Title

Using the DNA language model, GROVER, to parse sequence and epigenetic effects on genome stability

Authors

Pierre M. Joubert[1,2,3] and Anna R. Poetsch[1,4*]

Affiliations

1. Biomedical Genomics, Biotechnology Center, Center for Molecular and Cellular Bioengineering, Technische Universität, Dresden, Germany
2. Center for Advanced Systems Understanding (CASUS), Görlitz, Germany
3. Helmholtz Zentrum Dresden-Rossendorf (HZDR), Dresden, Germany
4. National Center for Tumor Diseases (NCT) partner site Dresden, German Cancer Research Center (DKFZ), Dresden, Germany

*Correspondence to arpoetsch@gmail.com

Abstract

Genome stability is shaped by DNA sequence and epigenomic context, but their relative contributions to double-strand break (DSB) sensitivity remain unclear. We show that the DNA language model, GROVER, can predict DSBs using sequence features like short DNA sequences (tokens), which we linked to DSB modulation. Another model using epigenomic features outperforms the sequence-only model, highlighting complementary and cell-type specific information in the epigenome. Integrating sequence and epigenomic data yields the best performance, demonstrating their synergy. Analyzing this model revealed that genome stability information encoded in chromatin marks H3K36me3 and DNase-seq can be learned from the sequence, but not H3K27ac or H3K9me3. Embedding histone data directly into the GROVER architecture enabled cell-type specific modeling with performance matching the full epigenome model. Our results suggest that much of the information shaping DSB patterns is already encoded in the DNA sequence itself, with the epigenome acting as a fine-tuning layer.

Introduction

Organisms must protect their genomes from stress to function properly. Genome instability can lead to disease but also serves as a driver of evolution. Although it is well established that epigenomic features and DNA sequence context influence genome stability, the extent and the precise mechanisms remain unclear. One way to study genome stability is by mapping the locations of double-strand breaks (DSBs). Previous studies have established that DSBs are not randomly distributed across the genome but instead show enrichment or depletion in specific epigenomic and sequence contexts. For example, DSBs are frequently associated with genes, and especially promoters, which is attributed to topoisomerase activity[1] and R-loop formation[2,3], which often occurs at G-quadruplexes[4–6]. This association, along with the co-localization of CpG islands and promoters[7,8] and the presence of high levels of DSBs at

telomeric repeats[9], results in a correlation between DSBs and GC content. Similarly, Alu elements tend to be found in GC-rich regions[8] and are known to be a common repair substrate for homology-directed repair[10,11]. Other repeats, such as L1 elements[12], microsatellites[13,14], and palindromic AT-rich repeats[15] are also known as hotspots for DSBs. Common fragile sites, which are often found in long, AT-rich genes, are also thought to be enriched in DSBs[16]. Finally, CRISPR-induced DSBs have been shown to exhibit sequence-dependent repair[17–19]. Taken together, these studies indicate a strong sequence-specificity in DSB locations and repair. However, it is unclear to which extent the sequence alone is sufficient to understand and model the DSB landscape and which additional information from the epigenome and tissue identity is necessary.

Finding patterns in the epigenome that define where DSBs occur is complicated by the fact that sequence elements, like genes and repeats, have strong trends in which chromatin marks they exhibit. For example, active transcription, which occurs within genes, is associated with DSBs[20]. Active transcription is also associated with histone marks that themselves correlate with DSBs. These marks include the active enhancer and promoter mark H3K27ac[21–24] and the active promoter mark H3K4me3[21,24,25]. Similarly, H3K36me3, which marks actively transcribed gene bodies, plays a significant role in coordinating DSB repair[26]. H3K9me3[27] and H3K27me3[28–30], which mark constitutive and facultative heterochromatin, respectively, both play a role in DSB repair[27,31,32]. Additionally, CCCTC-binding factor (CTCF) is of particular interest, as it plays a direct role in DSB repair[33,34] and chromosome loop anchors bound by CTCF are prone to DSBs[35]. Replication timing plays a crucial role in DSB generation as well, both at late-replicating common fragile sites[36,37] and in early-replicating Poly(dA:dT) sites[38,39] and promoters[1–3]. Furthermore, recently developed DSB-sequencing technologies such as DSBCapture[40], BLISS[41], and its variant sBLISS[42] have enabled the rapid and straightforward genome-wide quantification of DSBs *in situ*. These methods provide a snapshot of DSB locations, which are shaped by repair pathway decisions and differences in repair speed. Using these methods, researchers have been able to characterize the overall, genome-wide association of DSBs with chromatin accessibility[40,43,44], CTCF-binding[40,43,44] and early replication[43].

These epigenomic features are directly affected by the genomic sequence, which makes it difficult to disentangle them. CRISPR-based experiments using the same sequence inserted into different epigenomic contexts have shown that chromatin profiles influence the pathway decisions of DSB repair in a sequence-agnostic way[45–47]. However, understanding how DSB sensitivity changes in different epigenomic contexts is a difficult task given the complexity of the underlying sequence. Therefore, powerful tools that can capture the non-linear interactions between features within these sequences are required.

Recent advancements in deep learning have enabled the development of DNA language models, which have the potential to interpret these complex interactions within the sequence. These models also have the potential to be highly interpretable thanks to their self-attention mechanism[48]. In practice, however, interpreting these models in a biomedical context is difficult[49–51], and special care must be given to choices like model architecture and vocabulary choice to enable these models to understand sequence context[52]. We have shown that our model, GROVER (Genome Rules Obtained Via Extracted Representations), learns sequence context and many features of DNA without supervision, including GC content, genes,

repetitive elements and replication timing[53]. GROVER can also be fine-tuned on additional data to perform many tasks, including predicting CTCF-binding and recognizing promoters. These findings make GROVER particularly well suited for understanding the relationship between the genome, the epigenome and the DSB landscape.

DSB locations are shaped by both the DNA sequence and the epigenome, with the epigenome itself being influenced in part by the underlying sequence (**Fig. 1A**). In this study, we aim to disentangle these intertwined contributions and assess the extent to which each shapes the DSB landscape using interpretable machine learning approaches. We show that GROVER can be fine-tuned to predict DSB sensitivity, indicating that DSBs can be modeled using the sequence alone. However, combining GROVER with epigenomic features results in a better-performing model, which demonstrates that the epigenome shapes the DSB landscape independent of the sequence. We also show, by using interpretable machine learning techniques, that certain epigenomic features overlap with what can be learned from the sequence, and others do not, indicating that these features could be important for determining the cell type-specificity of DSB locations. By integrating these non-redundant features into the GROVER framework, we learn about the interaction of the DNA sequence with the epigenome for genome instability. Such an approach is generally suited to disentangle the contributions of sequence and chromatin context to genome biology.

Results

The DNA sequence and the epigenome provide distinct information that can be used to model double strand breaks

To study the relationship between the DNA sequence, the epigenome and genome stability, we re-analyzed two previously published DSB sequencing datasets, one using sBLISS in MCF7 cells[43] and one using DSBCapture in NHEK cells[40]. While both technologies involve a blunting step to prepare DNA ends for adapter ligation, a key difference is that sBLISS incorporates unique molecular identifiers (UMIs) and uses T7-based in vitro transcription for linear amplification, enabling quantitative measurement of DSBs[41,42]. In contrast, DSBCapture relies solely on PCR amplification[40]. In our dataset, this likely results in DSBCapture being more sensitive to rare DSBs but prone to overestimating the frequency of recurrent breaks due to PCR duplicates, as well as introducing sequence-dependent amplification biases, which may distort the observed distribution of DSBs[54,55]. To avoid potential biases in DSB locations caused by DSB-inducing treatments[56], both datasets we used were generated from cells that were not treated.

To determine whether DSB counts can be predicted using just the genome sequence information, we fine-tuned GROVER to predict DSB counts. We computed the DSB counts from our datasets for sequence windows that are of length equal to the maximum input size of the model (2,075 base pairs on average). GROVER predicts DSB counts generated from sBLISS and DSBCapture datasets with a Spearman's R of 0.57 and 0.71, respectively (**Fig. 1B & C**), which shows that DSBs can be modeled using the DNA sequence data alone.

In parallel, we trained a model only on epigenomic features. Following previously published methods[43], we gathered various epigenomic datasets generated through ENCODE[57] and used

this data to train a random forest model for prediction of DSB counts on the same windows as we used for GROVER. Despite the small window sizes, this random forest model is still able to predict DSB counts generated by both datasets, with a Spearman's R of 0.76 for the sBLISS data and 0.82 for the DSBCapture data (**Fig. 1D & E**). This shows that the epigenome is more predictive of the DSB sequencing data than the model that is based on genome sequence alone.

To understand the synergy between sequence and epigenome, we combined both of these models by using the fine-tuned GROVER model's predictions as a feature of the random forest model trained on epigenomic data. This combined model performs better than either model, with a Spearman's R of 0.80 for the sBLISS data and 0.86 for the DSBCapture data (**Fig. 1F & C)**. These results indicate that the genome sequence and the epigenome contain both overlapping and distinct information that can be used to predict DSBs. We therefore sought to understand where these overlaps and distinctions might be by further analyzing the features and the models that we trained them on.

DNA sequences display features that are associated with double strand breaks

To better understand what information about DSBs can be encoded within the genome directly, we explored sequence-encoded features that are associated with DSBs. As expected, we found that GC content is correlated with DSBs, with a Spearman's R of 0.23 for the sBLISS data and 0.62 for the DSBCapture data (**Fig. 2A & B**), which is unlikely to be purely of technical origin[40]. These correlation coefficients are meaningfully lower than GROVER's performance, indicating that the sequence contribution to DSB counts is greater and more complex than just its GC content alone. Additionally, we found that the presence of a promoter in a sequence results in a doubling of DSBs on average (**Fig. 2C & D**), in line with previously published results[1–3]. The presence of short interspersed nuclear elements (SINEs), including Alus, and genes are also associated with increased DSBs, which is expected given that these three features are correlated with the presence of promoters[8]. These results hint that DSB hotspots are characterized by specific patterns within the sequence itself. While such complex patterns can be difficult to learn using conventional machine learning models, DNA language models such as GROVER can learn their representations even during pre-training[53]. This likely explains why GROVER is able to model DSB counts using the sequence alone.

Before a sequence can be used as an input to GROVER for fine-tuning, it must be separated into tokens using the model's tokenizer (**Fig. 2E**). These tokens are of variable lengths, which results in variable lengths of the input sequences and requires the DSB counts to be normalized by sequence length in our dataset. An enrichment of specific GROVER token types within high DSB sequences could indicate a pattern of enrichment of specific short sequences in areas of the genome that frequently experience DSBs. When we analyzed these enrichments, we found that, across both datasets, GC-rich tokens like "AGCCC" are frequently associated with increased DSBs while AT-rich tokens like "ATATT" are anticorrelated (**Fig. 2F & G**).

During pre-training, GROVER learns patterns in sequences in a self-supervised fashion, *i.e.*, without any additional biological information[53]. Using DNA sequence as input for GROVER without fine-tuning the model results in a vector of numbers, known as an embedding, which

can be analyzed to interpret what patterns exist within the sequence (**Fig. 2E**). Generating a uniform manifold approximation and projection (UMAP) of these embeddings, we observed separation between sequences with low and high DSBs, which indicates that regions of the genome that are particularly prone to instability can be differentiated by their sequence alone (**Fig. 2H & I**). This analysis also shows that sequences containing similar features are DSB hotspots in both DSBCapture and sBLISS datasets. Overall, our results support the presence of patterns within the sequence itself that are reflected in the associated DSB counts.

Patterns within the epigenome are associated with double strand breaks

To understand what patterns within the epigenome might be associated with DSBs, we calculated the correlation between the epigenomic features we used for our random forest model and our DSB sequencing data (**Fig. 3A & B**). We found that sBLISS data is correlated with H3K36me3 and H3K27ac histone marks as well as RNA polymerase 2 (Pol2) pausing and RepliSeq data (Spearman's R of 0.43, 0.42, 0.37 and 0.34, respectively), while the DSBCapture data is correlated with RepliSeq, DNase-Seq and Pol2 pausing data (Spearman's R of 0.65, 0.64 and 0.51, respectively). Despite using smaller sequence windows, our results align with previously published results[40,43]. The differences between the two datasets may be caused by the specifics of the sequencing technologies used. However, given the fact that these details should largely only affect PCR duplicates, it is more likely that the differences in the epigenomic features can be attributed to differences in the DSB and epigenomic profiles of the two cell lines.

We then asked whether the variation in epigenetic features across the genome could separate sequences experiencing different numbers of DSBs. We performed a principal component analysis (PCA) and found that sequences with high DSB levels are linearly separated from others along the first principal component (PC1), which explains more than 25% of the variance in the data (**Fig. 3C & E**). The correlation between PC1 and DSB levels suggests that some of the dominant patterns in epigenetic variation are aligned with DSB locations. An analysis of PC loadings shows that PC1 for both datasets is primarily driven by DNase-seq, Pol2, H3K27ac and H3K4me3, indicating that a sizeable part of the variance in these datasets can be attributed to the same epigenetic features, though their relative contributions differ (**Fig. 3D & F**). Additionally, we used UMAP to visualize the epigenetic data in non-linear space and again observed separation between low and high DSB sequences (**Fig. S1**). These results confirm previous findings that certain epigenomic features align well with DSBs and that these features can be used to characterize DSB hotspots.

With the aim of understanding which epigenomic features have a central role in shaping the DSB landscape in the genome, we assessed the relative importance of these features within our random forest models by calculating the decrease in the mean-squared error (MSE) of the model when permuting each feature in the training data. We found that H3K36me3, H3K9me3, and H3K27ac are the most important features for the sBLISS data (**Fig. 3A**), while DNase-seq is the most important for the DSBCapture model by far, followed by H3K27ac and RepliSeq (**Fig. 3B**). While the importances we observed reflect the ones published by Ballinger *et al.*[43], one major difference is the smaller importance of RepliSeq in our models. We observed the same results when we repeated our analysis using analogous 50,000 base pair windows (**Fig. S2A, B, E, & F**).

Though we did not include this feature in our random forest model because it is not expressly biologically relevant, while quality checking our epigenomic data, we noticed that the corresponding ChIP-Seq input datasets are correlated with both sBLISS and DSBCapture data (Spearman's R of 0.45 and 0.27, respectively, **Fig. S3A & B**). When we added this data to new random forest models, it became the top feature in the sBLISS model, dwarfing all other features and substantially boosting the performance of the model (**Fig. S2C, D, G, H, S3C, D, E & F**). Assuming there is no major sequencing bias in sBLISS, which is supported by the original publication[41], this result could suggest overlap in sequences that are prone to DSBs and sequences that are prone to being over-sequenced in a ChIP-Seq input sample due to a bias in fragmentation and subsequent size selection.

In summary, our analysis of these epigenomic features and our random forest models demonstrates that patterns in these epigenomic features are associated with DSBs and that these features can be used to model DSBs.

The epigenome and the genome contain non-overlapping patterns associated with double strand breaks

After identifying patterns within the sequence and the epigenome that are associated with DSBs, we sought to compare these patterns to determine their similarities and differences. When we added GROVER's predictions to our random forest model that previously only contained epigenomic features, we found that the GROVER predictions are in the top 4 most important features for the model (**Fig. 4C & D**). This indicates that GROVER learns important DSB information from the sequence that is not present in the epigenetic datasets. This is expected given the increase in performance when GROVER is included in the epigenomic model (**Fig. 1F & G**).

In this importance analysis, features encoding similar information tend to "steal" importance from each other[58], so analyzing how these importances change when GROVER is included in the model could give us a better idea of how DSB sensitivity is encoded at the genomic level and at the epigenomic level. For the sBLISS data, we observed that H3K36me3, H3K27me3, and especially CTCF decrease in importance, indicating that the information learned by GROVER in the sequence is partially redundant with the information encoded in these features, or completely redundant, in the case of CTCF (**Fig. 4A & C**). CTCF binding can be predicted by GROVER very well[53], which suggests that it learned the link between CTCF binding and DSBs[35,59], from the sequence alone. We also found that H3K9me3, H3K4me3, and H3K27ac retain the same importance when GROVER predictions are included, hinting that these epigenetic datasets contain patterns associated with DSBs that cannot be learned from the sequence alone.

For the DSBCapture data, the ranks of the epigenetic features remain largely the same after the inclusion of the GROVER predictions, though DNase-seq and RepliSeq lose some of their importance, which indicates that some of the information contained within these features overlaps with the patterns found within the sequence (**Fig. 4B & D**). These results are likely influenced by the dominance of DNase-seq as a feature in the DSBCapture models. However, we observed once again that the importance of H3K27ac does not change for this dataset,

which supports that this epigenetic mark encodes information that cannot be learned from the sequence.

To provide GROVER with complementary epigenetic information, we incorporated these epigenetic features directly within the architecture of the GROVER model. This differs from using the GROVER predictions per sequence as a feature of the random forest model since these features are treated independently by the models. Here, we allowed GROVER to be fine-tuned while taking into account the non-redundant epigenomic features. This approach should improve the model's ability to interpret the sequence in a cell-type specific way. We used H3K9me3 and H3K27ac data for the sBLISS model and H3K27ac data for the DSBCapture model. These models have improved performance over the original fine-tuned GROVER models for both sBLISS and DSBCapture, with a Spearman's R of 0.76 and 0.79, respectively (**Fig. 4E & F**). For the sBLISS data, the fine-tuned GROVER model even matches the performance of the epigenomic model (**Fig. 1D**). This indicates that the DNA sequence combined with two minimal epigenomic features that encode cell-type specificity are enough to model DSBs as well as the full epigenomic panel.

Discussion

In this study, we sought to disentangle and understand the influences of the epigenome and the DNA sequence on genome stability. To this end, we successfully fine-tuned GROVER to predict DSBs using only DNA sequences. This result shows that DSBs can be modeled directly using the genome sequence information, and our analysis of sequence-encoded features shows how these features appear to influence DSB sensitivity. To identify whether the sequence alone is enough to fully model DSBs, we compared our fine-tuned GROVER model's performance to the performance of an epigenome-only model. We found that GROVER performs worse than this model, showing that the epigenome contains additional useful information for modeling DSBs that cannot be learned from the sequence. Besides technical reasons, this difference in performance can be explained by the presence of cell-type specific information that cannot be interpreted from the DNA sequence, which is largely shared across all cell types within an organism. We also found that the two datasets differ in which features are particularly predictive of DSBs. This may be due to differences in the sequencing technologies or biological differences between the cell types. The minor differences between the two technologies point towards the latter, though this is difficult to fully answer given the current availability of datasets. This highlights the need for more high-quality DSB-sequencing datasets in untreated cells. Additionally, technologies such as DSBCapture and sBLISS provide a snapshot of DSBs and their repair, and it is challenging to differentiate the two and interpret the associated patterns accordingly. Selective DNA repair rather than DNA damage impact could explain why H3K36me3 is such a strong predictor for the sBLISS data, for example, given the histone mark's important role in DNA repair[26].

When we combined the two models, we found that the combined model's performance is greater than either of its parts, indicating that the sequence and epigenome provide synergistic information. Furthermore, we found that the sequence itself, in the form of GROVER's predictions, is one of the most important features of this combined model. These results show that the genome and the epigenome capture both shared and distinct information. Analyzing the importances of this combined model and how they differ from the

epigenome-only model revealed that some of the DSB information provided by epigenomic features like H3K36me3 can be directly learned from the sequence, which suggests that the sequence is related to the location of these epigenetic marks. This analysis also shows that certain features like H3K27ac do not overlap with what GROVER learns from the sequence, hinting that these features are not influenced by the sequence in the context of DSBs. Broadly, these results underscore the importance of the epigenome in shaping the DSB landscape, while also further raising the question of which epigenomic features act directly or merely reflect underlying DNA sequence patterns.

Our findings reveal that the sequence alone is not sufficient to fully model DSBs. However, when we added the sequence-independent epigenomic features directly to the GROVER architecture and fine-tuned the model, we found that the model's performance is competitive with the epigenomic model. Interpreting the importances of random forest models is a powerful tool to identify the genome-wide relevance of these features. However, sophisticated approaches for the interpretation of DNA language models, such as attention analysis, have been developed[49–51] that could allow the fine-grained interpretation of these patterns and identify in which sequence contexts particular epigenomic features are important for DSBs. We found that specific tokens are associated with DSBs and confirmed previously published results regarding the association of high GC content and promoters with DSBs[1–5,7–9]. A fine-grained analysis could help determine which specific combinations of these elements with epigenomic features result in DSB sensitivity. However, these interpretability analyses are still maturing, and their use in a biomedical context remains challenging[49–51]. Because the epigenomic features we integrated within GROVER appear to be sequence-independent, upscaling such an approach may allow the associated models to interpret the sequence in a cell-type specific manner. In addition, this strategy of multimodal learning improves performance, which is often associated with reduced noise and greater interpretability, without introducing redundant features that overlap much with the sequence. As interpretation methods continue to develop, incorporating a minimal set of epigenomic features may offer a practical way to include cell-type specific information while preserving model clarity.

In conclusion, our study advances the understanding of how both DNA sequence and epigenomic context contribute to genome stability. By leveraging interpretable machine learning approaches, we demonstrate that while DNA language models like GROVER can capture important sequence-level information related to DSBs, the epigenome provides synergistic information. Importantly, we show that incorporating a minimal set of sequence-independent epigenomic features can substantially improve performance without sacrificing interpretability, offering a promising strategy for building biologically meaningful models. This work highlights the potential of combining deep learning with genomic and epigenomic data to understand the relationship between DNA sequence and genome biology. We hope these insights will inspire future efforts to use DNA language models not just for prediction but for revealing the principles by which sequence and epigenomics shape the genome's stability and function across diverse biological settings.

Methods

Machine learning and data preparation scripts were written in Python (v3.11.5) with scikit-learn (v1.2.2), numpy (1.26.2), pytorch (v2.7.1+cu126), pandas (v2.1.1), and pyranges (v0.1.2). Plotting and statistical testing scripts were written in R (v4.4.0) with ggplot2 (v3.5.1), dplyr (v1.1.4), readr (v2.1.5), scales (v1.3.0), viridis (v0.6.5), reshape2 (v1.4.4), tidyverse (v2.0.0), ggrastr (v1.0.2), ggsignif (v0.6.4), RColorBrewer (v1.1.3), patchwork (v1.3.0) and reticulate (v1.40.0).

Double strand break data

The previously published, GROVER-tokenized *Homo sapiens* (human) hg19 (GRCh37) genome was divided into 510 token windows[53]. Any windows containing Ns were removed from the dataset. DSB counts were then determined according to previously published methods[40,41,43] and by adapting the code obtained from Ballinger *et al*.[43]. For the sBLISS data, FASTQ files were downloaded from the Sequence Read Archive (SRP150602) and filtered to verify that each read contained the expected 8 bp UMI and the 8 bp sample barcode while allowing one mismatch per barcode using *scan_for_matches* (https://blog.theseed.org/servers/2010/07/scan-for-matches.html). Filtered reads were then mapped to the hg19 genome using BWA-MEM (v0.7.17-r1198-dirty)[60] and alignments with a mapping quality lower than 30 were removed using SAMtools (v1.3.1)[61]. Read alignments were used along with their UMIs to produce a strand-aware, deduplicated list of breakpoint locations and counts as a bedGraph file which was then filtered for PCR duplicates using umi_filter.py[43] before it was converted to a bigWig file using bedGraphToBigWig (v4) from kentUtils (https://github.com/ucscGenomeBrowser/kent). For the DSBCapture data, bedGraph files containing DSB locations were directly obtained from the Gene Expression Omnibus (GSE78172) and converted to bigWig files using bedGraphToBigWig (v4). Three biological replicate samples (SRR7346494, SRR7346495, and SRR7346496) were summed for the sBLISS data, and two biological replicate samples (GSM2068755 and GSM2068756) were averaged for the DSBcapture data, in accordance with previously published methods[43]. DSB counts per 510 token window were then calculated by using bigWigAverageOverBed (v4) from kentUtils (https://github.com/ucscGenomeBrowser/kent) which calculates the average number of DSBCapture reads or sBLISS UMIs per base pair, therefore normalizing for the differences in window lengths. Genomic windows overlapping centromeric or telomeric regions (downloaded from the UCSC genome browser, https://www.genome.ucsc.edu/),or difficult to map regions (ENCODE blacklist version 1[62]) were removed from the dataset using BEDtools intersect[63] (v2.26.0). Finally, genomic windows with DSB counts greater than the 99.99th percentile were removed from the dataset to stabilize model training.

GROVER fine-tuning

Genomic windows were split into train, validation and test sets using an 80:10:10 split for model training. All performance metrics shown were calculated using the unseen test set alone. DSB counts were scaled using StandardScaler from scikit-learn (v1.2.2) before training and testing. This scaling was later reversed for plotting. The pre-trained GROVER model was downloaded from HuggingFace (https://huggingface.co/PoetschLab/GROVER) using huggingface_hub (v0.20.3) and fine-tuned using Pytorch (v3.11.5) with an MSE loss function, using the AdamW optimizer with an epsilon of 10^{-8}, a learning rate of 10^{-6}, no warmup, a batch size of 32, and a dropout probability of 0.2. Fine-tuning hyperparameters were optimized

using a randomized gird search on 1% of the dataset. The models were trained for 5 epochs but the performance on the validation set was assessed after each epoch and only the best performing model was kept.

Density plots

Density plots were created using R (v4.4.0) with ggplot2 (v3.5.1). The densities were calculated using the stat_density_2d function of ggplot2, with n = 1000 and adjust = 1.5. A random sample of 10% of the dataset is shown in the density plots.

Epigenomic data

Epigenomic datasets from ENCODE for MCF7 and NHEK cells was obtained for the hg19 genome from UCSC genome browser in the form of bigWig files. RepliSeq data was downloaded from the UWRepliSeq dataset (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeUwRepliSeq) and DNAse-seq data was downloaded from the OpenChromDnase dataset (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeOpenChromDnase). RNAseq data for MCF7 cells was downloaded from the CaltechRnaseq dataset (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeCaltechRnaSeq) and RNAseq data for NHEK cells was downloaded from the RegTxn dataset (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRegTxn). The remaining datasets for NHEK were downloaded from the BroadHistone dataset (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeBroadHistone). H3K9me3, H3K27ac, H3K27me3, H3K36me3, and input data for MCF7 cells was downloaded from the SydhHistone dataset (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeSydhHistone). H3K4me3 data for MCF7 was downloaded from the UwHistone dataset (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeUwHistone). Pol2 and CTCF datasets for MC7 cells were downloaded from the OpenChromChip dataset (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeOpenChromChip) . BigWig files were then averaged over the genomic windows using bigWigAverageOverBed (v4) from kentUtils (https://github.com/ucscGenomeBrowser/kent) to obtain a table of epigenomic markers per genomic window.

Epigenomic model and feature importances

Genomic windows were split into train, validation and test sets using an 80:10:10 split for model training. All performance metrics shown were calculated using the unseen test set alone. Random forest models were trained to predict DSB counts per window using the epigenomic features for each window. DSB counts were scaled using StandardScaler from scikit-learn (v1.2.2) before training and testing. This scaling was later reversed for plotting. Epigenomic features were also scaled in the same way for model training. The models were trained with scikit-learn (v1.2.2) using 500 estimators with no minimum samples per split or per leaf. Feature importances were calculated by training 10 models, measuring their performance using MSE, measuring the decrease in MSE when each feature is shuffled across sequences, and then averaging the decreases in MSE across the 10 models.

Genome annotations

The gene and SINE annotation files for the hg19 genome were obtained from the UCSC genome browser (https://www.genome.ucsc.edu/). Promoter annotations were generated from the chromatin colors for NHEK cells (downloaded from the UCSC genome browser) and for MCF7 cells (downloaded from the Gene Expression Omnibus, GSE57498).

Epigenomic model with GROVER predictions

To incorporate GROVER predictions into the random forest model, we fine-tuned GROVER on two-thirds of the data in each fold and used it to generate predictions for the remaining one-third, following the standard three-fold cross-validation procedure. This process was repeated across all folds to produce a single prediction for each sequence in the entire dataset without redundancy. These predictions were then included within the table of epigenomic features per window, and this table was then used to train the random forest models as before.

Dimensionality reduction and GROVER embeddings

UMAP was performed using the uwot package (v0.2.2) in R (v4.4.0) with 15 neighbors, a minimum distance of 0.25 and the Euclidian distance metric. PCA was performed with centering and scaling using the stats package (v4.4.0) in R (v4.4.0). The data for the UMAP of pre-trained GROVER embeddings was obtained by using the 510-token genomic windows as input into the pre-trained GROVER model to obtain the 512 by 768-dimension matrices that make up the outputs of the final transformer layer of the model for each sequence. The embeddings from the first and last tokens, which are special tokens added to the sequence by the model, were removed from these matrices, and the embeddings were averaged across all tokens to obtain one 768-dimensional vector per genomic window which was then used for dimensionality reduction. To perform these analyses on the epigenomic data, the tables of epigenomic features for each genomic window that were used in the epigenomic random forest models were also used for dimensionality reduction.

GROVER model with integrated epigenomic features

The fine-tuned GROVER model's default classifier head was replaced with a neural network with one hidden layer containing 10000 nodes and a number of input nodes equal to 768 (length of the output of the pre-trained GROVER model) plus one extra input node per epigenomic feature. These models were fine-tuned as before except this time with a learning rate of 10^{-5}. As previously described, the hyperparameters were optimized using a randomized gird search on 1% of the dataset.

Statistics

Correlation coefficients were calculated in R (v4.4.0) using the cor function of the stats package (v4.4.0). Wilcoxon rank-sum test was also calculated in R (v4.4.0) using the wilcox.test function of the stats package (v4.4.0).

Code availability

All scripts are available on Zenodo: https://doi.org/10.5281/zenodo.16375951.

Data availability

Input and analysis files are also available on Zenodo: https://doi.org/10.5281/zenodo.16375951.

References

1. Pommier, Y., Nussenzweig, A., Takeda, S. & Austin, C. Human topoisomerases and their roles in genome stability and organization. *Nature Reviews Molecular Cell Biology* **23**, 407–427 (2022).
2. Castillo-Guzman, D. & Chédin, F. Defining R-loop classes and their contributions to genome instability. *DNA Repair* **106**, 103182 (2021).
3. Crossley, M. P., Bocek, M. & Cimprich, K. A. R-Loops as Cellular Regulators and Genomic Threats. *Molecular Cell* **73**, 398–411 (2019).
4. Puget, N., Miller, K. M. & Legube, G. Non-canonical DNA/RNA structures during Transcription-Coupled Double-Strand Break Repair: Roadblocks or Bona fide repair intermediates? *DNA Repair* **81**, 102661 (2019).
5. Técher, H., Koundrioukoff, S., Nicolas, A. & Debatisse, M. The impact of replication stress on replication dynamics and DNA damage in vertebrate cells. *Nature Reviews Genetics* **18**, 535–550 (2017).
6. Kotsantis, P. *et al.* RTEL1 Regulates G4/R-Loops to Avert Replication-Transcription Collisions. *Cell Rep* **33**, 108546 (2020).
7. Larsen F *et al.* CpG islands as gene markers in the human genome.
8. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
9. Crosetto, N. *et al.* Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nature Methods* **10**, 361–365 (2013).
10. Morales, M. E. *et al.* The Contribution of Alu Elements to Mutagenic DNA Double-Strand Break Repair. *PLOS Genetics* **11**, e1005016 (2015).
11. Lehrman, M. A. *et al.* Mutation in LDL Receptor: Alu-Alu Recombination Deletes Exons Encoding Transmembrane and Cytoplasmic Domains. *Science* **227**, 140–146 (1985).
12. Gasior, S. L., Wakeman, T. P., Xu, B. & Deininger, P. L. The Human LINE-1 Retrotransposon Creates DNA Double-strand Breaks. *Journal of Molecular Biology* **357**, 1383–1393 (2006).
13. Gadgil, R. Y. *et al.* Replication stress at microsatellites causes DNA double-strand breaks and break-induced replication. *Journal of Biological Chemistry* **295**, 15378–15397 (2020).
14. Neil, A. J., Kim, J. C. & Mirkin, S. M. Precarious maintenance of simple DNA repeats in eukaryotes. *Bioessays* **39**, (2017).
15. Correll-Tash, S. *et al.* Double strand breaks (DSBs) as indicators of genomic instability in PATRR-mediated translocations. *Human Molecular Genetics* **29**, 3872–3881 (2020).
16. Glover, T. W., Wilson, T. E. & Arlt, M. F. Fragile sites in cancer: more than meets the eye. *Nature Reviews Cancer* **17**, 489–501 (2017).
17. Chakrabarti, A. M. *et al.* Target-Specific Precision of CRISPR-Mediated Genome Editing. *Molecular Cell* **73**, 699-713.e6 (2019).

564 18. Shou, J., Li, J., Liu, Y. & Wu, Q. Precise and Predictable CRISPR Chromosomal
565 Rearrangements Reveal Principles of Cas9-Mediated Nucleotide Insertion. *Molecular Cell*
566 **71**, 498-509.e4 (2018).
567 19. Shen, M. W. *et al.* Predictable and precise template-free CRISPR editing of pathogenic
568 variants. *Nature* **563**, 646–651 (2018).
569 20. Milano, L., Gautam, A. & Caldecott, K. W. DNA damage and transcription stress. *Molecular*
570 *Cell* **84**, 70–79 (2024).
571 21. Gothe, H. J. *et al.* Spatial Chromosome Folding and Active Transcription Drive DNA Fragility
572 and Formation of Oncogenic MLL Translocations. *Molecular Cell* **75**, 267-283.e12 (2019).
573 22. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and
574 predicts developmental state. *Proceedings of the National Academy of Sciences* **107**,
575 21931–21936 (2010).
576 23. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-
577 specific gene expression. *Nature* **459**, 108–112 (2009).
578 24. Duttke, S. H. C. *et al.* Human Promoters Are Intrinsically Directional. *Molecular Cell* **57**,
579 674–684 (2015).
580 25. Wang, H. & Helin, K. Roles of H3K4 methylation in biology and disease. *Trends in Cell*
581 *Biology* **35**, 115–128 (2025).
582 26. Sharda, A. & Humphrey, T. C. The role of histone H3K36me3 writers, readers and erasers
583 in maintaining genome stability. *DNA Repair* **119**, 103407 (2022).
584 27. Li, Z. & Zhang, Z. The silent guardian: unraveling the roles of H3K9me3 in genome
585 maintenance. *Genome Instability & Disease* **5**, 133–153 (2024).
586 28. Bernstein, B. E. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in
587 Embryonic Stem Cells. *Cell* **125**, 315–326 (2006).
588 29. Kundu, S. *et al.* Polycomb Repressive Complex 1 Generates Discrete Compacted Domains
589 that Change during Differentiation. *Molecular Cell* **65**, 432-446.e5 (2017).
590 30. Hosogane, M., Funayama, R., Shirota, M. & Nakayama, K. Lack of Transcription Triggers
591 H3K27me3 Accumulation in the Gene Body. *Cell Reports* **16**, 696–706 (2016).
592 31. Ayrapetov, M. K., Gursoy-Yuzugullu, O., Xu, C., Xu, Y. & Price, B. D. DNA double-strand
593 breaks promote methylation of histone H3 on lysine 9 and transient formation of
594 repressive chromatin. *Proceedings of the National Academy of Sciences* **111**, 9169–9174
595 (2014).
596 32. Zhang, Y. *et al.* Histone H3K27 methylation modulates the dynamics of FANCD2 on
597 chromatin to facilitate NHEJ and genome stability. *Journal of Cell Science* **131**, jcs215525
598 (2018).
599 33. Hilmi, K. *et al.* CTCF facilitates DNA double-strand break repair by enhancing homologous
600 recombination repair. *Science Advances* **3**, e1601898.
601 34. Lang, F. *et al.* CTCF prevents genomic instability by promoting homologous recombination-
602 directed DNA double-strand break repair. *Proceedings of the National Academy of*
603 *Sciences* **114**, 10912–10917 (2017).
604 35. Canela, A. *et al.* Genome Organization Drives Chromosome Fragility. *Cell* **170**, 507-521.e18
605 (2017).
606 36. Helmrich, A., Ballarino, M. & Tora, L. Collisions between Replication and Transcription
607 Complexes Cause Common Fragile Site Instability at the Longest Human Genes. *Molecular*
608 *Cell* **44**, 966–977 (2011).
609 37. Wilson, T. E. *et al.* Large transcription units unify copy number variants and common
610 fragile sites arising under replication stress. *Genome Res* **25**, 189–200 (2015).

38. Barlow, J. H. *et al.* Identification of early replicating fragile sites that contribute to genome instability. *Cell* **152**, 620–632 (2013).
39. Tubbs, A. *et al.* Dual Roles of Poly(dA:dT) Tracts in Replication Initiation and Fork Collapse. *Cell* **174**, 1127-1142.e19 (2018).
40. Lensing, S. V. *et al.* DSBCapture: in situ capture and sequencing of DNA breaks. *Nat Methods* **13**, 855–857 (2016).
41. Yan, W. X. *et al.* BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nat Commun* **8**, 15058 (2017).
42. Bouwman, B. A. M. *et al.* Genome-wide detection of DNA double-strand breaks by in-suspension BLISS. *Nature Protocols* **15**, 3894–3941 (2020).
43. Ballinger, T. J. *et al.* Modeling double strand break susceptibility to interrogate structural variation in cancer. *Genome Biology* **20**, 28 (2019).
44. Mourad, R., Ginalski, K., Legube, G. & Cuvier, O. Predicting double-strand DNA breaks using epigenome marks or DNA at kilobase resolution. *Genome Biology* **19**, 34 (2018).
45. Schep, R. *et al.* Impact of chromatin context on Cas9-induced DNA double-strand break repair pathway balance. *Molecular Cell* **81**, 2216-2230.e10 (2021).
46. Vergara, X. *et al.* Widespread chromatin context-dependencies of DNA double-strand break repair proteins. *Nature Communications* **15**, 5334 (2024).
47. Schep, R., Leemans, C., Brinkman, E. K., van Schaik, T. & van Steensel, B. Protocol: A Multiplexed Reporter Assay to Study Effects of Chromatin Context on DNA Double-Strand Break Repair. *Frontiers in Genetics* **Volume 12-2021**, (2022).
48. Vaswani, A. *et al.* Attention Is All You Need. *CoRR* **abs/1706.03762**, (2017).
49. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods* **18**, 1196–1203 (2021).
50. Consens, M. E., Li, B., Poetsch, A. R. & Gilbert, S. Genomic language models could transform medicine but not yet. *npj Digital Medicine* **8**, 212 (2025).
51. Benegas, G., Ye, C., Albors, C., Li, J. C. & Song, Y. S. Genomic language models: opportunities and challenges. *Trends in Genetics* **41**, 286–302 (2025).
52. Sanabria, M., Hirsch, J. & Poetsch, A. R. Distinguishing word identity and sequence context in DNA language models. *BMC Bioinformatics* **25**, 301 (2024).
53. Sanabria, M., Hirsch, J., Joubert, P. M. & Poetsch, A. R. DNA language model GROVER learns sequence context in the human genome. *Nat Mach Intell* **6**, 911–923 (2024).
54. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports* **2**, 666–673 (2012).
55. Van Gelder, R. N. *et al.* Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proceedings of the National Academy of Sciences* **87**, 1663–1667 (1990).
56. van Bueren, M. A. E. & Janssen, A. The impact of chromatin on double-strand break repair: Imaging tools and discoveries. *DNA Repair* **133**, 103592 (2024).
57. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
58. Beware Default Random Forest Importances. http://explained.ai/decision-tree-viz/index.html.
59. Chen, B. *et al.* Stratifying TAD boundaries pinpoints focal genomic regions of regulation, damage, and repair. *Briefings in Bioinformatics* **25**, bbae306 (2024).
60. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).

61. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
62. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports* **9**, 9354 (2019).
63. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
64. UW Repli-seq Track Settings. https://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeUwRepliSeq.
65. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences* **107**, 139–144 (2010).

Acknowledgements

This work was supported by the Center for Scalable data analytics and artificial intelligence (Scads.AI) Dresden-Leipzig. This work was partially funded by the Center for Advanced Systems Understanding (CASUS) which is financed by Germany's Federal Ministry of Education and Research (BMBF) and by the Saxon Ministry for Science, Culture, and Tourism (SMWK) with tax funds on the basis of the budget approved by the Saxon State Parliament.
ARP was supported by the Mildred Scheel Early Career Center Dresden P2, funded by the German Cancer Aid. The authors gratefully acknowledge the computing time made available to them on the high-performance computer at the NHR Center of TU Dresden. This center is jointly supported by the Federal Ministry of Education and Research and the state governments participating in the NHR (www.nhr-verein.de/unsere-partner). We thank Melissa Sanabria for her help with implementing the GROVER model, Jakub Zastapiło and Mario Aguilar-Herrador for their feedback on data visualization, and the Biomedical Genomics group at TU Dresden for helpful discussions throughout the research process.

Author contributions

ARP conceptualized the study. **PMJ** trained the models. **PMJ** analyzed the data with input from **ARP. PMJ** and **ARP** wrote the manuscript.

Declaration of interests

The authors declare no competing interests.

Figure 1

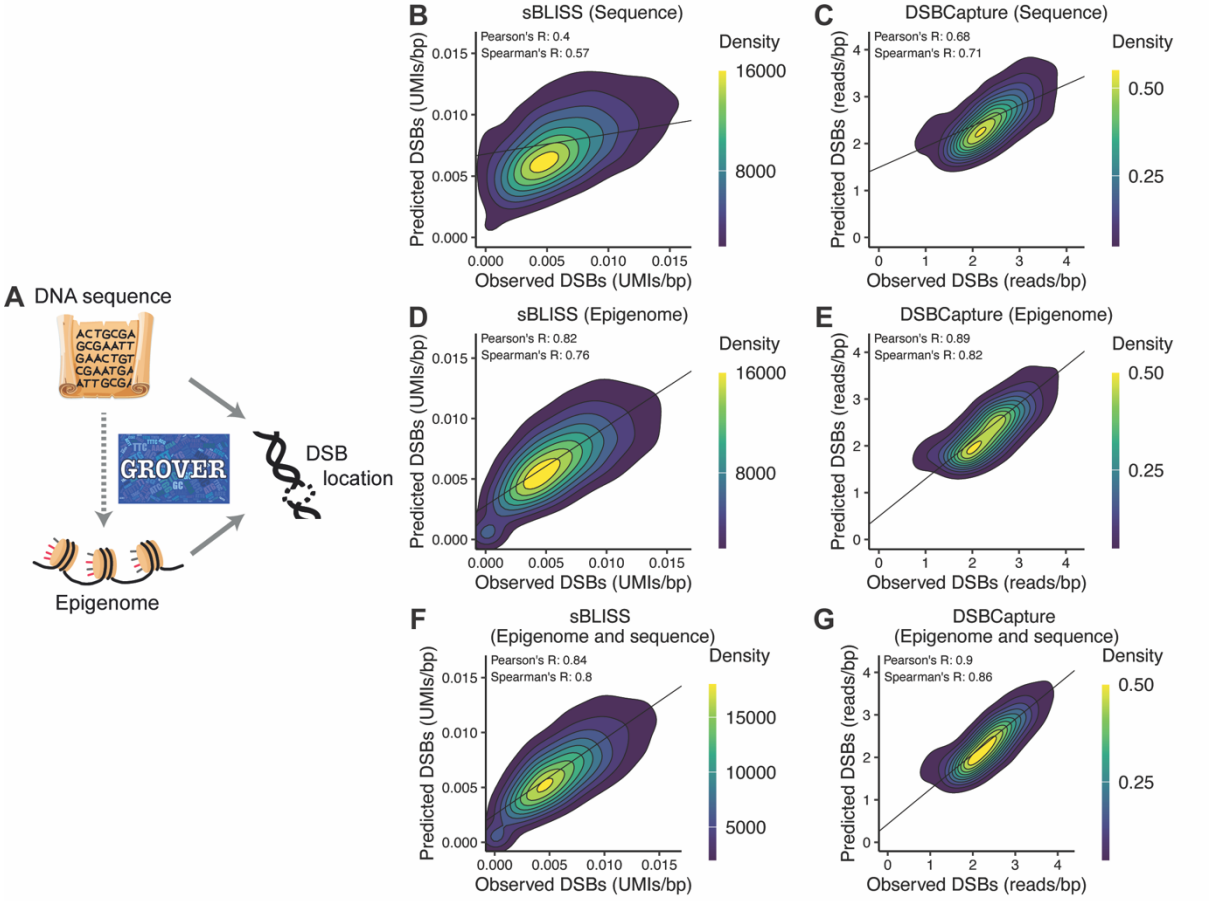

Fig. 1. The DNA sequence and the epigenome both contain information that can help model double strand breaks (DSBs). A. DSB locations are shaped by both the DNA sequence and the epigenome, with the epigenome itself being influenced in part by the underlying sequence. DNA language models like GROVER, combined with interpretable machine learning approaches, can be used to disentangle these interactions. **B and C.** Performance of GROVER model fine-tuned to predict sBLISS data (**B**) and DSBCapture data (**C**) using the DNA sequence alone. **D and E.** Performance of random forest model trained on epigenomic features to predict sBLISS data (**D**) and DSBCapture data (**E**). **F and G.** Performance of random forest model trained on epigenomic features combined with the predictions of the GROVER fine-tuned model to predict sBLISS data (**F**) and DSBCapture data (**G**). DSB data is shown as molecular identifier (UMI) counts (sBLISS) or read counts (DSBCapture) averaged per base pair over the length of each window.

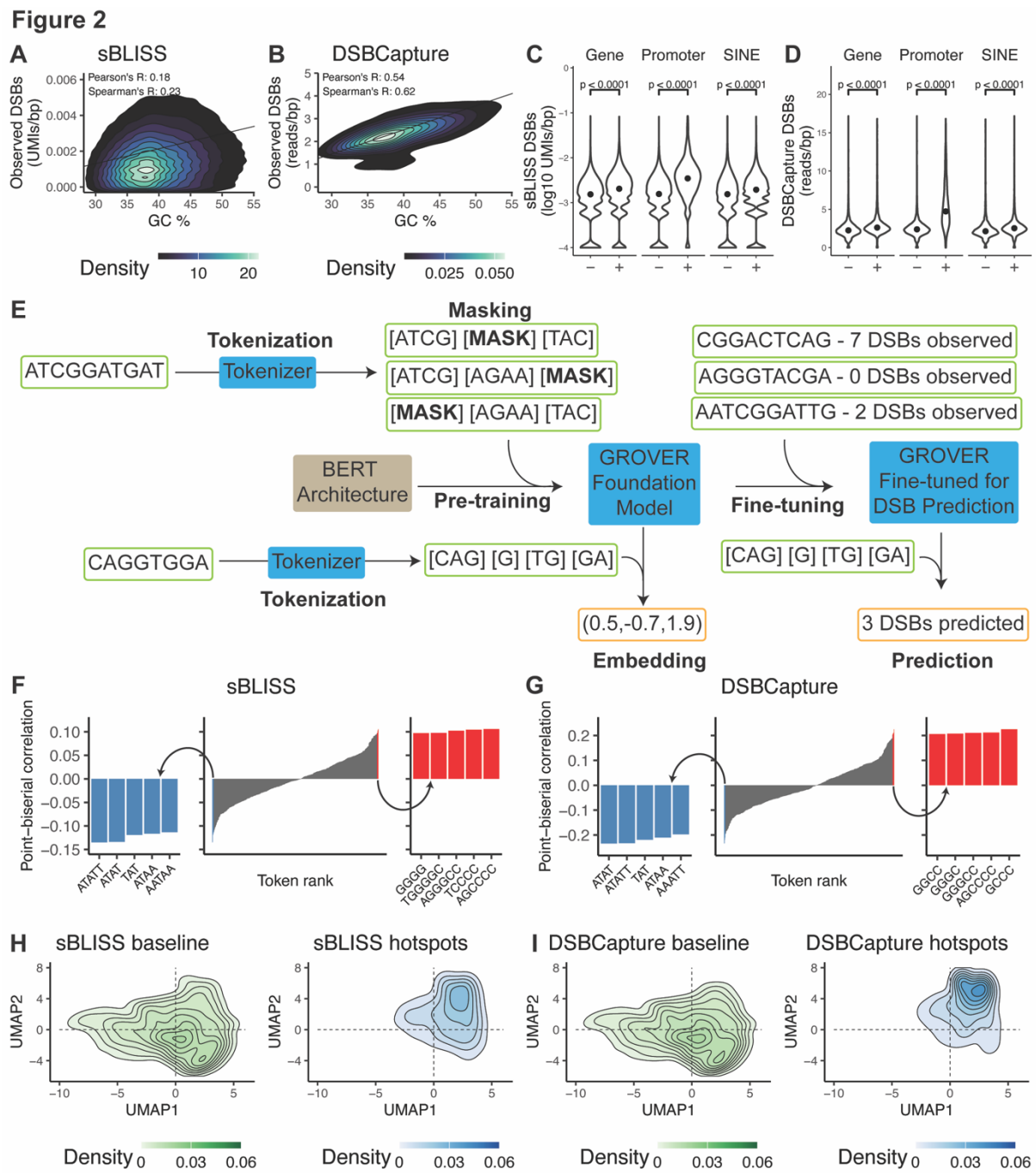

Fig. 2. Sequence features are associated with double strand breaks (DSBs). A and B. Correlation between sequence GC content and sBLISS data (**A**) or DSBCapture data (**B**). **C and D.** Comparison of sBLISS data (**C**) and DSBCapture data (**D**) across sequences categorized by the presence of a gene, promoter, or short interspersed nuclear element (SINE) ($p < 0.0001$ by Wilcoxon rank test). Points represent median values. **E.** Explanation of GROVER model data processing, pre-training and fine-tuning workflow (see main text). For the sake of visual clarity, a few simplifications have been made. The embedding output from the pre-trained model has 768 dimensions, but only three are shown here. Sequences input into the model are 510 tokens long (or 2,075 base pairs on average), but only a few tokens are shown here. Additionally, DSB counts are normalized for sequence window length in our data, which is not shown here. **F and G.** Point-biserial correlation between the presence of each token within the GROVER vocabulary and sBLISS data (**F**, UMIs per base pair in each window) or DSBCapture data (**G**, reads per base pair in each window). **H and I.** Uniform manifold approximation and projection (UMAP) analysis of pre-trained GROVER embeddings separated by DSB level, according to sBLISS data (**H**) and DSBCapture data (**I**). Sequences are separated based on whether they are in the top 10% (hotspots) or bottom 90% (baseline) of sequences.

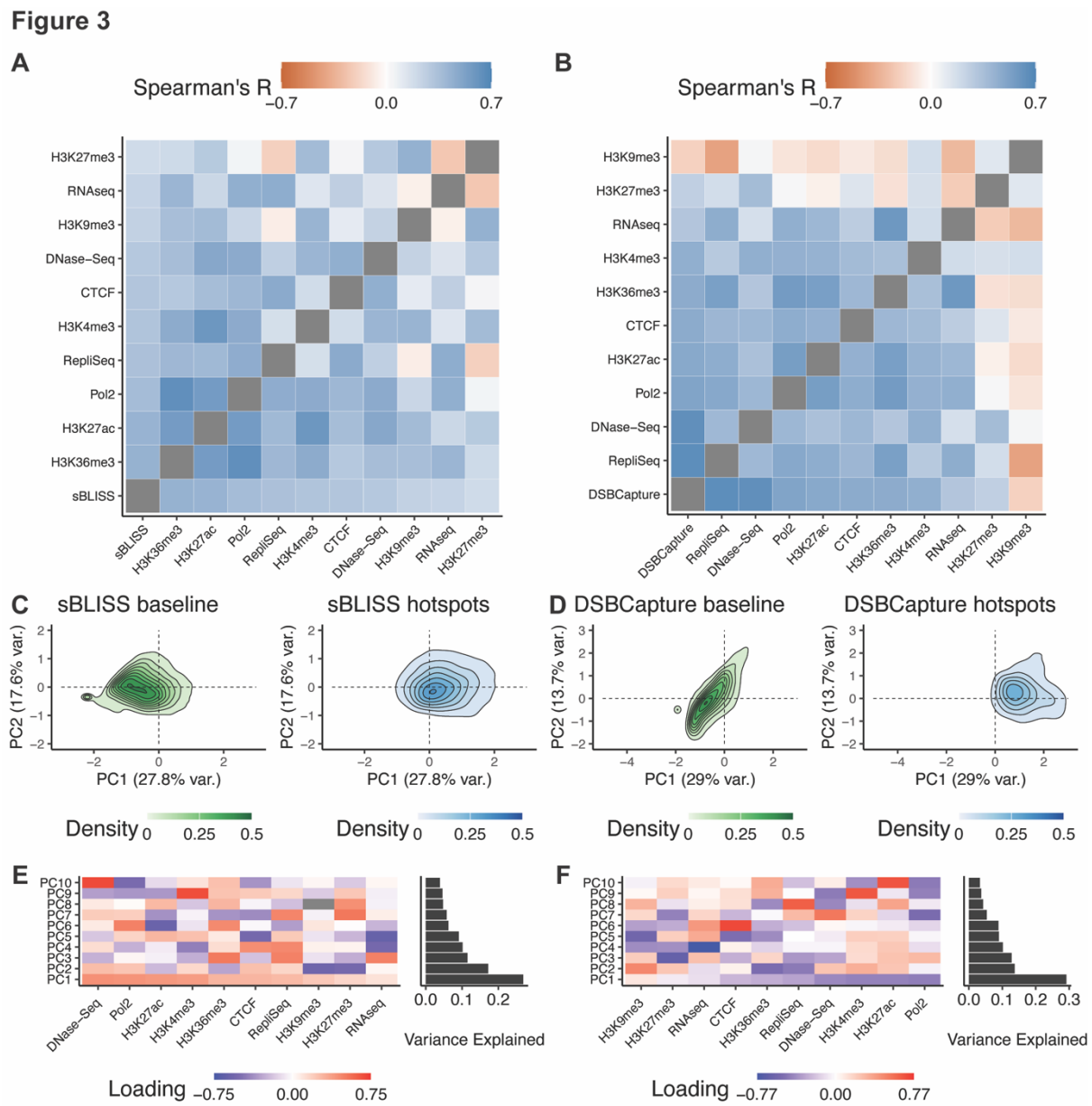

Fig. 3. Epigenomic features are correlated with double-strand breaks (DSBs). A. Spearman correlation between MCF7 epigenomic features and sBLISS data (**A**, unique molecular identifiers per base pair) or DSBCapture data (**B**, reads per base pair). **C and D.** Principal components analysis (PCA) of MCF7 (**C**) and NHEK (**D**) epigenomic features. Sequences are separated based on whether they are in the top 10% (hotspots) or bottom 90% (baseline) of sequences. **E and F.** Loading of each MCF7 (**E**) and NHEK (**F**) epigenomic feature within each PC, as well as the variance explained by each PC.

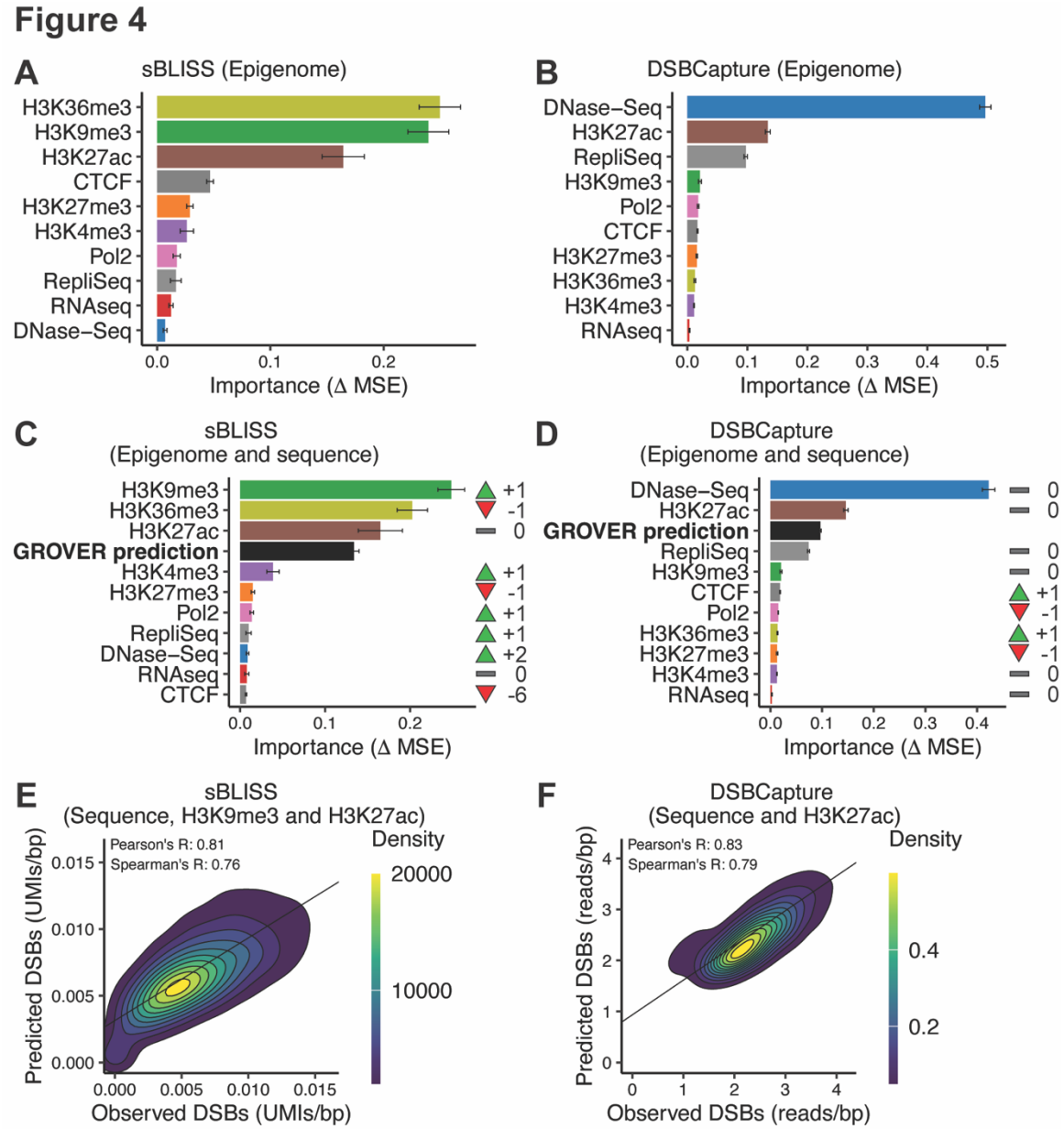

Fig. 4. Epigenomic features predict double strands breaks (DSBs). A and B. Importances of features used by the random forest model trained on epigenomic features to predict sBLISS data (**A**) and DSBCapture data (**B**). **C and D.** Importances of features used by the random forest model trained on epigenomic features combined with the predictions of the fine-tuned GROVER model to predict sBLISS data (**C**) and DSBCapture data (**D**). Arrows represent changes in rank of feature importances compared to panel **A** and **B**, respectively. **E and F.** Performance of GROVER model fine-tuned on DNA sequence and limited epigenomic features to predict sBLISS data (**E**) and DSBCapture data (**F**). Error bars represent standard deviation from the mean. Importances are measured using the change in mean squared error (MSE) of the model when a feature is shuffled in the test data. DSB data is shown as molecular identifier (UMI) counts (sBLISS) or read counts (DSBCapture) averaged per base pair over the length of each window.

Figure S1

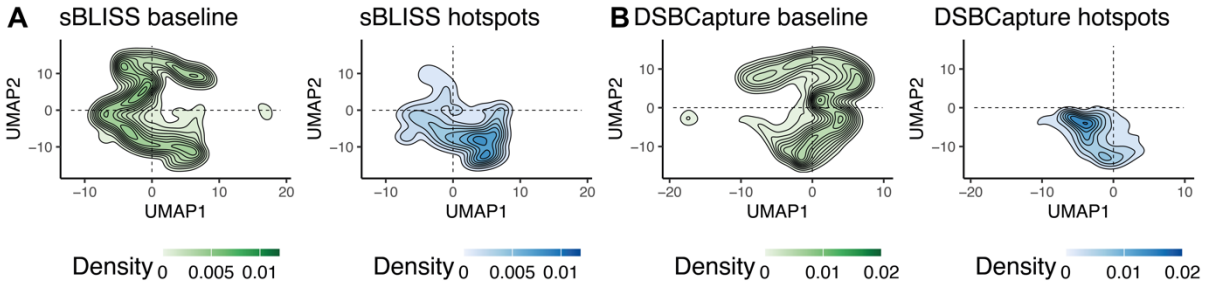

Fig. S1. Uniform manifold approximation and projection (UMAP) analysis of epigenomic features. A and B. UMAP of MCF7 (**A**) and NHEK (**B**) epigenomic features. Sequences are separated based on whether they are in the top 10% (hotspots) or bottom 90% (baseline) of sequences.

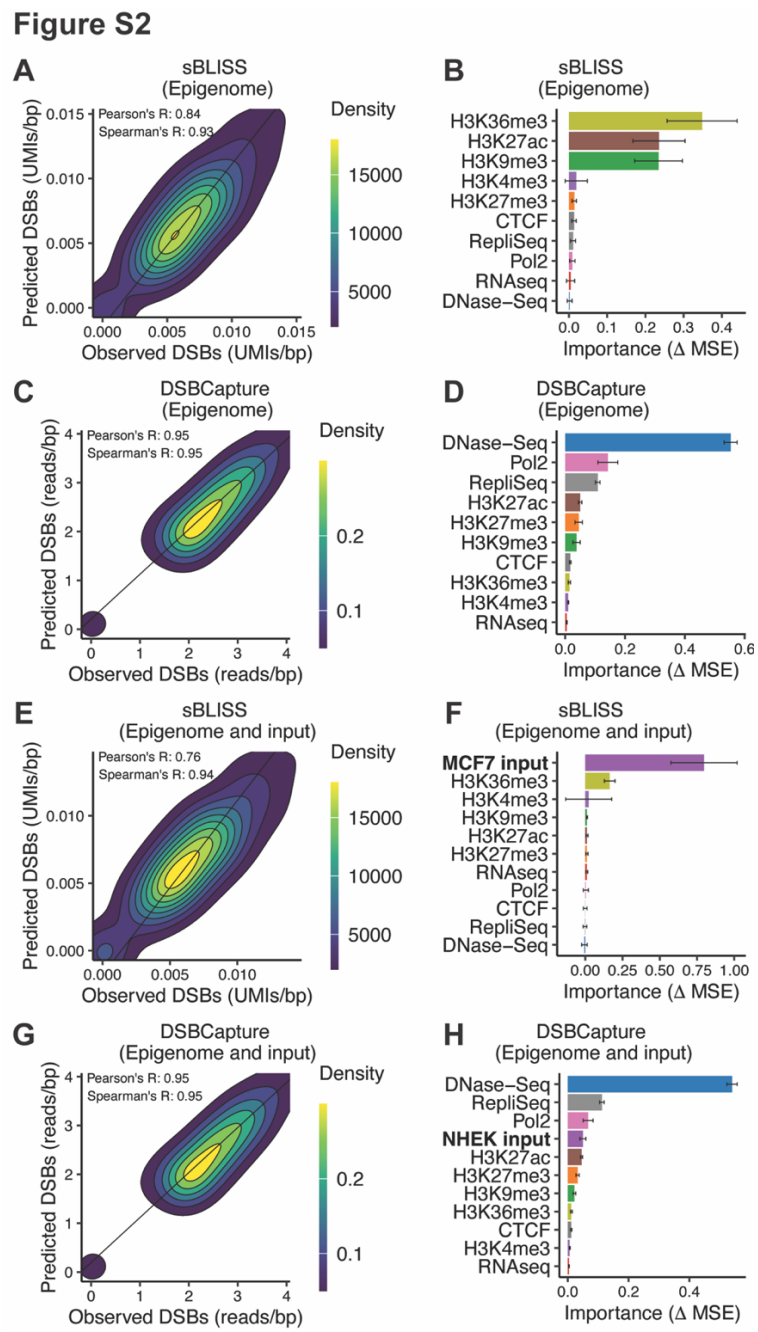

Fig. S2. Performance and importances of random forest models trained to predict double strand breaks (DSBs) within 50kbp genomic windows. A and C. Performance of random forest model trained on epigenomic features to predict sBLISS data (**A**) and DSBCapture data (**C**). **B and D.** Importances of features used by the random forest model trained on epigenomic features to predict sBLISS data (**B**) and DSBCapture data (**D**). **E and G.** Performance of random forest model trained on epigenomic features and ChIP-Seq input data to predict sBLISS data (**E**) and DSBCapture data (**G**). **F and H.** Importances of features used by the random forest model trained on epigenomic features and ChIP-Seq input data to predict sBLISS data (**F**) and DSBCapture data (**H**). DSB data is shown as molecular identifier (UMI) counts (sBLISS) or read counts (DSBCapture) averaged per base pair over the length of each window. Error bars represent standard deviation from the mean. Importances are measured using the change in mean squared error (MSE) of the model when a feature is shuffled in the test data. Note: While the results we present here largely reflect the ones published by Ballinger et al., one major difference is the reduced importance of RepliSeq in our models. While we used data from the same published datasets as Ballinger et al., our analysis used the Wavelet-smooth Signal tracks rather than the Summed Densities tracks from ENCODE. The description of these file types as well as the original RepliSeq publication support our choice [64,65]. The importance of the ChIP-Seq input data in our models could explain why Ballinger et al. found that RepliSeq is such an important feature while we did not.

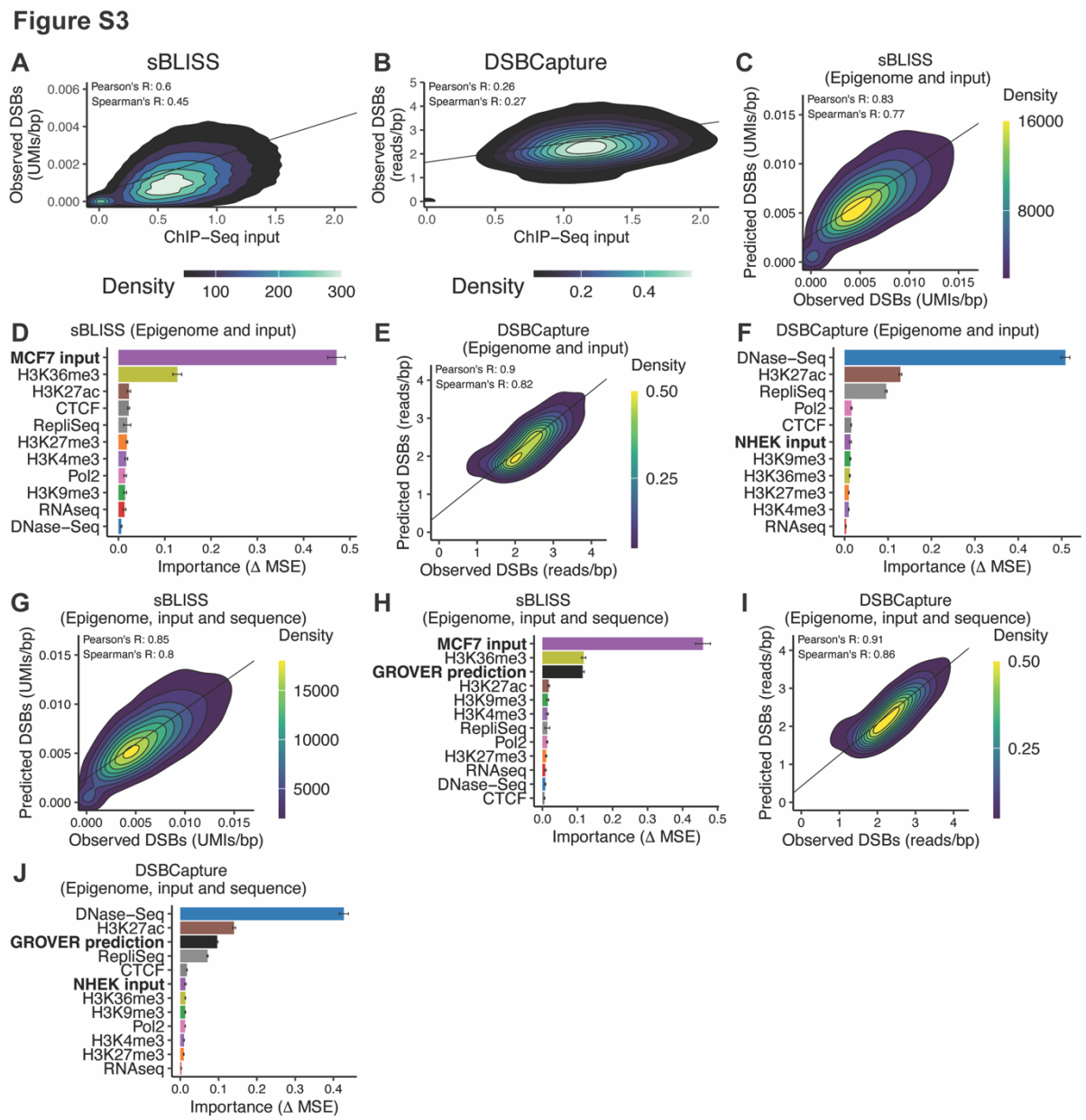

Fig. S3. Performance and importances of random forest models including ChIP-Seq input data, trained to predict double strand breaks (DSBs). A and B. Correlation between ChIP-Seq input data and observed sBLISS data (**A**) or DSBCapture data (**B**). **C and E.** Performance of the random forest model trained on epigenomic features and ChIP-Seq input data to predict sBLISS data (**C**) and DSBCapture data (**E**). **D and F.** Importances of features used by the random forest model trained on epigenomic features and ChIP-Seq input data to predict sBLISS data (**D**) and DSBCapture data (**F**). **G and I.** Performance of the random forest model trained on epigenomic features, ChIP-Seq input data and fine-tuned GROVER model predictions to predict sBLISS data (**G**) and DSBCapture data (**I**). **H and J.** Importances of features used by the random forest model trained on epigenomic features, ChIP-Seq input data and fine-tuned GROVER model predictions to predict sBLISS data (**H**) and DSBCapture data (**J**). DSB data is shown as molecular identifier (UMI) counts (sBLISS) or read counts (DSBCapture) averaged per base pair over the length of each window. Error bars represent standard deviation from the mean. Importances are measured using the change in mean squared error (MSE) of the model when a feature is shuffled in the test data.