Organisation of Transcriptomes

Searching for Regulatory DNA Elements Involved in the Correlated Expression of Genomic Neighbours

Master Thesis
in Bioinformatics

carried out at the
Max Planck Institute for Molecular Genetics
Sperling Group (Cardiovascular Genetics)
Department Lehrach (Vertebrate Genomics)
Berlin, Germany

submitted to the
Freie Universität Berlin
Department of Mathematics and Computer Science
Berlin, Germany

by

Markus Schüler

Berlin, November 2006

Acknowledgement

First and foremost I am very grateful for the advice and support of my supervisor, Dr. Silke Sperling, for her inspiration, her feedback and for keeping me focussed on my research. It was a great pleasure for me to conduct this thesis under her supervision.

I would especially like to thank Jenny Fisher, who did a wonderful and doubtlessly nerve-wracking job reading this thesis, as well as Tammo Krüger, who was the calm *in* the storm. Both have greatly contributed to improve this work and it was really exciting to spend this time with them.

In addition I would like to thank all my colleagues from the "Cardiovascular Genetics" Group of the Max Planck Institute for Molecular Genetics. Working together with them has not only been a good learning experience, but also a great pleasure.

Lastly, and most importantly, I thank those who loved and supported me and encouraged me to become a researcher.

Contents

List of Figures **IV**

List of Tables **V**

1. Introduction **1**

Evidence for Clusters of Co-expressed Genes Throughout the Genome . . . 2
1.1. Finding Clusters of Correlated Genes . . . 2
1.1.1. Clusters of Co-expressed Genes . . . 2
1.1.2. Clusters of Co-functional Genes . . . 2
1.1.3. Spatial Organisation Versus Clustering . . . 3
1.2. Our Previous Results in Investigating Highly Co-expressed Genomic Neighbours . . . 3

Levels of Eukaryotic Genome Regulation . . . 4
1.3. The Sequence Level . . . 4
1.4. The Chromatin Level . . . 5
1.4.1. ATP-dependent Chromatin Remodeling Complexes Increase the Mobility and Fluidity of Nucleosomes . . . 5
1.4.2. Replacement of Core Histones by Special Histone Variants . 6
1.4.3. Histone Tail Modifications . . . 7
1.5. The Nuclear Level . . . 9
1.6. Links Between the Three Hierarchical Levels . . . 9

"Active Chromatin Hubs" Mediate Correlated Gene Expression . . 9
1.7. Known Gene Clusters Driven by Active Chromatin Hubs . . . 10
1.8. Proposed Genomic Elements of Active Chromatin Hubs . . . 10
1.8.1. Cluster Control Elements . . . 11
1.8.2. Enhancer and Promoter . . . 12

1.8.3. Boundary Elements (Insulators) 12
1.9. Different Models of Active Chromatin Hub Establishment 13

2. Methods **17**

Definition of Sequence Datasets 18
2.1. Definition of Highly Co-expressed and Uncorrelated Gene Pairs 18
2.2. Extraction of Phylogenetically Conserved Pairs 19
2.3. Definition of Positive and Negative Dataset 21
2.3.1. Selection of Pairs for the Positive/Negative Groups 21
2.3.2. Further Preparation of Pairs of the Positive/Negative Group 21
2.3.3. Definition of the final mouse datasets "*2K-2K*" and "*2K-next*" 22
2.3.4. Definition of Orthologous Human Datasets "H2K-2K" and "H2K-next" 23

Retrieve and Process of Sequence Data and Features 24
2.4. Sequence Extraction 24
2.5. Feature Extraction 24
2.5.1. Transcribed Regions and Transcriptional Start Sites 25
2.5.2. Repeats 25
2.5.3. Regions with Regulatory Potential 26
2.5.4. CpG Island/Regions 28
2.5.5. Specific Binding Sites 29
2.5.6. Final Feature Extraction Output 31
2.6. Sequence Masking 32
2.6.1. Without Regulatory Potential Information 33
2.6.2. With Regulatory Potential Information 33

Motif Finding & Processing 34
2.7. Motif Finding Algorithms 34
2.7.1. MEME 34
2.7.2. BioProspector 35
2.7.3. AlignACE 35
2.7.4. Improbizer 36
2.8. Representation of Found Motifs 37
2.9. Search Found Motifs in the Dataset using MAST 38
2.10. Score Found Motifs 39
2.10.1. Group Count and Frequency 39
2.10.2. Ratio of Group Frequencies (+/− Ratio) 40
2.10.3. Group Specificity Score 40
2.11. Comparison and Selection of Found Motifs 41
2.12. Draw Sequence Logos for Found Motifs 42
2.13. Assign Known TFBS Matrices to Found Motifs 43
2.14. Search for TRANSFAC Matrices in the Datasets 44

2.14.1. Search for All Vertebrate Matrices 44
2.14.2. Searching Matrices of Nuclear Receptors 44

Investigation of the Distributions of Certain Genomic Features . . 46
2.15. Median, Mean and Density Computation 46
2.16. Representation of Feature Distributions Over Genomic Regions 47

3. Results & Discussion **51**
3.1. Sequence Datasets 52
3.2. Bindings Site Analysis 56
3.2.1. Motif Search 56
3.2.2. Vertebrate Matrices Matching 62
3.3. Feature Distributions 66
3.3.1. CpG Islands/Regions 66
3.3.2. Specific Transcription Factor Binding Sites 68
3.3.3. Repeats 71
3.3.4. Co-appearance of Genomic Features 74

4. Conclusion **77**

5. Outlook **81**

Appendix **82**

A. Datasets **82**
A.1. Mouse Datasets (mm8) 82
A.1.1. 2K-2K 82
A.1.2. 2K-next 86
A.2. Human Orthologous Datasets (hg18) 89
A.2.1. H2K-2K 89
A.2.2. H2K-next 91

B. Description & Overview of Used Scripts **95**
B.1. Description of Scripts 95
B.2. Interaction Diagram 99

C. Feature Distribution Figures **100**
C.1. Mean, Median and Density 100
C.2. FeaturePlotter Plots 106

List of Abbreviations **115**

Bibliography **117**

List of Figures

1.1. Scheme of nuclear DNA packaging and 3D representation of a nucleosome containing histone H2A, H2B, H3, and H4 and wrapped DNA 6
1.2. Selection of histone tail modifications and their positions at the tails of histone H2A, H2B, H3, and H4 7
1.3. "States" of chromatin caused by histone acetyltransferases (HAT) and histone deacetylases (HDAC) 8
1.4. Genomic organization of the mouse β-globin cluster 11
1.5. Different models for enhancer activation of genes over a large distance . . 13
1.6. An active chromatin hub "knot" of active genes and hypersensitive sites in the mouse β-globin cluster 14
1.7. An explanation for the expression of overlapping gene loci drawn by ACHs 15

2.1. Scheme of the assignment of Fantom transcripts to homologous Symatlas transcripts, and *vice versa*. 20
2.2. Illustration of the included sequence for the *2K* and *next* datasets. 22
2.3. Sequence logo visualising the TRANSFAC matrix of the Sp1 protein (*V$SP1_Q6_01*). 31
2.4. Sequence logo visualising the TRANSFAC matrix of the *TATA-Box* (*V$TATA_C*). 31
2.5. Sequence logo visualizing the TRANSFAC matrix of the *myogenic MADS factor MEF-2* (*V$MEF2_04*) 43
2.6. Illustration of adjustment and mapping of two arbitrary feature distributions of different length. 48
2.7. Exemplary illustration of a **FeaturePlotter** plot with 3 regions and 3 features. 50

3.1. Boxplots of the intergenic distances for pairs of the positive and negative datasets. **A:** Mouse Datasets **B:** Human Datasets 52

3.2. Chromosomal position (relative to total number of genes on each chromosome) of gene pairs belonging to the positive and negative dataset. **A:** Mouse Datasets **B:** Human Datasets 55
3.3. Amount of pairexpression of pairs belonging to the positive and negative dataset. **A:** Mouse Datasets (13 tissues) **B:** Human Datasets (79 tissues) 56
3.4. Average distribution of CpG island over the positive and negative sequence in the mouse *2K-2K* datasets. 67
3.5. Average distribution of CpG island over the positive and negative sequence in the human *H2K-2K* datasets. 68
3.6. Distributions of predicted occurences of the SP1 binding site in the positive and negative sequence in the mouse **2K-2K** datasets. 69
3.7. Distributions of predicted occurences of the CTCF binding site in the positive and negative sequence in the mouse **2K-next** datasets. 71
3.8. Mean, median and density for the procentage coverage of **SINE** repeats. . 72
3.9. Mean, median and density for the procentage coverage of **LINE** repeats. . 73
3.10. Mean, median and density for the procentage coverage of **simple** repeats (**micro-satellites**). 74

B.1. Diagram of the interaction between the different scripts and programs used in this master thesis. 99

C.1. Mean, median and density for the procentage coverage of **SINE** repeats. . 100
C.2. Mean, median and density for the procentage coverage of **LINE** repeats. . 101
C.3. Mean, median and density for the procent coverage of **Simple** repeats. . . 102
C.4. Mean, median and density for the procentage coverage of **Low Complexity** repeats. 103
C.5. Mean, median and density for the procentage coverage of **LTRs**. 104
C.6. Mean, median and density for the procentage coverage of **DNA** repeats. . 105
C.7. Average distribution of CpG islands over the positive (red) and negative (blue) sequence. 106
C.8. Average distribution of CpG regions over the positive and negative sequence. 107
C.9. Average distribution of SP1 binding sites over the positive and negative sequence. 108
C.10. Average distribution of GC Box hexanucleotides over the positive and negative sequence. 109
C.11. Average distribution of CTCF binding sites over the positive and negative sequence. 110
C.12. Average distribution of SINE repeats over the positive and negative sequence. 111
C.13. Average distribution of LINE repeats over the positive and negative sequence. 112
C.14. Average distribution of simple repeats over the pxositive and negative sequence. 113

List of Tables

2.1. Thresholds for the group definition in FANTOM3 and GNF Symatlas. . . 19
2.2. Resulting amount of gene pairs for each dataset after group definition. . . 19
2.3. Distribution of phylogenetically conserved genomic pairs among co-expression groups in FANTOM3 (mouse) and Symatlas (human). 21
2.4. Full list of masked features and their assigned signs 32
2.5. Table of nuclear receptors and additional TFs of special interest and their associated TRANSFAC matrices used in the search for nuclear receptor binding sites. 46
2.6. Used mapping lengths of each region used in the **FeaturePlotter** plots according to the selected dataset. 49

3.1. Statistics for the *2K-2K* dataset. 53
3.2. Statistics for the *2K-next* dataset. 53
3.3. Statistics for the *H2K-next* dataset. 54
3.4. Statistics for the *H2K-next* dataset. 54
3.5. Percentage of gene pairs that have a certain genomic orientation for the human and mouse dataset. 55
3.6. The 10 best-ranking motifs resulting from the motif search process using the *2K-2K* dataset. 58
3.7. The 10 best-ranking motifs resulting from the motif search process using the *2K-next* dataset. 60
3.8. The 10 best-ranking TFBS resulting from the vertebrate TFBS search process using the **2K-2K** dataset. 62
3.9. The 10 best-ranking TFBS resulting from the tfbs search process using the **2K-next** dataset. 64
3.10. The 5 best-ranking nuclear receptor TFBS resulting from the nuclear receptor TFBS search process using the **2K-2K** dataset. 64
3.11. The 5 best-ranking nuclear receptor TFBS resulting from the nuclear receptor TFBS search process using the **2K-next** dataset. 65

3.12. Mean and significance level for the number of predicted SP1 binding sites contained in the sequences of the datasets. 69
3.13. Mean and significance level for the number of GC Boxes contained in the sequences of the datasets. 70
3.14. Mean and significance level for the number of predicted CTCF binding sites contained in the sequences of the datasets. 70
3.15. Number of SP1-associated CpG islands for positive and negative sequences in the mouse *2K-2K* and the human *H2K-2K* datasets and assigned pvalue. 74
3.16. Number of CpG island-associated SP1 binding sites for positive and negative sequences in the two *2K* datasets and assigned pvalue. 75
3.17. Statistics for the association between CTCF binding sites and CpG islands for the positive/negative sequences in all four dataset. 76

Chapter 1

Introduction

Since the thesis that every gene acts as a single unit which transcription is solely regulated by promoter-binding transcription factors (TF) - irrespective of the surrounding genomic landscape - has been rejected, transcriptional regulation of genes has become a field of ever-growing complexity.
Factors like the "state" of chromatin and DNA positioning inside the nucleus have been shown to have a major impact on the activation and repression of the transcription of genes [1],[2],[3]. Furthermore it was discovered that the expression of individual adjacent genes in the genome is not independent, but genomic neighbours are co-expressed more often than what would be expected by chance [4],[5]. These neighbours form clusters of co-expressed genes that can be found all over the genome containing from two to several adjacent entities. In this thesis a possible explanation of this observation was investigated, namely the active alteration of chromatin state by possible interaction of transcription factors or other genomic features. Sequence analysis methods were used to search for possible DNA specific factors that could form "active chromatin hubs (ACH)" [6] in the region of those co-expressed genes and therefore could lead to the revealed correlated expression. The thesis is based on our earlier analysis of the expression of genomic neighbours in mouse/human and proceeds these investigations [7].

Evidence for Clusters of Co-expressed Genes Throughout the Genome

The following pages present an overview of the previous studies analysing the existence of clusters of co-expressed genes in the genomes of eukaryotic and prokaryotic organisms, including the results from our group investigating the level of correlated expression of genomic neighbours and their genomic properties.

1.1 Finding Clusters of Correlated Genes

1.1.1 Clusters of Co-expressed Genes

Co-expression of genomic neighbours on a genomic scale was first discovered in yeast for genes involved in the mitotic cell cycle [4]. In this analysis 25% of the genes that were expressed in cell-cycle-dependent manner lay adjacent to each other. Another analysis in the genome of Drosophila melanogaster reveals that testes genes were found in clusters of at least four genes [5], which could also be extended using a looser definition of clusters (allowing for intervening genes). In addition to those one-tissue-clusters, other groups analysed genes that are expressed in a broader range of tissues and found genes with high expression levels (housekeeping genes) to be clustered in the human genome [8],[9]. However, our own analysis also postulated a high number of co-expressed genes in the human and mouse genome that are expressed in a broader range of tissues (from housekeeping pairs to pairs that are exclusively expressed in only one tissue). Regarding the full genome expression analysis published so far and the increasing number of finished genomic sequences, there is growing evidence for the existence of clusters of co-expressed genes across all eukaryotic organisms.

1.1.2 Clusters of Co-functional Genes

It has been shown that genes encoding for proteins that are involved in the same metabolic pathway have the tendency to cluster along the genome of several organisms (including human, worm, fly, A. thaliana and yeast) [10]. Nevertheless, a relationship between co-functionality and co-expression in higher vertebrates has not been shown satisfactorily. Most of the well-studied clusters (e.g. Hox cluster, growth hormone cluster) show high co-functionality but fail to show high co-expression or even deny it, because of highly different expression patterns (e.g. resulting from different times of expression in development). An analysis of common GO categories for co-expressed genomic

neighbours resulted in only a rare number of clusters [11]. This was also suggested by our own analysis, even when the number of genomic neighbours, irrespective of their co-expression, which share common GO categories, is significantly higher compared to random pairing [7]. A very recent analysis investigating gene cluster in human and mouse also support this non-correlation between co-functionality and co-expression, proposing *transcriptional leakage* (e.g. driven by unspecific "opening" of a whole chromatin region) to be one of the major factors leading to coordinate expression of genomic neighbours [12]. This model would suggest gene expression in tissues without any functional need.

1.1.3 Spatial Organisation Versus Clustering

A higher order of gene arrangement in genomes must not solely mean the occurrence of clusters. The possibility of a spatial distribution of co-expressed genes over the chromosomes was primarily postulated by several groups performing analysis of microarray experiments on yeast [13],[14]. In contrast to this other groups strongly deny such regular spacing [15]. They propose the existence of periodicity to be artifacts caused by the printing of yeast chips in genomic order. Moreover, they postulate that there is currently no statistically significant evidence that transcription factor binding sites in yeast tend to be regularly spaced. Nevertheless, they found striking significance for co-expression and transcription factor binding site sharing for genes of close proximity [15].

1.2 Our Previous Results in Investigating Highly Co-expressed Genomic Neighbours

In our previous analysis [7] we focused on adjacent genes in the human and mouse genome, using the **FANTOM3 [16] Mouse** and **GNF Symatlas [17] Human** datasets to annotate these genes with expression data (13 tissues in FANTOM3 mouse and 79 tissues in GNF Symatlas human). Using a measurement of the ratio of co-expression over those tissues, we extracted genomic clusters which we called "highly co-expressed"[1]. Those clusters mainly consists of pairs and triplets of genes and can be located all over the genome. Analysing the amount of tissues the individual genes of those clusters are expressed in, we found a wide range of clusters from one-tissue-clusters to housekeeping-clusters. We showed these clusters to be limited in size (measured in nucleotides) and individual highly co-expressed pairs to have a smaller intergenic distance than overall genomic neighbours (median of 7,662bp versus 18,665 bp for all genomic pairs in mouse; *p-value* of $3 * 10^{-5}$ in Wilcoxon Rank Sum Test). Analysing genomic orientations of pairs, we could not find a difference in their distributions between highly

[1] See section 2.1 for the full definition of "highly co-expressed" gene clusters

co-expressed gene pairs and overall gene pairs. Further assessing the sharing of GO terms, protein domains and TFBS we found that highly co-expressed gene pairs share those features to a lesser extent than overall genomic neighbours (with the exception of transcription factor binding sites that were shared to the same extent). All our findings which were mainly based on mouse data, could be confirmed using the human data. Additionally we found a high number of highly co-expressed pairs that are phylogenetically conserved between these two species.
From our analysis we suggested that the high amount of highly co-expressed genomic neighbours could be a result of large-scale chromatin alterations that lead to "open" regions that allow the correlated expression of several genes, additionally regulated by individual transcription factor binding sites (TFBS). The aim of this master thesis was to find possible mediators of these opening events in the sequence of our postulated gene pairs.

Levels of Eukaryotic Genome Regulation

The following pages present an overview of the eukaryotic gene expression as a framework with three hierarchical "levels" [1] of genomic regulation, from the level of individual gene regulation via regulation of chromatin regions to the nuclear level.

1.3 The Sequence Level

The sequence level is the best-studied level of transcriptional control in eukaryotes. It involves elements that lead to regulation of individual genes, so called **trans-acting and cis-acting elements**.

- **Trans-acting elements**
  Trans-acting elements include the **RNA polymerase 2**, which transcribes genetic DNA into messenger RNA, as well as several **co-factors**. These are directed to specific transcriptional start sites (TSS) by a huge amount of **transcription factors** that governs tissue specific transcription of individual genes. Furthermore chromatin-remodelling systems that give access to transcribed regions play a role in this basic transcriptional machinery (those will be further discussed in the *chromatin level*).

- **Cis-acting elements**
  Cis-acting elements are sequence elements that guide the specific transcriptional machinery. They are normally sub-divided into **promoters** (which enable gene transcription), **enhancers** (which increase transcriptional level) and **silencers**

(which are bound by repressing transcription factors and therefore prevent genes from being transcribed). Those cis-acting elements are not exclusively localised in the nearer environment of a gene (e.g. near its TSS), but are also found several kb apart of its controlled gene (e.g. 40-60kb apart in case of the β-globin gene cluster)[6].

The regulation of gene transcription at sequence level is highly complex in itself and the scheme presented here is therefore only a fragmentary overview.

1.4 The Chromatin Level

As all eukaryotic genomes are found to be packed in nucleosomes that are furthermore condensed to finally reach a compression of 10,000-fold rendering it inaccessible to the transcriptional machinery, the chromatin level is likely to play a role in transscriptional regulation. The basic units of such nucleosomes is an octamer of the histone molecules H2A, H2B, H3, and H4, the linker histone H1 and an appropriate DNA double-helix which is tightly wound around the complex in 1.75 turns per nucleosome [18] (see figure 1.1). A good review of the up-to-date knowledge of chromatin modifications and their transcriptional impact is given in [19].
A deeper understanding of the chromatin level is of high interest for our data analysis because we expect the postulated existence of co-expressed genomic neighbours to be mainly a result of higher order changes in chromatin states over large regions.
In terms of gene expression, chromatin structures that made genes accessible for transcriptional assessment by transcription factors and RNA polymerase 2 transcriptional initiation machinery are called **"open chromatin"** or **"euchromatin"**, whereas structures that prevent genes from being transcribed are called **"condensed chromatin"** or **"heterochromatin"** [19]. At least three distinct types of nucleosomal alteration have been proposed and proven to change transcription level of targeted genes: **chromatin remodeling**, **core histone replacement**, and **histone tail modifications**.

1.4.1 ATP-dependent Chromatin Remodeling Complexes Increase the Mobility and Fluidity of Nucleosomes

The same complex that forms chromatin structure in replication are found to be relevant for sliding of the histone octamer. This is mediated by ATP hydrolysis to rearrange nucleosomal arrays and free specific regions for later accession by the basic transcriptional machinery [2],[21],[22].
Using ChIP-Chip experiments in yeast it was shown that there is a positive correlation between the presence of those **nucleosome-free regions (NRF)** of approximately **150bp** which are located in promoter regions and the rate of gene transcription [2],[22].

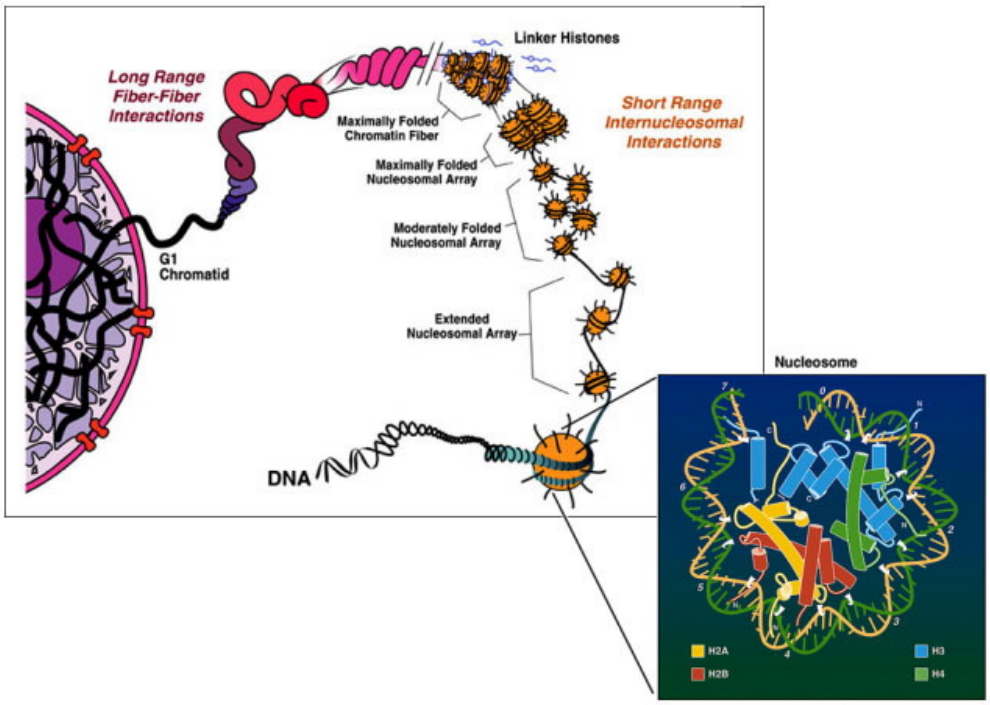


Figure 1.1: **Scheme of nuclear DNA packaging and 3D representation of a nucleosome containing histone H2A, H2B, H3, and H4 and wrapped DNA**. [20]

Reports on an corresponding promoter NRF association in human are contradictory [21],[23].

1.4.2 Replacement of Core Histones by Special Histone Variants

Replacement of histone particles by specialised variants have been shown to occur near transcribed regions and could influence the transcriptional machinery.
Using ChIP-Chip experiments in yeast the histone variant H2A.Z was shown to replace the core histone H2A preferentially near promoter regions [24],[25]. It is strongly suggested that this variant flanks NRFs and blocks the spreading of activating histonemarks, thereby preventing euchromatin formation [26],[27].
Histone H3 was also shown to be replaced by a variant called H3.3. This typically happens in genomic regions and marks actively transcribed genes, because H3.3 is gradually enriched with every round of transcription [2],[22]. Furthermore, a slight enrichment of H3.3 can be found upstream of the TSS and of NFRs [22].

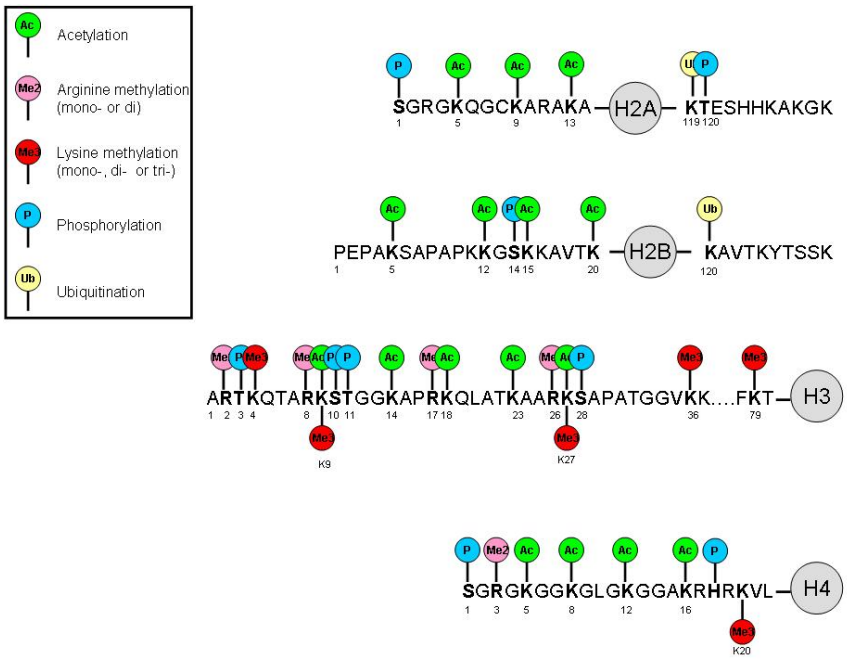


Figure 1.2: **Selection of histone tail modifications and their positions at the tails of histone H2A, H2B, H3, and H4**. A spot above the tail indicates postulated transcriptional activation or unknown function, a spot below the tail indicates repression. [19]

1.4.3 Histone Tail Modifications

Studies investigating chromatin state alterations have so far mainly focused on on post-translational histone tail modifications. These modifications can influence the wrapping of DNA around the histone core and thereby lead to an altered transcriptional accessibility. Known histone modifications are: *acetylation* [28], *methylation* [29], *phosphorylation* [30], *ubiquitiniation* [31], *sumoylation* [32], *ADP ribosylation* [33], *glycosilation* [34], *biotinylation* [35] and *carbonylation* [36]. The distributions of these modifications along the histone tails and their influence on the transcriptional machinery is called the **histone code** [37]. A graphical overview of some of these modifications and their position and transcriptional function at the histone tails is presented in figure 1.2. **Acetylation** and **methylation** are the best-known of these modifications:

- **Acetylation**
  Acetylation marks are placed by a group of enzymes called **histone acetyltransferases (HAT)**. The acetylation of the histone tail is widely proposed to lead to an alteration in charge and lower the electrochemical coupling between the histone

octamer and the wrapped DNA making the DNA more accessible for the transcriptional machinery [38]. Correspondingly histone acetylation is tightly linked to an increase in transcription ("euchromatic state") [39].
Deacetylation on the other hand is associated with an decrease in transcriptional level ("heterochromatic state") . It is mediated by **histone deactelyase (HDAC)** co-repressor complexes.
An overview of these two modifications and their influence on the condensation state of chromatin is given in figure 1.3.

- **Methylation**
  In contrast to acetylation methylation is not clearly correlated with transcriptional activation. Moreover the position of the methylation at the histone tail seems to be the major factor of its effect [19]. A methylation of the lysin at position 4 of histone 3 (H3K4) was for example shown to be associated with chromatin structures that allows for transcriptional activation [21]. In contrast, methylation of K9

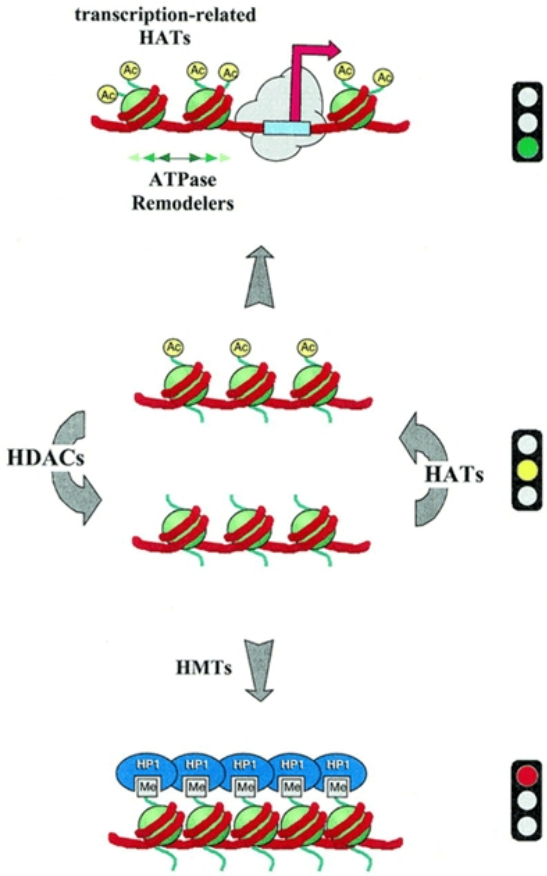


Figure 1.3: **"States" of chromatin caused by histone acetyltransferases (HAT) and histone deacetylases (HDAC)**. [3]

at the same histone is thought to be linked to heterochromatin formation [40]. Furthermore the degree of methylation of H3K4 was shown to be dependent on its position in the genomic region [2]. It decreases continuously from 5' to 3' with trimethylation at 5', dimethylation in the middle and monomethylation at the 3' end.

In contrast to the two chromatin alterations mentioned above, these histone modifications do not only occur locally but can spread along the chromatin fibre, thereby inducing a change in the functional state of whole chromatin domains containing one or more genes. These regions, also called "Active Chromatin Hubs" [6], could be a fundamental architecture of highly co-expressed gene clusters and are therefore discussed further in section 1.8.

1.5 The Nuclear Level

The knowledge of the nuclear level of transcriptional regulation is so far very limited. It includes location of chromosomal parts throughout the **nucleus** as well as 3D convergence of very distant (or even chromosome spanning) gene regions. In yeast it was shown that **nuclear areas** exist, that differentially influence transcriptional level - from repressive to boosted transcription. For instance genes that are located near the yeast cell periphery are silenced [1],[41]. A correlation has also been shown between the number of genes on eukaryotic chromosomes and their position in the nucleus, with gene-rich chromosomes residing more frequently in the center and gene-poor chromosomes located in the periphery [42]. Additionally, transcription factors like SATB1 have been shown to form "networks" that specifically link targeted DNA sequences and therefore change **nuclear architecture** [43].

1.6 Links Between the Three Hierarchical Levels

While the presented framework is only a model of different levels of transcriptional regulation, the real procedures in the cell are much more linked. Several sequence specific transcription factors (e.g. REST [44], CBP [24]) are known to recruit the activating/repressing HAT/HDAC complexes and therefore initiate chromatin "opening" or "closing" [45],[46]. A class of transcription factors which are called **nuclear receptors** have recently been shown to be able to bind to histones and activate the remodelling machinery [21].
This link between the hierarchical levels is also true for the nuclear level, as the already mentioned transcription factor SATB1 does not solely form its own nuclear architecture but also attracts both enhancing and repressing chromatin alteration enzymes [43].

"Active Chromatin Hubs" Mediate Correlated Gene Expression

The following pages present an overview of the concept of "Active Chromatin Hubs" (ACH) - regions of "open" chromatin that could lead to a correlated expression of genes in close genomic regions - and their proposed control elements. Furthermore several models that could explain correlated expression of even more distant genes are discussed.

1.7 Known Gene Clusters Driven by Active Chromatin Hubs

Irrespective of their level of co-expression, several **conserved gene clusters** have been identified that share a high level of regulated expression that is guided by **chromatin state "switches"**. The best characterized clusters so far are the β-globin [47], the growth hormone [48] and the multiform Hox gene clusters [49].
The Hox gene family for example, which is responsible for controlling the genetic system that specifies structures along animal body axes in mammals [50], is grouped into so far four known genomic clusters: HoxA, HoxB, HoxC and HoxD. HoxB genes have been shown to have a strict expressional *order* that depends on the developmental stage of the organism and is guided by chromatin modifications [51]. Biochemical experiments using embryonic stem cells showed transcription of HoxB1 at day 2-4 after treatment with retinoic acid (which initiates cell differentiation) whereas HoxB9 was expressed at day 10, at which HoxB1 was no longer expressed. Consistent with the chromatin alteration model of transcriptional activation/silencing, an acetylation of lysin 9 of histone 3 and a methylation of lysin 4 also at histone 3 simultaneously at the HoxB locus at day 4. These signals disappeared until day 10 (when the gene is silenced). Nevertheless, they also found HoxB9 to be associated with the same modifications in chromatin, but already at day 4 and continuously afterwards [49].

1.8 Proposed Genomic Elements of Active Chromatin Hubs

Clusters which result from chromatin alteration / active chromatin hubs are supposed to consist of the following three classes of genomic elements: **Cluster Control Elements**, **Enhancer and Promoter**, and **Boundary Elements (Insulators)**.

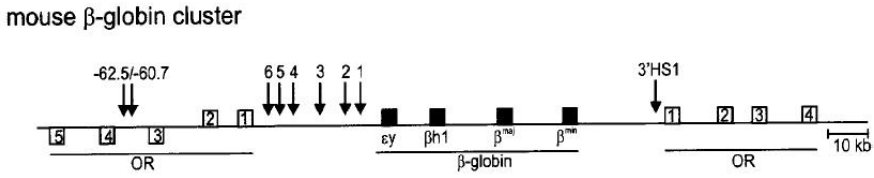


Figure 1.4: **Genomic organization of the mouse β-globin cluster**. β-globin cluster genes are indicated as black rectangles, genes belonging to olfactory receptor clusters in white. Rectangles above the line represent genes on the positive strand, below the line at the negative strand. Arrows indicate known HS. [6]

1.8.1 Cluster Control Elements

The cluster control element is responsible for switching the genomic domain between its active and inactive state. These elements might recruit histone-modifying enzymes complexes containing HATs and HDACs. The initiated chromatin change could then spread along the genomic region to "open" or "close" chromatin structure and make associated genes accessible/unaccessible for transcription.

Candidates for these cluster control elements are:

- **Locus Control Regions (LCR)**
  Locus control regions consist of a set of of cis-acting elements that have the competence to fully activate a transgene[2] (e.g. in a tissue-specific and copy-number-dependent manner) at any location in the genome [52]. In normal LCRs, each cis-acting element forms a **DNAse I hypersensitive site (HS)** and contains several transcription factor binding sites [48] (see figure 1.4 for an example of different HS in the mouse β-globin cluster).
  Several transcription factors have been annotated to have LCR binding properties and can therefore initiate (e.g. tissue) specific gene regulation. One example is the transcription factor REST (RE-1 silencing transcription factor) which was allready mentioned in chapter 1.6. It is a zinc-finger gene-specific repressor element that restricts the activity of genes in non-neural tissues due to recruitment of HDACs that repress expression. Trough the recruitment of CoREST (associated co-repressor) it expands its silencing influence to genes in the near genomic environment that have no own REST response element [44]. Also SATB1 (already introduced in chapter 1.5 to regulate gene expression at the nuclear level by inducing its own "networks") was shown to upregulate the transcription of its targeted and neighbouring genes by binding to SBS-T4 which initiates hyperacetylation of adjacent regions of chromatin [43]. Another known factor having LCR binding properties is the CREB-binding protein (CBP). It binds the cAMP-response binding protein

[2] A transgene is a gene which has been transferred into genomic DNA from a different source.

and recruits HATs [53].
Besides sequence specific transcription factors, **nuclear receptors (NR)** have been reported to recruit chromatin modifying complexes. NRs were shown to activate target gene expression by recruiting co-activators (among others HATs) in a ligand dependent manner. On the other hand the same receptor diminishes transcription in the absence of a ligand by recruiting co-repressors (amongst others HDACs) [21].

- **Repetitive DNA Elements and RNAi**
  Repetitive DNA elements, so called **interspersed sequence repeats**, which comprises ∼**50%** of the genome of mice and humans, have been suggested to function as cluster control elements. Pairing among those repeats was proposed to introduced secondary DNA structures that can act as nucleation sites for the establishment of heterochromatin like configurations [54], [55].
  Furthermore RNA interference (RNAi) was shown to mediate heterochromatin formation in yeast [56], arabidopsis [57], drosophila [58] and chicken [59].
  A connection between **RNAi pathway** and the assembly of silent chromatin on (and spreading from) nearby long terminal repeats was proposed but couldn't be confirmed in human [60].
  But repetitive elements are not exclusively associated with transcriptional silencing, as for example **Alu repeats** are found to contain many binding sites for transcription factors that might mediate developmental processes [61]. Furthermore chromosomal regions that are transcriptionally very active were shown to have a high SINE repeat density [9].

1.8.2 Enhancer and Promoter

Additionally to the superordinate chromatin alteration that changes the accessibility of genomic regions containing several genes, individual expression is furthermore regulated by gene specific enhancers and promoters (see section 1.3).

1.8.3 Boundary Elements (Insulators)

The existence of insulators, that separate gene clusters by limiting the control range of long-distance regulatory elements, is controversial.
If affirmed, their location is proposed at the borders of ACHs to stop surrounding heterochromatin marks from entering the active region [48], [62]. The best known mammalian protein that has insulating activity is **CTCF** [63], which was also shown to mediates long-range chromatin looping and local histone modification in the β-globin gene cluster [64].

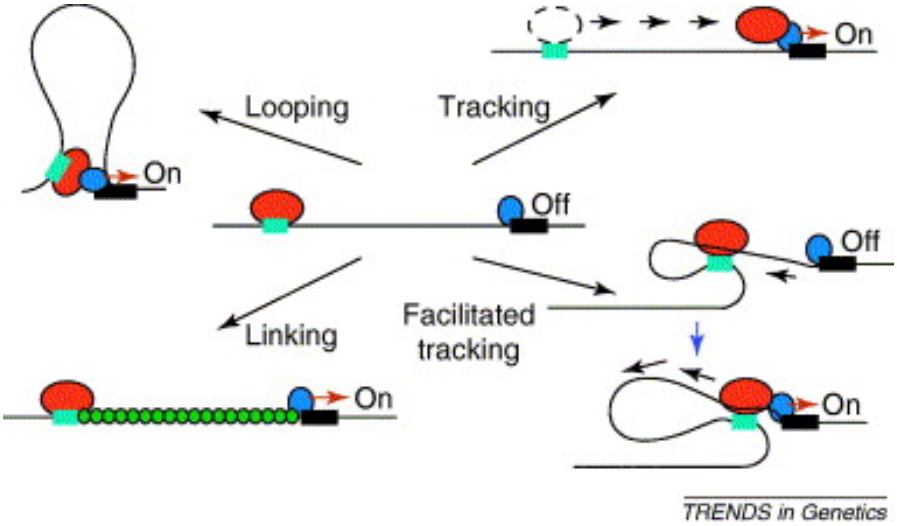


Figure 1.5: **Different models for enhancer activation of genes over a large distance**. Blue rectangles represent an enhancer, the red ellipses its recruited activation complex. The genes are represented as black rectangles and the promoter-binding complex as blue ellipses. Linker proteins are indicated as green circles. [67]

On the other hand examples of transgene-induced heterochromatin were reported to fail to enter euchromatic regions without the necessary existence of any insulatory element [65], [66]. A possible explanation for the stopped re-repression would be the **accumulation of transcription factors** and associated chromatin-modifying complexes (containing e.g. HATs) resulting in **hyperacetylation** which is proposed to be a mechanism that avoids heterochromatin silencing [48].

1.9 Different Models of Active Chromatin Hub Establishment

The fact that enhancer elements have the possibility to influence genomic regions that are up to 800 kb apart has long been disputed [67]. Current models favoured are the **tracking model** [68], the **looping model** [69], the **linking model** [70], and the **facilitated tracking model** [71] (for an overview of all these models see figure 1.5).

- **Tracking Model**
  The tracking model (or scanning model) proposes the tracking/scanning of an transcription-activation complex that was initially recruited by an enhancer along the DNA until it reaches a promoter, meanwhile opening the whole stretch of chromatin between these element, but does not alter their proximity.

- **Looping Model**
  In the looping model the enhancer and promoter regions are directly brought together in the nucleus by binding of their associated complexes.

- **Linking Model**
  In the linking model and enhancer binding protein is iteratively bound by **facilitator proteins** until the protein chain reaches a promoter region where it enhances transcriptional activity.

- **Facilitated Tracking Model**
  In the facilitated tracking model both, the tracking model and the looping model, are incorporated. It suggest that an enhancer-bound activation complex migrates along the DNA until it reaches a promoter, meanwhile forming a loop which is progressively enlarged during the process.

Because of its explanatory power to some aspects of gene cluster regulation, the looping model (and the related facilitated tracking model) have recently found higher support [67]. Studies investigating enhancer-promoter proximity of the β-globin cluster in erythroid cells revealed that those are closely located in 3D [72]. This might indicate the formation of chromatin loops which co-locates specific sequence sites. The looping model also explains **different expression levels of genes that belong to the same gene cluster** [6]. It is suggested that different HS are co-located by the established loops and form a "**knot**", which causes a high enrichment of transcription factors and associated HATs. Genes that lay close to this knot can interact with these factors leading to an increased transcriptional level, while genes that lay in the outer part of the

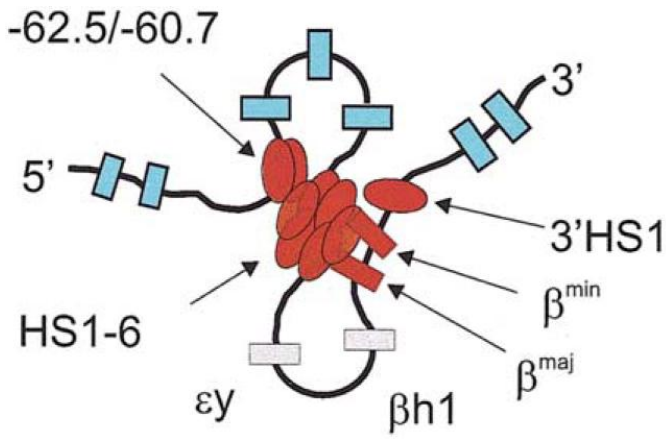


Figure 1.6: **An active chromatin hubs "knot" of active genes and hypersensitive sites in the mouse beta-globin locus**. HS (ellipses) and genes (rectangles) in red are activated, those in grey are not transcribed. Rectangles in blue mark the surrounding olfactory receptor genes. [6]

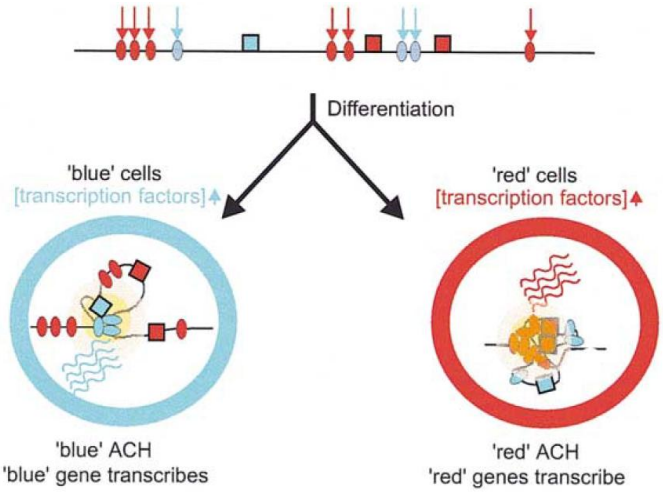


Figure 1.7: **An explanation for expression of overlapping gene loci drawn by ACHs**. Presentation of two hypothetical, differentially regulated, gene loci (red and blue) that overlap, with cis-regulatory sequences as ellipses and genes as rectangles. Depending on transcription factor binding competition the 'blue' ACH is formed in the 'blue' cells, which results in expression of the 'blue' genes. A similar mechanism applies to the formation of 'red' ACH in 'red' cells resulting in the expression of 'red' genes. [6]

loop remains untranscribed. The reverse is true if the knot includes silencing regions. Figure 1.6 demonstrates such a possible knotting structure for the β-globin gene cluster. During development this loop formation might be rearranged, now containing the new targeted genes located near the knot while the old genes are silenced [73]. An explanation for this rearrangement could be a change in **chromatin flexibility**, which in turn depends on chromatin modifications (especially acetylation) [67]. The model predicts the degree of acetylated chromatin to determine the size of the established loops and has been used to explain the linear decrease in expression of the HoxD cluster genes and the volatile expression in the human β-globin cluster.
As presented in figure 1.7 the looping model provides a possible explanation for the coherence of ACHs and the expression of overlapping gene loci [6]. Several distinct promoters in one loci might compete with each other, leading to distinct formations of chromatin loops and therefore distinct expression patterns. The distance between the promoter and the HS might affect the result of these competition, but presence of specific transcription factors could also provide an important contribution.

Chapter 2

Methods

The methods used in this thesis consist of four parts:
first, the **Definition of Sequence Datasets** part contains the constructions of sequence datasets that represent the previously defined highly co-expressed and uncorrelated gene pairs [7], respectively, in mouse together with an orthologous human dataset to verify our results.
Furthermore, scripts for the **Retrieval of Sequence Data and Features** were implemented to build a base for the following analysis.
The first analysis searches for overrepresented motifs in the sequence set of highly co-expressed gene pairs which could point to possible transcription factors involved in co-ordinated expression. This part is called **Motif Finding & Processing**.
The second set of analysis comprises the **Investigation of Distribution of Certain Genomic Features** of the sequences as a whole and over individual regions.

For the exact chromosomal position and included Ensembl genes of each defined dataset refer to **Appendix A**

For an overview of the used scripts, their description and interactions see **Appendix B**.

Definition of Sequence Datasets

The following pages contain detailed information about the choice of genomic regions that were used to search for regulatory elements to reveal the regulatory mechanisms that might lead to correlated co-expression of genomic neighbours. The description of the initial selection of gene pairs to a set of highly co-expressed/uncorrelated gene pairs is given, along with the different setups that have been defined based on this selection.

2.1 Definition of Highly Co-expressed and Uncorrelated Gene Pairs

In our previous work [7] we analysed the amount of co-expression of genomic neighbours in two dataset: 1. the **FANTOM3 [16] Mouse** datasets, which consists of **39593 genes** with expression values for **13 tissues** and 2. the **GNF Symatlas [17] Human** dataset, which consists of **19358 genes** with expression values for **79 tissues**.

Previously we grouped gene pairs (genomic neighbours) into categories called **highly co-expressed (HCP)**, **uncorrelated (UCP)**, **housekeeping**, and **silenced** according to the amount of contiguous expression relative to overall expression.
More precisely, we defined two **coefficients** A, which is the proportion of tissues from all n tissues, in which both genes of a genomic pair are expressed together, and Ω, which is is the proportion of tissues from all n tissues, in which either one or both genes of a genomic pair are expressed. Both coefficient lay in the interval [0,1] and by definition $A \leq \Omega$. We used the **ratio** $\frac{A}{\Omega}$ to access the degree of co-expression for each genomic pair. This ratio is close to or equal 1, if the genes are expressed together in almost all cases (irrespective of the total number of tissues they are expressed in) and is close to or equal 0 if they are never or rarely expressed together.

Transcripts were assigned to the above categories following two **thresholds** θ_{coex} **and** θ_{uncor}:

1. A gene pair is defined as **highly co-expressed** if $\frac{A}{\Omega} \geq \theta_{coex}$ and $A < 1$

2. A gene pair is defined as **uncorrelated** if $\frac{A}{\Omega} \leq \theta_{uncor}$ and $\Omega > 0$

3. A gene pair is defined as **housekeeping** if $A = 1$ (both gene are expressed in all n tissues)

4. A gene pair is defined as **silenced** if $\Omega = 0$ (both genes are never expressed)

Due to the difference in the distribution of expression over the total number of tissues, the threshold θ_{coex} and θ_{uncor} were set differently for the FANTOM3 and GNF Symatlas dataset. The defined threshold values are reported in table 2.1.

Dataset	θ_{coex}	θ_{uncor}
FANTOM3 Mouse	0.75	0.5
GNF Symatlas Human	0.5	0.33

Table 2.1: Thresholds for the group definition in FANTOM3 and GNF Symatlas.

After applying these group definitions to the two datasets, the amount of gene pairs that belong to each group were obtained as shown in table 2.2.

Dataset	***highly co-expressed***	***uncorrelated***	***housekeeping***	***silenced***
FANTOM3 Mouse	3,230	27,287	154	36
GNFSymatlas Human	1,800	14,886	21	1,370

Table 2.2: Resulting amount of gene pairs for each dataset after group definition.

2.2 Extraction of Phylogenetically Conserved Pairs

We aimed to design a set of sequences that provide the possibility to analyse the proposed mechanism of regulated co-expression of genomic neighbours. The datasets provided by FANTOM3 and GNF Symatlas are large and noisy. To obtain gene pairs with **stable expression properties** these datasets were reduced to those gene pairs existing in both sets, and hence contain two **human-mouse orthologs**.

Data To define human/mouse gene homologs the current table of orthologous genes (represented by Ensembl.Gene.IDs) was downloaded from Ensembl[1] via its "BioMart" tool. (Date: 05.06.06; Ensembl 39; Mouse: NCBI m36 Assembly (Dec 2005) mm8 Genebuild Ensembl (Jun 2006); Human: NCBI 36 (Oct 2005) hg18 Genebuild Ensembl (Mar 2006)).
Furthermore, two tables that assign Mouse.Ensembl.Gene.IDs (for the FANTOM3 Mouse dataset) and Human.Ensembl.Gene.IDs (for the GNF Symatlas Human dataset), respectively, to the transcripts in the appropriate dataset were used. These tables are provided with the Fantom3 and GNF Symatlas dataset. From all **39593** transcripts in the FANTOM3 dataset **20837** have a Mouse.Ensembl.Gene.ID and for GNF Symatlas Human it is **10795** out of **19358**.

[1] `http://www.ensembl.org`

Approach

1. **Compute a one-to-one ortholog assignment table**
   Using both Ensembl orthologous gene tables, all Ensembl.Gene.IDs for mouse/human, that refer to more than one Ensembl.Gene.ID in the opposed species, were removed. Additionally, inconsistency among the two Ensembl tables was verified. These procedure resulted in a one-to-one assignment table between Mouse.Ensembl.Gene.IDs and Human.Ensembl.Gene.IDs (or vice versa).

2. **Compute the orthologous Symatlas transcript(s) for every FANTOM3 transcript**
   The assigment of orthologous Symatlas transcripts to Fantom transcripts is schematically outlined in Figure 2.1.

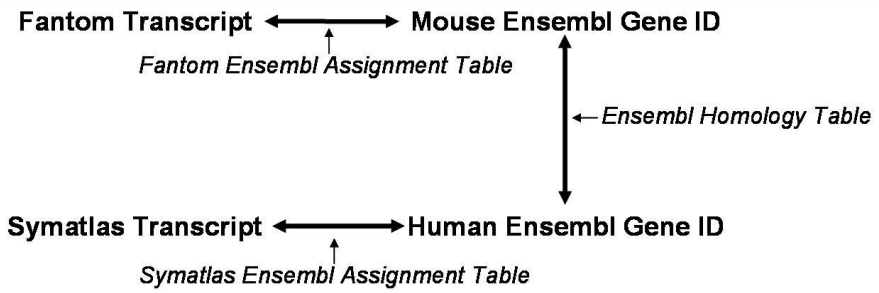


Figure 2.1: Scheme of the assignment of Fantom transcripts to homologous Symatlas transcripts, and *vice versa*.

   Afterwards, this assignment was revised for **A)** FANTOM3 transcripts that refer to several Symatlas transcripts (due to duplicated Human.Ensembl.Gene.IDs for different Symatlas transcripts) and **B)** Symatlas transcripts that refer to several FANTOM3 transcripts (due to duplicated Mouse.Ensembl.Gene.IDs for different FANTOM3 transcripts).

   This procedure resulted in **8269** distinct FANTOM3 transcripts that could be assigned to unique Symatlas transcripts (or vice versa).

3. **Extract pairs of mouse genes with paired human homologs**
   To identify **phylogenetically conserved pairs** for all adjacent genomic neighbours (pairs) it was determind that A) they consist of two transcripts assigned to orthologous Symatlas transcripts and B) these orthologs are also adjacent (paired) in the Symatlas human dataset. The specific order of the transcripts in the pair was neglected to allow for evolutionary inversion.

Result **1667** phylogenetically conserved (between mouse and human) FANTOM3 pairs were identified. Those 1667 pairs were distributed among the defined co-expression groups as listed in table 2.3.

	Symatlas.HCP	**Symatlas.UCP**	**Symatlas.Misc**	**FANTOM Sum**
FANTOM.HCP	168	278	90	536
FANTOM.UCP	39	453	41	533
FANTOM.Misc	109	416	73	598
Symatlas Sum	316	1147	204	1667

Table 2.3: Distribution of phylogenetically conserved genomic pairs among co-expression groups in FANTOM3 (mouse) and Symatlas (human). "Misc" = gene pairs not belonging to HCPs or UCPs.

2.3 Definition of Positive and Negative Dataset

2.3.1 Selection of Pairs for the Positive/Negative Groups

Definition To identify regulatory elements that lead to a high level of co-expression a **positive group** was defined. Sequences contained in this group are proposed to contain these elements. The **negative group** was defined as a set of genes with low co-expression, as these are unlikely to contain these regulatory elements.
We categorised phylogenetically conserved pairs, that are **highly co-expressed in FANTOM3 AND in Symatlas** (in total **168**) into the positive group and those phylogenetically conserved pairs, that appear to be **uncorrelated in FANTOM3 AND in Symatlas** (in total **453**) into the negative group (compare to table 2.3).

As basis for all following computations and analysis the mouse sequence of the appropriate pairs were used. Importantly, the results were verified using a **human orthologous dataset**.

2.3.2 Further Preparation of Pairs of the Positive/Negative Group

Data For all FANTOM3 transcripts belonging to one of the two defined groups the following features were extracted via "BioMart": Mouse.Ensembl.Transcript.IDs (by their assigned Mouse.Ensembl.Gene.IDs), their chromosome, genomic start/end position and strand information using current annotations provided by Ensembl (Date: 20.05.06; Ensembl 39; Mouse: NCBI m36 Assembly (Dec 2005) Genebuild Ensembl (Jun 2006)).

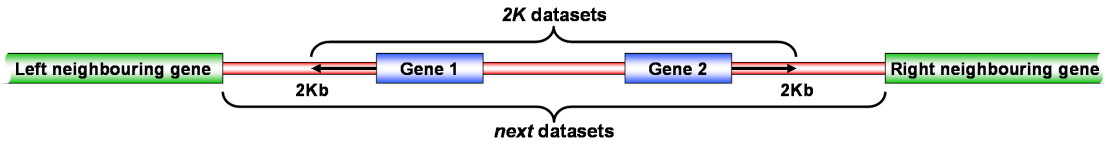


Figure 2.2: Illustration of the included sequence for the *2K* and *next* datasets.

Approach From the selected groups following pairs were removed:

1. **FANTOM3 transcripts that have multiple Mouse.Ensembl.Transcript.IDs**
   An assignment of multiple IDs can occur, as gene IDs may have several transcript IDs (because a single gene can have multiple transcripts). By removing these a one-to-one relation was obtained.

2. **Ensembl transcripts overlapping other Ensembl transcript or with other Ensembl transcripts included**
   The Fantom3 dataset does not inlcude all current Ensembl transcript IDs. To ensure genomic adjacency, pairs that overlap or that have other transcripts laying in between were excluded.

After removing all affected pairs, **93** pairs for the positive group and **226** pairs for the negative group remained. These pairs provide the basis for the following definition of specific datasets (see below).

2.3.3 Definition of the final mouse datasets "2K-2K" and "2K-next"

To ensure a real "clustering" of the pair the **distance to the next left/right transcript was required to be at least 2kb** (as annotated by Ensembl). Two datasets differing in the amount of surrounding sequence were defined:

1. **2K-2K** This dataset includes the sequence of all genes pairs of the positive/negative dataset which are at least 2,000bp distant to their next adjacent transcript and includes 2,000 bp around the pair.

2. **2K-next** This dataset includes the sequence of all genes pairs of the positive/negative dataset which are at least 2,000bp distant to their next adjacent transcript and the total sequence that spans the distance to the next right/left transcript. If regional overlaps occured, one of the overlapping pairs was skipped to avoid duplicated sequences.

The amount of sequence included in the datasets is illustrated in Figure 2.2.

Dataset *2K-2K* comprises **185 sequences** (**51 positive** and **134 negative**). Dataset *2K-next* comprises **181 sequences** (**51 positive** and **130 negative**). The difference in numbers is a result of overlapping regions.[2]

2.3.4 Definition of Orthologous Human Datasets "H2K-2K" and "H2K-next"

Based on the two datasets defined for mouse, two **datasets of orthologous gene pairs in human** called "H2K-2K" and "H2K-next" were computed. The annotated Ensembl human homologous pairs were extracted for all 185/181 gene pairs using the Ensembl homology table described above. Again gene pairs where reviewed for their distance to the left/right neighbour to assure at least 2kb distance. Furthermore pairs were supervised for overlapping or intermediate Ensembl transcripts. The existance of multiple transcripts for the orthologous genes remained uninspected. For *H2K-2K* a distance of 2,000bp was added left/right around each pair while *H2K-next* includes the full sequence up to the neighbouring transcripts (as annotated by Ensembl). Again, if regional overlaps occured, one of the overlapping pairs was skipped to avoid duplicated sequences.

Dataset *H2K-2K* comprises **130 sequences** (**35 positive** and **96 negative**). Dataset *H2K-next* comprises **128 sequences** (**35 positive** and **93 negative**)[2].

[2]For a full annotation of the datasets see **Appendix A**.

Retrieve and Process of Sequence Data and Features

The following pages contain detailed information about the used procedures to extract nucleotide sequence for the defined datasets and the extraction of several features (e.g. repeats, phylogenetic conserved regions) that are distributed over the regions of interest, as well as the combination of these two procedures to generate masked sequences.

2.4 Sequence Extraction

The script ***SequenceExtractor.pl*** extracts the genomic mouse or human sequence of a specified region on a chromosome.

The current Mouse February 2006 (mm8) assembly from NCBI (Build 36) and the current Human March 2006 (hg18) assembly from NCBI (Build 35) was downloaded from the UCSC website[3] and stored in fasta format with one file per chromosome.
The script **extracts nucleotide sequences from all autosomes plus X and Y chromosome**. Sequence contained in the "*_random*" files and the "*M(itochondrial)*" and "*Un(mapped clone contigs)*" files is not included.
To extract the appropriate sequence the following attributes are required by the script: assembly (mm8 or hg18), the chromosome (e.g. 1, X), the *inclusive* start and end positions and strand annotation (+ or -). If strand is specified as '-', the sequence will be returned as its reverse complement. It is possible to format the sequence to upper or lower case letters and to output it in fasta format (containing 50 chars per line).

2.5 Feature Extraction

The feature extraction procedure is accomplished by the script ***FeatureExtractor.pl***[3] which extracts annotations of transcripts, repeats, regulatory potential, and other features for a specified region on a chromosome and returns these annotations as a list and/or masking string.

[3] `http://genome.ucsc.edu/`

2.5.1 Transcribed Regions and Transcriptional Start Sites

We assume it to be unlikely, that transcribed regions contain bindings sites for TFs that could lead to the observed co-expression of genomic neighbours. Therefore these were excluded (masked) from the motif finding process. However, TFs that bind to transcribed regions (e.g. introns) are known. Nevertheless we suppose these to be of minor impact to the regulatory mechanisms we wanted to examine.

Data
Transcriptional information (including chromosome, start/end and strand annotation) was obtained for all transcripts (represented by Ensembl.Transcript.IDs) from Ensembl (Ensembl 39; Mouse: NCBI m36 Assembly (Dec 2005) & Human: NCBI 36 Assembly (Oct 2005); Genebuild Ensembl (Jun 2006)) via "BioMart".
Transcript annotations were stored one file per chromosome and included the following fields:

Field	Example	Description
genoName	chr1	Genomic sequence name
genoStart	3000001	Start in genomic sequence
genoEnd	3000156	End in genomic sequence
strand	-1	Relative orientation 1 or -1
id	ENSMUST00000015346	Ensembl.Transcript.ID (mouse)

Approach The extraction process searches trough the whole data file of the specified chromosome for transcripts that are localized in the region of interest. Every transcript with the start and/or end position (*genoStart* and *genoEnd*) between the start and end position of the specified region is extracted. If a transcript overlaps either the start or the end of the region or both, its start/end positions are "cut" to that of the specified region.
The **TSS** is annotated using the start/end position as annotated by Ensembl, depending on the strand annotation of the transcript. It is possible to add a specified number of n nucleotides to the left/right of the annotated TSS to obtain a **TSS window**.

2.5.2 Repeats

Repeats are repetitive sequence elements that can occur in multiple regions of the genome. Some groups propose a masking of these interspersed elements prior to the motif search to **reduce the noise level in the sequence data** [74]. Nevertheless, the presence of certain repeats is **likely to play a biological role in transcriptional control** [61],[75].

Data The repeat information was extracted from UCSC RepeatMasker annotation track (mm8/hg18). It was created using the Arian Smit's RepeatMasker program[4], which screens DNA sequences for interspersed repeats and low complexity DNA sequences. Repeats were classified into several subgroups and the masking function uses the "repeat class" tag to annotate the repeats existing in our sequences. Furthermore the possibility to **exclude specific repeats from the masking process** was added.
The downloaded annotation files (one per chromosome) contains repeat annotations in the following format[5]:

Field	Example	Description
bin	607	Indexing field to speed chromosome range queries.
swScore	687	Smith Waterman alignment score
milliDiv	174	Base mismatches in parts per thousand
milliDe	l0	Bases deleted in parts per thousand
milliIns	0	Bases inserted in parts per thousand
genoName	chr1	Genomic sequence name
genoStart	3000001	Start in genomic sequence
genoEnd	3000156	End in genomic sequence
genoLeft	-194069806	Size left in genomic sequence
strand	-	Relative orientation + or -
repName	L1_Mur2	Name of repeat
repClass	LINE	Class of repeat
repFamily	L1	Family of repeat
repStart	-4310	Start in repeat sequence
repEnd	1551	End in repeat sequence
repLeft	1397	Size left in repeat sequence
id	1	First digit of id field in RepeatMasker .out file.

Approach Repeat positions were extracted in the same fashion as for transcribed regions (see above).

2.5.3 Regions with Regulatory Potential

Including conservational information, also called **phylogenetic footprinting**, into the search for regulatory elements is a widely recommended approach [76],[77],[74],[78],[79]. It is based on the assumption that regulatory elements (e.g. TFBS) are evolutionary stable, while bulk DNA is free to mutate.

[4] `http://www.repeatmasker.org`
[5] Description taken from UCSC website

Several approaches have been made to extract phylogenetically conserved sequence parts, ranging from straightforward two-species alignment and percentage-conservation windows [80],[81] to more complex approaches as the PhastCons score derived from a phylogenetic hidden Markov model [82].
To extract sequence stretches of potential regulatory function the **Regulatory Potential (RP) Score** [83] which was developed by members of the *Comparative Genomics and Bioinformatics Center* at *Penn State University* was used. The RP score (together with the PhastCons score) has already been shown to **successfully extract cis-regulatory modules** in the β-globin gene cluster[84].

RP scores are derived from the comparison of two **hidden Markov models (HMMs)** which were trained using frequencies of short multiple alignment patterns in **regions of known regulatory elements and ancestral repeats**. In this approach the ancestral repeats act as a model of neutral DNA. The multiple alignments used to build the HMMs were calculated using the following assemblies of 7 vertebrate species:

- **human** (Feb 2006, hg18)
- **chimpanzee** (Jan 2006, panTro2)
- **macaque** (Jan 2006, rheMac2)
- **mouse** (Feb 2006, mm8)
- **rat** (Nov 2004, rn4)
- **dog** (May 2005, canFam2)
- **cow** (Mar 2005, bosTau2)

Each resulting alignment column was represented using a **collapsed alphabet** (collapsed means that two distinct alignment columns might share a certain alphabet symbol) and hidden Markov models were trained on short k-mers of the resulting sequence. The composition and frequency of these short k-mers is supposed to differ between multiple-alignments of real regulatory sequence elements and neutral DNA. The RP score is calculated from the **log-ratio of the transition probabilities of the two hidden Markov models**.
The calibration study performed by King et al [84] suggested a threshold of >0 for identifying potential regulatory elements.

Data RP scores are availible at the UCSC Genome Browser for all of the included assemblies. For the analysis the mm8 RP score data was downloaded, which exist in a very simple format, that displays increasing genomic positions and their appropriate RP scores in one line. Before accomplishing further extractions every position in the DNA that had a RP score of 0 was removed (to decrease running time).

Approach The RP extraction function searches through the specific chromosome file until a position is reached that is localised inside of or directly at the start of the region of interest. Then, for every continuous stretch of position-score tuples, it stores the inclusive start and end position, until a position is encountered that is localised beyond the end of the region of interest.

2.5.4 CpG Island/Regions

CpG island are regions that comprise a **high C+G content and a higher-than-average number of the CpG dinucleotides** (which is significantly underrepresented in vertebrates genomes [85]). CpG islands are present in the promoter and exonic regions of **approximately 40-60% of the mammalian genes** [86], and have been proposed to play a role in processes such as housekeeping gene functionality [87]. As the definition of a CpG island is somewhat arbitrary **two different approaches** were used. The first is a strict approach called "**CpG Islands**" the second approach, "**CpG Regions**" uses less constraints.

CpG Islands

Data To extract CpG islands the existing CpG island annotation available from UCSC Genome Browser for mm8/hg18 was downloaded. It is derived from the **CpG island definition by Garden-Gardiner** [87].
The downloaded annotation file contains CpG annotations in the following format[6]:

Field	Example	Description
chrom	chr1	Reference sequence chromosome or scaffold
chromStart	18598	Start position in chromosome
chromEnd	19673	End position in chromosome
name	CpG: 116	Name of CpG island
length	1075	Island length
cpgNum	116	Number of CpGs in island
gcNum	787	Number of C and G in island
perCpg	21.6	Percentage of island that is CpG
perGc	73.2	Percentage of island that is C or G
obsExp	0.83	Ratio of observed(cpgNum) to expected(numC*numG/length) CpG in island

[6]Description taken from UCSC website

Approach CpG island positions were extracted in the same fashion as for transcribed regions (see above).

CpG Regions

The second approach also follows the definition from Garden-Gardiner, but uses less constraints for the existence of a CpG island (which is therefore called CpG region). A sliding window of 100bp length was analysed over the whole (unmasked) sequence derived by *SequenceExtractor.pl*(see above). A region was marked as "CpG regions" if it fulfilled the following conditions:

- The **GC content** is greater than 50%
- The **length of the region** is at least 200 bp
- The **ratio between the observed number of CG dinucleotides and the expected number** is greater or equal to 0.6

The ratio between observed and expected GC dinucleotides is computed using the formula by Gardiner-Garden [87]:

$$\frac{Obs}{Exp}CpG = \frac{\text{Number of CpG dinucleotides} \times N}{\text{Number of Cs} \times \text{Number of Gs}}$$

where N is the length of the sliding window.
Following this definition, every CpG island is also (at least a subset of) a CpG region.

2.5.5 Specific Binding Sites

GC Boxes

The GC Box is the hexanucleotide sequence **"GGGCGGG"** (or it reverse complement **"CCCGCCC"**), which is also the consensus sequence for the transcription factor **SP1** [87]. Because this signal is much easier to locate in genomic sequences than the appropriate transcription factor bindings site motif of SP1 (see below), it was located in addition to the search for possible Sp1 binding sites as described below. The disadvantage of a search for this fixed nucleotide sequence instead of using a weigthed matrix model is the larger number of probable false positive sites because of the shorter sequence length (compared to the Sp1 binding site).

Approach To extract GC Box positions in our regions of interest the sequence of the region (as obtained from the script *SequenceExtractor.pl*) was scanned for the substring "GGGCGGG" and "CCCGCCC" using **regular expression matching** and found sites were recorded. A window of specified size can be added to each site of the found GC Box position to amplify the signal.

CTCF Binding Sites

The protein **CTCF - a 11-zinc finger protein - is a known insulator** which represses heterochromatin from entering euchromatic regions and is therefore supposed to reside at the edges of open chromatin. It has also been shown to **block the advance of RNA polymerase II** [88]. The main difficulty of locating CTCF binding sites in DNA sequences using *in-silico* techniques is it's affinity to bind **different binding sites** engaging different subsets of zinc fingers [89]. Nevertheless a binding site for CTCF has been derived by several groups. It consists of the consensus sequence **"CCGCNNGGNGGCAG"** (or its reverse complement **"CTGCCNCCNNGCGG"**) [90],[91].

Approach To extract possible binding sites of CTCF, the consensus sequence and its reverse complement were located in the sequences. As proposed by the authors of [90] every match with at least **13 matching positions** ("N" is always a match) was stated as a possible binding site. The search was performed using regular expression matching for all sequences that could be derived from the consensus (and its reverse complement) by changing one more nucleotide into "N".

Specific TFBS Using TRANSFAC Matrices

In addition to the location of possible binding sites using consensus sequences, a search for specific **transcription factor bindings site motifs** present in the **TRANSFAC database**[7] was implemented. The two binding sites investigated were V$SP1_Q6_01 (Figure 2.3) and V$TATA_01 (Figure 2.4) representing the transcription factor **Sp1** and the **TATA box**, a motif common in eukaryotic gene promoters.

Approach To search for the two presented motifs the motif search program **MAST** [92] was used which will be presented in detail in section 2.9. Mast can find transcription factor binding site motifs in nucleotide sequences. The appropriate motifs for SP1 and the TATA box were extracted by hand and converted into a format that is readable by MAST. The search was performed on the sequences without any masking. As MAST uses **pvalues** and **Evalues** to secure credible matching a pvalue of *0.1* and an Evalue

[7] `http://www.biobase.de/cgi-bin/biobase/transfac/start.cgi`

Figure 2.3: Sequence logo visualising the TRANSFAC matrix of the Sp1 protein (*V$SP1_Q6_01*).

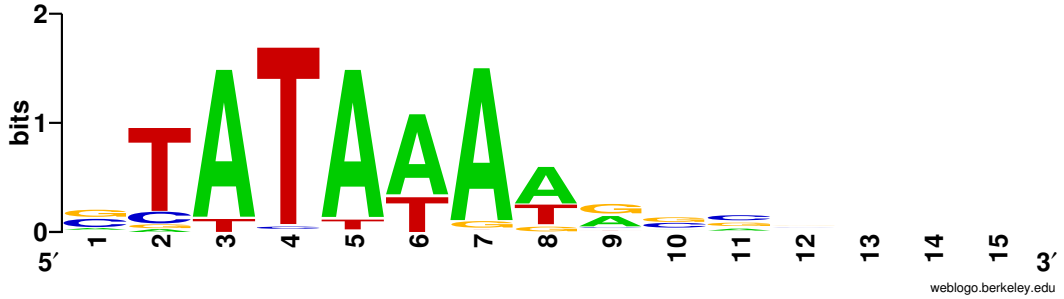


Figure 2.4: Sequence logo visualising the TRANSFAC matrix of the *TATA-Box* (*V$TATA_C*).

of *100* was used to allow the finding of possible binding sites even in the probably long sequences.

2.5.6 Final Feature Extraction Output

Each potential annotation feature has an assigned symbol which is a **single character representation of that feature**. The whole list of features and their appropriate symbols can be found in table 2.4.

All feature extraction processes output a list with inclusive start/end annotation together with the appropriate features. In concatenating and sorting all resulting lists by their start annotation, a **full feature list** is produced, which represents all extracted features of the region of interest. Due to the fact that a genomic region can be annotated by several distinct features, this list can **include overlaps** between the annotated regions. The full masking list is then used to build a **feature string**, which is a **base by base representation** of the sequence features. In this feature string every base position is assigned a single character that stands for its feature according to the regions specified in the full feature list. **Overlaps** between feature annotations (causing a single base

position to be annotated by several features) lead to the assignment of '∼' as the specific overlap symbol. Nucleotide positions that lack any feature are represented by a '-'. Afterwards, this feature string is again retranslated into a **non-overlapping feature list**, which now contains no overlapping annotations anymore, but inclusive start/end positions for non-overlapping feature regions and overlap regions. Regions that lack any features are not contained in this list.

All three output formats, the **non-overlapping feature list**, **feature string**, and **full (potentially overlapping) feature list**, are returned for further analysis. In addition some meta data is returned, providing the assembly, chromosome, inclusive start and end positions of the region of interest and the features that have been extracted together with the used parameters.

Transcript Features		*Repeats*	
Transcript	#	SINE	B
TSS	!	Simple-repeats	D
Phylogenetic Features		LINE	E
		LTR	F
Regulatory Potential	$	Low_complexity	H
CpG Features		DNA	I
		Other	J
CpG island	§	scRNA	K
CpG Region	?	tRNA	L
Binding Site Features		snRNA	M
		Unknown	0
GC Box	°	rRNA	P
Motif 1-9	1-9	Satellite	Q
Insulators		RNA	R
		srpRNA	R
CTCF Binding Site	ˆ	*Overlap*	
No Feature		Overlap	∼
No Feature	-		

Table 2.4: Full list of masked features and their assigned signs

2.6 Sequence Masking

The sequence masking procedure, merges the data from the "Sequence Extraction" and "Feature Extraction" procedures into one masked sequence, using the script ***Sequence-Masker.pl***[8].

[8]See **Appendix B** for a description of scripts

After assuring that sequence data and masking data correspond (using the meta data saved along with the computed results) the non-overlapping masking list is used to successively mask regions annotated by a certain genomic feature to pretend these regions from being included into the motif search.

Depending on the usage of regulatory potential information, the sequence masking procedure returns different output:

2.6.1 Without Regulatory Potential Information

In the case of missing regulatory potential information the function masks every region that is assigned with a feature as transcript or a specific repeat class. All blocks not assigned with such a feature contain the appropriate nucleotide sequences. The masked parts are replaced according to the user-specified **masking mode**:

Masking Mode	Cut-Out Replacement
1	The appropriate number (length of cut-out region) of repetitions of the appropriate feature character
2	A single feature character for the whole region spanned by the feature
3	The appropriate number (length of cut-out region) of repetitions of an unspecific wildcards (assigned by the user or default to 'N')
4	A single unspecific wildcard (assigned by the user or default to 'N')

2.6.2 With Regulatory Potential Information

If regulatory potential information is present, the sequence masking procedure masks all those parts, that are not assigned to have a regulatory potential. Because the non-overlapping masking list is used, even parts that are conserved, but overlapped by other features such as transcripts or repeats, are masked. The masked parts are replaced according to the masking mode as described above.

Retrieving and Processing Sequence Data and Features

The following pages contain detailed information about what motif finding algorithms were used and how the resulting motifs were scored and compared to gain a set of unique and overrepresented motifs for the positive dataset. Furthermore a description of how the found motifs were compared to known vertebrate transcription factor binding site matrices is given. The whole searching procedure was performed on the two datasets *2K-2K* and *2K-next* using different masking conditions.

2.7 Motif Finding Algorithms

As proposed by a motif finding tool competition and already performed by other groups ([76],[93]) **several motif finding algorithms** were integrated into the motif search to increase the number of identified motifs. Motif finding algorithms that were **based on different finding strategies** were used to overcome possible loss of motifs resulting from specific characteristics of certain searching strategies. All used motif finding algorithms are freely available for academical purpose and were downloaded and installed as a local copy. All four programs use FASTA-formated files of the positive dataset as input, which were generated using the masked/processed sequences of the different datasets by the perl script ***PreMotifFinder.pl***[9].

2.7.1 MEME

MEME was developed by Bailey and Elkan [94] and is provided by the *Department of Computer Science and Engineering* at the *University of California at San Diego.*

It uses a modified form of the **expectation maximisation (EM) algorithm** to fit a two-component finite mixture model to a given set of (nucleotide) sequences. The two-component finite mixture model [94] consists of one component that represents a motif (multiple occurrences of a specific subsequence) of variable length and a second component that models the background. For the second component an optionally Markov background model of any order can be provided by the user (if not, a 0th-order background model will be estimated from the given sequences). MEME allows the specification of a model for the distribution of the motifs to search for, which can either be contained exactly **one time** in every sequence, **one or zero times** in every sequence or a **user-defined number of repetitions** in every sequence. MEME also provides

[9]See **Appendix B** for a description of scripts

the possibility to weight the input sequences. Furthermore, MEME is capable of finding multiple motifs by applying its algorithm several times to the dataset starting from different initial points in the search space.

MEME was used with the default parameters together with the "-dna" switch which indicates the use of a DNA alphabet and the "-revcomp" switch which allows the motif to occur on either + or - strand. The "zero or one occurrence per sequence" motif distribution was selected. The maximum motif width was set to 25 and the number of output motifs was set to 10. The default sequence-estimated 0th-order background model was used and no weights have been specified for the input sequences.

2.7.2 BioProspector

BioProspector was developed by Liu, Liu, and Brutlag [95] and is provided by *Stanford Medical Informatics* at the *Stanford University*.

BioProspector uses a **Gibbs sampler algorithm** to find overrepresented motifs of a certain fixed size in a database of (nucleotide) sequences. Gibbs sampling[10] strategies in motif finding are a heuristic and probabilistic method to optimize local multiple alignments in a dataset of sequences using a strategy that is very close to the **Monte Carlo Markov chain** algorithm. BioProspector uses a **3rd-order Markov model** to model the background, which is generated by a user-specified database of sequences, which can be equal to the input sequences. It overcomes the Gibbs sampling problem of the proposed occurence of the motif in every single input sequence by using a "two threshold strategy". It separates *sure* and *unsure* subsequences from *improbable* ones and is therefore also called a **threshold sampler**.
BioProspector can find multiple motifs by repeated runs from different start points in the search space.

The BioProspector program was used with the default parameters. As background sequences the whole negative sequence set was used, as provided in FASTA-format by ***PreMotifFinder.pl*** (see above). Because the motif width was unknown but must be specified for BioProspector three runs defining the motif width as 10, 15, and 20 nucleotides, respectively, were performed.

2.7.3 AlignACE

AlignACE was developed by Hughes et al. [96] and is provided by the *Department of Genetics* at *Harvard Medical School*.

[10]The algorithm was firstly introduced 1984 by S. Geman and D. Geman for the use in pattern analysis

AlignACE is another **Gibbs sampler** and works in a similar fashion as BioProspector, but uses **GC content** to approximate the background. It needs no fixed motif width, but can be started with a user-defined expectation. It only uses the **10 most informative positions** to sample a motif and lets the other positions in the motifs evolve unoptimised. AlignACE is capable of finding multiple motifs by successive masking of most informative sites for found motifs in the sequences and then iterate its search, pretending to use these masked sites.

AlignACE was used with the default parameters. The background CG content was set to the calculated GC content from the appropriate negative dataset. The "oversample" parameter was set to 5, leading to an exhaustive search but increasing runtime.

2.7.4 Improbizer

Improbizer was written by Kent [97] and is provided by *University of California Santa Cruz*.

The program is another **expectation maximisation algorithm** which determines DNA motifs - represented by position specific weight matrices - that are overrepresented in a given sequence database, compared to the background distribution which is specified by a 2nd-order Markov model, estimated from a user-specified sequence database. In the first step, an **initial-motif** is produced by using all subsequences of the first 10 sequences, match these to the first 20 sequences and keep the most promising subsequences for further improvement. In the next step it then iteratively collects the matches and near matches for all motifs over all sequences and averages them together to create a new motif. The algorithm stops after it converges.
The Improbizer program is capable of finding multiple motifs by starting from different initial points in the search space.

Improbizer was used with the default parameters. The sequences of the whole negative sequence set, as provided in FASTA-format by ***PreMotifFinder.pl*** (see above), were assignes as negative dataset. The "ignoreLocation" and "rcToo" switches were set to "on" to allow for motifs on both strand and in arbitrary locations in the sequence. The number of ouput motifs was set to 5.

2.8 Representation of Found Motifs

All motif finding algorithms output representations of their found motifs, but using different formats. The script ***PostMotifFinder.pl***[11] contains functions to parse all these motif finder output files into **one common motif file format** (called *mot(if)* format), which was designed for an easy access of the found motifs for further processing. It contains the number and three representations of every found motif:

1. **Multiple alignment of sites**
   This representation is the output of the most motif finding algorithms used. It represents the motif as a **multiple alignment of found sites** in the input sequences using a fixed width.

   Example:
   ```
   CCCCCGCCCA
   GCCCCGCCCC
   CGCCCGCCGC
   GCCCCGCCCC
   GCCCCGCCCC
   CCCCCGCCCG
   ...
   ```

   If no multiple alignment of found sites was present in the motif finder output (as, for example, in the case of *Improbizer*), a set of sequences that closely approximate the presented motif was generated. The approximation bases on the fact that each column of a position-dependent frequency matrix (see below) is independent of all the others. For every column a defined-length set of nucleotides was generated that follows the distribution of that column of the motif. Afterwards the shuffled sets were concatenated into a multiple alignment.

2. **Position-specific frequency matrix**
   A **position-specific frequency matrix (PSFM)** (also called **position frequency matrix (PFM)**) shows the fraction of each of the nucleotides A, C, G, and T at a specific position in the motif. Two different representations are known, one shows the **total count of the appropriate nucleotide**, the other its **frequency**. Each found motif in the *mot* files was stored using the second representation.

   Example:

Column	A	C	G	T
1	0.000000	0.283582	0.716418	0.000000
2	0.000000	0.701493	0.298507	0.000000
3			...	

[11] See **Appendix B** for a description of scripts

3. **Position-specific scoring matrix**
   The **position-specific scoring matrix (PSSM)** (also called **position weight matrix** (PWM)) is calculated from a PSFM using logarithmic values instead of probabilities. Two distinct forms exist, the first using **log-likelihood values** and the second using **log-odds scores**, which additionally include the **nucleotide background distribution** into the PSSM. Elements in such a log-odds PSSM are calculated in the following way:

   $$m_{ij} = \log\left(\frac{p_{ij}}{b_i}\right)$$

   where p_{ij} is the probability of observing nucleotide i at position j in the motif, and b_i is the background probability of nucleotide i.
   This representation was used to store found motifs in the *mot* files. Because a percentage of "0" would lead to log(0), each zero entry of the PSFM was assigned the lowest possible value before calculating the PSSM.

   Example:

Column	A	C	G	T
1	-1791	18	152	-1794
2	-1791	148	25	-1794
3		...		

2.9 Search Found Motifs in the Dataset using MAST

After a specific motif has been found by one of the motif finding algorithms, its localisation in the sequences of the whole dataset must be determined. Several tools exist for these localisation process, ranging from very easy scoring using only the PSFM or PSSM, to advanced motif localisers that additionally include statistical considerations (see for example [92] and [96]).

In this analysis the program **MAST** written by Timothy L. Bailey and Michael Gribskov [92] was used which is provided by the *San Diego Supercomputer Center*.

MAST searches motifs represented by PSSM in a provided sequence database. Instead of scoring a match solely based on the score derived from the match of the PSSM against the sequence, it uses a **Fisher "omnibus" procedure** (see [92]) to provide statistical pvalues for a motif occurrence in a sequence. These pvalues are calculated from a **random sequence model** based on the average nucleotide frequencies of the provided sequences. Using the *-comp* switch the letter frequencies are adjusted for every sequence instead of the whole database. MAST provides three different types of pvalues:

1. **Position pvalue** = The probability that a randomly selected position in a randomly generated sequence of the same length as the matching sequence has a match score at least as large as the match score of the best matching position of that sequence to the motif.

2. **Sequence pvalue** = The probability that a randomly generated sequence of the same length as the input sequence would achieve a match score that is at least as large as the match score of the best matching position of that sequence to the motif.

3. **Evalue** = The Evalue is the expected number of sequences in a randomly generated database of the same length that would match the motif as good as the sequence does.

MAST output includes a list of all sequences that match a given motif with a position/sequence pvalue and Evalue less than specified by the user. The name of each sequence is provided along with the position of each match. For evaluation of reliable matches of found motifs to the sequence database **sequence pvalues** ranging from *0.01* to *0.05* and **Evalues** ranging from *2* to *10* together with the *-comp* switch were used.

2.10 Score Found Motifs

The motif finding process was designed to identify potential TFBS that reveal the mechanism of highly correlated co-expression of genomic neighbours. The positive and negative datasets were build up of sequences that should include or lack these TFBS, respectively. To **score the significance of a found motif** in respect to its distribution between the positive and negative dataset, several distinct scores were used, which are presented in the following. The script ***PostMotifFinder.pl*** (see above) calculates these scores for every motif found by any of the used motif finding algorithms using the script ***ScoreMotifs.pl***[12]. While only these scores were included into the final PDF output, the matching sequences for each motif and the appropriate MAST output files were stored as ancillary data files.

2.10.1 Group Count and Frequency

The **Group Count** of a found motif is the total number of sequences in a dataset that contain at least one site that matches the motif as calculated by MAST. Multiple occurrences of a single motif in one sequence are counted as single occurrences. In

[12]See **Appendix B** for a description of scripts

our setup, each found motif is assigned such a group count for its occurrences in the positive ("+"-count) and in the negative ("−"-count) dataset. The respective **Group Frequencys** are calculated by

$$f_{pos} = \frac{\text{“+”-count}}{\text{\# of sequences in positive dataset}}$$

and

$$f_{neg} = \frac{\text{“-”-count}}{\text{\# of sequences in negative dataset}}$$

2.10.2 Ratio of Group Frequencies (+/− Ratio)

The **Ratio of Group Frequencies (+/− Ratio)** is the ratio between f_{pos} and f_{neg} and is calculated by

$$R_{+/-} = \frac{f_{pos}}{f_{neg}}$$

The value measures the proportional difference ("fold") of the group frequencies for a found motif between the positive and negative dataset. The advantage of **comparing frequencies instead of total group counts** is the unequal number of sequences in the two distinct datasets. This score becomes "1" if the frequencies are equal and increases/decreases if the motif is more frequent in the positive/negative dataset. Because we are searching for motifs that are very specific for the positive dataset but are less frequently present in the negative dataset, a candidate motif should have a +/− Ratio greater 1.

2.10.3 Group Specificity Score

The **Group Specificity Score (GSS)** was introduced for motif finding by Hughes et al. [96] and was used to score the significance of found motifs in several studies [76],[96],[98],[99]. Its usability in discriminating real binding sites from background noise (or in our case uniformly distributed occurrences in the whole positive and negative dataset) was shown for example in [100].

The Group Specificity Score measures the affiliation of a found motif to the sequences it was computed from, in respect to all possible sequences. It is calculated using the hypergeometric distribution:

$$S_{group} = \sum_{i=x}^{\min(s_1,s_2)} \frac{\binom{s_1}{i}\binom{N-s_1}{s_2-i}}{\binom{N}{s_2}}$$

where N is the total number of sequences in the whole (positive and negative) dataset, s_1 is the number of sequences used to find the motif (this is the number of sequences in the positive dataset), s_2 is the number of sequences in the whole dataset that have an occurrence of the motif, and x is the number of sequences in the intersection of s_1 and s_2 (the number of sequences in the positive dataset that have an occurrence of the motif). Each term of the sum calculates the probability of having obtained an intersection of i sequences between the set of sequences containing the motif and those used to find it assuming a random sampling of the two sets. The sum S_{group} is therefore the probability of observing the actual intersection or a greater one. It ranges between 0 and 1.
Briefly, the Group Specificity Scores gives an **impression how specific a found motif is for the positive dataset**, in terms of probability of observing this distribution between positive and negative dataset. A candidate motif is expected to have a very low Group Specificity Score, meaning it is very improbable to see these distribution leading in the direction of the positive dataset by chance.

Some groups have used the **Site Specificity Score** instead of the GSS to score the significance of their found motifs (see for example [101]). The Site Specificity Score is calculated using the same distribution but with N standing for the total number of sites in the dataset (total number of nucleotides), s_1 for the number of nucleotides in the sequences used to find the motif and s_2 for the number of sites targeted by the found motif. The Site Specificity Score accounts for multiple occurrences of a single found motif in the input sequences and might become a better choice, if most of the sequences in the negative and positive dataset have at least one occurrence of the found motif [76]. We decided to use Group Specificity Score instead of Site Specificity Score because we did not a priori expect multiple occurrences of single TFBS in our positive dataset.

2.11 Comparison and Selection of Found Motifs

Using four different motif finding algorithms it is very probable that a specific motif will be postulated more than one time in eventually slightly different form. Therefore it is highly recommended to compare the found motifs and select the best scoring member of every group of similar motifs and eliminate the poorer scoring redundant motifs.

To compare two found motifs the tool **CompareACE** was used which is also provided by Hughes et al. [96] together with their motif finding algorithm **AlignACE**. It calculates the **Pearson correlation coefficient** between the PSFM of two motifs using only the 6 most informative positions of the first motif. The script ***PostMotifFinder.pl*** (see above) calculates this Pearson correlation coefficient for every possible pair of found motifs and then uses **Tree**, which is provided along with CompareACE to hierarchically cluster the found motifs. The cluster-cluster score is set to be the average of all pairwise scores of motifs between the two clusters. A cut-off correlation coefficient must be set

to determine final distance of clusters. For the clustering of found motifs a correlation coefficient cut-off of *0.6* was used.

After clusters were calculated and stored in text files, the best scoring motif in every cluster was extracted and the others were removed. GSS was used to select the best motif (lowest score). For every selected motif, the script ***PostMotifFinder.pl*** (see above) returns its number together with all calculated scores (Group Counts, Group Frequencies, $R_{+/-}$, and S_{group}).

2.12 Draw Sequence Logos for Found Motifs

After selecting the best scoring motifs and removing redundancy, a sequence logo is drawn for every final motif. Sequence logos for multiple alignments have been developed by Tom Schneider and Mike Stephens [102]. An exemplary sequence logo is shown in Figure 2.5.

A sequence logo is a graphical representation of a multiple alignment and illustrate three position-specific information:

1. **Relative frequency of each nucleotide**

2. **Order of predominance of each nucleotide**

3. **Information content in bits**

The relative nucleotide frequency at each position represented by the height of the four letter "A", "C", "G", and "T", which is calculated

$$h_{ij} = p_{ij} * I_j$$

where h_{ij} is the height of the nucleotide letter i at position j in the motif, p_{ij} is the appropriate probability of observing these nucleotide at that position in the sequence and I_j is the information content of the sequence at position j. The information content is defined as

$$I_j = 2 - U_j + e(n)$$

where 2 is the maximal possible uncertainty at a position based on 4 possible letters, $e(n)$ is a correction factor that is required if only a few samples alignment sequences are present and U_j is the uncertainty at position j, which is evaluated using the formula

$$U_j = -\sum_{i \in \{\mathrm{A,C,G,T}\}} p_{ij} \log_2 p_{ij}$$

Figure 2.5: Sequence logo visualizing the TRANSFAC matrix of the *myogenic MADS factor MEF-2* (*V$MEF2_04*)

Letters are sorted to put the most frequent nucleotide letter at top.

Sequence logos for all found motifs were drawn using a local installation of the web-tool **WebLogo**[13][103]. The program was used with the standard parameters but adding colours and axis labels.

2.13 Assign Known TFBS Matrices to Found Motifs

To compare the found motifs to already known TFBS current matrix data from TRANSFAC Professional 10.2 (BIOBASE Biological Databases, Germany) was downloaded.

To obtain a large set of known TFBS, all present **TFBS matrices** that were generated from **vertebrate data** (indicated by the "V" designator in the matrix identifier - e.g. *V$AP1_Q1*) were filtered. Even if TRANSFAC identifier include a quality code for their provided TFBS matrices (in the example above "*Q1*" means quality 1 which is the highest quality) all provided matrices were used in the comparison to the found motifs, regardless of their quality. From a total of 811 present TFBS matrices **584 vertebrate matrices** were extracted and all of them were converted to the *mot* file format, which is described above.

After these preprocessing steps, each final motif was compared to all vertebrate TFBS matrices using the same strategy as used in the comparison of all motif finding algorithm output motifs. A Pearson correlation coefficient of at least **0.6** was used to assign known TFBS to the found motifs. If several TFBS match a certain motif, the one with the **highest correlation coefficient** was included in the final PDF output together with the total number of matching TFBS. Nevertheless, all TFBS matching with a correlation coefficient of at least **0.6** were stored in the ancillary data files.

[13] `http://weblogo.berkeley.edu/`

To manually compare the similarity of a found motif to its assigned TFBS matrix, Sequence Logos for the appropriate TFBS matrix were drawn using the WebLogo tool and included in the PDF output.

2.14 Search for TRANSFAC Matrices in the Datasets

In addition to searching for common motifs using several motif finding algorithms (as described in section 2.7), the search for known TFBS from the TRANSFAC database was implemented. This approach is somewhat **opponsite to the motif search** described previously. In this case the search is directed from existing PSSM for known TFBS that are searched for in the sequence datasets (instead of searching for overrepresented motifs in the sequences and afterwards comparing these to known TFBS).

2.14.1 Search for All Vertebrate Matrices

A search for all existing vertebrate TRANSFAC (TRANSFAC Professional 10.2) PSSM was performed using a similar strategy as described in the sections 2.9 to 2.13. The TRANSFAC PSSM were used as input files for the script **PostMotifFinder.pl** (see above).
Using **MAST** a search for all **584 previously extracted vertebrate matrices**, both (positive and negative) dataset, was performed. Sequence pvalues ranging from *0.01* to *0.05* and Evalues ranging from *2* to *10* together with the *-comp* switch were used for the evaluation of reliable matches of vertebrate PSSM to the sequence database.
After computing all matches of a specific PSSM using MAST the same scores as described in section 2.10 were computed and PSSMs were clustered with a correlation coefficient cut-off of *0.7*. For the best-ranking (due to Group Specificity Score) motif in each cluster a sequence logos was drawn. Finally a PDF output file was created in the same fashion as described previously.

2.14.2 Searching Matrices of Nuclear Receptors

Nuclear receptors (NRs) have been shown to interact in a sequence specific manner with histones and histone modifying proteins and therefore to influence expression of targeted genes [21]. Because of the **possible role of NRs in the establishment of euchromatic regions** an additional TFBS search for the subgroup of TRANSFAC PSSM that represent NR TFBS (shown in Table 2.5) was performed with the same settings used in the overall TFBS search.

NR Symbol	**NR Name**	**TRANSFAC Matrices**
Esr1	estrogen receptor 1 alpha	VER_Q6, VER_Q6_02
Esr2	estrogen receptor 2 beta	V$ER_Q6_02
Esrra	estrogen related receptor, alpha	V$ERR1_Q2
Hnf4a	hepatic nuclear factor 4, alpha	V$DR1_Q3,V$HNF4_DR1_Q3, V$HNF4_Q6,V$HNF4_Q6_01, V$HNF4_Q6_02, V$HNF4_Q6_03, V$HNF4ALPHA_Q6
Hnf4g	hepatocyte nuclear factor 4, gamma	V$DR1_Q3,V$HNF4_DR1_Q3, V$HNF4_Q6
Nr3c1	nuclear receptor subfamily 3, group C, member 1	VGR_Q6,VGR_Q6_01, VGRE_C,VPR_Q2, V$GR_01
Ppara	peroxisome proliferator activated receptor alpha	V$PPAR_DR1_Q2, V$PPARA_01, V$PPARA_02
Rora	RAR-related orphan receptor alpha	V$RORA1_01, V$RORA2_01
RARA	retinoic acid receptor, alpha	V$DR4_Q2, V$T3R_Q6
RARB	retinoic acid receptor, beta	V$T3R_Q6,V$DR4_Q2, V$T3R_Q6
Rarg	retinoic acid receptor, gamma	V$DR4_Q2, V$T3R_Q6
RXRA	retinoid X receptor alpha	V$DR4_Q2,V$PPARA_02, V$T3R_Q6, V$DR3_Q4
RXRB	retinoid X receptor beta	V$DR3_Q4,V$T3R_Q6, V$DR4_Q2
RXRG	retinoid X receptor gamma	V$T3R_Q6
Thra	thyroid hormone receptor alpha	V$T3R_Q6
Nr2f1	nuclear receptor subfamily 2, group F, member 1	V$COUP_01, V$COUP_DR1_Q6, V$COUPTF_Q6,V$DR1_Q3, V$DR4_Q2, V$HNF4_Q6
Nr2f2	nuclear receptor subfamily 2, group F, member 2	V$ARP1_01, V$COUP_DR1_Q6, V$COUPTF_Q6,V$DR1_Q3, V$DR4_Q2, V$HNF4_Q6
Nr1h3	nuclear receptor subfamily 1, group H, member 3	V$DR4_Q2,V$LXR_DR4_Q3, VLXR_Q3, VPXR_Q2
Nr1h2	nuclear receptor subfamily 1, group H, member 2	V$DR4_Q2, V$LXR_Q3

NR Symbol	NR Name	TRANSFAC Matrices
*Mef2a**	myocyte enhancer factor 2 alpha	V$MEF2_01,V$MEF2_02, V$MEF2_03,V$MEF2_04, V$MEF2_Q6_01, V$MMEF2_Q6,V$HMEF2_Q6, V$MEF2_01,V$AMEF2_Q6, V$RSRFC4_01,V$RSRFC4_Q
*Mef2b**	myocyte enhancer factor 2 beta	V$MEF2_Q6_01
*Mef2c**	myocyte enhancer factor 2 gamma	V$MEF2_Q6_01
*gata4**	GATA-box binding factor 4	V$GATA_Q6, V$GATA4_Q3
*srf**	serum responsive factor;	VSRF_01,VSRF_C, VSRF_Q5_02,VSRF_Q6, VSRF_Q4, VSRF_Q5_01
*nkx2.5**	cardiac-specific homeobox protein	V$NKX25_01,V$NKX25_02, V$NKX25_Q5

Table 2.5: Table of nuclear receptors and additional TFs of special interest and their associated TRANSFAC matrices used in the search for nuclear receptor binding sites. TF assigned with a "*" are no nuclear receptors.

Investigation of the Distributions of Certain Genomic Features

The following pages present the methods used to investigate the distributions of several genomic features (e.g. CpG islands, repeats) in the positive/negative dataset as well as their distribution over individual sequences.

2.15 Median, Mean and Density Computation

Certain genomic features as CpG island or specific TFBS are known to influence expression of specific genes and chromatin environment. Therefore analyses examining the **distributions of these features** in the positive and negative datasets were implemented. The annotation of the genomic regions with these features was computed using the script ***FeatureExtractor.pl***[14] (see section 2.5 for further details of the feature ex-

[14]See **Appendix B** for a description of scripts

traction process). The analysis was performed using the scripts ***FeatureStatistics.pl*** **and** ***FeatureStatistics.R***[14].

The following list contains all analysed features:

- CpG island
- CpG region
- SP1 bindings site (TRANSFAC)
- GC Box
- CTCF binding site
- TATA Box (TRANSFAC)
- All repeat classes as annotated by UCSC genome browser

For each of these features the following properties were analysed for every sequence:

1. Absolute number of occurrences
2. The amount of nucleotides that is covered (in percentage of the total region)
3. Absolute number of occurrences in the regions *Left*[15], *Gene 1*, *Intergenic*, *Gene 2*, and *Right*[15]
4. The amount of nucleotides that is covered in the regions *Left*[15], *Gene 1*, *Intergenic*, *Gene 2*, and *Right*[15](in percentage of the individual region)

After extracting these properties for each individual sequence, the **mean**, the **median**, and the **density** (only points 2 and 4) for each feature is computed for the positive and negative dataset. Additionally a *Wilcoxon rank sum test* is performed to analyse the significance of differences in the feature distributions between the two datasets. The nonparametric *Wilcoxon test* was used instead of the *t test* as the distributions of the extracted features does not need to be Gaussian.

2.16 Representation of Feature Distributions Over Genomic Regions

The analysis performed on distributions (see section 2.15) describes the presence of a specific feature in the region of interest but does not account for its position in that particular sequence. To detect a regional enrichment or depletion of a feature over the positive/negative dataset a package called **FeaturePlotter** was implemented.

This tools comprises the following two submodules:

1. **Adjustment/Mapping of Regions of Different Length**
   A maijor problem in investigating similarities in the distributions of certain features

[15] *Left* and *Right* are used for better understandability; the region itself has no orientation

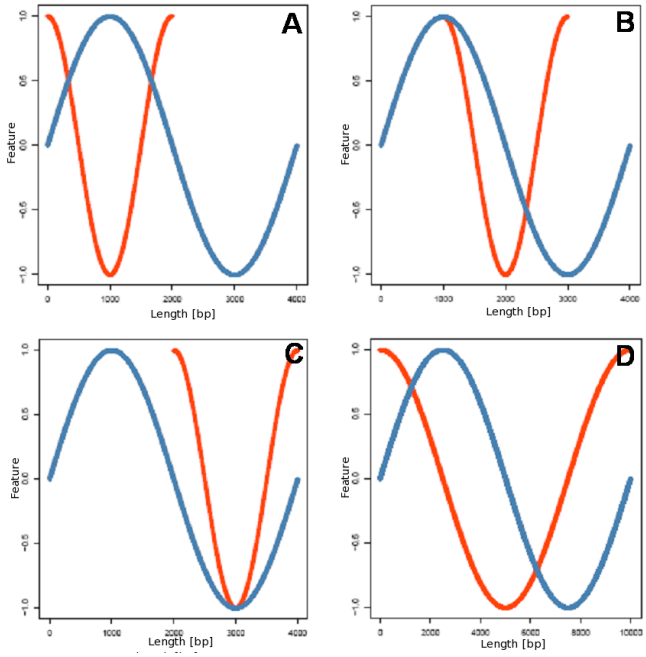


Figure 2.6: Illustration of adjustment and mapping of two arbitrary feature distributions of different length. **A:** Adjustment to the **left B:** Adjustment to the **middle C:** Adjustment to the **right D:** Mapping.

over multiple genomic regions is the different length of these regions. For example, the intergenic distance ranges from **-184bp** to **116,174bp** in the *2K-2K* positive dataset. Therefore, functions that adjust or map feature distributions over different sized regions are contained in the **FeaturePlotter** package:

- **Adjusting** several feature distributions to a (longer) region of specified length will insert zeros at positions that are not covered by the (shorter) region.
- In contrast the **mapping** process will not insert zeros but enlarge or reduce the length of the mapped region while maintaining its feature distribution.

The different possibilities of adjusting/mapping two or more feature distributions of different length is illustrated in Figure 2.6. After all regions and their appropriate feature distributions are adjusted/mapped to a given length, the mean value of every position in that region is computed. This mean reflects the probability of finding the investigated feature in a specific proportion of all sequences contained in the dataset.

2. **Plot Feature Distributions over Several Region**
   After adjusting/mapping all selected regions to a given length and computing the overall mean for each position, the sequence of this mean is used as input for the plotting script. The number of regions and the number of (mean) feature

distributions is not fixed to keep the **FeaturePlotter** flexible for multiple settings. The resulting plot consist of one line per passed feature that represents its distribution over all regions, together with a summary line which contains all feature distributions in one graph. For each feature distribution, values for line style, width and color can be specified to adapt the output for personal needs, as well as y- and x-axis definitions. Furthermore, a graphical illustration of the represented regions is drawn below each feature plot to enable the reader to interpret the feature distribution in its context. The illustration consists of boxes and lines that are defined and coloured by user-defined parameters. An example for a FeaturePlotter plot is given in Figure 2.7.

The **FeaturePlotter** was used to represent the distribution of all the extracted features after preprocessing the feature annotations with the script ***DistributionExtractor.pl***[16]. Every individual region (*Gene 1*, *Intergenic*, etc.) from a single pair of the dataset was therefore mapped to a specified size. The obtained region lengths are shown in Table 2.6.

	Left	*Gene 1*	*Intergenic*	*Gene 2*	*Right*
2K-2K	2,000	10,000	10,000	10,000	2,000
2K-next	10,000	10,000	10,000	10,000	10,000
H2K-2K	2,000	10,000	10,000	10,000	2,000
H2K-next	10,000	10,000	10,000	10,000	10,000

Table 2.6: Used mapping lengths of each region used in the **FeaturePlotter** plots according to the selected dataset.

[16]See **Appendix B for a description of scripts**

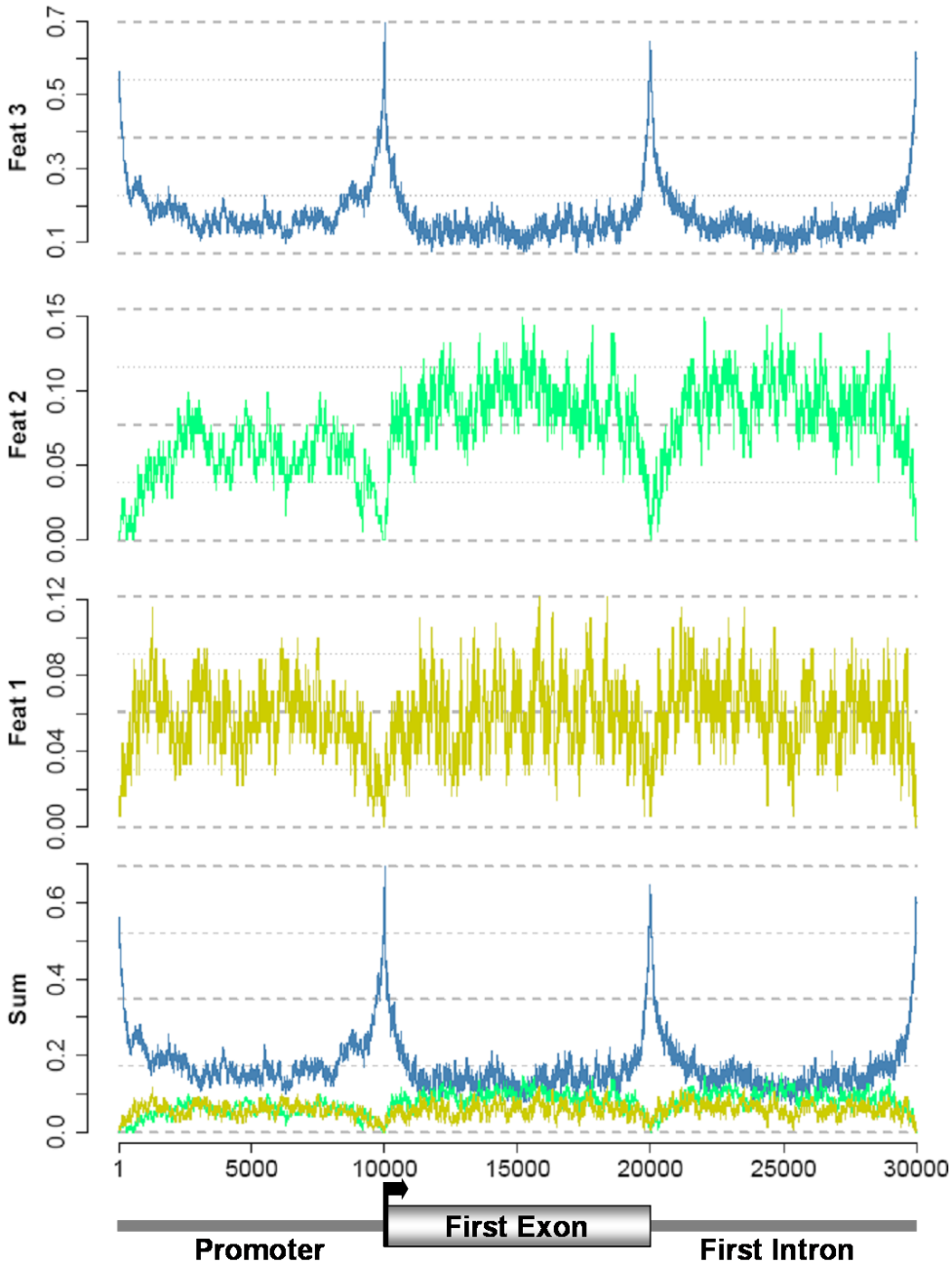


Figure 2.7: Exemplary illustration of a **FeaturePlotter** plot with 3 regions and 3 features.

Chapter 3

Results & Discussion

The following pages contain the results from all analysis performed in this master thesis. The first and second part depict the results of the Motif search and the complementary TFBS search process together with a discussion of the results. Furthermore, the distribution of genomic features is depicted together with an analysis of selected co-occurences.

A collection of all Figures included in this chapter together with some additional Figures that further illustrates the obtained results can be found in **Appendix C**.

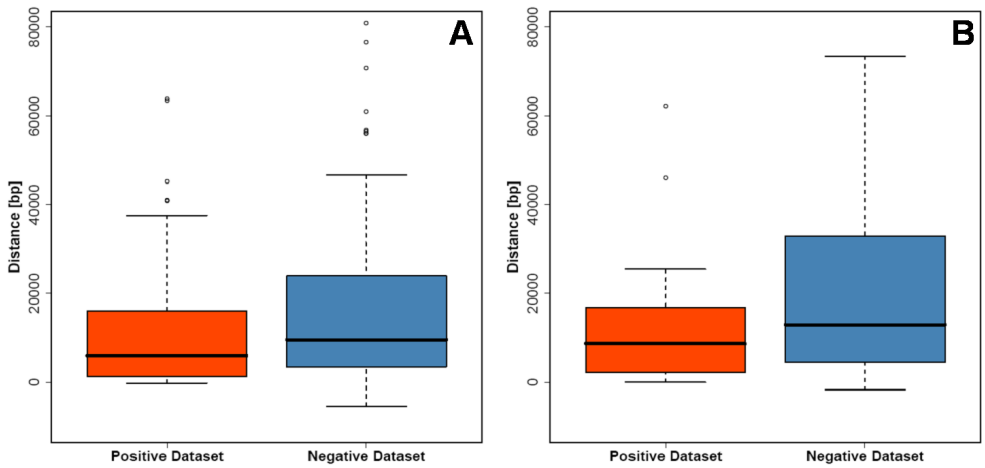


Figure 3.1: Boxplots of the intergenic distances for pairs of the positive and negative datasets. **A:** Mouse Datasets **B:** Human Datasets

3.1 Sequence Datasets

The aim of our investigations was to examine regulatory elements that could lead to the observed coordinated expression of adjacent genes. The basis of the analyses performed are **two mouse (mm8) sequence datasets**, *2K-2K* and *2K-next*. Both are derived from two groups of gene pairs that share a high amount of correlated expression (**highly co-expressed gene pairs - HCPs - postive dataset**) and an uncorrelated expression patterns (**uncorrelated gene pairs - UCP - negative dataset**), respectively. To assure stable expression properties for each individual gene pair, only those pairs have been extracted that belong to the appropriate co-expression group according to both **Fantom3 mouse** and **GNF Symatlas human** dataset. Furthermore, these pairs have been revised for intermediate Ensembl transcripts and a unique Fantom/Ensembl transcript assignment. Finally, **only gene pairs that are at least 2kb distant from their sourrounding genes** have been included into the dataset. In case of *2K-2K* **2kb on either side** of the pair was included whereas in case of *2K-next* the **entire sequence up to the neighbouring left/right genes** was added. The tables 3.1 and 3.2 show some statictisc of the two datasets defined.

Moreover, the **two homologous human (hg18) sequence datasets** *H2K-2K* and *H2K-next* have been defined on the two datasets to verify the results. The appropriate dataset statistics are presented in table 3.3 and 3.4.

For a full description of the co-expression groups and the dataset definition process refer to **sections 2.1 to 2.3**. **Appendix A** contains individual definition of all pairs included in the datasets.

Dataset: **2K-2K**					
Organism: **Mouse**					
Positive:	**51**	Negative:	**134**		
Total number of bp in the positive dataset:					3,644,565
Total number of bp in the negative dataset:					8,287,500
Distribution of sequence lengths:					
		Min	Avg		Max
Positive dataset:		11,435	71,462		294,357
Negative dataset:		8,605	61,847		262,771
Distribution of GC content:					
		Min	Avg		Max
Positive dataset:		38.78%	44.13%		55.03%
Negative dataset:		37.90%	45.34%		55.84%
Specific nucleotide content:					
		A	C	G	T
Positive dataset:		27.86%	22.07%	22.06%	28.01%
Negative dataset:		27.40%	22.71%	22.64%	27.25%

Table 3.1: Statistics for the *2K-2K* dataset.

Dataset: **2K-next**					
Organism: **Mouse**					
Positive:	**51**	Negative:	**130**		
Total number of bp in the positive dataset:					6,882,780
Total number of bp in the negative dataset:					20,331,575
Distribution of sequence lengths:					
		Min	Avg		Max
Positive dataset:		16,657	134,956		404,800
Negative dataset:		11,534	156,396		1,278,380
Distribution of GC content:					
		Min	Avg		Max
Positive dataset:		39.77%	42.85%		55.62%
Negative dataset:		33.99%	42.70%		52.72%
Specific nucleotide content:					
		A	C	G	T
Positive Dataset:		28.00%	21.91%	21.98%	28.11%
Negative Dataset:		28.10%	21.91%	21.90%	28.09%

Table 3.2: Statistics for the *2K-next* dataset.

Dataset: **H2K-2K**					
Organism: **Human**					
Positive:	**35**	Negative:	**96**		
Total number of bp in the positive dataset:					3,058,003
Total number of bp in the negative dataset:					6,965,054
Distribution of sequence lengths:					
		Min	Avg		Max
Positive dataset:		11,593	87,371		295,498
Negative dataset:		12,678	72,552		249,948
Distribution of GC content:					
		Min	Avg		Max
Positive dataset:		38.42%	43.59%		63.45%
Negative dataset:		35.09%	44.19%		60.73%
Specific nucleotide content:					
		A	C	G	T
Positive dataset:		27.56%	21.67%	21.92%	28.85%
Negative dataset:		27.81%	22.07%	22.11%	28.01%

Table 3.3: Statistics for the *H2K-next* dataset.

Dataset: **H2K-next**					
Organism: **Human**					
Positive:	**35**	Negative:	**93**		
Total number of bp in the positive dataset:					5,123,961
Total number of bp in the negative dataset:					18,771,424
Distribution of sequence lengths:					
		Min	Avg		Max
Positive dataset:		24,230	146,399		433,351
Negative dataset:		19,785	201,843		1,714,699
Distribution of GC content:					
		Min	Avg		Max
Positive dataset:		38.55%	43.24%		63.14%
Negative dataset:		35.39%	42.14%		59.43%
Specific nucleotide content:					
		A	C	G	T
Positive dataset:		27.97%	21.58%	21.66%	28.79%
Negative dataset:		28.72%	21.03%	21.10%	29.15%

Table 3.4: Statistics for the *H2K-next* dataset.

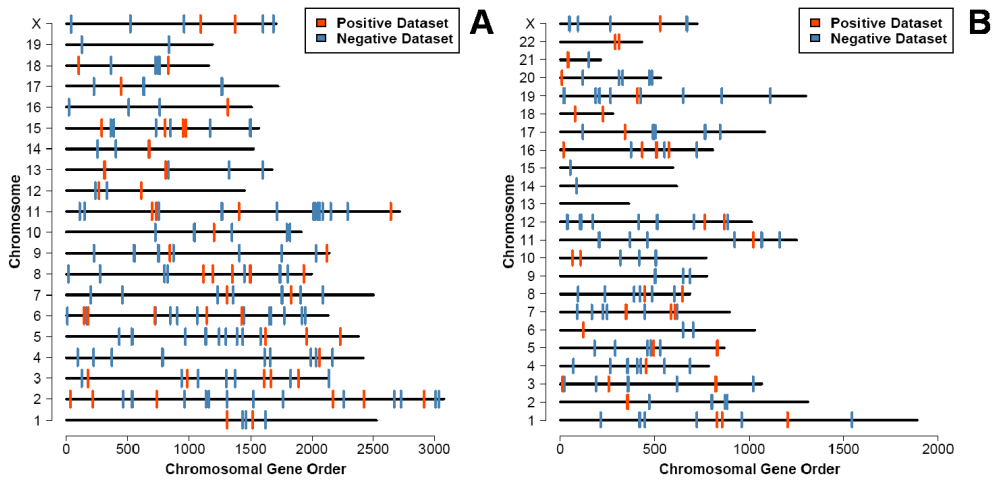


Figure 3.2: Chromosomal position (relative to total number of genes on each chromosome) of gene pairs belonging to the positive and negative dataset. **A:** Mouse Datasets **B:** Human Datasets

In our previous analysis [7] we showed that **HCPs are limited in size**. Investigating the intergenic distance between the pairs contained in the positive and negative datasets reveals the proposed difference in length, as genes in the mouse positive dataset have an intergenic distance median of **5,924bp** (human: **8,727bp**) compared to **9,593bp** (human: **12,896bp**) in the negative dataset ($p_{mouse} = 0.08901$ and $p_{human} = 0.03819$ in *Wilcoxon rank sum test*). At least for the mouse data, these median values are even smaller than suggested from our previous analysis. However, this is true for both the positive and negative dataset. Furthermore, the distance between the means of the two defined group is smaller. This indicates that there may also be a certain mechanism that effects clustering of (strongly) uncorrelated gene pairs. The distribution of intergenic lengths is shown in Figure 3.1.

Mouse datasets

Dataset	**Convergent**	**Divergent**	**Unidirectional**
Positive	33.33%	13.72%	52.95%
Negative	22.39%	17.91%	59.70%

Human datasets

Dataset	**Convergent**	**Divergent**	**Unidirectional**
Positive	34.29%	11.43%	54.28%
Negative	20.83%	17.71%	61.46%

Table 3.5: Percentage of gene pairs that have a certain genomic orientation for the human and mouse dataset.

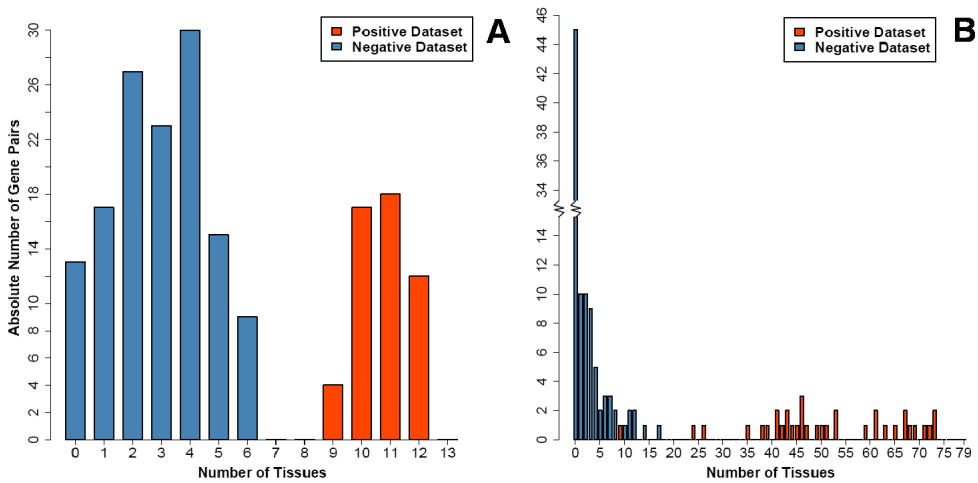


Figure 3.3: Amount of pairexpression of pairs belonging to the positive and negative dataset. **A:** Mouse Datasets (13 tissues) **B:** Human Datasets (79 tissues)

The gene pairs tend to be **randomly distributed over the chromosomes** as indicated in Figure 3.2. This finding agrees with our pervious results. Furthermore, the gene pairs in the positive and negative dataset have a **similar distribution** in respect to the three possible **genomic orientations** (see table 3.5).

We previously could show that HCPs are not solely the result of a clustering of housekeeping pairs [7]. Accordingly the pairs in the defined positive datasets **does not solely comprises of housekeeping pairs** (see Figure 3.3).

3.2 Bindings Site Analysis

3.2.1 Motif Search

To find motifs overrepresented in the positive mouse datasets (compared to the negative dataset) several motif finding alggorithms were used. Motif finding programs search for overrepresented short subsequences that could **represent TFBS**. In our case, we included the **regulatory potential** information (as indicated by a *regulatory potential score* > 0) into the searching process to specifically reduce the initial sequences to conserved and potentially regulatory subsequences. This methods is also called **phylogenetic footprinting**. Furthermore, other elements like specific repeats classes can be masked from the sequences prior to the motif finding process. For all analysis transcribed regions were excluded.

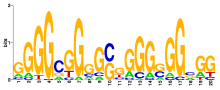
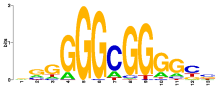
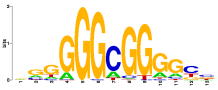
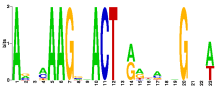
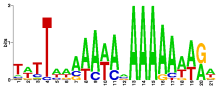
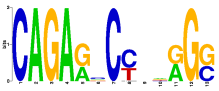
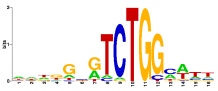
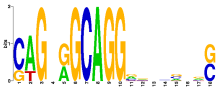
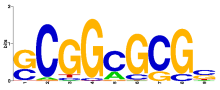
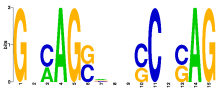
Motif	+	−	+/− Ratio	Group Specificity Score	Motif Sequence Logo	Matching TFBS (≥0.6)	Best TFBS	Best TFBS Sequence Logo
1	37	38	2.558308	5.529381e-08		9	Sp1 (V$SP1_Q6) *0.723513*	
2	32	38	2.212590	2.025750e-05		24	Sp1 (V$SP1_Q6) *0.979574*	
3	12	6	5.254902	3.074113e-04		0	/	/
4	25	30	2.189542	4.965302e-04		14	CIZ (V$CIZ_01) *0.703133*	
5	20	20	2.627451	5.279426e-04		8	AP-4 (V$AP4_Q6) *0.734285*	
6	22	26	2.223228	1.233805e-03		2	Hand1:E47 (V$HAND1E47_01 *0.663604*	
7	26	35	1.951821	1.376354e-03		21	KAISO (V$KAISO_01) *0.762017*	
8	9	4	5.911765	1.435477e-03		7	E2F (V$E2F_Q2) *0.838049*	
9	19	22	2.269162	2.785642e-03		0	/	/
10	16	17	2.472895	3.892884e-03		13	Ik-3 (V$IK3_01) *0.77656*	

Table 3.6: **The best-ranking motifs resulting from the motif search process using the 2K-2K dataset.**
The + and - columns contain the absolute number of occurences in the appropriate dataset. *+/- Ratio* and the *Group Specificity Score* are described in section 2.10. The total number of matching TRANSFAC matrices that match the motif with a correlation coefficient of at least 0.6 is given in column *Matching TFBS*. The best-matching TRANSFAC matrix and the appropriate correlation coefficient is presented in column *Best TFBS*. The WebLogo tool (described in section 2.12) was used to draw the sequence logo of the found motif and the best-matching TRANSFAC TFBS.
The following additional attributes were used in the search: **Tree cluster distance:** *0.6*, **MAST Sequence pvalue:** *0.05*, **MAST Evalue:** *10*

Performing the search on the two mouse datasets using regulatory potential information and masking all repeat classes other than *simple repeat* and *low complexity* results in the motif ranking presented in Table 3.6 and 3.7. The ranking was based on *Group Specificity Score* (GSS) which is decribed in section 2.10.3 and only the best 10 motifs are shown.

Significant overrepresentation of a GC-rich motif indicating SP1 binding

In both datasets the motif search detects a significantly overrepresented GC-rich motif (compare Table 3.6 motif 1 and Table 3.7 motif 1), which is present in almost all sequences of the positive dataset (**37** and **41** hits in a total of **51** sequences, respectively) but underrepresented in the negative dataset (**38** and **49** hits in a total of **134** and **130** sequences, respectively). This distribution leads to a more than **doubled frequency** (as indicated by the +/- ratio of >2) and a **very low GSS**.

Motif 2 of Table 3.6 seems to be a shorter but very similar version of the first detected motif. Both are highly similar to the TF binding matrix of SP1.

SP1 (identified in the early 1980s) was the first transcription factor shown to bind to GC Boxes (GGGGCGGGG) and GT/CACC boxes (GGTGTGGGG) via its three Cys_2His_2 zinc-finger motifs [104]. These GC/GT boxes are commonly found in **CpG-rich methylation-free islands** [105]. It is a member of a large family of Sp1-like/KLF (Krüppel-like factor) genes that can either activate or repress their target genes. SP1 was shown to control the expression of **housekeeping genes** as well as **tissue-specific** and **viral genes**. Sp1 binding has been reported in **promoters**, **enhancers** and **locus control regions**. The family member EKLF for example has a functional target site located in the main regulatory element of the β-globin locus. Wheter SP1 activates or represses it's target genes is suggested to be controled by interacting corepressors and coactivators [104]. One example is the **CREB-binding protein (CBP) homolog p300** and the CBP/p300-associated factor (PCAF) that were shown to have **acetyltransferase (HAT) activity** [106].

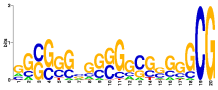
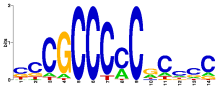
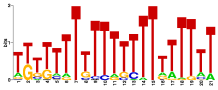
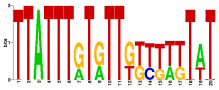
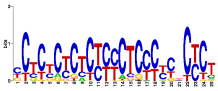
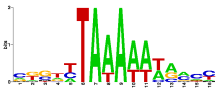
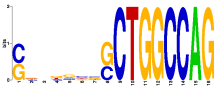
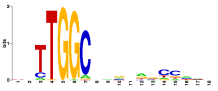
Motif	+	−	+/− Ratio	Group Specificity Score	Motif Sequence Logo	Matching TFBS (≥0.6)	Best TFBS	Best TFBS Sequence Logo
1	41	49	2.132853	1.503406e-07		1	KROX (V$KROX_Q6) *0.714317*	
2	15	12	3.186275	1.086099e-03		1	TFII-I (V$TFIII_Q6) *0.676122*	
3	19	19	2.549020	1.094421e-03		3	AREB6 (V$AREB6_03) *0.629228*	
4	45	85	1.349481	1.233841e-03		10	FOXP1 (V$FOXP1_01) *0.686917*	
5	51	112	1.160714	1.805383e-03		0	/	/
6	21	25	2.141176	2.589291e-03		0	/	/
7	14	12	2.973856	2.592593e-03		0	/	/
8	18	20	2.294118	3.683190e-03		2	Lyf-1 (V$LYF1_01) *0.704622*	
9	25	35	1.820728	4.252573e-03		21	MEF-2 (V$MMEF2_Q6) *0.81539*	
10	21	28	1.911765	7.267082e-03		2	NF-1 (V$NF1_Q6) *0.63069*	

Table 3.7: **The 10 best-ranking motifs resulting from the motif search process using the 2K-next dataset.**
The following additional attributes were used in the search: **Tree cluster distance:** *0.6*, **MAST sequence pvalue:** *0.05*, **MAST Evalue:** *10*

The first motif in Table 3.7 is the binding site of **KROX**. The binding sites of SP1 and KROX as annotated by TRANSFAC are **highly similar** and share a correlation coefficient of **0.9261**. Because of their very similar binding site, an interaction of this factors has been suggested [107]. We therefore propose a **SP1 dependent mechanism** for coexpression of genes.

Other found motifs

The motif search process reveals additional possible binding sites that could contribute to the observed level of co-expression of genomic neighbours.
A large fraction of the found TFBS belong to TFs that have been associated to chromatin remoddeling. These TFs are **AP-4**, **KAISO**, **E2F**, **MEF-2**, and the transcription factor **Ikaros 3 (Ik-3)/Lyf-1** which belongs to the 10 top-ranking TFBS in both datasets. Ikaros can participate in **chromatin remoddeling** by interaction with HATs and nuclear receptor [108]. Similarly, the other mentioned TFs have been described to contribute in HAT and HDAC recruitment [109],[110],[111],[112].
Furthermore, the TF **NF-1** is a **nuclear receptor**. These proteins bind histones and activate the remodelling machinery [21].
For the remaining TFs **CIZ**, **AREB6**, **FOXP1**, **Hand1:E47** and **TFII-I** interactions with HAT/HDAC has not been described in the literature.
Finally, 5 motifs are found that can not be associated to any known vertebrate TFBS.

All in all, the other found TFBS scored worse compared to the SP1 binding sites mentioned above. However, all used motif finding tools are developed for the search in relatively short (promoter) sequences. Because the exact position of our proposed regulatory elements was unknown, we needed to **include a large amount of sequence**. This might have increased the noise level and therefore led to missed motifs. To overcome the noise problem, we included the conservational RP information. However, as the **position of regulatory elements might have shifted** during evolution, the use of conservational information could again have led to a loss of binding sites in the masked sequences.

Investigation on the influence of masking conditions in the motif search

The above analysis was repeated with **modified masking conditions** (e.g. the exclusion of regulatory potential information or modified repeat masking conditions). Additional masking of the *low complexity* and *simple repeats* leads to a diminished occupancy of SP1 binding sites in the sequences of both datasets. Nevertheless, the results were **found to be robust in regard to the masking parameters**.

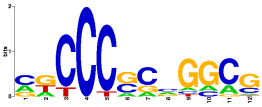
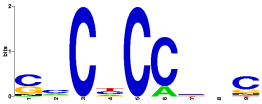
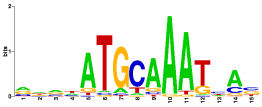
	TFBS	+	−	+/− Ratio	Group Specificity Score	TFBS Sequence Logo
1	c-Ets-1(p54) (V$CETS1P54_01)	12	2	15.764706	3.883440e-06	
2	GCbox (V$GC_01)	34	55	1.624242	1.509024e-03	
3	Nrf-1 (V$NRF1_Q6)	9	6	3.941176	6.165381e-03	
4	AP-2 (V$AP2_Q6)	6	3	5.254902	1.425194e-02	
5	HIF-1 (V$HIF1_Q5)	4	0	/	2.095589e-02	
6	WT1 (V$WT1_Q6)	10	11	2.388592	3.084261e-02	
7	SMAD-4 (V$SMAD4_Q6)	8	9	2.335512	5.883073e-02	
8	Cdc5 (V$CDC5_01)	3	1	7.882353	6.436312e-02	
9	HEB (V$HEB_Q6)	5	4	3.284314	6.679776e-02	
10	Oct-1 (V$OCT1_Q6)	5	4	3.284314	6.679776e-02	

Table 3.8: **The 10 best-ranking TFBS resulting from the TFBS search process using the 2K-2K dataset.**
The + and - columns contain the absolute number of occurences in the appropriate dataset. *+/- Ratio* and the *Group Specificity Score* are described in section 2.10. The WebLogo tool (described in section 2.12) was used to draw the sequence logo of the found TFBS.
The following additional attributes were used in the search: **Tree cluster distance:** *1*, **MAST Sequence pvalue:** *0.05*, **MAST Evalue:** *10*

3.2.2 Vertebrate Matrices Matching

In contrast to searching for overrepresented subsequences and comparing these to known vertebrate TFBS we also **directly matched all known vertebrate TFBS** extracted from TRANSFAC to our two mouse datasets. Again, several search conditions were sampled including the use of regulatory potential information and the masking of repetetive elements.

Search for overrepresented vertebrate TFBS

Using regulatory potential information and masking all repeat classes other than *simple repeat* and *low complexity* in the search the rankings shown in table 3.8 and 3.9 were obtained. Again, the ranking was based on GSS and only the best 10 motifs are shown. All in all the search for known vertebrate matrices resulted in a much lower number of hits and higher GSS.

Compared to the motif finding results, the **SP1 binding site** is the only TFBS that is present in both rankings. In this case it is represented by the GC Box. All other TFBS had **low frequencies** in the positive dataset. In no case a high-ranking TFBS was found in more than a fifth of all sequences included in the positive sequence datasets (with the exception of c-Ets-1 which is only slightly more frequent). In summary, the TFBS search revealed no overrepresented known vertebrate TFBS other than SP1.

Investigation on the influence of masking conditions in the vertebrate TFBS search

As in the previous analysis we checked for the influence of masking conditions. In contrast to the motif finding process, the **masking conditions show a stronger influence on the results**. According to the used pre-processing of the sequences of the positive and negative datasets the TFBS score differently, leading to shifts in the ranking. Still, all best-ranking TFBS do not cover the positive dataset to a greater extent than shown before. This might also be the reason for the observed shifting effect as minor changes in the number of matching in sequences of the positive dataset influence the (allready very similar) GSS strongly.

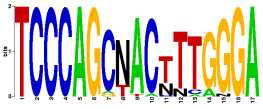
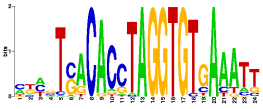
	TFBS	+	−	+/− Ratio	Group Specificity Score	TFBS Sequence Logo
1	Nrf-1 (V$NRF1_Q6)	9	2	11.470588	2.168464e-04	
2	GCbox (V$GC_01)	33	48	1.752451	6.435278e-04	
3	Stra13 (V$STRA13_01)	6	2	7.647059	6.825753e-03	
4	HIC1 (V$HIC1_02)	6	2	7.647059	6.825753e-03	
5	C/EBPdelta (V$CEBPDELTA_Q6)	5	0	/	7.200213e-03	
6	LUN-1 (V$LUN1_01)	9	7	3.277311	1.289559e-02	
7	c-Ets-1(p54) (V$CETS1P54_01)	5	2	6.372549	1.973023e-02	
8	OCT-x (V$OCT_C)	9	8	2.867647	2.130718e-02	
9	NF-kappaB(p65) (V$NFKAPPAB65_01)	7	6	2.973856	3.946231e-02	
10	Brachyury (V$BRACH_01)	4	2	5.098039	5.385452e-02	

Table 3.9: **The 10 best-ranking TFBS resulting from the TFBS search process using the 2K-next dataset.**
The following additional attributes were used in the search: **Tree cluster distance:** *1*, **MAST sequence pvalue:** *0.05*, **MAST Evalue:** *10*

The binding site of SP1 stayed among the highest-scoring TFBS irrespective of the used masking conditions.

When masking all repeats but neglecting regulatory potential information, the TFBS Aire was significantly overrepresented. The **autoimmune regulator (Aire)** is a transcription factor that controls the self-reactivity of the T cell repertoire. The TFBS was present **13** times in the positive dataset compared to **2** times in the negative dataset, which resulted in a +/- ratio of >**17** and a GSS of $1.018x10^6$.
Aire was proposed to have **clustered target genes** [113] though frequently interspersed with genes that are independent of Aire regulation [114]. The presence of Aire binding sites in our group of highly co-expressed gene pairs substanciate the proposition of clustered target genes.

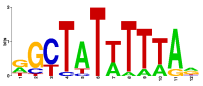
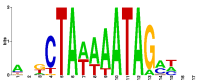
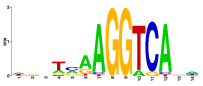
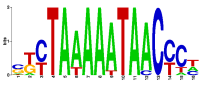
	NR TFBS	+	−	+/− Ratio	Group Specificity Score	NR TFBS Sequence Logo
1	NKX25 (V$NKX25_Q5)	7	9	2.043573	1.129541e-01	
2	MEF-2 (V$MEF2_Q6_01)	3	2	3.941176	1.294740e-01	
3	RSRFC4 (V$RSRFC4_Q2)	3	2	3.941176	1.294740e-01	
4	ERRalpha (V$ERR1_Q2)	4	4	2.627451	1.475419e-01	
5	MEF-2 (V$HMEF2_Q6)	2	1	5.254902	1.846189e-01	

Table 3.10: **The 5 best-ranking nuclear receptor TFBS resulting from the nuclear receptor TFBS search process using the 2K-2K dataset.**
The following additional attributes were used in the search: **Tree cluster distance:** *1*, **MAST sequence pvalue:** *0.05*, **MAST Evalue:** *10*

	NR TFBS	+	−	+/− Ratio	Group Specificity Score	NR TFBS Sequence Logo
1	RORalpha2 (V$RORA2_01)	5	6	2.124183	1.651820e-01	
2	MEF-2 (V$MEF2_Q6_01)	5	6	2.124183	1.651820e-01	
3	RSRFC4 (V$RSRFC4_Q2)	3	3	2.549020	2.197365e-01	
4	aMEF-2 (V$AMEF2_Q6)	4	5	2.039216	2.251930e-01	
5	SRF (V$SRF_Q5_02)	4	5	2.039216	2.251930e-01	

Table 3.11: **The 5 best-ranking nuclear receptor TFBS resulting from the nuclear receptor TFBS search process using the 2K-next dataset.** The following additional attributes were used in the search: **Tree cluster distance:** *1*, **MAST sequence pvalue:** *0.05*, **MAST Evalue:** *10*

Search for nuclear receptor binding sites

The analysis was repeated using only binding sites that belong to NRs and some additional TFBS of specific interest. Those were annotated to guide **chromatine remodelling** and could be associated to several resulting TFBS of the motif search process. The 5 best-ranking TFBS (according to GSS) are shown in Table 3.10 and 3.11. In accordance with the results for all known vertebrate TFBS the search **does not reveal a strong overrepresentation of a specific NR binding site** in the sequences of the positive datasets. These findings are somewhat contrary to the results of the motif finding process. A possible explantion of this discrepancy could be bad quality of the TRANSFAC matrices. Another possible reason could be the number of different possible binding sites of a single NR. The glucocorticoid receptor for example comprises 4 different binding sites in TRANSFAC. As our analysis only reveals the presence of single DNA binding sites and does not account for the total number of different binding sites of a specific factor present in the dataset, its real number of occurences might be underestimated.

Just as the motif finding process, the TFBS search is affected by the large sequences used and a potential influence of the used conservational information.

3.3 Feature Distributions

Based on the results of the prior analyses we investigated the distribution of certain genomic features over the sequences in the datasets. As SP1 binding sites are known to reside inside **CpG islands** this features was included into the analysis as well as the **SP1 binding site** itself (as represented by a TRANSFAC matric and the GC Box motif). Additionally, two further specific sequence motifs were included: the **TATA Box** and the binding site for the insulator **CTCF**. Furthermore the different classes of repeats as annotated by the *RepeatMasker* program by Arian Smith[1] were analysed. The repeat information was included, as a recent report suggests certain **repetitive elements** to include TFBS for several TF that control specific expression [61].
The search for these genomic features was performed using the two mouse datasets as well as the orthologous human datasets.

3.3.1 CpG Islands/Regions

CpG island are very GC-rich stretches of DNA that contain the CpG dinucleotide with higher frequency than suggested by whole-genome sequence analyses. In our analysis we used two definitions of a CpG island. The first was called **CpG island** while the other was called **CpG Region** (both are described in section 2.5.4), the latter having lower constraints.

Striking evidence for the presence of CpG islands in the positive dataset

Searching for CpG islands revealed a striking association of promoter-associated CpG islands to genes included in the positive datasets. **70%** of the sequences in the positive mouse dataset *2K-2K* and **86%** of the positive human dataset *H2K-2K* contained **at least 2 CpG islands**. On the contrary, only **22%** of the sequences in the negative mouse dataset *2K-2K* and **47%** of the negative human dataset *H2K-2K* contained 2 CpG islands. A *Wilcoxon rank sum test* assigned the differences in median with a significance level of $8.3x10^{-9}$ for the mouse dataset and $3.4x10^{-3}$ for the human dataset.
The distribution over the sums of all sequences in the positive and negative datasets in mouse and human are shown in Figure 3.4 and 3.5. In this plots the sequences have been sorted due to the orientation of their genes for better understandability. Looking in detail on the regional distribution of all CpG island discloses that almost all of them **reside at the transcriptional start site (TSS)**. As these plots show the average CpG island coverage for every position they reflect the enhanced number of TSS-associated CpG island of the positive datasets in contrast to the negative datasets.

[1] `http://www.repeatmasker.org`

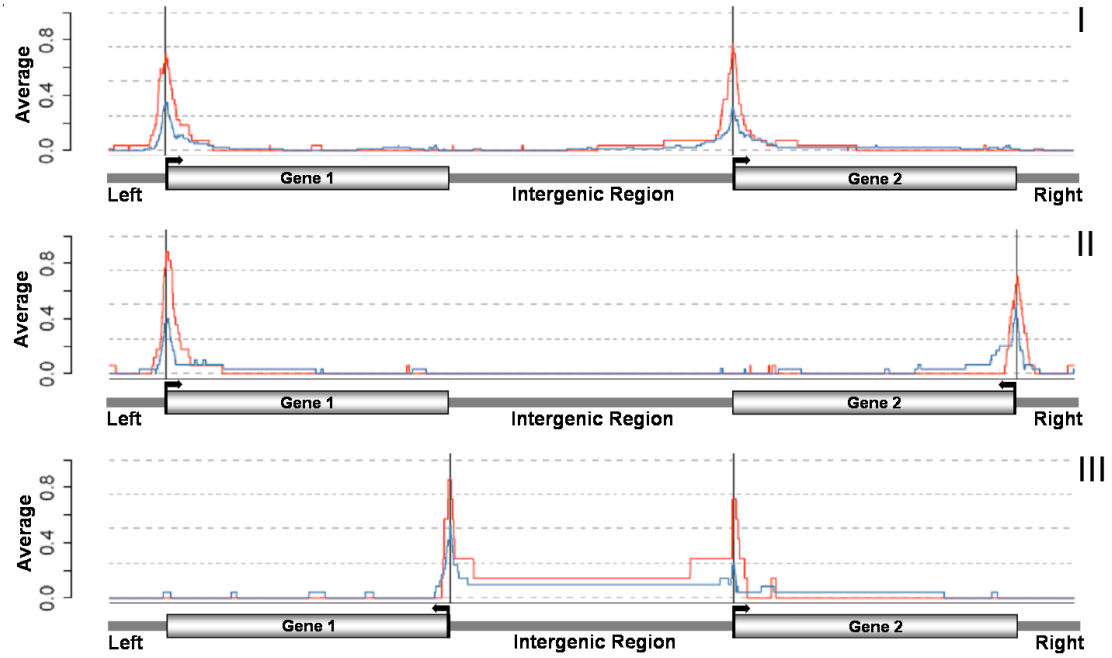


Figure 3.4: Average distribution of CpG island over the positive (red) and negative (blue) sequence in the mouse *2K-2K* datasets. **I:** Unidirectional Pairs **II:** Convergent Pairs **III:** Divergent Pairs

Impact of using the lower-constrained CpG region definition

Almost the same conclusions can be drawn when the **less constrained CpG regions** are taken into account. As the presence of a CpG region is **more probable** than a CpG island, these regions are found much more often. The mouse *2K-2K* dataset has a **CpG region mean** of 6.02 compared to 5.57 ($p = 0.32$ in *Wilcoxon rank sum test*). The human *H2K-2K* has a mean of 14.47 to 11.48 ($p = 0.023$ in *Wilcoxon rank sum test*). The loss of significance in the mouse dataset might be the result of one outlying sequence in the negative dataset (which comprises **40** CpG regions). Furthermore, CpG regions are not exclusively located at the TSS (data shown in **Appendix C**).

In respect to the results, the use of lowered constraints for the definition of CpG islands **increases the number of CpG islands and the noise level simultaneously**. Therefore it does not contribute to a better understanding of the differences in the used datasets.
CpG island analysis was not perfomed on the *2K-next* and *H2K-next* dataset, because these sequences might include CpG islands referring to the promoters of the neighbouring genes.

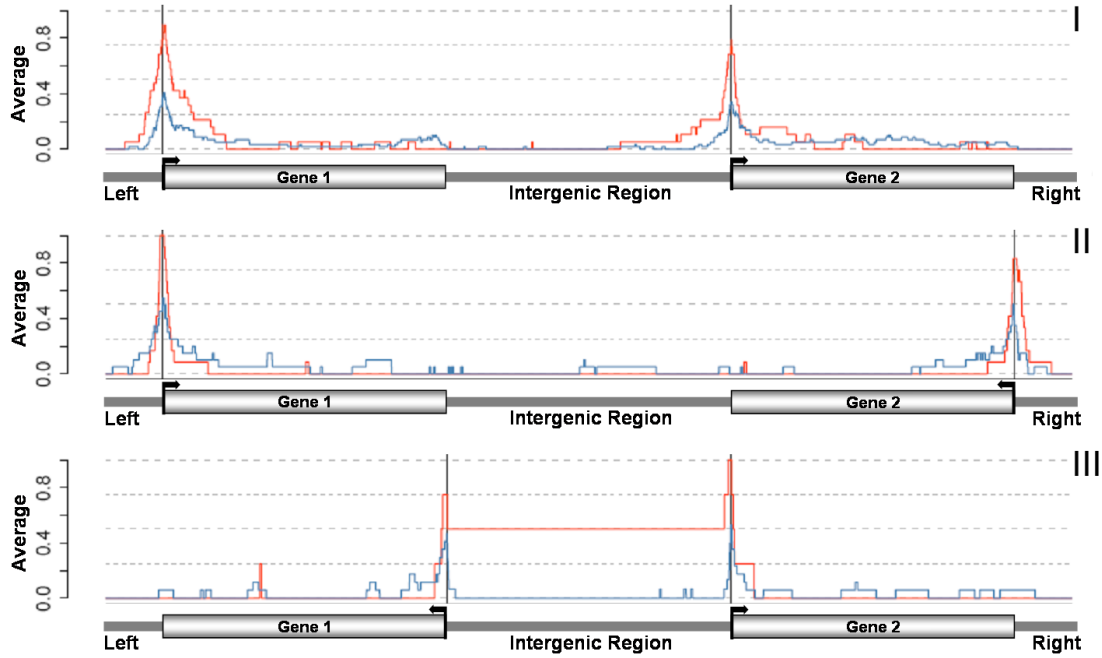


Figure 3.5: Average distribution of CpG island over the positive (red) and negative (blue) sequence in the human **H2K-2K** datasets. **I:** Unidirectional Pairs **II:** Convergent Pairs **III:** Divergent Pairs

3.3.2 Specific Transcription Factor Binding Sites

The TRANSFAC SP1 bindings matrix is again highly overrepresented

Searching for the SP1 binding site in the mouse *2K-2K* and *2K-next* dataset again revealed a higher number in the positive sequences (see Table 3.12 for mean and significance values). **66%** of the positive *2K-2K* dataset contains **at least one SP1 binding site** compared to **44.3%** in the negative dataset. Furthermore, the found binding sites do mainly **reside near the TSS** as indicated in Figure 3.6. In the positive mouse *2K-next* dataset we found **55%** of the sequences to contain at least one SP1 binding site compared to **35%** in the negative dataset. The loss of bindings sites in the longer *2K-next* dataset is very likely to be the influence of the enlarged sequences. Because MAST scores TFBF matches in respect to p- and Evalues, weak matches might drop out if a sequence becomes longer as the probability of random matches increases.

Analysing the human *H2K-2K* and *H2K-next* datasets ended up in less significant results, although the distribution is still different.

Using the **GC Box** "GGGCGGG" and its reverse complement the search resulted in higher number of sites per sequence. Mean values and significance level are shown in Table 3.13.

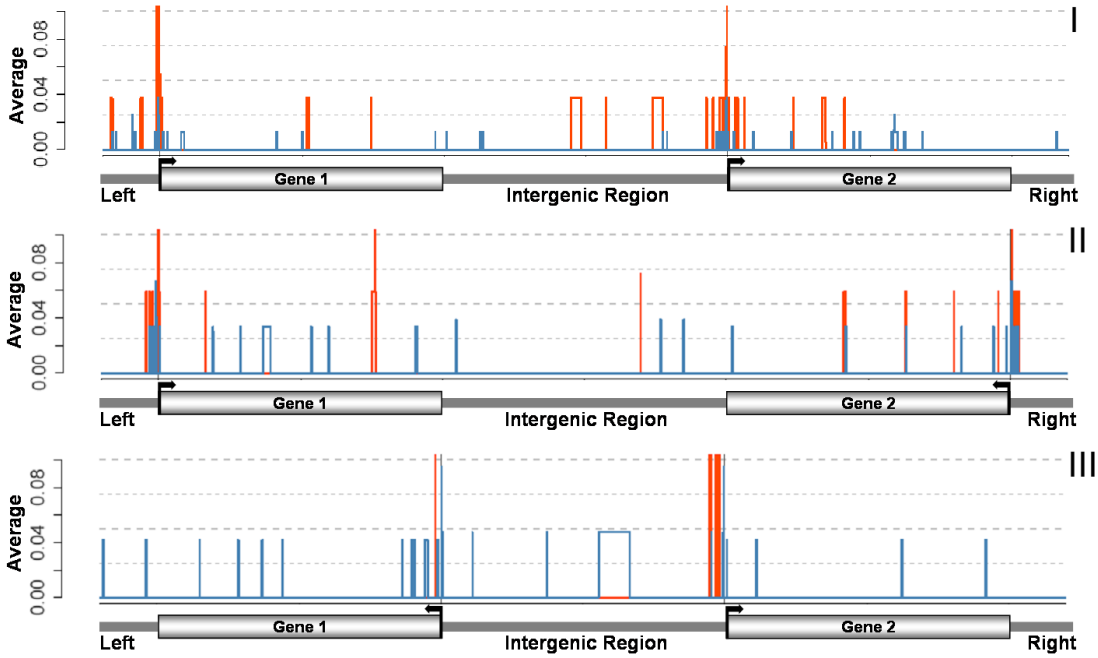


Figure 3.6: Distributions of predicted occurences of the SP1 binding site (represented by the TRANSFAC matrix *V$SP1_Q6_01*) in the positive (red) and negative (blue) sequence in the mouse **2K-2K** datasets. **I:** Unidirectional Pairs **II:** Convergent Pairs **III:** Divergent Pairs

The use of the GC Box hexanucleotide did lead to higher pvalues, especially in the case of the *next* datasets. This might reflect the fact that the GC Box pattern itself is very unspecific and occurs at several random sites throughout the genome. Because the *next* datasets are larger than the *2K* datasets, the GC Box motif occurs more often and in a more randomizes fashion, leading to the **loss of significance**.

	Mean +	*Mean -*	*pvalue*
2K-2K	1.29	0.76	$p = 8.1x10^{-3}$
2K-next	1.06	0.63	$p = 6.72x10^{-3}$
H2K-2K	1.39	0.82	$p = 0.15$
H2K-next	1.19	0.49	$p = 0.19$

Table 3.12: Mean and significance level (*Wilcoxon rank sum test*) for the number of predicted SP1 binding sites contained in the sequences of the datasets. "+" stands for the positive set of sequences (HCPs) and "-" for the negative set (UCPs).

	Mean +	*Mean -*	*pvalue*
2K-2K	7.94	6.34	$p = 0.0138$
2K-next	12.92	11.56	$p = 0.17$
H2K-2K	10.97	9.23	$p = 0.0344$
H2K-next	15.22	16.2	$p = 0.439$

Table 3.13: Mean and significance level (*Wilcoxon rank sum test*) for the number of GC Boxes contained in the sequences of the datasets.

The TATA box matrix as provided by TRANSFAC is not present in the datasets

We also performed a search for the TATA box represented by the TRANSFAC matrix *V$TATA_C*. However, we hardly find any occurences of this matrix in any sequence irrespective of the dataset. This could be a consequence of the used binding site matrix provided by TRANSFAC. Therefore an interpretation of the presence of TATA boxes in respect to the genes contained in our sequence datasets is not possible.

CTCF binding sites are enriched at the borders of HCPs

Searching for occurences of CTCF insulator binding sites in our sequences lead to different results in the mouse/human *2K-2K*/*H2K-2K* and *2K-next*/*H2K-next* datasets (see Table 3.14 for mean values and significances as determined by *Wilcoxon rank sum test*). While both the *2K-2K* and *H2K-2K* datasets do not show a significant overrepresentation of CTCF binding sites, it is **slightly overrepresented in the mouse *2K-next* dataset**. This would fit with the thesis that the CTCF insulatory protein resides at the **edges of euchromatic regions**. As the mouse *2K-next* and human *H2K-next* datasets includes more sequence around the gene pair, it would be more likely to contain this border and therefore to contain CTCF binding sites. Figure 3.7 illustrates the positions of the predicted CTCF sites and highlights a slight enrichment of binding sites in the left/right genomic regions. However, the orthologous human *H2K-next* dataset shows a counterdirected distribution of predicted binding sites (more bindings sites in the negative sequences). Therefore it is hard to interprete these results in a definite manner.

	Mean +	*Mean -*	*pvalue*
2K-2K	0.92	0.66	$p = 0.121$
2K-next	1.71	1.47	$p = 0.0841$
H2K-2K	1.44	1.44	$p = 0.401$
H2K-next	2.14	2.85	$p = 0.2261$

Table 3.14: Mean and significance level (*Wilcoxon rank sum test*) for the number of predicted CTCF binding sites contained in the sequences of the datasets.

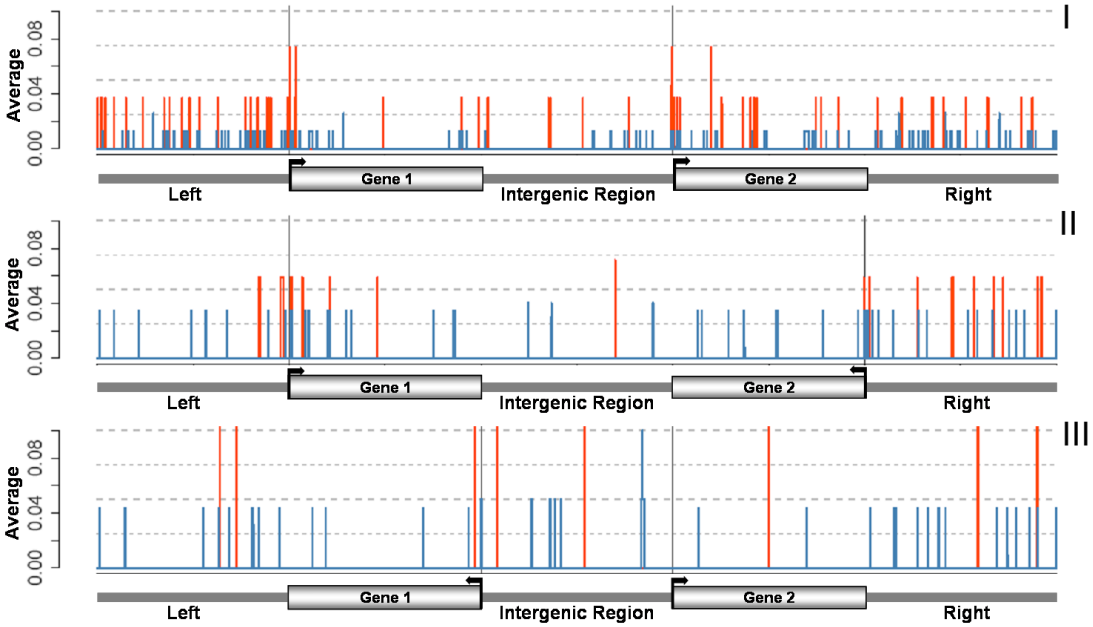


Figure 3.7: Distributions of predicted occurences of the CTCF binding site (represented by the consensus sequence "CTGCCNCCNNGCGG") in the positive (red) and negative (blue) sequence in the mouse **2K-next** datasets. **I:** Unidirectional Pairs **II:** Convergent Pairs **III:** Divergent Pairs

3.3.3 Repeats

Repeats cover almost 50% of the human genome (as stated by the RepeatMasker website). They consist of severel distinct classes[2]:

- Short interspersed nuclear elements (SINEs), which include ALUs
- Long interspersed nuclear elements (LINEs)
- Long terminal repeat elements (LTRs), which include retroposons
- DNA repeat elements (DNA)
- Simple repeats (micro-satellites)
- Low complexity repeats
- Satellite repeats
- RNA repeats
- Other repeats

[2]Description taken from the UCSC website

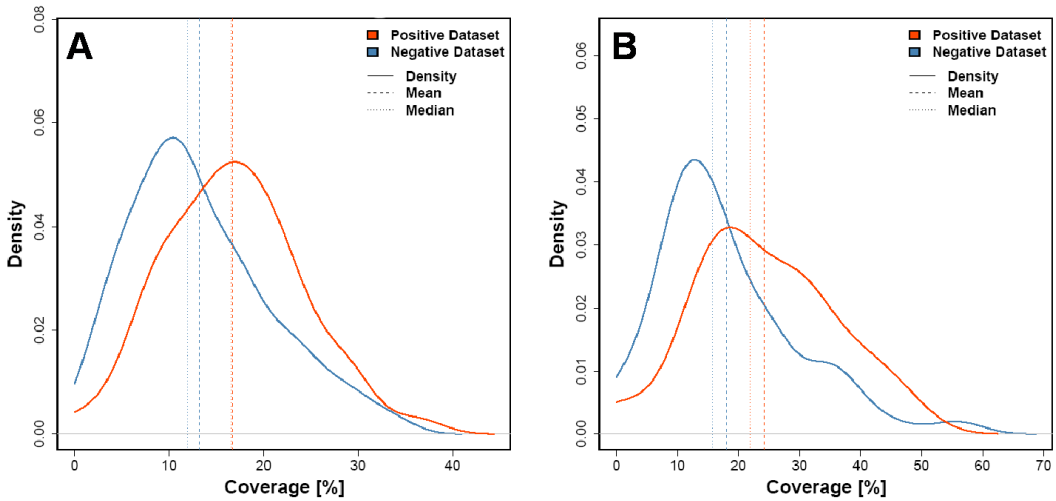


Figure 3.8: Mean, median and density for the procentage coverage of **SINE** repeats. **A:** Mouse *2K-2K* dataset $p = 2.3x10^{-3}$, **B:** Human *H2K-2K* dataset $p = 4.19x10^{-3}$

Enrichment of SINE repeats in positive sequences and stable LINE content

When investigating the distributions of these repetitive elements over our positive and negative sequences an astonishing differences in the amount of **sequence covered by SINE repeats** can be detected in all four datasets. The mean, median and density and significance (as indicated by *Wilcoxon rank sum test*) of this distribution are given in Figure 3.8. For the two datasets *2K-2K* and *H2K-2K* a **significant enrichment** of this repeat in sequences contained in the positive datasets is shown. The same is true for the other two datasets *2K-next* ($p = 3.57x10^{-3}$) and *H2K-next* ($p = 1.33x10^{-3}$) (data shown in **Appendix C**).

This enrichment with SINE repeats is a very interesting fact as those SINE repeats have been proposed to **contain a variaty of TFBS** [61]. The mean coverage of SINE repeats in the negative sequences is very close to the overall genomic SINE content in mouse (**8.22%**[115]) and human (**13.64%**[116]), respectively, while the mean coverage in the positive dataset exceeds this value by far (**17.35%** in the mouse *2K-2K* dataset and **23.93%** in the human *H2K-2K* dataset). This has been previously described for very gene dense regions in the human genome [9]. In contrast, the same analysis showed a simultaneously depletion of LINE repeats in the same regions. However, we can not find this depletion in our own positive datasets compared to the negative dataset (as shown by mean, median, density and significance level for *2K-2K* and *H2K-2K* in Figure 3.9). Nevertheless, we find a decreased level of **LINE repeats** in all positive/negative sequences compared to the average mouse (19.2%)[115] and human (20.99%)[116] LINE

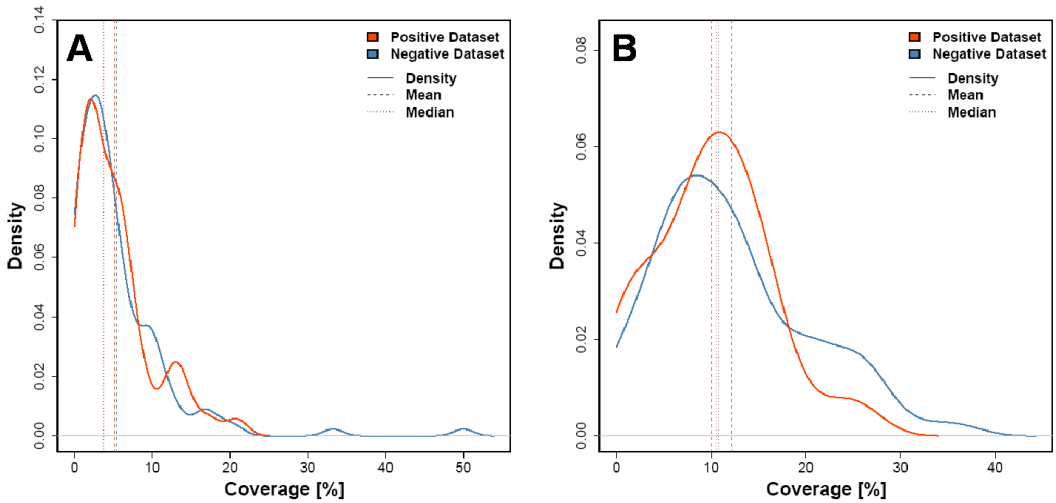


Figure 3.9: Mean, median and density for the procentage coverage of **LINE** repeats. **A:** Mouse *2K-2K* dataset $p = 0.915$, **B:** Human *H2K-2K* dataset $p = 0.165$

content, respectively. This findings suggests that sequences close to existing genes might be generally depleted of LINE repeats. Another possible explanation is the **size** of LINE repeats, which have an average length of **900bp**[3] compared to **100-400bp** for SINE repeats[3]. As we limit our sequences around the pairs these might contain less large repeats.

Depletion of simple repeats (micro-sattelites) in the positive datasets

In addition to the enrichment with SINE repeats, the analysis shows a depletion of **simple repeats** in the positive dataset. The mean, median, density and significance of their distribution is shown in Figure 3.10 for the mouse *2K-2K* dataset. Simple repeats (also called **micro-satellites**) consists of two, three or four nucleotides (di-, tri-, and tetranucleotide repeats respectively), and are repeated 10 to 100 times. Today, it is **controversial**, if these repeats have a biological meaning. It has been proposed that they are associated with regulation of gene activity as well as metabolic DNA processes (like replication and recombination) and chromatin organisation [117]. For our purpose the possible association of micro-satellites to chromatin structure are of main interest. Micro-satellites are thought to **induce DNA secondary structures** like loops and hairpins that may have an influence on gene expression [117]. Triplet repeats that are located in the UTRs or intron can induce **heterochromatin-mediated-like gene silencing** [118]. Furthermore, satellite repeats are associated to **heterochromatin that forms centromeric chromosome structures** [119].

All other repeats do not show a significant different distribution between the positive and negative datasets (for all plots see **Appendix C**).

[3]Average number for human genome taken from [116]

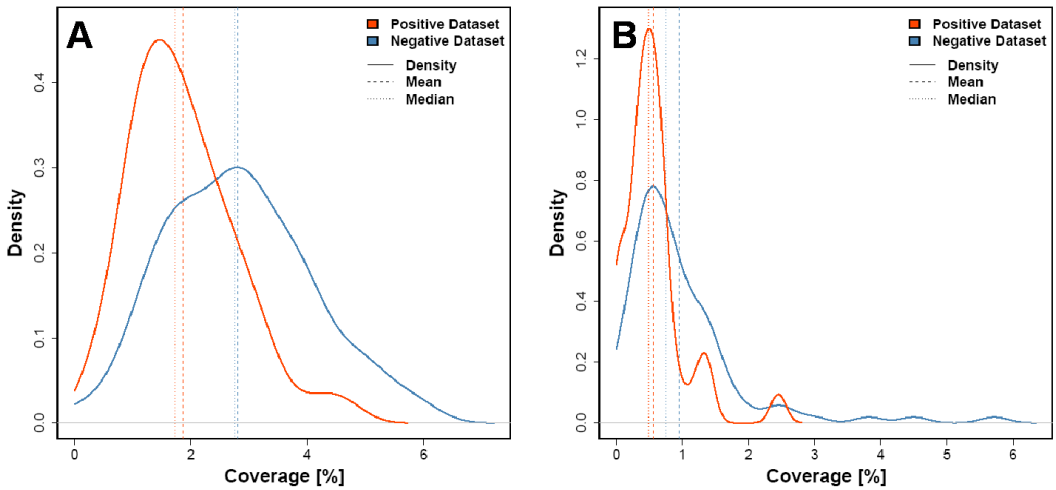


Figure 3.10: Mean, median and density for the procentage coverage of **simple** repeats (**micro-satellites**). **A:** Mouse *2K-2K* dataset $p = 9.77x10^{-7}$, **B:** Human *H2K-2K* dataset $p = 4.27x10^{-4}$

3.3.4 Co-appearance of Genomic Features

In addition to the distribution of single features we also investigated in the co-occurence of some of the features mentioned above.

Association of SP1 to CpG islands is stable in the positive and negative dataset
We examined if the association of SP1 to CpG islands is stronger in the positive datasets than in the negative. We therefore computed the number of SP1-associated CpG islands (CpG island with at least one SP1 binding site) and the number of CpG island-associated

2K-2K ($p_{greater} = 0.191$)

Dataset	Total CpGI	Associated	% of total	Non-associated	% of total
positive	99	23	23.23	76	76.77
negative	151	27	17.88	124	82.12

H2K-2K ($p_{greater} = 0.313$)

Dataset	Total CpGI	Associated	% of total	Non-associated	% of total
positive	84	14	16.67	70	83.33
negative	191	26	13.61	165	86.39

Table 3.15: Number of SP1-associated CpG islands (CpGI) for positive and negative sequences in the mouse *2K-2K* and the human *H2K-2K* datasets and assigned pvalues (*exact Fisher test*).

2K-2K ($p_{greater} = 0.111$)

Dataset	Total SP1	Associated	% of total	Non-associated	% of total
positive	66	31	46.97	35	53.03
negative	102	37	36.27	65	63.73

H2K-2K ($p_{greater} = 0.116$)

Dataset	Total SP1	Associated	% of total	Non-associated	% of total
positive	50	29	58.00	21	42.00
negative	79	36	45.57	43	54.43

Table 3.16: Number of CpG island-associated SP1 for positive and negative sequences in the two *2K* datasets and assigned pvalues (*exact Fisher test*).

SP1 binding sites (SP1 binding sites that are located in CpG islands), respectively. This analysis was performed on the mouse *2K-2K* and the human *H2K-2K* datasets. The results are shown in table 3.15 and 3.16.

Using an *exact Fisher test* to check for a nonrandom distribution, these results **do not indicate a stronger association** of the SP1 binding site and CpG islands in the positive dataset compared to the negative dataset. Indeed, the percentage of CpG islands associated to SP1 binding sites is increased in the positive dataset, however, as both the average number of SP1 bindings sites and CpG islands is higher in the positive dataset, a higher cooccurrence was expected. The same is true for the CpG island-associated SP1 binding sites. Nevertheless, approximately **40-50% of the SP1 binding sites reside in CpG islands** in both the positive and negative dataset. This confirms the proposed SP1-association to CpG islands. However, the reverse is not true as only **13% to 23%** of the found CpG islands include a SP1 binding site.

CTCF binding sites are randomly associated to CpG islands

We also questioned if there is a dependency in the distribution of CTCF binding sites and the presence of CpG islands in the positive and negative dataset, respectively. We therefore categorised each single sequence into one of the four categories "*includes both features*", "*includes only CpG island(s)*", "*includes only CTCF binding site(s)*" and "*included none of these features*". The results for the postive and negative sequences in all four datasets are shown in table 3.17.

Again, the *exact Fisher test* was used to check for a nonrandom distribution between these two features. However, **none of the datasets shows any significant positive or negative association**.

2K-2K

Dataset	**Total**	**both**	**CpGI**	**CTCF**	**none**	**pvalue**
positive	51	30	1	20	0	~ 1
negative	134	51	14	51	18	0.551

2K-next

Dataset	**Total**	**both**	**CpGI**	**CTCF**	**none**	**pvalue**
positive	51	40	1	10	0	~ 1
negative	130	79	12	31	8	0.299

H2K-2K

Dataset	**Total**	**both**	**CpGI**	**CTCF**	**none**	**pvalue**
positive	35	24	0	11	0	~ 1
negative	96	57	8	25	6	0.37

H2K-next

Dataset	**Total**	**both**	**CpGI**	**CTCF**	**none**	**pvalue**
positive	35	27	0	7	0	~ 1
negative	93	76	7	7	3	0.0724

Table 3.17: Statistics for the association between CTCF binding sites and CpG islands for the positive/negative sequences in all four dataset.
both = sequences including both features, **CpGI** = sequences including only CpG island(s), **CTCF** = sequences including only CTCF binding site(s), **none** = sequences including none of these features; pvalues are computed with an *exact Fisher test*

Chapter 4

Conclusion

The aim of this master thesis was to investigate regulatory sequence elements that could lead to a high degree of co-expression of genomic neighbours. This co-expression was stated previously by many groups [4],[5],[7] and it was shown that these pairs have a reduced intergenic length.

The methods used in this thesis consists of sequence analysis techniques that searches for overrepresented subsequences (motifs) in the sequences of these **highly co-expressed gene pairs (HCPs)** and compare their occurences to sequences of **uncorrelated gene pairs (UCPs)**. These motifs may reflect bindings sites of TFs that might contribute to the observed level of correlated expression. Additionally, a search for known vertebrate TFBS (extracted from the TRANSFAC database) was performed. Furthermore, the distribution of certain genomic features like CpG islands and repetetive elements have been investigated over all sequences as well as specific regions.

All these analysis were performed on **two mouse datasets** and **two orthologous human datasets**, that consists of positive (derived from HCPs) and negative (derived from UCPs) sequences and differ in the amount of sequences included around the pair.

Clues for the Existence of Active Chromatin Hubs

Based on our own analysis [7] as well as previous data (see e.g. [6],[120]), we suggest that the cause of the observed clustering of HCPs are large **open chromatin regions (active chromatin hubs - ACHs)**. These regions, that include several genes, are accessible for the basal transcriptional machinery as well as individual TFs. The consequent "opening" of these regions in specific cell types would therefore lead to the correlated expression of genes included in these ACHs. These genes may be additionally regulated on a single-gene level by **individual TFs**.

Indeed, our results point in that direction. Both, the motif and the TFBS search, resulted in a high overrepresentation of TFBS for the transcription factor **SP1**. SP1 is a common TF that was shown to interact with the **histone acetyltransferase (HAT) p300** to induce hyperacetylated chromatin states [106]. Histone acetylation is highly associated with an increased transcriptional activity [39]. The motif search revealed additional overrepresented binding sites that are associated with other TFs that also recruit or co-act with HATs and **histone deacetylases (HDACs)**. However, SP1 was the only factor that could be found in both the motif and the TFBS search.

As SP1 is known to reside in **CpG islands** [87] we also investigated the occurence of these in our sequence datasets. We found CpG islands to reside at the **promoters** of a large fraction of genes in HCPs, while they rarely occur in the promoters of genes from UCPs. CpG islands have been annotated to 40-60% of all mammalian genes and were found in almost all housekeeping gene promoters [87]. Histones in the region of CpG islands were shown to be **highly acetylated** [121] and **H3K4 methylated** [122]. Investigating the association of SP1 binding sites to CpG island, we found a high number of this bindings sites to reside in CpG islands. However, this association was equally strong in the positive and negative datasets.

Furthermore, we found an **enrichment of SINE repeats** in the HCP sequences. These SINE repeats were shown to contain TFBS of several transcription factors [61] that might contribute to the co-expression of genomic pairs. This enrichment of SINE repeats as well as a high GC content was previously shown for very gene dense regions known as **ridges** [9]. About 30 domains in the human genome have been defined as ridges. These contain highly expressed genes with short intron lengths. In addition to the enrichment of SINE repeats a depletion of LINE repeats has been reported in these ridges. However, we did **not find a depletion of LINE repeats** in our datasets. Furthermore, in contrast to the reported 30 gene dense regions, we showed that our HCPs are equally distributed over the mouse/human chromosomes.
However, the direct influence of increased coverage with SINE repeats on correlated expression of gene pairs is presently a matter of speculation.

While investigating the distribution of SINE and LINE repeats we also found a **reduced number of simple repeats** (also called **micro-satellites**) in the HCP sequences com-

pared to UCP sequences. This finding is in accordance with to the theory of active chromatin hubs, as satellite repeats are commonly associated with **heterochromatinic structures**. Therefore these might be subject to negative selection in euchromatic regions, such as the proposed ACHs.

Finally, we found an enrichment for binding sites of the transcription factor **CTCF**, a protein that is known to have **insulatory function** [63]. It is proposed to reside at the edges of euchromatinic regions to prevent heterochromatin from entering. Corresponding to the hypothesis of ACHs we found an increased enrichment of CTCF binding sites at the edges of our sequences. An investigation on the association of CTCF binding sites and CpG island in the same sequences revealed no unusual distribution.

In summary, our results point in the direction of the proposed ACHs. As they could be confirmed using an orthologous human dataset, they may hold true for mammals in general. However, most findings stated by our analysis function at the level of individual genes rather than gene clusters. As no specific regulatory element explaning the observed co-expression could be identified, we suggest that co-expression is a highly complex phenomenon. Our data propose the following theory.

The Shared Systems Strategy

The fact that genomic neighbours share common expression patterns has been shown for many organisms. However, no satisfactory explanation for this observation has been found so far. While single gene clusters as the growth hormone and Hox gene clusters have been analysed in detail, the majority of co-expressed gene pairs in mammalian genomes remains unexplored. For the first mentioned clusters, individual **global and locus control region (LCRs)** and **hypersensitive sites (HSs)** were shown to regulate even distant genes [48],[123],[124]. However, LCRs or HSs have not been identified in other clusters [125]. The theory of open chromatine regions has been proposed as possible explanation for the remaining clusters and in fact **open chromatin fibres** have been shown to correlate to gene dense regions. However, genes in these fibres were not particularly highly expressed [126]. Based on the results of this master thesis, we formulated a novel model for the explanation of gene clustering, termed *shared system strategy*.

The gene **CD74B** is switched on because of its proximity to an actively transcribed gene located in the growth hormon cluster [127]. This behaviour is called "**bystander effect**". We believe this effect to be present not only in pairs consisting of a low expressed gene proximal to a high expressed gene but also between neighbouring genes of common expression patterns. The proximity of these genes might lead to a lowered regulatory and transcriptional "cost" as these genes could **share regulatory** (e.g. LCRs, HATs,

HDACs) **and transcriptional elements** (e.g. specific TFBS). If a gene is located next to another gene that is already active, due to the presence of factors like HATs and SP1, the proximal gene does not need to recruit these factors on its own, but might "share" it with its neighbour. This could **decrease the amount of energy** needed for recruitment and would therefore be an **evolutionary advantage** for the two genes as well as the cell. The same is true for the reverse case of two tissue-specific expressed genes as these could benefit from the local enrichment of common TFs as well as suppressive factors. Chromosomal looping that further reduces the distance between the gene promoters [6] might contribute to this sharing as well as the enrichment of e.g. SINE repeats. An extreme of this mechanism would be the use of **bidirectional promoters** which have been characterised in mammalian genomes [128]. Nevertheless, the clustering of genes should be a **dynamic process** whith an overall **equilibrium**.

The proposed model suggests a clustering not only for housekeeping but also for tissue-specific and all correlated genes in general and is confirmed by the observed existence of such clusters [8],[17],[129],[130]. Furthermore, this proposed the existance of open chromatin regions to be the effect of this clustering process rather than its cause. The open chromatin region present to one gene might be enlarged to include the proximal gene instead of establishing two individual open chromatin regions. This would also explain the existance of large open chromatin fibres in very gene dense regions. Factors like repetetive elements, in particular the observed SINE repeats, might contribute to the sustainment of these open chromatine region by additionally recruiting TFs.

So far, many investigations have been performed both on clusters of housekeeping and on tissue-specific genes. According to our model, the existence of housekeeping and tissue-specific gene clusters would be two sides of the same coin. We believe the *sharing* of common regulatory and transcriptional elements with close genomic neighbours to be a possible explanation for the observed clustering of co-expressed gene pairs as this process could be evolutionary favoured.

In summary, individual genes and their regulatory elements should not be seen as isolated entities but as a dynamic system in the context of neighbouring genes, and *vice versa*.

Chapter 5

Outlook

Today, the influence of the chromatin environment on the expression of single genes or gene clusters is still only partially understood. Transcription factors that influence the chromatin environment directly or indirectly by recruiting other proteins are subject of many investigations which will provide further insight into the mechanisms governing chromatin structure. The impact of certain chromatin states on (correlated) gene expression as well as a clearer understanding on the real processes involved will be gained by high resolution techniques like "Chromatin Immunoprecipitation (ChIP)"-chip analysis. Analysing those results with bioinformatic methods will be needed to enhance current knowledge.

Investigation of the influence of repetetive elements (e.g. SINE) in the shown regulatory mechanisms is of maijor interest, as these have long been thought to be solely parasitic. Furthermore an *in vitro/vivo* analysis could confirm binding of TFs to SINE repeats.

As discussed in the result part of this work, the used motif search programs are developed for relatively short sequences. However, regulatory elements can reside in large distance from their genes and their exact position might change between organisms. Techniques that could overcome this problem (e.g. using conservational information in a more direct and problem-oriented fashion) would highly contribute to *in silico* analyses of such sequence elements.

Finally, these investigations could shed light on the model of shared system strategies implicating that correlated gene pairs share regulatory elements to decrease the transcriptional "costs".

Appendix A

Datasets

The following pages contain name, dataset membership, chromosome, start/end position of the analysed regions and the Ensembl.Gene.IDs for the two genes contained in the pairs. The annotations reflect the datasets *2K-2K* (mouse), *2K-next* (mouse), *H2K-2K* (human), and *H2-next* (human) that were analysed in this master thesis. The dataset definition is described in detail in sections 2.1 to 2.3.
Ensembl.Gene.IDs were extracted from Ensembl 39. The positions are based on the NCBI m36 Assembly (Dec 2005) mm8 (mouse) and NCBI 36 (Oct 2005) hg18 (human).

A.1 Mouse Datasets (mm8)

A.1.1 2K-2K

Name	Set	Chr.	Start	End	Ensembl.Gene.ID.1	Ensembl.Gene.ID.2
2K-2K+1	+	1	108549445	108625994	ENSMUSG00000009905	ENSMUSG00000009907
2K-2K+2	+	1	133692838	133762518	ENSMUSG00000026433	ENSMUSG00000026434
2K-2K+3	+	2	4796836	4858001	ENSMUSG00000026662	ENSMUSG00000026664
2K-2K+4	+	2	18588191	18606385	ENSMUSG00000051154	ENSMUSG00000026739
2K-2K+5	+	2	38828205	38889415	ENSMUSG00000026755	ENSMUSG00000026754
2K-2K+6	+	2	131928935	131946621	ENSMUSG00000027341	ENSMUSG00000027342
2K-2K+7	+	2	151983317	152026076	ENSMUSG00000027465	ENSMUSG00000027466
2K-2K+8	+	2	172641076	172679673	ENSMUSG00000027509	ENSMUSG00000027510
2K-2K+9	+	3	20247962	20346903	ENSMUSG00000002428	ENSMUSG00000019528
2K-2K+10	+	3	88663807	88734841	ENSMUSG00000041355	ENSMUSG00000028059
2K-2K+11	+	3	116478970	116543982	ENSMUSG00000000339	ENSMUSG00000000340
2K-2K+12	+	3	122235614	122279293	ENSMUSG00000028124	ENSMUSG00000039756
2K-2K+13	+	3	138378551	138428769	ENSMUSG00000028138	ENSMUSG00000005813

Definition of mouse 2K-2K *dataset*

Name	Set	Chr.	Start	End	Ensembl.Gene.ID.1	Ensembl.Gene.ID.2
2K-2K+14	+	4	140714760	140812322	ENSMUSG00000006215	ENSMUSG00000040761
2K-2K+15	+	5	117616335	117652063	ENSMUSG00000029364	ENSMUSG00000029363
2K-2K+16	+	5	135432317	135468456	ENSMUSG00000005374	ENSMUSG00000029681
2K-2K+17	+	5	145365783	145392524	ENSMUSG00000029622	ENSMUSG00000029623
2K-2K+18	+	6	17229310	17293324	ENSMUSG00000000058	ENSMUSG00000007655
2K-2K+19	+	6	21897647	22192003	ENSMUSG00000029670	ENSMUSG00000062980
2K-2K+20	+	6	54911680	55011078	ENSMUSG00000002797	ENSMUSG00000029777
2K-2K+21	+	6	86638803	86700988	ENSMUSG00000001158	ENSMUSG00000001157
2K-2K+22	+	6	108624406	108791494	ENSMUSG00000030103	ENSMUSG00000030105
2K-2K+23	+	7	84459344	84557575	ENSMUSG00000030630	ENSMUSG00000030629
2K-2K+24	+	7	114004740	114067641	ENSMUSG00000030754	ENSMUSG00000030751
2K-2K+25	+	8	86496298	86556894	ENSMUSG00000005483	ENSMUSG00000019433
2K-2K+26	+	8	88149235	88200175	ENSMUSG00000031696	ENSMUSG00000031697
2K-2K+27	+	8	97553780	97587371	ENSMUSG00000031776	ENSMUSG00000031775
2K-2K+28	+	8	108492410	108574051	ENSMUSG00000038604	ENSMUSG00000005698
2K-2K+29	+	8	127793494	127922282	ENSMUSG00000031987	ENSMUSG00000056820
2K-2K+30	+	9	53340917	53423096	ENSMUSG00000032047	ENSMUSG00000032030
2K-2K+31	+	9	123213638	123372215	ENSMUSG00000035202	ENSMUSG00000025239
2K-2K+32	+	10	82971771	83080463	ENSMUSG00000034560	ENSMUSG00000020263
2K-2K+33	+	10	126915927	126927384	ENSMUSG00000040280	ENSMUSG00000025403
2K-2K+34	+	11	51461196	51481319	ENSMUSG00000001056	ENSMUSG00000001054
2K-2K+35	+	11	53101239	53178745	ENSMUSG00000020361	ENSMUSG00000018239
2K-2K+36	+	11	76540797	76665191	ENSMUSG00000010392	ENSMUSG00000020841
2K-2K+37	+	12	31849443	31939471	ENSMUSG00000002900	ENSMUSG00000020664
2K-2K+38	+	12	70213335	70277771	ENSMUSG00000020978	ENSMUSG00000020982
2K-2K+39	+	13	24823420	24851611	ENSMUSG00000006717	ENSMUSG00000035958
2K-2K+40	+	13	55600765	55621576	ENSMUSG00000058569	ENSMUSG00000021504
2K-2K+41	+	14	53835600	53854324	ENSMUSG00000022194	ENSMUSG00000022198
2K-2K+42	+	15	34180615	34382376	ENSMUSG00000022257	ENSMUSG00000022324
2K-2K+43	+	15	76152720	76164154	ENSMUSG00000034259	ENSMUSG00000022561
2K-2K+44	+	15	79496136	79517074	ENSMUSG00000022427	ENSMUSG00000022426
2K-2K+45	+	15	80059375	80086794	ENSMUSG00000022412	ENSMUSG00000042406
2K-2K+46	+	16	87342207	87387053	ENSMUSG00000025616	ENSMUSG00000025613
2K-2K+47	+	17	25838917	25856211	ENSMUSG00000024180	ENSMUSG00000024181
2K-2K+48	+	18	10615794	10708694	ENSMUSG00000002477	ENSMUSG00000002475
2K-2K+49	+	18	64623034	64788369	ENSMUSG00000024587	ENSMUSG00000039529
2K-2K+50	+	X	102174891	102227167	ENSMUSG00000031232	ENSMUSG00000031231
2K-2K+51	+	X	135800977	135891549	ENSMUSG00000031432	ENSMUSG00000031431
2K-2K-1	-	1	127502539	127741760	ENSMUSG00000026343	ENSMUSG00000026344
2K-2K-2	-	1	130069727	130157849	ENSMUSG00000026353	ENSMUSG00000026354
2K-2K-3	-	1	137613850	137650362	ENSMUSG00000026418	ENSMUSG00000041782
2K-2K-4	-	2	28375831	28407255	ENSMUSG00000026818	ENSMUSG00000026816
2K-2K-5	-	2	30665290	30727306	ENSMUSG00000039476	ENSMUSG00000050737
2K-2K-6	-	2	62274551	62376362	ENSMUSG00000000394	ENSMUSG00000000392
2K-2K-7	-	2	62396289	62496948	ENSMUSG00000026896	ENSMUSG00000026893
2K-2K-8	-	2	73082711	73140646	ENSMUSG00000041777	ENSMUSG00000008226
2K-2K-9	-	2	74473926	74489855	ENSMUSG00000001823	ENSMUSG00000042499
2K-2K-10	-	2	84779300	84796725	ENSMUSG00000027073	ENSMUSG00000027072
2K-2K-11	-	2	93441303	93625384	ENSMUSG00000040310	ENSMUSG00000027198
2K-2K-12	-	2	113575206	113651495	ENSMUSG00000023236	ENSMUSG00000041219
2K-2K-13	-	2	142752501	142775154	ENSMUSG00000008333	ENSMUSG00000027416
2K-2K-14	-	2	162620982	162667966	ENSMUSG00000016921	ENSMUSG00000035576
2K-2K-15	-	2	164079097	164108653	ENSMUSG00000016995	ENSMUSG00000017007
2K-2K-16	-	2	180427384	180517698	ENSMUSG00000027568	ENSMUSG00000027569
2K-2K-17	-	2	181114712	181134196	ENSMUSG00000016344	ENSMUSG00000038751
2K-2K-18	-	3	14838253	14879513	ENSMUSG00000027559	ENSMUSG00000027562
2K-2K-19	-	3	87930938	87973507	ENSMUSG00000028071	ENSMUSG00000004895

Definition of mouse 2K-2K *dataset*

Name	Set	Chr.	Start	End	Ensembl.Gene.ID.1	Ensembl.Gene.ID.2
2K-2K-20	-	3	90693737	90702341	ENSMUSG00000001023	ENSMUSG00000001025
2K-2K-21	-	3	97712280	97760488	ENSMUSG00000028088	ENSMUSG00000038205
2K-2K-22	-	3	102861831	102913775	ENSMUSG00000027858	ENSMUSG00000027857
2K-2K-23	-	3	135148046	135243813	ENSMUSG00000045328	ENSMUSG00000028167
2K-2K-24	-	3	159428738	159562092	ENSMUSG00000028175	ENSMUSG00000028174
2K-2K-25	-	4	11629604	11747042	ENSMUSG00000028214	ENSMUSG00000028217
2K-2K-26	-	4	25288533	25374390	ENSMUSG00000028259	ENSMUSG00000040359
2K-2K-27	-	4	41219917	41288844	ENSMUSG00000028427	ENSMUSG00000028435
2K-2K-28	-	4	63211443	63349645	ENSMUSG00000050395	ENSMUSG00000028362
2K-2K-29	-	4	125527618	125559843	ENSMUSG00000028859	ENSMUSG00000028861
2K-2K-30	-	4	126811725	126834897	ENSMUSG00000050234	ENSMUSG00000042367
2K-2K-31	-	4	126851390	126862468	ENSMUSG00000046623	ENSMUSG00000042357
2K-2K-32	-	4	138020609	138056219	ENSMUSG00000028749	ENSMUSG00000041202
2K-2K-33	-	4	140015941	140084724	ENSMUSG00000025330	ENSMUSG00000025328
2K-2K-34	-	4	146827776	146847193	ENSMUSG00000029019	ENSMUSG00000041616
2K-2K-35	-	5	33528375	33593839	ENSMUSG00000037379	ENSMUSG00000037373
2K-2K-36	-	5	38021808	38114832	ENSMUSG00000062329	ENSMUSG00000048450
2K-2K-37	-	5	78277883	78359759	ENSMUSG00000053030	ENSMUSG00000029249
2K-2K-38	-	5	93299876	93355381	ENSMUSG00000029410	ENSMUSG00000029413
2K-2K-39	-	5	93421001	93442685	ENSMUSG00000034855	ENSMUSG00000060183
2K-2K-40	-	5	100893023	100961985	ENSMUSG00000029319	ENSMUSG00000035273
2K-2K-41	-	5	104409005	104456404	ENSMUSG00000053268	ENSMUSG00000029307
2K-2K-42	-	5	108626694	108676688	ENSMUSG00000029491	ENSMUSG00000050856
2K-2K-43	-	5	110512597	110579755	ENSMUSG00000029499	ENSMUSG00000007080
2K-2K-44	-	5	115725265	115746110	ENSMUSG00000029522	ENSMUSG00000029524
2K-2K-45	-	6	3910585	3949436	ENSMUSG00000029664	ENSMUSG00000029663
2K-2K-46	-	6	18796636	18831583	ENSMUSG00000044155	ENSMUSG00000029517
2K-2K-47	-	6	55264010	55320108	ENSMUSG00000004655	ENSMUSG00000004654
2K-2K-48	-	6	65518186	65638508	ENSMUSG00000044162	ENSMUSG00000049001
2K-2K-49	-	6	71301062	71371997	ENSMUSG00000053977	ENSMUSG00000002222
2K-2K-50	-	6	83674922	83726468	ENSMUSG00000034777	ENSMUSG00000006269
2K-2K-51	-	6	112423278	112457574	ENSMUSG00000062694	ENSMUSG00000049112
2K-2K-52	-	6	122477962	122531799	ENSMUSG00000030116	ENSMUSG00000040627
2K-2K-53	-	6	122784311	122823624	ENSMUSG00000003154	ENSMUSG00000040552
2K-2K-54	-	6	127036593	127077160	ENSMUSG00000000182	ENSMUSG00000038028
2K-2K-55	-	6	135029383	135084600	ENSMUSG00000046733	ENSMUSG00000030205
2K-2K-56	-	6	136810647	136842001	ENSMUSG00000030217	ENSMUSG00000030218
2K-2K-57	-	7	24169260	24215187	ENSMUSG00000046223	ENSMUSG00000054793
2K-2K-58	-	7	30309852	30336911	ENSMUSG00000006313	ENSMUSG00000036751
2K-2K-59	-	7	80929113	81045929	ENSMUSG00000038763	ENSMUSG00000025726
2K-2K-60	-	7	89739588	89818377	ENSMUSG00000039391	ENSMUSG00000062797
2K-2K-61	-	7	109516642	109545440	ENSMUSG00000035951	ENSMUSG00000031021
2K-2K-62	-	7	109587826	109753599	ENSMUSG00000007279	ENSMUSG00000035901
2K-2K-63	-	7	119891575	119938439	ENSMUSG00000030917	ENSMUSG00000030911
2K-2K-64	-	7	127192816	127215175	ENSMUSG00000045757	ENSMUSG00000045251
2K-2K-65	-	8	3629139	3661817	ENSMUSG00000004626	ENSMUSG00000012705
2K-2K-66	-	8	24087588	24120966	ENSMUSG00000031535	ENSMUSG00000031536
2K-2K-67	-	8	69606954	69637097	ENSMUSG00000044014	ENSMUSG00000036437
2K-2K-68	-	8	71964697	72044699	ENSMUSG00000036330	ENSMUSG00000006273
2K-2K-69	-	8	97632719	97673552	ENSMUSG00000031779	ENSMUSG00000031778
2K-2K-70	-	8	107501270	107522445	ENSMUSG00000031881	ENSMUSG00000031880
2K-2K-71	-	8	119979831	120093161	ENSMUSG00000031845	ENSMUSG00000052557
2K-2K-72	-	8	123981919	124006867	ENSMUSG00000031816	ENSMUSG00000046714
2K-2K-73	-	9	20518452	20578565	ENSMUSG00000004098	ENSMUSG00000053773
2K-2K-74	-	9	39940897	40044188	ENSMUSG00000025602	ENSMUSG00000049281
2K-2K-75	-	9	48649624	48718922	ENSMUSG00000032269	ENSMUSG00000008590
2K-2K-76	-	9	48805564	48869555	ENSMUSG00000032264	ENSMUSG00000032268
2K-2K-77	-	9	54807480	54848727	ENSMUSG00000032303	ENSMUSG00000035200

Definition of mouse 2K-2K *dataset*

Name	Set	Chr.	Start	End	Ensembl.Gene.ID.1	Ensembl.Gene.ID.2
2K-2K-78	-	9	83573418	83671100	ENSMUSG00000032262	ENSMUSG00000038379
2K-2K-79	-	9	107183804	107209784	ENSMUSG00000032579	ENSMUSG00000037977
2K-2K-80	-	9	119330109	119545058	ENSMUSG00000032511	ENSMUSG00000034533
2K-2K-81	-	10	61007366	61071522	ENSMUSG00000020090	ENSMUSG00000020089
2K-2K-82	-	10	79194845	79235619	ENSMUSG00000035863	ENSMUSG00000035852
2K-2K-83	-	10	79287464	79297783	ENSMUSG00000020125	ENSMUSG00000061780
2K-2K-84	-	10	92881228	92948542	ENSMUSG00000015889	ENSMUSG00000020017
2K-2K-85	-	10	127106196	127136716	ENSMUSG00000025401	ENSMUSG00000025400
2K-2K-86	-	10	127668946	127697798	ENSMUSG00000047631	ENSMUSG00000040033
2K-2K-87	-	11	5794639	5813027	ENSMUSG00000020469	ENSMUSG00000041798
2K-2K-88	-	11	7095782	7115909	ENSMUSG00000020429	ENSMUSG00000020427
2K-2K-89	-	11	53408826	53464067	ENSMUSG00000018395	ENSMUSG00000000869
2K-2K-90	-	11	53490993	53570526	ENSMUSG00000020380	ENSMUSG00000036117
2K-2K-91	-	11	71856697	71953973	ENSMUSG00000020808	ENSMUSG00000040543
2K-2K-92	-	11	87607776	87644310	ENSMUSG00000009350	ENSMUSG00000009356
2K-2K-93	-	11	87667445	87693726	ENSMUSG00000034121	ENSMUSG00000052234
2K-2K-94	-	11	98959048	99046412	ENSMUSG00000037944	ENSMUSG00000037935
2K-2K-95	-	11	99242493	99311214	ENSMUSG00000035775	ENSMUSG00000006777
2K-2K-96	-	11	99894947	99914276	ENSMUSG00000046095	ENSMUSG00000048013
2K-2K-97	-	11	99954900	100011336	ENSMUSG00000020911	ENSMUSG00000051617
2K-2K-98	-	11	100148497	100174218	ENSMUSG00000017165	ENSMUSG00000006930
2K-2K-99	-	11	100559701	100581021	ENSMUSG00000035355	ENSMUSG00000045471
2K-2K-100	-	11	101914262	101925442	ENSMUSG00000017316	ENSMUSG00000017311
2K-2K-101	-	11	106125431	106167480	ENSMUSG00000040592	ENSMUSG00000001027
2K-2K-102	-	12	29178925	29224193	ENSMUSG00000036655	ENSMUSG00000061477
2K-2K-103	-	12	36501173	36554755	ENSMUSG00000020581	ENSMUSG00000020577
2K-2K-104	-	13	56258820	56308173	ENSMUSG00000048904	ENSMUSG00000021508
2K-2K-105	-	13	96449413	96547471	ENSMUSG00000021681	ENSMUSG00000021680
2K-2K-106	-	13	96610414	96721173	ENSMUSG00000021678	ENSMUSG00000048376
2K-2K-107	-	13	114290752	114340984	ENSMUSG00000042385	ENSMUSG00000042379
2K-2K-108	-	14	25743872	25831210	ENSMUSG00000040760	ENSMUSG00000040726
2K-2K-109	-	14	33197367	33234834	ENSMUSG00000023064	ENSMUSG00000041445
2K-2K-110	-	14	53662568	53738091	ENSMUSG00000052435	ENSMUSG00000022180
2K-2K-111	-	15	37910458	38173228	ENSMUSG00000037487	ENSMUSG00000061923
2K-2K-112	-	15	39574006	39689541	ENSMUSG00000022303	ENSMUSG00000022304
2K-2K-113	-	15	74546489	74557289	ENSMUSG00000056665	ENSMUSG00000022596
2K-2K-114	-	15	76840742	76886362	ENSMUSG00000018893	ENSMUSG00000033576
2K-2K-115	-	15	89148739	89166776	ENSMUSG00000054136	ENSMUSG00000022613
2K-2K-116	-	15	101970396	101993489	ENSMUSG00000023046	ENSMUSG00000023045
2K-2K-117	-	16	3950371	3995038	ENSMUSG00000005980	ENSMUSG00000005981
2K-2K-118	-	16	26270000	26400125	ENSMUSG00000022512	ENSMUSG00000038148
2K-2K-119	-	16	38264238	38299346	ENSMUSG00000046516	ENSMUSG00000022803
2K-2K-120	-	17	17546242	17590557	ENSMUSG00000003665	ENSMUSG00000045551
2K-2K-121	-	17	31373793	31412057	ENSMUSG00000061613	ENSMUSG00000024041
2K-2K-122	-	17	31847599	31920169	ENSMUSG00000038146	ENSMUSG00000037577
2K-2K-123	-	17	56829330	56881512	ENSMUSG00000019489	ENSMUSG00000005824
2K-2K-124	-	18	34812721	34879482	ENSMUSG00000024366	ENSMUSG00000044201
2K-2K-125	-	18	58679609	58805609	ENSMUSG00000024600	ENSMUSG00000024601
2K-2K-126	-	18	60684137	60719157	ENSMUSG00000024604	ENSMUSG00000049173
2K-2K-127	-	18	61142231	61212428	ENSMUSG00000024619	ENSMUSG00000024620
2K-2K-128	-	19	5458629	5470499	ENSMUSG00000024911	ENSMUSG00000024910
2K-2K-129	-	19	41898527	41951484	ENSMUSG00000047604	ENSMUSG00000035049
2K-2K-130	-	X	6829587	6869216	ENSMUSG00000031147	ENSMUSG00000031148
2K-2K-131	-	X	53556927	53687513	ENSMUSG00000031132	ENSMUSG00000031133
2K-2K-132	-	X	96790873	96823106	ENSMUSG00000044359	ENSMUSG00000060890
2K-2K-133	-	X	155732830	155937827	ENSMUSG00000031298	ENSMUSG00000031295
2K-2K-134	-	X	162647036	162728804	ENSMUSG00000044583	ENSMUSG00000025742

Definition of mouse 2K-2K *dataset*

A.1.2 2K-next

Name	Set	Chr.	Start	End	Ensembl.Gene.ID.1	Ensembl.Gene.ID.2
2K-next+1	+	1	108541822	108688726	ENSMUSG00000009905	ENSMUSG00000009907
2K-next+2	+	1	133676410	133790512	ENSMUSG00000026433	ENSMUSG00000026434
2K-next+3	+	2	4751883	4859445	ENSMUSG00000026662	ENSMUSG00000026664
2K-next+4	+	2	18567101	18616776	ENSMUSG00000051154	ENSMUSG00000026739
2K-next+5	+	2	38828178	38897018	ENSMUSG00000026755	ENSMUSG00000026754
2K-next+6	+	2	131906727	131984684	ENSMUSG00000027341	ENSMUSG00000027342
2K-next+7	+	2	151973282	152029647	ENSMUSG00000027465	ENSMUSG00000027466
2K-next+8	+	2	172636507	172737605	ENSMUSG00000027509	ENSMUSG00000027510
2K-next+9	+	3	20227436	20407666	ENSMUSG00000002428	ENSMUSG00000019528
2K-next+10	+	3	88663069	88737824	ENSMUSG00000041355	ENSMUSG00000028059
2K-next+11	+	3	116440569	116552959	ENSMUSG00000000339	ENSMUSG00000000340
2K-next+12	+	3	122223219	122411819	ENSMUSG00000028124	ENSMUSG00000039756
2K-next+13	+	3	138368184	138463811	ENSMUSG00000028138	ENSMUSG00000005813
2K-next+14	+	4	140708182	140818505	ENSMUSG00000006215	ENSMUSG00000040761
2K-next+15	+	5	117600079	117675029	ENSMUSG00000029364	ENSMUSG00000029363
2K-next+16	+	5	135423012	135471995	ENSMUSG00000005374	ENSMUSG00000029681
2K-next+17	+	5	145362289	145393923	ENSMUSG00000029622	ENSMUSG00000029623
2K-next+18	+	6	17121834	17386305	ENSMUSG00000000058	ENSMUSG00000007655
2K-next+19	+	6	21802517	22195460	ENSMUSG00000029670	ENSMUSG00000062980
2K-next+20	+	6	54901992	55019664	ENSMUSG00000002797	ENSMUSG00000029777
2K-next+21	+	6	86634731	86703397	ENSMUSG00000001158	ENSMUSG00000001157
2K-next+22	+	6	108531366	108794428	ENSMUSG00000030103	ENSMUSG00000030105
2K-next+23	+	7	84247088	84651887	ENSMUSG00000030630	ENSMUSG00000030629
2K-next+24	+	7	113908963	114206461	ENSMUSG00000030754	ENSMUSG00000030751
2K-next+25	+	8	86464649	86556972	ENSMUSG00000005483	ENSMUSG00000019433
2K-next+26	+	8	88121541	88214405	ENSMUSG00000031696	ENSMUSG00000031697
2K-next+27	+	8	97549405	97634718	ENSMUSG00000031776	ENSMUSG00000031775
2K-next+28	+	8	108457428	108579535	ENSMUSG00000038604	ENSMUSG00000005698
2K-next+29	+	8	127788018	127940284	ENSMUSG00000031987	ENSMUSG00000056820
2K-next+30	+	9	53338361	53446504	ENSMUSG00000032047	ENSMUSG00000032030
2K-next+31	+	9	123109244	123378103	ENSMUSG00000035202	ENSMUSG00000025239
2K-next+32	+	10	82963862	83331644	ENSMUSG00000034560	ENSMUSG00000020263
2K-next+33	+	10	126911765	126928421	ENSMUSG00000040280	ENSMUSG00000025403
2K-next+34	+	11	51450271	51486511	ENSMUSG00000001056	ENSMUSG00000001054
2K-next+35	+	11	53044482	53194255	ENSMUSG00000020361	ENSMUSG00000018239
2K-next+36	+	11	76514858	76718353	ENSMUSG00000010392	ENSMUSG00000020841
2K-next+37	+	12	31845125	32004814	ENSMUSG00000002900	ENSMUSG00000020664
2K-next+38	+	12	70202567	70279787	ENSMUSG00000020978	ENSMUSG00000020982
2K-next+39	+	13	24821402	24852601	ENSMUSG00000006717	ENSMUSG00000035958
2K-next+40	+	13	55579592	55630034	ENSMUSG00000058569	ENSMUSG00000021504
2K-next+41	+	14	53830827	53860949	ENSMUSG00000022194	ENSMUSG00000022198
2K-next+42	+	15	34085969	34385143	ENSMUSG00000022257	ENSMUSG00000022324
2K-next+43	+	15	76134501	76164828	ENSMUSG00000034259	ENSMUSG00000022561
2K-next+44	+	15	79494916	79518149	ENSMUSG00000022427	ENSMUSG00000022426
2K-next+45	+	15	80042775	80087868	ENSMUSG00000022412	ENSMUSG00000042406
2K-next+46	+	16	87329763	87438270	ENSMUSG00000025616	ENSMUSG00000025613
2K-next+47	+	17	25826822	25866333	ENSMUSG00000024180	ENSMUSG00000024181
2K-next+48	+	18	10610108	10725622	ENSMUSG00000002477	ENSMUSG00000002475
2K-next+49	+	18	64614436	65012094	ENSMUSG00000024587	ENSMUSG00000039529
2K-next+50	+	X	102131272	102230042	ENSMUSG00000031232	ENSMUSG00000031231
2K-next+51	+	X	135779974	136011319	ENSMUSG00000031432	ENSMUSG00000031431
2K-next-1	-	1	127423361	127742288	ENSMUSG00000026343	ENSMUSG00000026344
2K-next-2	-	1	130065282	130159136	ENSMUSG00000026353	ENSMUSG00000026354
2K-next-3	-	1	137585545	137657285	ENSMUSG00000026418	ENSMUSG00000041782

Definition of mouse 2K-next *dataset*

Name	Set	Chr.	Start	End	Ensembl.Gene.ID.1	Ensembl.Gene.ID.2
2K-next-4	-	2	28375091	28431463	ENSMUSG00000026818	ENSMUSG00000026816
2K-next-5	-	2	30652343	30774980	ENSMUSG00000039476	ENSMUSG00000050737
2K-next-6	-	2	62212145	62398288	ENSMUSG00000000394	ENSMUSG00000000392
2K-next-7	-	2	73076166	73144298	ENSMUSG00000041777	ENSMUSG00000008226
2K-next-8	-	2	74470976	74492808	ENSMUSG00000001823	ENSMUSG00000042499
2K-next-9	-	2	84758796	84796726	ENSMUSG00000027073	ENSMUSG00000027072
2K-next-10	-	2	93365132	93636272	ENSMUSG00000040310	ENSMUSG00000027198
2K-next-11	-	2	113559486	113701036	ENSMUSG00000023236	ENSMUSG00000041219
2K-next-12	-	2	142548869	143237619	ENSMUSG00000008333	ENSMUSG00000027416
2K-next-13	-	2	162600055	162673800	ENSMUSG00000016921	ENSMUSG00000035576
2K-next-14	-	2	164064152	164115451	ENSMUSG00000016995	ENSMUSG00000017007
2K-next-15	-	2	180404261	180518866	ENSMUSG00000027568	ENSMUSG00000027569
2K-next-16	-	2	181086423	181134969	ENSMUSG00000016344	ENSMUSG00000038751
2K-next-17	-	3	14779328	15348524	ENSMUSG00000027559	ENSMUSG00000027562
2K-next-18	-	3	87902029	87992292	ENSMUSG00000028071	ENSMUSG00000004895
2K-next-19	-	3	90691968	90740228	ENSMUSG00000001023	ENSMUSG00000001025
2K-next-20	-	3	97695609	97775228	ENSMUSG00000028088	ENSMUSG00000038205
2K-next-21	-	3	102650067	102969922	ENSMUSG00000027858	ENSMUSG00000027857
2K-next-22	-	3	134901426	135245133	ENSMUSG00000045328	ENSMUSG00000028167
2K-next-23	-	3	158496650	159775029	ENSMUSG00000028175	ENSMUSG00000028174
2K-next-24	-	4	11577631	11886957	ENSMUSG00000028214	ENSMUSG00000028217
2K-next-25	-	4	25099864	25398459	ENSMUSG00000028259	ENSMUSG00000040359
2K-next-26	-	4	41213850	41303098	ENSMUSG00000028427	ENSMUSG00000028435
2K-next-27	-	4	63098339	63381905	ENSMUSG00000050395	ENSMUSG00000028362
2K-next-28	-	4	125210393	125560904	ENSMUSG00000028859	ENSMUSG00000028861
2K-next-29	-	4	126750113	126853389	ENSMUSG00000050234	ENSMUSG00000042367
2K-next-30	-	4	138016650	138071319	ENSMUSG00000028749	ENSMUSG00000041202
2K-next-31	-	4	140014715	140085056	ENSMUSG00000025330	ENSMUSG00000025328
2K-next-32	-	4	146789283	146850283	ENSMUSG00000029019	ENSMUSG00000041616
2K-next-33	-	5	33479555	33695405	ENSMUSG00000037379	ENSMUSG00000037373
2K-next-34	-	5	37937397	38327483	ENSMUSG00000062329	ENSMUSG00000048450
2K-next-35	-	5	78189899	78368948	ENSMUSG00000053030	ENSMUSG00000029249
2K-next-36	-	5	93290609	93359209	ENSMUSG00000029410	ENSMUSG00000029413
2K-next-37	-	5	93403318	93463342	ENSMUSG00000034855	ENSMUSG00000060183
2K-next-38	-	5	100812528	101002453	ENSMUSG00000029319	ENSMUSG00000035273
2K-next-39	-	5	104354036	104537489	ENSMUSG00000053268	ENSMUSG00000029307
2K-next-40	-	5	108613529	108681355	ENSMUSG00000029491	ENSMUSG00000050856
2K-next-41	-	5	110510185	110580494	ENSMUSG00000029499	ENSMUSG00000007080
2K-next-42	-	5	115715202	115750457	ENSMUSG00000029522	ENSMUSG00000029524
2K-next-43	-	6	3672974	3953986	ENSMUSG00000029664	ENSMUSG00000029663
2K-next-44	-	6	18775060	19268246	ENSMUSG00000044155	ENSMUSG00000029517
2K-next-45	-	6	55247890	55381557	ENSMUSG00000004655	ENSMUSG00000004654
2K-next-46	-	6	65387729	65687513	ENSMUSG00000044162	ENSMUSG00000049001
2K-next-47	-	6	71263640	71424171	ENSMUSG00000053977	ENSMUSG00000002222
2K-next-48	-	6	83643458	83728301	ENSMUSG00000034777	ENSMUSG00000006269
2K-next-49	-	6	112404402	112529418	ENSMUSG00000062694	ENSMUSG00000049112
2K-next-50	-	6	122451942	122543410	ENSMUSG00000030116	ENSMUSG00000040627
2K-next-51	-	6	122756500	122840175	ENSMUSG00000003154	ENSMUSG00000040552
2K-next-52	-	6	126990338	127091326	ENSMUSG00000000182	ENSMUSG00000038028
2K-next-53	-	6	134989465	135103220	ENSMUSG00000046733	ENSMUSG00000030205
2K-next-54	-	6	136804704	136871582	ENSMUSG00000030217	ENSMUSG00000030218
2K-next-55	-	7	24142151	24215940	ENSMUSG00000046223	ENSMUSG00000054793
2K-next-56	-	7	30300664	30342377	ENSMUSG00000006313	ENSMUSG00000036751
2K-next-57	-	7	80916232	81087250	ENSMUSG00000038763	ENSMUSG00000025726
2K-next-58	-	7	89729552	89829867	ENSMUSG00000039391	ENSMUSG00000062797
2K-next-59	-	7	109515056	109549231	ENSMUSG00000035951	ENSMUSG00000031021
2K-next-60	-	7	109573203	109763363	ENSMUSG00000007279	ENSMUSG00000035901
2K-next-61	-	7	119852182	119965024	ENSMUSG00000030917	ENSMUSG00000030911

Definition of mouse 2K-next *dataset*

Name	Set	Chr.	Start	End	Ensembl.Gene.ID.1	Ensembl.Gene.ID.2
2K-next-62	-	7	127190149	127233285	ENSMUSG00000045757	ENSMUSG00000045251
2K-next-63	-	8	3625533	3665755	ENSMUSG00000004626	ENSMUSG00000012705
2K-next-64	-	8	24059307	24124757	ENSMUSG00000031535	ENSMUSG00000031536
2K-next-65	-	8	69447325	69789961	ENSMUSG00000044014	ENSMUSG00000036437
2K-next-66	-	8	71836438	72064491	ENSMUSG00000036330	ENSMUSG00000006273
2K-next-67	-	8	107485995	107528967	ENSMUSG00000031881	ENSMUSG00000031880
2K-next-68	-	8	119968427	120143123	ENSMUSG00000031845	ENSMUSG00000052557
2K-next-69	-	8	123974087	124013938	ENSMUSG00000031816	ENSMUSG00000046714
2K-next-70	-	9	20426129	20619102	ENSMUSG00000004098	ENSMUSG00000053773
2K-next-71	-	9	39938627	40048477	ENSMUSG00000025602	ENSMUSG00000049281
2K-next-72	-	9	48587865	48737316	ENSMUSG00000032269	ENSMUSG00000008590
2K-next-73	-	9	48800215	49020503	ENSMUSG00000032264	ENSMUSG00000032268
2K-next-74	-	9	54805917	54947514	ENSMUSG00000032303	ENSMUSG00000035200
2K-next-75	-	9	83391495	83678931	ENSMUSG00000032262	ENSMUSG00000038379
2K-next-76	-	9	107160886	107257833	ENSMUSG00000032579	ENSMUSG00000037977
2K-next-77	-	9	119312929	119602456	ENSMUSG00000032511	ENSMUSG00000034533
2K-next-78	-	10	60977600	61075677	ENSMUSG00000020090	ENSMUSG00000020089
2K-next-79	-	10	79190660	79257796	ENSMUSG00000035863	ENSMUSG00000035852
2K-next-80	-	10	79286304	79297837	ENSMUSG00000020125	ENSMUSG00000061780
2K-next-81	-	10	92740635	92953136	ENSMUSG00000015889	ENSMUSG00000020017
2K-next-82	-	10	127103997	127142486	ENSMUSG00000025401	ENSMUSG00000025400
2K-next-83	-	10	127658851	127699088	ENSMUSG00000047631	ENSMUSG00000040033
2K-next-84	-	11	5776936	5855834	ENSMUSG00000020469	ENSMUSG00000041798
2K-next-85	-	11	7078510	7692313	ENSMUSG00000020429	ENSMUSG00000020427
2K-next-86	-	11	53381343	53474746	ENSMUSG00000018395	ENSMUSG00000000869
2K-next-87	-	11	53489805	53613500	ENSMUSG00000020380	ENSMUSG00000036117
2K-next-88	-	11	71853654	71970124	ENSMUSG00000020808	ENSMUSG00000040543
2K-next-89	-	11	87602123	87669444	ENSMUSG00000009350	ENSMUSG00000009356
2K-next-90	-	11	98905393	99048852	ENSMUSG00000037944	ENSMUSG00000037935
2K-next-91	-	11	99238347	99330713	ENSMUSG00000035775	ENSMUSG00000006777
2K-next-92	-	11	99866642	99918298	ENSMUSG00000046095	ENSMUSG00000048013
2K-next-93	-	11	99952019	100019251	ENSMUSG00000020911	ENSMUSG00000051617
2K-next-94	-	11	100138187	100177805	ENSMUSG00000017165	ENSMUSG00000006930
2K-next-95	-	11	100554281	100582421	ENSMUSG00000035355	ENSMUSG00000045471
2K-next-96	-	11	101904606	101940163	ENSMUSG00000017316	ENSMUSG00000017311
2K-next-97	-	11	106117657	106181561	ENSMUSG00000040592	ENSMUSG00000001027
2K-next-98	-	12	29168752	29235907	ENSMUSG00000036655	ENSMUSG00000061477
2K-next-99	-	12	36460003	36600641	ENSMUSG00000020581	ENSMUSG00000020577
2K-next-100	-	13	56188424	56372522	ENSMUSG00000048904	ENSMUSG00000021508
2K-next-101	-	13	96438541	96577999	ENSMUSG00000021681	ENSMUSG00000021680
2K-next-102	-	13	96579355	96727873	ENSMUSG00000021678	ENSMUSG00000048376
2K-next-103	-	13	114229213	114463425	ENSMUSG00000042385	ENSMUSG00000042379
2K-next-104	-	14	25742143	25865965	ENSMUSG00000040760	ENSMUSG00000040726
2K-next-105	-	14	33184486	33240157	ENSMUSG00000023064	ENSMUSG00000041445
2K-next-106	-	14	53644950	53770691	ENSMUSG00000052435	ENSMUSG00000022180
2K-next-107	-	15	37905645	38182898	ENSMUSG00000037487	ENSMUSG00000061923
2K-next-108	-	15	39512528	39700676	ENSMUSG00000022303	ENSMUSG00000022304
2K-next-109	-	15	74544325	74559507	ENSMUSG00000056665	ENSMUSG00000022596
2K-next-110	-	15	76777990	76895045	ENSMUSG00000018893	ENSMUSG00000033576
2K-next-111	-	15	89142935	89178770	ENSMUSG00000054136	ENSMUSG00000022613
2K-next-112	-	15	101964252	102005031	ENSMUSG00000023046	ENSMUSG00000023045
2K-next-113	-	16	3914581	3997428	ENSMUSG00000005980	ENSMUSG00000005981
2K-next-114	-	16	26021143	26497061	ENSMUSG00000022512	ENSMUSG00000038148
2K-next-115	-	16	38261017	38315355	ENSMUSG00000046516	ENSMUSG00000022803
2K-next-116	-	17	17543690	17592441	ENSMUSG00000003665	ENSMUSG00000045551
2K-next-117	-	17	31365876	31572966	ENSMUSG00000061613	ENSMUSG00000024041
2K-next-118	-	17	31829644	31924979	ENSMUSG00000038146	ENSMUSG00000037577
2K-next-119	-	17	56792849	56889308	ENSMUSG00000019489	ENSMUSG00000005824

Definition of mouse 2K-next *dataset*

Name	Set	Chr.	Start	End	Ensembl.Gene.ID.1	Ensembl.Gene.ID.2
2K-next-120	-	18	34776538	34884879	ENSMUSG00000024366	ENSMUSG00000044201
2K-next-121	-	18	58335175	58962132	ENSMUSG00000024600	ENSMUSG00000024601
2K-next-122	-	18	60684132	60719359	ENSMUSG00000024604	ENSMUSG00000049173
2K-next-123	-	18	61139569	61230940	ENSMUSG00000024619	ENSMUSG00000024620
2K-next-124	-	19	5457505	5474700	ENSMUSG00000024911	ENSMUSG00000024910
2K-next-125	-	19	41884121	41965233	ENSMUSG00000047604	ENSMUSG00000035049
2K-next-126	-	X	6826580	6879182	ENSMUSG00000031147	ENSMUSG00000031148
2K-next-127	-	X	53452231	53733141	ENSMUSG00000031132	ENSMUSG00000031133
2K-next-128	-	X	96780608	96825604	ENSMUSG00000044359	ENSMUSG00000060890
2K-next-129	-	X	155614069	155985934	ENSMUSG00000031298	ENSMUSG00000031295
2K-next-130	-	X	162607912	162790093	ENSMUSG00000044583	ENSMUSG00000025742

Definition of mouse 2K-next *dataset*

A.2 Human Orthologous Datasets (hg18)

A.2.1 H2K-2K

Name	Set	Chr.	Start	End	Ensembl.Gene.ID.1	Ensembl.Gene.ID.2
H2K-2K+1	+	18	59146813	59242673	ENSG00000119537	ENSG00000119541
H2K-2K+3	+	10	13357806	13432303	ENSG00000107537	ENSG00000086475
H2K-2K+4	+	10	22642909	22662419	ENSG00000148444	ENSG00000168283
H2K-2K+7	+	20	334694	393197	ENSG00000125826	ENSG00000125875
H2K-2K+9	+	3	150190024	150289007	ENSG00000163754	ENSG00000071794
H2K-2K+10	+	1	154181269	154259374	ENSG00000116584	ENSG00000163479
H2K-2K+11	+	1	100430324	100532913	ENSG00000137992	ENSG00000137996
H2K-2K+12	+	1	94102427	94149600	ENSG00000067334	ENSG00000023909
H2K-2K+13	+	4	100133903	100227399	ENSG00000164024	ENSG00000197894
H2K-2K+15	+	12	116936893	116985334	ENSG00000111445	ENSG00000176871
H2K-2K+16	+	7	72586622	72632908	ENSG00000106635	ENSG00000106638
H2K-2K+18	+	7	115924680	115990466	ENSG00000105971	ENSG00000105974
H2K-2K+21	+	2	69908336	69987852	ENSG00000087338	ENSG00000124380
H2K-2K+22	+	3	4994208	5199701	ENSG00000134107	ENSG00000134108
H2K-2K+25	+	19	14447572	14492201	ENSG00000123159	ENSG00000132002
H2K-2K+26	+	16	45249095	45291806	ENSG00000069329	ENSG00000091651
H2K-2K+27	+	16	55834539	55878077	ENSG00000102931	ENSG00000102934
H2K-2K+28	+	16	66118221	66232584	ENSG00000039523	ENSG00000102974
H2K-2K+30	+	11	107382618	107525485	ENSG00000166266	ENSG00000075239
H2K-2K+31	+	3	45403072	45698569	ENSG00000011376	ENSG00000144791
H2K-2K+32	+	12	104023622	104156138	ENSG00000136051	ENSG00000136044
H2K-2K+34	+	5	177488603	177515567	ENSG00000145916	ENSG00000145912
H2K-2K+35	+	5	132358579	132470608	ENSG00000155329	ENSG00000170606
H2K-2K+36	+	17	25728110	25879957	ENSG00000108582	ENSG00000108587
H2K-2K+37	+	7	107316847	107433040	ENSG00000091140	ENSG00000091136
H2K-2K+39	+	6	24756184	24811921	ENSG00000111802	ENSG00000112304
H2K-2K+40	+	5	176949814	176971918	ENSG00000184840	ENSG00000027847
H2K-2K+42	+	8	98854461	99119116	ENSG00000104341	ENSG00000132561
H2K-2K+43	+	8	145203510	145215102	ENSG00000178896	ENSG00000197858
H2K-2K+44	+	22	37405900	37428405	ENSG00000100216	ENSG00000100221
H2K-2K+45	+	22	38226230	38250637	ENSG00000100335	ENSG00000128272
H2K-2K+46	+	21	29316809	29369881	ENSG00000156256	ENSG00000156261
H2K-2K+47	+	16	355437	373979	ENSG00000086504	ENSG00000129925
H2K-2K+48	+	18	17444235	17540724	ENSG00000167088	ENSG00000158201
H2K-2K+51	+	X	106756385	106907858	ENSG00000147224	ENSG00000157514

Definition of human H2K-2K *dataset*

Name	Set	Chr.	Start	End	Ensembl.Gene.ID.1	Ensembl.Gene.ID.2
H2K-2K-3	-	1	199614589	199667617	ENSG00000159166	ENSG00000159173
H2K-2K-4	-	9	134893897	134939069	ENSG00000148308	ENSG00000170835
H2K-2K-5	-	9	131465741	131557165	ENSG00000167157	ENSG00000148344
H2K-2K-6	-	2	162706779	162810291	ENSG00000115263	ENSG00000078098
H2K-2K-7	-	2	162829836	162927875	ENSG00000115267	ENSG00000115271
H2K-2K-8	-	2	174919126	175003971	ENSG00000138433	ENSG00000144306
H2K-2K-9	-	2	176670776	176684562	ENSG00000170178	ENSG00000128713
H2K-2K-10	-	11	56898819	56916684	ENSG00000156575	ENSG00000186652
H2K-2K-11	-	11	44071675	44290195	ENSG00000151348	ENSG00000052850
H2K-2K-12	-	15	30692983	30778584	ENSG00000198826	ENSG00000166922
H2K-2K-13	-	20	16656629	16682809	ENSG00000125870	ENSG00000125879
H2K-2K-14	-	20	41517932	41605949	ENSG00000124193	ENSG00000185513
H2K-2K-15	-	20	43353499	43381876	ENSG00000124159	ENSG00000124232
H2K-2K-16	-	20	60808634	60904390	ENSG00000101188	ENSG00000101189
H2K-2K-17	-	20	61620521	61641151	ENSG00000125534	ENSG00000101213
H2K-2K-18	-	8	86535710	86582945	ENSG00000164879	ENSG00000104267
H2K-2K-22	-	1	115371938	115435644	ENSG00000134200	ENSG00000134198
H2K-2K-24	-	1	68665033	68737386	ENSG00000116745	ENSG00000024526
H2K-2K-25	-	8	95206566	95345733	ENSG00000079112	ENSG00000164949
H2K-2K-26	-	6	97074405	97173233	ENSG00000014123	ENSG00000112214
H2K-2K-28	-	9	116589421	116734591	ENSG00000181634	ENSG00000106952
H2K-2K-29	-	1	36691906	36723466	ENSG00000116898	ENSG00000119535
H2K-2K-30	-	1	35017377	35035935	ENSG00000188910	ENSG00000187513
H2K-2K-31	-	1	34991235	35003912	ENSG00000189280	ENSG00000189433
H2K-2K-32	-	1	20309019	20351466	ENSG00000117215	ENSG00000158786
H2K-2K-36	-	4	4910307	5074100	ENSG00000163132	ENSG00000170891
H2K-2K-37	-	4	57368784	57498767	ENSG00000128040	ENSG00000084093
H2K-2K-38	-	4	76998052	77083126	ENSG00000156194	ENSG00000138744
H2K-2K-40	-	4	84402003	84477329	ENSG00000173085	ENSG00000173083
H2K-2K-41	-	4	88749085	88806529	ENSG00000152591	ENSG00000152592
H2K-2K-44	-	12	119222546	119251975	ENSG00000089163	ENSG00000170890
H2K-2K-45	-	7	93350663	93380421	ENSG00000105825	ENSG00000127928
H2K-2K-46	-	7	117609455	117671977	ENSG00000128534	ENSG00000106013
H2K-2K-47	-	7	30915993	30992114	ENSG00000106125	ENSG00000106128
H2K-2K-48	-	4	122174232	122306926	ENSG00000173376	ENSG00000050730
H2K-2K-49	-	2	86798995	86873578	ENSG00000153561	ENSG00000153563
H2K-2K-51	-	3	8748253	8788300	ENSG00000182533	ENSG00000180914
H2K-2K-52	-	12	8646147	8708700	ENSG00000111732	ENSG00000197614
H2K-2K-53	-	12	8074626	8112280	ENSG00000065970	ENSG00000171860
H2K-2K-54	-	12	4298632	4361155	ENSG00000078237	ENSG00000118972
H2K-2K-56	-	12	14871512	14932025	ENSG00000111339	ENSG00000111341
H2K-2K-57	-	19	48816362	48868539	ENSG00000105767	ENSG00000011422
H2K-2K-58	-	19	40828995	40863207	ENSG00000126267	ENSG00000105668
H2K-2K-60	-	11	85688936	85813798	ENSG00000149196	ENSG00000149201
H2K-2K-61	-	11	8913695	8944578	ENSG00000176009	ENSG00000175348
H2K-2K-64	-	16	30470587	30493556	ENSG00000169951	ENSG00000197162
H2K-2K-65	-	19	7606010	7643340	ENSG00000076944	ENSG00000104918
H2K-2K-66	-	8	42313131	42355832	ENSG00000070501	ENSG00000104371
H2K-2K-67	-	4	164462567	164494534	ENSG00000164128	ENSG00000164129
H2K-2K-68	-	8	20044646	20125485	ENSG00000036565	ENSG00000147416
H2K-2K-69	-	16	55948219	55978455	ENSG00000102962	ENSG00000006210
H2K-2K-70	-	16	65497526	65518939	ENSG00000166589	ENSG00000166592
H2K-2K-71	-	16	79827797	79973441	ENSG00000135697	ENSG00000127688
H2K-2K-73	-	19	9929237	9995954	ENSG00000080573	ENSG00000080511
H2K-2K-74	-	11	123003107	123119573	ENSG00000166257	ENSG00000166261
H2K-2K-75	-	11	113278729	113368241	ENSG00000149305	ENSG00000166736
H2K-2K-76	-	11	113061483	113151635	ENSG00000166682	ENSG00000086827

Definition of human H2K-2K *dataset*

Name	Set	Chr.	Start	End	Ensembl.Gene.ID.1	Ensembl.Gene.ID.2
H2K-2K-78	-	6	80679248	80810958	ENSG00000118402	ENSG00000112742
H2K-2K-79	-	3	50568467	50599426	ENSG00000088543	ENSG00000114735
H2K-2K-80	-	3	38562558	38812505	ENSG00000183873	ENSG00000185313
H2K-2K-81	-	10	71630592	71698154	ENSG00000180817	ENSG00000148734
H2K-2K-82	-	19	658002	717318	ENSG00000099864	ENSG00000099812
H2K-2K-83	-	19	801291	816606	ENSG00000197561	ENSG00000197766
H2K-2K-84	-	12	94889273	94955496	ENSG00000084110	ENSG00000111144
H2K-2K-85	-	12	55688053	55732160	ENSG00000166863	ENSG00000166866
H2K-2K-87	-	7	44142990	44197563	ENSG00000106631	ENSG00000106633
H2K-2K-88	-	7	45892484	45929396	ENSG00000146678	ENSG00000146674
H2K-2K-89	-	5	132035272	132103229	ENSG00000113520	ENSG00000131437
H2K-2K-90	-	5	131903035	132009651	ENSG00000113525	ENSG00000113522
H2K-2K-91	-	17	6286483	6402601	ENSG00000129195	ENSG00000091622
H2K-2K-92	-	17	53668847	53715281	ENSG00000167419	ENSG00000005381
H2K-2K-93	-	17	53623088	53650606	ENSG00000121053	ENSG00000011143
H2K-2K-95	-	17	36283721	36349362	ENSG00000171431	ENSG00000108244
H2K-2K-96	-	17	36867588	36893194	ENSG00000108759	ENSG00000197079
H2K-2K-101	-	17	59357832	59406010	ENSG00000007312	ENSG00000007314
H2K-2K-103	-	7	16757853	16813133	ENSG00000106537	ENSG00000106541
H2K-2K-104	-	5	134896566	134944868	ENSG00000181965	ENSG00000145824
H2K-2K-105	-	5	76282436	76398815	ENSG00000145708	ENSG00000164252
H2K-2K-107	-	5	54307449	54368155	ENSG00000164283	ENSG00000113088
H2K-2K-109	-	10	88683277	88714997	ENSG00000173269	ENSG00000173267
H2K-2K-110	-	14	22654387	22724689	ENSG00000092067	ENSG00000092068
H2K-2K-112	-	8	105419228	105550453	ENSG00000164935	ENSG00000147647
H2K-2K-113	-	8	143803623	143822831	ENSG00000130193	ENSG00000126233
H2K-2K-116	-	12	51775703	51806589	ENSG00000167779	ENSG00000167780
H2K-2K-118	-	3	191504197	191613027	ENSG00000163347	ENSG00000113946
H2K-2K-119	-	3	120841596	120880933	ENSG00000121577	ENSG00000138495
H2K-2K-120	-	19	56906177	56948912	ENSG00000105509	ENSG00000171051
H2K-2K-121	-	21	43384136	43467982	ENSG00000160201	ENSG00000160202
H2K-2K-122	-	19	15129445	15206231	ENSG00000074181	ENSG00000105131
H2K-2K-123	-	19	6534867	6623599	ENSG00000125726	ENSG00000125735
H2K-2K-125	-	5	128326720	128479612	ENSG00000113396	ENSG00000066583
H2K-2K-129	-	10	99080244	99153090	ENSG00000181274	ENSG00000052749
H2K-2K-130	-	X	48855278	48913766	ENSG00000068394	ENSG00000017621
H2K-2K-131	-	X	135556002	135693913	ENSG00000102245	ENSG00000129675
H2K-2K-133	-	X	18818339	19052676	ENSG00000044446	ENSG00000173698
H2K-2K-134	-	X	12717452	12820420	ENSG00000101911	ENSG00000196664

Definition of human H2K-2K *dataset*

A.2.2 H2K-next

Name	Set	Chr.	Start	End	Ensembl.Gene.ID.1	Ensembl.Gene.ID.2
H2K-next+1	+	18	59137594	59295198	ENSG00000119537	ENSG00000119541
H2K-next+3	+	10	13316338	13520642	ENSG00000107537	ENSG00000086475
H2K-next+4	+	10	22596142	22674404	ENSG00000148444	ENSG00000168283
H2K-next+7	+	20	326206	411339	ENSG00000125826	ENSG00000125875
H2K-next+9	+	3	150188144	150294832	ENSG00000163754	ENSG00000071794
H2K-next+10	+	1	154179365	154271715	ENSG00000116584	ENSG00000163479
H2K-next+11	+	1	100416389	100590539	ENSG00000137992	ENSG00000137996
H2K-next+12	+	1	94085295	94160128	ENSG00000067334	ENSG00000023909
H2K-next+13	+	4	100070140	100263854	ENSG00000164024	ENSG00000197894
H2K-next+15	+	12	116777725	116985782	ENSG00000111445	ENSG00000176871
H2K-next+16	+	7	72574552	72645459	ENSG00000106635	ENSG00000106638

Definition of human H2K-next *dataset*

Name	**Set**	**Chr.**	**Start**	**End**	**Ensembl.Gene.ID.1**	**Ensembl.Gene.ID.2**
H2K-next+18	+	7	115686072	116099694	ENSG00000105971	ENSG00000105974
H2K-next+21	+	2	69906411	69993330	ENSG00000087338	ENSG00000124380
H2K-next+22	+	3	4895387	5204226	ENSG00000134107	ENSG00000134108
H2K-next+25	+	19	14447175	14501354	ENSG00000123159	ENSG00000132002
H2K-next+26	+	16	45218252	45298959	ENSG00000069329	ENSG00000091651
H2K-next+27	+	16	55831888	55950218	ENSG00000102931	ENSG00000102934
H2K-next+28	+	16	66075218	66236530	ENSG00000039523	ENSG00000102974
H2K-next+30	+	11	107339417	107534878	ENSG00000166266	ENSG00000075239
H2K-next+31	+	3	45272117	45705467	ENSG00000011376	ENSG00000144791
H2K-next+32	+	12	104002367	104248543	ENSG00000136051	ENSG00000136044
H2K-next+34	+	5	177485714	177545325	ENSG00000145916	ENSG00000145912
H2K-next+35	+	5	132356792	132560050	ENSG00000155329	ENSG00000170606
H2K-next+36	+	17	25711718	25908255	ENSG00000108582	ENSG00000108587
H2K-next+37	+	7	107230907	107451231	ENSG00000091140	ENSG00000091136
H2K-next+39	+	6	24754363	24813067	ENSG00000111802	ENSG00000112304
H2K-next+40	+	5	176914115	176978508	ENSG00000184840	ENSG00000027847
H2K-next+42	+	8	98853828	99123129	ENSG00000104341	ENSG00000132561
H2K-next+43	+	8	145186924	145221967	ENSG00000178896	ENSG00000197858
H2K-next+44	+	22	37404351	37431752	ENSG00000100216	ENSG00000100221
H2K-next+45	+	22	38215386	38255047	ENSG00000100335	ENSG00000128272
H2K-next+46	+	21	29313564	29380052	ENSG00000156256	ENSG00000156261
H2K-next+47	+	16	352534	376763	ENSG00000086504	ENSG00000129925
H2K-next+48	+	18	17434844	17558077	ENSG00000167088	ENSG00000158201
H2K-next+51	+	X	106733260	106924106	ENSG00000147224	ENSG00000157514
H2K-next-3	-	1	199613432	199701242	ENSG00000159166	ENSG00000159173
H2K-next-4	-	9	134886375	134946516	ENSG00000148308	ENSG00000170835
H2K-next-5	-	9	131444266	131605252	ENSG00000167157	ENSG00000148344
H2K-next-6	-	2	162639299	162831835	ENSG00000115263	ENSG00000078098
H2K-next-8	-	2	174910398	175004620	ENSG00000138433	ENSG00000144306
H2K-next-9	-	2	176668047	176689745	ENSG00000170178	ENSG00000128713
H2K-next-10	-	11	56894126	56931004	ENSG00000156575	ENSG00000186652
H2K-next-11	-	11	44062146	44543716	ENSG00000151348	ENSG00000052850
H2K-next-12	-	15	30681857	30797496	ENSG00000198826	ENSG00000166922
H2K-next-13	-	20	16599388	16911475	ENSG00000125870	ENSG00000125879
H2K-next-14	-	20	41513315	41621021	ENSG00000124193	ENSG00000185513
H2K-next-15	-	20	43316621	43384414	ENSG00000124159	ENSG00000124232
H2K-next-16	-	20	60802278	60906621	ENSG00000101188	ENSG00000101189
H2K-next-17	-	20	61600950	61642606	ENSG00000125534	ENSG00000101213
H2K-next-18	-	8	86478496	86742205	ENSG00000164879	ENSG00000104267
H2K-next-22	-	1	115339515	115630059	ENSG00000134200	ENSG00000134198
H2K-next-24	-	1	68481327	69806668	ENSG00000116745	ENSG00000024526
H2K-next-25	-	8	95007471	95453364	ENSG00000079112	ENSG00000164949
H2K-next-26	-	6	96837461	97204049	ENSG00000014123	ENSG00000112214
H2K-next-28	-	9	116448524	116822633	ENSG00000181634	ENSG00000106952
H2K-next-29	-	1	36688640	37033714	ENSG00000116898	ENSG00000119535
H2K-next-30	-	1	35016833	35091718	ENSG00000188910	ENSG00000187513
H2K-next-31	-	1	34907883	35016748	ENSG00000189280	ENSG00000189433
H2K-next-32	-	1	20290249	20363070	ENSG00000117215	ENSG00000158786
H2K-next-36	-	4	4594677	5104427	ENSG00000163132	ENSG00000170891
H2K-next-37	-	4	57367055	57524272	ENSG00000128040	ENSG00000084093
H2K-next-38	-	4	76971481	77090082	ENSG00000156194	ENSG00000138744
H2K-next-40	-	4	84254936	84547526	ENSG00000173085	ENSG00000173083
H2K-next-41	-	4	88669411	88884129	ENSG00000152591	ENSG00000152592
H2K-next-44	-	12	119215424	119263515	ENSG00000089163	ENSG00000170890
H2K-next-45	-	7	93059023	93388973	ENSG00000105825	ENSG00000127928
H2K-next-46	-	7	117300798	119015496	ENSG00000128534	ENSG00000106013
H2K-next-47	-	7	30898146	31058666	ENSG00000106125	ENSG00000106128

Definition of human H2K-next *dataset*

Name	Set	Chr.	Start	End	Ensembl.Gene.ID.1	Ensembl.Gene.ID.2
H2K-next-48	-	4	122063464	122333401	ENSG00000173376	ENSG00000050730
H2K-next-49	-	2	86703521	86895972	ENSG00000153561	ENSG00000153563
H2K-next-51	-	3	8668761	8893759	ENSG00000182533	ENSG00000180914
H2K-next-52	-	12	8641502	8743708	ENSG00000111732	ENSG00000197614
H2K-next-53	-	12	8059052	8126104	ENSG00000065970	ENSG00000171860
H2K-next-54	-	12	4284778	4413568	ENSG00000078237	ENSG00000118972
H2K-next-56	-	12	14868054	14958240	ENSG00000111339	ENSG00000111341
H2K-next-57	-	19	48815847	48912077	ENSG00000105767	ENSG00000011422
H2K-next-58	-	19	40827614	40895669	ENSG00000126267	ENSG00000105668
H2K-next-60	-	11	85667427	85829797	ENSG00000149196	ENSG00000149201
H2K-next-61	-	11	8911104	8960969	ENSG00000176009	ENSG00000175348
H2K-next-64	-	16	30453696	30497794	ENSG00000169951	ENSG00000197162
H2K-next-65	-	19	7604637	7647513	ENSG00000076944	ENSG00000104918
H2K-next-66	-	8	42309131	42368546	ENSG00000070501	ENSG00000104371
H2K-next-67	-	4	164307524	164612455	ENSG00000164128	ENSG00000164129
H2K-next-68	-	8	19867913	20131569	ENSG00000036565	ENSG00000147416
H2K-next-70	-	16	65479357	65523468	ENSG00000166589	ENSG00000166592
H2K-next-71	-	16	79811477	80036395	ENSG00000135697	ENSG00000127688
H2K-next-73	-	19	9908229	10014155	ENSG00000080573	ENSG00000080511
H2K-next-74	-	11	122998973	123129497	ENSG00000166257	ENSG00000166261
H2K-next-75	-	11	113251467	113435524	ENSG00000149305	ENSG00000166736
H2K-next-76	-	11	112986680	113173807	ENSG00000166682	ENSG00000086827
H2K-next-78	-	6	80470092	80873082	ENSG00000118402	ENSG00000112742
H2K-next-79	-	3	50516033	50618924	ENSG00000088543	ENSG00000114735
H2K-next-80	-	3	38558442	38862263	ENSG00000183873	ENSG00000185313
H2K-next-81	-	10	71600286	71728734	ENSG00000180817	ENSG00000148734
H2K-next-82	-	19	646484	748410	ENSG00000099864	ENSG00000099812
H2K-next-83	-	19	799176	818960	ENSG00000197561	ENSG00000197766
H2K-next-84	-	12	94886501	95019859	ENSG00000084110	ENSG00000111144
H2K-next-85	-	12	55686498	55735697	ENSG00000166863	ENSG00000166866
H2K-next-87	-	7	44128239	44207102	ENSG00000106631	ENSG00000106633
H2K-next-88	-	7	45849867	45981500	ENSG00000146678	ENSG00000146674
H2K-next-89	-	5	132024702	132111035	ENSG00000113520	ENSG00000131437
H2K-next-90	-	5	131854390	132021763	ENSG00000113525	ENSG00000113522
H2K-next-91	-	17	6279244	6422374	ENSG00000129195	ENSG00000091622
H2K-next-92	-	17	53648607	53733594	ENSG00000167419	ENSG00000005381
H2K-next-95	-	17	36276989	36368194	ENSG00000171431	ENSG00000108244
H2K-next-96	-	17	36851123	36896153	ENSG00000108759	ENSG00000197079
H2K-next-101	-	17	59349887	59433707	ENSG00000007312	ENSG00000007314
H2K-next-103	-	7	16712663	16865553	ENSG00000106537	ENSG00000106541
H2K-next-104	-	5	134815888	135198314	ENSG00000181965	ENSG00000145824
H2K-next-105	-	5	76252608	76403652	ENSG00000145708	ENSG00000164252
H2K-next-107	-	5	53875479	54372431	ENSG00000164283	ENSG00000113088
H2K-next-109	-	10	88674926	88715477	ENSG00000173269	ENSG00000173267
H2K-next-110	-	14	22639501	22798513	ENSG00000092067	ENSG00000092068
H2K-next-112	-	8	105333264	105565772	ENSG00000164935	ENSG00000147647
H2K-next-113	-	8	143782603	143828630	ENSG00000130193	ENSG00000126233
H2K-next-116	-	12	51759438	51818595	ENSG00000167779	ENSG00000167780
H2K-next-118	-	3	191365998	191629141	ENSG00000163347	ENSG00000113946
H2K-next-119	-	3	120831342	120902675	ENSG00000121577	ENSG00000138495
H2K-next-120	-	19	56900255	56955994	ENSG00000105509	ENSG00000171051
H2K-next-121	-	21	43369542	43573005	ENSG00000160201	ENSG00000160202
H2K-next-122	-	19	15117123	15209300	ENSG00000074181	ENSG00000105131
H2K-next-123	-	19	6524069	6628877	ENSG00000125726	ENSG00000125735
H2K-next-125	-	5	127901635	128824001	ENSG00000113396	ENSG00000066583
H2K-next-129	-	10	99071663	99170943	ENSG00000181274	ENSG00000052749
H2K-next-130	-	X	48845004	48915188	ENSG00000068394	ENSG00000017621
H2K-next-131	-	X	135552255	135726303	ENSG00000102245	ENSG00000129675

Definition of human H2K-next *dataset*

Name	**Set**	**Chr.**	**Start**	**End**	**Ensembl.Gene.ID.1**	**Ensembl.Gene.ID.2**
H2K-next-133	-	X	18756098	19271967	ENSG00000044446	ENSG00000173698
H2K-next-134	-	X	12652564	12834678	ENSG00000101911	ENSG00000196664

Definition of human H2K-next *dataset*

Appendix B

Description & Overview of Used Scripts

B.1 Description of Scripts

1. **SequenceExtractor.pl**
   The SequenceExtractor script extract the nucleotide sequence for a specified region on any chromosome of the Mouse February 2006 (mm8) assembly from NCBI (Build 36) and the actual Human March 2006 (hg18) assembly from NCBI (Build 35). It provides some attributes to format the output sequence. The script is described in section 2.4.

2. **FeatureExtractor.pl**
   The FeatureExtractor script extracts specific features for a specified region on any chromosome of the Mouse February 2006 (mm8) assembly from NCBI (Build 36) and the actual Human March 2006 (hg18) assembly from NCBI (Build 35).
   The following features can be annotated by the script:
   - CpG island
   - CpG region
   - SP1 bindings site (TRANSFAC)
   - GC Box
   - CTCF binding site
   - TATA Box (TRANSFAC)
   - All repeat classes as annotated by UCSC genome browser
   - Regions with "Regulatory Potential"

The annotation is returned using three different formats:

a) **Full (Potentially Overlapping) Masking List**
The full masking list contains the exact and sorted start/end positions of all extracted features in the specified region (using chromosome based coordinates). As several features could overlap with each other, this list might contain overlapping regions.

b) **Masking String**
The masking string is a one-to-one representation of the annotation. Each nucleotide of the sequence is assigned by an appropriate symbol that represents the feature that is annotated to that particular position. An overlap of features is indicated by the symbol '∼'. Nucleotide positions that lack any feature assignment are represented by a '-'.

c) **Non-overlapping Masking List**
The non-overlapping masking list contains sorted start/end positions of all extracted features in the specified region (using chromosome based coordinates). In contrast to the full masking list, overlaps between distinct features are indicated by "overlap" entries.

The script is described in section 2.5.

3. **SequenceMasker.pl**
The SequenceMasker script takes the sequence extracted by *SequenceExtractor.pl* and the appropriate feature annotation provided by *FeatureExtractor.pl* to compute a masked sequence that can be used as input to the motif finding programs. The masking process comprises of changing every nucleotide, that is masked by a certain feature - as indicated by the full masking list - into a user-specified character. The script provides the possibility to include regulatory potential information into the masking. It is described in section 2.6.

4. **ProjectHandler.pl**
The ProjectHandler script was implemented to build a user-friendly interface to the three scripts *SequenceExtractor.pl*, *FeatureExtractor.pl*, and *SequenceMasker.pl*. It is controlled by *project files* which include the position of the *dataset definition file* (which itself contains the start/end positions and additional localisation parameters for every single region of the positive/negative dataset) and the output directory. The files are structured according to the key-value principal. It is possible to enable/disable every single feature from the extraction procedure and to define the appropriate parameters individually. The process is then guided by ProjectHandler using the SequenceExtractor script to extract the sequence of every defined region. Afterwards the FeatureExtraction script is started to extract the appropriate annotation of that regions. The masking step is fulfilled by the SequenceMasker script again for every region defined in the *dataset definition file*.

5. **PreMotifFinder.pl**
   The PreMotifFinder script concatenates all masked sequences from the positive dataset gained from the sequence masking procedure into one file in FASTA format that can afterwards be used as input for the motif finding programs. The script is mentioned in section 2.7.

6. **Motif Finding Programs**
   Motif finding programs try to find overrepresented subsequences (motifs) in an input set of sequences. We used different programs that have been developed by several groups:

   - **AlignACE**
   - **MEME**
   - **BioProspector**
   - **Improbizer**

   For further details on these motif finding programs refer to sections 2.7.1 to 2.7.4.

7. **PostMotifFinder.pl**
   The PostMotifFinder script converts the output (the found motifs) of the single motif finding programs into a common file format (called *mot file format*) that is then used as input for MAST. MAST screens a database of nucleotide sequences for occurrences of one motif provided by the user. The searching process is based on a statistical model to ensure significance of the found positions and therefore to search for reliable occurrences of this motifs. Mast is described in section 2.9. The PostMotifFinder script also manages the call for *ScoreMotifs.pl* that computes motif statistics like the ratio of group frequencies and the group specificity score (see section 2.10) for a single motif in the whole (positive and negative) sequence database. *PostMotifFinder.pl* guides the pairwise comparison between motifs and the clustering of similar motifs using the CompareACE and TREE programs. Furthermore it calls the sequence logo drawing program (WebLogo) and the comparison of the found motifs to the TRANSFAC vertebrate PSSM. Finally it outputs a latex-style ASCII file which is afterwards translated into a PDF file. The script and the appropriate programs and approaches are described in section 2.9 to 2.13.

8. **FeatureStatistics.pl & FeatureStatistics.R**
   The two FeatureStatistics script read the extracted features prepared by the FeatureExtraction script and compute median, mean and density for each single feature. The perl scripts converts the start/end annotations of the *FeatureExtraction.pl* output into the total number of a particular feature contained in the region of interest as well as the percentage coverage of that region. These values are additionally computed for each subregion (like the intergenic or transcript regions).

The appropriate R script takes these values and computes means, medians and densities using the standard R functions (*density* is called with *from=0*). Additionally, a Wilcoxon rank sum test is performed on each individual distribution using the standard *wilcox.test* function with the two data vectors (one from the positive dataset and one from the negative) to investigate difference in distribution between these two sets. Finally, the density, mean and median are plotted for each dataset separately. The approach is described in section 2.15.

9. **DistributionExtractor.pl**
   The DistributionExtractor computes 0-vectors in the length of the region. For each features to be analysed one vector is computed and filled with 1s at positions that are annotated by the specific feature. These 0-1-vectors can then be imported into R and used by the FeaturePlotter.

10. **FeaturePlotter Package**
    The FeaturePlotter package consist of two kinds of R functions:

    a) **Functions that adjust/map distributions over regions of different size to the same length**
    b) **The plotFeature function that draws the FeaturePlotter plots**

    Both kinds of functions are described in detail in section 2.16.

B.2 Interaction Diagram

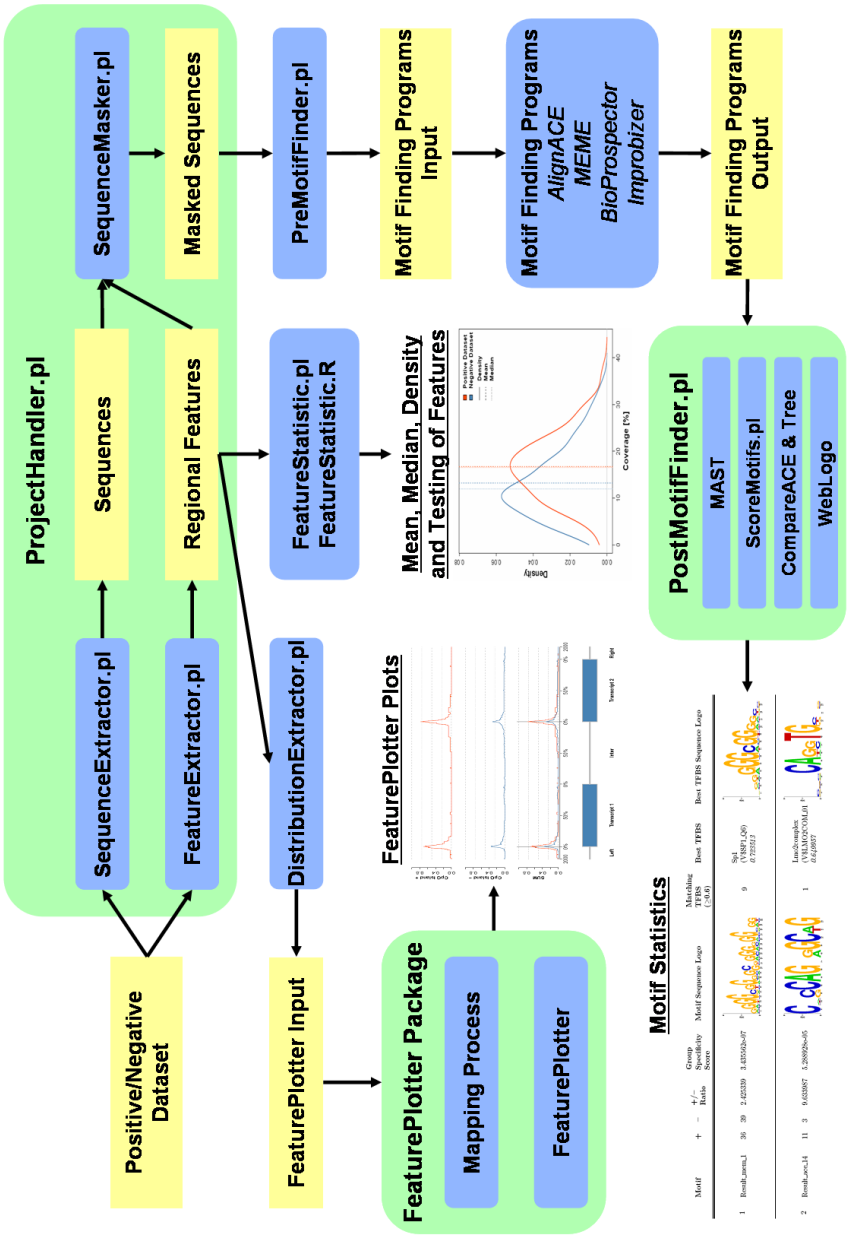


Figure B.1: Diagram of the interaction between the different scripts and programs used in this master thesis. Yellow boxes represent data files, blue rounded boxes represent scripts and programs. Green rounded boxes depict packages or scripts that are programmed to call other scripts. The arrows illustrate the dataflow.

Appendix C

Feature Distribution Figures

C.1 Mean, Median and Density

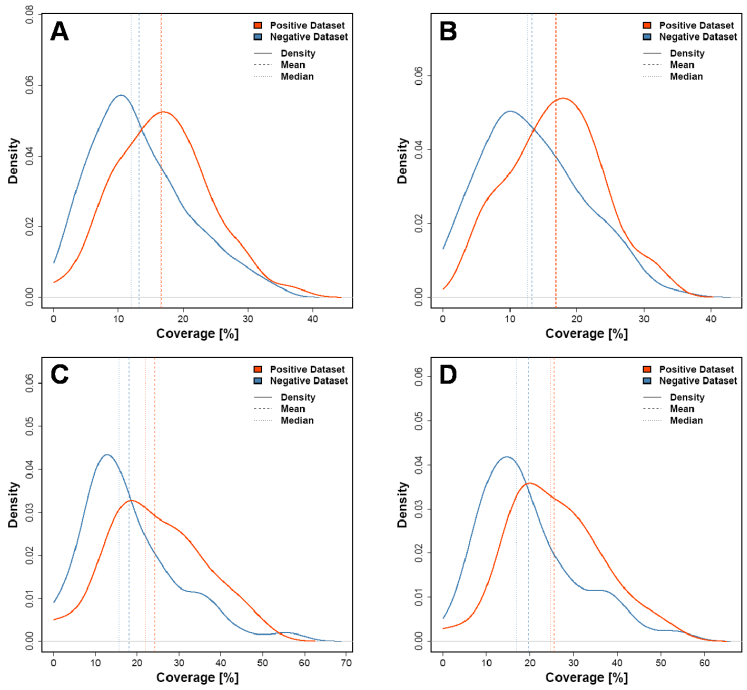


Figure C.1: Mean, median and density for the procentage coverage of **SINE** repeats. **A:** *2K-2K* $p = 2.3x10^{-3}$ **B:** *2K-next* $p = 3.57x10^{-3}$ **C:** *H2K-2K* $p = 4.19x10^{-3}$ **D:** H2K-next $p = 1.34x10^{-3}$.

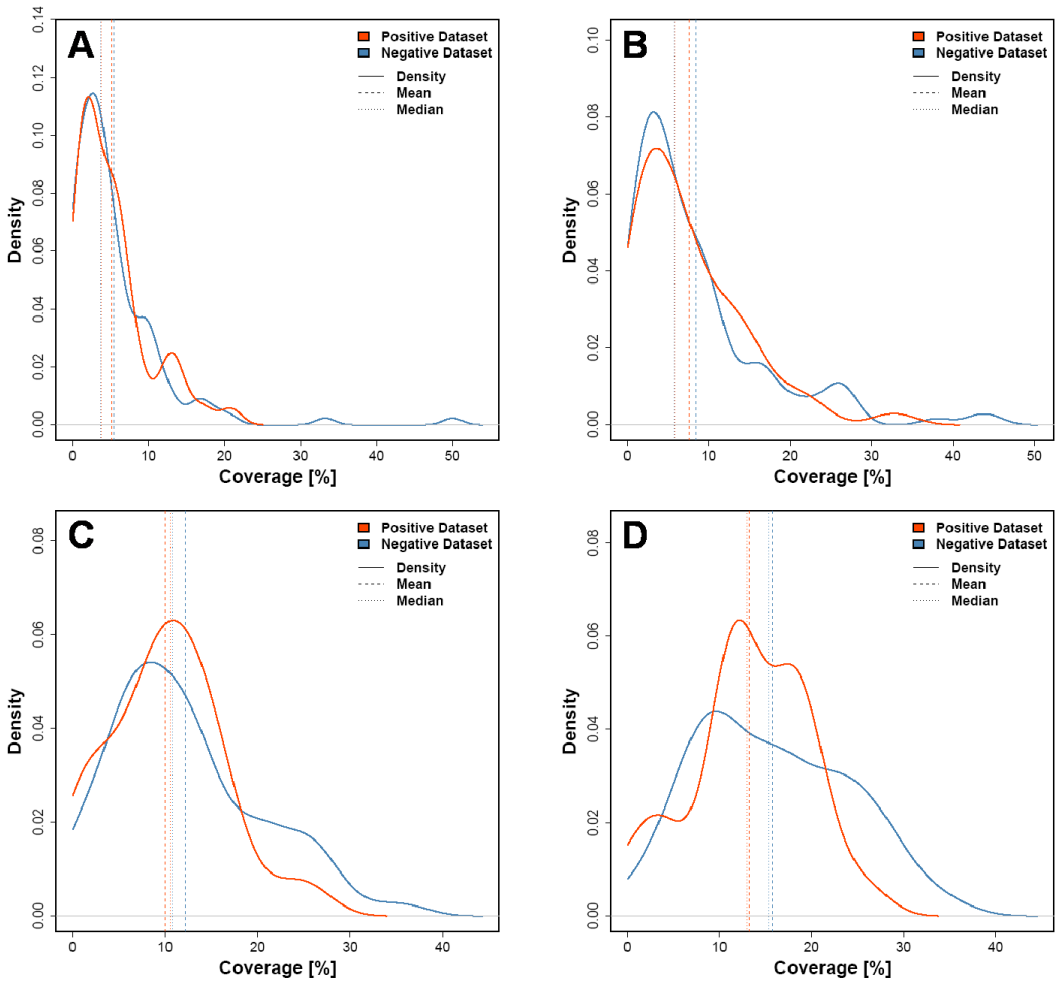


Figure C.2: Mean, median and density for the procentage coverage of **LINE** repeats. **A:** *2K-2K* $p = 0.916$ **B:** *2K-next* $p = 0.908$ **C:** *H2K-2K* $p = 0.165$ **D:** *H2K-next* $p = 0.245$.

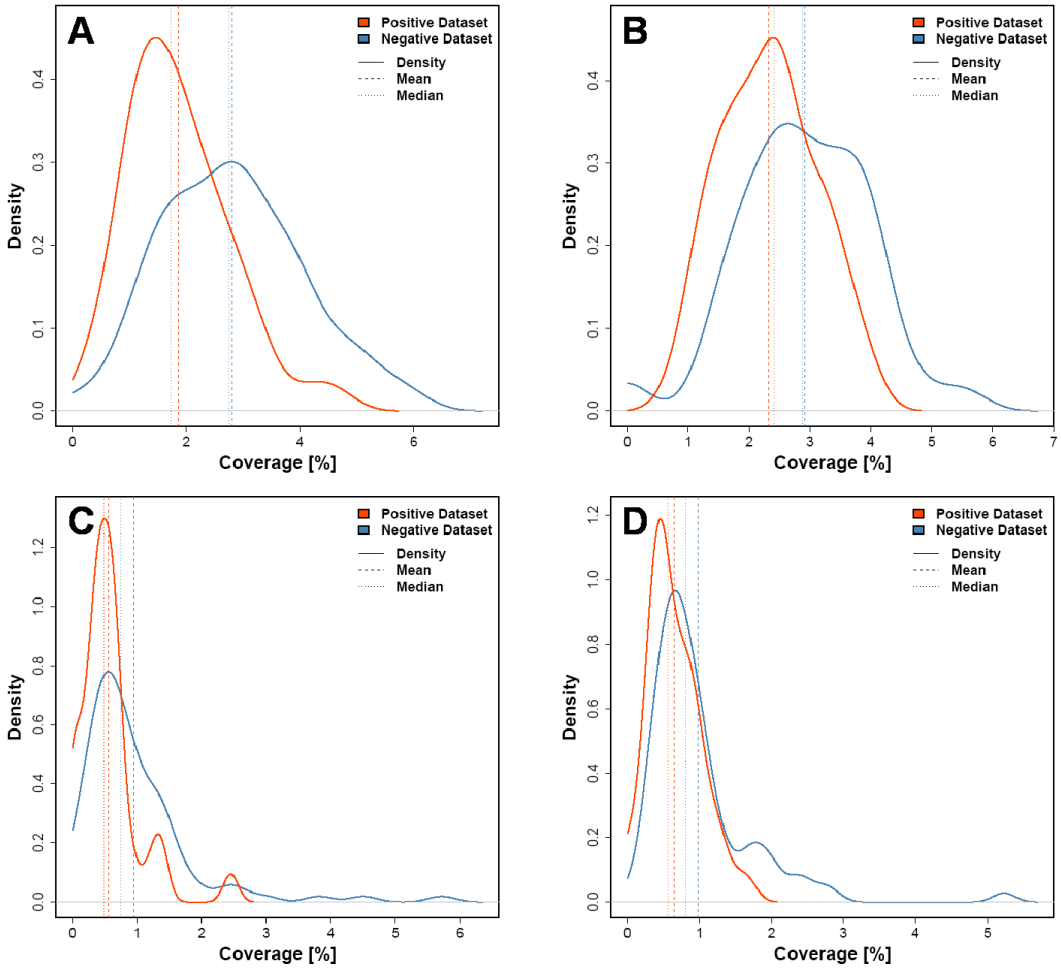


Figure C.3: Mean, median and density for the procent coverage of **Simple** repeats. **A:** *2K-2K* $p = 9.776x10^{-7}$ **B:** *2K-next* $p = 1.79x10^{-4}$ **C:** *H2K-2K* $p = 4.265x10^{-4}$ **D:** *H2K-next* $p = 6.874x10^{-3}$.

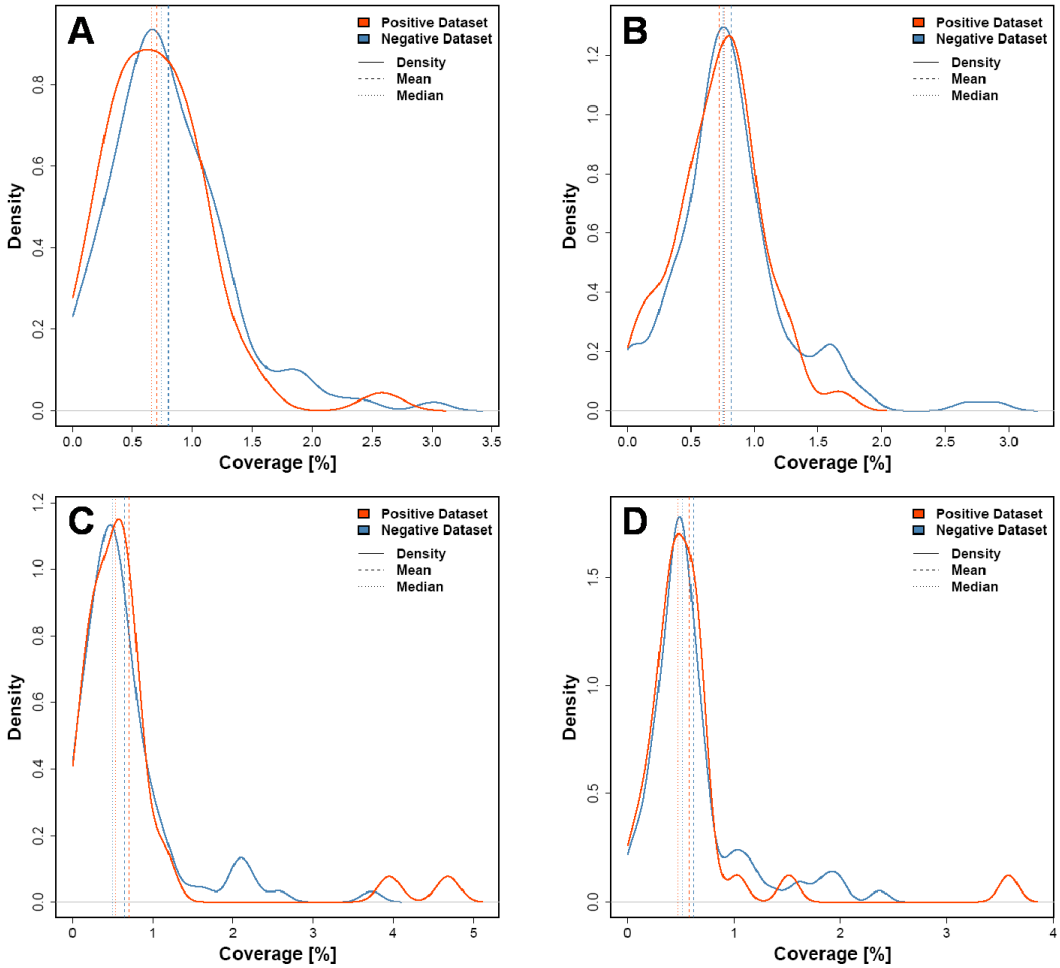


Figure C.4: Mean, median and density for the procentage coverage of **Low Complexity** repeats. **A:** *2K-2K* $p = 0.263$ **B:** *2K-next* $p = 0.424$ **C:** *H2K-2K* $p = 0.666$ **D:** *H2K-next* $p = 0.482$.

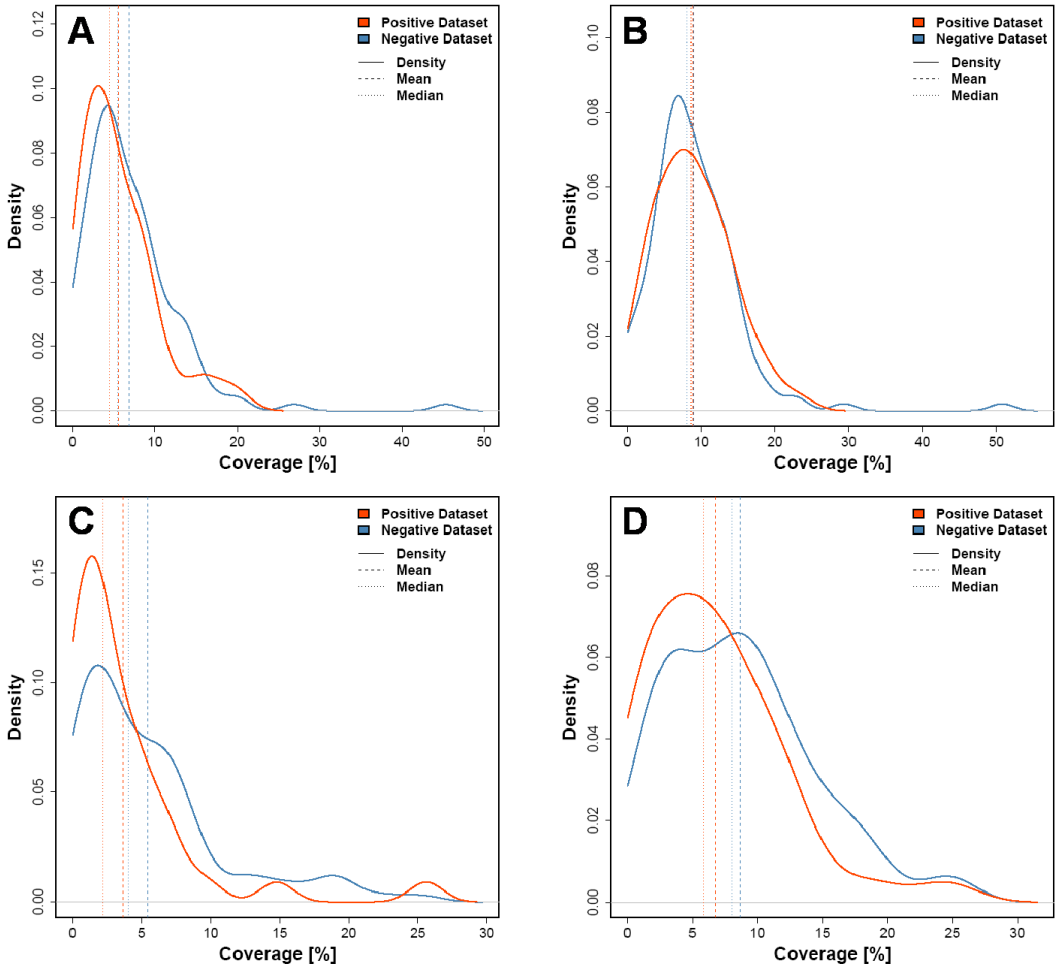


Figure C.5: Mean, median and density for the procentage coverage of **LTRs**. **A:** *2K-2K* $p = 0.0989$ **B:** *2K-next* $p = 0.885$ **C:** *H2K-2K* $p = 0.0105$ **D:** *H2K-next* $p = 0.0593$.

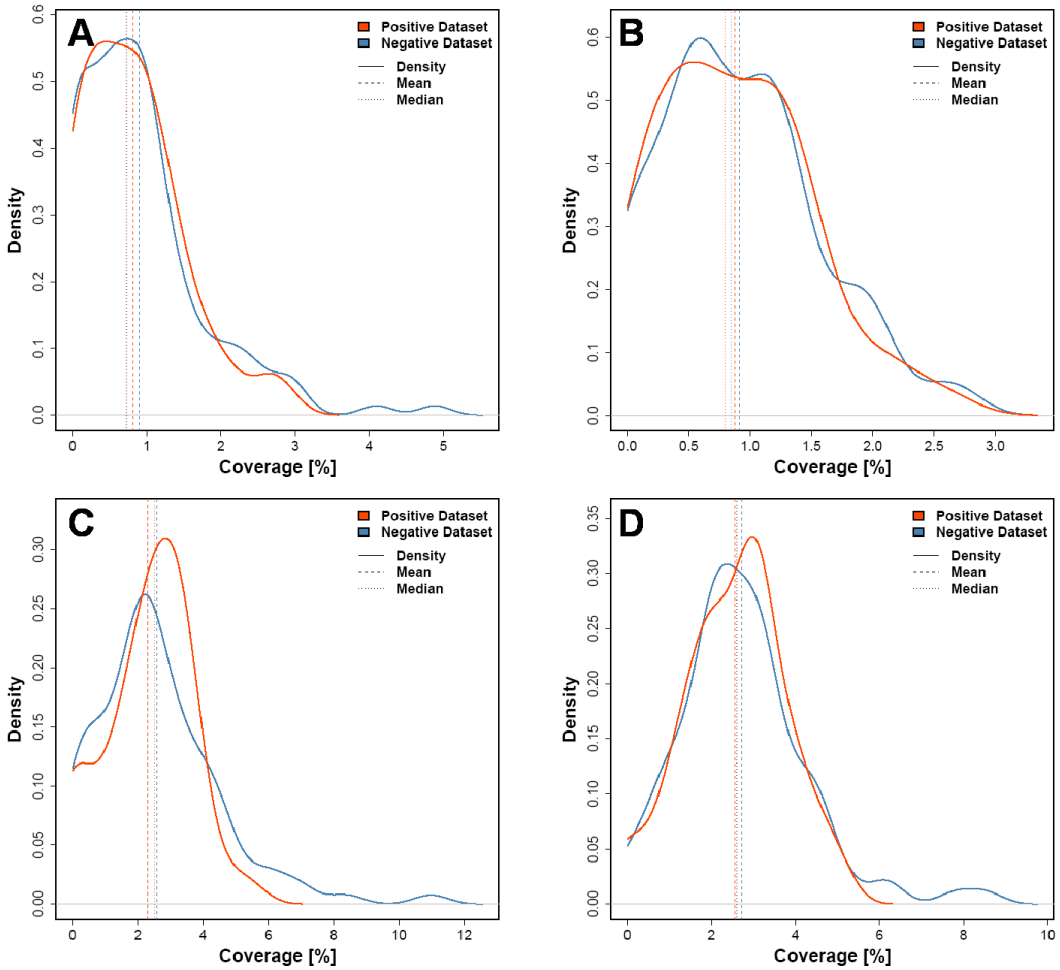


Figure C.6: Mean, median and density for the procentage coverage of **DNA** repeats. **A:** *2K-2K* $p = 0.828$ **B:** *2K-next* $p = 0.799$ **C:** *H2K-2K* $p = 0.604$ **D:** *H2K-next* $p = 0.777$.

C.2 FeaturePlotter Plots

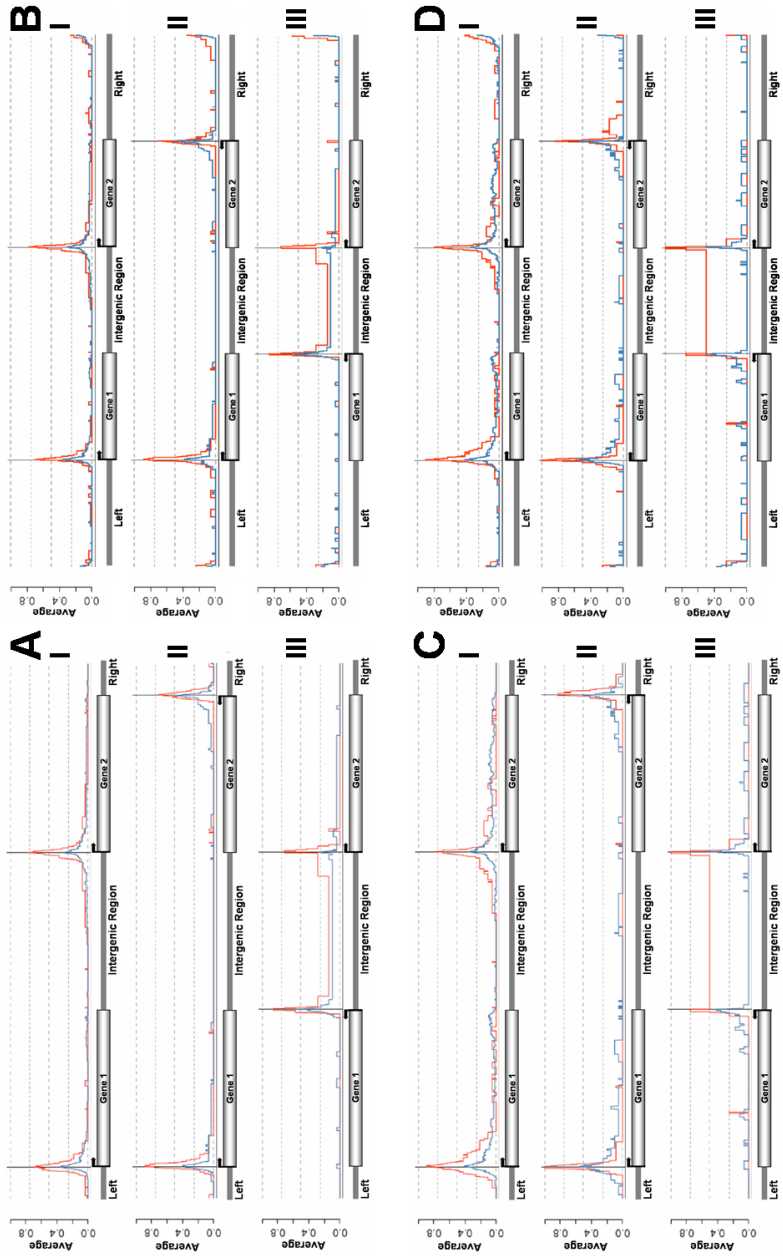


Figure C.7: Average distribution of **CpG islands** over the positive (red) and negative (blue) sequence. **A:** *2K-2K* **B:** *2K-next* **C:** *H2K-2K* **D:** *H2K-next* **I:** Unidirectional Pairs **II:** Convergent Pairs **III:** Divergent Pairs

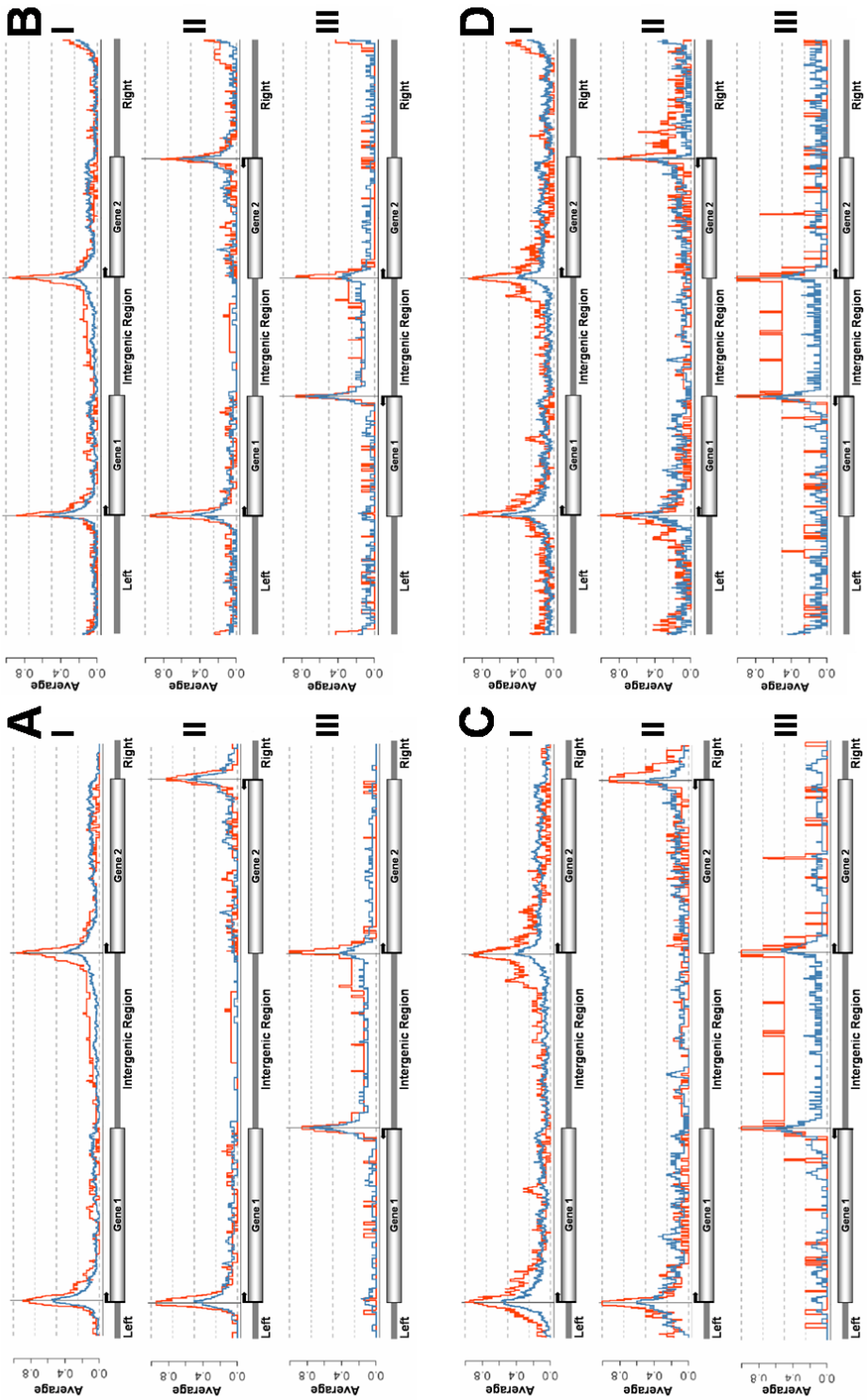


Figure C.8: Average distribution of **CpG regions** over the positive (red) and negative (blue) sequence. **A:** *2K-2K* **B:** *2K-next* **C:** *H2K-2K* **D:** *H2K-next* **I:** Unidirectional Pairs **II:** Convergent Pairs **III:** Divergent Pairs

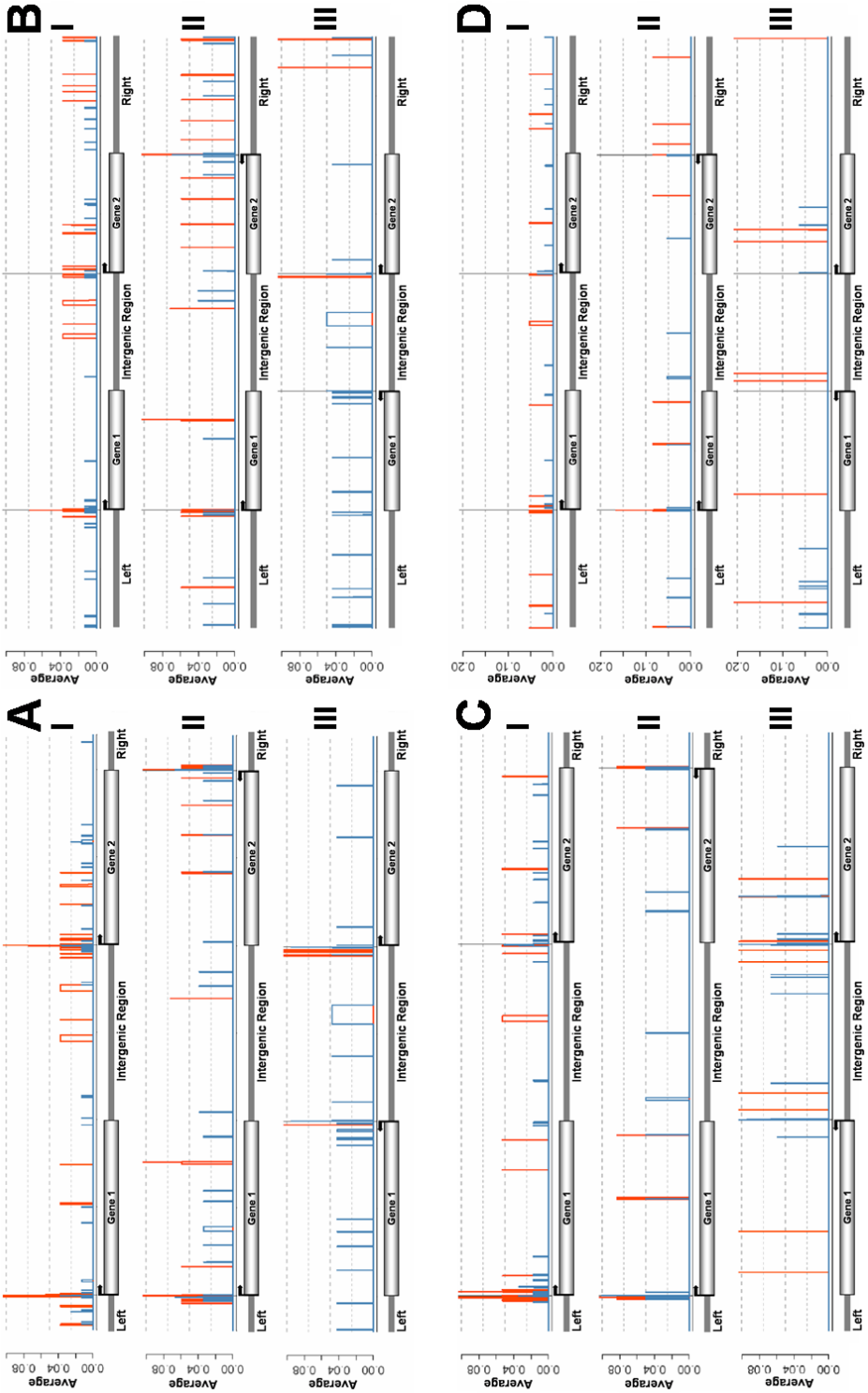


Figure C.9: Average distribution of **SP1 binding sites** over the positive (red) and negative (blue) sequence. **A:** *2K-2K* **B:** *2K-next* **C:** *H2K-2K* **D:** *H2K-next* **I:** Unidirectional Pairs **II:** Convergent Pairs **III:** Divergent Pairs

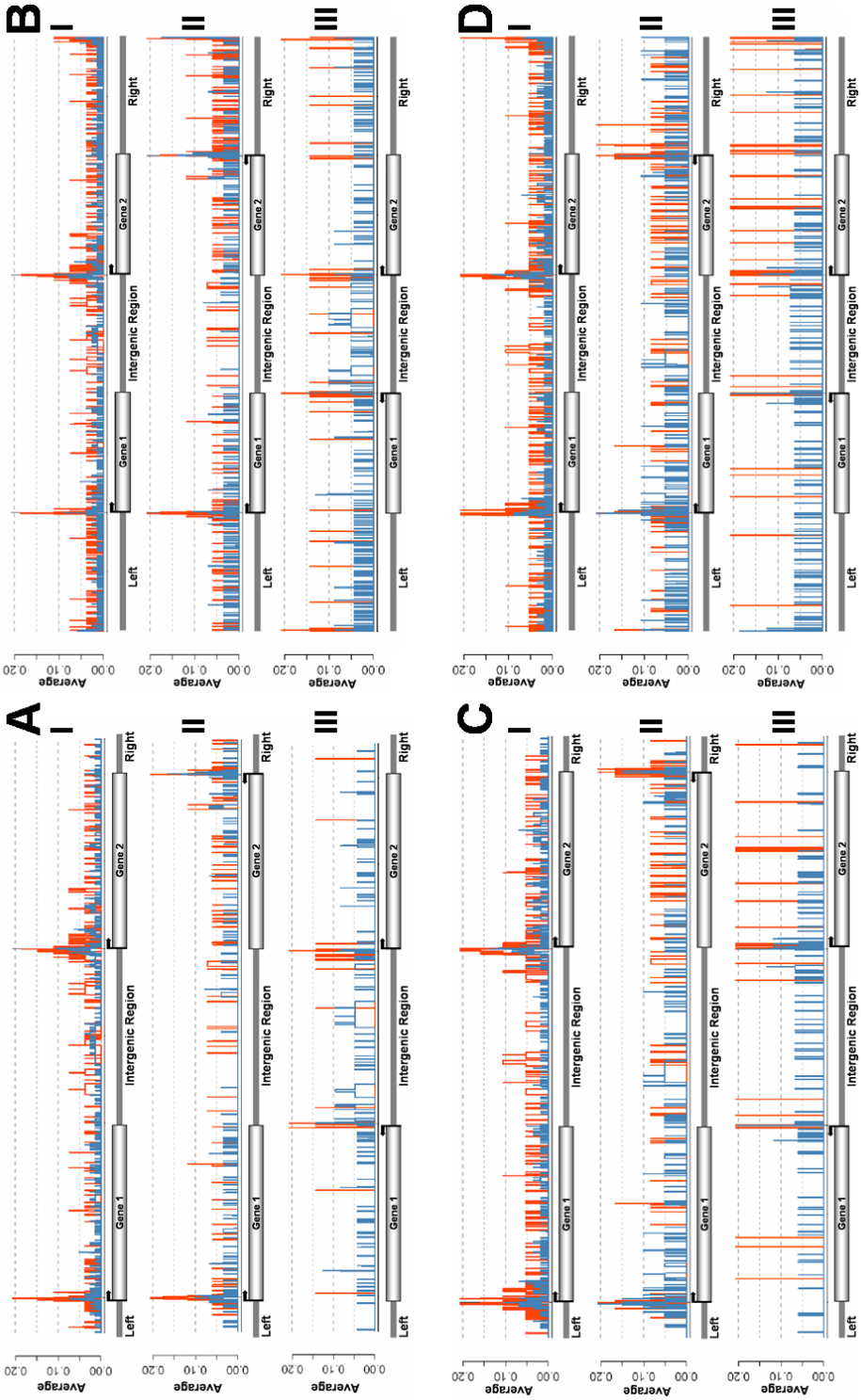


Figure C.10: Average distribution of **GC Box hexanucleotides** over the positive (red) and negative (blue) sequence. **A:** *2K-2K* **B:** *2K-next* **C:** *H2K-2K* **D:** *H2K-next* **I:** Unidirectional Pairs **II:** Convergent Pairs **III:** Divergent Pairs

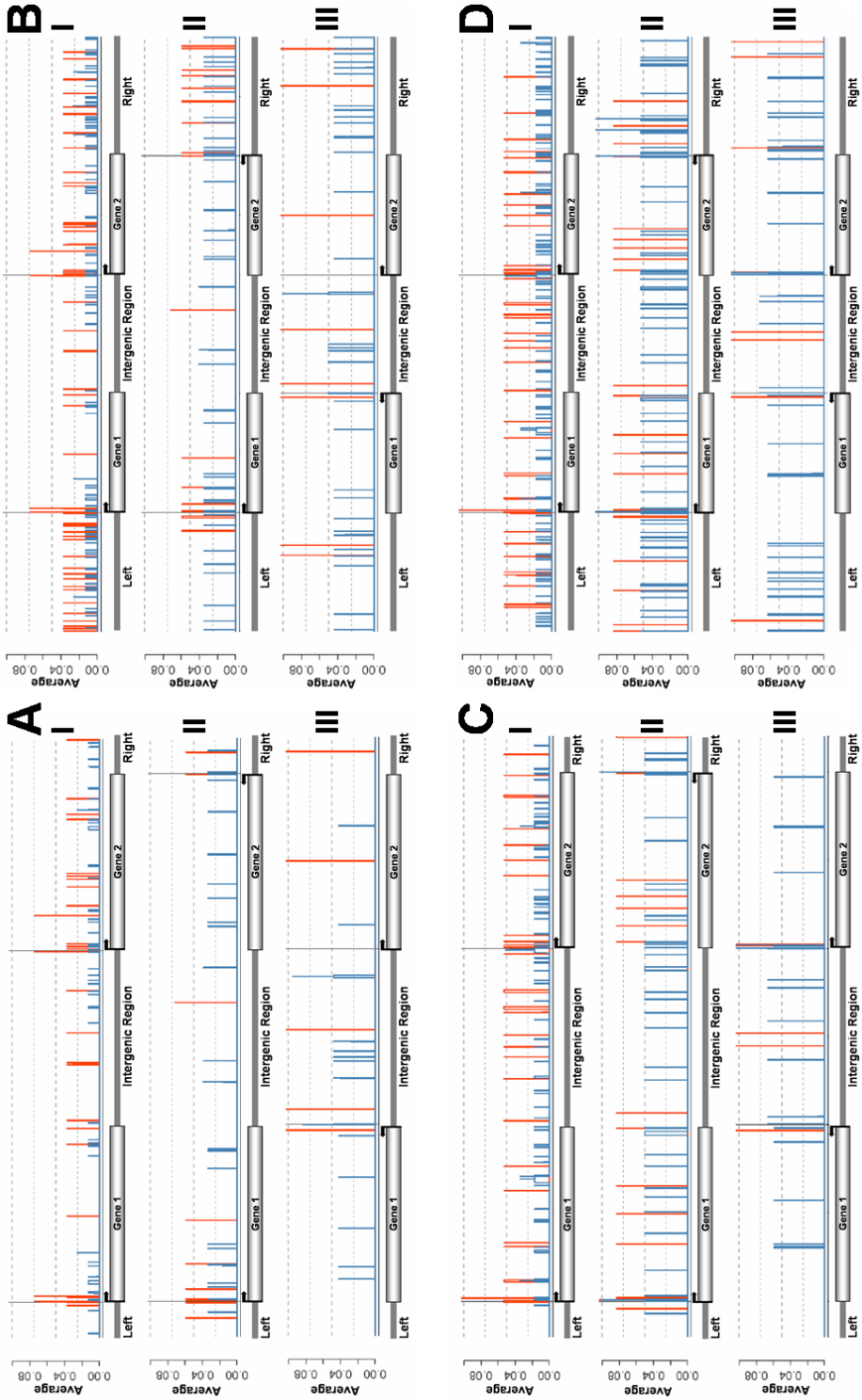


Figure C.11: Average distribution of **CTCF binding sites** over the positive (red) and negative (blue) sequence. **A:** *2K-2K* **B:** *2K-next* **C:** *H2K-2K* **D:** *H2K-next* **I:** Unidirectional Pairs **II:** Convergent Pairs **III:** Divergent Pairs

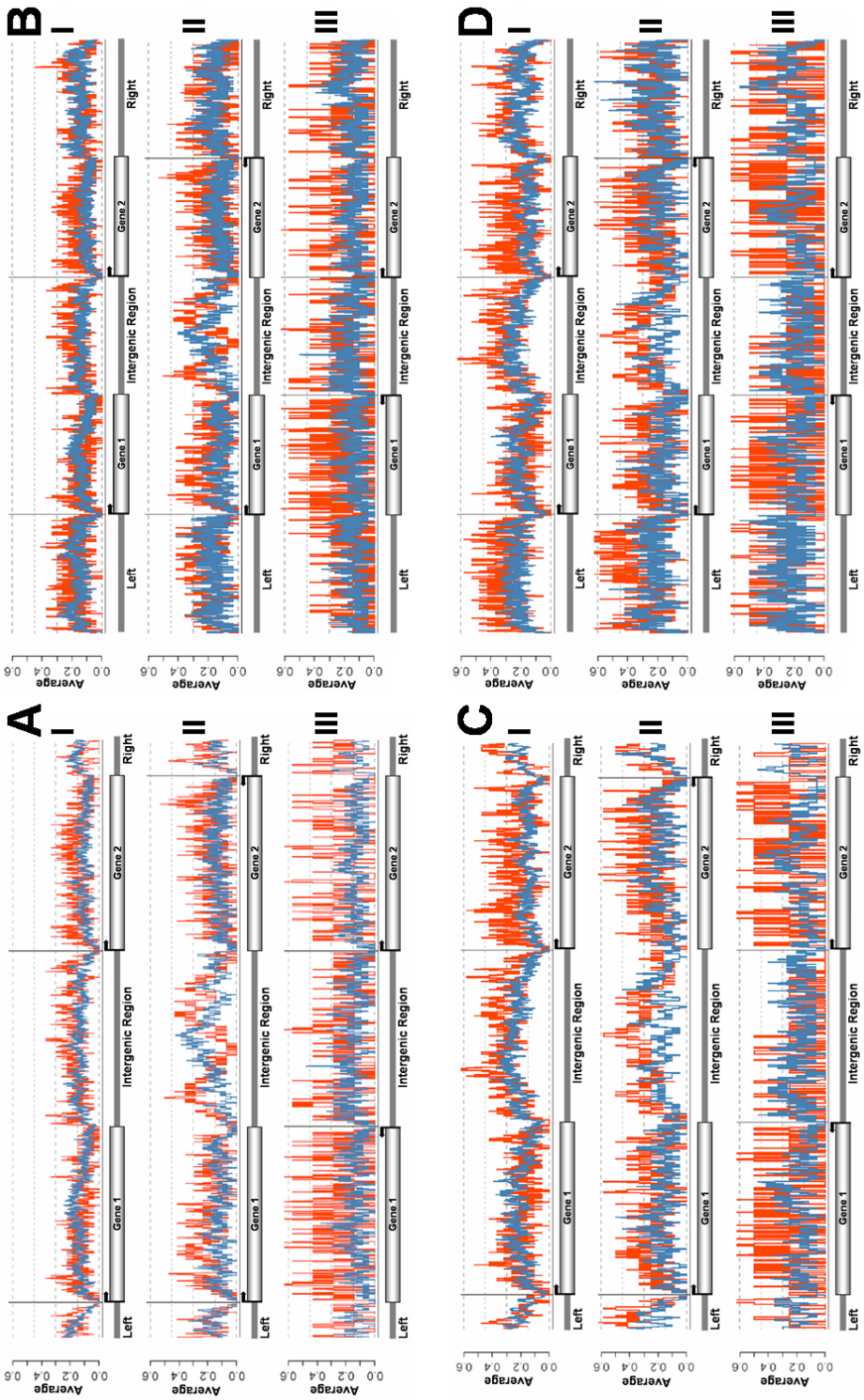


Figure C.12: Average distribution of **SINE repeats** over the positive (red) and negative (blue) sequence. **A:** *2K-2K* **B:** *2K-next* **C:** *H2K-2K* **D:** *H2K-next* **I:** Unidirectional Pairs **II:** Convergent Pairs **III:** Divergent Pairs

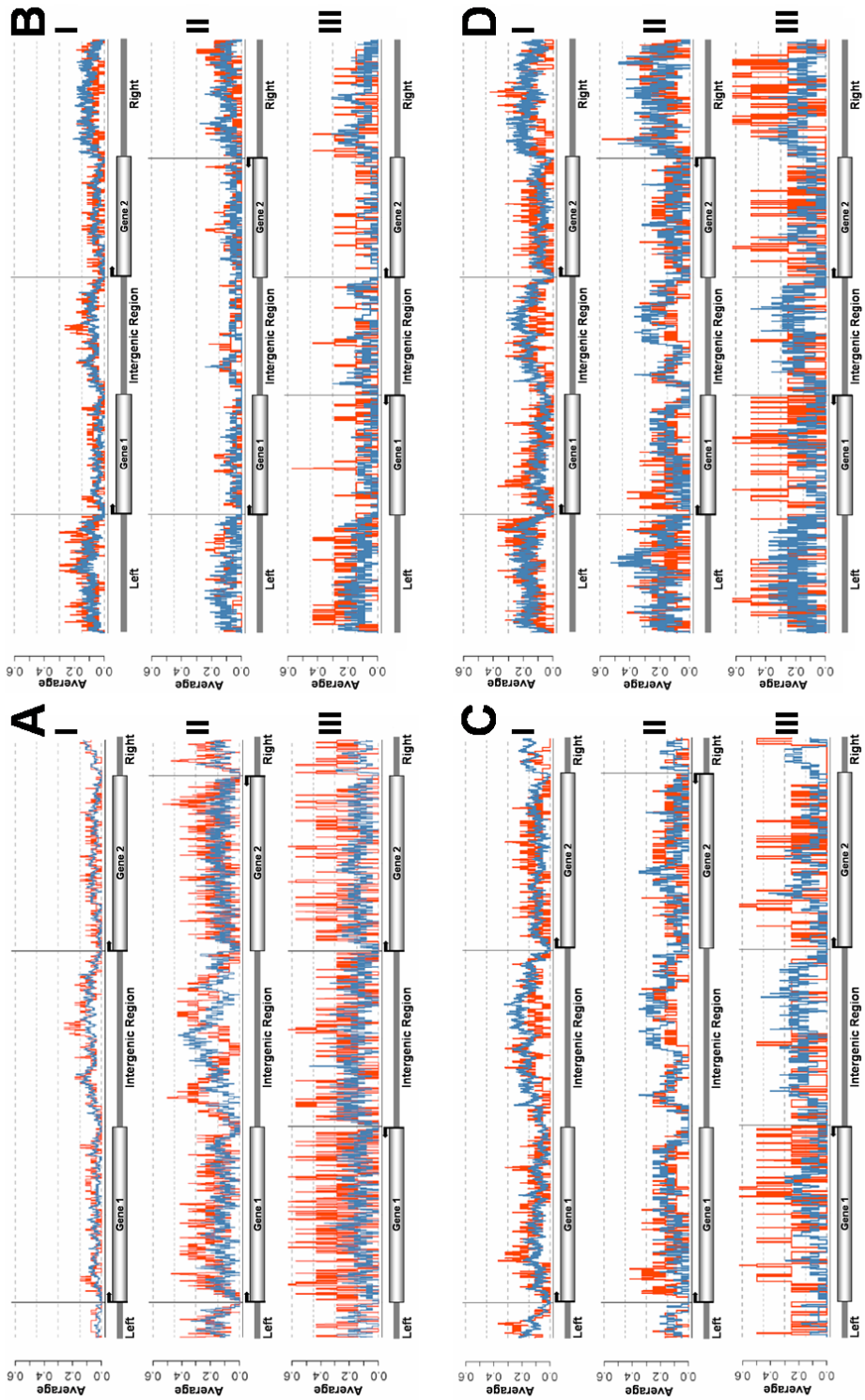


Figure C.13: Average distribution of **LINE repeats** over the positive (red) and negative (blue) sequence. **A:** *2K-2K* **B:** *2K-next* **C:** *H2K-2K* **D:** *H2K-next* **I:** Unidirectional Pairs **II:** Convergent Pairs **III:** Divergent Pairs

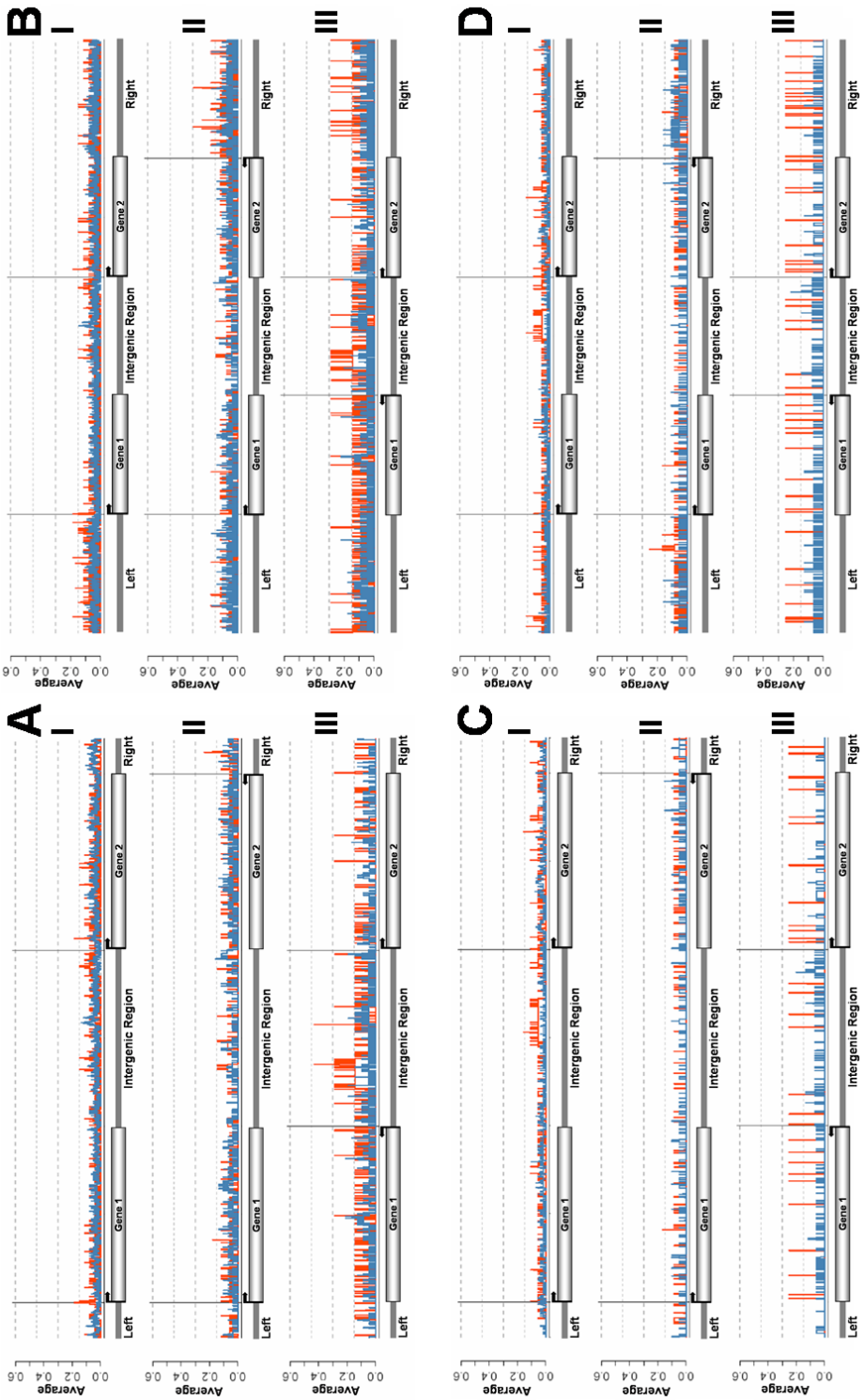


Figure C.14: Average distribution of **simple repeats** over the positive (red) and negative (blue) sequence. **A:** *2K-2K* **B:** *2K-next* **C:** *H2K-2K* **D:** *H2K-next* **I:** Unidirectional Pairs **II:** Convergent Pairs **III:** Divergent Pairs

List of Abbreviations

ACH	active chromatin hub
EM	expectation maximisation
GSS	group specificity score
HAT	histone acetyltransferase
HDAC	histone deacetylase
HCP	highly co-expressed gene pair
HMM	hidden Markov model
HS	hypersensitive site
LCR	locus control region
LINE	long interspersed nuclear elements
LTR	long terminar repeat
NR	nuclear receptor
NRF	nucleosome-free regions
PFM	position frequency matrix
PSFM	position-specific frequency matrix
PSSM	position-specific scoring matrix
PWM	position weight matrix
RNAi	RNA interference
RP	regulatory potential
SINE	short interspersed nuclear elements
TF	transcription factor
TFBS	transcription factor binding site
TSS	transcriptional start site
UCP	uncorrelated gene pair

Bibliography

[1] van Driel, R., Fransz, P. F., and Verschure, P. J. The eukaryotic genome: a system regulated at different hierarchical levels. *J Cell Sci* **116**(Pt 20), 4067–75 Oct (2003).

[2] Hawkins, R. D. and Ren, B. Genome-wide location analysis: insights on transcriptional regulation. *Hum Mol Genet* **15 Spec No 1**, R1–7 Apr (2006).

[3] Eberharter, A. and Becker, P. B. Histone acetylation: a switch between repressive and permissive chromatin. Second in review series on chromatin dynamics. *EMBO Rep* **3**(3), 224–9 Mar (2002).

[4] Cho, R., Campbell, M., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart, D., and Davis, R. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **2**(1), 65–73 Jul (1998).

[5] Boutanaev, A. M., Kalmykova, A. I., Shevelyov, Y. Y., and Nurminsky, D. I. Large clusters of co-expressed genes in the Drosophila genome. *Nature* **420**(6916), 666–9 Dec (2002).

[6] de Laat, W. and Grosveld, F. Spatial organization of gene expression: the active chromatin hub. *Chromosome Res* **11**(5), 447–59 (2003).

[7] Purmann, A., Toedling, J., Schueler, M., Carninci, P., Lehrach, H., Hayashizaki, Y., Huber, W., and Sperling, S. Genomic organization of transcriptomes in mammals: Co-regulation and co-functionality.

[8] Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M., van Asperen, R., Boon, K., Voûte, P., Heisterkamp, S., van Kampen,

A., and Versteeg, R. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291**(5507), 1289–92 Feb (2001).

[9] Versteeg, R., van Schaik, B. D. C., van Batenburg, M. F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H. J., and van Kampen, A. H. C. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res* **13**(9), 1998–2004 Sep (2003).

[10] Fukuoka, Y., Inaoka, H., and Kohane, I. S. Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. *BMC Genomics* **5**(1), 4 Jan (2004).

[11] Lee, J. M. and Sonnhammer, E. L. L. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res* **13**(5), 875–82 May (2003).

[12] Yanai, I., Korbel, J. O., Boue, S., McWeeney, S. K., Bork, P., and Lercher, M. J. Similar gene expression profiles do not imply similar tissue functions. *Trends Genet* **22**(3), 132–8 Mar (2006).

[13] Cohen, B., Mitra, R., Hughes, J., and Church, G. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet* **26**(2), 183–6 Oct (2000).

[14] Mannila, H., Patrikainen, A., Seppänen, J. K., and Kere, J. Long-range control of expression in yeast. *Bioinformatics* **18**(3), 482–3 Mar (2002).

[15] Lercher, M. J. and Hurst, L. D. Co-expressed yeast genes cluster over a long range but are not regularly spaced. *J Mol Biol* **359**(3), 825–31 Jun (2006).

[16] Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V., Brenner, S., Batalov, S., Forrest, A. R. R., Zavolan, M., Davis, M., Wilming, L., Aidinis, V., Allen, J., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R., Bailey, T., Bansal, M., Baxter, L., Beisel, K., Bersano, T., Bono, H., Chalk, A., Chiu, K., Choudhary, V., Christoffels, A., Clutterbuck, D., Crowe, M., Dalla, E., Dalrymple, B., de Bono, B., Gatta, G. D., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T., Gojobori, T., Green, R., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P. T., Kruger, A., Kummerfeld, S., Kurochkin, I., Lareau, L., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Babu, M. M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F.,

Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K., Pavan, W., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J., Ring, B., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S., Sandelin, A., Schneider, C., Schönbach, C., Sekiguchi, K., Semple, C. A. M., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S., Tang, S., Taylor, M., Tegner, J., Teichmann, S., Ueda, H., van Nimwegen, E., Verardo, R., Wei, C., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S., Teasdale, R., Liu, E., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J., Hume, D., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., Hayashizaki, Y., Consortium, F., Group, R. G. E. R., and Group), G. S. G. G. N. P. C. The transcriptional landscape of the mammalian genome. *Science* **309**(5740), 1559–63 Sep (2005).

[17] Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., Orth, A. P., Vega, R. G., Sapinoso, L. M., Moqrich, A., Patapoutian, A., Hampton, G. M., Schultz, P. G., and Hogenesch, J. B. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* **99**(7), 4465–70 Apr (2002).

[18] Kornberg, R. Chromatin structure: a repeating unit of histones and DNA. *Science* **184**(139), 868–71 May (1974).

[19] Fischer, J. and Sperling, S. The histone code - looking beyond yeast. *In Preparation* .

[20] Shogren-Knaak, M. `http://www.bb.iastate.edu/FacultyResearchFolder/ShogrenKnaak/Research.h%tml`, September (2006).

[21] Kishimoto, M., Fujiki, R., Takezawa, S., Sasaki, Y., Nakamura, T., Yamaoka, K., Kitagawa, H., and Kato, S. Nuclear receptor mediated gene regulation through chromatin remodeling and histone modifications. *Endocr J* **53**(2), 157–72 Apr (2006).

[22] Barrera, L. O. and Ren, B. The transcriptional regulatory code of eukaryotic cells–insights from genome-wide analysis of chromatin organization and transcription factor binding. *Curr Opin Cell Biol* **18**(3), 291–8 Jun (2006).

[23] Bernstein, B. E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D. K.,

Huebert, D. J., McMahon, S., Karlsson, E. K., Kulbokas, E. J., Gingeras, T. R., Schreiber, S. L., and Lander, E. S. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**(2), 169–81 Jan (2005).

[24] Zhang, H., Roberts, D. N., and Cairns, B. R. Genome-wide dynamics of Htz1, a histone H2A variant that poises repressed/basal promoters for activation through histone loss. *Cell* **123**(2), 219–31 Oct (2005).

[25] Guillemette, B., Bataille, A. R., Gévry, N., Adam, M., Blanchette, M., Robert, F., and Gaudreau, L. Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. *PLoS Biol* **3**(12), e384 Dec (2005).

[26] Raisner, R. M., Hartley, P. D., Meneghini, M. D., Bao, M. Z., Liu, C. L., Schreiber, S. L., Rando, O. J., and Madhani, H. D. Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin. *Cell* **123**(2), 233–48 Oct (2005).

[27] Meneghini, M. D., Wu, M., and Madhani, H. D. Conserved histone variant H2A.Z protects euchromatin from the ectopic spread of silent heterochromatin. *Cell* **112**(5), 725–36 Mar (2003).

[28] Grunstein, M. Histone acetylation in chromatin structure and transcription. *Nature* **389**(6649), 349–52 Sep (1997).

[29] Zhang, Y. and Reinberg, D. Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes Dev* **15**(18), 2343–60 Sep (2001).

[30] Nowak, S. J. and Corces, V. G. Phosphorylation of histone H3: a balancing act between chromosome condensation and transcriptional activation. *Trends Genet* **20**(4), 214–20 Apr (2004).

[31] Davie, J. and Murphy, L. Level of ubiquitinated histone H2B in chromatin is coupled to ongoing transcription. *Biochemistry* **29**(20), 4752–7 May (1990).

[32] Nathan, D., Sterner, D. E., and Berger, S. L. Histone modifications: Now summoning sumoylation. *Proc Natl Acad Sci U S A* **100**(23), 13118–20 Nov (2003).

[33] Adamietz, P. and Rudolph, A. ADP-ribosylation of nuclear proteins in vivo. Identification of histone H2B as a major acceptor for mono- and poly(ADP-ribose) in dimethyl sulfate-treated hepatoma AH 7974 cells. *J Biol Chem* **259**(11), 6841–6 Junier (1984).

[34] Liebich, H., Gesele, E., Wirth, C., Wöll, J., Jobst, K., and Lakatos, A. Non-enzymatic glycation of histones. *Biol Mass Spectrom* **22**(2), 121–3 Feb (1993).

[35] Hymes, J., Fleischhauer, K., and Wolf, B. Biotinylation of histones by human serum biotinidase: assessment of biotinyl-transferase activity in sera from normal individuals and children with biotinidase deficiency. *Biochem Mol Med* **56**(1), 76–83 Oct (1995).

[36] Wondrak, G., Cervantes-Laurean, D., Jacobson, E., and Jacobson, M. Histone carbonylation in vivo and in vitro. *Biochem J* **351 Pt 3**, 769–77 Nov (2000).

[37] Strahl, B. and Allis, C. The language of covalent histone modifications. *Nature* **403**(6765), 41–5 Jan (2000).

[38] Shogren-Knaak, M., Ishii, H., Sun, J.-M., Pazin, M. J., Davie, J. R., and Peterson, C. L. Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science* **311**(5762), 844–7 Feb (2006).

[39] Allfrey, V., Faulkner, R., and Mirsky, A. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proc Natl Acad Sci U S A* **51**, 786–94 May (1964).

[40] Dillon, N. and Festenstein, R. Unravelling heterochromatin: competition between positive and negative factors regulates accessibility. *Trends Genet* **18**(5), 252–8 May (2002).

[41] Andrulis, E., Neiman, A., Zappulla, D., and Sternglanz, R. Perinuclear localization of chromatin facilitates transcriptional silencing. *Nature* **394**(6693), 592–5 Aug (1998).

[42] Cremer, M., von Hase, J., Volm, T., Brero, A., Kreth, G., Walter, J., Fischer, C., Solovei, I., Cremer, C., and Cremer, T. Non-random radial higher-order chromatin arrangements in nuclei of diploid human cells. *Chromosome Res* **9**(7), 541–67 (2001).

[43] Cai, S., Han, H.-J., and Kohwi-Shigematsu, T. Tissue-specific nuclear architecture and gene expression regulated by SATB1. *Nat Genet* **34**(1), 42–51 May (2003).

[44] Lunyak, V. V., Burgess, R., Prefontaine, G. G., Nelson, C., Sze, S.-H., Chenoweth, J., Schwartz, P., Pevzner, P. A., Glass, C., Mandel, G., and Rosenfeld, M. G. Corepressor-dependent silencing of chromosomal regions encoding neuronal genes. *Science* **298**(5599), 1747–52 Nov (2002).

[45] Jenuwein, T. and Allis, C. Translating the histone code. *Science* **293**(5532), 1074–80 Aug (2001).

[46] Turner, B. M. Cellular memory and the histone code. *Cell* **111**(3), 285–91 Nov (2002).

[47] Forrester, W., Thompson, C., Elder, J., and Groudine, M. A developmentally stable chromatin structure in the human beta-globin gene cluster. *Proc Natl Acad Sci U S A* **83**(5), 1359–63 Mar (1986).

[48] Ho, Y., Liebhaber, S. A., and Cooke, N. E. Activation of the human GH gene cluster: roles for targeted chromatin modification. *Trends Endocrinol Metab* **15**(1), 40–5 (2004).

[49] Chambeyron, S. and Bickmore, W. A. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes Dev* **18**(10), 1119–30 May (2004).

[50] Shashikant, C., Utset, M., Violette, S., Wise, T., Einat, P., Einat, M., Pendleton, J., Schughart, K., and Ruddle, F. Homeobox genes in mouse development. *Crit Rev Eukaryot Gene Expr* **1**(3), 207–45 (1991).

[51] Roelen, B. A. J., de Graaff, W., Forlani, S., and Deschamps, J. Hox cluster polarity in early transcriptional availability: a high order regulatory level of clustered Hox genes in the mouse. *Mech Dev* **119**(1), 81–90 Nov (2002).

[52] Li, Q., Harju, S., and Peterson, K. Locus control regions: coming of age at a decade plus. *Trends Genet* **15**(10), 403–8 Oct (1999).

[53] Zhang, X., Odom, D. T., Koo, S.-H., Conkright, M. D., Canettieri, G., Best, J., Chen, H., Jenner, R., Herbolsheimer, E., Jacobsen, E., Kadam, S., Ecker, J. R., Emerson, B., Hogenesch, J. B., Unterman, T., Young, R. A., and Montminy, M. Genome-wide analysis of cAMP-response element binding protein occupancy, phosphorylation, and target gene activation in human tissues. *Proc Natl Acad Sci U S A* **102**(12), 4459–64 Mar (2005).

[54] Dorer, D. and Henikoff, S. Transgene repeat arrays interact with distant heterochromatin and cause silencing in cis and trans. *Genetics* **147**(3), 1181–90 Nov (1997).

[55] Hodgetts, R. Eukaryotic gene regulation by targeted chromatin re-modeling at dispersed, middle-repetitive sequence elements. *Curr Opin Genet Dev* **14**(6), 680–5 Dec (2004).

[56] Volpe, T. A., Kidner, C., Hall, I. M., Teng, G., Grewal, S. I. S., and Martienssen, R. A. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297**(5588), 1833–7 Sep (2002).

[57] Zilberman, D., Cao, X., and Jacobsen, S. E. ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science* **299**(5607), 716–9 Jan (2003).

[58] Pal-Bhadra, M., Leibovitch, B. A., Gandhi, S. G., Rao, M., Bhadra, U., Birchler, J. A., and Elgin, S. C. R. Heterochromatic silencing and HP1 localization in Drosophila are dependent on the RNAi machinery. *Science* **303**(5658), 669–72 Jan (2004).

[59] Fukagawa, T., Nogami, M., Yoshikawa, M., Ikeno, M., Okazaki, T., Takami, Y., Nakayama, T., and Oshimura, M. Dicer is essential for formation of the heterochromatin structure in vertebrate cells. *Nat Cell Biol* **6**(8), 784–91 Aug (2004).

[60] Wang, F., Koyama, N., Nishida, H., Haraguchi, T., Reith, W., and Tsukamoto, T. The assembly and maintenance of heterochromatin initiated by transgene repeats are independent of the RNA interference pathway in mammalian cells. *Mol Cell Biol* **26**(11), 4028–40 Jun (2006).

[61] Polak, P. and Domany, E. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics* **7**, 133 (2006).

[62] Valenzuela, L. and Kamakaka, R. T. Chromatin Insulators. *Annu Rev Genet* Sep (2006).

[63] Bell, A., West, A., and Felsenfeld, G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98**(3), 387–96 Aug (1999).

[64] Splinter, E., Heath, H., Kooren, J., Palstra, R.-J., Klous, P., Grosveld, F., Galjart, N., and de Laat, W. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev* **20**(17), 2349–54 Sep (2006).

[65] Ellis, J., Talbot, D., Dillon, N., and Grosveld, F. Synthetic human beta-globin 5'HS2 constructs function as locus control regions only in multicopy transgene concatamers. *EMBO J* **12**(1), 127–34 Jan (1993).

[66] Sabbattini, P., Georgiou, A., Sinclair, C., and Dillon, N. Analysis of mice with single and multiple copies of transgenes reveals a novel arrangement for the lambda5-VpreB1 locus control region. *Mol Cell Biol* **19**(1), 671–9 Jan (1999).

[67] Li, Q., Barkess, G., and Qian, H. Chromatin looping and the probability of transcription. *Trends Genet* **22**(4), 197–202 Apr (2006).

[68] Tuan, D., Kong, S., and Hu, K. Transcription of the hypersensitive site HS2 enhancer in erythroid cells. *Proc Natl Acad Sci U S A* **89**(23), 11219–23 Dec (1992).

[69] Ptashne, M. and Gann, A. Transcriptional activation by recruitment. *Nature* **386**(6625), 569–77 Apr (1997).

[70] Bulger, M. and Groudine, M. Looping versus linking: toward a model for long-distance gene activation. *Genes Dev* **13**(19), 2465–77 Oct (1999).

[71] Blackwood, E. and Kadonaga, J. Going the distance: a current view of enhancer action. *Science* **281**(5373), 60–3 Jul (1998).

[72] Tolhuis, B., Palstra, R. J., Splinter, E., Grosveld, F., and de Laat, W. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell* **10**(6), 1453–65 Dec (2002).

[73] Duboule, D. and Deschamps, J. Colinearity loops out. *Dev Cell* **6**(6), 738–40 Jun (2004).

[74] Bulyk, M. L. Computational prediction of transcription-factor binding site locations. *Genome Biology* **5**(201) December (2003).

[75] Sch?nbach, C. From masking reoeats to identifying functional repeats in the mouse transcriptome. *Briefings in Bioinformatics* **5**(2), 107 – 117 June (2004).

[76] Huber, B. R. and Bulyk, M. L. Meta-analysis discovery of tissue-specific dna sequences motifs from mammalian gene expression data. *BMC Bioinformatics* **7**, 229 (2006).

[77] Jegga, A. G., Gupta, A., Gowrisankar, S., Deshmukh, M. A., Conolly, S., Finly, K., and Aronow, B. J. Cismols analyzer: identification of compositionally similar cis-element clusters in ortholog conserved regions of coordinately expressed genes. *Nucleic Acids Research* **33**, W408 – W411 (2005).

[78] Wassermann, W. W. and Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics* **5**, 276 – 287 April (2004).

[79] Wassermann, W. W. and Fickett, J. W. Identification of regulatory regions which confer muscle-specific gene expression. *Journal of Molecular Biology* **278**, 167 – 181 (1998).

[80] Dermitzakis, E. T., Reymond, A., Lyle, R., Scamuffa, N., Ucla, C., Deutsch, S., Stevenson, B. J., Flegel, V., Bucher, P., Jongeneel, C. V., and Antonarakis, S. E. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420**(6915), 578–82 Dec (2002).

[81] Nobrega, M. A., Ovcharenko, I., Afzal, V., and Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302**(5644), 413 Oct (2003).

[82] Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**(8), 1034–50 Aug (2005).

[83] Kolbe, D., Taylor, J., Elnitski, L., Eswara, P., Li, J., Miller, W., Hardison, R., and Chiaromonte, F. Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res* **14**(4), 700–7 Apr (2004).

[84] King, D. C., Taylor, J., Elnitski, L., Chiaromonte, F., Miller, W., and Hardison, R. C. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res* **15**(8), 1051–60 Aug (2005).

[85] JOSSE, J., KAISER, A., and KORNBERG, A. Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *J Biol Chem* **236**, 864–75 Mar (1961).

[86] Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. CpG islands as gene markers in the human genome. *Genomics* **13**(4), 1095–107 Aug (1992).

[87] Gardiner-Garden, M. and Frommer, M. CpG islands in vertebrate genomes. *J Mol Biol* **196**(2), 261–82 Jul (1987).

[88] Gaszner, M. and Felsenfeld, G. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* **7**(9), 703–13 Sep (2006).

[89] Zhao, H. and Dean, A. An insulator blocks spreading of histone acetylation and interferes with RNA polymerase II transfer between an enhancer and gene. *Nucleic Acids Res* **32**(16), 4903–19 (2004).

[90] Ishihara, K. and Sasaki, H. An evolutionarily conserved putative insulator element near the 3' boundary of the imprinted Igf2/H19 domain. *Hum Mol Genet* **11**(14), 1627–36 Jul (2002).

[91] Hark, A., Schoenherr, C., Katz, D., Ingram, R., Levorse, J., and Tilghman, S. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* **405**(6785), 486–9 May (2000).

[92] Bailey, T. L. and Gribskov, M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**, 48 – 54 (1998).

[93] Tompa, M., Li, N., Bailey, T. L., Church, G. M., Moor, B. D., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. Assessing computational tools for the discovery of transcription factor binsing sites. *Nature Biotechnology* **23**, 137 – 144 January (2005).

[94] Bailey, T. L. and Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* , 28–36 August (1994).

[95] Liu, X., Brutlag, D., and Liu, J. Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing* **6**, 127–138 (2001).

[96] Hughes, J. D., Estep, P. W., Tavazoie, S., and Church, G. M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae. *J. Mol. Biol.* **296**, 1205–1214 (2000).

[97] Ao, W., Gaudet, J., Kent, W. J., Muttumu, S., and Mango, S. E. Environmentally induced foregut remodeling by ph-4/foxa and daf-12/nhr. *Science* **305**, 1743 – 1746 September (2004).

[98] McGuire, A. M. and Church, G. M. Predicting regulons and their cis-regulatory motifs by comparative genomics. *Nucleic Acids Research* **28**(22), 4523 – 4530 (2000).

[99] Pilpel, Y., Sudarsanam, P., and Munch, G. M. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics* **29**, 153 – 159 October (2001).

[100] Friberg, M., von Rohr, P., and Gonnet, G. Scoring functions for transcription factor binding site prediction. *BMC Bioinformatics* **6**(84) April (2005).

[101] McGuire, A. M., Hughes, J. D., and Church, G. M. Conservation of dna regulatory motifs and discovery of new motifs in microbial genomes. *Genome Research* **10**, 744 – 757 (2000).

[102] Schneider, T. D. and Stephens, R. M. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).

[103] Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. Weblogo: A sequence logo generator. *Genome Research* **14**, 1188 – 1190 (2004).

[104] Kaczynski, J., Cook, T., and Urrutia, R. Sp1- and Krüppel-like transcription factors. *Genome Biol* **4**(2), 206 (2003).

[105] Philipsen, S. and Suske, G. A tale of three fingers: the family of mammalian Sp/XKLF transcription factors. *Nucleic Acids Res* **27**(15), 2991–3000 Aug (1999).

[106] Sun, H.-J., Xu, X., Wang, X.-L., Wei, L., Li, F., Lu, J., and Huang, B.-Q. Transcription factors Ets2 and Sp1 act synergistically with histone acetyltransferase p300 in activating human interleukin-12 p40 promoter. *Acta Biochim Biophys Sin (Shanghai)* **38**(3), 194–200 Mar (2006).

[107] Beauchef, G., Kypriotou, M., Chadjichristos, C., Widom, R. L., Porée, B., Renard, E., Moslemi, S., Wegrowski, Y., Maquart, F.-X., Pujol, J.-P., and Galéra, P. c-Krox down-regulates the expression of UDP-glucose dehydrogenase in chondrocytes. *Biochem Biophys Res Commun* **333**(4), 1123–31 Aug (2005).

[108] Westman, B. J., Mackay, J. P., and Gell, D. Ikaros: a key regulator of haematopoiesis. *Int J Biochem Cell Biol* **34**(10), 1304–7 Oct (2002).

[109] Kim, M.-Y., Jeong, B. C., Lee, J. H., Kee, H. J., Kook, H., Kim, N. S., Kim, Y. H., Kim, J.-K., Ahn, K. Y., and Kim, K. K. A repressor complex, AP4 transcription factor and geminin, negatively regulates expression of target genes in nonneuronal cells. *Proc Natl Acad Sci U S A* **103**(35), 13074–9 Aug (2006).

[110] Yoon, H.-G., Chan, D. W., Reynolds, A. B., Qin, J., and Wong, J. N-CoR mediates DNA methylation-dependent repression through a methyl CpG binding protein Kaiso. *Mol Cell* **12**(3), 723–34 Sep (2003).

[111] McKinsey, T., Zhang, C., and Olson, E. Control of muscle development by dueling HATs and HDACs. *Curr Opin Genet Dev* **11**(5), 497–504 Oct (2001).

[112] Taubert, S., Gorrini, C., Frank, S. R., Parisi, T., Fuchs, M., Chan, H.-M., Livingston, D. M., and Amati, B. E2F-dependent histone acetylation and recruitment of the Tip60 acetyltransferase complex to chromatin in late G1. *Mol Cell Biol* **24**(10), 4546–56 May (2004).

[113] Johnnidis, J. B., Venanzi, E. S., Taxman, D. J., Ting, J. P.-Y., Benoist, C. O., and

Mathis, D. J. Chromosomal clustering of genes controlled by the aire transcription factor. *Proc Natl Acad Sci U S A* **102**(20), 7233–8 May (2005).

[114] Anderson, M. S., Venanzi, E. S., Chen, Z., Berzins, S. P., Benoist, C., and Mathis, D. The cellular mechanism of Aire control of T cell tolerance. *Immunity* **23**(2), 227–39 Aug (2005).

[115] Consortium, M. G. S., Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigó, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W. J., Kirby, A., Kolbe, D. L., Korf, I., Kucherlapati, R. S., Kulbokas, E. J., Kulp, D., Landers, T., Leger, J., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D. R., Mardis, E. R., Matthews, L., Mauceli, E., Mayer, J. H., McCarthy, M., McCombie, W. R., McLaren, S., McLay, K., McPherson, J. D., Meldrim, J., Meredith, B., Mesirov, J. P., Miller, W., Miner, T. L., Mongin, E., Montgomery, K. T., Morgan, M., Mott, R., Mullikin, J. C., Muzny, D. M., Nash, W. E., Nelson, J. O., Nhan, M. N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M. J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K. H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C. S., Poliakov, A., Ponce, T. C., Ponting, C. P., Potter, S., Quail, M., Reymond, A., Roe, B. A., Roskin, K. M., Rubin, E. M., Rust, A. G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M. S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J. B., Slater, G., Smit, A., Smith, D. R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J. P., Niederhausern, A. C. V., Wade, C. M., Wall, M., Weber, R. J., Weiss, R. B., Wendl, M. C., West, A. P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R. K., Winter, E., Worley, K. C., Wyman, D., Yang, S., Yang, S.-P., Zdobnov,

E. M., Zody, M. C., and Lander, E. S. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**(6915), 520–62 Dec (2002).

[116] Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R., Wilson, R., Hillier, L., McPherson, J., Marra, M., Mardis, E., Fulton, L., Chinwalla, A., Pepin, K., Gish, W., Chissoe, S., Wendl, M., Delehaunty, K., Miner, T., Delehaunty, A., Kramer, J., Cook, L., Fulton, R., Johnson, D., Minx, P., Clifton, S., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R., Muzny, D., Scherer, S., Bouck, J., Sodergren, E., Worley, K., Rives, C., Gorrell, J., Metzker, M., Naylor, S., Kucherlapati, R., Nelson, D., Weinstock, G., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R., Federspiel, N., Abola, A., Proctor, M., Myers, R., Schmutz, J., Dickson, M., Grimwood, J., Cox, D., Olson, M., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G., Athanasiou, M., Schultz, R., Roe, B., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D., Burge, C., Cerutti, L., Chen, H., Church, D., Clamp, M., Copley, R., Doerks, T., Eddy, S., Eichler, E., Furey, T., Galagan, J., Gilbert, J., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L., Jones, T., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W., Kitts, P., Koonin, E., Korf, I., Kulp, D., Lancet, D., Lowe, T., McLysaght, A., Mikkelsen, T., Moran, J., Mulder, N., Pollara, V., Ponting, C., Schuler, G., Schultz, J., Slater, G., Smit, A., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y., Wolfe, K., Yang, S., Yeh, R., Collins, F., Guyer, M., Peterson, J., Felsenfeld, A., Wetterstrand, K., Patrinos, A., Morgan, M., de Jong, P., Catanese, J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y., Szustakowki, J., and Consortium, I. H. G. S. Initial sequencing and analysis of the human genome. *Nature* **409**(6822), 860–921

Feb (2001).

[117] Li, Y.-C., Korol, A. B., Fahima, T., Beiles, A., and Nevo, E. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol Ecol* **11**(12), 2453–65 Dec (2002).

[118] Li, Y.-C., Korol, A. B., Fahima, T., and Nevo, E. Microsatellites within genes: structure, function, and evolution. *Mol Biol Evol* **21**(6), 991–1007 Jun (2004).

[119] Guenatri, M., Bailly, D., Maison, C., and Almouzni, G. Mouse centric and pericentric satellite repeats form distinct functional heterochromatin. *J Cell Biol* **166**(4), 493–505 Aug (2004).

[120] Zhan, S., Horrocks, J., and Lukens, L. N. Islands of co-expressed neighbouring genes in Arabidopsis thaliana suggest higher-order chromosome domains. *Plant J* **45**(3), 347–57 Feb (2006).

[121] Roh, T.-Y., Cuddapah, S., and Zhao, K. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev* **19**(5), 542–52 Mar (2005).

[122] Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L., and Lander, E. S. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**(2), 315–26 Apr (2006).

[123] Li, Q., Peterson, K. R., Fang, X., and Stamatoyannopoulos, G. Locus control regions. *Blood* **100**(9), 3077–86 Nov (2002).

[124] Spitz, F., Gonzalez, F., and Duboule, D. A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell* **113**(3), 405–17 May (2003).

[125] Sproul, D., Gilbert, N., and Bickmore, W. A. The role of chromatin structure in regulating the expression of clustered genes. *Nat Rev Genet* **6**(10), 775–81 Oct (2005).

[126] Gilbert, N., Boyle, S., Fiegler, H., Woodfine, K., Carter, N. P., and Bickmore, W. A. Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* **118**(5), 555–66 Sep (2004).

[127] Cajiao, I., Zhang, A., Yoo, E. J., Cooke, N. E., and Liebhaber, S. A. Bystander gene activation by a locus control region. *EMBO J* **23**(19), 3854–63 Oct (2004).

[128] Trinklein, N. D., Aldred, S. F., Hartman, S. J., Schroeder, D. I., Otillar, R. P., and Myers, R. M. An abundance of bidirectional promoters in the human genome. *Genome Res* **14**(1), 62–6 Jan (2004).

[129] Barrans, J. D., Ip, J., Lam, C.-W., Hwang, I. L., Dzau, V. J., and Liew, C.-C. Chromosomal distribution of the human cardiovascular transcriptome. *Genomics* **81**(5), 519–24 May (2003).

[130] Gabrielsson, B., Carlsson, B., and Carlsson, L. Partial genome scale analysis of gene expression in human adipose tissue using DNA array. *Obes Res* **8**(5), 374–84 Aug (2000).