

PREDICTING CONTEXT SPECIFIC  
ENHANCER-PROMOTER  
INTERACTIONS FROM CHIP-SEQ  
TIME COURSE DATA

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER  
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY  
IN THE FACULTY OF BIOLOGY MEDICINE AND HEALTH

2016

By  
Tomasz Dzida  
School of Biological Sciences

# Contents

<b>Abstract</b>	<b>22</b>
<b>Declaration</b>	<b>23</b>
<b>Copyright</b>	<b>24</b>
<b>Acknowledgements</b>	<b>25</b>
<b>1 Introduction</b>	<b>26</b>
1.1 Thesis outline . . . . .	27
1.2 Reproducibility . . . . .	28
<b>2 Background</b>	<b>30</b>
2.1 Transcription and its Regulation . . . . .	30
2.2 Nuclear Receptors and ER- $\alpha$ . . . . .	31
2.3 Topologically Associating Domains . . . . .	34
2.4 Discovery of protein-DNA TF binding by ChIP-seq . . . . .	35
2.5 RNA-seq, eRNA, and cellular program . . . . .	38
2.6 Discovery of elongating polymerase - GRO-seq . . . . .	39
2.7 Empirical discovery of chromatin interactions . . . . .	40
2.8 Bayesian Inference . . . . .	43
2.8.1 Parameter inference . . . . .	44
2.8.2 MCMC algorithms . . . . .	45
2.8.3 Metropolis-Hastings . . . . .	46
2.8.4 Gibbs Sampler . . . . .	46
2.8.5 Convergence of Gibbs sampler . . . . .	47
2.8.6 Supervised and unsupervised learning . . . . .	49
2.8.7 Naive Bayes . . . . .	50

2.8.8	Mixture of Gaussians and hidden variable models . . . . .	50
2.8.9	Kernel Density Estimation . . . . .	51
2.9	Related work . . . . .	51
<b>3</b>	<b>ChIP-seq Time Series Data Processing</b>	<b>54</b>
3.1	The experiment - studying ER- $\alpha$ responsive genes . . . . .	54
3.1.1	Alignment to a reference human genome . . . . .	56
3.1.2	Discovery of ER- $\alpha$ bindings . . . . .	56
3.1.3	ChIP-seq time series data . . . . .	58
3.2	Clustering Pol-II and ER- $\alpha$ time series data . . . . .	59
3.2.1	Linking Pol II dynamics with TF occupancy . . . . .	61
3.3	Gene-enhancer links confirmed by ChIA-PET . . . . .	64
3.3.1	Comparison of ChIA-PET confirmed and unconfirmed enhancers	67
3.3.2	Links within and outside domains . . . . .	67
3.4	Discussion . . . . .	67
<b>4</b>	<b>Supervised learning</b>	<b>70</b>
4.1	Enhancer-centric Naive Bayes model . . . . .	71
4.1.1	The definition of the model . . . . .	71
4.1.2	Approximate joint likelihood . . . . .	72
4.1.3	Posterior enhancer-gene allocations . . . . .	74
4.2	Training of the model . . . . .	74
4.2.1	Set of data-specific correlations and genomic distances . . . .	75
4.2.2	Set of interacting and non-interacting pairs . . . . .	75
4.2.3	Training, test and predictive sets . . . . .	75
4.2.4	Kernel density estimation of distributions of attributes . . . .	76
4.3	Evaluation of the model . . . . .	76
4.3.1	Precision-Recall curves . . . . .	77
4.3.2	MAP curves . . . . .	77
4.3.3	The test and training errors . . . . .	77
4.3.4	Performance within and outside TADs . . . . .	78
4.3.5	Validation of model's predictions with GRO-seq and RNA-seq data . . . . .	78
4.4	Results . . . . .	79
4.4.1	Time series correlation and distance-based features are infor- mative about enhancer-promoter interactions . . . . .	79

4.4.2	Fisher's Linear Discriminant performance . . . . .	80
4.4.3	Naive Bayes classifier performance . . . . .	80
4.4.4	Inter-domain and Intra-domain predictions . . . . .	82
4.4.5	Alternative data processing strategies . . . . .	83
4.4.6	Naive Bayes predicts ER- $\alpha$ -regulated transcriptionally active genes . . . . .	86
4.4.7	GO enrichment . . . . .	86
4.5	Summary . . . . .	90
<b>5</b>	<b>Unsupervised learning</b>	<b>93</b>
5.1	Gene-centric Latent Variable Allocation model . . . . .	93
5.1.1	Generative Latent Variable model . . . . .	94
5.1.2	Latent Variable Allocation with a Dirichlet prior . . . . .	96
5.1.3	Latent Variable Allocation with a separation-based prior . . . . .	97
5.1.4	Estimation of the separation-based prior with ChIA-PET data . . . . .	98
5.1.5	Inference with a Gibbs sampler . . . . .	98
5.1.6	Estimation of the frequency of enhancer-gene contacts . . . . .	101
5.2	Results . . . . .	102
5.2.1	Model inference - technical details . . . . .	102
5.2.2	Performance . . . . .	103
5.2.3	Testing the effect of ER- $\alpha$ enhancers with unknown status . . . . .	104
5.3	Summary . . . . .	104
<b>6</b>	<b>Conclusions</b>	<b>111</b>
6.1	Achieved Results . . . . .	111
6.2	Limitations of the models . . . . .	112
6.3	Limitations of the data . . . . .	112
6.4	Comparison between NB and LVA and study of bias . . . . .	113
6.5	Future work . . . . .	114
	<b>Bibliography</b>	<b>117</b>
	<b>Appendix A Supplementary figures for Chapter 3</b>	<b>132</b>
	<b>Appendix B Supplementary figures for Chapter 4</b>	<b>143</b>



# List of Tables

3.1	The table shows the cluster-specific patterns of TF bindings across ER- $\alpha$ enhancers for the corresponding Pol II clusters in figure 3.4a. The count column indicates the size of the cluster. The amplitude column shows the difference between Pol II occupancies in E2-deprived (0 min) and E2-stimulated samples (40 min) in the mean dynamics of each cluster. (dynamics of orange replicate) . . . . .	65
3.2	The table shows the cluster-specific patterns of TF bindings across 2000bp-long TSS-centred regions for the corresponding Pol II clusters in figure 3.4c. The count column indicates the size of the cluster. The amplitude column shows the difference between Pol II occupancies in E2-deprived (0 min) and E2-stimulated samples (40 min) in the mean dynamics of each cluster. (dynamics of orange replicate) . . . .	66
4.1	Table shows the number of predicted links by distance-alone and distance-assisted models and varying test errors. . . . .	82
4.2	The table shows significantly enriched GO classes for drugs and diseases for the predicted targets of ER- $\alpha$ enhancers, FDR = 0.25. . . . .	90
4.3	The table shows significantly enriched GO classes for biological processes for the predicted targets of ER- $\alpha$ enhancers, FDR = 0.25. . . . .	91
A.1	The table shows the cluster-specific patterns of TF bindings across 300bp-upstream-extended-genes for the corresponding Pol II clusters in figure 3.4c. The count column indicates the size of the cluster. The amplitude column shows the difference between Pol II occupancies in E2-deprived (0 min) and E2-stimulated samples (40 min) in the mean of each cluster. (dynamics of orange replicate) . . . . .	136

A.2	The table shows the cluster-specific patterns of TF bindings across enhancers for the corresponding ER- $\alpha$ clusters in figure 3.4b. The count column indicates the size of the cluster. The amplitude column shows the difference between ER- $\alpha$ occupancies in E2-deprived (0 min) and E2-stimulated samples (40 min) in the mean dynamics of each cluster.	137
A.3	The table shows the cluster-specific patterns of TF bindings across 2000bp-long TSS-centred regions (around promoters) for the corresponding ER- $\alpha$ clusters in figure 3.4d. The count column indicates the size of the cluster. The amplitude column shows the difference between ER- $\alpha$ occupancies in E2-deprived (0 min) and E2-stimulated samples (40 min) in the mean dynamics of each cluster. . . . .	138
A.4	The table shows the cluster-specific patterns of TF bindings across across 300bp-upstream-extended-genes for the corresponding ER- $\alpha$ clusters in figure 3.4d. The count column indicates the size of the cluster. The amplitude column shows the difference between ER- $\alpha$ occupancies in E2-deprived (0 min) and E2-stimulated samples (40 min) in the mean dynamics of each cluster. . . . .	139
A.5	The table shows the cluster-specific patterns of TF bindings across enhancers for the corresponding joint Pol II and ER- $\alpha$ clusters in figure A.4. The count column indicates the size of the cluster. The amplitude column shows the difference between Pol II occupancies in E2-deprived (0 min) and E2-stimulated samples (40 min) in the mean Pol II dynamics of each cluster. . . . .	141
A.6	The table shows the cluster-specific patterns of TF bindings across enhancers for the corresponding joint Pol2 and ER- $\alpha$ clusters in figure A.4. The count column indicates the size of the cluster. The amplitude column shows the difference between ER- $\alpha$ occupancies in E2-deprived (0 min) and E2-stimulated samples (40 min) in the mean ER- $\alpha$ dynamics of each cluster. . . . .	142
B.1	Significant GO terms for drugs . . . . .	164
B.2	The tables show the gene ontology annotations for predicted ER- $\alpha$ regulated genes and three categories: biological process, disease, and drugs. The genes consist of NB predicted targets of ER- $\alpha$ distal enhancers (enhancer-gene links with FDA of 0.25) and the genes with an intra-genic ER- $\alpha$ binding. . . . .	164

B.3 The tables show the gene ontology annotations for predicted ER- $\alpha$  regulated genes and three categories: biological process, disease, and drugs. The genes consist of NB predicted targets of ER- $\alpha$  distal enhancers (enhancer-gene links with FDA of 0.25) and the genes with an intra-genic ER- $\alpha$  binding. . . . . 165



# List of Figures

2.1	(a) shows the most common genomic elements involved in transcription. ChIP-seq method (Section 2.4) measures genome-wide occupancy of a selected protein. (b) shows the loop mechanism involved in transcriptional regulation in Eukaryotes. The mechanism enables spacial contacts of distal enhancers, promoters, and other loci and interactions of their associated proteins. Chromatin Conformation Capture experiments (Section 2.7) allow identification of spatially (in 3D) proximal loci. ChIA-PET allows identification of all loci whose contact is mediated by a particular protein. The figure was adapted from [6]. . . . .	32
3.1	Preprocessing pipeline . . . . .	55
3.2	The upper part of the figure shows the ChIP-seq data across GREB region 5 min after the incorporation of E2 to the estrogen deprived MCF7 cell-line. The peaks correspond to putative ER- $\alpha$ bindings. The second row shows ChIP-seq data of 0 min unstimulated sample. The third row shows the corresponding local chromatin interactions mediated by ER- $\alpha$ and captured by ChIA-PET. A large fraction of the ER- $\alpha$ bindings coincide with links from ChIA-PET. . . . .	58
3.3	Cartoon shows the process of merging individual MACS-called peaks with the objective of finding approximate locations of time persistent ER- $\alpha$ bindings. In the process MACS-detected time varying peaks from [0], 5, . . . , 320 min time points (0 is optional and by default not included) which co-occur at least twice across time points are merged by union operation to produce the approximate consensus locations of a single binding. The single occurrences of peaks are discarded. . . .	60

3.4	The first column of the figure shows the results of the clustering of joint time series of both Pol II replicates at enhancers (top row) and genes (bottom row) with Affinity Propagation. The second column shows the corresponding clustering for ER- $\alpha$ time series at the enhancer (top row) and gene regions (bottom row). The remaining clusters can be seen in A.1, A.2, A.3. . . . .	62
3.5	The graphs (a, b, c, d, e) show distributions of sums of ChIP-seq tags across time series of ChIA-PET-confirmed interacting enhancers (green) and the ones with unknown status (red) collected across all 23 chromosomes. The tags are calculated over enhancer bodies for ER- $\alpha$ , both Pol II replicates, H2AZ and H3K4me3 ChIP-seq datasets and normalised by the total size of each set. . . . .	68
4.1	The graphs (a, b, c, d) show positive (green) and negative (yellow) distributions of correlations between pairs of time series of 300bp-upstream-extended-gene regions and enhancer bodies for ER- $\alpha$ , Pol II, H2AZ and H3K4me3 collected across all odd chromosomes. The figure (e) shows the distribution of genomic distances between centres of distal enhancers and 300bp-upstream-shifted-TSS of genes. The set of positive and negative pairs was constructed using 300bp-upstream-extended-genes and distal enhancers. . . . .	81
4.2	The cartoon shows the NB inferred interactions in three classes with decreasing test errors of 0.2, 0.25, 0.3 FDR levels, and corresponding lower bound probability cut-offs of 0.66, 0.47, 0.35. The class membership of each predicted link and its confidence level is indicated by its darker/lighter shading of its colour (more/less confident). The green/grey colour indicates whether each predicted link is confirmed/unconfirmed by ChIA-PET. . . . .	83
4.3	Figure shows the performance of the model for training/test data, measured by Precision-TPR and MAP scores, for selected combinations of datasets in the model. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores (posterior probabilities), and values above corresponding thresholds. . . . .	84
4.4	Figure shows the performance of the model for training/test data. As in Fig. 4.3 but only considering interactions within TADs. . . . .	84

4.5	Figure shows the performance of the model for training/test data. As in Fig. 4.3 but only considering interactions spanning two TADs. . . .	85
4.6	The PR curves assess the performance of Naive Bayes and proximity criterium on the ability to predict differentially expressed genes as detected by GRO-seq experiment and extracted at 3 different q-values (confirdence levels) of 0.001, 0.01, 0.05. The predictor for the transcriptional activity of each gene of the NB predicted ER- $\alpha$ regulated genes is the probability that at least one of its NB assigned distal regulators indeed controls it. The Second predictor for the transcriptional activity of each gene of the genes with at least one ER- $\alpha$ binding within 40kB from their canonical TSS is the absolute value of the distance (proximity) between its canonical TSS and its closest ER- $\alpha$ binding. .	87
4.7	The PR curves assess the performance of Naive Bayes and proximity criterium on the ability to predict differentially expressed genes as detected by RNA-seq experiment and extracted at 3 different q-values (confirdence levels) of 0.001, 0.01, 0.05. The predictor for the transcriptional activity of each gene of the NB predicted ER- $\alpha$ regulated genes is the probability that at least one of its NB assigned distal regulators indeed controls it. The Second predictor for the transcriptional activity of each gene of the genes with at least one ER- $\alpha$ binding within 40kB from their canonical TSS is the absolute value of the distance (proximity) between its canonical TSS and its closest ER- $\alpha$ binding. .	88
5.1	Directed factor graph representation of: a) standard latent Dirichlet allocation of mixture of Gaussians generating data with independent dimensions and b) our latent variable gene-centric mixture of Gaussians with the separation-based prior. . . . .	95
5.2	Figure shows the distributions of separations between ChIA-PET-detected enhancer-gene interactions (green) and gene allocations for the ChIA-PET-confirmed interacting enhancers sampled from LVA's prior (purple).	99
5.3	Gelman $\hat{R}$ statistics for the samples of $\mathbf{Z}$ generated by the Gibbs Sampler of the LVA model with $\kappa_0 = 1$ , $\alpha_0 = 2$ , $\beta_0 = 2$ . . . . .	103

5.4	The comparison of the performance between the Naive Bayes algorithm and the Latent Variable Allocation models with different parametrisation of $\kappa_0$ on all even chromosomes. The parametrisation of the model in the first row is $\kappa_0 = 1, \alpha_0 = 2, \beta_0 = 2$ , the second $\kappa_0 = 3, \alpha_0 = 2, \beta_0 = 2$ . The Precision-TPR curves show the accuracy for the predictions with the highest 10%, 20%, 30% posterior probabilities. . . . .	105
5.5	Figure shows the performance of the model, as in Fig. 5.4 but only considering interactions within TADs. . . . .	106
5.6	Figure shows the performance of the models, as in Fig. 5.4 but only considering interactions spanning two TADs. . . . .	107
5.7	Comparison of performance between the separation-based prior of Latent Variable Allocation model and distance-only Naive Bayes model. . . . .	108
5.8	Figure shows the comparison of the performance between NB and two LVA models on discovery of ChIA-PET-detected links from odd even chromosomes. LVA in the first row was inferred from time course data of all enhancers. LVA in the second row was inferred from the time course data of only those enhancers with ChIA-PET-confirmed links. . . . .	109
A.1	Figure shows subsequent clusters of Fig. 3.4. The first column of the figure shows the results of the clustering of joint time series of both Pol II replicates at enhancers (top row) and genes (bottom row) with Affinity Propagation. The second column shows the corresponding clustering for ER- $\alpha$ time series at the enhancer (top row) and gene regions (bottom row). . . . .	133
A.2	Figure shows subsequent clusters of Fig. 3.4. The first column of the figure shows the results of the clustering of joint time series of both Pol II replicates at enhancers (top row) and genes (bottom row) with Affinity Propagation. The second column shows the corresponding clustering for ER- $\alpha$ time series at the enhancer (top row) and gene regions (bottom row). . . . .	134
A.3	Figure shows subsequent clusters of Fig. 3.4. The first column of the figure shows the results of the clustering of joint time series of both Pol II replicates at enhancers (top row) and genes (bottom row) with Affinity Propagation. The second column shows the corresponding clustering for ER- $\alpha$ time series at the enhancer (top row) and gene regions (bottom row). . . . .	135

A.4	The figure shows the clustering of the joint time course of Pol II and ER- $\alpha$ at enhancers with Affinity Propagation. The clustering involves only the time series which individually possess a sum of at least 100 tags across all time point. . . . .	140
B.1	The graphs show positive (blue) and negative (red) class size-normalised histograms of projections of the training set of Fisher's linear discriminant analysis when applied on five selected variants of vectors of features (see sub-titles). . . . .	144
B.2	The graphs show positive (blue) and negative (red) histograms of projections of the training set of Fisher's linear discriminant analysis when applied on five selected variants of vectors of features (see sub-titles). Due to a large difference in sizes of the two classes, the histograms of the positives are invisible. . . . .	145
B.3	Precision-Recall curves measuring performance of Fisher's linear discriminant analysis for increasing values of cut-off levels (i.e. from the most negative to the most positive value of projection) for (a) training and (b) test sets. Each colour shows performance of the classifier for a different selected combination of features. . . . .	146
B.4	The graphs (a, b, c, d) show positive (green) and negative (yellow) distributions of correlations between pairs of time series of 300bp-upstream-extended-gene regions and enhancer bodies for ER- $\alpha$ , PolIII, H2AZ and H3K4me3 collected across all 24 chromosomes. The figure (e) shows the distribution of genomic distances between centres of distal enhancers and 300bp-upstream-shifted-TSS of genes. The set of positive and negative pairs was constructed using 300bp-upstream-extended-genes and distal enhancers. . . . .	147
B.5	The graphs (a, b, c, d) show positive (green) and negative (yellow) distributions of correlations between time series of 300bp-upstream-extended-gene regions and enhancer bodies for ER- $\alpha$ , PolIII, H2AZ and H3K4me3 collected across all odd chromosomes (training data). The figure (e) shows the distribution of genomic distances between centres of distal enhancers and 1500bp-upstream-shifted-TSS of genes. The set of positive and negative pairs was constructed using 1500bp-upstream-extended-genes and distal enhancers. . . . .	148

B.6 The graphs (a, b, c, d) show positive (green) and negative (yellow) distributions of correlations between time series of 300bp-upstream-extended-gene regions and enhancer bodies for ER- $\alpha$ , PolIII, H2AZ and H3K4me3 collected across all 24 chromosomes. The figure (e) shows the distribution of genomic distances between centres of distal enhancers and 1500bp-upstream-shifted-TSS of genes. The set of positive and negative pairs was constructed using 1500bp-upstream-extended-genes and distal enhancers. . . . . 149

B.7 The graphs (a, b, c, d) show positive (green) and negative (yellow) distributions of correlations between time series of 300bp-upstream-extended-gene regions and enhancer bodies for ER- $\alpha$ , PolIII, H2AZ and H3K4me3 collected across all odd chromosomes. The figure (e) shows the distribution of genomic distances between centres of distal enhancers and 1500bp-upstream-shifted-TSS of genes. The set of positive and negative pairs for the first model was constructed using TSS-centred 3000bp-long regions and distal enhancers. . . . . 150

B.8 The graphs (a, b, c, d) show positive (green) and negative (yellow) distributions of correlations between time series of 300bp-upstream-extended-gene regions and enhancer bodies for ER- $\alpha$ , PolIII, H2AZ and H3K4me3 collected across all 24 chromosomes. The figure (e) shows the distribution of genomic distances between centres of distal enhancers and 1500bp-upstream-shifted-TSS of genes. The set of positive and negative pairs for the first model was constructed using TSS-centred 3000bp-long regions and distal enhancers. . . . . 151

B.9 Figure shows the comparison of performance of the NB model on odd chromosomes (training data) measured by Precision-TPR and MAP scores. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using 300bp-upstream-extended-genes and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 300bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers. . . . . 152

B.10 Figure shows the comparison of performance of the NB model on even chromosomes (test data) measured by Precision-TPR and MAP scores. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using 300bp-upstream-extended-genes and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 300bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers. . . . . 153

B.11 Figure shows the comparison of performance of the NB model on odd chromosomes (training data) measured by Precision-TPR and MAP scores. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using 1500bp-upstream-extended-genes and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 1500bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers. . . . . 154

B.12 Figure shows the comparison of performance of the NB model on even chromosomes (test data) measured by Precision-TPR and MAP scores. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using 1500bp-upstream-extended-genes and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 1500bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers. . . . . 155



B.13 Figure shows the performance of the TSS-centric NB model on odd chromosomes (training data) measured by Precision-TPR and MAP scores. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using TSS-centred 3000bp-long regions and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 1500bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers. . . . . 156

B.14 Figure shows the performance of the TSS-centric NB model on even chromosomes (test data) measured by Precision-TPR and MAP scores. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using TSS-centred 3000bp-long regions and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 1500bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers. . . . . 157

B.15 Figure shows the comparison of performance of the NB model between odd and even chromosomes (training and test data) measured by Precision-TPR and MAP scores for selected combinations of datasets. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using 300bp-upstream-extended-genes and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 300bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers. . . 158

B.16 Figure shows the comparison of performance of the NB model between odd and even chromosomes (training and test data) measured by Precision-TPR and MAP scores for selected combinations of datasets. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using 1500bp-upstream-extended-genes and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 1500bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers. . . . . 159

B.17 Figure shows the comparison of performance of the TSS-centric NB model between odd and even chromosomes (training and test data) measured by Precision-TPR and MAP scores for selected combinations of datasets. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using TSS-centred 3000bp-long regions and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 1500bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers. . . . . 160

B.18 Figure shows the performance of the NB model of training data and for selected combinations of datasets under two different parametrisations of MACS peak-calling. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The first column of the figure shows the performance of the NB model trained on the stringent time persistent merged MACS-called peaks (i.e. distal ER- $\alpha$  bindings) from the scan with the  $p$ -value of  $1e-11$  and the local control switched off, in which case the search is done with  $\lambda_{BG}$ . In the second column we see the performance under the alternative peak calling with the  $p$ -value of  $1e-05$  (MACS' default), no control and the local control flag on. The set of positive and negative pairs for the first model was constructed using 1500bp-upstream-extended-genes and distal enhancers. The correlation-based attributes of the model were estimated using pairs of 300bp-upstream-extended-genes, and enhancers (merged distal MACS-called peaks). The separation-based from 1500bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers. 161

- B.19 Figure shows the comparison of the performance between promoter-extended-gene and TSS-centric models on odd chromosomes (training data) measured by Precision-TPR and MAP scores and for selected datasets. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using 1500bp-upstream-extended-genes and distal enhancers, whereas for the second using TSS-centred 3000bp-long regions and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 1500bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers. . . . . 162
- B.20 The PR curves show the comparison of the performance between the 1500bp-upstream-extended-gene and 1500bp-TSS-centric Naive Bayes models on their ability to predict differentially expressed genes. The positive set consists of GRO-seq-detected genes for 3 different confidence levels of 0.001, 0.01, 0.05. Score of each tested gene is a cumulative posterior probability of the set of its NB-predicted regulators. The graph also shows the accuracy of using an absolute value of the separation (proximity) between its closest ER- $\alpha$  binding and a canonical TSS of a gene as a predictor of gene activity. Gene was regarded as ER- $\alpha$  regulated if the distance to its putative regulator was within 40kB. 163
- C.1 Gelman  $\hat{R}$  statistics for the samples of  $Z$  generated by the Gibbs Sampler of the LVA model inferred from the time series of the enhancers with ChIA-PET evidence and parameters  $\kappa_0 = 1$ ,  $\alpha_0 = 2$ ,  $\beta_0 = 2$ . . . . 167
- C.2 Figure shows the comparison of the performance between NB and two LVA models on discovery of ChIA-PET-detected links from all odd chromosomes. LVA in the first column was inferred from time course data of all enhancers. LVA in the second column was inferred from the time course data of only those enhancers with ChIA-PET-confirmed links. . . . . 168

C.3 Figure shows the comparison of the performance between NB and two LVA models on discovery of ChIA-PET-detected links from all even chromosomes. LVA in the first column was inferred from time course data of all enhancers. LVA in the second column was inferred from the time course data of only those enhancers with ChIA-PET-confirmed links. . . . . 169

# Abstract

## PREDICTING CONTEXT SPECIFIC ENHANCER-PROMOTER INTERACTIONS FROM CHIP-SEQ TIME COURSE DATA

Tomasz Dzida

A thesis submitted to the University of Manchester  
for the degree of Doctor of Philosophy, 2016

We develop machine learning approaches to predict context specific enhancer-promoter interactions using evidence from changes in genomic protein occupancy over time. Occupancy of estrogen receptor alpha (ER- $\alpha$ ), RNA polymerase (Pol II) and histone marks H2AZ and H3K4me3 were measured over time using ChIP-Seq experiments in MCF7 cells stimulated with estrogen.

Two Bayesian classifiers were developed, unsupervised and supervised. The supervised approach uses the correlation of temporal binding patterns at enhancers and promoters and genomic proximity as features and predicts interactions. The method was trained using experimentally determined interactions from the same system and achieves much higher precision than predictions based on the genomic proximity of nearest ER- $\alpha$  binding. We use the method to identify a confident set of ER- $\alpha$  target genes and their regulatory enhancers genome-wide. Validation with publicly available GRO-Seq data shows our predicted targets are much more likely to show early nascent transcription than predictions based on genomic ER- $\alpha$  binding proximity alone.

Accuracy of the predictions from the supervised model was compared against the second more complex unsupervised generative approach which uses proximity-based prior and temporal binding patterns at enhancers and promoters to infer protein-mediated regulatory complexes involving individual genes and their networks of multiple distant regulatory enhancers.

# Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

# Copyright

- i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.
- ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.
- iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.
- iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see <http://documents.manchester.ac.uk/display.aspx?DocID=24420>), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see <http://www.library.manchester.ac.uk/about/regulations/>) and in The University’s policy on presentation of Theses



# Acknowledgements

Firstly, I would like to thank Prof. Magnus Rattray for great supervision, support and encouragement, lots of constructive criticism and uncountable amount of good sarcastic jokes. Besides I would like to thank my lab mates. Peter for showing me how to set up ssh server without which the study would never be finished. Panos, Rebecca and Lijing for being great drinking companions. Jing for her everlasting sense of humour. Maria for our countless enriching discussions. Mudassar and Ciira for good collaboration and advice. All my friends, especially Wojciech "The Chemist", and my irreplaceable bicycle without which my time in Manchester would not be the same. The University of Manchester for the funding. I would also like to thank my family, especially my mum, for all their support and good word. Lastly, Friedrich Nietzsche for his quote: *"What doesn't kill you make you stronger"*.

# Chapter 1

## Introduction

Gene expression requires the binding of transcription factor (TF) proteins to genomic regions which regulate transcriptional initiation [83]. In eukaryotic cells these regulatory genomic regions are referred to as promoters and enhancers and their activity is associated with specific changes to chromatin such as histone modifications [11, 7, 138, 111]. Due to the mechanism of action as well as the cell- and context-specific activity of enhancers, their location and associated target genes are however difficult to predict. Specifically, the enhancers can act upstream or downstream of their target gene promoters and are often distal, separated by large inter-genic regions [100, 97, 105]. Their enhancer-promoter interactions require protein-mediated physical contacts through formation of chromatin loops [120]. The contacts can affect the rate of transcription and may be associated with paused RNA polymerase (Pol II) [38]. Some enhancers can mediate contacts between promoters and the body of the genes [2, 32]. For instance in [61], enhancers were shown to connect promoter, gene-bodies and follow a precise location of actively elongating Pol II. Although most contacts are intra-chromosomal, there are some interactions between loci from different chromosomes [34, 63, 64]. Transcriptional regulation may be mediated by large multi-gene and multi-enhancer complexes [34, 64].

Recent progress in experimental techniques such as ChIA-PET, 3C and its derivatives 4C, 5C, and Hi-C [34, 23, 41, 137, 29, 108, 124, 82, 52] have mapped large numbers of chromatin interactions including enhancer-promoter interactions. However, these methods are technically challenging and genome-wide methods such as Hi-C typically lack the genomic resolution required to identify individual interacting enhancer elements. Some methods are also thought to produce a high false negative rate (in case of ChIA-PET, 5C) [64, 46] or cannot be applied on a genome-wide scale

(3C, 4C) [108]. Capture-HiC methods have recently been developed [78] to improve genomic resolution through focussing on pre-determined genomic regions, e.g. promoters or known enhancers, and show promise but are not yet widely used. Data from these technologies can also be noisy and subject to various sources of bias which can be difficult to correct for [124]. In addition, the physical contact between two chromatin regions does not determine a functional interaction [106] and the stimulus-dependant behaviour of chromatin looping adds another layer of complexity [30, 123].

To allow for a better insight and correct for some of these pitfalls, experimental methods can be usefully complemented by computational approaches that exploit other more readily available sources of genomic data such as ChIP-seq and RNA-seq data [87, 73]. The existing methods, however, typically require data from multiple cell-types or tissues and therefore do not allow discovery of interactions given data from one cell-type. The methods also do not take into account cell-type specific evidence from time course ChIP-seq data across gene bodies. Most existing methods also assume a stringent maximum distance constraint and are therefore unable to discover distal links beyond this constraint.

In contrast, in this thesis we investigate whether cell-line specific ChIP-seq time course data measuring TF and RNA polymerase occupancy changes after a cellular stimulation can be used to accurately predict cell-line-specific enhancer-promoter and enhancer-intra-gene interactions within chromosomes. Our hypothesis is that time course data from a single cell-type is sufficient to predict many enhancer-promoter interactions and we develop methods to carry out this task.

## 1.1 Thesis outline

Our analysis involves several steps, and here we summarise the structure of the thesis and outline of the study.

In Chapter 2 we provide a general overview of the gene regulation including the role of regulatory elements such as promoters and enhancers, transcription factors and higher order chromatin topologies. We review experimental and computational methods for discovery of chromatin interactions and their inference. We introduce underlying theoretical aspects of the Bayesian methodology used for the design of our computational methods in the remaining part of the thesis.

In Chapter 3 we describe our pre-processing of multiple time course ChIP-seq

datasets from MCF7 breast cancer cells after the stimulation with estradiol. The analysis includes alignment of raw ChIP-seq reads, TF binding discovery, normalisation of time course data and preliminary clustering of the time series data for the analysis of their dynamics.

In Chapter 4 we propose a supervised classification model that combines evidence from the correlation of ChIP-seq time course data at enhancers and promoters or across gene bodies with genomic separation data, to predict the probability of putative enhancer-gene contacts. We benchmark performance against publicly available ChIA-PET data from the same cell-line and stimulation. We use the method to find gene targets of the ER- $\alpha$  enhancers which lack assignments in ChIA-PET data and provide a highly confident list of directly ER- $\alpha$  regulated genes.

In Chapter 5 we introduce a generative unsupervised model. The model combines genomic separation and shapes of time series of multiple ChIP-seq datasets, to provide us with the posterior probability of enhancer-gene contacts. As opposed to the supervised approach, the model allows for the involvement of multiple enhancers in regulating each gene. We validate the model with ChIA-PET data and compare its performance against the one in the previous chapter.

In Chapter 6 we summarise achieved results and discuss potential limitations of the data and of our models. We also discuss possible future extensions of the proposed approaches.

## 1.2 Reproducibility

In recent years, a lack of reproducibility of many scientific findings has been recognised to be a serious issue. To address the problem in computational science, many journals now require that each submission should be accompanied by the code and data used in the study [28]. Those allow a later verification of scientific findings and potentially further development of alternative methods. Recognising the importance of the practice, we decided to make the research presented in this thesis fully reproducible and satisfying the criteria given by [96].

Our analysis is implemented in Python and all figures and output results can be reproduced by running appropriately named scripts. The supervised Bayesian model in chapter 4 can be run on a standard desktop machine. For the unsupervised model in chapter 5, we recommend using a multi-core node and for each chromosome the model can be run as a separate process. Additionally, thanks to use of numexpr library, the

computations for each process can be run on multiple cores in a cluster. The code relies on the bedtools package, numpy and the numexpr. The package along with scripts is located in GitHub at: [https://github.com/ManchesterBioinference/EP\\_Bayes](https://github.com/ManchesterBioinference/EP_Bayes). The repository also contains a readme file with the list of scripts and their descriptions.

# Chapter 2

## Background

In this section we will provide an insight into the role of various layers of the cellular system which play a part in the control of gene expression in eukaryotic cells. We elucidate how functional DNA sequences, such as promoters and enhancers, higher order structural units and properties of chromatin encode various cellular programs and determine cellular identity. We describe the role of transcription factors in the regulation of transcription. We also summarise and review current experimental and computational methods which identify elements and associations in gene regulatory networks.

### 2.1 Transcription and its Regulation

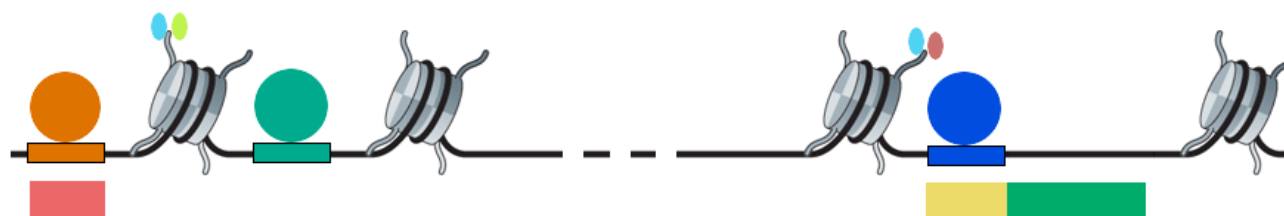
Transcription is the initial process of gene expression which involves a transfer of genetic information from a DNA sequence into an RNA molecule by RNA polymerase. Fig. 2.1a illustrates the most common elements involved in regulation of transcription. The process requires binding of RNA polymerase at a promoter region located upstream of the transcription start site (TSS) of a gene. Proteins are produced from messenger RNA (mRNA) which is synthesised by RNA polymerase II (Pol II) in eukaryotes. The efficiency of transcription is regulated by DNA-sequence-specific transcription factors (TFs), protein molecules which aid or repress the process and whose recognition motif sequences and binding hubs are often located away from the controlled TSS, at so called enhancer regions [132]. Enhancers can be located upstream or downstream of the TSS, in introns of genes or even on different chromosomes from their target promoters [56, 84]. The enhancer-promoter interactions require protein-mediated physical contacts enabled via chromatin loops which bring enhancers in close

proximity of their target gene promoters (see Fig. 2.1b). Mediator proteins play various roles, examples of which are stabilisation of loop structures by cohesin [109] or activation/repression of Pol II by mediator complex [40]. CTCF protein was shown to restrict the action of enhancers and thus can play the role of an insulator [93, 85], however in general it assists in tethering of distant enhancers to their target promoters [85]. Interactions can also exist as part of large multi-gene and multi-enhancer complexes, which enable a coordinated gene expression [34, 64].

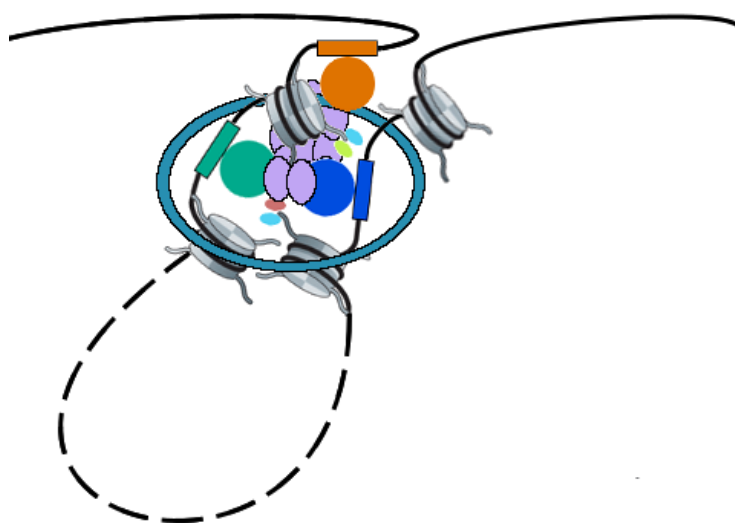
The activity of enhancers, repressors, insulators as well as other functional structures encode a transcriptional program of a cell, and therefore varies between cell-types, tissues, and in time [47]. Regulation of transcriptional activity is achieved by combination of reduction or widening of chromatin accessibility, achieved by action of proteins such as pioneer factors, affecting TF binding in those regions, as well as a pattern of histone modifications (chromatin marks) such as acetylations and methylations of N-terminal tails [134]. Examples include the chromatin marks present in the vicinity of active enhancers and promoters. Although active enhancer regions lack histones, which widens their accessibility to TF bindings, the regions are flanked by histones enriched in H3K4me1 and H3K27ac modifications but displaying low levels of H3K4me3 [47, 15]. In contrast, active promoters are depleted of H3K4me1, show different levels of H3K27ac and are enriched in H3K4me3 [105]. Some of the chromatin marks are recognised by other regulatory proteins, e.g. Tri-methylation of histone 3 lysine 27 H3K27me3 was shown to be associated with silencing of enhancers by polycomb proteins (PcG)[31]. The exact mechanism of action of PcG proteins remains to be elucidated however the proteins are known to be crucial in differentiation and maintenance of cell-type identity [5].

## 2.2 Nuclear Receptors and ER- $\alpha$

The activity and binding location of transcription factors is controlled spatially by recognition motif sequences, epigenetic modifications of chromatin, developmental stage or a presence of extracellular stimuli. The latter type of regulation is particularly vivid in the case of TFs called nuclear receptors which harbour ligand binding domains. The nuclear receptors, in contrast to other receptors, are classified as TFs due to their ability to directly interact with genomic DNA. Although the majority of nuclear receptors only bind DNA in the presence of their ligands, some nuclear receptors display a basal weak ability to bind DNA even in a ligand free environment. In



(a) regulatory elements of gene regulation



(b) loop formation

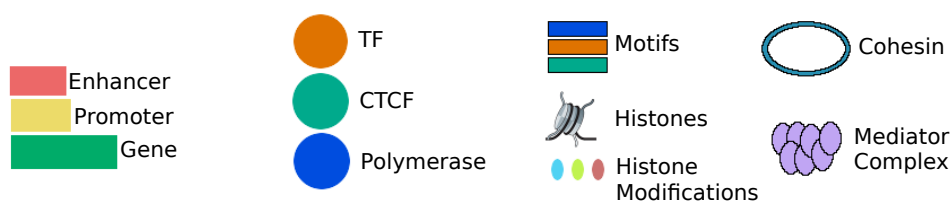


Figure 2.1: (a) shows the most common genomic elements involved in transcription. ChIP-seq method (Section 2.4) measures genome-wide occupancy of a selected protein. (b) shows the loop mechanism involved in transcriptional regulation in Eukaryotes. The mechanism enables spatial contacts of distal enhancers, promoters, and other loci and interactions of their associated proteins. Chromatin Conformation Capture experiments (Section 2.7) allow identification of spatially (in 3D) proximal loci. ChIA-PET allows identification of all loci whose contact is mediated by a particular protein. The figure was adapted from [6].



the event of extracellular stimulation and binding of the ligand to its domain a conformation of the receptor alters. This results in either a formation of new or stabilisation of existing nuclear receptor-DNA bindings, which in consequence, depending on the nature of the ligands, either elevate or suppress transcription of regulated genes.

A particularly well studied example of a nuclear receptor, partly due to its role in breast cancer development, is estrogen-receptor-alpha (ER- $\alpha$ ) encoded by gene ESR1. Its genome-wide binding pattern under stimulation with estrogen has been established through ChIP-seq experiments (see Section 2.4) [68, 72, 94]. The receptor can also exhibit a basal function in a ligand-free environment [14, 112]. Although the actual pattern and strength of bindings under estrogen stimulation is cell-type specific and determined by epigenetic marks and accessibility of chromatin, theoretically the estrogen-bound-receptor could bind to at least 70,000 of its motifs, so called estrogen responsive elements (ERE)[13]. The actual number of possible locations of bindings, however is restricted, cell-type specific, and orchestrated by pioneer and other TFs [51]. Pioneer TFs possess the ability to bind to hetero-chromatin and increase its accessibility [106]. Removal of the pioneer factors results in chromatin compaction and either weakening of binding or total inability of binding at affected locations. Pioneer factors are also involved in chromatin interactions [16].

### **ER- $\alpha$ and Pol II-mediated chromatin loops**

ER- $\alpha$  preferentially binds at distal regions away from the vicinity of their target promoters [17]. The experiments with ChIA-PET (Section: 2.7) show that ER- $\alpha$  takes active part in formation of chromatin loops which bring distal ER- $\alpha$  bound enhancers to their target genes [34]. The loop formation is induced by estrogen which is recruited to ER- $\alpha$  [79]. Additionally, enrichment of Pol II as well as pattern of epigenetic marks at loci involved in loops, suggest that the structures are involved in regulation of transcription [64]. Some of the Pol II contacts were shown to be further involved in Pol II-mediated interactions. Although the majority of the ER- $\alpha$  and Pol II-mediated interactions were intra-chromosomal, some of them also involved loci from different chromosomes [34, 64, 136]. In addition both types of interactions formed duplex and more complex interactions involving several loci and protein mediators. The distribution of distances between interacting loci was far from uniform. According to [34], the majority (86%) of ER- $\alpha$ -mediated links lay within distance of less than 100Kb, 13% from 100Kb to 1Mb and 1% above the region [34]. Among the ER- $\alpha$  bindings which were involved in the chromatin interactions with a promoter, 83% were distal and 17%

proximal to TSS (based on a cut-off of 5kB from the closest TSS). The same ratio hold true also for the ER- $\alpha$  bindings uninvolved in the interactions. Additionally in [17], their genome-wide approach, shows that only 4% of estrogen receptor bindings occur within 1-Kb promoter-proximal regions. Further experiments with 5C technique (Section: 2.7) and comparative analysis of cell-lines showed that the majority of loops are cell-type-specific, which suggests that chromatin loops define a unique transcriptional program of each cell [97].

### **Pioneer factors are present prior to ER- $\alpha$ recruitment and are involved in chromatin interactions**

The first reported pioneer factor which is present at more than a half of ER- $\alpha$  binding sites and involved in estrogen associated loops is FOXA1 [72, 115]. Due to its role in regulation of estrogen associated transcription and modulation of chromatin accessibility of EREs, its malfunctional behaviour is associated with breast cancer [10]. Other pioneer factors present individually or in clusters at ER- $\alpha$  bindings and implicated in ER-mediated loops are AP2 $\gamma$ , PBX1 and GATA3 [54, 71, 118]. The factors increase chromatin openness, collectively bind near estrogen responsive genes and impact their transcription [72].

Other factors such as chromatin modifiers, act on histones around the binding site in the event of estrogen recruitment. These factors may further acetylate or methylate histones or may trigger other histone modification. An example of such factor is P300 [140], a form of histone acetyl transferase (HAT). HAT appears at the sites of ESR1 bindings and plays a role of co-activator, therefore increasing accessibility of chromatin [72]. Binding of P300 is commonly used as a mark for active enhancers [126, 131]. The binding also changes the activity of some of enhancers in the vicinity of differentially expressed genes under estrogen stimulation. Other factors such as Mi2 can act as co-repressors [22].

## **2.3 Topologically Associating Domains**

Recent maps of chromatin interactions have revealed that the majority of chromatin contacts take place within distinct domains called Topologically Associating Domains (TADs) of variable sizes ranging from a few kilobases up to megabase structures, with the average size of 1 Mb [26, 20]. TADs segregate chromatin into non-overlapping

adjacent regions with clear boundaries. A cross-species comparison of the structures shows that the domains are evolutionarily conserved between mammals and even more distant species [26]. The domains differ from each other in their chromatin accessibility, replication time, epigenetic marks, as well as transcriptional activity [90, 25]. In addition, it was shown that the majority of enhancer-promoter interactions occur within domains, thus domains serve as a spatial constraint on the range of enhancer action [75]. The above observations indicate that the domains may function as autonomous regulatory units which provide a spatially restricted compartment for coordinated gene regulation, utilising intra-domain distal regulators. Depending on their epigenetic profile, the domains can be transcriptionally active or silent. The active domains have the characteristics of euchromatin, and are enriched in H3K4me3 [104, 103]. In contrast, the inactive TADs are more heterochromatin-like, and enriched in H3K27me3 [5]. Transcription of some TADS is additionally controlled by their location within the cell. Their association with peripheral nuclear lamina was shown to be linked with transcriptional repression [5]. The exact mechanism of an establishment of the structural units is unknown, although sub-domains within TADs are known to be important in cell differentiation and determine cell-identity [25]. Studies show that the boundaries of TADS are enriched in structural proteins such as cohesin and CTCF proteins [110, 26]. However, CTCF was shown to be present also within TADS [85]. Restrictive domains are also enriched for restrictive forms of PcG proteins [5]. These factors are likely contributing to the overall chromatin conformation. However, a depletion of the putative structural proteins, although altering the frequency of inter-domain interactions, does not abrogate the structures and they are unlikely to be sufficient for maintenance or formation of the structures. Thus, their role is likely supportive, and an interplay between the elements is more complex [139, 85].

## **2.4 Discovery of protein-DNA TF binding by ChIP-seq**

High throughput methods such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) [87], enable a genome-wide *in vivo* detection of antibody specific DNA-protein bindings. ChIP-seq is routinely used to identify the genomic location of TFs, RNA polymerase, and histone modifications. In the method cells are mixed with formaldehyde, which fixes native DNA-protein bindings, by formation of cross-links. The step is followed by sonication which shears the chromatin into short 200-600 bp long fragments. The fragmented chromatin is then immunoprecipitated with a selected

antibody-coated-magnetic beads, to enable an isolation of DNA-bound-protein complexes from the unbound sequences in the process of purification. The cross-links in the purified sample are reversed which results in the separation of DNA fragments from the bound proteins. The fragments may then be processed with subsequent steps such as sequencing, mapping and detection of genome-wide bindings and occupancy of a selected TF or protein.

Treatment of cells with formaldehyde results in cross-linking, i.e. formation of covalent bonds in protein-protein and protein-DNA contacts. Among the cross-linked protein-DNA contacts are bindings of TFs to their recognition sites. Some cross-linked protein-protein contacts, on the other hand, are associated with proteins which either directly interact with the enhancer-bound protein of our interest or are in close proximity to the protein. Due to the nature of treatment with formaldehyde, ChIP-seq method is believed to detect not only directly TF-bound loci but also those in genomic proximity to the selected TF, and thus retain information about chromatin conformation [76] (see Section 2.7).

## Sequencing and Alignment

After purification, the recovered fragments undergo the sequencing procedure which reveals the nucleotide content of the DNA fragments (templates). The most widespread sequencing method is the one provided by Illumina [87]. In this massively parallel technique, after initial size selection, each fragment is consecutively separated and immobilised, amplified by bridge PCR, and sequenced. The sequencing is a cyclic process which consists of incorporation of four colour dyed fluorescence nucleotides, singular extension of each fragment's complementary strand by polymerase, detection of a light emitted after each successful synthesis via a camera device, and the removal of remaining unattached nucleotides. Usually one end of a fragment is sequenced and the cycle is repeated a fixed number of times until a desired read length is achieved. The resultant reads are aligned to a reference genome.

The reads can be aligned to a single location or multiple locations depending on the frequency of appearance of a given read sequence in the reference genome. Due to the vast number of reads produced by the above protocols the process requires a vast amount of processing power and specialised algorithms. Examples of alignment programs are BWA, Bowtie and STAR [65, 59, 27]. In order to reduce the number of mismatches and increase the number of uniquely aligned reads, the reads can be aligned to a reference genome using paired-end reads. In this case, the sequencing

is repeated from the opposite end, and both complementary reads are used for mapping. Sample-specific differences between a reference genome and sequences, such as single nucleotide polymorphisms also need to be addressed. The aligners deal with individual-specific variations by allowing two nucleotides to differ between the reference and 35 bp long reads.

## Peak calling

The real challenge is to locate the precise binding sites of the regulatory proteins and histone modifications. The aligned reads will constitute not only the true signal, but will also contain a substantial amount of noise reads originating from both protein-DNA and antibody-DNA random binding events. Additionally, the enrichment of true signal as well as noise will be affected by factors such as chromatin openness, PCR amplification, nucleic acid isolation, sequencing artefacts or mappability [117, 24, 101]. In terms of mappability bias, it is a common practice to discard the reads which map to multiple locations. Those reads, however, often originate from the regions with frequently occurring sequences. GC content of fragments was also shown to affect read density, i.e fragments with a high GC-content tend to be overrepresented in ChIP-seq fragment pool [9].

The presence of these biases pose a substantial challenge to distinguish the true especially weak signal from noise, and any classification method should take those into account and model their effect on the predictions. Control experiments enable the tracking of the bias, and the types of control, each addressing different bias, are: input DNA, which is obtained by removing a sample of DNA prior to immunoprecipitation, DNA immunoprecipitation with non-specific antibody or DNA immunoprecipitation without antibodies. In general it is assumed that the ChIP treatment and the control samples have identical biases. A treatment set is usually compared against its control set in order to find true TF bindings. However in some cases, for instance, in the problem of determination of differential bindings between experiments performed at different times or under different conditions, and when the biases between experiments are comparable, one of the experiments would serve as a control and the preparation of a control could be avoided.

Several computational methods has been designed for detection of TF bindings (so called peak calling algorithms), each with a different signal to noise model [129, 48, 135]. In this study we use the popular method MACS [135] which uses a Poisson model. The advantage of using a Poisson model over a simple fold enrichment is that

it associates a higher probability to e.g. 400 reads in treatment and 100 reads in control than it would to 40 and 10. The peak searching procedures in MACS is divided into two stages. In the first stage, MACS scales the number of reads in a selected control to be equal to the number of reads in a treatment ChIP-seq experiment and estimates the  $\lambda_{BG}$  parameter which is a ratio of the total number of reads and genome size, thus the expected number of reads per bp. It then shifts a frame of a fixed short size  $d$  along the genome, for each estimates read enrichment and the corresponding  $p$  value, and calls a peak if the  $p$  value is below a user defined (default  $10^{-5}$ ) threshold. The adjacent significantly enriched frames are merged together to form a single peak. In the second stage MACS corrects for the local chromatin structure of each detected peak and biases, using the parameters  $\lambda_{local}$  which is defined as either a maximum of a set of  $\{\lambda_{BG}, \lambda_{1k}, \lambda_{5k}, \lambda_{10k}\}$  if control is used, or  $\{\lambda_{BG}, \lambda_{5k}, \lambda_{10k}\}$  otherwise. The  $\lambda$  is estimated from a window of  $[1kb], 5kb, 10kb$ , centred at the location of maximal read pile up, and captures the above average background enrichment of regions surrounding a peak. The peak is considered significant if its  $p$  value under the new Poisson( $\lambda_{local}$ ) model is below a set threshold (default  $10^{-5}$ ).

## Enhancer discovery

ChIP-seq experiments enable the discovery of the genomic location of transcriptionally relevant proteins such as TFs, RNA polymerase and modified histones. Accordingly, the technique is usually able to detect previously reported enhancers. The method however also has a high false positive rate. Some of the reported events correspond to non-functional enhancers, probably due to a lack of other co-factors at these locations, others to transient bindings caused by general affinity of TFs to even non-specific DNA sequences. One solution to overcome the problem is to investigate dynamics of the bindings of one or ideally also other complementary TFs. Alternatively as in [31, 138] multiple ChIP-seq datasets can be combined with data from other relevant genomic assays to identify active promoters and enhancers using genomic segmentation algorithms.

## 2.5 RNA-seq, eRNA, and cellular program

The set of all messenger RNA produced by the particular cell population defines a transcriptome, which in turn is determined by the transcriptional program of a cell.

Eucaryotic protein-coding genes consists of coding exonic and non-coding intronic regions, which enables alternative splicing and further increase in combinatorial complexity of produced transcripts. RNA-seq is a genome-wide method which enables quantification of all expressed transcripts, on a base-pair precision level [73]. The high-throughput DNA sequencing allows RNA analysis through reverse-transcription onto cDNA and sequencing. The content of the transcriptome undergoes dynamic changes caused by selective enhancer action, and is tissue and context specific. This is further complicated by the fact that transcription is not only limited to promoters but can also occur at enhancers. In the latter case, enhancer regions recruit RNA polymerase and produce so called enhancer RNA (eRNA) [55]. The role of enhancer transcripts on gene expression is unclear. However, the eRNA expression is positively correlated with the expression of targeted genes, and can be triggered by the presence of specific TFs, including the estrogen receptor ER- $\alpha$  [66, 80]. Additionally the targets of eRNA producing enhancers have higher expression [19, 66, 80].

## 2.6 Discovery of elongating polymerase - GRO-seq

Although RNA polymerase ChIP-seq (Pol II) allows the discovery of Pol II binding location and density, the method does not discriminate between its various transcriptional states and dynamics. As part of transcriptional regulation the molecule can be promoter-bound, paused downstream from the promoter or can elongate unconstrained over a gene. The GRO-seq [62] technique enables a genome-wide detection of transcriptionally engaged polymerases at a certain time point. The method uses nuclear run-on assays to tag growing RNA chains of elongating polymerases. New polymerase initiations are prevented by Sarkosyl. The method thus enables identification of a location and orientation of active transcription. When used to investigate nascent transcription in the experiment involving stimulation of MCF7 cells with estradiol, the method showed firstly that a large proportion of the transcripts are unannotated, noncoding, and enhancer centric (eRNA), and secondly that in that cellular context, ER- $\alpha$  can be involved in eRNA transcription [42]. The method was also used in the same study to identify early estradiol-responsive genes.

## 2.7 Empirical discovery of chromatin interactions

The first method which allowed the identification of chromatin interactions was C3 [23]. The method led to development of several of its derivatives, such as C4, C5 and Hi-C [137, 29, 67], each with growing ability to map interactions on a genome-wide scale but sharing most experimental steps in common. The steps involve: treatment of cells with formaldehyde to cross-link spatially proximal chromatin, digestion of chromatin with restriction enzymes, possessing either 4 or 6 bp recognition sequences and producing fragments (restriction fragments) with cohesive ends located at sites of cuts, ligation of spatially proximal cross-linked DNA fragments via their cohesive ends and under very diluted conditions to favour proximal intra-molecular ligations but avoid non-specific inter-molecular ones, and reversal of cross-links to produce linear ligations which can be detected with a method-specific procedure. The 1D sequences offer a method to investigate population-averaged 3D chromatin conformation, as well as a relative frequency of the contacts.

### 3C

In the 3C [23] method the detection of chromatin contacts is performed via PCR with two pre-designed sequence-specific primers and quantification of relative speed of fragment amplification. The primer design requires a previous knowledge of candidate interacting sequences hence the method is unsuitable for mass scale genome-wide detection of interactions. The method however is often used for validation of interactions.

### 4C

A big step forward was offered by 4C-seq [137] method, which enabled creation of genome-wide contacts profile of one selected locus, so called bait. The protocol involves PCR amplification and high-throughput sequencing of all bait ligated partners. The PCR is initiated with two bait-specific pre-designed primers, each containing a sequence of single end of bait.

### 5C

The 5C [29] method in contrast to 3C allows a simultaneous use of thousands of loci-specific primers per single run of PCR. The amplification, similarly to C3, is limited



only to the fragments representing the contacts between pre-selected loci. Such fragments contains two primer recognition sites at their ligation junctions, which can be targeted by the pre-designed primers. Additionally, the primers contain sequences which allow quantification of amplified fragments by NGS.

## Hi-C

All above methods require a design of loci-specific primers therefore cannot be applied in a genome wide fashion. The Hi-C [67] method circumvents the requirement by introducing biotin-marked ligation junctions and paired-end sequencing of ligation products. The first modification turns cohesive ends of restriction fragments into blunt ends by filling the restriction sites with nucleotides, where one of them is biotinylated. After ligation of blunt-ends and additional round of chromatin fragmentation, the biotin-marked ligation junctions can be pulled down with streptavidin, the cross-links reversed and fragments sequenced using paired-end sequencing, with starting points close to ligation junction.

## Promoter Capture Hi-C

Although the Hi-C method enables a proximity scan of the chromatin conformation, the resultant maps are very complex and require a depth of tens of billions of reads to accurately distinguish true interactions from random non-specific bindings. To circumvent that requirement and in consequence increase statistical significance of findings, the Capture Hi-C [78] technique concentrates on promoter-centric interactions. The experiment mirrors the steps of Hi-C experiment with an additional filtering step. In the process ligated fragments in Hi-C libraries are treated with multiple biotinylated RNA oligomers with complementary sequences of all promoters. When bound to their recognition sequences, the oligomers enable the extraction of promoter-containing fragments, in consequence increasing the specificities of the libraries. After sequencing of the fragment, the technique can be used for genome-wide capture of promoter-centric links with other promoters, enhancers and potentially other functional genomic loci. The resolution of the technique is higher than 5kB, and provides a coverage of long-distance interactions of on average 250 kb [78].

## ChIA-PET protocol

ChIA-PET [63], similarly to Hi-C offers a genome-wide assessment of proximity between genomic regions. However, the ChIA-PET method concentrates on the contacts mediated by a protein of interest and as a result its predictions are of higher resolution (100 bp, which due to read specificity is 10 fold higher than corresponding Hi-C), and statistical confidence. Additionally, the method, as opposed to Hi-C, enables a statistical assessment of the amount of inter-ligations.

Although ChIA-PET is experimentally similar to C3-based methods, it differs in several steps. After cross-linking, the fragmentation of chromatin involves sonication instead of digestion. Next, protein-bound fragments are purified via immunoprecipitation with an antibody of interest. The cohesive ends of fragments are filled in with nucleotides, and extended by half-linkers which contain restriction enzyme cutting sites and biotinylated nucleotide. Following the ligation, the digestion cuts fragments 20 bp away from the ligation junctions. The fragments are pulled down with streptavidin, the cross-links reversed and resultant short fragments sequenced using paired-end sequencing.

The ChIA-PET protocol, apart from the additional ligation step which joins coprecipitated fragments, mimics that of ChIP-seq method. Thus in contrast to ChIP-seq which is only informative of close proximity of a DNA fragment to a selected protein, ChIA-PET method is able to identify regions which coprecipitate because of simultaneously laying in close proximity of the protein. The crucial step of the process is initial treatment of cells with formaldehyde which results in cross-linking, i.e. formation of covalent bonds, of protein-protein and protein-DNA contacts. Some contacts involve functional bindings between TFs and their recognition sites as well as contacts of those TFs with other mediator proteins which take part in their regulatory complexes. Other contacts are believed to be less specific and originate from polymer-like character of chromatin fibres which by law of thermodynamics would collide at frequency inversely proportional to linear distance between loci. These interactions are believed to occur on the peripheries of the regulatory complexes and would form less covalent bonds under treatment with formaldehyde. The gentle restriction enzyme digestion, a common step of C3 and its derivatives, is unable to break the non-specific bonds. Those methods rely on elaborate control experiments. In contrast, the vigorous sonication step of ChIA-PET (and ChIP-seq) is believed to remove the non-specific weaker bonds, and hence the need for extra control experiments [35].

The immunoprecipitation step of ChIA-PET is used to reveal all sequences involved in a molecular process associated with a particular protein. For instance in [64] immunoprecipitation with Pol II antibody revealed all Pol II-mediated contacts and in consequence loci involved in transcription. Similarly, in [34] and [43] immunoprecipitation with ER- $\alpha$  and CTCF antibodies revealed all contacts mediated by ER- $\alpha$  and CTCF, respectively. Those experiments enabled in consequence better understanding of mechanisms affecting gene expression and chromatin folding.

## 2.8 Bayesian Inference

Bayesian inference is a statistical procedure to make inferences from data. The method provides means to update our prior beliefs in the light of new evidence, which results in an updated posterior belief. In mathematical formulation, the prior assumptions, evidence, and likelihood of observed evidence given our initial beliefs are expressed and encoded as probability distributions. To simplify the notation we will use symbols  $E$ ,  $B$  for evidence and belief, respectively. The Bayes' formula states that:

$$P(B|E) = \frac{P(E|B)P(B)}{P(E)} \quad (2.1)$$

Here,

- $P(B)$  is the prior distribution of beliefs.
- $P(E|B)$  is the likelihood, the probability distribution of observing some evidence given our belief.
- $P(E, B) = P(E|B)P(B)$  is the joint likelihood and the product of the prior and the likelihood  $P(E|B)$  functions.
- $P(E)$  is the probability distribution of observed evidence irrespective of any belief that we may apriori possess. The function is sometimes called the marginal distribution and is either a sum or an integral over beliefs, fixed  $E$ , and over the set of all probabilities  $P(B, E)$ .
  1.  $P(E) = \sum P(E, B_i)$  for countable set of beliefs.
  2.  $P(E) = \int P(E, B)dB$  for uncountable set of beliefs.
- $P(B|E)$  is the updated posterior distribution of our beliefs given observed evidence.

Regardless of application, the evidence will usually correspond to observed values of data. In contrast an interpretation of belief  $B$  will depend on the problem tackled. For instance, in clustering our belief will represent a membership class of an object, while in inference it will usually be equivalent to a distribution of parameter values. Similarly, the posterior distribution will describe class memberships of clustered objects given their value of features, or in regression a distribution of probable output response variables of a learnt function of predictor features.

In many situations we are not interested in a distribution over many possible values of learnt posterior distribution but in a single estimate. In such cases we will often take a maximum a posteriori value i.e the optimum of a function of response variable, value of inferred parameter or membership of an object.

### 2.8.1 Parameter inference

The search for a set of parameters  $\theta$  consistent with the data  $X$  is called parameter inference,

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{\int P(X|\theta)P(\theta)d\theta} \propto P(X|\theta)P(\theta) \quad . \quad (2.2)$$

Here,  $P(\theta)$  denotes a prior distribution of a parameter  $\theta$ .  $P(X|\theta)$  denotes a likelihood given the value of parameter and  $P(\theta|X)$  an updated version of the parameter given observed data  $X$ . The  $\int P(X|\theta)P(\theta)d\theta$  is the normalisation factor which ensures that the posterior probabilities sum up to one. The normalisation factor is a function of data  $X$  and does not depend on the parameter  $\theta$ . This quantity is however usually intractable. In such cases we use approximate inference methods. Various methods have been developed to tackle the problem. One of the approaches is the Expectation Maximisation or variational bayes algorithms which provide computationally fast but partially inaccurate approximations to the true posterior distribution. Alternatively one can use slower but more accurate iterative sampling methods such as Markov chain Monte Carlo (MCMC). The accuracy of the MCMC methods increases with the number of iterations and theoretically after infinite time the sampler would sample from a true posterior distribution. However, due to computational limitations the inference is always based on a finite number of samples.

### 2.8.2 MCMC algorithms

Monte Carlo methods form a general class of numerical algorithms which rely on Monte Carlo simulation to approximate a desired function  $P(\theta)$  from its samples. The methods are particularly useful due to their ability to sample from a target distribution  $P(\theta)$  when the distribution is only known up to a normalisation constant. In such cases the unnormalised distribution serves as a proposal distribution  $Q(\theta)$ , which is used by the methods to sample from the desired distributions. The problem of intractable normalisation constants often occurs in bayesian inference where the evaluation of the marginal distribution of data, the normalisation constant of a posterior distribution, is often difficult or impossible. In MCMC algorithms the proposal distribution is equal to  $Q(\theta'|\theta)$  and therefore the draws form a Markov chain. A Markov chain is a stochastic process, a sequence of random variables  $\theta^0, \theta^1, \dots, \theta^n$  with the property that

$$P(\theta^n | \theta^{n-1}, \theta^{n-2}, \dots, \theta^0) = P(\theta^n | \theta^{n-1}) \quad . \quad (2.3)$$

Hence the value of the random variable  $\theta^n$  only depends on the value of the last value in the sequence. Each Markov chain can be characterised by its initial distribution  $P(\theta^0)$  over a finite or infinite state space  $S$ , and a set of transition probabilities  $P_{ij} = P(\theta^{n+1} = j | \theta^n = i)$  i.e. the set of probabilities of moving from state  $i$  to state  $j$  which can be summarised in a  $|S|$ -dimensional square matrix  $P$ , or kernel  $K(\theta^{n+1} | \theta^n)$  in case of continues state space. As the chain progresses, moving from state  $\theta^n = i$  to  $\theta^{n+1} = j$ , the initial distribution  $P(\theta^0)$  and each consecutive distribution  $P(\theta^n)$  evolves, according to the equations:

$$P(\theta^n) = \sum_{\theta^{n-1} \in S} P(\theta^{n-1}) P(\theta^n | \theta^{n-1}) \quad (2.4)$$

or,

$$P(\theta^n) = \int P(\theta^{n-1}) K(\theta^n | \theta^{n-1}) d\theta^{n-1} \quad (2.5)$$

hence each update depends only on the distribution of the last step  $P(\theta^{n-1})$  and fixed transition probabilities  $P_{ij}$ .

When the chain is ergodic i.e. aperiodic (any state  $i$  does not return to  $i$  in a finite  $k$  number of iterations), and irreducible (any state is reachable from any state in a finite number of steps), then

$$\forall P(\theta^0), \text{ as } n \rightarrow \infty, P(\theta^n) \rightarrow \pi(\theta) \quad (2.6)$$

$$\pi(\theta') = \sum_{\theta \in S} \pi(\theta) P(\theta' | \theta) \quad (2.7)$$

for large  $N$ ,  $P(\theta^n) = P(\theta^{n-1}) = \pi$ , the distribution over  $S$  converges to a fixed distribution  $\pi$ , so called steady state or invariant distribution.

### 2.8.3 Metropolis-Hastings

The Metropolis-Hastings algorithms [77, 45] is a class of MCMC algorithms which can be used to sample from an arbitrary posterior distribution  $P(\theta)$ , as long as it is known up to a normalisation constant. The methods often rely on an iterative two step procedure. Firstly a proposal sample  $\theta^*$  is drawn from an appropriate proposal distribution  $Q(\theta^* | \theta^{n-1})$ , secondly the sample is either accepted or rejected according to acceptance/rejection rule,

$$\theta^n = \begin{cases} \theta^* & \text{with } P_{accept} \\ \theta^{n-1} & \text{otherwise,} \end{cases}$$

where,

$$P_{accept} = \min \left( 1, \frac{P(\theta^*) Q(\theta^{n-1} | \theta^*)}{P(\theta^{n-1}) Q(\theta^* | \theta^{n-1})} \right) . \quad (2.8)$$

The  $P_{accept}$  depends on the ratio  $P(\theta^*)/P(\theta^{n-1})$  thus any normalisation term of the distribution  $P(\theta)$  cancels out. The two step rate adjusted random walk is an ergodic Markov chain which in long-term converges to a target stationary distribution  $P(\theta)$ . The choice of  $Q(\theta^* | \theta^{n-1})$  is crucial since it determines the convergence rate of the methods. A good distribution should possess a high degree of mobility to minimise a possibility of getting stacked in a local node of a state space of the target distribution, and a low sample rejection rate which would result in slow convergence.

### 2.8.4 Gibbs Sampler

The Gibbs Sampler [37] is a special case of Metropolis-Hastings algorithm which samples a new  $\theta^n$  sequentially updating  $d$  sub-vectors  $\theta_j$ . Each update is conditional on the

values of all the other components  $\theta_{-i}$ . The steps are:

$$\begin{aligned}\theta_1^n &\sim P(\theta_1|\theta_2^{n-1}, \theta_3^{n-1}, \dots, \theta_d^{n-1}) \\ \theta_2^n &\sim P(\theta_2|\theta_1^n, \theta_3^{n-1}, \dots, \theta_d^{n-1}) \\ &\vdots \\ \theta_d^n &\sim P(\theta_d|\theta_1^n, \theta_2^n, \dots, \theta_{d-1}^n)\end{aligned}\tag{2.9}$$

At each sub-step  $i$  the Markov random walk makes a restricted transition:

$$\begin{aligned}\theta^{(n-1,n),i-1} &\rightarrow \theta^{(n-1,n),i} \\ &= \\ (\theta_1^n, \dots, \theta_{i-1}^n, \theta_i^{n-1}, \theta_{i+1}^{n-1}, \dots, \theta_d^{n-1}) &\rightarrow (\theta_1^n, \dots, \theta_{i-1}^n, \theta_i^n, \theta_{i+1}^{n-1}, \dots, \theta_d^{n-1})\end{aligned}\tag{2.10}$$

where the notation  $(n-1, n)$  in the exponent of  $\theta$  emphasizes that the resultant random vector is a transient state between  $\theta^n$  and  $\theta^{n-1}$ . The random vectors  $\theta^{(n-1,n),i-1}, \theta^{(n-1,n),i}$  share a fixed  $\theta_{-i}$  and differ in  $\theta_i$  part.

Let's denote the new proposed state as  $\theta^*$ . The transitions in the above procedure are made according to the proposal distributions

$$Q\left(\theta^{*(n-1,n)}|\theta^{(n-1,n),i-1}\right) = \begin{cases} P\left(\theta_i^{*(n-1,n)}|\theta_{-i}^{(n-1,n)}\right) & \text{if } \theta_{-i}^{*(n-1,n)} = \theta_{-i}^{(n-1,n)} \\ 0 & \text{otherwise} \end{cases}$$

Substituting the transition probabilities into eq. 2.8, rewriting the  $P(\theta)$  as  $P(\theta_i)P(\theta_{-i})$ , and for readability dropping the superscripts  $(n-1, n)$ , yields

$$P_{accept} = \frac{P(\theta^*)Q(\theta|\theta^*)}{P(\theta)Q(\theta^*|\theta)} = \frac{P(\theta_i^*|\theta_{-i}^*)P(\theta_{-i}^*)P(\theta_i|\theta_{-i}^*)}{P(\theta_i|\theta_{-i})P(\theta_{-i})P(\theta_i^*|\theta_{-i}^*)} = 1\tag{2.11}$$

and shows that the set of proposal distributions guarantees that none of the samples is rejected. Due to its reliance on conditional distributions the scope of application of the algorithm is limited to the models with conditional distributions of standard and tractable form.

### 2.8.5 Convergence of Gibbs sampler

One of the most important aspects of Markov Chain Monte Carlo methods is their convergence to the true target distribution and the rate of that convergence. In general, the

faster the chain converges, measured by the number of samples, the better. The convergence is dependent on the initial conditions, and since the sampling procedures are usually started from a different, usually random, initial conditions, therefore different chains will need different time to reach the stationary distribution. Here we employ the Gelman's and Rubin's multi-chain diagnostic to assess the convergence (reviewed by Gelman [36]).

### Burn-in

To resolve the issue with the initial bias due to a starting position, we apply burn-in. In burn-in one discards  $B$  initial samples before starting to collect the samples. Burn-in allows the Markov Chain to reach its steady-state equilibrium distribution, and prevents retaining samples from regions which are very rare under equilibrium distribution of Markov chains. Although the burn-in is arbitrary in practice we run the chain for  $2S$  samples and discard the first  $S$  samples. The sampling is then either doubled or stopped if the chain converged.

### Gelman's and Rubin's multi-chain diagnostic

Multiple diagnostics can be run to establish the convergence of Markov Chains. Here we assess the convergence using Gelman's  $\hat{R}$  statistics which relies on comparison of within chain and between chain variability. The diagnostic requires running  $M \geq 2$  independent chains with over-dispersed initial conditions for  $S$  iterations and then assessing within-chains ( $W$ ) and between-chains ( $B$ ) variances of the parameter  $\theta$ , defined as:

$$W = \frac{1}{M} \sum_{m=1}^K \left( \frac{1}{S-1} \sum_{s=1}^L (\theta_{m,s} - \bar{\theta}_m)^2 \right) \quad (2.12)$$

$$\bar{\theta}_m = \frac{1}{S} \sum_{s=1}^L (\theta_{m,s}) \quad (2.13)$$

$$B = \frac{S}{M-1} \sum_{m=1}^M (\bar{\theta}_m - \bar{\bar{\theta}})^2 \quad (2.14)$$

$$\bar{\bar{\theta}} = \frac{1}{M} \sum_{m=1}^M \bar{\theta}_m \quad (2.15)$$

Due to initial non-stationarity of the chains we can expect that a number of initial draws will be biased to a particular subspaces of the sampling space of the target distribution



and that the variance  $W$  estimated on the draws would provide underestimated estimates of the true variance of the target distribution. If we knew the starting values which would result in immediate stationarity we could expect that the variance would be unbiased, however the information is in practice unknown. As the number of samples increases the sampling methods explores the sampling space to a wider extent, reaches more states, and loses the initial bias. We can expect that once the chains have converged to the same distributions the within and between chain variances should be approximately equal. The intuition can be summaries in  $\hat{R}$  which can be shown to converge to 1 as  $S \rightarrow \infty$ .

$$\hat{R} = \sqrt{\frac{\widehat{Var}(\theta)}{W}} \quad (2.16)$$

where,

$$\widehat{Var} = \left( \frac{S-1}{S} \right) W + \frac{1}{S} B \quad (2.17)$$

which is a weighted average of  $W$  and  $B$ .

### **Gelman's and Rubin's multi-chain diagnostic - practice**

In practice we iteratively double the number of samples, removing the first half of the produced samples as burn-in and repeat the process until the convergence criterion is met. The convergence is assumed to be met if for each parameter the  $\hat{R}$  is close to 1. We test the convergence for every inferred parameter.

## **2.8.6 Supervised and unsupervised learning**

Two of the main learning strategies in machine learning are supervised and unsupervised learning. In a supervised learning problem a ML algorithm receives a labelled training data to learn an underling function which could be used to assign unobserved new data points to their labels. The typical examples of supervised learning algorithms are SVM, Naive Bayes, linear regression, logistic regression, neural networks and decision trees (reviewed by Hostie [44] and in Bishop [12]) each with its strengths and weaknesses. On the contrary in unsupervised problem the task is to learn the function from unlabelled data. Examples of the algorithms include mixture models, hierarchical clustering, and AP algorithm [44, 12, 33]. The problem in supervised learning usually encompasses problems of finding similarities between feature vectors i.e clustering or finding density function. Regardless of the modelling approach, the factors

which needs to be taken into account when building a model are the feature representation of classified objects, number of features, function complexity, and amount of data available. In general, the more complex the underlying function the more data is needed.

### 2.8.7 Naive Bayes

Naive Bayes is a supervised learning algorithm which is commonly used in sparse data problems when the number of parameters of inferred model - suggested solution outnumbers the amount of available training data. The model can also be used to combine data from different experiments. The model approximates a posterior distribution by assuming conditional independence assumption between data points in the likelihood i.e.

$$P(C)P(X|C) = P(C)P(X_1, \dots, X_n|C) = P(C) \prod_{i=1}^N P(X = x_i|C) \quad (2.18)$$

Thus, the Bayes' formula becomes,

$$P(C|X) = \frac{P(C)P(X|C)}{\sum P(C)P(X|C)} = \frac{P(C) \prod_{n=1}^N P(X_n|C)}{\sum P(C) \prod_{n=1}^N P(X_n|C)} \quad (2.19)$$

Although the assumption of conditional independence is quite strong, the algorithm is shown to give a good performance in a number of applications.

### 2.8.8 Mixture of Gaussians and hidden variable models

The mixture of Gaussians (eq: 2.20) can be used in unsupervised learning scenario and is an example of more general class of models called mixture models. Mixture models are linear combinations of probability density functions (PDF), which can be used to generate and approximate more complex functions. In that specific mixture combining the Gaussian PDFs with different means and covariance matrices let us approximate almost any continuous PDF to any desired accuracy.

$$P(\mathbf{X}|\theta) = \sum_{k=1}^N P(k)p(\mathbf{X}|k) = \sum_{k=1}^N \pi_k \mathcal{N}(\mathbf{X}|\mu_k, \Sigma_k) \quad (2.20)$$

where  $k$  is a class variable with probability of occurrence given by mixing components  $\pi_i$ , satisfying  $\sum^K \pi_i = 1$ , and  $\theta$  is a vector of all parameters  $\pi_i, \mu_i, \Sigma_i$ .

Mapping  $k$  onto 1-of- $K$  representation i.e binary vector  $\mathbf{Z}_k$  with 1 at  $k^{th}$  position

and 0 elsewhere, and defining  $P(\mathbf{Z}_k) = \prod_{k=1}^K \pi_k^{z_k}$  and  $P(\mathbf{X}|\mathbf{Z}_k) = \prod_{k=1}^K \mathcal{N}(\mathbf{X}|\mu_k, \Sigma_k)^{z_k}$

$$\begin{aligned} P(\mathbf{X}|\theta) &= \sum_{\mathbf{Z}} P(\mathbf{Z})p(\mathbf{X}|\mathbf{Z}) = \sum_{\mathbf{Z}} \prod_{k=1}^K (\pi_k \mathcal{N}(\mathbf{X}|\mu_k, \Sigma_k))^{z_k} = \\ &= \sum_{j=1}^K \prod_{k=1}^K (\pi_k \mathcal{N}(\mathbf{X}|\mu_k, \Sigma_k))^{I_{kj}} = \sum_{j=1}^N \pi_j \mathcal{N}(\mathbf{X}|\mu_j, \Sigma_j) \end{aligned} \quad (2.21)$$

shows (eq: 2.21) that each occurrence of  $\mathbf{Z}_k$ , so called called hidden variable, determines which Gaussian distribution  $P(\mathbf{X}|\mathbf{Z})$  generates  $\mathbf{X}_k$ , and that the two variables are inseparable. Although, we usually observe only the marginalised version  $P(\mathbf{X}|\theta)$  of  $P(\mathbf{X}, \mathbf{Z})$ , and as the name suggests the value of  $\mathbf{Z}_k$  remains hidden, we can however use the posterior  $P(\mathbf{Z}|\mathbf{X}_n)$  to infer the origin of  $\mathbf{X}_n$ . In practice we will usually be interested in the values of the parameters of each component function. These will be usually obtained from the posterior of parameters  $P(\theta|\mathbf{X}, \mathbf{Z})$ .

### 2.8.9 Kernel Density Estimation

$$P(X) = \frac{1}{N} \sum_{n=1}^N \frac{1}{h^D} k\left(\frac{X - X_n}{h}\right) \quad (2.22)$$

Kernel Density Estimation (KDE) [81, 88] is an example of non-parametric probability density estimation method which as opposed to parametric approaches makes very little assumptions about underlying functional form of a modelled distribution. The simplest example of non-parametric approach is a histogram function. KDE has an advantage over the histogram that it provides smooth continuous PDFs. In non-parametric approaches we still select a kernel function and bandwidth, where the choices determine the smoothness of the output function. The common choice of kernel function is a gaussian, and the bandwidth can be obtained either via cross-validation or Scotts or Silverman's rules [102, 107].

$$P(X) = \frac{1}{N} \sum_{n=1}^N \frac{1}{(2\pi h^2)^{1/2}} \exp\left\{-\frac{\|X - X_n\|^2}{2h^2}\right\} \quad (2.23)$$

## 2.9 Related work

In addition to the standard already covered applications of ChIP-seq data, and more relevant to the present study, others have used ChIP-seq data to infer enhancer-promoter

interactions. For example, Ernst et al. [31] used histone mark data from multiple cell-types to identify active enhancers and promoters and then enhancer-associated data was correlated with expression data from genes within 125kbp to identify likely interactions [31]. Thurman et al. used DNase I hypersensitivity (DHS) data from multiple cell-types to correlate and link distal DNase hypersensitivity sites (within 500kbp) to those within putative gene targets [119]. Similarly, in Andersson et al. [3], enhancer-promoter links were predicted by correlating CAGE enhancer RNA to CAGE promoter RNA.

Approaches for discovering cell-type specific interactions include PreSTIGE [21], RIPPLE [95], and the method developed in [74]. PreSTIGE uses a method based on the Shannon entropy to identify cell-type specific enhancers and genes using H3K4me1 and RNA-seq data respectively. The regions are linked within promoter-centric domains, bounded on each side by the minimal distance of 100kbp up to the first CTCF binding site from a TSS. RIPPLE uses four cell-lines and per each 11 ENCODE datasets (RNA-seq, CTCF, RAD21, DNase1, TBP and histone marks) to train random forest classifier which predicts enhancer-gene interactions within 1MB distance. The features used are two joint binary vectors of presence/absence of dataset signal peak over a promoter and enhancer, correlation of entries of the vectors, as well as gene expression of the promoter controlled gene. The method from [74] aggregates RNA-seq data over genes and DHS data over  $\pm 200$ kb regions surrounding them for twenty different cell lines. The method searches through each gene and cell-line for unexpected DHS/RNA-seq ratios and once found, scanned across the gene vicinities in search of causal, local DHS variabilities. Lastly, a method proposed by He et al. uses a random forest classifier to find enhancer-gene interactions [46]. The method uses three features: evolutionary conservation, correlation of enhancer scores derived from histone marks with RNA-seq data, as well as an average of correlations between TF ChIP-seq and gene expression across 12 cell-types. A distance constraint is also imposed to aid inference.

In summary, firstly the great part of the methods presented above require data from multiple cell-types and therefore do not allow discovery of interactions given data from one cell-type. Secondly, the imposed stringent distance constrain of most of the methods prevents a discovery of more distant enhancer links. Thirdly, the methods above do not take into account an evidence from time course data over gene region nor consider evidence from TFs across gene bodies.

Here, we attempt to address the drawbacks. We build statistical models for the

inference of the enhancer-gene links mediated by ER- $\alpha$  and Pol II within MCF7 cell-line. For that we utilise time course data from the same cellular context and cell-line, one which is complementary to the chromatin conformation capture experiments such as ChIA-PET.

# Chapter 3

## ChIP-seq Time Series Data Processing

In this section we describe the experimental data used and the initial processing of the ChIP-seq data which transforms the raw data into a form which is more appropriate for the modelling in chapters 4 and 5. The pre-processing steps include: alignment of short reads originating from ChIP-seq experiments to a reference human genome, discovery of TF bindings via identification of TF-specific peaks with MACS, merging of time persistent peaks to reduce the amount of noise and to create shared regions for counting of the enrichments of the ER- $\alpha$  datasets and other ER- $\alpha$ -binding-associated ChIP-seq datasets. Due to a non-unique mapping of some reads, differences in read depth and quantity of starting material, the alignment of the ChIP-seq reads results in a variable total number of reads across datasets, and thus in a loss of comparability of the enrichments between time points. Here, we show how we normalise the data to deal with this issue. Additionally, the chapter covers the initial analysis of variability in the ChIP-seq time series data using the Affinity Propagation (AP) clustering algorithm. The clustering reveals underlying dynamics of the ER- $\alpha$  bindings and other associated datasets, and provides initial insight into the usefulness of the time series for the modelling of chromatin contacts.

### 3.1 The experiment - studying ER- $\alpha$ responsive genes

The experiment which produced our data was designed to uncover a response of genes to estrogen, more specifically estradiol (E2), in MCF7 breast cancer cells, and was already presented and described in [49, 127]. Thus, the first step was to create a reference sample in a ligand free environment. For that, the cells were placed into estradiol free media for 3 days, which reduced the binding between ER- $\alpha$  and E2. The cells

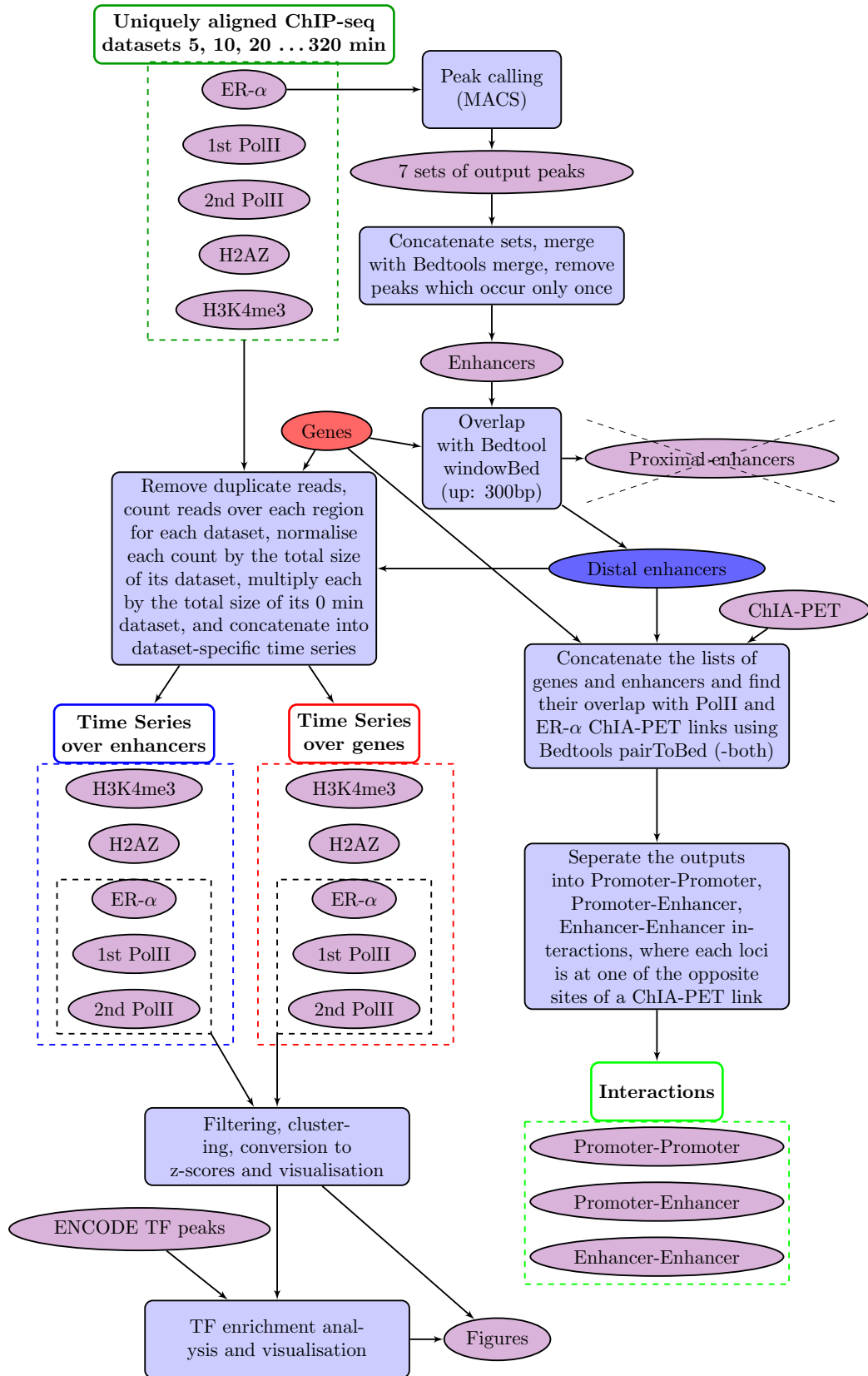


Figure 3.1: Preprocessing pipeline

were then ready to be re-exposed to E2. Following the introduction of E2, the resultant changes were tracked by multiple ChIP-seq experiments. The experiments were performed at 0, 5, 10, 20, . . . , 1280 minutes after the stimulation. Each ChIP-seq experiment was carried out with a different antibody to measure genome-wide fluctuations in levels of their specific protein targets. Specifically, the studied protein factors and histone modifications were: ER- $\alpha$ , Pol-II (two replicates), H3K4me3, and H2AZ. To enable an analysis of gene expression, time series RNA-seq data was also produced. The experiments at 0 min correspond to the untreated sample. To ensure the comparability across time points and datasets, the cells used for the measurements were taken from the same initial population.

### 3.1.1 Alignment to a reference human genome

Raw reads from the experiments were mapped onto the human reference genome (NCBI\_build37) using the Genomatix Mining Station (version 3.5.2) to enable further analysis. The sequencing depth, i.e. the total number of sequenced reads, was the same for each dataset, however, on average only 81%, 76%, 67%, 61%, 64% of ER- $\alpha$ , Pol-II (rep 1), Pol-II (rep 2), H3K4me3, and H2AZ ChIP-seq reads respectively were mapped uniquely to the genome. Reads which mapped to several locations were discarded from further analysis. Using the statistical criterion provided by MACS, we established that our sequencing depth allows for no duplicates of reads, thus we discarded any duplicated reads. The duplicated reads are potentially an artefact in ChIP-seq.

### 3.1.2 Discovery of ER- $\alpha$ bindings

We wish to identify the binding sites of ER- $\alpha$ , occurring after the stimulation with E2. Bound regions of the genome are accompanied by aggregations of ChIP-seq reads, so called peaks. Fig. 3.2 (first row) shows an example of sharp ER- $\alpha$  peaks at 5 min after stimulation around the GREB gene locus. Given a set of peaks, the question is which of these originate from random noise and which corresponds to true binding sites. To answer this question, we used MACS (v1.4.2) [135] (for peak calling challenges, procedure of MACS and technical details see Section 2.4)). The algorithm scans across a reference genome comparing enrichments of background and treatment reads and probabilistically assesses adjacent regions for the presence of potential peaks. It is often considered that control datasets play a vital role in this task. A control dataset



enables modelling of the background distribution of reads, which provides a reference point for the assessment of significance of the peaks. More precisely, after the initial search for peaks using rate parameter  $\lambda_{BG}$ , i.e. genome-wide average enrichment of reads per bp estimated from treatment or control (i.e. MACS normalises the total size of reads of each set to be the same), and significance under a Poisson model, MACS can be chosen to run a local background correction. This process re-assesses each putative peak from the first run using averages  $[\lambda_{1k}], \lambda_{5k}, \lambda_{10k}$  estimated over  $[1k], 5k, 10k$  regions. The  $\lambda$ s are either estimated using control (if a control is available), or treatment dataset (if it is not available), with the  $\lambda_{1k}$  estimated only in the presence of a control dataset. The most relaxed case corresponds to switching the lambda flag off. In such case the local estimates of  $\lambda$  are ignored.

Our initial choice was to use the 0 min untreated dataset as the control. However, visual inspection of the density of reads revealed a large amount of pile ups at loci which are known to play important functional roles. Figure 3.2 (second row) shows examples of the peaks in the E2 depleted sample, prior to stimulation and around the GREB gene locus. To confirm the significance of the peaks in the E2 depleted sample, we overlapped them with chromatin interactions mediated by ER- $\alpha$  and detected by ChIA-PET in an independent MCF7 sample. Figure 3.2 (third row) shows that many of the ChIA-PET links overlap with the ER- $\alpha$  bindings. We conclude that the reduction of the level of E2 does not abolish all of ER- $\alpha$  bindings, equivalently that not all of the ER- $\alpha$  bindings require presence of the ligand. Our conclusion is based on the phenomena described in Section 2.2 i.e. ER- $\alpha$  possesses known ability to bind to chromatin even in the complete absence of its ligands and therefore we can expect that a portion of the bindings are present prior to the stimulation. As shown in the Figure 3.2, even if the peaks in the untreated sample may have generally lower signals, running MACS on our treatment datasets with that set as the control would render many functional peaks, which could be involved in chromatin interactions, as statistically insignificant.

We therefore tried two alternative approaches, which do not use control datasets to estimate the local backgrounds. Firstly, we used MACS on each of our ER- $\alpha$  ChIP-seq datasets, with the stringent  $p$ -value of  $1e-11$  (default  $1e-05$ ) and the local control switched off. Secondly, we compared the first parametrisation against the search with no control, default  $p$ -value, and the local control flag on (Section 3.1.3). In subsequent analysis of the study and in building of our models we choose the first parametrisation. However for the sake of comparison, in Section 4.4.5 we benchmark the model based

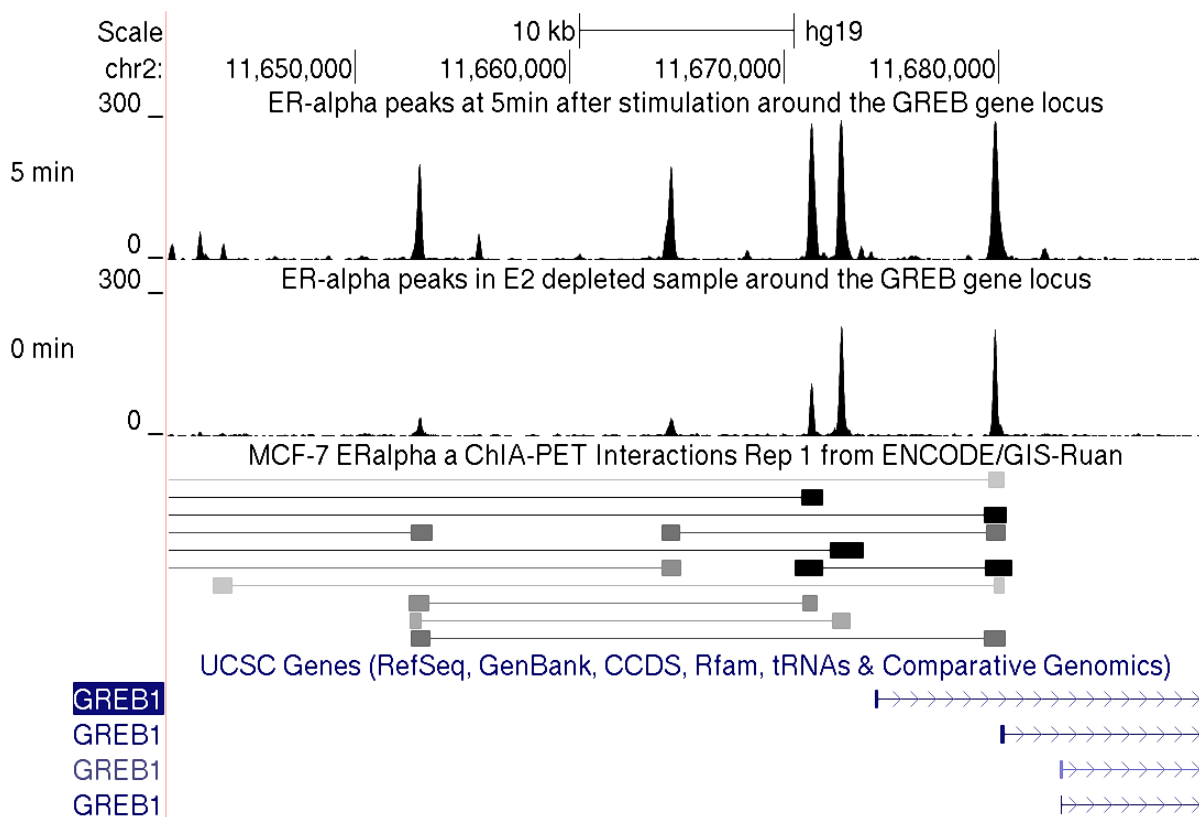


Figure 3.2: The upper part of the figure shows the ChIP-seq data across GREB region 5 min after the incorporation of E2 to the estrogen deprived MCF7 cell-line. The peaks correspond to putative ER- $\alpha$  bindings. The second row shows ChIP-seq data of 0 min unstimulated sample. The third row shows the corresponding local chromatin interactions mediated by ER- $\alpha$  and captured by ChIA-PET. A large fraction of the ER- $\alpha$  bindings coincide with links from ChIA-PET.

on the default and alternative parametrisations and present the results in the appendix. Before using the resultant multiple sets of MACS-found peaks in our model, we pre-processed them to find a set of time persistent ER- $\alpha$  bindings. In the next section we describe the process and creation of time series at each of the resultant regions.

### 3.1.3 ChIP-seq time series data

We ran MACS on 0, 5, 10, 20, ..., 320 min time course datasets. The remaining 640 and 1280 min datasets were omitted from any further analysis due to their relatively lower total number of tags and putatively also quality of the datasets. Next, we checked which peaks co-occurred in time. Since the locations of bindings are imprecise and vary between time points, we sought to identify consensus regions. For that, we merged

the resultant sets in a fashion similar to the mergeBED method from BEDTools [91]. That is, treating each region as a set we merged by union operation the regions which co-occurred in time and were present at least twice across time points. The single occurrences of peaks were discarded. The method is illustrated in Figure 3.3. We tried the approach on the MACS outputs from the default and the second parametrisation (see Section 3.1.2). When merging the peaks from the default parametrisation, we did not include peaks from the unstimulated sample (0 min). In case of the alternative parametrisation, to test a potential effect of adding peaks from the 0 min time point dataset on the total number of produced consensus regions, we first included and then discarded the set of peaks from the process.

Merging of the peaks from the default parametrisation resulted in 45598 regions, 20652 overlapped with a known gene or 300bp region upstream from its TSS while 24946 were distant from genes (distal enhancers). We extended the genes from their canonical ENCODE (Homo Sapiens GRCh37.75 - hg19) annotated TSS to account for their promoter region. Only the distant (non-overlapping) peaks are used in the further analysis as our aim is to link distal enhancers to genes. Merging of the outputs from the alternative parametrisation produced 56844 (31214 distal, 25630 overlapping) or 56407 (31014 distal, 25393 overlapping) depending on whether we included or ignored peaks from the unstimulated sample.

We calculated the tag enrichment of each of our ChIP-seq datasets over promoter-extended-gene bodies and over our resultant distant non-overlapping consensus ER- $\alpha$  binding sites to create time series data for genes and enhancers. To create the time series for each of our ChIP-seq datasets, we first counted the number of reads at our consensus ER- $\alpha$  binding sites and at promoter-extended genes for each time point. In order to make the time points comparable we normalized each count by dividing it over the total number of uniquely mapped and non-duplicated tags of its corresponding dataset. The normalised counts were then concatenated into a data-specific time series for each ChIP-seq antibody.

## 3.2 Clustering Pol-II and ER- $\alpha$ time series data

To help visualise the occupancy dynamics at enhancers and genes we clustered the data with the R-implementation of Affinity Propagation (AP) [33]. AP is a clustering method based on belief propagation and works iteratively by passing messages between data points until exemplars (cluster centres) automatically emerge. A preference

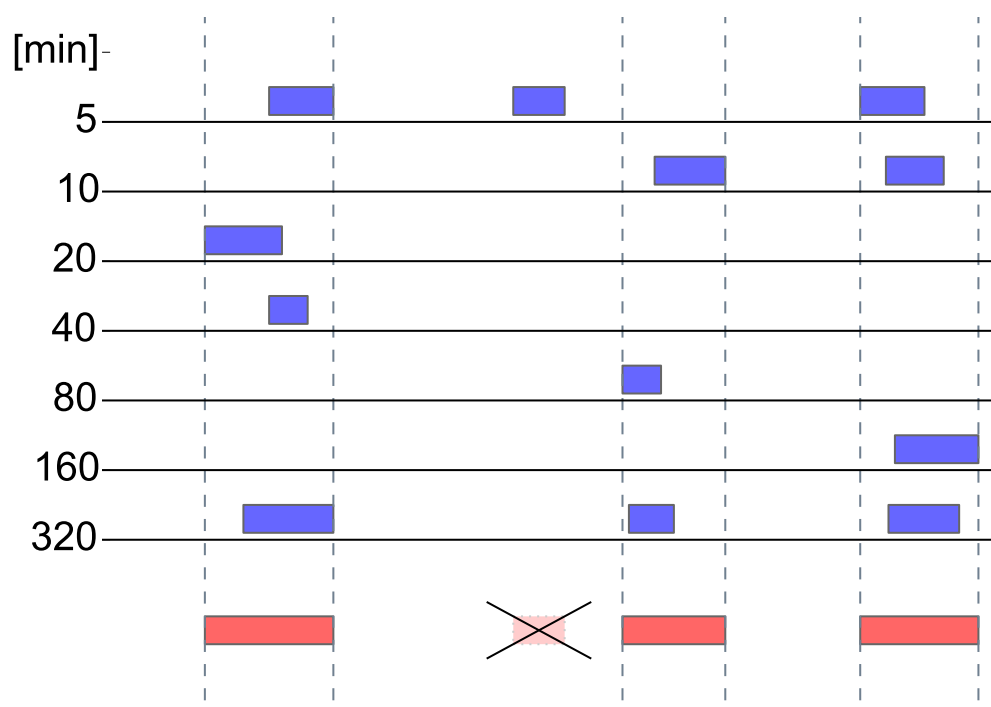


Figure 3.3: Cartoon shows the process of merging individual MACS-called peaks with the objective of finding approximate locations of time persistent ER- $\alpha$  bindings. In the process MACS-detected time varying peaks from [0], 5,  $\dots$ , 320 min time points (0 is optional and by default not included) which co-occur at least twice across time points are merged by union operation to produce the approximate consensus locations of a single binding. The single occurrences of peaks are discarded.

parameter  $p$  has an effect on the final number of clusters. The R implementation of AP can search through values of  $p$  to achieve an approximately pre-specified number of clusters. The method is similar to k-means but can achieve much better optimisation of the k-means objective function than the standard EM algorithm.

Prior to the clustering we standardized each time series to z-scores to bring all time series onto the same scale. Figure 3.4 shows the clustering of the time series of Pol II replicates and ER- $\alpha$  over enhancers and genes. To reduce the effect of noise, for Pol II we clustered only the pairs of the time series for which the Pearson correlation coefficient was at least 0.2 between replicates and the total enrichment of each time series at least 30. For ER- $\alpha$ , due to lack of replicates, we only clustered the time series with total enrichments of more than 100 reads.

The clusters show substantial differences in occupancy dynamics across both genes and enhancers. This is expected for Pol II which is known to show a broad range of response profiles in this system [49] but we also see some differences in ER- $\alpha$  profiles suggesting that occupancy is not solely determined by the nuclear concentration of ER- $\alpha$  but is also influenced by other cofactors.

### 3.2.1 Linking Pol II dynamics with TF occupancy

In [127] the Pol II ChIP-seq time course data considered here was used to model and infer the Pol II dynamics across selected genes. Clustering of the inferred dynamics identified clusters associated with ER- $\alpha$ -regulated early responsive genes following the simulation with E2, characterised by early peaking (maximum) of Pol II occupancy in their mean dynamics. Subsequently, the genes in each cluster and their Pol II dynamics were linked to cluster-specific combinations of TF bindings occurring in vicinity of their promoters. The analysis confirmed that FOXA1 binding is associated with ER- $\alpha$  [51] and that it is involved in early transcriptional response of ER- $\alpha$  responsive genes under E2 stimulation. Analogously, in this section we attempt to identify cluster-specific patterns of TF-bindings of our AP-clustered Pol II and ER- $\alpha$  time series of genes and enhancers using complementary publically available ChIP-seq data from Cistrome Project's database. Among the data sets which we analysed are ER- $\alpha$  [128], FoxA1 [70], c-MYC [50], c-Jun [53], c-Fos [53], SRC-3 [60], TRIM24 [122], RAD21 [98], CTCF [98] and STAG1 [98]. The datasets come from the same cell-line and stimulation, from between 45 to 60 min after E2 stimulation. Tables 3.1 and 3.2 show the results of our analysis for Pol II clusters, the examples of which are presented in Figures 3.4a (over enhancer) and 3.4c (over genes). In the appendix, Tables A.2 and

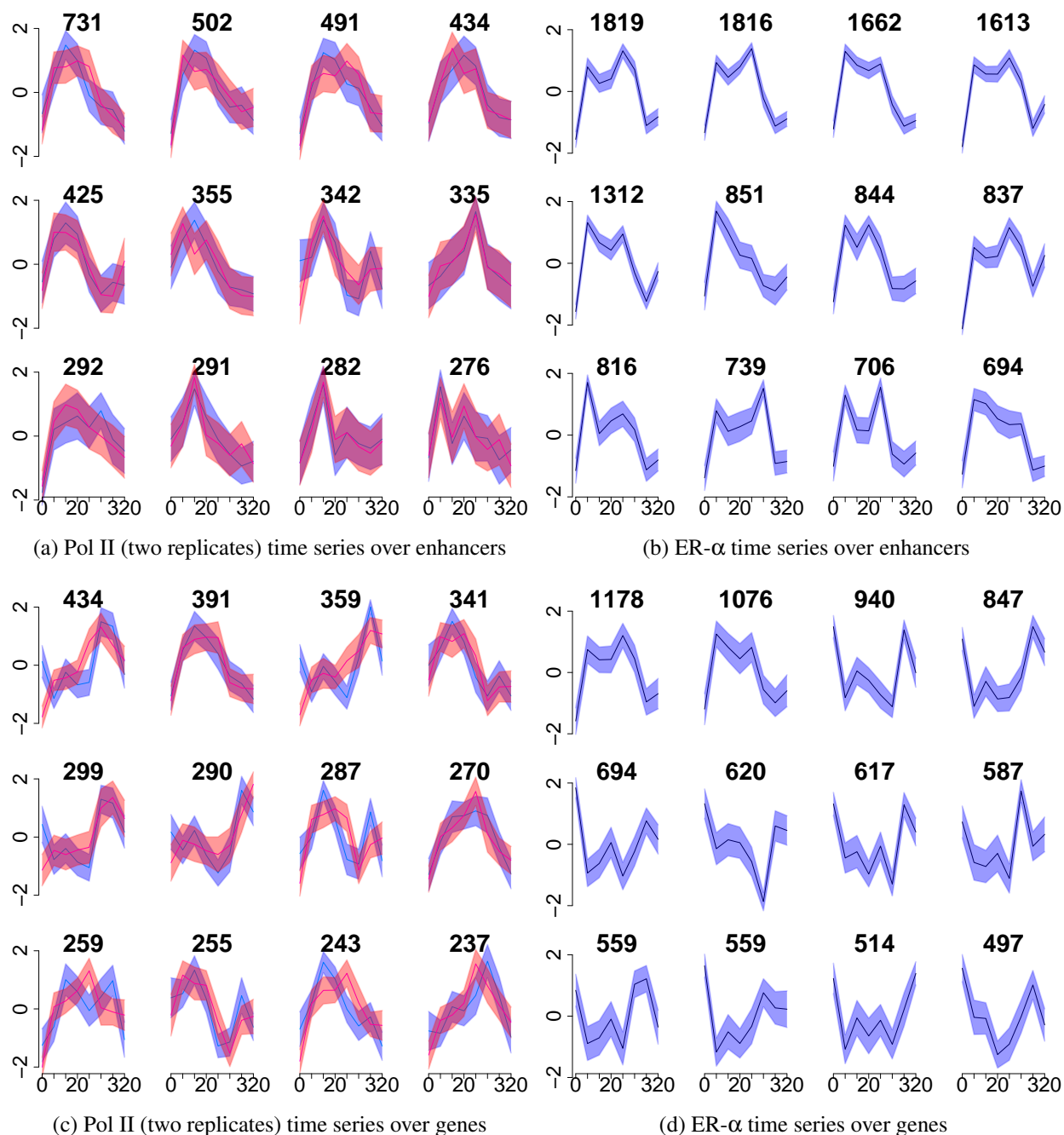


Figure 3.4: The first column of the figure shows the results of the clustering of joint time series of both Pol II replicates at enhancers (top row) and genes (bottom row) with Affinity Propagation. The second column shows the corresponding clustering for ER- $\alpha$  time series at the enhancer (top row) and gene regions (bottom row). The remaining clusters can be seen in A.1, A.2, A.3.

A.3 show the corresponding analysis for our clustered ER- $\alpha$  time series, the examples of which are presented in Figures 3.4b and 3.4d. In each row of the tables we assess the statistical significance of an observed number of overlaps of each TF with distal enhancers or TSS-centred 2000bp-long regions to the expected number of overlaps (i.e. the same size drawn at random from the set of all distal enhancers or 2000bp-long region around each TSS). Each cluster is enriched or depleted in a TF, if the probability of observing its fraction of overlaps with that TF is below  $p$ -value of 0.01. The value determined from the two-tailed binomial test. The last two columns of each row correspond in turn to the size of each cluster and difference in occupancy between time points 0 min (E2-deprived sample) and 40 min of the mean of each cluster, multiplied by 100 and rounded up to the first integer for easier visualisation.

The cluster-specific patterns of bindings around promoters for Pol II time series in Table 3.2 show that the largest clusters tend to be enriched in ER- $\alpha$  bindings as well as most other TFs and show elevated levels of Pol II 40 min after the stimulation with E2, thus their profiles are associated with early transcription. In contrast the majority of the clusters with smaller sizes which are depleted in ER- $\alpha$  and most of the other bindings are associated with lower Pol II occupancy after the simulation, which suggest lower response to stimulation with E2 or no changes comparing to the basal level of transcription of the genes. In Table A.1 we also considered overlaps of the TFs across the whole 300bp-extended genes. The patterns of bindings for the genes are similar to the ones around the promoters. In case of the distal enhancers and their Pol II dynamics, Table 3.1 shows that the five largest clusters are not only enriched in ER- $\alpha$  and FoxA1 beyond average level but also all other considered TFs, and show high occupancy of Pol II 40 min after the stimulation, suggesting that the enhancers may be associated with active transcription occurring at their target ER- $\alpha$  responsive genes. The profiles with lower than average enrichments of ER- $\alpha$  are linked with occupancies of Pol II which are lower or comparable to the levels before the stimulation. Comparison of the cluster-specific binding patterns around the promoters and at the enhancers, suggests that ER- $\alpha$  and FoxA1 bindings tend to co-occur more often at the enhancers than around the promoters.

We also considered cluster-specific bindings for ER- $\alpha$ . In Tables A.3 and A.4 we observe that for the largest ER- $\alpha$  clusters their binding-patterns over promoter and genes are enriched in FoxA1 and ER- $\alpha$  as well as most of other considered TFs. Their links with ER- $\alpha$  occupancy at time 40 min after the stimulation is less clear. For completeness, in Table A.2 we also considered the clusters of ER- $\alpha$  time series over

enhancers however we do not draw any conclusions from their TF-binding patterns.

### 3.3 Gene-enhancer links confirmed by ChIA-PET

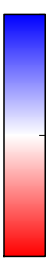
Next, we determined how many of our distal ER- $\alpha$ -bound enhancers are known to form links with promoter-extended genes. Before proceeding, in order to refine signal to noise ratio of our data, we removed genes and distal enhancers which possessed a total of less than 30 tags across all time points of their time series in each of our time course datasets. Next, we collected publically available ChIA-PET dataset from ENCODE/GIS-Ruan [64] for PolII and ER- $\alpha$  and MCF7. The datasets can be downloaded from <http://rohshdb.cmb.usc.edu/GBshape/cgi-bin/hgFileUi?db=hg19&g=wgEncodeGisChiaPet> or GEO accession numbers: GSM970209 and GSM970212, and viewed in UCSC browser at <http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeGisChiaPet>. The overall design and processing of the datasets can be found under GEO accession number GSE39495. The sources contain the high-confidence binding sites and protein-mediated chromatin interactions for ER- $\alpha$  and Pol II, including respectively 3 and 4 replicates for ChIA-PET with each of the antibodies. For the analysis we concatenated the replicates together.

In order to establish which of our promoter-extended-genes and distal ER- $\alpha$  bindings are confirmed to interact with ChIA-PET-detected interactions, we overlapped the joint set of regions of the both types (file *B*) with the concatenated set of ChIA-PET interactions (file *A*). For that we used bedtools' `pairToBed`. The function reports each interaction in file *A*, i.e. the file which stores coordinates of head and tail of each ChIA-PET interaction, which overlaps via both its head and its tail with at least two regions in file *B*, i.e. the file which stores coordinates of genomic regions. The output of the function consists of ChIA-PET links and per each a list of regions which overlap with that link. Using each list we generated all possible pairs of its elements and retained those pairs whose elements laid on the opposite sites of its ChIA-PET link. That process revealed a total of 2449 enhancer-promoter links, and shows that 1864 of our distal enhancers interact with at least one promoter. We suppose that many of our ER- $\alpha$  distal enhancers and their associations are most likely not picked up by the ChIA-PET method due to its limited sensitivity, and that the real number of interactions is much higher. The ChIA-PET detected interactions are used as a positive set for the purpose of developing our classifiers in chapter 4 and our prior in chapter 5 as well as in evaluation of the performance of the models.



	CTCF	ESR1	FoxA1	SRC-3	STAG1	TRIM24	c-Fos	c-jun	c-MYC	Count	Amplitude
1	170	390	220	556	436	153	220	66	3	731	199
2	128	303	175	378	317	111	165	54	4	502	197
3	128	310	128	344	279	106	129	38	1	491	264
4	72	224	100	260	171	59	61	17	4	434	165
5	87	168	127	277	210	72	77	22	5	425	73
6	42	113	118	221	152	39	67	21	3	355	-22
7	126	114	90	215	230	76	96	28	3	342	105
8	52	109	49	143	93	28	35	10	2	335	256
9	44	168	78	193	125	45	59	14	3	292	184
10	39	103	72	158	114	32	47	6	0	291	8
11	40	104	57	131	89	28	22	3	0	282	95
12	16	101	52	132	54	18	32	1	0	276	57
13	55	72	88	160	122	34	82	10	2	265	-27
14	24	49	96	148	95	23	46	3	0	263	-34
15	27	123	42	119	74	18	28	4	0	258	206
16	52	97	69	167	115	44	80	15	1	253	233
17	21	70	33	71	40	14	6	0	0	233	-8
18	17	66	31	65	37	6	8	4	0	221	-2
19	14	46	38	51	27	8	9	1	0	219	8
20	26	52	36	72	42	13	25	3	1	218	68
21	40	81	48	110	75	28	41	8	0	215	105
22	60	63	29	96	81	33	43	6	1	213	116
23	42	50	53	93	75	18	46	12	4	212	211
24	26	39	35	69	49	16	28	1	0	211	136
25	25	68	25	65	43	12	31	7	1	211	171
26	12	44	23	49	19	5	2	0	0	203	-15
27	10	49	27	41	28	14	20	5	0	203	-206
28	30	54	35	89	57	14	21	2	1	202	112
29	29	38	29	59	45	22	36	5	1	198	-2
30	22	66	54	94	51	14	21	5	0	198	137
31	15	47	48	75	36	12	9	0	0	190	120
32	22	32	49	91	67	14	40	12	2	183	-88
33	11	45	25	49	15	5	5	0	0	182	51
34	15	40	30	47	27	15	7	4	0	180	61
35	8	33	43	48	34	14	17	3	0	174	62
36	5	30	14	45	18	10	8	1	2	170	-51
37	10	29	25	40	18	7	12	1	0	167	-82
38	14	37	34	64	43	12	9	1	1	163	45
39	8	31	13	26	16	17	11	5	0	160	-127
40	16	47	23	38	24	10	6	1	0	160	177
41	13	46	9	25	13	14	3	1	0	143	-145

- enriched,  $p < 0.01$
- enriched,  $p < 0.005$
- enriched,  $p < 0.001$
- depleted,  $p < 0.01$
- depleted,  $p < 0.005$
- depleted,  $p < 0.001$
- neither



rises between 0-40min

stationary

drops between 0-40min

Table 3.1: The table shows the cluster-specific patterns of TF bindings across ER- $\alpha$  enhancers for the corresponding Pol II clusters in figure 3.4a. The count column indicates the size of the cluster. The amplitude column shows the difference between Pol II occupancies in E2-deprived (0 min) and E2-stimulated samples (40 min) in the mean dynamics of each cluster. (dynamics of orange replicate)

	CTCF	ESR1	FoxA1	SRC-3	STAG1	TRIM24	c-Fos	c-Jun	c-MYC	Count	Amplitude
1	244	64	18	84	305	77	32	24	50	434	258
2	216	95	32	77	255	79	28	12	3	391	216
3	206	44	18	57	235	81	36	25	30	359	185
4	215	57	28	59	242	69	31	14	2	341	82
5	160	27	11	34	185	39	10	3	15	299	78
6	154	38	27	41	189	44	21	6	7	290	28
7	187	48	31	43	225	78	34	9	8	287	224
8	165	49	15	42	187	62	24	11	7	270	301
9	152	56	23	68	199	64	22	9	6	259	331
10	152	26	31	36	184	55	17	9	9	255	-24
11	166	52	15	58	198	65	19	9	5	243	302
12	135	59	14	42	165	50	16	12	7	237	313
13	47	7	5	10	52	14	3	2	2	227	-238
14	123	33	14	42	144	50	21	11	3	226	14
15	115	26	15	32	121	32	15	11	12	218	264
16	140	34	16	37	159	50	20	6	5	208	206
17	93	20	24	36	118	42	25	9	2	196	2
18	111	30	8	38	142	43	18	13	7	194	229
19	63	12	7	13	87	19	5	0	1	194	-27
20	127	23	22	41	134	53	25	14	7	192	120
21	50	10	4	6	54	21	2	1	0	190	-174
22	36	6	7	6	41	9	2	0	0	161	-271
23	49	5	6	11	63	13	2	1	2	154	-117
24	86	15	18	31	102	26	12	2	2	152	-186
25	51	9	7	11	64	13	1	1	1	146	-219
26	34	10	2	5	40	7	5	1	0	146	-228
27	35	6	2	5	36	9	1	0	1	142	-39
28	56	10	10	9	71	20	6	0	0	140	-99
29	59	7	7	10	59	17	9	3	2	136	-106
30	86	17	9	24	99	33	11	7	4	133	215
31	58	11	7	18	67	19	3	3	3	118	57
32	50	9	7	10	63	22	1	1	0	117	-186
33	43	5	2	4	40	10	2	1	1	99	-45
34	36	12	6	9	43	10	4	0	1	96	63
35	22	7	6	4	22	9	0	0	1	93	-7
36	25	2	1	2	20	13	0	0	0	93	-99
37	15	6	1	3	14	3	1	1	0	85	133
38	17	4	1	1	17	5	1	0	0	85	-79
39	27	9	3	6	20	10	1	0	1	74	97
40	22	7	2	2	30	8	1	0	1	73	214
41	13	6	0	5	13	5	4	1	0	73	-24

enriched, $p < 0.01$
enriched, $p < 0.005$
enriched, $p < 0.001$
depleted, $p < 0.01$
depleted, $p < 0.005$
depleted, $p < 0.001$
neither

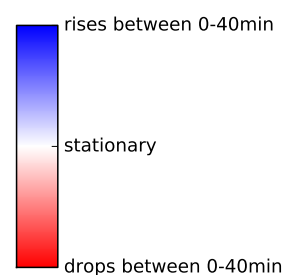


Table 3.2: The table shows the cluster-specific patterns of TF bindings across 2000bp-long TSS-centred regions for the corresponding Pol II clusters in figure 3.4c. The count column indicates the size of the cluster. The amplitude column shows the difference between Pol II occupancies in E2-deprived (0 min) and E2-stimulated samples (40 min) in the mean dynamics of each cluster. (dynamics of orange replicate)

### 3.3.1 Comparison of ChIA-PET confirmed and unconfirmed enhancers

We analysed the enrichments across the time series of ChiA-PET-confirmed interacting enhancers and those with no ChIA-PET data to check whether there are any underlying differences in the number of tags across their time series, and thus to assess the relative strength of their signals. The histograms of their enrichments are shown in Fig. 3.5, we observe that the distributions of each of the two classes are comparable. This analysis was carried out to provide evidence that the results derived on the basis of the interacting enhancers generalise to the set of enhancers with unknown status. This will be important in the predictive phase in Chapter 4 and in the inference of the model in Chapter 5.

### 3.3.2 Links within and outside domains

Since Topologically Associating Domains (TADS, see Section 2.3) are shared across mammalian genomes, we therefore use the TADS from [26] (Table S3 - Domains in mESC, mouse Cortex, hESC, IMR90) to stratify our interactions to inter-domain interactions and intra-domain interactions. The majority (82%) of enhancer-gene interactions lay within domains. In case when one of the interacting loci lays on a border of two domains, the interaction is considered intra-genic if the other interacting element is in any of the two adjacent domains.

## 3.4 Discussion

In this chapter we present the experimental protocol and methodology designed to investigate the effect of estradiol treatment on ER- $\alpha$  receptor and its resultant profile of genome-wide binding. We include the description of the pre-processing and transformation of the raw ChIP-seq data into a form which will be used as features in the supervised and unsupervised models covered in the Chapters 4 and 5. Among aspects of the pre-processing of data discussed, we include the alignment of sequenced reads of ER- $\alpha$  and other associated ChIP-seq datasets, discovery of estrogen binding sites with MACS and our approach to estimate the locations of time persistent bindings of ER- $\alpha$  given their time-dependent variability and noise in the data. The regions were used for the evaluation of time point read counts, and creation of time series data. The counts were normalised by a total number of reads in each dataset to correct for a

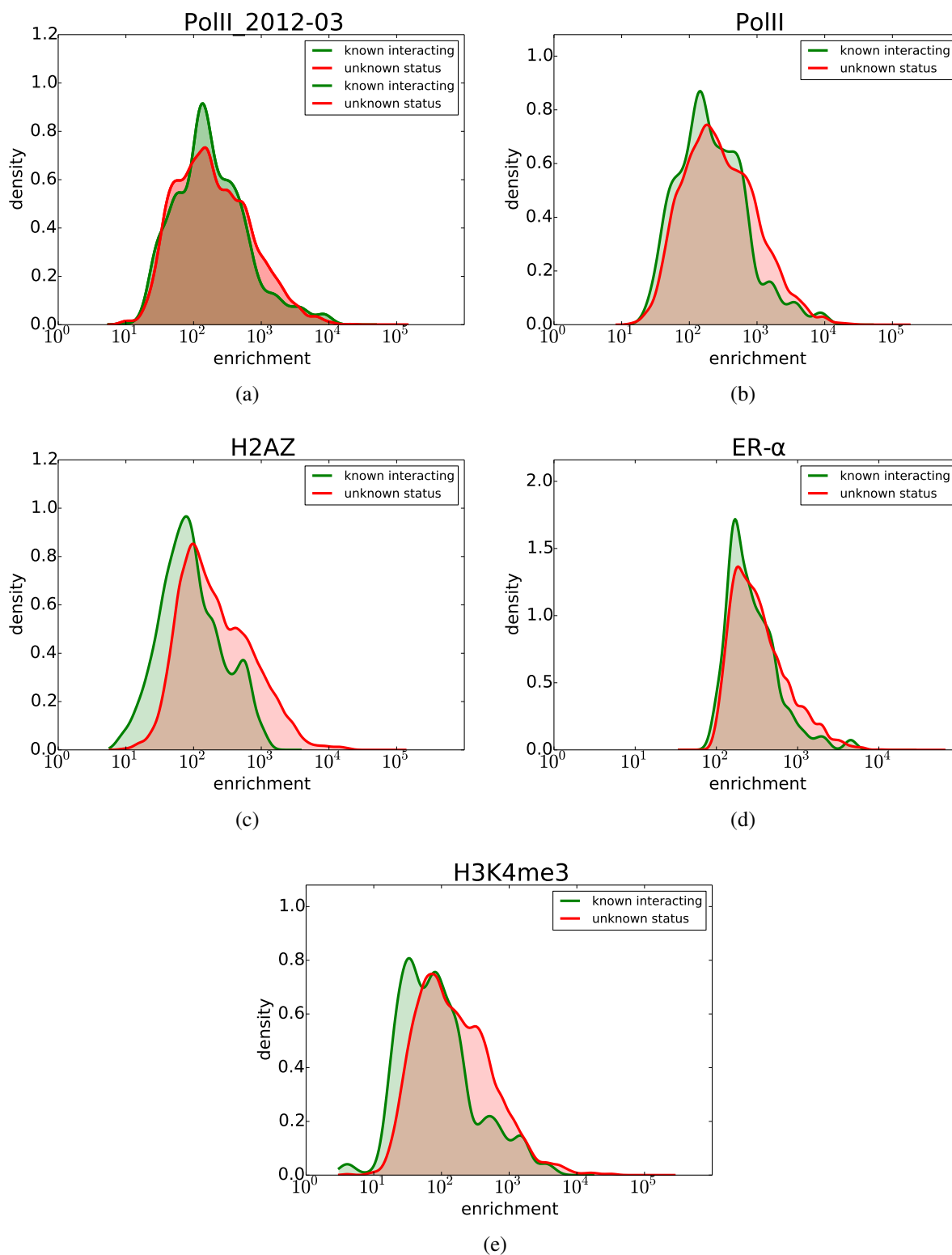


Figure 3.5: The graphs (a, b, c, d, e) show distributions of sums of ChIP-seq tags across time series of ChIA-PET-confirmed interacting enhancers (green) and the ones with unknown status (red) collected across all 23 chromosomes. The tags are calculated over enhancer bodies for ER- $\alpha$ , both Pol II replicates, H2AZ and H3K4me3 ChIP-seq datasets and normalised by the total size of each set.

lack of comparability between enrichment of time points in time series, resulting from non-unique mappings of some reads, differences in read depth and quantity of starting material.

We subjected the time series to a preliminary clustering analysis with the AP algorithm to reveal dynamics of time series across genes and distal ER- $\alpha$  bindings, and their putative usefulness for association of ER- $\alpha$  enhancers to their target genes. The clusters show substantial differences in occupancy dynamics across both genes and enhancers. This is expected for Pol II which is known to show a broad range of response profiles in this system [49] but we also see some differences in ER- $\alpha$  profiles suggesting that occupancy is not solely determined by the nuclear concentration of ER- $\alpha$  but could also be influenced by other cofactors. We attempted to link the clusters of Pol II across genes and enhancers to patterns of TF bindings at those two types of regions. The enrichments of several ER- $\alpha$  associated TFs from the system, stimulation and similar timing show that some early transcription could be associated with patterns of TFs. In general the higher the levels of enrichments of the TFs the more likely that the genes are transcriptionally active after the stimulation with E2. For enhancer time series of Pol II we observe that the largest clusters and their enhancers are significantly enriched for most of the considered TFs which suggests their involvement in active transcription, presumably occurring at their target genes. Also, the simultaneous binding of ER- $\alpha$  and FoxA1 is more likely to occur at the enhancers than at the promoters.

Overlapping the set of pairs of ER- $\alpha$  distal bindings sites and gene and promoter regions with ChIA-PET data revealed that the majority of enhancers lack experimentally confirmed empirical assignments most likely due to the limited sensitivity of the experimental technique. We showed that the interaction data were not biased towards enhancers, with higher number of ChIP-seq tags. Using experimentally validated pairs and publicly available locations of TADs, we confirm that the majority of ER- $\alpha$ -mediated enhancer-promoter contacts occur within the domains. The unknown targets of the ER- $\alpha$  enhancers could be addressed and inferred by computational models which use other complementary sources of evidence of chromatin contacts. Our efforts to build such models are presented in the next two chapters.

# Chapter 4

## Supervised learning

In this chapter we attempt to address the problem of linking ER- $\alpha$ -bound enhancers to their target genes by applying a Naive Bayes algorithm which is a supervised classification algorithm. We investigate whether similarities between time series data at enhancers and genes, quantified by the Pearson correlation coefficient, can be used to discriminate interacting from non-interacting enhancer-gene pairs. An important feature of our approach is that we combine time course ChIP-seq data with genomic distance to greatly improve performance.

In order to classify each pair of the loci, the method approximates the joint distribution of the features through an assumption of their conditional independence. Thus, each input dataset contributes equally to the binary classification task. This property is especially useful for our task since it allows densities to be modelled using simple one-dimensional estimates. In order to make the classifier robust to zero frequencies in test mode, corresponding to unobserved values in the training set, we apply kernel density estimation to model densities.

We train the method on the ChIA-PET interactions from all odd chromosomes, and test the model on interactions from all even chromosomes. Using the test data, we estimate FDR for previously unseen instances and in consequence make predictions for the ER- $\alpha$  enhancers which lack assignments in the ChIA-PET data.

Finally, we associate target genes with their chance of being differentially expressed under stimulation with estradiol, and validate our predictions using independent publicly available GRO-seq and RNA-seq data from the same system.

## 4.1 Enhancer-centric Naive Bayes model

Suppose that an enhancer  $j = 1, \dots, J$  regulates a gene  $k = 1, \dots, K$  at a number of time points, and that their contact is mediated by a protein. We can expect that the time course data at an enhancer  $j$  i.e.  $\mathbf{X}_j = (x_{j,1}, \dots, x_{j,D})$  and gene  $k$  i.e.  $\mathbf{Y}_k = (y_{k,1}, \dots, y_{k,D})$  would on average be more correlated for interacting pairs than their non-interacting counterparts. Here, we intend to learn the underlying distribution of correlations of the two classes of pairs for four complementary datasets and on their basis jointly classify a new unobserved instance. In addition, we combine the time course derived attributes with the corresponding distribution of genomic separation for interacting and non-interacting elements.

We propose an enhancer-centric model of contacts, where each enhancer can interact with only one gene. The model takes into account the ratios of products of two distributions to provide a probabilistic assessment that an enhancer  $j$  is more likely to be interacting with a particular gene  $k$  than any other gene. The gene-centric model is simplified in a sense that it does not take into account the promoter-promoter interactions which are known to play important part in the higher order gene regulatory complexes. There is also no model of the relationship between enhancers regulating the same promoter. This simplification leads to a particularly simple classification model.

### 4.1.1 The definition of the model

Our model is defined in terms of two  $K$ -dimensional random variables  $\mathbf{I}_j = I_{j,1}, \dots, I_{j,K}$  and  $\mathbf{D}_j = D_{j,1}, \dots, D_{j,K}$ . The first variable  $\mathbf{I}_j$  encodes a structure of simultaneous contacts of a given enhancer  $j$  with its surrounding  $K$  putative target genes. The random variable  $\mathbf{I}_j$  is composed of  $K$  binary entries  $I_{j,k}$  indicating whether the  $(E_j, G_k)$  forms an interacting ( $I_{j,k} = 1$ ) or non-interacting pair ( $I_{j,k} = 0$ ). The random variable  $\mathbf{D}_j$  is a  $K \times N$ -dimensional matrix of observed attributes with each row  $(D_{j,k})$  consisting of  $N$  values of pair-wise comparisons between time series of an enhancer  $j$  and a gene  $k$ , and their genomic location. The first set of comparisons rely on Pearson correlation and involves calculating its value  $c_{j,k,n}$  for each pair  $(E_j, G_k)$ , i.e. its time series  $(\mathbf{X}_{j,n}, \mathbf{Y}_{k,n})$ , and for each dataset  $n \in N$ , where  $N$  is a number of time course ChIP-seq datasets. Additionally, the vector also contains the Euclidean distance  $d_{j,k}$  calculated between the genomic coordinates of the ENCODE defined canonical TSS (Homo.Sapiens.GRCh37.75 (hg19)) of a gene  $k$  to the centre of an enhancer  $j$ . The

distance serves as another attribute.

The joint likelihood of the model can be written as:

$$P(\mathbf{D}_j, \mathbf{I}_j) = P(\mathbf{D}_j | \mathbf{I}_j) P(\mathbf{I}_j) . \quad (4.1)$$

The model provides a probability of observing a particular  $\mathbf{D}_j$  under a given structure  $\mathbf{I}_j$ . Due to its regulatory role, an enhancer is unlikely to regulate a high number of genes, thus we can expect that the true  $P(\mathbf{I}_j)$ , which in the Bayesian treatment is a prior distribution over the structures, would be sparse. Moreover, we could expect that  $D_{j,k}$  and  $D_{j,k'}$  of any two interacting pairs  $k, k'$  would be interlinked, as correlations between gene-enhancer pairs are not independent variables. These dependencies would be reflected in a true form of the likelihood  $P(\mathbf{D}_j | \mathbf{I}_j)$ . Lastly, we could also expect that the  $N + 1$  attributes i.e correlations  $c_{j,k,n}$  and distance  $d_{j,k}$  of a pair  $j, k$  of the vector  $D_{j,k}$  would also be correlated.

### 4.1.2 Approximate joint likelihood

The modelling of all dependencies however is infeasible given the sparsity of our training data. Here, we restrict the form of the joint distribution and construct an approximate joint probability of enhancer-gene contacts. Pairwise correlations allows a valid likelihood if we restrict ourselves to one gene per enhancer. To model multiple enhancers per gene requires a likelihood over the data vectors as considered in Chapter 5.

#### a) The joint distribution factorises

We assume that the likelihood  $P(\mathbf{D}_j | \mathbf{I}_j)$  can be factorised and written in the form:

$$P(\mathbf{D}_j | \mathbf{I}_j) = \prod_{\{k: I_{j,k}=1\}} P(D_{j,k} | I_{j,k} = 1) \prod_{\{k: I_{j,k}=0\}} P(D_{j,k} | I_{j,k} = 0) \quad (4.2)$$

where  $\mathbf{I}_j = I_{j,1}, \dots, I_{j,N}$  and  $\mathbf{D}_j = D_{j,1}, \dots, D_{j,N}$ . Hence the distribution of each  $D_{j,k}$  is conditionally independent of other allocations and conditional only on the indicator variable  $I_{j,k}$ .



### b) An enhancer regulates a single gene

We assume further, that an enhancer  $j$  can interact with only one gene  $k$ . We restrict the event space of  $P(D_j, I_j)$  to its subspace  $P(D_j, \mathbf{I}_{j,k}^{(1)})$ , where  $\mathbf{I}_{j,k}^{(1)} = 0, \dots, \underset{kt\text{th}}{1}, \dots, 0$ . From 4.2 the events are given by:

$$P(\mathbf{D}_j | \mathbf{I}_{j,k}^{(1)} = 0, \dots, \underset{kt\text{th}}{1}, \dots, 0) = P(D_{j,k} | I_{j,k} = 1) \prod_{\{l:l \neq k\}} P(D_{j,l} | I_{j,l} = 0). \quad (4.3)$$

The prior distribution  $P(\mathbf{I}_j)$  follows a multivariate Bernoulli distribution, and thus the restriction is equivalent to setting the probabilities of all the structures  $\mathbf{I}_j$  with non-singular number of contacts i.e.  $\mathbf{I}_j^{(2)}, \mathbf{I}_j^{(3)}, \dots, \mathbf{I}_j^{(K)}$  to zero. For the remaining  $\mathbf{I}_{j,k}^{(1)}$  we assume that the prior is uniform across these sparse vector, i.e.

$$P(\mathbf{I}_{j,k}^{(1)} = 0, \dots, \underset{kt\text{th}}{1}, \dots, 0) = 1/K. \quad (4.4)$$

hence each  $\mathbf{I}_{j,k}^{(1)}$  is equally likely.

### c) The distribution of attributes is independent

Assuming that the attributes are conditionally independent, the likelihood component  $P(D_{j,k} | I_{j,k})$  becomes:

$$P(D_{j,k} | I_{j,k}) = P(d_{j,k}, c_{j,k,1}, \dots, c_{j,k,N} | I_{j,k}) = P(d_{j,k} | I_{j,k}) \prod_{n \in \mathcal{N}} P(c_{j,k,n} | I_{j,k}) \quad (4.5)$$

where  $d_{j,k}$  is a distance from the centre of an enhancer  $j$  to the TSS of a gene  $k$ , whereas  $c_{j,k,n}$  is a correlation between the time series of the  $n^{\text{th}}$  time course dataset between an enhancer  $j$  and gene  $k$ .

The above assumption of conditional independence in 4.5 and the fact that each vector  $\mathbf{I}_{j,k}^{(1)}$  is a 1-of-K (i.e one-to-one relation) representation of  $K$  class indicators makes this algorithm a special case of Naive Bayes (NB) model and hence we will refer to our model as Naive Bayes.

## The likelihood

Combining the assumption of the factorisable likelihood (4.2) with the conditional independence of attributes (4.5) yields,

$$P(\mathbf{D}_j | \mathbf{I}_j) = \prod_{k=1}^K P(\mathbf{D}_{j,k} | I_{j,k}) = \prod_{k=1}^K \left[ P(d_{j,k} | I_{j,k}) \prod_{n \in N} P(c_{j,k,n} | I_{j,k}) \right] \quad (4.6)$$

Restricting the event space to single enhancer-gene events (4.3) results in,

$$P(\mathbf{D}_j | \mathbf{I}_{j,k}^{(1)}) = \left[ P(d_{j,k} | I_{j,k} = 1) \prod_{n \in N} P(c_{j,k,n} | I_{j,k} = 1) \right] \prod_{\{l: l \neq k\}} \left[ P(d_{j,l} | I_{j,l} = 0) \prod_{n \in N} P(c_{j,l,n} | I_{j,l} = 0) \right] \quad (4.7)$$

### 4.1.3 Posterior enhancer-gene allocations

The posterior distribution of the model is:

$$P(\mathbf{I}_{j,k}^{(1)} | \mathbf{D}_j) = \frac{P(\mathbf{D}_j | \mathbf{I}_{j,k}^{(1)}) P(\mathbf{I}_{j,k}^{(1)})}{\sum_{k=1}^K P(\mathbf{D}_j | \mathbf{I}_{j,k}^{(1)}) P(\mathbf{I}_{j,k}^{(1)})} \quad (4.8)$$

The posterior distribution can be used to find the probability of each structure  $\mathbf{I}_{j,k}^{(1)}$  given the pair-wise comparisons in  $\mathbf{D}_j$ , i.e. the values of the data-specific correlations and distance for each pair  $(E_j, G_k)$  and all complementary pairs  $(E_j, G_{\{l: l \neq k\}})$ . The posterior probabilities can be used to infer the most likely target of an enhancer  $j$  out of  $K$  genes.

## 4.2 Training of the model

Since the  $P(\mathbf{D}_j | \mathbf{I}_j)$  is factorisable (4.2) and the attributes are conditionally independent (4.5), thus we only need to estimate the univariate distributions of attributes for positive ( $I_{j,k} = 1$ ) and negative ( $I_{j,k} = 0$ ) pairs, i.e.  $P(d_{j,k} | I_{j,k})$  and  $P(c_{j,k,n} | I_{j,k})$  where  $n \in N$ , to train the model. The product of the attributes i.e.  $P(\mathbf{D}_{j,k} | I_{j,k})$ , along with the prior  $P(I_{j,k})$  completely characterises each interacting and non-interacting pair  $(E_j, G_k)$ .

### 4.2.1 Set of data-specific correlations and genomic distances

The method uses five attributes: correlations of Pol II, ER- $\alpha$ , H2AZ, and H3K4me3 ChIP-seq time course data as well as one distance-based. In order to create the first four sets, we correlated each individual normalized time series from distal ER- $\alpha$  enhancers with their corresponding time series from promoter-extended gene bodies (i.e. data aggregated over a gene region and the 300bp-long upstream region). To insure robustness, for the Pol II replicates we replaced each pair of correlations with their average value. The other feature is a genomic distance between each (E, G) pair, calculated from an enhancer's centre to a gene's canonical TSS shifted upstream by 300 bp. We transformed the absolute values of the distances to  $\log_{10}$  space and used the sign to distinguish upstream (positive sign) and downstream (negative sign) connections.

### 4.2.2 Set of interacting and non-interacting pairs

In order to estimate the distributions of correlation  $P(c_{j,k,n}|I_{j,k})$  and distance  $P(d_{j,k}|I_{j,k})$  as well as to validate the model, we required a labelled dataset consisting of examples of interacting and non-interacting enhancer-gene pairs. For the positive set, we overlapped the combined set of MCF7 ChIA-PET of Pol II and ER- $\alpha$  with the set of pairs of non-overlapping ER- $\alpha$  enhancers and extended-genes (refer to Section 3.3 for more technical details).

To define the negative set, we restricted ourselves to all enhancer-gene pairs involving known interacting enhancers coming from the positive set and all the remaining non-targeted genes. Enhancers without any confirmed interactions from ChIA-PET data were not used for training as we have no information about their target genes.

### 4.2.3 Training, test and predictive sets

We partitioned the total set of positive and negative interactions into training and test sets. The training set consisted of full set of enhancer-promoter pairs from odd chromosomes where enhancers have a known interaction partner from ChIA-PET evidence. Similarly, the test set was formed from all corresponding pairs from even chromosomes. When training and testing the classifier, we have not included enhancers that do not have any interactions according to the ChIA-PET data. These enhancers are most likely not picked up by the ChIA-PET method due to its limited sensitivity and

would introduce many false negatives into our training and testing data. However, we do apply the classifier to all enhancers when making target gene predictions.

#### 4.2.4 Kernel density estimation of distributions of attributes

We used KDE with a Gaussian kernel and interacting (positive) and non-interacting (negative) samples from the training set to estimate the distributions  $P(c_{j,k,n}|I_{j,k})$  and  $P(d_{j,k}|I_{j,k})$ . KDE provides smooth PDFs over a pre-specified region which ensures that the frequencies for unseen instances, lying in-between observed values, are non-zero. That provides the advantage that any distributions estimated in the training phase can be also used for testing and in predictive phase.

We used the data-specific correlations and distance of our interacting (positive) and non-interacting (negative) pairs to estimate each distribution. To ensure that the bandwidths of positive distributions are biologically meaningful and robust, we used cross-validation. As part of the approach, we iteratively ignored all pairs of each individual chromosome from our total and calculated the likelihood of KDE of the resultant reduced set. We then used the maximum of the sum of the log-likelihood for multiple values of the bandwidth as an estimate of its value. In contrast, due to a large number of negative examples and computational cost associated with the KDE, employing the same approach for negative distributions would be infeasible. Their size, however, also entails less requirement for optimised fitting, and thus to select the bandwidth we resorted to the Scott's rule [102].

### 4.3 Evaluation of the model

Following the training of the model on the training data, we evaluated the posterior probabilities for each enhancer-promoter interaction in the training and test sets. We could then make predictions based on posterior probabilities and estimate precisions for different posterior probability thresholds and different combinations of attributes. The precisions were then visualised using precision-recall (PR) curves. The training and test errors could also be established for different cut-offs. Since the test data was not used to train the model we consider the test errors to be reasonable estimates of performance. Even though in Section 3.3.1 we established that the average ChIP-seq tag counts across time courses of ChIA-PET-confirmed interacting enhancers and unconfirmed enhancers and thus signal strengths are comparative, we note that the

performance assessed for enhancers with grand truth ChIA-PET data could be unrepresentative for those enhancers with no data.

### 4.3.1 Precision-Recall curves

In order to create the curves, we sorted the probabilities of positive and negative elements. Next using the ordered scores, we estimated precisions at a number of monotonously decreasing probability cut-offs. The precisions are defined as a number of positives to the number of negatives and positives, with posterior probabilities over a threshold value. We also obtained a corresponding true positive rate (TPR), defined as ratio of the number of predicted positives to the total number of predicted positives.

To measure the performances of the model, we plotted the TPR values of 10%, 20%, and 30% against their precisions. Both values corresponded to some levels of threshold. We measured the model's performance for a number of selected combinations of input data, to choose the most informative combination.

### 4.3.2 MAP curves

Additionally, we considered an alternative global MAP measure. Under our model each enhancer possesses a maximum a posteriori (MAP) choice of gene. This gene is our best guess of an enhancer's target. The MAP measure is the fraction of times the MAP inferred target genes could be found among positive set of interactions.

### 4.3.3 The test and training errors

The test and training errors (FDRs) were computed at each posterior probability threshold as,

$$\text{FDR} = 1 - \text{Precision}(\text{TPR}) . \quad (4.9)$$

Since the error function is not monotonously decreasing, we searched for a unique lowest threshold which satisfied a selected error rate while maximising the TPR (a right most TPR value of PR curves). This value of test error is used in the predictive phase for inferring missing links of the remaining non-interacting (according to ChIA-PET) ER- $\alpha$  enhancers.

#### 4.3.4 Performance within and outside TADs

We stratified our predicted interactions at 10%, 20%, and 30% TPR thresholds into those that lie within domain and those that crossed domain boundaries. Each TPR threshold maps to a subsets of negative and positive links, and therefore each subset was partitioned into inter- and intra- domain interactions. We then tested precisions at each of the thresholds for each of the two subsets.

#### 4.3.5 Validation of model's predictions with GRO-seq and RNA-seq data

We used our model to infer gene targets of interacting and non-interacting enhancers. We only accepted the enhancer-gene pairs with probability values exceeding a test-data determined threshold with the  $FDR \geq 0.25$ . To summarise the links, we combined the filtered posterior probabilities according to the rule in eq 4.10

$$P(\text{card}(\{j \in J : I_{j,k} = 1\}) > 0) = 1 - \prod_{\{j \in J : I_{j,k} = 1\}} (1 - P(I_{j,k}^{(1)} | \mathbf{D}_j)) \quad (4.10)$$

to provide the probabilities that a gene  $k$  is regulated by at least one enhancer. The product in the equation is equal to the probability that none of the predicted regulators  $\{j \in J : I_{j,k} = 1\}$  of the gene  $k$  regulates the gene. The higher the number of the assigned regulators the lower the product in the equation and thus the higher the probability that at least one of the predicted enhancers regulates the gene  $k$ , increasing our confidence that the gene  $k$  is ER- $\alpha$  regulated.

The high-precision, publically available GRO-seq data from [42], described and accessible at GEO accession number GSM678536, aimed to detect the primary estrogen target genes, i.e., those genes which are being actively transcribed shortly after treatment with E2. This experiment detects only the early changes in gene expression of the primary targets, while later changes in the expression of genes which are not directly regulated by ER- $\alpha$  are not detected.

The experiments were performed in the same cell-line and context as ours. Using the data and our cumulative scores from 4.10, we assessed how many of our predicted distally regulated genes were differentially expressed at early time points. Using the EdgeR processed GRO-seq data from GSE27463 we filtered the GRO-seq determined DE genes at 10, 40, 160 min after E2 stimulation with q-value (multiple hypotheses testing adjusted p-values from EdgeR) of less than 0.05, 0.01 and 0.001. For each

q-value, we combined the DE genes from each of the time points into a single list.

Similarly, we repeated the assessment with our in-house RNA-seq data determined DE genes. In order to determine the DE genes from our RNA-seq data, we first mapped our  $t = 0, 30, 60,$  and  $90$  (two replicates for each time-point) RNA-seq datasets to hg19 genome using Tophat [121] and counted the reads over genes using HTseq-count [1]. We then used DESeq2 [69] to perform differential expression for  $t = 30, 60,$  and  $90$  against time 0 (control). We filter the genes with q-value of less than 0.05, 0.01, 0.001, and as previously combined the DE genes from each time point into one list of early responsive DE genes. The data will be used for validation of our predictions (see Section 4.4.6).

## 4.4 Results

We demonstrate our method using ChIP-seq time course data collected from the MCF7 breast cancer cell-line stimulated by estrogen. After stimulation, the ER- $\alpha$  TF translocates into the nucleus where it binds numerous enhancers to regulate transcription of target genes. The genome-wide occupancy of ER- $\alpha$  along with RNA polymerase (Pol II) and two histone marks (H3K4me3 and H2AZ) were measured via ChIP-seq at eight consecutive times after exposure of cells to estradiol. Public ChIA-PET data are also available in this system for testing our method’s performance [34, 63, 64]. We train and test the model as described in Section 4.2. Here we show the results of the analysis.

### 4.4.1 Time series correlation and distance-based features are informative about enhancer-promoter interactions

Figure 4.1 shows the distributions of  $P(c_{j,k,n}|I_{j,k})$  for each dataset estimated on the training dataset. As expected, the distributions of positive interactions differ substantially from the non-interacting pairs for all four datasets, with interacting regions more highly correlated on average. The difference is most pronounced for ER- $\alpha$  and Pol II (Fig. 4.1a and Fig. 4.1b) while there is a much smaller difference for the histone marks H2AZ and H3K4me3 (Fig. 4.1c and Fig. 4.1d).

Similarly, we also compare the distribution of genomic separation for interacting and non-interacting pairs in Fig. 4.1e. Although a highly informative feature, there is a substantial overlap in the positive and background distance densities due to the large separation of many ER- $\alpha$  bound enhancers from their target promoters; distance alone

is insufficient for accurate prediction of interactions, it is however a useful addition to the other attributes.

#### 4.4.2 Fisher's Linear Discriminant performance

Prior to developing our models, we investigated whether Fisher's linear discriminant analysis, which is a commonly used classifier, could perform efficient discovery of interacting and non-interacting enhancer-promoter pairs using combinations of correlation and distance-based features. Fisher's discriminant analysis finds  $c - 1$  linear discriminant axis, where  $c$  equals to the number of class labels in a given problem, such that when data points of each class are projected on the axis, the between-class separation of the resultant projections is highest.

We trained Fisher's linear discriminant analysis on the data from odd chromosomes with 1305 interacting and 1,068,522 non-interacting pairs (i.e.  $c = 2$ ), and tested the method on all remaining 1144 of interacting and 713,790 of non-interacting pairs from all even chromosomes. Figures B.1 and B.2 show class-size-normalised (i.e. divided by their class size) and standard relative-size-reflecting histograms for the projections of the two classes of data in the training set. Fig. B.1 shows that size-normalised histograms of the projections of the two classes appear to be well separated. However, Fig. B.2 shows that the projections of interacting pairs account for a very small fraction in the total number of projections.

In order to formally test performance of the classifier, we used Precision-Recall (PR) curves with increasing values (i.e. from the most negative to the most positive value of the projections) of the cut-off levels. Fig. B.3 shows that performance of the classifier is very low. That finding highlighted the need for less generic and more problem-specific methods.

#### 4.4.3 Naive Bayes classifier performance

We developed a Naive Bayes classifier which integrates several discriminative features to estimate the probability of interactions between an enhancer and putative target genes. Fig. 4.2 shows predicted interactions with only a small number confirmed by ChIA-PET (green). We evaluated classifier performance using precision-recall (PR) curves (Fig. 4.3a and Fig. 4.3b). The classifier was trained on data from odd chromosomes and the results were used to establish which combination of features is most informative. Data from even chromosomes was then used as an unbiased test set to



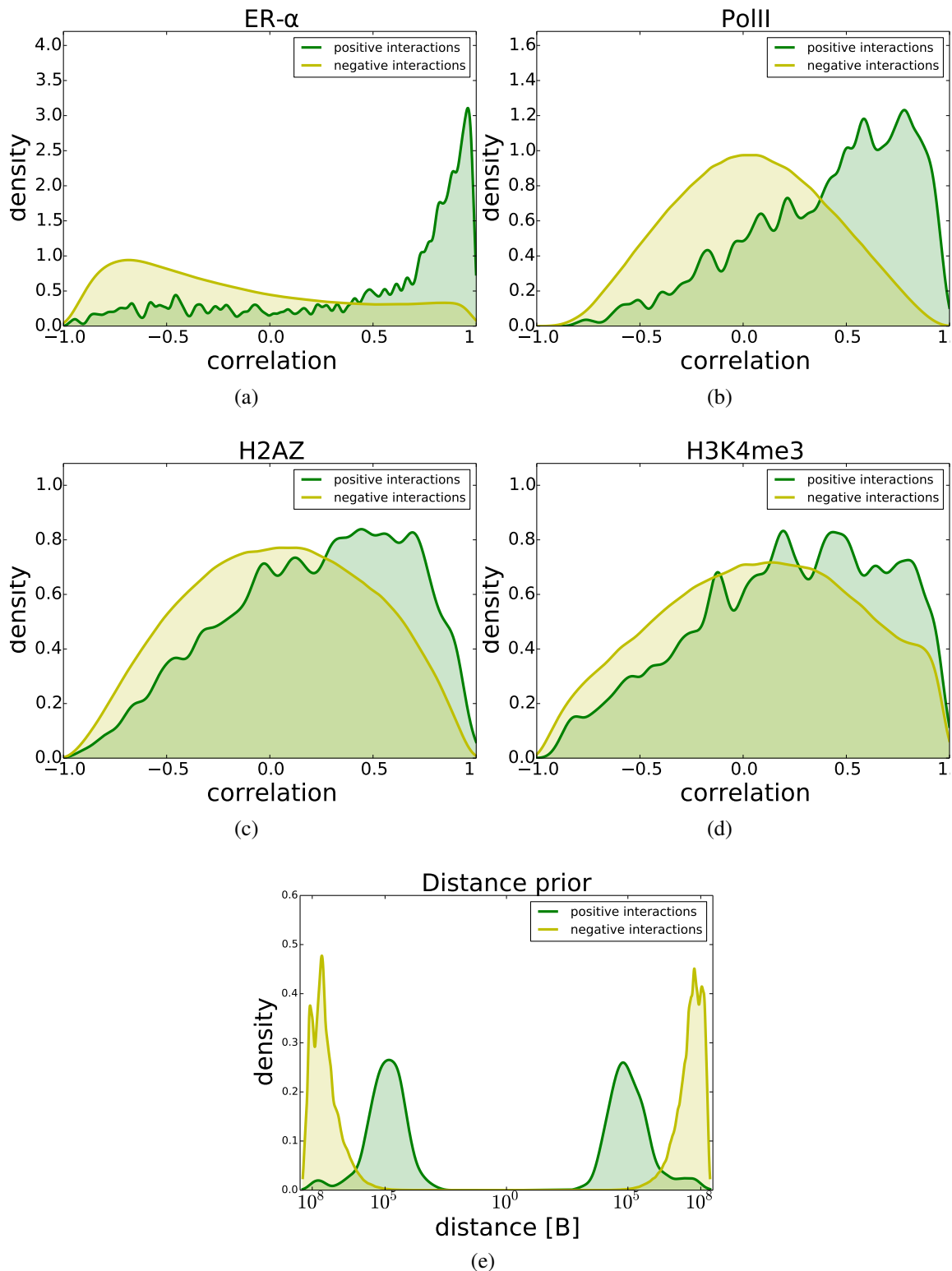


Figure 4.1: The graphs (a, b, c, d) show positive (green) and negative (yellow) distributions of correlations between pairs of time series of 300bp-upstream-extended-gene regions and enhancer bodies for ER- $\alpha$ , Pol II, H2AZ and H3K4me3 collected across all odd chromosomes. The figure (e) shows the distribution of genomic distances between centres of distal enhancers and 300bp-upstream-shifted-TSS of genes. The set of positive and negative pairs was constructed using 300bp-upstream-extended-genes and distal enhancers.

establish the performance of the selected model and to estimate decision cut-off levels. However, we do not observe significant over-fitting which is not surprising given the relatively small number of features used. Comparison of different combinations of correlations and distance features, including distance-alone and correlation-alone variants shows that data from ER- $\alpha$  or Pol II or both can be combined with distance to greatly enhance predictive performance. The H2AZ and H3K4me3 time course data were found not to be particularly informative, consistent with Fig. 4.1 which shows these histone marks to have a less pronounced difference in distribution for positive and negative links. Table 4.1 shows that using the probability cut-offs to infer links across all 23 chromosomes our model (combination of features: Pol II, ER, distance) consistently outperforms the distance-alone model in terms of the number of discovered true links. We show that at FDR equal to 0.25 our model infers 33-times more interactions than predictions based on proximity alone (see table 4.1). In addition to considering PR benchmark we also tested how often using maximum a posteriori probabilities (MAP) to link enhancers (in the training and test data) to their most probable promoters would result in correct assignments according to the ChIA-PET data (right-most column of plots in Fig. 4.3a and Fig. 4.3b).

FDR	data/distance	distance	ratio
0.4	20987	7947	2.6
0.3	12645	1611	7.8
0.25	8432	258	32.7
0.2	4359	230	19.0
0.1	1132	213	5.3

Table 4.1: Table shows the number of predicted links by distance-alone and distance-assisted models and varying test errors.

#### 4.4.4 Inter-domain and Intra-domain predictions

In order to assess the performance of the model on discovery of intra-domain interactions and the ones involving elements from two different domains, we stratified our predicted interactions into those two groups, and used PR and MAP measures.

The majority (81%) of enhancer-promoter interactions lay within domains. The PR curves in Fig. 4.4a and 4.4b show that combining ER- $\alpha$  data with distance information yields the best results. Interestingly, a comparison of distance-alone and data-alone cases within domains, showed us that the performances of those two are comparable,

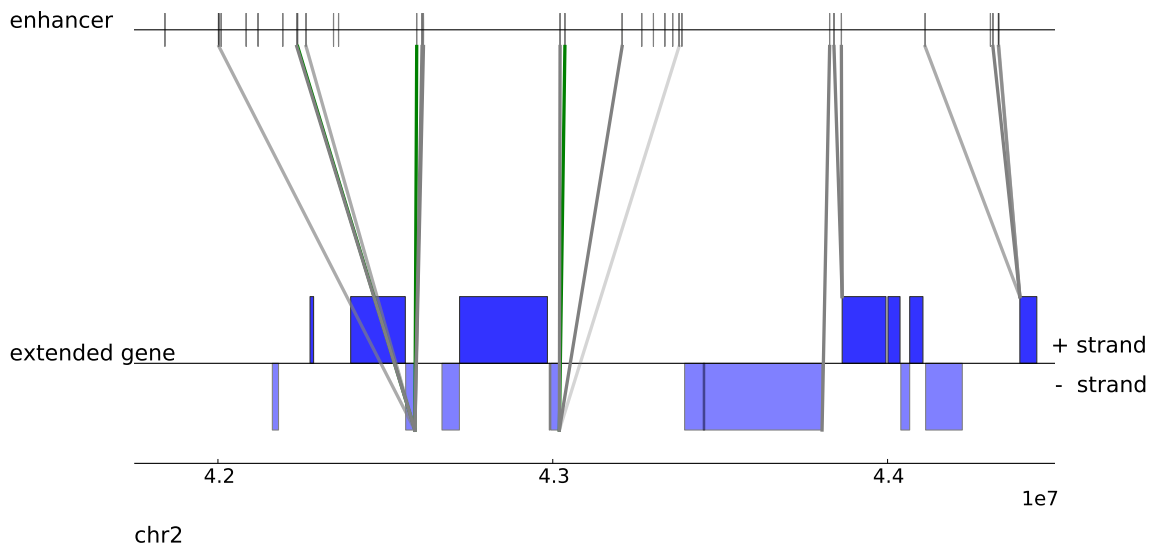


Figure 4.2: The cartoon shows the NB inferred interactions in three classes with decreasing test errors of 0.2, 0.25, 0.3 FDR levels, and corresponding lower bound probability cut-offs of 0.66, 0.47, 0.35. The class membership of each predicted link and its confidence level is indicated by its darker/lighter shading of its colour (more/less confident). The green/grey colour indicates whether each predicted link is confirmed/unconfirmed by ChIA-PET.

and that data alone possesses a very high predictive power. Thus, in summary, we deduce that these features are complementary to each other.

On the contrary, see Figures: 4.5a, 4.5b, focusing on the remaining inter-domain interactions we noticed that, due to a large number of negative interactions for feature spanning TADS, the correlation data alone is insufficient for efficient classification. The proximity criterion, despite being better than the data-alone, also does not offer the performance that we achieved in the intra-domain cases. Interestingly although the above are largely imprecise, the distance-assisted correlation variants still performs not only better than data but also distance-alone models, consistent with the previous section.

#### 4.4.5 Alternative data processing strategies

We investigated different promoter sizes for promoter-extended-gene regions and their effect on the performance of the model. The comparison between PR curves in Fig. B.9, B.10 and B.11, B.12 in the appendix shows that increasing the promoter sizes

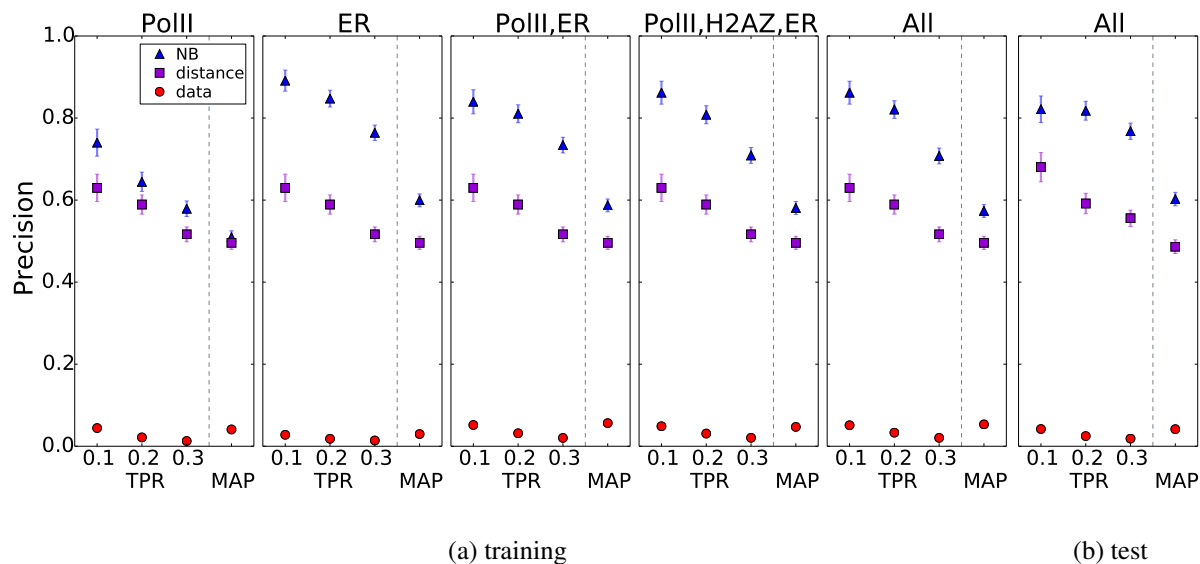


Figure 4.3: Figure shows the performance of the model for training/test data, measured by Precision-TPR and MAP scores, for selected combinations of datasets in the model. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores (posterior probabilities), and values above corresponding thresholds.

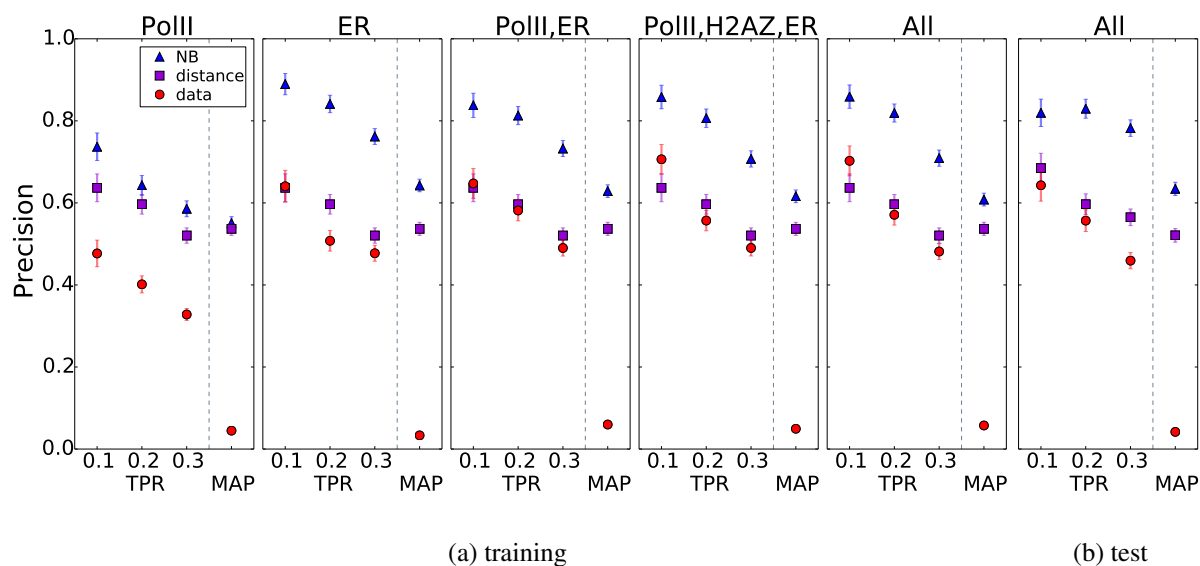


Figure 4.4: Figure shows the performance of the model for training/test data. As in Fig. 4.3 but only considering interactions within TADs.

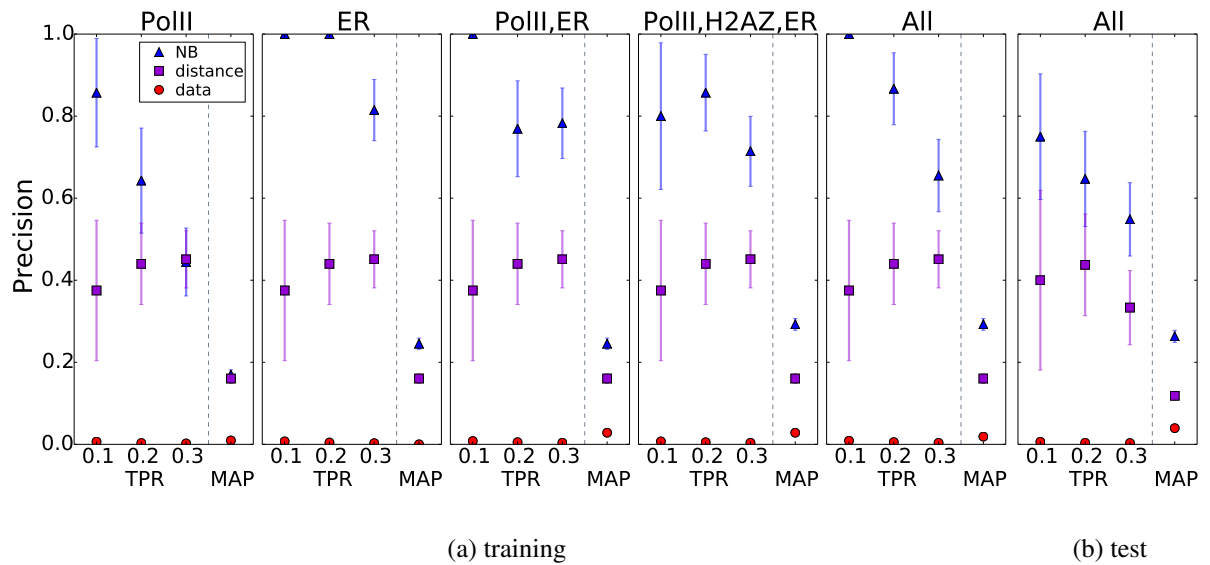


Figure 4.5: Figure shows the performance of the model for training/test data. As in Fig. 4.3 but only considering interactions spanning two TADs.

from 300bp to 1500bp produces no effect on the overall performance.

We also tested an alternative promoter-centric approach in which we overlapped the ER- $\alpha$  enhancers with 3000bp-long regions centred at TSS. That resulted in a positive set with reduced number of enhancer-intragenic links. We used the positive set and the unchanged negative set from the previous definition for training and testing of the model. Figures B.11, B.12 and B.13, B.14 show a performance comparison between the promoter-centric and the promoter-extended (1500 bp) gene models. In contrast to the original promoter-extended-gene model, the Fig. B.13 shows that the time course data of H3K4me3 and H2AZ, which are characteristic chromatin marks of promoter regions, improve the performance of the promoter-centric model, however the model has less coverage and ignores some functional intra-genic links. Comparison of the histograms of the features for the TSS-centric and the original promoter-extended model: B.5, B.6, B.7, B.8 shows that the enrichments of H3K4me3 for the TSS-centric model are more positively correlated.

We used GRO-seq data to compare the ability of inferring DE genes of the two model. The Fig. B.20 shows that the precision and the number of predictions of ER- $\alpha$  targeted genes differs, however the differences are not statistically significant.

#### 4.4.6 Naive Bayes predicts ER- $\alpha$ -regulated transcriptionally active genes

Finally, we used our method with Pol II and ER- $\alpha$  time course data to provide a highly confident ( $FDR \geq 0.25$ ) list of direct gene targets of ER- $\alpha$  in the context and cellular system under study. This list included 3146 genes with at least one enhancer link, and consisted of those genes which are regulated either by both distal and by our previous assumption intra-genic enhancers located within their bodies or the ones regulated solely by distal enhancers. For each gene on the list, using 4.10, we calculated the probability that at least one of its NB assigned regulators controls its gene. The score is higher for those genes for which our model assigns a higher number of distal enhancers.

We then investigated whether any of the genes from our list were actively being transcribed shortly after the stimulation with E2. For that we used GRO-seq and RNA-seq data-derived differentially expressed genes. The data came from the same context and cell line as our experiments, and early time points after stimulation with E2 (see Section 4.3.5).

In Fig. 4.6 we linked the scores of our genes with their transcriptional activity. In that figure, starting from the highest and gradually decreasing the value of cut-off level until it reached its minimum, we extracted subsets of genes from our list with values of scores above each cut-off level and at each of the steps compared against a list of differentially expressed (DE) genes obtained from GRO-seq data in [42]. PR curves showed that the larger the value of the score of a gene, the higher the chance that the gene is differentially expressed. Fig. 4.7 shows the corresponding DE genes as established from our in-house RNA-seq data. Both figures show also the accuracy of using the absolute value of the distance (proximity) between gene's canonical TSS and its closest ER- $\alpha$  binding as a predictor of gene's activity for the list of genes which possessed at least one ER- $\alpha$  binding within 40kB from their canonical TSS. Both GRO-seq and RNA-seq Figures show that NB is capable of discovering DE genes with a much higher precision than the closest gene approach.

#### 4.4.7 GO enrichment

We tested whether our predicted genes are enriched in gene ontology classes with ToppGene GO tool [18]. We assessed the ER- $\alpha$  regulated genes on the basis of their involvement in biological processes, diseases, and potential treatments. Table 4.2

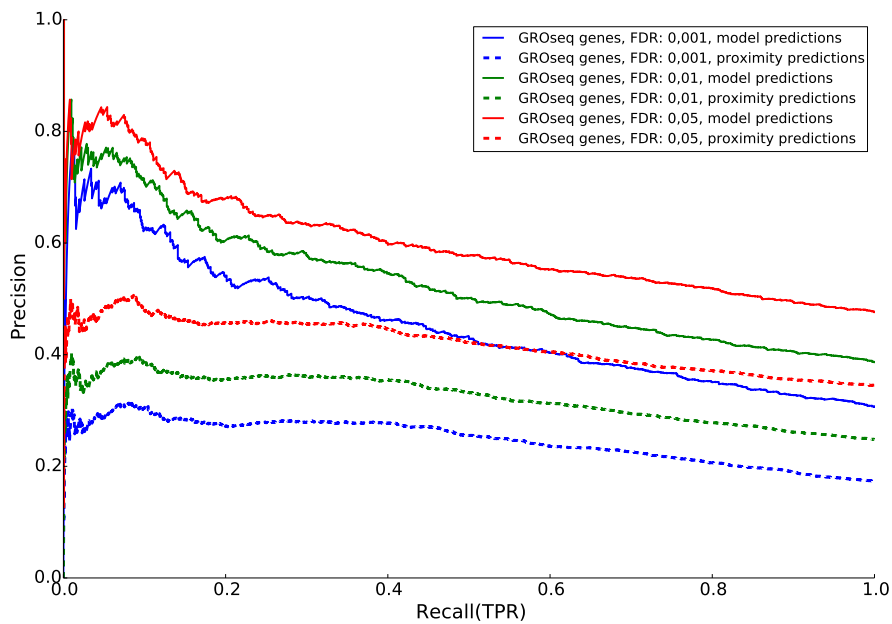


Figure 4.6: The PR curves assess the performance of Naive Bayes and proximity criterion on the ability to predict differentially expressed genes as detected by GRO-seq experiment and extracted at 3 different q-values (confidence levels) of 0.001, 0.01, 0.05. The predictor for the transcriptional activity of each gene of the NB predicted ER- $\alpha$  regulated genes is the probability that at least one of its NB assigned distal regulators indeed controls it. The Second predictor for the transcriptional activity of each gene of the genes with at least one ER- $\alpha$  binding within 40kB from their canonical TSS is the absolute value of the distance (proximity) between its canonical TSS and its closest ER- $\alpha$  binding.

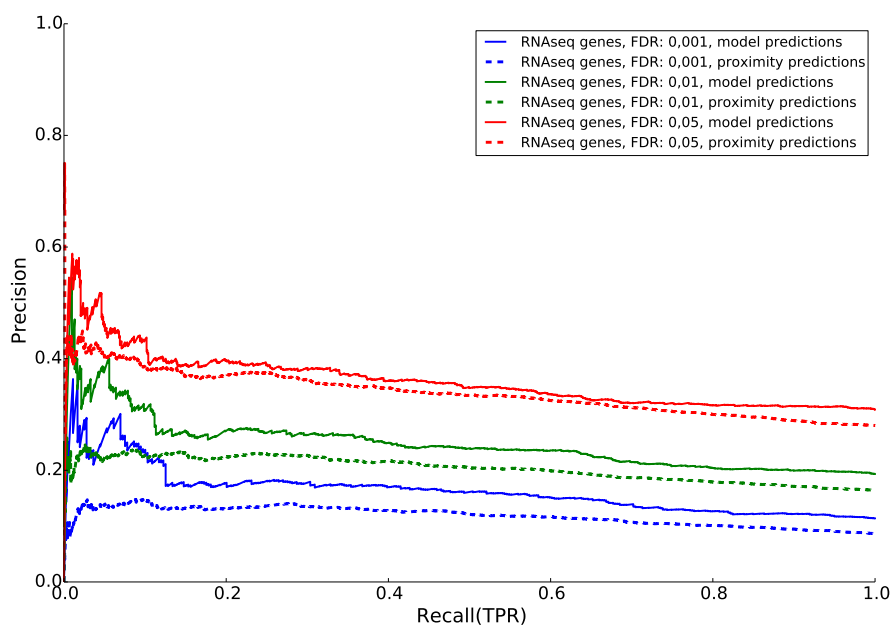


Figure 4.7: The PR curves assess the performance of Naive Bayes and proximity criterion on the ability to predict differentially expressed genes as detected by RNA-seq experiment and extracted at 3 different q-values (confidence levels) of 0.001, 0.01, 0.05. The predictor for the transcriptional activity of each gene of the NB predicted ER- $\alpha$  regulated genes is the probability that at least one of its NB assigned distal regulators indeed controls it. The Second predictor for the transcriptional activity of each gene of the genes with at least one ER- $\alpha$  binding within 40kB from their canonical TSS is the absolute value of the distance (proximity) between its canonical TSS and its closest ER- $\alpha$  binding.



shows the enrichment for the genes which were either targeted solely by distal enhancers (803, FDR = 0.25) or the ones which apart from being distally regulated also overlapped either with their gene body or 300bp-long promoter with intra-genic ER- $\alpha$  bindings (2343, FDR = 0.25). The intra-genic enhancers are assumed to control their host genes, and thus their targets are not inferred by the model. Although the assumption may hold for most genes, it is known that intronic enhancers can also target other not necessary their host genes [56, 132, 57]. The gene list was prepared using selected the 100 most significant GO classes with the lowest p-values and sorted by their proportions. The proportions are the ratios of the number of genes from our list which are associated with a given GO class to the total number of genes in that class. Table 4.2 shows that the two combined groups of genes are confidently predicted to be associated with MCF7 breast cancer, and show enrichments for several treatments of MCF7. Additionally, Table 4.3 shows the enrichments for biological processes such as E2-induced cell proliferation [58], regulation of cell differentiation [113], tissue morphology [125], and several developmental processes including vascularisation of MCF7 tumors [4], and gland (mammary) development [116]. The genes are enriched also for epithelial and motility classes [89].

The above list did not include genes without distal enhancer allocations. However, some of the genes still possessed at least one intra-genic or promoter-overlapping ER- $\alpha$  binding (5440, FDR = 0.25), which by our previous assumption are involved in regulation of their host genes. Combined with our previous list of genes that led to a total of 8586 ER- $\alpha$  targets. The gene ontologies (GO) for the genes can be seen in the table B.3. Most of the GO terms for biological functions of the extended gene list are unchanged, however, we additionally observed involvement of our genes in more specific Ras protein signal transduction and the enrichments in other 3 GTPase-related categories. All 216 of our genes in Ras category are found among genes in GTPase categories. Ras is a subfamily within a larger family of GTPases, which among other proteins includes R-Ras. In MCF7 R-Ras mediates interplay between estrogen and insulin signalling, and in consequence affects glucose metabolism and cell proliferation [133]. R-Ras was also shown to be involved in cell migration. Other Ras proteins (reviewed in [114]) regulate apoptosis, cell proliferation and differentiation. Mutations in their target genes are associated with cancers.

## Significant GO terms for diseases

Name	pValue	ER- $\alpha$ Genes	Genes in Annot.	Prop.
Noninfiltrating Intraductal Carcinoma	3.03E-11	85	281	30%
Fibroid Tumor	1.01E-10	85	287	30%
Ductal Carcinoma	9.50E-08	58	196	30%
Monosomy	2.64E-06	46	157	29%
Invasive breast carcinoma	1.66E-12	107	369	29%
Craniofacial Abnormalities	8.49E-07	53	184	29%

## Significant GO terms for drugs

Name	pValue	ER- $\alpha$ Genes	in Annot.	prop.
Retinoic acid; MCF7;	1.96E-22	77	178	43%
Thioridazine hydrochloride; MCF7;	1.11E-19	67	155	43%
Forskolin; MCF7;	2.55E-19	67	157	43%
Digitoxigenin; MCF7;	2.87E-19	69	165	42%
Niclosamide; MCF7;	1.84E-18	66	158	42%
Trichostatin A, Streptomyces sp.; MCF7;	9.51E-20	71	170	42%
Digoxin; MCF7;	1.97E-18	68	166	41%

Table 4.2: The table shows significantly enriched GO classes for drugs and diseases for the predicted targets of ER- $\alpha$  enhancers, FDR = 0.25.

## 4.5 Summary

In this chapter we showed how ChIP-seq time course data measuring TF and RNA polymerase occupancy changes after cellular stimulation can be used to predict enhancer-promoter/gene interactions within chromosomes. We developed a Bayesian classifier that combines the correlation of ChIP-seq time course data at enhancers and across gene bodies and genomic proximity as features. We applied our method to time course data from MCF7 breast cancer cells after stimulation with estradiol and we benchmark performance against publically available ChIA-PET data from this system. We showed that our method performs much better than association by proximity, identifying 33 times more interactions at a False Discovery Rate (FDR) of 0.25 than predictions based on proximity alone. Estrogen Receptor (ER- $\alpha$ ) and RNA polymerase (Pol II) ChIP-seq time course data were shown to be highly informative for predicting interactions. We also stratified our predicted interactions to those that lie within Topologically Associating Domains (TADs [26]) and those that span TADs, showing

## Significant GO terms for biological processes

Name	pValue	ER- $\alpha$	in Annot,	prop.
Gland morphogenesis	6.06E-14	62	154	40%
Epithelial cell differentiation	1.17E-12	60	155	39%
Morphogenesis of a branching epithelium	6.32E-16	83	222	37%
Placenta development	1.04E-11	61	166	37%
Stem cell development	4.93E-13	70	192	36%
Branching morphogenesis of an epithelial tube	2.21E-12	67	185	36%
Morphogenesis of a branching structure	2.01E-15	86	238	36%
Mesenchymal cell development	2.94E-12	67	186	36%
Mesenchymal cell differentiation	1.50E-12	71	200	36%
Ossification	7.43E-14	121	406	30%
Gland development	7.23E-17	151	508	30%
Epithelial cell differentiation	1.38E-19	193	670	29%
Morphogenesis of an epithelium	5.18E-16	171	613	28%
Tissue morphogenesis	3.54E-19	212	766	28%
Vasculature development	1.02E-15	184	680	27%
Blood vessel morphogenesis	7.33E-13	149	552	27%
Regulation of cell migration	9.80E-17	200	743	27%
Blood vessel development	1.38E-14	175	653	27%
Response to growth factor	5.39E-14	177	672	26%
Response to steroid hormone	3.95E-12	152	578	26%
Epithelium development	3.43E-26	343	1307	26%
Circulatory system development	4.51E-21	278	1062	26%
Cardiovascular system development	4.51E-21	278	1062	26%
Cellular response to growth factor stimulus	2.06E-12	166	643	26%
Cellular response to endogenous stimulus	3.72E-17	307	1280	24%
Response to hormone	6.08E-14	248	1035	24%
Regulation of cell development	1.33E-12	234	989	24%
Regulation of cell differentiation	2.03E-19	396	1706	23%
Regulation of cell proliferation	1.83E-14	369	1674	22%
Regulation of cell death	9.05E-13	359	1663	22%
Programmed cell death	2.72E-12	410	1963	21%

Table 4.3: The table shows significantly enriched GO classes for biological processes for the predicted targets of ER- $\alpha$  enhancers, FDR = 0.25.

that our classifier can make useful predictions in both categories. Finally, we used our predictions to provide a highly confident list of directly ER-regulated target genes in this system and validated it against GRO-seq data. The validation showed that our predicted targets are much more likely to show early nascent transcription than predictions based on genomic ER- $\alpha$  binding proximity alone. Gene Ontology showed that our ER- $\alpha$ -bound enhancers and their predicted targets are involved in many biological processes associated with breast cancer. Our models can thus offer biologically meaningful insight into the transcriptional regulation involving ER- $\alpha$ .

The enhancer-centric model is simplified in a sense that it does not take into account the similarities of all enhancers which interact at a common target gene, as part of a larger gene regulatory complex. In the next section we propose another model which addresses that drawback.

# Chapter 5

## Unsupervised learning

We introduce a generative unsupervised model which combines distance and shape of time course of multiple TF ChIP-seq datasets to probabilistically assign regulatory ER- $\alpha$ -bound enhancers to their target genes. At each assignment, in order to model the complex multi-enhancer aspect of protein-mediated transcriptional regulation of genes, the model assesses similarity between the time series of not only a single enhancer and its putative target gene but also of all the enhancers which are part of the inferred regulatory complex of the candidate gene.

The inference of the special form of the latent variable mixture model presented here requires us to apply approximate Bayesian techniques. Here we employ a Gibbs sampler to sample from the posterior distribution of our model. The main variable of interest is an assignment vector which is used to calculate the relative contact frequencies of enhancers and genes.

We validate the model with ChIA-PET Pol II/ER- $\alpha$  links and compare the performance of the model against the classification approach in the previous chapter. We investigate whether the integration of multiple datasets improves the performance of the model, we test alternative parametrisations of the model which corresponds to reweighing the importance of gene and enhancer time series in the assignment. We also investigate the effect of time course data from the enhancers with no ChIA-PET-confirmed links on the inference of the model.

### 5.1 Gene-centric Latent Variable Allocation model

Suppose that an enhancer  $j = 1, \dots, J$  regulates a gene  $k = 1, \dots, K$  at a number of time points, and that their contact is mediated by a protein. In the previous chapter we have

observed that the time varying ChIP-seq signals at the interacting loci are correlated to a higher extent than their non-interacting counterparts for multiple datasets. We can therefore expect that the time series at the enhancer  $j$  i.e.  $\mathbf{X}_j = (x_{j,1}, \dots, x_{j,D})$  and the gene  $k$  i.e.  $\mathbf{Y}_k = (y_{k,1}, \dots, y_{k,D})$  would exhibit similarity. Here we would like to exploit the similarities to find the interacting pairs and propose an alternative generative gene-centric model of multi-enhancer-gene contacts. According to our model an enhancer time series  $\mathbf{X}_j$  is generated from one of  $K$  underlying gene-based patterns  $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ , where the prior mean  $\boldsymbol{\mu}_k$  of each profile is equal to the time series of gene  $\mathbf{Y}_k$ . As in the previous chapter, the gene-centric model is simplified in a sense that it does not take into account the promoter-promoter interactions which are known to play an important part in the higher order gene regulatory complexes. However, we do model the potential interaction of multiple enhancers with one target gene. The main variable of interest is the indicator variable  $\mathbf{Z} = (z_1, \dots, z_J)$ , which indicates the source of an enhancers' time series and specifies the clustering of enhancers at their target genes. In our model the samples  $\mathbf{X}_{\{j:z_j=k\}}$  which shares the value of the indicator variable are generated from one of  $K$  multivariate Gaussians  $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k I)$ , where  $I$  is the  $D \times D$  identity matrix.

### 5.1.1 Generative Latent Variable model

Here, we review a standard latent variable mixture of Gaussians which generates vectors of data  $\mathbf{X}$  with independent dimensions (for a more general model refer to [39]). The graphical representation of the model can be seen in Fig. 5.1(a). The assumption of independent entries reduces the number of parameters of the model and therefore reduces the computational cost of its inference. The model is hierarchical and consists of multiple layers. The first layer,

$$p(\mathbf{X}_j | z_j = k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2 I) = \mathcal{N}(\mathbf{X}_j | \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2 I) \quad (5.1)$$

states that the vectors of observations  $\mathbf{X}$  are generated from one of conditionally independent clusters  $K$ , characterised by  $\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1 I, \dots, \boldsymbol{\mu}_K, \boldsymbol{\sigma}_K I$ . Given data, we would usually be interested in finding a set of patterns which generated our observations, and the origin of each observation. The origin of observations  $\mathbf{X}$  is specified by the indicator variables  $z_j$  (one per  $\mathbf{X}_j$ ) which follows a categorical distribution with probabilities

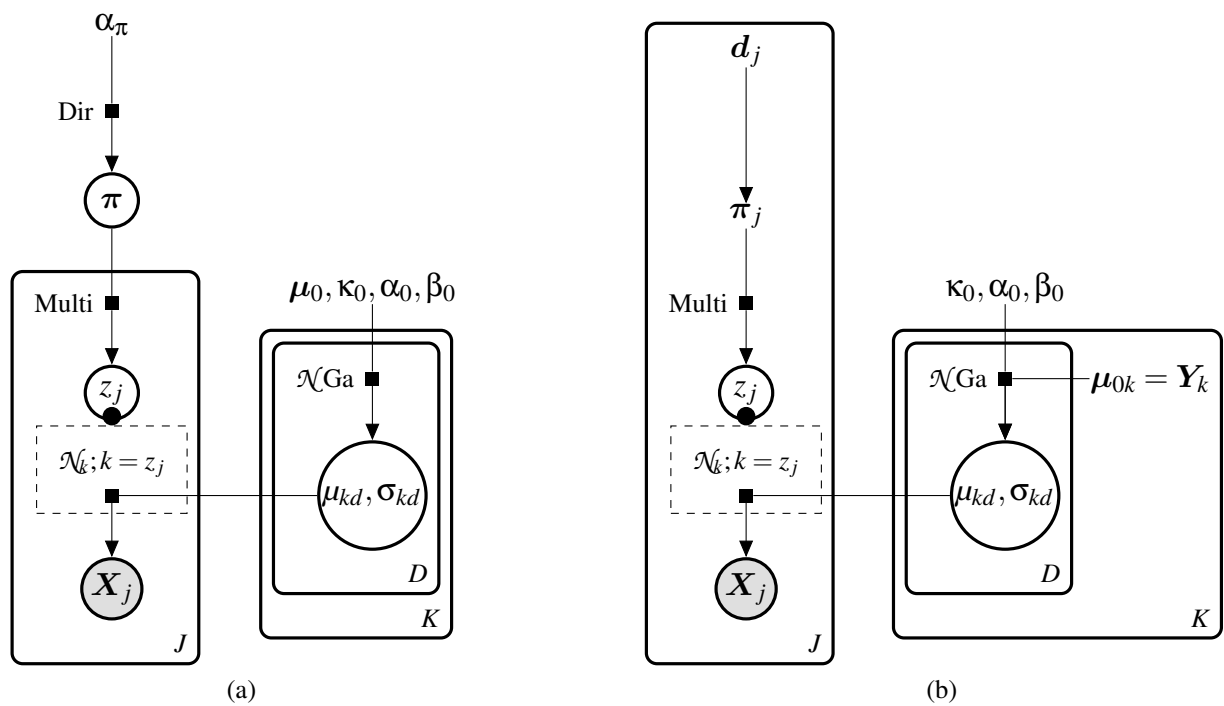


Figure 5.1: Directed factor graph representation of: a) standard latent Dirichlet allocation of mixture of Gaussians generating data with independent dimensions and b) our latent variable gene-centric mixture of Gaussians with the separation-based prior.

$\pi$  (mixture components), that is

$$p(z_j|\pi_1, \dots, \pi_K) = \text{Cat}(z_j|\pi) \quad (5.2)$$

The next layer is the one defining the means and standard deviations. In the Bayesian setting the means and standard deviations are also random variables which follow,

$$p(\mu_{k,d}|\mu_{0,k,d}, \sigma_{k,d}^2, \kappa_0) = \mathcal{N}(\mu_{k,d}|\mu_{0,k,d}, \sigma_{k,d}^2/\kappa_0) \quad (5.3)$$

$$p(\sigma_{k,d}^{-2}|\alpha_0, \beta_0) = \text{Ga}(\sigma_{k,d}^{-2}|\alpha_0, \beta_0) \quad (5.4)$$

where the sampling is repeated  $D$  times. The gamma function is a conjugate prior to the normal distribution and the conjugacy requires that the  $\mu_{k,d}$ , depends on  $\sigma_{k,d}^2$ . The pair of functions which generates the parameters is jointly referred to as normal-gamma ( $\mathcal{NGa}$ ) prior. Here the parameter  $\kappa_0$  reflects how strongly we believe in a prior mean  $\mu_0$  of a cluster. The parameters  $\alpha_0$  and  $\beta_0$  regulate the shape of the gamma function and thus control the value of the variances.

### 5.1.2 Latent Variable Allocation with a Dirichlet prior

The standard prior for mixture components is a uniform Dirichlet,

$$p(\pi_1, \dots, \pi_K) = \text{Dir}(\alpha/K, \dots, \alpha/K) \quad (5.5)$$

where parameter  $\pi_k$  controls how strongly we believe that a cluster  $k$  is *a priori* responsible for generating the observed data. Given a sample of mixture components  $\pi$ , occupation numbers  $n_k = |\{j : z_j = k\}|$ , which are equal to the number of indicators  $z_j$  equal to  $k$ , follow a multinomial distribution

$$p(n_1, \dots, n_K) = \text{multi}(n_1, \dots, n_K|N, \pi) \quad (5.6)$$

and the joint distribution of  $z_j$  is

$$p(z_1, \dots, z_J|\pi) = \prod_{k=1}^K \pi_k^{n_k}. \quad (5.7)$$



In order to remove the need to sample the parameters, the component variables  $\pi$  may be integrated out. This results in the collapsed model

$$p(z_1, \dots, z_J | \alpha) = \int p(z_1, \dots, z_J | \pi) d\pi = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \prod_{k=1}^K \frac{\Gamma(n_k + \alpha/K)}{\Gamma(\alpha/K)}. \quad (5.8)$$

Once the variables are collapsed, the indicator variables become dependent on each other. The conditional prior is obtained after fixing all but one  $z_j$ , the remaining  $\mathbf{Z}_{-j} = z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_J$  are assumed to be known. That results in:

$$p(z_j = k | \mathbf{Z}_{-j}, \alpha) = \frac{n_{-j,k} + \alpha/K}{N - 1 + \alpha}. \quad (5.9)$$

Note that the form of the distribution indicates that the model will favour allocations of observed  $\mathbf{X}_j$  to the clusters with a higher number of assignments.

### 5.1.3 Latent Variable Allocation with a separation-based prior

To address the problem of assigning regulatory enhancers to their target genes we changed the prior of the model to one which we believe is better suited to tackle the task of identifying enhancer-gene links. We will refer to the altered model as LVA. Specifically, we make use of a non-uniform distribution of separations between enhancers and their target genes. Fig. 5.1(b) illustrates the form of the generative model.

Firstly, since each cluster  $k$  corresponds to a unique gene  $G_k$ , we encode each  $\mathbf{Y}_k$  into the model by setting the prior mean  $\mu_{0_k}$  of each cluster to  $\mathbf{Y}_k$ . Thus, the equation 5.3 becomes:

$$p(\mu_k | \mathbf{Y}_k, \sigma_k^2 I, \kappa_0) = \mathcal{N}(\mu_k | \mu_{0_k} = \mathbf{Y}_k, \sigma_k^2 I / \kappa_0) \quad (5.10)$$

Secondly, we estimate the average frequencies of the contacts at a number of regions  $b_m$  away from an enhancer. We take the ratio of the average density of ChIA-PET confirmed ER- $\alpha$  targeted genes to the average density of genes at that location, summed up across all gene distributions of ER- $\alpha$  enhancers, i.e:

$$freq|b_m \approx \frac{\rho(\text{target genes}|b_m)}{\rho(\text{genes}|b_m)} \quad (5.11)$$

Each enhancer  $j$  possesses its own unique genomic location and a set of  $K$  distances  $d_j$  to its putative target genes. The distances fall into bins  $\{b_m\}$  where  $m \in \{1, \dots, M\}$ , so that if  $d_{j,k} \in b_m$  then  $\pi_{j,k} | d_{j,k} \propto freq|b_m$ . Normalising the frequencies by their sum

leads to,

$$\pi_{j,k}|d_{j,1},\dots,d_{j,K} = \frac{\pi_{j,k}|d_{j,k}}{\sum_K \pi_{j,k}|d_{j,k}} \quad (5.12)$$

so that the  $\pi_j$  sum up to one, and take into account the genomic distribution of genes around an enhancer  $j$ . The vector of the resulting  $\pi_j|d_j$  serves as mixture components for the enhancer  $j$ . Thus the prior probability that the enhancer  $j$  targets a gene  $k$  located at  $d_{j,k}$  is,

$$p(z_j = k|\pi_j, d_j) = \text{Cat}(z_j = k|\pi_j, d_j) \quad (5.13)$$

#### 5.1.4 Estimation of the separation-based prior with ChIA-PET data

In order to estimate the relative contact frequencies of the chromatin between ER- $\alpha$  enhancers and ENSEMBL-annotated genes at a distance  $d$  (5.11), we divide the local density of positive contacts by the local density of genes. The ratio can be interpreted as an average number of successes (targeted genes, i.e positives from Chapter 4) at that location to the total average number of trials (genes, i.e positives and negatives from Chapter 4) at that location.

The nominator of the equation is equal to the density of distance of positive enhancer-gene pairs  $P(d_{j,k}|I_{j,k} = 1)$  from Chapter 4, except that we estimate it on all chromosomes. The denominator however, as opposed to the distribution of negative pairs  $P(d_{j,k}|I_{j,k} = 0)$ , is a KDE of a total number of positive and negative pairs. Due to the large sample size, the procedure for estimating the distribution, mirrors that of non-interacting negative pairs (Section 4.2.4). The estimation of those two distributions with KDE ensures that the frequencies for potential unobserved instances of distance are non-zero.

As preliminary test, we checked whether the samples from our prior could recover the distribution of contacts as established from the ChIA-PET-confirmed links. For that, we sampled one candidate indicator variable  $z_j$  per each interacting enhancer according to our prior 5.13 and estimated the distribution of frequencies of the corresponding distances via KDE. Fig. 5.2 shows that the approach correctly approximates the frequencies of the data-confirmed contacts.

#### 5.1.5 Inference with a Gibbs sampler

Due to the requirement to know the marginal distribution (denominator of the Bayes equation) of the latent variable mixture of Gaussians, the exact posterior distribution of

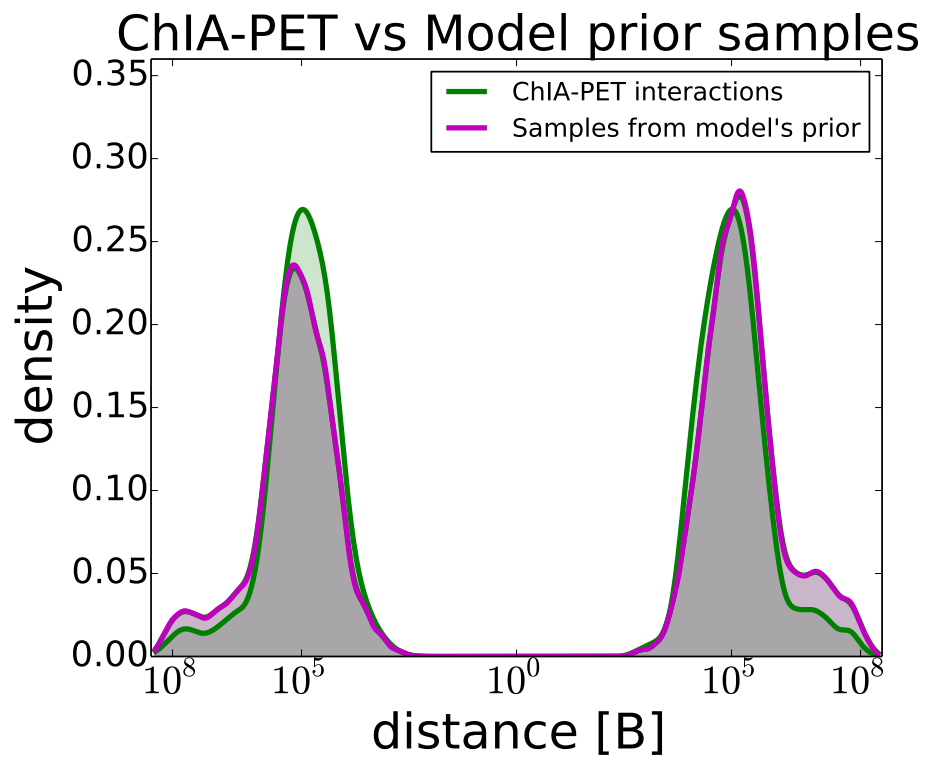


Figure 5.2: Figure shows the distributions of separations between ChIA-PET-detected enhancer-gene interactions (green) and gene allocations for the ChIA-PET-confirmed interacting enhancers sampled from LVA's prior (purple).

the model is intractable. One solution is to use approximate inference method such as Gibbs sampling which approximate the posterior distribution by sequentially drawing samples from conditional distributions of the model. The sampling scheme for this model can be seen in Algorithm 1. The conditional distributions enable the algorithm to draw subsets of parameters of the target distribution, conditional on the remaining parameters of the model. Although Gibbs conditions each update of a variable on the

---

**Algorithm 1** Gibbs sampler for the latent variable mixture of Gaussians
 

---

Input:  $X$   
 Initialize:  $\mu, \sigma, Z$   
**for** each sample  $s \in \{1, 2, \dots, S\}$  **do**

**for** each sample  $j \in \{1, 2, \dots, J\}$  **do**  
 Draw  $z_j | \mu, \sigma, Z_{-j} \sim$  Eq: 5.14 or 5.15

**for** each sample  $k \in \{1, 2, \dots, K\}$  **do**

**for** each sample  $d \in \{1, 2, \dots, D\}$  **do**  
 Draw  $\mu_{kd}, \sigma_{kd} | \mu_{-k}, \sigma_{-k}, Z, X \sim$  Eq: 5.16

---

remaining variables, many variables of the model are conditionally independent. This greatly reduces the complexity and the form of the posterior distributions. Each iteration involves reallocating each enhancer  $X_j$  to a new gene cluster, that is re-sampling the full  $Z$ . In mixture of Gaussians with Dirichlet prior the updates follow:

$$\begin{aligned}
 & p(z_j = k | Z_{-j}, \mu_k, \sigma_k, \alpha) \\
 & \propto p(z_j = k | Z_{-j}, \alpha) p(X_j | \mu_k, \sigma_k I) \\
 & \propto \frac{n_{-j,k} + \alpha/K}{n - 1 + \alpha} \mathcal{N}(X_j | \mu_k, \sigma_k I)
 \end{aligned} \tag{5.14}$$

In our model with the distance-based prior,

$$\begin{aligned}
 & p(z_j = k | \mu_j = Y_k, \sigma_k, \pi_j) \\
 & \propto p(z_j = k | \pi_j) p(X_j | \mu_k = Y_k, \sigma_k I) \\
 & \propto \pi_j \mathcal{N}(X_j | \mu_k = Y_k, \sigma_k I)
 \end{aligned} \tag{5.15}$$

where  $\pi_j$  depends on the distribution of distances from an enhancer  $j$  to its potential target genes and can no longer be integrated out to obtain a collapsed sampler. Once the vector  $Z$  is updated, the parameters controlling the shape of the gene clusters have

to be re-determined. The updates follow the equation:

$$p(\mu_{j,d}, \sigma_{j,d}^{-2}) = \mathcal{N}Ga(\mu_{j,d}, \sigma_{j,d}^{-2} | \mu'_{0,j,d}, \kappa'_{0,j}, \alpha'_{0,j}, \beta'_{0,j,d}). \quad (5.16)$$

where,

$$\begin{aligned} \kappa'_{0,j} &= \kappa_0 + n_j \\ \mu'_{0,j,d} &= \frac{\kappa_0 \mu_{0,j,d} + n_j \bar{x}_{j,d}}{\kappa_j} \\ \alpha'_{0,j} &= \alpha_0 + \frac{n_j}{2} \\ \beta'_{0,j,d} &= \beta_0 + \frac{1}{2} \sum_{\{k|z_k=j\}} (x_{k,d} - \bar{x}_{j,d})^2 + \frac{\kappa_0 n_j (\bar{x}_{j,d} - \mu_{0,j,d})^2}{2\kappa_j}. \end{aligned} \quad (5.17)$$

Notice that the updated pairs  $(\sigma_j, \mu_j)$  take into account all enhancer time series clustered at a gene  $j$  and the time series at the gene to a degree specified by the parameter  $\kappa_0$ . The  $\kappa_0$  weights the importance of the gene time series against that of the associated enhancer time series.

As the number of samples drawn from our model increases, the samples approximate the underlying joint multivariate posterior distribution of all parameters. It is important to discard a number of initial samples as part of so called burn-in period, since they may be erroneous.

### 5.1.6 Estimation of the frequency of enhancer-gene contacts

The sampled indicator vectors  $\mathbf{Z}^B, \dots, \mathbf{Z}^S$ , where the first  $B$  samples are discarded as part of the burn-in period, can be used to estimate the frequency of contacts between an enhancer  $j$  and a gene  $k$  such that,

$$P(z_j, k) = \frac{\sum_{s=B}^S \delta(z_{j,s}, k)}{S - B} \quad (5.18)$$

where  $\delta$  is the Kronecker delta function which is equal to 1 if  $z_{j,s} = k$  or 0 otherwise. The expression is equal to the average number of times an enhancer  $j$  targets a gene  $k$  across all  $S - B$  samples .

## 5.2 Results

We demonstrate our method using ChIP-seq time course data collected from MCF7 breast cancer cell-line stimulated by estrogen (E2) as discussed in Section 3.1.3. The genome-wide occupancy of ER- $\alpha$  along with RNA polymerase (Pol II) and two histone marks (H3K4me3 and H2AZ) were measured via ChIP-seq at eight consecutive times after exposure of cells to estradiol. ChIA-PET data are available in this system for testing our method’s performance [34, 63, 64]. The estimation of the prior of the model is described in Section 5.1.4. Here, we measure the performance of the model and compare it to the model in chapter 4.

### 5.2.1 Model inference - technical details

Similarly as prior to the AP clustering in Section 3.2, the normalised time course data of ER- $\alpha$  along with RNA polymerase (Pol II) and two histone marks (H3K4me3 and H2AZ) at enhancers and 300bp-upstream-extended-genes were standardized to z-scores to bring all time series onto the same scale. In order to refine signal to noise ratio of our data, we removed genes and distal enhancers which possessed a total of less than 30 tags across all time points of their time series in each of our time course datasets.

We set the prior parameters of the model to  $\kappa_0 = 1$ ,  $\alpha_0 = 2$ ,  $\beta_0 = 2$ . This choice of  $\kappa_0$  set the importance of gene time series to be equal to that of each enhancer. The choice of  $\alpha_0$  and  $\beta_0$ , ensured that the prior variance of each time point is centred around 1. In order to benchmark performance of the model with  $\kappa_0 = 1$  against the one with an alternative parametrization we repeated the sampling with  $\kappa_0 = 3$ . The initial values for vectors  $\mathbf{Z}$  and  $\boldsymbol{\sigma}$  were random.

We aimed to compare the performances of the model with five selected combinations of the input data (the same combinations as in Section 4.3). For each combination we ran three independent chains to assess their convergence with 62,000, 62,000, 240,000, 240,000, and 600,000 samples respectively. In each case we discarded half of the samples as burn-in. Fig. 5.3, shows that most of the values of the multi-chain-based  $\hat{R}$  statistics are relatively low and lie in the interval between 1. and 1.1, which suggest that the chains have reached equilibrium.

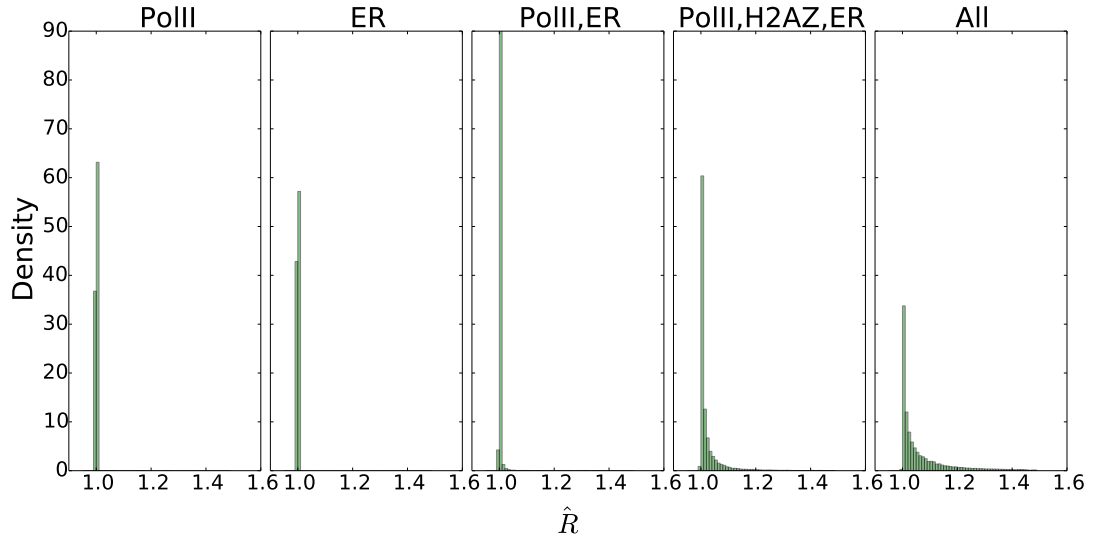


Figure 5.3: Gelman  $\hat{R}$  statistics for the samples of  $\mathbf{Z}$  generated by the Gibbs Sampler of the LVA model with  $\kappa_0 = 1$ ,  $\alpha_0 = 2$ ,  $\beta_0 = 2$ .

### 5.2.2 Performance

We present performance of the model validated on the ChIA-PET-detected promoter-enhancer links from all even chromosomes (as described in Section 3.3). As opposed to the NB in Chapter 4, the model is inferred using the time course data from all enhancers, not only the ones with the ChIA-PET data. We validated the predictions of the model using PR curves (precisions against the true positive rate of 10%, 20%, and 30%) of the posterior probabilities of our predictions, using cut-offs of 0.92, 0.56, 0.18. The low value of the last cut-off suggests that the events in the group are relatively infrequent and therefore harder to detect. Additionally, we compared performance of LVA to that of NB model. Fig. 5.4 shows that the evidence from the ER- $\alpha$  dataset improves the performance of the model. Nevertheless, the model is neither able to take advantage of the data to the degree observed in the Naive Bayes model, nor gains enhanced performance with an inclusion of additional datasets. In fact, we observe that combining datasets decreases its performance. Fig. 5.7 shows that the separation-based prior of our LVA model performs similarly to the distance-only NB, thus its lower performance does not stem from its ability to tackle the evidence from separation. Additionally, Fig. 5.4 shows that increasing the value of  $\kappa_0$  and thus elevating the importance of gene time series does not lead any significant performance changes, which suggests a more vital role of individual enhancers and their time series in the discovery of their target genes.

## Performance for intra-domain and inter-domain links

Fig. 5.5 shows the performance of the model on intra-domain interactions. In contrast to NB, where integrating Pol II and ER- $\alpha$  led to the best performance, the best performance is achieved when the model is inferred solely from the ER- $\alpha$  time course data. Other datasets inclusions typically decrease the accuracy of the model. As for the NB model, the data leads to improved predictions over using separation alone. Fig. 5.6 shows that LVA performs also worse than NB on links passing TAD boundaries. The data aided LVA achieves worse accuracy than only using a distance-based prior in this case. Both figures show that changing the parameter  $\kappa_0$  does not lead to significant changes in the performance of the model for any of the two groups.

### 5.2.3 Testing the effect of ER- $\alpha$ enhancers with unknown status

We investigated whether the relatively poor performance of the LVA model relative to the NB method can be attributed to the incorporation of time course data from the enhancers with no ChIA-PET-confirmed links. In the construction of the NB model we only used data from enhancers with ChIA-PET evidence of interaction.

Fig. 5.8 shows that inclusion of the data from enhancers without ChIA-PET evidence lowers the ability of LVA to predict interactions from the odd chromosomes. Fig. C.2 show that the observation is true regardless of whether we measure the accuracy for inter- or intra-domain subsets of predictions or global-wise, however the decrease is most clearly observed for the intra-domain interactions. Interestingly, we show that for the within-domain predictions the performance of the model inferred from the enhancer-restricted ER- $\alpha$  time course data can be similar to that of the corresponding NB model. In the ER- $\alpha$ -inferred enhancer-restricted LVA model Pol II seems to either have a small or at least no negative effect on the accuracy of the predictions.

We draw similar conclusions when we test the performance for the interactions located on the even chromosomes in Fig. C.3. However, here the within-domain performance is lower than that achieved by NB.

## 5.3 Summary

In this chapter we developed a generative gene-centric model which combines observed time course ChIP-seq data over enhancers and gene bodies with the prior based on the genomic separation of interacting elements and local gene densities to infer



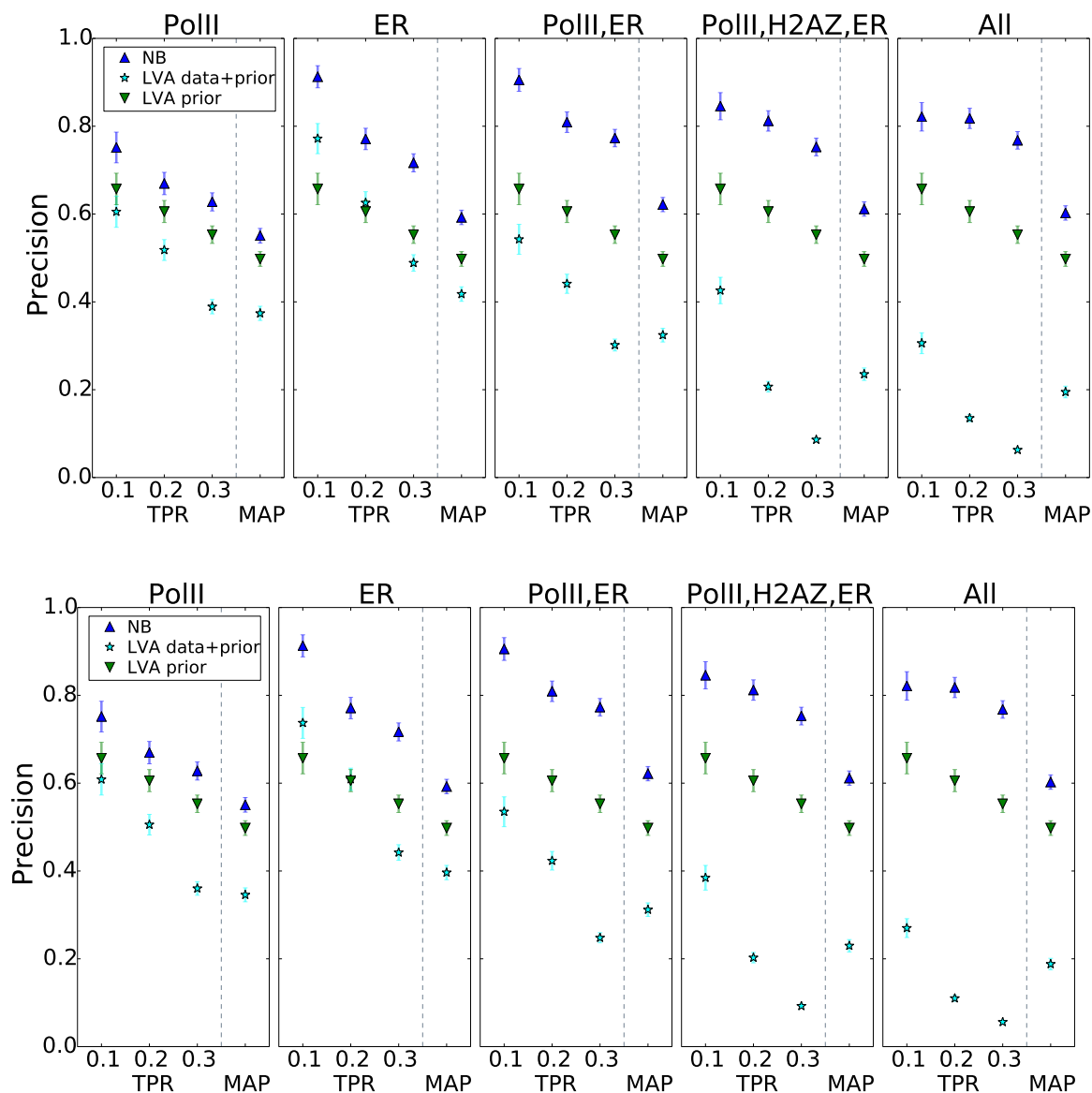


Figure 5.4: The comparison of the performance between the Naive Bayes algorithm and the Latent Variable Allocation models with different parametrisation of  $\kappa_0$  on all even chromosomes. The parametrisation of the model in the first row is  $\kappa_0 = 1$ ,  $\alpha_0 = 2$ ,  $\beta_0 = 2$ , the second  $\kappa_0 = 3$ ,  $\alpha_0 = 2$ ,  $\beta_0 = 2$ . The Precision-TPR curves show the accuracy for the predictions with the highest 10%, 20%, 30% posterior probabilities.

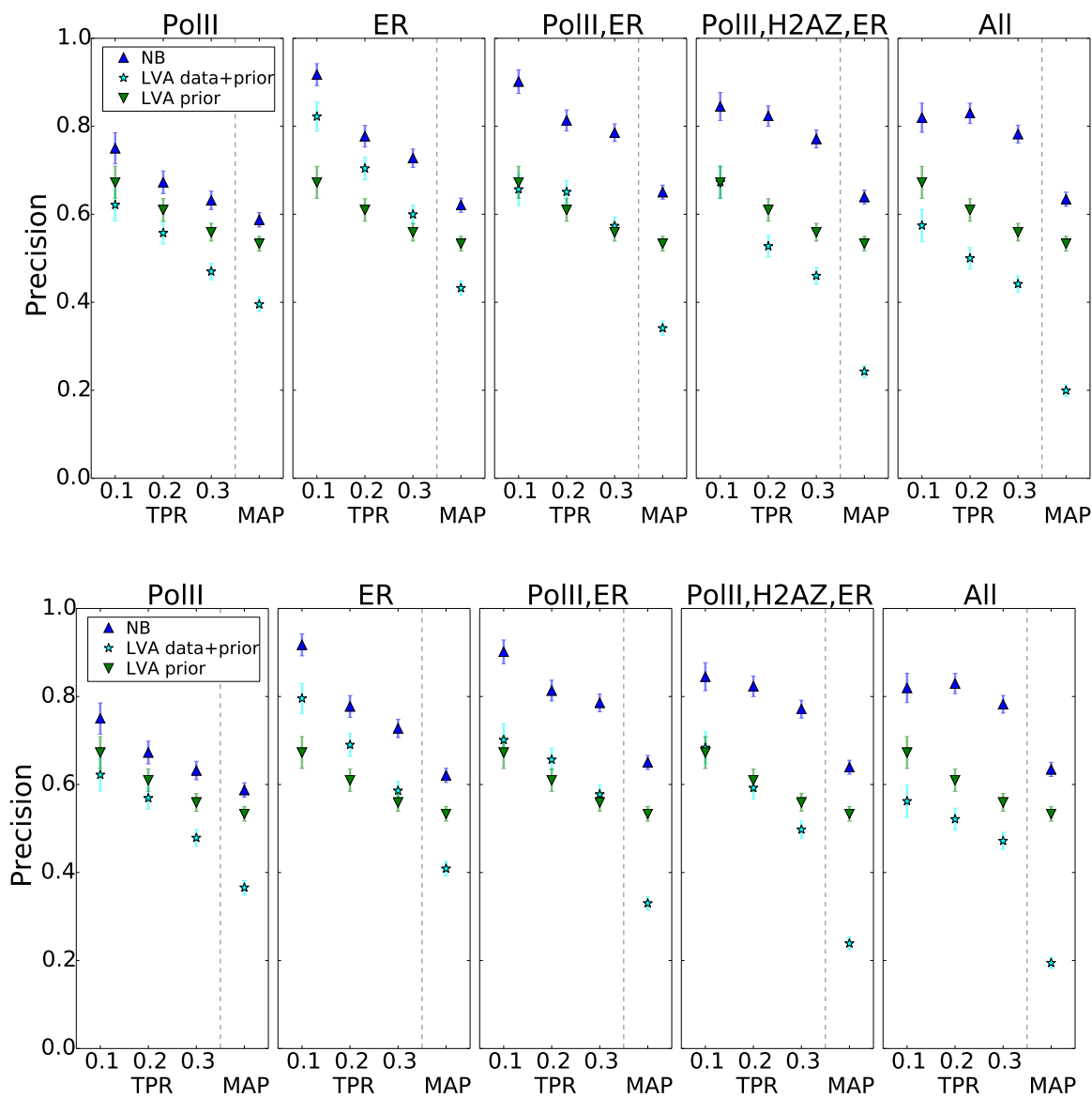


Figure 5.5: Figure shows the performance of the model, as in Fig. 5.4 but only considering interactions within TADs.

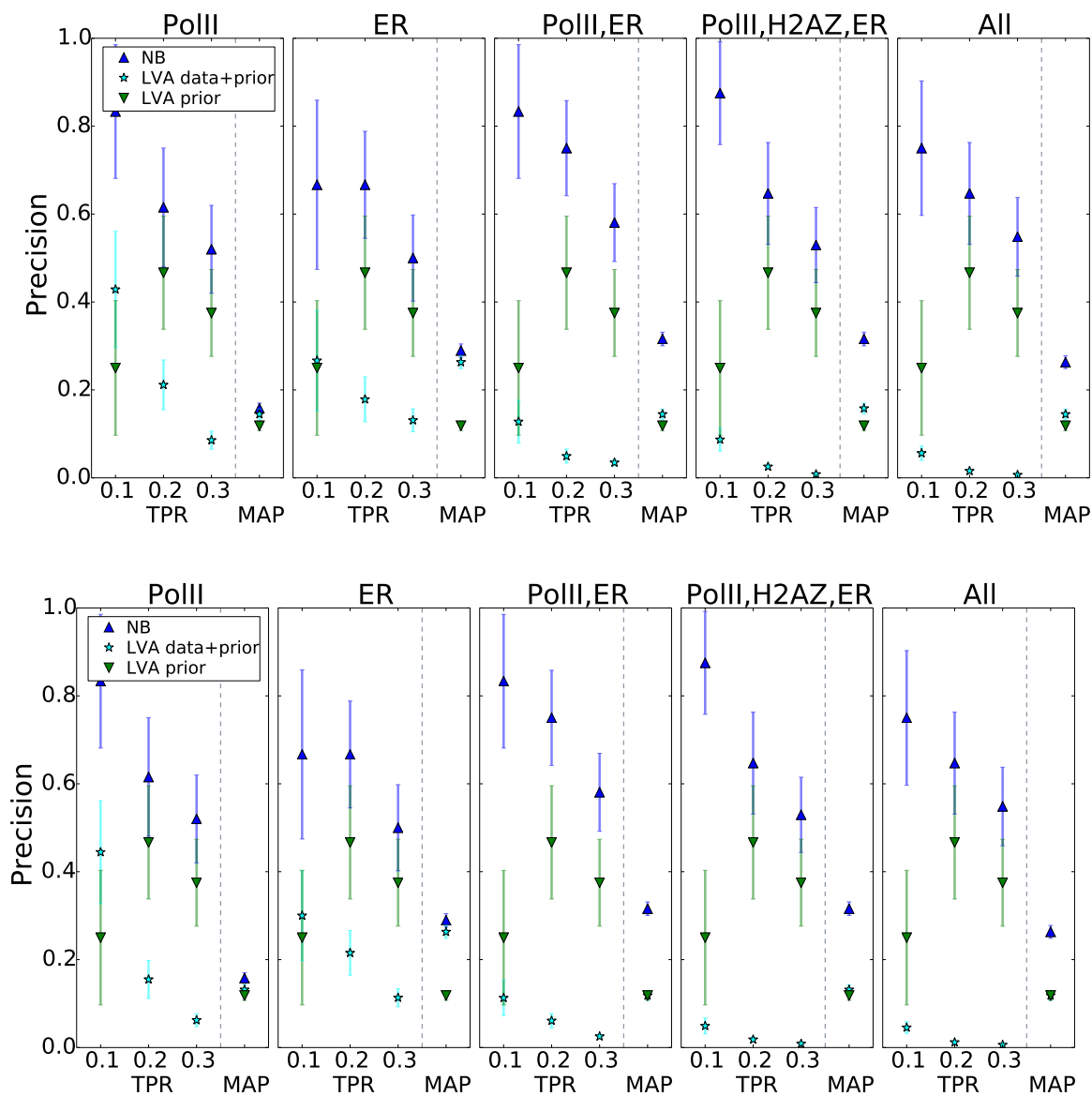


Figure 5.6: Figure shows the performance of the models, as in Fig. 5.4 but only considering interactions spanning two TADs.

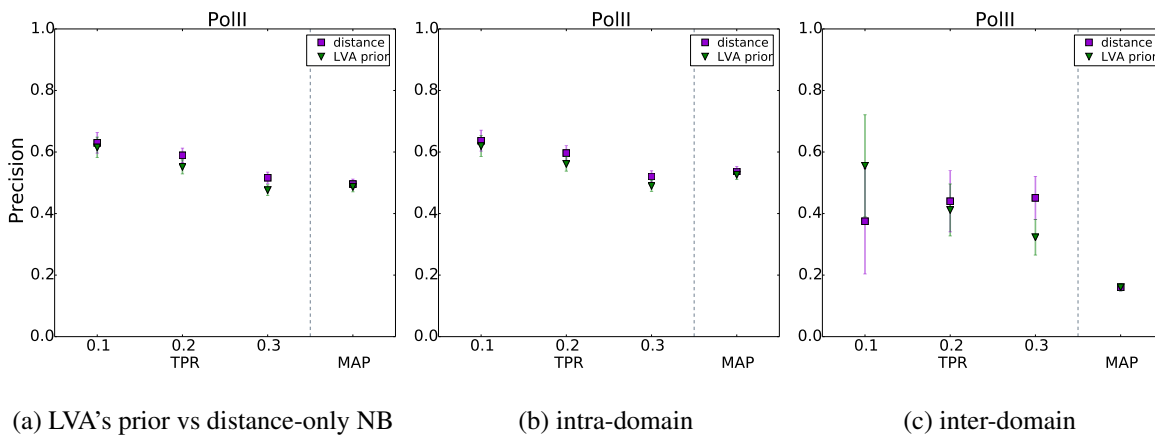


Figure 5.7: Comparison of performance between the separation-based prior of Latent Variable Allocation model and distance-only Naive Bayes model.

protein-mediated regulatory complexes involving individual genes and their networks of multiple distant regulatory enhancers.

The inference of the generative model presented here required us to employ approximate Bayesian inference techniques. Here we chose to use a Gibbs sampler which produces samples from the posterior distribution of our model. The convergence of our MCMC chains was established via multi-chain  $\hat{R}$  statistics. Due to efficient multi-core implementation, the sampler achieves a speed of approximately one second per sample when applied to the human genome.

Here, we showed that LVA inferred from the data of all enhancers performs significantly worse than the NB model from Chapter 4. However in order to allow a more comparative benchmark, since the NB model was trained only from the time course data of the enhancers with ChIA-PET-confirmed links, we re-inferred LVA using the same restricted set of data and re-assessed it against NB and the previous LVA. We observed a significant increase in the accuracy between LVA inferred from the full and the restricted dataset, however NB was still considerably better. We conclude that using collective similarities of multiple time series of putative regulatory enhancers of a gene complex to predict individual enhancer-gene contacts may lower confidence of at least some of the true contacts.

The higher performance of the data-restricted LVA may be associated with a potential bias of our ChIA-PET-derived prior or be a result of different underlying characteristic of the enhancers with no ChIA-PET evidence. In the next chapter we will discuss means to clarify the findings (Section 6.4), consider potential solutions and propose

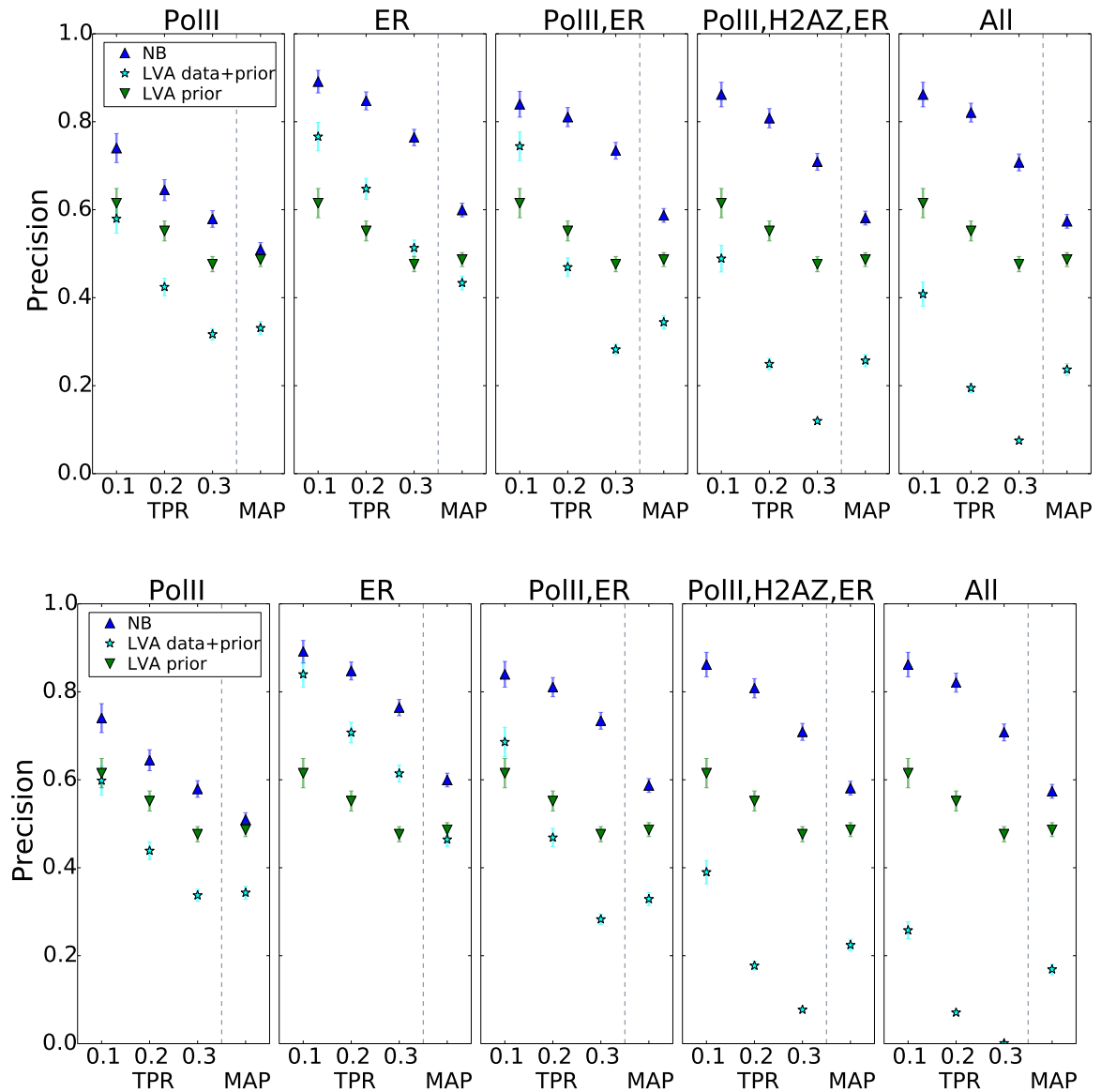


Figure 5.8: Figure shows the comparison of the performance between NB and two LVA models on discovery of ChIA-PET-detected links from odd even chromosomes. LVA in the first row was inferred from time course data of all enhancers. LVA in the second row was inferred from the time course data of only those enhancers with ChIA-PET-confirmed links.

further development of the study.

# Chapter 6

## Conclusions

### 6.1 Achieved Results

In this study we presented two methods which can utilise ChIP-Seq time course data measuring TF and RNA polymerase occupancy changes and histone modifications after cellular stimulation to predict enhancer-promoter/gene interactions within chromosomes. The methods combine observed time course ChIP-Seq data over enhancers and gene bodies with the genomic separation of interacting elements to predict putative cell-type-specific and protein-mediated associations of enhancer-promoter/gene pairs. We applied our method to time course data from MCF7 breast cancer cells after stimulation with estradiol and we benchmark performance against publically available ChIA-PET data from this system.

We show that our methods performs much better than association by proximity and estrogen receptor (ER- $\alpha$ ) ChIP-Seq time course data is shown to be highly informative for predicting interactions. We also stratify our predicted interactions to those that lie within Topologically Associating Domains (TADs [26]) and those that span TADs, showing that our classifiers can make useful predictions in both categories. The method thus is a complementary approach to protein-specific chromosome conformation capture methods such as ChIA-PET.

Using complementary GRO-seq data from the same cell-line and context we show that our supervised approach can accurately predict distally regulated, early responsive genes under stimulation with estrogen. Gene Ontology analysis showed our ER- $\alpha$ -bound enhancers and their predicted targets are involved in many biological processes associated with breast cancer. Our models can thus offer biologically meaningful insight into the transcriptional regulation involving ER- $\alpha$ .

## 6.2 Limitations of the models

Our models possess a number of limitations, the majority of which are shared between the first and the second model. Both methods are unable to model the interactions involving promoter-promoter contacts. If available, this would likely enhance the model's predictions and incorporate more realism to the model [64]. In addition, the NB model assumes that each enhancer interacts with only one gene, which is unlikely to be realistic for all contacts. The models are also not optimised to take the advantage of the time course data. That is, neither the choice of Pearson correlation as a measure of association nor the diagonal covariance matrices in the latent variable model allows capture of covariance structures in the data. The assumption of a conditionally independent structure in the likelihood was motivated by good performance of such models in practice. It is suggested in [130] that making the independence assumption in modelling of temporal data can be often more justified if one uses differences between data of adjacent time points rather than time point data itself, and it shows examples of models where it can naturally hold. However, their findings were drawn under the assumption of stationary character of time series data which is unlikely to be met in the experiment which produced our data. Another issue of practical relevance is the uneven logarithmic spacing of our time points. The spacing has proved to be problematic when we attempted to fit models based on Gaussian Processes to our data. In the next section we will discuss some other sources of limitations resulting from the quality of the available experimental data and possible future solutions.

## 6.3 Limitations of the data

The major factor which entailed the simplification of assumptions was the availability and quality of public experimental data from the cellular context and cell-line, as well as limitations of the ChIA-PET method. Firstly, the ChIA-PET data which was used for the design and validation of our models is inherently very conservative and possesses a very low coverage of chromatin associations. Secondly, the contacts which were established by the method correspond to a single time point, thus the resulting data is unlikely to reveal the whole spectrum of chromatin contacts. Thirdly, PETs separated by less than 10Kb are often the result of self-ligation, and thus often unreliable. The FDA of the method is likely to be associated with distance between PETs. In consequence, the majority of our distal MACS-detected ER- $\alpha$  binding sites lack



experimentally confirmed enhancer links, and thus our training and validation datasets are very sparse. The degree of sparsity is even higher for the promoter-promoter interactions, where the amount of confirmed interactions is around one percent of that of enhancer-promoter links. Those factors are prohibitive for building a more complex model and especially limit our ability to model multi-loci complexes.

An alternative in the form of Hi-C technique, provides a genome-wide snapshot of proximity between loci of chromatin in a selected cell-line, however, the links are, in comparison with ChIA-PET, less likely to be functional and are in general not mediated by a specific protein of interest. That combined with relatively low resolution of publicly available Hi-C data for the MCF7 cell line [8], i.e. the resolution of 40kB, the highest which is available for this cell-type, is too low to distinguish individual enhancers and promoters/genes which take part in the interactions.

## 6.4 Comparison between NB and LVA and study of bias

Comparison of performance of LVA inferred from the data of ChIA-PET-confirmed enhancers which possessed the ability to jointly assess the similarities between the time series of multiple potentially inter-linked regulatory enhancers and their individual putative target genes with that of the supervised approach which lacked that ability showed that predictions of the less complex NB model can be more accurate. Comparing LVA inferred from the data of the same enhancers as the used in the construction of NB ensured a comparability of our two models, and enabled a fairer comparison of their performance.

The form of the LVA allowed us to study the effect of inclusion of the time course data from enhancers with no ChIA-PET-confirmed links on the predictions of ChIA-PET-confirmed interacting enhancers, and study of potential bias of the models which ignores that data. The time course data from the enhancers with no ChIA-PET evidence could not be used and thus was ignored in training and testing of the NB model 4.2.3, however the data was used in the predictive phase of the modelling to find the most likely targets of the enhancers with missing links. In that phase we assumed that the test error estimated on the test set which was not used to train the model was a reasonable estimate of the accuracy of the predictions for the group of enhancers with no ChIA-PET confirmed links. Inferring LVA from all data of all enhancers, and comparing its performance to the model inferred solely from the data of enhancers with ChIA-PET evidence allowed us to indirectly test the validity of our assumption. LVA inferred from

the extended set of all data of all enhancers showed decreased performance relative to the enhancer-restricted LVA, as validated on the ChIA-PET-detected links. Since the distributions of enrichments of the two groups of enhancers (Section 3.3.1) are similar, the difference in performance are unlikely to be caused by lower signal to noise ratio and quality of the ChIA-PET-undetected enhancers. However, the lack of associations in ChIA-PET data for the ER-alpha-bound enhancers could suggest that at least some of the enhancers may not be engaged with a nearby promoter in the cellular context (we propose how to test this hypothesis in Section 6.5), in such case the prior of the LVA based on ChIA-PET confirmed links would be unrepresentative for the group of enhancers. The prior could lead to erroneous allocations for the group of enhancers with no ChIA-PET data, which could in turn affect the mean time series of its assigned gene complexes, and in consequence reduce similarity between time series of the gene and its true regulatory enhancers, lowering confidence of some true pairs. To address the issue the LVA would need to be equipped with additional clusters to address the existence of non-interacting enhancers in the system. Alternatively, to study the potential bias we could compare our prior derived from ChIA-PET to an analogous prior derived from high resolution Hi-C data. The prior would be built as in Section 3.3 from separations of Hi-C-confirmed pairs of interacting ER-bound enhancers and genes.

## 6.5 Future work

The future development would likely concentrate on the second model since its design allows more flexibility. One of the layers which could be used to improve the model's performance and enable a more realistic modelling of the system of context-specific gene regulation could be the one capturing interactions between pairs of promoters. That combined with the ability of LVA to model assemblies of multiple regulatory enhancers at a single gene, could lead to full-scale realistic predictions of regulation involving multi-enhancers-multi-genes regulatory complexes. However, the relative scarceness of the links involving multiple-promoters in ChIA-PET data can be prohibitive if the prior of the promoter-promoter layer was to be estimated analogously to that in Chapter 5. The problem could be addressed via more appropriate chromatin conformation data such as promoter capture HiC (PCHi-C) or HiCap experiments, similar to the one available for mouse ES cells [99], which are specifically designed for capture of promoter-centred contacts.

In order to investigate potential bias of the ChIA-PET-derived prior, we could use an analogous Hi-C-derived prior estimated from contacts of ER- $\alpha$ -bound enhancers and genes confirmed by Hi-C data with a resolution higher than 40kb.

The decreased performance of the LVA caused by inclusion of time course data from enhancers with no ChIA-PET confirmed links could be likely solved by incorporation of additional clusters which could aggregate true non-interacting potentially inactive ER- $\alpha$ -bound enhancers. In order to test the hypothesis of the existence of the class of inactive enhancers, we could design multiple 4C experiments centred at some instances of the class of enhancers, to show that the enhancers do not form loops in the studied context. If successfully confirmed, we could attempt to characterise and distinguish enhancer activity using complementary ChIP-seq datasets of other TFs and patterns of their bindings at the enhancer.

Although, the currently available resolution of the Hi-C experiment (up to 40kB) is insufficient for accurate and unambiguous detection of loci involved in chromatin contacts between the windows of that size. We believe that our model could potentially be used to establish which of the loci within windows of 40kB are most likely engaged in the protein mediated enhancer-gene contacts.

The unrealistic assumption of the independence of the time points of the time course data in our mixture model, could potentially have a high impact on the performance of the mixture model. The non-diagonal covariance matrix structures, could be investigated to confront the assumption, at the cost of the more difficult and slower inference procedure. Alternatively, we could also use a form of gaussian processes model with an appropriate time series-aware kernel function.

Although the linear-distance prior is experimentally confirmed to be a limiting factor for the range of plausible enhancer-gene, gene-enhancer-gene interactions, with the enhancer sharing as an underlying mechanism for coordinated gene expression [92], it would be interesting to investigate whether other distance priors and measures of association could be more suitable for capture of multi-enhancer-multi-gene network topologies, such as the associativity measure applied genome-wide in [86].

Lastly, it would be interesting to investigate how the performance of the models changes after the introduction of additional publicly available cell-line specific single time point assays from other stimulation, cell types and their replicates from ENCODE and BLUEPRINT consortium. Among the datasets we can find experiments with the same stimulation of E2 as our time course data, such as c-MYC and CTCF ChIP-seq datasets (analysed in Section 3.2.1), as well as forty others from MCF-7 cell-line but

with basal unchanged concentration of ER- $\alpha$  ligands.

# Bibliography

- [1] S. Anders, P. T. Pyl, and W. Huber. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 2015.
- [2] R. Andersson. Promoter or enhancer, what’s the difference? Deconstruction of established distinctions and presentation of a unifying model. *BioEssays*, 37(3):314–323, 2015.
- [3] R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, et al. An atlas of active enhancers across human cell types and tissues. *Nature*, 507:455–461, 2014.
- [4] M. P. Applanat, H. Buteau-Lozano, M. A. Herve, and A. Corpet. Vascular endothelial growth factor is a target gene for estrogen receptor and contributes to breast cancer progression. *Adv Exp Med Biol*, 617:437–444, 2008.
- [5] S. Aranda, G. Mas, and L. Di Croce. Regulation of gene transcription by Polycomb proteins. *Science advances*, 1(11):e1500737, 2015.
- [6] S. D. Bailey, X. Zhang, K. Desai, M. Aid, O. Corradin, R. Cowper-Sal Lari, B. Akhtar-Zaidi, P. C. Scacheri, B. Haibe-Kains, M. Lupien, R. Cowper-Sal-lari, B. Akhtar-Zaidi, P. C. Scacheri, B. Haibe-Kains, and M. Lupien. ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nature Communications*, 2:6186, 2015.
- [7] A. J. Bannister and T. Kouzarides. Regulation of chromatin by histone modifications. *Cell research*, 21(3):381–395, 2011.
- [8] A. Barutcu, B. Lajoie, R. McCord, C. Tye, D. Hong, T. Messier, G. Browne, A. van Wijnen, J. Lian, J. Stein, J. Dekker, A. Imbalzano, and G. S. Stein.

- Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome Biology*, 16(1):214, 2015.
- [9] Y. Benjamini and T. P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10):1–14, 2012.
- [10] G. M. Bernardo, G. Bebek, C. L. Ginther, S. T. Sizemore, K. L. Lozada, J. D. Miedler, L. A. Anderson, A. K. Godwin, F. W. Abdul-Karim, D. J. Slamon, and R. A. Keri. FOXA1 represses the molecular phenotype of basal breast cancer cells. *Oncogene*, 32(5):554–63, 2013.
- [11] B. E. Bernstein, M. Kamal, K. Lindblad-Toh, S. Bekiranov, D. K. Bailey, D. J. Huebert, S. McMahon, E. K. Karlsson, E. J. Kulbokas, T. R. Gingeras, S. L. Schreiber, and E. S. Lander. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, 120(2):169–181, 2005.
- [12] C. M. Bishop. *Pattern recognition and Machine Learning*, volume 128. Springer, 2006.
- [13] V. Bourdeau, J. Deschênes, R. Métivier, Y. Nagai, D. Nguyen, N. Bretschneider, F. Gannon, J. H. White, and S. Mader. Genome-wide identification of high-affinity estrogen response elements in human and mouse. *Molecular Endocrinology*, 18(6):1411–1427, 2004.
- [14] L. Caizzi, G. Ferrero, S. Cutrupi, F. Cordero, C. Ballaré, V. Miano, S. Reineri, L. Ricci, O. Friard, A. Testori, D. Corà, M. Caselle, L. Di Croce, and M. De Bortoli. Genome-wide activity of unliganded estrogen receptor- $\alpha$  in breast cancer cells. *Proceedings of the National Academy of Sciences of the United States of America*, 111(13):1–6, 2014.
- [15] E. Calo and J. Wysocka. Modification of Enhancer Chromatin: What, How, and Why? *Molecular Cell*, 49(5):825–837, 2013.
- [16] J. S. Carroll, X. S. Liu, A. S. Brodsky, W. Li, C. A. Meyer, A. J. Szary, J. Eeckhoutte, W. Shao, E. V. Hestermann, T. R. Geistlinger, E. A. Fox, P. A. Silver, and M. Brown. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell*, 122(1):33–43, 2005.

- [17] J. S. Carroll, C. a. Meyer, J. Song, W. Li, T. R. Geistlinger, J. Eeckhoute, A. S. Brodsky, E. K. Keeton, K. C. Fertuck, G. F. Hall, Q. Wang, S. Bekiranov, V. Sementchenko, E. a. Fox, P. a. Silver, T. R. Gingeras, X. S. Liu, and M. Brown. Genome-wide analysis of estrogen receptor binding sites. *Nature genetics*, 38(11):1289–1297, 2006.
- [18] J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*, 37(SUPPL. 2):305–311, 2009.
- [19] J.-H. Cheng, D. Z.-C. Pan, Z. T.-Y. Tsai, and H.-K. Tsai. Genome-wide analysis of enhancer RNA in gene regulation across 12 mouse tissues. *Scientific reports*, 5:12648, 2015.
- [20] F. Ciabrelli and G. Cavalli. Chromatin-driven behavior of topologically associating domains. *Journal of Molecular Biology*, 427(3):608–625, 2015.
- [21] O. Corradin, A. Saiakhova, B. Akhtar-Zaidi, L. Myeroff, J. Willis, R. Cowper-Sallari, M. Lupien, S. Markowitz, and P. C. Scacheri. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Research*, 24:1–13, 2014.
- [22] Y. Cui, A. Niu, R. Pestell, R. Kumar, E. M. Curran, Y. Liu, and S. A. W. Fuqua. Metastasis-associated protein 2 is a repressor of estrogen receptor alpha whose overexpression leads to estrogen-independent growth of human breast cancer cells. *Molecular endocrinology (Baltimore, Md.)*, 20(9):2020–35, 2006.
- [23] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science (New York, N.Y.)*, 295(5558):1306–1311, 2002.
- [24] A. Diaz, A. Nellore, and J. S. Song. CHANCE: comprehensive software for quality control and validation of ChIP-seq data. *Genome Biology*, 13(10):R98, 2012.
- [25] J. R. Dixon, I. Jung, S. Selvaraj, Y. Shen, J. E. Antosiewicz-Bourget, A. Y. Lee, Z. Ye, A. Kim, N. Rajagopal, W. Xie, Y. Diao, J. Liang, H. Zhao, V. V. Lobanenkova, J. R. Ecker, J. A. Thomson, and B. Ren. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331–336, 2015.

- [26] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.
- [27] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.
- [28] D. L. Donoho. An invitation to reproducible computational research. *Biostatistics*, 11(3):385–388, 2010.
- [29] J. Dostie, T. a. Richmond, R. a. Arnaout, R. R. Selzer, W. L. Lee, T. a. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*, 16(10):1299–1309, 2006.
- [30] R. Drissen, R. Drissen, R.-j. Palstra, R.-j. Palstra, N. Gillemans, N. Gillemans, E. Splinter, E. Splinter, F. Grosveld, F. Grosveld, S. Philipsen, S. Philipsen, W. D. Laat, and W. D. Laat. The active spatial organization of the  $\beta$ -globin locus requires the transcription factor EKLF. *Genes & Development*, pages 2485–2490, 2004.
- [31] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shores, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011.
- [32] A. Feuerborn and P. R. Cook. Why the activity of a gene depends on its neighbors. *Trends in Genetics*, 31(9):483–490, 2015.
- [33] B. Frey and D. Dueck. Clustering by passing messages between data points. *science*, VOL 315(February):972–977, 2007.
- [34] M. J. Fullwood, M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, et al. An oestrogen-receptor- $\alpha$ -bound human chromatin interactome. *Nature*, 462(7269):58–64, 2009.



- [35] M. J. Fullwood and Y. Ruan. ChIP-based methods for the identification of long-range chromatin interactions. *Journal of Cellular Biochemistry*, 107(1):30–39, 2009.
- [36] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2004.
- [37] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.
- [38] Y. Ghavi-Helm, F. a. Klein, T. Pakozdi, L. Ciglar, D. Noordermeer, W. Huber, and E. E. M. Furlong. Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*, 2014.
- [39] D. Görür and C. E. Rasmussen. Dirichlet process Gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology*, 25(July):615–626, 2010.
- [40] C. E. Grueter. Mediator Complex Dependent Regulation of Cardiac Development and Disease. *Genomics, Proteomics and Bioinformatics*, 11(3):151–157, 2013.
- [41] H. Hagège, P. Klous, C. Braem, E. Splinter, J. Dekker, G. Cathala, W. de Laat, and T. Forné. Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nature protocols*, 2(7):1722–33, 2007.
- [42] N. Hah, C. Danko, L. Core, J. Waterfall, A. Siepel, J. Lis, and W. Kraus. A Rapid, Extensive, and Transient Transcriptional Response to Estrogen Signaling in Breast Cancer Cells. *Cell*, 145(7):1156, 2011.
- [43] L. Handoko, H. Xu, G. Li, C. Y. Ngan, E. Chew, M. Schnapp, C. W. H. Lee, C. Ye, J. L. H. Ping, F. Mulawadi, et al. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature genetics*, 43(7):630–638, 2011.
- [44] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. *Elements*, 1:337–387, 2009.
- [45] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika Vol*, 57:97–109, 1970.

- [46] B. He, C. Chen, L. Teng, and K. Tan. Global view of enhancer-promoter interactome in human cells. *Proceedings of the National Academy of Sciences of the United States of America*, 111:E2191—9, 2014.
- [47] N. D. Heintzman, G. C. Hon, R. D. Hawkins, P. Kheradpour, A. Stark, L. F. Harp, Z. Ye, L. K. Lee, R. K. Stuart, C. W. Ching, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112, 2009.
- [48] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, 38(4):576–589, 2010.
- [49] A. Honkela, J. Peltonen, H. Topa, I. Charapitsa, F. Matarese, K. Grote, H. G. Stunnenberg, G. Reid, N. D. Lawrence, and M. Rattray. Genome-wide modeling of transcription kinetics reveals patterns of RNA production delays. *Proceedings of the National Academy of Sciences*, page 201420404, 2015.
- [50] S. Hua, R. Kittler, and K. P. White. Genomic Antagonism between Retinoic Acid and Estrogen Signaling in Breast Cancer. *Cell*, 137(7):1259–1271, 2009.
- [51] A. Hurtado, K. A. Holmes, C. S. Ross-Innes, D. Schmidt, and J. S. Carroll. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nature Genetics*, 43(1):27–33, 2011.
- [52] F. Jin, Y. Li, J. R. Dixon, S. Selvaraj, Z. Ye, A. Y. Lee, C.-A. Yen, A. D. Schmitt, C. a. Espinoza, and B. Ren. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 503(7475):290–294, 2013.
- [53] R. Joseph, Y. L. Orlov, M. Huss, W. Sun, S. L. Kong, L. Ukil, Y. F. Pan, G. Li, M. Lim, J. S. Thomsen, et al. Integrative model of genomic factors for determining binding site selection by estrogen receptor- $\alpha$ . *Molecular Systems Biology*, 6(456):456, 2010.
- [54] K. M. Jozwik and J. S. Carroll. Pioneer factors in hormone-dependent cancers. *Nature Reviews Cancer*, 12(6):381–385, 2012.

- [55] T.-K. Kim, M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. A. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–7, 2010.
- [56] M. S. Kowalczyk, J. R. Hughes, D. Garrick, M. D. Lynch, J. A. Sharpe, J. A. Sloane-Stanley, S. J. McGowan, M. De Gobbi, M. Hosseini, D. Vernimmen, et al. Intragenic Enhancers Act as Alternative Promoters. *Molecular Cell*, 45(4):447–458, 2012.
- [57] E. Z. Kvon, T. Kazmar, G. Stampfel, J. O. Yáñez-Cuna, M. Pagani, K. Scherhuber, B. J. Dickson, and A. Stark. Genome-scale functional characterization of *Drosophila* developmental enhancers in vivo. *Nature*, 512(7512):91–5, 2014.
- [58] P. La Rosa, V. Pesiri, M. Marino, and F. Acconcia. 17-Estradiol-induced cell proliferation requires estrogen receptor (ER) monoubiquitination. *Cellular Signalling*, 23(7):1128–1135, 2011.
- [59] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3):R25, 2009.
- [60] R. Lanz, Y. Bulyanko, A. Malovannaya, P. Labhart, L. Wang, W. Li, J. Qin, M. Harper, and B. O’Malley. Global characterization of transcriptional impact of the SRC-3 coregulator. *Molecular endocrinology (Baltimore, Md.)*, 24(4):859–872, 2010.
- [61] K. Lee, C. C. Hsiung, P. Huang, A. Raj, and G. A. Blobel. Dynamic enhancer Gene body contacts during transcription elongation. *Genes & Development*, pages 1992–1997, 2015.
- [62] . J. T. L. Leighton J. Core,\* Joshua J. Waterfall. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *science*, 322(December):0–4, 2008.
- [63] G. Li, M. J. Fullwood, H. Xu, F. H. Mulawadi, S. Velkov, V. Vega, P. N. Ariyaratne, Y. B. Mohamed, H.-S. Ooi, C. Tennakoon, et al. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome biology*, 11(2):R22, 2010.

- [64] G. Li, X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1-2):84–98, 2012.
- [65] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.
- [66] W. Li, D. Notani, Q. Ma, B. Tanasa, E. Nunez, A. Y. Chen, D. Merkurjev, J. Zhang, K. Ohgi, X. Song, et al. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*, 498(7455):516–20, 2013.
- [67] E. Lieberman-Aiden, N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragozy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.
- [68] M. H. Liu and E. Cheung. Estrogen receptor-mediated long-range chromatin interactions and transcription in breast cancer. *Molecular and cellular endocrinology*, 382(1):624–632, 2014.
- [69] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550, 2014.
- [70] M. Lupien, J. Eeckhoute, C. A. Meyer, Q. Wang, Y. Zhang, W. Li, J. S. Carroll, X. S. Liu, and M. Brown. FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-Specific Transcription. *Cell*, 132(6):958–970, 2008.
- [71] L. Magnani, E. B. Ballantyne, X. Zhang, and M. Lupien. PBX1 genomic pioneer function drives ER $\alpha$  signaling underlying progression in breast cancer. *PLoS Genetics*, 7(11):1–15, 2011.
- [72] L. Magnani and M. Lupien. Chromatin and epigenetic determinants of estrogen receptor alpha (ESR1) signaling. *Molecular and cellular endocrinology*, 382(1):633–641, 2014.
- [73] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517, 2008.

- [74] T. T. Marstrand and J. D. Storey. Identifying and mapping cell-type-specific chromatin programming of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 111:E645—54, 2014.
- [75] N. Matharu and N. Ahituv. Minor loops in major folds: Enhancer–promoter looping, chromatin restructuring, and their association with transcriptional regulation and disease. *PLoS Genetics*, 11(12):1–14, 2015.
- [76] T. R. Mercer and J. S. Mattick. Understanding the regulatory and transcriptional complexity of the genome through structure. *Genome Research*, 23(7):1081–1088, 2013.
- [77] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal Chemical Physics*, 21(6):1087–1092, 1953.
- [78] B. Mifsud, F. Tavares-Cadete, A. N. Young, R. Sugar, S. Schoenfelder, L. Ferreira, S. W. Wingett, S. Andrews, W. Grey, P. A. Ewels, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, 47(6):598–606, 2015.
- [79] R. Mourad, P.-Y. Hsu, L. Juan, C. Shen, P. Koneru, H. Lin, Y. Liu, K. Nephew, T. H. Huang, and L. Li. Estrogen induces global reorganization of chromatin structure in human breast cancer cells. *PloS one*, 9(12):e113354, 2014.
- [80] K. Mousavi, H. Zare, S. Dell’Orso, L. Grontved, G. Gutierrez-Cruz, A. Derfoul, G. Hager, and V. Sartorelli. ERNAs Promote Transcription by Establishing Chromatin Accessibility at Defined Genomic Loci. *Molecular Cell*, 51(5):606–617, 2013.
- [81] Murray Rosenblatt. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, Volume 27:832–837, 1956.
- [82] T. Nagano, Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E. D. Laue, A. Tanay, and P. Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013.
- [83] S. Nagarajan, T. Hossan, M. Alawi, Z. Najafova, D. Indenbirken, U. Bedi, H. Taipaleenmäki, I. Ben-Batalla, M. Scheller, S. Loges, et al. Bromodomain

- protein BRD4 is required for estrogen receptor-dependent enhancer activation and gene transcription. *Cell reports*, 8(2):460–469, 2014.
- [84] D. Noordermeer, E. De Wit, P. Klous, H. Van De Werken, M. Simonis, M. Lopez-Jones, B. Eussen, A. De Klein, R. H. Singer, and W. De Laat. Variegated gene expression caused by cell-specific long-range DNA interactions. *Nature cell biology*, 13(8):944–951, 2011.
- [85] C.-T. Ong and V. G. Corces. CTCF: an architectural protein bridging genome topology and function. *Nature reviews. Genetics*, 15(4):234–46, 2014.
- [86] V. Pancaldi, E. Carrillo-de Santa-Pau, B. M. Javierre, D. Juan, P. Fraser, M. Spivakov, A. Valencia, and D. Rico. Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. *Genome biology*, 17(1):152, 2016.
- [87] P. J. Park. ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, 10(10):669–80, 2009.
- [88] E. Parzen. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.
- [89] M. D. Planas-Silva and P. K. Waltz. Estrogen promotes reversible epithelial-to-mesenchymal-like transition and collective motility in MCF-7 breast cancer cells. *Journal of Steroid Biochemistry and Molecular Biology*, 104(1-2):11–21, 2007.
- [90] B. D. Pope, T. Ryba, V. Dileep, F. Yue, W. Wu, O. Denas, D. L. Vera, Y. Wang, R. S. Hansen, T. K. Canfield, et al. Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515(7527):402–405, 2014.
- [91] A. R. Quinlan and I. M. Hall. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.
- [92] P. Quintero-Cadena and P. Sternberg. Enhancer sharing promotes neighborhoods of transcriptional regulation across eukaryotes. *bioRxiv*, page 061218, 2016.
- [93] C. Ribeiro de Almeida, R. Stadhouders, M. J. W. De Bruijn, I. M. Bergen, S. Thongjuea, B. Lenhard, W. Van IJcken, F. Grosveld, N. Galjart, E. Soler,

- and R. W. Hendriks. The dna-binding protein ctf limits proximal  $\nu\kappa$  recombination and restricts  $\kappa$  enhancer interactions to the immunoglobulin  $\kappa$  light chain locus. *Immunity*, 35(4):501–513, 2011.
- [94] C. S. Ross-Innes, R. Stark, A. E. Teschendorff, K. A. Holmes, H. R. Ali, M. J. Dunning, G. D. Brown, O. Gojis, I. O. Ellis, A. R. Green, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, 481(7381):389–393, 2012.
- [95] S. Roy, A. F. Siahpirani, D. Chasman, S. Knaack, F. Ay, R. Stewart, M. Wilson, and R. Sridharan. A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Research*, 43(18):gkv865, 2015.
- [96] G. K. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig. Ten Simple Rules for Reproducible Computational Research. *PLoS Computational Biology*, 9(10):1–4, 2013.
- [97] A. Sanyal, B. R. Lajoie, G. Jain, and J. Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113, 2012.
- [98] D. Schmidt, P. C. Schwalie, C. S. Ross-Innes, A. Hurtado, G. D. Brown, J. S. Carroll, P. Flicek, and D. T. Odom. A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Research*, 20(5):578–588, 2010.
- [99] S. Schoenfelder, M. Furlan-Magaril, B. Mifsud, F. Tavares-Cadete, R. Sugar, B.-M. Javierre, T. Nagano, Y. Katsman, M. Sakthidevi, S. W. Wingett, et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Research*, 25(4):582–597, 2015.
- [100] S. Schoenfelder, T. Sexton, L. Chkalova, N. F. Cope, A. Horton, S. Andrews, S. Kurukuti, J. A. Mitchell, D. Umlauf, D. S. Dimitrova, et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nature genetics*, 42(1):53–61, 2010.
- [101] S. Schwartz, R. Oren, and G. Ast. Detection and removal of biases in the analysis of next-generation sequencing reads. *PLoS ONE*, 6(1), 2011.
- [102] D. W. Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

- [103] T. Sexton and G. Cavalli. Review The Role of Chromosome Domains in Shaping the Functional Genome. *Cell*, 160(6):1049–1059, 2015.
- [104] T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay, and G. Cavalli. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*, 148(3):458–472, 2012.
- [105] Y. Shen, F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, and B. Ren. A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116–120, 2012.
- [106] D. Shlyueva, G. Stampfel, and A. Stark. Transcriptional enhancers: from properties to genome-wide predictions. *Nature reviews. Genetics*, 15(4):272–86, 2014.
- [107] B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.
- [108] M. Simonis, J. Kooren, and W. de Laat. An evaluation of 3C-based methods to capture DNA interactions. *Nature methods*, 4(11):895–901, 2007.
- [109] S. Sofueva and S. Hadjur. Cohesin-mediated chromatin interactions into the third dimension of gene regulation. *Briefings in Functional Genomics*, 11(3):205–216, 2012.
- [110] S. Sofueva, E. Yaffe, W.-C. Chan, D. Georgopoulou, M. Vietri Rudan, H. Mira-Bontenbal, S. M. Pollard, G. P. Schroth, A. Tanay, and S. Hadjur. Cohesin-mediated interactions organize chromosomal domain architecture. *The EMBO journal*, 32(24):3119–29, 2013.
- [111] T. J. Stasevich, Y. Hayashi-Takanaka, Y. Sato, K. Maehara, Y. Ohkawa, K. Sakata-Sogawa, M. Tokunaga, T. Nagase, N. Nozaki, J. G. McNally, and H. Kimura. Regulation of RNA polymerase II activation by histone acetylation in single living cells. *Nature*, 516(7530):272–275, 2014.
- [112] C. Stellato, I. Porreca, D. Cuomo, R. Tarallo, G. Nassa, and C. Ambrosino. The “busy life” of unliganded estrogen receptors. *Proteomics*, 16(2):288–300, 2016.



- [113] O. A. Sukocheva, Y. Yang, and J. F. Gierthy. Estrogen and progesterone interactive effects in postconfluent MCF-7 cell culture. *Steroids*, 74(4-5):410–418, 2009.
- [114] T. Takai, Yoshimi and Sasaki, Takuya and Matozaki. Small GTP-binding proteins. *Physiological Reviews*, 81(1):153—208, 2001.
- [115] S. K. Tan, Z. H. Lin, C. W. Chang, V. Varang, K. R. Chng, Y. F. Pan, E. L. Yong, W. K. Sung, W. K. Sung, and E. Cheung. AP-2 $\gamma$  regulates oestrogen receptor-mediated long-range chromatin interaction and gene transcription. *The EMBO journal*, 30(13):2569–81, 2011.
- [116] T. Tanos, L. Rojo, P. Echeverria, and C. Brisken. ER and PR signaling nodes during mammary gland development. *Breast cancer research : BCR*, 14(4):210, 2012.
- [117] L. Teytelman, B. Özaydin, O. Zill, P. Lefrançois, M. Snyder, J. Rine, and M. B. Eisen. Impact of chromatin structures on DNA processing for genomic analyses. *PLoS ONE*, 4(8):1–11, 2009.
- [118] V. Theodorou, R. Stark, S. Menon, and J. S. Carroll. GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. *Genome Research*, 23(1):12–22, 2013.
- [119] R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, et al. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 2012.
- [120] B. Tolhuis, R. J. Palstra, E. Splinter, F. Grosveld, and W. De Laat. Looping and interaction between hypersensitive sites in the active  $\beta$ -globin locus. *Molecular Cell*, 10(6):1453–1465, 2002.
- [121] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.
- [122] W.-W. Tsai, Z. Wang, T. T. Yiu, K. C. Akdemir, W. Xia, S. Winter, C.-Y. Tsai, X. Shi, D. Schwarzer, W. Plunkett, B. Aronow, O. Gozani, W. Fischle, M.-C. Hung, D. J. Patel, and M. C. Barton. TRIM24 links a non-canonical histone signature to breast cancer. *Nature*, 468(7326):927–932, 2010.

- [123] C. R. Vakoc, D. L. Letting, N. Gheldof, T. Sawado, M. a. Bender, M. Groudine, M. J. Weiss, J. Dekker, and G. a. Blobel. Proximity among distant regulatory elements at the  $\beta$ -globin locus requires GATA-1 and FOG-1. *Molecular Cell*, 17(3):453–462, 2005.
- [124] B. van Steensel and J. Dekker. Genomics tools for unraveling chromosome architecture. *Nature Biotechnology*, 28(10):1089–1095, 2010.
- [125] M. M. Vantangoli, S. Wilson, S. J. Madnick, S. M. Huse, and K. Boekelheide. Morphologic effects of estrogen stimulation on 3D MCF-7 microtissues. *Toxicology Letters*, 248:1–8, 2016.
- [126] A. Visel, M. J. Blow, Z. Li, T. Zhang, J. a. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, V. Afzal, B. Ren, E. M. Rubin, and L. a. Pennacchio. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–8, 2009.
- [127] C. wa Maina, A. Honkela, F. Matarese, K. Grote, H. G. Stunnenberg, G. Reid, N. D. Lawrence, and M. Rattray. Inference of RNA Polymerase II Transcription Dynamics from Chromatin Immunoprecipitation Time Course Data. *PLoS Computational Biology*, 10(5):e1003598, 2014.
- [128] W.-J. Welboren, M. A. Van Driel, E. M. Janssen-Megens, S. J. Van Heeringen, F. C. Sweep, P. N. Span, and H. G. Stunnenberg. ChIP-Seq of ER $\alpha$  and RNA polymerase II defines genes differentially responding to ligands. *The EMBO Journal*, 28(10):1418–1428, 2009.
- [129] E. G. Wilbanks and M. T. Facciotti. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE*, 5(7), 2010.
- [130] C. K. I. Williams. How to pretend that correlated variables are independent by using difference observations. *Neural computation*, 17(1):1–7, 2005.
- [131] S. Witte, A. Bradley, A. J. Enright, and S. A. Muljo. High-density P300 enhancers control cell state transitions. *BMC genomics*, 16(1):903, 2015.
- [132] L. Yao, B. P. Berman, and P. J. Farnham. Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes. *Critical Reviews in Biochemistry and Molecular Biology*, 9238(October):1–24, 2015.

- [133] Y. Yu, Y. Hao, and L. A. Feig. The R-Ras GTPase mediates cross talk between estrogen and insulin signaling in breast cancer cells. *Molecular and cellular biology*, 26(17):6372–80, 2006.
- [134] G. E. Zentner and S. Henikoff. Regulation of nucleosome dynamics by histone modifications. *Nature structural & molecular biology*, 20(3):259–66, 2013.
- [135] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nussbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, 2008.
- [136] Y. Zhang, C.-H. Wong, R. Y. Birnbaum, G. Li, R. Favaro, C. Y. Ngan, J. Lim, E. Tai, H. M. Poh, E. Wong, et al. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature*, 504(7479):306–310, 2013.
- [137] Z. Zhao, G. Tavoosidana, M. Sjölander, A. Göndör, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. S. Sandhu, U. Singh, et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature genetics*, 38(11):1341–7, 2006.
- [138] Y. Zhu, L. Sun, Z. Chen, J. W. Whitaker, T. Wang, and W. Wang. Predicting enhancer transcription and activity from chromatin modifications. *Nucleic Acids Research*, 41(22):10032–10043, 2013.
- [139] J. Zuin, J. R. Dixon, M. I. van der Reijden, Z. Ye, P. Kolovos, R. W. Brouwer, M. P. van de Corput, H. J. van de Werken, T. A. Knoch, W. F. van IJcken, et al. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci USA*, 111(3):996–1001, 2014.
- [140] W. Zwart, V. Theodorou, M. Kok, S. Canisius, S. Linn, and J. S. Carroll. Oestrogen receptor-co-factor-chromatin specificity in the transcriptional regulation of breast cancer. *Embo J*, 30(23):4764–4776, 2011.

# **Appendix A**

## **Supplementary figures for Chapter 3**

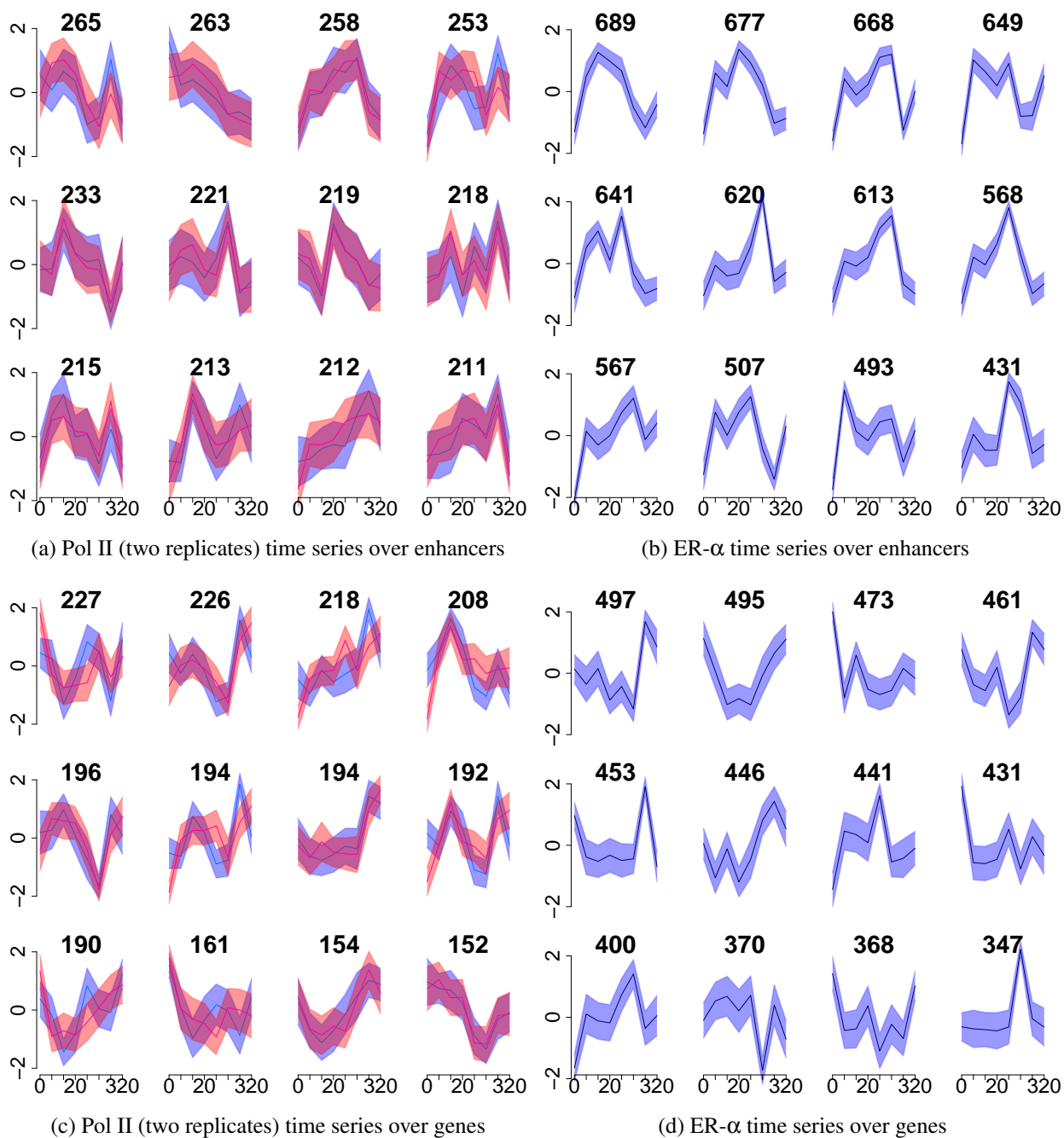


Figure A.1: Figure shows subsequent clusters of Fig. 3.4. The first column of the figure shows the results of the clustering of joint time series of both Pol II replicates at enhancers (top row) and genes (bottom row) with Affinity Propagation. The second column shows the corresponding clustering for ER- $\alpha$  time series at the enhancer (top row) and gene regions (bottom row).

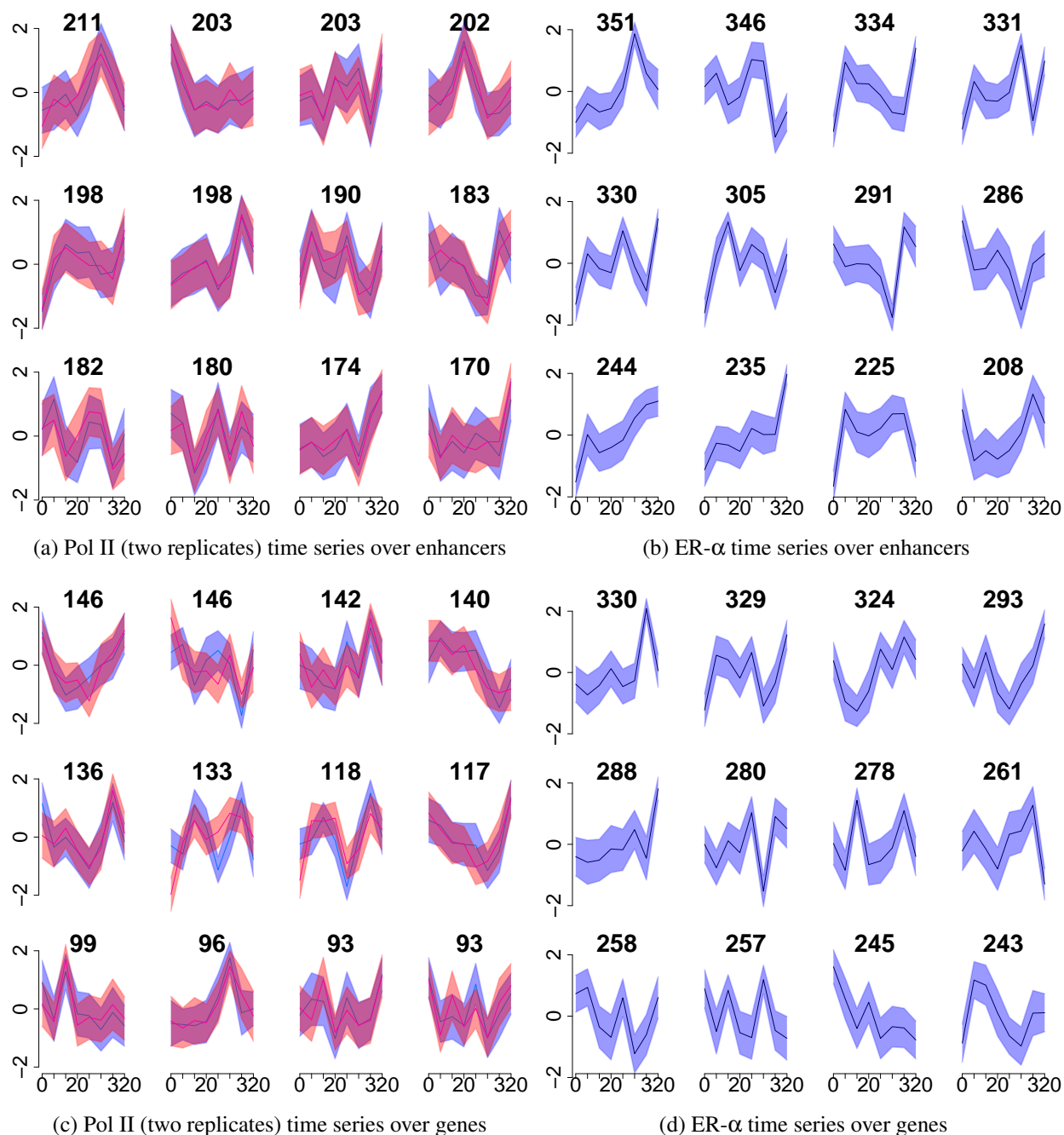
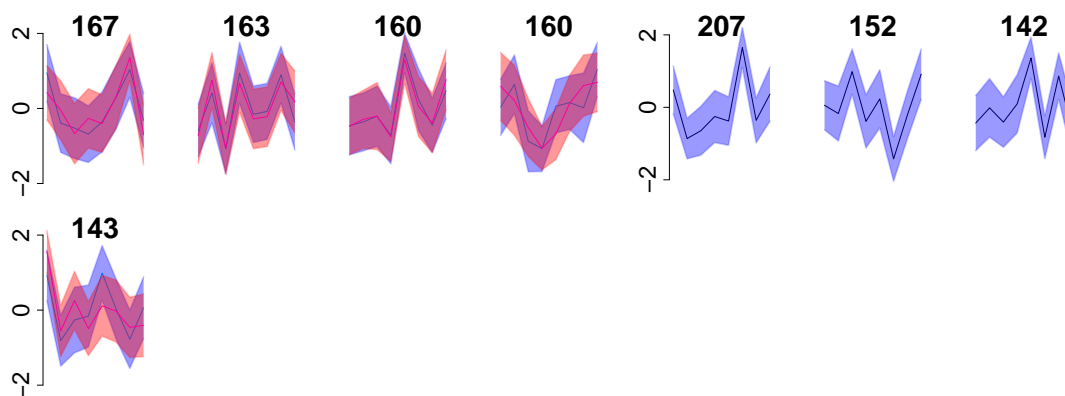
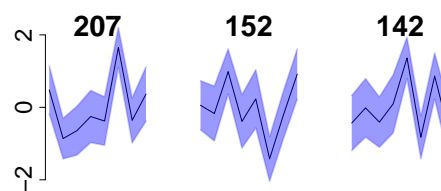
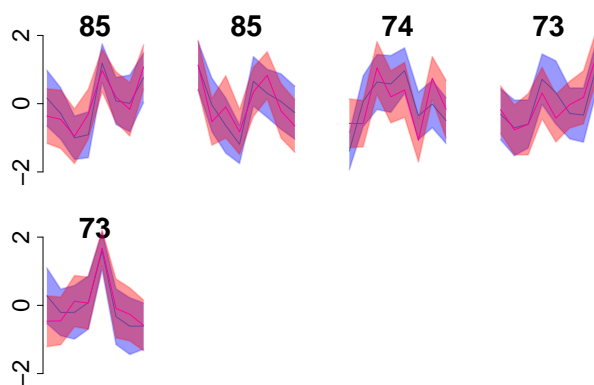


Figure A.2: Figure shows subsequent clusters of Fig. 3.4. The first column of the figure shows the results of the clustering of joint time series of both Pol II replicates at enhancers (top row) and genes (bottom row) with Affinity Propagation. The second column shows the corresponding clustering for ER- $\alpha$  time series at the enhancer (top row) and gene regions (bottom row).



(a) Pol II (two replicates) time series over enhancers

(b) ER- $\alpha$  time series over enhancers

(c) Pol II (two replicates) time series over genes

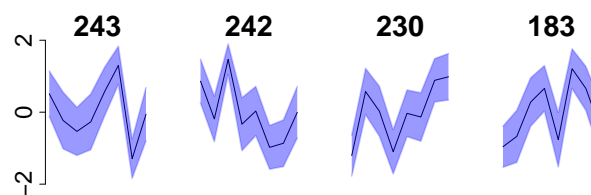







(d) ER- $\alpha$  time series over genes

Figure A.3: Figure shows subsequent clusters of Fig. 3.4. The first column of the figure shows the results of the clustering of joint time series of both Pol II replicates at enhancers (top row) and genes (bottom row) with Affinity Propagation. The second column shows the corresponding clustering for ER- $\alpha$  time series at the enhancer (top row) and gene regions (bottom row).

	CTCF	ESR1	FoxA1	SRC-3	STAG1	TRIM24	c-Fos	c-Jun	c-MYC	Count	Amplitude
1	324	164	69	152	363	230	71	36	65	434	258
2	348	284	177	287	355	261	154	56	28	391	216
3	285	123	53	119	303	198	62	34	41	359	185
4	293	195	154	214	307	221	121	36	21	341	82
5	240	108	69	105	248	165	39	7	24	299	78
6	210	80	48	89	238	161	41	13	20	290	28
7	240	158	113	162	262	200	104	41	25	287	224
8	227	189	112	190	238	183	115	44	24	270	301
9	219	156	97	148	233	174	82	40	18	259	331
10	213	108	108	137	228	154	78	20	23	255	-24
11	212	140	67	139	227	140	67	25	5	243	302
12	193	148	80	134	200	156	73	37	28	237	313
13	123	73	43	53	130	98	22	7	10	227	-238
14	176	88	60	93	188	126	54	21	14	226	14
15	170	71	52	79	169	113	41	27	23	218	264
16	168	92	47	85	190	98	45	19	6	208	206
17	147	76	72	100	153	122	54	20	7	196	2
18	145	74	27	79	167	113	38	20	12	194	229
19	138	75	35	60	143	110	34	7	11	194	-27
20	150	49	31	62	158	89	37	23	10	192	120
21	118	72	38	64	120	93	32	5	8	190	-174
22	76	35	32	45	83	59	19	3	4	161	-271
23	103	58	42	57	104	77	28	8	10	154	-117
24	121	67	71	83	130	93	49	8	11	152	-186
25	95	46	38	53	95	70	24	2	8	146	-219
26	72	35	39	42	77	50	22	5	9	146	-228
27	80	37	16	33	74	51	11	3	4	142	-39
28	100	74	65	75	103	72	45	7	16	140	-99
29	90	29	23	37	87	53	22	5	7	136	-106
30	99	37	14	38	112	56	17	11	7	133	215
31	88	36	23	43	95	70	15	7	9	118	57
32	81	40	40	46	83	63	16	5	5	117	-186
33	60	20	5	13	57	24	5	1	1	99	-45
34	61	41	26	35	66	40	19	2	3	96	63
35	49	23	10	18	43	27	4	0	3	93	-7
36	53	18	15	22	51	47	5	4	4	93	-99
37	37	16	10	20	41	22	6	2	1	85	133
38	29	14	8	11	31	28	7	0	1	85	-79
39	54	28	9	32	55	29	9	3	2	74	97
40	51	25	21	30	52	38	13	4	5	73	214
41	31	11	3	12	31	26	6	0	1	73	-24

	enriched, p < 0.01
	enriched, p < 0.005
	enriched, p < 0.001
	depleted, p < 0.01
	depleted, p < 0.005
	depleted, p < 0.001
	neither

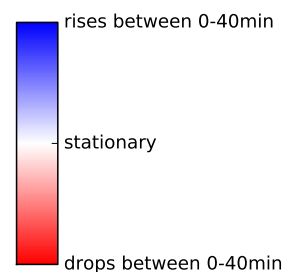


Table A.1: The table shows the cluster-specific patterns of TF bindings across 300bp-upstream-extended-genes for the corresponding Pol II clusters in figure 3.4c. The count column indicates the size of the cluster. The amplitude column shows the difference between Pol II occupancies in E2-deprived (0 min) and E2-stimulated samples (40 min) in the mean of each cluster. (dynamics of orange replicate)



	CTCF	ESR1	FoxA1	SRC-3	STAG1	TRIM24	c-Fos	c-Jun	c-MYC	Count	Amplitude
1	137	852	350	920	446	135	249	29	5	1819	286
2	162	982	111	508	287	106	50	15	4	1816	272
3	228	710	174	526	420	149	108	34	4	1662	210
4	221	918	359	945	565	215	294	69	7	1613	287
5	254	558	252	631	501	168	202	55	11	1312	251
6	200	132	131	265	329	80	72	22	8	851	121
7	108	194	45	132	156	55	22	4	0	844	171
8	124	392	273	517	358	147	242	66	3	837	327
9	80	180	147	267	192	48	59	4	3	816	183
10	61	141	201	316	168	41	125	11	3	739	184
11	104	141	74	207	163	51	32	10	0	706	255
12	72	175	90	189	158	48	45	6	4	694	160
13	105	167	38	118	130	37	18	8	1	689	197
14	46	183	48	116	81	23	18	4	2	677	229
15	39	173	222	364	178	38	124	8	4	668	269
16	163	152	93	261	277	102	86	30	2	649	260
17	87	143	58	144	122	39	26	14	4	641	264
18	24	44	269	324	137	24	172	10	0	620	159
19	51	82	172	290	125	20	98	2	2	613	239
20	34	146	78	182	81	23	33	10	1	568	309
21	64	154	172	298	187	44	186	33	0	567	281
22	58	106	86	179	119	26	28	2	1	507	252
23	70	106	167	250	202	48	117	17	1	493	218
24	31	36	155	219	97	27	72	6	2	431	277
25	32	26	133	159	87	21	105	12	4	351	110
26	38	28	144	238	94	17	47	12	1	346	88
27	100	35	69	119	158	37	44	14	4	334	116
28	27	28	122	174	85	21	70	8	1	331	118
29	62	47	99	167	120	38	61	12	0	330	238
30	53	47	59	124	87	23	36	5	1	305	221
31	54	53	31	69	61	108	7	12	0	291	-106
32	48	45	20	68	63	75	19	22	1	286	-157
33	34	26	66	93	72	32	67	14	2	244	135
34	55	28	65	105	103	32	56	14	2	235	134
35	32	22	47	87	60	10	32	5	1	225	187
36	19	8	13	118	29	38	29	19	1	208	-131
37	8	9	37	100	30	12	31	9	0	207	-85
38	32	23	23	60	47	22	18	9	1	152	17
39	26	6	33	60	43	16	14	7	0	142	179

enriched,  $p < 0.01$

enriched,  $p < 0.005$

enriched,  $p < 0.001$

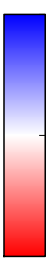
depleted,  $p < 0.01$

depleted,  $p < 0.005$

depleted,  $p < 0.001$

neither

rises between 0-40min



stationary

drops between 0-40min

Table A.2: The table shows the cluster-specific patterns of TF bindings across enhancers for the corresponding ER- $\alpha$  clusters in figure 3.4b. The count column indicates the size of the cluster. The amplitude column shows the difference between ER- $\alpha$  occupancies in E2-deprived (0 min) and E2-stimulated samples (40 min) in the mean dynamics of each cluster.

	CTCF	ESR1	FoxA1	SRC-3	STAG1	TRIM24	c-Fos	c-Jun	c-MYC	Count	Amplitude
1	598	216	81	166	716	192	71	30	32	1178	278
2	596	183	65	182	661	227	79	35	12	1076	201
3	589	112	75	187	705	255	87	64	37	940	-222
4	482	103	46	108	556	170	52	33	23	847	-190
5	383	78	38	88	459	112	33	20	21	694	-287
6	350	88	50	131	408	179	54	39	23	620	-189
7	394	73	37	111	442	166	59	56	35	617	-137
8	196	24	13	22	221	54	6	3	5	587	-184
9	197	41	11	48	252	69	19	13	12	559	-189
10	278	47	26	50	343	86	23	17	13	559	-197
11	308	57	33	86	356	112	33	26	22	514	-136
12	264	49	25	65	315	93	26	10	16	497	-247
13	257	50	19	66	285	105	35	17	16	497	-56
14	236	49	24	59	279	98	18	18	17	495	-216
15	262	54	29	80	316	83	37	24	19	473	-269
16	219	31	22	41	256	68	24	16	15	461	-214
17	217	42	23	54	260	82	22	16	11	453	-145
18	192	32	11	34	224	46	16	6	13	446	-51
19	228	88	29	63	269	98	36	16	5	441	305
20	232	45	21	81	291	94	36	31	20	431	-139
21	181	56	23	51	216	63	24	7	9	400	238
22	198	52	20	41	215	64	21	20	8	370	81
23	158	21	14	26	191	50	13	3	11	368	-252
24	100	23	10	9	123	26	7	1	2	347	0
25	134	25	11	42	149	42	15	4	7	330	-7
26	199	44	20	55	213	81	17	12	8	329	188
27	138	26	15	35	167	51	14	9	7	324	37
28	126	22	15	27	133	51	14	8	6	293	-146
29	120	19	17	21	150	43	6	6	5	288	22
30	143	32	14	42	170	64	20	17	12	280	103
31	119	24	7	37	129	35	17	10	11	278	-57
32	96	21	8	13	119	24	3	0	4	261	54
33	151	33	15	35	166	61	20	16	4	258	-13
34	84	19	13	22	101	30	2	3	5	257	-159
35	106	19	11	23	128	35	11	6	5	245	-233
36	110	31	13	34	133	43	13	5	6	243	23
37	97	25	12	19	112	30	6	3	8	243	5
38	128	25	6	22	141	38	10	11	7	242	-83
39	96	18	16	27	104	35	13	8	8	230	117
40	50	8	6	8	60	12	0	2	0	183	19

enriched, $p < 0.01$
enriched, $p < 0.005$
enriched, $p < 0.001$
depleted, $p < 0.01$
depleted, $p < 0.005$
depleted, $p < 0.001$
neither

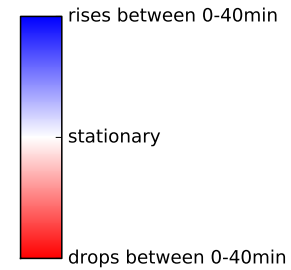


Table A.3: The table shows the cluster-specific patterns of TF bindings across 2000bp-long TSS-centred regions (around promoters) for the corresponding ER- $\alpha$  clusters in figure 3.4d. The count column indicates the size of the cluster. The amplitude column shows the difference between ER- $\alpha$  occupancies in E2-deprived (0 min) and E2-stimulated samples (40 min) in the mean dynamics of each cluster.

	CTCF	ESR1	FoxA1	SRC-3	STAG1	TRIM24	c-Fos	c-Jun	c-MYC	Count	Amplitude
1	996	973	548	871	1023	765	449	131	105	1178	278
2	905	700	395	695	924	672	351	148	60	1076	201
3	800	401	258	450	853	649	208	115	81	940	-222
4	686	287	177	288	729	501	141	63	80	847	-190
5	559	235	176	249	597	419	110	35	48	694	-287
6	511	296	198	311	549	425	138	79	52	620	-189
7	530	222	147	249	544	393	122	80	52	617	-137
8	423	199	166	161	449	323	103	16	49	587	-184
9	382	151	103	139	417	279	62	23	42	559	-189
10	417	154	96	139	451	288	63	27	36	559	-197
11	408	137	101	161	430	276	71	49	34	514	-136
12	378	120	85	144	407	255	56	22	24	497	-247
13	395	168	89	183	408	265	97	44	37	497	-56
14	344	113	82	131	369	252	54	25	27	495	-216
15	361	139	82	162	392	217	65	32	36	473	-269
16	335	119	87	141	367	241	74	32	35	461	-214
17	327	105	62	114	348	206	50	27	18	453	-145
18	310	110	76	127	330	212	58	16	19	446	-51
19	358	287	141	258	376	261	135	54	29	441	305
20	315	106	63	147	344	201	57	47	28	431	-139
21	314	253	192	253	328	241	137	20	32	400	238
22	312	202	109	185	322	235	84	41	23	370	81
23	255	99	71	84	280	174	35	8	22	368	-252
24	251	157	116	133	257	191	59	8	27	347	0
25	211	76	41	75	224	150	36	15	13	330	-7
26	276	151	79	164	290	174	87	35	23	329	188
27	207	66	48	82	231	134	39	16	14	324	37
28	206	60	40	68	208	127	28	11	12	293	-146
29	183	59	57	76	201	129	36	11	12	288	22
30	204	101	61	112	216	144	47	29	18	280	103
31	171	59	27	62	184	98	34	16	11	278	-57
32	160	79	50	67	179	108	26	7	16	261	54
33	205	74	54	87	214	121	40	27	11	258	-13
34	151	49	43	61	158	102	21	5	12	257	-159
35	152	53	40	60	170	90	28	9	10	245	-233
36	177	93	49	93	185	110	46	17	12	243	23
37	149	74	63	99	164	90	44	10	15	243	5
38	171	64	31	62	183	97	29	19	14	242	-83
39	152	62	56	85	159	110	42	17	13	230	117
40	106	44	25	43	106	55	8	3	2	183	19

<span style="background-color: #c8e6c9; border: 1px solid black; padding: 2px;"> </span>	enriched, $p < 0.01$
<span style="background-color: #a5d6a7; border: 1px solid black; padding: 2px;"> </span>	enriched, $p < 0.005$
<span style="background-color: #81c784; border: 1px solid black; padding: 2px;"> </span>	enriched, $p < 0.001$
<span style="background-color: #ffcdd2; border: 1px solid black; padding: 2px;"> </span>	depleted, $p < 0.01$
<span style="background-color: #ffb74d; border: 1px solid black; padding: 2px;"> </span>	depleted, $p < 0.005$
<span style="background-color: #ff8a65; border: 1px solid black; padding: 2px;"> </span>	depleted, $p < 0.001$
<span style="background-color: #f5f5f5; border: 1px solid black; padding: 2px;"> </span>	neither

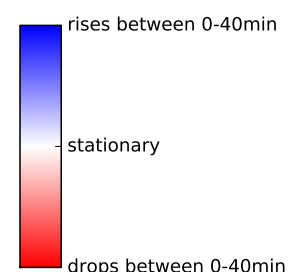


Table A.4: The table shows the cluster-specific patterns of TF bindings across across 300bp-upstream-extended-genes for the corresponding ER- $\alpha$  clusters in figure 3.4d. The count column indicates the size of the cluster. The amplitude column shows the difference between ER- $\alpha$  occupancies in E2-deprived (0 min) and E2-stimulated samples (40 min) in the mean dynamics of each cluster.

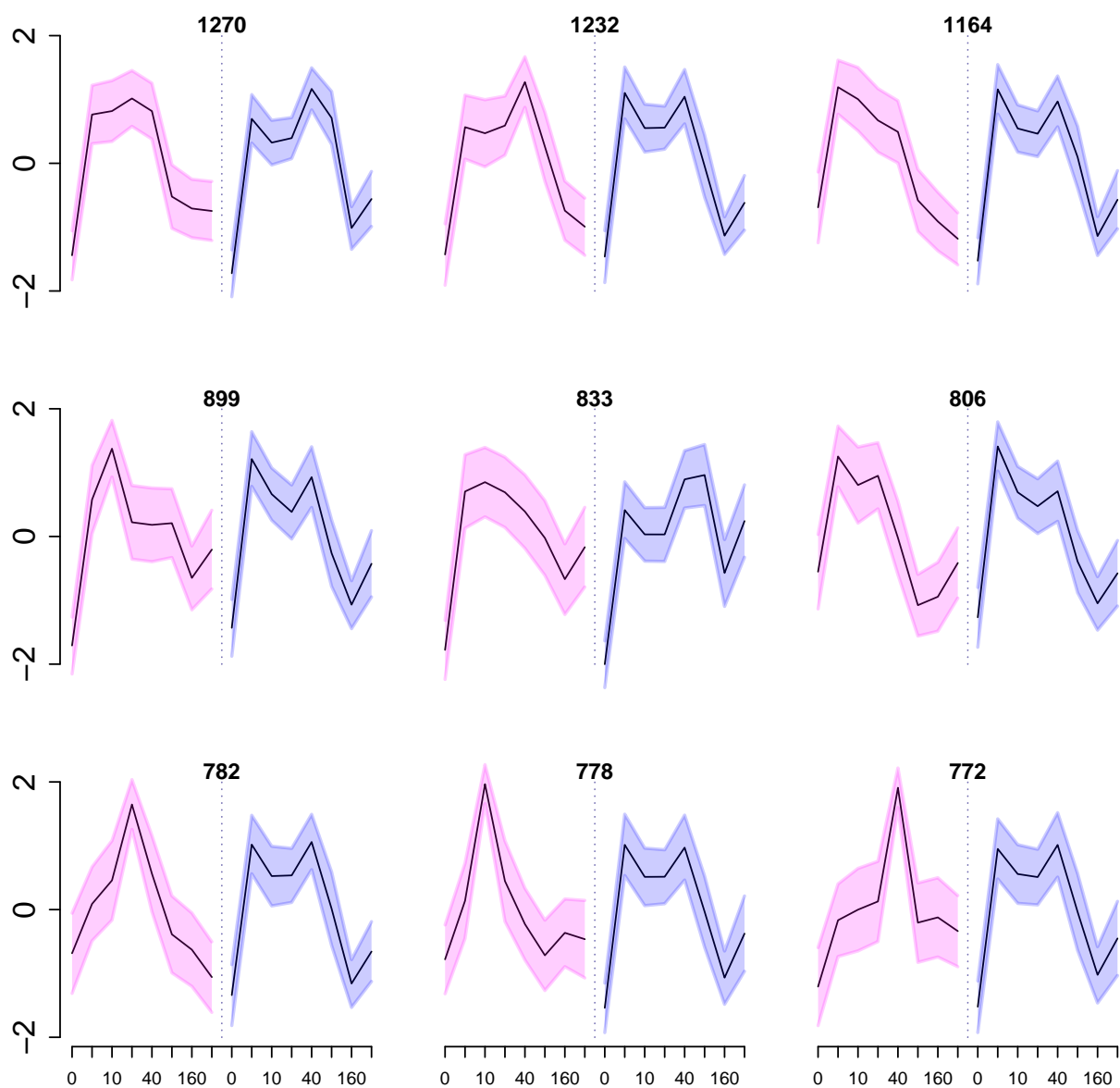


Figure A.4: The figure shows the clustering of the joint time course of Pol II and ER- $\alpha$  at enhancers with Affinity Propagation. The clustering involves only the time series which individually possess a sum of at least 100 tags across all time point.

	CTCF	ESR1	FoxA1	SRC-3	STAG1	TRIM24	c-Fos	c-Jun	c-MYC	Count	Amplitude
1	220	775	503	1033	655	257	415	94	11	1270	225
2	309	683	297	799	650	226	246	68	9	1232	270
3	217	582	440	862	632	212	299	79	11	1164	118
4	362	379	197	506	549	183	201	47	3	899	188
5	181	428	410	645	477	168	364	101	8	833	216
6	179	281	269	536	410	117	153	42	4	806	53
7	105	303	160	466	260	68	106	17	3	782	124
8	142	319	213	482	302	97	112	29	2	778	55
9	139	320	145	355	244	78	88	22	6	772	311
10	140	279	219	465	312	88	150	31	5	719	107
11	109	296	221	456	294	83	152	25	3	713	238
12	130	227	222	408	294	87	140	26	4	634	109
13	143	247	155	360	271	101	139	24	3	611	210
14	307	212	84	213	390	121	72	37	15	603	193
15	52	136	255	435	172	50	181	21	4	587	278
16	39	87	35	67	51	183	14	13	1	580	-239
17	93	198	135	285	169	45	78	13	3	552	195
18	57	92	233	379	173	38	111	12	3	551	92
19	99	178	114	281	181	53	71	13	2	526	147
20	38	142	174	298	134	39	76	15	1	499	132
21	43	52	257	356	150	19	171	9	2	491	134
22	80	193	126	253	159	50	57	12	1	489	138
23	274	112	90	236	365	127	93	51	12	486	180
24	47	145	142	252	112	25	46	7	1	466	16
25	42	59	203	271	117	20	171	12	3	449	135
26	58	113	171	246	174	46	76	14	2	443	-147
27	27	61	186	257	106	10	121	8	3	423	144
28	28	81	216	261	136	32	127	17	4	419	-69
29	119	64	58	140	144	136	37	30	0	403	133
30	78	72	132	208	157	59	152	28	8	400	159
31	54	42	168	221	134	36	129	16	1	364	112
32	61	44	174	208	123	41	126	17	5	351	66
33	10	44	159	215	87	18	74	5	3	346	-88
34	43	83	83	139	90	44	34	6	1	343	-54
35	57	104	74	150	93	31	35	2	2	335	-56
36	48	89	70	108	72	21	35	6	3	326	18
37	30	92	39	88	53	31	17	3	1	317	-147
38	44	39	26	55	40	123	12	15	0	305	-187
39	65	44	112	194	139	34	82	10	2	293	174
40	31	22	21	136	41	58	32	23	1	227	-19
41	107	36	38	98	131	57	66	48	16	209	195
42	3	8	40	89	18	19	31	15	0	207	-120

<span style="background-color: #c8e6c9; border: 1px solid black; display: inline-block; width: 15px; height: 10px;"></span> enriched, $p < 0.01$
<span style="background-color: #a5d6a7; border: 1px solid black; display: inline-block; width: 15px; height: 10px;"></span> enriched, $p < 0.005$
<span style="background-color: #81c784; border: 1px solid black; display: inline-block; width: 15px; height: 10px;"></span> enriched, $p < 0.001$
<span style="background-color: #ffcdd2; border: 1px solid black; display: inline-block; width: 15px; height: 10px;"></span> depleted, $p < 0.01$
<span style="background-color: #ffb74d; border: 1px solid black; display: inline-block; width: 15px; height: 10px;"></span> depleted, $p < 0.005$
<span style="background-color: #ff8a65; border: 1px solid black; display: inline-block; width: 15px; height: 10px;"></span> depleted, $p < 0.001$
<span style="background-color: #bdbdbd; border: 1px solid black; display: inline-block; width: 15px; height: 10px;"></span> neither

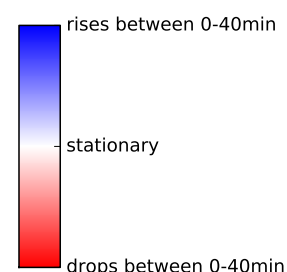


Table A.5: The table shows the cluster-specific patterns of TF bindings across enhancers for the corresponding joint Pol II and ER- $\alpha$  clusters in figure A.4. The count column indicates the size of the cluster. The amplitude column shows the difference between Pol II occupancies in E2-deprived (0 min) and E2-stimulated samples (40 min) in the mean Pol II dynamics of each cluster.

	CTCF	ESR1	FoxA1	SRC-3	STAG1	TRIM24	c-Fos	c-Jun	c-MYC	Count	Amplitude
1	220	775	503	1033	655	257	415	94	11	1270	288
2	309	683	297	799	650	226	246	68	9	1232	250
3	217	582	440	862	632	212	299	79	11	1164	249
4	362	379	197	506	549	183	201	47	3	899	236
5	181	428	410	645	477	168	364	101	8	833	289
6	179	281	269	536	410	117	153	42	4	806	197
7	105	303	160	466	260	68	106	17	3	782	239
8	142	319	213	482	302	97	112	29	2	778	251
9	139	320	145	355	244	78	88	22	6	772	253
10	140	279	219	465	312	88	150	31	5	719	261
11	109	296	221	456	294	83	152	25	3	713	263
12	130	227	222	408	294	87	140	26	4	634	250
13	143	247	155	360	271	101	139	24	3	611	247
14	307	212	84	213	390	121	72	37	15	603	230
15	52	136	255	435	172	50	181	21	4	587	237
16	39	87	35	67	51	183	14	13	1	580	-88
17	93	198	135	285	169	45	78	13	3	552	249
18	57	92	233	379	173	38	111	12	3	551	199
19	99	178	114	281	181	53	71	13	2	526	242
20	38	142	174	298	134	39	76	15	1	499	254
21	43	52	257	356	150	19	171	9	2	491	164
22	80	193	126	253	159	50	57	12	1	489	249
23	274	112	90	236	365	127	93	51	12	486	177
24	47	145	142	252	112	25	46	7	1	466	251
25	42	59	203	271	117	20	171	12	3	449	189
26	58	113	171	246	174	46	76	14	2	443	213
27	27	61	186	257	106	10	121	8	3	423	224
28	28	81	216	261	136	32	127	17	4	419	216
29	119	64	58	140	144	136	37	30	0	403	-112
30	78	72	132	208	157	59	152	28	8	400	207
31	54	42	168	221	134	36	129	16	1	364	176
32	61	44	174	208	123	41	126	17	5	351	145
33	10	44	159	215	87	18	74	5	3	346	176
34	43	83	83	139	90	44	34	6	1	343	225
35	57	104	74	150	93	31	35	2	2	335	233
36	48	89	70	108	72	21	35	6	3	326	224
37	30	92	39	88	53	31	17	3	1	317	220
38	44	39	26	55	40	123	12	15	0	305	-118
39	65	44	112	194	139	34	82	10	2	293	155
40	31	22	21	136	41	58	32	23	1	227	-120
41	107	36	38	98	131	57	66	48	16	209	-130
42	3	8	40	89	18	19	31	15	0	207	-38

enriched, $p < 0.01$
enriched, $p < 0.005$
enriched, $p < 0.001$
depleted, $p < 0.01$
depleted, $p < 0.005$
depleted, $p < 0.001$
neither

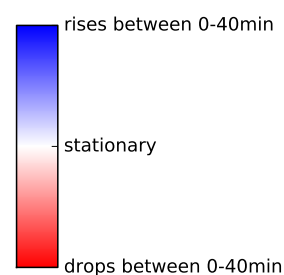


Table A.6: The table shows the cluster-specific patterns of TF bindings across enhancers for the corresponding joint Pol2 and ER- $\alpha$  clusters in figure A.4. The count column indicates the size of the cluster. The amplitude column shows the difference between ER- $\alpha$  occupancies in E2-deprived (0 min) and E2-stimulated samples (40 min) in the mean ER- $\alpha$  dynamics of each cluster.

# **Appendix B**

## **Supplementary figures for Chapter 4**

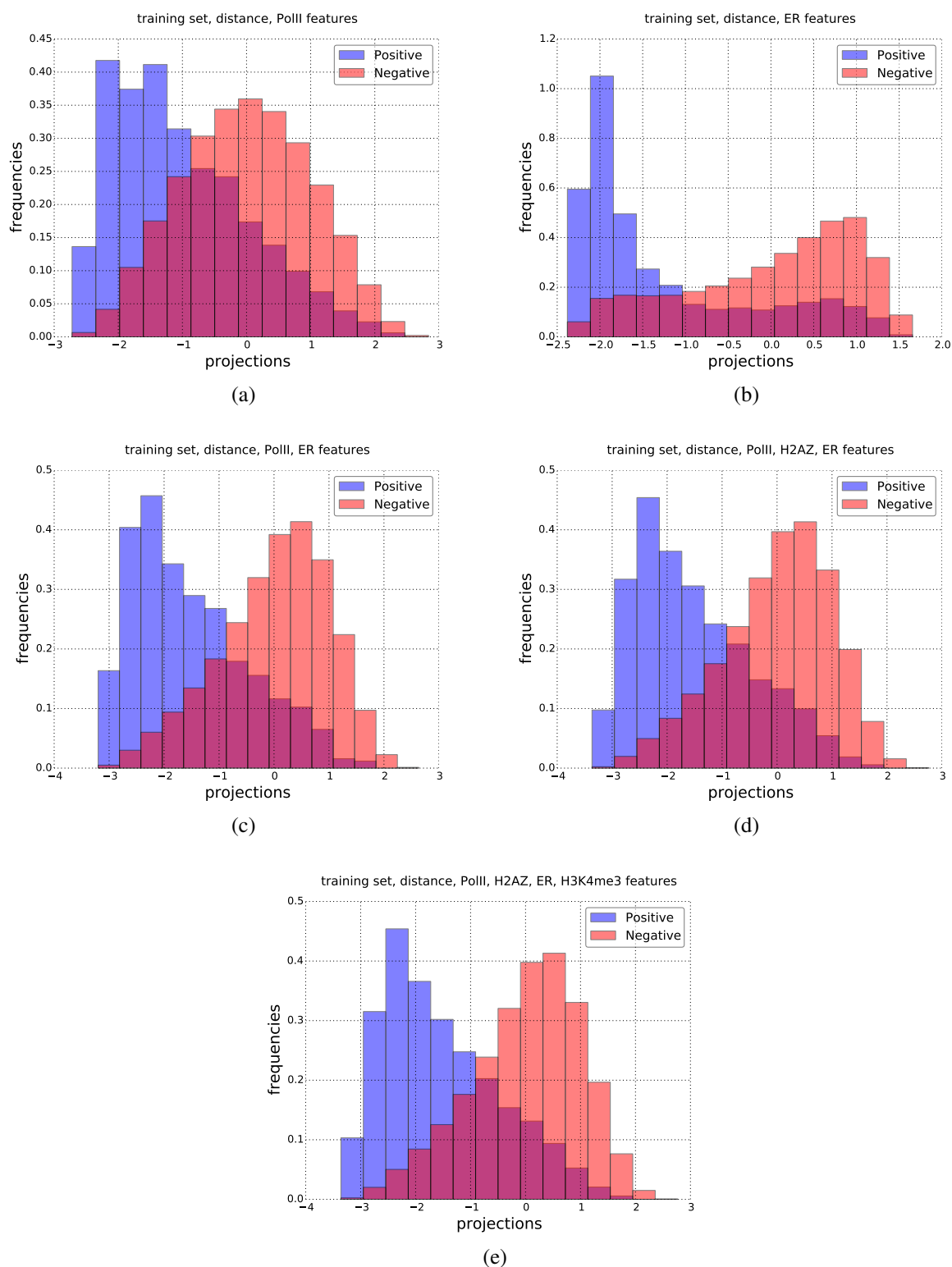


Figure B.1: The graphs show positive (blue) and negative (red) class size-normalised histograms of projections of the training set of Fisher's linear discriminant analysis when applied on five selected variants of vectors of features (see sub-titles).



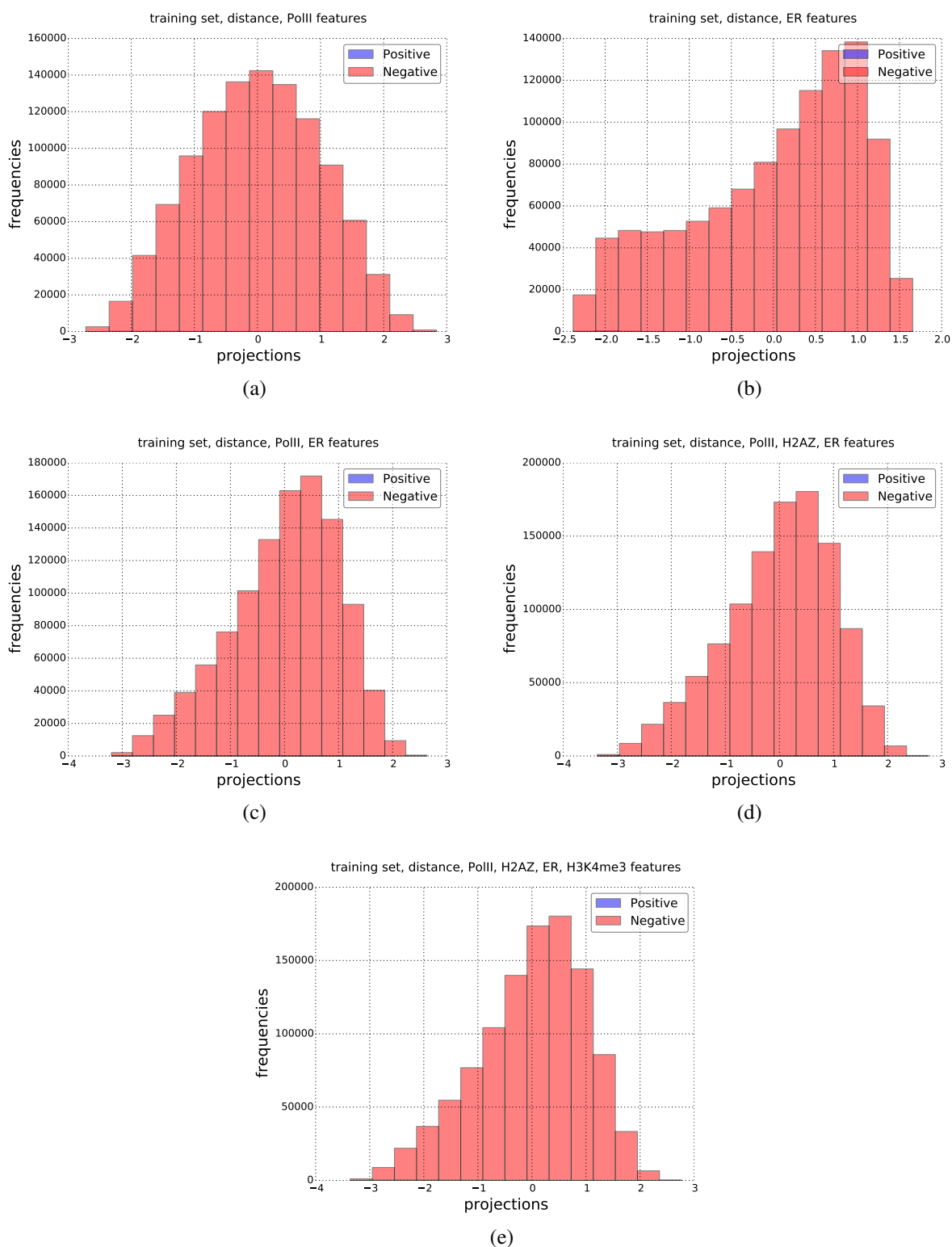


Figure B.2: The graphs show positive (blue) and negative (red) histograms of projections of the training set of Fisher's linear discriminant analysis when applied on five selected variants of vectors of features (see sub-titles). Due to a large difference in sizes of the two classes, the histograms of the positives are invisible.

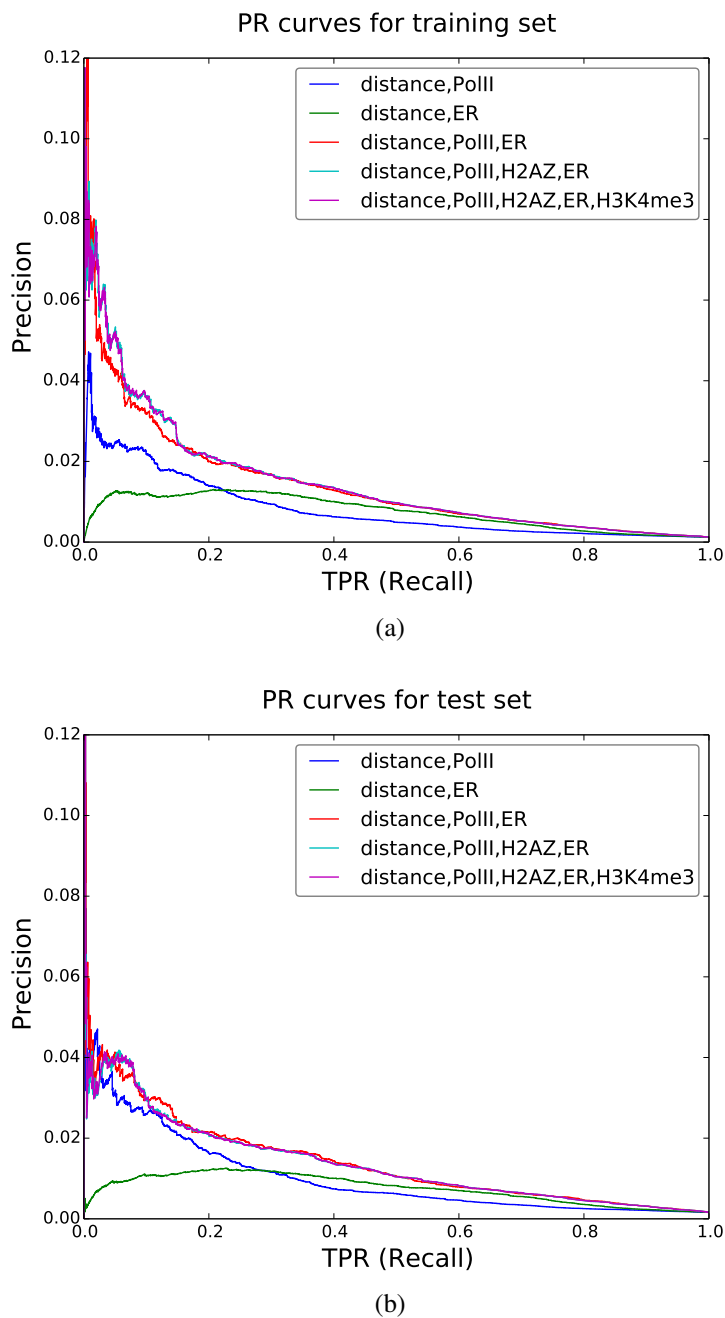


Figure B.3: Precision-Recall curves measuring performance of Fisher's linear discriminant analysis for increasing values of cut-off levels (i.e. from the most negative to the most positive value of projection) for (a) training and (b) test sets. Each colour shows performance of the classifier for a different selected combination of features.

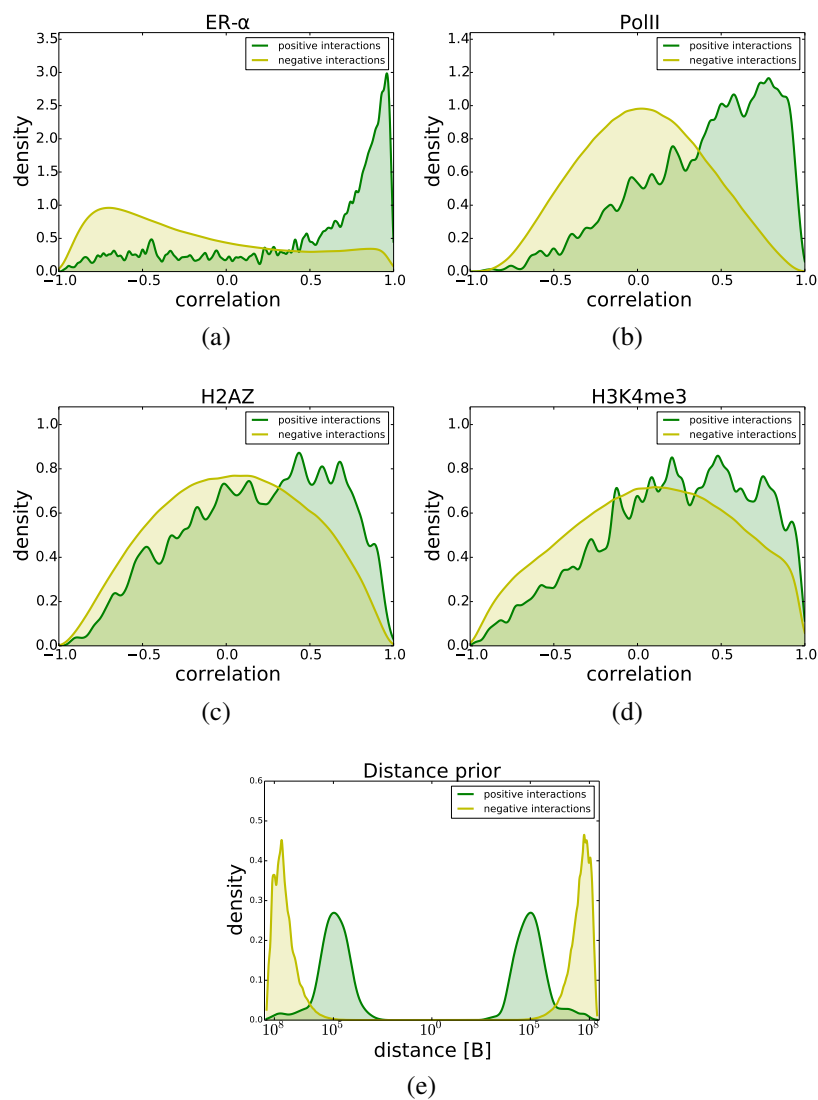


Figure B.4: The graphs (a, b, c, d) show positive (green) and negative (yellow) distributions of correlations between pairs of time series of 300bp-upstream-extended-gene regions and enhancer bodies for ER- $\alpha$ , PolII, H2AZ and H3K4me3 collected across all 24 chromosomes. The figure (e) shows the distribution of genomic distances between centres of distal enhancers and 300bp-upstream-shifted-TSS of genes. The set of positive and negative pairs was constructed using 300bp-upstream-extended-genes and distal enhancers.

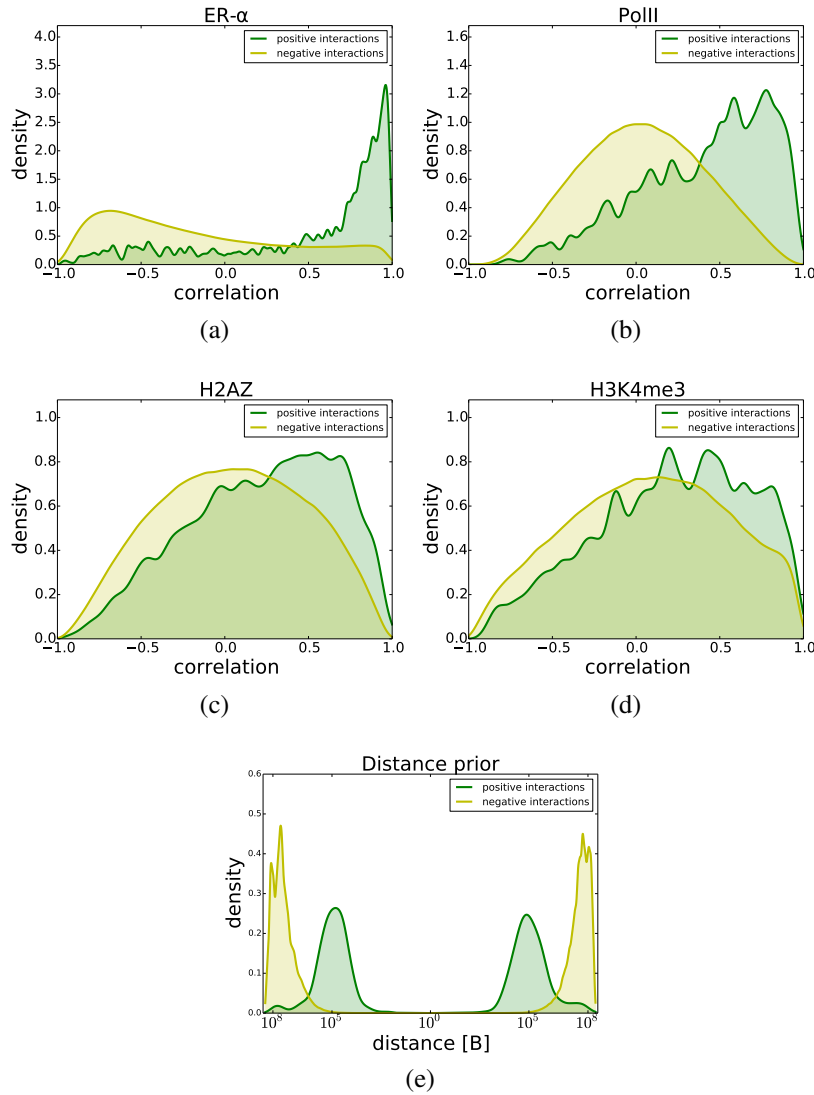


Figure B.5: The graphs (a, b, c, d) show positive (green) and negative (yellow) distributions of correlations between time series of 300bp-upstream-extended-gene regions and enhancer bodies for ER- $\alpha$ , PolII, H2AZ and H3K4me3 collected across all odd chromosomes (training data). The figure (e) shows the distribution of genomic distances between centres of distal enhancers and 1500bp-upstream-shifted-TSS of genes. The set of positive and negative pairs was constructed using 1500bp-upstream-extended-genes and distal enhancers.

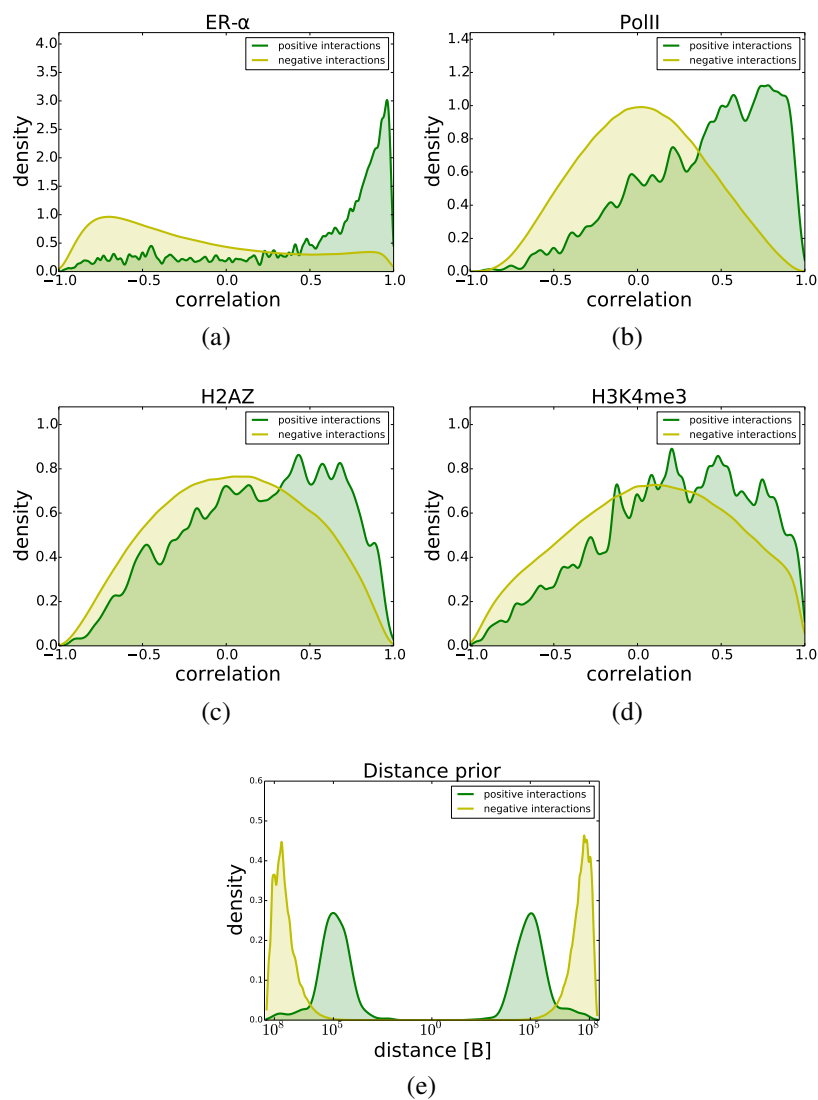


Figure B.6: The graphs (a, b, c, d) show positive (green) and negative (yellow) distributions of correlations between time series of 300bp-upstream-extended-gene regions and enhancer bodies for ER- $\alpha$ , PolII, H2AZ and H3K4me3 collected across all 24 chromosomes. The figure (e) shows the distribution of genomic distances between centres of distal enhancers and 1500bp-upstream-shifted-TSS of genes. The set of positive and negative pairs was constructed using 1500bp-upstream-extended-genes and distal enhancers.

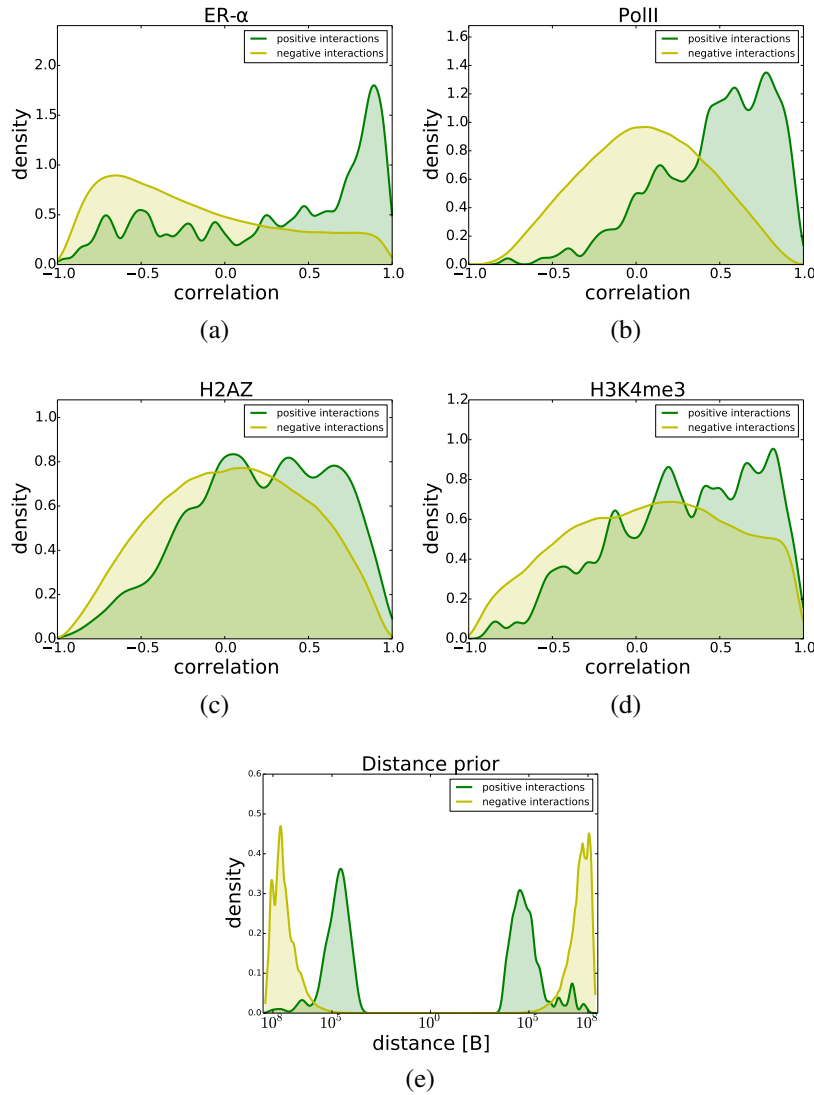


Figure B.7: The graphs (a, b, c, d) show positive (green) and negative (yellow) distributions of correlations between time series of 300bp-upstream-extended-gene regions and enhancer bodies for ER- $\alpha$ , PolII, H2AZ and H3K4me3 collected across all odd chromosomes. The figure (e) shows the distribution of genomic distances between centres of distal enhancers and 1500bp-upstream-shifted-TSS of genes. The set of positive and negative pairs for the first model was constructed using TSS-centred 3000bp-long regions and distal enhancers.

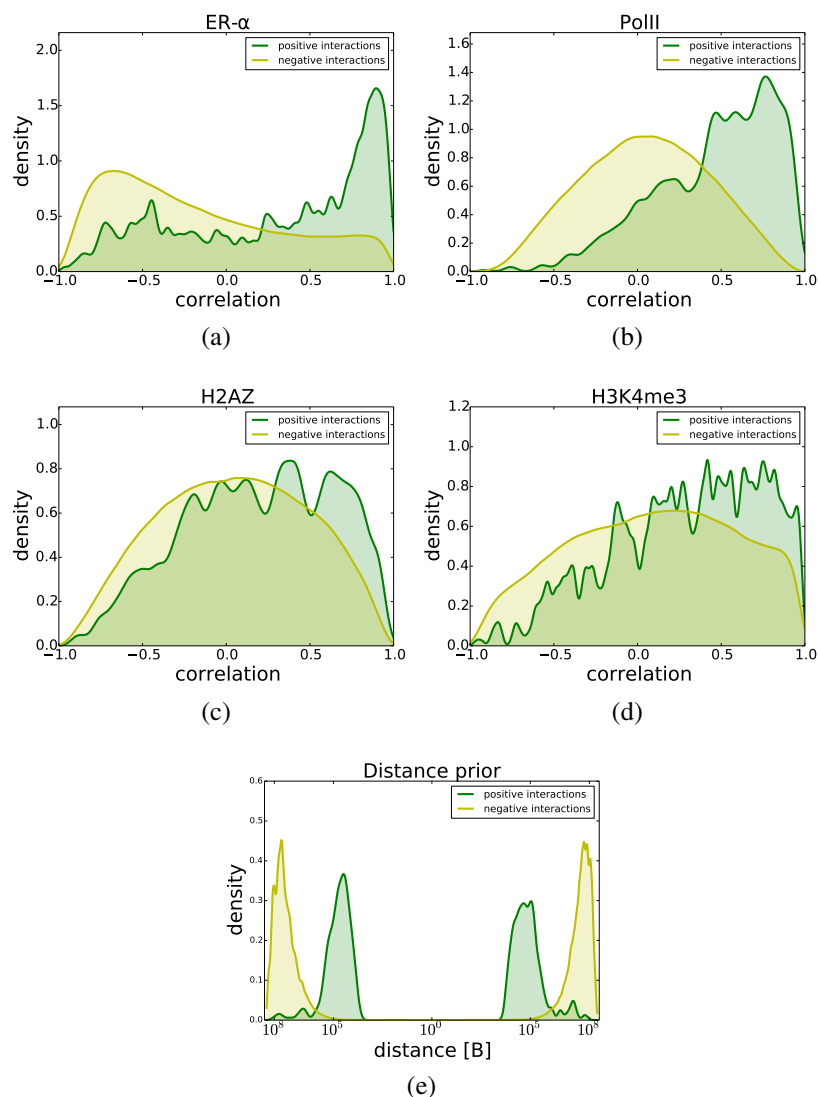
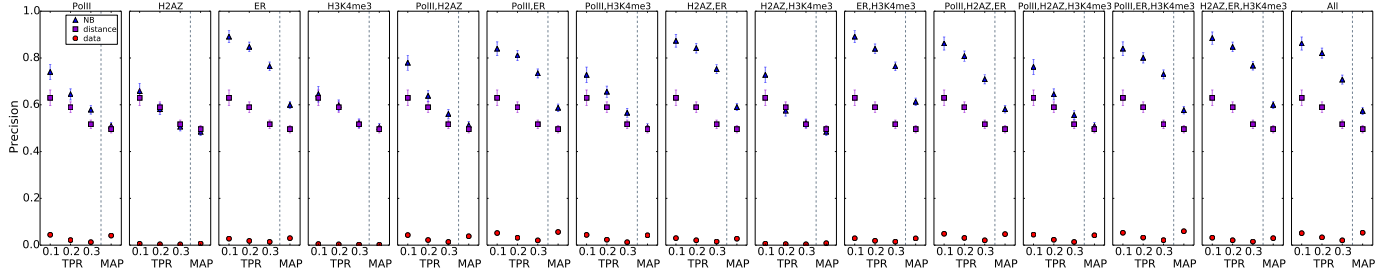
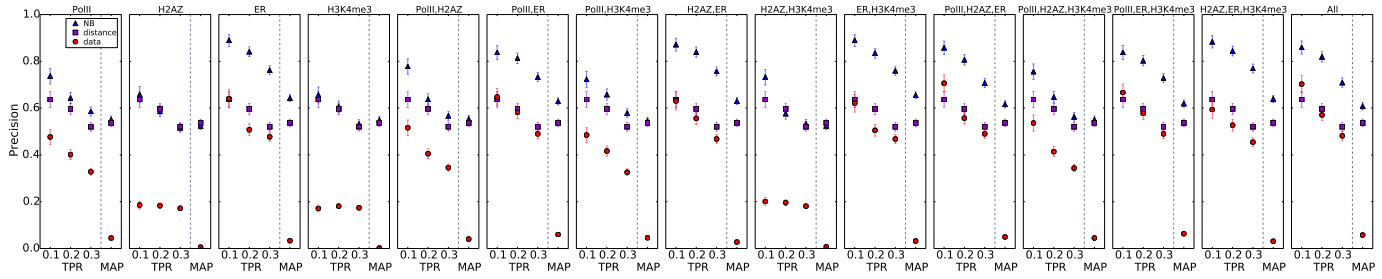


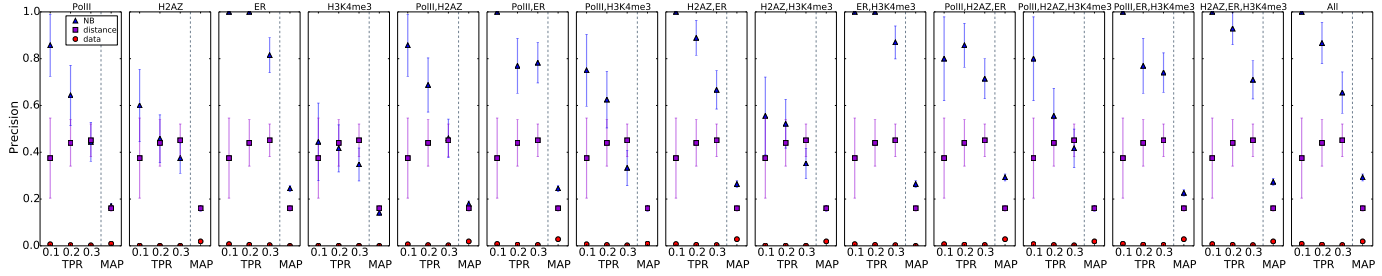
Figure B.8: The graphs (a, b, c, d) show positive (green) and negative (yellow) distributions of correlations between time series of 300bp-upstream-extended-gene regions and enhancer bodies for ER- $\alpha$ , PolII, H2AZ and H3K4me3 collected across all 24 chromosomes. The figure (e) shows the distribution of genomic distances between centres of distal enhancers and 1500bp-upstream-shifted-TSS of genes. The set of positive and negative pairs for the first model was constructed using TSS-centred 3000bp-long regions and distal enhancers.



(a) training data performance



(b) intra-domain performance



(c) inter-domain performance

Figure B.9: Figure shows the comparison of performance of the NB model on odd chromosomes (training data) measured by Precision-TPR and MAP scores. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using 300bp-upstream-extended-genes and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 300bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers.



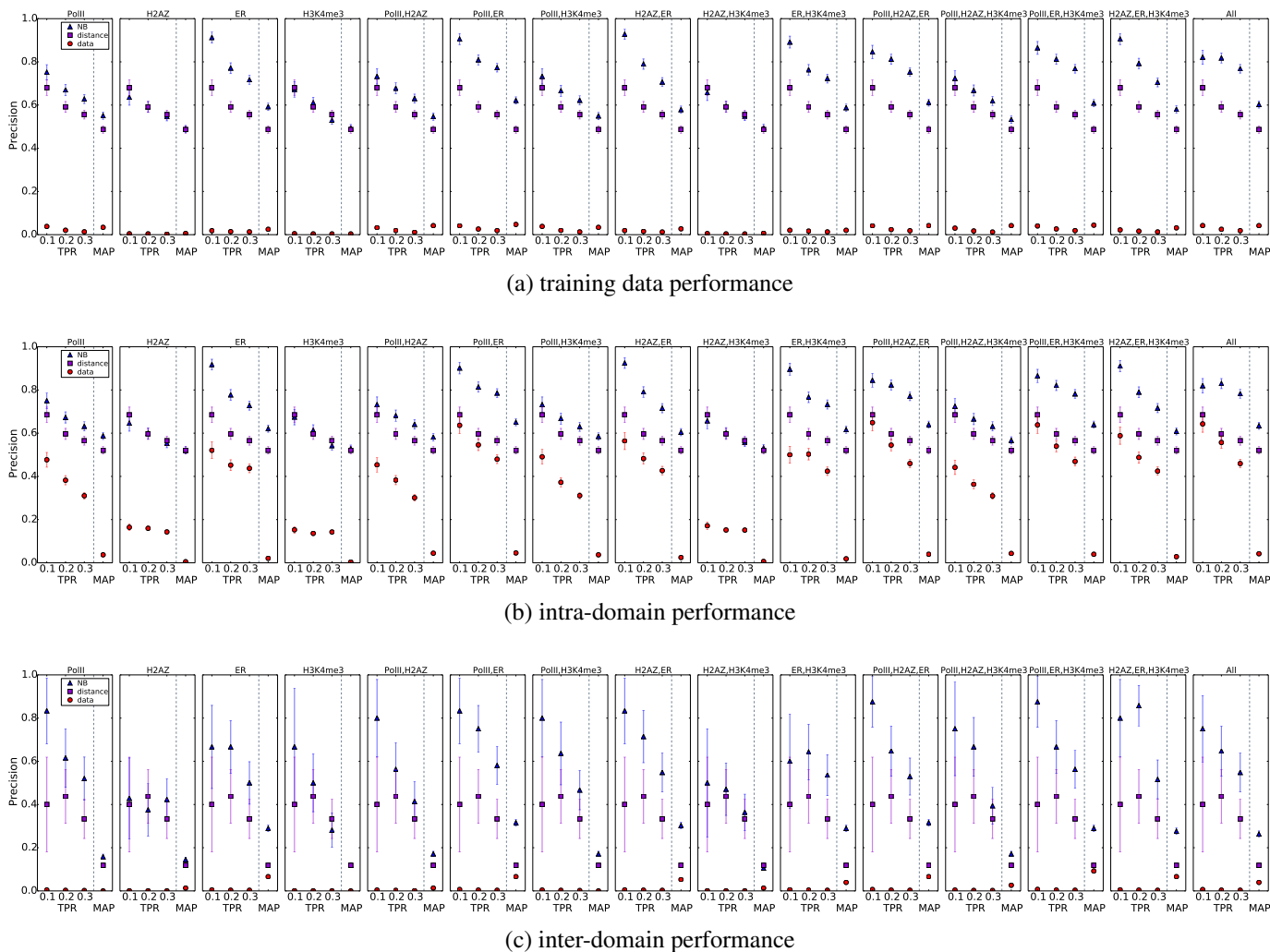


Figure B.10: Figure shows the comparison of performance of the NB model on even chromosomes (test data) measured by Precision-TPR and MAP scores. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using 300bp-upstream-extended-genes and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 300bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers.

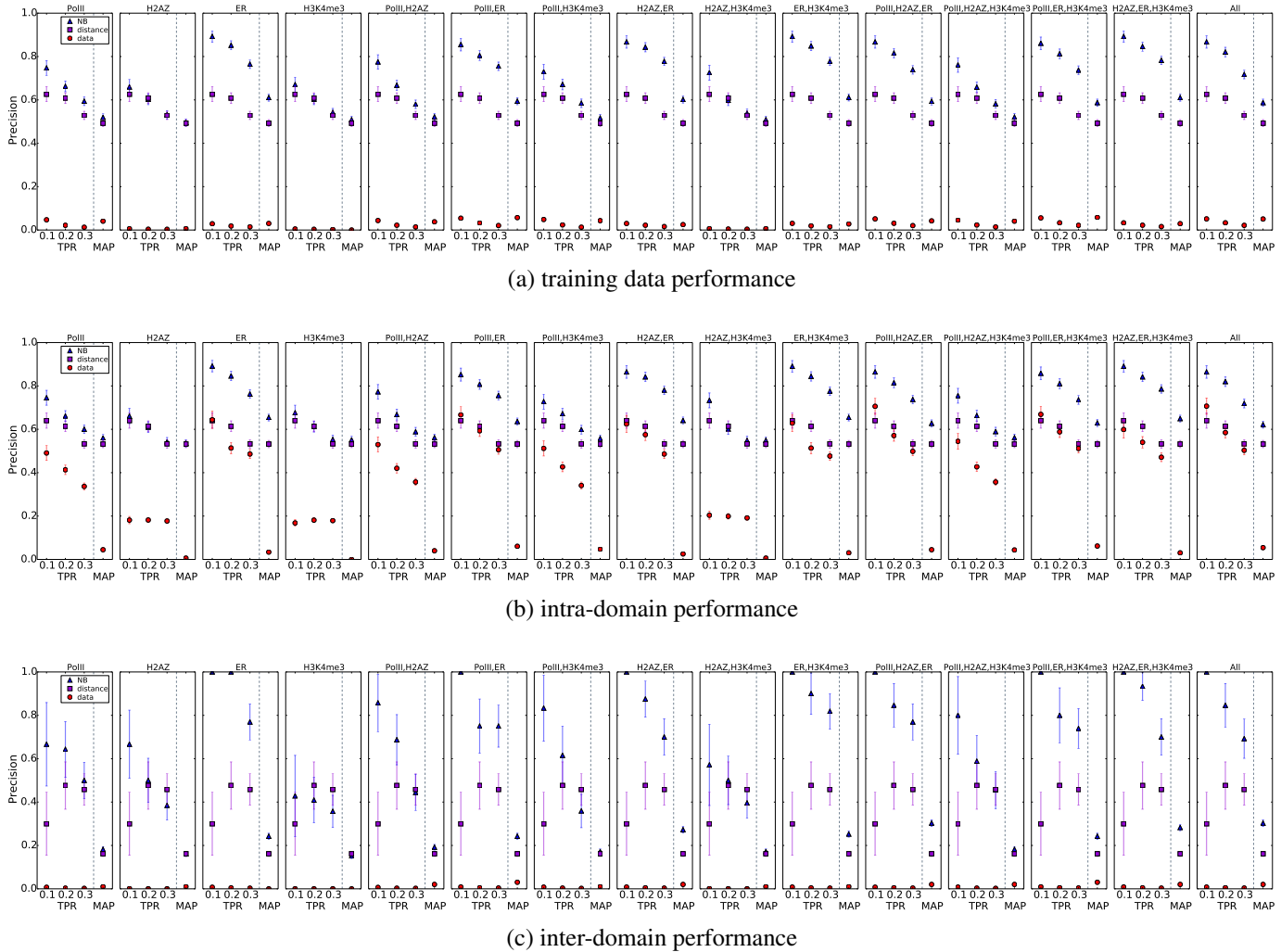


Figure B.11: Figure shows the comparison of performance of the NB model on odd chromosomes (training data) measured by Precision-TPR and MAP scores. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using 1500bp-upstream-extended-genes and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 1500bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers.

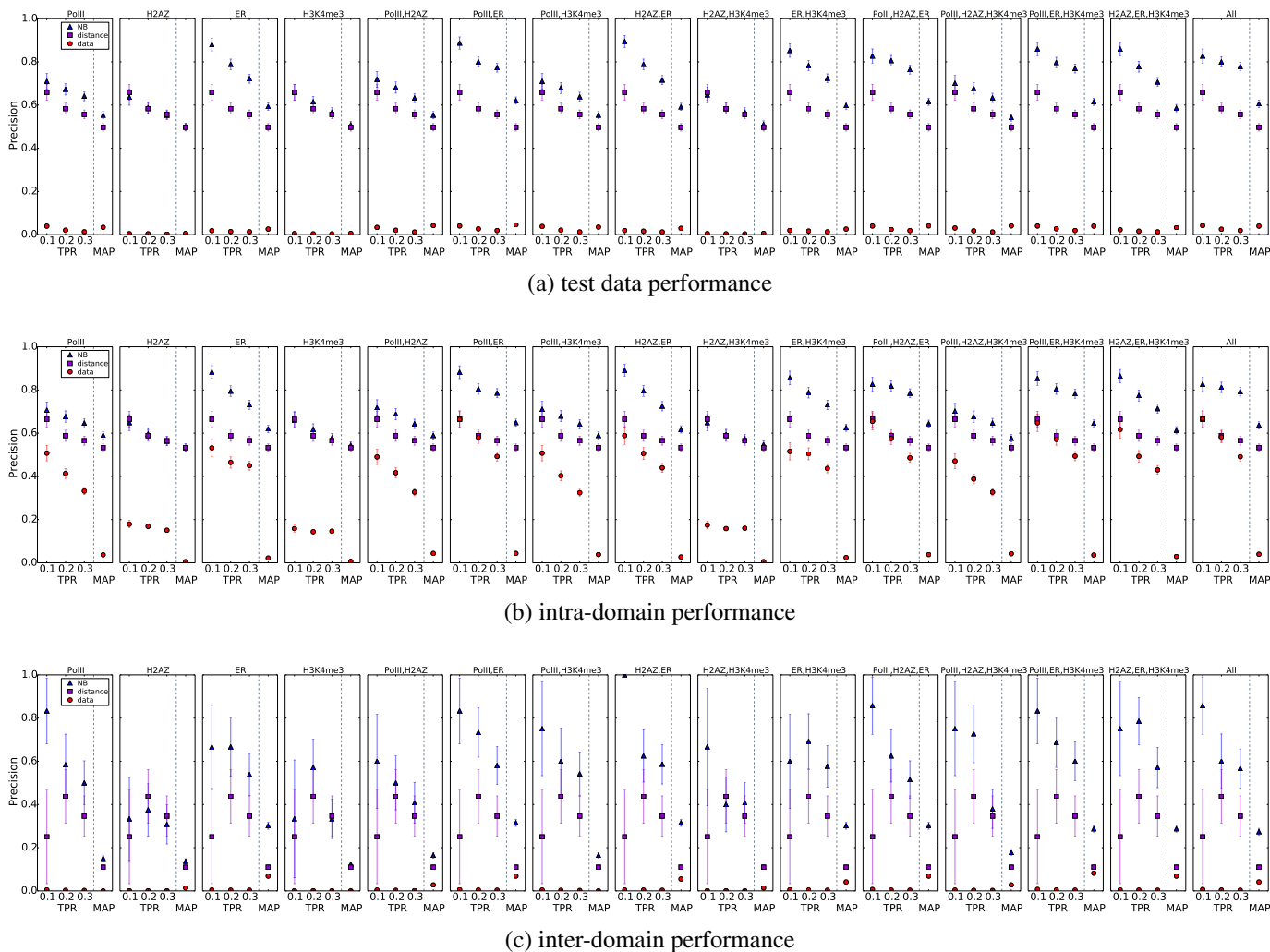


Figure B.12: Figure shows the comparison of performance of the NB model on even chromosomes (test data) measured by Precision-TPR and MAP scores. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using 1500bp-upstream-extended-genes and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 1500bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers.

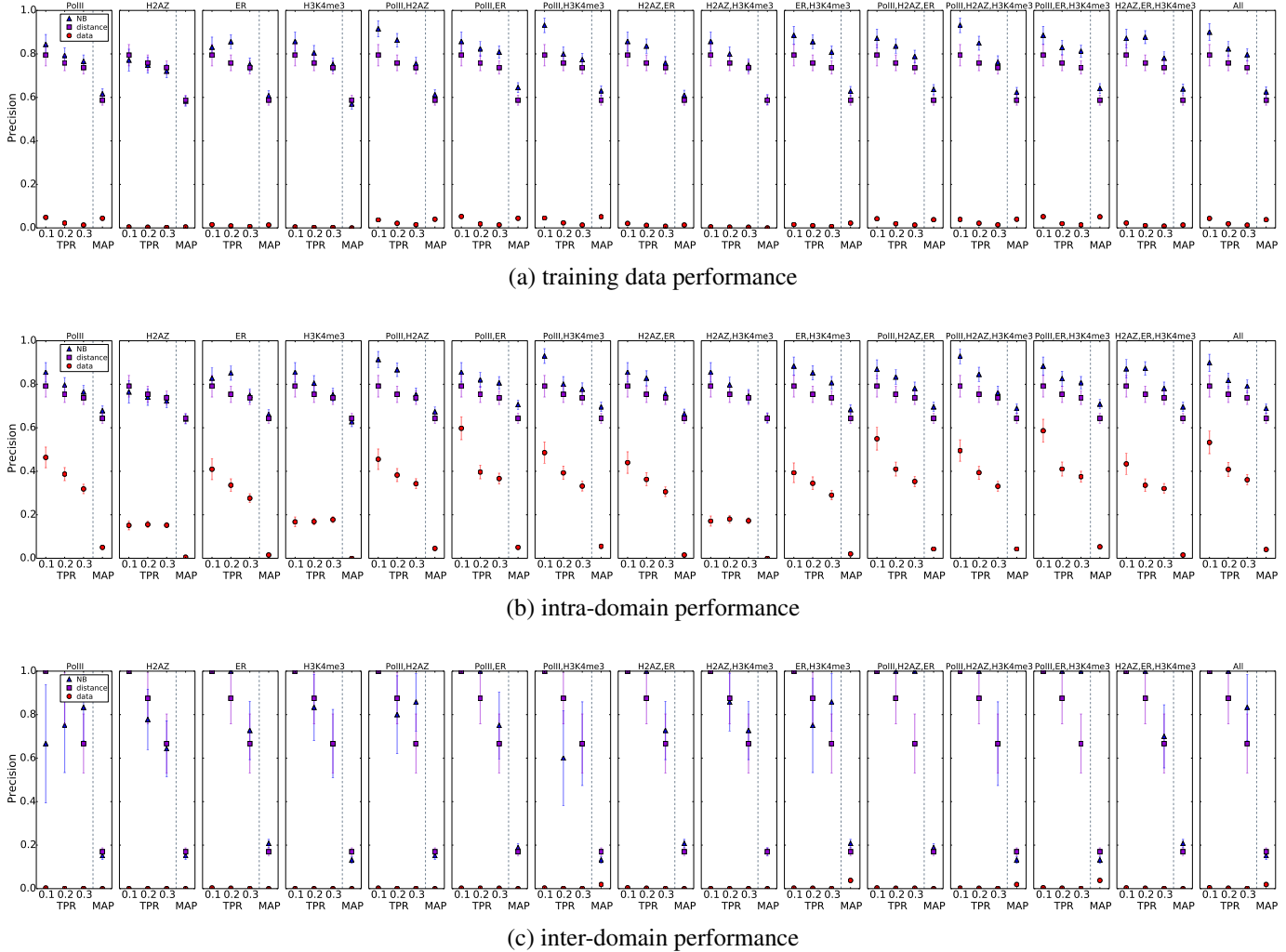


Figure B.13: Figure shows the performance of the TSS-centric NB model on odd chromosomes (training data) measured by Precision-TPR and MAP scores. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using TSS-centred 3000bp-long regions and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 1500bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers.

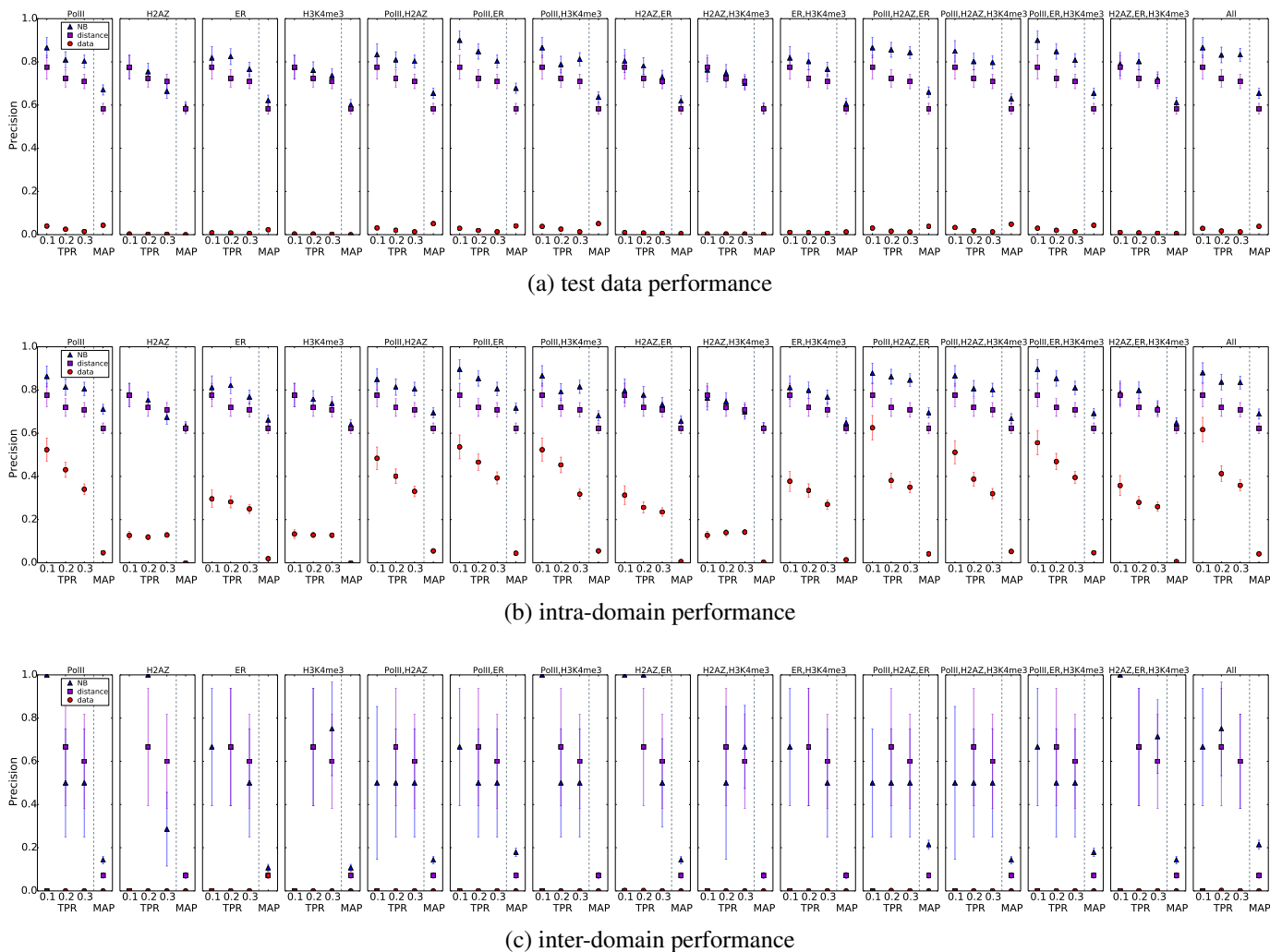


Figure B.14: Figure shows the performance of the TSS-centric NB model on even chromosomes (test data) measured by Precision-TPR and MAP scores. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using TSS-centred 3000bp-long regions and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 1500bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers.

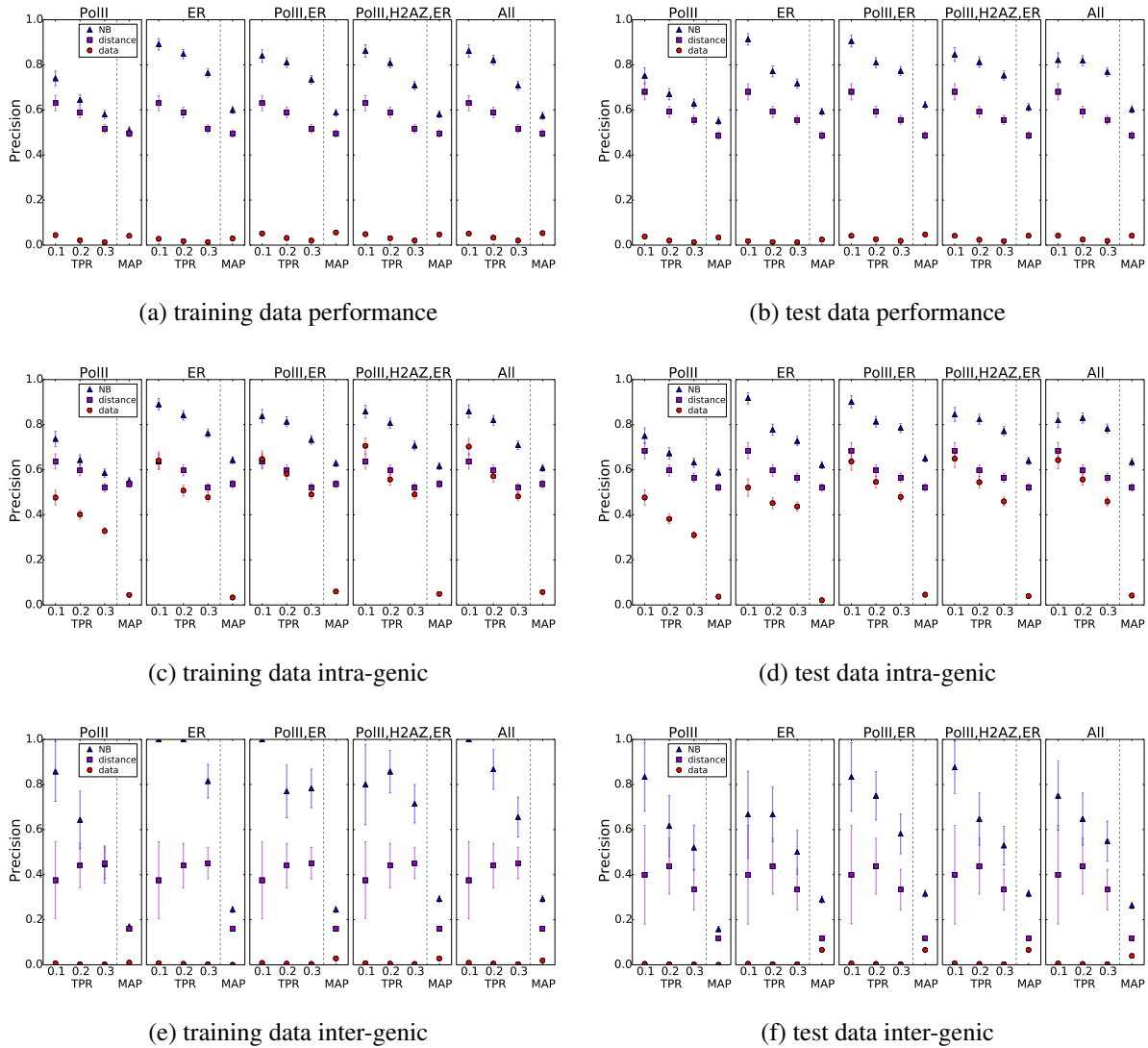


Figure B.15: Figure shows the comparison of performance of the NB model between odd and even chromosomes (training and test data) measured by Precision-TPR and MAP scores for selected combinations of datasets. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using 300bp-upstream-extended-genes and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 300bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers.

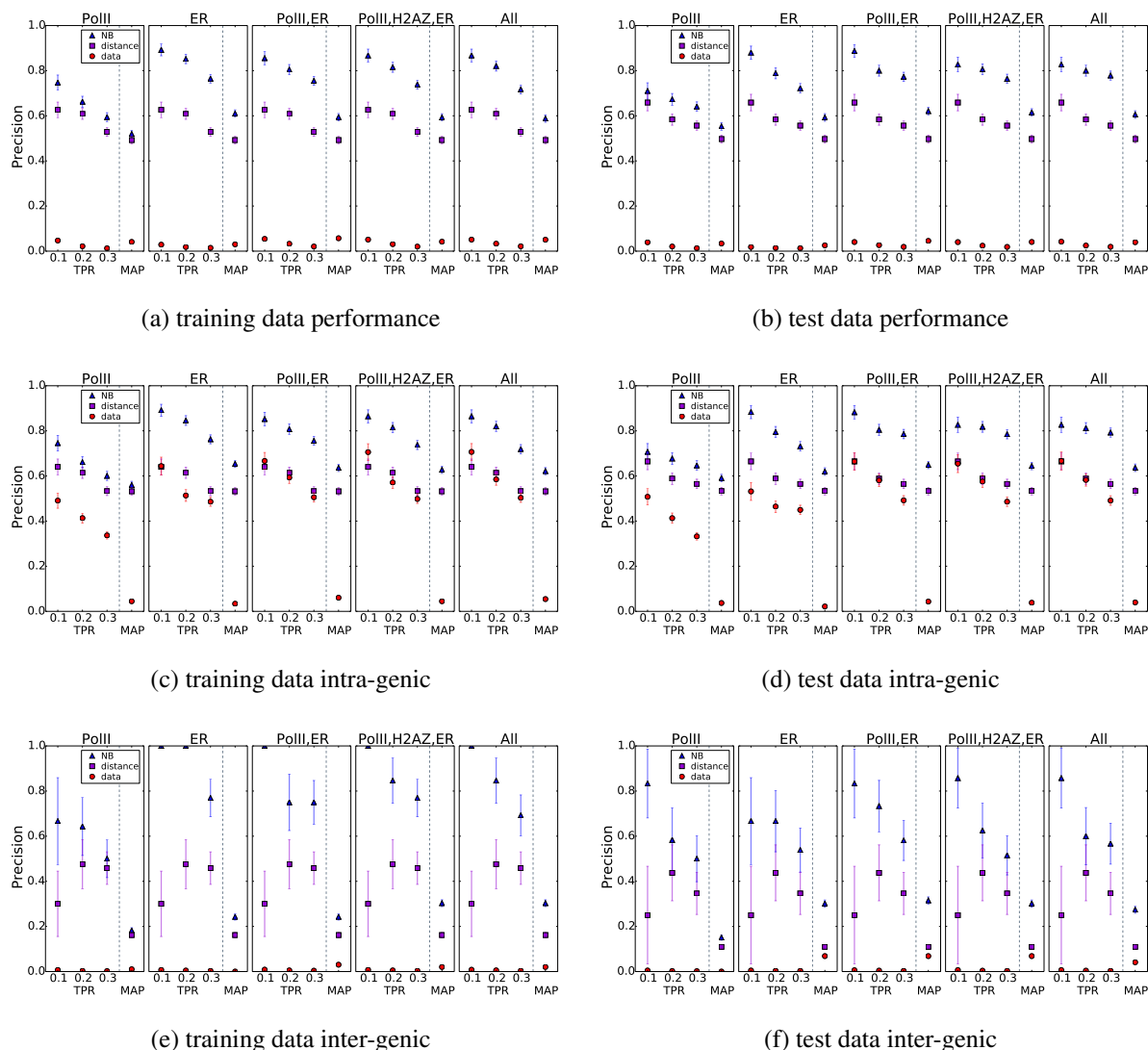


Figure B.16: Figure shows the comparison of performance of the NB model between odd and even chromosomes (training and test data) measured by Precision-TPR and MAP scores for selected combinations of datasets. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using 1500bp-upstream-extended-genes and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 1500bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers.

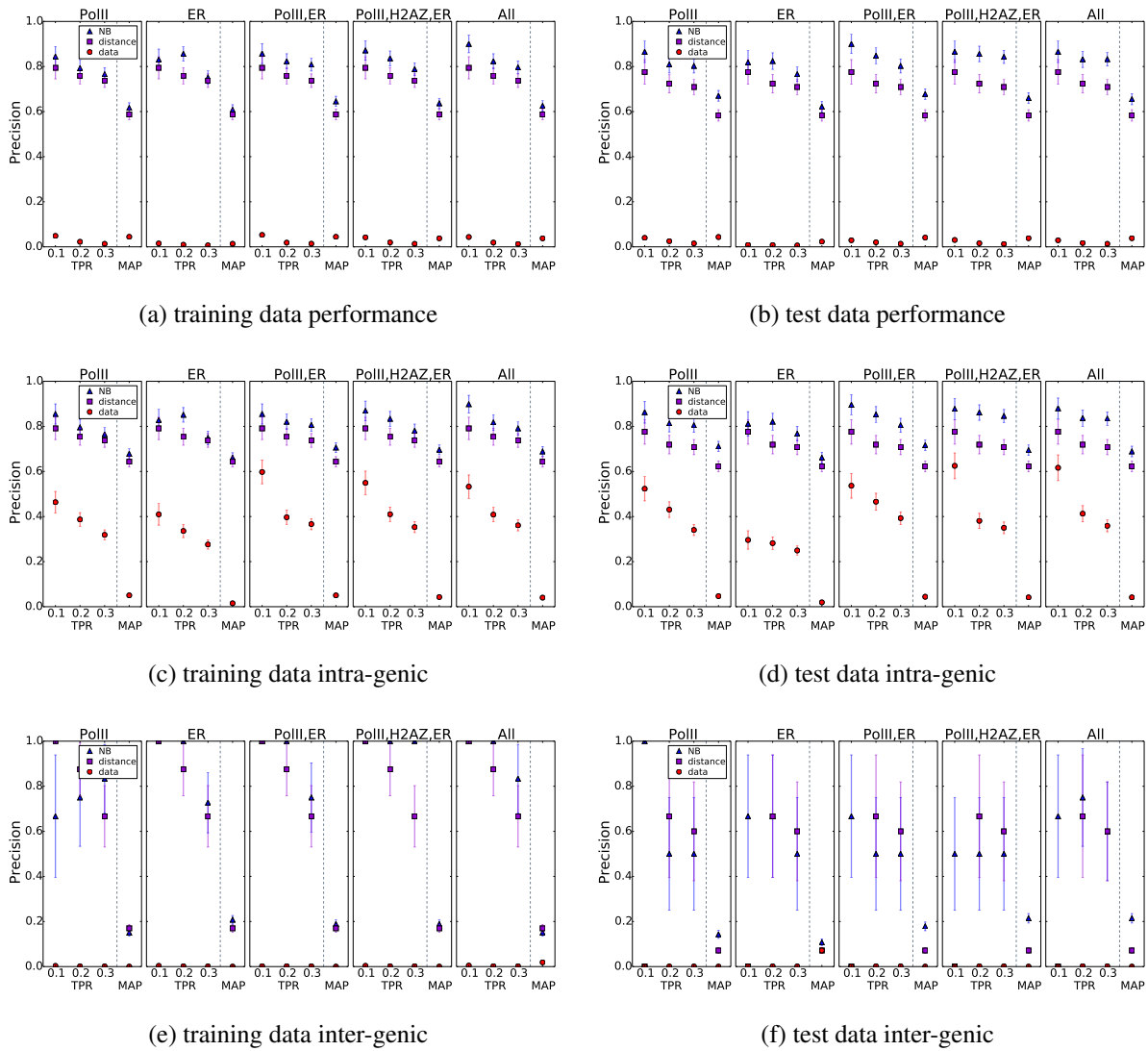


Figure B.17: Figure shows the comparison of performance of the TSS-centric NB model between odd and even chromosomes (training and test data) measured by Precision-TPR and MAP scores for selected combinations of datasets. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using TSS-centred 3000bp-long regions and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 1500bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers.



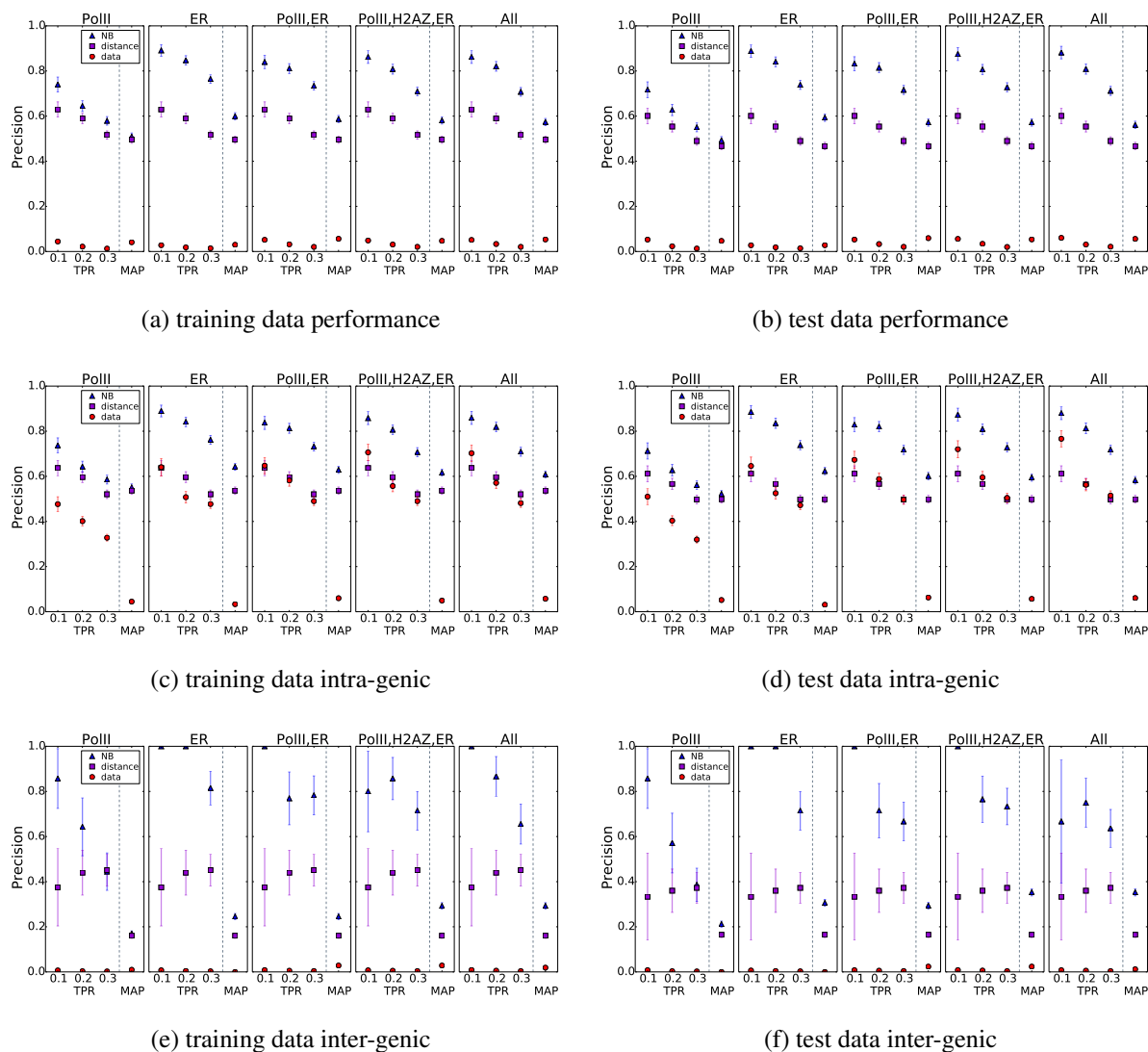


Figure B.18: Figure shows the performance of the NB model of training data and for selected combinations of datasets under two different parametrisations of MACS peak-calling. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The first column of the figure shows the performance of the NB model trained on the stringent time persistent merged MACS-called peaks (i.e. distal ER- $\alpha$  bindings) from the scan with the  $p$ -value of  $1e-11$  and the local control switched off, in which case the search is done with  $\lambda_{BG}$ . In the second column we see the performance under the alternative peak calling with the  $p$ -value of  $1e-05$  (MACS' default), no control and the local control flag on. The set of positive and negative pairs for the first model was constructed using 1500bp-upstream-extended-genes and distal enhancers. The correlation-based attributes of the model were estimated using pairs of 300bp-upstream-extended-genes, and enhancers (merged distal MACS-called peaks). The separation-based from 1500bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers.

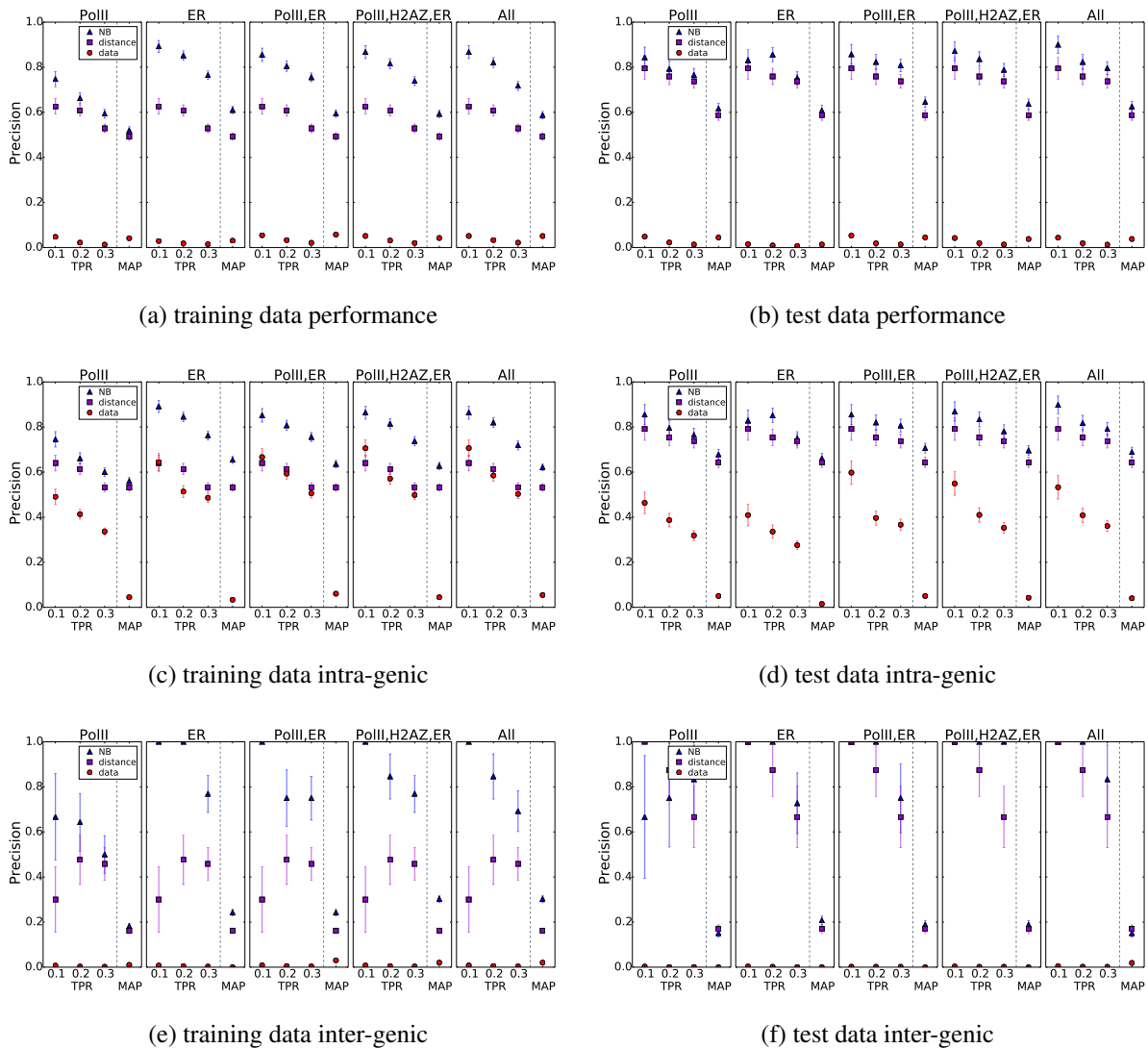


Figure B.19: Figure shows the comparison of the performance between promoter-extended-gene and TSS-centric models on odd chromosomes (training data) measured by Precision-TPR and MAP scores and for selected datasets. The Precision-TPR curves show the accuracy of the predictions with the highest 10%, 20%, 30% scores i.e. posterior probabilities. The second and the third rows stratify predictions at each of the thresholds into those which take place within domains and those involving inter-domain contacts. The set of positive and negative pairs for the first model was constructed using 1500bp-upstream-extended-genes and distal enhancers, whereas for the second using TSS-centred 3000bp-long regions and distal enhancers. The correlation-based attributes of the two models were estimated using signals (time series) aggregated over 300bp-upstream-extended-genes, and distal enhancer bodies. For separation-based from 1500bp-upstream-shifted TSS to the centres of the ER- $\alpha$  enhancers.

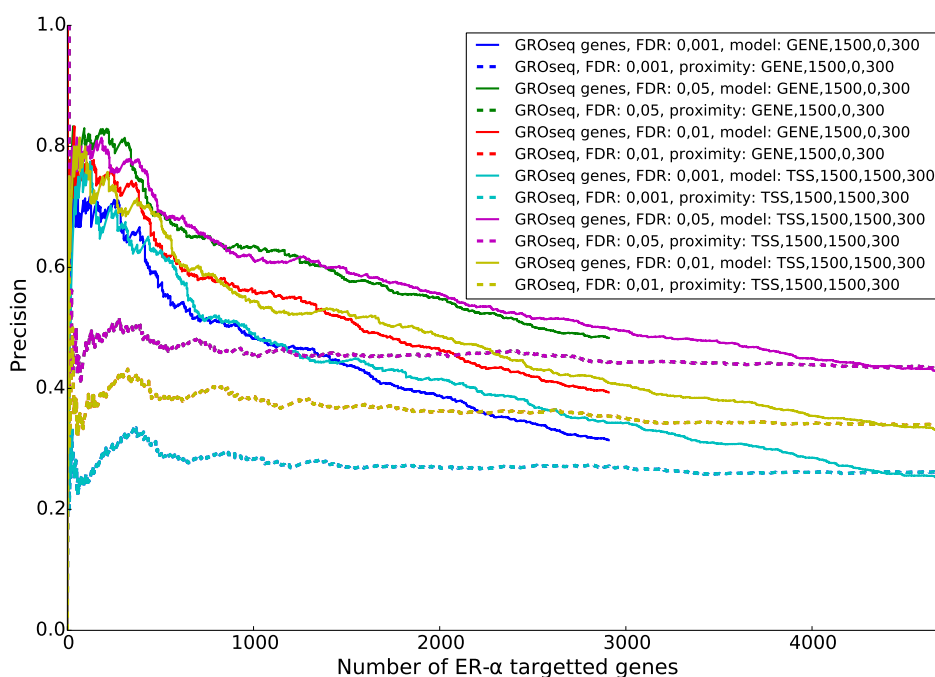


Figure B.20: The PR curves show the comparison of the performance between the 1500bp-upstream-extended-gene and 1500bp-TSS-centric Naive Bayes models on their ability to predict differentially expressed genes. The positive set consists of GROseq-detected genes for 3 different confidence levels of 0.001, 0.01, 0.05. Score of each tested gene is a cumulative posterior probability of the set of its NB-predicted regulators. The graph also shows the accuracy of using an absolute value of the separation (proximity) between its closest ER- $\alpha$  binding and a canonical TSS of a gene as a predictor of gene activity. Gene was regarded as ER- $\alpha$  regulated if the distance to its putative regulator was within 40kB.

## Significant GO terms for diseases

Name	pValue	ER- $\alpha$ Genes	Genes in Annot.	Prop.
Adenoid Cystic Carcinoma	7.20E-08	108	185	58%
Fibroid Tumor	3.19E-10	164	287	57%
Invasive breast carcinoma	3.82E-12	209	369	57%
Noninfiltrating Intraductal Carcinoma	1.58E-09	159	281	57%
Carcinoma, Papillary	1.05E-06	105	186	56%
Ductal Carcinoma	1.76E-06	109	196	56%

Table B.1: Significant GO terms for drugs

Name	pValue	ER- $\alpha$ Genes	Genes in Annot.	Prop.
Irinotecan; MCF7;	9.41E-29	136	174	78%
Verteporfin; MCF7;	4.76E-26	125	161	78%
Retinoic acidMCF7;	1.81E-23	121	160	76%
Colcemid; MCF7;	5.28E-25	140	190	74%
Afimoxifene	2.18E-37	313	478	65%

Table B.2: The tables show the gene ontology annotations for predicted ER- $\alpha$  regulated genes and three categories: biological process, disease, and drugs. The genes consist of NB predicted targets of ER- $\alpha$  distal enhancers (enhancer-gene links with FDA of 0.25) and the genes with an intra-genic ER- $\alpha$  binding.

## Significant GO terms for biological processes

Name	pValue	ER- $\alpha$	in Annot	prop
Regulation of Ras protein signal transduction	5.48E-13	135	200	68%
Cell-cell junction organization	3.72E-14	154	230	67%
Epithelial cell development	7.57E-15	167	251	67%
Ras protein signal transduction	1.151E-16	216	334	65%
Regulation of small GTPase mediated signal transduction	8.84E-14	188	295	64%
Small GTPase mediated signal transduction	5.742E-18	352	589	60%
Positive regulation of GTPase activity	4.854E-18	374	632	59%
Regulation of GTPase activity	1.685E-19	407	688	59%
Regulation of cell morphogenesis	2.21E-15	354	610	58%
Blood vessel development	7.91E-16	377	653	58%
Circulatory system development	3.735E-25	613	1062	58%
Cardiovascular system development	3.735E-25	613	1062	58%
Vasculature development	5.43E-16	391	680	58%
Blood vessel morphogenesis	7.88E-13	316	552	57%
Tissue morphogenesis	3.347E-17	438	766	57%
Morphogenesis of an epithelium	7.38E-14	350	613	57%
Cellular response to growth factor stimulus	1.20E-13	364	643	57%
Response to growth factor	4.24E-14	380	672	57%
Regulation of cell development	2.011E-17	548	989	55%
Epithelial cell differentiation	4.24E-12	371	670	55%
Cell migration	6.435E-22	718	1302	55%
Epithelium development	1.405E-21	719	1307	55%
Cell motility	7.426E-23	785	1433	55%
Cell morphogenesis	2.952E-21	734	1341	55%
Cellular response to endogenous stimulus	6.260E-20	699	1280	55%
Response to hormone	2.22E-13	552	1035	53%
Regulation of cell differentiation	4.680E-20	902	1706	53%

Table B.3: The tables show the gene ontology annotations for predicted ER- $\alpha$  regulated genes and three categories: biological process, disease, and drugs. The genes consist of NB predicted targets of ER- $\alpha$  distal enhancers (enhancer-gene links with FDA of 0.25) and the genes with an intra-genic ER- $\alpha$  binding.

# **Appendix C**

## **Supplementary figures for Chapter 5**

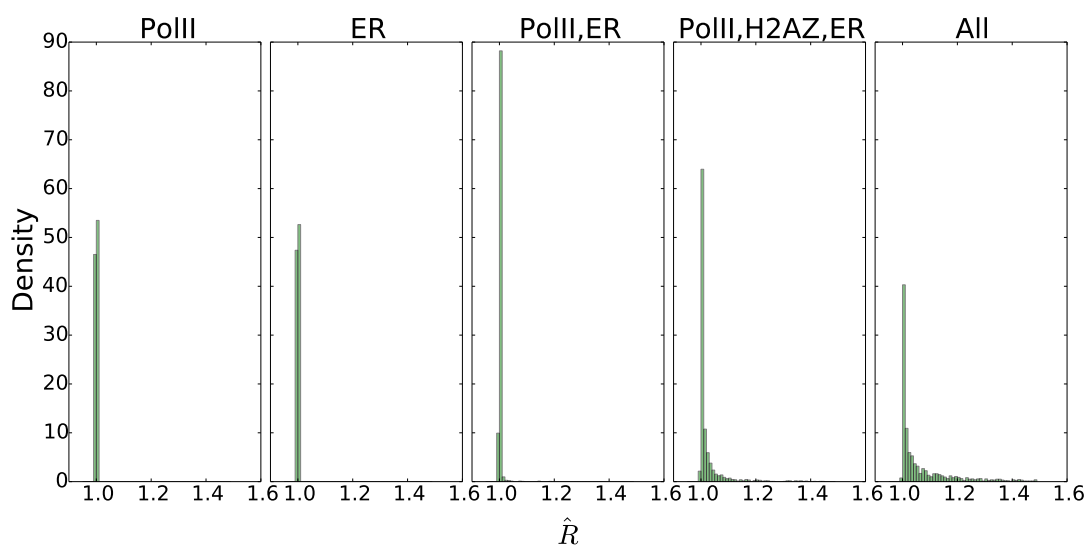


Figure C.1: Gelman  $\hat{R}$  statistics for the samples of  $\mathbf{Z}$  generated by the Gibbs Sampler of the LVA model inferred from the time series of the enhancers with ChIA-PET evidence and parameters  $\kappa_0 = 1$ ,  $\alpha_0 = 2$ ,  $\beta_0 = 2$ .

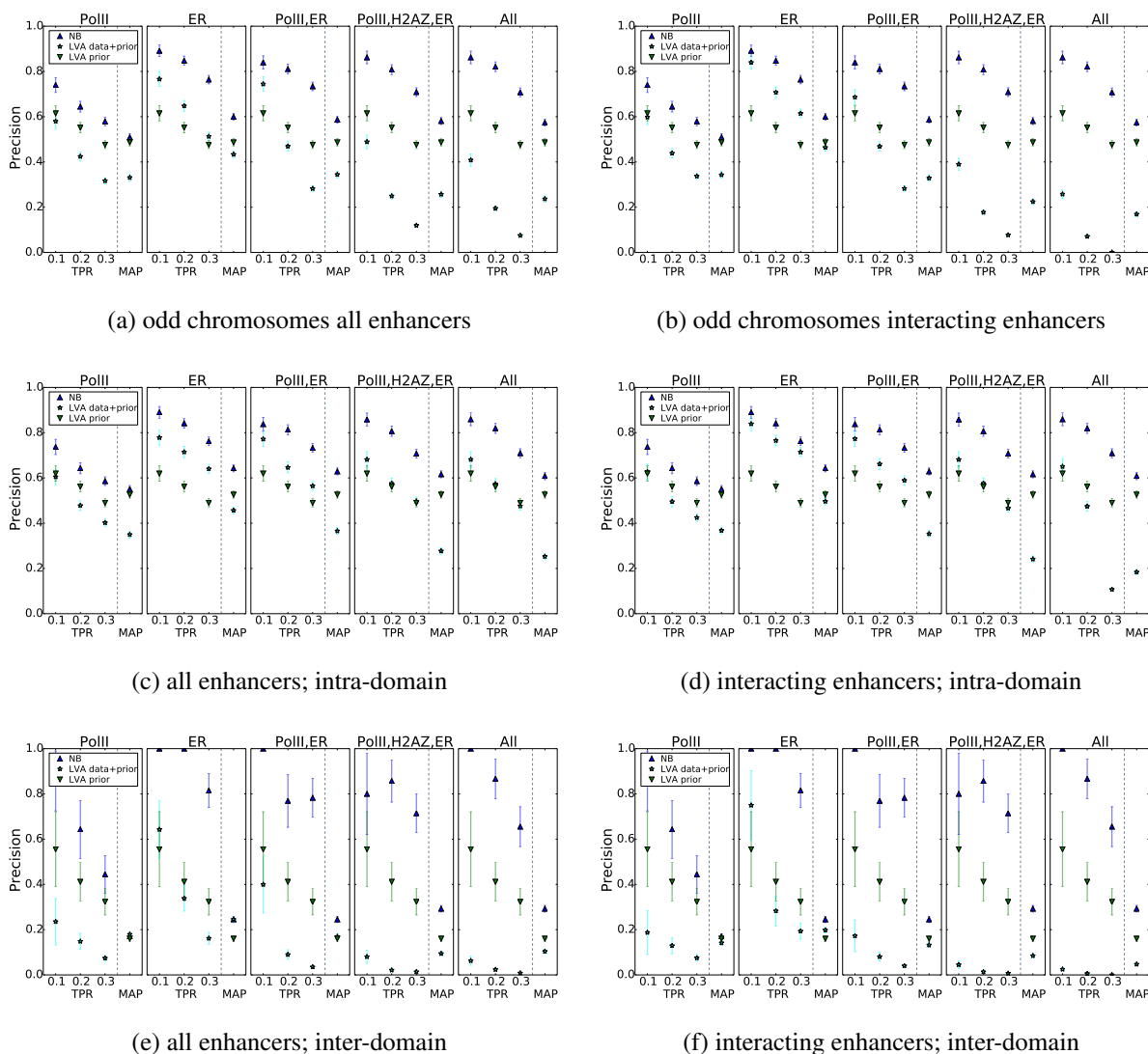


Figure C.2: Figure shows the comparison of the performance between NB and two LVA models on discovery of ChIA-PET-detected links from all odd chromosomes. LVA in the first column was inferred from time course data of all enhancers. LVA in the second column was inferred from the time course data of only those enhancers with ChIA-PET-confirmed links.



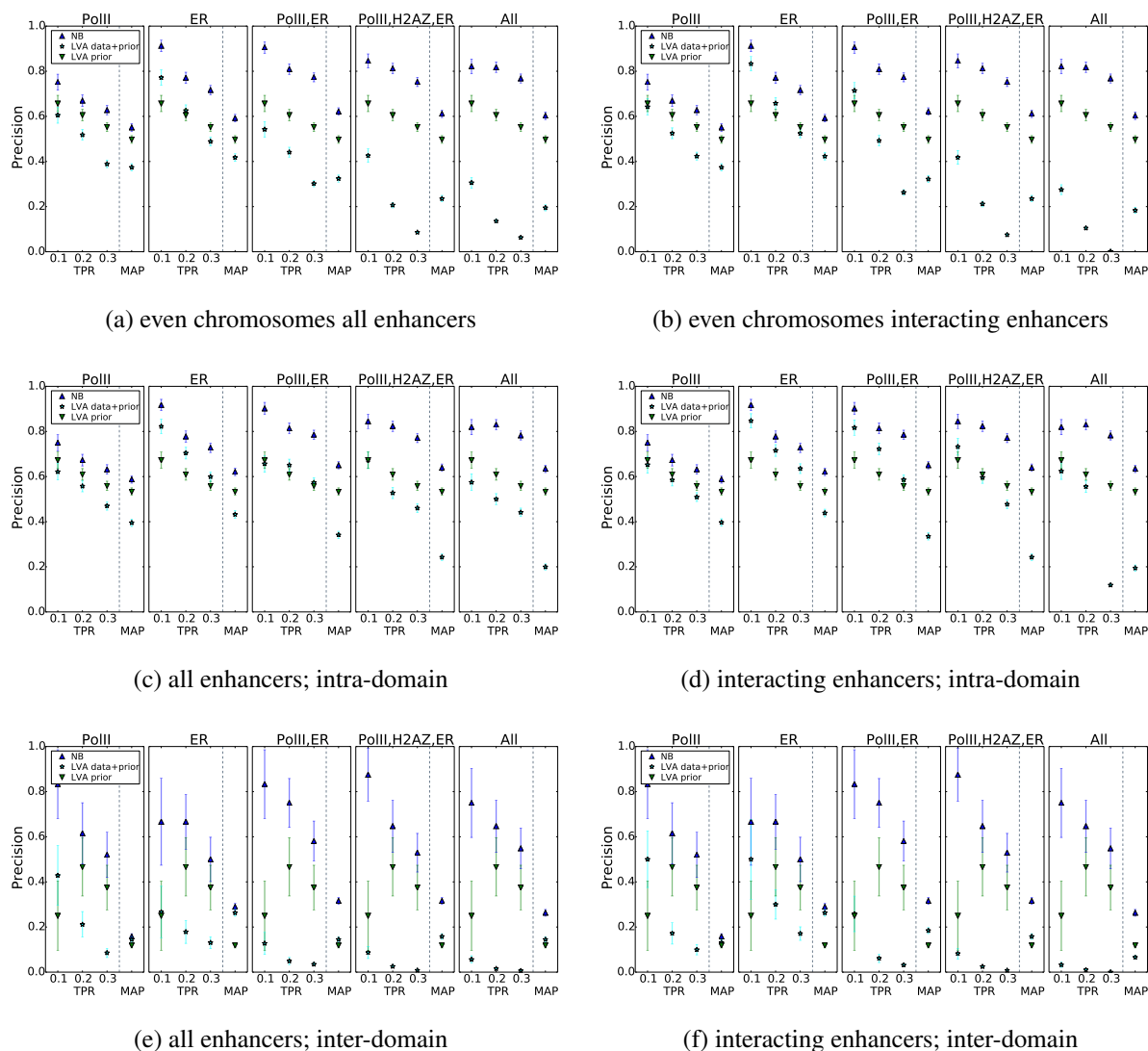


Figure C.3: Figure shows the comparison of the performance between NB and two LVA models on discovery of ChIA-PET-detected links from all even chromosomes. LVA in the first column was inferred from time course data of all enhancers. LVA in the second column was inferred from the time course data of only those enhancers with ChIA-PET-confirmed links.