Predicting double-strand DNA breaks using epigenome marks or DNA at kilobase resolution

Raphaël Mourad[1] and Olivier Cuvier[1]

April 10, 2017

[1] Laboratoire de Biologie Moléculaire Eucaryote (LBME), CNRS, Université Paul Sabatier (UPS), 31000 Toulouse, France

Abstract

Double-strand breaks (DSBs) result from the attack of both DNA strands by multiple sources, including exposure to ionizing radiation or reactive oxygen species. DSBs can cause abnormal chromosomal rearrangements which are linked to cancer development, and hence represent an important issue. Recent techniques allow the genome-wide mapping of DSBs at high resolution, enabling the comprehensive study of DSB origin. However these techniques are costly and challenging. Hence we devised a computational approach to predict DSBs using the epigenomic and chromatin context, for which public data are available from the ENCODE project. We achieved excellent prediction accuracy ($AUC = 0.97$) at high resolution (< 1 kb), and showed that only chromatin accessibility and H3K4me1 mark were sufficient for highly accurate prediction ($AUC = 0.95$). We also demonstrated the better sensitivity of DSB predictions compared to BLESS experiments. We identified chromatin accessibility, activity and long-range contacts as best predictors. In addition, our work represents the first step toward unveiling the "*cis*-DNA repairing" code underlying DSBs, paving the way for future studies of *cis*-elements involved in DNA damage and repair.

1 Introduction

Double-strand breaks (DSBs) arise when both DNA strands of the double helix are severed. DSBs are caused by the attack of deoxyribose and DNA bases by reactive oxygen species and other electrophilic molecules [22]. DSBs are particularly hazardous to the cell because they can lead to deletions, translocations, and fusions in the DNA, collectively referred as chromosomal rearrangements [23]. DSBs are most commonly found in cancer cells. Several high-throughput sequencing techniques have been developped for the genome-wide mapping of DSBs *in situ* such as GUIDE-seq [34], BLESS [8] and DSBCapture [17]. The most recent technique, DSBCapture, allowed to map more than 80 thousand endogenous DSBs at a lower than 1 kb resolution in human. To date, DSBs have been mapped at high resolution only for a few number of cell lines, because of high sequencing costs and experimental difficulties. This prevented the comprehensive study of the double-strand break landscape in the human genome across diverse cell lines and tissues.

Chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq) and DNase I hypersensitive site sequencing (DNase-seq) data are publicly available for dozens of cell lines with the

ENCODE [31] and Roadmap Epigenomics [7] projects. On the one hand, recent studies have shown that the mapping of regulatory elements such as enhancers or promoters can be accurately predicted using available epigenome and chromatin data [9, 14]. Other studies have shown that the epigenome can in its turn be predicted by combinations of DNA motifs and DNA shape [21, 29, 36, 38]. On the other hand, DSBs and the resulting DNA repair mechanisms were shown to be linked to epigenome marks, including H3K4me1/2/3 and chromatin accessibility [17]. Accordingly, PRDM9-mediated trimethylation of H3K4 (H3K4me3) was originally shown to play a critical role in regulating DSBs associated with meiotic recombination hotspots [1, 11, 24]. Moreover the repair of DSBs involves both posttranslational modification of histones, in particular γ-H2AX, and concentration of DNA-repair proteins at the site of damage [13, 28]. It remains unclear to what extent DNA motifs or histone modifications predict or regulate the cellular response to DSBs in other developmental stages. Here we thus sought to test whether publicly available epigenome and chromatin data, or DNA motifs and shape could be used to possibly predict DSBs.

In this article, we demonstrate, for the first time, that endogenous DSBs can be computationally predicted using the epigenomic and chromatin context, or using DNA sequence and DNA shape. Our predictions achieve excellent accuracy ($AUC > 0.97$) at high resolution (<1kb) using available ChIP-seq and DNase-seq data from public databases. DNase, CTCF binding and H3K4me1/2/3 are among the best predictors of DSBs, reflecting the importances of chromatin accessibility, activity and long-range contacts in determining DSB sites and subsequent repairing. Another important predictor is p63 binding, a member of the p53 gene family known to be involved in DNA repair. We also successfully predict DSB sites using DNA motif occurences only ($AUC = 0.839$), supporting a "*cis*-DNA repairing" code of DSBs involving numerous DNA damage and repair protein binding motifs including members of the transcription factor complex AP-1 and p53 family. In addition, DNA shape analysis further reveals the importance of the structure-based readout in determining DSB sites, complementary to the sequence-based readout (motifs).

2 Results and Discussion

2.1 Double-strand break prediction approach

Our computational approach to predict DSBs is schematically illustrated in Figure 1. In the first step, we analyzed public DSBCapture data from Lensing *et al.* [17], which provided the most sensitive and accurate genome-wide mapping of double-strand breaks to date (Figure 1a). DSBCapture is a technique that captures DSBs *in situ* and that can directly map them at the single-nucleotide resolution. DSBCapture peaks were called with less than 1 kb resolution (median size of 391 bases). The DSBCapture peaks obtained from two biological replicates were intersected to yield more reliable DSB sites. Endogeneous breaks were captured for NHEK cells, for which numerous ChIP-seq and DNase-seq data were made publicly available by the ENCODE project [31]. In the second step, we integrated and mapped different types of data within DSB sites and non-DSB sites. To prevent bias effects, non-DSB sites were randomly drawn from the human genome with sizes, GC and repeat contents similar to those of DSB sites [10] (Figure 1b). ChIP-seq and DNase-seq peaks in NHEK cells as obtained from the ENCODE project were mapped corresponding to DSB and non-DSB sites [31]. We also mapped p63 ChIP-seq peaks from keratinocyte cells (HKC) [15]. We further searched for potential protein binding sites at DSB and non-DSB sites using JASPAR 2016 database motif position weight matrices [20], and predicted DNA shape at DSB and non-DSB sites using Monte Carlo simulations [6]. In the third step, a random forest classifier was built to discriminate between DSB sites and non-DSB sites based on epigenome marks or DNA (Figure 1c). Random forest variable importances were used to estimate the predictive importance of a feature. We also compared random forest predictions with another popular method, lasso logistic regression [32]. Using lasso regression, we assessed the positive, negative or null contribution of a feature

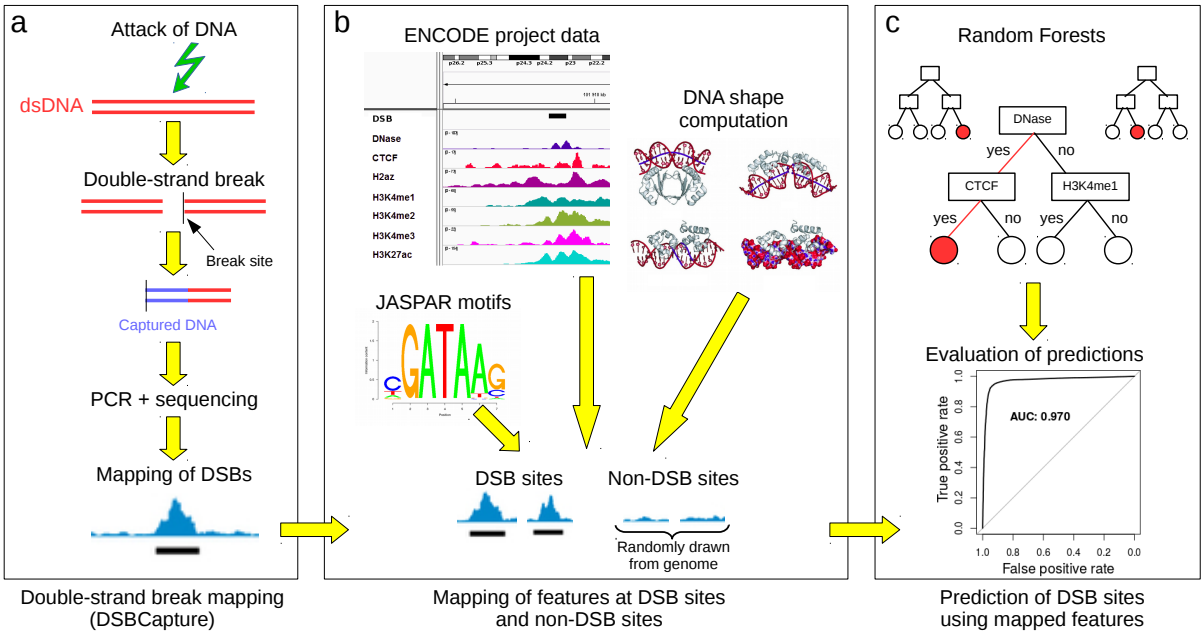


Figure 1. Double-strand break (DSB) prediction using epigenome mark or DNA. The prediction approach consisted in three steps. a) Mapping of DSBCapture sequencing data and DSB peak calling. b) Mapping of features at DSB and non-DSB sites. Features included epigenomic and chromatin data from ENCODE project, DNA motifs from JASPAR database and DNA shape predictions. c) Prediction of DSB sites using features.

to DSBs. We then split the DSB dataset into a training set to learn model parameters by cross-validation, and into a testing set to compute receiver operating characteristic (ROC) curve and the area under the curve (AUC) for prediction accuracy evaluation.

2.2 Double-strand breaks are enriched with epigenome marks and DNA motifs

We first sought to comprehensively assess the link between DSBs and epigenome marks or DNA motifs. As previously shown [17,30], several epigenomic and chromatin marks colocalized at double-strand breaks (Figure 2a). Among the most enriched marks were DNase I hypersensitive sites, H3H4 methylations and CTCF (Figure 2b). For instance, 91% of DSBs colocalized to a DNase site, whereas this percentage dropped to 11% for non-DSB regions. This corresponded to an odds ratio (OR) of 89.3. Similarly a high enrichment was found for H3K4me2 (74% versus 11%; $OR = 22.4$) and for the insulator protein CTCF (25% versus 2%; $OR = 19$), which may involve its interactions with the insulator-related cofactor cohesin that has been shown to protect genes from DSBs [5]. As such, DSBs mostly localized within open and active regions that were often implicated in long-range contacts [27]. Interestingly, DSBs also colocalized with tumor protein p63 binding (19.4% versus 1%; $OR = 23.8$), a member of the p53 gene family involved in DNA repair [19,37]. In addition, we could distinguish DNase and CTCF sites that were enriched at the center of DSBs, from histone marks that were found at the edges of DSB sites (Figure 2c). Therefore the strong enrichments of epigenomic and chromatin marks at DSB sites suggested that DSB regions could be accurately predicted using available ChIP-seq and DNase-seq data from public databases including ENCODE and Roadmap Epigenomic.

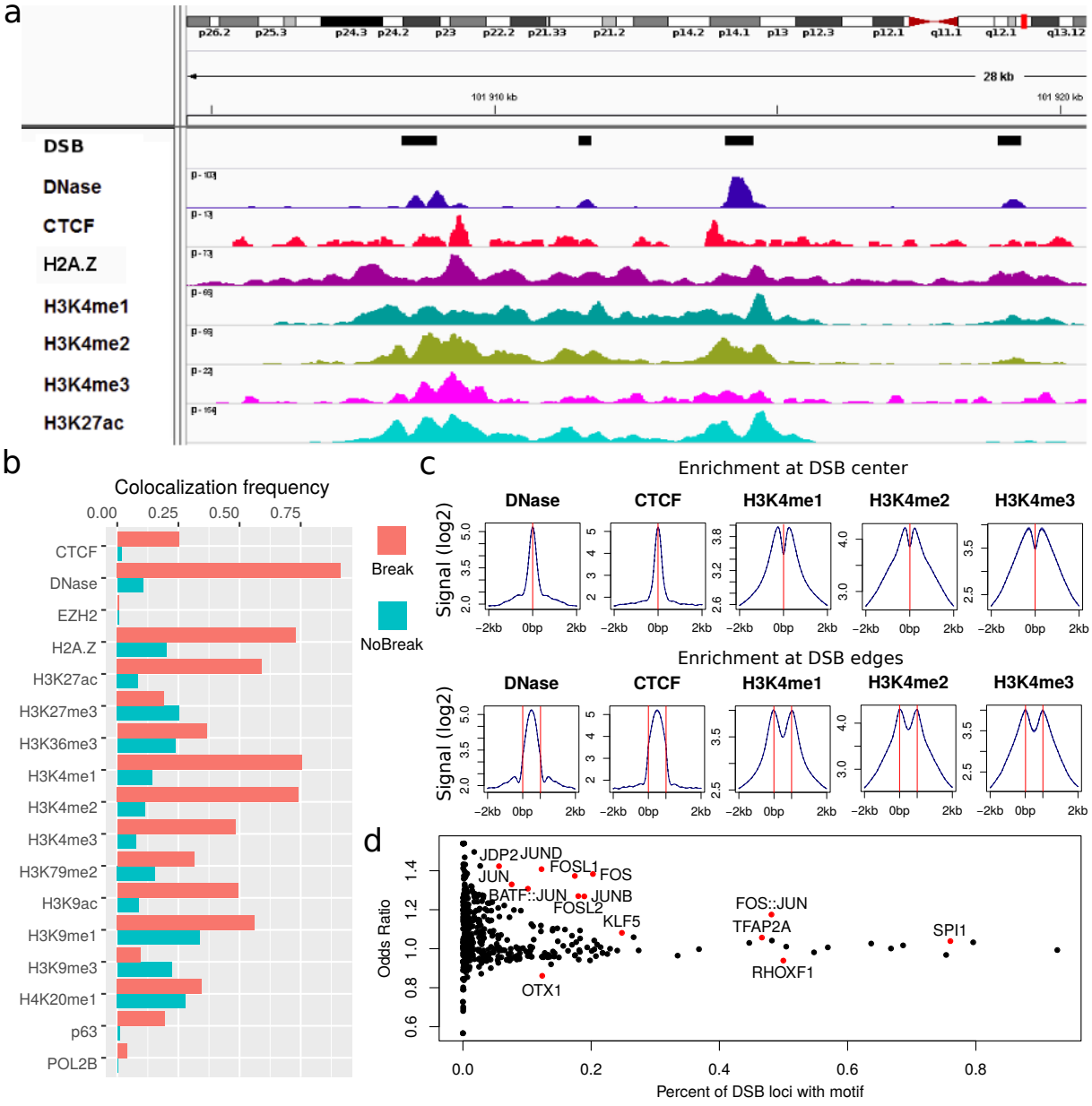


Figure 2. Epigenomic, chromatin and DNA motif profiles of double-strand breaks (DSBs). a) A genome browser view of DSBs with histone marks, chromatin openess (DNase-seq) and DNA binding proteins. b) Colocalization frequencies of epigenomic marks and DNA binding proteins at DSB sites, as compared to non-DSB sites. c) Average profiles of epigenomic marks and DNA binding proteins at DSB sites. d) Enrichment of DNA motifs at DSB sites, as measured by the odds ratio and the percent of DSB loci with motif.

Previous enrichment analyses of DNA-binding proteins were limited by available ChIP-seq data. Hence we sought which DNA motifs may be enriched at DSB sites as a way to obtain a more comprehensive

list of candidate DNA-binding proteins. Over the 434 available motifs from JASPAR database, 134 were significantly enriched ($p < 0.05$, Bonferroni correction), indicating that DSBs were associated with a large number of protein binding sites (Figure 2d). Among the most enriched and frequent motifs, we identified numerous motifs specifically recognized by protein cofactors of the transcription factor complex AP-1 whose activity has been shown to be induced by genotoxic agents. This included JUND ($OR = 1.40$, 12% of DSBs), JUNB ($OR = 1.27$, 19% of DSBs), the heterodimer BATF::JUN ($OR = 1.31$, 10% of DSBs), and also FOS ($OR = 1.37$, 20% of DSBs), FOSL1 ($OR = 1.37$, 17% of DSBs) and FOSL2 ($OR = 1.27$, 18% of DSBs). Among the most enriched but less frequent motifs, we found as expected CTCF ($OR = 1.54$, 1.7% of DSBs), as well as the members of the tumor protein family p53, *i.e.* p53 itself ($OR = 1.54$, 0.2% of DSBs), p63 ($OR = 1.49$, 0.3% of DSBs) and p73 ($OR = 1.54$, 0.1% of DSBs), whose cofactors are specifically involved in the response to DNA damage [19, 37]. Such enrichments of DNA motifs at DSB sites therefore supported the view that DNA sequence could already predict some of the DSBs encountered.

2.3 Prediction using epigenomic and chromatin data

Given the strong link between DSBs and epigenomic and chromatin marks, we sought to build a classifier to discriminate between DSB sites from non-DSB sites based on the presence/absence of such marks. For this purpose, we used random forests that represent very efficient classifiers to predict a feature, and that can capture non-linear and complex interaction effects [3]. Using this classifier, we obtained excellent predictions of DSBs based on the epigenomic and chromatin marks available (AUC=0.970; Figure 3a). We could also compute the "variable importance" (VI) that reflects the importance of a mark as a predictor (Figure 3b). Among the marks, DNase showed the highest variable importance ($VI = 0.180$), reflecting known higher chromatin accessibility after DNA damage [28] or the involvement of chromatin remodeling complexes in processing of DSBs [12]. Other good predictors were CTCF ($VI = 0.042$), p63 ($VI = 0.031$), H3K4me1 ($VI = 0.028$), H3K4me2 ($VI = 0.019$), H3K4me3 ($VI = 0.012$) and H3K27ac ($VI = 0.010$), highlighting the roles of active chromatin, but also long-range contacts and DNA damage response in predicting DSB sites. A drawback of variable importance lies in its inability to distinguish between the positive or negative contribution of the predictive mark on DSBs. For this reason, we also used lasso logistic regression to predict DSBs [32]. With this second model, we obtained excellent predictions, although slightly less accurate (AUC=0.967, Supplemental Figure 1). From lasso regression, we could assess the positive or negative contributions of the predictive marks using beta coefficients (Figure 3c). We identified positive predictive contributions of DNase, CTCF, p63, H3K4me1 and H3K4me2 marks, as previously revealed by enrichment analysis. We also uncovered negative predictive contributions of H3K9ac, H3K36me3 and H3K79me2. In agreement, H3K9ac was shown to be rapidly and reversibly reduced in response to DNA damage [33]. Moreover, H3K36me3 may negatively impede on DSBs by restricting chromatin accessibility through nucleosome positioning [18] or more directly by favoring the repair of DSBs [26].

We next sought to build a classifier using only one or two epigenomic marks, because it could allow to predict DSB sites even for cells for which only a few data may be available. We found that DNase I sites alone were sufficient to achieve good prediction accuracy (AUC=0.919), whereas H3K4me2 was not sufficient (AUC=0.816). Combinations of DNase with H2A.Z or H3K4me1 yielded very accurate predictions (AUC=0.952 and AUC=0.951, resp.), close to the model including all marks. All results demonstrated that DSBs can be accurately predicted at less than 1 kb resolution using just a few data.

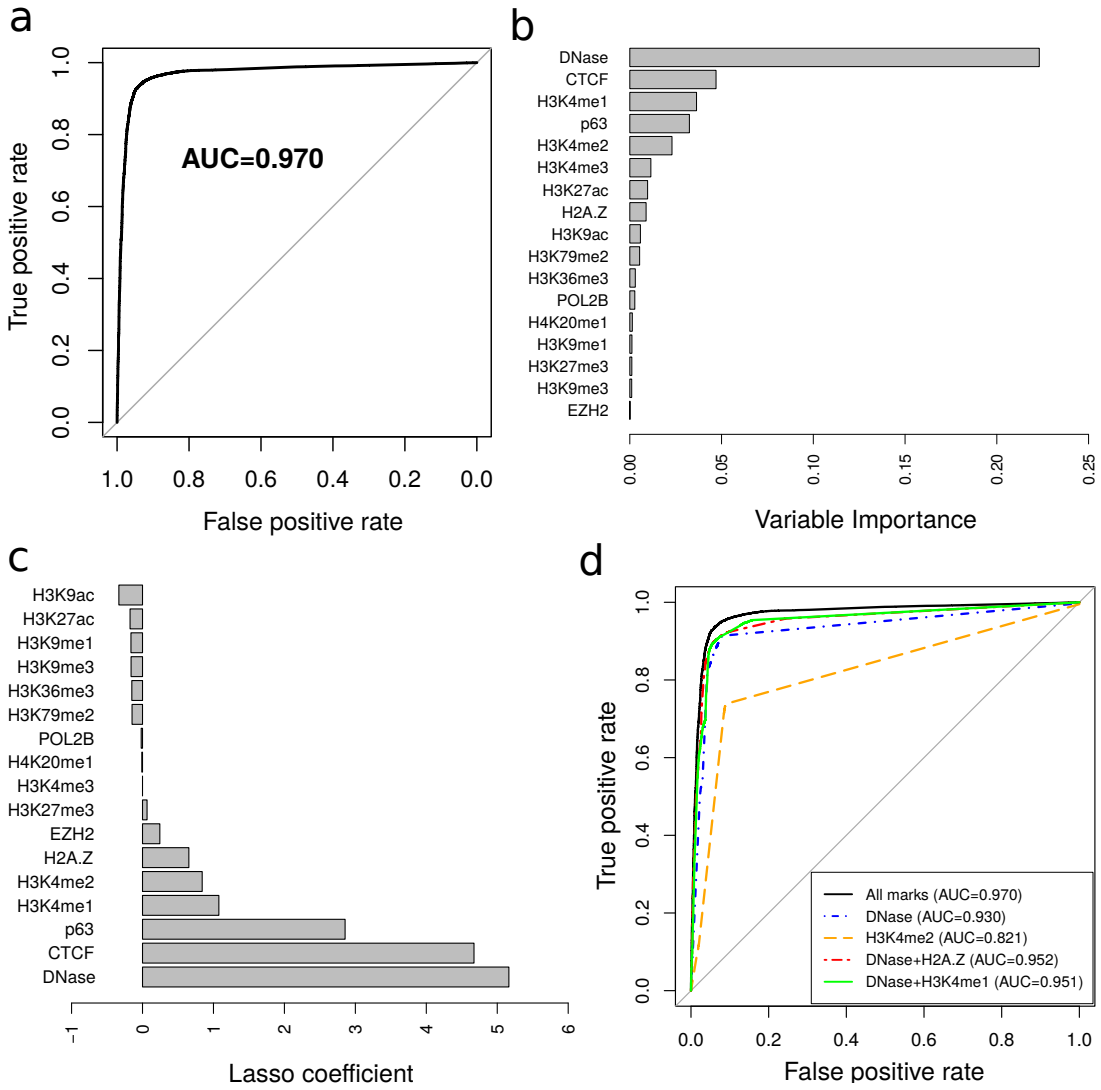


Figure 3. Prediction of double-strand breaks (DSBs) using epigenomic marks and random forests. a) Receiver operating characteristic (ROC) curve for the prediction of DSBs. Area under the curve (AUC) is plotted. b) Variable importances of epigenomic marks. c) Lasso logistic regression coefficients. d) Different predictive models including all marks, DNase only, H3K4me2 only, DNase+H2A.Z, or DNase+H3K4me1.

2.4 Model predictions outperformed BLESS experiment and were validated using independent dataset

We then compared previous DSB predictions with DSBs identified by BLESS experiments [8, 17]. We also included in the comparison DSBCapture DSBs as gold standard. We first looked at predicted DSB sites surrounding the two genes MYC and MAP2K9 (Figure 4a). For MYC, random forests correctly identified the 4 DSBs that were detected by DSBCapture, but erroneously predicted one DSB (yellow circle), whereas BLESS only identified one DSB out of four. For MAP2K3, random forests successfully predicted all DSBs detected by DSBCapture, whereas BLESS only identified three DSBs out of 11. We

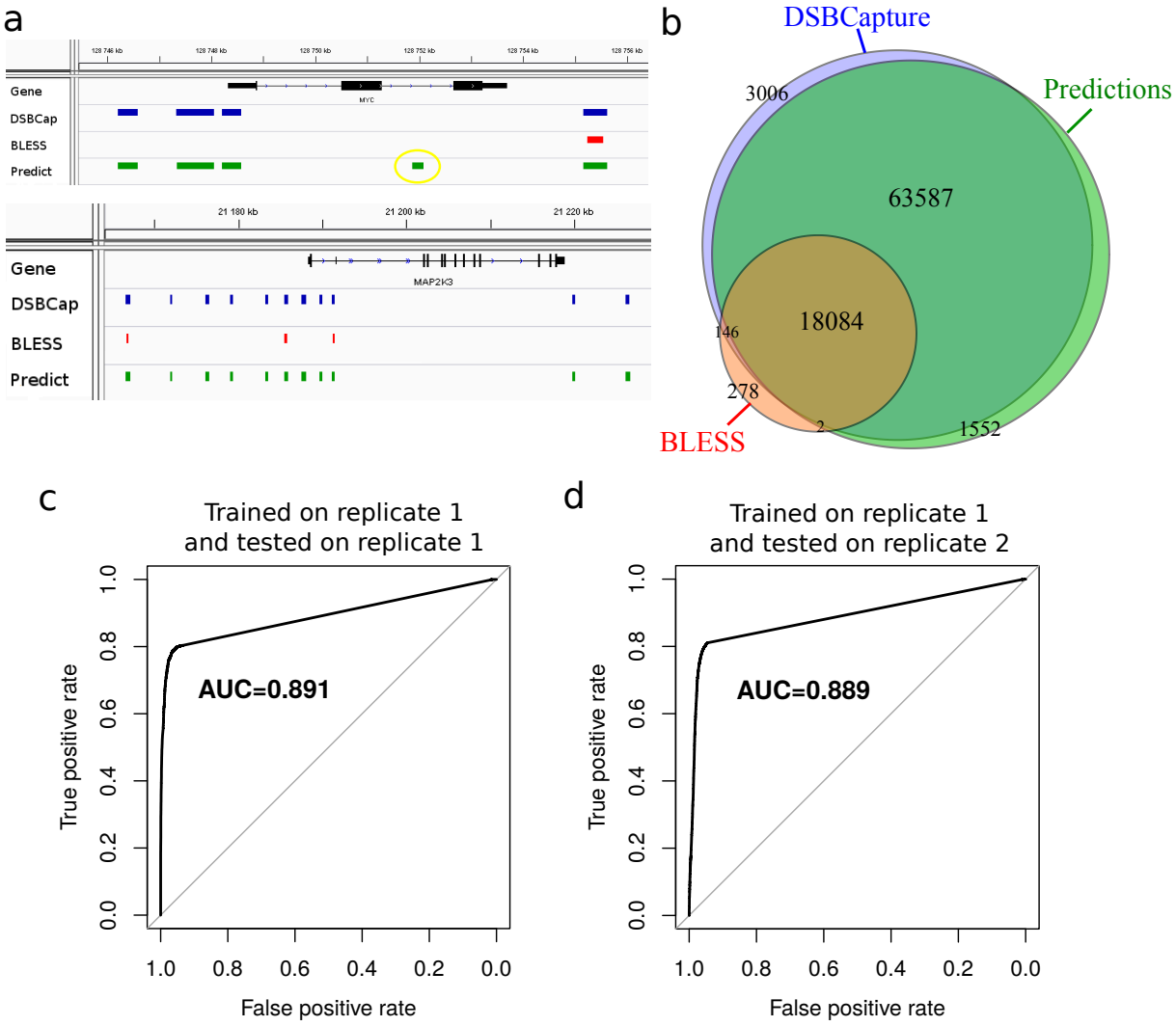


Figure 4. Comparison of predicted and BLESS double-strand breaks (DSBs) and validation with an independent dataset. a) Comparison for the MYC and MAP2K9 genes. b) Venn diagram illustrating the overlaps between DSBCapture (gold standard), random forest predictions and BLESS DSBs. c) Receiver operating characteristic (ROC) curve for the prediction of DSBs trained on replicate 1 and tested on same replicate. Area under the curve (AUC) is plotted. d) ROC curve for the prediction of DSBs trained on replicate 1 and tested on replicate 2.

then compared predictions with BLESS at the genome-wide level (Figure 4b). We observed that random forests correctly predicted 18084 out of 18510 DSB sites (97.70%) found by BLESS, while it also successfully identified additional 63587 out of 66593 DSB sites (95.48%) found by DSBCapture that were not detected by BLESS. The model only misclassified 1552 sites as DSBs. Such comparisons thus revealed the better sensitivity of random forest predictions in detecting DSBs compared to BLESS experiments.

In the previous subsection, we evaluated the accuracy of model predictions using a testing dataset that was from the same data as the training data (DSBs that overlapped between two replicates were split

into a training and a testing datasets). Here we assessed model predictions by training random forests on a first biological replicate and by testing prediction accuracy on a second biological replicate. For this purpose, we used the two available DSBCapture biological replicates [17]. Accordingly, we used ENCODE epigenomics data for which two biological replicates were available: DNase, CTCF, H3K4me3, H3K27me3 and H3K36me3. The first (resp. second) replicates of ENCODE data were associated with the first (resp. second) DSBCapture replicate. Using only those 5 DNase-seq and ChIP-seq data, the model learned with the first replicate achieved accurate predictions on tested data from the first replicate (AUC=0.891) (Figure 4c). It is noteworthy that the observed lower accuracy compared to previous subsection (Figures 3a and d) can be explained by the small number of epigenomic data, and the lower reliability of DSBs identified using only one DSBCapture replicate. To validate the model on an independent dataset, we predicted DSBs from the second replicate using the model trained on the first replicate together with DNase-seq and ChIP-seq data of the second replicate. We obtained accurate predictions (AUC=0.889) close to the one obtained for the first replicate (Figure 4d). These accurate predictions demonstrated that using a classifier trained with epigenome and chromatin data represented a reliable strategy for predicting DSBs.

2.5 Prediction from DNA motifs and shape

There is growing evidence supporting the importance of the *cis*-regulatory code in shaping the epigenome [36]. Moreover, we previously identified a large number of DNA motifs that were enriched or depleted at DSB sites, suggesting a "*cis*-DNA repairing" code. Hence, we explored the possibility to predict DSBs based on the occurrence of DNA motifs from JASPAR database. We built a random forest classifier using 434 available motifs and obtained good prediction accuracy (AUC=0.827; Figure 5a). Several motifs from the transcription factor complex AP-1 represented good predictors such as FOS::JUN ($VI = 0.016$) and FOS ($VI = 0.009$) (Figure 5b), which were previously shown to be enriched at DSB sites. We also uncovered TFAP2A motif as good predictor ($VI = 0.011$), corresponding to a protein that has not been linked to DNA repair yet. Using lasso regression, we improved previous predictions (AUC=0.839; Figure 5a). Based on lasso regression, we found that the CTCF motif presented the highest beta coefficient ($\beta = 3.22$), corresponding to an odds ratio $OR = 25$ (Figure 5c), supporting recent evidence showing that long-range contacts are involved in DNA repair [2, 30]. Furthermore we found many motifs recognized by proteins that participate to DNA damage and repair. For instance, all tumor proteins p53, p63 and p73 motifs showed high coefficients ($\beta > 2.03$, $OR > 7.6$). Interestingly, we also found motifs recognized by factors highly related to DSB pathways such as those involved in heavy metal response (MTF-1: $\beta = 2.08$, $OR = 8$), in oxidative stress response (NRF1: $\beta = 0.93$, $OR = 2.53$; REST: $\beta = 1.75$, $OR = 5.75$), in endoplasmic reticulum stress (ATF4: $\beta = 0.97$, $OR = 2.64$) and in estrogen-induced DNA damage (ESR1: $\beta = 0.88$, $OR = 2.41$). Many of the abovementionned proteins were actually shown to interact with each other. For instance, NRF1 associates with Jun proteins of AP-1 complex [35]. ESR1 associates with AP-1/JUN and FOS to mediate estrogen element response (ERE)-independent signaling [16], which may thus participate to coordinate multiple functions along with the processing of DSBs.

DNA shape was recently shown to predict transcription factor binding sites and gene expression [21, 25]. We thus assessed if DNA shape could similarly serve to predict DSBs together with motifs. For this purpose, we predicted four DNA shape features using simulations: minor groove width (MGW), propeller twist (ProT), roll (Roll) and helix twist (HelT) of DSB sites at base resolution. From each feature, we computed 12 predictors including quantiles (0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100%) and the variance to describe the distribution of the feature within a DSB site. We used the resulting 48 variables combined with motif occurences to predict DSBs with random forests and obtained better accurary ($AUC = 0.838$) compared to using motifs alone ($AUC = 0.827$; Figure 5d). Among the DNA shape variables, ProT median and MGW variance presented the highest variable importances ($VI = 0.01$ and $VI = 0.01$, resp.). Using lasso regression, we also obtained better predictions

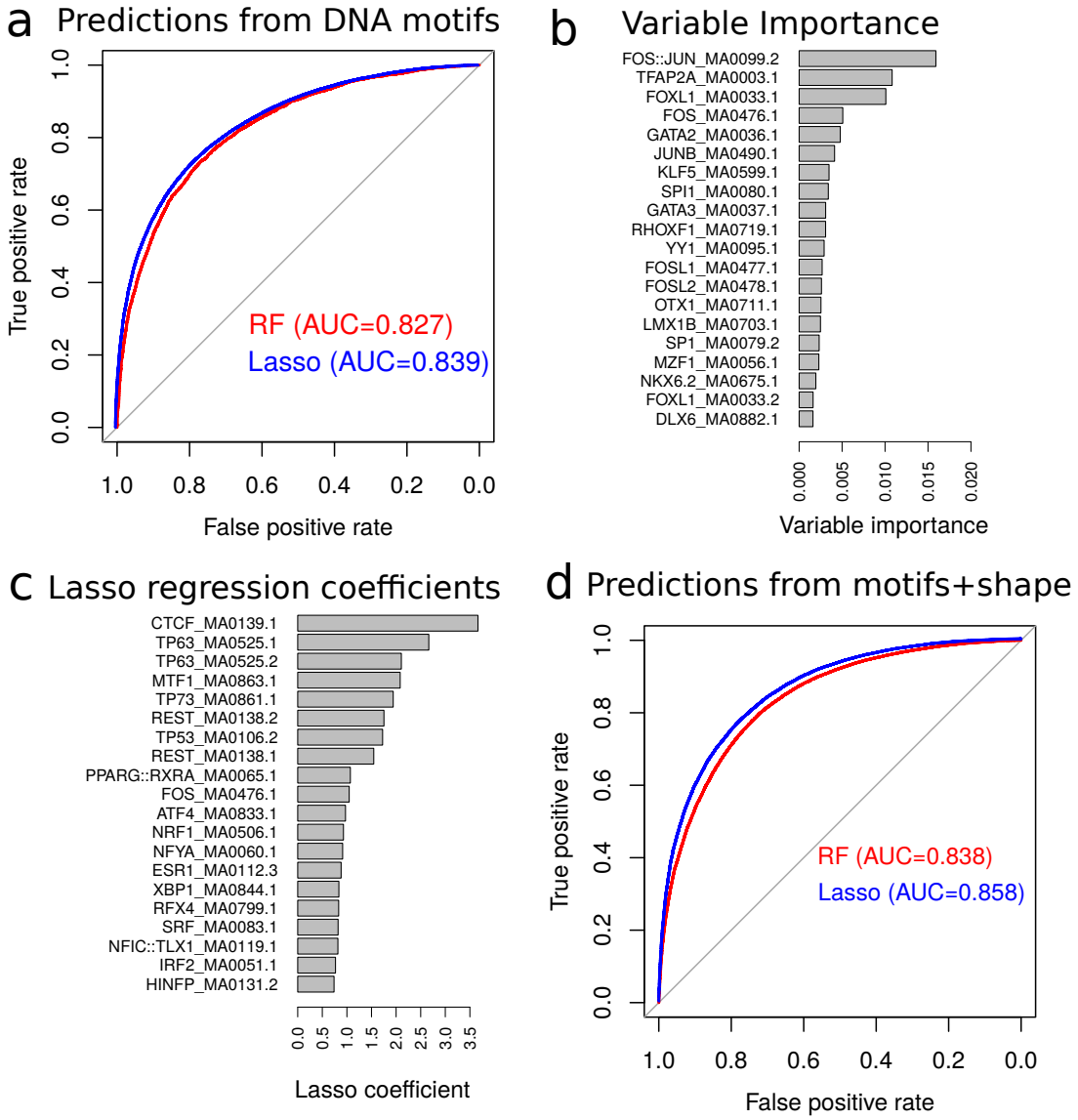


Figure 5. Prediction of double-strand breaks (DSBs) using DNA motifs and shape. a) Receiver operating characteristic (ROC) curve for the DSB predictions using DNA motifs from JASPAR database. Random forest (RF) and lasso logistic regression (Lasso) were compared. b) The 20 highest DNA motif variable importances. c) The 20 highest DNA motif lasso coefficients. d) ROC curve for the DSB predictions using DNA motifs with DNA shape.

($AUC = 0.858$), compared to using motifs only ($AUC = 0.839$) (Figure 5d). These results reflected the importance of DNA shape in determining DSB sites, in agreement with studies showing that narrow minor grooves (created by either sequence context or DNA bending) limit access by reactive oxygen species [4].

3 Conclusion

Double-strand breaks represent a major threat to the cell, and they are associated with cancer development. Here we show, for the first time, that such DSBs can be computationally predicted using public epigenomic data, even when the availability of data is limited (e.g. DNase I and H3K4me1). By using state-of-the-art computational models, we achieve excellent prediction accuracy, paving the way for a better understanding of DSB formation depending on developmental stage or cell-type specific epigenetic marks. In addition, our work represents the first step toward unveiling the "*cis*-DNA repairing" code underlying DSBs, which is composed of numerous DNA motifs for binding of key regulators of DNA repair, and could guide further locus- specific genome editing.

4 Materials and Methods

4.1 Double-strand breaks

We used double-strand breaks mapped by DSBCapture in human epidermal keratinocytes (NHEK) cells [17]. DSBCapture peaks were called using MACS 2.1.0 on human genome assembly hg19 (https://github.com-/taoliu/MACS). The peaks obtained from two biological replicates were intersected to yield more reliable DSB sites used for model predictions.

4.2 ChIP-seq and DNase-seq data

We used ChIP-seq (CTCF, POL2B, EZH2, H3K4me1/me2/me3, H3K9me1/me3/ac, H3K27me3/ac, H3K36me3, H4K20me1) and DNase-seq data for NHEK cells from the ENCODE project [31]. We also used p63 ChIP-seq of keratinocyte cells (HKC) from Kouwenhoven *et al.* [15].

4.3 DNA motifs

We used transcription factor binding site (TFBS) motif position frequency matrices from the JASPAR database (http://jaspar.genereg.net/). We called transcription factor binding sites over the human genome using the position weight matrices and a minimum matching score of 80%.

4.4 DNA shape

We predicted four DNA shape features using Monte Carlo simulations: minor groove width (MGW) and propeller twist (ProT) at base pair (bp) resolution and values of roll (Roll) and helix twist (HelT) at base pair step resolution using R package DNAshapeR (https://bioconductor.org/packages/release/bioc/html/DNAshapeR.html).

4.5 Random forest and lasso regression

We used R package ranger (https://cran.r-project.org/web/packages/ranger/) to efficiently compute random forest classification [3]. We used the default package parameters: $num.trees = 500$ and $mtry$ is the square root of the number variables. Variable importance was computed using the mean decrease in accuracy in the out-of-bag sample. To discriminate between DSB and non-DSB sites, we randomly selected genomic sequences that matched sizes, GC and repeat contents of DSB sites using R package gkmSVM (https://cran.r-project.org/web/packages/gkmSVM/). To learn the model, we mapped epigenomic data, DNA motifs and DNA shape as follows. For epigenomic data including ChIP-seq and DNase-seq data, we used peak genomic coordinates of a feature (for instance CTCF binding sites) and considered the

presence ($x = 1$) or absence ($x = 0$) of the corresponding feature at the DSB site. If a feature peak only overlapped 60% of the DSB site, then $x = 0.6$. For DNA motifs, we computed the number of motif occurrence within DSB and non-DSB sites. For DNA shape, we computed 4 features including MGW, ProT, Roll and HelT of DSB sites at base resolution. For each DNA shape feature, we then computed 12 predictors including quantiles (0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100%) and the variance to describe the distribution of the feature within a DSB site. The DSB data were next split into two sets: the training set used for learning the model and a test set used for assessing prediction accuracy. We also used R package glmnet (https://cran.r-project.org/web/packages/glmnet/index.html) to compute lasso logistic regression with cross-validation. To estimate prediction accuracy of random forests and lasso regression, we computed the receiver operating characteristic (ROC) curve and the area under the curve (AUC).

Acknowledgements

The authors are grateful to the Balasubramanian lab (Babraham Institute, UK) for data availability and for help in processing them. The authors are also grateful to the genotoul bioinformatics platform Toulouse Midi-Pyrenees for providing computing resources.

References

[1] F. Baudat, J. Buard, C. Grey, A. Fledel-Alon, C. Ober, M. Przeworski, G. Coop, and B. de Massy. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, 327(5967):836–840, 2010.

[2] Simon Bekker-Jensen and Niels Mailand. Assembly and function of DNA double-strand break repair foci in mammalian cells. *DNA Repair*, 9(12):1219–1228, 2010.

[3] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001.

[4] Wendy J. Cannan and David S. Pederson. Mechanisms and consequences of double-strand DNA break formation in chromatin. *Journal of Cellular Physiology*, 231(1):3–14, 2016.

[5] Pierre Caron, Francois Aymard, Jason S. Iacovoni, Sbastien Briois, Yvan Canitrot, Beatrix Bugler, Laurent Massip, Ana Losada, and Galle Legube. Cohesin protects genes against γ-H2AX induced by DNA double-strand breaks. *PLOS Genetics*, 8(1):1–17, 01 2012.

[6] Tsu-Pei Chiu, Federico Comoglio, Tianyin Zhou, Lin Yang, Renato Paro, and Remo Rohs. DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics*, 32(8):1211–1213, April 2016.

[7] The Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, Viren Amin, John W. Whitaker, Matthew D. Schultz, Lucas D. Ward, Abhishek Sarkar, Gerald Quon, Richard S. Sandstrom, Matthew L. Eaton, Yi-Chieh Wu, Andreas R. Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R. Alan Harris, Noam Shoresh, Charles B. Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R. David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J. Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K. Canfield, R. Scott Hansen, Rajinder Kaul, Peter J. Sabo, Mukul S. Bansal, Annaick Carles, Jesse R. Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R. Mercer, Shane J. Neph, Vitor Onuchic, Paz

Polak, Nisha Rajagopal, Pradipta Ray, Richard C. Sallari, Kyle T. Siebenthall, Nicholas A. Sinnott-Armstrong, Michael Stevens, Robert E. Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E. Beaudet, Laurie A. Boyer, Philip L. De Jager, Peggy J. Farnham, Susan J. Fisher, David Haussler, Steven J. M. Jones, Wei Li, Marco A. Marra, Michael T. McManus, Shamil Sunyaev, James A. Thomson, Thea D. Tlsty, Li-Huei Tsai, Wei Wang, Robert A. Waterland, Michael Q. Zhang, Lisa H. Chadwick, Bradley E. Bernstein, Joseph F. Costello, Joseph R. Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A. Stamatoyannopoulos, Ting Wang, Manolis Kellis, and Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, February 2015.

[8] Nicola Crosetto, Abhishek Mitra, Maria J. Silva, Magda Bienko, Norbert Dojer, Qi Wang, Elif Karaca, Roberto Chiarle, Magdalena Skrzypczak, Krzysztof Ginalski, Philippe Pasero, Maga Rowicka, and Ivan Dikic. Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nature Methods*, 10(4):361–365, April 2013.

[9] Jason Ernst and Manolis Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9(3):215–216, February 2012.

[10] Mahmoud Ghandi, Morteza Mohammad-Noori, Narges Ghareghani, Dongwon Lee, Levi Garraway, and Michael A. Beer. gkmSVM: an R package for gapped-kmer SVM. *Bioinformatics*, 32(14):2205, 2016.

[11] Katsuhiko Hayashi, Kayo Yoshida, and Yasuhisa Matsui. A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. *Nature*, 438(7066):374–378, November 2005.

[12] Karine Jacquet, Amlie Fradet-Turcotte, Nikita Avvakumov, Jean-Philippe Lambert, Cline Roques, RajK. Pandita, Eric Paquet, Pauline Herst, Anne-Claude Gingras, TejK. Pandita, Galle Legube, Yannick Doyon, Daniel Durocher, and Jacques Ct. The TIP60 complex regulates bivalent chromatin recognition by 53BP1 through direct H4K20me binding and H2AK15 acetylation. *Molecular Cell*, 62(3):409–421, 2016.

[13] Andrea Kinner, Wenqi Wu, Christian Staudt, and George Iliakis. γ-H2AX in recognition and signaling of DNA double-strand breaks in the context of chromatin. *Nucleic Acids Research*, 36(17):5678, 2008.

[14] Dimitrios Kleftogiannis, Panos Kalnis, and Vladimir B. Bajic. DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Research*, 43(1):e6, 2014.

[15] Evelyn N Kouwenhoven, Martin Oti, Hanna Niehues, Simon J van Heeringen, Joost Schalkwijk, Hendrik G Stunnenberg, Hans van Bokhoven, and Huiqing Zhou. Transcription factor p63 bookmarks and regulates dynamic enhancers during epidermal differentiation. *EMBO reports*, 16(7):863–878, 2015.

[16] P. J. Kushner, D. A. Agard, G. L. Greene, T. S. Scanlan, A. K. Shiau, R. M. Uht, and P. Webb. Estrogen receptor pathways to AP-1. *The Journal of steroid biochemistry and molecular biology*, 74(5):311–317, November 2000.

[17] Stefanie V. Lensing, Giovanni Marsico, Robert Hansel-Hertsch, Enid Y. Lam, David Tannahill, and Shankar Balasubramanian. DSBCapture: in situ capture and sequencing of DNA breaks. *Nature Methods*, 13(10):855–857, August 2016.

[18] Priscillia Lhoumaud, Magali Hennion, Adrien Gamot, Suresh Cuddapah, Sophie Queille, Jun Liang, Gael Micas, Pauline Morillon, Serge Urbach, Olivier Bouchez, Dany Severac, Eldon Emberly, Keji Zhao, and Olivier Cuvier. Insulators recruit histone methyltransferase dmes4 to regulate chromatin of flanking genes. *The EMBO Journal*, 33(14):1599–1613, 2014.

[19] Yu-Li Lin, Shomit Sengupta, Katherine Gurdziel, George W. Bell, Tyler Jacks, and Elsa R. Flores. p63 and p73 transcriptionally regulate genes involved in DNA repair. *PLOS Genetics*, 5(10):1–13, 10 2009.

[20] Anthony Mathelier, Oriol Fornes, David J. Arenillas, Chih-yu Chen, Grgoire Denay, Jessica Lee, Wenqiang Shi, Casper Shyr, Ge Tan, Rebecca Worsley-Hunt, Allen W. Zhang, Franois Parcy, Boris Lenhard, Albin Sandelin, and Wyeth W. Wasserman. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 44(D1):D110, 2015.

[21] Anthony Mathelier, Beibei Xin, Tsu-Pei Chiu, Lin Yang, Remo Rohs, and WyethW. Wasserman. DNA shape features improve transcription factor binding site predictions invivo. *Cell Systems*, 3(3):278–286.e4, 2016.

[22] Peter J. McKinnon and Keith W. Caldecott. DNA strand break repair and human genetic disease. *Annual Review of Genomics and Human Genetics*, 8(1):37–55, 2007.

[23] Anuja Mehta and James E. Haber. Sources of DNA double-strand breaks and models of recombinational DNA repair. *Cold Spring Harbor Perspectives in Biology*, 6(9), 2014.

[24] Simon Myers, Rory Bowden, Afidalina Tumian, Ronald E. Bontrop, Colin Freeman, Tammie S. MacFie, Gil McVean, and Peter Donnelly. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science*, 327(5967):876–879, 2010.

[25] Pei-Chen Peng and Saurabh Sinha. Quantitative modeling of gene expression using DNA shape features of binding sites. *Nucleic Acids Research*, 44(13):e120, July 2016.

[26] SophiaX. Pfister, Sara Ahrabi, Lykourgos-Panagiotis Zalmas, Sovan Sarkar, Franois Aymard, CsandZ. Bachrati, Thomas Helleday, Galle Legube, NicholasB. LaThangue, AndrewC.G. Porter, and TimothyC. Humphrey. SETD2-dependent histone H3K36 trimethylation is required for homologous recombination repair and genome stability. *Cell Reports*, 7(6):2006 – 2018, 2014.

[27] Jennifer E. Phillips-Cremins, Michael E. G. Sauria, Amartya Sanyal, Tatiana I. Gerasimova, Bryan R. Lajoie, Joshua S. K. Bell, Chin-Tong Ong, Tracy A. Hookway, Changying Guo, Yuhua Sun, Michael J. Bland, William Wagstaff, Stephen Dalton, Todd C. McDevitt, Ranjan Sen, Job Dekker, James Taylor, and Victor G. Corces. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, 153(6):1281–1295, June 2013.

[28] Brendan D. Price and Alan D. D'Andrea. Chromatin Remodeling at DNA Double-Strand Breaks. *Cell*, 152(6):1344–1354, March 2013.

[29] Sean D. Taverna, Haitao Li, Alexander J. Ruthenburg, C. David Allis, and Dinshaw J. Patel. How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nature Structural & Molecular Biology*, 14(11):1025–1040, November 2007.

[30] Nickolai A. Tchurikov, Daria M. Fedoseeva, Dmitri V. Sosin, Anastasia V. Snezhkina, Nataliya V. Melnikova, Anna V. Kudryavtseva, Yuri V. Kravatsky, and Olga V. Kretova. Hot spots of DNA double-strand breaks and genomic contacts of human rDNA units are involved in epigenetic regulation. *Journal of Molecular Cell Biology*, 7(4):366–382, October 2014.

[31] The ENCODE Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.

[32] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, January 1996.

[33] Jorrit V Tjeertes, Kyle M Miller, and Stephen P Jackson. Screen for DNA-damage-responsive histone modifications identifies H3K9Ac and H3K56Ac in human cells. *The EMBO Journal*, 28(13):1878–1889, 2009.

[34] Shengdar Q Tsai, Zongli Zheng, Nhu T Nguyen, Matthew Liebers, Ved V Topkar, Vishal Thapar, Nicolas Wyvekens, Cyd Khayter, A John Lafrate, Long P Le, Martin J Aryee, and J Keith Joung. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnology*, 33:187197, December 2015.

[35] Radjendirane Venugopal and Anil K. Jaiswal. Nrf2 and Nrf1 in association with Jun proteins regulate antioxidant response element-mediated expression and coordinated induction of genes encoding detoxifying enzymes. *Oncogene*, 17(24):3145–3156, December 1998.

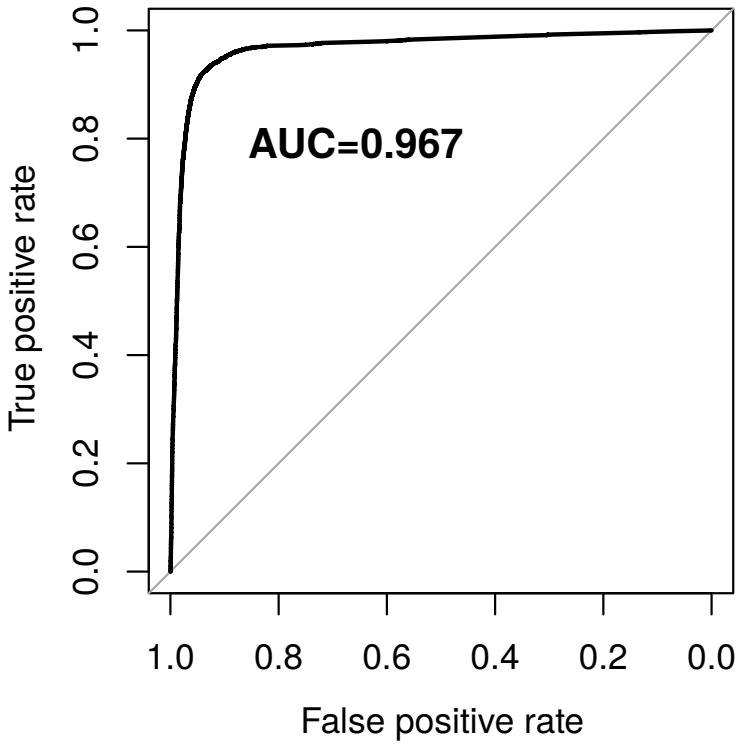
[36] John W. Whitaker, Zhao Chen, and Wei Wang. Predicting the human epigenome from DNA motifs. *Nature Methods*, 12(3):265–272, March 2015.

[37] Ashley B. Williams and Björn Schumacher. p53 in the DNA-Damage-Repair Process. *Cold Spring Harbor perspectives in medicine*, 6(5), 2016.

[38] Jian Zhou and Olga G. Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, August 2015.

5 Supplemental Files

Supplemental Figure 1



Prediction accuracy of double-strand breaks (DSBs) using epigenomic marks and lasso logistic regression. Receiver operating characteristic (ROC) curve of the prediction of DSBs. Area under the curve (AUC) is plotted.