bioRxiv preprint doi: https://doi.org/10.1101/796383; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Computational identification of cell-specific variable regions in ChIP-seq data

Tommaso Andreani[1,2], Steffen Albrecht[1], Jean-Fred Fontaine[1] and Miguel A. Andrade-Navarro[1]*

[1] Faculty of Biology, Johannes Gutenberg University of Mainz, 55128 Mainz, Germany
[2] Institute of Molecular Biology (IMB), 55128 Mainz, Germany
* To whom correspondence should be addressed. Tel: +49-6131-39-21582; Email andrade@uni-mainz.de

ABSTRACT

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is used to identify genome-wide DNA regions bound by proteins. Several sources of variation can affect the reproducibility of a particular ChIP-seq assay, which can lead to a misinterpretation of where the protein under investigation binds to the genome in a particular cell type. Given one ChIP-seq experiment with replicates, binding sites not observed in all the replicates will usually be interpreted as noise and discarded. However, the recent discovery of high-occupancy target (HOT) regions suggests that there are regions where binding of multiple transcription factors can be identified. To investigate these regions, we developed a reproducibility score and a method that identifies cell-specific variable regions in ChIP-seq data by integrating replicated ChIP-seq experiments for multiple protein targets on a particular cell type. Using our method, we found variable regions in human cell lines K562, GM12878, HepG2, MCF-7, and in mouse embryonic stem cells, defined as protein binding regions with non-reproducible results across replicated experiments. These variable-occupancy target (VOT) regions are CG dinucleotide rich, and show enrichment at promoters and R-loops. They overlap significantly with HOT regions, but are not blacklisted regions producing non-specific binding ChIP-seq peaks. Interestingly, among various genomic features, DNA accessibility is a better predictor of VOTs than CpG islands or epigenetic marks. Our method can be useful to point to such regions along the genome in a given cell type of interest, to improve the downstream interpretative analysis before follow up experiments.

bioRxiv preprint doi: https://doi.org/10.1101/796383; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

INTRODUCTION

A series of genome-wide experiments are largely adopted to study biological systems in relation to a given protein. They contribute to our understanding of particular molecular mechanisms at the basis of biological processes such as transcription and development, just to mention a few. In particular, ChIP-seq evaluates the genomic positions bound by a protein in the genome. Standard ChIP-seq experiments typically include replicated measurements in the experimental design in order to have the proper statistical power for the identification of reliable binding sites (or ChIP-seq peaks).

Previous results have indicated for several model organisms such as yeast, *Drosophila* and *Caenorhabditis elegans* the existence of genomic regions that are bound more often with respect to others, even in genomic positions in which a binding site is not expected for the protein under investigation. These regions have been previously characterized and described as “hyper-ChIPable” in yeast (1) and confirmed later in *Drosophila, C. elegans* and mouse and referred as “phantom peaks” (2). Furthermore, other regions defined here as variable regions, have protein binding that tends to variate stochastically and is difficult to interpret because their inconsistency in the reproducibility of the results. Current approaches to analyse ChIP-seq experiments do not report to the users regions that misbehave before downstream interpretative analysis; this might lead to the misinterpretation of the ChIP-seq results in terms of the function associated to the protein under investigation.

Here, we present a method that uses replicated ChIP-seq data for several proteins on the same cell line to detect regions that misbehave in ChIP-seq experiments. We assigned the term variable for a given genomic region if a protein binding site (or ChIP-seq peak) was not consistently detected in several experimental replicates of the same protein and for several independent proteins in a given cell type. These assignments can increase the value of ChIP-seq experiments by categorizing certain peaks as having cell-specific variability. Possible reasons for this variation might be the adoption of variable genomic structures (3), the high expression of a nearby gene (2), the specificity of the antibody used and the conformation of the chromatin during the immunoprecipitation. By finding variable regions, we expect to be able to characterize the origins of this variability and its potential relation to biological processes.

During the last years, the ENCODE consortium (4) addressed the problem of data collection for ChIP-seq experiments as well as other sequencing datasets creating the metadata of all the experiments. This effort is praiseworthy because at the time of reusing specific datasets it is important to know in detail how the data were produced, from which laboratory and according to which experimental criteria. This information allows controlling possible confounding factors in our study that focuses on local variability potentially caused by local genomic structural conformation or activity. Thus, we used

bioRxiv preprint doi: https://doi.org/10.1101/796383; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

data from the ENCODE consortium, and we controlled how the experiments were performed, from which laboratory and the bioinformatics tools used for data handling among other parameters.

In this work, we took advantage of the metadata provided by the ENCODE consortia, as indicated above, to select experiments in a consistent and comparable manner to implement a sliding window approach to classify genomic regions as variable or not.

Our results show that the method can identify variable regions for every cell line tested and that, particularly for the K562 cell line, for which many datasets are currently available, it improves the separation of the samples in a PCA to promote a better downstream interpretative analysis. Method and scripts can be found online in this link: https://github.com/tAndreani/IPVARIABLE.

MATERIAL AND METHODS

Collection of ChIP-seq data

The ENCODE data portal provides comprehensive information about the meta-data of each experiment generated by the ENCODE consortium. We selected experiments according to specific parameters in order to avoid unwanted variability and to maintain consistency on the parameters of the downloaded data. The experiments were selected according to the following criteria: (i) laboratory producing the data as Snyder, (ii) identical untreated isogenic human cell lines (K562, MCF-7, GM12878, and HepG2) and ES-E14 mouse embryonic stem cells (mESCs), (iii) data processed with the standard ENCODE pipeline that uses the optimal IDR threshold as statistical method to obtain the significant peaks (5), (iv) status as released corresponding to a possible usage of the data, (v) the experiments of each biological replicate correspond to a peak file compared with appropriate input control experiment and (vi) peaks significance selected with a false discovery rate (FDR) lower than or equal to 5%. The metadata presented in JSON format was extracted and stored in a relational SQL database (See Supplementary File 1, Fig. S1). For every cell we selected the following targets: for HepG2 we used MAFK, MNT, TBX3 and ZNF24 with two biological replicates; for MCF-7 we used CREB1, CLOCK, NFIB and ZNF512B with two biological replicates; for GM12878 we used BHLHE40, EP300, IKFZ2 and ZNF143 with two biological replicates; for K562 we used ARNT, NCOR1, MNT and ZNF24 with three biological replicates; for ES-14 mESCs we used HCFC1, MAFK, ZC3H11A and ZNF384 with two biological replicates (see Supplementary Table S1 for details).

Reproducibility score implementation

After the identification of suitable experiments, the genome is binned in consecutive segments of 200 base pairs (bp) and the experimental ChIP-seq peaks are mapped to each segment. We formalized the reproducibility and not reproducibility of the segments for a given protein as illustrated in Fig. 1A and as follows:

bioRxiv preprint doi: https://doi.org/10.1101/796383; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

```
Let S be the genomic segments for a given genome;
Let N be the number of replicate ChIP-seq experiments for a given protein;
For each segment in S;
    Let P be the number of peaks detected in the segment;
    Reproducibility score = NA if P = 0;
    Else Reproducibility score = 1 if the segment itself or one of its neighbours* has P=N;
    Otherwise Reproducibility score = 0
* Neighbours are all consecutive segments with P > 0
```

In the following paragraph, we explain the procedure described by the pseudocode above in words. For our study, segments of the genome are defined considering a window size of 200 base pairs, N represents the number of replicates for each protein under investigation in a given cell type, and P is the number of replicated ChIP-seq peaks detected in a genomic segment (the signal). Consecutive segments without any signal (P=0; no peaks) are assigned with a NA. Consecutive segments in between two NA segments with a signal P reaching as a maximum value N are considered as reproducible regions and assigned a value of 1. On the contrary, consecutive segments in between two NA segments reaching a maximum value lower than N are considered as variable regions and assigned a value of 0 (Fig. 1A). The results of each protein under investigation are aggregated in a Reproducibility Score Matrix (RSM) (Fig. 1B) where rows show segments and columns show their reproducibility score for each protein and a final score (FS) defined as the average value of the row (or NA if more than 1 reproducibility score equals NA).

Statistical test of scored regions

To assess whether the number of reproducible or variable regions associated with a particular score is significant, a suitable control had to be identified. The appropriate null distribution was built by randomizing the RSM. We performed this task using the “sample” function in R. The randomization was performed 1000 times, and regions at particular scores (0, 0.25, 0.33, 0.5, 0.66, 0.75 and 1) of the null distribution were counted. Afterwards, a *z-score* was computed according to this formula:

$$\text{z-score} = \frac{S - S^{rand}}{\sigma S^{rand}}$$

bioRxiv preprint doi: https://doi.org/10.1101/796383; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where S is the number of regions observed with a particular score, and S^{rand} and σS^{rand} are the mean value and the standard deviation of the null distribution, respectively. Assuming normality of the null distribution, it is possible to analytically calculate the corresponding *p-value* for a given *z-score* with significant level $\alpha = 0.05$. The regions for each particular score were subjected to the test.

Principal Component Analysis and Euclidean distances

Principal Component Analysis (PCA) was performed using the Python package scikit-learn version 0.19.1. The dots represented in the PCA are biological replicates for a given protein. Each colour represents a specific protein and the features set used to perform the PCA are all the segments detected in all the proteins. In order to test the effect of the removal of the variable regions in the PCA, segments within the variable regions were removed from the features set. The similarity distances between replicates of the same protein in the PCA were computed with the Python package SciPy version 0.19.1 using as a metric the Euclidean distance. Boxplot and Dotplot were performed using the Python library Matplotlib version 2.2.2.

Enrichment analysis at regulatory elements

We collected genomic coordinates of the following gene related features from the UCSC table browser database in hg19 and mm10 annotations: promoter, 5UTR, coding exon, intron and 3UTR. Furthermore, we also used regions with R-loops (6) since they were previously reported as a potential feature associated with misbehaving ChIP peaks (3). For a set of regions (e.g. variable), the enrichment for each feature is obtained by dividing the number of regions overlapping a regulatory feature by the number of randomized regions overlapping the same feature.

$$Enrichment = \frac{VariableRegions \bigcap Feature}{RandomRegions \bigcap Feature}$$

Randomized regions were obtained using bedtools version v2.25.0 shuffleBed (7).

CG and AT dinucleotide frequency calculation

The percentage of CG and AT dinucleotides in the mouse and human genomes was calculated with the *nuc* function in the bedtools version v2.25.0 toolkit. To compute the CG and AT enrichment in the variable regions of mESCs and K562 cells we used a set of control regions using the shuffleBed function in bedtools version v2.25.0 (7). The differences in dinucleotide composition between the variable regions and the set of control regions were tested for significance using a t-test.

bioRxiv preprint doi: https://doi.org/10.1101/796383; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Prediction of variable regions in K562 and mESCs using genomic features

In order to found out whether different genomic features (active chromatin marks, repressive chromatin marks, DNA accessible regions, CpG islands, etc.) could be used by a random forest classifier to predict variable regions in mESCs or in K562 cells we used a large panel of datasets.
For mouse ESCs, we also considered regions undergoing TET oxidation and bivalent domains. We used the following published data: 5hmC, 5fC and 5caC (8); CpG islands extracted from UCSC table browser for mm10 annotation; H3K4me1, H4K4me3, H3K79me2, H3K27ac, H3K27me3, H3K36me3 (9), LMR (10), DNAse-seq from ID:ENCSR000CMW experiment in the ENCODE portal. mm9 genome features were converted to mm10 using the Batch Coordinate Conversion (liftOver) tool from the UCSC Genome Browser Utilities (https://genome.ucsc.edu/cgi-bin/hgLiftOver).
For K562 cells, we used the following data downloaded from the ENCODE data portal: H3K27ac (ID: ENCFF044JNJ), H3K27me3 (ID: ENCFF145UOC), H3K4me1 (ID: ENCFF183UQD), H3K4me3 (ID: ENCFF261REY), H3K79me3 (ID: ENCFF350GQM), H3K36me3 (ID: ENCFF537EUG), DNAse-seq (ID: ENCFF856MFN). CpG islands were extracted from UCSC table browser for hg19 annotation.

To train and test the random forest model we used the function randomForest from the R package randomForest version 4.6-14 (DOI:10.1023/A:1010933404324). As a positive set we have used the variable regions estimated by our method and as a negative set we used a set of regions obtained with the package gkmSVM version 2.0 (11). This package has a function named genNullSeqs capable of using the positive set of sequences and learning their nucleotide composition. Subsequently, the function generates a set of genomic locations with sequences of nucleotide composition and length similar to those in the positive test set. This approach was successfully used previously for the prediction of double strand breaks at CTCF and accessible chromatin sites (17). The train and test set were obtained using the function sample.split in the R package caTools version 1.17. The function is used to split the data used during classification into train and test subsets. We decided to split train and test in a 30:70 ratio. We also split the dataset using other ratios such as 10:90, 20:80 and 40:60 and this did not change the performance of the prediction. The evaluation of the prediction was performed by computing the ROC curve using the package pROC version 1.13.0 (https://cran.r-project.org/web/packages/pROC/pROC.pdf).

RESULTS

Variable regions can be detected in all tested cell lines

Our method to detect variable regions in ChIP-seq datasets for a given cell line relies on having several proteins tested for the particular cell line with replicates coming from the same laboratory and using the same platform (see Methods for details). Currently, the number of such sets in the ENCODE database is limited. While we were able to obtain a suitable set for K562 with four proteins and three replicates each, it is more usual to have a lower number of replicates, typically two.

bioRxiv preprint doi: https://doi.org/10.1101/796383; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For this reason, we applied our method to identify variable regions using just two replicates per protein for the human cell lines K562, GM12878, HepG2, MCF-7, and for mouse ESCs (mESCs). We found that K562 cells have a higher and significant number of variable regions for a total amount of 483 (p-val = 6.3e-103; see Methods for details) whereas the other three human cell lines GM12878, HepG2, MCF-7 have similar lower but also significant numbers: 61, 76 and 62, respectively (p-val = 1.1e-07, 5.9e-6, or 4e-5, respectively) (Fig. 2A-D). This might be due to the instability of the K562 cell's genome as reported in previous publications (12). Furthermore, we used another popular cell line used for developmental studies, mouse embryonic stem cells ES-14. Also in this cell line, we identified a number of variable regions for a total amount of 332 (p-val = 6.9e-3) (Fig 2E).

Finally, we wanted to see to which extent increasing the number of replicates in the experimental design would increase the number of variable regions. We were able to design a proper experiment only for K562 cell lines, which is the cell line with the higher number of experiments in the ENCODE project. Using three replicates per protein and four proteins (see Methods for details), the number of variable regions detected was drastically higher (a total of 3012).

Variable regions are rich in CG dinucleotides and enriched along gene body features

Next, we tested the CG dinucleotide frequency of the variable regions in K562 and mESCs. We found a higher frequency of CG dinucleotides compared to a random set of control genomic regions in K562 (p val 3.8e-4) and mESCs (p val 8.4e-8) (Fig. 3A and 3B, respectively). The protein targets in the ChIP-seq experiments used do not have DNA-binding motifs particularly affine for CG dinucleotides (motifs from the Jaspar database (13); Supplementary Fig. S2), hence the CG composition is specifically related to the variable behavior and not to the DNA-motifs bound by the proteins selected.

Furthermore, variable regions were highly enriched for different features among K562 and mESCs cells with K562 showing a high enrichment for promoters and R-loops (Fig. 3C) and mESCs showing a high enrichment for 3UTR and Promoters (Fig. 3D). In a recently published work (3), the authors reported that previously characterized regions as "high occupancy target" (HOT) (18,19) are likely to be a ChIP-seq artifact. Among the properties of these regions, they reported GC/CpG-rich kmers and RNA–DNA hybrids (R-loops). Since also in our work we found these characteristics for the variable regions, we downloaded the regions reported in (3) and checked for a possible enrichment. We found significant enrichment for all the cell lines tested except for GM12878. Especially for the K562 and HepG2 cell lines the enrichment was of 52 and 17 fold change, respectively (observed vs expected, two-sided Fisher test p val. 1.46e-87, 3.9e-3, respectively). Furthermore, also mESCs showed a significant enrichment of 8 fold change (observed vs expected, two-sided Fisher test p val. 3.2e-3) (Fig. 3E).

bioRxiv preprint doi: https://doi.org/10.1101/796383; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Variable regions are not blacklisted regions from ENCODE project

The ENCODE consortia provides a detailed description about the meaning of "blacklisted sites". These genomic positions often produce artifact signal in certain loci mainly because of excessive unstructured anomalous sequences. Reads mapping to them are uniquely mappable so simple mappability filters do not remove them. These regions are often found at specific types of repeats such as centromeres, telomeres and satellite repeats. Given the high variability of the ChIP-seq peaks of the regions described in this manuscript we thought to check whether our method was detecting the already described and characterized "blacklisted regions" or not. To answer this question, we analysed the overlap of the variable regions obtained in all the cell lines we have used (K562, HepG2, MCF-7, GM12878 and mESC) with the public available ENCODE blacklisted regions (20). We found no overlap except for K562 (significant depletion, 35 observed vs 261 in random model, two-sided fisher test p val. 2.92e-46) and mESC (significant depletion, 7 observed vs 29 random, p val. 2.1 e-4). These results confirm that our variable regions are not associated with the ENCODE blacklisted regions, hence need to be considered for new detection methods.

The removal of variable regions improves the interpretation of the PCA in K562 cell lines

Variable regions may reflect cell-specific effects that are not target-specific. While this information might be indicating biological function, we hypothesized that the removal of such target non-specific data could result in an improvement of the separation of the replicates points in a PCA. In order to test such potential benefit in removing the variable regions for downstream interpretative analysis, we performed a PCA of the ChIP-seq samples obtained for the K562 cell line (Fig. 4A). The PCA was performed using (i) all the segments bound by each protein in the respective replicates in the original datasets and (ii) without the segments within the variable regions. We found that the separation along the components improves after removing the segments within the variable regions. Furthermore, the replicates of the proteins tend to cluster better without the segments within the variable regions and this is reflected with a lower Euclidean distances in pairwise comparisons between replicates of the same proteins (Supplementary Fig. S3). We note that this does not mean that data from these regions should be discarded, but that they should be considered differently. Further research is needed to characterize these regions and find out if they have a cell-specific biological function.

DNA accessible regions are predictive of the variable behavior in K562 and mESCs

Finally, we searched for genomic features that can be predictive of variable behaviour. We evaluated the possible association of different genomic features with variable regions using a random forest classifier.

bioRxiv preprint doi: https://doi.org/10.1101/796383; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The classifier (random forest) was trained with a positive set consisting of the variable regions detected in mESCs (332 variable regions), and with a negative set consisting of genomic sequences with size and nucleotide composition similar to those of the positive set (see Methods for details about the training and about the set of genomic features).

The algorithm was able to classify the variable regions (area under ROC curve = 0.82) and returned as best predictors DNA accessible regions, together with regions lowly methylated and oxidative products of TET enzymes 5hmC and 5caC (Fig. 5A and 5B). These modifications are highly frequent at distal regulatory elements (8) and promoters (14) and we speculate that the turnover of these modifications might affect the binding of the proteins to the DNA leading to stochastic variation of the binding sites.

To observe the reproducibility of these results, we studied next data from K562 cells (483 variable regions). Again, as best predictor of variable regions we found DNA accessible regions, together with K3K4me1 and H3K27ac chromatin marks (area under the ROC curve = 0.97) (Fig. 5C and 5D).

bioRxiv preprint doi: https://doi.org/10.1101/796383; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DISCUSSION

During the latest years, several laboratories tried to study the regulation of gene expression in different model organisms. For this scope, ChIP-seq was adopted as a standard technique but the extent of its usage raised some questions in terms of reliability (1,2,15,16). In particular, in a previous work (3), Wreczycka and colleagues presented a method that considers the nature of phantom peaks and hyper-ChIPable regions to define high-occupancy target (HOT) regions where un-specific binding to multiple targets would be found even in the absence of expected binding motifs. They concluded that the un-specificity of binding sites in HOT regions is associated with CG dinucleotide rich regions and as a consequence at R-loops (that are CG rich) and DNA tertiary structures. Though, this is a common concern for ChIP-seq assays and since the beginning the technique was known to be biased toward GC-rich contents during fragment selection in the steps of the library preparation and amplification during the sequencing (15). Here we have found evidence that supports that such regions could also be responsible for variable behaviour in ChIP-Seq in a cell-specific fashion. Our method evaluates replicated ChIP-seq experiments for multiple targets in a cell type, to find regions where target binding is not reproduced in all replicates for multiple targets. These variable-occupancy target (VOT) regions are cell-specific and share structural features with HOT regions. However, differently to HOTs, VOTs do not produce consistent un-specific target binding. Accordingly, VOTs do not overlap blacklisted ENCODE regions. Together, the cell-specificity of VOTs and our finding that VOTs can be predicted using DNA accessibility suggest their dependency on gene expression and epigenetic state.

While the most consistent enrichment of variable regions observed was for promoters and 5UTR regions in both K562 cells and mESCs, the differences observed for variable regions at R-loops suggest that it is not possible to drive a certain conclusion relating where exactly this variation occurs. On the other hand, the fact that we are able to predict variable regions using genomic features alone with relatively high accuracy indicates that there is certainly a relation between genomic features and variability that could be eventually detected. Taking these results together, we assume that with the future availability of further ChIP-Seq datasets testing multiple proteins in the same cell lines it will be possible to assess the sources of variability in ChIP-Seq with more certainty.

With our method, we propose a systematic approach using ChIP-seq experiments and replicated measurements in a given cell-type to identify misbehaving DNA regions that have to be treated differently in the post processing downstream analysis. We have shown that discarding data from these regions can improve studies focusing on target specific effects. However, further research is needed to study potential cell-specific functions of VOTs as hubs or sponges for transcriptional regulatory complexes, which could be verified with other experimental assays like ChIP-qPCR. We suggest applying our approach as a post processing quality check of the data before starting follow up experiments and driving biological conclusions.

We must point out that our method requires enough replicates for the same protein in a given tissue. For example, predictions for organisms like the fly *Drosophila melanogaster* using data from

bioRxiv preprint doi: https://doi.org/10.1101/796383; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

modENCODE and modERN are not yet possible, At this point, the only datasets we find suitable for our analyses are in ENCODE.

Finally, we note that similar approaches to the one used for our method to point to variable regions in ChIP-seq datasets could be eventually developed and applied to any type of next generation sequencing datasets that uses replicated measurements under various conditions (for example, ATAC-seq from multiple cell types). This could open avenues for the discovery of other types of variability leading to a more informed use of sequencing-based data. The study of the similarities and differences between variable regions obtained with different techniques might be crucial to increase our understanding of the inter-relation between genomic structural flexibility and regulatory function.

DATA AVAILABILITY

The code to reproduce this manuscript is available in the GitHub repository (https://github.com/tAndreani/IPVARIABLE).

SUPPLEMENTARY DATA

Supplementary Data are available online.

ACKNOWLEDGEMENTS

We thank Katerina Taškova for fruitful discussions. Tommaso Andreani and Steffen Albrecht were affiliated with and financially supported by the International PhD Programme on *Gene Regulation, Epigenetics and Genome Stability*, Mainz, Germany.

REFERENCES

1. Teytelman, L., Thurtle, D. M., Rine, J., & van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proceedings of the National Academy of Sciences*, *110*(46), 18602–18607.

2. Jain, D., Baldi, S., Zabel, A., Straub, T., & Becker, P. B. (2015). Active promoters give rise to false positive “Phantom Peaks” in ChIP-seq experiments. *Nucleic Acids Research*, *43*(14), 6959–6968.

3. Wreczycka K., F. Vedran, U. Bora, R. Wurmus, S. Bulut, B. Tursun, A. Akalin (2019). HOT or not: examining the basis of high-occupancy target regions, *Nucleic Acids Research*, 47(11), 5735-5745.

4. Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., ... & Chen, Y. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research*, *22*(9), 1813-1831.

bioRxiv preprint doi: https://doi.org/10.1101/796383; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5. Li, Q., Brown, J. B., Huang, H., & Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *Annals of Applied Statistics*, *5*(3), 1752–1779.

6. Sanz, L. A., Hartono, S. R., Lim, Y. W., Steyaert, S., Rajpurkar, A., Ginno, P. A., ... & Chédin, F. (2016). Prevalent, dynamic, and conserved R-loop structures associate with specific epigenomic signatures in mammals. *Molecular cell*, *63*(1), 167-178.

7. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842.

8. Shen, L., Wu, H., Diep, D., Yamaguchi, S., D'Alessio, A. C., Fung, H. L., ... & Zhang, Y. (2013). Genome-wide analysis reveals TET-and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell*, *153*(3), 692-706.

9. Whyte, W. A., Bilodeau, S., Orlando, D. A., Hoke, H. A., Frampton, G. M., Foster, C. T., ... & Young, R. A. (2012). Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. *Nature*, *482*(7384), 221.

10. Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., ... & Tiwari, V. K. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, *480*(7378), 490.

11. Ghandi, M., Mohammad-Noori, M., Ghareghani, N., Lee, D., Garraway, L., & Beer, M. A. (2016). gkmSVM: an R package for gapped-kmer SVM. *Bioinformatics*, *32*(14), 2205-2207.

12. Ramachandran, P., Palidwor, G. A., & Perkins, T. J. (2015). BIDCHIPS: bias decomposition and removal from ChIP-seq data clarifies true binding signal and its functional correlates. *Epigenetics & chromatin*, *8*(1), 33.

13. Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., .. & Baranasic, D. (2017). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic acids research*, *46*(D1), D260-D266.

14. Neri, F., Incarnato, D., Krepelova, A., Rapelli, S., Anselmi, F., Parlato, C., ... & Oliviero, S. (2015). Single-base resolution analysis of 5-formyl and 5-carboxyl cytosine reveals promoter DNA methylation dynamics. *Cell reports*, *10*(5), 674-683.

15. Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nature reviews genetics*, *10*(10), 669.

16. Park, D., Lee, Y., Bhupindersingh, G., & Iyer, V. R. (2013). Widespread misinterpretable ChIP-seq bias in yeast. *PloS one*, *8*(12), e83506.

17. Mourad, Raphaël, et al. "Predicting double-strand DNA breaks using epigenome marks or DNA at kilobase resolution." *Genome biology* 19.1 (2018): 34.

18. Xie D., Dan X., Boyle A.P., Linfeng W., Jie Z., Trupti K. et al. . Dynamic trans-acting factor colocalization in human cells. *Cell*. 2013; 155:713–724.

19. Boyle A.P., Araya C.L., Cathleen B., Philip C., Chao C., Yong C., Gardner K., Hillier L.W., Janette J., Jiang L. et al. . Comparative analysis of regulatory information and circuits across distant species. *Nature*. 2014; 512:453–456.

20. Haley M. Amemiya, Anshul Kundaje and Alan P. Boyle. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports* 9.1 (2019): 9354.

bioRxiv preprint doi: https://doi.org/10.1101/796383; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

FIGURE LEGENDS

Figure 1. Method to annotate genomic regions with a reproducibility score. (**A**) From the genomic segments S, segments with ChIP-seq peaks for a given protein 1 in a given cell, in N=3 replicates, are converted to a binary format (Rep1 to Rep3). The sum at each segment of the values for the replicates (SUM or P) allows to define blocks of consecutive segments between zero-scored segments devoid of peaks (here two blocks; green and blue). All segments in a block are identified as indicating a reproducible region (Reproducibility=1; green) if the block holds at least one segment with value 3. Otherwise they are given a value indicating a non-reproducible region (Reproducibility=0; blue). (**B**) Average reproducibility values for ChIP-seq experiments from four different proteins in cell type A are combined in a final score that ranges from 0 (not reproduced in the four proteins) to 1 (reproduced in the four proteins). Only segments with values for at least 3 proteins were considered.

Figure 2. Significance of the observed number of variable regions for different cell types. The number of variable regions observed in each cell line (red line) is significantly higher than the corresponding computed null distribution (blue). (**A**) K562, (**B**) GM12878, (**C**) HepG2, (**D**) MCF-7 and (**E**) mouse ESCs cells;

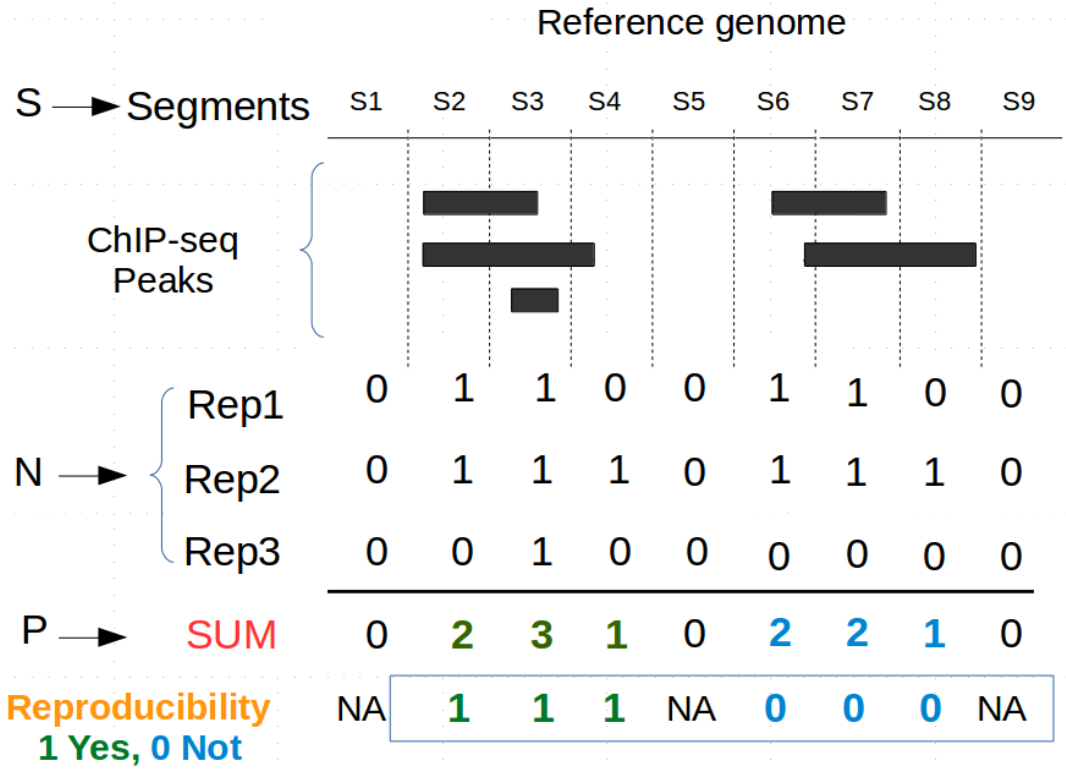
Figure 3. Properties of variable regions in K562 and mESCs. CG dinucleotide composition in K562 human cell lines (**A**) and in mouse ESCs (**B**). Control regions are a set of randomly sampled genomic regions of similar size. Enrichment of variable regions in K562 human cell lines (**C**) and mouse ESCs (**D**) at several gene body features and R-loops. (**E**) Enrichment of variable regions at HOT regions (3).

Figure 4. Effect of removing variable regions. PCA using presence of peaks in the set of 200 bp segments on the genome as features with the original data and without 200 bp segments within the variable regions in K562 cell lines (**A** and **B**, respectively). Each dot represents a biological replicate and each colour the protein target.

Figure 5. Prediction of variable regions in K562 and mESCs. (**A**) Receiver operating characteristic (ROC) as a quality measure of the predictability of the variable regions in mESCs and (**B**) importance of the features for predicting the variable regions in mESC measured as mean decrease Gini in the random forest. See text for details about every feature. (**C**) Receiver operating characteristic (ROC) as a quality measure of the predictability of the variable regions in K562 cells and (**D**) importance of the features for predicting the variable regions in K562 cells measured as mean decrease Gini in the random forest.

Figure 1

A



B

Reproducibility Score Matrix

Segments	Reproducibility Protein1	Reproducibility Protein2	Reproducibility Protein3	Reproducibility Protein4	Final Score
s1	NA	NA	NA	NA	NA
s2	1	0	1	0	0.50
s3	1	1	1	1	1
s4	1	1	1	0	0.75
s5	NA	0	0	0	0
s6	0	1	1	0	0.50
s7	0	NA	0	0	0
s8	0	0	0	0	0
s9	NA	0	0	1	0.33

Figure 2

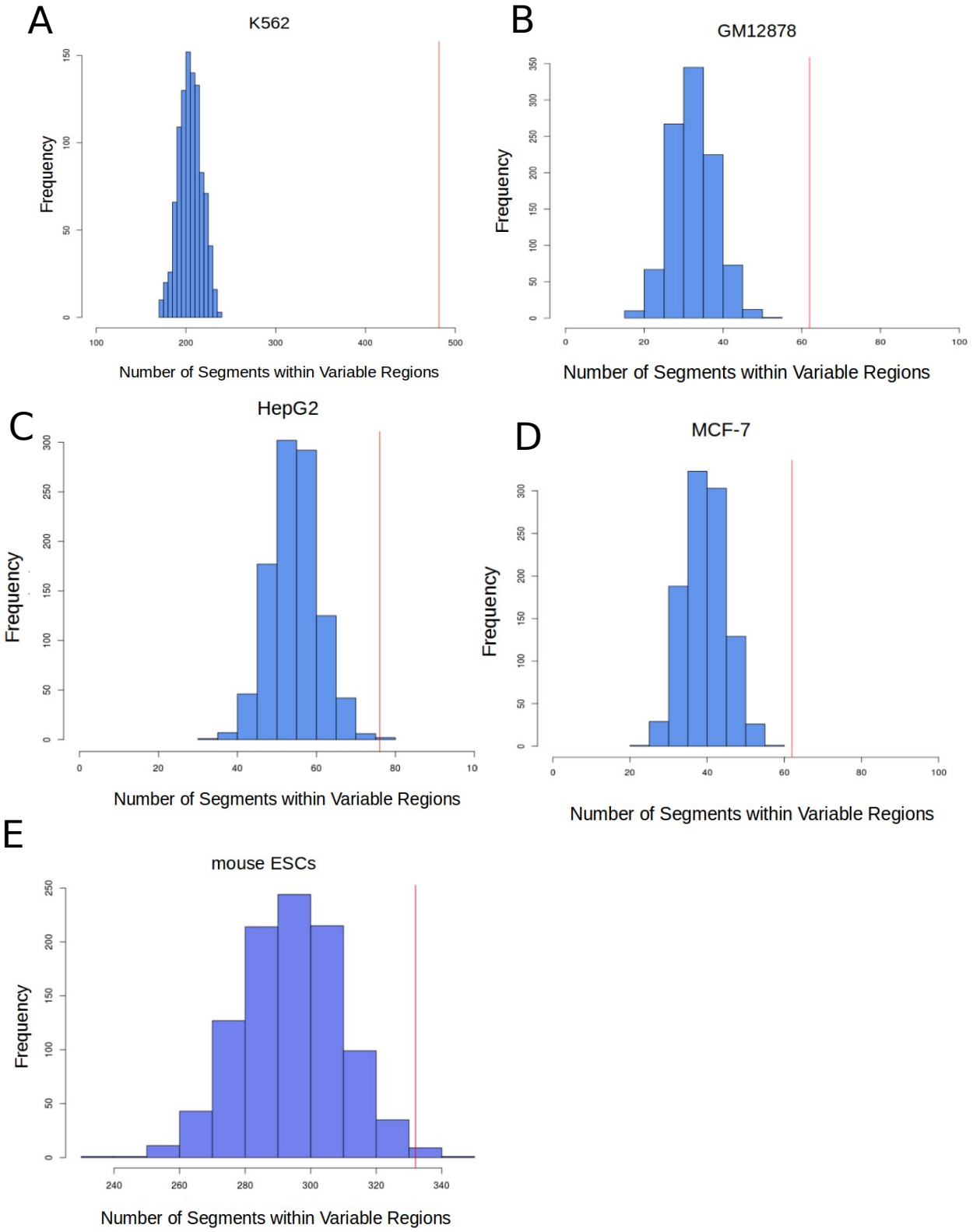
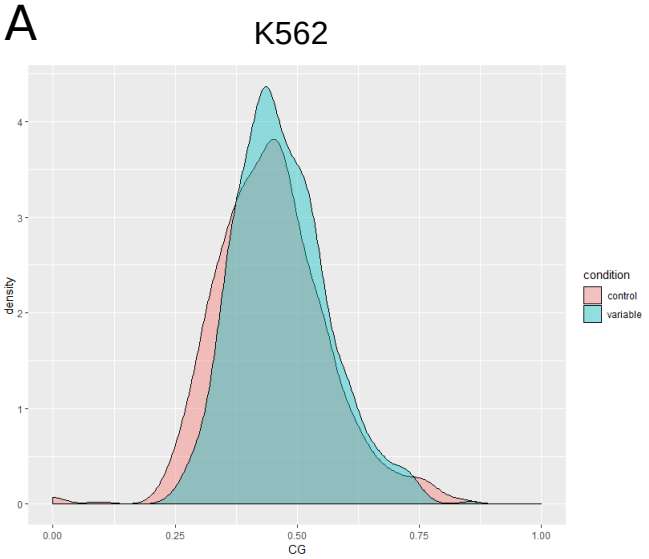
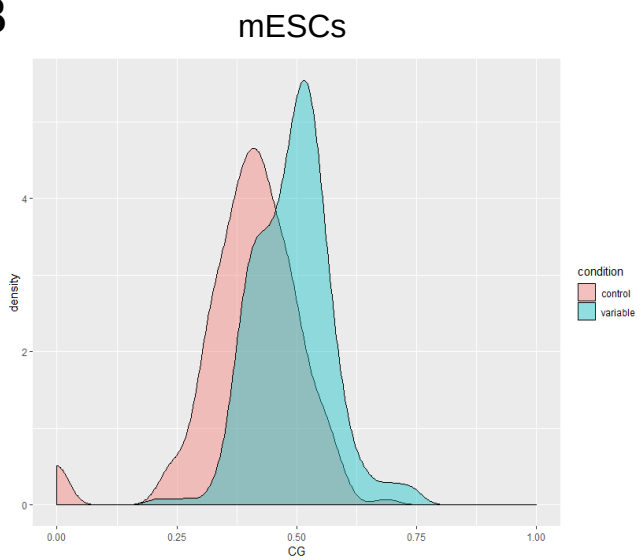

A
K562
Frequency
Number of Segments within Variable Regions
B
GM12878
Frequency
Number of Segments within Variable Regions
C
HepG2
Frequency
Number of Segments within Variable Regions
D
MCF-7
Frequency
Number of Segments within Variable Regions
E
mouse ESCs
Frequency
Number of Segments within Variable Regions

Figure 3

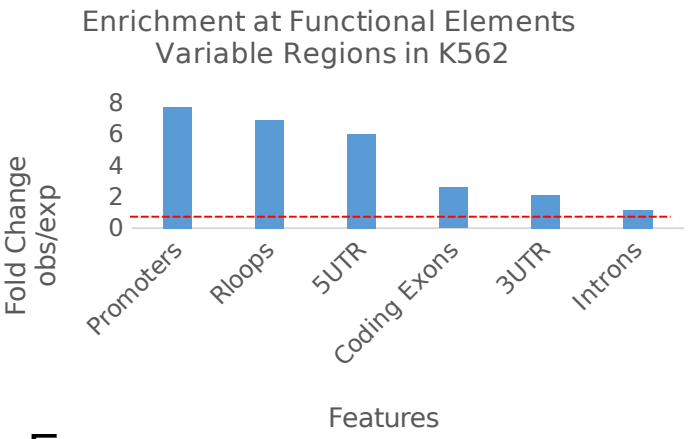
A



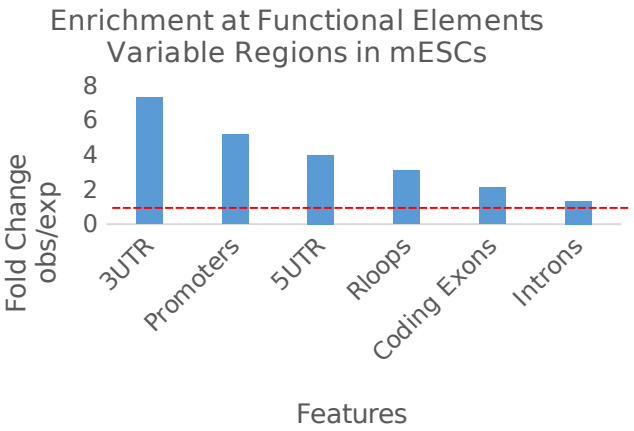
B



C



D



E

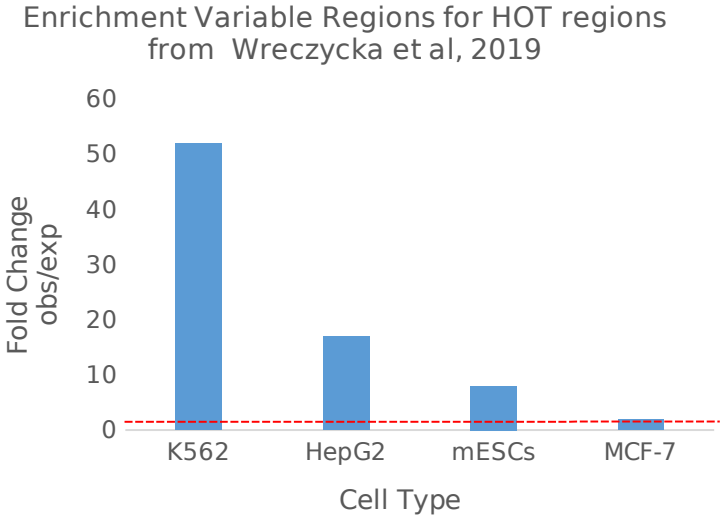


Figure 4

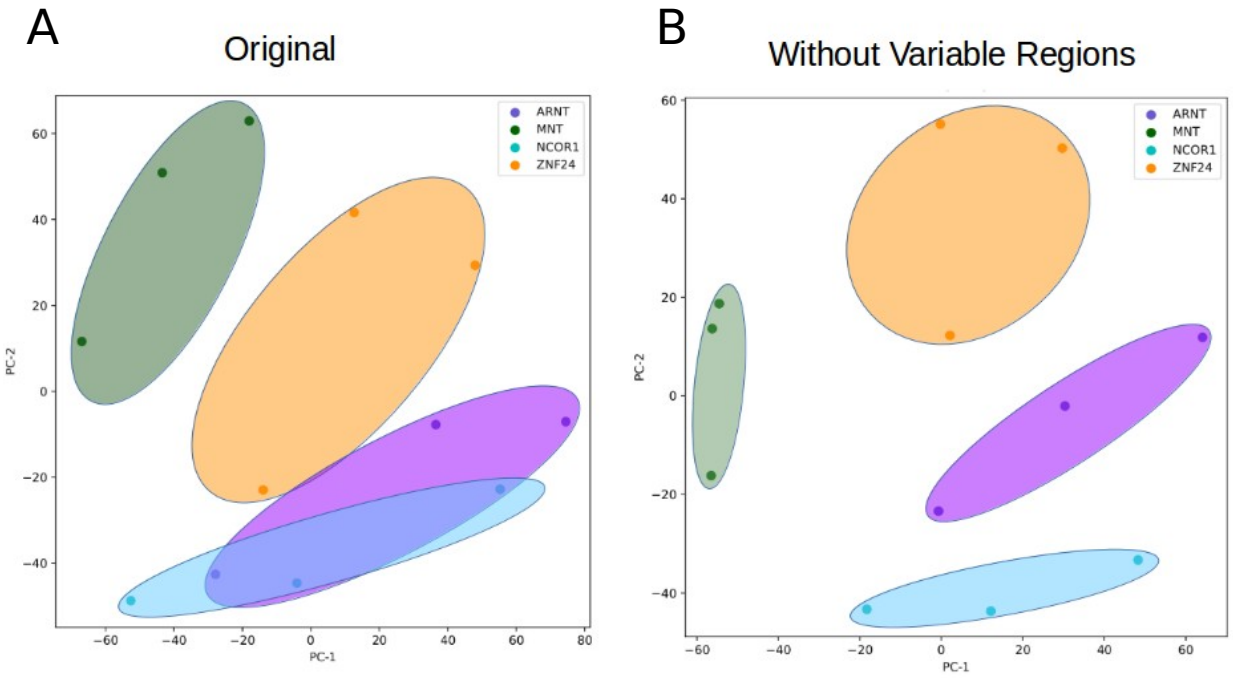
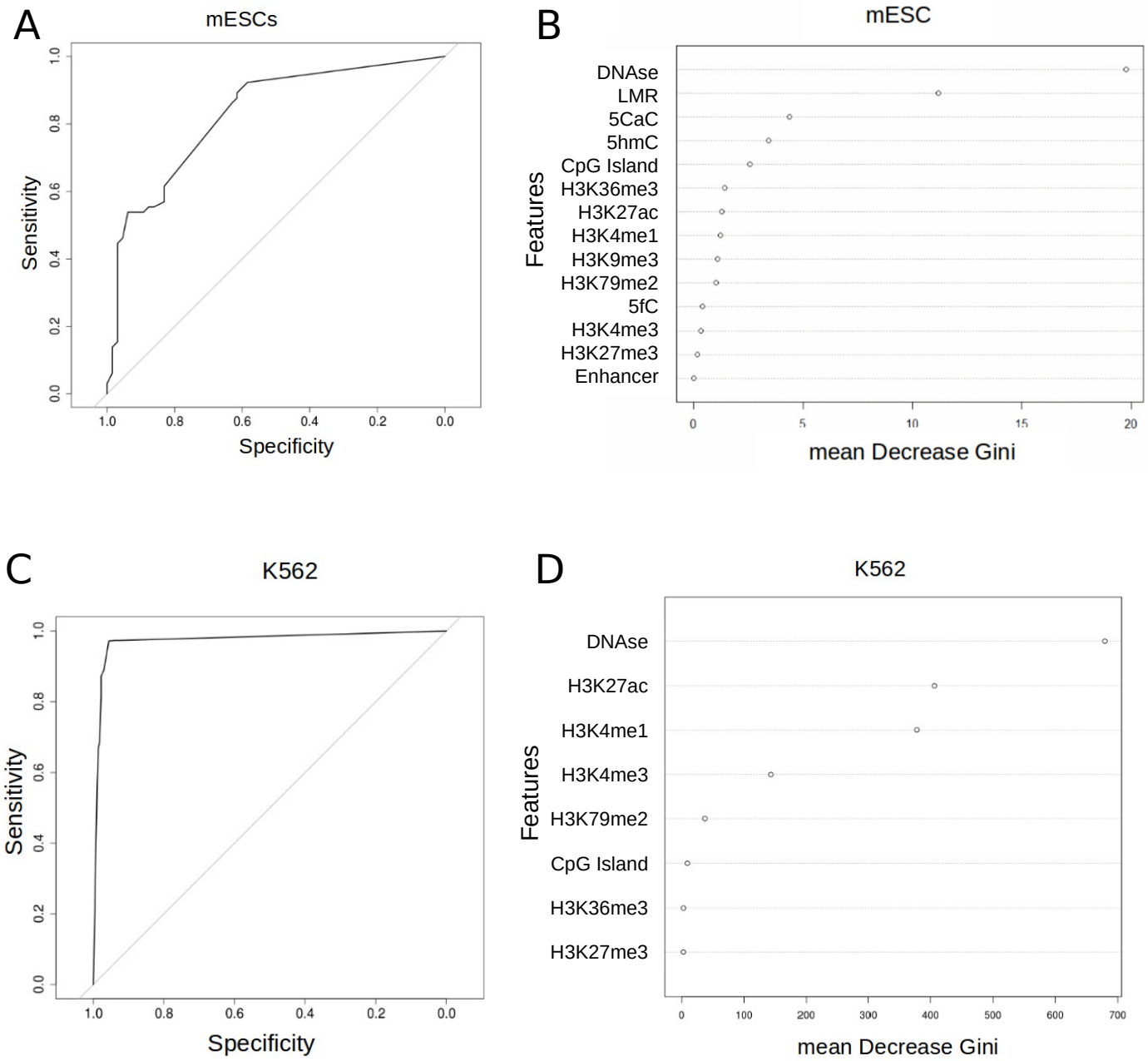

A
Original
ARNT
MNT
NCOR1
ZNF24
PC-2
PC-1
B
Without Variable Regions
ARNT
MNT
NCOR1
ZNF24
PC-2
PC-1

Figure 5

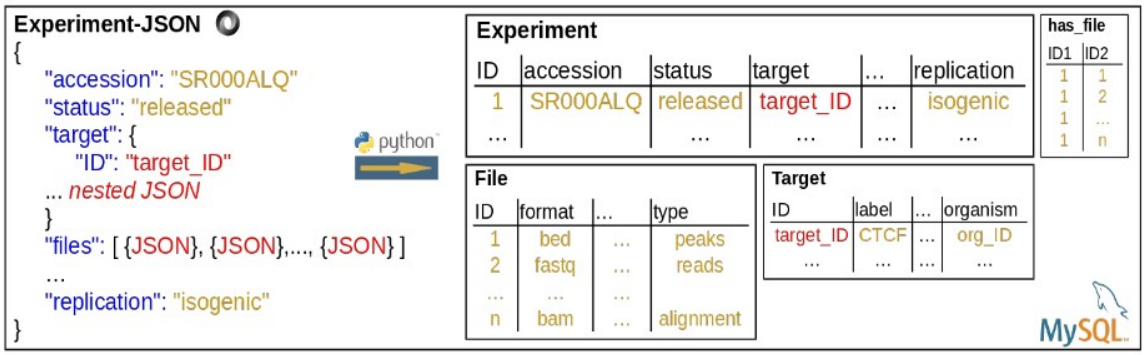

A
mESCs
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
B
mESC
Features
DNAse
LMR
5CaC
5hmC
CpG Island
H3K36me3
H3K27ac
H3K4me1
H3K9me3
H3K79me2
5fC
H3K4me3
H3K27me3
Enhancer
0
5
10
15
20
mean Decrease Gini
C
K562
Sensitivity
Specificity
D
K562
Features
DNAse
H3K27ac
H3K4me1
H3K4me3
H3K79me2
CpG Island
H3K36me3
H3K27me3
0
100
200
300
400
500
600
700
mean Decrease Gini

Supplementary Material

Conversion of JSON-based files to a relational SQL database

The metadata in ENCODE is represented in JSON-format. There is an API (Application Program Interface) that enables downloading JSON files for many experiments automatically. In addition, it is possible to restrict the experiments to be downloaded by specifying conditions the experimental metadata has to comply with. However, a JSON file for one experiment can include up to 11,000 lines and this can cause two problems. The first is that the extraction of information can be complicated and requires further implementation of small scripts / programs. The second is that parsing such files can be computationally expensive especially if there are more than 1000 files to be processed. To avoid these problems and enabling the analysis of the metadata in a flexible and fast way we decided to convert the information from JSON into a MySQL database. For objects like experiment, target, organism, replicate, etc. we downloaded all JSON files. A Python script was implemented that extracts information encoded in each JSON file and automatically fills the SQL database with this information. The procedure in the Python script takes also cares about creating the tables for each object and furthermore creating also relational tables for the connection between two tables (objects). Relational information is also automatically extracted and stored into SQL. Figure_1 represents a simplified case with one experiment. The brackets "{}" define the beginning and end of a JSON. There are simple features like accession and status for which the value is directly given. Those simple values can also appear in lists represented by the brackets "[]". However, one JSON can include other JSONs. This is shown by the target in the experiment that is completely described by a so called nested JSON. It is also possible that one JSON includes a list of JSONs. The example in Figure_S1 shows that for each file that is related to the experiment there is one JSON within the experiment. These nested JSONs are the reason why those files are getting so large. Another advantage coming with the conversion from JSON into SQL is that objects are not stored redundantly. Taking as example the target that describes the protein that is targeted in a ChIP-seq experiment. All JSONs for experiments with the same target store the same information for the same target multiple times. In SQL there is only one entry for a specific target in the target table and experiments with this target are just linked to this entry by the target_ID. An example for such a link is represented by the target in Supplementary Figure S1 below and has_file is a relational table.

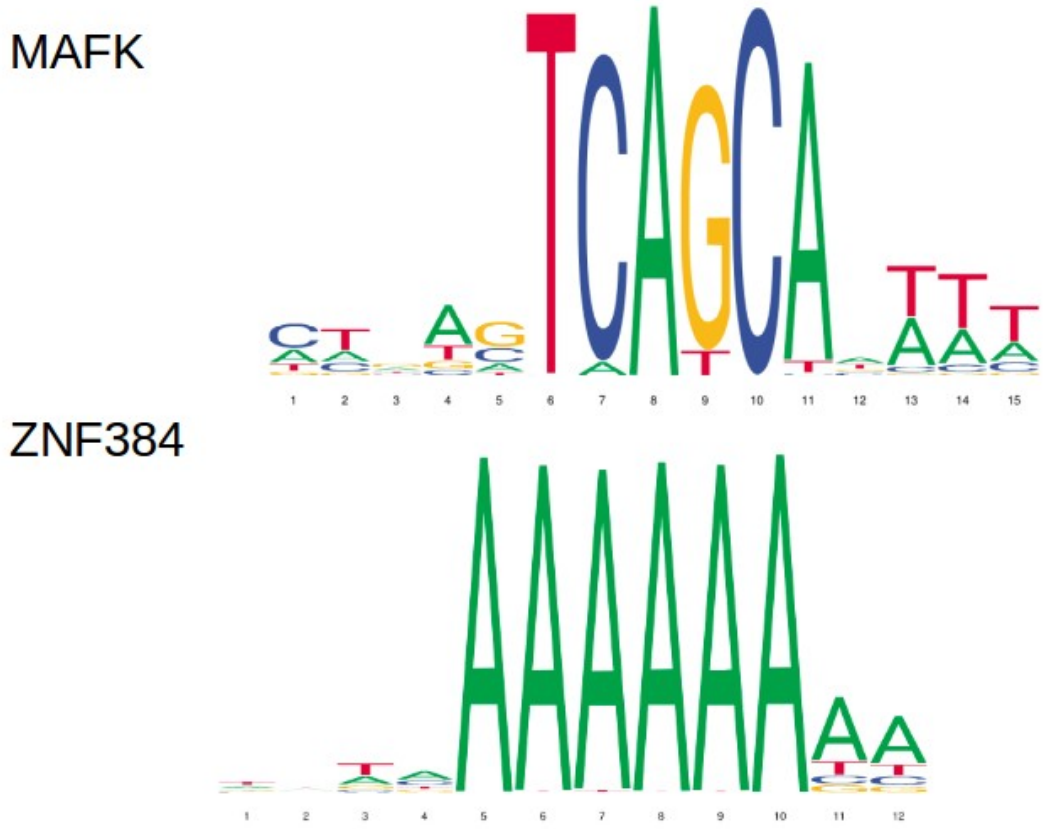
Figure S1



1) Same Cell Isogenic Line (K562)
2) Same condition/treatment
3) Same Laboratory (Snyder)
4) Same Bioinformatics ENCODE Pipeline
5) Same Statistical Test (IDR threshold)

Figure S2

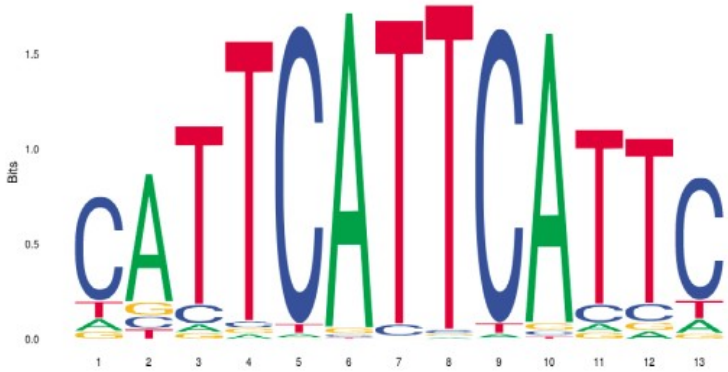
(A) DNA binding motifs of the protein targets of the ChIP-seq samples used to detect variable regions in mouse ESCs from the Jaspar database. No data for ZC3H11A and HCFC1.



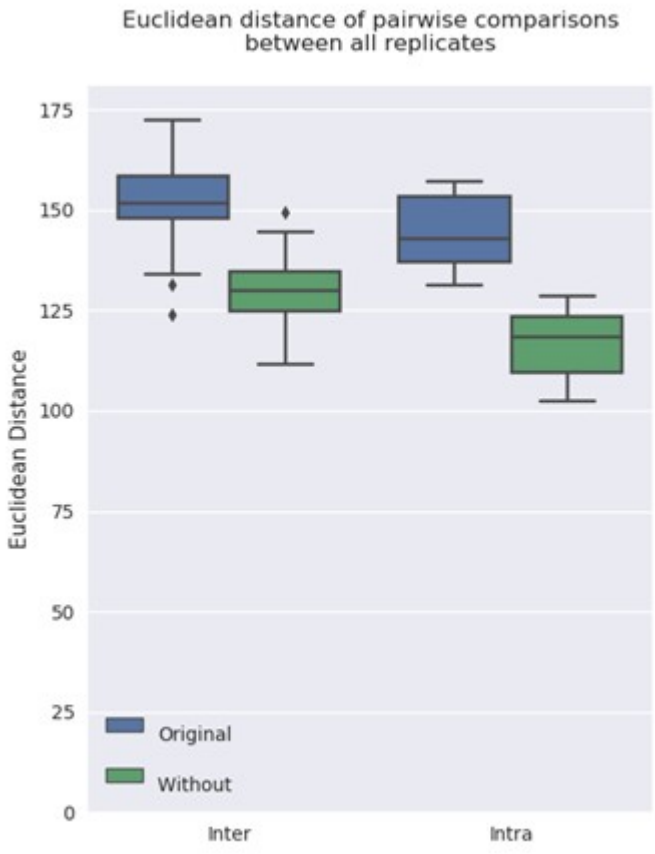
(B) DNA binding motifs of the protein targets of the ChIP-seq samples used to detect variable regions in K562 cell lines

ARNT

ZNF24



MNT

Figure S3

Pairwise comparisons of Euclidean distances between replicates of different proteins (inter) and within replicates of the same proteins (intra) in the original dataset and after removing the segments within the variable regions for the K562 ChIP-seq dataset (four proteins with three replicates).