BIO-463: Report

Lisa Dratva (SV)

Paper 8: Predicting effects of noncoding variants with deep learning-based sequence model.
29 May 2020

***Abstract*—We first analyze the paper of Zhou & Troyanskaya introducing their chromatin feature prediction software DeepSEA. DeepSEA uses sequence information to infer functional effects from non-coding variants. Non-coding variants make up the majority of disease-associated variants, yet their impact on transcription remains very difficult to evaluate. DeepSEA is able to predict the impact on many chromatin features down to the single-nucleotide substitution level. We then reproduce a figure from the paper and further use DeepSEA to identify a transcription factor (TF) consensus motif. Lastly, we evaluate the effect of known single-nucleotide polymorphisms associated with disease on TF binding affinity with DeepSEA.**

Introduction

The following report is based on the paper *Predicting effects of non-coding variants with deep learning–based sequence model* by Zhou & Troyanskaya [1]. They have developed a deep learning-based model that can predict the probability of specific chromatin features being present on any given nucleotide sequence, called DeepSEA. This implies being able to predict the functional effect of a mutation in a non-coding sequence, a task so challenging that no reliable models for it existed until 2015. This knowledge is of great importance to medicine and human health in general, as most trait and disease-associated variants are found in those non-coding parts. Our report will analyze their model as well as some of their mentioned results, and use the model to investigate further some pathological variants.

Chromatin features

Chromatin is the material that makes up a cell's chromosomes, comprised of DNA and supporting proteins. Chromatin feature is a term introduced by the authors of [1] and can be broadly resumed as parts of a genomic sequence that provide additional organization to the cell other than mRNA-coding sequences, for instance in form of binding sites for transcription factors (TFs). They play key roles in genetic regulation, but are hard to analyze because they affect genes, and not directly measurable gene products. The presence of chromatin features tells of the proteins that a sequence can give rise to: If DNA cannot be accessed by TFs, then the underlying genes might not be expressed, since the TFs can act as on-switches for genes.

DeepSEA

The authors report that their model achieves the best accuracy out of all existing chromatin effect prediction models (as of 2015), while being exclusively trained on genetic sequences, meaning no genome annotation data was used. The software predicts the probability of select TFs binding onto a sequence, but can also identify DNase I hypersensitive sequences or predict histone marks. They have trained the model to recognize these chromatin features in different cell types and can effectively predict the epigenetic state of a sequence.

Experimental data

To check the performance of the model, we have to compare its output to known TF binding sites. These can be obtained from ChIP-Seq experiments, where interacting protein-DNA complexes are analyzed. The technique of chromatin immunoprecipitation (ChIP) is applied on many short sequences of DNA with proteins of interest added to the mixture. Then all DNA not bound to the proteins is removed and only the bound DNA is sequenced, revealing the binding sites [2]. ChIP-seq data for TFs is available from the ENCODE project. Those results show peaks for chromatin regions where the protein binds. Thus, all the sequences contained in the ChIP-seq data are already filtered for confirmed binding sites and can be used to find the true binding sites of each TF.

Code

All code used in this report is hosted on Github and can be accessed via this link, including instructions to reproduce the results. Everything is coded in Python using Jupyter. The libraries `pandas`, `sklearn` and `matplotlib` are needed to run the code. The DeepSEA output files are too large to be put on Github (7.5GB), they can be accessed upon request, and pickle files are uploaded to the repository instead.

Metrics

The main metrics used to assess model performance in our case are Receiver operating characteristic (ROC) curve and the corresponding area under the ROC curve (AUROC or short AUC). The ROC curve is a measure of how well the model performs on a binary classification task [3]. The true positive rate and the false positive rate are calculated for a range of threshold values and plotted to yield the ROC curve. How the decision cutoff impacts those rates is illustrated in Figure 1. With perfect predictive power, the false positive rate will be zero while the true positive rate is at 100%, making the ROC curve adhere to the upper left corner with an area under the curve of 100%. For a random guess, the true and false positive rates will be equal for all threshold values, yielding an AUC of 50%.

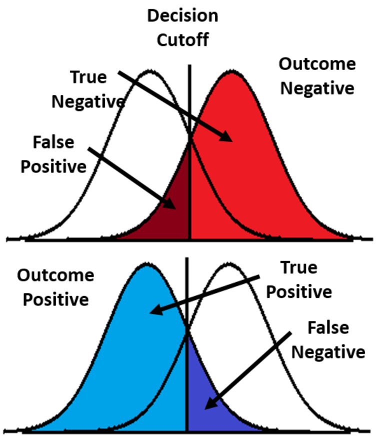

Figure 1: Illustration of thresholding to define true and false positive rates.

Task

For the first task, the figure chosen to be reproduced form the original paper is shown below. It depicts the ROC curves of all TFs analyzed in [1]. The second task was split into the two subtasks of *in silico* mutagenesis and predicting disease-associated phenotypes on new data.

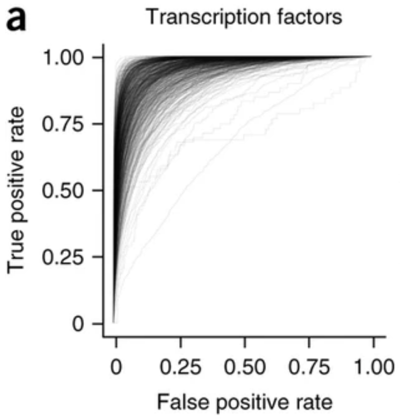

Figure 2: Chosen figure from [1] to be reproduced for first task.

Performance evaluation for 10 TFs

DeepSEA was trained with the GRCh37/hg19 genome assembly using all chromosomes except chromosome 8 and 9, which were kept for testing. This means we can feed the model ChIP-seq data from those two chromosomes to assess its performance on data it hasn't seen before.

Selection of TFs

10 sample TFs were chosen to replicate the paper's analysis. In the online supplementary material, the authors share a table showing the performance they have achieved for each separate chromatin feature and condition (combination of feature, cell type and cell culture treatment) and a table listing all publicly available chromatin feature profile files used for training DeepSEA. We randomly choose the ten TFs listed in Table I, which are well distributed throughout DeepSEA's performance curve. The AUC ranges from 80% to 100% for this selection, which corresponds to the lower and upper performance bounds reported by the authors.

Transcription factor Cell Type, Treatment	Short description of activity
BRCA1 GM12878, None	Involved in DNA repair, can directly regulate transcription.
CTCF GM12801, None	Repressor. Defines the boundaries of heterochromatic DNA.
E2F4 K652, None	Suppression of proliferation-associated genes.
c-Fos GM12878, None	Proto-oncogene involved in proliferation, differentiation, survival.
IRF3 HeLa-S3, None	Activates the transcription of interferons alpha and beta.
SP4 H1-hESC, None	Binds to GC promoter region of the photoreceptor signaling system.
TAF1 GM12891, None	Facilitates complex assembly and transcription initiation.
STAT2 K562, IFNa6h	Activates transcription in response to cytokines and growth factors.
ERα T-47D, Genistein-100nM	Important for hormone and DNA binding, activation of transcription.
KAP1 K652, None	Interaction with the Krüppel-associated box repression domain.

Table I: List of TFs used for testing

Preparing DeepSEA input for chromosome 8

DeepSEA accepts file inputs of the formats FASTA, BED and VCF. A FASTA file contains the actual nucleotide (acgt) sequence, whereas a BED file designates positions on the genome. The input sequence should always be 1000 nucleotides long, and then DeepSEA will make accurate chromatin feature predictions for the central 200 nucleotides. The authors note that analyzing a stretch of 1000bp instead of just a small sequence significantly improved predictive power of the model, presumably because the chromatin context around the central basepairs influences DNA accessibility and thus TF binding.

The next task is to write a text file where the entire chromosome is divided into packets using a sliding window of length 1000bp and sliding 200bp on each iteration. This way we intend to capture predictions for all regions of the chromosome.

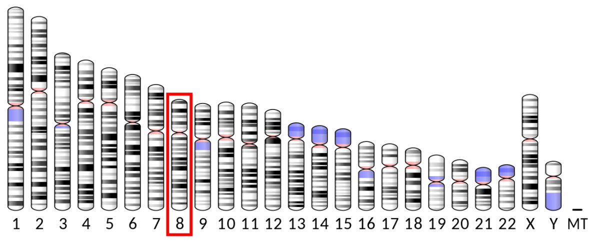

Figure 3: Organisation of the human genome on 23 chromosomes

Chromosome 8 has a reported length of 146,364,022 nucleotides in the hg19 genome assembly, and is shown for comparison in Figure 3.

The total chromosome length is then divided into suitable DeepSEA input sequences and saved as a text file, as shown in Table II. Each row of the table corresponds to an independent analysis and will receive a prediction output.

Chromosome name	Start position	End position
chr8	1	1001
chr8	201	1201
chr8	401	1401
chr8	601	1601
chr8	...	...

Table II: BED input to DeepSEA

DeepSEA has a size limit to the text file that can be provided, which lies at 2MB. A BED file can list around 74'000 rows before exceeding the size limit. This means that chromosome 8 has to be divided into 10 separate BED files for analysis. DeepSEA takes a few hours to analyze a 2MB file and saves the results in the cloud for later access.

Processing DeepSEA output

The model outputs a text file with 922 columns and a row for each prediction (the same number of rows as in the input BED file). Each chromatin feature has its own column and analysis is always performed for all 919 chromatin features, including DNase I hypersensitivity and histone mark analysis. The output is in the form of a probability value for each feature. The result of a 2MB input file is a 725MB output file, of which 10 are produced for analysis of the full chromosome 8.

A Python script was written to extract the columns for the ten sample TFs from the output files and merge them into one dataframe.

Normalizing probabilities

The overall scale of predicted probabilities for a chromatin feature depends on the proportion of positive examples in the DeepSEA training set (uniformly processed peaks from ENCODE and Roadmap Epigenomics projects). The output probabilities have to be normalized according to the positive proportion reported. The formula applied is

$$P_{norm} = \frac{1}{1 + \exp(-\log\frac{P}{1-P} - \log\frac{0.05}{0.95} + \log\frac{c}{1-c})}$$

where c designates the proportion of positive examples for this chromatin feature in the training data and P is the output probability. Uniform positive proportion of 5% is assumed across all chromatin features. After normalizing all probabilities, the resulting dataframe is saved to a pickle file for later use.

Downloading and labeling experimental data

We then download the TF ChIP-seq files, which contain ENCODE narrow peak information for each TF. This relates to the signal peak observed during the ChIP-seq experiment. We are only interested in the position of beginning and end of the peak signal on the chromosome and can drop the other columns of the file (we know that the peaks have already been filtered for significance and can thus ignore signal strength).

Since the experimental data corresponds to the true labels of TF binding sites, they have to be labeled to compare against the model's output. This means we have to create a vector of true labels with a label for each prediction from the model, where the positive label `1` is assigned each time the ChIP-seq data shows a peak, and the negative label `0` elsewhere.

Evaluating predictions

The predictions are kept as normalized probabilities for constructing the ROC curves. For each TF, the prediction is compared to the experimentally found true labels for a range of thresholds using the sci-kit learn function `sklearn.metrics.roc_curve`. The true positive rate (TPR) is then plotted against the false positive rate (FPR) for all TFs. The AUC is also evaluated.

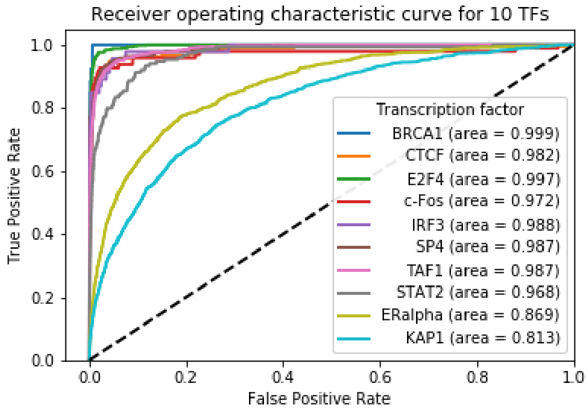

Figure 4: ROC curve for the ten sample TFs and associated AUC values.

Result and discussion

We obtain Figure 4 for the ten TFs analyzed with DeepSEA. The black dashed line represents an AUC of 50% and corresponds to a randomly guessing model. Table III lists the reported AUC values versus the ones we obtained, as well as the absolute difference of the two.

TF	reported	obtained	abs. diff.
BRCA1	0.997	0.999	+0.2 %
CTF	0.996	0.982	−1.4 %
E2F4	0.991	0.997	+0.6 %
c-Fos	0.989	0.972	−1.7 %
IRF3	0.987	0.988	+0.1 %
SP4	0.975	0.987	+1.2 %
TAF1	0.964	0.987	+2.3 %
STAT2	0.946	0.968	+2.2 %
ERα	0.877	0.869	−0.8 %
KAP1	0.817	0.813	−0.4 %

Table III: Author's reported and our obtained AUC values for the 10 sample TFs, and absolute difference.

Overall the resulting Figure 4 is very similar to the figure we intended to reproduce (Figure 2). Our ROC curves seem more rugged, which could indicate that the authors tested more thresholding values. The order of TFs was chosen to reflect the performance reported by the authors (TF BRCA1 had the highest AUC, KAP1 the lowest of the ten TFs, see Table III), but our results no longer show the exact same ordering of performance. Our obtained AUC values are however all within 3% of the originally reported AUC values. The discrepancy could be due to the experimental ChIP-seq data being updated since 2015, or because they used both chromosome 8 and 9 for testing their model, whereas we only use chromosome 8, or further because we used fewer thresholding values for evaluating the AUC. We conclude that the performance reported by the authors for transcription factors is confirmed by our analysis.

In silico saturated mutagenesis

DeepSEA has an inbuilt tool called Sequence Profiler. It is an extension of the model that mutates each base of a sequence and predicts the effect on the binding affinity of a TF to the central basepairs of the sequence, a process named *in silico* saturated mutagenesis analysis. To see if the model has learnt the TF binding motifs, we test the DeepSEA sequence profiler with a known TF binding site inserted centrally into a random sequence of 1000bp.

Sequence logo of E2F4

We choose the TF E2F4 already used in the earlier analysis of the model. The sequence logo for this transcription factor is shown in Figure 5, according to the database Jaspar. If for a position of the logo only one letter is shown, then the TF is only able to bind if this letter is present on that position (it is essential for binding). Thus we expect a high loss of binding affinity if one of those positions is mutated [4].

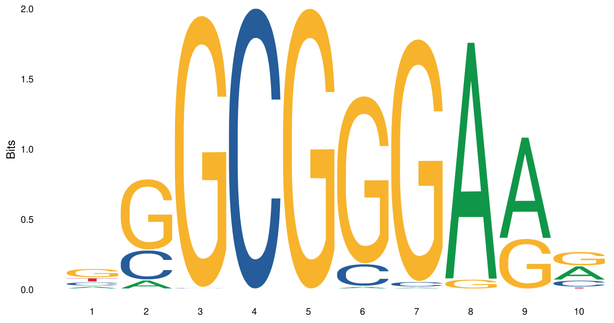

Figure 5: Sequence logo of TF E2F4

Next a random genetic sequence is chosen of length 1000bp, and the central 10 bases are replaced with the consensus motif from Figure 5. Then the sequence is given to the DeepSEA sequence profiler for analysis, with the TF E2F4 selected, cell line GM12878 and no culture treatment.

Sequence profiler results and discussion

We obtain the resulting mutagenesis analysis in form of Figure 6. The three rows from bottom to top represent the three possible base substitutions following $A \to G \to C \to T \to A$ order. For example, if the original sequence has base G, then the three rows from bottom to top represent C, T, and A. The $\log_2$ fold changes of odds ($\frac{P}{1-P}$) are shown using a heatmap.

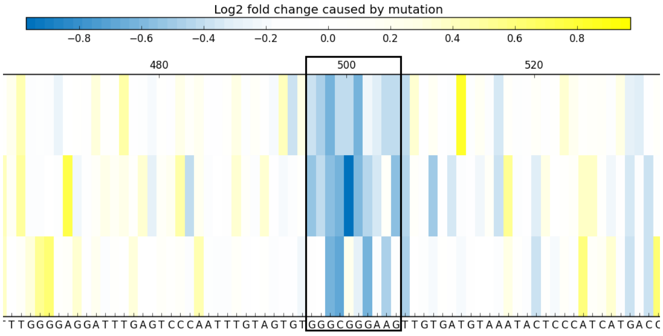

Figure 6: DeepSEA sequence profiler result for E2F4 in GM12878 cell line without culture treatment. Consensus motif from sequence logo is marked.

The blue parts of Figure 6 designate base substitutions that lead to a decrease in binding affinity. For comparison: Position 500 in Figure 6 is the same G as position 5 of the sequence logo (Figure 5). Clearly the sequence logo is marked in blue for substitutions, meaning that the substitution leads to a decrease in affinity. This was expected for changing the essential parts of the TF binding motif. We further notice that no yellow is visible in the marked range, implying the present sequence is optimal for binding. It can be pointed out that some mutations aren't caught by the model as being disruptive for binding (eg. $G \to C$ in position 500 should not be possible according to the sequence logo, but DeepSEA predicts an unchanged binding affinity). Furthermore DeepSEA predicts a loss of affinity for any mutation of base T in position 506, which would correspond to position 11 in the sequence logo, but the sequence logo does not include an 11th position. We conclude that DeepSEA has learnt the majority of important bases for this TF correctly.

Predicting disease-associated phenotypes

The authors report they are able to predict how TF binding affinity is impacted by a point mutation. They provide positive results for several pathology-associated variants known to disrupt or create TF binding sites. We now intend to test DeepSEA on such single nucleotide polymorphisms (SNPs) found in literature.

SNP rs2238631

A variant of the thromboxane A2 receptor (TBXA2R) gene has been correlated with childhood-onset asthma in Asians [5]. SNP rs2238631 has been identified as linked to the condition by a genome-wide association study (GWAS). The variant is shown to change the binding site of several TFs, of which ELK1's only binding site for that gene lies on the minus strand (see supplementary material). We find the variant in the genome browser and extract the 500bp on the left and right of it, and then run the reverse complement of the sequence through the DeepSEA sequence profiler, checking for binding affinity for TF ELK1.

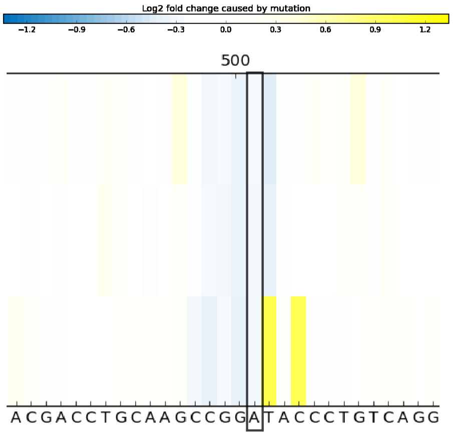

Figure 7: Mutagenesis screen of asthma variant rs2238631 for ELK1.

Figure 7 shows several affinity-decreasing mutations around base 500, but only a slight decrease for variant rs2238631 (framed in black): This variant is an $A \rightarrow G$ mutation in position 501, and the $\log_2$ fold change predicted by DeepSEA is around -0.5. Since this result could be considered not significant, we retry the analysis for another disease-associated variant.

Treacher Collins syndrome

Treacher Collins syndrome (TCS) is an autosomal dominant craniofacial malformation caused by null mutations in the TCOF1 gene. An SNP was shown to be functionally related, as it decreased the promoter activity by 38%. Electrophoretic mobility shift assay (EMSA) analysis was then performed, a common affinity electrophoresis technique used to study protein–DNA interactions. This procedure can determine if a protein or mixture of proteins is capable of binding to a given DNA sequence. It demonstrated that the variant allele impairs DNA-binding to the YY1 transcription factor [6]. This study does not mention the name of the variant they tested for, but instead shows the sequence in one of their figures ($C \rightarrow T$ at 346bp upstream of TCOF1). The 51bp sequence could then be identified in the genome browser using BLAST, and the SNP was located at position 149,736,964 on chromosome 5. The location was then centered in a 1000bp sequence and analyzed by the DeepSEA sequence profiler for binding affinity for TF YY1.

According to [6], there should be a marked decrease in YY1 binding on position 149,736,964 for an $A \rightarrow C$ substitution. The DeepSEA prediction however marks no predicted change for any substitution at that location (framed in Figure 8): the whole column is white, indicating no effect on binding affinity.

SNP rs4784227

After the ambiguous results obtained above for two independent conditions, we checked the original claims by [1] to see if we could at least reproduce the mutagenesis results for the SNPs they name. The authors report the following on variant analysis in their paper:

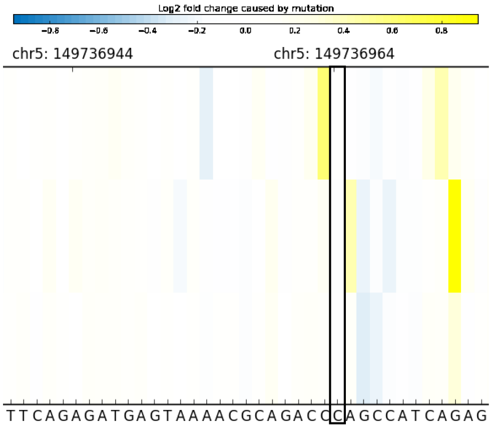

Figure 8: Mutagenesis screen of TCS-associated variant on YY1

> … for the breast cancer risk locus SNP rs4784227, we identified the increased affinity of FOXA1 as the strongest effect of C-to-T alteration consistently in all five cell types for which we have learned predictors for FOXA1.

SNP rs4784227 is located on position 52,599,188 on chromosome 16, and associated with breast cancer [7]. When feeding DeepSEA the sequence including the SNP and the reference sequence, we obtain the following normalized TF binding probabilities for FOXA-1, for different cell lines:

TF (cell line, culture treatment)	rs4784227	reference	factor
FOXA-1 (A549, DEX100nM)	17.76 %	13.08 %	1.36
FOXA-1 (HepG2, None)	16.13 %	11.49 %	1.40
FOXA-1 (HepG2, None)	14.42 %	10.03 %	1.44
FOXA-1 (T-47D, DMSO0.02pct)	14.33 %	10.17 %	1.41

Table IV: DeepSEA normalized binding probabilities for rs4784227 and reference genome, and increase factor from reference to variant.

While Table IV does show an increase in binding probability, this increase might not be significant, with an average of 1.4X fold change. FOXA-1 (HepG2, None) appears twice in the table because it also appears twice in DeepSEA, without clear distinction between the entries.

Discussion

For this subtask, the results were not as conclusive as for the other analyses we carried out. For two disease-associated SNPs, DeepSEA was not able to predict the impact on binding affinity reported in the literature. We have to however be cautious with the asthma variant, as this variant was identified through GWAS analysis alone and the effect on ELK1 binding not experimentally validated. But even for the variant mentioned in the original paper DeepSEA did not perform too convincingly.

Conclusion

We have analyzed the DeepSEA software developed by Zhou & Troyanskaya and tested it on both their own data and for new analysis. We confirm that DeepSEA is very useful in predicting binding affinity for a range of chromatin features, of which we concentrated on a select few TFs, and obtain similar results as the authors. The software's sequence profiler tool performs computational saturated mutagenesis analysis and could be successfully used to determine the consensus motif of a TF. We then tried to reproduce findings from literature where a genetic variant disrupts TF binding, but did not get conclusive results. This might be due to the literature using a different genome assembly or DeepSEA not generalizing well enough. Either way, the software has the advantages of being very convenient and free of charge and can be used to gain a quick overview of possible binding sites in any sequence. It also shows that the amount of data currently available is too limited to make the model broadly applicable.

References

[1] Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nature methods, 12(10), 931-934.

[2] Ma, W., & Wong, W. H. (2011). The analysis of ChIP-Seq data. In Methods in enzymology (Vol. 497, pp. 51-73). Academic Press.

[3] Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7), 1145-1159.

[4] Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic acids research, 18(20), 6097-6100.

[5] Buroker, N. E. (2014). TBXA2R rSNPs, transcriptional factor binding sites and asthma in Asians. Open Journal of Pediatrics, 2014.

[6] Masotti, C., Armelin-Correa, L. M., Splendore, A., Lin, C. J., Barbosa, A., Sogayar, M. C., & Passos-Bueno, M. R. (2005). A functional SNP in the promoter region of TCOF1 is associated with reduced gene expression and YY1 DNA–protein interaction. Gene, 359, 44-52.

[7] Cowper-Sal, R., Zhang, X., Wright, J. B., Bailey, S. D., Cole, M. D., Eeckhoute, J., ... & Lupien, M. (2012). Breast cancer risk–associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. Nature genetics, 44(11), 1191.