Beyond Blacklists: A Critical Assessment of Exclusion Set Generation Strategies and Alternative Approaches

Brydon P. G. Wall[1,2], Jonathan D. Ogata[1,2], My Nguyen[2], Joseph L. McClay[3], J. Chuck Harrell[4,5], \ Mikhail G. Dozmorov[1,2,4*]

[1]These authors contributed equally

[2]Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, 23298, USA.

[3]Department of Pharmacotherapy and Outcomes Science, Virginia Commonwealth University, Richmond, VA, 23298, USA.

[4]Department of Pathology, Virginia Commonwealth University, Richmond, VA, 23284, USA.

[5]Massey Comprehensive Cancer Center, Virginia Commonwealth University, Richmond, VA 23298, USA.

[*]To whom correspondence should be addressed: mdozmorov@vcu.edu

Abstract

Short-read sequencing data can be affected by alignment artifacts in certain genomic regions. Removing reads overlapping these exclusion regions, previously known as Blacklists, help to potentially improve biological signal. Tools like the widely used Blacklist software facilitate this process, but their algorithmic details and parameter choices are not always clearly documented, affecting reproducibility and biological relevance. We examined the Blacklist software and found that pre-generated exclusion sets were difficult to reproduce due to variability in input data, aligner choice, and read length. We also identified and addressed a coding issue that led to over-annotation of high-signal regions. We further explored the use of "sponge" sequences—unassembled genomic regions such as satellite DNA, ribosomal DNA, and mitochondrial DNA—as an alternative approach. Aligning reads to a genome that includes sponge sequences reduced signal correlation in ChIP-seq data comparably to Blacklist-derived exclusion sets while preserving biological signal. Sponge-based alignment also had minimal impact on RNA-seq gene counts, suggesting broader applicability beyond chromatin profiling. These results highlight the limitations of fixed exclusion sets and suggest that sponge sequences offer a flexible, alignment-guided strategy for reducing artifacts and improving functional genomics analyses.

Introduction

Alignment of short-read sequencing data to reference genome assemblies poses significant challenges due to the presence of low-complexity regions, centromeres, telomeres, satellite repeats, chromatin accessibility biases, and other artifacts [1]. These regions often result in abnormal read pileups caused by ambiguous alignments or biases introduced during library preparation steps, such as cell fixation or PCR amplification after adapter ligation [2,3]. Such artifacts have been observed across multiple species [4,5] and are particularly pronounced in chromatin-targeting sequencing technologies, including ChIP-seq [6,7,5], ChIP-exo [8], and CUT&RUN [9]. Removing these artifact signals has been shown to reduce noise and improve peak calling, signal normalization, and motif analysis accuracy [8,10,11,9,2]. Hence, defining strategies for removing such problematic signals is crucial for accurate genomic data analysis and interpretation.

Lists of exclusion regions (formerly known as blacklist regions), defined as genomic coordinates of problematic regions, have become a commonly accepted method for excluding artifact signals [12]. Several exclusion region sets have been developed for humans and model organisms, with an overview provided in [8]. The ENCODE project and others have compiled exclusion sets for various genome assemblies [13–15]. With technological advancements, such as long-read sequencing, sequencing data for newer organisms, and updated genome

assemblies for existing organisms—such as the Telomere-to-Telomere (T2T) human genome assembly [16]—generating and updating these exclusion sets remains a priority. Notably, the commonly used liftOver tool for coordinate conversion has been deemed unsuitable for translating exclusion set coordinates across genome assemblies [13], highlighting the need for standardized exclusion set definitions and reproducible methods for generating them.

Methods for creating exclusion sets range from ad hoc approaches [17,5,9] to dedicated tools. Among these, the Blacklist software [13] is widely used. This tool employs BAM files from ChIP-seq input experiments (expected to have uniform non-specific sequencing coverage) and mappability files generated by Umap [18]. However, several key parameters of its algorithmic implementation are not fully documented, including the number of BAM files required, the read length and aligner specifications, and the k-mer length for selecting low-mappability files. The impact of these parameters on exclusion set generation remains unclear. Additionally, the software merges regions within a fixed distance of 20,000 bp—a parameter that appears excessive and is not easily adjustable. While these simplifications allow for straightforward use as a black-box tool, they raise concerns about the robustness and biological relevance of the resulting regions, limiting its applicability across sequencing technologies and reference genomes.

Alternative methods also have limitations. PeakPass employs a random forest model trained on hg19 ENCODE blacklist regions to predict excludable regions in hg38. It identifies assembly gaps and genome complexity as key predictors but lacks a user-friendly implementation. Similarly, Greenscreen [19] uses a traditional approach by processing ChIP-seq data through standard pipelines and calling peaks with MACS2 [20]. While Greenscreen claims 99% agreement with Blacklist-generated regions, its reliance on fixed parameters and a technically challenging Docker-based setup hampers its utility. The GreyListChIP R package [21], developed in 2015, defines excludable regions using a tiled genome and merges regions within a fixed distance. However, its unpublished algorithm and untested parameters further limit its reliability, particularly for less-studied organisms.

An alternative approach involves aligning sequencing data to reference genomes that include so-called "decoy" or "sponge" sequences. Such sequences typically consist of regions not included in the standard human reference genome. For example, the hs38d1 genome assembly (GCA_000786075.2) includes such sequences for the GRCh38/hg38 human genome assembly. A more comprehensive approach uses sponge sequences, which encompass unmapped and uncharacterized regions, satellite repeats, ribosomal sequences, and mitochondrial sequences. Incorporating these sponge sequences into the genome assembly during alignment has been shown to reduce signal in Blacklist exclusion regions and mitigate other alignment artifacts [1]. Despite its advantages, this approach is less widely adopted, with only 42 citations compared to 1,410 citations for the Blacklist manuscript (Google Scholar, January 2025).

In this study, we systematically benchmarked the performance of the Blacklist software [13] and related factors, such as aligner choice, to develop recommendations for exclusion set generation. Focusing on human and mouse genomes, we identified reproducibility and quality issues in existing exclusion sets. We provide a corrected, configurable version of the Blacklist software and propose improvements for defining exclusion regions. Rather than relying on fixed lists of exclusion regions, our results emphasize the importance of using "sponge" sequences at the alignment step. Our findings underscore the importance of transparent algorithms and adaptable methodologies to address the ongoing challenges in sequencing data analysis.

Results

Filtering signal overlapping excludable (also known as Blacklist) regions is a standard practice in genomic data analysis. These regions were originally defined using input ChIP-seq data [8,13,22]. The simplicity of excludable regions (genomic coordinates in BED format) has enabled their application in analyzing genomic data generated by technologies based on assumptions other than ChIP-seq, such as ATAC-seq [23,24], chromatin conformation capture technologies [25], and their single-cell variants [26–28], as well as in pipelines and tools for genomic data analysis [29–31].

The ENCODE Blacklist exclusion set is arguably one of the most cited and frequently used exclusion sets for

the human genome, and the Blacklist software implementation has been employed to generate exclusion sets for model organisms [13]. However, we have observed instances where signals expected to be associated with functionally relevant genes were missed, and such genes often overlapped with or were located in proximity to Blacklist exclusion regions. These observations prompted us to investigate the properties of the ENCODE Blacklist exclusion sets and to develop a set of recommendations for their optimal use.

Challenges in reproducing pre-generated Blacklist exclusion sets.

The Blacklist GitHub repository offers pre-generated exclusion sets (also referred to as lists) for human (hg19, hg38), mouse (mm10), Drosophila (dm3, dm6), and worm (cd10, ce11). We noted that version 1 of the hg38 exclusion set contained 38 regions, in contrast to 636 regions in version 2. In an effort to reproduce the hg38 "GitHub Blacklist," we ran the software using the same set of 250 BAM files described in the original publication, generating what we refer to as the "Generated Blacklist."

Our results indicated some differences compared to the GitHub version. Specifically, the "Generated Blacklist" contained more regions (1,273 vs. 636 in the GitHub version, Table 1), which were generally narrower (mean width approximately 213 Kbp vs. 357 Kbp in the GitHub version). The total genome coverage was also higher (approximately 271 Mbp vs. 227 Mbp in the GitHub version, Figure 1A), with similar telomere, centromere, and short-arm coverage (referred to hereafter as gap coverage), except for the short arms, which were primarily covered by our generated list (Figure 1B). Most regions from the "GitHub Blacklist" overlapped with the "Generated Blacklist" (79.4%, covering 97.1% of the same bases, Figure 1B). However, both lists showed notable differences from the "Kundaje Unified" list, which is recommended as a manually curated gold standard [12]. Only 18.1% of regions from the "GitHub Blacklist" overlapped with regions from the "Kundaje Unified" list, with overlap by width similarly limited (28.8%, Figure 1C).

For the mm10 list, our results showed fewer regions compared to the GitHub version (2,970 vs. 3,435), but the regions were wider on average (49,076 bp vs. 41,235 bp) and covered more bases (approximately 253 Mbp vs. 238 Mbp in the GitHub version) (Supplementary Figure S1A, Supplementary Table S2). The overall gap coverage between the two mouse Blacklists was comparable (Supplementary Figure S1A). While approximately 30% of regions were unique to each list, the shared regions accounted for over 80% of common base coverage (Supplementary Figure S1C).

To support reproducibility, we have made our analysis scripts available (Code Availability). These findings highlight some variability in exclusion sets generated by the Blacklist software and suggest the value of cross-comparison with curated references like the "Kundaje Unified" list.

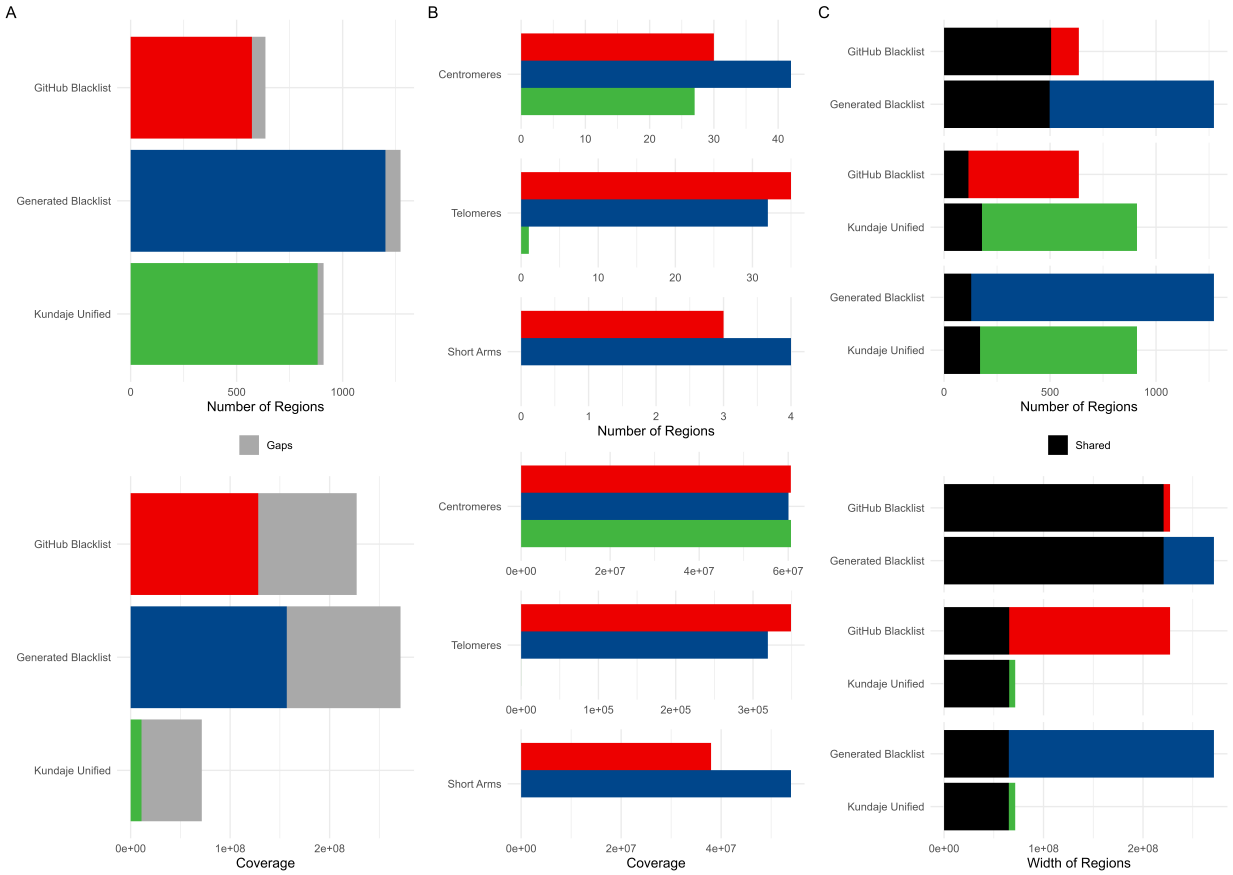

Figure 1. Differences between the GitHub version of the hg38 exclusion set, our hg38 exclusion set generated with the Blacklist software, and the reference Kundaje Unified set. (A) Differences in count, coverage, and width distribution. (B) Differences in gap coverage. (C) Pairwise overlaps of region counts and genome coverage between sets.

Table 1. Characteristics of the hg38 exclusion sets. Total number of regions, average region width, number of regions overlapping gaps, and proportion of gap coverage.

	Total		Summary of region widths, bp			Number of regions / Coverage proportion of		
List	**Count**	**Coverage, bp**	**Minimum**	**Mean / Median**	**Max**	**Centromeres**	**Telomeres**	**Short Arms**
GitHub Blacklist	636	227162400	1200	357174 / 10150	30590100	30 / 97.6%	35 / 72.7%	3 / 58.9%
hg38 Generated List	1273	271267100	1000	213093 / 6300	30602600	42 / 96.7%	32 / 66.5%	4 / 83.7%
mm10 GitHub List	3435	238977200	1000	69571 / 8100	50585400	72 / 2.7%	17 / 35.4%	86 / 23.4%
mm10 Generated List	2970	253654600	1000	85406 / 12600	91744600	74 / 4.0%	17 / 37.5%	76 / 24.0%
hg38 Kundaje Unified	910	71570285	19	78649 / 384	5407756	27 / 97.7%	1 / 0.0%	0 / 0.0%
High Signal + Centromere	2831	71535766	1000	25269 / 3083	5400309	24 / 100.0%	1 / 0.0%	2 / 0.0%
Low Mappability	3691	153220769	1001	41512 / 1823	18171354	984 / 28.7%	46 / 95.8%	5 / 100.0%
HS + LM + CM	5409	206519467	1000	38181 / 2885	18223524	23 / 99.6%	47 / 95.8%	5 / 100.0%

	Total		Summary of region widths, bp			Number of regions / Coverage proportion of		
List	**Count**	**Coverage, bp**	**Minimum**	**Mean / Median**	**Max**	**Centromeres**	**Telomeres**	**Short Arms**
hg38 Nordin CUT&RUN	884	4463850	2	5050 / 2878	93434	555 / 4.2%	7 / 1.4%	0 / 0.0%
GreyListChIP STAR 1k	22480	9270576	1	412 / 36	140768	5673 / 4.5%	0 / 0.0%	0 / 0.0%

Supplementary Figure S1. Differences between the GitHub version of the mm10 exclusion set and our mm10 exclusion set generated with the Blacklist software. (A) Differences in count, coverage, and width distribution. (B) Differences in gap coverage. (C) Pairwise overlaps of region counts and genome coverage between sets.

To explore the differences between the exclusion sets provided on GitHub and those we generated, we examined the properties of the 1,255 BAM files used to create the hg38 "GitHub Blacklist." Our analysis indicated that a significant portion of the BAM files (37.6%) originated from 36 bp-long reads, followed by 101 bp-long reads (21.8%). However, read length varied considerably, with files generated from reads of 28 bp, 50 bp, 76 bp, and other lengths (Supplementary Figure S2A, Supplementary Table S1). Additionally, 365 out of 1,255 files (29.1%) were derived from paired-end experiments and aligned using bwa sampe v.0.7.10, while single-end files were aligned with bwa samse. The average number of mapped reads also differed, with paired-end files having approximately 89 million mapped reads, compared to 38 million for single-end files (Supplementary Figure S2B).

We observed further variation in the dataset, including 256 files that combined multiple (pooled) FASTQ files and 60 that used cropped FASTQ files, where the cropping parameters were not specified. Of the total BAM files, 135 were restricted (FASTQ sequences unavailable), 52 contained duplicated FASTQ files, and 4 were both restricted and duplicated (Supplementary Table S1). These BAM files were merged into 250 donor-specific files in the original publication [13], with the number of BAM files per donor ranging from 1 to 149 (average of 5) (Supplementary Figure S2C). We also noted variability in donor-specific BAM files, which differed in read length, mixed single- and paired-end sequencing, and were derived from various cell and tissue types (Supplementary Figure S2D, Supplementary Table S1). To assess whether merging files affected exclusion set generation, we compared exclusion sets generated from 274 101 bp paired-end BAM files (spanning 21 donors) to those produced from a single BAM file containing all merged 101 bp data. The differences between the resulting exclusion sets were minimal (Supplementary Figure S3).

These observations suggest that the diversity in input data characteristics, along with differences in sequencing and alignment protocols, may influence the reproducibility and consistency of exclusion sets generated by the Blacklist software. By highlighting these factors, we hope to encourage further consideration of input data variability in exclusion set generation.

Supplementary Figure S2. The diversity of BAM files used to generate the hg38 exclusion set. (A) Read length proportions among 1,225 BAM files. (B) Number of mapped reads for single- and paired-end files. (C) Number of single- and paired-end files per donor. (D) Total number of mapped and unmapped reads per donor.

Supplementary Table S1. Diverse characteristics of human sequencing data used by the Blacklist software. "bam_accession_table" – Accession numbers of 1,255 BAM files and the associated FASTQ files ("R1 FASTQs", "R2 FASTQs"), their paired-end status, and indicators of whether the data was cropped, pooled, duplicated, or restricted. Also includes BAM read length and the number of mapped and unmapped reads. "donor_accession_table" – Accession numbers of 250 donors and the associated BAM files, total mapped and unmapped reads per donor, number of BAM files per donor, read length, and tissue ontology. "fastq_accession_table" – Characteristics of the original FASTQ files used to generate sample-specific BAM files.

Supplementary Figure S3. Exclusion set characteristics generated from the 274 unmerged 101bp BAM files and the corresponding merged BAM file. (A) Differences in count, coverage, and

width distribution. (B) Differences in gap coverage. (C) Pairwise overlaps of region counts and genome coverage between sets.

Manually defined High Signal and Low Mappability regions are different from other exclusion sets.

To evaluate the reproducibility of exclusion sets in the absence of a definitive gold standard for excludable regions, we chose the manually curated "Kundaje Unified" list as a reference for known hg38 exclusion regions. This list includes a larger number of regions compared to the GitHub-provided exclusion set (910 regions in the "Kundaje Unified" list vs. 636 in the GitHub version, Figure 2A). Of these, 179 regions overlapped with the GitHub list (Figure 2B, Supplementary Figure S4B). The "Kundaje Unified" regions were generally narrower (mean width of 78.6 Kbp compared to 357.2 Kbp in the GitHub list, Figure 2C) and covered less of the genome (71.6 Mbp for the "Kundaje Unified" list vs. 227.2 Mbp for the GitHub list, Table 1). While the "Kundaje Unified" list provided near-complete coverage of centromeres (97.7%), it did not extend to telomeres or short arms (0.0%, Supplementary Figure S4C). In contrast, the GitHub list covered 97.6% of centromeres, 72.7% of telomeres, and 58.9% of short arms (Supplementary Figure S4C). For our benchmarking, we used the "Kundaje Unified" list as the primary reference, recognizing its role as a curated and widely referenced set of exclusion regions.

High Signal (HS) regions are another well-characterized category of exclusion regions [24]. Using 274 101 bp paired-end BAM files, we identified "High Signal" regions by applying MACS3 to call peaks with fold changes exceeding the 99th percentile. We merged regions within 1,000 bp and excluded those smaller than 1,000 bp (Methods). Our analysis showed that the "High Signal" regions had limited coverage of centromeres compared to the "Kundaje Unified" list (2.1% vs. 97.7%). To address this, we incorporated centromeric regions into our "High Signal" list, producing a final set of 2,657 regions (Figure 2A). This adjusted list covered a comparable portion of the genome and gap regions as the "Kundaje Unified" list (71.0 Mbp vs. 71.6 Mbp, Supplementary Figures S4A and S4B). The "High Signal" regions were generally narrower than those in the "Kundaje Unified" list and the GitHub Blacklist (mean width of 26.7 Kbp, Figure 2B). While 422 "High Signal" regions overlapped with the GitHub Blacklist, accounting for 27.3% of its total coverage, the overlap with the "Kundaje Unified" list was smaller in number (113 regions) but encompassed a significant portion of its base coverage (85.2%, Figure S4C). This "High Signal" list offers a conservative representation of high-signal regions derived from input ChIP-seq data.

Low Mappability (LM) regions represent another category considered by the Blacklist software. Following an approach similar to that used for "High Signal" regions, we defined "Low Mappability" regions as those falling below the 1st percentile of the mappability range. Regions within 1,000 bp were merged, and regions smaller than 1,000 bp were excluded (Methods). Despite applying this conservative strategy, we identified a substantial number of "Low Mappability" regions (12,455, Figure 2A), which tended to be wide (mean width of 41.5 Kbp, Figure 2B). These regions covered a significant portion of the genome (153.2 Mbp, Figure S4A), including centromeres (28.7% coverage), telomeres (95.8%), and short arms (100%, Supplementary Figure S4C). The "Low Mappability" regions overlapped with 228 regions from the GitHub Blacklist, accounting for 51.3% of its coverage (Figure 2C). They also intersected with 45 regions from the "Kundaje Unified" list, covering 26.7% of its bases (Figure 2C). This "Low Mappability" list reflects a conservative set of regions that are challenging to map, and it appears to differ from both the GitHub Blacklist and the "Kundaje Unified" lists.

To align with the characteristics of the Blacklist-generated list, which includes both high signal and low mappability regions, we combined the "High Signal" (with centromeres) and "Low Mappability" lists, referred to as "HS + LM". This resulted in 5,409 regions with a mean width of 38.1 Kbp, covering 206.5 Mbp of the genome and nearly complete gap coverage (Figure 2A–B, Supplementary Figure S4A–B, Table 1). Although 2,078 (38.4%) of the "HS + LM" regions overlapped with the GitHub Blacklist, they accounted for 70.6% of its total coverage. Similarly, while only 173 (3.2%) regions overlapped with the "Kundaje Unified" list, they covered 87.0% of it, largely driven by centromeric regions (Figure 2C, Supplementary Figure S4C). The "HS + LM" list, created as a single reference for addressing genomic regions prone to mapping artifacts, highlights distinctions from both the GitHub Blacklist and the "Kundaje Unified" list.

To explore similarities between exclusion lists, we visualized Jaccard overlap indexes among them using multidimensional scaling (Methods). The "GitHub Blacklist" and our "Generated Blacklist" showed similarity, while the "Kundaje Unified," "High Signal," "Low Mappability," and "HS + LM" lists appeared more distinct (Figure 2D). To account for differences in list size, region number, and genome coverage, we applied the Forbes width overlap coefficient, known for its robustness in such contexts [32]. This analysis indicated that the "Kundaje Unified" and "High Signal" lists share some similarities and align more closely with the "GitHub Blacklist," whereas the "Low Mappability" and "HS + LM" lists diverge (Supplementary Figure S4D). These findings highlight notable heterogeneity among exclusion lists, suggesting that no single reference set universally defines problematic regions. Moving forward, we benchmark the Blacklist software using the "Kundaje Unified" list alongside the "High Signal," "Low Mappability," and "HS + LM" lists to provide a broader evaluation of performance.

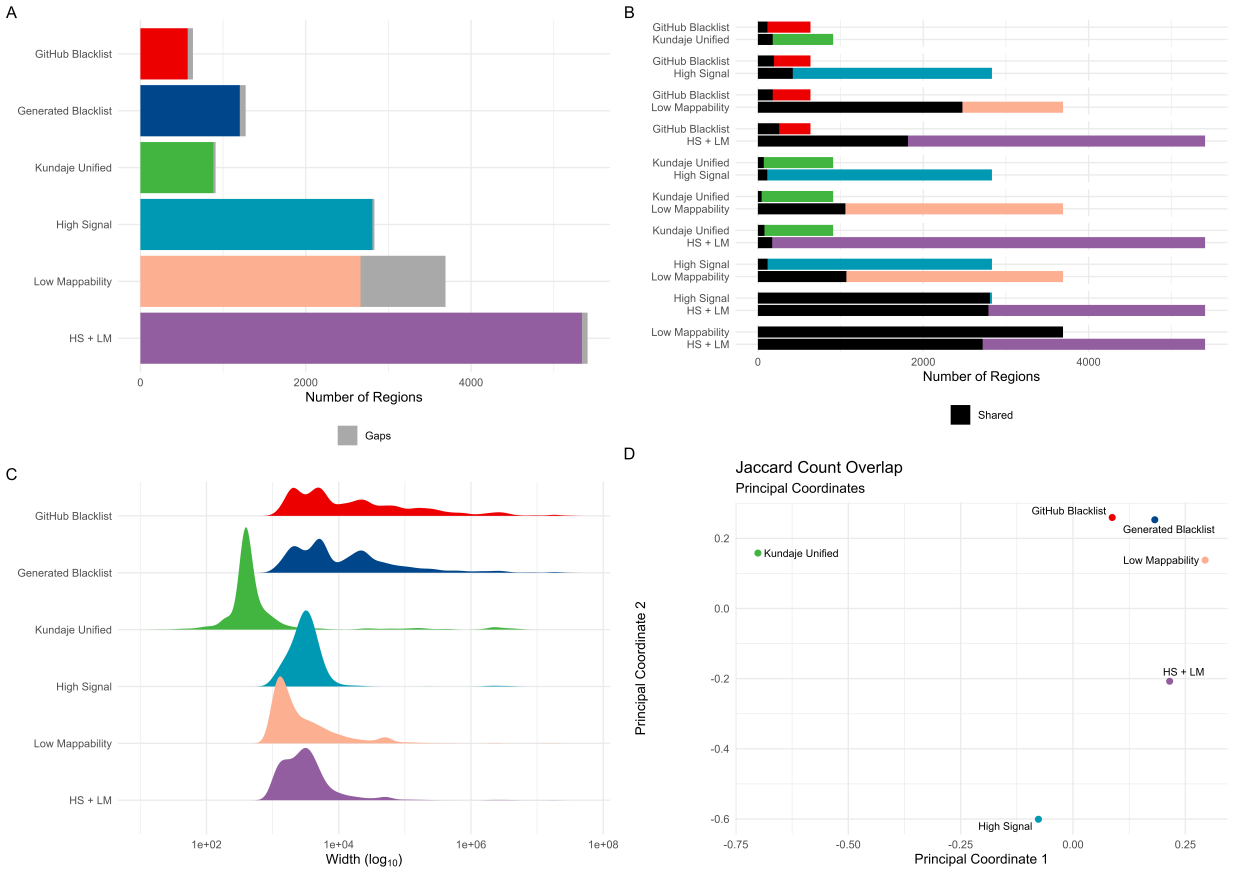

Figure 2. Differences between the GitHub version of the hg38 exclusion set and manually defined gold-standard exclusion sets. (A) Differences in count. (B) Pairwise overlaps of region counts between the GitHub exclusion set and manually defined gold-standard exclusion sets. (C) Differences in width distribution. (D) Multidimensional scaling plot of Jaccard count similarity among the GitHub exclusion set and manually defined gold standards.

Supplementary Figure S4. Differences between the GitHub version of the hg38 exclusion set and manually defined gold-standard exclusion sets. (A) Differences in width. (B) Pairwise width overlaps between the GitHub exclusion set and manually defined gold-standard exclusion sets. (C) Differences in gap coverage. (D) Multidimensional scaling plot of Forbes width similarity among the GitHub exclusion set and manually defined gold standards.

Excessive High Signal annotation of Blacklist-generated regions.

Exclusion sets produced by the Blacklist software include annotations indicating whether regions are classified as "High Signal" or "Low Mappability," with no overlap between categories. We observed that "High Signal" annotations were predominant across all organisms' lists ($83.75 \pm 9.54\%$), with the hg38 "GitHub Blacklist" showing a particularly strong bias (93.40%, Supplementary Table S2). To investigate this pattern, we examined the Blacklist software's algorithm and original C code (Supplementary Note, Supplementary Figure S5). Our analysis revealed that the software defaulted to "High Signal" annotations in cases of ambiguity, as it could not output combined classifications. We addressed this by modifying the code to allow for a new annotation, "High Signal, Low Mappability," and have made the updated version available (Code Availability).

The modified code produced consistent results (Supplementary Table S2) and, for the hg38 genome assembly, identified the same 65 "Low Mappability" regions spanning 7.4 Mbp. Of the 1,208 "High Signal" regions (covering 263.8 Mbp), 183 were reclassified as "High Signal, Low Mappability." Although these reclassified regions were few, they accounted for 219.4 Mbp, representing 80.88% of the total genome coverage in the Blacklist-generated hg38 exclusion set. Similar trends were observed for the mm10 assembly (Supplementary Table S2). These findings underscore the importance of recognizing multi-annotation regions in exclusion set generation and highlight the potential for annotation biases to influence downstream analyses.

Supplementary Figure S5. The Blacklist Algorithm. A) **Binning parameters:** Two internal parameters, "binSize" and "binOverlap," define how bins are created. B) **Read sorting:** The start position of each read is sorted according to the read's length and compared using the given pre-calculated mappability vector. C) **The "binsMap" vector:** This vector is generated using the "uniqueLength" threshold. D) **Normalization:** The signal ("Unique") and mappability ("Multi") vectors are divided, quantile normalized, and the median is taken per bin. E) **Region calling:** This step uses several predefined internal parameters/thresholds to call and annotate excludable regions.

Supplementary Table S2. Dominant High Signal annotation counts and proportions in exclusion sets generated by the Blacklist software. The "Corrected List" refers to exclusion sets generated using our modified version of the Blacklist software, which enables dual high-signal and low-mappability annotations.

The choice of aligner and read length affects exclusion set generation.

We hypothesized that the choice of aligner could influence the resulting exclusion sets. To assess this effect, we used STAR, bwa-mem2, and bowtie2 aligners (Methods), along with the originally used bwa samse, to realign 381 36 bp single-end files and generate exclusion sets using the Blacklist software. The 36 bp read length was selected to minimize variability from read length differences and to align with the internal k-mer setting of 36 (Supplementary Note). We compared these exclusion sets with the "GitHub Blacklist" and "Generated Blacklist," which were derived from all files aligned using bwa samse, and included the "Kundaje Unified" list as a reference. Realigned sets generally contained a larger number of regions. Notably, the STAR-realigned set had a region count closest to the "GitHub Blacklist" (749 vs. 636), while the bwa-mem2-realigned set had the highest number of regions (1,174, Supplementary Table S3). The STAR-realigned set also exhibited a width distribution most similar to the "GitHub Blacklist" (mean width 381 kbp vs. 357 kbp), whereas sets from other aligners had narrower distributions (Figure 3A). Realigned sets covered a broader portion of the genome, including gap regions (Figure 3B). They overlapped approximately 60–80% of the regions in the "GitHub Blacklist" and covered up to 99% of its width (Figure 3C). Jaccard count multi-dimensional scaling and hierarchical clustering demonstrated that the STAR-aligned set closely resembled the "GitHub Blacklist," while sets from other aligners were more similar to each other and distinct from both the "GitHub Blacklist" and the "Kundaje Unified" list (Supplementary Figure S6A, B). Forbes width overlap plots also highlighted differences between the realigned sets and the reference lists (Supplementary Figure S6C, D). These findings underscore the potential influence of aligner choice on exclusion set generation and highlight the importance of considering alignment tools in such analyses.

We further hypothesized that read length could influence the resulting exclusion sets. To test this, we realigned 274 101 bp paired-end files using the same set of aligners and compared the resulting exclusion sets. The use of 101 bp files was associated with a larger number of regions that were generally narrower and covered a broader portion of the genome (Supplementary Figure S7A, Supplementary Table S3). Examining the impact of aligners, bwa-mem2 consistently identified a larger number of regions, which covered a greater portion of the genome but tended to be narrower. Conversely, STAR identified fewer regions that covered a smaller portion of the genome but were generally wider (Supplementary Table S3). Sets generated with STAR were distinct from those generated using bwa-mem2 and bowtie2. However, the segregation of sets was primarily influenced by read length, with 36 bp STAR-generated exclusion sets showing the highest similarity to the "GitHub Blacklist" (Supplementary Figure S7B). These findings suggest that the Blacklist software is sensitive to the read length of input BAM files, emphasizing the importance of considering read length in exclusion set analyses.

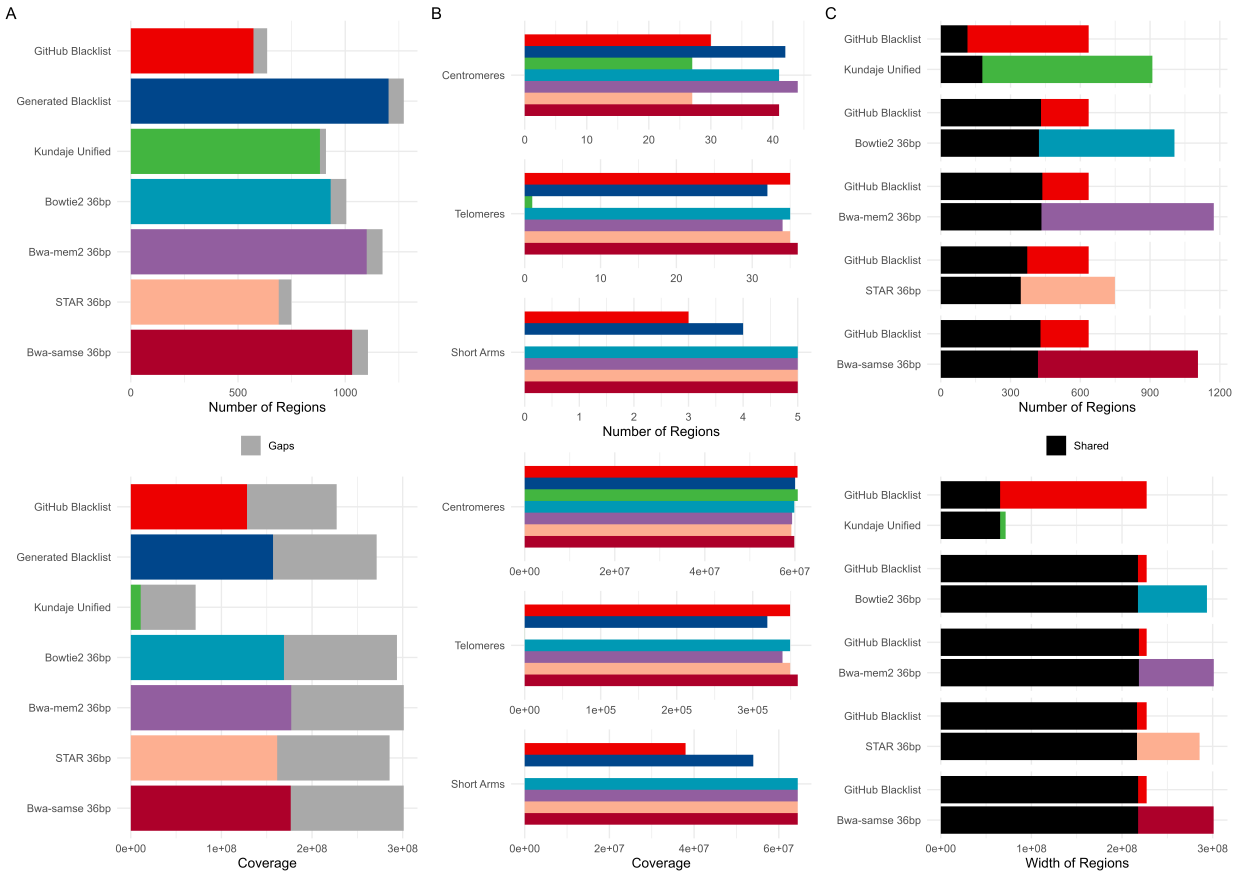

Figure 3. Differences between the hg38 exclusion sets, the reference Kundaje Unified set, and sets generated from 36bp single-end reads realigned with different aligners. A) Count, coverage, and width distribution differences; B) Gap coverage differences; C) Pairwise overlaps of counts and coverage between sets.

Supplementary Figure S6. Relationships among realigned 36bp exclusion sets. A, B) Jaccard count overlap, multidimensional scaling, and hierarchical clustering; C, D) Forbes width overlap, MDS, and clustering.

Supplementary Figure S7. Relationships among realigned 36bp and 101bp exclusion sets. A) Number and width distribution; B) Jaccard count overlap and Forbes width overlap MDS plots.

Supplementary Table S3. Characteristics of exclusion sets aligned with different aligners. Total number and average width of the regions, number of regions overlapping gaps, and proportion of gap coverage.

Optimal parameters for exclusion set calling

Recognizing the sensitivity of the Blacklist software to input settings, we systematically evaluated the effects of three parameters on the resulting exclusion sets. These parameters included the number of files (ranging from 10 to 300) to determine the optimal number needed for exclusion set generation, the "bridge" parameter (tested at 1,000, 10,000, and the default 20,000) to assess the optimal merging distance for nearby exclusion regions, and the "k-mer" parameter (default 36, 50, and 100) to evaluate the ideal k-mer size. For this analysis, we used the "Kundaje Unified," "High Signal," "Low Mappability," and "HS + LM" sets as references, along with the 381 36 bp single-end STAR-aligned BAM files, which produced exclusion sets most similar to the hg38 Blacklist (Supplementary Figure S7B, D). We measured the effects of these parameters on the number of regions, total coverage, and mean/median width.

Increasing the number of files and the "bridge" parameter generally resulted in fewer but wider regions that covered a larger portion of the genome. In contrast, increasing the k-mer parameter led to a greater number of narrower regions that also covered a larger genomic portion (Supplementary Figure S8, Supplementary Table S4). Jaccard count overlap and Forbes width similarity analyses revealed distinct parameter effects. The k-mer parameter had the most significant impact on Jaccard count overlap similarity (Supplementary Figure S9C), followed by the number of BAM files (Supplementary Figure S9A). The "bridge" parameter had the greatest influence on Forbes width similarity (Supplementary Figure S9F), followed by the k-mer parameter (Supplementary Figure S9G). Interestingly, none of the parameter combinations produced exclusion sets closely resembling the "Kundaje Unified" or the "High Signal," "Low Mappability," and "HS + LM" reference lists. However, exclusion sets generated using only 10 BAM files were unexpectedly the closest to the GitHub list by Jaccard count overlap similarity (Supplementary Figure S9D, H). These findings underscore the sensitivity of the Blacklist software to input parameters and highlight its limitations in replicating classical definitions of excludable regions, such as high-signal or low-mappability peaks.

Supplementary Figure S8. The effect of varying the number of BAM files, the "bridge" parameter, and the "k-mer" parameter on the output of the Blacklist software. The effect on the number of exclusion regions, sum of region widths, and mean and median region widths.

Supplementary Figure S9. Jaccard count overlap (A-D) and Forbes similarity (E-H) MDS plots colored by parameters. A), E) Number of BAM files; B), F) Bridge parameter; C), G) k-mer parameter; D), H) Parameter sweep results (gray) vs. reference exclusion sets.

Supplementary Table S4. Characteristics of exclusion sets generated with different parameters. Total number and average width of the regions, number of regions overlapping gaps, and proportion of gap coverage. The list names include the bridge parameter (e.g., b1000), the k-mer parameter (e.g., k36), and the number of files (e.g., n50).

Ribosomal genes are most affected by exclusion regions.

To assess the impact of different exclusion sets on known transcripts, we quantified the extent to which these sets overlapped with transcripts (Supplementary Table S5). Our analysis included two newer exclusion sets: the "Nordin CUT&RUN" list [33] and the "GreyListChIP" list, generated using the GreyListChIP R package with a 1,000 bp merging setting. These two lists exhibited similarity to each other but were distinct from other sets, except for the "Nordin CUT&RUN" list, which showed some similarity to the "GitHub Blacklist" (Supplementary Figure S10). We observed that the "GreyListChIP" list and the "High Signal" list affected the largest number of protein-coding genes, whereas the "Generated Blacklist" covered the greatest portion of gene bases (Figure 4A). In contrast, the "Kundaje Unified" list had the smallest impact on the number and coverage of protein-coding genes, likely reflecting its manual curation, which intentionally avoids regions overlapping transcripts. Similarly, the "Nordin CUT&RUN" list showed a low impact on both the number and coverage of protein-coding genes. Comparable trends were observed for long noncoding RNAs and other transcript types (Figure 4A).

We visualized a representative region of ubiquitin-specific peptidase 17-like family member genes to illustrate the large size of the Blacklist-generated regions, the "Low Mappability" regions that harbor protein-coding

genes, the tendency of "High Signal" regions to overlap genes, and the conservative approach of the "Kundaje Unified" and the "Nordin CUT&RUN" lists, which avoid overlapping genes (Figure 4B, upper panel). These results indicate that the Blacklist-generated list, "High Signal," "Low Mappability," and the combined lists may remove peaks overlapping gene regions.

To determine whether specific gene sets or pathways are affected by exclusion regions, we performed KEGG enrichment analysis (Supplementary Table S5). Pathways related to "Ribosome" and "Ribosome biogenesis in eukaryotes" were enriched in genes affected by the "Nordin CUT&RUN" list, the "High Signal" regions, the combined "HS + LM" regions, and our "Generated Blacklist" (Figure 4C). Genes from these pathways were also affected by the "Low Mappability" and "GitHub Blacklist" regions, though the observed enrichments were not statistically significant. This result was expected, as ribosomal RNA gene (rDNA) repeats are a well-characterized source of contamination [34–36], and ribosomal RNA (rRNA) constitutes approximately 80–90% of total cellular RNA [37]. However, the "GitHub Blacklist," "Kundaje Unified," and "GreyListChIP" lists did not show any significant enrichment.

We visualized a cluster of ribosomal genes affected by the "Nordin CUT&RUN" list (Figure 4B, lower panel), highlighting a key distinction between the "GitHub Blacklist" and the "Generated Blacklist." The latter covered ribosomal genes, whereas the former did not, suggesting that the "GitHub Blacklist" may have been filtered to minimize overlap with gene regions. These findings underscore the unique characteristics of the "Nordin CUT&RUN" list, which, despite affecting fewer genes overall, appears to exhibit exclusive enrichment in ribosomal genes.

We examined the impact of exclusion sets on cancer-associated genes using oncoEnrichR analysis [38]. The analysis identified five oncogenes and tumor suppressors (collectively referred to as cancer drivers) with moderate to very strong evidence of being affected by the "GitHub Blacklist," with KDM5A (lysine demethylase 5A) and MLH1 (mutL homolog 1) being the most notable oncogene and tumor suppressor, respectively (Supplementary Table S6). The "Generated Blacklist" impacted 15 cancer drivers, including the NOTCH2 oncogene and tumor suppressor, with very strong confidence. The "High Signal" list affected 24 cancer drivers, encompassing key genes such as PIK3CA, MTOR, RAF1, and JUN. The "Low Mappability" list impacted six cancer drivers, including NOTCH2 and SSX2 (SSX family member 2). The combined "HS + LM" list affected the highest number of cancer drivers (32), while the "GreyListChIP" list impacted 17, including NOTCH2 and TP53. In contrast, the "Kundaje Unified" and "Nordin CUT&RUN" lists did not overlap with cancer driver genes. These observations suggest that the "High Signal" and "Low Mappability" regions should not be excluded and position the "Kundaje Unified" list as a viable option for analyses aiming to minimize the impact on functional pathways and oncogenes.

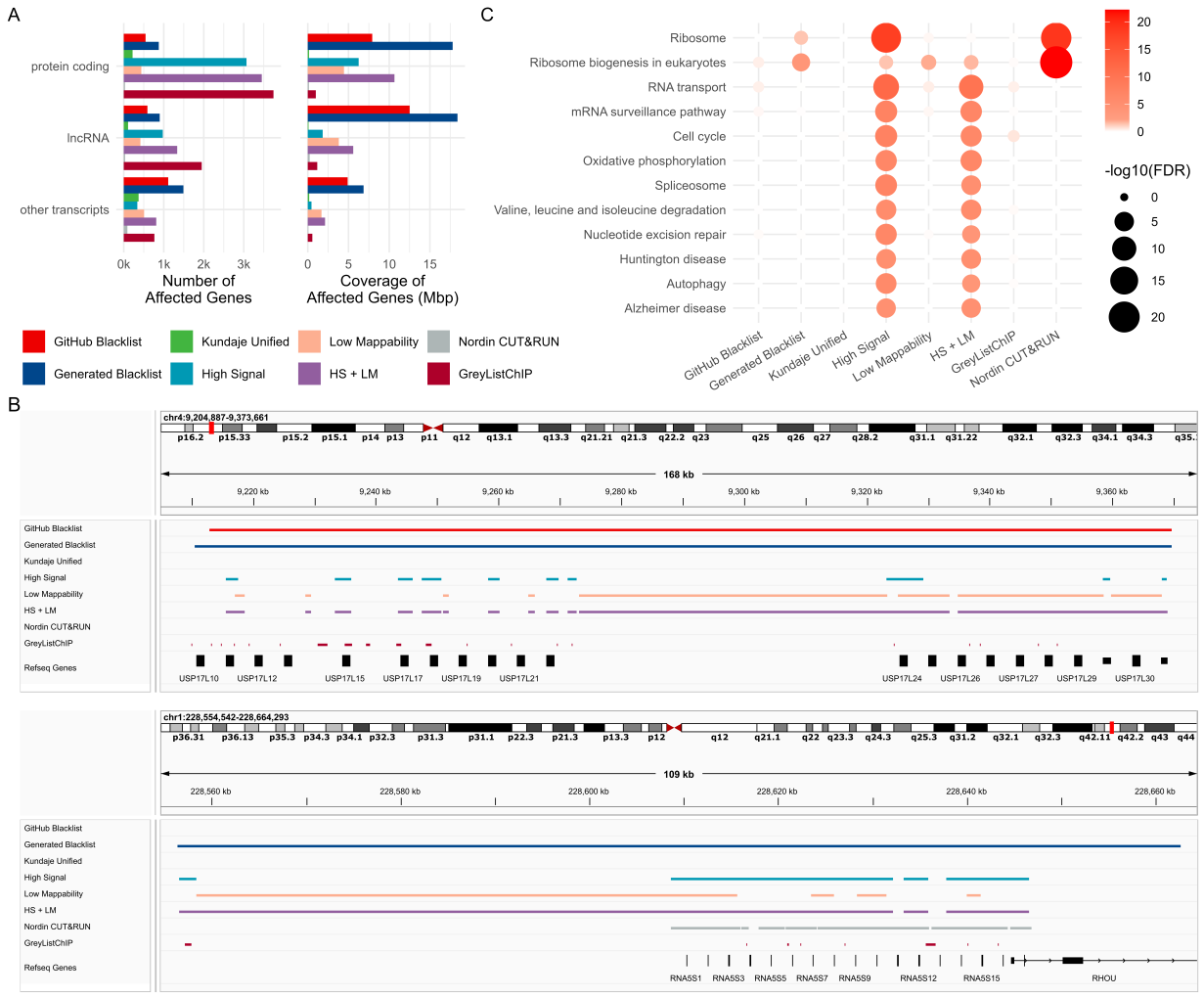

Figure 4. Biological characterization of genes affected by exclusion regions. A) Count and coverage of protein-coding, long noncoding, and other transcripts affected by exclusion sets; B) Representative comparison of exclusion set coverage over a cluster of ubiquitin-specific peptidase 17-like family member genes; C) KEGG pathways enriched in genes overlapped by exclusion sets.

Supplementary Figure S10. Relationships among Blacklist, manually created, and Nordin CUT&RUN and GreyListChIP exclusion sets. A, B) Jaccard count overlap, multi-dimensional scaling, and hierarchical clustering; C, D) Forbes width overlap, MDS, and clustering.

Supplementary Table S5. Genes affected by exclusion sets. Protein-coding genes, sorted alphabetically, are listed first, followed by lncRNAs and other categories.

Supplementary Table S6. Oncogenes and tumor suppressors affected by exclusion sets. "Cancer Association" results from the oncoEnrichR analysis are shown for each list, sorted by oncogenic and tumor suppressor confidence levels.

Using "sponge" sequences decreases artificial correlation in ChIP-seq data comparable to the Blacklist-generated exclusion sets.

The original Blacklist publication suggested that removing reads overlapping exclusion regions would enhance the biological interpretability of ChIP-seq data by reducing signal correlation among ChIP-seq profiles. The authors hypothesized that read pileups in exclusion regions lead to spurious correlations and that

removing them would allow for better biological interpretability. In this study, we evaluated this hypothesis by correlating ChIP-seq transcription factor signals for the GM12878 cell line before and after removing reads overlapping exclusion regions. Notably, removing reads overlapping the "High Signal," "Low Mappability," and "Kundaje Unified" lists resulted in the smallest decrease in correlations (Figure 5A). We confirmed that removing reads overlapping the "Generated Blacklist" and the "GitHub Blacklist" lists resulted in the largest decrease in signal correlation for the REST, SREBF2, and CTCF transcription factors (Figure 5B).

Aligning sequencing reads to a genome that includes so-called "sponge" sequences represents another strategy to reduce artifact signals. These sequences include satellite DNA, ribosomal DNA, mitochondrial sequences, etc., and constitute roughly 8% of the human genome omitted from the reference assembly [1]. We hypothesized that including these "sponge" sequences at the alignment step would reduce read pileups and signal correlation compared to reads aligned to the assembled chromosomes only. We found that this was indeed the case, with the overall "sponge"-aligned signal correlation reduction being second only to the "Generated Blacklist" and "GitHub Blacklist" lists (Figure 5). With minor exceptions, these observations were consistent for other transcription factors (Supplementary Figure S11).

We investigated the alignment characteristics of reads aligned to the "sponge" sequences and found that 48.78% of 36 bp-long reads were aligned to them (Supplementary Figure S12A, Supplementary Table S6). In contrast, only 6.21% of 101 bp-long reads were aligned to the "sponge" sequences, suggesting that the higher information content in longer reads improves alignment to the autosomal reference. We also investigated the number and proportion of reads derived from excludable regions that overlapped the "sponge" sequences and found that the "Nordin CUT&RUN" and "GreyListChIP" lists had the largest proportion of such reads, followed by the "Kundaje Unified" list (80.38%, 74.30%, and 45.66%, respectively; Supplementary Table S6). We further hypothesized that using the "sponge" sequences for RNA-seq data would result in a minimal effect on gene counts due to the high complexity and conservation of gene regions. We found that the total number of reads overlapping gene regions remained nearly unchanged (Supplementary Figure S12B; nonsignificant t-test, Wilcoxon test, and Kolmogorov-Smirnov test p-values). We identified 74 differentially expressed genes (FDR < 0.05) when aligning with "sponge," with 63 being downregulated and 11 upregulated. Among these, 23 were mitochondrial, 18 were lncRNAs, 11 were TEC (To be Experimentally Confirmed), and others included pseudogenes, novel transcripts, and rRNAs (Supplementary Table S7).

We also investigated the effect of "sponge" sequences on Single Nucleotide Polymorphism (SNP) and InDel calls from whole-genome sequencing data. Out of 3,872,204 SNPs and 814,696 InDels called from the hg38-aligned data, 2.52% of SNPs and 1.85% of InDels were lost, while 0.51% of SNPs and 0.52% of InDels were gained when calling variants from the hg38+sponge-aligned data (Supplementary Figure S12C). We hypothesized that most lost or gained variants would be false positives. To test this, we examined the proportion of lost or gained SNPs and InDels that were annotated as benchmark variants by the Genome in a Bottle (GIAB) consortium. Only a small fraction of lost and gained SNPs (3.08% and 1.52%, respectively) were annotated as benchmark GIAB SNPs (out of 3,357,884 total). Similarly, only a small fraction of lost and gained InDels (9.50% and 3.15%, respectively) were annotated, suggesting that the vast majority of lost or gained variants are likely false positives. Together, these results suggest that using "sponge" sequences may represent an alignment-guided approach to removing peak pileups and improving biological signal and that such an approach may be universally applicable to technologies other than chromatin profiling.

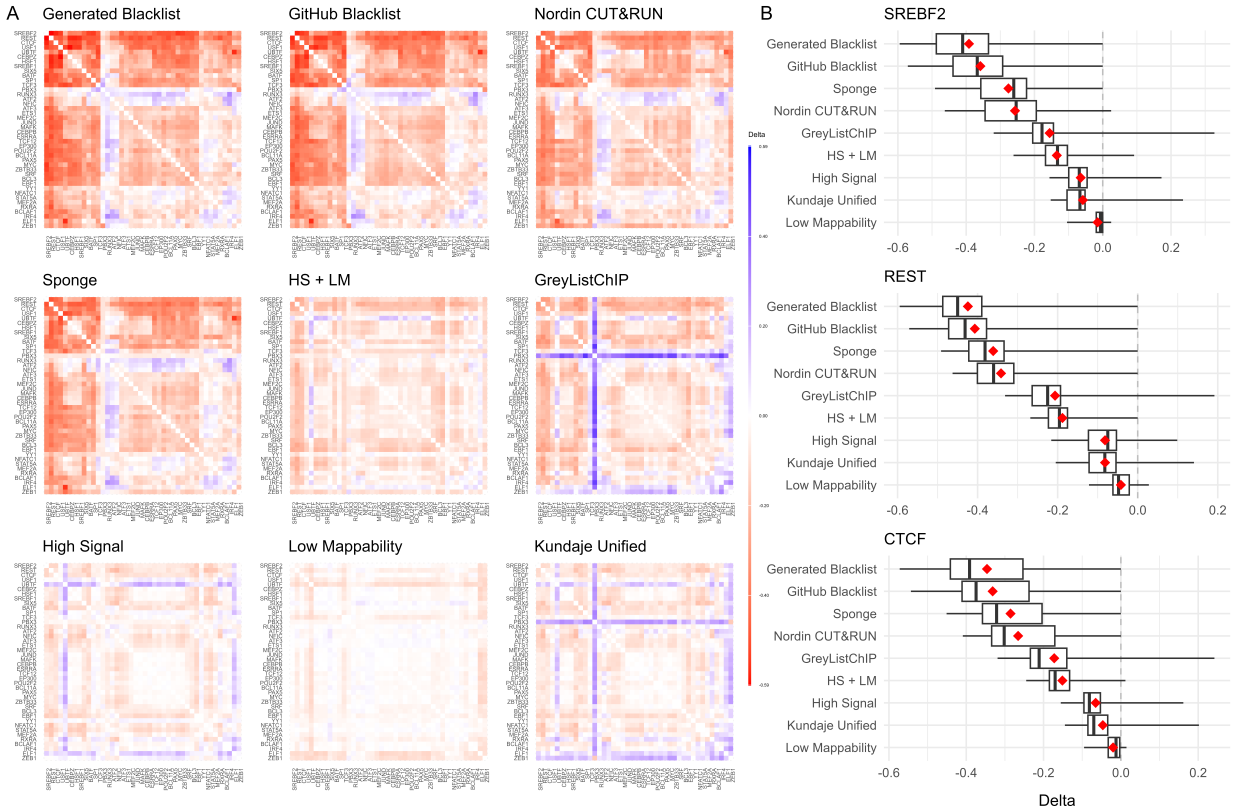

Figure 5. Changes in ChIP-seq signal correlation with and without reads overlapping exclusion sets or aligned to the sponge. Data for the Gm12878 cell line is shown. A) Heatmaps of correlation differences for each exclusion set, sorted by mean (red/blue gradient corresponds to decreases/increases in correlations, respectively); B) Correlation difference distributions for the top three most affected transcription factors.

Supplementary Figure S11. Transcription factor-specific changes in ChIP-seq signal correlation with and without reads overlapping exclusion sets or aligned to the sponge. Data for the Gm12878 cell line is shown. Correlation difference distributions, red transcription factor names highlights cases where exclusion lists other than "Blacklist" and "Sponge" have the strongest decrease in correlations.

Supplementary Figure S12. The effect of "sponge" sequences on alignment. A) The number and proportion of 36 bp and 101 bp reads aligned to "sponge" sequences from the whole genome and from various excludable sets; B) The total number of reads overlapping annotated gene regions when aligned with and without the "sponge" sequences.

Supplementary Table S6. Alignment statistics of ChIP-seq input sequencing data to the "sponge" sequences and other contigs comprising the human genome reference.

Supplementary Table S7. Differentially expressed genes between samples aligned to the hg38 autosomal reference and to the hg38 reference with the "sponge" sequences.

Discussion

The original publication describing the Blacklist software provides only limited methodological details, which hinders the reproducibility of its results. For instance, key algorithmic parameters such as k-mer size, "binSize," "binOverlap," and the single- versus multi-read configuration of Umap mappability files are not described or benchmarked. Additionally, the correlation analysis demonstrating reduced artificial correlations

among ChIP-seq data lacks transparency regarding data sources and correlation methods, prompting us to implement our own version of this analysis. Moreover, the data properties influencing exclusion set generation remain unexplored, leaving room for variability across different datasets and software configurations. In this study, we addressed these gaps by reverse-engineering the Blacklist software, fixing an annotation error, providing corrected C++ and Python code, and systematically investigating the properties of various exclusion sets and algorithmic choices. Despite our efforts, we were unable to fully reproduce the exclusion sets provided in the Blacklist GitHub repository.

While we thoroughly benchmarked the properties of exclusion sets, the Blacklist software contains numerous hidden parameters that may influence results. Comprehensive evaluation of these parameters would require advanced parameter optimization methods, such as grid search algorithms [39]. Data preprocessing steps, such as adapter trimming, also warrant investigation. Although we assumed the BAM-derived FASTQ files were pre-trimmed, testing raw FASTQ files could reveal the impact of trimming software. Similarly, peak caller choice is critical in defining high-signal regions. We used MACS3 [20], the latest iteration of MACS2, but alternatives such as Genrich and deep learning-based tools (e.g., LanceOtron [40]) could be evaluated to determine their influence on results.

Exhaustive testing of exclusion set generation tools is beyond the scope of this work. Greenscreen software, despite being Docker-wrapped, was technically challenging to run and featured numerous parameters affecting its pipeline. The algorithm underlying the GreyListChIP R package is unpublished, with internal parameters comparable to those in Blacklist software. While GreyListChIP performed well in reducing artificial correlation among ChIP-seq peaks, we observed some unexplained results, such as missing calls on specific chromosomes. Benchmarking GreyListChIP remains an area for future exploration.

We recommend caution when selecting exclusion sets. While our "High Signal" and "Low Mappability" regions are potential candidates for exclusion, they differ significantly from Blacklist-generated lists and overlap many protein-coding genes and cancer drivers. The "Nordin CUT&RUN" and GreyListChIP-generated lists appear effective at reducing artificial correlations but also overlap cancer driver genes. We provide GreyListChIP-generated exclusion sets for hg38, T2T, mm10, and mm39, together with other exclusion sets generated in this work, in our excluderanges R package [12] but advise applying GreyListChIP to experiment-specific input data.

Using "sponge" sequences at the alignment step appears to be a viable alternative to exclusion sets. We demonstrate that including "sponge" sequences at the alignment step reduces artifact signals comparable to Blacklist-generated lists. The fact that some reads overlapping excludable regions remain aligned even in the presence of "sponge" sequences (Supplementary Figure S12A) suggests that removing all reads from excludable regions may lead to loss of biologically relevant signal. "Sponge" sequences avoid the problem of poor reproducibility of exclusion sets and potential loss of biological signal by targeting artifact reads at the alignment step. We observed their effectiveness in RNA-seq alignment (negligible effect on gene expression) and in whole genome sequencing settings (minimal loss of genomic variants that are likely false positives), suggesting that including "sponge" sequences as part of genome references may improve biological signal from genomic data generated by any short-read technology.

Besides using "sponge" sequences, we encourage researchers to prioritize data processing steps, including QC, adapter trimming, aligner selection, and peak calling. To alleviate artifact signal in ribosomal gene repeats, wet-lab and computational ribosomal RNA depletion methods have been recommended [36,37]. Hard-masking well-defined mitochondrial DNA sequences [41] and gap elements (e.g., telomeres, centromeres) is also advisable. Testing the effect of these steps in conjunction with including "sponge" sequences in genome references is one of the goals of our future work.

Although highly promising, using "sponge" sequences remains a method that is less broadly applicable. This is due to limited understanding of sequences that constitute "sponge" or "decoy" sequences. While these sequences have been assembled for the hg38 human genome [1], they haven't been created for genome assemblies of other model organisms. Moreover, with the development of the telomere-to-telomere versions of human [16] and mouse [42] genomes these "sponge" sequences may not be necessary. The development of long-read sequencing (Oxford Nanopore Technologies, Pacific Biosciences) may also alleviate spurious alignments, as we have shown when comparing the 36bp and 101bp alignments (Supplementary Figure S12A).

Furthermore, the growing amount of whole genome data and the development of pangenome graph assembly methods has been shown to improve genomic variant discovery, RNA-seq and chromatin immunoprecipitation read mapping [43]. It represents a promising way to alleviate alignment artifacts; however, methods for using graph assembly are still relatively less widely adopted than those using linear genome assemblies. Our future work includes defining "sponge" sequences for other genomes and model organisms as well as exploring the use of the pangenome graph assembly for improving biological signal and eliminate the need for exclusion sets.

Methods

Data sources

The original Blacklist exclusion lists for human, mouse, and other species were obtained from the Blacklist GitHub repository (https://github.com/Boyle-Lab/Blacklist/tree/master/lists), using version 2 of the lists. The Kundaje Unified list was retrieved from the ENCODE Project (accession number ENCFF356LFX). The Nordin CUT&RUN exclusion set was originally obtained from Additional File 2 of the corresponding publication [33] and sourced from the excluderanges R package [12]. Unless otherwise indicated, we used lists based on the hg38 human genome assembly.

To replicate the Blacklist exclusion lists, accession numbers of the corresponding BAM files were obtained using metadata provided in the Blacklist GitHub repository (https://github.com/Boyle-Lab/Blacklist/tree/master/lists/metadata) and downloaded from ENCODE. The "Umap" mappability files were obtained from the Hoffman Lab Umap/Bismap Project page (https://hoffmanlab.org/proj/bismap/). Accession numbers of the FASTQ files used to generate these BAM files were identified via the [**PG?**] tag in the header lines of each BAM file and downloaded from ENCODE.

The Gm12878 transcription factor FASTQ files used for correlation analysis were retrieved from ENCODE using the following metadata link: (https://www.encodeproject.org/metadata/?type=Experiment&cart=%2Fcarts%2F4b50a1ed-d002-4c76-8401-5ac42d3f2228%2F&files.output_category=raw+data).

The whole genome sequencing data for the NA12878 sample were downloaded from https://www.internationalgenome.org/data-portal/sample/NA12878. The GIAB benchmark SNP/InDel dataset was downloases from https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/NA12878_HG001/latest/GRCh38/. Known variant sites for base quality score recalibration (BQSR) were obtained from dbSNP build 138 (https://storage.cloud.google.com/genomics-public-data/resources/broad/hg38/v0/Homo_sapiens_assembly38.dbsnp138.vcf), and indels from GATK (https://storage.cloud.google.com/genomics-public-data/resources/broad/hg38/v0/Homo_sapiens_assembly38.known_indels.vcf.gz). All data were accessed on February 1, 2025.

Generated exclusion sets

Blacklist exclusion sets for the hg38 and mm10 genome assemblies were generated using the BAM files listed in the metadata on GitHub, relevant Umap mappability files, and the original Blacklist software.

To define High Signal regions for the hg38 genome assembly, we generated a signal track by merging 274 paired-end 100/101 bp BAM files (aligned with bwa 0.7.10 sampe) via samtools merge (v1.2). We called local peaks using: `macs3 callpeak --treatment ${in_file} --name 101_local --outdir ${out_dir} --gsize hs --slocal 10000 --llocal 100000 --keep-dup all` (v3.0.1) and selected those with a fold change greater than 99.0% of the fold change range. Regions within 1,000 bp were merged, and regions smaller than 1,000 bp were discarded. This strict filtering ensured the selection of highly significant and long high-signal regions. After generating our High Signal regions, we merged these regions with centromeres from the UCSC Genome Browser, available via the excluderanges R package [12].

To define Low Mappability regions for the hg38 genome assembly, we downloaded the kmer-100 multi-read mappability bedGraph file from the Hoffman Lab website (https://bismap.hoffmanlab.org), which contains base-level mappability values ranging from 0 to 1 (with 1 being highly mappable). We first defined the mappable universe by selecting regions with mappability greater than 0.01. Using bedtools (v2.31.1), regions within 1,000 bp were merged. Inverting the mappable universe resulted in a set of regions with mappability ≤ 0.01. These regions were then merged within 1,000 bp, and regions smaller than 1,000 bp were discarded. This strategy preserved mappable regions interspersed with short unmappable segments while conservatively selecting long low-mappability regions.

The GreyListChIP R package generates a BAM file-specific exclusion set. We generated GreyListChIP lists for humans by combining 381 36 bp single-end FASTQ files used in the original Blacklist publication, realigning them with STAR to the hg38 genome assembly, and setting the maxGap parameter to 1,000. This list was used in our benchmarks. Additionally, we combined 274 paired-end 100 bp and 101 bp FASTQ files and generated lists for the hg38 and T2T genome assemblies. Similarly, for mice, we used 106 36 bp single-end FASTQ files and 176 50 bp single-end FASTQ files to generate lists for the mm10 and mm39 genome assemblies. All GreyListChIP lists have been added to our excluderanges R package [12].

Comparison of exclusion sets

We compare two exclusion sets $A = \{a_1, a_2, \ldots, a_i\}$ and $B = \{b_1, b_2, \ldots, b_j\}$ using the Jaccard count overlap, which is defined as the ratio between the average number of regions in A and B that overlap with any region in the other set, and the total number of unique regions in the union of A and B. In mathematical terms, this is given by

$$J_c(A, B) = \frac{\frac{1}{2}\Big(|\{a \in A : a \cap B \neq \emptyset\}| + |\{b \in B : b \cap A \neq \emptyset\}|\Big)}{|A \cup B|}.$$

Similarly, Forbes width overlap formula:

$$F_w(A, B) = \frac{\text{Genome Size} \cdot W(A \cap B)}{W(A) \cdot W(B)}$$

Where $W(A)$ and $W(B)$ are the total widths in bases of sets A and B, respectively, $W(A \cap B)$ is the total width of the intersection of A and B, and the "Genome Size" is the total reference genome size (e.g., 2,875,001,522 base pairs for hg38).

Similarity among exclusion sets was visualized using Classical Multidimensional Scaling (MDS) by converting the Jaccard count overlap and Forbes width overlap matrices into distance matrices and plotting exclusion sets using the first two principal coordinates. These distance matrices were also used for hierarchical clustering with the Ward clustering method. All other comparisons (e.g., number and width of overlapping regions) were performed using the GenomicRanges (v1.56.1) R package [44], R (v4.4.0), and Bioconductor (v3.19) [45].

Aligner test

The original hg38 BAM files were aligned using version 0.7.10 of the bwa sampe (paired-end) and bwa samse (single-end) aligners [46]. In our comparison of exclusion sets generated from BAMs aligned with different aligners, we used the following: bwa-mem2 2.2.1 [46]: `bwa-mem2 mem references/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz <R1.fastq.gz> <R2.fastq.gz> | samtools view -bS - | samtools sort -o <bwamem.bam> -`; bowtie2 2.5.4 [47]: `bowtie2 -x references/GCA_000001405.15_GRCh38_no_alt_analysis_set -U <R1.fastq.gz> <R2.fastq.gz> --local | samtools view -bS - | samtools sort -o <bowtie2.bam> -`; STAR 2.7.11b [48]:

`STAR --genomeDir <references/star> --readFilesIn <(bgzip -cd <R1.fastq.gz>) <(bgzip -cd <R2.fastq.gz>) --outSAMtype BAM SortedByCoordinate --alignEndsType Local`, as well as htslib and samtools 1.20 [49].

Correlation analysis

FASTQ files for each transcription factor were merged, and the resulting files were aligned using STAR 2.7.11b (–alignEndsType Local) to GRCh38_no_alt_analysis_set_GCA_000001405.15.fasta (ENCODE ID: ENCSR425FOI) and to the same reference concatenated with "sponge" sequences [1].

Read count bins for each exclusion set or sponge were computed using deepTools v3.5.6 [31] with the multiBamSummary bins command (default settings). The Pearson correlation matrix was then obtained using plotCorrelation (–corMethod pearson), and each matrix was clustered and visualized in R using Euclidean distance and complete linkage.

RNA-seq analysis

We realigned our RNA-seq data [50,51] (GSE235167) to two versions of the hg38 human genome assembly from the UCSC Genome Browser. One version included only autosomal, sex, and mitochondrial chromosomes, while the other also contained additional random and unplaced contigs, as well as "sponge" sequences [1]. The data were processed using the Nextflow rnaseq v3.18.0 pipeline [52] with the STAR-RSEM quantification strategy.

WGS analysis

Whole-genome sequencing (WGS) data were processed using the nf-core/sarek v3.5.0 pipeline with the BWA-MEM2 setting. Reads were aligned to the hg38 human genome assembly (ENCODE ID: ENCSR425FOI) and, separately, to the hg38 genome assembly with "sponge" sequences concatenated. Variant calling was performed using GATK HaplotypeCaller within the nf-core pipeline. Variants (including the GIAB benchmark set) were normalized using BCFtools v1.19, removing duplicates and standardizing variant representation. Intersections were taken with bcftools isec, and variants were counted with bcftools view -v snps "$vcf" -H | wc -l.

Functional enrichment analysis

Hypergeometric enrichment analysis of transcripts with a non-zero sum of exon widths covered by exclusion regions was performed using the enrichR 3.2 R package [53] and the "KEGG_2019_Human" signature database. Analysis of cancer drivers affected by exclusion sets was conducted using the oncoEnrichR 1.5.2 R package [38], considering the "Cancer associations" analysis.

excluderanges R package update

We added the GreyListChIP-generated sets to our excluderanges R package [12], making them available within the R ecosystem. We used STAR-aligned sequencing data with 36 bp and 101 bp read lengths for the human genome and 36 bp and 50 bp read lengths for the mouse genome, applying a maxGap = 1,000 merge setting. Specifically, lists for the hg38 and T2T human genome assemblies and the mm10 and mm39 mouse genome assemblies were added.

Code availability

The modified Blacklist code and scripts to reproduce our analyses and figures are available at https://github.com/dozmorovlab/excluderanges_supplementary

Funding

This work was supported in part by the George and Lavinia Blick Research Scholarship to MD, the NIH/NCI (R01CA246182, R21CA273779, U54CA283762) grants to JCH, and the NIH R56 MH107879 JLM.

References

1. Miga KH, Eisenhart C, Kent WJ: **Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments**. *Nucleic Acids Res* 2015, **43**:e13310.1093/nar/gkv671.

2. Tanaka N, Takahara A, Hagio T, Nishiko R, Kanayama J, Gotoh O, Mori S: **Sequencing artifacts derived from a library preparation method using enzymatic fragmentation**. *PLoS One* 2020, **15**:e022742710.1371/journal.pone.0227427.

3. Haile S, Corbett RD, Bilobram S, Bye MH, Kirk H, Pandoh P, Trinh E, MacLeod T, McDonald H, Bala M, Miller D, Novik K, Coope RJ, Moore RA, Zhao Y, Mungall AJ, Ma Y, Holt RA, Jones SJ, Marra MA: **Sources of erroneous sequences and artifact chimeric reads in next generation sequencing of genomic DNA from formalin-fixed paraffin-embedded samples**. *Nucleic Acids Res* 2019, **47**:e1210.1093/nar/gky1142.

4. Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, Gardner K, Hillier LW, Janette J, Jiang L, Kasper D, Kawli T, Kheradpour P, Kundaje A, Li JJ, Ma L, Niu W, Rehm EJ, Rozowsky J, Slattery M, Spokony R, Terrell R, Vafeados D, Wang D, Weisdepp P, Wu Y-C, Xie D, Yan K-K, Feingold EA, Good PJ, et al.: **Comparative analysis of regulatory information and circuits across distant species**. *Nature* 2014, **512**:453–610.1038/nature13668.

5. Wreczycka K, Franke V, Uyar B, Wurmus R, Bulut S, Tursun B, Akalin A: **HOT or not: Examining the basis of high-occupancy target regions**. *Nucleic Acids Res* 2019, **47**:5735–574510.1093/nar/gkz460.

6. Kharchenko PV, Tolstorukov MY, Park PJ: **Design and analysis of ChIP-seq experiments for DNA-binding proteins**. *Nat Biotechnol* 2008, **26**:1351–910.1038/nbt.1508.

7. Xie D, Boyle AP, Wu L, Zhai J, Kawli T, Snyder M: **Dynamic trans-acting factor colocalization in human cells**. *Cell* 2013, **155**:713–2410.1016/j.cell.2013.09.043.

8. Carroll TS, Liang Z, Salama R, Stark R, Santiago I de: **Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data**. *Front Genet* 2014, **5**:7510.3389/fgene.2014.00075.

9. Nordin A, Zambanini G, Pagella P, Cantu C: **The CUT&RUN blacklist of problematic regions of the genome**. *bioRxiv* 2022,.

10. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, Hardison RC, Dunham I, Kellis M, Noble WS: **Integrative annotation of chromatin elements from ENCODE data**. *Nucleic Acids Res* 2013, **41**:827–4110.1093/nar/gks1284.

11. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, Madrigal P, Taslim C, Zhang J: **Practical guidelines for the comprehensive analysis of ChIP-seq data**. *PLoS Comput Biol* 2013, **9**:e100332610.1371/journal.pcbi.1003326.

12. Ogata JD, Mu W, Davis ES, Xue B, Harrell JC, Sheffield NC, Phanstiel DH, Love MI, Dozmorov MG: **Excluderanges: Exclusion sets for T2T-CHM13, GRCm39, and other genome assemblies**. *Bioinformatics* 2023, **39**10.1093/bioinformatics/btad198.

13. Amemiya HM, Kundaje A, Boyle AP: **The ENCODE blacklist: Identification of problematic regions of the genome**. *Sci Rep* 2019, **9**:935410.1038/s41598-019-45839-z.

14. Wimberley CE, Heber S: **PeakPass: Automating ChIP-seq blacklist creation**. *J Comput Biol* 2019, 10.1089/cmb.2019.0295.

15. Klasfeld S, Wagner D: **Greenscreen decreases type I errors and increases true peak detection in genomic datasets including ChIP-seq**. *bioRxiv* 2022,.

16. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, Aganezov S, Hoyt SJ, Diekhans M, Logsdon GA, Alonge M, Antonarakis SE, Borchers M, Bouffard GG, Brooks SY, Caldas GV, Chen N-C, Cheng H, Chin C-S, Chow W, Lima LG de, Dishuck PC, Durbin R, Dvorkina T, Fiddes IT, Formenti G, et al.: **The complete sequence of a human genome**. *Science* 2022, **376**:44–5310.1126/science.abj6987.

17. Scheinin I, Sie D, Bengtsson H, Wiel MA van de, Olshen AB, Thuijl HF van, Essen HF van, Eijk PP, Rustenburg F, Meijer GA, Reijneveld JC, Wesseling P, Pinkel D, Albertson DG, Ylstra B: **DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly**. *Genome Res* 2014, **24**:2022–3210.1101/gr.175141.114.

18. Karimzadeh M, Ernst C, Kundaje A, Hoffman MM: **Umap and bismap: Quantifying genome and methylome mappability**. *Nucleic Acids Res* 2018, **46**:e12010.1093/nar/gky677.

19. Klasfeld S, Roulé T, Wagner D: **Greenscreen: A simple method to remove artifactual signals and enrich for true peaks in genomic datasets including ChIP-seq data**. *Plant Cell* 2022, **34**:4795–481510.1093/plcell/koac282.

20. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS: **Model-based analysis of ChIP-seq (MACS)**. *Genome Biol* 2008, **9**:R13710.1186/gb-2008-9-9-r137.

21. Brown G: **GreyListChIP: Grey lists–mask artefact regions based on ChIP inputs**. *R package version* 2015, **1**.

22. ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, **489**:57–7410.1038/nature11247.

23. Yan F, Powell DR, Curtis DJ, Wong NC: **From reads to insight: A hitchhiker's guide to ATAC-seq data analysis**. *Genome Biol* 2020, **21**:2210.1186/s13059-020-1929-3.

24. Yang H, Luan Y, Liu T, Lee HJ, Fang L, Wang Y, Wang X, Zhang B, Jin Q, Ang KC, Xing X, Wang J, Xu J, Song F, Sriranga I, Khunsriraksakul C, Salameh T, Li D, Choudhary MNK, Topczewski J, Wang K, Gerhard GS, Hardison RC, Wang T, Cheng KC, Yue F: **A map of cis-regulatory elements and 3D genome structures in zebrafish**. *Nature* 2020, **588**:337–34310.1038/s41586-020-2962-9.

25. Huang N, Seow WQ, Appert A, Dong Y, Stempor P, Ahringer J: **Accessible region conformation capture (ARC-c) gives high-resolution insights into genome architecture and regulation**. *Genome Res* 2022, **32**:357–36610.1101/gr.275669.121.

26. Chen H, Lareau C, Andreani T, Vinyard ME, Garcia SP, Clement K, Andrade-Navarro MA, Buenrostro JD, Pinello L: **Assessment of computational methods for the analysis of single-cell ATAC-seq data**. *Genome Biol* 2019, **20**:24110.1186/s13059-019-1854-5.

27. Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, Lücken MD, Strobl DC, Henao J, Curion F, Single-cell Best Practices Consortium, Schiller HB, Theis FJ: **Best practices for single-cell analysis across modalities**. *Nat Rev Genet* 2023, **24**:550–57210.1038/s41576-023-00586-w.

28. Vitak SA, Torkenczy KA, Rosenkrantz JL, Fields AJ, Christiansen L, Wong MH, Carbone L, Steemers FJ, Adey A: **Sequencing thousands of single-cell genomes with combinatorial indexing**. *Nat Methods* 2017, **14**:302–30810.1038/nmeth.4154.

29. Crowley C, Yang Y, Qiu Y, Hu B, Abnousi A, Lipiński J, Plewczyński D, Wu D, Won H, Ren B, Hu M, Li Y: **FIREcaller: Detecting frequently interacting regions from hi-c data**. *Comput Struct Biotechnol J* 2021, **19**:355–36210.1016/j.csbj.2020.12.026.

30. Stansfield J, Dozmorov MG: **HiCcompare: A method for joint normalization of hi-c datasets and differential chromatin interaction detection**. *bioRxiv* 2017, 10.1101/147850Available: https://www.biorxiv.org/content/early/2017/08/28/147850.

31. Ramirez F, Dündar F, Diehl S, Grüning BA, Manke T: **deepTools: A flexible platform for exploring deep-sequencing data**. *Nucleic Acids Res* 2014, **42**:W187–9110.1093/nar/gku365.

32. Salvatore S, Dagestad Rand K, Grytten I, Ferkingstad E, Domanska D, Holden L, Gheorghe M, Mathelier A, Glad I, Kjetil Sandve G: **Beware the jaccard: The choice of similarity measure is important and non-trivial in genomic colocalisation analysis**. *Brief Bioinform* 2020, **21**:1523–153010.1093/bib/bbz083.

33. Nordin A, Zambanini G, Pagella P, Cantù C: **The CUT&RUN suspect list of problematic regions of the genome**. *Genome Biol* 2023, **24**:18510.1186/s13059-023-03027-3.

34. Warmerdam DO, Wolthuis RMF: **Keeping ribosomal DNA intact: A repeating challenge**. *Chromosome Res* 2019, **27**:57–7210.1007/s10577-018-9594-z.

35. Kobayashi T: **Ribosomal RNA gene repeats, their stability and cellular senescence**. *Proc Jpn Acad Ser B Phys Biol Sci* 2014, **90**:119–2910.2183/pjab.90.119.

36. Kopylova E, Noé L, Touzet H: **SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data**. *Bioinformatics* 2012, **28**:3211–710.1093/bioinformatics/bts611.

37. Chen L, Yang R, Kwan T, Tang C, Watt S, Zhang Y, Bourque G, Ge B, Downes K, Frontini M, Ouwehand WH, Lin J-W, Soranzo N, Pastinen T, Chen L: **Paired rRNA-depleted and polyA-selected RNA sequencing data and supporting multi-omics data from human t cells**. *Sci Data* 2020, **7**:37610.1038/s41597-020-00719-4.

38. Nakken S, Gundersen S, Bernal FLM, Polychronopoulos D, Hovig E, Wesche J: **Comprehensive interrogation of gene lists from genome-scale cancer screens with oncoEnrichR**. *Int J Cancer* 2023, **153**:1819–182810.1002/ijc.34666.

39. Bergstra J, Bardenet R, Bengio Y, Kégl B: **Algorithms for hyper-parameter optimization**. *Ad-*

vances in neural information processing systems 2011, **24**.

40. Hentges LD, Sergeant MJ, Cole CB, Downes DJ, Hughes JR, Taylor S: **LanceOtron: A deep learning peak caller for genome sequencing experiments**. *Bioinformatics* 2022, **38**:4255–426310.1093/bioinformatics/btac525.

41. Calabrese FM, Simone D, Attimonelli M: **Primates and mouse NumtS in the UCSC genome browser**. *BMC Bioinformatics* 2012, **13 Suppl 4**:S1510.1186/1471-2105-13-S4-S15.

42. Liu J, Li Q, Hu Y, Yu Y, Zheng K, Li D, Qin L, Yu X: **The complete telomere-to-telomere sequence of a mouse genome**. *Science* 2024, **386**:1141–114610.1126/science.adq8191.

43. Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, Buonaiuto S, Chang XH, Cheng H, Chu J, Colonna V, Eizenga JM, Feng X, Fischer C, Fulton RS, Garg S, Groza C, Guarracino A, Harvey WT, Heumos S, Howe K, Jain M, Lu T-Y, Markello C, Martin FJ, Mitchell MW, et al.: **A draft human pangenome reference**. *Nature* 2023, **617**:312–32410.1038/s41586-023-05896-x.

44. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ: **Software for computing and annotating genomic ranges**. *PLoS Comput Biol* 2013, **9**:e100311810.1371/journal.pcbi.1003118.

45. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J: **Bioconductor: Open software development for computational biology and bioinformatics**. *Genome Biol* 2004, **5**:R8010.1186/gb-2004-5-10-r80.

46. Li H, Durbin R: **Fast and accurate short read alignment with burrows-wheeler transform**. *Bioinformatics* 2009, **25**:1754–6010.1093/bioinformatics/btp324.

47. Langmead B, Salzberg SL: **Fast gapped-read alignment with bowtie 2**. *Nat Methods* 2012, **9**:357–910.1038/nmeth.1923.

48. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: Ultrafast universal RNA-seq aligner**. *Bioinformatics* 2013, **29**:15–2110.1093/bioinformatics/bts635.

49. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H: **Twelve years of SAMtools and BCFtools**. *Gigascience* 2021, **10**10.1093/gigascience/giab008.

50. Boyd DC, Zboril EK, Olex AL, Leftwich TJ, Hairr NS, Byers HA, Valentine AD, Altman JE, Alzubi MA, Grible JM, Turner SA, Ferreira-Gonzalez A, Dozmorov MG, Harrell JC: **Discovering synergistic compounds with BYL-719 in PI3K overactivated basal-like PDXs**. *Cancers (Basel)* 2023, **15**10.3390/cancers15051582.

51. Zboril EK, Grible JM, Boyd DC, Hairr NS, Leftwich TJ, Esquivel MF, Duong AK, Turner SA, Ferreira-Gonzalez A, Olex AL, Sartorius CA, Dozmorov MG, Harrell JC: **Stratification of tamoxifen synergistic combinations for the treatment of ER+ breast cancer**. *Cancers (Basel)* 2023, **15**10.3390/cancers15123179.

52. Ewels PA, Peltzer A, Fillinger S, Alneberg J, Patel H, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S: **Nf-core: Community curated bioinformatics pipelines**. *bioRxiv* 2019, :61074110.1101/610741Available: http://biorxiv.org/content/early/2019/05/26/610741.abstract.

53. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A: **Enrichr: A comprehensive gene set enrichment analysis web server 2016 update**. *Nucleic Acids Res* 2016, **44**:W90–710.1093/nar/gkw377.