUT-RUN_greenlist.

Validations and statistics

For comparisons between lists, correlation was calculated as Spearman's rank correlation coefficient, ranking bins by entropy (for greenlists) or highest median signal (for blacklists). Overlap comparisons were calculated by finding intersects with bedtools [57] and calculating *F*-scores for each pair of lists. Likewise, *F*-scores were also used to quantify similarity between peaksets for validation of replicate pairs.

The Monte Carlo test to validate our analysis of normalization efficiency of different thresholds (Figures 2C and 3B) was done with base R, the doParallel (v1.0.16) package for parallelization and doRNG (v1.8.6) package to ensure random seed consistency across worker threads. Tests were done by selecting 0.1% of bins (2789) with the sample function and testing their effect on entropy of our test set (10% of lowest entropy bins, as previously), with 100 000 trials. Efficiency was also assessed after filtering out sampled regions that were close/overlapping gene bodies. A similar methodology was employed for assessing greenlist robustness over subsampling (Figure 3F), randomly sampling at intervals of 2% of the total number of samples. To speed computations, and since larger samplings are expected to present less variability of results than smaller samplings, the number of trials per interval was inversely proportional to the size of the sampling, i.e. smaller subsamplings were repeated more times, following the expression:

$$\text{Trials} = 2 \times (100 - \text{sampling percentage})$$

For analyzing the association of meta-factors to the overall variation of the datasets, PCA was performed, and a linear regression model was built for each of the first five PCs. Significance was tested with analysis of covariance (ANCOVA) *F*-test, with a threshold of $P \leq 0.01$ for significance with the R package car (v3.0-12) [60]. It is worth noting that this statistical design is by nature unbalanced, as we cannot ensure that all categorical variables are equally represented without greatly downsampling our dataset. As such, some groups are incomplete (i.e. not every combination of variable values is present) and some values have a small representation (e.g. antibodies that were only used once across the whole dataset), slightly limiting the estimations; nevertheless, ANCOVA should be resilient enough to this imbalance.

Additional statistical analyses were done with R (v4.1.0) [61], and plotting was done with ggplot2 (v3.4.2) [62].

Key Points

- CUT&RUN experiments always yield consistent nonspecific noise over a few genomic regions, regardless of experimental conditions; we have named these regions a 'Greenlist'.
- Greenlist regions show consistent sequencing representation and thus are effective endogenous normalizing factors for genome-wide quantitative epigenome mapping.
- Greenlist normalization outperforms current normalization standards and requires no additional experimental work.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

FUNDING

The São Paulo Research Foundation [grant numbers 18/23693-5, 22/11192-7, 20/02976-9]. The funder had no role in the study design, data collection, analysis, the decision to publish or the preparation of the manuscript.

DATA AVAILABILITY

The data underlying this article are available at https://github.com/fndemello/CUT-RUN_greenlist, along with the scripts used for its generation. Public datasets analyzed are available at the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) and are fully identified in Supplementary Table S3.

REFERENCES

1. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 2007;**316**(5830):1497–502.
2. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009;**10**(10):669–80.
3. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet* 2012;**13**(12):840–52.
4. Skene PJ, Henikoff S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* 2017;**6**:e21856.
5. Schmid M, Durussel T, Laemmli UK. ChIC and ChEC; genomic mapping of chromatin proteins. *Mol Cell* 2004;**16**(1):147–57.
6. Meers MP, Bryson TD, Henikoff JG, Henikoff S. Improved CUT&RUN chromatin profiling tools. *Elife* 2019;**8**:e46314. https://doi.org/10.7554/eLife.46314.
7. Meers MP, Tenenbaum D, Henikoff S. Peak calling by sparse enrichment analysis for CUT&RUN chromatin profiling. *Epigenetics Chromatin* 2019;**12**(1):42.
8. Salma M, Andrieu-Soler C, Deleuze V, Soler E. High-throughput methods for the analysis of transcription factors and chromatin modifications: low input, single cell and spatial genomic

Downloaded from https://academic.oup.com/bib/article/25/2/bbad538/7590321 by Universidade de São Paulo user on 22 April 2024

technologies. *Blood Cells Mol Dis* 2023;**101**:102745. https://doi.org/10.1016/j.bcmd.2023.102745.
9. Klein DC, Hainer SJ. Genomic methods in profiling DNA accessibility and factor localization. *Chromosome Res* 2020;**28**(1):69–85.
10. Leo L, Colonna Romano N. Emerging single-cell technological approaches to investigate chromatin dynamics and centromere regulation in human health and disease. *Int J Mol Sci* 2021;**22**(16):8809.
11. Sarkar A, Liu NQ, Magallanes J, *et al.* STAT3 promotes a youthful epigenetic state in articular chondrocytes. *Aging Cell* 2023;**22**:e13773.
12. Wolf BK, Zhao Y, McCray A, *et al.* Cooperation of chromatin remodeling SWI/SNF complex and pioneer factor AP-1 shapes 3D enhancer landscapes. *Nat Struct Mol Biol* 2022;**30**:10–21.
13. Baniulyte G, Durham SA, Merchant LE, Sammons MA. Shared gene targets of the ATF4 and p53 transcriptional networks. *Mol Cell Biol* 2023;**43**:426–49.
14. Kaya-Okur HS, Wu SJ, Codomo CA, *et al.* CUT&tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun* 2019;**10**(1):1930.
15. Gökbuget D, Lenshoek K, Boileau RM, *et al.* Transcriptional repression upon S phase entry protects genome integrity in pluripotent cells. *Nat Struct Mol Biol* 2023;**30**:1561–70.
16. Lu DY, Ellegast JM, Ross KN, *et al.* The ETS transcription factor ETV6 constrains the transcriptional activity of EWS–FLI to promote Ewing sarcoma. *Nat Cell Biol* 2023;**25**:285–97.
17. Douse CH, Tchasovnikarova IA, Timms RT, *et al.* Tasor is a pseudo-PARP that directs hush complex assembly and epigenetic transposon control. *Nat Commun* 2020;**11**:4940.
18. Janssens DH, Greene JE, Wu SJ, *et al.* Scalable single-cell profiling of chromatin modifications with sciCUT&Tag. *Nat Prot* 2024;**19**:83–112. https://doi.org/10.1038/s41596-023-00905-9.
19. Patty BJ, Hainer SJ. Transcription factor chromatin profiling genome-wide using ulicut&run in single cells and individual blastocysts. *Nat Protoc* 2021;**16**:2633–66.
20. Bartosovic M, Kabbe M, Castelo-Branco G. Single-cell CUT&TAG profiles histone modifications and transcription factors in complex tissues. *Nat Biotech* 2021;**39**:825–35.
21. Zambanini G, Nordin A, Jonasson M, *et al.* A new CUT&RUN low volume-urea (LOV-U) protocol optimized for transcriptional co-factors uncovers WNT/β-catenin tissue-specific genomic targets. *Development* 2022;**149**:dev201124.
22. Koidl S, Timmers HT. GreenCUT&RUN: efficient genomic profiling of gfp-tagged transcription factors and chromatin regulators. *Curr Protoc* 2021;**1**:e266. https://doi.org/10.1002/cpz1.266.
23. Janssens DH, Otto DJ, Meers MP, *et al.* CUT&Tag2for1: a modified method for simultaneous profiling of the accessible and silenced regulome in single cells. *Genome Biol* 2022;**23**:81.
24. Henikoff, S. CUT&Tag-Direct for Whole Cells with CUTAC V2. protocols.io, 2023. http://dx.doi.org/10.17504/protocols.io.bmbfk2jn
25. Meers MP, Janssens DH, Henikoff S. Pioneer factor-nucleosome binding events during differentiation are motif encoded. *Mol Cell* 2019;**75**(3):562–575.e5.
26. Yu F, Sankaran VG, Yuan GC. CUT&RUNTools 2.0: a pipeline for single-cell and bulk-level CUT&RUN and CUT&Tag data analysis. *Bioinformatics* 2021;**38**(1):252–4.
27. Boyd J, Rodriguez P, Schjerven H, Frietze S. ssvQC: an integrated CUT&RUN quality control workflow for histone modifications and transcription factors. *BMC Res Notes* 2021;**14**(1):366.
28. Orlando DA, Chen MW, Brown VE, *et al.* Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. *Cell Rep* 2014;**9**(3):1163–70.
29. Ghosh D, Qin ZS. Statistical issues in the analysis of ChIP-Seq and RNA-Seq data. *Genes (Basel)* 2010;**1**(2):317–34.
30. Dickson BM, Tiedemann RL, Chomiak AA, *et al.* A physical basis for quantitative ChIP-sequencing. *J Biol Chem* 2020;**295**(47):15826–37.
31. Chen K, Hu Z, Xia Z, *et al.* The overlooked fact: fundamental need for spike-in control for virtually all genome-wide analyses. *Mol Cell Biol* 2015;**36**(5):662–7.
32. Grzybowski AT, Chen Z, Ruthenburg AJ. Calibrating ChIP-Seq with nucleosomal internal standards to measure histone modification density genome wide. *Mol Cell* 2015;**58**(5):886–99.
33. Bonhoure N, Bounova G, Bernasconi D, *et al.* Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. *Genome Res* 2014;**24**(7):1157–68.
34. Amemiya HM, Kundaje A, Boyle AP. The ENCODE blacklist: identification of problematic regions of the genome. *Sci Rep* 2019;**9**(1):9354.
35. Carroll TS, *et al.* Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Front Genet* 2014;**5**:75.
36. Wimberley CE, Heber S. PeakPass: automating ChIP-Seq blacklist creation. *J Comput Biol* 2020;**27**(2):259–68.
37. Stark, R. and G.D. Brown. *DiffBind: Differential Binding Analysis of ChIP-Seq Peak Data*. 2011; Available from: http://bioconductor.org/packages/release/bioc/html/DiffBind.html.
38. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J* 1948;**27**(4):623–56.
39. Hausser J, Strimmer K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *J Mach Learn Res* 2009). Available from;**10**:1469–84. https://jmlr.csail.mit.edu/papers/v10/hausser09a.html.
40. Lai WKM, Mariani L, Rothschild G, *et al.* A chip-exo screen of 887 protein capture reagents program transcription factor antibodies in human cells. *Genome Res* 2021;**31**:1663–79.
41. Nordin A, Zambanini G, Pagella P, Cantù C. The CUT&RUN suspect list of problematic regions of the genome. *Genome Biol* 2023;**24**:185.
42. Skene PJ, Henikoff JG, Henikoff S. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nat Protoc* 2018;**13**(5):1006–19.
43. Singh S, Abu-Zaid A, Jin H, *et al.* Targeting KDM4 for treating Pax3-FOXO1–driven alveolar rhabdomyosarcoma. *Sci Transl Med* 2022;**14**:eabq2096.
44. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**(12):550.
45. Agrawal-Singh S, Bagri J, Giotopoulos G, *et al.* HOXA9 forms a repressive complex with nuclear matrix-associated protein SAFB to maintain acute myeloid leukemia. *Blood* 2023;**141**(14):1737–54.
46. Zou H, Poore B, Brown EE, *et al.* A neurodevelopmental epigenetic programme mediated by SMARCD3-DAB1-Reelin signalling is hijacked to promote medulloblastoma metastasis. *Nat Cell Biol* 2023;**25**(3):493–507.
47. Xu C, Li C, Chen J, *et al.* R-loop-dependent promoter-proximal termination ensures genome stability. *Nature* 2023;**621**(7979):610–9.
48. Weigel B, Tegethoff JF, Grieder SD, *et al.* MYT1L haploinsufficiency in human neurons and mice causes autism-associated phenotypes that can be reversed by genetic and pharmacologic intervention. *Mol Psy* 2023;**28**(5):2122–35.
49. Vinjamur DS, Yao Q, Cole MA, *et al.* ZNF410 represses fetal globin by singular control of CHD4. *Nat Genetics* 2021;**53**(5):719–28.

Downloaded from https://academic.oup.com/bib/article/25/2/bbad538/7590321 by Universidade de São Paulo user on 22 April 2024

50. Kong NR, Chai L, Tenen DG, Bassal MA. A modified CUT&RUN protocol and analysis pipeline to identify transcription factor binding sites in human cell lines. *STAR Protoc* 2021;**2**(3):100750. https://doi.org/10.1016/j.xpro.2021.100750.
51. *EpiCypher*. Available from: https://www.epicypher.com/content/documents/protocols/cutana-cut&run-protocol-2.1.pdf.
52. *Cell Signaling Technology*. Available from: https://www.cellsignal.com/learn-and-support/protocols/cut-and-run-protocol.
53. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**(17):i884–90.
54. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**(4):357–9.
55. Frankish A, Diekhans M, Jungreis I, *et al*. Gencode 2021. *Nucleic Acids Res* 2021;**49**(D1):D916–23.
56. Danecek P, Bonfield JK, Liddle J, *et al*. Twelve years of SAMtools and BCFtools. *Gigascience* 2021;**10**(2):giab008.
57. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**(6):841–2.
58. Ramirez F, *et al*. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 2016;**44**(W1):W160–5.
59. Broman K, Tian J. *kbroman/broman: Version 0.80 (0.80) Zenodo*, 2022, Available from: https://doi.org/10.5281/zenodo.6811647.
60. Fox, J. and S. Weisberg. *An R Companion to Applied Regression,* 3rd ed. 2019; SAGE Publications, Inc., USA. Available from: https://socialsciences.mcmaster.ca/jfox/Books/Companion/.
61. R Core Team. *R: A Language and Environment for Statistical Computing*. 2021; Available from: https://www.R-project.org/.
62. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*, 2nd edn. Springer, Switzerland, 2016, ISBN 978-3319242750.

Downloaded from https://academic.oup.com/bib/article/25/2/bbad538/7590321 by Universidade de São Paulo user on 22 April 2024