



PhD-FSTC-2012-25
The Faculty of Sciences, Technology and Communication

DISSERTATION

Defense held on 28/09/2012 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN *BIOLOGIE*

by

Anke Katrin WIENECKE-BALDACCHINO

Born on 10th September 1977 in Wiesbaden (Germany)

IN SILICO PREDICTION OF TRANSCRIPTION FACTOR
BINDING SITES BY PROBABILISTIC MODELS

Dissertation defense committee

Dr Rudi Balling, dissertation supervisor

Professor, Université du Luxembourg

Dr Yesim Aydin Son

Assistent Professor, Middle East Technical University, Ankara

Dr Patrick May, Chairman

Collaborateur scientifique, Université du Luxembourg

Dr Garry Wong

Professor, University of Eastern Finland

Dr Francisco Azuaje, Vice Chairman

Senior Researcher, CRP Santé

IN SILICO PREDICTION OF TRANSCRIPTION FACTOR BINDING SITES BY PROBABILISTIC MODELS

A Dissertation

By

Anke Katrin Wienecke-Baldacchino

University of Luxembourg, Life Science Research Unit

&

Luxembourg Centre for Systems Biomedicine

In partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Dissertation Defense Committee:

Chair of committee: Dr. Patrick May

Committee members: Assist. Prof. Dr. Yesim Aydın Son

Dr. Francisco Azuaje

Prof. Dr. Rudi Balling

Prof. Dr. Garry Wong

Supervisor: Prof. Dr. Rudi Balling

“Wenn wir nach langem Suchen und peinlicher Ungewissheit uns endlich einen bestimmten Sachverhalt erklären zu können glauben, kann unser darin investierter emotionaler Einsatz so gross sein, dass wir es vorziehen, unleugbare Tatsachen, die unserer Erklärung widersprechen, für unwahr oder unwirklich zu erklären, statt unsere Erklärung diesen Tatsachen anzupassen.” (Watzlawick, 1976)

For David

I hereby declare that this dissertation has been written only by the undersigned and without any assistance from third parties. Furthermore, I confirm that no sources have been used in the preparation of this thesis other than those indicated herein.

Nennig, 31.08.2012

Anke Katrin Wienecke-Baldacchino

ACKNOWLEDGMENTS

First of all I would like to thank Dr. Merja Heinäniemi. I can certainly say that without her scientific and mental support this thesis would not have been possible! Her effort to give scientific suggestions to improve, her tremendous time investment to read the different chapters felt 100 times I cannot acknowledge enough. The last five years that we worked together broadened my professional and personal horizon strongly. Her talent to catch up scientific problems and her overwhelming motivation to explore the same is simply admirable. I am very grateful that I had the pleasure to get to know “this Finnish young girl” and hope that our personal contact will last, also in case our scientific work might diverge. Merja, thanks for all the mental energy you send me to Nennig and let me know if I can ever reciprocate for what you did for me. I wish you all the best, that you reach all your scientific and personal goals. And if not – call me ;-).

Secondly, I would like to thank Dr. Nikos Vlassis. He was the person who did an incredible job by inspiring me with machine learning methods, which I honestly would have never touched by choice. His calm character and loyalty is enviable. I never met a person like him, able to explain very complex statistical and mathematical problems with such a respect and patience. For me as a non-machine-learning person it was absolutely essential to meet someone like Nikos considering his didactical qualities. Without his input, suggestions and personal efforts to improve the methodology, this work would not have been possible. His ideas and power of comprehension - also of biological aspects not belonging to his core competence – are amazing. I hope that we will stay in touch whatever happens on our scientific way.

Special thanks I would like to express for Prof. Dr. Rudi Balling, who was there as times have been hard and offered me a place to finish my dissertation. He gave me all freedom needed to develop independently and supported me by any means.

I would like to thank Dr. Francisco Azuaje, Assist. Prof. Dr. Yesim Aydın Son, Dr. Partick May and Prof. Dr. Garry Wong for accepting to participate in my dissertation jury. It is a great honor for me to have my work evaluated by them.

I would like to thank the University of Luxembourg, in particular the Life Science Research Unit, as well as the Fonds National de la Recherche for financial and material means to perform my dissertation.

I also would like to acknowledge John Reid, the developer of STEME for tremendous assistance to install and run STEME, even providing a modified version to accomplish an automating of the analysis pipeline.

Special thanks to the former CC-group, which finally turned out to be a great team! I share some very impressive experiences with you, which sharpened my way of judging people and research habits strongly in a positive manner. For fruitful, funny, intensive discussions and

ACKNOWLEDGMENTS

mental support I would like to thank my “room mates” Cathy and Martine. Also Janine, Elisabeth, Anna and Ola who accompanied me on my way, I would like to gratefully thank for fruitful discussions.

My sincere gratitude I would like to express to the FSTC and LCSB secretaries, Stephanie, Caroline, Veronique and Aurelia for supporting me whenever I had an administrative problem.

Also Serge, Kirsten, Jochen, Cédric and Fotis I would like to thank for honest, funny and serious discussions.

Special thanks to the proof-reading task force – Dr. Cathy Fiatte, Dr. Lasse Sinkkonen, Elisabeth John, Dr. Monique Wiesinger and Dr. Stephanie Kreis. All of you did a great job in such a short time. Your suggestions improved the work significantly.

Also a special thanks to Steph, Claude and Monique for supporting me morally when times have been hard, always (successfully) motivating me to continue when I was close to chuck the thesis.

Ralf, my “personal IT dealer” I would like to thank for really funny conversations and great IT support.

Special thanks also to Prof. Dr. Hans Pfeiff, my personal mentor since several years. He inspired me to come to Luxembourg and supported me with “brain food” and moral energy, always standing beside me offering me the possibility to express myself in a free and irrespective manner.

Ein riesen Dankeschön an meine kleine Familie - Uschi, Christoph, Petra and Heinz - die mich immer unterstützt hat in all meinen Entscheidungen egal wie “wahnsinnig” sie gewesen sein mögen. Sie waren und sind immer für mich da, in guten, wie in schlechten Zeiten. Also my big “French-Italian Connection” – Maria, Filippo, Lucy, José, Luca, Julia, Rosanna, Pier, Jessica and Matteo – I would like to thank for amusing family parties reflecting a nice change during the hard time of thesis execution and writing.

Finally I would like to thank my husband David! I cannot express how much he supported me during the time of the thesis and in particular the final phase, taking over selflessly all tasks of daily life – cooking, cleaning, shopping etc. Without him a writing of the thesis in the given time frame would have been impossible. Thanks for your leniency referring to my bad moods, stressful behavior and overnight working sessions. I think I can say, that I won Euromillion already! Je voudrais te remercier pour tout, mon cher! Je t'embrasse. ILYVVMML!

ABBREVIATIONS

#	number of / amount
1000G	1000 Genomes Project
Ago	Argonaute protein family
ASB	Allele-Specific Binding
ASE	Allele-Specific Expression
ATP	Adenosine-5'-Triphosphate
AUC	Area Under the Curve
bp	base pair
BS	Binding Site
CBP	CREB-binding protein
CDP	Combined Density Profile
CEPH	Centre d'Étude du Polymorphisme Humain
CEU	HapMap/1000G population: Utah residents with Northern and Western <i>European</i> ancestry
CHD	Chromodomain Helicase DNA-binding
ChIP	Chromatin Immunoprecipitation
CLTree	Chow-Liu Tree
CNP	Copy Number Polymorphism
CNV	Copy Number Variation
<i>CRM</i>	<i>Cis</i> -Regulatory Module
CRT	Cyclic Reversible Termination
CTCF	CCCTC-binding factor
DGCR8	DiGeorge syndrome critical region gene
DNA	Deoxyribonucleic Acid
DNMT	DNA methyltransferase
DOE	U.S. department of energy
dsRBD	double-stranded RNA-binding domain
E2F2	E2F transcription factor 2
E2F3	E2F transcription factor 3
EM	Expectation Maximization
EMBL	European Molecular Biology Laboratory
ENCODE	ENCyclopedia Of DNA Elements
Eomes	Eomesodermin
ERG1	Early growth response 1
Esrra	Estrogen-related receptor alpha
ET	Ensemble of Trees
ETPx	Probability Score based on ET model
ETS1	v-ets erythroblastosis virus E26 oncogene homolog 1 (avian)
FAIRE	Formaldehyde-Assisted Isolation of Regulatory Elements
FDR	False Discovery rate
FM index	Ferragina and Manzini index

ABBREVIATIONS

FMM	Feature Motif Models
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
GAD	Genetic Association Database
Gcm1	Glial cells missing homolog 1 (Drosophila)
GEO	Gene Expression Omnibus
GO	Gene Ontology
GTF	General TF
GWAS	Genome Wide Association Studies
GWM	Generalized Weight Matrix
H2BK5ac	Histone 2B lysine 5 acetylation
H3K27	Histone 3 lysine 27
H3K27ac	Histone 3 lysine 27 acetylation
H3K36	Histone 3 lysine 36
H3K4	Histone 3 lysine 4
H3K4me3	Histone 3 lysine 4 trimethylation
H3K79	Histone 3 lysine 79
H3K79me1	Histone 3 lysine 79 monomethylation
H3K9	Histone 3 lysine 9
H4K20	Histone 3 lysine 20
H4K20me1	Histone 4 lysine 20 monomethylation
HapMap	Halpotype Map
HAT	Histone acetyltransferases
Hbp1	HMG-box transcription factor 1
HDAC	Histone deacetylase
HDM	Histone demethylases
hg18	corresponds to NCBI36
HGP	Human Genome Project
HMT	Histone methyltransferases
IGV	Integrated Genome Browser
indel	Insertion/Deletions as genetic variation
IRF4	Interferon regulatory factor 4
Irf5	Interferon regulatory factor 5
ISWI	Imitation Switch
kb	kilo bases
KDE	Kernel Density Estimation
KLHDC7B	Kelch domain containing 7B
MAF	Minor Allele Frequency
MAX	MYC associated factor X

ABBREVIATIONS

MBD	Methyl-CpG-Binding Domain
MeCP2	Member of the methyl-CpG-binding domain proteins
MEME	Multiple EM for Motif Elicitation
MI	Mutual Information
mio	million
miRNA/miR	micro RNA
MM	an extension of the EM technique
mRNA	messenger RNA
Myb	v-myb myeloblastosis viral oncogene homolog (avian)
Mybl1	v-myb myeloblastosis viral oncogene homolog (avian)-like 1
NAR	Nucleic Acid Research
NCBI36	National Center for Biotechnology Information Human Genome Version/Assembly
NGS	Next Generation Sequencing
NIH	U.S. National Institute of Health
NP-hard	Non-Deterministic polynomial-time hard
Nr2f2	Nuclear receptor subfamily 2, group F, member 2
nt	nucleotide (in context of sequence length)
ODF3B	Outer dense fiber of sperm tails 3B
PACT	PKR activating protein
PBM	Protein Binding Microarray
PCR	Polymerase Chain Reaction
PFM	Position Frequency Matrix
PHRED	quality scores are assigned to each base call in automated sequencer trace
PKR	RNA-dependent protein kinase
POLII	RNA polymerase II
POU2F2	POU class 2 homeobox 2
pri-miRNAs	primary miRNAs
PVLMM	Permutated Variable Length Markov Model
PWM	Position Weight Matrix
PWMPx	Probability Score based on PWM model
qPCR	quantitative PCR
QuEST	Quantitative Enrichment of Sequence Tags
Rara	Retinoic acid receptor, alpha
RISC	RNA-induced silencing complex
RNA	Ribonucleic Acid
RNase III	ribonuclease III
ROC	Receiver Operating Characteristic
RPTOR	Regulatory associated protein of MTOR, complex 1
rSNPs	regulatory SNPs
RXRA/Rxra	Retinoid X receptor, alpha

ABBREVIATIONS

SNE	Single Nucleotide Exchange
SNP	Single Nucleotide Polymorphism
Sox11	SRY (sex determining region Y)-box 11
Sox4	SRY (sex determining region Y)-box 4
Sox7	SRY (sex determining region Y)-box 7
Sox8	SRY (sex determining region Y)-box 8
Spdef	SAM pointed domain containing ets transcription factor
SPI1	Spleen focus forming virus (SFFV) proviral integration oncogene spi1
SRA	Sequence Read Archive
SRF	Serum response factor (c-fos serum response element-binding transcription factor)
STEME	Suffix Tree EM for Motif Elicitation
SV	Structural Variation
SWI/SNF	SWItch/Sucrose NonFermentable
SYCE3	Synaptonemal complex central element protein 3
TAR	Trans-Activation Response
TBP	TATA box-binding protein
Tcf2a	TCF3 transcription factor 3
TF	Transcription Factors
TFBS	Transcription Factor Binding Site
TN	True Negatives
TP	True Positives
TPR	True Positive Rate
TRBP	Trans-activation Response RNA-binding protein
TreePx	Probability Score based on CLTree model
TSS	Transcription Start Site
TYMP	Thymidine phosphorylase
UCSC	University of California, Santa Cruz
UTR	Untranslated Region
vcf	variation call format
WGS	Whole Genome Shotgun (sequencing)
Zfp281	Zinc finger protein 281

Summary

The characterization of *in silico* detected transcription factor binding sites represents a fundamental problem in the field of regulatory gene expression analysis. Several approaches have been proposed to model DNA-protein-interactions, composed by two main classes: qualitative models considering a consensus sequence and quantitative models providing a measure of binding affinity. The latter can be further subdivided in models assuming an independent contribution of the nucleotides forming a potential binding site and more flexible ones implicating a positional interdependence.

In this work the applicability of three probabilistic models to predict transcription factor binding sites has been investigated: (i) the simple position weight matrix (PWM), assuming independence, and two flexible models capturing positional interdependencies represented by a (ii) Chow-Liu Tree and (iii) Ensemble of Trees model. The training and validation of the models on the *Mus musculus* subset of the UniPROBE database revealed that complex models provide a better predictive power suggesting a high amount of transcription factors binding motifs being affected by positional interdependencies. Additionally, numerous transcription factors were detected, for which the Ensemble of Trees model outperformed both, the Chow-Liu Tree and PWM model.

The UniPROBE-based trained models have been applied in a biological context - the prediction of differential binding profiles in five different ChIP-seq samples, followed by the detection of causative regulatory SNPs. The chosen set-up involved family trio data, meaning genotype data from a family composed of father, mother and daughter, providing internal validation. The models provide strong power to correctly classify true negatives in an independent biological sample, represented by a high specificity. The applied approach to detect causative regulatory SNPs, resulted in a candidate list of 20 SNPs. Those gain strong support by epigenetic markers and both, model-based predicted binding affinity of the comprising binding site and significant p-values, describing the effect of the nucleotide exchange.

1. Introduction.....	1
1.1. A historical review - the 19 th and 20 th Century.....	2
1.2. The Human Genome Project – the Beginning of the Genome Era	5
1.3. The ENCODE - ENCyclopedia Of DNA Elements - Project.....	8
1.4. The Haplotype Map Project.....	9
1.5. The 1000 Genomes Project	10
1.6. Levels of Gene Expression Regulation	12
1.6.1. Transcriptional Regulation by Transcription Factors	13
1.6.2. Post-transcriptional Regulation by miRNAs.....	16
1.6.3. Epigenetic Modifications.....	18
1.6.3.1. Nucleosome Positioning and Chromatin Remodeling	18
1.6.3.2. DNA Methylation	20
1.6.3.3. Histone Modification	22
1.7. Next Generation Sequencing	25
1.7.1. Technology	25
1.7.1.1. Template Preparation	25
1.7.1.2. Genome Alignment and Assembly	27
1.7.2. Application of NGS - Chromatin Immunoprecipitation Sequencing	28
1.8. Modeling Transcription Factor Binding Sites.....	32
1.9. Data Integration for Regulatory Element Detection.....	39
1.10. Approaches for regulatory SNP Detection	40
2. Objectives	44
3. Material and Methods.....	45
3.1. UniPROBE – Protein Binding Array Data.....	45
3.2. Selection of TFs	47
3.3. Generating Training Input and <i>de novo</i> Motif Discovery	47
3.4. Screening and Scoring of Sequences.....	50
3.5. Probabilistic TF Binding Models.....	51
3.5.1. Position Weight Matrix.....	51
3.5.2. Chow-Liu Tree	54
3.5.3. Ensemble of Trees	59
3.6. Model Validation.....	64
3.7. Probabilistic Models in causative SNP Detection in Family Data.....	67
3.7.1. SNP Data.....	68
3.7.2. Family Trio Data	69

TABLE OF CONTENTS

3.7.3.	Selection of TFs.....	70
3.7.4.	Determination of Significance of TF Binding	70
3.7.5.	Assigning Significance to SNP Effect by SNE Distributions	71
3.7.6.	ChIP-seq Analysis and Peak Detection	73
3.7.6.1.	Sequence Read Preprocessing	73
3.7.6.2.	Alignment of Reads	74
3.7.6.3.	Peak Detection	76
3.7.7.	Training Probabilistic Models.....	79
3.7.8.	Screening of Peaks	80
3.7.8.1.	Detection of Common and “Missing” Peaks	80
3.7.8.2.	Screening for best Binding Site	80
3.7.9.	Assigning merged data with Significance	81
3.7.10.	Reference set Generation of Differential Parental Binding	82
3.7.11.	Sensitivity and Specificity Analysis for Peak Profile Classification	85
3.7.12.	Determination of potentially Causative SNPs.....	86
4.	Results.....	87
4.1.	General Overview and Classification of the UniPROBE mouse TF set	87
4.2.	Detailed Results for the Probabilistic Models for selected TFs	93
4.2.1.	Probabilistic Model Training and Validation using UniPROBE data	93
4.2.1.1.	PWM model	93
4.2.1.2.	CLTree model	94
4.2.1.3.	ET model	96
4.2.1.4.	AUC-profiles of Model Validation	97
4.3.	Application of Probabilistic Models for Differential Peak Profile and rSNP detection... ..	101
4.3.1.	Transcription Factor Selection	101
4.3.2.	Training Probabilistic Models.....	101
4.3.2.1.	PWM model	101
4.3.2.2.	CLTree model	102
4.3.2.3.	Model Validation	103
4.3.3.	ChIP-seq Analysis and Peak Detection	105
4.3.3.1.	Sequence Read Preprocessing	105
4.3.3.2.	Alignment of Reads	106
4.3.3.3.	Peak Detection	106
4.3.4.	Descriptive Analysis of Maternal and Paternal Peaks	107
4.3.4.1.	Detection of Common and Parent Specific Peaks	107
4.3.4.2.	Detection of Peak-SNP-co-location	108
4.3.5.	Detection of Differential TF Binding Profiles in Parental Peaks.....	109
4.3.5.1.	Sensitivity and Specificity	110

TABLE OF CONTENTS

4.3.5.2. Winning Parent	111
4.3.6. Detection of Causative SNPs	113
4.3.6.1. SNP-BS co-location and BS Sequence Differences	113
4.3.6.2. Candidate List of Causative SNPs.....	114
5. Discussion	123
5.1. Probabilistic Transcription Factor Binding Models	123
5.2. Detection of Differential Binding in ChIP-seq Experiments	126
5.3. Detection of causative SNPs.....	131
6. Conclusions and Future Work	138
7. References	142
8. ANNEX.....	155

1. Introduction

In these days researchers are discussing if complex disease mechanisms are caused by probably interacting but rarely occurring single nucleotide polymorphisms (SNPs). These SNPs can be investigated in individual genomes, driven by the hope to successfully develop treatments for diseases like Alzheimer's disease or cancer. Different approaches are applied, ranging from gene networks, the analysis metabolic pathways by computational approaches, or the modeling gene-environment interactions. Soon a complete human genome sequence might be resolved within hours for 100\$ (Schadt, Linderman, Sorenson, Lee, & Nolan, 2010).

Recent discoveries and project initiatives pioneered bioinformatics, a highly dynamically evolving field of biology, which is shifting to an integrative science, joining physics, mathematics, informatics and classical molecular biology in a new discipline better described as systems biology.

Focusing on the regulation of gene expression as one of the major topics to elucidate complex disease mechanisms and their genetic causation patterns arises following questions:

Which levels of gene expression regulation are known and which are the main biological players?

Which technologies are existing today, enabling science to investigate gene regulation with a resolution of a single base pair?

Which computational approaches exist to model Deoxyribonucleic Acid (DNA)-binding events?

The classical approach, still willingly applied by life scientists, to analyze data with Excel, will soon become obsolete, due to the 1,048,576 row and 16,384 column limit. How to deal with these massive data amounts provided by new technologies?

The aim of this chapter is to give an insight and some answers to these questions.

1.1. A historical review - the 19th and 20th Century

In 1859 Darwin published his book “*On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle of Life*”, based on ideas he gained during a five-year journey in the 1830’s including a visit to the Galapagos Islands. Due to observation of the variety of living organisms and fossils, he concluded that species change through natural selection. Not even ten years later (1865) Mendel presented results derived from his experiments with peas, which lead to the basic principles of hereditary transmission in genetics which are still valid today. DNA has been isolated for the first time in the same decade by Friedrich Miescher, followed by the description of chromosomal behavior in 1879 by Walter Flemming. The term “gene” has been defined in 1909, to describe the Mendelain units of heredity by Wilhelm Johannsen, who also made the first distinction between phenotype and genotype. In 1933 Thomas Hunt Morgan received the Nobel Prize in Physiology or Medicine for his achievements in establishing the chromosome theory of inheritance. He and his students, showed by studies with the fruit fly *Drosophila melanogaster*, that genes are seeded on chromosomes, and described for the first time chromosome recombination (Morgan, Sturtevant, Muller, & Bridges, 1915).

In 1944 Barbara McClintock discovered by experiments using *Zea Mays*, that certain genes “are mobile” on chromosomes, showing an unexpected flexibility of the genome (McClintock, 1944). The respective structures, today called “transposons”, have been found later to be distributed over all kinds of organisms – eukaryotes as well as prokaryotes.

The 1950’s two hallmark discoveries were obtained: the elucidation of the DNA structure and the discovery of the DNA replicating enzyme “DNA polymerase”. The competitive work of different scientists on analyzing the DNA structure peaked with the epoch-making publication of the model of the DNA-Alpha-Helix by Francis Crick and James Watson in 1953 (J. Watson & Crick, 1953). Two years later Arthur Kornberg, working on *E. coli*, discovered the DNA polymerase, and therefore the mechanisms of DNA synthesis (Kornberg, 1974). This finding pioneered all modern kinds of recombinant DNA technologies and sequencing.

In the 60’s the principle of gene regulation has been described for the first time by Jacob and Monod (Jacob & Monod, 1961) referring to the lac-operon in *E.coli*. Some years later and ~25 years after McClintock, John Britten showed, that also eukaryotic genomes have many repetitive, non-coding DNA sequences. In 1969 he published a theoretical paper “Gene Regulation for Higher Cells: A Theory” (Britten & Davidson, 1969), which defined the basics for the modern understanding of the regulation of gene expression. Also the genetic code was deciphered by Nirenberg and colleagues (Nirenberg, 1963).

While Watson and Crick resolved the DNA structure, or Britten discovered eukaryotic gene regulation, computational scientists worked on information theory (Shannon & Weaver, 1962) random strings (Martin-Löf, 1966) or the theory of games (Neumann & Morgenstern, 1953). Looking from today's point of view, it seems obvious that the majority of biological discoveries at that time represented optimal playgrounds for computer science to test or improve their models. Thus, the step to explore methods developed in computer science on biological problems can be seen as the birth of bioinformatics or computational biology.

In 1977 introns have been discovered (Berk & Sharp, 1977; L. Chow, Roberts, Lewis, & Broker, 1977). At that time, these non-coding regions have been thought of being junk DNA with the only function to “fill up” the sequence without any other purpose. As known today, this assumption was wrong.

The first half of the century was mainly characterized by hallmark discoveries on basic structures involved in genetics. This trend continued in the second half, but strongly accompanied by the development of technologies and approaches to analyze or manipulate, these structures. One example is the development of Sanger- and Maxam-Gilbert DNA sequencing (Maxam & Gilbert, 1977; Sanger *et al.*, 1977; Sanger & Coulson, 1975), leading to rapid developments in sequencing. As a matter of fact, one of the first bioinformatics task being addressed was sequence alignment. Given a poorly characterized protein, it is possible to search for homologues that are better understood. With caution, the knowledge of the better understood homologue can be explored to the poorly characterized protein (Luscombe, Greenbaum, & Gerstein, 2001).

One of the first algorithms for sequence comparison on primary structures was published in 1970 by Needleman and Wunsch (Needleman & Wunsch, 1970), representing the first implementation of dynamic programming for protein sequence comparison. However, also other biological questions such as genetic evolution by gene duplication or the derivation of preference for amino acid residues in secondary structure and many more, defined a catalog of problems to be addressed by bioinformatics approaches during the following decades (Ouzounis & Valencia, 2003).

Pioneer work was performed by providing public data resources, curation and storage referring to protein sequence data with the “Atlas of Protein Sequence and Structure” (Dayhoff, 1978) and “The Protein Data Bank” (Bernstein *et al.*, 1977). Compilations of data within public data resources amplified tremendously in direct future, and keeps on going until today.

In 1981 and 1982, when the first transgenic mice and fruit flies have been generated, a new era in analyzing gene-functions in eukaryotic systems started. In the same decade the first

gene involved in a disease has been mapped by means of DNA polymorphisms, represented by a small segment on chromosome 4, revealing a primary genetic defect in Huntington disease (Gusella *et al.*, 1983). In 1983 automated polymerase chain reaction (PCR)-technique had been invented, which dramatically boosted the pace of genetic research (Saiki *et al.*, 1985). From this year on, it was possible to amplify DNA segments within a few hours to billions of copies. This invention pioneered the next generation sequencing technologies as they are in use today.

The expectation that a better understanding of biology will be driven by better computational analysis of nucleotide sequences (Gingeras & Roberts, 1980) this field continued to benefit from parallel developments in computer science. Beside theoretical developments in sequence analysis, as for instance the computation of evolutionary distances (Sellers, 1980) or approximate string matching (Ukkonen, 1985), key algorithms were developed, like Smith and Waterman (Smith & Waterman, 1981) representing a dynamic programming sequence alignment algorithm or the FASTA algorithm (Lipman & Pearson, 1985).

As a hallmark in database development and quality control in bioinformatics the formation of two major resources for nucleotide data submission, namely the GenBank database (Bilofsky & Burks, 1988) and European Molecular Biology Laboratory (EMBL) Data Library (Hamm & Cameron, 1986), can be considered. At the time the GenBank paper appeared, GenBank contained already 15,000 DNA and Ribonucleic Acid (RNA) entries that have been reported since 1967.

Until today databases serve as one of the major resources for hypothesis driven biology, including literature databases like PubMed. However, the amount of biological databases grew exponentially. Today there are even journals existing just dedicated to the development and publication of newly launched databases (Landsman, Gentleman, Kelso, & Francis Ouellette, 2009). As to take from Figure 1, in 2012 the number of databases being tracked by Nucleic Acid Research (NAR) (Galperin & Fernández-Suárez, 2012) has grown from 1,330 in 2011 to 1,380 in 2012 (Todd Smith, 2012).

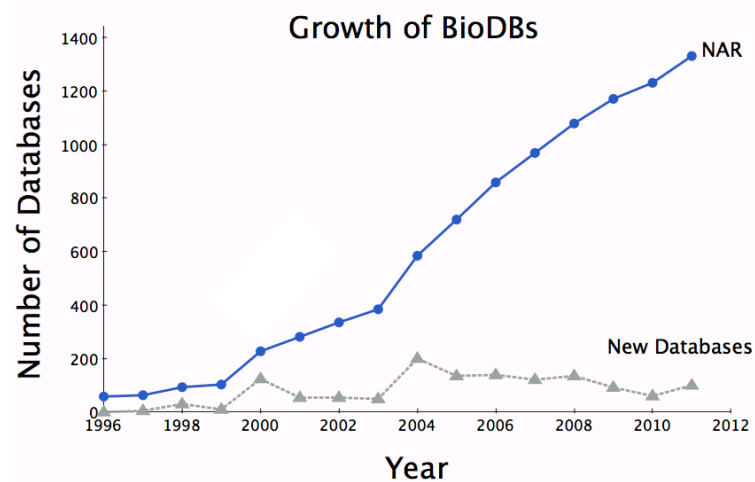


Figure 1: Development of the number of biological databases from 1996 to 2011. New databases (triangle, grey) are the difference between the number of existing databases (circle, blue) for each year. NAR: Database Issue of the journal *Nucleic Acid Research*. Taken and adapted from (Todd Smith, 2012).

Though data curation is an essential task, the information stored in these databases is essentially useless until analyzed. Accordingly, the need to develop tools and resources in an organized manner is required (Luscombe *et al.*, 2001). Therefore, it is not surprising, that in the late 80's scientific programs were launched by institutions like the EMBL and NIH, forming departments exclusively directed to computational biology. Database development evolved considering relational database systems to facilitate querying and experimentation with hardware platforms for more efficient sequence analysis started.

1.2. The Human Genome Project – the Beginning of the Genome Era

In 1990 the Human Genome Project (HGP) had been launched, initially based on the interest of the U.S. department of energy (DOE) to detect mutations that nuclear radiation might cause. This interest was also shared, with the goal of advancing medicine, by the U.S. National Institute of Health (NIH). The project was projected to be completed in 15 years, by the year 2005. The initial set of goals included beside ethical, legal and social implications, the mapping of the human genome, eventually accompanied by the determination of the entire sequence, the sequencing of other organism's genomes to gain basic biological knowledge and to accelerate technological development related to analyzing DNA (J. D. Watson & Jordan, 1989).

In parallel first internet-like structures appeared in mid 1990s, allowing accessing a limited amount of web sites. Distributed databases, like GenBank and Medline, offered public access, however, simply their availability was problematic, mainly realized by means of CD-ROMs (Ouzounis & Valencia, 2003). Interpreter languages like Perl¹ or Python² appeared, inspired

¹ www.perl.org

by the Unix utility *awk* as well as first sophisticated gene-prediction programs (Guigó, Knudsen, Drake, & Smith, 1992) while the community was yielding the first successful results in protein docking (Walls & Sternberg, 1992). In 1992 the genomics information era celebrated its first computational re-annotation of the first ever entire chromosome sequence, the yeast chromosome III (Bork *et al.*, 1992) and the mouse genetic map had been completed (Dietrich *et al.*, 1996).

In 1996, the Human DNA sequencing started by pilot projects to find efficient strategies for completely sequencing the human genome. Two years later, a private company Celera Genomics was founded, with the aim to sequence the Human genome within three years, based on shotgun sequencing (Venter *et al.*, 1998).

In 2001 the first draft, covering more than 90%, of the human genome sequence had been released by both, the official Human Genome Sequencing Consortium as well as from Celera Genomics (Lander *et al.*, 2001; Venter *et al.*, 2001). The total number of genes in the human genome was much lower than expected with “only” 30,000-35,000, later further reduced to little over 25,000 (after a revision in 2003), where only 2% of the sequence coded for proteins. For over half of the genes the function was not known, and more than 3 mio SNPs have been detected.

Parallel efforts to sequence the mouse genome, thought to serve as an important resource to disclose the mysteries of the human genome, succeeded in 2002 (Waterston *et al.*, 2002). Since sequences, conserved between the two organisms are presumed to be functionally important, mice served as an experimental animal model to investigate gene functions, where in many cases the results of the studies are transferable to the Human Genome.

In 2003 the International Human Genome Sequencing Consortium announced the Human Genome project being completed, two years earlier than projected (Collins, Green, Guttmacher, & Guyer, 2003). The finished sequence covered 99% of the genome with 99.9% accuracy.

New sequencing technologies accelerated the deciphering of the genomes of numerous other organisms and led to an ongoing exponential growth of sequences deposited in sequence databases like GenBank (see Figure 2).

In April 2011 Genbank hosted approximately 126,551,501,141 bases in 135,440,924 sequence records in the traditional GenBank divisions and 191,401,393,188 bases in 62,715,288 sequence records in the whole genome shotgun (WGS) division.

² www.python.org

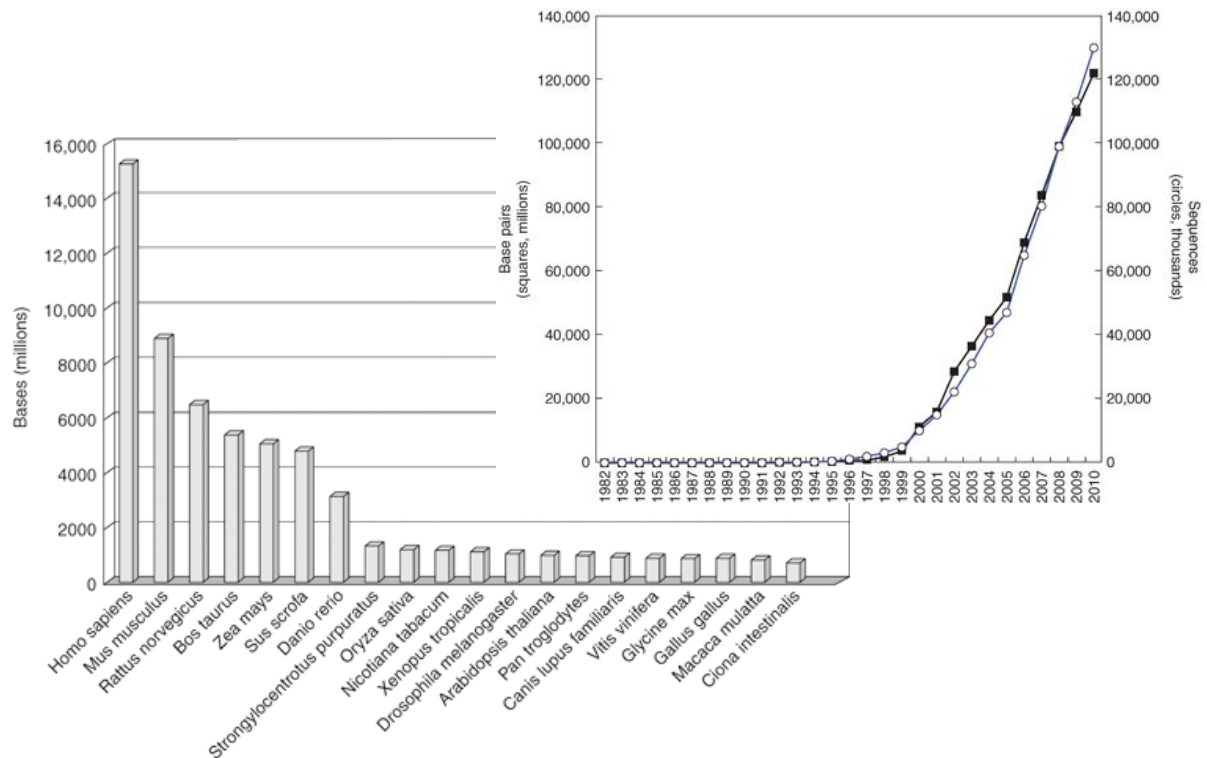


Figure 2: (upper right) Exponential growth of GenBank. Period of accelerated growth after 1997 coincides with the completion of the HGP's genetic and physical mapping goals, setting the stage for high-accuracy, high-throughput sequencing, as well as the development of new sequencing technologies. (lower left) The number of nucleotide bases currently in GenBank for the 20 most-sequenced organisms, excluding chloroplast or mitochondrial sequences, number can exceed the actual size of the organism's genome. Taken from (Baxeavanis, 2011)

Starting from “simple” Sanger or Maxam-Gilbert sequencing, nowadays researchers have access to technologies classified under next generation sequencing (NGS), referring to approaches allowing massive parallel sequencing of million DNA fragments in one sequencing run (see chapter 1.7, Figure 3).

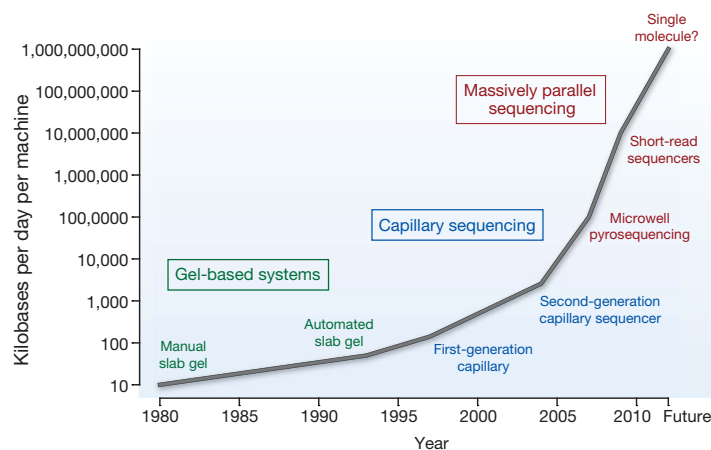


Figure 3: Improvements in the rate of DNA sequencing over the past 30 years and into the future. From slab gels to capillary sequencing and second-generation sequencing technologies, there has been a more than a million-fold improvement in the rate of sequence generation over this time scale. Taken from (Stratton, Campbell, & Futreal, 2009)

1.3. The ENCODE - ENCyclopedia Of DNA Elements - Project

The ENCODE (**ENC**yclopedia **O**f **DNA** **E**lements) project was launched in 2003, as the follow-up of the Human genome project (NIH, 2003) to identify and locate of all functional elements in the human genome. These elements comprise all protein coding and non-protein coding genes as well as so-called junk DNA or non-coding regions (e.g. introns, transposons), assumed to play an important role in gene expression regulation.

In 2007 the first results from the pilot phase have been published, reporting the generation and analysis of functional data from multiple, diverse experiments performed on a targeted 1% of the human genome. Some of the main findings reported are (Birney *et al.*, 2007):

- The majority of the bases of the human genome are associated with at least one primary transcript and many transcripts link distal regions to protein-coding loci.
- Many regions in the genome thought to be transcriptionally silent have been identified as non-protein-coding transcripts.
- Measures for chromatin accessibility and histone modifications are highly predictive for the presence and activity of transcription start sites.
- Distal DNaseI hypersensitive sites show marks consistent with insulator function and provide histone modification patterns reliably distinguishing them from promoters.
- Functional elements can vary greatly in their sequence variability across the human population and their likelihood of being placed within a structurally variable region of the genome.
- Many functional elements appear to be unconstrained across mammalian evolution suggesting a large pool of neutral elements that are biologically active but provide no specific benefit to the organism. It is assumed that this pool may serve as a 'warehouse' for natural selection, potentially acting as the source of lineage-specific elements and functionally conserved but non-orthologous elements between species.

Very recently³ the ENCODE project published numerous papers presenting latest results. The results refer to the analysis of 1,640 data sets and have been integrated from diverse experiments within 147 different cell types and ENCODE data with other resources including GWAS or evolutionary information. For the sake of shortness only the main findings will be quoted in the following (Dunham *et al.*, 2012):

- "The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type. Much of the genome

³ in fact one week after the official submission of this work

lies close to a regulatory event: 95% of the genome lies within 8 kb of a DNA–protein interaction (as assayed by bound ChIP-seq motifs or DNaseI footprints), and 99% is within 1.7 kb of at least one of the biochemical events measured by ENCODE.”

- “Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.”
- “Classifying the genome into seven chromatin states indicates an initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoter-like features, as well as hundreds of thousands of quiescent regions. High-resolution analyses further subdivide the genome into thousands of narrow states with distinct functional properties.”
- “It is possible to correlate quantitatively RNA sequence production and processing with both chromatin marks and transcription factor binding at promoters, indicating that promoter functionality can explain most of the variation in RNA expression.”
- “Many non-coding variants in individual genome sequences lie in ENCODE-annotated functional regions; this number is at least as large as those that lie in protein-coding genes.”
- “SNPs associated with disease by GWAS are enriched within non-coding functional elements, with a majority residing in or near ENCODE-defined regions that are outside of protein-coding genes. In many cases, the disease phenotypes can be associated with a specific cell type or transcription factor.”

1.4. The Haplotype Map Project

In 1998, the SNP initiative was started by the NIH, with the aim to use SNPs as markers on a genetic map⁴ (Collins, Brooks, & Chakravarti, 1998). Four years later, in 2002, the Haplotype Map (HapMap) project had been announced (Tanaka, 2005), with the aim to provide data on human genetic variation, with minor allele frequency of >5%, to be used to accelerate the identification of genes associated with common and chronic diseases such as cancer, diabetes, or heart disease. In order to gain representative data, taking into account ethnic differences and ancestry, the DNA had been sampled from an African (Nigeria), two Asian (China, Japan) and an American (with European ancestry) group, with 269 individuals in total. This project was also the starting point for genome wide association studies (GWAS), aiming to find common genetic variants being associated with traits, like a common complex disease or phenotype.

⁴ There characteristic of a SNP in comparison to a point mutation is that it persists in the populations, so that it can serve as a general biomarker. A point mutation in turn can occur as an event in a single individual without any impact in the general population.

In 2005 the HapMap project completed phase I, publishing ~1 mio SNPs, followed by phase II in 2007 adding further 2 mio SNPs. In 2010 the results of phase III have been published. In comparison to the previous phases the number of individuals genotyped had been increased on 1,184 from 11 global populations. 1.6 mio common SNPs had been genotyped and ten 100-kilobase ENCODE regions in 692 individuals had been sequenced. The resulting dataset includes both SNPs and copy number polymorphisms (CNPs). All over, the Human Genome Project, the SNP Consortium and the International HapMap Project collectively identified ~10 million common DNA variants, primarily SNPs, in a limited set of DNA samples (Altshuler *et al.*, 2010). The amount of detected SNPs is increasing tremendously. The latest build of dbSNP – the online resource administrating SNPs submitted by numerous projects (1000 genomes project, HapMap, personal genomes, etc) – reports 187,852,828 submissions and 53,558,214 RefSNP clusters, wherein 38,077,993 have been validated⁵.

1.5. The 1000 Genomes Project

The most recent and outstanding effort in genetics is the 1000 genomes project (1000G). It has been launched in 2008 as an international effort to discover, genotype and accurate haplotype information on all forms of human DNA polymorphisms, including SNPs, indels and CNVs, in multiple populations. In more detail, the aim is to characterize >95% of variants with a minor allele frequency of >1%, that are in genomic regions accessible with high-throughput sequencing technologies, in five major populations (Europe, East Asia, South Asia, West Africa and America) reflected by 2,500 individuals. Additionally, in the catalogue functional alleles in coding regions and lower frequency alleles down to 0.1% are recorded. The results from the pilot phase have been published in 2010 as summarized in Table 1 (The 1000 Genomes Project Consortium, 2010), providing data for three phases considered: 1) low-coverage whole-genome sequencing of 179 individuals from four populations, 2) high-coverage sequencing of two mother–father–child trios and 3) exon-targeted sequencing of 697 individuals from seven populations.

The table below lists the results extracted for the European samples, including deep sequencing data from the CEU family trio used in this work. The NCBI36 reference genome (hg18) served as a reference for sequence read alignment. For the low-coverage analysis in the CEU samples the accessible genome contains on average 86% of the reference sequence and from nearly 8 million detected SNPs, 33% are novel, disclosing the power for sequencing technologies in comparison to genotyping chips. The higher amount of novel

⁵ http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi

SNPs in the low-coverage and exon project is based on the larger sample size in comparison to the family trio, allowing to detect a higher number of low frequency variants (Minor Allele Frequency (MAF) < 5%).

Table 1: Variants discovered by the 1000 genome project for the CEU samples by project (low-coverage, Trios, Exon) (The 1000 Genomes Project Consortium, 2010), ND: not determined

	Low Coverage	Trios	Exon
Samples	60	3	90
Total raw bases (Gb)	1,402	560	151
Total mapped bases (Gb)	817	369	-
Mean mapped depth (X)	4.62	43.14	73
Bases accessed (% of genome)	2.43 Gb (86%)	2.26 Gb (79%)	1.4 Mb
No. of SNPs (% novel)	7,943,827 (33%)	3,646,764 (11%)	3,489 (34%)
Mean variant SNP sites per individual	2,918,623	2,741,276	715
No. of indels (% novel)	728,075 (39%)	411,611 (25%)	23 (No. novel: 10)
Mean variant indel sites per individual	354,767	322,078	3 (not mean)
No. of deletions (% novel)	ND	6,593 (41%)	ND
No. of genotyped deletions (% novel)	ND	ND	ND
No. of duplications (% novel)	259 (90%)	187 (93%)	ND
No. of mobile element insertions (% novel)	3,202 (79%)	1,397 (68%)	ND
No. of novel sequence insertions (% novel)	ND	111 (96%)	ND

The main aim of the three pilot projects was to develop and evaluate access different strategies for genome-wide sequencing with high-throughput platforms. The results from the pilot phase provided robust protocols for whole-genome and target sequence data generation and validated algorithms to detect variants for each project design tested. Additionally it has been shown, that low-coverage sequencing represents an efficient approach to detect variation genome wide, while targeted sequencing (like exon regions) offers an efficient approach to detect and accurately genotype rare variants in regions of functional interest (The 1000 Genomes Project Consortium, 2010).

1.6. Levels of Gene Expression Regulation

In order to keep cellular life going an organism has to be able to recognize and react to internal and external signals. The appropriate answer on such stimuli is ensured by an expression of specific genes, controlled by the transcriptional regulatory system. Therefore, this regulatory system plays a central role in administering biological processes. Accordingly it is quite obvious, that the disturbance or even breakdown of this system results in an abnormal phenotype or disease.

A rough classification reveals two major levels of gene expression regulation: transcriptional regulation and post-transcriptional regulation (see Figure 4). For each class certain mechanisms can be assigned, accompanied by different key players. Transcriptional regulation applies on the level transcription for example due to the activity of transcription factors resulting in differentially expressed genes. However, even if transcription as the necessary first step in gene expression is undeniable, this does not imply that transcriptional regulation has the largest effect on the final concentration of the active gene product, which has a higher impact on the phenotype (Krol, Loedige, & Filipowicz, 2010).

Post-transcriptional regulation works on the stability and distribution of the produced transcript involving for example short RNA molecules, termed as micro RNAs (miRNAs, miR). Besides miRNAs there are many other mechanisms of post-transcriptional gene regulation not being introduced here, including: cell signaling, mRNA splicing, polyadenylation and localization, mechanisms of protein localization, modification and degradation.

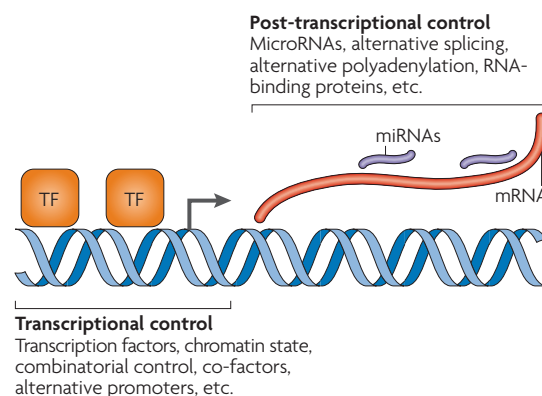


Figure 4: Gene regulation by transcription factors and microRNAs. One or more transcription factors activate transcription by binding to *cis*-regulatory sites, which are often, although not always, situated upstream of protein-coding genes. After transcription, one or more microRNAs bind to *cis*-regulatory sites, usually in the 3' UTR of the messenger RNA (mRNA), and repress protein translation. Taken from (Chen & Rajewsky, 2007)

Epigenetic modifications through chromatin remodeling and covalent modifications of DNA is the third class of gene expression regulation. These processes do not directly interfere with

the transcriptional mechanism, but regulate gene expression on a higher level like chromatin accessibility or gene silencing.

Following sections will concentrate on those entities exposed to strong research efforts, including transcription factors, miRNAs, chromosome remodeling, DNA methylation and histone modification.

1.6.1. Transcriptional Regulation by Transcription Factors

Transcription Factors (TF) are a group of proteins that is essential in the transcriptional machinery. Together with the RNA polymerase and a multiprotein complex called mediator, functioning as a co-activator, general TFs (GTFs, e.g. TFIIB) form the basal transcriptional apparatus that initiates the transcription. In addition to GTFs, a battery of more specific transcription factors exist, directing the transcription initiation to specific promoters (Vaquerizas, Kummerfeld, Teichmann, & Luscombe, 2009). The function of these TFs is to activate (or inhibit) the transcription of DNA, coding for specific target genes. In this regard they bind via a defined DNA binding domain to specific DNA sequences (see Figure 5). These interacting DNA sequences usually range in size from 8-21 bp (Wang et al., 2012). The binding affinity between these highly conserved DNA sequences and the TF binding domain is up to 10^6 -fold higher than to the remainder DNA strand (Phillips, 2008).

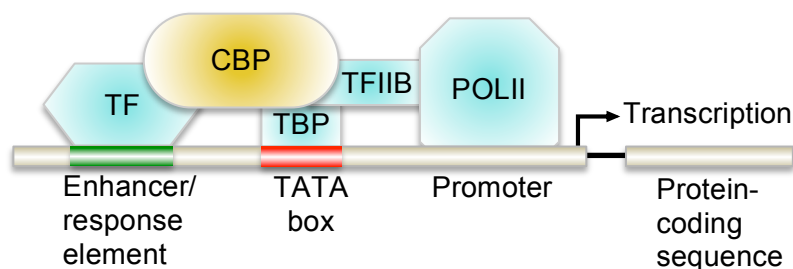


Figure 5: Simplified model of a complex for POLII catalysed transcription. A bridging protein such as CREB-binding protein (CBP) closely contacts sequence-specific TFs, the TATA box-binding protein (TBP) and the general transcription factor IIB (TFIIB). The latter does not contact DNA directly but complexes with POLII. The CBP-connected TF binds DNA at TF-specific sequence motifs. Adapted from (Tata, 2002).

These transcription factor binding sites (TFBS) – also called *cis*- or response-elements - are either distal (see below) or proximal to the genes, whose expression is regulated by the respective TF. Depending whether the TF is activating (enhancing) or inhibiting the expression of its target gene, these *cis*-elements are annotated as enhancer or silencer. The genomic region comprising these TFBS and the transcription machinery is called promoter region. In

complex multicellular organisms, transcription factors generally do not work in isolation, but instead, together with co-regulators. They form large networks of cooperating and interacting transcription factors (Krol *et al.*, 2010). In case of cooperative interactions between TFs, the respective TFBS are represented by clusters of sites within *cis*-regulatory modules (CRMs).

In order to interact with distal *cis*-elements, the transcriptional machinery can be engaged in a loop structure allowing for interaction between involved proteins and the distant binding sites. “While not long ago a distance of 50kb has been considered as an unreasonable large distance for long range enhancer, today such a distance can be classified as medium-range” (Osborne *et al.*, 2004; Wright, Brown, & Cole, 2010).

Besides *cis*-regulatory mechanisms often mediated by *cis*-regulatory elements as described above also *trans*-regulatory changes can result in differences in gene expression. In general *trans*-regulatory elements are factors as proteins, which can modify the expression of a distant gene. For example a TF regulating the expression of a gene on chromosome 8 might be transcribed by itself from a gene on chromosome 21, or more general *trans*-regulatory modification can alter the activity or expression levels of TFs (Chang *et al.*, 2008). Examples for *trans*-elements might be TFs, but also insulators or small interfering RNA. Furthermore, *cis*-regulatory elements have an allele-specific effect on gene expression, while *trans*-elements affect the regulation of both alleles (Gilad, Rifkin, & Pritchard, 2008). The relative contributions of these regulatory changes are under investigation and remain to be explored in more detail.

TF functions are mainly specific to certain biological processes, as for example cell cycle control or development. Furthermore, TF target genes might be specifically needed in a certain cell type at a certain time point to react on a certain stimuli. Accordingly, the regulation of TF itself is self-evident.

Their own expression is often regulated by another TF, but if a TF is regulating its own expression, a complex regulatory mechanism like negative feedback loop can exist. A different level of regulation is the relocation to their place of action, which is the nucleus. For example within the class of nuclear receptors, some members need first to bind a ligand in the cytoplasm before they can enter the nucleus (Whiteside & Goodbourn, 1993). Since those ligands might be dependent on other signal cascades, triggered by external stimuli, this is a common way for a cell, to express a certain protein as a reaction on the environment. Other levels of TF regulations are, that they might need to bind a ligand at their signal-sensing domain to be activated, need to be covalently modified (e.g. phosphorylated) or need to form homo- or heterodimers with other TFs. An additional regulation of the TF activity takes place

on the level of the chromatin structure or in other words, the accessibility of the DNA sequence (see section 1.6.3.1).

In summary (see Figure 6) the regulation of transcription is determined by “the interactions between TFs that bind *cis*-regulatory elements in DNA, additional co-factors and the influence of chromatin structure” (Wasserman & Sandelin, 2004).

The initial human genome sequencing projects (Lander *et al.*, 2001; Venter *et al.*, 2001) estimated ~200-300 genes being involved in the basal transcriptional machinery and up to 2,000-3,000 sequence specific DNA-binding TFs (Vaquerizas *et al.*, 2009). In 2012 the biological process of “sequence-specific DNA transcription factor activity” in the Gene Ontology (GO) database (Gene & Consortium, 2000) identifies 2,107 gene products, wherein only 244 are assigned with an experimental evidence code. ~2,200 genes code for TFs, which corresponds to nearly 10% of genes in the whole Human Genome coding for transcription factors. The resources for TF data in form of public databases are numerous, wherein one of the best established and with high level of curation but commercial, is TRANSFAC® (Wingender *et al.*, 2000).

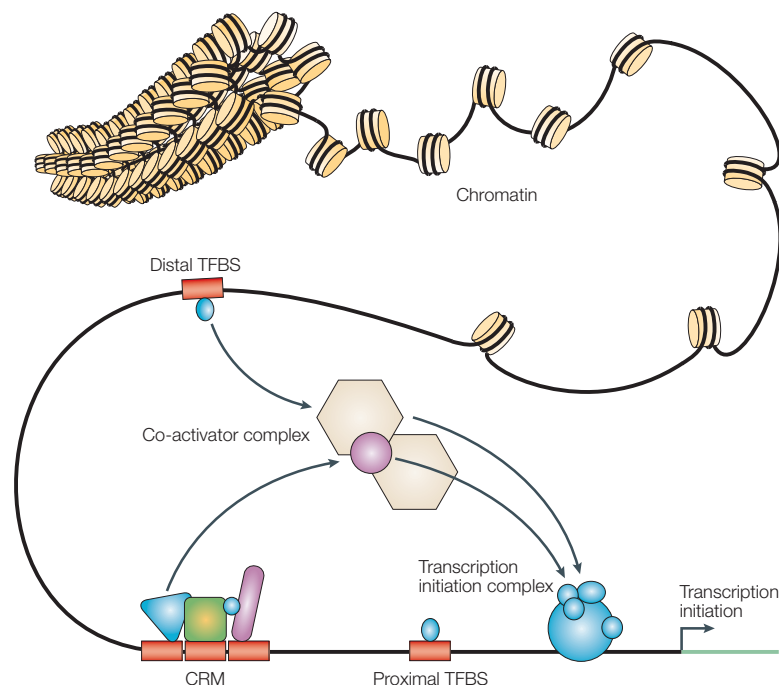


Figure 6: Components of transcriptional regulation. TFs bind to TFBS that are either proximal or distal to a transcription start site. Sets of TFs can operate in functional CRMs to achieve specific regulatory properties. Interactions between bound TFs and cofactors stabilize the transcription-initiation machinery to enable gene expression. The regulation that is conferred by sequence-specific binding TFs is highly dependent on the three-dimensional structure of chromatin. Taken from (Wasserman & Sandelin, 2004)

1.6.2. Post-transcriptional Regulation by miRNAs

In 1993, a new player in gene regulation has been identified: miRNAs. They have been discovered by Ambros and colleagues, Rosalind Lee and Rhonda Feinbaum who performed investigations on *C. elegans* larval development. They found that the *lin-4* gene did not code for a protein but for a pair of small RNAs (Lee, Feinbaum, & Ambros, 1993). One molecule was approximately 22bp long, while the other had an approximate length of 61nt. The longer molecule had been considered to be the precursor of the shorter one being predicted to form a stem loop.

Comparable with non-coding sequences in the middle of the last century, these RNAs were believed to play a minor role. Seven years later, as another miRNA, namely *let-7*, another gene in the *C. elegans* heterochronic pathway, encoding a further ~22nt RNA (Bartel, 2004), was discovered. Shortly after, cloning experiments from several labs with flies, worms and human cells, reported a total of over one hundred additional genes for tiny non-coding RNAs (Lagos-Quintana, Rauhut, Lendeckel, & Tuschl, 2001; Lau, Lim, Weinstein, & Bartel, 2001; Lee & Ambros, 2001).

Today it is known that as a class, miRNAs constitute about 1%–2% of genes in worms, flies, and mammals (Bartel, 2009). More than 2,000 mature miRNAs have now been described in humans (miRbase v.19). In the meantime, the biogenesis of miRNAs has been elucidated (see Figure 7).

miRNAs function in gene regulation as repressors of the post-transcriptional control (Chen & Rajewsky, 2007) usually by binding to the 3' UTR of their targeted mRNA. The so-called “seed region” of miRNAs (7-8 nt at the 5' end of the miRNA) interacts with the short complementary sequence in the target mRNA. The interaction between the miRNA and its target mRNA is mediated by the RNA-induced silencing complex (RISC) containing Dicer and other proteins like members of the Argonaute (Ago) protein family. The Ago proteins are central to the RISC function. They bind the mature miRNA and orientate it for the interaction with its target mRNA (Pratt & MacRae, 2009). As a consequence miRNAs can repress the initiation of translation or lead to direct degradation of their mRNA targets (see Figure 7).

The fact that only 7-8 nt are required for miRNA-mRNA interactions results in the assumption that a single miRNA can regulate many different genes (Ghildiyal & Zamore, 2009). More than 60% of protein-coding genes are computationally predicted to be miRNA targets based on conserved base-pairing between the mRNA 3' UTR and miRNA seed regions (Friedman, Farh, Burge, & Bartel, 2009).

Even though most miRNAs and their target binding sites are highly conserved, which suggests important biological functions, a typical miRNA- target interaction results in a relatively low reduction (<2-fold) in protein levels, and according to Sharp *et al.*, “many miRNAs can be deleted without creating any obvious phenotype” (Ebert & Sharp, 2012).

Another role, not as extreme as the loss- or gain-of-function idea, is assumed with regard to robustness, which is “the ability of a system to maintain its function in spite of internal or external perturbations” (Kitano, 2004). In comparison to “classical” transcriptional regulators, miRNAs are executing their regulatory activity in the cytoplasm. Thereby they can intervene later in the pipeline of gene expression to counteract variation from the upstream processes of transcription, splicing, and nuclear export (Ebert & Sharp, 2012).

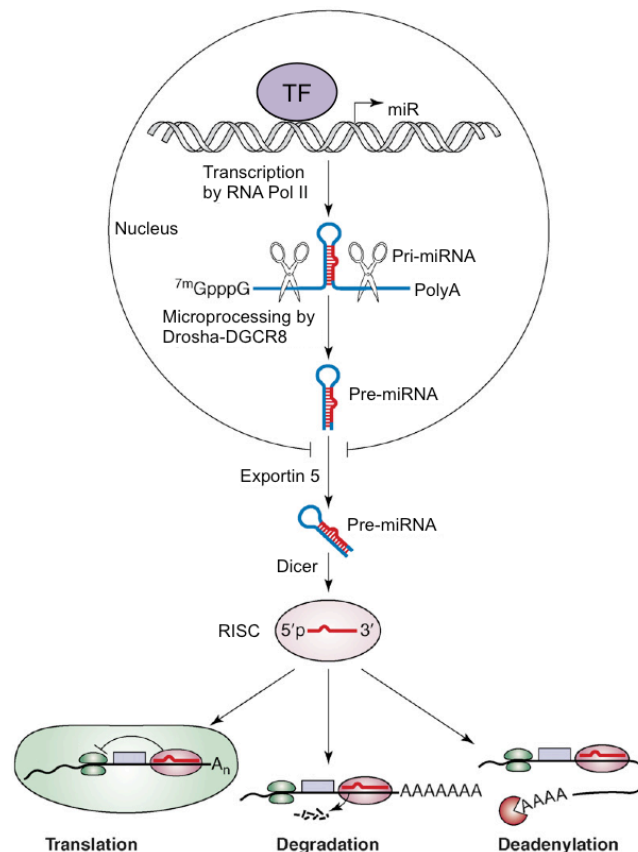


Figure 7: Schematic representation of miRNA biogenesis and function. The initial RNA is typically transcribed by POLII as primary miRNAs (pri-miRNAs), which range from a few hundred to thousands of nucleotides (nt) in length. The pri-miRNA of each miRNA has a characteristic stem-loop structure that can be recognized and cleaved by the ribonuclease III (RNase III) endonuclease Drosha within the nucleus. Efficient pri-miRNA cleavage by Drosha requires a protein partner, DGCR8 (DiGeorge syndrome critical region gene; also known as Pasha), which has a double-stranded RNA-binding domain (dsRBD). The cleavage product, a ~70-nt stem-loop pre-miRNA, is exported from the nucleus to the cytoplasm by Exportin 5. In the cytoplasm, another evolutionarily conserved RNase III enzyme, Dicer, together with its dsRBD protein partner, trans-activation response (TAR) RNA-binding protein (TRBP) and PKR (RNA-dependent protein kinase)-activating protein (PACT), further process pre-miRNA into mature miRNA (~21 nt). The mature miRNA is then unwound and a single strand is incorporated into RISC SRF: serum response factor; TF: transcription factor. Taken and adapted from (Yong Zhao & Srivastava, 2007)

In 2004, a registry has been set up to catalog the miRNAs and facilitate the naming of newly identified genes (Griffiths-Jones, 2004), still serving as the main comprehensive resource of microRNA target predictions and expression profiles (<http://www.mirbase.org/>). The current release 19 (August 2012) contains 1,600 precursor and 2,042 mature miRNA entries.

As in all “omics” research fields, also here, next generation sequencing is being increasingly used to profile small RNA expression across the stages of development and in different tissues and disease states. Profiling by deep sequencing provides not only quantitative information about small RNA expression levels, as would quantitative PCR (qPCR) or microarray-based approaches, but it can also precisely detect subtle changes in small RNA sequence or length (Ghildiyal & Zamore, 2009).

1.6.3. Epigenetic Modifications

In 1939 Conrad Waddington created the term “epigenetics” pointing to molecular mechanisms that convert this genetic information into observable traits or phenotypes (Van Speybroeck, 2002). It has been known for a long time, that even if all cells in an organism share the same genetic information, not every gene is active in each cell at each time. Epigenetic gene expression patterns and associated phenotypes are conserved through mitosis or even meiosis, although no change in the primary DNA sequence has occurred. Thus, epigenetics is generally understood to be “the study of mechanisms that control gene expression in a potentially heritable way” (Portela & Esteller, 2010).

1.6.3.1. Nucleosome Positioning and Chromatin Remodeling

In order to fit in the nucleus, DNA has to be packed in a highly condensed manner (see Figure 8). To achieve this, DNA is wound around protein complexes, themselves composed of a histone octamer. These structures are called nucleosomes, forming a complex called chromatin. Histones are positively charged which allows them to associate with the negatively charged DNA. Each octamer is composed of the histone proteins H2A, H2B, H3 and H4, grouped into two H2A-H2B dimers and one H3-H4 tetramer to form the nucleosome. The core histones are predominantly globular except for their N-terminal tails, which are unstructured (Kouzarides, 2007). After a chain of nucleosomes is wrapped into a 30nm spiral, a fifth histone, namely H1 (called linker histone), is involved to maintain the chromatin structure (see Figure 8).

As known now, this complex does not only serve as a way to condense DNA, but also controls the usage of DNA. Genes can only be expressed, if their structural elements are accessible for the RNA polymerase as well as transcription factors. The default status of chromatin is tightly coiled, being therefore not accessible. This in turn means, that chromatin has to open to allow gene expression, an event called chromatin remodeling modification (Phillips & Shaw, 2008). In fact the condensed DNA packing into nucleosomes appears to affect all stages of transcription, thereby regulating gene expression. Additionally they have been reported to play an important role in shaping the methylation landscape (Chodavarapu *et al.*, 2010).

Nucleosomes are a kind of bulwark, blocking the access of activators and transcription factors to their DNA binding sites and at the same time inhibiting the elongation of the transcripts by engaged polymerases. The precise position of nucleosome around the transcription start site (TSS) has a strong impact on the initiation of transcription. At any genomic locus a preferential positioning of nucleosomes can be described and slight nucleosome shifts (~30bp) at TSS results in changes in the activity of RNA polymerase II (POLII). The 5' and 3' ends of genes show regions free of nucleosomes to provide space for the assembly and disassembly of the transcription machinery. The loss of a nucleosome directly upstream of the TSS is tightly correlated with gene activation, whereas the occlusion of the TSS by a nucleosome is associated with gene repression (Cairns, 2009; Schones *et al.*, 2008).

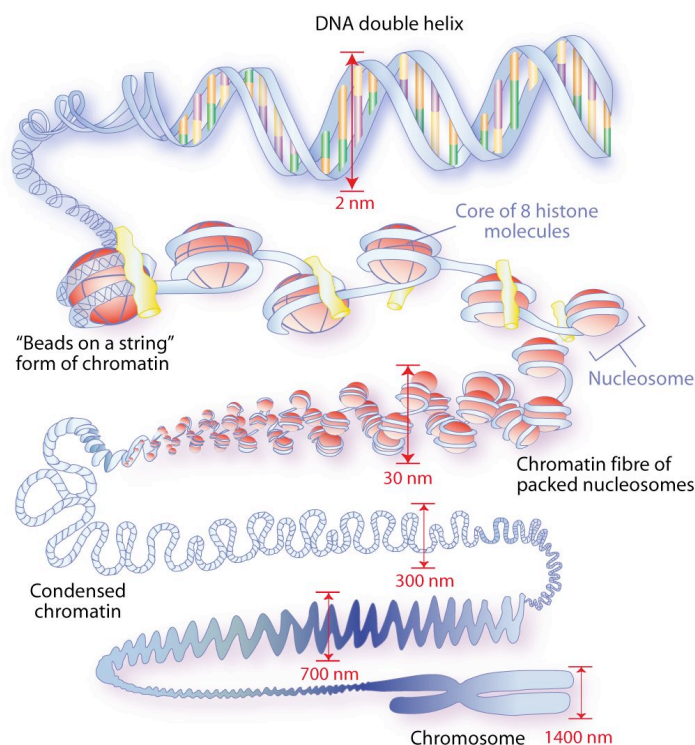


Figure 8: Illustration of different “packing” levels of DNA (taken from <http://pleasanton.k12.ca.us/avhsweb/kawashimae/video/chp5and8.html>)

The factor influencing the nucleosome positioning and thereby the gene expression is the composition of histones incorporated in a nucleosome. Accordingly, histone variants are finally the main regulators (Zilberman, Gehring, Tran, Ballinger, & Henikoff, 2007). Furthermore, DNA methylation and specific histone modifications have been linked with the nucleosome remodeling machinery (Wysocka *et al.*, 2006). miRNAs have also been shown to play a role in histone variant replacement (Lal *et al.*, 2009; Yoo, Staahl, Chen, & Crabtree, 2009).

Several groups of large macromolecular complexes are known to move, destabilize, eject or restructure nucleosomes in an Adenosine-5'-Triphosphate (ATP) hydrolysis-dependent manner. These complexes, known as chromatin remodeling complexes or chromatin remodeler, can be classified into four families (SWI/SNF, ISWI, CHD and INO80) that share similar ATPase domains but differ in the composition of their unique subunits (Ho & Crabtree, 2010).

SWItch/Sucrose NonFermentable (SWI/SNF) family complexes are master regulators of gene expression. They regulate the expression of numerous genes and have been reported to modulate alternative splicing. Additionally, they show the unique ability to restructure the nucleosome, removing the H2A-H2B dimers and replacing them with H2A.Z-H2B dimers, resulting in a protection against DNA methylation (Clapier & Cairns, 2009; Zilberman, Coleman-Derr, Ballinger, & Henikoff, 2008).

Many members of the second class, the Imitation Switch (ISWI) family, have been reported to promote chromatin assembly and to repress transcription. In the chromodomain helicase DNA-binding (CHD) family, some members participate in the sliding and ejection of nucleosomes, promoting transcription (Portela & Esteller, 2010). Finally members of the fourth group, the INO80 family, have been reported to participate in multiple cellular processes: transcriptional activation, DNA repair, telomere regulation, chromosome segregation and DNA replication among others (Ho & Crabtree, 2010).

1.6.3.2. DNA Methylation

DNA methylation, or more precisely the methylation of cytosine nucleotide, is the most studied epigenetic modification (see Figure 9). It occurs almost exclusively in the context of CpG dinucleotides, which tend to cluster to so-called CpG islands. These islands, accounting for ~1% in mammalian genomes, are defined as regions with a length of more than 200nt holding a GC-content of at least 50% and a C/G-ratio of at least 0.6. About 60% of human genome promoters are associated with CpG islands, being usually unmethylated. Their methylation is in general connected to gene silencing and plays a key role in genomic imprinting, meaning

that a hypermethylation at one of the parental alleles leads to monoallelic expression (Portela & Esteller, 2010).

Gene expression can be influenced by DNA methylation on different levels. For example methylated DNA can promote the recruitment of methyl-CpG-binding domain (MBD) proteins, which in turn recruit histone- modifying and chromatin-remodeling complexes to methylated sites (Esteller, 2007; Lopez-Serra & Esteller, 2008). Furthermore, the recruitment of DNA binding proteins can be directly precluded, resulting in an inhibited transcription (Kuroda *et al.*, 2009). Unmethylated CpG islands for instance are generating a chromatin structure being favorable for gene expression by recruiting proteins associating with histone methyl transferases, creating domains rich in the histone methylation mark H3K4 trimethylation (H3K4me3, see below) (Thomson *et al.*, 2010).

Recent research revealed that most of the tissue-specific DNA methylation seems to occur not at CpG islands but at so-called CpG island shores, regions lying in close proximity (~2 kb) of CpG islands, also being associated with transcriptional inactivation (Doi *et al.*, 2009).

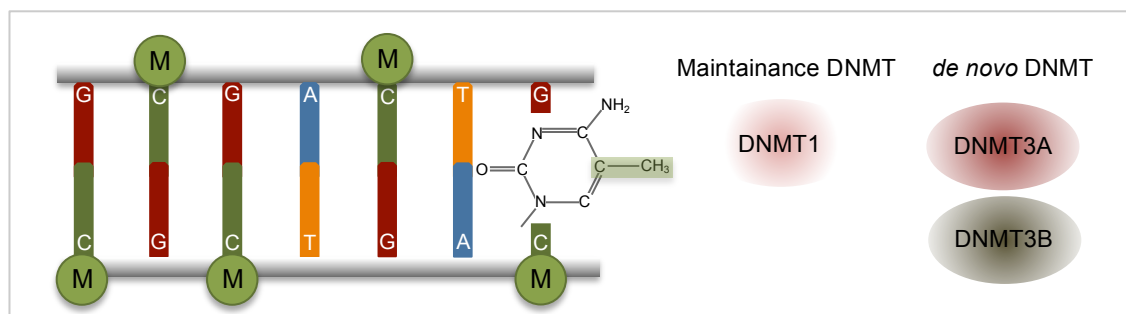


Figure 9: DNA methylation is the addition of a methyl group (M) to the DNA base cytosine (C). Different Enzymes are involved in DNA methylation (DNMTs) either for maintenance or *de novo*.

DNA methylation is mediated by the DNA methyltransferase (DNMT) family, catalyzing the transfer of a methyl group from S-adenosyl methionine to DNA. From five reported members of the DNMT family in mammals, only three, namely DNMT1, DNMT3a and DNMT3b possess methyltransferase activity. These are classified into *de novo* DNMTs (DNMT3A and DNMT3B) and maintenance DNMTs (DNMT1). While DNMT3A and DNMT3B are thought to be involved in the establishment of methylation patterns during embryonic development (Esteller, 2007), DNMT1 has a 30- to 40-fold preference for hemimethylated DNA, having also a *de novo* DNMT activity (Portela & Esteller, 2010).

1.6.3.3. Histone Modification

In the year 2000 a first formally proposition with regard to the possible impact of histone modifications has been made by Brian Strahl and David Allis (Strahl & Allis, 2000) - “*We propose that distinct histone modifications, on one or more tails, act sequentially or in combination to form a ‘histone code’ that, read by other proteins to bring about distinct downstream events.*” Twelve years passed and their proposition became a major research field to understand the different levels of gene regulation.

All histones are highly conserved and affected by post-transcriptional modification (see Figure 10). Most of these modifications occur in the histone tails and have an important role in transcriptional regulation, DNA repair, DNA replication, alternative splicing and chromosome condensation. The most prominent ones are acetylation, methylation, phosphorylation, ubiquitination, SUMOylation and ADP-ribosylation (Kouzarides, 2007; Rando & Chang, 2009) (see Figure 12).

Depending on the transcriptional state, the human genome can be subdivided in two main classes: euchromatin, which is actively transcribed and heterochromatin being transcriptionally inactive. Euchromatin is characterized by high levels of acetylation and trimethylated H3K4, H3K36 and H3K79, while heterochromatin shows low levels of acetylation and high levels of H3K9, H3K27 and H4K20 methylation (Li, Carey, & Workman, 2007).



H3K27ac, H2BK5ac and H4K20me1 in the promoter and H3K79me1 and H4K20me1 along the gene body (Karlić, Chung, Lasserre, Vlahovicek, & Vingron, 2010).

Beside covalent post-transcriptional modification as described above, a new type of modification has been found recently (Santos-Rosa *et al.*, 2009), representing the first massive clearing of histone marks, by clipping the histone tail of H3 after the Ala21 residue. This leads to a loss of the remaining 21 N-terminal residues associated with post-transcriptional modifications. Histone modification represents a very complex level of gene expression regulation, characterized by the fact that histones can be modified at different sites simultaneously, accompanied by cross-talk (see Figure 11). This communication in turn can occur within the same site, the same histone and among different histones, so that a single histone mark does not determine transcriptional regulation alone. It is rather the contrary, namely the combination of all marks in a nucleosome or region that specifies the outcome (Portela & Esteller, 2010). In 2010 Ernst and Kellis published a study revealing up to 51 distinct 'chromatin states' based on the enrichment of specific combinations of histone modifications (Ernst & Kellis, 2010).

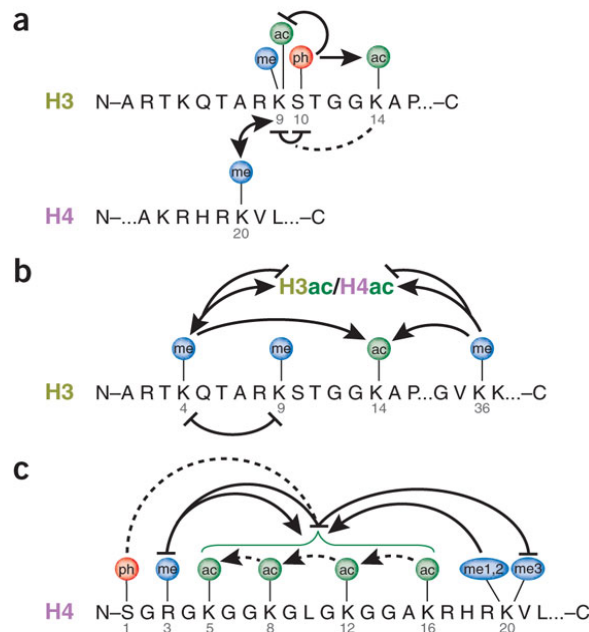


Figure 11: Cross-talk between H3K9, H3S10 and H3K14. (b) Cross-talk with H3K4me. H3ac/H4ac refers to acetylation of H3 and H4 at multiple sites. (c) Cross-talk on the H4 N-terminal tail. Dotted lines connecting modifications indicate possible cross-talk. me1, monomethylation. (taken from (Latham & Dent, 2007))

Several histone methyltransferases (HMT) have been reported to direct DNA methylation to specific genomic targets by recruiting DNMTs, supporting to set the silenced state established

by the repressive histone marks. HMTs and histone demethylases (HDMs) can also regulate DNA methylation levels, by directly modulating the stability of DNMT proteins.

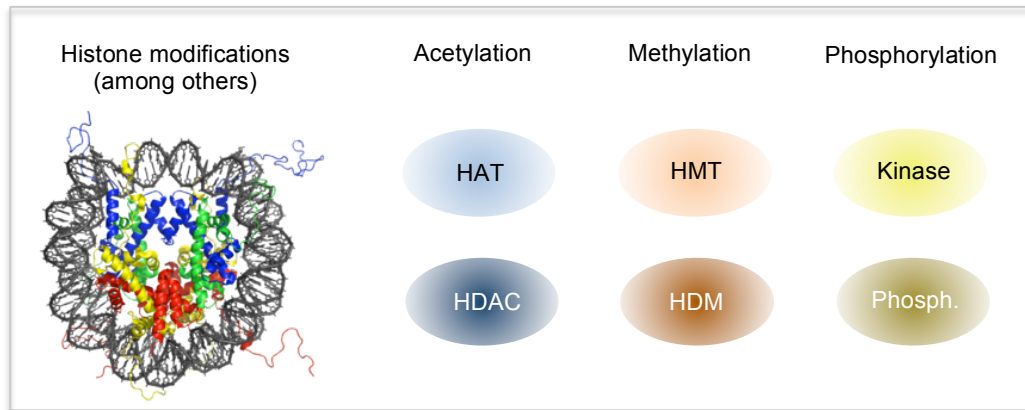


Figure 12: Epigenetic regulation depends on the interplay among the different players: histone marks (adapted from (Portela & Esteller, 2010))

In turn DNA methylation can also directly affect histone modifications. For example methylated DNA mediates H3K9me through MeCP2 (member of the methyl-CpG-binding domain proteins) recruitment. MeCP2 is also known to silence gene expression partly by recruiting the histone deacetylase (HDAC) repressive machinery, which removes acetyl groups from histones resulting in gene silencing (Fuks *et al.*, 2003).

HMTs, HDMs and kinases (phosphorylation) are those from the histone modifying enzymes being most specific to individual histone subunits and residues (Kouzarides, 2007), while most of the histone acetyltransferases (HATs) and HDACs are not highly specific and modify more than one residue. Certainly, the list of histone modifications and related enzymes is probably not complete yet.

1.7. Next Generation Sequencing

Within the last years a fundamental shift from the automated Sanger sequencing (also called “first generation technology”) to methods referred as next generation sequencing (NGS) can be observed.

Methods like gene-expression microarrays are more and more replaced by “seq-based methods”⁶ allowing identification and quantification of rare transcripts without the prior knowledge about their sequence, usually necessary to design tools like microarrays (Wold & Myers, 2008).

The new technologies join template preparation, sequencing and imaging, genome alignment and assembly methods. At the same time these technologies allow generation of enormous amounts of data at reasonable costs, with up to one billion short reads per instrument run. However, whole-genome sequencing is still hardly affordable and associated with challenges like *de novo* assembly. As an interim solution NGS technologies are currently mostly applied to target specific regions of interest, as beyond others the exome, epigenetic markers, TFBS or specific gene families (Metzker, 2010).

1.7.1. Technology

In the meantime several different platforms have been developed all combining specific protocols in a certain manner determining the data produced. These achievements came with new challenges regarding the comparison of results from different platforms: although all manufacturers provide quality scores and accuracy estimates, there is no consensus that the ‘quality standards’ from one platform is equivalent to that from another platform.

A presentation of all platforms in detail would be out of the scope of this work. Therefore the introduction is limited to one of the technology platforms used in 1000G, the Solexa platform from Illumina Inc.⁷.

1.7.1.1. Template Preparation

In context of Chromatin Immunoprecipitation sequencing (ChIP-seq) (see chapter 1.7.2) a template is a recombinant DNA molecule comprising a known region like a primer or adaptor sequence. To this adapter sequence a universal primer can bind, which is attached to the unknown target to be sequenced (see Figure 13). Following the general consideration that a good output depends on a good input, robust methods to generate a representative, unbiased

⁶ Assays that use next-generation sequencing technologies. They include methods for determining the sequence content and abundance of mRNAs, non-coding RNAs and small RNAs (collectively called RNA-seq) and methods for measuring genome-wide profiles of immunoprecipitated DNA-protein complexes (ChIP-seq), methylation sites (methyl-seq) and DNase I hypersensitivity sites (DNase-seq).

⁷ http://www.illumina.com/technology/solexa_technology.ilmn

source of nucleic acid material from the sample to be investigated are essential. Methods applied currently generally involve a random fragmenting of genomic DNA into smaller sizes ($< 1\text{kb}$) used to create fragment templates.

Because most imaging systems have not been designed to detect single fluorescent events, the templates need to be amplified to detect a fluorescent signal in the imaging step. The Illumina platform is using an approach called solid-phase amplification. Here, the sample fragments are randomly distributed and clonally amplified in clusters on a glass slide producing up to 100–200 million spatially separated template clusters. The ends of the templates are free that a universal sequencing primer can be hybridized to initiate the NGS reaction.

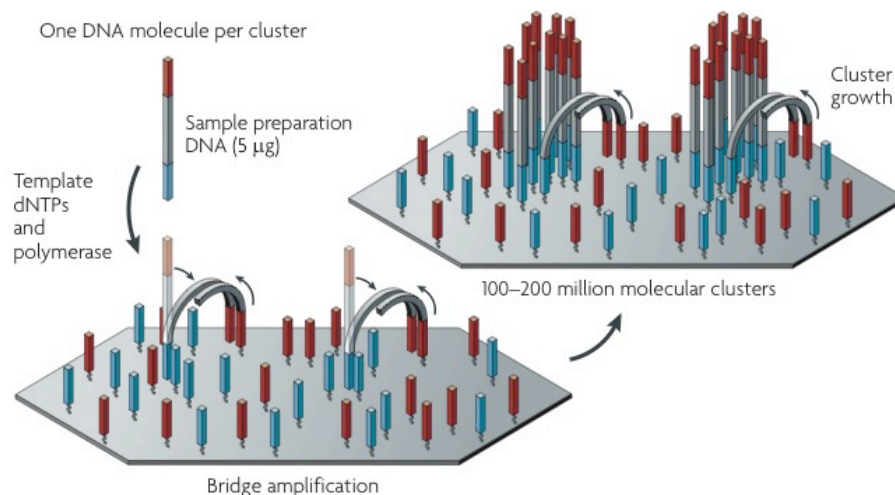


Figure 13: Solid-phase amplification is composed of two basic steps: initial priming and extending of the single-stranded, single-molecule template, and bridge amplification of the immobilized template with immediately adjacent primers to form clusters. (Taken and adapted from (Metzker, 2010))

The signal detection is based on a so-called four-color cyclic reversible termination (CRT). This method is based on cyclic nucleotide incorporation using reversible terminator, in each step followed by a fluorescence image and terminator cleavage (Metzker, 2010). Each base is assigned by a different color (see Figure 14).

In the initial step, a DNA polymerase, which is bound to the primed template, adds one nucleotide being fluorescently modified and representing the complement of the template base. Next, the free, labeled nucleotides are washed away and the imaging takes place, taking a snapshot to record the identity of the single nucleotide added. Then, the labeled nucleotide, which is terminating the reactions and which inhibits further attachment, is cleaved.

Before the next cycle, which adds the next nucleotide, starts, another washing step is performed. The sequence can be iteratively derived from the color-coding for each cluster in each cycle.

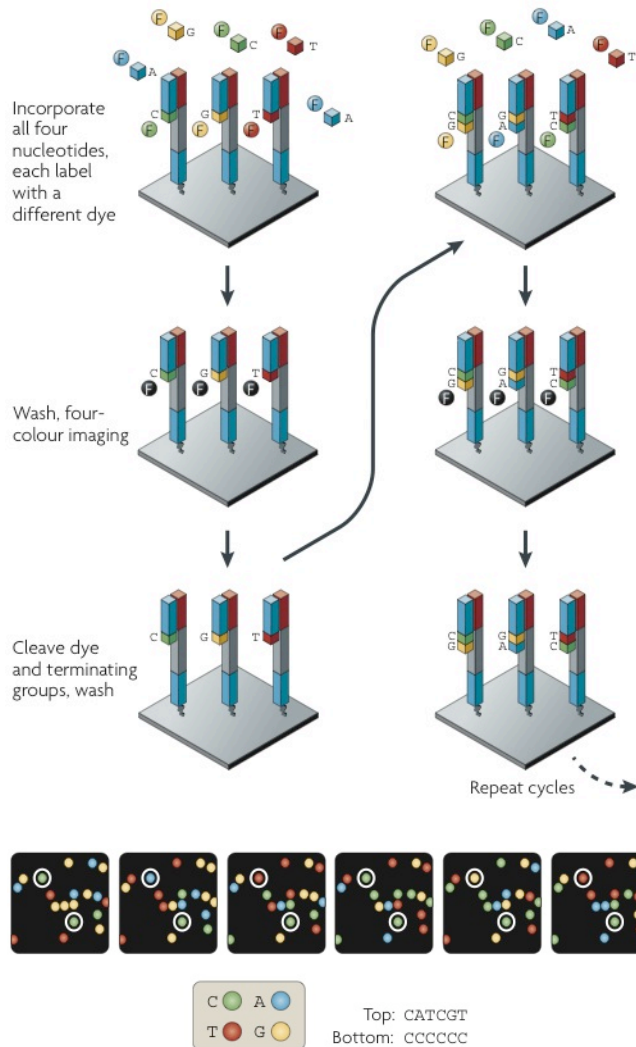


Figure 14: The four-color CRT method used by Illumina/Solexa using solid-phase-amplified template clusters (shown as single templates for illustrative purposes). After imaging, a cleavage step removes the fluorescent dyes and regenerates termination site with a reducing agent. The four-colour images highlight the sequencing data from two clonally amplified templates (Taken and adapted from (Metzker, 2010))

1.7.1.2. Genome Alignment and Assembly

After NGS reads have been generated the experimental part is completed and bioinformatics analysis takes over. In order to gather the sequence data all reads need to be put in order. There are two general possibilities to achieve this: either by alignment to a reference genome or by *de novo* assembly (Chaisson, Brinza, & Pevzner, 2009; Pop & Salzberg, 2008; Trapnell & Salzberg, 2009). Which strategy is chosen depends on the research question as well as cost, required effort and time considerations. For example, identifying and cataloging genetic variations in multiple strains of highly related genomes can be accomplished by aligning NGS reads to their reference genomes. However, alignment has its limitations, for example in the case of repetitive regions a read cannot be unambiguously aligned. Another problem are

genome regions in the target genome not being existent in the reference genome, due to the presence of structural variants (SV) or gaps in the reference genome. Nevertheless, apart from the trend directed to personal genomes, the sequencing of big sample sizes as for example in context of 1000G might help to overcome these limitations (Rosenfeld, Mason, & Smith, 2012).

With regard to human genomes substantial challenges exist for *de novo* assembly, resulting in a still rather rare application. Metaphorically speaking one can imagine the target genome as a puzzle to be solved without knowing the picture to be reconstructed. With regard to mathematics, *de novo* assembly represents an NP-hard problem, which is a computational problem for which no efficient solution is known (Pop & Salzberg, 2008). “*De novo* assembly requires overlapping pairs, and for example $6 \cdot 10^9$ reads as an output of a sequencing experiments would require the consideration of $3.6 \cdot 10^{19}$ potentially overlapping reads” (Rahmann, 2011). Repetitive sequences are also here a major issue, in particular when they are longer than the length of a read. Beside time resources and the need of specialized hardware other aspects like data transfer, programming tools and storage needs have to be addressed.

1.7.2. Application of NGS - Chromatin Immunoprecipitation Sequencing

Currently numerous applications of NGS exist. One of the early ones, ChIP-seq, is applied to detect genome wide maps for transcription factor binding or histone modification to analyze chromatin structure, revealing regions being generally transcriptionally active or inhibited (Mikkelsen *et al.*, 2007). Besides ChIP-seq other applications emerged, like Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) sequencing or DNase-seq (Giresi, Kim, McDaniel, Iyer, & Lieb, 2007; Simon, Giresi, Davis, & Lieb, 2012; Song *et al.*, 2011) directed to the genome-wide association of accessible DNA regions, or the sequencing of mRNA (RNA-seq) for gene expression profiling (Kim *et al.*, 2007; Nagalakshmi *et al.*, 2008; Wilhelm *et al.*, 2008).

In the following ChIP-seq as an example application will be described in more detail (see Figure 15). The investigation of protein-DNA interactions and epigenetic marks (see section 1.6.3) can provide essential knowledge to understand transcriptional regulation. The mapping of TFBS, meaning the binding of the transcriptional machinery and other involved DNA-binding proteins on a genome-wide level can help to decipher the TFBS landscape (Park, 2009). The currently established tool to achieve this information is chromatin immunoprecipitation (ChIP), a general technique for assaying protein-DNA binding *in vivo* (Solomon, Larsen, & Varshavsky, 1988). The basic principle is that antibodies are used to selectively elute DNA-fragments specifically bound by proteins like TFs. Due to recent

developments of NGS technologies, the experimental setup changed from a microarray based approach, called ChIP-chip, to ChIP-seq. In ChIP-chip, the DNA fragments obtained from ChIP are hybridized to a microarray, enabling a genome-wide scale view of DNA-protein interactions (Ren *et al.*, 2000).

However, even if those arrays can be tiled at a high density, the required number of probes is tremendous. ChIP-seq combines genome-wide sequencing with ChIP, resulting in a tool able to sequence tens or hundreds of millions of short DNA fragments in a single run, providing specific genomic regions of interest in base pair resolution (Park, 2009). Its usage was published for the first time in 2007 (Barski *et al.*, 2007; Johnson, Mortazavi, Myers, & Wold, 2007; Mikkelsen *et al.*, 2007; Robertson *et al.*, 2007).

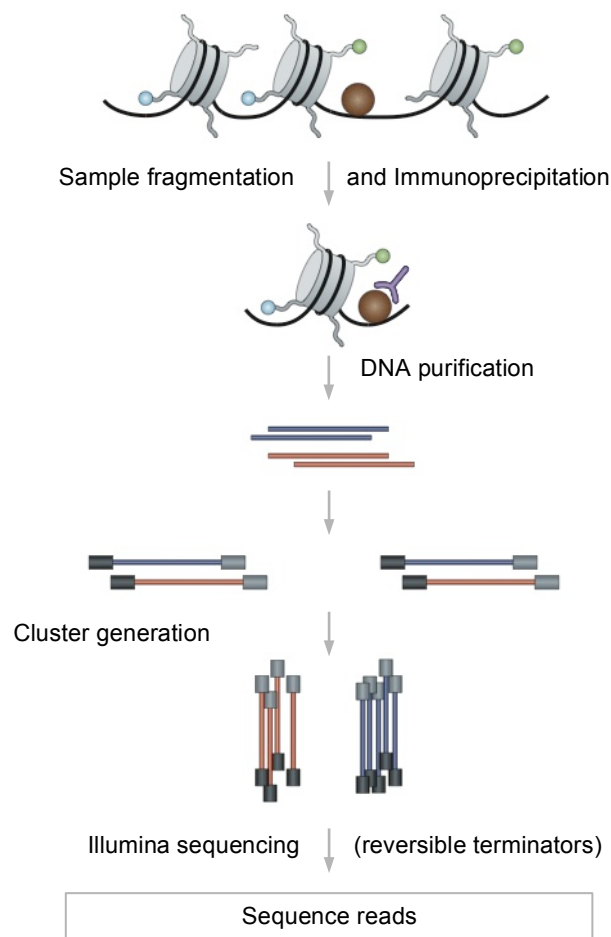


Figure 15: Schematic overview ChIP-seq experiment (illustrated for the case of transcription factor binding investigation). The ChIP process enriches the cross-linked proteins or modified nucleosomes of interest using an antibody specific to the protein or the histone modification. Purified DNA can be sequenced on any of the next-generation platforms. The basic concepts are similar for different platforms: common adaptors are ligated to the ChIP DNA and clonally clustered amplicons are generated. The sequencing step involves the enzyme-driven extension of all templates in parallel. After each extension, the fluorescent labels that have been incorporated are detected through high-resolution imaging. (Adapted from (Park, 2009))

The main difference to ChIP-chip experiments is that the enriched DNA fragments are not hybridized to an array, but directly sequenced. The information content and quality of the resulting data provide a higher resolution, fewer artifacts, greater coverage and a larger dynamic range. Furthermore, the serious problem of short read length (~35bp) in *de novo* assembly for genome sequencing are acceptable for ChIP-seq (Park, 2009).

In a ChIP-seq experiment for TFBS mapping, the DNA fragments associated with the TF are enriched. By treating cells with formaldehyde the DNA is cross-linked *in vivo* with the TF-protein, followed by sonication shearing, resulting in fragments that show a general size from 200-600bp. Then, an antibody being specific to the TF-protein is added to immunoprecipitate the DNA-TF-protein-complex. Finally the crosslinks are reversed, the DNA is released and sequenced. Admittedly, each improvement might come with its certain costs. In ChIP-seq experiments sequencing errors have been strongly reduced over time, but they are still present, especially towards the end of each read. This problem can be addressed during the data analysis and preliminary quality control (see methods). Furthermore, the experiment is losing sensitivity and specificity in the detection of captured regions, if the number of reads is insufficient (Park, 2009). Besides technical issues need to be considered when the experiment is performed as for example sample preparation or loading the correct amount of sample (too little sample results in few reads, too much sample results in fluorescent labels being too close to each other leading to lower quality data).

Everything finally depends on the specificity and sensitivity of the antibody. Hence it represents a very crucial entity in a ChIP-seq experiment, demanding preliminary validation experiments, to check its reactivity in general (western blotting) and cross-reactivity.

Another very important aspect in a ChIP-seq is the control experiment, addressing the problem of artifacts generated during the experimental steps. For example the fragments, resulting from DNA shearing are not uniformly distributed over the genome, because open chromatin regions tend to be fragmented more easily than closed regions. As a consequence, the peak from a ChIP-seq profile needs to be compared to the same region in a matched control sample to determine its significance.

In this regard three types of control samples are used: a) input DNA, which is a portion of the DNA sample, removed prior to immunoprecipitation (IP), b) mock IP DNA, which is obtained from IP performed without antibody and c) DNA from non-specific IP, which is an IP performed with an antibody like immunoglobulin G, known to bind a protein which is not involved in transcription regulation on DNA or chromatin level. Each of these control samples tests for

different artifacts and until now there is no consensus which control is the appropriate one (Park, 2009) (see Figure 16).

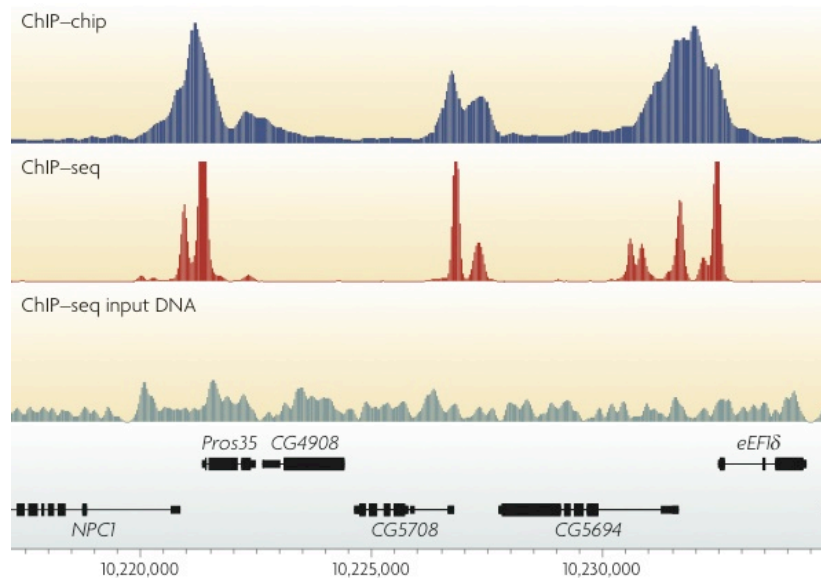


Figure 16: Example of ChIP profiles generated by chromatin immunoprecipitation followed by sequencing (ChIP-seq) or by microarray (ChIP-chip). Shown is a section of the binding profiles of the chromodomain protein Chromator, as measured by ChIP-chip (unlogged intensity ratio; blue) and ChIP-seq (tag density; red) in the *Drosophila melanogaster* S2 cell line. The tag density profile obtained by ChIP-seq reveals specific positions of Chromator binding with higher spatial resolution and sensitivity. The ChIP-seq input DNA (control experiment) tag density is shown in grey for comparison (taken and adapted from (Park, 2009))

This in turn should be considered when published ChIP-seq experiments are analyzed in a comparative manner. Input DNA is probably the mostly used control sample, correcting for errors due to variable solubility of different region, DNA fragmentation and amplification. A main issue of control experiments is that a large amount of sequencing is required, since many of the sequences tags are spread evenly across the genome. Thus, to prevent large errors of fold enrichments at the peaks due to sampling bias, accurate measurements throughout the genome are necessary with a sufficient numbers of tags at each point. Nevertheless, if one is only interested in differential binding patterns between different biological conditions or different time points and the variation in chromatin preparations is small, the sequencing of a control sample might be avoided (Park, 2009).

Beside the antibody and control experiment, also the sequencing depth has a strong impact. In contrary to ChIP-chip, the number of fragments to be sequenced is in ChIP-seq determined by the investigator. The question is, whether the sequencing depth is sufficient to detect the complete space of the regulatory element under investigation. One approach to determine a sufficient sequencing depth would be to chose the threshold for the number of reads, from which on the results do not change anymore, meaning to find a kind of saturation point. Though, studies have shown contradictory results, on the one hand, the number of sites

discovered increased continuously with additional sequencing (Kharchenko, Tolstorukov, & Park, 2008) on the other hand (for example for POLII) the signal saturated very quickly (Rozowsky *et al.*, 2009). An alternative approach is to choose a fixed threshold on the fold enrichments between the peaks in the ChIP-experiment and the peaks in the control experiment. Then, saturation occurs when only prominent peaks (defined by minimum enrichment) are considered.

Costs are not a major issue anymore. Due to the increasing number of manufacturers and the increasing demand the sequence cost per base pair is decreasing (see Figure 17). However, the analysis of ChIP-seq experiments emerges to a kind of expert knowledge, so that in most cases, the researcher performing the experiment is different from the one executing the analysis.

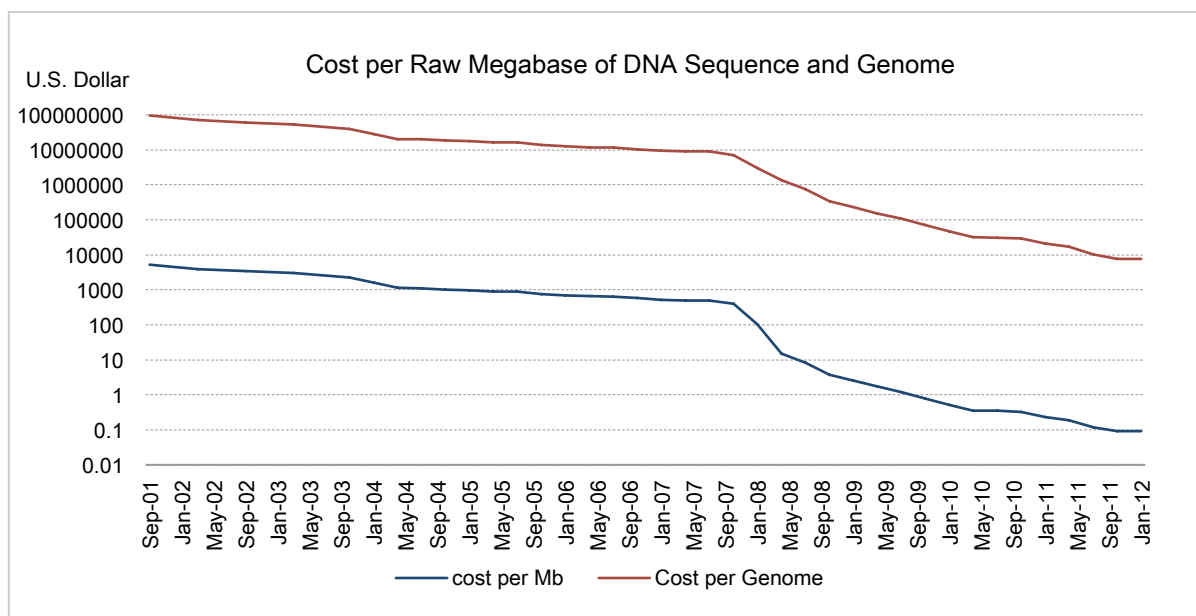


Figure 17: Cost per raw megabase of DNA Sequence and per genome (2001-2012), data taken from the NHGRI Large-Scale Genome Sequencing Program

At this point no description of the analysis of ChIP-seq experiments is provided, as this is specified in detail in methods section 3.7.6.

1.8. Modeling Transcription Factor Binding Sites

“The full understanding of the interplay between transcription factors and *cis*-sequences would revolutionize biological research, providing the means to interpret, model and modulate the response of cells to diverse stimuli. Deciphering the regulatory control mechanisms directing gene expression might enable simplified interpretation of the complex data that now flood

computers” (Wasserman & Sandelin, 2004). However, the initial step of gene expression, namely transcription, one of the most strongly studied mechanism in cellular biology does not only depend on transcription factors, but also other players like miRNAs, chromatin remodeling, histone modification and cofactors. Accordingly the elucidation of transcription initiation does not reveal the whole picture. Thus, “the mastery of the entire network of gene regulation will remain a distant hope and aspiration”, but also strong motivation (Wasserman & Sandelin, 2004). As known already TFs have in general distinct preferences toward specific DNA sequences. The sequences to which these TFs bind are very short and show considerable variability (Van Loo & Marynen, 2009).

Based on that, it is possible to construct models describing the properties of the target sequence or binding motif, to predict potential binding sites in an unknown sample or genomic sequence. In this regard two requirements are existing: on the one hand, the kind of model to develop needs to be selected and on the other hand these models need to be transformed to methods, being able to predict TFBS in an efficient way either on selected genomic regions or even genome-wide.

The methods used currently for TFBS prediction underlie certain assumptions, wherein the most strongly violated one is the so-called independence assumption. This assumption simplifies the complexity of TFBS by claiming, that the binding domain of a TF occurs independently from adjoining sequences and other proteins in its proximity. This is fundamentally incorrect. Transcriptional regulatory sequences are often composed of multiple binding sites for multiple transcription factors. By this concerted binding of a specific combination of TFs and co-factors gene regulation can be tightly controlled (Balmer & Blomhoff, 2006; Van Loo & Marynen, 2009). Furthermore, this assumption leads to a severe limitation, termed as futility theorem. This theorem describes the inability to specifically distinguish between TFBS being functional *in vivo* and those with no functional role.

The simplest binding model is the consensus sequences. Here, binding site sequences are aligned together and the nucleotide occurring mostly for a position is assigned as the consensus nucleotide (see Figure 18).

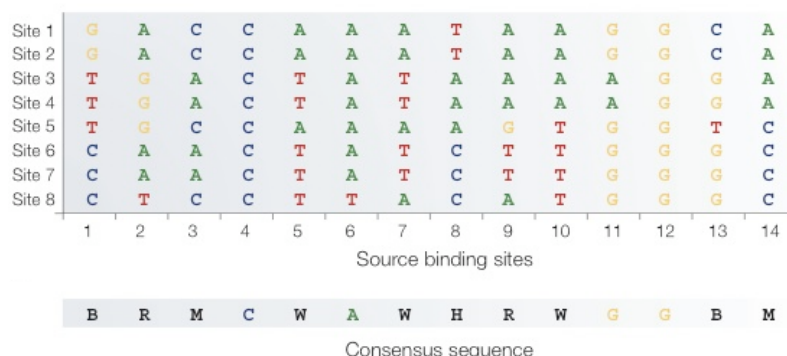


Figure 18: Consensus sequence model: a consensus sequence is defined by selecting a degeneracy nucleotide symbol for each position (column) in the alignment (upper panel). Unusual binding sites can have an extreme effect on the consensus (see, for example, site eight). (taken and adapted from (Wasserman & Sandelin, 2004))

The advantages of the consensus model are: it summarizes several sequences, it is easy to build and it allows fast visual comparison. Thus, a binding motif can be scored based on the differences to the consensus motif. The disadvantage is, that it “fails to reflect the quantitative characteristics of TF binding” (Wasserman & Sandelin, 2004).

This issue has been overcome by the next model described, which represents the most widely used one – the Position Weight Matrix (PWM). A PWM constitutes a profile providing quantitative description of the known binding sites for a TF (Stormo, 2000).

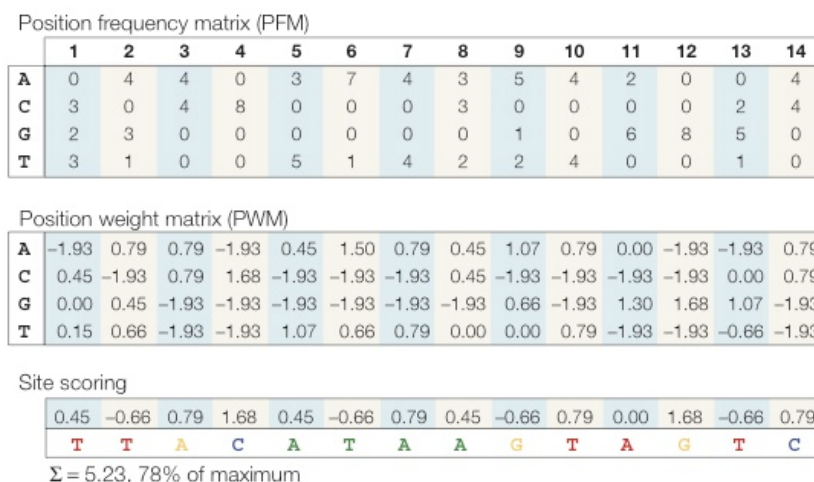


Figure 19: (in continuation to Figure 18) To more accurately reflect the characteristics at each position, a matrix that contains the number of observed nucleotides at each position is created. The frequency matrix is usually converted to a position weight matrix by converting normalized frequency values to a log-scale. Using a matrix model, a quantitative score for any DNA sequence can be generated by summing the values that correspond to the observed nucleotide at each.). (taken and adapted from (Wasserman & Sandelin, 2004))

As for the consensus model, it is based on an alignment of all known binding sequences. Different from the consensus model, the total number of observed nucleotides per site position is recorded, serving as the basis to provide a position frequency matrix (PFM, see Figure 19). Finally, if the frequencies are normalized (each column sums up to 1) and the PFM is

transformed to a probability matrix providing for each nucleotide on each position a probability of occurrence in a TFBS. This framework allows assigning to each arbitrary sequence, with a length correlating with the number of columns in a PFM, a probability score to determine whether the scored sequence is a BS for the TF corresponding to the PFM.

The underlying mathematics of the PWM model is described in detail in the methods section.

The PWM scores are directly related to the binding affinity of the DNA-TF-binding event. Hence a PWM can be considered as both, a statistical and energy-based model (Berg & von Hippel, 1987; Stormo, 2000). However, the PWM model makes two additional assumptions.

The first assumption is that a TF does not tolerate variable spacing or gaps in its binding site. Indeed, it is known that for example the nuclear receptor family allows variable spacing. Thus, the appropriateness of PWMs in TFBS prediction for such TFs is questionable (Wasserman & Sandelin, 2004). The second and widely discussed assumption is that the contribution of a nucleotide to the overall binding affinity of the TF at one position of the site does not depend on the nucleotides that appear on other positions of the site (also referred as additivity) (Sharon, Lubliner, & Segal, 2008). That this assumption does not hold, has been shown for example by Badis and colleagues (Badis *et al.*, 2009), stating that reasonable amount of TFs capture positional interdependencies in their binding motifs.

Figure 20 should illustrate in a fictive example what positional interdependency means (adapted from (Sharon *et al.*, 2008)).

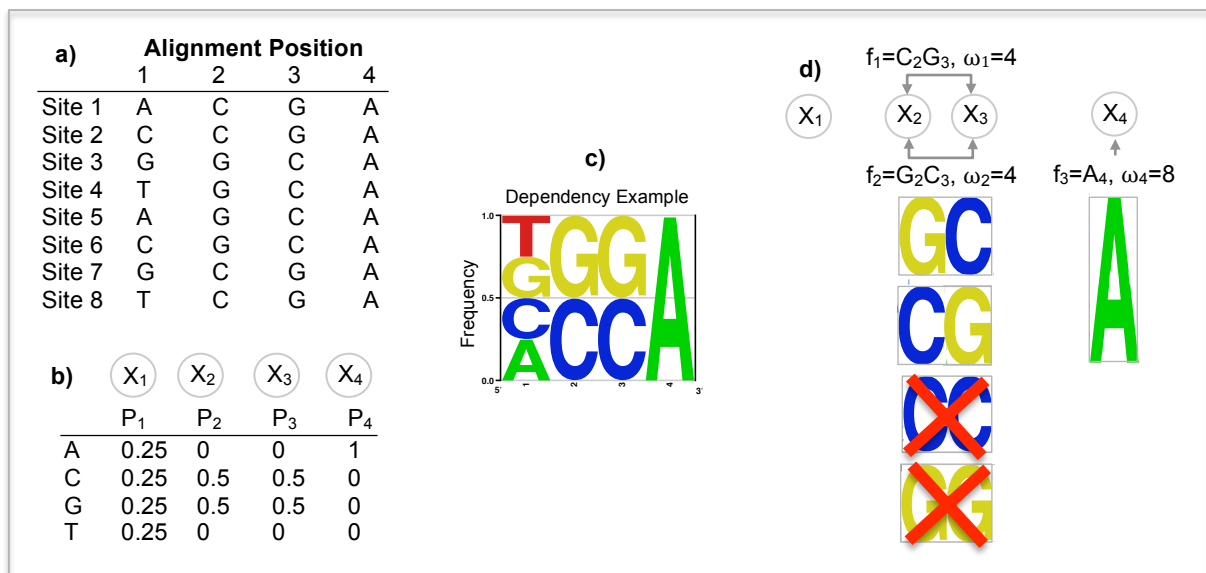


Figure 20: Illustration of positional interdependency based on a fictive example (adapted from (Sharon *et al.*, 2008)). a) aligned known TFBSs, b) PWM based on a), c) sequence motif derived from the PWM, d) consideration of positional interdependency to describe a sequence motif, with ω denoting the weight (corresponding with a probability) of the nucleotide or nucleotide pair showing dependency.

Starting from a set of eight known TFBS (Figure 20 a) a PWM is constructed (Figure 20 b). The TFBS sites show only “CG” or “GC” in the center positions. The PWM learned from the data, that nucleotide “A” on position 4 has a probability of 1, followed by “C” and “G” on position 2 and 4. The PWM does not distinguish between “CG”, “GG”, “CC” and “GC”. Thus, a high probability will be assigned to “CC” and “GG” in the center positions (Figure 20 c), even if this pairing does not occur in the measured TFBS set.

An alternative is to allow assigning probabilities to multiple nucleotides at multiple positions. In such case it is possible to give an exact description of the binding specificities over the center position, namely that “GC” and “CG” have a high probability, while “CC” and “GG” do not (Sharon *et al.*, 2008) (Figure 20 d).

Several approaches implemented in different models have been developed to overcome this clearly violated dependency assumption, wherein all of them include the PWM model intrinsically. However, these models are very complex, integrating different ideas from machine learning and probability theory. Here, only a very brief description annotated with references for the interested reader will be given (the following description of approaches is not claimed to be complete).

The approaches developed can be separated in two main classes: Markov chains and Bayesian network based models. As already coded in the term, the first class of model applies a Markov model of some order.

A Markov chain of order m , wherein m is finite, satisfies

$$P(X_n=x_n \mid X_{n-1}=x_{n-1}, X_{n-2}=x_{n-2}, \dots, X_{n-m}=x_{n-m}) \text{ for } n > m$$

Intuitively spoken, a Markov chain of order k states that the future depends on the past k observations only.

For example in terms of a nucleotide sequence, a Markov model of order zero relies in that the current nucleotide does not depend on the previous one. There is “no memory” and every nucleotide is untied. In fact a Markov model of order zero is represented by a PWM model, assuming independence.

The multinomial model would be: $p(A) + p(C) + p(G) + p(T) = 1$

A first order Markov model introduces dependencies, actually saying that the current nucleotide is dependent on the previous one.

The multinomial model is defined as:

$$\begin{aligned}
 p(A/A) + p(C/A) + p(G/A) + p(T/A) &= 1.0 \\
 p(A/C) + p(C/C) + p(G/C) + p(T/C) &= 1.0 \\
 p(A/G) + p(C/G) + p(G/G) + p(T/G) &= 1.0 \\
 p(A/T) + p(C/T) + p(G/T) + p(T/T) &= 1.0
 \end{aligned}$$

Accordingly the second order Markov model considers a dependency of the current nucleotide on the two previous nucleotides etc.

Zhao *et al* developed a permuted variable length Markov model (PVLMM) showing two improvements over the classical Markov approach. “It searches for the best permutation of the motif positions, and reduces the number of parameters by using a context tree representation for the Markov model representation” (X. Zhao, Huang, & Speed, 2005).

However, even if Markov chain based models perform well, their ability to model dependencies between more distant positions is limited and with an increasing order the size of the model representation is growing exponentially. Therefore, Zhao and colleagues suggested searching for a permutation of the binding site positions that produces the best model, considering a maximum motif length of 9bp (Sharon *et al.*, 2008).

The other approach is using Bayesian networks to represent higher order dependencies between motif positions and it has been shown that they outperform PWMs in predicting putative TFBSs in ChIP-chip data. Here, a directed acyclic graph G has been used to represent the dependencies. “The nodes of the graph correspond to random variables X_1, \dots, X_k and a parameterization describing a conditional distribution for each variable, given its immediate parent in G . The corresponding joint probability distribution decomposes into the product form” (Barash, Elidan, Friedman, & Kaplan, 2003):

$$P(X_1 \dots X_K) = \prod_{i=1}^K P(X_i | Pa_i^G)$$

with Pa_i^G denoting the (possibly empty) set of parents of X_i in G

“The formal semantics of Bayesian networks is in term of conditional independence statements that each variable X_i is independent on its non-descendants in the graph G given its parents in G . In general, the more edges are in a graph G the more complex the dependencies between the positions are. The most simple graph has no edges, corresponding to a PWM model” (Barash *et al.*, 2003).

Figure 21 illustrates a few examples of Bayesian networks and their associated form of probability distribution. In Bayesian tree networks each position has only one parent, so that G

is denoted as a forest. Referring to Figure 21, examples a) and b) refer to this kind of network. “These networks generalize a first order Markov chain, for which efficient algorithms exist to learn the best tree structure as for example a Chow-Liu tree” (C. K. Chow & Liu, 1968) (see methods).

A mixture of tree model (see Figure 21 c) is a tree structure network combining the benefits of a tree structure with the added richness of hidden mechanisms, leading to a natural extension similar to a mixture of PWMs. Here each X_i has as parent the hidden variable T and maximum one other nucleotide position (Barash *et al.*, 2003).

Further extensions of this approach have been made, like using a context dependent representation of the conditional probability distributions (Ben-Gal *et al.*, 2005), to only consider dependencies between non-overlapping positions, suggesting a simple Bayesian network model called generalized weight matrix (GWM) (Zhou & Liu, 2004), to add structural DNA features (Pudimat, Schukat-Talamazzini, & Backofen, 2005).

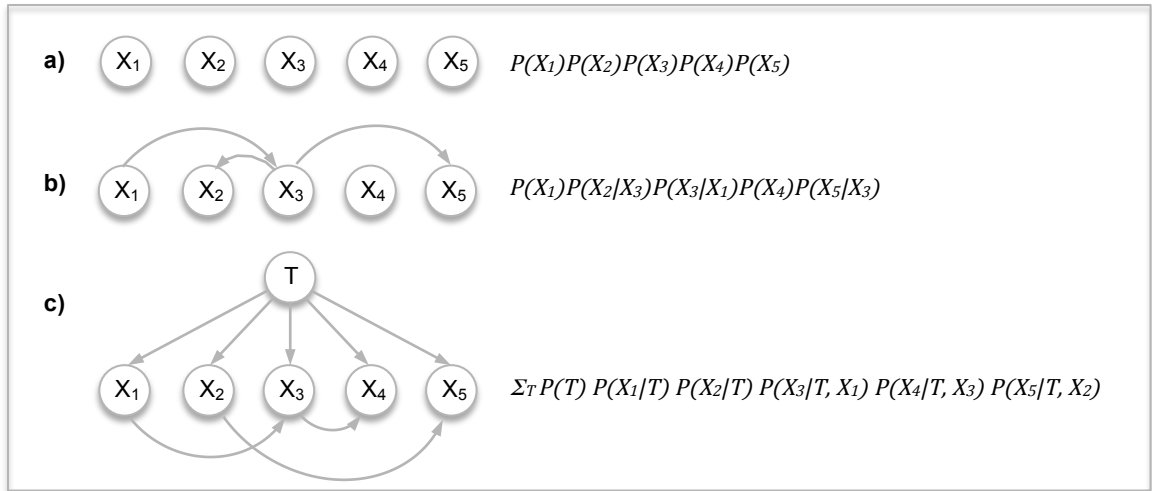


Figure 21: Examples of different Bayesian network models for a sequence motif with 5 positions. For each model an example of a Bayesian network structure is provided plus the corresponding representation in a joint distribution. a) PWM, b) Tree, c) Mixture of Trees (adapted from (Barash *et al.*, 2003))

However, even if all extensions described above outperform the classical PWM model (at least in the datasets tested), in practice the probabilistic models for binding site recognition, such as the PWM, are popular because of their simplicity, intuitive appeal and because they can be easily implemented in motif discovery algorithms (Yue Zhao, Granas, & Stormo, 2009). Additionally, its wide usage and successful application, indicates that the simpler PWM model is adequate in most cases even if the additivity of models is clearly violated.

The details and mathematical representations of the models applied in this work are provided in the methods section.

1.9. Data Integration for Regulatory Element Detection

Based on these new sequencing technologies, researchers are able to approach the sequencing 2,500 genomes in a reasonable time with reasonable costs, measure gene expression and transcription factor binding on a genome wide scale or score hundreds of thousands of SNPs in individual samples. Automated bibliographic searches are applied to extract biological information from literature and to model biological interaction, hoping to infer by the analysis of one type of data the observations for another type of data.

The flood of data, still lacking standardized formats, results in the fact that challenges in biology are now challenges in computing. For example 1000G will collectively generate data in a scale of petabytes just for raw data, not including the subsequent integrative analysis. One main goal of biological researchers is to integrate diverse, large-scale data sets to construct models that can predict complex phenotypes such as disease, so that the task of data integration as essential need becomes more and more evident. Consequently, “the information coming from numerous technologies must be stored in large-scale databases and data mining in such high-dimensional data sets becomes more complex, requiring multiple scoring schemes” (Schadt *et al.*, 2010).

In the following, one very recent example of data modeling, integration and mining efforts in the context of gene regulation analysis will be provided.

CENTIPEDE (Pique-Regi *et al.*, 2011) constitutes the approach of massive data integration with the aim to infer gene regulatory elements on a genome wide-scale. It represents a probabilistic framework that combines genome sequence information with cell-specific experimental data. The model applied by CENTIPEDE is a hierarchical Bayesian mixture model to infer regions from the genome that are bound by particular TFs.

The general workflow of CENTIPEDE is

- 1) Genome-wide scan for all approximate matches to a target PWM, wherein each site that matches the PWM is considered as a candidate region. The original publication considered 756 from ~1,000 available matrices from JASPAR (Bryne *et al.*, 2008) and TRANSFAC®.
- 2) Collecting relevant data for the candidate region for prior and likelihood estimation. The original publication used for the prior estimation the results from the PMW screening, an average PhastCons⁸ conservation score and the distance to the closest annotated Ensemble (Birney *et al.*, 2004) TSS. The experimental data for the

⁸ PhastCons is a program for identifying evolutionarily conserved elements in a multiple alignment, given a phylogenetic tree.

likelihood estimation comprised three DNase-seq and seven histone modification ENCODE (Birney *et al.*, 2007) datasets.

- 3) Fit a Bayesian mixture model (see Annex).
- 4) Report candidate binding sites with a high posterior probability (e.g. > 0.99) of being bound.
- 5) Validation of candidate binding sites from 4) using ChIP-seq or if not available sequence conservation (in such case to be excluded from the model fitting)

The result of the first application of CENTIPEDE is a genome-wide map of 827,000 TFBS in human lymphoblastoid cell lines, based on 239 PWM of known TF binding motifs and 49 novel sequence motifs (Pique-Regi *et al.*, 2011). The authors were able to show that a model integrating chromatin accessibility, measured by DNaseI, and histone modification data agrees very closely with empirical ChIP-seq measurements of TF binding at candidate motif sites. The strength of CENTIPEDE is the ability to identify binding sites of many factors from a single experimental assay, because chromatin accessibility and histone modification are only tissue or cell type, but not TF-specific measures. The weakness of CENTIPEDE might be, that independence between the different experimental data is assumed in the modeling approach, which is rather false than true. However, according to the authors, the merge of a high amount of different datasets (here seven histone marks) might reduce a possible violation of reality. The best results were gained by combining the prior genomic data and DNase-seq data only. The authors conclude that, when DNaseI data are integrated no further predictive power is gained by adding the histone sets. The predictions of CENTIPEDE show a precise resolution of binding locations and a quantitative measurement of potential binding occupancy. It can be considered as a complementary tool, extending ChIP-seq, which can provide exhaustive information about binding for factors of special interest, but may miss sites that do not contain a recognizable motif.

1.10. Approaches for regulatory SNP Detection

Several approaches have been developed in recent past to analyze the role of regulatory SNPs (rSNPs), wherein the main focus lies on SNP-effects on TF binding, leading in its final consequence to an altered gene expression. The need for *in silico* methods is on hand, as the wet-lab discovery of such variants is practically not feasible on a genome-wide scale.

The most simple approach is to overlap *in silico* predicted TFBS with SNP locations (Ponomarenko, 2003). Though, this approach produces a high amount of false positives, since the effect a SNP might have on the binding affinity is not included. Thus, the logical

improvement is to integrate SNP-generated score changes based on a model scoring TFBS as for example a PWM (Andersen *et al.*, 2008). The underlying assumption is that SNPs generating a “larger” score difference are more likely to be an rSNP than those generating a “smaller” difference. This approach was only successfully applied in combination with additional information like phylogenetic footprinting, revealing that the score difference alone does not represent a valid descriptor. In 2010 the first method was published using a modified affinity score. The calculation of the complete distribution of affinity scores, allows the computation of a p-value. The ratio of the p-values associated with the two allelic variants of the sequence can be used to indicate, whether the binding affinity of the BS is disrupted. The result of this method is a ranked list of p-value ratios, not indicating the significance of the ratio that is why the distinction of false and true positives is problematic. The most recent approach published by MacIntyre and colleagues (Macintyre, Bailey, Haviv, & Kowalczyk, 2010) provides predictions with statistical significance. Here, the PWM scores are used directly and all possible PWM scores are calculated. Then, the distribution of score differences is used to gain a p-value (log-rank method) allowing the determination of significant SNP effects. However, although useful for identifying mutations overlapping known TFBSs, in the absence of additional information, such comparisons have limited value for predicted TFBSs (including all cases of *de novo* generation of TFBSs) (Worsley-Hunt, Bernard, & Wasserman, 2011). The remaining problem is the high rate of false positives, resulting in a poor specificity.

Nevertheless, *in silico* predicted regulatory elements can be overlaid with genome annotations or experimental data to focus attention on the regions that are more likely to be functional (see Figure 22). The main purpose of such additional information is to gain supporting evidence that a predicted regulatory element is functional, increasing specificity. Such filters, applied individually or in combination, to improve the biological plausibility are for example topology based (gene structure), based on phylogenetic footprinting (sequence conservation), experimentally determined TFBS (e.g. ChIP-seq) or chromatin accessibility (see section 1.6.3).

A further step, feasible due to the increasing affordability of whole-genome sequencing, is the investigation of *cis*-regulatory-element-related traits in a familial context. Being able to analyze related genomes, therefore including the segregation of sequence variants possibly related with a phenotype, can improve the ability to predict associated regulatory variants dramatically (Worsley-Hunt *et al.*, 2011).

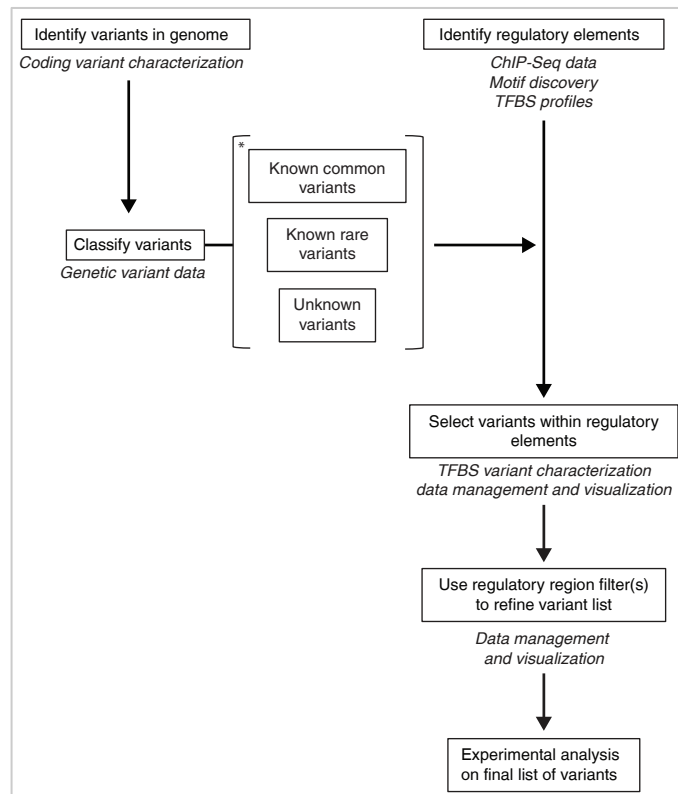


Figure 22: Overview of a workflow for *cis*-regulatory variant detection. The boxes represent steps in the workflow, and the italicized descriptions under the boxes correspond to analysis resources. For identification of regulatory elements and regions, the order in the workflow may be changed without loss of information. Common variants, flagged by an asterisk, may be eliminated from the analysis or alternatively flagged for later tracking (taken from (Worsley-Hunt *et al.*, 2011))

Very recently Boyle *et al.* published a database called RegulomeDB which aims to annotate and guide the interpretation of functional variants in personal human genomes (Boyle *et al.*, 2012). This database combines high-throughput experimental data sets from ENCODE as well as other sources to a tool meant to score variants with respect to their regulatory potential. The data sets used are manually curated regions experimentally characterized of being involved in regulation, ChIP-seq data for a variety of TFs gained from numerous different cell types, chromatin state and accessibility data from over 100 cell types and expression quantitative trait loci information. Additionally to experimentally derived data machine learning methods have been applied to improve the predictive capabilities for regulatory variants. The methods applied in this regard are DNase footprinting to detect protein binding sites and binding motif alteration detection by nucleotide variants. The final scoring of a query variant is based on the amount of biological and computation support and reflected by different categories as illustrated in Table 2.

Table 2: RegulomeDB variant classification scheme (taken from Boyle *et al.*, 2012))

Category Scheme	
Category	Description
	Likely to affect binding and linked to expression of a gene target
1a	eQTL + TF binding + matched TF motif + matched DNase footprint + DNase peak
1b	eQTL + TF binding + any motif + DNase footprint + DNase peak
1c	eQTL + TF binding + matched TF motif + DNase peak
1d	eQTL + TF binding + any motif + DNase peak
1e	eQTL + TF binding + matched TF motif
1f	eQTL + TF binding/DNase peak
	Likely to affect binding
2a	TF binding + matched TF motif + matched DNase footprint + DNase peak
2b	TF binding + any motif + DNase footprint + DNase peak
2c	TF binding + matched TF motif + DNase peak
	Less likely to affect binding
3a	TF binding + any motif + DNase peak
3b	TF binding + matched TF motif
	Minimal binding evidence
4	TF binding + DNase peak
5	TF binding or DNase peak
6	Motif hit

Loss of heterozygosity

One special version of regulatory SNP is the so-called mechanism of “Loss of heterozygosity”. This term describes an event where the normal function of one allele is lost, when the other one has been already inactivated. For example a mother has a gene-expression-inactivating mutation in a germline cell which is passed to the offspring. If the child receives a functional allele from the paternal side, the offspring is heterozygous for this allele. If now the remaining functional allele in the child is affected by a mutation, the child will lose this heterozygosity resulting in a loss of gene function. This mechanism is known in context of cancer (Clarke *et al.*, 2006; Jones & Nakamura, 1992), where a loss of heterozygosity has been identified to silence tumor suppressor genes. However, this mechanism is difficult to detect, since access to germline cells and for example cancer cells from the affected individual as well as its parents is necessary.

2. Objectives

The ability to extract the genomic specificity of a TF allows to chart the affinity landscape of a TF, which describes how a TF interprets a nucleotide sequence. A fundamental problem involves the characterization and the *in silico* prediction of the genomic specificity of a given TF. The dominant model in the literature for TF specificity is the PWM, which assumes that the binding affinity of the factor decomposes additively over the positions of the binding site — an often violated approximation. More flexible probabilistic models, outperforming the classical PWM, have been proposed in the literature, but their use has been limited in favor of the typically much simpler PWM. However, accurate TF binding models, with high specificity and sensitivity, would allow to predict in large-scale the genomic binding sites of a TF, as well as possible effects of sequence variations like SNPs on transcriptional binding. The latter can have significant impact in biomedical research, since the ability to detect *in silico* the effects of genomic variations on transcriptional programs could help to elucidate the biology of diseases and lay the foundations for identifying therapeutic targets (Segal & Widom, 2009).

Therefore, the specific aims of this thesis were:

1. To advance the state of the art by proposing a flexible probabilistic model that is very easy to infer from data, termed “Ensemble of Trees” (ET), for transcription factor binding site detection.
2. To compare the predictive performance of the ET model with other state-of-the-art models on an independent, experimentally derived and complete, sequence validation set.
3. To test the specificity and sensitivity of the proposed models in a biological context by detecting differential TF binding profiles in ChIP-seq data in a family based set-up.
4. To apply the proposed probabilistic models to detect causative regulatory SNPs.

3. Material and Methods

Learning models to predict TF binding requires training as well as validation sets. For both datasets, the need for correct assignment of binding *versus* non-binding is very demanding. While a misclassification in the training set would lead to a biased model, the same in the validation set would lead, if the training was correct, to biased evaluation of the model. Accordingly, the optimal sets provide correct assignment of binding and non-binding in a comparable manner.

The public database UniPROBE (Newburger & Bulyk, 2009) provides for the majority of its TFs, two independent replications. The array design follows a certain algorithm (see next section), resulting in replicates that are in general well comparable with regard to sequence structure and completeness.

It has been decided to use UniPROBE's Protein Binding Microarray (PBM) data to investigate the training of probabilistic models for TF binding site detection, allowing to train and validate and therefore to assign binding affinity on a comparable set of probes.

3.1. UniPROBE – Protein Binding Array Data

The UniPROBE PBMs have a highly compact and synthetic DNA sequence design, representing all possible 10-mers (DNA sequence with given length of 10) on a single array.

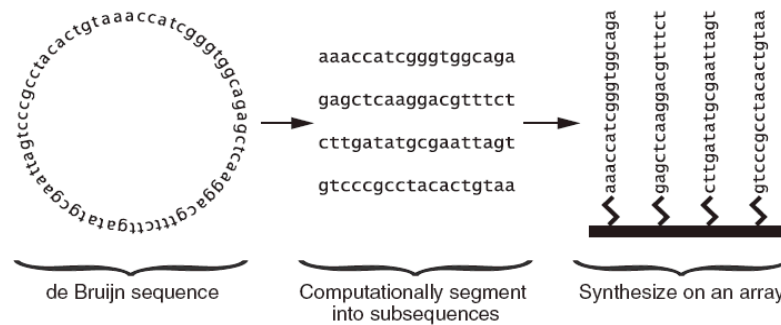


Figure 23: Schematic illustration of a de Bruijn sequence with order 3 (left). This sequence contains all 3-mer variants once (64 3-mers). The circular sequence is separated into subsequences, with an overlap of two nucleotides preserving all 3-mers (middle). The subsequences are spotted on the array (right). Taken and adapted from (Berger *et al.*, 2006)

The basis to achieve this complete set of small oligonucleotides is a de Bruijn sequence of order 10 (see Figure 23). This sequence represents a circular string of length 4^k ($k=10$) covering all contiguous 10-mers and all gapped 10-mers spanning total 11 positions. This circular sequence is partitioned in subsequences with a length of 36bp, leading to ~44,000 single-stranded features on a PBM. Each feature (in the following also called probe) is

composed of a 24nt primer and a 36nt variable sequence, summing up to a feature length of 60nt with 26 distinct, overlapping 10-mers (see Figure 24). The arrays are constructed in a way, that all spotted single-strands oligonucleotides covering all 10-mers are converted to double-stranded DNA probes (Berger *et al.*, 2006).

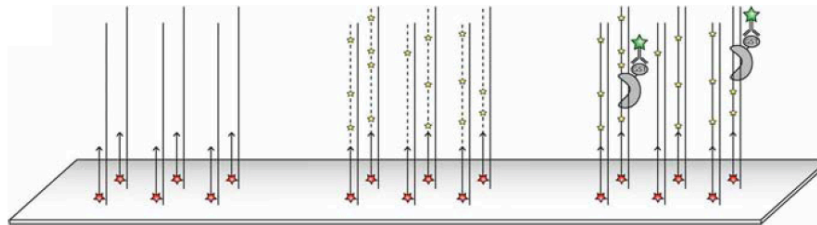


Figure 24: Schematic illustration of the three major stages of a PBM experiment: primer annealing (left), primer extension (middle) and protein binding (right). Taken from (Berger *et al.*, 2006)

Considering statistics, a single instance per sequence is not sufficient concerning quality control. Thus, the length of the k-mer to be considered has to be reduced in order to assure reproducibility. Due to the sequence design every 8-mer is present at least 16 times per array. Nonpalindromic versions occur at least 32 times (including reverse complements). In the work at hand a k-mer length of 9 has been chosen to have an appropriate compromise between binding site length and occurrences of k-mers on the array. According to the authors each possible 9-mer is present in either orientation in at least eight probes, which has been considered to be sufficient. The UniPROBE database provides already analyzed data and position weight matrices directly usable for TFBS screening. However, to develop new methods to better capture nucleotide dependencies, the processed intensity values of UniPROBE data provided are used, namely the complete probes and 36-mers respectively, with the measured intensity values. The normalized probe intensities are derived from the raw data after adjusting for relative DNA concentrations at each spot and for spatial non-uniformities within the microarray (Newburger & Bulyk, 2009).

For the majority of TFs at least one replication array is provided, based on a de Bruijn sequence uncorrelated with the one on the first array, but providing the same sequence properties regarding 10-mer coverage. The authors show that a "striking correlation between experiments, not only for the highest affinity binding sites, but also for moderate- and low-affinity binding sites" exists. Thus, this second array represents an optimal validation sample after model training.

Currently the UniPROBE database hosts DNA binding data from a diverse collection of organisms, including amongst others the prokaryote *Vibrio harveyi*, the yeast *Saccharomyces cerevisiae*, the worm *Caenorhabditis elegans*, *Mus musculus*, and *Homo sapiens* (Newburger & Bulyk, 2009).

3.2. Selection of TFs

The TFs considered in this work come from the complete mouse collection. A certain focus is directed to TFs suspected to have a nucleotide dependency within their binding motif, since this feature should be better captured by a tree structure than a PWM. This sub-group of TFs is based on Bulyk *et al* (Badis *et al.*, 2009) and covers 19 different factors, namely E2F2, E2F3, Eomes, Esrra, Gcm1, Hbp1, Irf5, Myb, Mybl1, Nr2f2, Rara, Rxra, Sox4, Sox7, Sox8, Sox11, Spdef, Tcf2a and Zfp281.

3.3. Generating Training Input and *de novo* Motif Discovery

Motif finders can be classified in combinatorial or probabilistic ones. The output of the combinatorial approach is a consensus sequence and TF binding sites are predicted by the number of mismatches with the consensus sequence. The result of a probabilistic motif finder is for example a PWM, representing the distribution of the bases for each position in the binding site (Reid & Wernisch, 2011). Probabilistic models are more complex and flexible and well established in the field of TFBS prediction. These can be trained from a ready provided binding site sequence collection or after an initial motif inference step that starts with longer sequences that are expected to contain a specific TF motif. Two popular algorithms for inferring a motif applied are the expectation-maximization (EM) algorithm and Gibbs sampling (Zambelli, Pesole, & Pavesi, 2012).

To achieve an aligned input set for the training of the probabilistic models used in this work, a motif finder step has been introduced at the beginning of the training phase. The probably most popular and well-known motif finder is MEME (Multiple EM for Motif Elicitation) (Bailey & Elkan, 1994). This software is an implementation of the MM algorithm estimating the parameters of a probabilistic model, which could have generated a dataset of (unaligned) sequences with a common protein-binding site or other sequence feature. This MM algorithm is an extension of the EM technique able to model also sequences containing zero, one or many occurrences of a motif. MEME is improving a model of the motif iteratively. In each iteration, the locations of the binding sites are estimated using the current model of the motif and the motif is updated using the predicted sites weighted by their likelihoods (Bailey & Elkan, 1994).

In this work STEME (Suffix Tree EM for Motif Elicitation) has been chosen as a motif discovery tool (Reid & Wernisch, 2011). STEME represents an approximation of the EM algorithm in form of a suffix tree, which is a data structure able to efficiently index a set of sequences, as for example input sequences coming from ChIP-seq peaks or UniPROBE

probes. Additionally, these suffix trees are fitting well in an environment where content is more important than position only allowing efficient access to subsequences by their content.

The decision to select STEME as a motif finder has been done based on three facts:

- In order to train models aligned binding site sequences are required and STEME provides this information in an easy-to-integrate manner.
- STEME is implemented in C++ as an open source library and provides also a Python scripting interface. This made its results easier to integrate and access for the project, which is mostly implemented in Python as well.
- STEME has been shown to be very fast in comparison to other motif finders, making it more suitable as an integrated pipeline item.

In order to get an idea and to evaluate if the output of STEME can be used as a valuable input for the model training, the generated PWMs have been compared with those of MEME and if available from UniPROBE and TRANSFAC®.

Starting from the raw UniPROBE data, the best 800 and worst 800 probes are selected based on sorted intensity values as STEME input (see Figure 25). To address the problem of an appropriate background model, STEME was applied in a discriminative manner, meaning that the background model is built from different sequences, determined by worst 800 probes. In this way STEME finds motifs that are less likely in the background sequences (personal communication with John Reid). The motif length to be detected has been restricted to 9.

The following parameters have been used:

```
--minw=9, --maxw=9, --min-sites=800, --bg-model-order=2, --bg-fasta-file=<worst800.fa> <best800.fa>
```

The output of STEME is a list of ideally 800 sequences. If STEME did not detect one motif per sequence then the number would be lower than 800. Beside the detected 9-mer motif, a probability of binding, a p-value, the Hamilton distance to the consensus, the sequence input-number and offset of the detected motif within the screened sequence is provided. The input-number of the sequence and offset are of special relevance. This information is needed to reassign the measured binding signal to the respective input-sequence or to perform signal correction (e.g. distance from glass slide of motif for UniPROBE, see below).

If $signal = (signal_1, \dots, signal_z)$ denotes the intensity signal of a probe i , containing the 9-mer sequence X , within z probes on the array, then the normalized intensity $\omega(X)$ of a probe i calculates by

$$\omega(X) = \frac{signal_i}{\sum_{j=1}^Z signal_j}$$

Furthermore, to not bias the training with mixed signals, all output sequences from STEME comprising more than one detected motif are discarded. Additionally, it has been hypothesized that the distance to the glass slide of the array has an influence on the signal strength, meaning that binding events, more close to the glass slide, have a confounded lower signal than those more far away (Yue Zhao et al., 2009).

The algorithm to correct for the glass slide distance has been adapted from (Yue Zhao et al., 2009) for 9-mers and was integrated in the training process. The correction factors are calculated for the different positions of the motif within the probe and the offset of the detected motif provided by the STEME output is a necessary item to integrate the correction procedure in an automatized manner.

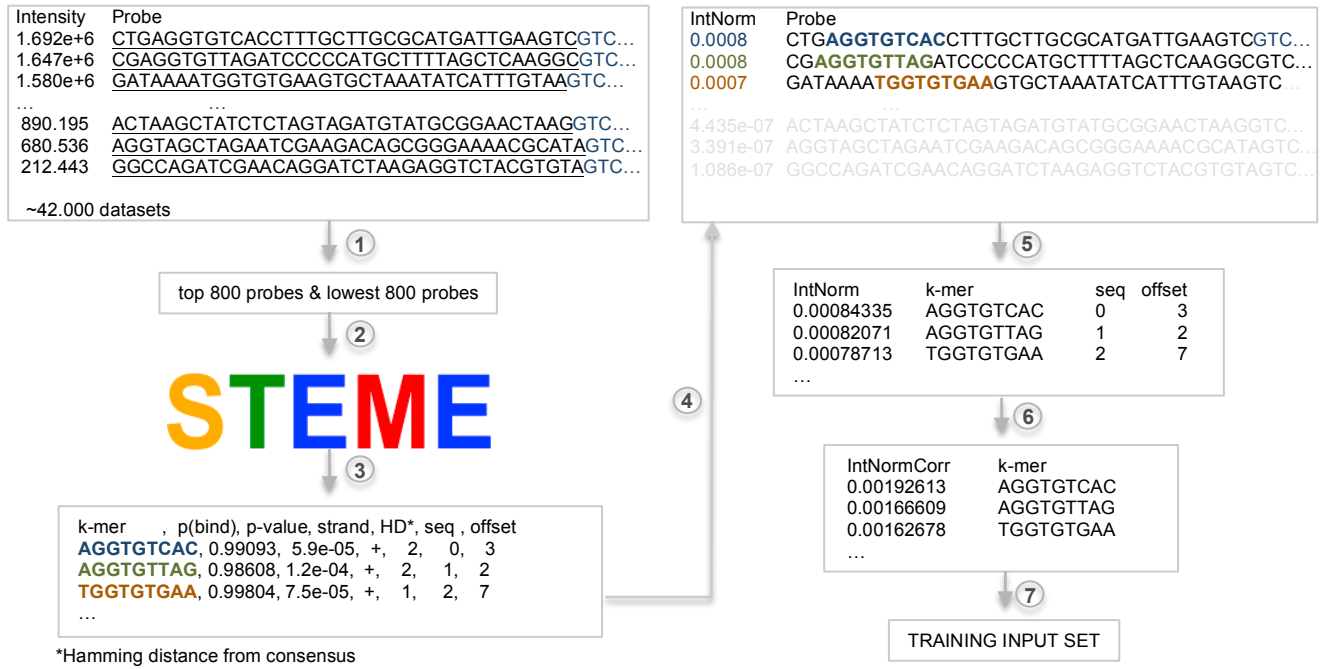


Figure 25: Schematic illustration of the generation of the training input set. 1. Selection of the best and worst binding probes and extraction of the variable regions, 2. Discriminative motif discovery by STEME, 3. Achieving a STEME output with the detected motifs, a binding probability, a p-value, the strand of detection, a Hamilton distance from consensus, the probe number and motif offset within the probe, 4./5. Re-assigning motifs with the normalized intensity of the respective probe, 6. Correcting the normalized intensity for the distance to the glass slide, 7. Receiving final training input set.

Finally the training input set holds only 9-mers which can be assigned to the corresponding probe. This means, that the weight of each 9-mer is the normalized intensity signal, corrected for the location of the 9-mer within the corresponding probe. If a 9-mer occurs in y probes it

will be y -times present within the training input, each time with a different weight depending on the comprising probe intensity signal.

3.4. Screening and Scoring of Sequences

When describing the screening and scoring of a sequence, two different cases need to be addressed:

- scoring a sequence with the same length as the motif to be detected
- scoring a sequence with a length greater than the motif to be detected

In the latter, the sequence has first to be screened for all possible, here ungapped, motifs. Since a model provides a score for each possible sequence one has to select which of the scored motifs (if at all) should be kept in the dataset. In this work, only the best scoring motif (highest score) is kept for further processing. Different approaches might be possible, like keeping all motifs, determined by its offset within the sequence. The procedure was applied for all three models and is illustrated in the following Figure 26.

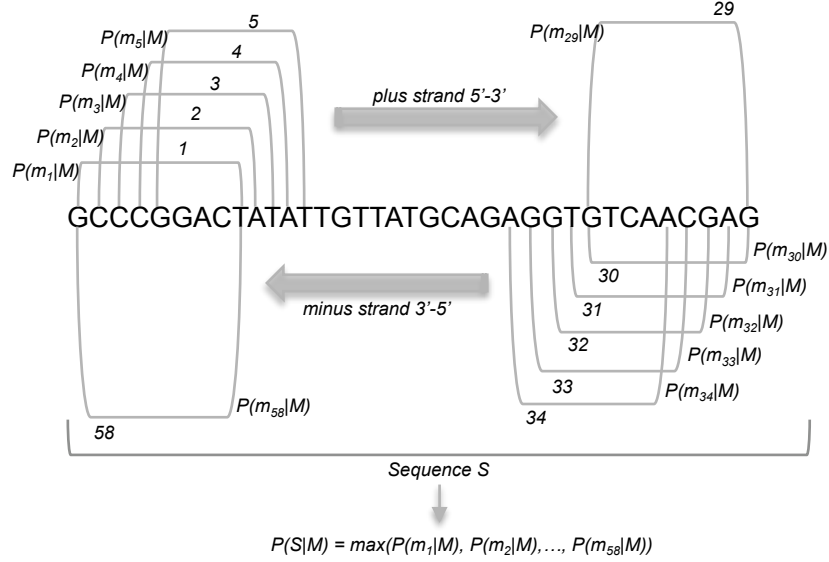


Figure 26: Schematic illustration of the screening algorithm. If the sequence to be score is greater the motif to detect, the algorithm is moving with a sliding window of length l over the sequence, calculating for each sequence-window the model score. This procedure is applied on both, the plus and minus strand. The sequence window with the highest model score, its offset within the probe and strand info is kept for further processing.

3.5. Probabilistic TF Binding Models

3.5.1. Position Weight Matrix

Three different items determine a binding site motif, its start, end and sequence. If multiple binding sites have been found experimentally for a factor, the PWM represents an expressive model to summarize the residue conservation. If a PWM is generated for DNA sequences each row represents a nucleotide and each column a position of the aligned sequences. In most cases, the matrix values indicate the integer counts, frequencies or probabilities of each nucleotide to occur on the respective position (Van Helden, 2005).

The PWM is the dominant model for TF binding site description (referred also as “motif”) and detection in biology and is effectively a Naive Bayes probabilistic model that assumes that all positions in a binding site are statistically mutually independent given the binding event.

The calculation of a score for a sequence X under a PWM model can be formalized as:

$$p(X|M) = \prod_{i=1}^L p(x_i)$$

with

$p(x_i)$ denoting for the probability of nucleotide nt on position i within the motif.

$p(X|M)$ is the product of the probabilities for each nucleotide in sequence X ; $P(X|M) = p(x_1) \cdot p(x_2) \dots$

L is defined as the length of sequence X

For efficient computational analysis, the values of the PWM are converted to a log-scale, so that the product turns to a sum (Wasserman & Sandelin, 2004).

Practically illustrated the calculation of a PWM score for a sequence X with length L is as follows:

$X = ATCGTTGAA$

$$\begin{aligned} p(X|PWM) &= p(A_1) \times p(T_2) \times p(C_3) \times p(G_4) \times p(T_5) \times p(T_6) \times p(G_7) \times p(A_8) \times p(A_9) \\ &= 0.611237 \times 0.148949 \times 0.060187 \times \dots \times 0.480218 \times 0.765885 \times 0.437807 \\ &= 1.415e-6 \end{aligned}$$

A PWM can be summarized in a sequence logo (Schneider, Stormo, Gold, & Ehrenfeucht, 1986). Sequence logos provide a more valuable and precise representation of a sequence similarity that a consensus sequence is doing (Crooks, Hon, Chandonia, & Brenner, 2004). They reveal features of the sequence in a self-speaking and intuitive way. With regard to the appearance of the nucleotides, different possibilities, how a sequence logo is build, are possible (see Figure 27).

The most common representation is the one using the relative entropy, which is formalized as follows (GuhaThakurta, 2006):

$$I(p) = \sum_{j=1}^L \sum_{i=A}^T f_{i,j} \log \frac{f_{i,j}}{P_i}$$

with

$I(p)$ denoted as information content also called relative entropy for the PWM representing a pattern p
 L denoting the pattern length, i denotes for the index of the base {range A through T} at position j of the PWM, $f_{i,j}$ denotes for the frequency of base i at position j of the PWM, P_i denoted the probability of observing that base in the data (e.g. for uniform nucleotide distribution P_i was 0.25 for each nucleotide)

Other representations are using the frequency or probability directly, where the height of the letters sums to one (see Figure 27). The y-axis is arbitrary and depends on the choice for the letter height descriptor. For example, when the probability is used directly, the maximum of the y-axis is one. The x-axis is always the position of the nucleotide within the motif.

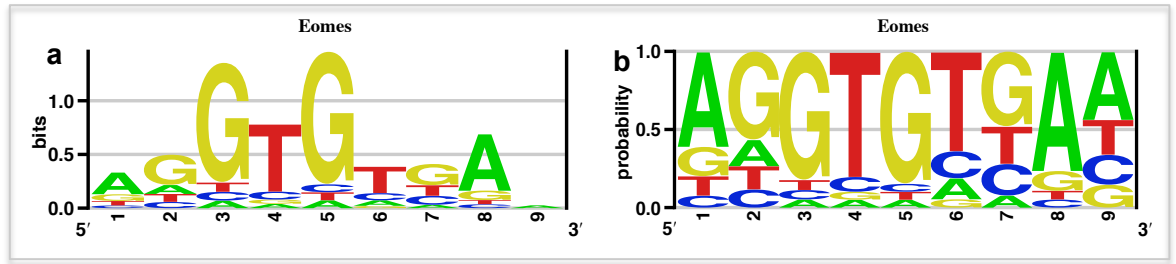


Figure 27: Two different PWM representations are shown: a) displaying the nucleotide height as relative entropy and b) displaying the height as probabilities as to find in Table 6. Generated by means of enoLOGOS (Workman et al., 2005)

Training the PWM model

In this work, the PWM is trained from an input set of aligned sequences resulting from the *de novo* motif detection performed by STEME. Each motif is assigned with the normalized intensity of the respective probe. Accordingly, here instead of an integer count, a weighted count, namely the probe intensity value, is used for the nucleotides.

Table 3: Illustration of the training input set as is used to learn a PWM.

Weighted count	Input 9-mer								
0.00193	A	G	G	T	G	T	C	A	C
0.00167	A	G	G	T	G	T	T	A	G
0.00163	T	G	G	T	G	T	G	G	A
...

As a first step the simple weighted count $\omega_{nt,i}$ for each nucleotide on each position of the matrix is calculated. If X is a 9-mer in the training input set, then $\omega(X)$ is the weighted count of

9-mer X . The calculation of the weighted count of a single nucleotide can be formally expressed as

$$\omega_{nt,i} = \sum_{X: X_i=nt}^{\#X} \omega(X)$$

with

$nt \in \{A, G, T, C\}$ denoting for the different nucleotides on position i

$i = (1, 2, \dots, L)$ denoting for the positions of the single nucleotides within a motif of length L

$\omega(X)$ denoting the weighted count of a 9-mer within the training input set

The following Table 4 illustrates the weight calculation in more detail.

Table 4: Illustration of the initial step for PWM training. The weight for each nucleotide is calculated by summing up the corrected normalized intensities of the respective probe over all probes in the input training set (upper table). The lower table holds the sum of the weighted counts over all probes in the training set.

position	1	2	3	4	5	6	7	8	9
A	0.0019 + 0.0017	0.0019 + 0.0017 ...	0.0016 ...
T	0.00163	0.0019 + 0.0017 + 0.0016	0.0019 + 0.0017 + 0.0016 ...	0.0017
C	0.0019	0.0019...
G	...	0.0019 + 0.0017 + 0.0016 ...	0.0019 + 0.0017 + 0.0016	0.0019 + 0.0017 + 0.0016	0.0016 ...	0.0016 ...	0.0017 ...
position	1	2	3	4	5	6	7	8	9
A	2.77532	0.58266	0.00700	0	0	0.41433	0.13337	3.54075	1.91693
T	0.36403	0.48723	0.06457	3.72298	0	2.91165	0.93428	0.02293	0.85619
C	0.13007	0.32882	0.04789	0.21973	0	0.61099	0.75501	0.00828	0.70104
G	0.68008	2.5508	3.83004	0	3.94957	0.01253	2.12684	0.37754	0.47534

As one can take from the upper table, some cells in the matrix have the value 0, meaning that for this position in none of the input sequences the respective nucleotide has been observed. To eliminate these null values before log-conversion, and in part to correct for small samples of binding sites, a sampling correction, known as pseudocount ps , is added to each cell of the PWM (Wasserman & Sandelin, 2004). This pseudocount will be “shared” between all the residues of each column of the matrix in order to obtain the probability $p(x_i=nt)$ (also referred as single marginal in the following) (Sand, Turatsinze, & Helden, 2008).

$$p(x_i = nt) = \frac{\omega_{nt,i} + \frac{ps}{4}}{\sum_{nt'}^{(A,C,T,C)} \omega_{nt',i} + ps} \quad \text{with} \quad \sum p(x_i) = 1$$

After adding the pseudocount the individual positions are normalized, so that each value in the matrix represents a relative frequency or probability. Consequently, each column needs to sum up to 1.0.

Table 5: The weight for each nucleotide is corrected by adding a pseudocount. *ps*: pseudocount =1/4

position	1	2	3	4	5	6	7	8	9
A	2.77532+ <i>ps</i>	0.58266+ <i>ps</i>	0.00700+ <i>ps</i>	0+ <i>ps</i>	0+ <i>ps</i>	0.41433+ <i>ps</i>	0.13337+ <i>ps</i>	3.54075+ <i>ps</i>	1.91693+ <i>ps</i>
T	0.36403+ <i>ps</i>	0.48723+ <i>ps</i>	0.06457+ <i>ps</i>	3.72298+ <i>ps</i>	0+ <i>ps</i>	2.91165+ <i>ps</i>	0.93428+ <i>ps</i>	0.02293+ <i>ps</i>	0.85619+ <i>ps</i>
C	0.13007+ <i>ps</i>	0.32882+ <i>ps</i>	0.04789+ <i>ps</i>	0.21973+ <i>ps</i>	0+ <i>ps</i>	0.61099+ <i>ps</i>	0.75501+ <i>ps</i>	0.00828+ <i>ps</i>	0.70104+ <i>ps</i>
G	0.68008+ <i>ps</i>	2.5508+ <i>ps</i>	3.83004+ <i>ps</i>	0+ <i>ps</i>	3.94957+ <i>ps</i>	0.01253+ <i>ps</i>	2.12684+ <i>ps</i>	0.37754+ <i>ps</i>	0.47534+ <i>ps</i>

Table 6: After adding a pseudocount, the corrected normalized intensities of the respective probe are calculated by summing up over all probes in the input training set. Finally the weighted corrected counts are normalized to achieve relative frequencies or probabilities.

position	1	2	3	4	5	6	7	8	9
A	0.611237	0.16823	0.051925	0.050510	0.050510	0.134221	0.077456	0.765885	0.437807
T	0.124059	0.148949	0.063557	0.804076	0.050510	0.638781	0.239273	0.055144	0.223497
C	0.076789	0.116945	0.060187	0.094904	0.050510	0.173956	0.203053	0.052183	0.192148
G	0.187915	0.565875	0.824332	0.050510	0.848470	0.053042	0.480218	0.126789	0.146548
Σ	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Finally this matrix can be summarized in a sequence logo as shown above.

3.5.2. Chow-Liu Tree

The Chow-Liu tree (CLTree) is a single parent-node, maximum spanning tree to approximate optimally an n -dimensional discrete probability distribution by a product of second-order distributions (C. K. Chow & Liu, 1968).

The calculation of a score for a sequence X under the CLTree model can be formalized as:

$$p(X|M) = p(x_{root}) \prod_i^L p(x_i|x_{parent_i})$$

with

$p(x_{root})$ denoting for the single marginal of the root nucleotide

$p(x_i|x_{parent_i})$ denoting for the conditional probability of x_i given x_{parent_i} representing a child parent relationship with x_{parent_i} being the parent and x_i the child.

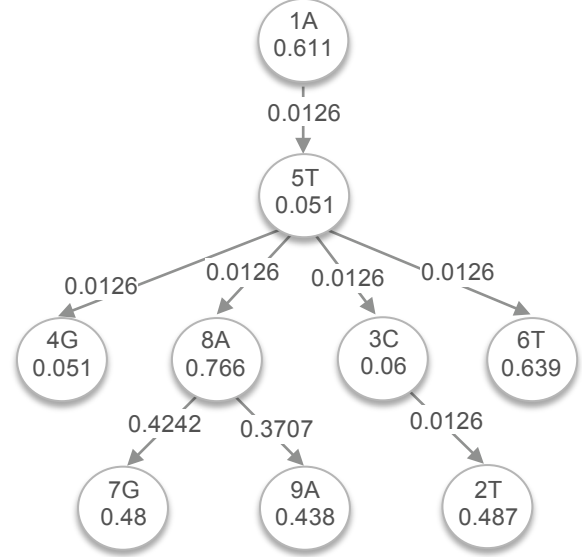
L denoting for the length of the sequence

If the CLTree shown below has been learned based on the training data (described in more detail below) the calculation of the CLTree score is performed as follows:

$$X = ATCGTTGAA$$

$$p(X|CLTree) = p(A_1) \cdot p(T_2|C_3) \cdot p(C_3|T_5) \cdot p(G_4|T_5) \cdot p(T_5|A_1) \cdot p(T_6|T_5) \cdot p(G_7|A_8) \cdot p(A_8|T_5) \cdot p(A_9|A_8)$$

$$p(X|CLTree) = p(A_1) \cdot \frac{p(T_2, C_3)}{p(C_3)} \cdot \frac{p(C_3, T_5)}{p(T_5)} \cdot \frac{p(G_4, T_5)}{p(T_5)} \cdot \frac{p(T_5, A_1)}{p(A_1)} \cdot \frac{p(T_6, T_5)}{p(T_5)} \cdot \frac{p(G_7, A_8)}{p(A_8)} \cdot \frac{p(A_8, T_5)}{p(A_8)} \cdot \frac{p(A_9, A_8)}{p(A_8)}$$



$$p(X|CLTree) = 0.611 \cdot \frac{0.0126}{0.051} \cdot \frac{0.0126}{0.051} \cdot \frac{0.0126}{0.051} \cdot \frac{0.0126}{0.611} \cdot \frac{0.0126}{0.051} \cdot \frac{0.4242}{0.766} \cdot \frac{0.0126}{0.051} \cdot \frac{0.3707}{0.766}$$

$$p(X|CLTree) = 3.058e - 06$$

Training the Chow-Liu Tree

The problem is to find an optimum set of $n - 1$ first order dependence relationship among the n variables. It has been shown, that this methods, when applied to empirical observations from an unknown distribution of tree dependence, is the maximum likelihood estimate of the distribution (C. K. Chow & Liu, 1968).

The main measure to learn the tree model is the mutual information (MI), that is a measure quantifying the mutual dependence of two random variables (Zare-Mirakabad, Ahrabian, Sadeghi, Nowzari-Dalini, & Goliaei, 2009). Intuitively spoken, it measures for instance how much knowing of nucleotide “A” on position i reduced the uncertainty about nucleotide “T” on position j .

In order to calculate the mutual information paired and single marginals are required.

If X is a 9-mer in the training input set, then $\omega(X)$ is the weighted count of 9-mer X . The calculation of the weighted count $\omega_{i,j}$ of a nucleotide pair can be formally expressed as

$$\omega_{i,j}^{nt_1,nt_2} = \sum_{\substack{X: X_i=nt_1 \\ X_j=nt_2}}^{\#X} \omega(X)$$

with

$nt_1 \in \{A, G, T, C\}$ denoting for the different nucleotides on position i

$nt_2 \in \{A, G, T, C\}$ denoting for the different nucleotides on position j

$i, j = (1, 2, \dots, L)$ denoting for the positions of the single nucleotides within a motif of length L

$\omega(X)$ denoting the weighted count of a 9-mer within the training input set

$\#X$ denoting the number of sequences in the training input set

According to the PWM a pseudocount ps is added, so that the probability $p(x_i, x_j)$ of a nucleotide pair (in the following also referred as paired marginal) is calculated by

$$p(x_i = nt_1, x_j = nt_2) = \frac{\omega_{i,j}^{nt_1,nt_2} + \frac{ps}{16}}{\sum_{nt'_1, nt'_2} \omega_{i,j}^{nt'_1,nt'_2} + ps} \quad \text{with } \sum_i \sum_j p(x_i, x_j) = 1$$

The single marginals for the CLTree are exclusively deduced from the pair-matrix, formalized as follows:

$$p(x_i) = \sum_{x_j} p(x_i, x_j)$$

with

$p(x_i)$ denoting for the single marginal and $p(x_i, x_j)$ denoting for the paired marginals holding $p(x_i)$

If no pseudocount is added these single marginals are correlating with those from the PWM. However, depending on the pseudocount chosen, the single nucleotide probabilities of the tree model can slightly differ from those of the PWM.

Once having the paired and single marginals the mutual information can be calculated according to

$$MI(i, j) = \sum_{i,j}^L p(x_i, x_j) \log \left(\frac{p(x_i, x_j)}{p(x_i)p(x_j)} \right)$$

with

$MI(i, j)$ denoting for the non-negative mutual information of the position pair i, j

$p(x_i, x_j)$ denoting for the probability of the nucleotide pair x, y on position pair i, j

$p(x_i)$ denoting for the probability of nucleotide x on position i

$p(x_j)$ denoting for the probability of nucleotide x on position j

L denoting for the length of sequence X

The maximum likelihood CLTree $(\{X_i\}, E)$ maximizes

$$\sum_{(i,j) \in E} MI(i,j)$$

where E denotes for all edges of the tree and can be found via a maximum weight spanning tree on a complete graph with edge weights $MI(i,j)$.

Starting from the input dataset, the corrected normalized intensity of the motif-comprising probe is summarized for all possible pairs and position.

Table 7: Illustration of the initial step for Tree training. The weight for each pair of nucleotides is calculated by summing up the corrected normalized intensities of the respective probe over all probes in the input training set. For example the nucleotide pair AG occurs on position 1 and 3 (A.G.....) in probes with a corrected normalized intensity of 0.0019 and 0.0017.

Position pairs	12	13	14	...	85	86	87	89
AA
AT	0.0019 +0.0017
AC	0.0019 ...
AG	0.0019 +0.0017 ...	0.0019 +0.0017	0.0017 ...
TA	0.0016
...
GG

The resulting matrix is corrected for unobserved events and to prevent numerical instabilities by adding a pseudocount, followed by a normalization step over all position pairs.

Table 8: Sum of the weighted paired counts over all probes in the training set. The sum of the weight for each nucleotide pair is corrected by adding a pseudocount ps (lower table). The weighted corrected counts are normalized to achieve relative frequencies or probabilities for nucleotide pairs (paired marginals). The table is continued on the following page.

Position pairs	12	13	14	...	85	86	87	89
AA	0.3609	0.0070	0	...	0	0.3853	0.1334	1.7725
AT	0.3634	0.0646	2.5970	...	0	2.5347	0.7788	0.7885
AC	0.2632	0.0479	0.1783	...	0	0.6082	0.5911	0.5854
AG	1.7878	2.6558	0	...	3.5407	0.0125	2.0375	0.3943
TA	0.0743	0	0	...	0	0	0	0.0089
...
GC	0.0344	0	0.0258	...	0	0.0028	0.1590	0.1087
GG	0.4891	0.6801	0	...	0.3775	0	0.0811	0.0780

MATERIAL AND METHODS

Position pairs	12	13	14	...	85	86	87	89
AA	0.3609+ <i>ps</i>	0.0070+ <i>ps</i>	0+ <i>ps</i>	...	0+ <i>ps</i>	0.3853+ <i>ps</i>	0.1334+ <i>ps</i>	1.7725+ <i>ps</i>
AT	0.3634+ <i>ps</i>	0.0646+ <i>ps</i>	2.5970+ <i>ps</i>	...	0+ <i>ps</i>	2.5347+ <i>ps</i>	0.7788+ <i>ps</i>	0.7885+ <i>ps</i>
AC	0.2632+ <i>ps</i>	0.0479+ <i>ps</i>	0.1783+ <i>ps</i>	...	0+ <i>ps</i>	0.6082+ <i>ps</i>	0.5911+ <i>ps</i>	0.5854+ <i>ps</i>
AG	1.7878+ <i>ps</i>	2.6558+ <i>ps</i>	0+ <i>ps</i>	...	3.5407+ <i>ps</i>	0.0125+ <i>ps</i>	2.0375+ <i>ps</i>	0.3943+ <i>ps</i>
TA	0.0743+ <i>ps</i>	0+ <i>ps</i>	0+ <i>ps</i>	...	0+ <i>ps</i>	0+ <i>ps</i>	0+ <i>ps</i>	0.0089+ <i>ps</i>
...
GC	0.0344+ <i>ps</i>	0+ <i>ps</i>	0.0258+ <i>ps</i>	...	0+ <i>ps</i>	0.0028+ <i>ps</i>	0.1590+ <i>ps</i>	0.1087+ <i>ps</i>
GG	0.4891+ <i>ps</i>	0.6801+ <i>ps</i>	0+ <i>ps</i>	...	0.3775+ <i>ps</i>	0+ <i>ps</i>	0.0811+ <i>ps</i>	0.0780+ <i>ps</i>

Position pairs	12	13	14	...	85	86	87	89
AA	0.08555	0.01404	0.01262	...	0.01262	0.09047	0.03957	0.37074
AT	0.08604	0.02567	0.53732	...	0.01262	0.52473	0.16998	0.17194
AC	0.06580	0.02230	0.04865	...	0.01262	0.13551	0.13204	0.13090
AG	0.37383	0.54921	0.01262	...	0.72800	0.01515	0.42428	0.09228
TA	0.02763	0.01262	0.01262	...	0.01262	0.01262	0.01262	0.01442
...
GC	0.01956	0.01262	0.01784	...	0.01262	0.01318	0.04475	0.03459
GG	0.11144	0.15003	0.01262	...	0.08890	0.01262	0.02901	0.02837
Σ	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Next, based on the pairwise marginals, the mutual information is calculated.

Table 9: Mutual information of a 9-mer. The maximum *MI* of each position pair is highlighted in grey.

<i>MI</i>	1	2	3	4	5	6	7	8	9
1	0	0.01645	0.04129	0.04371	0.05931	0.01610	0.00547	0.03361	0.00248
2	0.01645	0	0.04237	0.02752	0.03849	0.01263	0.00467	0.02032	0.00227
3	0.04129	0.04237	0	0.12684	0.15881	0.05965	0.02555	0.10840	0.01050
4	0.04371	0.02752	0.12684	0	0.15199	0.06056	0.02903	0.10386	0.01047
5	0.05931	0.03849	0.15881	0.15199	0	0.07882	0.03688	0.13229	0.01560
6	0.01610	0.01263	0.05965	0.06056	0.07882	0	0.02498	0.05183	0.00593
7	0.00547	0.00467	0.02555	0.02903	0.03688	0.02498	0	0.04733	0.01376
8	0.03361	0.02032	0.10840	0.10386	0.13229	0.05183	0.04733	0	0.01989
9	0.00248	0.00227	0.01050	0.01047	0.01560	0.00593	0.01376	0.01989	0

The maximum *MI* is only used to build the tree structure. When it comes to the tree score estimation the calculation is based on the single or paired marginals, depending on the sequence to be evaluated.

In the final step the maximum spanning CLTree is detected, based on the *MI* data and by means of the python package PYGRAPH (version 1.8.1-py2.7). The score of sequence *X* is calculated as formalized above.

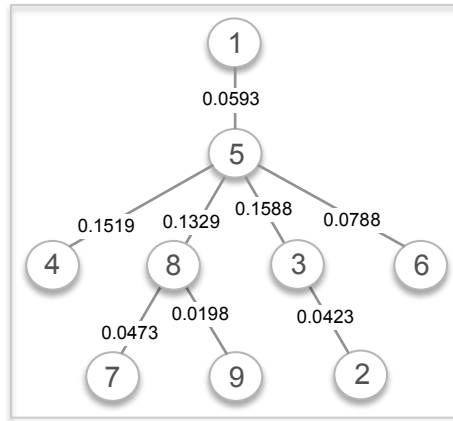


Figure 28: Generation of the tree structure based on mutual information for all possible pairings (upper table). For each nucleotide position the “partner nucleotide” is detected, by searching the maximum pair-MI for this nucleotide, determining one edge in the tree. Finally the tree is composed of all position pairs with the maximum total sum of MIs. The root of the tree can be picked freely.

3.5.3. Ensemble of Trees

The Ensemble of Trees (ET) model is a relatively novel approach, first introduced in the year 2006 by Meila and Jaakkola (Meila & Jaakkola, 2006). In general the ET model is approximating a Markov network and represents a multivariate distribution using a mixture of all possible spanning trees over a complete graph (see Figure 29).

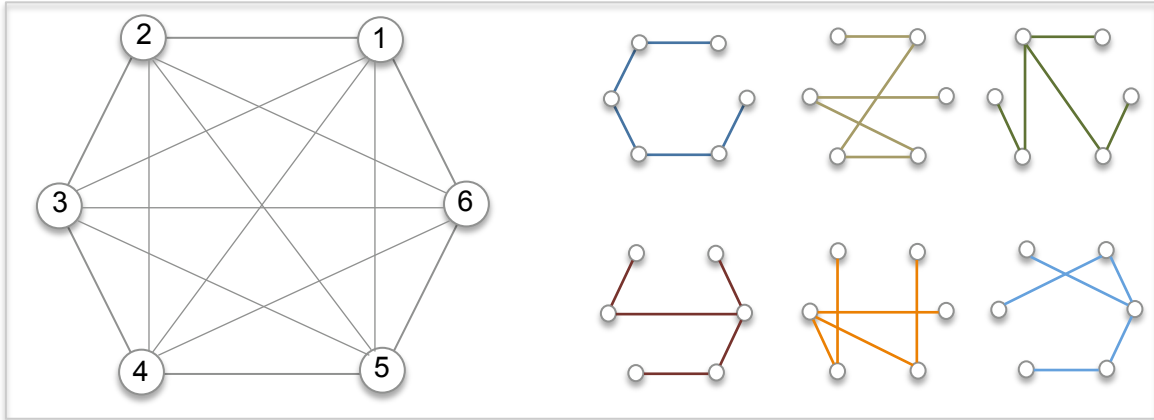


Figure 29: Illustration of a complete graph with 6 nodes and 15 edges (left); Six examples of $n^{(n-2)}$ possible spanning trees summarized by the ET model (right).

Thus, the ET model is a probability mixture model, wherein each component is a probabilistic tree that is able to capture high-order dependencies within the input data, here sequence-nucleotides. The ET model tends to fit data by a number of spanning trees being less restrictive as the single maximum spanning tree described before (Lin, Zhu, Lee, & Taskar, 2009).

Considering a genomic sequence with $X = (x_1, \dots, x_L)$ of length L , the probability of X in the ET model is the summation over all possible spanning trees.

$$p(X) = \sum_T p(T)p(X|T)$$

with

$p(T)$ denoting the mixing weight (prior probability) of each spanning tree T

$p(X|T)$ denoting the probability of a sequence X under a spanning tree T

Even if it has to sum over super-exponentially many trees, the ET model is computable in a closed form.

First, the mixing prior $p(T)$ is parameterized as

$$p(T) = \frac{1}{Z} \prod_{e_{i,j} \in T} \beta_{e_{i,j}}$$

with

$Z = \sum_T \prod_{e \in T} \beta_e$, the partition function of the model that sums over all spanning trees.

$\beta = \{\beta_e \geq 0\}$, a parameter to be chosen for all edges e of the complete graph of X .

The parameter β is not predefined and can be chosen, based on the question to be addressed. In this work β is assigned by $e^{\alpha \times MI(i,j)}$ for fixed α (to be learned from data).

The partition function in turn can be solved by means of the general matrix theorem (Meila & Jaakkola, 2006), an extension of the matrix theorem, if the parameter β is not binary and is non-negative.

The closed form of the partition function $Z = \det[Q(\beta)]$ is the determinant of the first $L-1$ rows and columns of the Laplacian matrix $M(\beta)$ ($L \times L$) given by

$$M_{i,j} = \begin{cases} -\beta_{i,j} & \text{if } i \neq j \text{ (off diagonal)} \\ \sum_k \beta_{i,k} & \text{if } i = j \text{ (diagonal)} \end{cases}$$

Since the parameter β has been chosen to be determined by the mutual information, tuned by the parameter α , the Laplacian matrix MI_j has the following shape, with $\beta_{i,j} = e^{\alpha \times MI(i,j)}$.

		i								
		1	2	3	4	5	6	7	8	9
j	1	$\Sigma\beta_{ij}$	$-\beta_{12}$	$-\beta_{13}$	$-\beta_{14}$	$-\beta_{15}$	$-\beta_{16}$	$-\beta_{17}$	$-\beta_{18}$	$-\beta_{19}$
	2	$-\beta_{21}$	$\Sigma\beta_{ij}$
	3	$-\beta_{31}$...	$\Sigma\beta_{ij}$
	4	$-\beta_{41}$	$\Sigma\beta_{ij}$
	5	$-\beta_{51}$	$\Sigma\beta_{ij}$
	6	$-\beta_{61}$	$\Sigma\beta_{ij}$
	7	$-\beta_{71}$	$\Sigma\beta_{ij}$
	8	$-\beta_{81}$	$\Sigma\beta_{ij}$...
	9	$-\beta_{91}$	$\Sigma\beta_{ij}$

Thus, the respective matrix $Q(\beta) = M(\beta)$ ($L-1 \times L-1$) is determined by

		i								
		1	2	3	4	5	6	7	8	9
j	1	$\Sigma\beta_{ij}$	$-\beta_{12}$	$-\beta_{13}$	$-\beta_{14}$	$-\beta_{15}$	$-\beta_{16}$	$-\beta_{17}$	$-\beta_{18}$	$-\beta_{19}$
	2	$-\beta_{21}$	$\Sigma\beta_{ij}$
	3	$-\beta_{31}$...	$\Sigma\beta_{ij}$
	4	$-\beta_{41}$	$\Sigma\beta_{ij}$
	5	$-\beta_{51}$	$\Sigma\beta_{ij}$
	6	$-\beta_{61}$	$\Sigma\beta_{ij}$
	7	$-\beta_{71}$	$\Sigma\beta_{ij}$
	8	$-\beta_{81}$	$\Sigma\beta_{ij}$...
	9	$-\beta_{91}$	$\Sigma\beta_{ij}$

Second, the ET model assumes that the single and paired marginals p are shared over all trees, so that the observation model $p(X)$ can be parameterized as:

$$p(X|T) = \prod_{e_{i,j} \in T} \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \prod_{i=1}^L p(x_i)$$

with

$p(x_i, x_j)$ denoting for the pairwise marginals

$p(x_i)$ denoting for the single marginals

L denoting for the length of sequence X

In summary the likelihood of X can be written as:

$$\begin{aligned}
 P(X) &= \sum_T \frac{1}{Z} \prod_{e_{i,j} \in T} \beta_{e_{i,j}} \prod_{e_{i,j} \in T} \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \prod_{i=1}^L p(x_i) \\
 &= p(X|PWM) \frac{1}{Z} \sum_T \prod_{e_{i,j} \in T} \beta_{e_{i,j}} w_{i,j}(x) \\
 &= p(X|PWM) \frac{\det [Q(\beta \otimes w(x))]}{\det [Q(\beta)]}
 \end{aligned}$$

with

$p(X|PWM)$ denoted the score of sequence X under the PWM model

$$w_{i,j}(x) = \frac{p(x_i, x_j)}{p(x_i)p(x_j)}$$

$\det[Q(\beta)]$ denotes for the determinant of the truncated Laplacian matrix $M(\beta)$ ($(L-1) \times (L-1)$)

L denoting for the length of sequence X

\otimes denotes pairwise multiplication

Training of α and calculation of an ET score

For the ensemble of tree score calculation different parameters are needed:

- Paired marginals (already gained during CLTree learning)
- Single marginals (already gained during CLTree learning)
- Parameter β chosen to be determined by $e^{\alpha \times MI(i,j)}$.
- α – a parameter to tune the model fitting, by amplifying MI -differences

As one can take from above, all sequence dependent variables needed for the ET model are already generated during the CLTree model learning, except α . This parameter has rather a “tuning” function and has to be determined empirically, randomly or by learning. Since no *a priori* information is available, how to chose α best, it has been decided to determine α empirically. For each α chosen, an ET model screening has been applied on the top 800 probe sequences of the training array, followed by the production of 30 Receiver Operating Characteristic (ROC) curves and a summary AUC-profile over all ROC curves. This procedure is in principle described in more detail in section 3.6. From the α -specific AUC-profile, the sum over all AUC-values is generated. The α providing the maximal sum over all AUC-values is chosen as “tuning” factor for the respective TF. The range for α has been chosen intuitively between 0 and 250. Within this range, α , starting from 0, is iteratively increased with a step size of 10. Correspondingly, all over 26 different α have been considered.

Calculation of an ET Score

For $X = ATCGTTGAA$ the ET score is calculated as follows:

1. Calculation of $p(X|PWM)$

$$\begin{aligned} p(X|PWM) &= p(A_1) \times p(T_2) \times p(C_3) \times p(G_4) \times p(T_5) \times p(T_6) \times p(G_7) \times p(A_8) \times p(A_9) \\ &= 0.611237 \times 0.148949 \times 0.060187 \times \dots \times 0.480218 \times 0.765885 \times 0.437807 \\ &= 1.415e-6 \end{aligned}$$

2. Calculation of $\det[Q(\beta)]$

First the different items of the matrix $M(\beta)$ ($L \times L$) have to be calculated. For example $-\beta_{12}$ is calculated with $\alpha=50$ by $-e^{\alpha \times M(1,2)} = -e^{50 \times 0.01645} = -2.276$. Then the last column and last row are deleted, so that $Q(\beta) = M(\beta)$ ($L-1 \times L-1$).

The determinant of $Q(\beta)$ has been calculated by means of the python package NUMPY (version 1.6.1). Following the example above $\det[Q(\beta)]$ is $7.795e+17$ ($\alpha=50$).

3. Calculation of $\det [Q(\beta \otimes w(x))]$, with $w_{i,j}(x) = \frac{p(x_i, x_j)}{p(x_i)p(x_j)}$ and $\beta_{i,j} = e^{\alpha \times M(i,j)}$:

The respective Laplacian matrix $M(\beta \otimes w(x))$ is constructed as follows:

		i					
		1	2	3	4	5	...
j	1	$\sum \frac{\theta_{i,j}(x_i, x_j)}{\theta_i(x_j)\theta_i(x_j)} \beta_{ij}$	$-\frac{p(A_1, T_2)}{p(A_1)p(T_2)} \beta_{12}$	$-\frac{p(A_1, C_3)}{p(A_1)p(C_3)} \beta_{13}$	$\frac{p(A_1, G_4)}{p(A_1)p(G_4)} \beta_{14}$	$-\frac{p(A_1, T_5)}{p(A_1)p(T_5)} \beta_{15}$...
	2	$\frac{p(T_2, A_1)}{p(T_2)p(A_1)} \beta_{21}$	$\sum \frac{\theta_{i,j}(x_i, x_j)}{\theta_i(x_j)\theta_i(x_j)} \beta_{ij}$
	3	$-\frac{p(C_3, A_1)}{p(C_3)p(A_1)} \beta_{31}$...	$\sum \frac{\theta_{i,j}(x_i, x_j)}{\theta_i(x_j)\theta_i(x_j)} \beta_{ij}$
	4	$\frac{p(G_4, A_1)}{p(G_4)p(A_1)} \beta_{41}$	$\sum \frac{\theta_{i,j}(x_i, x_j)}{\theta_i(x_j)\theta_i(x_j)} \beta_{ij}$
	5	$-\frac{p(T_5, A_1)}{p(T_5)p(A_1)} \beta_{51}$	$\sum \frac{\theta_{i,j}(x_i, x_j)}{\theta_i(x_j)\theta_i(x_j)} \beta_{ij}$...
...	

In order to get $Q(\beta \otimes w(x))$ the last column and last row are deleted and the determinant is calculated by means of NUMPY (version 1.6.1).

For the example sequence, the ET score is

$$p(X|PWM) \frac{\det [Q(\beta \otimes w(x))]}{\det [Q(\beta)]} = 1.415e - 6 \frac{\det [Q(\beta \otimes w(x))]}{7.795e+17} = 3.30132e - 06$$

3.6. Model Validation

As mentioned before, nearly all UniPROBE TFs are measured by two independently designed arrays, providing the same sequence properties, with regard to variability and completeness. Thus, the second array represents an appropriate validation set to test the goodness of model learning or predictive power. In more detail, the validation starts with the screening for TFBS in the validation probes, wherein the screening follows the procedure described in section 3.4. After the binding probabilities for the highest scoring motif within the probes have been found the screening results are sorted by the intensity of the probe.

Next a threshold is selected, based on the number of probes, to classify “good” and “bad” sites. For example a threshold of 2 means, that the 2 highest-ranking probes and respective motifs are classified as “positive” and all the others as “negative” (see Figure 30).

In order to calculate ROC curves, a classical method to compare models (Fawcett, 2006), the classified dataset is now sorted by the respective model score. With the new order of probes and the “knowledge” of the “true” positive (“good” sites) and negative ones (“bad” sites) based on the original intensity, it is possible to calculate *false* and *true positive rates* (*FPR* and *TPR*), or sensitivity and specificity respectively, which are the parameters determining a ROC curve.

The statistical measure sensitivity and specificity describe the performance of a binary classification. Both measures are directed to the proportion of correct classifications. Specificity (or true negative rate; 1-FPR) measures the amount of as negative classified items being really negative, and sensitivity (TPR) respectively the amount of positives correctly being identified as positive. A perfect classification algorithm or a perfect prediction tool would provide 100% for both, sensitivity and specificity.

In more detail, sensitivity is defined as

$$sensitivity = \frac{TP}{P} = \frac{TP}{(TP + FN)}$$

with

True positives (TP): number of “good” sites to find above the intensity threshold

Positives (P): true positives + false negatives (FN), equals the size of the median intensity interval

The specificity is defined as

$$specificity = \frac{TN}{N} = \frac{TN}{(TN + FP)}$$

with

False positives (FP): number of “good” sites to find below of the intensity threshold

Negatives (N): false positives + true negatives (TN), equals the sum of the number of “good” sites below the intensity threshold and the “bad” sites detected below the intensity threshold

If a point lies along the diagonal line of a ROC curve (also called line of no-discrimination) a completely random guess is given.

The procedure of classifying the validation set based on intensity values is repeated 50 times, with increasing thresholds (see Figure 30). The step size for the increase of the threshold is determined dynamically by an algorithm ensuring that (i) the validation sample holds in each iteration two classes with at least one negative or positive set and (ii) the first fourth of the validation set is sampled with a higher resolution than the rest, as in the middle range intensities are getting fuzzy.

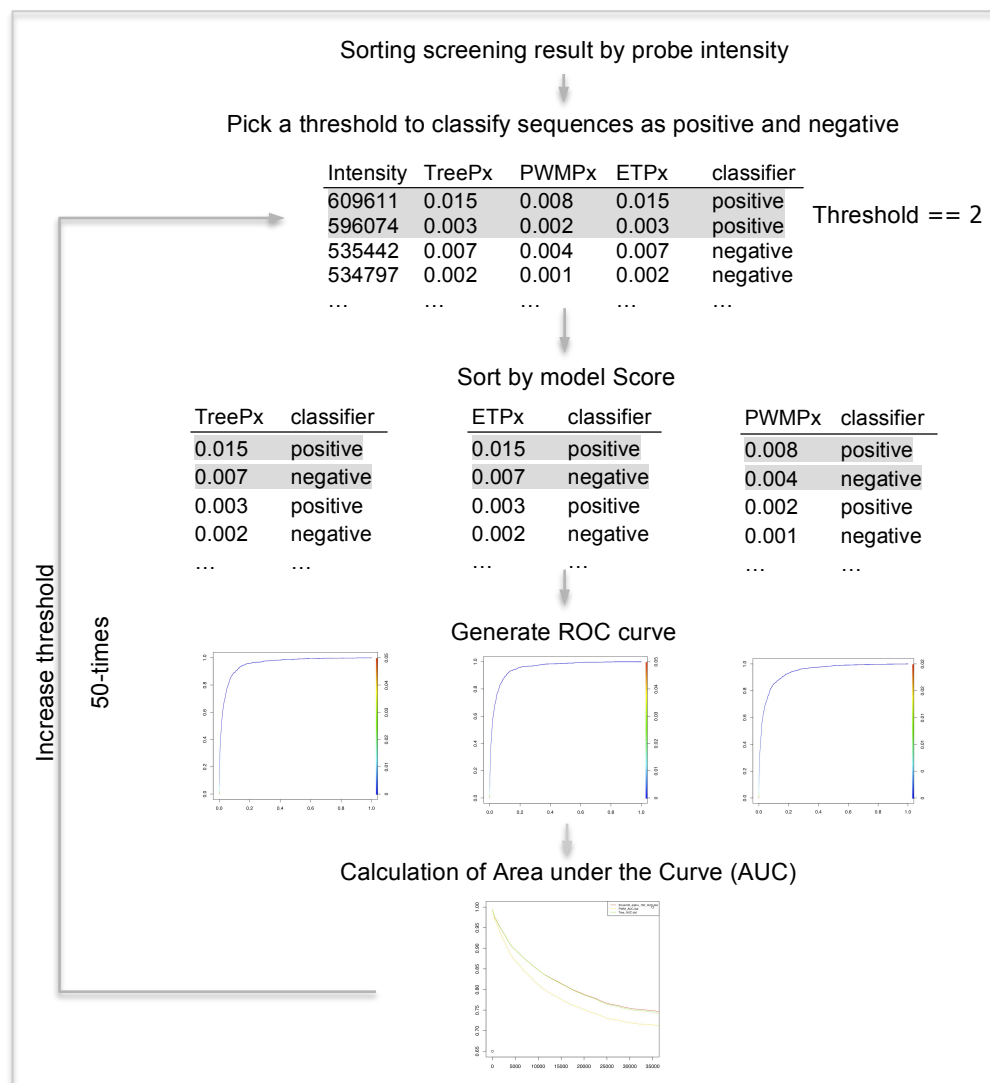


Figure 30: Schematic illustration of the model validation procedure.

In order to compare the performance of the different models, the Area Under the Curve (AUC, in this work also called AUC-profile) is calculated for each ROC Curve and respectively for each threshold chosen. The AUC of a classifier is equivalent to the probability that the

classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. This is equivalent to the Wilcoxon test of ranks (Fawcett, 2006).

An AUC-value close to 1 corresponds to a true positive rate of 1 and a false positive rate of 0, meaning that the classification, based on the “gold standard” has been completely recovered by the model score considered. In contrary an AUC-value of 0.5 represents a ROC curve where all points are located along the diagonal line, displaying randomness. Thus the higher the AUC-curve the better the model will rank a randomly chosen “good” site over a randomly chosen “bad” site.

In order to get an impression for different values represented by the AUC profile, in the following (see Figure 31) different ROC curves are shown (R-package ROCR (Sing, Sander, Beerenwinkel, & Lengauer, 2005)). By means of an AUC-profile the different models can be represented in one single graph providing the possibility to detect which model performed best and by which degree of accuracy. In particular for the ET model this representation is effective to easily choose the best tuning parameter α described above during the α -training procedure, following the same principle.

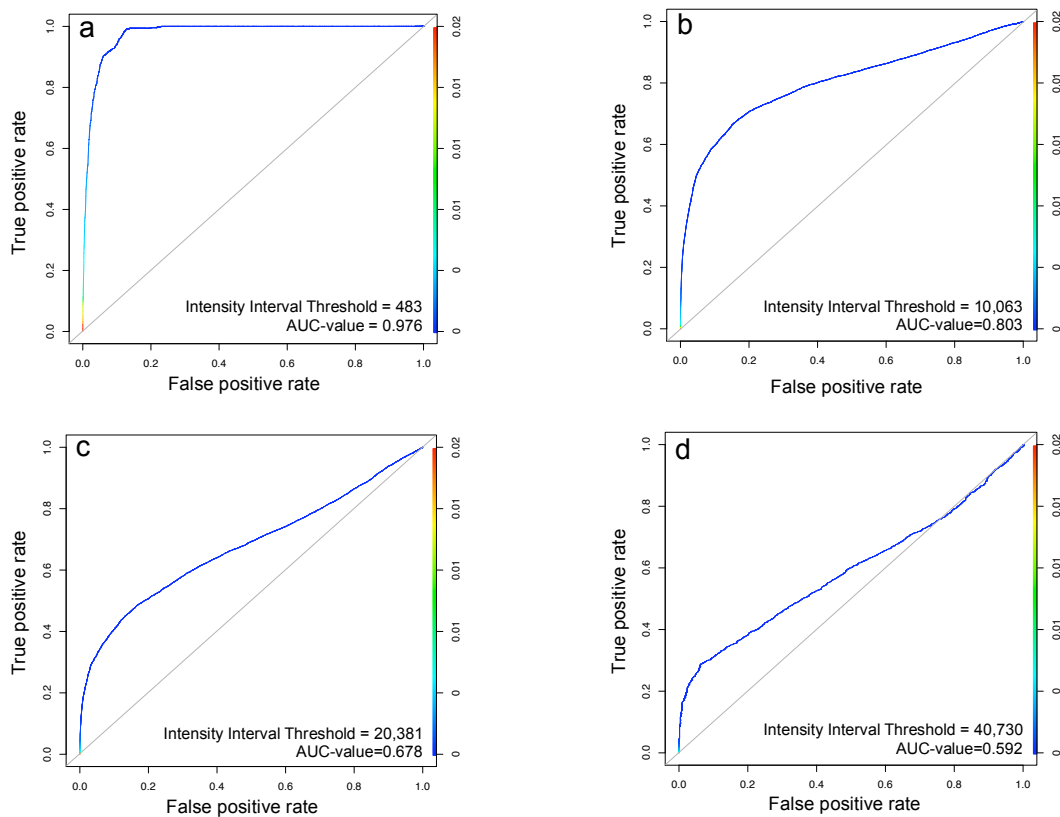


Figure 31: ROC curves for 4 distinctive intensity thresholds. a) ROC curve for a low intensity threshold of 483, the AUC-value is 0.976 (maximum is 1), b) ROC curve for a intensity threshold in the upper third of 10,063, the AUC-value is 0.803, c) ROC curve for a middle ranged intensity threshold of 20,381, the AUC-value is 0.678 (maximum is 1), d) ROC curve for a high intensity threshold of 40,730, the AUC-value is 0.592 (0.5 corresponds to randomness)

3.7. Probabilistic Models in causative SNP Detection in Family Data

In order to evaluate whether the consideration of positional interdependence is advantageous to detect regulatory variants, datasets available from one 1000G (also HapMap) individual have been analyzed. The detection is performed in parallel for all three models described above.

The approach used is motivated by a recently published idea to map ChIP-seq results generated in a child sample on the parental genomes (Rozowsky *et al.*, 2011). If a peak is detected in a child, it is in fact a mixed signal based on the two inherited haplotypes, coming from the mother and the father. Accordingly, by mapping the sequence reads on the parental genomes, it is possible to separate the signal for the two haplotypes. If the reads map better to one of the two parental haplotypes, one could assume that a genetic variant present in different variations in the two haplotypes might be responsible for the difference in read alignment or peak detection.

Starting from the model training, the detected peaks are screened for TFBSs. Besides a simple BS-screening of the peak sequences, the following aspect in detecting causative variants have to be considered:

Does the peak co-locate with a SNP, and if it does is the SNP also co-located with the best BS within the peak?

The underlying assumption is that a SNP or variant is regulatory due to its potential to disrupt a TFBS, which results in an altered gene expression. Here, loss of BS-significance or a significant change in BS-significance will be taken as a measure. However, the variation below a peak can consist of a combination of several SNPs or a combination of a SNP with an indel or only an indel. To evaluate whether the models are able to detect differential binding within the parents as a pre-requirement to detect rSNPs, a reference positive set is generated, by applying Alleleseq, which is measuring differential binding based on significant read count differences. After it has been investigated whether the detected differential binding can be assigned to a causative variant that directly affects the DNA binding of the TF in question, by focusing on the set of true positives with regard to the reference set.

To clarify, in Figure 32 the different steps of the causal variant detection are illustrated.

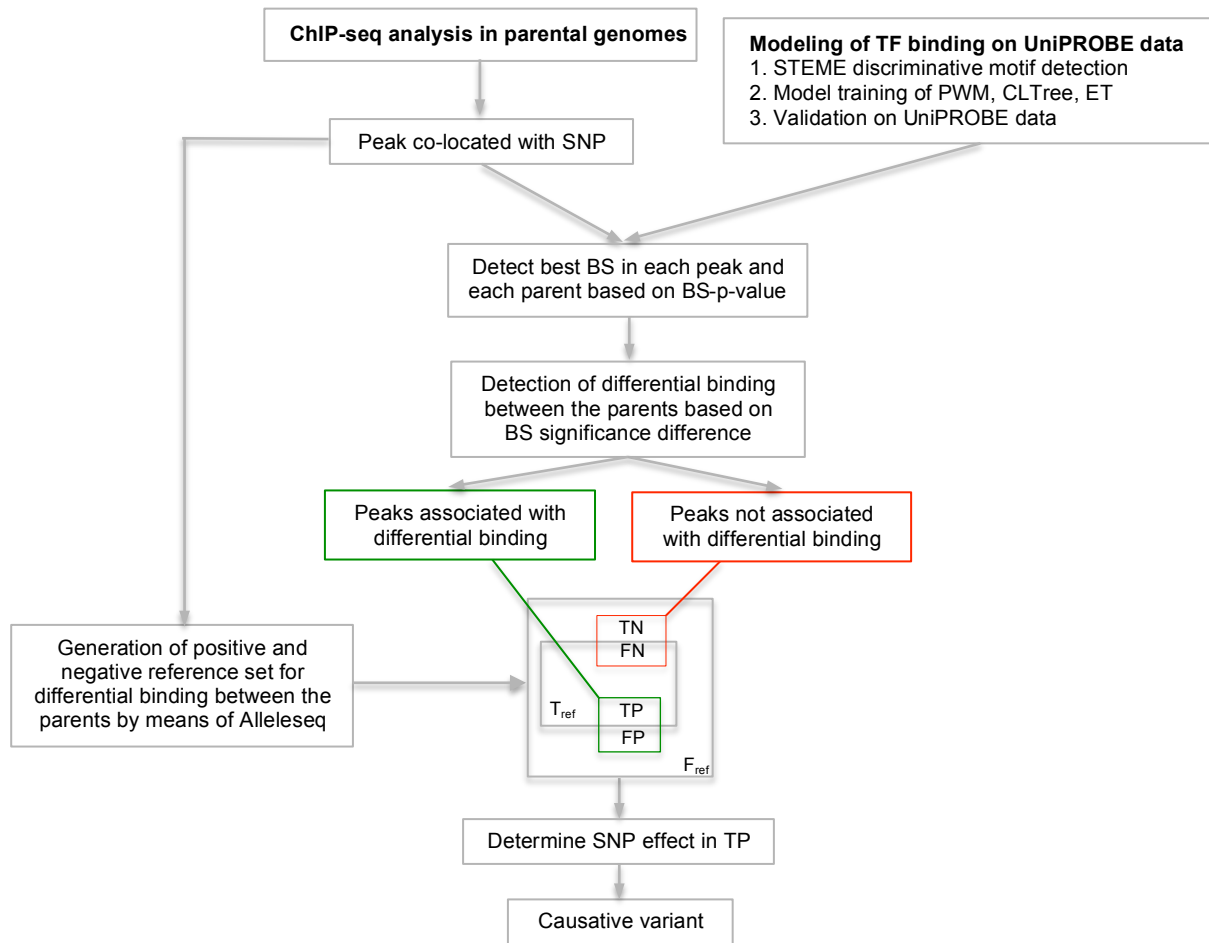


Figure 32: Schematic illustration of the different algorithmic steps performed to detect causal

3.7.1. SNP Data

The underlying basic SNP dataset used has been taken from 1000G, namely the 1000 Genomes Pilot 2 SNP calls. These calls are based on the pilot 2 phase deep coverage whole genome DNA sequence data collected in 2008 achieved by high throughput sequencing on two family trios in the CEPH Utah and HapMap sample collections (released for each trio in March 2010). Accordingly, for these data sets a complete set of heterozygous variants (SNPs, indels, and SVs) exists, which can mostly be phased into maternal and paternal variants by comparing against the parents sequences. This is important for assessing the genome-wide amount of allele-specific behavior, which is severely limited by the number of identified heterozygous SNPs available (Rozowsky *et al.*, 2011).

The DNA used was isolated from lymphoblastoid cell lines. The trio genotype data used in this work are coming from the CEU family trio composed of the individuals NA12892 (mother), NA12891 (father) and NA12878 (daughter) and hold over all chromosomes 3,646,764 SNPs. All variant-coordinates refer to the genome version hg18.

3.7.2. Family Trio Data

The approach to detect regulatory variants in this work is based on differences in TF binding between two parents. Accordingly, a separate analysis of parental genomes within a family trio takes place, requiring the maternal and paternal genome (also referred as haplotypes in the following). A straightforward possibility to generate the same is to apply a tool called *vcf2diploid* described in (Rozowsky *et al.*, 2011). In short, this tool can read in all variations to be considered (like SNPs, indels, and SVs) available for an individual of interest in the form of a variant call format (vcf) file (see Figure 33). The outputs are fasta-sequences for each chromosome for each allelic variant, along with equivalence map files allowing a mapping of nucleotide positions between paternal, maternal, and reference haplotypes. This mapping becomes very important during the variant prediction approach, and can be realized by using the liftOver tool (Fujita *et al.*, 2011) and respective chain-files.

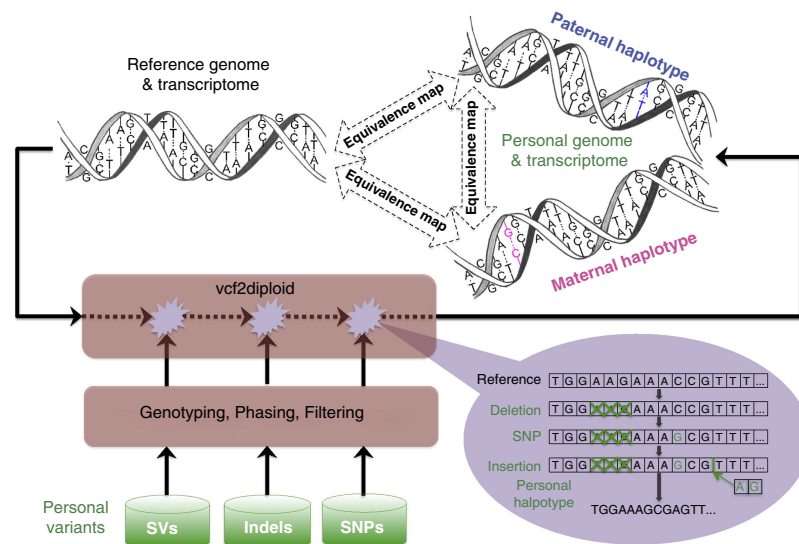


Figure 33: Construction of a personal genome by *vcf2diploid* tool is made by incorporating personal variants into the reference genome. Personal variants may require additional pre-processing, that is, filtering, genotyping, and/or phasing. The output is the two (paternal and maternal) haplotypes of personal genome. Taken from (Rozowsky *et al.*, 2011)

All individual genome processing has been performed already by Gerstein *et al.* in context of Alleleseq (see section 3.7.10) for the CEU family trio used. Thus, the respective data have been directly taken from *alleleseq.gersteinlab.org*⁹. The missing chain files, enabling a liftover from a parental to the reference genome have been generated in-house, by converting provided chain files for a mapping from reference to parental genome. These individual genomes play a significant role in the *de novo* subsequent ChIP-seq analysis, the TFBS screening and the prediction of allele-specific binding events.

⁹ <http://alleleseq.gersteinlab.org/downloads.html>

3.7.3. Selection of TFs

In contrary to the selection of TFs for the first part of the thesis, the selection done here is based on the fact, whether ChIP-seq experiments are available, which

- are provided by the ENCODE consortium (Birney *et al.*, 2007; Rosenbloom *et al.*, 2010)
- provide raw sequence read archive files for a *de novo* analysis
- are generated on the European family trio of the 1000G, namely on the child sample referred as GM12878
- have a shared TF with the UniPROBE database

These criteria applied at the time point the data for the thesis have been generated for eight different TFs, namely EGR1, ETS1, IRF4, MAX, SPI1, SRF, POU2F2 and RXRA.

3.7.4. Determination of Significance of TF Binding

By screening a sequence and generating a model score, as described before, only half of the work is done. What is missing is an indication of its goodness, relevance or significance. A widely used measure to assign significance is to provide a p-value, a parameter informing how probable a finding is due to a random effect. A p-value is about testing a null hypothesis. Thus, when a p-value is small, some evidence exists for rejecting the null hypothesis. In this work the significance of a TFBS has been assigned by an absolute p-value (see Figure 34). The underlying null hypothesis is, that a randomly taken 9-mer sequence belongs to genomic background and is therefore not a binding site.

Therefore, all possible 9-mers are generated by permutation, leading to a set of 4^9 unique items. Then, each 9-mer is scored by means of probabilistic model, in this work PWM, CLTree and ET for the respective TF as described previously. The resulting scores are sorted from the largest to the smallest and for each score a p-value is calculated by

$$p(9 - mer_i) = \frac{\#(Scores \geq Score_i)}{\#Scores_{total}}$$

with

$\#(Scores \geq Score_i)$ denoting for the number of Scores greater than probabilistic model score of 9-mer i
 $\#Scores_{total}$ number of total scores in the permutation set (here equals 262,144)

In the following (see Figure 34) the procedure is schematically illustrated for the CLTree model.

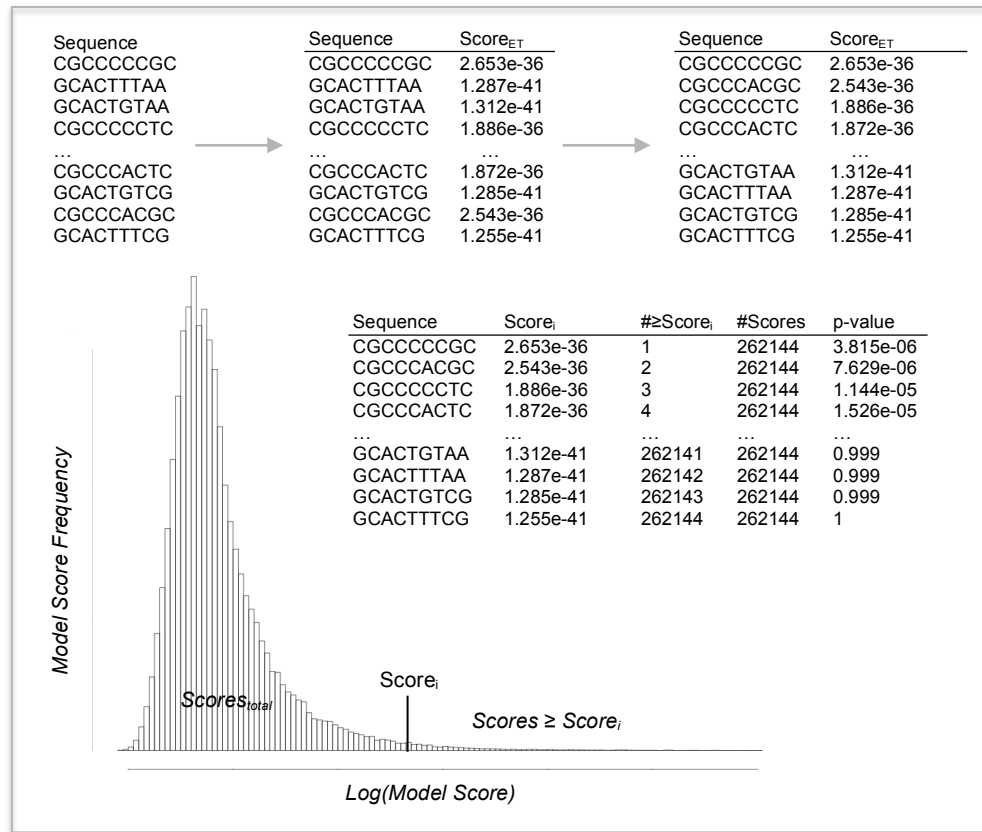


Figure 34: Graphical illustration of the procedure to generate a p-value for TF binding

3.7.5. Assigning Significance to SNP Effect by SNE Distributions

A single nucleotide exchange (SNE) is defined in context of this thesis as a flip of a single nucleotide in a sequence, comparable with a SNP. The sequences considered are all possible 9-mers. The reason to generate this distribution is, according to the previous chapter, to achieve p-values.

By considering all possible sequences and all possible SNEs, chosen uniformly at random from the whole space of sequences and SNEs, the null hypothesis is that a new observation is due to a random event. Here the item to be assigned is not a binding site or sequence anymore, but the effect of a single nucleotide exchange on the TFBS model score or binding affinity respectively.

To follow the approach from above a dataset has been generated holding for all possible 9-mers all possible SNEs. Additionally to a “simple” SNE, also SNEs in the reverse complements for all 9-mers are considered. The aim is to include also those cases, where a SNP might lead to a switch of the potential binding site to the reverse complement. In these cases the difference between two sequences is greater than 1, contrary to the “normal” SNE pairs. Therefore, these events need to be explicitly considered.

Here, p-values are calculated by the log-rank method (Macintyre et al., 2010). In this regard the difference of the negative log of all SNE-paired sequence p-values is calculated, sorted or ranked, and processed as above according to the following formula.

$$p(\text{toggle}_i) = \frac{\#(\Delta p \geq \Delta p_i)}{\# \Delta p_{\text{total}}}$$

with

$\Delta p = \text{abs}(-\log_{10}(p(x)) + \log_{10}(p(y)))$ with $p(x)$ denoting for the p-value of 9-mer x and $p(y)$ denoting for the p-value of 9-mer y , wherein x and y differ only by a single nucleotide (SNE)

$\#(\Delta p \geq \Delta p_i)$ denoting for the number of Δp greater than Δp of 9-mer i

$\# \Delta p_{\text{total}}$ number of total Δp in the SNE set (here equals 7,077,195)

The following Figure 35 should clarify this approach in more detail.

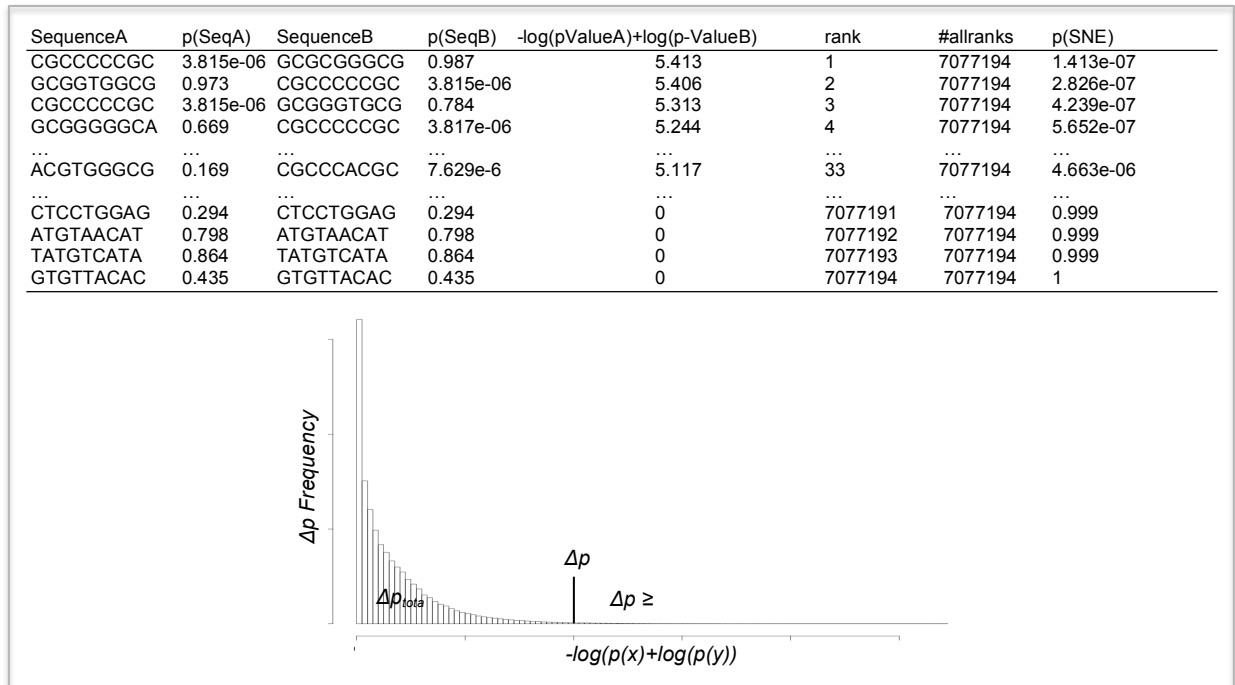


Figure 35: Illustration of the calculation the SNE – or SNP effect - p-value. In the upper table representation one can see an example for the possible case, that a SNE within a sequence leads to a switch from the orientation of a binding site to the reverse complement, but keeping the general binding coordinates (see sequence pair on rank 33).

3.7.6. ChIP-seq Analysis and Peak Detection

The ChIP-seq analysis in this work was in principle performed according to the following scheme illustrated in Figure 36:

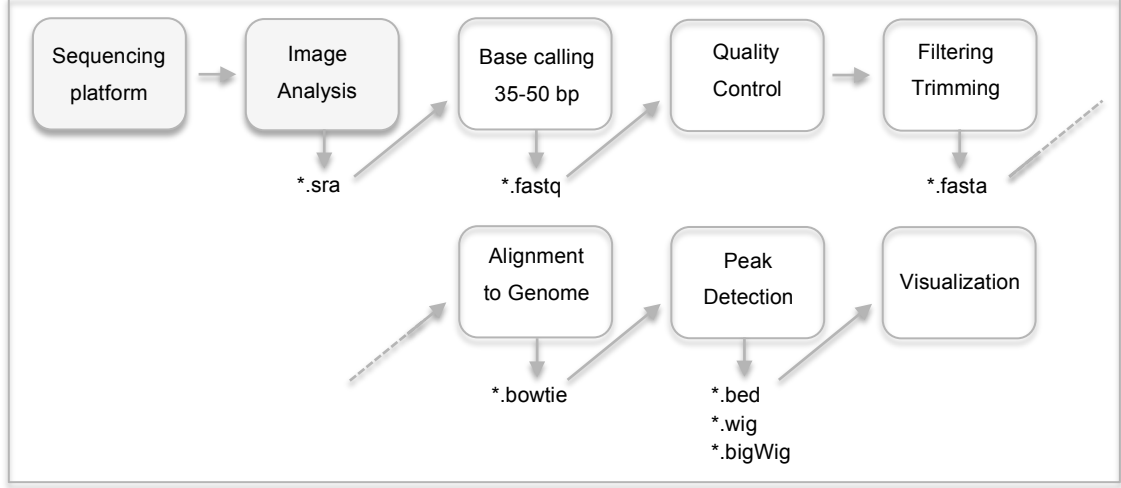


Figure 36: Schematic illustration of the principle ChIP-seq analysis including a peak detection step. The analysis performed starts with the output of the image analysis step, the SRA-files, and the respective base calling.

For all TFs considered, the same INPUT control sample has been used.

3.7.6.1. Sequence Read Preprocessing

The analysis starts with processing the raw sequence reads by means of the sequence read archive (SRA)-toolkit (sratoolkit.2.1.10-mac32). The SRA-files are converted to fastq-files by *fastq-dump*, resulting in raw reads represented in a readable fasta-format plus corresponding quality scores in the Illumina 1.3+ FASTQ-format (see Figure 37).

@SRR351698.878 ILLUMINA-EAS45_45:8:1:13280:998:0:1:1 length=36	→	header
NGTAATTTCTTAGTGACAGAGAGCATATACGTTTA	→	sequence
+SRR351698.878 ILLUMINA-EAS45_45:8:1:13280:998:0:1:1 length=36	→	header for quality scores
BKMHIQTTTQQQQ_____B	→	quality scores

Figure 37: Example FASTQ-format

The Illumina 1.3+ FASTQ variant encodes PHRED scores with an ASCII offset of 64, (PHRED scores from 0 to 62 (ASCII 64–126)) (Cock, Fields, Goto, Heuer, & Rice, 2010). The PHRED score is calculated as follows:

$$Q_{PHRED} = -10 \times \log_{10}(P_e)$$

with P_e denoting the estimated error probability for that base-call.

For example if a base has a PHRED quality score of 40, the chance that this base is called incorrectly is 1 in 10,000 (Ewing, Hillier, Wendl, & Green, 1998).

Next, the produced fastq-files are used for subsequent quality control steps, performed by means of the FASTX and FASTQC toolkit.

- a) *fastx_quality_stats*: scans a FASTQ-file, and produces some statistics about the quality and the sequences in the file
- b) *fastx_artifacts_filter*: removes some sequencing artefacts from FASTA/Q files
- c) *fastq_quality_filter*: removes low-quality sequences from FASTQ files, determined by minimum quality score and minimum percentage of bases that should have such a quality score
- d) *fastqc*: generating some graphical representations for quality checks (see (Cock et al., 2010), Figure 39).

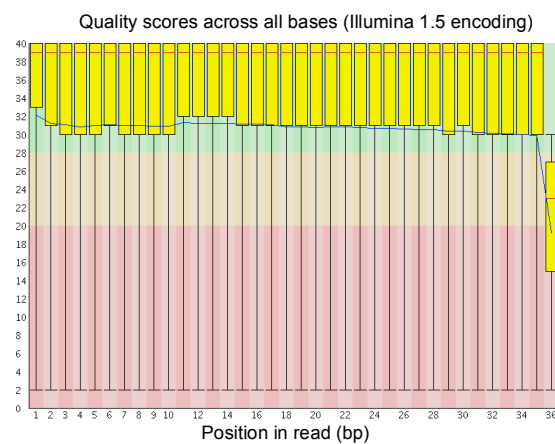


Figure 38: Example output of FASTQC representing the quality score across all bases. According to this result, base 36 shows a comparably bad quality score, so that a trimming of the sequence to base 1-35 would be performed for further processing.

At this point the only user input for the ChIP-seq analysis is requested. In order to decide whether the sequence reads needs to be trimmed for further processing (see Figure above), a visual inspection of FASTQC plots is necessary. In case a trimming is necessary, the sequence reads are trimmed, meaning subsequences are extracted, by *fastx_trimmer*. This step is followed by a new quality control run as already done before, to assure that an improvement has been achieved.

After, a collapsing of identical sequences in the FASTQ-file into a single sequence (while maintaining reads counts, *fastx_collapser*) and the assignment of unique sequence-IDs is performed, followed by the generation of a FASTA-file, serving as sequence alignment input.

3.7.6.2. Alignment of Reads

Once the preprocessing is done, the reads can be aligned to the genome of interest. In this work, the respective genomes are the reference genome (hg18), the maternal and paternal

genome. The program used in this regard is Bowtie (version 0.12.8). Bowtie is a sequence aligner being ultrafast and memory-efficient. It is indexing a genome by using a scheme based on the Burrows-Wheeler transform (BWT) and the Ferragina and Manzini (FM) index (Langmead, Trapnell, Pop, & Salzberg, 2009).

The general method for searching in an FM index is the exactmatch-algorithm of Ferragina and Manzini (Ferragina & Manzini, 2000). In Figure 39 indexing and exact matching is exemplary illustrated.

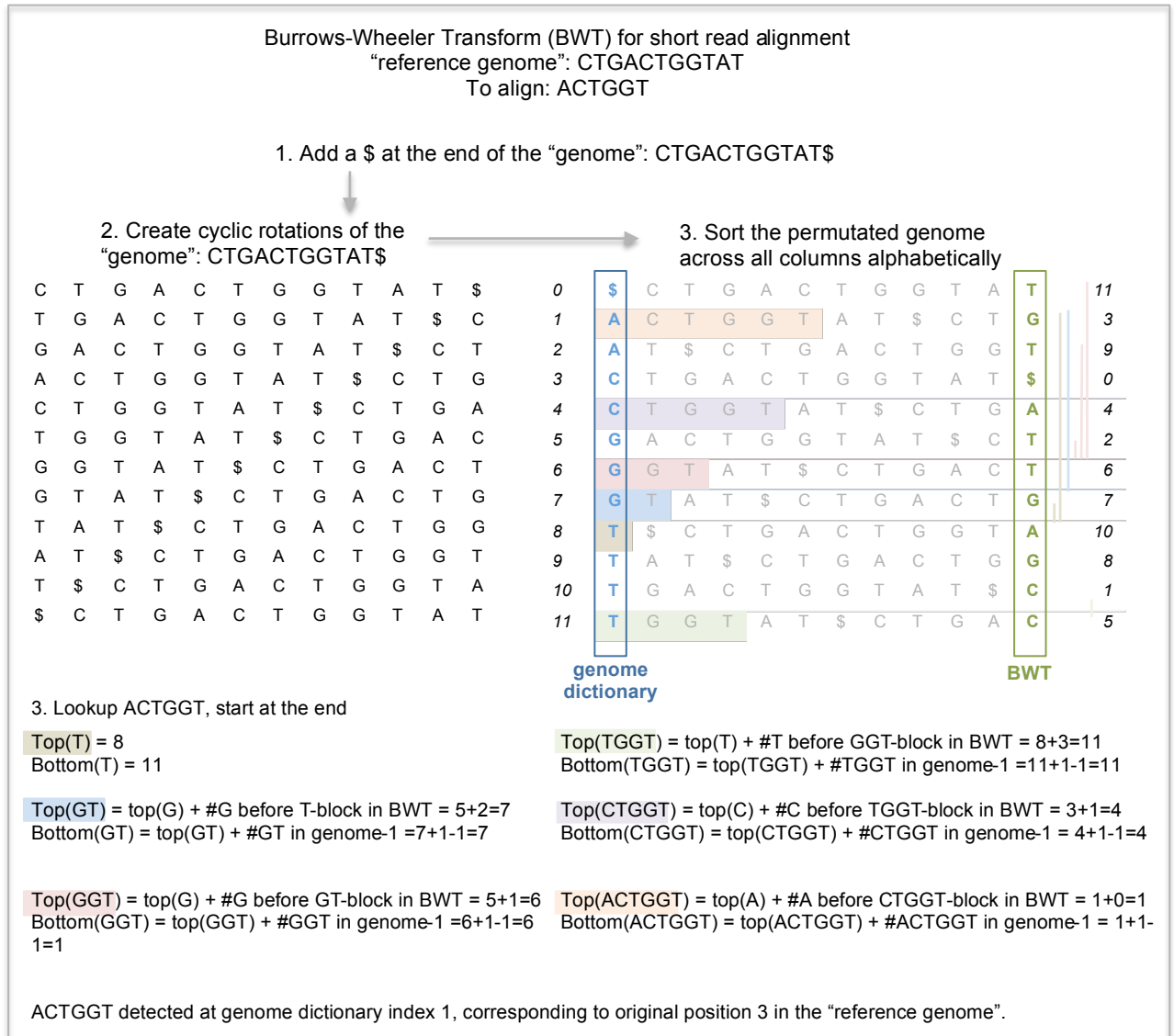


Figure (Ferragina & Manzini, 2000): Schematic illustration of the Burrows-Wheeler Transform (BWT) for short read alignment. The "reference genome" sequence is represented by CTGACTGGTAT and the sequence to align ACTGGT. Following the different steps of the algorithm, the sequence to align will be located to start at position 3 (0-based) of the reference genome".

In Bowtie, two novel extensions are introduced, namely a quality-aware backtracking algorithm that allows mismatches and favors high-quality alignments and a double indexing strategy to avoid excessive backtracking.

For the reference genome an existing public bowtie index has been used¹⁰, while for the maternal and paternal genomes new indexes are generated. Considering mismatches, Bowtie is applied allowing two mismatches, to not loose reads which might be affected by an allelic SNP variant, different from the genome used as reference for alignment.

A Bowtie output has the following format:

```
SRR351698-4599 - chr17 21933837 CACGGGAGCTCT... IIIIIIIIII... 03:A>G,4:A>C
SRR351698-1562 + chr5 99361102 CTACAGGGATGA... IIIIIIIIII... 010:T>G,25:A>G
```

It reports (from the left to the right) the name of the read, the strand where the read has been detected, the name of the reference sequence, the 0-based offset in the forward reference sequence, the sequence as itself, read qualities, the number of alignments of the identical read sequence (without mismatches) and a mismatch descriptor with the format “position”:”referenceBase”>”readBase”, for several mismatches separated by “;”.

3.7.6.3. Peak Detection

The peak detection is performed by means of QuEST (Quantitative Enrichment of Sequence Tags) (Valouev, Johnson, Sundquist, & Medina, 2008). The final aim of a ChIP-seq analysis is to find those regions in the genome that are specifically enriched in DNA fragments or aligned sequence reads. Those regions, showing a high density of aligned reads, are denoted as peaks. The output, a peak detection program reports, is a list of so called “peak calls” determined by genomic location and some statistical measures. QuEST is using a kernel density estimation (KDE) approach as statistical framework, which is a non-parametric way to estimate the probability density function of a random variable. It is data smoothing problem where inferences about the population (here TF binding) are made, based on a finite data sample (here aligned reads) (Parzen, 1962).

Since the sequencing starts strand specifically from one end of the tag towards its middle, those tags in the direct proximity of the TFBS are underrepresented and the reads from the forward and backward strand cluster on opposite sides of the TFBS.

QuEST builds initially two different profiles, one for each strand, wherein the individual density profiles are given by

¹⁰ ftp://ftp.cbcb.umd.edu/pub/data/bowtie_indexes/hg18.ebwt.zip

$$H_{+,-}(i) = \frac{1}{h} \sum_{j=i-3h}^{i+3h} K\left(\frac{j-i}{h}\right) \times C_{+,-}(j)$$

with

h denoting for the kernel density bandwidth (QuEST uses $h=30$ bases, based on visual inspection)

K denoting the Gaussian kernel density function $K(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$

$C_{+,-}(j)$ denotes for the count of 5' read ends at position j for forward (+) and reverse (-) reads

After, the estimation of the distance between the forward and backward profiles is necessary to correctly combine the same. As this distance is an experiment-specific measure, QuEST is using a subset of the input data for its estimation and refers to half of this distance as peak shift (see Figure 38).

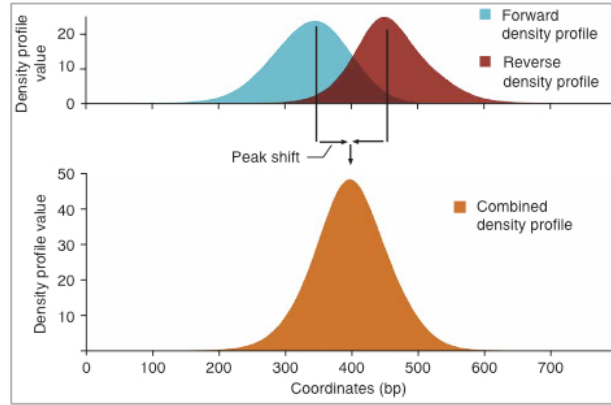


Figure 38: Forward and reverse read density profiles derived from the read data are merged to a combined density profile. Taken from (Valouev et al., 2008)

By means of this peak shift the two profiles are merged and summed to one combined density profile (CDP) by

$$H_{CDP}(i) = H_{+}(i - \lambda) + H_{-}(i + \lambda)$$

with

λ denoting for the peak shift parameter

H_{+} and H_{-} denoting for the forward and reverse strand density profiles

This combining procedure fulfills two key aspects of Quest's ChIP-seq analysis, namely that the forward and reverse profiles are represented by a single peak or classifier and that the local maxima of this classifier estimate for the location of the TFBS.

Within those CDPs, the peak calling - the search for enriched loci - is performed by identifying candidate peaks corresponding to local maxima of the CDP and providing a sufficient enrichment compared to the control data. The expression "sufficient" needs to be determined in more detail by assigning a CDP threshold for peak calling, which in turn can vary strongly

between different experiments. QuEST is performing a kind of calibration procedure, balancing between sensitivity and specificity. The user can add a threshold, without knowing the data in detail. In this calibration procedure, the negative control data are separated in two sets, wherein one serves as a pseudo-ChIP-sample and the other as the respective background signal. Every peak that is predicted in this Pseudo-ChIP-seq sample is considered as a false positive, so that the False Discovery Rate (FDR) of the real peak calling can be calculated as

$$FDR = \frac{\#peaks\ in\ Pseudo - ChIP - seq\ sample}{\#peaks\ in\ real\ ChIP - seq\ sample}$$

Accordingly, the user can set a threshold and determine the FDR.

Finally QuEST outputs for each peak a score, which quantifies the tag enrichment, and the respective genome coordinates. This score, which is kernel density estimation–derived, reports a proportional to the frequency at which the TFBS was present in the sequences library. Thus, the score is reflecting the supporting evidence for a peak.

The reported p-value of QuEST is calculated using a Poisson model, based on the number of tags within a region, the region size and local tag frequency¹¹.

$$p_i = 1 - \sum_{j=0}^{Y_i} \frac{\lambda e^{-\lambda}}{j!}$$

with

Y_i denoting the number of reads falling within region i

λ denoting the number of reads to map the region by random chance $\left(\frac{\text{region size} \times \text{total number of tags}}{\text{effective genome size}}\right)$

The effective genome size is determined as 0.8 x genome size (Nix, Courdy, & Boucher, 2008)

The q-value is calculated by multiple testing correction of the p-value. A conservative multiple testing correction is made, if the Bonferroni (Abdi, 2007) method is applied, simply multiplying each p-value by the number of regions tested (Nix et al., 2008). Both, the q- and p-value represent the statistical significance of regional enrichment, not the enrichment itself.

To consider the special situation in this work, of dealing with different parental haplotypes, the parental aligned reads are filtered before entering the peak detection step according to:

¹¹ [https://groups.google.com/forum/?fromgroups#!topic/chipseq/tcJxjnaOp4k\[1-25\]](https://groups.google.com/forum/?fromgroups#!topic/chipseq/tcJxjnaOp4k[1-25]) (last access 15.08.2012)

- If a successfully aligned read is detected on different chromosomes for the paternal and maternal genome and the mismatch descriptor is different, only the read with the lower mismatch descriptor for the respective genome is kept for further processing. For example, if read “Z” maps in the paternal genome on chromosome 2 with 2 mismatches and in the maternal genome on chromosome 6 with 0 mismatches, this read is kept for the peak detecting only in the maternal dataset and is discarded from the paternal one. This step should avoid biased mapping due to the mismatches allowed during read alignment.
- If a successfully aligned read maps to different chromosomes in both parents, but does not differ in the mismatch descriptor, it is kept for both parental datasets. In such cases other factors like indels could be the reason, representing established variants in the parental genomes.

Then, if genomes are processed, not being the reference genome, genome tables need to be provided. These tables do not provide any other information than the chromosome name and the length of the chromosome.

Finally, QuEST has been applied with the default parameters and therefore the “—silent” option. This choice has also been done to run the pipeline should as automatized as possible, without user input.

Two modifications have been provoked in comparison to the default parameters of Quest. First the mappable genome fraction has been adapted to be 0.8 for hg18 and second a FDR has not been applied for each TF separately. It has been decided to use a separate control sample (one complete input25bp sample), being the same for all TFs analyzed. This decision has been done, since the estimation of background is dependent on the number of reads left to the non-pseudo ChIP sample. Depending on the TF experiment analyzed this could be just a few million reads and additionally for the different TFs the number of reads is hardly the same. Then the FDR would always be different. Amongst others, QuEST outputs the ranked peak data as well as *.bed or *.wig files enabling a visualization with genome browsers, like the Integrated Genome Viewer (IGV, (Thorvaldsdóttir, Robinson, & Mesirov, 2012)) or the UCSC genome browser (Kent *et al.*, 2002).

3.7.7. Training Probabilistic Models

All selected TFs are part of the UniPROBE mouse TF set considered before, thus the learning of the models is based on UniPROBE data and has been done as described previously.

3.7.8. Screening of Peaks

The aim is to find variants leading to an allele-specific binding event in the parents of the CEU trio. Accordingly, each parent is screened for possible TFBS separately, followed by a merge of screening results to detect differences in binding affinity or significant SNP effects.

3.7.8.1. Detection of Common and “Missing” Peaks

As a first step of the pipeline, the occurrence of the peaks in the different genomes is detected. Sites bound by the TF are more likely to be located near the center of the peak or within a few base pairs from the maximum enrichment within the peak region itself. Consequently, not the whole peak, but only its center will be considered for further processing (Barski *et al.*, 2007). A peak center is defined as a sequence of 100bp located ± 50 bp from the position assigned with the peak maximum within the peak coordinates.

Since the coordinates of the parents are not directly comparable, due to indels and copy number variations (CNVs), the peak coordinates have to be lifted over to the reference genome. If the peak centers overlap by at least one nucleotide the peaks are considered as common and if a peak center of one parent does not overlap with any peak center of the other parent the peak is assigned to only occur in one parent. In the latter case, the respective “missing” peak center sequence is extracted from the respective parental genome using the reference genome coordinates as an anchor. In other words, a lift-over from the reference genome to the “missing” parent is performed to extract the nucleotide sequence. To clarify, the assigned coordinates for further processing are the matching reference genome coordinates (needed for SNP detection), while the peak center sequences are extracted from the respective parental genomes that can differ by SNPs, indels or CNVs introduced.

In order to detect peaks holding a SNP (see chapter 3.7.1), a binary search is performed, checking whether a SNP coordinate is placed within the range of a peak center. The CEU family SNP set is filtered based on the fact, whether the child genotype is heterozygous. This applies on 1,704,146 SNPs. Only in such cases, it is possible to deduce an allele-specific effect. The peak sequences are screened for the best binding site as they are extracted from the respective genome.

3.7.8.2. Screening for best Binding Site

After the peak center sequences to be screened are extracted, the actual screening for TFBS takes place. Each sequence is screened with each probabilistic model as described before. By means of the distribution of scores over all possible 9-mers a p-values is assigned to each

BS (see chapter 3.7.4) determining its significance, representing a measure of binding affinity. While in the initial description of the screening procedure, the site with the maximum score is considered as the best; here the BS with the smallest p-value is defined as the best BS of a peakcenter. As a threshold for BS significance $p \leq 0.05$ has been determined.

3.7.9. Assigning merged data with Significance

In order to evaluate the effect of a SNP, the results from the screening for the parents have to be considered in a comparative manner. Therefore the results for the maternal and paternal BS screening and SNP detection have to be merged to investigate possible differences between the parental peaks.

Accordingly the merging of screening results will reveal the parental peak combinations as illustrated in the following Figure 39 (showing the cases if only one SNP is sitting within the peak).

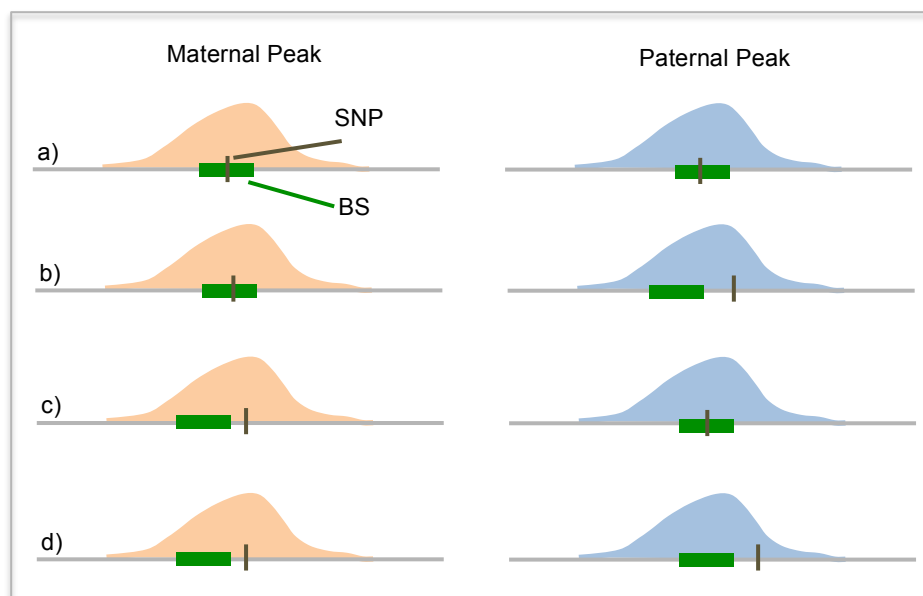


Figure 39: Graphical illustration of cases considered when merging screening data of common peaks. a) both parental BS are co-located with a SNP, b) only the maternal BS co-located with a SNP, c) only the paternal BS is co-located with a SNP, d) no parental BS is co-located with the best BS.

If a SNPs is co-located with a peak center sequence as well as with the best detected BS, it is considered as a potentially causal SNP. A p-value measuring the change of BS significance between the parental peaks, can be assigned for those BS detected, if they differ by exactly one nucleotide, as all these possibilities are captured by the SNE-distribution described before. This in turn means, that if two parental BS differ by more than one nucleotide, for example due to indels or a shifted BS, no p-value can be determined anymore. Nevertheless, a change in the BS significance might be observed and is considered in this work as a hint on a possible

effect on the binding affinity. Consequently, two aspects are taken into account when evaluating a possible change in BS affinity: if available, a SNE-p-value and if not, a change of significance in binding.

3.7.10. Reference set Generation of Differential Parental Binding

One main problem when predicting regulatory SNPs is the validation of the results. The biological validation would consist of the following steps: detection of change in TF binding, detection of altered gene expression, and finally possibly performing an association study to evaluate the functional effect of the altered gene regulation. Here, an approach is described to perform the first step. The same approach applied to RNA-seq data would also provide an evidence of differential gene regulation due to the SNPs. However, this is out of the scope of this work that focuses on TF binding.

To evaluate whether the models are able to detect differential binding, an additional approach has been integrated in the pipeline, published in 2011 (Rozowsky *et al.*, 2011). This approach is implemented in a tool called Alleleseq. The aim of Alleleseq is to detect allele-specific gene expression by analyzing allele-specific binding events using heterozygous SNPs to distinguish between maternal and paternal alleles. Even if it is not the main aim of their work, they state, that in some, but not in all location the heterozygous SNP might be the causative reason for the differential binding of the TF, if it is located within a TFBS (Rozowsky *et al.*, 2011).

Alleleseq constructs a diploid personal genome sequence using genomic sequence variants (SNPs, indels, and SVs). The construction of the two parental genomes has been described previously in chapter 3.7.2. Alleleseq deduces allele-specific binding events on the level of mapped reads between maternal and paternal alleles, initially not considering whether reads correlate with a peak. Accordingly, this tool is able to detect also slight differences in read mapping, not being dependent on the fact whether a real peak has been detected or not. The determination, whether a SNP can be involved in allele specific binding, is based on a significant difference in the counts of parental reads overlapping the SNP. Thus, Alleleseq can be considered as a complementary approach, to assign differential binding based on an independent, in fact upstream, measure compared to the approach based on TFBS screening.

The results from Alleleseq will be used to generate a positive set of differential binding events in order to determine the specificity and sensitivity of the probabilistic model screenings.

The original Alleleseq pipeline respects the flowchart illustrated in Figure 40 (Rozowsky *et al.*, 2011) which shows also the option available to detect differential gene expression. Here Alleleseq is only applied to detect differential binding and not its biological validity in context of

differential gene expression. The different steps executed are highlighted by green boxes in Figure 40. There is an option to provide binding sites as a filter to Alleleseq. The authors have used here ChIP-seq peak coordinates. The same approach has been adopted here: the BS submitted to Alleleseq to be overlapped with SNPs, are the peak centers resulting from the ChIP-seq analysis.

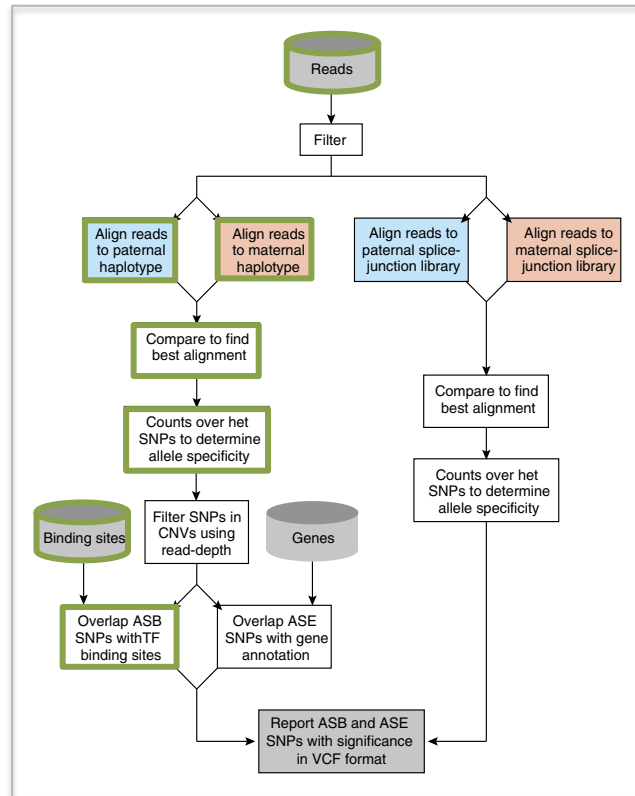


Figure 40: Alleleseq pipeline for determining allele-specific binding (ASB) and allele-specific expression (ASE) aligning reads against the personal diploid genome sequence as well as a diploid-aware gene annotation file. Taken and adapted from (Rozowsky *et al.*, 2011). In this work only one part of the pipeline has been applied, indicated by a green box. Additionally several steps have been modified, see text.

Deviant from the original Alleleseq pipeline, the following modifications have been applied:

- the step merging the two parental haplotypes has been re-implemented (see below)
- the filtering based on read depth of CNVs was not applied, since the necessary dataset was not available. Instead of a dummy CNV file has been used, indicating a sufficient read depth for all SNPs (no SNP has been filtered out based on a CNV threshold)

The input data for Alleleseq are bowtie files, meaning aligned reads. The subsequent step – the merging of the parental reads – has been modified for technical reasons. The newly implemented merging step merges the reads according to the following rules:

- if a read was mapped by bowtie to different chromosomes in the parental genomes the read has been discarded

- if a read was mapped to the same chromosome, but different locations in the parental genomes (using the reference genome as “carrier”) the read has been discarded
- if a read was mapped to both parental genomes on the same chromosome and the same location, only the one with the better mismatch descriptor has been kept for further processing
- if a read was mapped for both parental genomes on the same chromosome and same location with the same mismatch descriptor the read is kept and randomly assigned to either the maternal or the paternal read set
- if a read was mapped only to one parental genome, it has been kept

In the following, the SNP data and merged reads are used to generate allele counts for each SNP location. The resulting count-file contains the number of As, Cs, Gs, and Ts found in the reads mapped over each SNP location. Beside those counts, several other values are generated for each SNP location, including the parental genotypes, the reference allele, maternal/paternal allele (if determinable), major and minor allele, and a binomial p-value assuming a 50/50 probability of sampling each of the two alleles. An FDR is calculated, determining the number of false positives over the total number of observed positives, used as a threshold to report the final set of SNPs leading to an allele-specific binding. Finally, the SNP locations are overlapped with detected peak centres, serving as binding sites.

Alleleseq has been applied for all TFs also considered in the regulatory SNP detection described above. An exemplary output is shown below.

Table 10: An exemplary output of alleleseq. chr: chromosome, snppos: SNP position (reference genome), ref: reference allele, mat: maternal genotype, pat: paternal genotype, ch: child genotype, phase: SNP property, mata: alternative maternal allele, pata: alternative paternal allele, cA: count A, cC: count C, cG: count G, cT: count T, WP: winning parent (from which parent did the child inherit the allele), Cls: indicating the distribution of maternal and paternal alleles (Asym means, that the distribution between the maternal and paternal allele count is not equal), SymPval: binomial p-value assuming a 50/50 null, BS: determining whether SNP overlaps with BS (0:no overlap, 1: overlap)

chr	snppos	ref	mat	pat	ch	phase	mata	pata	cA	cC	cG	cT	WP	Cls	SymPval	BS
1	9258332	C	Y	C	Y	PHASED	T	C	0	15	0	0	P	Asym	0.000061	0
1	16812333	A	M	M	M	HETERO	None	None	28	5	0	0	?	Asym	0.000066	1
1	28444904	G	S	S	S	HETERO	None	None	0	0	9	0	?	Sym	0.003906	1
1	43628133	A	G	A	R	PHASED	G	A	10	0	1	0	P	Sym	0.011719	0
1	112099297	G	R	G	R	PHASED	A	G	59	0	114	0	P	Asym	0.000040	0
1	224897571	G	R	G	R	PHASED	A	G	13	0	2	0	M	Sym	0.007385	0

3.7.11. Sensitivity and Specificity Analysis for Peak Profile Classification

As already mentioned before, Allelseq is used to generate a positive and negative reference set of parental peaks that show differential read counts, which will be consider as “gold standard” to detect sequence variation that impacts TF binding. The input set are all parental peaks co-located with a child-heterozygous SNP.

To compare the models the sensitivity and specificity of the differential binding detections will be calculated. To do so, respective FP, FN, TN and TP need to be achieved. The merged peak set is classified based on BS significance calculations. The SNE distribution computed earlier can only be used to assign significance to BS differing by exactly one nucleotide, which does not apply to the majority of cases encountered here. Therefore, only the pure BS significance difference will be considered. To determine the best cut-off for BS difference to classify the input set, a ROC curve analysis by means of the R-package ROCR (Sing *et al.*, 2005) applying different BS significance thresholds will be performed. After the best threshold for BS difference is detected, the sensitivity and specificity for each model explicitly for this special cut-off to compare the models will be calculated.

In context of differential binding detection the classes to calculate the sensitivity and specificity are:

- TP: parental peak profiles classified as positive based on difference in BS significance being also classified as positive in the reference set
- FP: parental peak profiles classified as positive based on difference in BS significance being classified as negative in the reference set
- FN: parental peak profiles classified as negative with no difference in BS significance being classified as positive in the reference set
- TN: parental peak profiles classified as negative with no difference in BS significance being also classified as negative in the reference set

The following Figure 41 should explain the upper more illustratively.

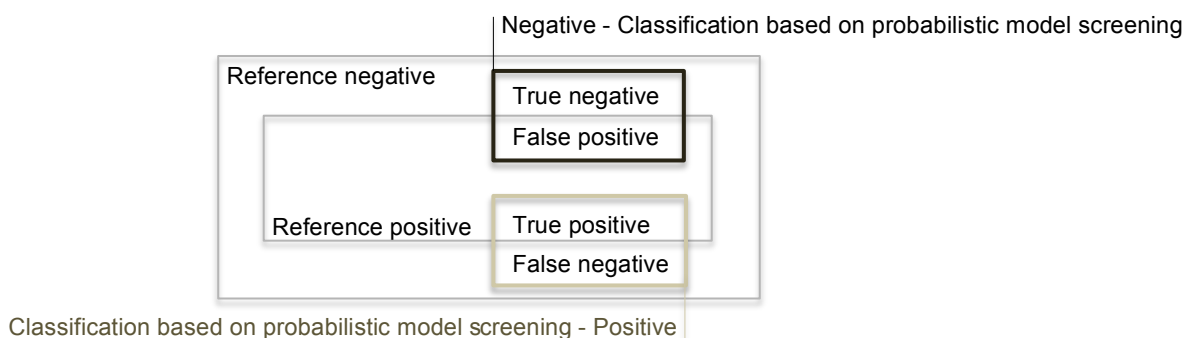


Figure 41: Graphical illustration of the relation of positive and negative reference data sets to true positives, true negatives, false positives and false negatives.

3.7.12. Determination of potentially Causative SNPs

The datasets to investigate for potentially causative SNPs is the true positive set as indicated in Figure 41. Only for this subgroup support of differential binding based on read counts and based on analysis of BS significance is existing. In context of this work a SNP is assigned as being causative if it co-located within a BS influencing the BS significance. In case of a single nucleotide exchange between the parental BS, a significance based on the SNE-p-value described before is assigned. However, for BS sequence differences of >1nt no statistically supported measure is available. Therefore, the final candidate list will be limited on SNPs sitting in BS differing by exactly one nucleotide and with a significant SNE-p-value being < 0.05.

4. Results

4.1. General Overview and Classification of the UniPROBE mouse TF set

As mentioned in the methods section, the UniPROBE set, investigated with regard to the binding site detection and ranking performance of the three probabilistic models, comprises all mouse TFs. However, only those TFs, for which at least data from two arrays are available, have been processed. The training set is composed of the 800 best and worst array probes, while the validation has been performed on a full array holding approximately 44,000 probes. Accordingly the number of probes used for training represents ~3.6% of the number used for model validation. A general classification of this subset is shown below. This classification is based on the performance the different models displayed for the validation procedure. Details for the different models are provided in the next section, for a representative subset.

All over, UniPROBE comprises currently 286 different mouse TFs. From those, for 104 different TFs the necessary data are available (corresponding to 115 datasets; some TFs offer for example 4 arrays, meaning that two datasets for a complete training and validation are available), for some TFs even more than two arrays. All three models have been learned for all available datasets, including a separate ET training to detect the best α in a range of 0-250. Depending on the model performance two different classes have been defined, namely “PWM better than Tree” and “Tree better than PWM”.

The classification of the different datasets has been done by calculating the difference of the mean over all AUC-values of the PWM and CLTree model. In general it has to be emphasized, that all AUC profiles shown in this work are based on 50 different data points corresponding to the AUC of 50 different ROC curves. Each ROC curve is calculated based on a different classification threshold referring to the same model training. In other words, the models have been trained once based on array one and the validation has been performed on array two by calculating 50 different ROC curves (with 50 different thresholds) summarized in one AUC profile. An example and a general overview over all datasets is shown in the following Figure 42.

From 115 datasets, the majority of 96 show a better¹² AUC profile with the Tree model than with PWM model. This corresponds to 86 out of 104 TFs. Only a relative small fraction of 18 TFs (represented by 19 datasets) show better results in the validation procedure for the PWM model.

¹² A better AUC profile at this point is not assigned to any threshold. It is simple determined by the higher ranging AUC curve, irrespectively of the size of the difference between AUC curves.

For those datasets where the tree model showed the better validation results, a further subgrouping has been done to “small difference”, “medium difference” and “strong difference”. Here, different thresholds have been chosen, based on the mean-AUC-difference between the PWM- and CLTree performed experiments.

A “small difference” is determined by a mean AUC-difference of <0.01 . The range to be a member of the class of “medium difference” has been chosen to be 0.01-0.026 and those TFs showing a mean AUC-difference of >0.026 are assigned to be strongly improved better represented by the CLTree model.

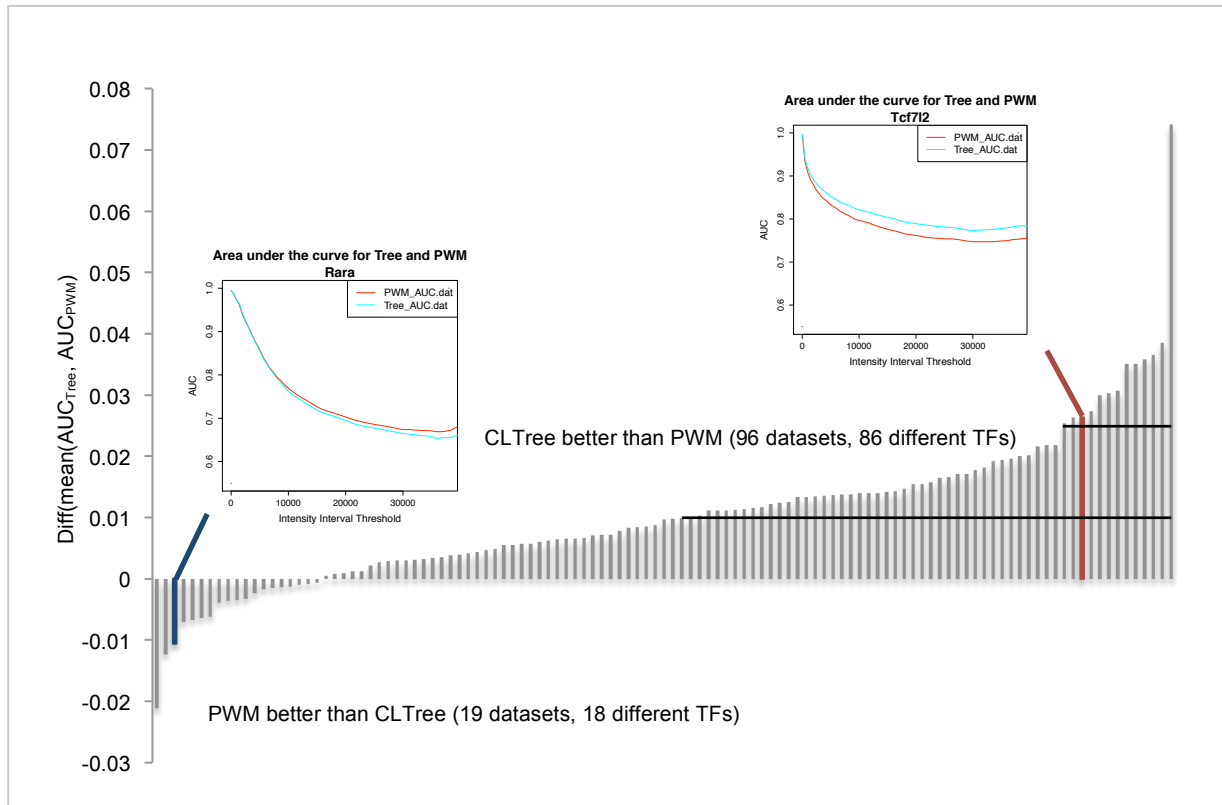


Figure 42: Bar chart illustrating the distribution of AUC-mean differences over all datasets. Two example validation profiles are shown in detail. Both profiles represent a summary of 50 ROC-curves as an Area under the curve (AUC) graph. (left) TF: Rara, the curve of the PWM deviates above the CLTree curve, representing a better relation of true positive to false positive predictions for the PWM. (right) TF: Tcf7l2, the curve of the CLTree deviates above the PWM curve, representing a better relation of true positive to false positive predictions for the CLTree. The black horizontal lines within the bar chart mark the different thresholds used to subgroup the datasets, performing better in the CLTree model.

In the following for each group one example will be provided.

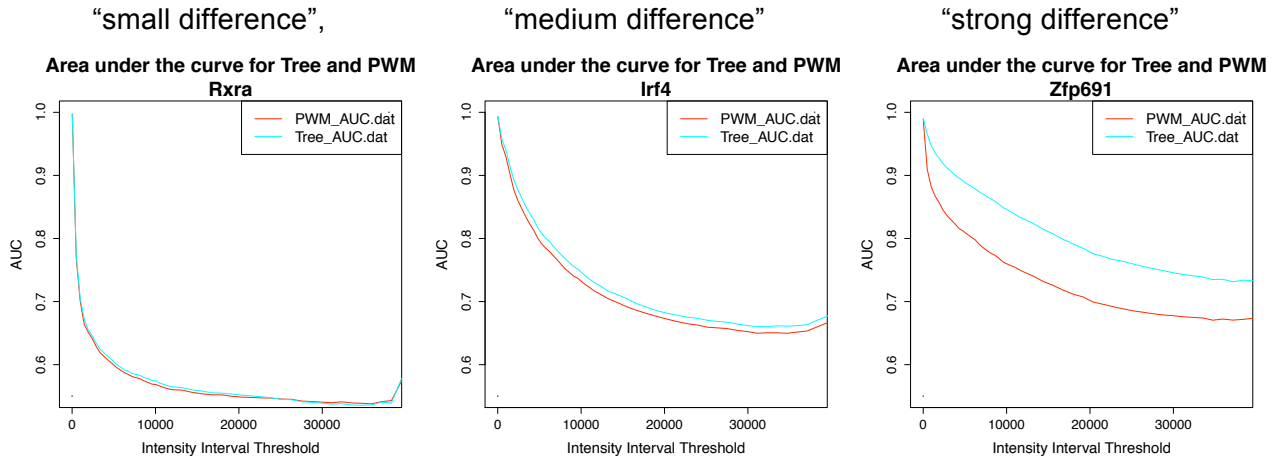


Figure 43: Three example validation profiles for the UniPROBE mouse dataset subgroup classification of “Tree better than PWM”. “small difference”: mean AUC-difference of <0.01 , “medium difference”: mean AUC-difference between 0.01 and 0.026, “strong difference”: mean AUC-difference of >0.026 . All three profiles represent a summary of 50 ROC-curves as an Area under the curve (AUC) graph. (left) TF: Rxra, the curve of the CLTree deviates slightly from the PWM curve, representing a small difference between the profiles. (middle) TF: Irf4, the curve of the CLTree deviates moderately from the PWM curve, representing a medium difference between the profiles. (right) TF: Zfp691, the curve of the CLTree deviates strongly from the PWM curve, representing a strong difference between the profiles.

The following Figure 44 summarizes the subgroups over all 96 “Tree better than PWM”-datasets.

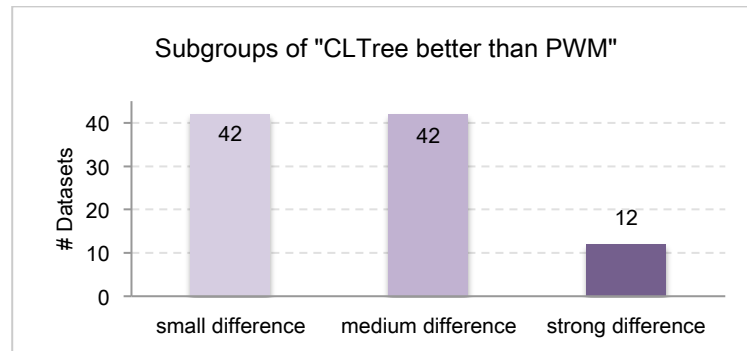


Figure 44: Bar chart representing the subgroups of those TF-sets showing a better validation profile with the CLTree in comparison to a PWM.

The smallest subgroup, with 12 datasets, is represented by those profiles showing a strong AUC-profile shift between CLTree and PWM. A moderate and small difference has been found in same amounts, with each 42 datasets. Thus, from originally 115 datasets $\sim 47\%$ (54 datasets) of the validation profiles show a considerable improvement by the CLTree over the PWM model.

With regard to the ET model, as a first step the parameter β has been selected for each TF as described in the methods chapter. In order to illustrate the selection better, one example plot is provided in the following on which basis the α selection has been done in general.

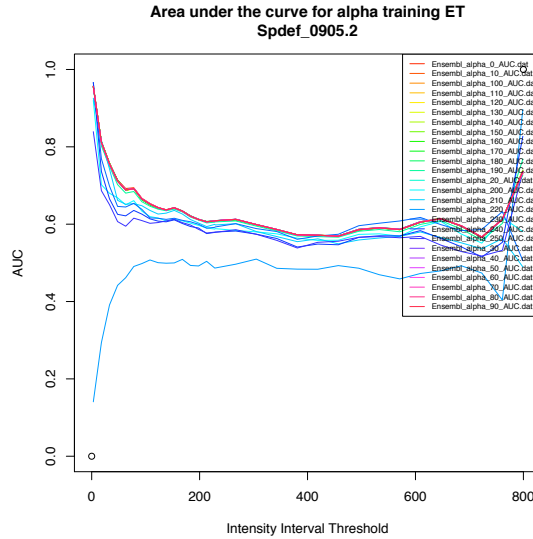


Figure 45: Exemplary AUC-profiles for the UniPROBE TF Spdef. For α s in a range of 0-250 AUC curves have been calculated, based on the best 800 probes of the training set. The best performing α has been selected for the validation screening. In this example case, the best α is 0.

Furthermore, it has been investigated, whether a trend of a range for best performing α can be observed. The following barchart represents the distribution of α over all TFs. As one can see in Figure 46, beside a slight trend of α to be located on the extrem towards 0 or > 190 , no special pattern can be detected. The lowest value of α , namely 0, occurs relatively often. Respectively, for a large portion of TFs (in numbers 26), the MI -amplifier α does not improve the system. In fact in such cases β has the value “1” and is therefore even not determined by the MI -level.

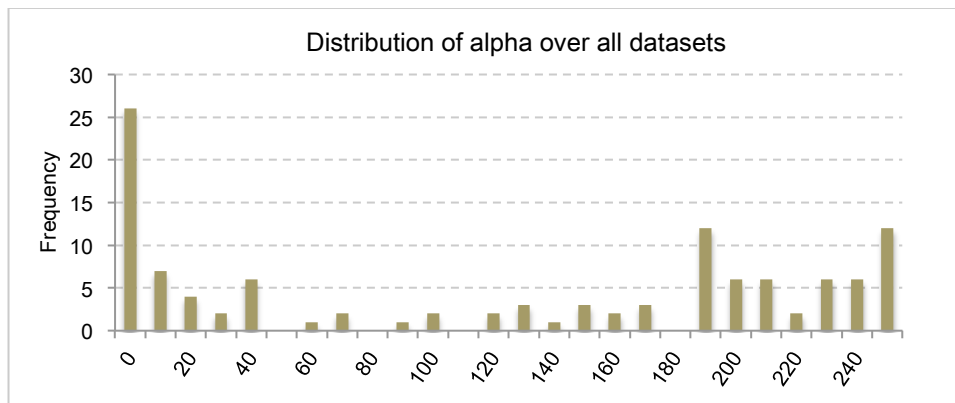


Figure 46: Distribution of best performing α over all TF-datasets analysed.

Next, the question is, whether the ET model can improve the all-over validation results. This has been tested for all TFs in the dataset. According to the results on hand the ET model performs for ~49% of the investigated UniPROBE datasets best, correlating to 56 out of 115

datasets. Interestingly, in particular, if the PWM provides better results than the CLTree, the ET model reaches for 68% of those datasets the best predictive power (see Figure 47).

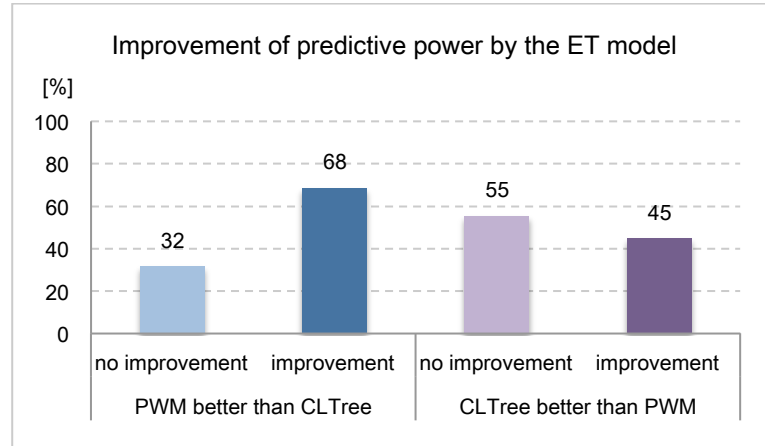


Figure 47: Barchart representing the percentage of improved validation profiles by means of the ET model for those datasets, where either the PWM or the CLTree showed the better validation result.

One of the best improvements by the ET model based on the mean-AUC difference, with an α of 140, is reached for Gcm1. By visual inspection also another dataset (Gata6) shows a high improvement. It has probably not been detected as the best, as the mean-AUC difference is smoothing out strong improvements over certain intensity interval ranges (see Figure 48).

In 2009, Bulyk *et al.* published a list of 19 TFs supposed to capture position interdependence within their binding motif. The 19 TFs are: E2F2, E2F3, Eomes, Esrra, Gcm1, Hbp1, Irf5, Myb, Mybl1, Nr2f2, Rara, Rxra, Sox4, Sox7, Sox8, Sox11, Spdef, Tcf2a and Zfp281.

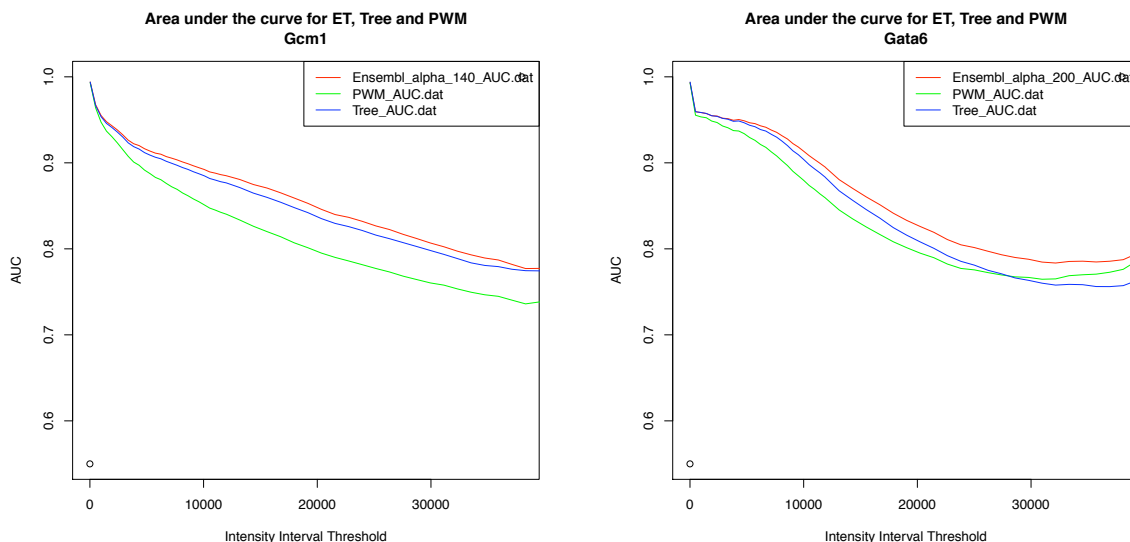


Figure 48: Validation profiles (AUC-curve) for Gcm1 (left) and Gata6 (right), showing a relatively strong improvement by the ET model. The CLTree curve (blue) ranges above the PWM curve (green). The predictive power of the ET model (red) is better than the same from the CLTree and PWM model.

A PWM assumes independence and trees are mirroring parent-child relationships – in other words dependencies. Hence, one would expect, that these 19 TFs are to find within the class of TFs where the CLTree and ET model provides better validation results than the PWM.

The Venn-Diagram (see Figure 49) shows the overlap of the Bulyk dependence-TFs and the group of TFs, for which the CLTree model had a higher predictive power than the PWM.

From those 19 TFs published as holding position interdependencies by Bulyk *et al* 15 (~79%) are to find within the group of TFs where the CLTree model provided better validation results. Additionally, for 7 out of those 15 TFs the ET model improved the predictive power further. From the remaining 4 Bulyk-TFs, not being detected directly by the CLTree training, also 2 validation profiles could be improved over the PWM model by the ET model. Only two Bulyk-TFs (Tcfe2a and Rara) the PWM model appeared in the analysis as the best performing. Accordingly, based on a tree-based model, it was possible to confirm 17 out of 19 TFs, representing 90%.

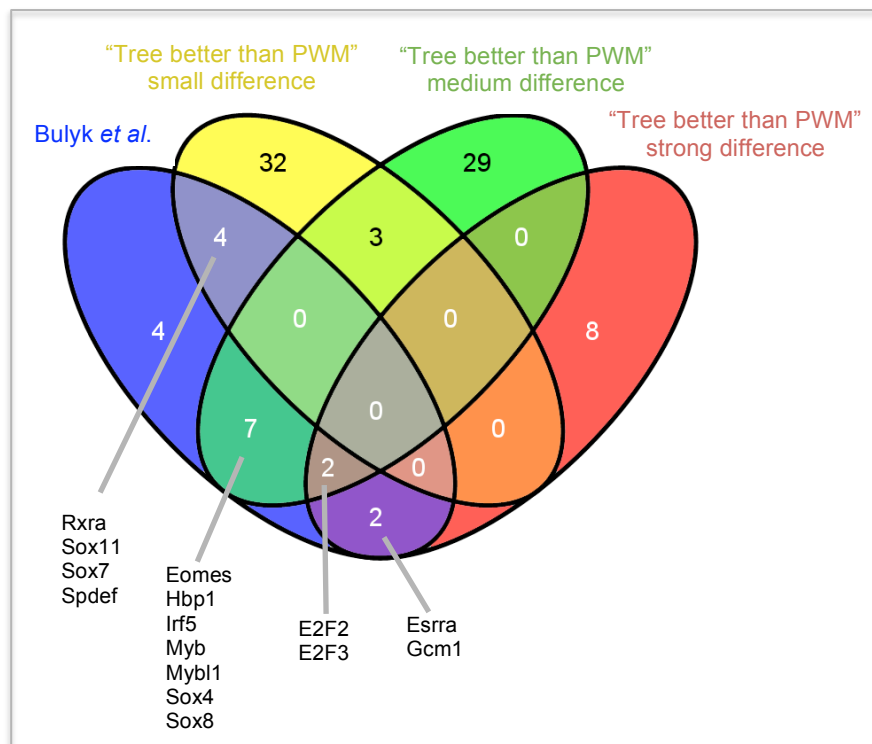


Figure 49: Venn-Diagram showing the overlap of the 19 TFs published from Bulyk *et al* with those TFs showing better validation profiles with a tree structure based model.

Claiming that by means of a tree based structure (CLTree or ET model), dependencies are better captured than by a PWM, the number of TFs, possibly affected by a positional interdependency increases from 19 (Bulyk *et al*) to 99 (correlating with 109/115 datasets). Not all of these 99 TFs show a tremendous improvement in predictive power for a tree-based model, however, for 51 a strong or moderate improvement has been observed.

4.2. Detailed Results for the Probabilistic Models for selected TFs

For the sake of shortness, detailed results will be provided for the TF-subset E2F2, Eomes, Esrra, Gcm1, Myb, selected from the overlap of TF with a considerable improvement in a tree-based model in the validation profile and Bulyk *et al.* In order inspect whether the trained PWM model corresponds to those publicly available by UniPROBE itself and TRANSFAC® the respective motifs have been compared (see section 4.2.1.1). In the section before, it has been hypothesized, that trees carry information about nucleotide dependencies. To investigate this aspect in more detail, the tree structures of the TF-subset listed above have been considered in more detail and if possible linked back to previous observations (by Bulyk *et al.*) (see section 4.2.1.2).

4.2.1. Probabilistic Model Training and Validation using UniPROBE data

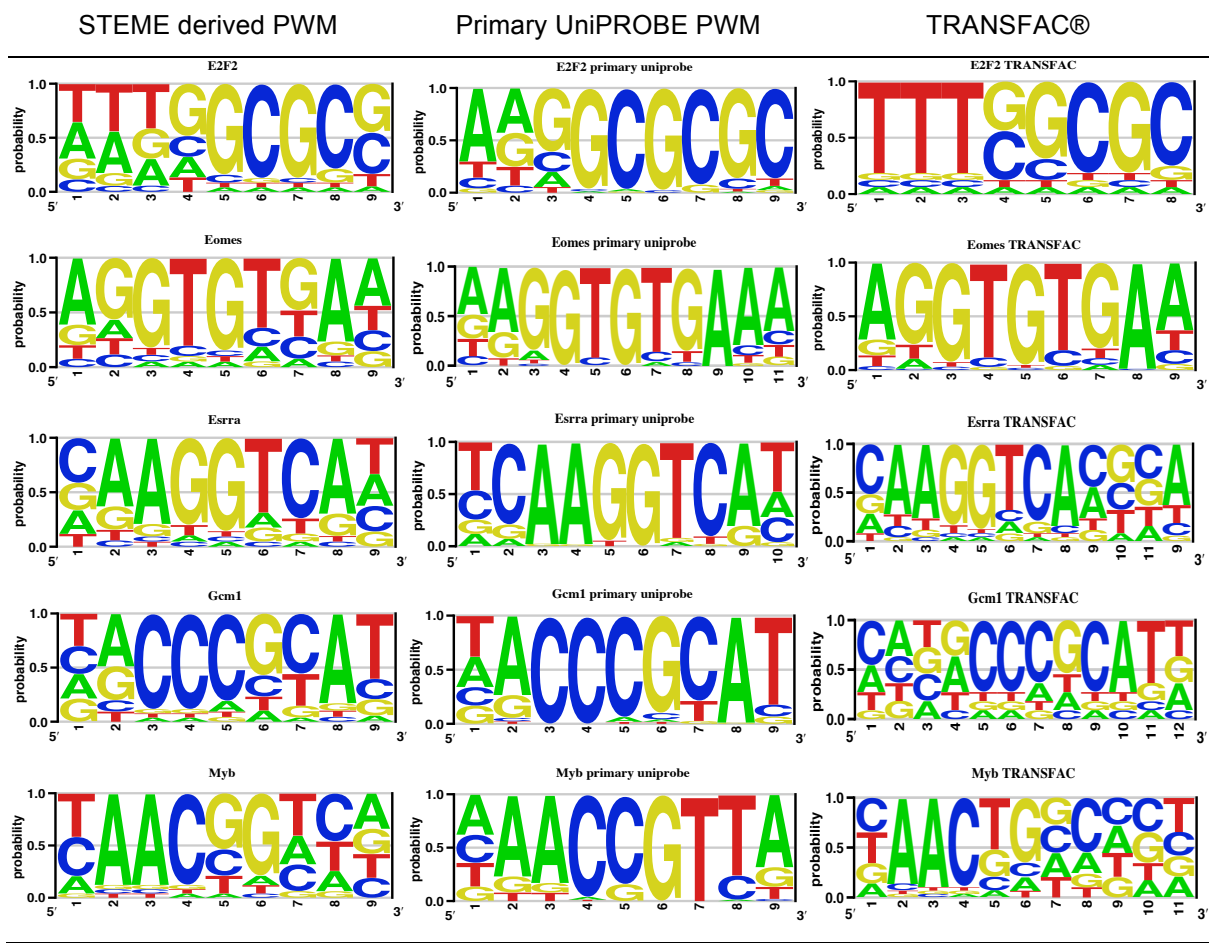
Since the AUC-curve progression is for all TFs selected described in principle the same, a very detailed description will be provided only for the first TF (E2F2), while for the remaining five a rather short statement, just pointing to the main characteristics, will be given.

4.2.1.1. PWM model

In this section those PWMs actually been used for the screening will be presented. The sequence logos for the newly trained PWMs based on STEME's motif discovery match very well published sequence motifs. The main core motifs are covered by all newly trained PWMs.

The main difference to observe is the amount of nucleotide conservation on the different positions, which might be due to the discriminative approach applied in STEME. However, a considerable overlap is to observe, suggesting a minor effect of the discriminative approach. In summary, the motif discovery step, including a data correction step considering glass-slide-distance-confounding, results in a representative sequence motif for all TFs.

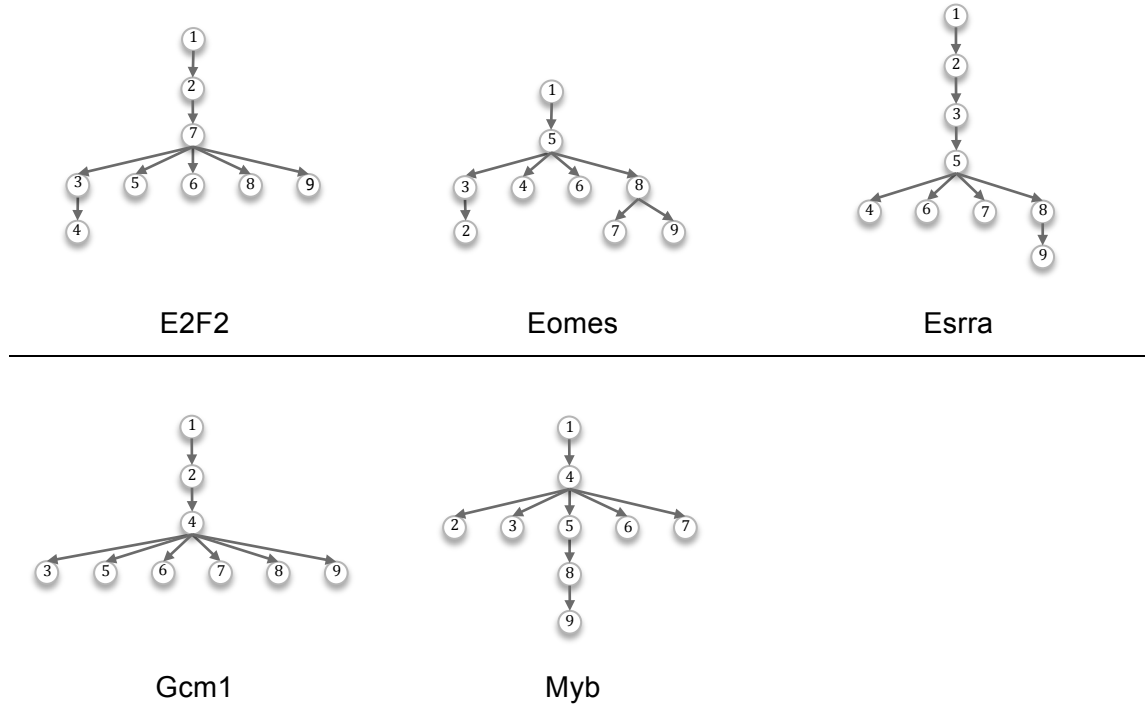
Table 11: Sequence logos for E2F2, Eomes, Esrra, Gcm1 and Myb. (left column) STEME derived PWM after glass slide distance correction and pseudocount adding, (middle column) published primary UniPROBE PWM, (right column) published TRANSFAC® sequence motif. The TRANSFAC® logo has been generated by extracting the single count PWM from TRANSFAC®, adding a pseudocount of 1 and final normalization. All logos have been generated by means of enoLOGOS.



4.2.1.2. CLTree model

Considering the CLTrees listed below, one can observe, that every tree holds at least one position serving as a parent for at least four children. The CLTree of Eomes displays, that positions 3, 5, 6, 8 and 9 in the binding motif seem to be dependent on position 7, being a highly conserved nucleotide in its sequence logo. This dependency cannot be derived from a simple PWM, considering all positions as being independent from each other. Intuitively spoken this means that a change on position 7 might have a stronger influence on the binding affinity than a change on position 6, which is also represented by a highly conserved nucleotide. Which nucleotide exchange correlates with the strongest impact on the binding affinity cannot be directly taken from the tree structure, but is coded within the paired marginals, used to calculate the individual score of a sequence.

Table 12: Graphical representation of the CLTree learned for E2F2, Eomes, Esrra, Gcm1 and Myb.



In fact positions 5, 6, 8 and 9 are identified as being independent from each other, suggesting that a change of these nucleotides might have minor impact on the remaining nucleotides. The trained CLTree for Eomes mirrors also the role of the centered “G” on position 5 in the STEME derived motif. As one can take from Table 12, this position, is the node connected to most other nodes in the tree. It represents the direct parent node of 4 different child-nodes, representing positions 3, 4, 6 and 8, being independent from each other. According to Badis *et al* (Badis *et al.*, 2009), who also identified Eomes as having positional interdependencies, Eomes binds most preferentially to AGGTGTGA and also binds AGGTGTCA or AGGTGTCG quite well, but relatively disfavors AGGTGTGG. In principle this observation can be confirmed. In the STEME derived sequence logo the underlined nucleotides correlate with positions 7 and 8, to be found within a small sub tree, reflecting a positional interdependence between positions 8 and 7, but also 8 and 9. The paired marginals report that the most frequent pairings on positions 7 and 8 within the training set are “TA”, “CA” and “GA”, whereby at least two of the nucleotide pairs identified by Badis *et al* are confirmed. However, according to the results on hand the strongest dependence reveals between position 5 and 3, assigned with the highest mutual information.

As already observed for Eomes, also in for the TF Esrra, the position with the highest nucleotide conservation, namely “G” on position 5 in the STEME derived motif, represents the node with the most positional dependencies. By itself being dependent on position 3, it serves

as a parent for nodes corresponding to positions 4, 6, 7 and 8. Position 9 is indirectly dependent on position 5 by being a child of position 8. Also for this TF Badis *et al.* provide preferred nucleotide combinations, namely strong preference for binding either CAAGGTCA or AGGGGTCA, but not CAGGGTCA or CGGGGTCA. The CLTree shows a parent-child relation of position 2 on position 1, and of position 3 on position 2, confirming the observation of Badis *et al.*. Furthermore, the highest paired marginals for the position pairs (1,2) and (1,3) are assigned to the nucleotide pair “CA”, corresponding to position triplet (1,2,3) of “CAA”.

The CLTree for Gcm1 appears to be less complex than the ones learned for the TFs described before, due to less branching. Position 4 in the STEME derived motif, occupies a very prominent role in the CLTree. It represents a node serving as parent for nearly all other nucleotide positions, except for position 2 and 1. Accordingly, the positional interdependencies seem to be centered in a manner, that a modification or change of the nucleotide placed on position 4 might lead to change in the binding affinity, not being able to be compensated by any other position. On the other side, the CLTree also implies, that nearly all other position (except position 2) are independent from each other.

Considering the CLTree for Myb it is obvious right away, that position 4 plays a central role, being the position with the highest number of dependent nodes. All together 5 positions (2, 3, 5, 6, 7) are directly and 2 positions (8, 9) are indirectly dependent on position 4, which is by itself only dependent on the root of the CLTree. The CLTree as a whole appears to be relatively complex. Badis *et al* say, that Myb’s interdependence results in a preference for binding either AACCGTCA or AACTGCCA. This would correlate with a parent-child relationship in the motif between position 5 and 7, which cannot be confirmed. Since these two positions are, according to the CLTree model, independent from each other. The position pair showing the highest MI within the applied training is (5,3).

4.2.1.3. ET model

For the ET model, no describable results are generated, giving any information to the reader helping to evaluate. A listing of the different α selected could be possible, based on the training procedure, or the value of determinant Q . Certainly, this numbers by themselves are meaningless. Thus, the only results visualizing the ET model are provided in the following section, presenting the AUC-profiles of the model validation.

4.2.1.4. AUC-profiles of Model Validation

The AUC-graphs below represent the summary of 50 ROC curves generated for the CLTree, PWM and ET model for each TF. Additionally, a comparative screening with the respective, published UniPROBE PWM, has been performed.

What is all AUC-profiles in common is, that a decreasing curve progression with an increasing intensity threshold is to observe. This means, that the more sites are classified as “good” the model loses predictive power. In other words, with higher intensity interval thresholds, the order of the scores does not correlate anymore with the order of the really measured binding signal. Thus, with a decreasing intensity threshold, the models diverge to randomness, meaning that the correct classification of “good” and “bad” sites is rather due to a random than due to a deterministic event. However, thinking about the space of analyzed data, and the respective signals, this observation is not surprising. There is a high amount of “fuzzy” sequences with medium and lower intensity values, ranging at the limit of being a good or bad binding site. Additionally, the models have been trained on to detect really good binding sites what is well mirrored in the AUC graphs. They start with a very high AUC-value, corresponding with very good predictive power for very good sites. Then the curve is falling continuously.

For all TFs shown here, the screening of array 2 with the UniPROBE PWM provides considerable less good AUC-profiles. These PWMs are generated differently, with an approach called Seed-and-Wobble and under the usage of 8-mer E-scores (Badis *et al.*, 2009). This in turn can affect the validation results for UniPROBE PWMs.

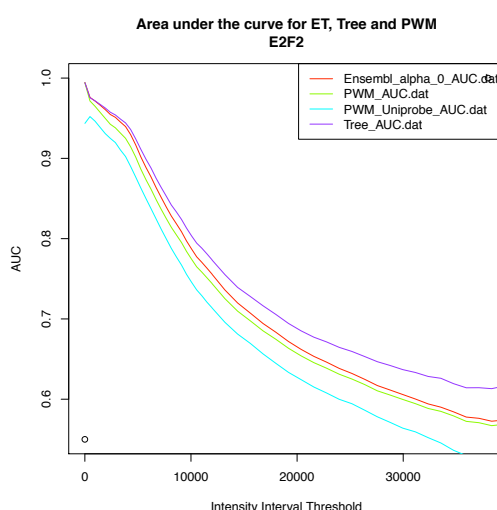


Figure 50: AUC-profile of E2F2 for the PWM, CLTree and ET model, including UniPROBE's primary PWM.

The AUC-profiles for E2F2 are ranging between nearly 1 and greater than 0.6 for the three models (see Figure 50). The AUC-curves of the tree-based models are all over settled in a higher range than the PWM-based one, so that the predictive power of the tree-based models can be considered as being better. This in turn leads to the conclusion, that the binding affinity of E2F2 is better represented by a tree-based model, considering parent-child relationships, meaning positional interdependence, than by a naïve probabilistic model like a PWM, neglecting the dependence assumption. The ET model (red curve) improves the CLTree further. Thus, it can be assumed that, the more complex structure of the ET model, capturing additional sequence features in different trees, could be even better suited to predict the binding affinity of a E2F2 binding site.

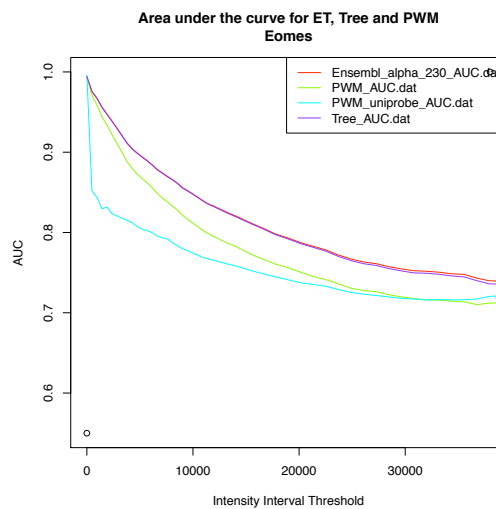


Figure 51: AUC-profile of Eomes for the PWM, CLTree and ET model, including UniPROBE's primary PWM.

Considering the AUC-curve of Eomes (see Figure 51), one can observe that the three models perform all over quite good, with a minimum AUC-value of the least well performing model – the PWM – still being above 0.7. The difference between the predictive power of the tree-based models and the PWM are relative high in comparison to the other TFs described in this section. Considering the CLTree learned and described above, its relatively complex structure, comprising three sub-trees, displays already, that the binding motif of Eomes is obviously affected by positional interdependencies. Accordingly, the result that the tree-structure is better performing in the classification of “good” binding sites is not surprising. However, the ET model, calculated with $\alpha=230$ does not improve too much over the CLTree. One reason for that might be, that for the binding prediction of Eomes, the more restrictive CLTree model is better suited and no more complex features, as captured by the ET model, are relevant.

Considering the branching of the CLTree structure of Esrra, its binding motif seems to be basically affected by positional interdependencies. Consequently, it is expectable that the

tree-based structures display a better curve progression in the AUC-profile, than the PWM model. Ranging from an AUC value of close to 1 for all models, the curve is falling on a minimum of ~ 0.65 for the PWM and ~ 0.68 for the tree-based models (see Figure 52). The ET model performs equal to the CLTree, leading to the assumption, that also here the higher complexity of the ET model is not relevant.

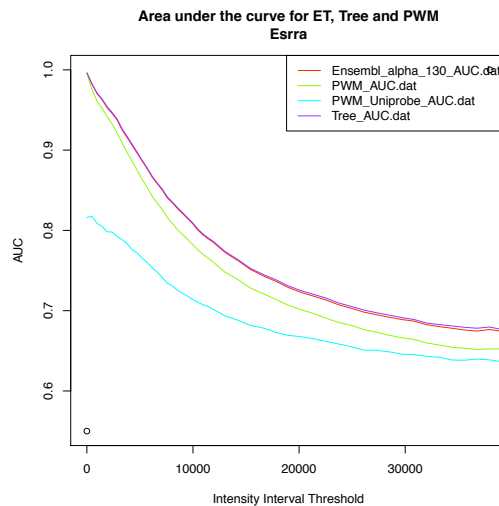


Figure 52: AUC-profile of Esrra for the PWM, CLTree and ET model, including UniPROBE's primary PWM.

In comparison to all other TFs described in this section, Gcm1 has the best AUC-profile, with a less strong falling curve progression (see Figure 53). The range auf AUC-values is relatively high settled. For the PWM model, the curve is in comparison to the tree-models falling faster, from a maximum of nearly 1 to a minimum of ~ 0.74 . The AUC-values for the tree-based models are mainly located between 0.9 and 0.8, wherein the lowest AUC-value is ~ 0.78 . Starting from an intensity threshold of $\sim 3,000$ till the highest one, nearly a curve linearity is to observe. Furthermore, the ET model, calculated with an α of 140, can be clearly distinguished from the CLTree curve, improving over its predictive power. This might be due to the fact, that the ET model captures additional underlying sequence features, which are relevant for Gcm1, improving the predictive power. Moreover, the ET model is less restrictive than a single maximum spanning tree, which might be an advantage, considering the CLTree structure highly centered on position 4, but with the highest amount of independent positions.

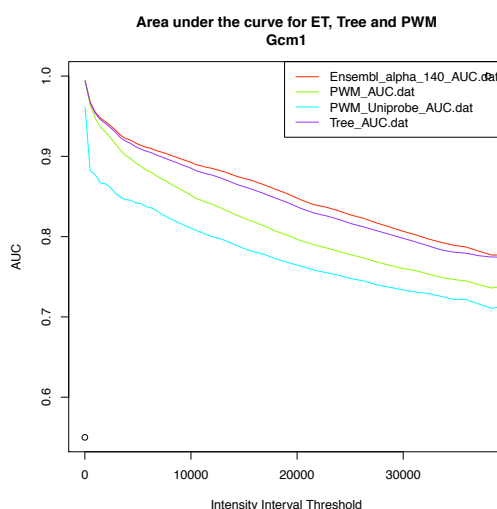


Figure 53: AUC-profile of Gcm1 for the PWM, CLTree and ET model, including UniPROBE's primary PWM.

The AUC-profile for Myb discloses that the ET model (calculated with $\alpha=220$) does not improve the predictive power over the CLTree (see Figure 54). The AUC-curve progression of all models is falling with an increased intensity threshold as already observed for all other TFs described. The AUC-values for the PWM model range between ~ 0.97 and ~ 0.69 . In comparison, the tree-based models perform better, starting from the same maximum, but having a minimum AUC-value of ~ 0.71 . As for the other TFs the results suggest the existence of positional interdependencies within the binding motif.

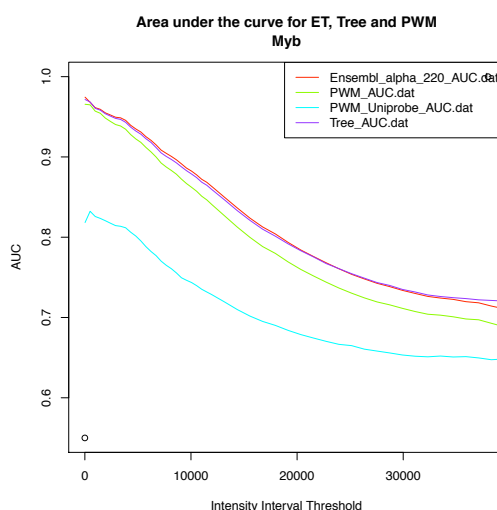


Figure 54: AUC-profile of Myb for the PWM, CLTree and ET model, including UniPROBE's primary PWM.

4.3. Application of Probabilistic Models for Differential Peak Profile and rSNP detection

4.3.1. Transcription Factor Selection

The selection criteria have been, as already described in the methods, the availability of raw sequence read files from ENCODE, as well as overlap with UniPROBE. Accordingly, the following TFs have been selected: EGR1, IRF4, MAX, SPI1, SRF, POU2F2, ETS1 and RXRA. As the screening will be run on a UniPROBE based training, it is a basic requirement, that the models, in fact trained on mouse derived datasets, are applicable on human data. In this regard amino acid and nucleotide sequence similarity have been consulted. Furthermore, for ETS1 and POU2F2 no second array is available, so that it has been decided to abandon them from further analysis, since no consistent model training can be applied. RXRA, that is known to bind mainly as a dimer (hetero or homodimer), has also been excluded. Its monomeric binding profile based on UniPROBE might not accurately model the dimeric binding expected to be observed in Chip-seq data. Accordingly, the differential binding profile prediction and rSNP-detection are executed for EGR1, IRF4, MAX, SPI1 and SRF.

Table 13: Amino acid and nucleotide similarities between the TFs selected and their *Mus musculus* homologue.

TF	Amino acid sequence similarity	Nucleotide sequence similarity
	(<i>Mus musculus</i> – <i>Homo sapiens</i>)	(<i>Mus musculus</i> – <i>Homo sapiens</i>)
EGR1	93.42%	87.59%
IRF4	93.44%	88.15%
MAX	98.12%	95.62%
SPI1	88.52%	88.02%
SRF	97.22%	95.17%.

In consideration of the similarity values listed below, the UniPROBE derived models have been considered as being applicable on a human dataset.

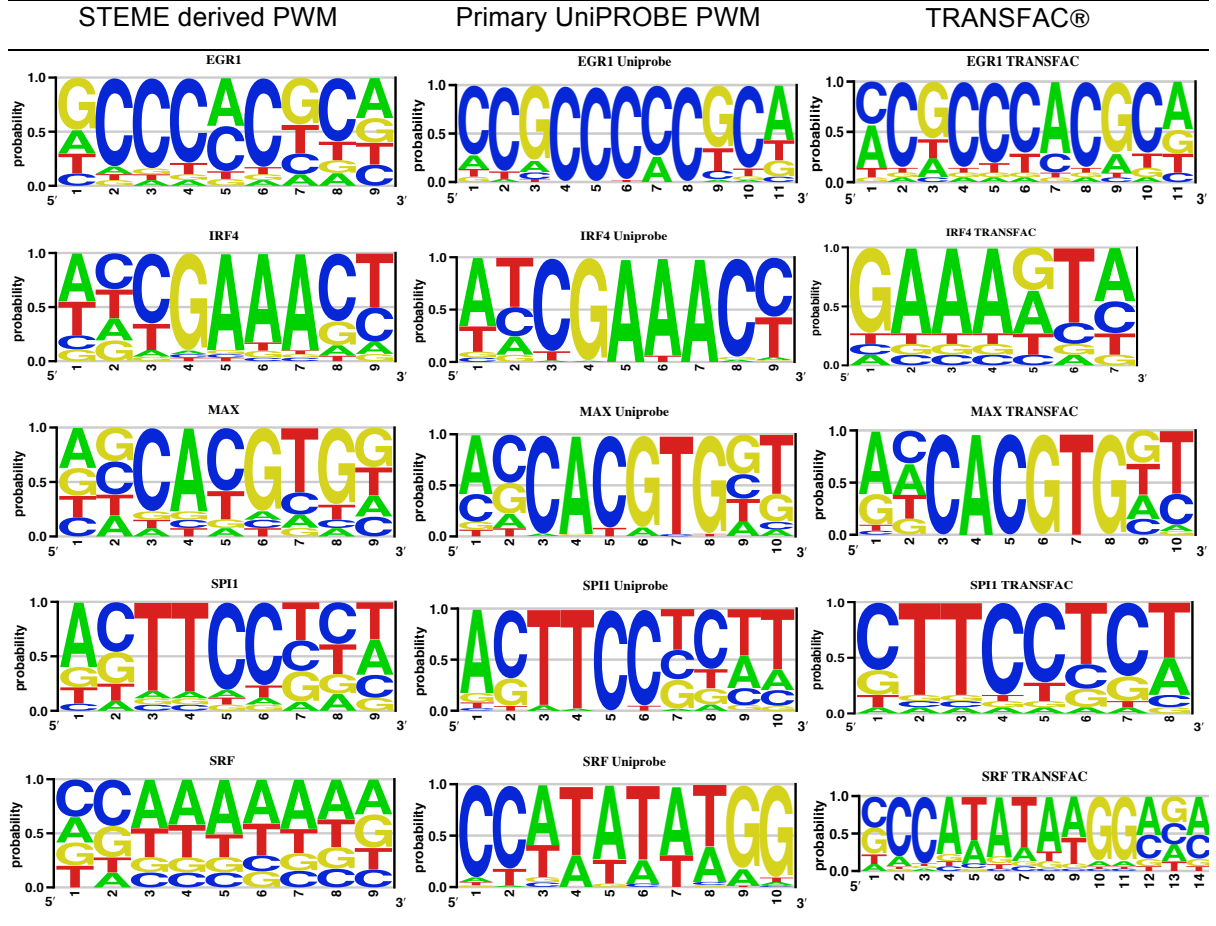
4.3.2. Training Probabilistic Models

4.3.2.1. PWM model

Considering the sequence logos listed in Table 14, it can be stated, that the STEME derived motif is very well in line with the published ones. The nucleotide with the highest conservation, maps nearly perfectly the respective ones in the well-established PWMs. The only TF, where a slight, but obvious, difference can be observed is SRF. Here the two starting “C” are

mapping the published motifs, while the rest is less well correlating. However, it is still considered as sufficient, since the main nucleotides, namely “A” and “T” are the two mostly conserved ones.

Table 14: Sequence logos for EGR1, IRF4, MAX, SPI1 and SRF. (left column) STEME derived PWM after glass slide distance correction and pseudocount adding, (middle column) published primary UniPROBE PWM, (right column) published TRANSFAC® sequence. The TRANSFAC® logo has been generated by extracting the single count PWM from TRANSFAC®, adding a pseudocount of 1 and final normalization. All logos have been generated by means of enOLOGOS.

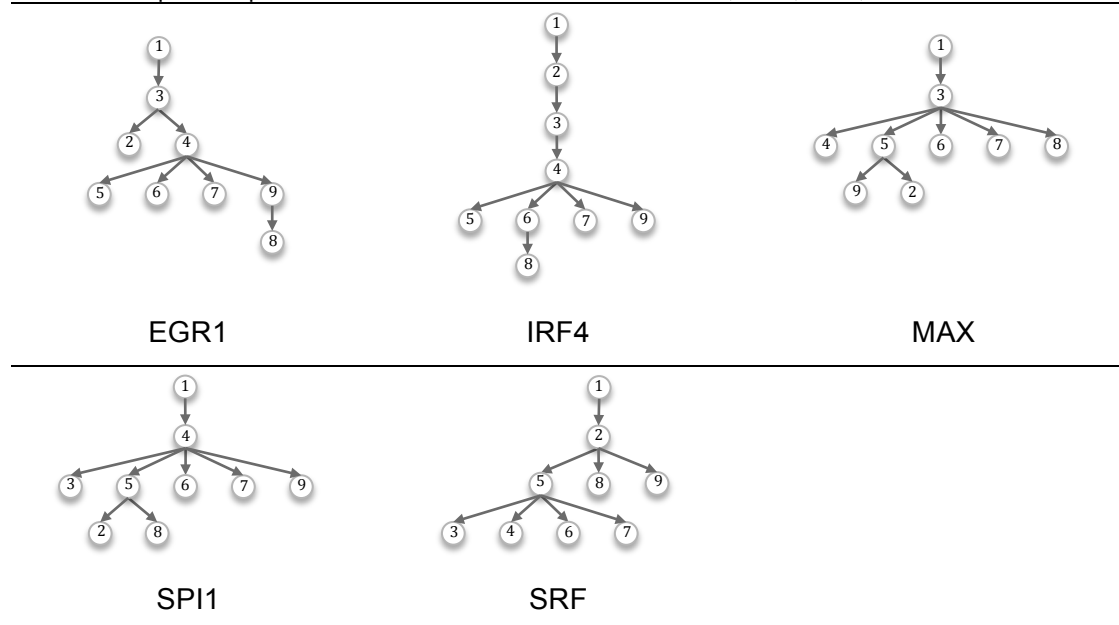


4.3.2.2. CLTree model

As already observed for the initial CLTree results presented, also here, each TF seems to be affected by positional interdependencies. For each CLTree one can detect one central parent, having at least 4 children. In general the position in the sequence motif showing the nucleotide with highest conservation can be identified as this central position. One exception is SRF; here a rather complex tree structure is constructed, which might be explained by the fact, that no real major position can be derived from the sequences used in the training. Sub-trees are existent for all TFs except IRF4. However, each tree structure reveals at least 4 nucleotide

positions are independent from each other. For example for SPI1, the positions 3,5,6,7 and 9 are independent from each other, all being dependent at the same time from position 4.

Table 15: Graphical representation of the CLTree learned for EGR1, IRF4, MAX, SPI1 and SRF.



4.3.2.3. Model Validation

As one can take from Figure 55, when neglecting the results for the screening with the primary UniPROBE PWM, for all TFs a tree-based model performed best, either the CLTree or ET model. The AUC-profiles start all with a high predictive power of nearly 1, followed by a falling curve progression to values ranging between ~0.65 and ~0.7.

Interestingly in SRF the ET model has the best AUC profile, even if the difference to the PWM model is very small. This might be due to a relative complex interaction within the motif, maybe mirrored by minor difference in the middle part occupied by A/T. Maybe the ET model is able to represent these minor differences by its higher complexity coded in numerous spanning trees, each capturing one additional feature.

For all TFs (see Figure 55), except EGR1, the published UniPROBE matrix provides consistently less good AUC-profiles than the newly trained ones. For EGR1 the UniPROBE-PWM shows better results from a certain threshold on. Therefore, it could be stated, that the UniPROBE PWM shows better performance for those BS, being in a “fuzzy” state, not being “really good” and not being “really bad”. However, in order to keep the subsequent procedures consistent, the subsequent steps for all TFs are executed based on the newly derived models.

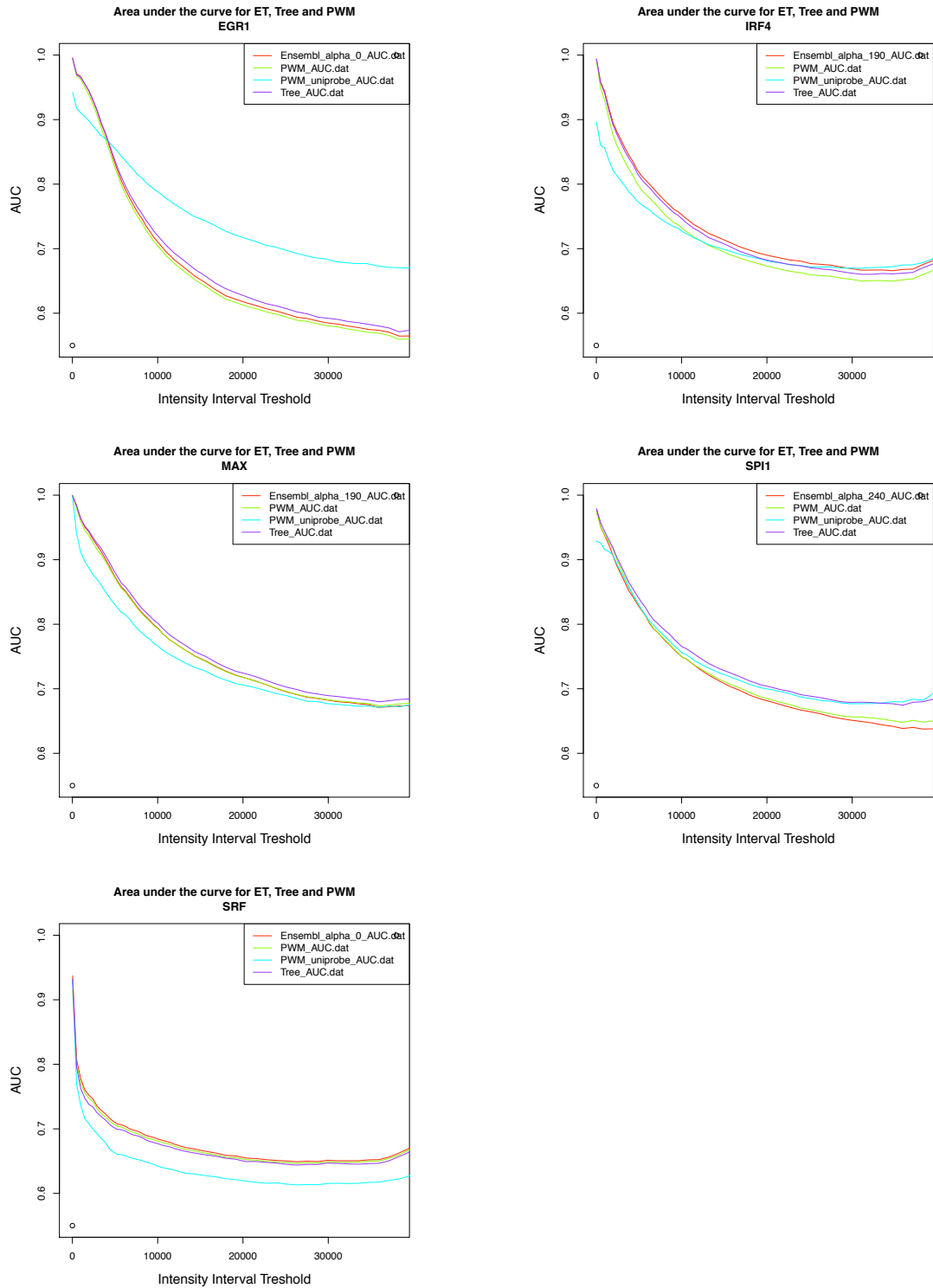


Figure 55: AUC-profiles for EGR1 (top row, left), IRF4 (top row, right), MAX (middle row, left), SPI1 (middle row, right) and SRF (bottom row)

4.3.3. ChIP-seq Analysis and Peak Detection

4.3.3.1. Sequence Read Preprocessing

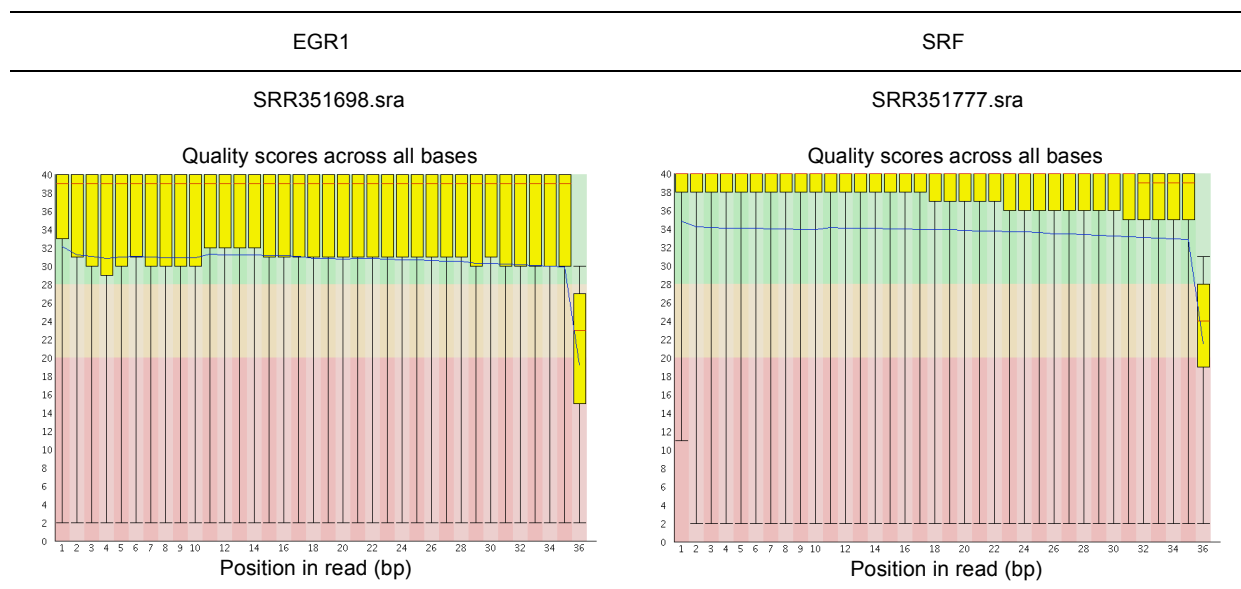
For all TFs except MAX original SRA-files have been downloaded from the Gene Expression Omnibus (GEO) repository. For MAX no SRA-files have been available. In order to keep MAX in the analysis the preprocessed files (fast-files, after trimming) from the original Alleleseq-publication¹³ have been used. Only for this file, the ChIP-seq analysis started directly with the read alignment, while the others passed through the preliminary steps of base calling, quality control and if necessary trimming.

Table 16: Listing of SRA-files used. As INPUT sample the following files have been taken: SRR351536.sra, SRR351537.sra, SRR351539.sra, SRR351660.sra, SRR351701.sra

EGR1	IRF4	SPI1	SRF
SRR351698.sra	SRR351616.sra	SRR351880.sra	SRR351533.sra
SRR351699.sra	SRR351617.sra	SRR351881.sra	SRR351534.sra
			SRR351777.sra
			SRR351778.sra

From all TFs processed only for SRF (SRR351777) and EGR1 (SRR351698) a trimming was considered of being necessary (see below).

Table 17: Listing of SRA-files for which trimming has been performed. Upper rows: name of the TF and SRA-file affected. Lower row “per-base-quality”-plots indicating the necessity to trim the read on 35bp.



¹³ <http://archive.gersteinlab.org/proj/AlleleSeq/Max/>

4.3.3.2. Alignment of Reads

As one can take from Table 18 the number of reads mapped to the parental genomes is for all TFs higher than for the reference genome. Taking the number of reads mapped to the reference genome, the increase of mapped reads to the parental genomes is around 0.3-0.4 %. The numbers of reads differ between the different TFs. While the ChIP-seq experiment for MAX yields the highest amount of mapped reads with ~46 mio, the TF with the smallest amount of mapped reads, namely SPI1, has only ~19.5 mio.

As mentioned in section 3.7.6.2, the reads to be used for peak detection have been sorted, according to inconsistent mapping behavior. For each TF the amount of inconsistently mapped read ranges around 0.002% of the number of mapped reads in the respective genome. Wherein for most TFs the amount of inconsistent mapping is higher in the maternal than in the paternal genome.

Table 18: Sequence reads mapped to the reference, maternal and paternal genome by means of Bowtie.

TF	Genome	No. of mapped reads		#reads sorted out		#reads used for peak detection	
ERG1	Reference	27,762,557				27,762,557	(100 %)
	Maternal	27,871,044	(+0,39 %)	548	(0,0020 %)	27,870,496	(99.998 %)
	Paternal	27,869,160	(+0,38 %)	490	(0,0018 %)	27,868,670	(99.998 %)
IRF4	Reference	20,273,126				20,273,126	(100 %)
	Maternal	20,348,666	(+0,37 %)	431	(0,0021 %)	20,348,235	(99.998 %)
	Paternal	20,347,622	(+0,37 %)	344	(0,0017 %)	20,347,278	(99.998 %)
MAX	Reference	46,601,788				46,601,788	(100 %)
	Maternal	46,736,493	(+0,29 %)	894	(0,0019 %)	46,735,599	(99.998 %)
	Paternal	46,733,167	(+0,28 %)	964	(0,0021 %)	46,732,203	(99.998 %)
SPI1	Reference	19,584,389				19,584,389	(100 %)
	Maternal	19,653,384	(+0,35 %)	360	(0,0018 %)	19,653,024	(99.998 %)
	Paternal	19,652,472	(+0,35 %)	272	(0,0014 %)	19,652,200	(99.999 %)
SRF	Reference	29,415,720				29,415,720	(100 %)
	Maternal	29,526,198	(+0,38 %)	595	(0,0020 %)	29,525,603	(99.998 %)
	Paternal	29,524,330	(+0,37 %)	580	(0,0020 %)	29,523,750	(99.998 %)

However, nearly 100% of reads have been used for the peak detection step for each TF.

4.3.3.3. Peak Detection

The peak detection has been performed for all selected TFs. In particular 2 TFs show a comparatively high number of reported peaks, namely EGR1 with ~10,100 and SPI1 with 28,000 (see table below).

The amount of reported peaks does not differ strongly between the different genomes, mainly varying in a range of +/- 10 peaks over all three genomes.

Table 19: Number of detected peaks per TF and genome

TF	Genome	#detected peaks
ERG1	Reference	10,155
	Maternal	10,145
	Paternal	10,134
IRF4	Reference	7,221
	Maternal	7,245
	Paternal	7,258
MAX	Reference	4,623
	Maternal	4,617
	Paternal	4,613
SPI1	Reference	28,070
	Maternal	28,087
	Paternal	28,079
SRF	Reference	5,165
	Maternal	5,174
	Paternal	5,164

4.3.4. Descriptive Analysis of Maternal and Paternal Peaks

In this section a general descriptive overview is given concerning the distribution of peaks with and without co-localized SNP between the maternal and paternal genome. Here, all SNPs are considered, not only the heterozygous SNPs, this limitation becomes only evident in the causative SNP detection.

As a preliminary note, the word “peaks” will be used as a synonym for “peak centers”.

4.3.4.1. Detection of Common and Parent Specific Peaks

The following Table 20 displays the absolute and relative amount of peaks in the parental genomes, separated in three groups, namely those peaks detected in both parents successfully overlapped (common), those only detected in the maternal genome (OnlyMat) and those only detected in the paternal genome (OnlyPat). As one can see, for all TFs the majority of peaks (~99%) has been detected in both genomes and was successfully overlapped. For example, from 10,145 peaks detected for EGR1 in the maternal genome, 10,049 have been successfully overlapped with a peak within the paternal genome, correlating to 99.1% of maternal peaks and 99.2% of all paternal peaks.

A very tiny number of peaks – ranging between 0.7% (SPI1, Only Mat) and 1.35 % (IRF4, Only Pat) do only occur in one of the parents. The percentages indicate that the relative amount of common peaks or those only mappable to one genome, does not vary strongly over all TFs, but seems to be rather stable.

Table 20: Detection of common and missing peaks in the maternal and paternal genome. peaksMat: peaks detected in maternal genome, peaksPat: peaks detected in paternal genome, Common: successfully overlapped peaks. %mat: percentage of maternal peaks paired with a paternal peak, %pat: percentage of paternal peaks paired with a maternal peak, OnlyMat: peaks detected only in the maternal genome, OnlyPat: peaks detected only in the paternal genome.

TF	peaksMat	peaksPat	Common	%mat	%pat	OnlyMat	OnlyPat
EGR1	10,145	10,134	10,049	99.1	99.2	96 (0.95 %)	85 (0.84 %)
IRF4	7,244	7,258	7,160	98.8	98.6	84 (1.16 %)	98 (1.35 %)
MAX	4,616	4,611	4,569	99.0	99.1	47 (1.02 %)	42 (0.91 %)
SPI1	28,087	28,079	27,891	99.3	99.3	196 (0.70 %)	188 (0.67 %)
SRF	5,174	5,164	5,123	99.0	99.2	53 (1.02 %)	43 (0.83 %)

4.3.4.2. Detection of Peak-SNP-co-location

As one can take from Figure 56, the relative amount of peaks, namely ~20%, co-located with one or several SNPs (referred as being SNPed in the following) is fairly similar over nearly all TFs. Only EGR1 shows moderately less SNPed peaks in both, the maternal and paternal genome.

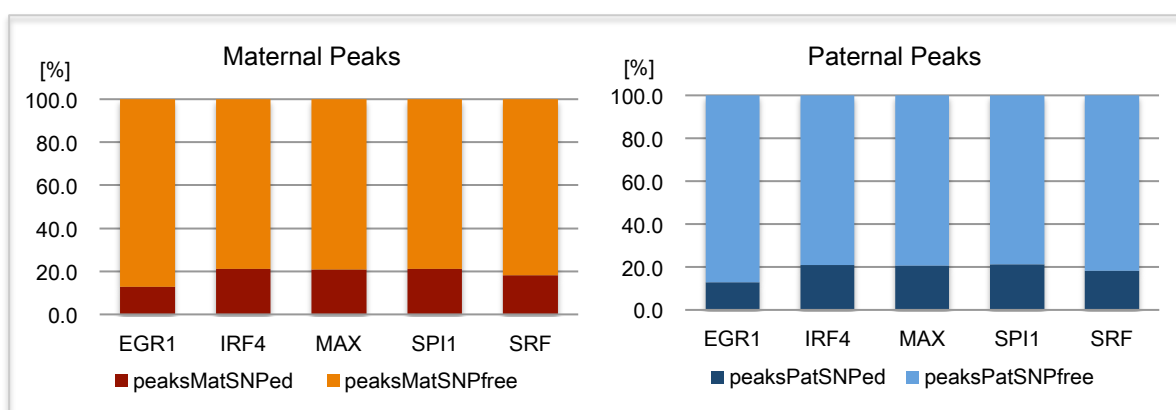


Figure 56: Relative distribution of SNPed versus not-SNPed peaks in the parental genomes (left: maternal peaks, right: paternal peaks)

Comparing the two parents with each other, very little differences are to observe (IRF4, shows a slightly smaller relative amount of SNPed peaks in the paternal genome). Assuming that the common peaks overlap over a wide range and not only in their extremes, only common peaks affected by indels or CNVs, leading to coordinate shift, might show a difference with regard to their “SNP-status” between the maternal and paternal genome. However, since common peaks have actually been successfully overlapped, the amount of common peaks with shifted

SNP coordinates should occur very rarely. Thus, the only peaks, expected to cause a difference in the upper graph, are those only detected in one genome.

Assuming that normally all peaks should be mappable to both genomes, it is to expect that in particular those peaks, only detected in one of the two parents, show an increased amount of SNPed peaks. This assumption is indeed reflected in the following graph (Figure 57), illustrating the relative distribution of SNPed versus not-SNPed peaks only detected in one genome. For all TFs and both parents, the relative amount of SNPed peaks is higher than in the overall peak set shown above. In the group of peaks only detected in the maternal genome EGR1 represents the TF with the smallest amount of SNPed peaks with 50%, while MAX has the highest amount of SNPed peaks in this group with even 77%. The distributions between the parental genomes are not as similar as for the results shown before.

However, the trend for EGR1, IRF4 and MAX compares very well between the two parents, but for SPI1 and SRF the trend reverses. Meaning, that in the maternal genome SPI1-peaks show a higher SNP-co-location than SRF peaks, while this correlation is inverted in the paternal genome.

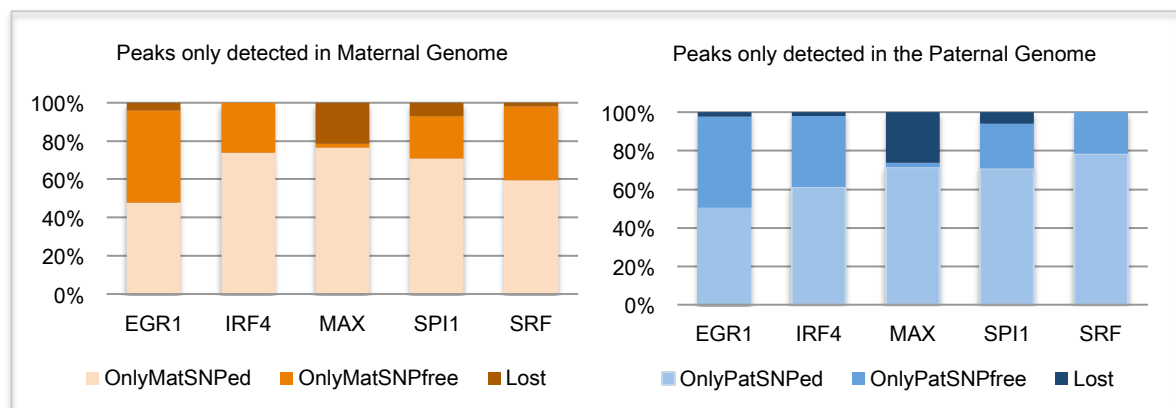


Figure 57: Relative distribution of SNPed versus not-SNPed and lost peaks in the subgroup of peaks only detected in one genome. Lost peaks are those, where coordinates have not been mappable to the reference genome, being lost during the overlap with SNP data. left panel: peaks detected only in maternal genome, right panel: peaks detected only in paternal genome.

4.3.5. Detection of Differential TF Binding Profiles in Parental Peaks

As a pre-requirement to detect causative SNPs the UniPROBE trained models should be able to detect those peak locations, where the binding profile between the parents is different. To consider the results for the different models in a comparative manner, the sensitivity and specificity of each model for a specific BS difference cut-off has been calculated and plotted as explained in detail in the methods section. A reference set has been generated by means of Alleleseq (see chapter 3.7.10). All BS which have been detected by screening peak regions co-located with heterozygous child SNPs are significant (see Annex).

4.3.5.1. Sensitivity and Specificity

The following figure shows the graphical representation of the sensitivity and specificity calculation. According to the ROC curve analysis (see Annex), the cut-off for the difference between the BS p-value in the parental genomes is p-value difference > 0. In other words, every BS significance difference, irrelevant from its size, is considered as an indicator for differential binding, so that only a binary cut-off, namely BS-significance difference “Yes/No” is applied. This analysis follows two aims: a) to detect how well the models are able to detect differential binding based that will be validated against observed read count differences and b) to compare the different models with each other.

The following Table 21 lists the respective values gained from the analysis, illustrated in Figure 58.

Table 21: Results of the Sensitivity and Specificity analysis for differential binding profiles between the maternal and paternal genome for EGR1, IRF4, MAX, SRF and SPI1 for the ET, PWM and CLTree model.

TF	Model	Model Based Classification				Reference Set		Sensitivity TP/(TP+FN)	Specificity TN/(TN+FP)
		TP	TN	FP	FN	P	N		
EGR1	ET	17	223	23	23	40	246	0.43	0.91
	PWM	17	225	21	23	40	246	0.43	0.91
	Tree	17	226	20	23	40	246	0.43	0.92
IRF4	ET	4	322	77	2	6	399	0.67	0.81
	PWM	4	329	70	2	6	399	0.67	0.82
	Tree	4	326	73	2	6	399	0.67	0.82
MAX	ET	6	195	34	23	29	229	0.21	0.85
	PWM	6	195	34	23	29	229	0.21	0.85
	Tree	6	194	35	23	29	229	0.21	0.85
SRF	ET	10	158	29	32	42	187	0.24	0.84
	PWM	10	158	29	32	42	187	0.24	0.84
	Tree	11	156	31	31	42	187	0.26	0.83
SPI1	ET	101	1110	80	141	242	1190	0.42	0.93
	PWM	104	1114	76	138	242	1190	0.43	0.94
	Tree	103	1115	75	139	242	1190	0.43	0.94

For all TFs the results for the specificity are in a much higher range than for the sensitivity. This means in turn, that the applied approach to detect differential binding is able to very well classify locations where no differential peak profiles within the parents occur. Over all TFs specificity values are equal or greater than 81%, for SPI1 even up to 94%. However, the sensitivity appears suboptimal. The highest sensitivity is found for IRF4 with ~67%, while the applied approach shows only a sensitivity of ~21% for MAX. This reveals that, assuming a causative event that disturbs direct DNA binding of the TF in question, the probability that the approach detects a differential binding in the parental peaks is only 67% maximum (IRF4).

A comparison of the probabilistic models reveals that all models perform rather equal.

Referring to the fact, that a good classification approach should have both, a high sensitivity as well as a high specificity, the sum of the sensitivity and specificity values can give an idea

about differences in model performance. For IRF4 the PWM model performs slightly better than the tree-based models, wherein the ET model shows slightly less good results than the CLTree.

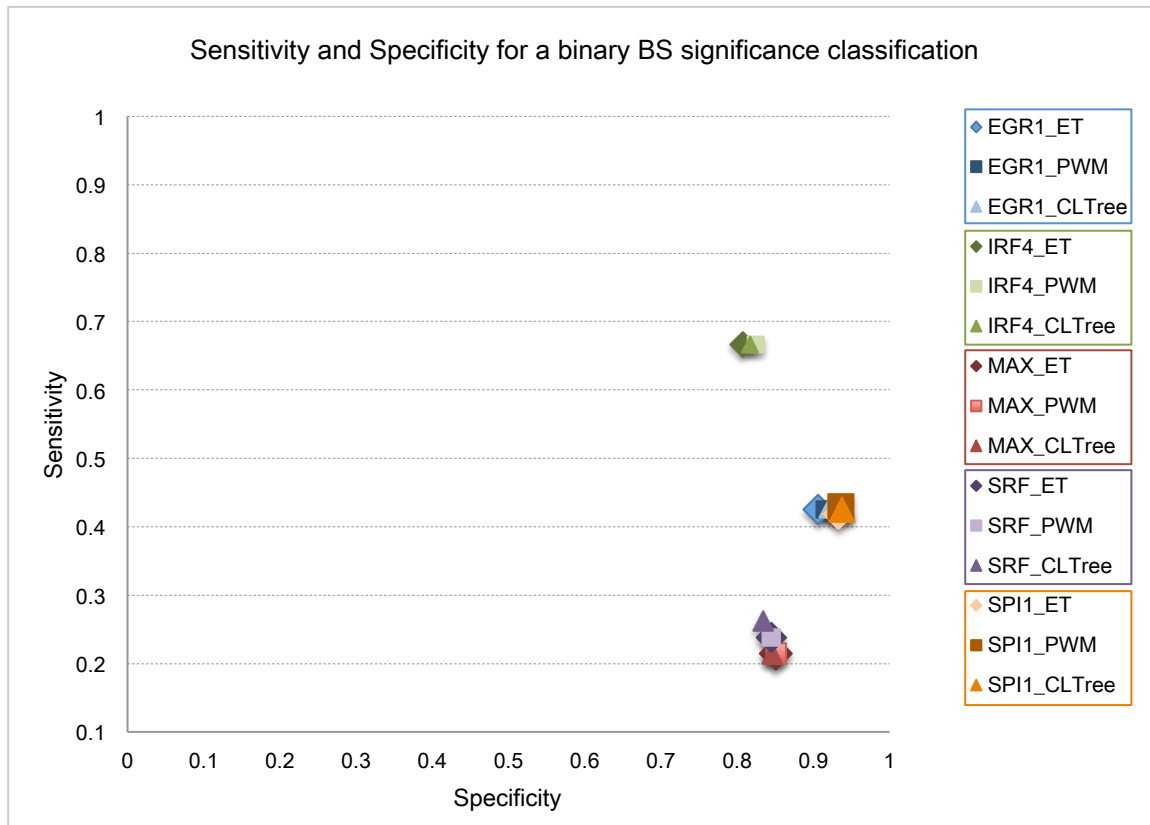


Figure 58: Sensitivity/Specificity plot for model comparison for EGR1, IRF4, MAX, SRF and SPI1

The difference determined is exclusively based on the level of specificity, the sensitivity value are for all three models the same. In SRF and SPI1 the CLTree shows the best performance due to a higher sensitivity, while in IRF4 the PWM shows the best performance. However, considering the results from a more general view, they are all very close to each other and an eye-catching difference between them cannot be detected (see discussion).

4.3.5.2. Winning Parent

Explored to the child, the winning parent determines the parent whose haplotype is the one mostly attributing to the ChIP-seq signal, and therefore TF binding, in the child's ChIP-seq experiment.

For the investigation, if the winning parent is predicted correctly with regard to the reference set, one needs to take into account the different approaches used. Alleleseq assigns the winning parent based on mapped-read-counts, while the other approach is based on BS significance difference. Thus, winning parent prediction can only be compared, for those

parental peak combinations, where Alleleseq is actually assigning a winning parent. Alleleseq does not assign a winning parent, when the genotypes of the parents are heterozygous, so that it is not possible to deduce, based on read counts, from where the actual allele, showing the higher read-count is inherited (see Figure 59).

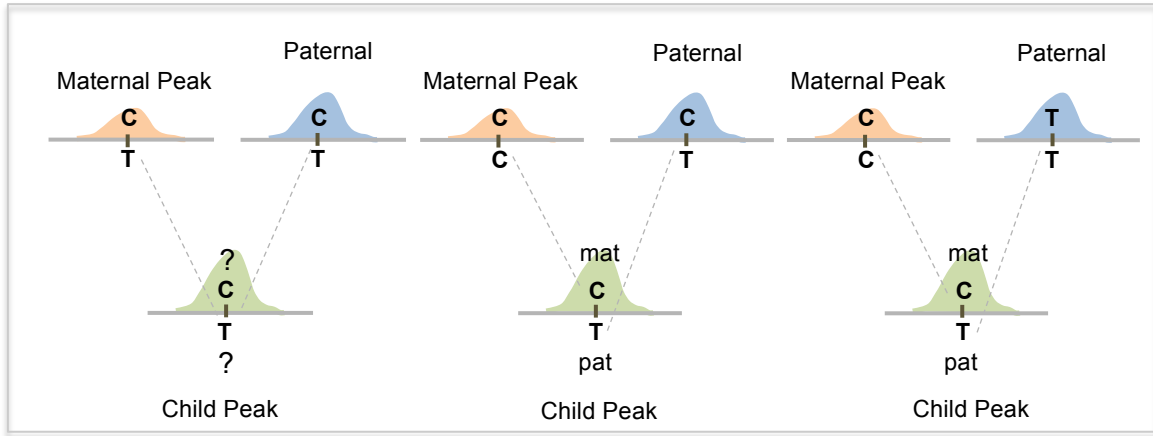


Figure 59: Graphical illustration of winning-parent assignment of Alleleseq

Consequently, in this regard only a respective subset of the positive reference set will be considered.

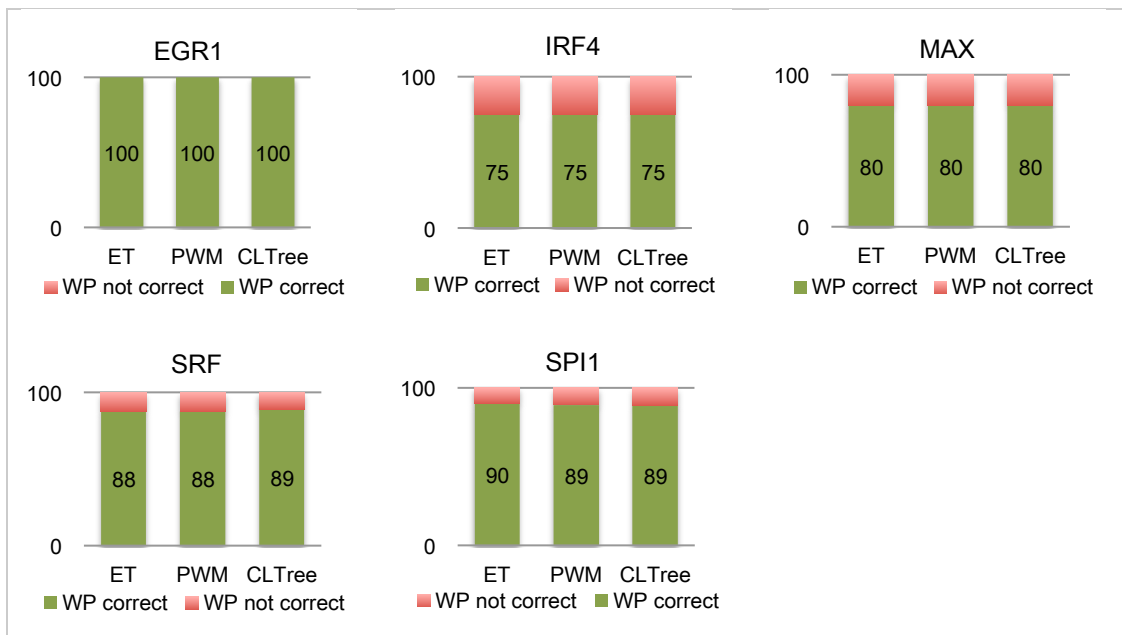


Figure 60: Column charts illustrating the relative amount of correctly and incorrectly predicted winning parent for all TFs and all models. WP match: correct assignment referring to the filtered reference set, WP no match: incorrect assignment of winning parent.

As one can take from Figure 60 the applied approach is performing very well in predicting the correct winning parent, with a correct prediction of maximum 100% in EGR1 and a minimum of 75% in IRF4. This result let suggests, that the UniPROBE trained models provide in general

reliable results. Furthermore it shows, that in at least the filtered subset, the number of mapped reads seems to correlate well with the significance of the best BS detected. Thus, it should be possible to develop a descriptor, being able to detect differential binding, based on probabilistic TFBS detection models.

4.3.6. Detection of Causative SNPs

4.3.6.1. SNP-BS co-location and BS Sequence Differences

That a genetic variant is co-located with a peak does not automatically mean that it has an impact on the binding affinity of the respective TF. In order to influence the binding of a TF the variant has to be placed within a BS, remodeling the binding affinity between the DNA and the binding domain of the TF.

From 3,646,764 SNPs listed for the 1000G family trio, 1,704,167 SNPs are heterozygous in the child. Accordingly, the available pool of SNPs, to investigate with regard to a possible causative effect, correlates with 46.7% of the complete 1000G CEU family SNP set. Next, the input set of parental peak pairs (detected in both parents and in only one) has been screened with each model for the BS with the highest BS significance. After, SNPs being co-located with the BS have been spotted, having the potential to influence the BS significance, so that a differential TF binding profile between the parents is observed. Figure 61 illustrates a summary graph for all TFs, displaying the relative amount of detected BS and their “SNP-status”.

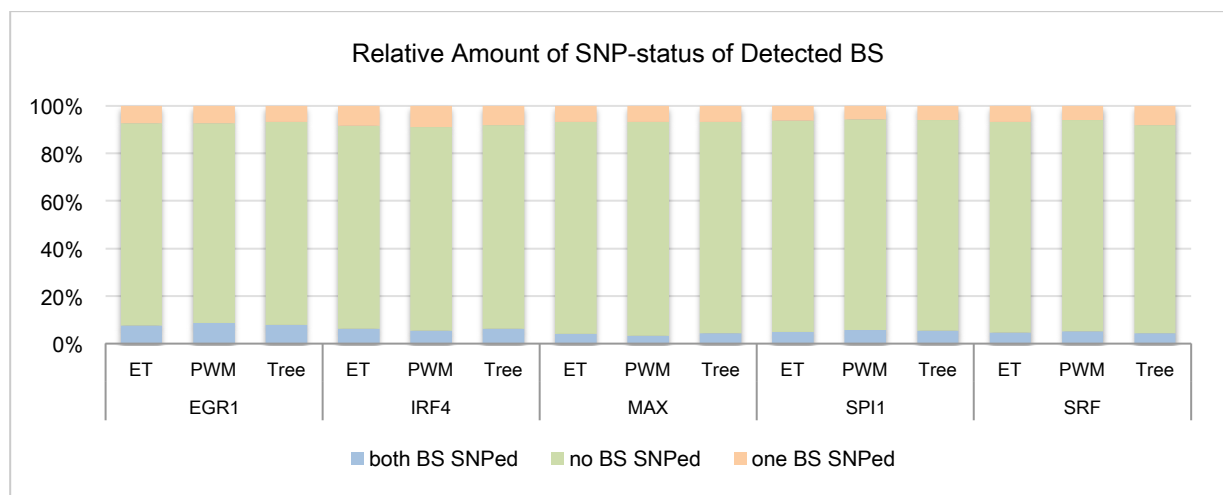


Figure 61: Bar chart illustrating the relative amount of parental BS, where both, only one or none BS are co-located with a SNP.

As one can take from the above, in the majority (~87%) of detected BS, no SNP is observed in both (“no BS SNPed”), the paternal or maternal BS. The amount is stable over all TFs

analyzed and models applied. The amount of BS, where only one parental BS has a SNP-co-localization, represents the second group, with on average 7%. Finally the smallest group, comprising those BS and SNPs, considered as high-confidence candidates, represents only ~6% of all BS detected. Furthermore, considering only SNPs affecting both parental BS as potential causative candidates, its number drastically reduced (see Table 22).

Table 22: Absolute number of SNPs with the potential of having a causative effect.

	EGR1	IRF4	MAX	SPI1	SRF
ET	23	27	11	75	11
PWM	26	24	9	87	12
CLTree	24	27	12	82	10

Referring to the input set of heterozygous child-SNPs, the candidate list represents, after filtering for BS-co-location in the parents, between 0.0005 % (MAX) and 0.005% (SPI1).

Finally, a significance measure can be assigned only for SNP effects where the parental BS differ by exactly one nucleotide, reducing the candidate SNP list further (see Figure 62).

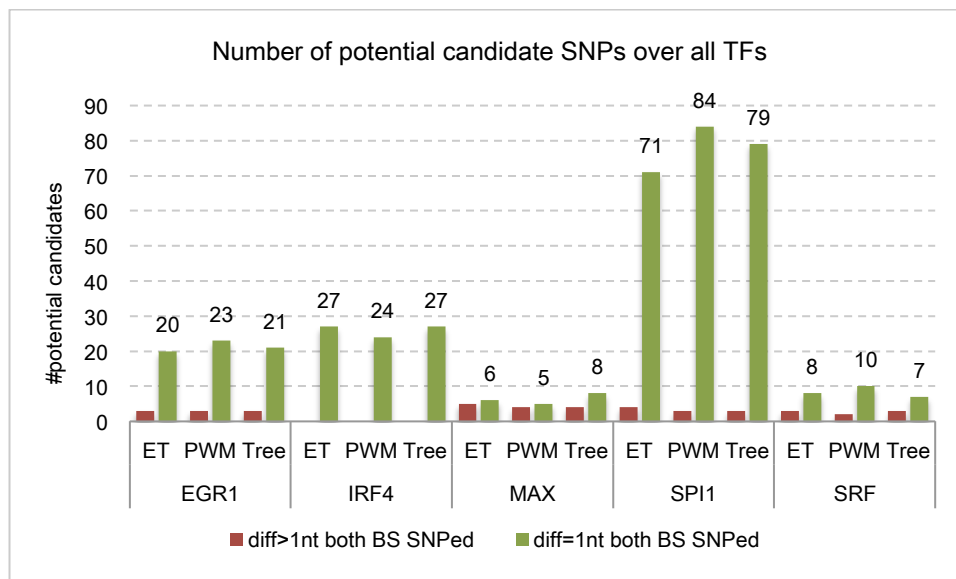


Figure 62: Graphical illustration of the final amount of potentially causative SNPs. Due to significance assignment of the SNP effect the dataset reduced slightly, neglecting those SNPs sitting in BS differing by more than one nucleotide. For example for SPI1 the number of potential causative SNPs is reduced from 75 to 7 (ET model).

4.3.6.2. Candidate List of Causative SNPs

The candidate list is generated by overlapping the TP predictions from the applied model screenings with the high-confidence candidates (see former section). By considering this overlap, support is on hand for a potentially causative SNP based on both, a significant difference in mapped-read-counts and a difference in the measured BS significance between

the parents due to a single nucleotide exchange. Furthermore, only those SNPs detected by all three models and with a SNE p-value ≤ 0.05 are considered.

For two TFs, namely MAX and IRF4, not a single SNP with a significant SNE-p-value has been found, therefore no potentially causative SNP can be reported. The following table holds respectively all potentially causative SNPs for EGR1, SRF and SPI1.

The model determining the number of predicted causative SNPs is the ET model, providing the smallest, but best confirmed, amount of significant SNPs. Over all TFs, except for one, all SNPs detected by the ET model are also detected by the other models. Only one SNP is not confirmed by the CLTree, but only the PWM model. Not a single SNP has been exclusively detected by tree-based models and most SNPs have been predicted by the PWM model (see Figure 63). Accordingly, the resulting candidate list provides quasi all significant predictions of the ET model, providing five potential causative SNPs for EGR1, two for SRF and 13 for SPI1.

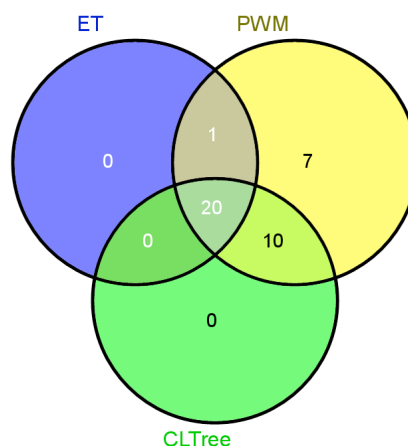


Figure 63: Venn Diagram illustrating that the resulting candidate list is quasi exclusively based on the ET model based rSNP detections.

None of the SNPs in the candidate table is listed in the GWAS catalogue, meaning, that none of those has been until now associated within a genetic association study with a phenotypic trait (LA Hindorff *et al.*, n.d.; Lucia a Hindorff *et al.*, 2009). Furthermore, the causative SNP candidates for EGR1 have been submitted to is-rSNP, a webservice- implementation¹⁴ of the regulatory SNP detection approach from MacIntyre and colleagues (Macintyre *et al.*, 2010). Is-rSNP assigned two out of five EGR1-candidates as regulatory SNPs (adjusted p-value cut-off 0.05), namely rs4842838 and rs55878408. For rs4842838 the list contained 90 hits for 77 different TFs, wherein EGR1 ranked on position 23 ($p=0.006$) and for rs55878408 the observed rank is 43 of 48 ($p=0.04$) within 43 different TFs respectively.

¹⁴ is-rSNP - SNP predictions made quick and easy
<http://www.genomics.csse.unimelb.edu.au/product-is-rSNP.php>

RESULTS

Table 23: List of predicted causative SNPs for EGR1, SPI1 and SRF. Criteria: Significant difference in mapped-read-count, difference in BS significance, SNP is co-localized in both parental BS, SNE p-value <0.05. chr: Chromosome, GT: genotype, BS seq: BS sequence, BSos: BS offset position within the peak sequence (1-based), str: strand, WP: winning parent.

TF	chr	SNP	Maternal Genome				Paternal Genome				SNE p-value	WP
			GT	BS seq	BSos	str	GT	BS seq	BSos	str		
EGR1	17	rs7502391	TC	GCCCCCGCA	33	+	TC	GCCCCTGCA	39	+	<=0.008	M
EGR1	15	rs4842838	TT	TCCCA A GCA	58	-	GG	TCCCAC G GCA	58	-	<=0.014	P
EGR1	19	rs4239605	GG	GCCCAC G GCA	54	+	AA	GCCCAC A CA	52	+	<=0.022	M
EGR1	17	rs4794799	TC	GCCC G CGCT	41	-	TC	GCCC A CGCT	29	-	<=0.026	P
EGR1	10	rs55878408	GG	GCCCC C TCA	43	-	AA	GCC C TCTCA	40	-	<=0.028	M
SRF	22	rs5770871	GA	CCAAATAAG	76	+	GG	CCAAATGAG	77	+	<=0.005	M
SRF	19	rs8107390	CC	CCT G AAAAG	55	-	CT	CCT A AAAAG	55	-	<=0.01	P
SPI1	17	rs2322709	CA	CCTTC C TCT	47	+	CA	ACTTC C TCT	45	+	<=0.004	P
SPI1	16	rs72777743	GA	AC C TCCCCT	49	-	GA	ACTT C CCCCT	51	-	<=0.005	P
SPI1	18	rs8096199	GA	ACTTC C TCT	69	-	AA	ACTT C TTCT	69	-	<=0.005	M
SPI1	6	rs9272536	GG	ACTT C CCCT	35	-	TG	ACTT C ACCT	35	-	<=0.007	M
SPI1	15	rs10519312	AG	AC C TCCCTA	48	-	AA	ACTT C CCCTA	33	-	<=0.013	P
SPI1	14	rs56247771	CC	ACTT C CCCT	48	+	CT	ACTT C TCCT	48	+	<=0.014	M
SPI1	2	rs55900716	GA	AC C TCCCCC	57	-	GA	ACTT C CCCC	73	-	<=0.015	P
SPI1	8	rs4288343	AT	TCTTC C TCA	55	+	AA	ACTT C TCTCA	53	+	<=0.02	P
SPI1	2	rs13394359	AG	ATTTC C TCT	9	-	GG	ACTT C TCTCT	9	-	<=0.027	P
SPI1	10	rs11597781	CA	ACTT A CCAT	49	+	CA	ACTT C CCAT	33	+	<=0.027	P
SPI1	7	rs2249189	CC	ACTTC C TCT	20	+	GG	ACTTC C T G T	28	+	<=0.031	M
SPI1	5	rs6874323	CC	C GTT C TCTCT	52	+	CA	A GTT C TCTCT	56	+	<=0.032	P
SPI1	20	rs386274	CC	C GTT C TCTCT	57	+	AA	A GTT C TCTCT	61	+	<=0.032	P

All 20 SNPs have been submitted to the RegulomeDB (see section 1.10). 15 out of 20 of the candidate SNPs gain support by the RegulomeDB analysis, with a minimum score of 2b.

Table 24: Results of the candidate SNP RegulomeDB analysis

TF	Chr	SNP pos	rs	RegulomeDBScore
EGR1	chr10	89877238	rs55878408	2b
EGR1	chr15	84582124	rs4842838	2b
EGR1	chr17	37213252	rs4794799	2a
EGR1	chr19	1099701	rs4239605	1b
SPI1	chr10	3849608	rs11597781	2b
SPI1	chr14	90348415	rs56247771	2a
SPI1	chr16	17288207	rs72777743	2b
SPI1	chr17	12843178	rs2322709	2a
SPI1	chr18	3747619	rs8096199	2a
SPI1	chr2	75831740	rs13394359	2a
SPI1	chr2	99109132	rs55900716	2b
SPI1	chr20	55030081	rs386274	2b
SPI1	chr7	83097657	rs2249189	2a
SPI1	chr8	99305539	rs4288343	1a
SRF	chr22	50978262	rs5770871	2a

In the following further details for the top SNP of EGR1 and SRF (SPI1 see Annex) are provided. To support the results with further biological data three histone modification tracks, namely H3K27ac, H3K4me1 and H3Kme3 and a POLII track have been added. The histone marker tracks have been taken directly from UCSC (Rosenbloom *et al.*, 2010) generated on the cell line GM12878, corresponding to the trio child. Even if no *de novo* analysis based on the parental genomes (according to section 3.7.6) has been performed, these tracks are considered as biological support. However, it might be possible that there is a differential signal in the parents, which cannot be observed. The biological impact of these histone modifications is as follows: H3K27ac has been shown to distinguish active enhancers from inactive/poised enhancer elements (Creyghton *et al.*, 2010) and H3K4me1 and H3Kme3 have been characterized as strongly enriched around the TSSs or in general to mark enhancer regions (Tian *et al.*, 2011). For POLII, marking transcriptionally involved and in particular TSS regions, a *de novo* analysis on the parental genomes based on raw ChIP-seq reads provided for the trio child has been done.

Referring to Figure 64, even though the top causative SNP candidate rs7502391 and the respective BS are located in intron 11 of the gene *RPTOR* (see Figure 64) the region seems to be an active enhancer, mirrored by the signal of the H3K27ac. Moreover, H3K4me1 and H3Kme3 support the detected BS as an enhancer and TSS region respectively. The SNPed BS it is closely located, but not overlapping, with a POLII peak, also supporting a biologically relevant region.

However, the signal of POLII does not reveal any eye-catching difference between the paternal and maternal tracks. Hence, an effect, like altered gene expression, is strongly unlikely. Furthermore, since no TSS is in the direct vicinity, this region seems to be affected by DNA looping, supported by several other POLII peaks in proximity. Beside the EGR1 peak directly overlapping with the binding site, several other EGR1 peaks appear, suggesting that in case of the “drop out” of one enhancer region another one might take over. Thus, based on those genetic marker tracks and considering the SNP on a higher level of biological consequences, it cannot be stated to expect a significant change regarding gene expression.

RESULTS

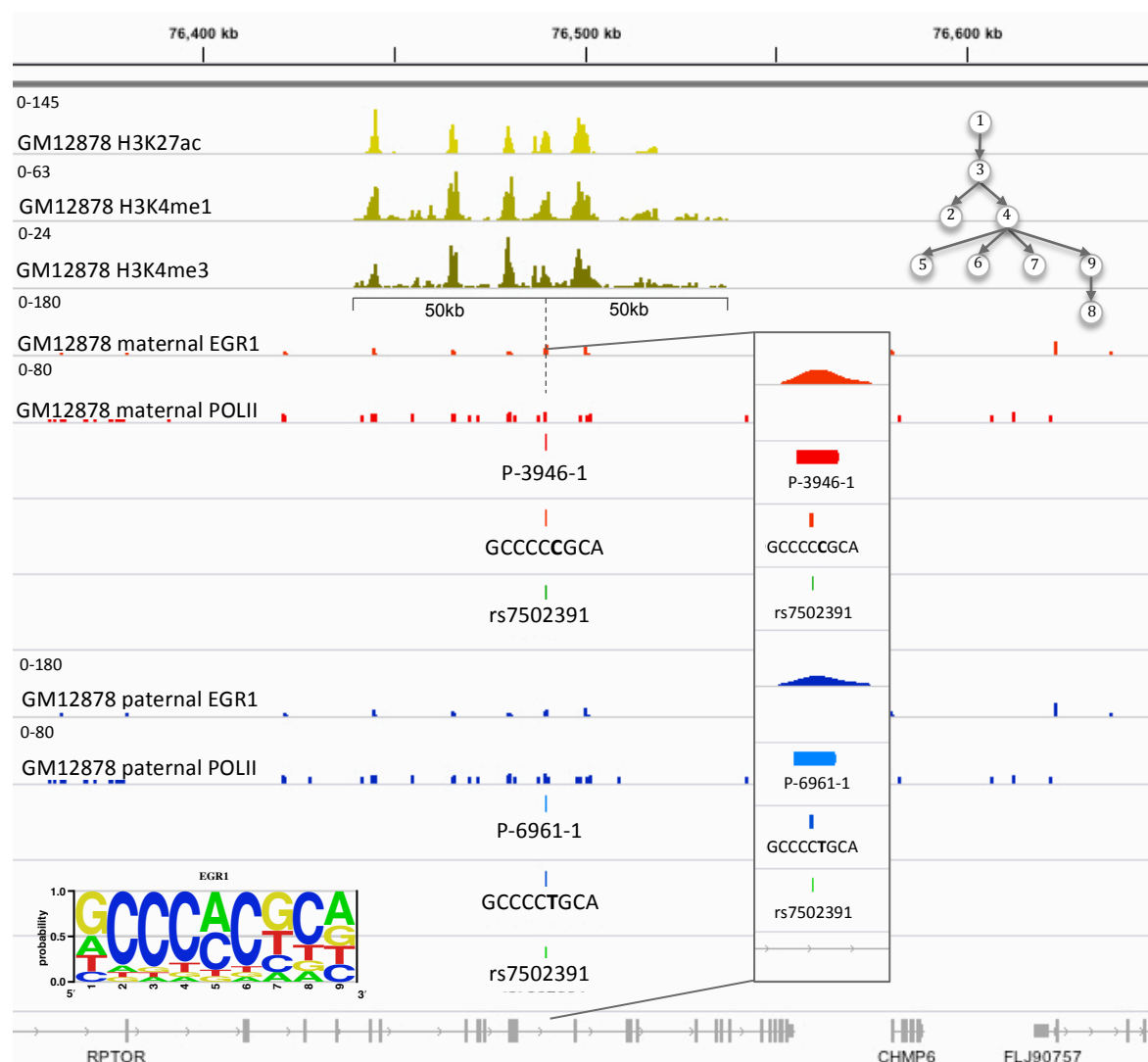


Figure 64: Top causative SNP candidate for EGR1. Illustration of genomic environment of rs7502391 including enhancer tracks for BS location +/- 50kb (histone acetylation and methylation). Lower left corner: Sequence logo for EGR1 trained PWM model, upper right corner: CLTree trained for EGR1

Nevertheless, that the SNP might be responsible for the differential TF binding profile in the parents is strongly supported by the fact that the paternal peak sequences are highly similar. They only differ by a tiny flanking sequence of 6 nucleotides and the SNP being placed close to the sequence centers (see Figure 65).

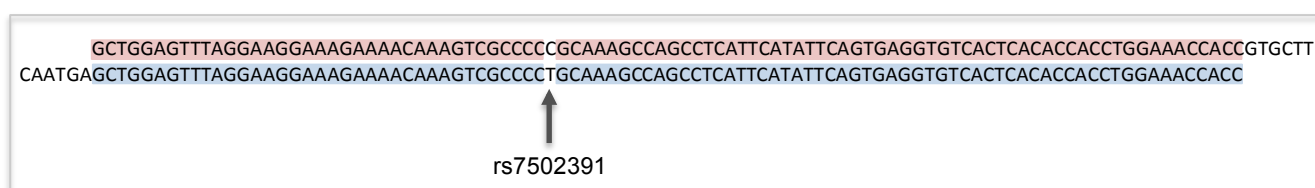


Figure 65: Paternal (blue) /maternal (red) peak sequence alignment, showing minor differences at the flanking regions and one differing nucleotide close to the centre represented by rs750239

The TFBS landscape (see Figure 66), below the peak center, reveals that the detected BS shows an outstanding significance in comparison to all other possible 9-mers screened, supporting a correct BS location based on the model prediction. Furthermore, the difference between the maternal and paternal BS peak, captured in the SNE-p-value, is highly significant. The SNE-p-value evaluating the nucleotide exchange "C" vs. "T" is for example in the ET model ranked on position 18,604 of 7,077,194 ranks.

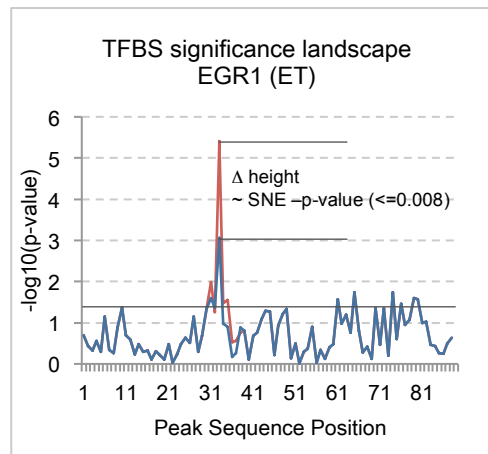


Figure 66: TFBS landscape for EGR1 with the ET model (representative for PWM, CLTree) and maternal and paternal peak sequences. Both the maternal (red) and paternal (blue) curves show their maximum on position 33 (truncated sequence). The horizontal line marks the BS significance level of $-\log_{10}(0.05)$. The peak sequences have been truncated, so provide a direct overlap of the maximum. This truncation (cut of non-aligned nucleotides) has no influence on results of TFBS detection.

The sequence logo of EGR1 shows that "C" on position 6 is highly conserved in contrary to "T". The consultation of the paired marginals used to build the tree models revealed that the nucleotide pairs (C_2, C_6), (C_3, C_6), (C_4, C_6) are highly conserved, even ranging in the top 5 of conserved nucleotide pairs. As one can take from the CLTree structure (see Figure 64) all three nucleotide pairs are captured within a sub-tree suggesting positional interdependencies. The values of the paired marginals, reflecting the probability that a nucleotide pair contributes to a "good" BS, revealed for (C_2, C_6) 77%, (C_3, C_6) 79% and (C_4, C_6) 73%. A nucleotide exchange from "C" to "T" on position 6 reduced this probability values to (C_2, T_6) 2%, (C_3, T_6) 1% and (C_4, T_6) 6%.

The BS covering rs5770871 (see Figure 67) is not placed within a gene body, but close to the center of a genomic region (~18kb) surrounded by four genes, namely *TYMP*, *SYCE3*, *KLHDC7B* and *ODF3B* (source UCSC). A search with the keywords "SRF" "TYMP", "SRF" "SYCE3", "SRF" "KLHDC7B" and "SRF" "ODF3B" in PubMed and KEGG (Kanehisa, Goto, Sato, Furumichi, & Tanabe, 2012) did not reveal any known direct target gene connection with any of those four genes to SRF.

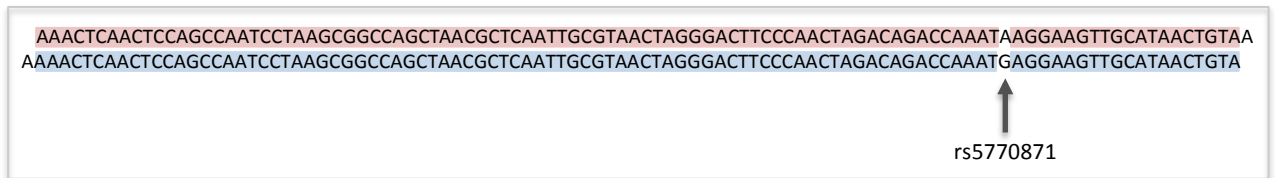


Figure 68: Paternal (blue) /maternal (red) peak sequence alignment, showing, apart from the location of rs5570871, nearly a perfect match.

This sequence similarity supports the claim, that the SNP detected is causal for the detected differential parental SRF binding profiles. Also the TFBS landscape (see Figure 69) for the peak regions reveals, a signal for the BS detected, strongly differing from the remaining 9-mers screened. This provides further support, that the detected BS is the "real" one. The assigned SNE-p-value corresponds with the rank 32,174 within 7,077,194 9-mer pairings and represents therefore a very strong statistical significance.

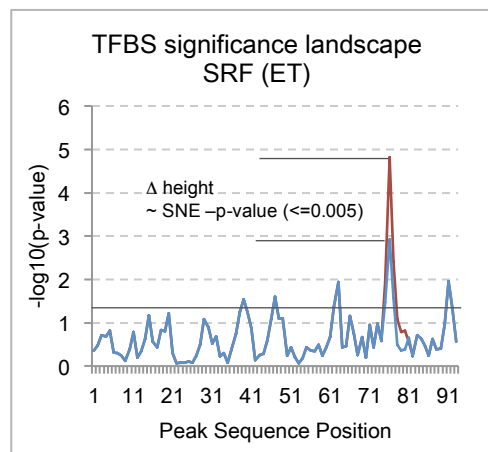


Figure 69: TFBS landscape for SRF with the ET model (representative for PWM, CLTree) and maternal and paternal peak sequences. Both the maternal (red) and paternal (blue) curves show their maximum on position 76 (truncated sequence). The horizontal line marks the BS significance level of $-\log_{10}(0.05)$. The peak sequences have been truncated, so provide a direct overlap of the maximum. This truncation (cut of non-aligned nucleotides) has no influence on results of TFBS detection.

The SNP leads in the parental BS to a switch from "A" (mother) to "G" (father) on position 7. With regard to the PWM a change from "A" to "G" would not have a tremendous effect, since the nucleotides on this positions (as many others for the STEME trained PWM) are relatively equally distributed. However, considering the official UniPROBE matrix, position 7 is assigned either with an "A" or a "T" but not with a "G".

Also for this SNP the paired marginals have been reviewed and it has been found that from all possible nucleotide pairings with position 6, the nucleotide pair (A_5, A_7) shows the highest probability ($\sim 29\%$) of being present in a "good" BS. The tree structure reveals further a dependency relation between position 5 and 7. The nucleotide exchange on position 7 from "A" to "G", resulting in the nucleotide pairing (A_5, G_7) shows only a probability of 5%. Furthermore,

the fact that the pairing (A_5, A_7) shows the highest probability suggests that any change within this specific nucleotide pairing, including (A_5, T_7) might have a strong effect.

5. Discussion

5.1. Probabilistic Transcription Factor Binding Models

The issues of modeling and inferring the sequence specificities of TFs from high throughput binding measurements provided by the UniPROBE database, in more detail the *Mus musculus* subset, have been addressed.

The majority of TFs tested, namely 86 out of 104 TFs, showed better AUC profiles for the tree-based models than for the simple naïve model of a PWM. This does not automatically mean, that all these TFs exhibit positional interdependencies, but it shows, that the application of tree models does not impair the predictive power and can be considered as a real alternative. In particular, if no prior knowledge about possible dependencies is available, the usage of tree models would make sure, that in case they are relevant, these dependencies would be detected. In fact the tree model includes the PWM model intrinsically, so that it is nearly surprising that for 18/104 TFs a better performance of the PWM model is observable. However, this might be an issue of the validation set or the training procedure in general.

This leads to one of the main aspects to discuss – the selection of the training and validation sets. The classical approach is based on cross-validation (Kohavi, 1995). This way of model training and validation is mainly used, if a predictive model should be developed and one wants to estimate how accurately it will perform in practice. The core procedure includes that the sample is partitioned into complementary subsets, wherein one of these subsets serves as a training set, while the other one serves as a validation set. The training set in turn can again be subdivided in two parts: one to learn the model and the other one to optimize its parameters. When the model learning is finished and optimal parameters are found, the validation is performed on the validation set. In order to reduce variability, this core procedure of training and validation is performed in multiple rounds, in each round using a different partition. Finally the validation results are averaged over all rounds.

In the work on hand, it has been decided to directly train and validate the models on “quasi real life” data provided by the UniPROBE database. This data source is providing nearly for all TFs, captured within the database, results for two independently performed PBM experiments. Moreover, not only processed but also pure raw data are available. The qualification as “quasi real life” is based on the design and experimental setup of these array experiments. The set of sequences is complete with regard nucleotide sequences with a length of ≤ 10 , achieved by a de Bruijn sequence of order 10. Furthermore, the focus of UniPROBE’s array experiments is

not to investigate the TF binding under certain biological conditions, but purely TF binding. Whether all these sequences are serving as real TFBS in a natural biological system is questionable, as it is not expected that nature offers a TF such a space of accessible TFBS. Nevertheless, for the purpose, to learn a probabilistic model, a complete set of sequences assigned with a measured binding signal, gained from two independent experiments, is highly qualified.

The design and thereby the properties of the two array sets are highly comparable. Consequently, data from the first array have been used to learn the models, while their predictive power has been tested on data from the second one. The UniPROBE data are provided as probe sequences composed of a 36bp long variable region holding the de Bruijn derived 10-mers and a constant region of 24bp, serving as a primer sequence to generate double stranded DNA molecules. In order to gain statistical power, the models have not been learned on 10-mers, as each of them is represented exactly once on the array, but on 9-mers, which should be 8-times available on an array. However, “real” biological data always come along with experimental biases, like measurement or systematic errors. Since no influence can be taken on the data as such, at least one known confounder within the training data pre-processing pipeline (see chapter 3.3), namely the distance of the 9-mer to the glass slide, has been addressed.

In order to gain the final input set, 9-mers need to be extracted from the variable region of the probe. The challenge in this regard is to extract exactly that 9-mer from the probe sequence responsible for the measured binding signal. This is a very complex problem and not within the scope of this work. Hence, it has been decided to use a *de novo* motif finder program called STEME to fulfill this task. Additionally, STEME has been applied in a specific manner, based on the top and worst 800 probe sequences with regard to the measured binding signal. Even if this approach reveals those 9-mers being with high probability a binding site, and therefore a superficially very good training set, a bias is introduced affecting the quality of the tree-based model learning. This bias is represented by the fact, that STEME is using a PWM model in its EM algorithm. Accordingly, the 9-mer sequences to train and improve the models, are based on the independence assumption. As a consequence, it has to be considered, that sequences showing a high positional interdependence are probably not reported, if they do not also match the consensus sequence found. Thus, it can be concluded that the *de novo* motif finder used, provides appropriate training sequences for the PWM model, but sub-optimal ones for the tree-based models. Apparently, there is no “ready-to-use” application available functioning as a motif discovery tool with regard to positional interdependencies. Thus, this bias has been identified and will be approached in future work. In case such a tool would exist, the optimal learning of PWM and tree-based models would have to be performed

on respectively different training sets and even better results for the tree-based models are to expect.

With regard to the model training, another issue, of probably not taking the full advantage of the ET model has to be reflected. While the training of the ET model can be in principle performed by using the same training set as for the other models, this does not apply to the model parameter α , determining the parameter β . Formally spoken a dataset being different from the main model training set has to be used. This requirement is not perfectly addressed in this work, as the dataset employed still includes parts of the model training set and is therefore biased. Nevertheless, the applied approach is considered as approximately statistically valid, since not the identical 9-mer training input is used to optimize α , but all top 800 36bp-variable probe sequences used to derive these 9-mers. The difference between these two sets should be satisfying, because STEME does not detect in each of the 800 probes a sequence. Hence, also probes are considered, not being touched during the main model learning. Additionally, another, more serious difficulty has to be solved, which would probably lead to a significant improvement of the ET model. This problem is not detectable, when optimizing α within a range of 0 to 250, as performed in this work intuitively. However, during pilot studies directed to improvements of the α -training in future, the way $\beta(e^{(\alpha \cdot MI)})$ has been determined turned out to be a non-convex problem being sensitive to numerical instabilities (see Annex).

A possible promising alternative seems to be an iterative approach where the β matrix is optimized simply based on MI -values. This approach is free of numerical instabilities, convex and no α -value is required. Moreover, the usage of the model training set would be statistically valid. This approach has been tested already in a small pilot set of TFs and promising results (data not shown) have been gained.

This approach, still to optimize, works as follows

1. Set all values in the diagonal within the β matrix on "0" and all others on a small value greater than "0", for example 0.00001
2. Sort MI descending
3. Take this first (highest) MI and assign the respective matrix cell with its value
4. Screen the training set (e.g. STEME output 9-mers)
5. Sum up all log-likelihoods and associate the value with the generated β matrix
6. Take the second highest MI and assign the respective cell with its value, etc

This procedure continues iteratively over all MI values and is finished when the lowest MI has been selected. Consequently, the β matrix is completely assigned with MI values (except the diagonal). Finally, the β matrix associated with the maximum sum of log-likelihoods is selected for the final ET model.

The most similar approach to the ET model applied in this work has been published by Sharon and colleagues (Sharon *et al.*, 2008), termed feature motif models (FMM). It suggests to describe a set of sequences (known TFBSs) based on sequence properties or features being relevant to the TF-DNA interaction. It is based on the assumption that the binding affinity of a given site to the TF increases if it contains more features being important for the TF to recognize its binding site. They consider binary features, like “C” on position 3 and “T” on position 6, multivalued features like “the number of “Gs” or “Cs” at positions 1 to 4”, as well as global features like “the sequence is palindromic”. To each feature a statistical weight corresponding with the importance of the feature to the DNA-TF interaction is assigned. The overall strength or score of a putative TFBS is assigned by summing up the contributions of all features mirrored by the TFBS. The model used is based on a Markov network, providing a natural framework for compact representation of a distribution as a set of feature functions.

The common aspect of the FMM and Ensemble of Trees is the modeling of features contributing to a TFBS. Intuitively spoken, the FMM aims to identify which of the features that are important for the TF–DNA interaction are present in a sequence, and to sum their contributions to obtain the overall affinity of the TF to the site. This means in turn that features have to be defined *a priori*. Besides, for each feature a maximization of a likelihood-function is performed to gain the specific weight of a feature. The advantage of the ET model to not require feature definition comes along with the disadvantage that the captured features are “invisible” for the researcher. However, in comparison to FMM the ET model appears positively trivial. It allows efficient implementation simply based on the inference of di-nucleotide pairs and intuitively captures dependency structure in form of a mixture of trees.

5.2. Detection of Differential Binding in ChIP-seq Experiments

The prediction of the winning parent, exclusively based on the difference between the parental BS scores, reveal very good results, correlating with peak profiles, reflecting normalized read counts.

The probabilistic models turned out to perform very well referring to specificity, while providing sub-optimal sensitivity. However, in this regard the group of FN, having a strong impact on the

sensitivity, should be considered more deeply. Here, the different approaches used to detect a differential peak profile have to be taken into account. A simplifying assumption was made that a significant read count difference is due to a distorted BS sequence (of the TF in question). As a consequence, when no BS significance difference is observed the sequence was always assigned to the negative category. The real explanation for the differential read count could very well lie upstream of the TF binding event, at the level of chromatin structure or histone modification patterns. Such misclassification leads to high FN counts. This issue could be addressed by extending the screening approach to evaluate differential histone modification levels between the parental alleles.

A comparative analysis of the screening performance of the three probabilistic models did not reveal tremendous differences. Considering the AUC graphs and the improvement gained by accounting for positional interdependencies by tree-based models, the TFs investigated are grouped as follows:

Table 25: Classification of TFs analysed based on CLTree- and ET model performance in UniPROBE data

	Improvement by CLTree	Improvement by ET
EGR1	medium	yes
IRF4	medium	yes
MAX	small	no
SP11	medium	no
SRF	none	yes

As one can take from the table above, for none of the TFs investigated a strong improvement of predicted power has been gained on UniPROBE data. The ET model improved only 3 out of five. Accordingly, tremendous differences were not to be expected. Nevertheless, since the incorporation of positional interdependencies could influence the power to detect rSNPs, all TFs and all models have been kept in the complete analysis. In fact it would have been very interesting to analyze data of Gcm1 or Myb, showing a strong improvement by the tree-based models, but the availability of ChIP-seq experiments for the CEU family trio limited the TF selection.

Furthermore, considering the fact that the models have been tested on ChIP-seq data, one main issue is on hand: Why not training the models directly from ChIP-seq peaks?

Even if this might sound as a trivial question, there are two aspects to consider:

1. Existence of an independent training set
2. Sufficient sequence variability to train models capturing positional interdependencies

With regard to the first point, one needs to face the problem that for the family trio used, only one ChIP-seq experiment per TF is currently available. Thus, the usage of a subset of peaks, to train and optimize models, would always bias the peaks in the dataset to be investigated for rSNPs. Meaning, the training set would be part of the set on which the trained models are applied. One could think about workarounds, like considering only peaks, not being co-localized with SNPs or an exclusion of the training peaks from the subsequent analysis. However, one question is: are the models trained on SNP-free peaks representative for all peaks? And, assuming that a SNP distorts a nucleotide pair being dependent from each other, do the BSs below SNP-free peaks provide such interdependencies in a representative amount?

To exclude those peaks, used for training, from the subsequent analysis is in so far critical, as the number of detected peaks varies over the different TFs investigated. If a comparable approach as used for UniPROBE was considered, one would need to decide how many peaks to use for the training. To determine a fixed number like 2,000 is questionable, as this represents for example for SPI1 ~7% of peaks and for SRF ~39%. To apply a relative training set size of 10% is also problematic, due to computational costs, representativeness and sequence variability. This 10% correlate for example with 2,800 peaks for SPI1 and with 500 for SRF. Furthermore, the running time of STEME reaches for 2,800x200bp a time frame, not being applicable anymore in a convenient manner, wherein STEME parameters could be adapted. Though, such an adaption, as for example to limit the running time or to only consider a certain amount of seeds, limits the space of model optimization for STEME.

With regard to representativeness and sequence variability the situation becomes more complicated. Here, one needs to deal with the two questions: “What is the aim of a ChIP-seq experiment?” and “What is needed to train a tree model aimed to catch positional interdependencies?”

The aim of a ChIP-seq experiment is to identify mammalian DNA sequences bound by transcription factors *in vivo* (Robertson *et al.*, 2007; Valouev *et al.*, 2008). *In vivo* the binding of a TF is underlying certain restraints, like DNA accessibility or a natural selection and competition. Accordingly, not each sequence theoretically being a TFBS has the same chance to interact with a TF.

ChIP experiments produce sequence sets which are ‘cleaner’ and more importantly highly redundant; in thousands of sequences one can assume to find several instances of binding sites highly similar to one another (Zambelli *et al.*, 2012). Thus, it is to expect that the BSs detected by a ChIP-seq experiment are strongly enriched for a certain subset of TFBS



sequences. This enrichment might be amplified by the ChIP-seq analysis, wherein a peak represents a genomic region showing a significant difference of mapped reads in comparison to a control experiment. In other words, TFBS not providing a very strong protein interaction might be lost. With regard to the biological question, to detect for example target genes being regulated by a TF of interest, a lack of medium or weak BS can be tolerable. In particular if the aim is to find true positive target genes and the occurrence of unknown false negatives is acceptable.

In contrary to a PWM, a relatively high variability of sequences is needed to train a tree model. Training-sequences are required, not necessarily representing the core sequence motif, but the core dependency motif. In order to catch dependencies, one needs data showing different levels of binding to distinguish whether a certain nucleotide pair has a higher impact than others. If a tree-based model will be trained on a dataset with low variability, its performance will be very close to a PWM model trained on the same data. Furthermore, the *de novo* motif finder, to extract the actual binding sequence would favor high frequent motifs, reducing the variability of binding sequences additionally.

In other words, the training of tree models based on a highly enriched BS training set as provided by ChIP-seq experiments would not be fair. This could result in a type III error, which mirrors the discrepancy between the research focus and the research question (Schwartz & Carpenter, 1999)- *"the error ... [of] choosing the wrong problem representation ... when one should have ... chosen the right problem representation"* (Mitroff & Featheringham, 1974) – with the consequence to discard the tree model for the wrong reasons.

To practically illustrate the problem respective data for EGR1 and SPI1 are provided below. A random selection of 2,000 peaks per TF has been made and a motif discovery with STEME has been performed. Assuming that all BS resulting from a ChIP-seq experiment are “good” sites, the discriminative approach as applied for the UniPROBE data has not been followed (ChIP-seq experiments do not provide data referring to “bad” or “no” binding sites). The random selection assured, that a peak was selected only once. Accordingly, the input sequence set held 2,000 different peaks. With regard to the number of detected peaks (EGR1 ~10,000, SPI1 ~28,000), this random selection and motif discovery has been repeated for EGR1 twice and for SPI1 five times. As a measure of peak goodness, to be considered as a weighted count, the enrichment score has been used. After, it has been investigated how many different binding sequences were detected and how the trained models perform in a randomly chosen validation set excluding training set peaks. With regard to the ET model, the α -values 10, 50, 100 and 150 have been considered.

Table 26: Number of different binding sequences for SPI1 and EGR1 extracted from ChIP-seq and UniPROBE data, functioning as input for model learning. Numbers in parenthesis refer to the amount of different training peaks.

	EGR1	SPI1
Per run	41 (2,000)	118 (2,000)
Merged runs	49 (2,598)	317 (7,461)
Sequence logo (2,000 peaks)		
800 UniPROBE probes	382	418

As one can take from Table 26, the number of different binding sequences below the peaks is very low. STEME discovered for EGR1 41 and for SPI1 118 different binding sequences from 2,000 peak sequences. A merge of the peak sets, representing $\sim 1/4$ of all available peaks, did not increase the number of different binding sequences for EGR1, while the ones for SPI1 doubled. Moreover, if several sequences are discovered within one peak, all of them are implementation wise deleted from the training set. Thereby, the number of sequences actually used for the training is even lower than the numbers listed in Table 26. For example, for EGR1 the number is reduced to 39. A comparison to UniPROBE shows, that here, only the top 800 probes, correlating with only $\sim 2\%$ of available datasets, reveal for example for EGR1 already 9 times more different sequences than a run with 2,000 randomly selected peaks correlating with $\sim 20\%$ of available peaks. Furthermore, considering the sequence logos, generated from the STEME detected motifs, the sequences are all very close to the consensus sequence and the core motif nucleotides are highly conserved.

As one can take from Figure 70, the model performance is all over not as good as observed for UniPROBE (on UniPROBE data). The reason for that might be, that below a peak several “good” binding motifs occur (Zambelli *et al.*, 2012), all-contributing to the enrichment score, while the screening program only evaluates one of those.

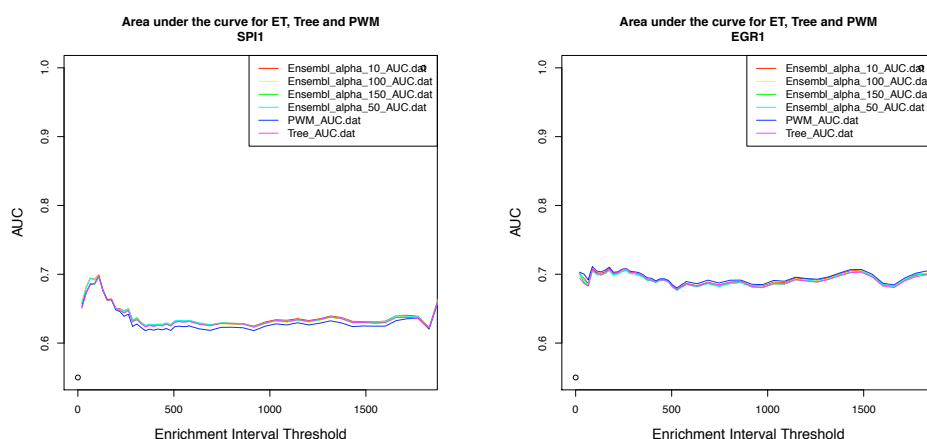


Figure 70: AUC profiles for EGR1 and SPI1 trained and validated on 2,000 randomly chosen ChIP-seq peaks.

However, for both TFs the AUC profile still ranges above 0.5, wherein the profile for EGR1 is better than the one for SPI1. As expected a difference between the three models is hard to determine, wherein this missing difference can be attributed to the low variability of training sequences and detected sequences.

Accordingly, in order to have the chance to observe differences between the three models to detect differential binding, the UniPROBE trained models have been applied to screen for BS below ChIP-seq peaks. Of course, the training data are not directly comparable, but they are independent from each other and UniPROBE provides a higher variability of training sequences, allowing also the training of tree-based models.

The resolution of a ChIP-seq experiment cannot be overcome. Thus, the possibilities to address the problem of low sequence variability in ChIP-seq peak sequences are limited. A relaxed peak detection threshold might increase the binding sequence variability, but might also introduce a bias of unknown size. A better way to approach the paradigm of a favored low false positive rate in ChIP-seq experiments and the required high sequence variability for tree-based models could be data fusion. Data fusion means to integrate various data and knowledge from different sources in a consistent and useful manner. In this regard the data and knowledge should be complementary directed to the same research question. Since the only improvable levels are either the training set or screening procedure, the various data and knowledge could be represented by a UniPROBE binding model integrated in the motif discovery or BS screening in ChIP-seq peak sequences. Practically this could be done by a Bayesian approach, wherein the likelihood for a TFBS derived from a model based on a complete sequence set like UniPROBE, could function as the prior.

5.3. Detection of causative SNPs

In order to understand how and why human diseases evolve a habit developed to focus strongly on genomic data from sources like GWAS. However, reality showed, that for most of the susceptibility loci detected, the biological meaning remains unsolved. This applies, in particular when the loci is placed in genetic deserts or islands, meaning in non-coding regions of the genome, which holds true for ~80% of detected variants. These regions harbor plenty of functional DNA elements, composed essentially of regulatory elements like promoters, enhancers, insulators or silencers (Urbach & Moore, 2011a). Investigation of these regions just started to elucidate their potential role in complex diseases and evidence for the etiological importance has accumulated (De Gobbi *et al.*, 2006; Wright *et al.*, 2010).

GWAS are directed to the population based genome wide analysis of common variations. In many cases, the results of such studies provide candidate genes characterized by being in the close vicinity of associated SNPs, neglecting for example mechanisms like looping. The detailed investigation of SNP-comprising sequences and possible regulatory mechanisms just started recently. Personalized genomics in comparison, using a trio study design allows the identification of candidate novel and rare polymorphisms that have the potential to cause disease. Trio studies can be applied on small sample sizes and provide internal quality control. For example environmental factors are shared and genetic background or population structure do not need to be explicitly addressed. Moreover, the issue of heritability and non-heritability can be explored, giving an idea about the impact of a detected variant in a wider population.

To study whether a SNP detected in a family based design might have an effect on a wider population could be performed, just focusing on selected SNPs but not on the genome-wide scale. The recent developments of NGS technologies provide further important improvements with regard to genetic association studies. For example the expression of genes can be allele-specifically mapped based on RNA-seq experiments, not being possible by using gene expression microarrays capturing a predefined set of RNA-templates. Additionally, a modification of the ChIP-seq approach, called ChIP-exo, might contribute to narrow down to actual TFBS placed within a peak (Rhee & Pugh, 2012). This might improve the detection of those TFBS to be considered in causal variant detection.

The reported missing overlap with the GWAS catalogue is not surprising, since the family trio is not meant to detect any disease-genetic trait associations. Furthermore, the family genomes used have been constructed based on common variant data. Accordingly, this work did not take the advantage of possible rare or novel variant detection as described above. Thus, the methodology applied, based on family trio data should be rather classified as a preliminary proof-of-principle. To explore the approach of causative SNP detection on disease-genetic trait associations would among others require detailed biological and phenotypic information from the family. Also *de novo* sequencing would be appropriate, not relying on the reference genome being based on a limited number of samples, neither representing the full range of human diversity, nor being complete (Rosenfeld et al., 2012). However, the usage of family trios belonging to the HapMap or 1000G samples as a start to elucidate phenotype-independent causal variants might contribute to develop approaches to solve the causality of intergenic SNPs.

The screening for TFBS in parental peak regions did not reveal any insignificant BS. Thus, quantitative changes in binding have to be evaluated, actually representing a more demanding task than predicting simple on/off situations (qualitative). The procedure to assign a SNP

effect based on a SNE-p-value is similar to the one of MacIntyre and colleagues (Macintyre *et al.*, 2010), with the difference, that the SNP effect is not evaluated based on model scores directly, but on their significance. Is-rSNP was able to assign two out of five EGR1 candidate SNPs as regulatory. Unfortunately, as all publicly available matrices are consulted, the list of possibly affected TFs is long, requesting the user to select, in case no prior knowledge is available. That not all SNPs have been identified as rSNPs might be based on the difference in SNP-significance assignment.

The analysis with the RegulomeDB resulted in the support of 15 out of 20 candidate SNPs. That not all SNPs have been found with a score of at least 2b can have various reasons. One might be that on this work the detection of candidate SNPs is based on a SNE-p-value, which is not considered in the RegulomeDB analysis. Another difference is that in the approach performed in this work not the amount of biological evidence has been measured as selection criteria, but differential binding profiles.

The approach to detect causative SNPs can be considered as the direct extension of the procedure suggested by (Worsley-Hunt *et al.*, 2011), filling the gap between available rSNP tools and biological evidence. The models for TFBS detection have been trained on independent, experimentally generated datasets (UniPROBE) and the screening datasets represented experimentally identified TF binding regions (ChIP-seq data) in a family trio. Since the SNP detection has been executed in a family based set up, providing comparable genomes, “natural” test cases were available, differing by a small number of nucleotides, coming from the same regulatory context. Differences in the parental peak profiles, based on raw read-count data, allowed filtering for genomic regions potentially exposed to regulatory variants.

The approach used to determine causative rSNPs, has not been applied in such a manner yet. The main advantage gained is a kind of internal validation, since confounders like a different regulatory context or genome comparability can be excluded. The filtering procedure and the strict underlying definition of a causative rSNP, namely that direct DNA binding by the TF is impacted due to a SNE in the TF binding site sequence, reduced the list of candidates dramatically, starting from > 1.7 million SNP ending up with 20, revealing that the vast majority of heterozygous child SNPs are not affecting BS regions at all. A relaxation of the rSNP definition might increase the number of candidates, accompanied by more complexity to be considered.

To generate the candidate list, only the case that both parental BS are co-located with the same SNP assuming a high sequence similarity between the parents, has been considered.

However, there are two more cases existing, namely (i) that only one parental BS is co-located with a SNP and (ii) that none of the parental BS is co-located with a SNP. Both cases, do not directly exclude the presence of a regulatory SNP, but maybe its causality.

With regard to (i) the case of an alternative BS in that parent, where the BS does not overlap with the SNP, is on hand. This alternative BS can occur either due to general more frequent sequence differences or due to an indirect effect of the SNP (in case of high sequence similarity). This indirect effect describes the event that the best BS might indeed be disrupted by the SNP, but another alternative BS is present, able to serve as a kind of “escape” strategy for the TF. However, those regions would implementation wise only be detected in differential peak profiles. Thus, even if a good alternative site might be occupied by the TF due to a SNP-disrupted original one, the signal difference is existing. Therefore, such a signal difference might be determined by an upstream mechanism, not being coded in the (TFBS) sequence itself, but for example by histone modifications. Accordingly, the SNP might have an effect, but its causality with regard to differential binding is questionable and not directly derivable from the used datasets. For (ii) one could follow the same argumentation as for (i), but assuming a higher probability of an upstream effect.

To gain further knowledge, if the SNP might play an important role in such cases, it would be interesting to investigate the hypothetical case of the alternative SNP variant (in case of heterozygosis). For example, a shift of the BS within a peak sequence equipped with the alternative SNP allele, so that both parents show a SNP-BS-co-location, could indicate a causative character of the SNP. However, such a result needs to be considered as a theoretical hypothesis, because it cannot be verified based on measured peak profiles associated with the actual (and not alternative) SNP variant.

For two selected examples, further biological evidence has been integrated, supporting the TFBS as such, required as a pre-condition to observe a causative SNP in context of the rSNP definition. In summary, the significance of both affected BS is supported by many factors describing TFBS in a biological manner (POLII, histone markers, differential parental peak profiles) as well as from the modeling point of view being equipped with a strongly lifted significance in comparison to all other possible 9-mers below the peak. The SNPs gain evidence by representing the central sequence difference between parental peak sequences in differentially appearing peak profiles, by being co-located with a biologically supported BS and by being assigned with highly significant p-value capturing the effect of the nucleotide exchange between the parents. However, the causality assumption holds currently only true for local different peak profiles. A phenotypic effect evaluated within a larger genomic region around the SNP based on epigenetic markers and POLII seems less likely. Furthermore, in

the examples considered, numerous other peaks of the respective TF in direct vicinity appear, which might compensate a disrupted TF binding.

Further extensions of the applied approach might be considered, as for example to examine the distance of BS and SNP locations to target gene TSSs. This challenging point has been so far neglected. As mentioned in the results section, multiple POLII peaks in the BS surrounding indicate DNA looping, whereby a BS being distal from its target gene TSS might be brought into proximity by the three-dimensional looping of chromatin (Worsley-Hunt *et al.*, 2011). As a result, regulatory sequences can act specifically on distant, even skipping intervening genes (Dean, 2011). Recently developed methods that detect such DNA proximity (Gavrilov *et al.*, 2009; Lieberman-Aiden *et al.*, 2009) may provide data suitable for integration into the approach applied. Additionally, the specification of the edges of accessible DNA regions, termed as insulators, could serve to narrow down the regions for possible target genes. A known insulator protein is the CCCTC-binding factor (CTCF) and the detection of its location (possible by means of ChIP-seq experiments) could be used to determine which promoter regions are accessible to a TF bound between insulator sequences (Cuddapah *et al.*, 2009; Gaszner & Felsenfeld, 2006). Actually, ENCODE provides CTCF ChIP-seq data for GM12878 CEU (trio child) and their integration will be applied in direct future.

Due to the very recent publication of ENCODE data¹⁵, new, additional datasets became available to be included in an improved analysis. For example for SRF a second ChIP-seq analysis has been done, so that a model training on ChIP-seq data might be considered. Furthermore, the new online resource “Factorbook”¹⁶ gives an easy and intuitively to access overview over all ENCODE datasets available stratified by cell type (Wang *et al.*, 2012). This presentation of the ENCODE experiments provides a well organized source offering additional information, like TF binding motifs based on ChIP-seq data and quality information, which will might help to improve the models used in this work (for example by adding seed sequences in the motif detection step). Additionally the search for appropriate datasets is strongly facilitated.

In order to gain a single descriptor for the probability whether a detected BS has an *in vivo* impact and to integrate epigenetic markers in an elegant way approaches like CENTIPEDE (see chapter 1.9 and Annex 8.3) could be very useful. Its extension by tree-based models and a strategy to overcome its independence assumption should be considered.

A further biological aspect also not included so far, are cooperative interactions between TFs. In such cases, several different TFs bind to clusters of DNA binding sites. This leads to a rather principle question, whether a dramatic phenotypic effect, like gene silencing, can be

¹⁵ <http://www.nature.com/encode/>

¹⁶ <http://www.factorbook.org/>

expected by the distortion of a single TFBS corresponding to a single TF. Many TFs act in combination possibly in a compensatory manner. The exploration of such regulatory strategies might be investigated by considering CRMs. The detection of SNPs, distorting the BS affinity of more than one TF, in context of a CRM, would strongly support a causal character (Hannah, Joshi, Wilson, Kinston, & Göttgens, 2011). In such a case a SNP might have the potential to knock out a complete regulatory control mechanism rather expected to show a tremendous phenotypic effect. Moreover, the investigation of CRMs might also help to overcome the futility theorem, meaning to improve the detection of BS sites with an impact *in vivo* and not only *in vitro* (Van Loo & Marynen, 2009). This task is in so far challenging, as, first experimental data for several TFs need to be on hand and, second those data would need to correspond to TFs in fact interacting with each other, forming a biological reasonable network. Thus, the selection of TFs to be analyzed needs to be based on prior knowledge, to gain for example by a pathway analysis or an overlap of all available ChIP-seq data for the family trio considered. Alternatively, one could screen the genome with all available TF binding models (mainly provided as PWMs), integrate the results with gene expression data, and follow the hypothesis, that genes being co-regulated or co-expressed in a specific process share regulatory signals (Van Loo & Marynen, 2009). A second principle to approach this futility theorem would be the integration of sequence conservation between related species. These comparative genomics are motivated by the observation that functional sequences accumulate fewer mutations during evolution than non-functional sequences (Van Loo & Marynen, 2009).

Beside the detection of affected TFBS and a possible impact on the expression of a single gene, one might ask, whether the alteration of a single gene expression profile might be causal for a common or complex disease? Here, disease networks or genetic disease maps might be required, integrating heritable genetic variants and affected genes in a cooperative manner to elucidate the genetic architecture of human diseases (Urbach & Moore, 2011b).

Finally, it has to be emphasized, that the probabilistic models applied have been trained on TF binding only and not to detect the effect of a SNP on TF binding. In principle, one can consider this as a task on its own. It requires the development of a model specifically trained on SNE-effects, including the specific sequences building a SNE-pair (including plus and minus strand), the specific nucleotide position, nucleotides being exchanged as well as a value representing a measure of the SNE. Furthermore, one should also consider the impact of a SNE within the motif on all nucleotide pairings within the motif, maybe even considering nucleotide sub-sequences going beyond two nucleotides. As an approximation one could use the dataset used to generate the SNE-p-value in this work, keeping in mind, that the scores used to calculate this p-value might be biased for the tree-based models, as mentioned before.

The question, if tree-based or a PWM model should be applied to screen for rSNPs, could be answered in two ways:

- If possible false positive predictions are considered as acceptable, and an investigator does not make a point to have additional support by other TFBS models, the PWM model could be sufficient.
- If false positive predictions are a critical issue and the support by other TFBS models is considered as an asset, the application of the ET model might be the better choice; since – at least in this work – the SNP candidate list has been exclusively determined by the ET model, being supported by both the CLTree and PWM model (intrinsically included in the ET model).

6. Conclusions and Future Work

In this work the potential of different probabilistic models trained on UniPROBE data with regard to their predictive power in an independent biological dataset has been investigated. It has been shown, that tree-based models provide, in the majority of TFs considered, a better predictive power in comparison to a naïve Bayesian model, represented by a PWM.

The tree-based models rarely perform less good than a PWM, and if so, only slightly. Hence, the general application of tree-based models as the default can be suggested and thereby the replacement of the well established PWM model. The training of tree-models is computationally not more expensive than a PWM and automatically captures positional interdependencies, if they are present.

Furthermore, this is the first work at all, using the ET model for TFBS detection showing a possible improvement over the CLTree model. Thus, in order to automatically include even more complex sequence features a switch to the ET model should be considered.

However, numerous improvements have to be resolved. The determination of the parameter β requires further research, in order to take full advantage from the ET model. First pilot studies to optimize β separately on the training set by maximizing the sum of the log-likelihood by an iterative approach based on MI data revealed promising results. Accordingly, this optimization in a systematic manner will be approached in direct future.

The main bias, introduced to the system, is the motif discovery step. Thus, the basic need to develop an algorithm able to detect motifs considering positional interdependencies as well as the related specific nucleotides is evident. However, this task is very demanding and probably requires a separate research project.

With regard to rSNP or causative variant detection the analysis in a family based set up provides essential improvements. The basic advantage is that these datasets fulfill the unsolved requirement of an internal and biological validation, if combined with additional data like gene expression or chromatin markers. The differences being observable between the parental data provide direct biological support whether a detected rSNP might be causal for differential TF binding. The resulting candidate list of 20 SNPs gained by the approach, earns strong evidence for causality on both levels, biologically and *in silico*.

The parallel application of different probabilistic models to detect rSNPs did not reveal major advantages for the tree-based models. However, the models used, have not been trained on directly comparable data. As discussed before, the training of tree-models on ChIP-seq data directly is not trivial, as the training method successfully applied on UniPROBE data cannot be

explored as such on ChIP-seq data. The main issue of sequence variability has to be approached in a smart manner, considering for example data fusion and the application of a Bayesian approach in the training set generation or screening step. Furthermore, the training in particular for rSNPs requires more complexity, to be done directly on SNE-data.

However, prioritizing a true positive causative SNP, being detected consistently by different TFBS models, and the provision for possible complex positional interdependencies, the application of the ET model as default should be taken into account, for both TFBS screening and rSNP detection.

Data integration to evaluate biological plausibility

With regard to biological plausibility additional knowledge should be integrated, improving the understanding of already published and newly detected disease-SNP or more general disease-regulatory element associations. The underlying biological experiments can be highly dependent on, for example, cell types, the related anatomical system or experimental setup. Additionally, information concerning already available data, like gene annotations in the SNP's vicinity, a possible entry in the genetic association database (GAD) or close miRNA coding regions should be considered.

The following Figure 71 should illustrate possible systematic dependencies when evaluating a detected rSNP out of the scope of a proof-of-principle experiment. For example, if an rSNP is located closely to a gene, one might check the annotation of this gene, its expression and protein profile if available. Moreover, when the project was directed to a certain phenotype, it would be essential to know, whether the possibly affected gene is expressed in a phenotype related cell type or organ and if this SNP has even been associated with the phenotype in a genetic association study already. If wet-lab validation is planned, one would need to check whether for the respective cell type expressing the gene, possibly associated with the rSNP, a comparable cell line or cellular model is existing. In short, the person in charge of an rSNP validation needs to consider numerous highly relevant biological connections to evaluate its biological plausibility and to plan possible wet-lab validation experiments.

In order to access the different data generated during rSNP detection and the outcomes in a comfortable manner a relational database would be very important. By this also researchers not involved in the data analysis, are able to access and request outputs. Furthermore, additional data to evaluate the biological relevance of a detected SNP or regulatory element could be reviewed directly.

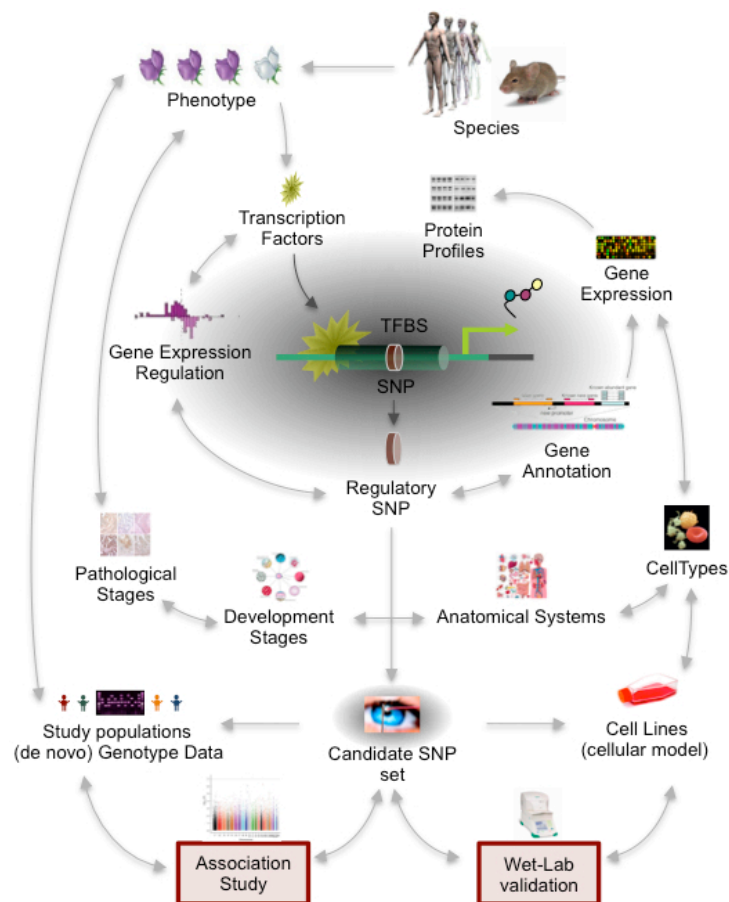
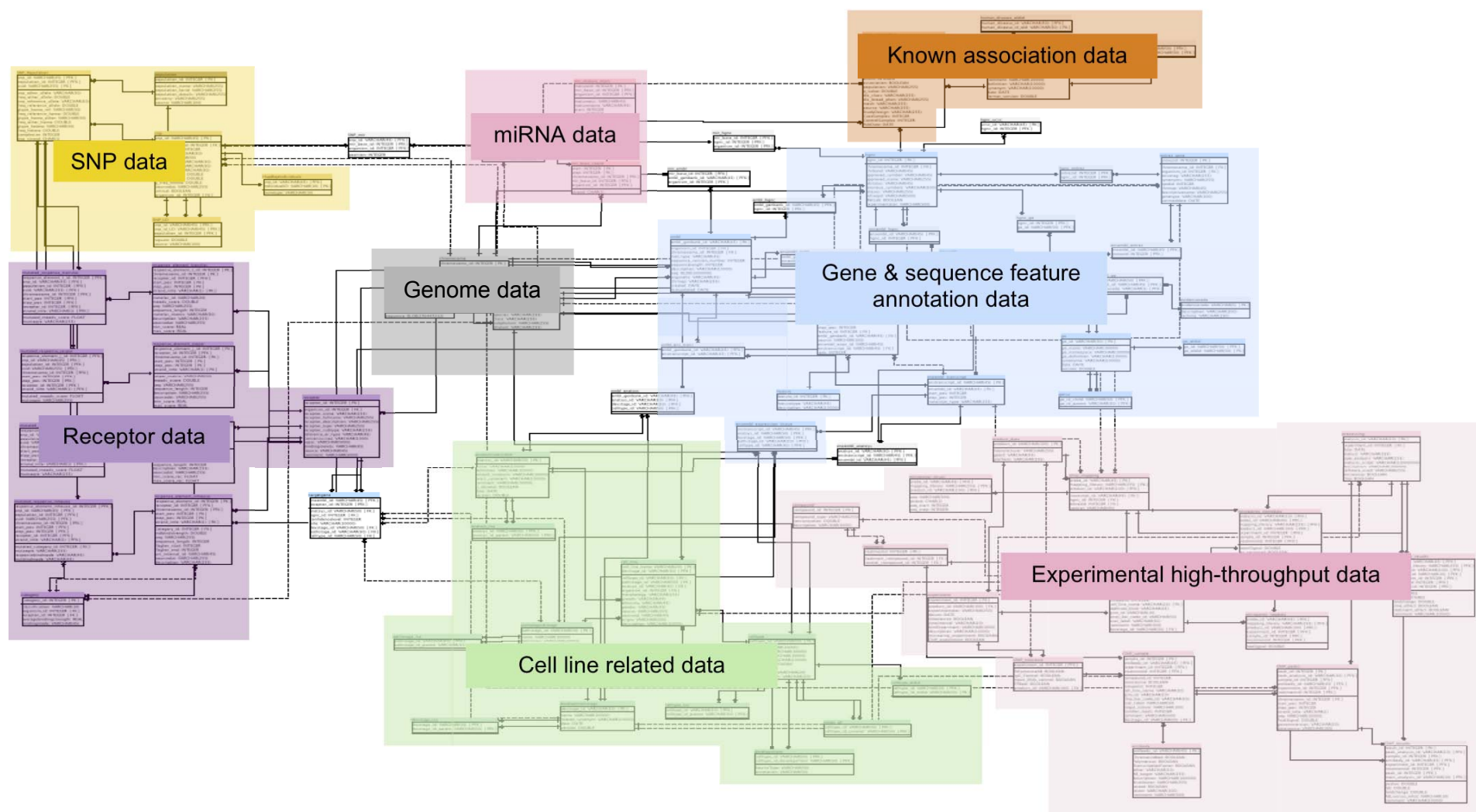


Figure 71: Schematic illustration of biological dependencies to be checked when investigating the biological plausibility of an rSNP.

The main intention here is to take advantage of the huge amount of data available and to transfer data to knowledge meaning to a reasonable biological hypothesis *a priori* or *a posteriori*. Thereby wet-lab or high throughput experiments are rather needed to confirm a hypothesis than generating the same. Main challenges are to curate the system and keep it up-to-date, especially in context of genome versions, coordinates, IDs and published high-throughput datasets.

A pilot database scheme as well as a prototype database has been set up as shown in Figure 72. The pilot scheme consists of 80 tables and 150 relations, reflecting the complex system of connections to be considered when the biological plausibility of an rSNP or TFBS should be investigated in detail. The prototype, still lacking experimental and tree-based model data, has a volume of ~470GB (further details see Annex). As future work, like a user interface or systematic and automated update functions have to be implemented.

Figure 72: Pilot relational database scheme developed to mirror biological connections to be considered to investigate biological plausibility.



7. References

- Abdi, H. (2007). The Bonferonni and Šidák Corrections for Multiple Comparisons. *Encyclopedia of Measurements and Statistics* (pp. 1–9).
- Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Bonnen, P. E., et al. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311), 52–8.
- Andersen, M. C., Engström, P. G., Lithwick, S., Arenillas, D., Eriksson, P., Lenhard, B., Wasserman, W. W., et al. (2008). In silico detection of sequence variations modifying transcriptional regulation. (G. Stormo, Ed.) *PLoS computational biology*, 4(1), e5.
- Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. *Science (New York, N.Y.)*, 324(5935), 1720–3.
- Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*, 37(Web Server issue), W202–8.
- Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers.
- Balmer, J. E., & Blomhoff, R. (2006). Anecdotes, data and regulatory modules. *Biology letters*, 2(3), 431–4.
- Barash, Y., Elidan, G., Friedman, N., & Kaplan, T. (2003). Modeling dependencies in protein-DNA binding sites. *Proceedings of the seventh annual international conference on Computational molecular biology - RECOMB '03*, 28–37.
- Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., et al. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4), 823–37.
- Bartel, D. P. (2004). MicroRNAs : Genomics , Biogenesis , Mechanism , and Function
Genomics : The miRNA Genes. *Cell*, 116, 281–297.
- Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2), 215–33.
- Baxeavanis, A. D. (2011). The importance of biological databases in biological discovery. *Current protocols in bioinformatics*, Chapter 1, Unit 1.1.
- Ben-Gal, I., Shani, a, Gohr, a, Grau, J., Arviv, S., Shmilovici, a, Posch, S., et al. (2005). Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics (Oxford, England)*, 21(11), 2657–66.
- Berg, O. G., & von Hippel, P. H. (1987). Selection of DNA binding sites by regulatory proteins. *Journal of Molecular Biology*, 193(4), 723–743.

- Berger, M. F., Philippakis, A. a, Qureshi, A. M., He, F. S., Estep, P. W., & Bulyk, M. L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology*, 24(11), 1429–35.
- Berk, A., & Sharp, P. (1977). Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. *Cell*, 12(3), 721–732.
- Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., et al. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of molecular biology*, 112(3), 535–42.
- Bilofsky, H. S., & Burks, C. (1988). The GenBank genetic sequence data bank, 16(5), 1861–1863.
- Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., et al. (2004). An overview of Ensembl. *Genome research*, 14(5), 925–8.
- Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146), 799–816.
- Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., & Sonnhammer, E. (1992). What's in a genome? *Nature*, 358(6384), 287.
- Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome research*, 22(9), 1790–7.
- Britten, R. J., & Davidson, E. H. (1969). Gene Regulation for Higher Cells: A Theory. *Science*, 165(3891), 349–357.
- Bryne, J. C., Valen, E., Tang, M.-H. E., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., et al. (2008). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic acids research*, 36(Database issue), D102–6.
- Cairns, B. R. (2009). The logic of chromatin architecture and remodelling at promoters. *Nature*, 461(7261), 193–8.
- Chaisson, M. J., Brinza, D., & Pevzner, P. A. (2009). De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome research*, 19(2), 336–46.
- Chang, Y.-W., Robert Liu, F.-G., Yu, N., Sung, H.-M., Yang, P., Wang, D., Huang, C.-J., et al. (2008). Roles of cis- and trans-changes in the regulatory evolution of genes in the gluconeogenic pathway in yeast. *Molecular biology and evolution*, 25(9), 1863–75.
- Chen, K., & Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. *Nature reviews. Genetics*, 8(2), 93–103.
- Chodavarapu, R. K., Feng, S., Bernatavichute, Y. V., Chen, P.-Y., Stroud, H., Yu, Y., Hetzel, J. A., et al. (2010). Relationship between nucleosome positioning and DNA methylation. *Nature*, 466(7304), 388–92.

- Chow, C. K., & Liu, C. N. (1968). Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Transactions on Information Theory*, *IT-14*(3), 462–467.
- Chow, L., Roberts, J., Lewis, J., & Broker, T. (1977). A map of cytoplasmic RNA transcripts from lytic adenovirus type 2, determined by electron microscopy of RNA:DNA hybrids. *Cell*, *11*(4), 819–836.
- Clapier, C. R., & Cairns, B. R. (2009). The biology of chromatin remodeling complexes. *Annual review of biochemistry*, *78*, 273–304.
- Clarke, C. L., Sandle, J., Jones, A. A., Sofronis, A., Patani, N. R., & Lakhani, S. R. (2006). Mapping loss of heterozygosity in normal human breast cells from BRCA1/2 carriers. *British journal of cancer*, *95*(4), 515–9.
- Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, *38*(6), 1767–71.
- Collins, F. S., Brooks, L. D., & Chakravarti, A. (1998). A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation, 1229–1231.
- Collins, F. S., Green, E. D., Guttmacher, A. E., & Guyer, M. S. (2003). A vision for the future of genomics research. *Nature*, *422*(6934), 835–47.
- Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(50), 21931–6.
- Crooks, G. E., Hon, G., Chandonia, J.-M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome research*, *14*(6), 1188–90.
- Cuddapah, S., Jothi, R., Schones, D. E., Roh, T.-Y., Cui, K., & Zhao, K. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome research*, *19*(1), 24–32.
- Dayhoff, M. (1978). Atlas of Protein Sequence and Structure. *National Biomedical Research Foundation*, *4*(3).
- De Gobbi, M., Viprakasit, V., Hughes, J. R., Fisher, C., Buckle, V. J., Ayyub, H., Gibbons, R. J., et al. (2006). A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science (New York, N.Y.)*, *312*(5777), 1215–7.
- Dean, A. (2011). In the loop: long range chromatin interactions and gene regulation. *Briefings in functional genomics*, *10*(1), 3–10.
- Dietrich, W. F., Miller, J., Steen, R., Merchant, M. A., Damron-Boles, D., Husain, Z., Dredge, R., et al. (1996). A comprehensive genetic map of the mouse genome. *Nature*, *380*(6570), 149–52.
- Doi, A., Park, I.-H., Wen, B., Murakami, P., Aryee, M. J., Irizarry, R., Herb, B., et al. (2009). Differential methylation of tissue- and cancer-specific CpG island shores distinguishes

- human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nature genetics*, 41(12), 1350–3.
- Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57–74.
- Ebert, M. S., & Sharp, P. a. (2012). Roles for microRNAs in conferring robustness to biological processes. *Cell*, 149(3), 515–24.
- Ernst, J., & Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology*, 28(8), 817–25.
- Esteller, M. (2007). Epigenetic gene silencing in cancer: the DNA hypermethylome. *Human molecular genetics*, 16 Spec No, R50–9.
- Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome research*, 8(3), 175–85.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.
- Ferragina, P., & Manzini, G. (2000). Opportunistic data structures with applications. *Proceedings 41st Annual Symposium on Foundations of Computer Science*, 390–398.
- Friedman, R. C., Farh, K. K., Burge, C. B., & Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs, 92–105.
- Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., et al. (2011). The UCSC Genome Browser database: update 2011. *Nucleic acids research*, 39(Database issue), D876–82.
- Fuks, F., Hurd, P. J., Wolf, D., Nan, X., Bird, A. P., & Kouzarides, T. (2003). The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. *The Journal of biological chemistry*, 278(6), 4035–40.
- Galperin, M. Y., & Fernández-Suárez, X. M. (2012). The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic acids research*, 40(Database issue), D1–8.
- Gaszner, M., & Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. *Nature reviews. Genetics*, 7(9), 703–13.
- Gavrilov, A., Eivazova, E., Priozhkova, I., Lipinski, M., Razin, S., & Vassetzky, Y. (2009). Chromosome conformation capture (from 3C to 5C) and its ChIP-based modification. *Methods in molecular biology (Clifton, N.J.)*, 567, 171–88.
- Gene, T., & Consortium, O. (2000). Gene Ontology : tool for the, 25(may), 25–29.
- Ghildiyal, M., & Zamore, P. D. (2009). Small silencing RNAs: an expanding universe. *Nature reviews. Genetics*, 10(2), 94–108.

- Gilad, Y., Rifkin, S. A., & Pritchard, J. K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in genetics : TIG*, 24(8), 408–15.
- Gingeras, T. R., & Roberts, R. J. (1980). Steps toward computer analysis of nucleotide sequences. *Science (New York, N.Y.)*, 209(4463), 1322–8.
- Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., & Lieb, J. D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome research*, 17(6), 877–85.
- Griffiths-Jones, S. (2004). The microRNA Registry. *Nucleic acids research*, 32(Database issue), D109–11.
- GuhaThakurta, D. (2006). Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic acids research*, 34(12), 3585–98.
- Guigó, R., Knudsen, S., Drake, N., & Smith, T. (1992). Prediction of gene structure. *Journal of molecular biology*, 226(1), 141–57.
- Gusella, J., Wexler, N., Conneally, P., Naylor, S., Anderson, M., Tanzi, R., Watkins, P., et al. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*, 306(5940), 234–238.
- Hamm, G. H., & Cameron, G. N. (1986). The EMBL data library Nucleic Acids Research, 14(1), 5–9.
- Hannah, R., Joshi, A., Wilson, N. K., Kinston, S., & Göttgens, B. (2011). A compendium of genome-wide hematopoietic transcription factor maps supports the identification of gene regulatory control mechanisms. *Experimental hematology*, 39(5), 531–41.
- Hindorff, LA, MacArthur, J., Wise, A., Junkins, H., Hall, P., Klemm, A., & Manolio, T. (n.d.). A Catalog of Published Genome-Wide Association Studies.
- Hindorff, Lucia a, Sethupathy, P., Junkins, H. a, Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. a. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23), 9362–7.
- Ho, L., & Crabtree, G. R. (2010). Chromatin remodelling during development. *Nature*, 463(7280), 474–84.
- Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3), 318–356.
- Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)*, 316(5830), 1497–502.
- Jones, M. H., & Nakamura, Y. (1992). Detection of loss of heterozygosity at the humanTP53 locus using a dinucleotide repeat polymorphism. *Genes, Chromosomes and Cancer*, 5(1), 89–90.

- Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40(Database issue), D109–14.
- Karlič, R., Chung, H.-R., Lasserre, J., Vlahovicek, K., & Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 107(7), 2926–31.
- Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, a. M., & Haussler, a. D. (2002). The Human Genome Browser at UCSC. *Genome Research*, 12(6), 996–1006.
- Kharchenko, P. V., Tolstorukov, M. Y., & Park, P. J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature biotechnology*, 26(12), 1351–9.
- Kim, J. B., Porreca, G. J., Song, L., Greenway, S. C., Gorham, J. M., Church, G. M., Seidman, C. E., et al. (2007). Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science (New York, N.Y.)*, 316(5830), 1481–4.
- Kitano, H. (2004). Biological robustness. *Nature reviews. Genetics*, 5(11), 826–37.
- Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 2(12), 1137–1143.
- Kornberg, A. (1974). *DNA Synthesis*. W. H. Freeman and company, San Francisco.
- Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, 128(4), 693–705.
- Krol, J., Loedige, I., & Filipowicz, W. (2010). The widespread regulation of microRNA biogenesis, function and decay. *Nature reviews. Genetics*, 11(9), 597–610.
- Kuroda, A., Rauch, T. a, Todorov, I., Ku, H. T., Al-Abdullah, I. H., Kandeel, F., Mullen, Y., et al. (2009). Insulin gene expression is regulated by DNA methylation. *PloS one*, 4(9), e6953.
- Lagos-Quintana, M., Rauhut, R., Lendeckel, W., & Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science (New York, N.Y.)*, 294(5543), 853–8.
- Lal, A., Pan, Y., Navarro, F., Dykxhoorn, D. M., Moreau, L., Meire, E., Bentwich, Z., et al. (2009). miR-24-mediated downregulation of H2AX suppresses DNA repair in terminally differentiated blood cells. *Nature structural & molecular biology*, 16(5), 492–8.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860–921.
- Landsman, D., Gentleman, R., Kelso, J., & Francis Ouellette, B. F. (2009). DATABASE: A new forum for biological databases and curation. *Database : the journal of biological databases and curation*, 2009, bap002.
- Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3), R25.

- Latham, J. A., & Dent, S. Y. R. (2007). Cross-regulation of histone modifications. *Nature structural & molecular biology*, 14(11), 1017–24.
- Lau, N. C., Lim, L. P., Weinstein, E. G., & Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science (New York, N.Y.)*, 294(5543), 858–62.
- Lee, R. C., & Ambros, V. (2001). An extensive class of small RNAs in *Caenorhabditis elegans*. *Science (New York, N.Y.)*, 294(5543), 862–4.
- Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, 75(5), 843–54.
- Li, B., Carey, M., & Workman, J. L. (2007). The role of chromatin during transcription. *Cell*, 128(4), 707–19.
- Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, 326(5950), 289–93.
- Lin, Y., Zhu, S., Lee, D. D., & Taskar, B. (2009). Learning Sparse Markov Network Structure via Ensemble-of-Trees Models, 5, 360–367.
- Lipman, D., & Pearson, W. (1985). Rapid and sensitive protein similarity searches. *Science*, 227(4693), 1435–1441.
- Lopez-Serra, L., & Esteller, M. (2008). Proteins that bind methylated DNA and human cancer: reading the wrong words. *British journal of cancer*, 98(12), 1881–5.
- Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics? An introduction and overview. *Yearbook of Medical Informatics*, 83–100.
- Macintyre, G., Bailey, J., Haviv, I., & Kowalczyk, A. (2010). is-rSNP: a novel technique for in silico regulatory SNP detection. *Bioinformatics (Oxford, England)*, 26(18), i524–30.
- Martin-Löf, P. (1966). The definition of random sequences. *Information and Control*, 9(6), 602–619.
- Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2), 560–4.
- McClintock, B. (1944). The Relation of Homozygous Deficiencies to Mutations and Allelic Series in Maize. *Geneti*, 2(September), 478–502.
- Meila, M., & Jaakkola, T. (2006). Tractable Bayesian Learning of Tree Belief Networks. *Statistics and Computing*, 16(1), 77–92.
- Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1), 31–46.
- Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153), 553–60.

- Mitroff, I. I., & Featheringham, T. R. (1974). On systemic problem solving and the error of the third kind. *Behavioral Science*, 19(6), 383–393.
- Morgan, T. H., Sturtevant, A. H., Muller, H. J., & Bridges, C. B. (1915). *The Mechanism of Mendelian Heredity*. Henry Holt and Company.
- NIH. (2003). 2003 Release: ENCODE Meeting. Retrieved August 3, 2012, from <http://www.genome.gov/10506706>
- Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)*, 320(5881), 1344–9.
- Needleman, S., & Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3), 443–453.
- Neumann, J. von, & Morgenstern, O. (1953). *Theory of Games and Economic Behavior* (p. 642). John Wiley & Sons Inc.
- Newburger, D. E., & Bulyk, M. L. (2009). UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic acids research*, 37(Database issue), D77–82.
- Nirenberg, M. (1963). The genetic code. II. *Scientific American*, 208, 80–94.
- Nix, D. a, Courdy, S. J., & Boucher, K. M. (2008). Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC bioinformatics*, 9, 523.
- Osborne, C. S., Chakalova, L., Brown, K. E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., et al. (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature genetics*, 36(10), 1065–71.
- Ouzounis, C. a., & Valencia, a. (2003). Early bioinformatics: the birth of a discipline--a personal view. *Bioinformatics*, 19(17), 2176–2190.
- Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, 10(10), 669–80.
- Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3), 1065–1076.
- Phillips, T. (2008). Regulation of transcription and gene expression in eukaryotes. *Nature Education*, 1(1).
- Phillips, T., & Shaw, K. (2008). Chromatin Remodeling in Eukaryotes. *Nature Education* 1(1).
- Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., & Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome research*, 21(3), 447–55.

- Ponomarenko, J. V. (2003). rSNP_Guide, a database system for analysis of transcription factor binding to DNA with variations: application to genome annotation. *Nucleic Acids Research*, 31(1), 118–121.
- Pop, M., & Salzberg, S. L. (2008). Bioinformatics challenges of new sequencing technology. *Trends in genetics : TIG*, 24(3), 142–9.
- Portela, A., & Esteller, M. (2010). Epigenetic modifications and human disease. *Nature biotechnology*, 28(10), 1057–68.
- Pratt, A. J., & MacRae, I. J. (2009). The RNA-induced silencing complex: a versatile gene-silencing machine. *The Journal of biological chemistry*, 284(27), 17897–901.
- Pudimat, R., Schukat-Talamazzini, E.-G., & Backofen, R. (2005). A multiple-feature framework for modelling and predicting transcription factor binding sites. *Bioinformatics (Oxford, England)*, 21(14), 3082–8.
- Rahmann, S. (2011). Today' s and Tomorrow' s Sequencing Technologies and their Bioinformatic Challenges.
- Rando, O. J., & Chang, H. Y. (2009). Genome-wide views of chromatin structure. *Annual review of biochemistry*, 78, 245–71.
- Reid, J. E., & Wernisch, L. (2011). STEME: efficient EM to find motifs in large data sets. *Nucleic acids research*, 39(18), e126.
- Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., et al. (2000). Genome-wide location and function of DNA binding proteins. *Science (New York, N.Y.)*, 290(5500), 2306–9.
- Rhee, H. S., & Pugh, B. F. (2012). Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*, 483(7389), 295–301.
- Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing, 4(8), 651–657.
- Rosenbloom, K. R., Dreszer, T. R., Pheasant, M., Barber, G. P., Meyer, L. R., Pohl, A., Raney, B. J., et al. (2010). ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic acids research*, 38(Database issue), D620–5.
- Rosenfeld, J. a., Mason, C. E., & Smith, T. M. (2012). Limitations of the Human Reference Genome for Personalized Genomics. (J.-S. Seo, Ed.) *PLoS ONE*, 7(7), e40294.
- Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., et al. (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology*, 7, 522.
- Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., et al. (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature biotechnology*, 27(1), 66–75.

- Saiki, R., Scharf, S., Faloona, F., Mullis, K., Horn, G., Erlich, H., & Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230(4732), 1350–1354.
- Sand, O., Turatsinze, J. V., & Helden, J. V. (2008). Evaluating the prediction of cis-acting regulatory elements in genome sequences. In A. Valencia & D. Frishman (Eds.), *Modern Genome Annotation* (pp. 55–89). Springer-Verlag/Wien.
- Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., et al. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596), 687–95.
- Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3), 441–8.
- Santos-Rosa, H., Kirmizis, A., Nelson, C., Bartke, T., Saksouk, N., Cote, J., & Kouzarides, T. (2009). Histone H3 tail clipping regulates gene expression. *Nature structural & molecular biology*, 16(1), 17–22.
- Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L., & Nolan, G. P. (2010). Computational solutions to large-scale data management and analysis. *Nature reviews. Genetics*, 11(9), 647–57.
- Schneider, T. D., Stormo, G. D., Gold, L., & Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology*, 188(3), 415–431.
- Schones, D. E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., et al. (2008). Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5), 887–98.
- Schwartz, S., & Carpenter, K. M. (1999). The right answer for the wrong question: consequences of type III error for public health research. *American journal of public health*, 89(8), 1175–80.
- Segal, E., & Widom, J. (2009). From DNA sequence to transcriptional behaviour: a quantitative approach. *Nature reviews. Genetics*, 10(7), 443–56.
- Sellers, P. H. (1980). The theory and computation of evolutionary distances: Pattern recognition. *Journal of Algorithms*, 1(4), 359–373.
- Shannon, C., & Weaver, W. (1962). *The mathematical theory of communication* (9. pr.). Urbana Ill.: Univ. of Illinois Pr.
- Sharon, E., Lubliner, S., & Segal, E. (2008). A feature-based approach to modeling protein-DNA interactions. (G. Stormo, Ed.) *PLoS computational biology*, 4(8), e1000154.
- Simon, J. M., Giresi, P. G., Davis, I. J., & Lieb, J. D. (2012). Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nature protocols*, 7(2), 256–67.
- Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCr: visualizing classifier performance in R. *Bioinformatics (Oxford, England)*, 21(20), 3940–1.

- Smith, TF, & Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1), 195–197.
- Smith, Todd. (2012). Bio Databases 2012. Retrieved August 4, 2012, from <http://finchtalk.geospiza.com/2012/01/bio-databases-2012.html>
- Solomon, M. J., Larsen, P. L., & Varshavsky, A. (1988). Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell*, 53(6), 937–47.
- Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B.-K., Sheffield, N. C., et al. (2011). Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome research*, 21(10), 1757–67.
- Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics (Oxford, England)*, 16(1), 16–23.
- Strahl, B. D., & Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, 403(January), 41–45.
- Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature*, 458(7239), 719–24.
- Tanaka, T. (2005). [International HapMap project]. *Nihon rinsho. Japanese journal of clinical medicine*, 63 Suppl 1, 29–34.
- Tata, J. R. (2002). Signalling through nuclear receptors. *Nature reviews. Molecular cell biology*, 3(9), 702–10.
- The 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), 1061–73.
- Thomson, J. P., Skene, P. J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A. R. W., et al. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature*, 464(7291), 1082–6.
- Tian, Y., Jia, Z., Wang, J., Huang, Z., Tang, J., Zheng, Y., Tang, Y., et al. (2011). Global mapping of H3K4me1 and H3K4me3 reveals the chromatin state-based cell type-specific gene regulation in human Treg cells. *PloS one*, 6(11), e27770.
- Trapnell, C., & Salzberg, S. L. (2009). How to map billions of short reads onto genomes. *Nature biotechnology*, 27(5), 455–7.
- Ukkonen, E. (1985). Algorithms for approximate string matching. *Information and Control*, 64(1-3), 100–118.
- Urbach, D., & Moore, J. H. (2011a). Mining beyond the exome. *BioData mining*, 4(1), 14.
- Urbach, D., & Moore, J. H. (2011b). Mining the diseasome. *BioData mining*, 4(1), 25.
- Valouev, A., Johnson, D., Sundquist, A., & Medina, C. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature methods*, (August), 1–6. doi:10.1038/NMETH.1246

- Van Helden, J. (2005). The Analysis of Regulatory Sequences.
- Van Loo, P., & Marynen, P. (2009). Computational methods for the detection of cis-regulatory modules. *Briefings in bioinformatics*, 10(5), 509–24.
- Van Speybroeck, L. (2002). From epigenesis to epigenetics: the case of C. H. Waddington. *Annals of the New York Academy of Sciences*, 981, 61–81.
- Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. a, & Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics*, 10(4), 252–63.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., et al. (2001). The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507), 1304–51.
- Venter, J. C., Adams, M. D., Sutton, G. G., Kerlavage, A. R., Smith, H. O., & Hunkapiller, M. (1998). Shotgun sequencing of the human genome. *Science (New York, N.Y.)*, 280(5369), 1540–2.
- Walls, P. H., & Sternberg, M. J. (1992). New algorithm to model protein-protein recognition based on surface complementarity. Applications to antibody-antigen docking. *Journal of molecular biology*, 228(1), 277–97.
- Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., Pierce, B. G., et al. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research*, 22(9), 1798–1812.
- Wasserman, W. W., & Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature reviews. Genetics*, 5(4), 276–87.
- Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915), 520–62.
- Watson, J., & Crick, F. (1953). Molecular structure of nucleic acids. *Nature*, (4356), 737.
- Watson, J. D., & Jordan, E. (1989). The Human Genome Program at the National Institutes of Health. *Genomics*, 5, 654–656.
- Whiteside, S. T., & Goodbourn, S. (1993). Signal transduction and nuclear targeting: regulation of transcription factor activity by subcellular localisation. *Journal of cell science*, 104 (Pt 4, 949–55.
- Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C. J., et al. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453(7199), 1239–43.
- Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., et al. (2000). TRANSFAC: an integrated system for gene expression regulation. *Nucleic acids research*, 28(1), 316–9.

- Wold, B., & Myers, R. M. (2008). Sequence census methods for functional genomics. *Nature methods*, 5(1), 19–21.
- Workman, C. T., Yin, Y., Corcoran, D. L., Ideker, T., Stormo, G. D., & Benos, P. V. (2005). enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic acids research*, 33(Web Server issue), W389–92.
- Worsley-Hunt, R., Bernard, V., & Wasserman, W. W. (2011). Identification of cis-regulatory sequence variations in individual genome sequences. *Genome medicine*, 3(10), 65.
- Wright, J. B., Brown, S. J., & Cole, M. D. (2010). Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. *Molecular and cellular biology*, 30(6), 1411–20.
- Wysocka, J., Swigut, T., Xiao, H., Milne, T. A., Kwon, S. Y., Landry, J., Kauer, M., et al. (2006). A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature*, 442(7098), 86–90.
- Yoo, A. S., Staahl, B. T., Chen, L., & Crabtree, G. R. (2009). MicroRNA-mediated switching of chromatin-remodelling complexes in neural development. *Nature*, 460(7255), 642–6.
- Zambelli, F., Pesole, G., & Pavesi, G. (2012). Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in bioinformatics*.
- Zare-Mirakabad, F., Ahrabian, H., Sadeghi, M., Nowzari-Dalini, A., & Goliaei, B. (2009). New scoring schema for finding motifs in DNA Sequences. *BMC bioinformatics*, 10, 93.
- Zhao, X., Huang, H., & Speed, T. P. (2005). Finding short DNA motifs using permuted Markov models. *Journal of computational biology*, 12(6), 894–906.
- Zhao, Yong, & Srivastava, D. (2007). A developmental view of microRNA function. *Trends in biochemical sciences*, 32(4), 189–97.
- Zhao, Yue, Granas, D., & Stormo, G. D. (2009). Inferring binding energies from selected binding sites. *PLoS computational biology*, 5(12), e1000590.
- Zhou, Q., & Liu, J. S. (2004). Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics (Oxford, England)*, 20(6), 909–16.
- Zilberman, D., Coleman-Derr, D., Ballinger, T., & Henikoff, S. (2008). Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. *Nature*, 456(7218), 125–9.
- Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., & Henikoff, S. (2007). Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nature genetics*, 39(1), 61–9.

8. ANNEX

8.1. Table of Tables	1
8.2. Table of Figures	1
8.3. Mixture Model of CENTIPEDE	4
8.4. Comparison of STEME detected Motifs with MEME	5
8.5. AUC profiles of UniPROBE Mus musculus set	6
8.6. Numerical Instabilities in ET- α -Training	27
8.7. Number of SNPs per Peak	28
8.8. BS Significance Distributions	29
8.9. ROC curve Analysis for BS significance Difference Threshold Detection.....	32
8.10. TFBS Landscapes Candidate SNP List	34
8.11. Top causative SNP Candidate for SPI1	41
8.12. A relational Database System to Investigate Biological Plausibility	42
8.13. Publications	43
8.14. Computing Resources Used and Source Code Snippets	45

8.1. Table of Tables

Table 1: Variants discovered by the 1000 genome project for the CEU samples by project	11
Table 2: Illustration of the training input set as is used to learn a PWM.	52
Table 3: Illustration of the initial step for PWM training	53
Table 4: Illustration of the initial step for PWM training II	54
Table 5: Illustration of the initial step for PWM training III.	54
Table 6: Illustration of the initial step for Tree training..	57
Table 7: Sum of the weighted paired counts over all probes in the training set.	57
Table 8: Mutual information of a 9-mer	84
Table 10: Sequence logos for E2F2, Eomes, Esrra, Gcm1 and Myb.	94
Table 11: Graphical representation of the CLTree learned for E2F2, Eomes, Esrra, Gcm1 and Myb.	95
Table 12: Amino acid and nucleotide similarities between the TFs selected and their <i>Mus musculus</i> homologue.	101
Table 13: Sequence logos for EGR1, IRF4, MAX, SPI1 and SRF.	102
Table 14: Graphical representation of the CLTree learned for EGR1, IRF4, MAX, SPI1 and SRF.	103
Table 15: Listing of SRA-files used.	105
Table 16: Listing of SRA-files for which trimming has been performed.	105
Table 17: Sequence reads mapped to the reference, maternal and paternal genome by means of Bowtie.	106
Table 18: Number of detected peaks per TF and genome	107
Table 19: Detection of common and missing peaks in the maternal and paternal genome.	108
Table 20: Results of the Sensitivity and Specificity analysis for differential binding profiles between the maternal and paternal genome for EGR1, IRF4, MAX, SRF and SPI1 for the ET, PWM and CLTree model.	110
Table 21: Absolute number of SNPs with the potential of having a causative effect.	114
Table 22: List of predicted causative SNPs for EGR1, SPI1 and SRF.	116
Table 23: Classification of TFs analysed based on CLTree- and ET model performance in UniPROBE data	127
Table 24: Number of different binding sequences for SPI1 and EGR1 extracted from ChIP-seq and UniPROBE data, functioning as input for model learning.	130
Table 25: Listing of STEME and MEME detected motif for E2F2, E2F3, Eomes, Esrra, Gcm1 and Myb. ...	5

8.2. Table of Figures

Figure 1: Development of the number of biological databases from 1996 to 2011.	5
Figure 2: Exponential growth of GenBank. The number of nucleotide bases currently in GenBank.	7
Figure 3: Improvements in the rate of DNA sequencing over the past 30 years and into the future.	7
Figure 4: Gene regulation by transcription factors and microRNAs.	12
Figure 5: Simplified model of a complex for POLII catalysed transcription.	13
Figure 6: Components of transcriptional regulation.	15
Figure 7: Schematic representation of miRNA biogenesis and function.	17
Figure 8: Illustration of different "packing" levels of DNA	19
Figure 9: DNA methylation.	21
Figure 10: Histone modifications.	22
Figure 11: Cross-talk between H3K9, H3S10 and H3K14.	23
Figure 12: Epigenetic regulation depends on the interplay among the different players	24
Figure 13: Solid-phase amplification	26
Figure 14: The four-color CRT method used by Illumina/Solexa	27
Figure 15: Schematic overview ChIP-seq experiment	29
Figure 16: Example of ChIP profiles generated by chromatin immunoprecipitation followed by sequencing (ChIP-seq) or by microarray (ChIP-chip).	31

Figure 17: Cost per raw megabase of DNA Sequence and per genome (2001-2012)	32
Figure 18: Consensus sequence model	34
Figure 19: (in continuation to Figure 18) The frequency matrix model	34
Figure 20: Illustration of positional interdependency based on a fictive example	35
Figure 21: Examples of different Bayesian network models for a sequence motif with 5 positions	38
Figure 22: Overview of a workflow for <i>cis</i> -regulatory variant detection.....	42
Figure 23: Schematic illustration of a de Bruijn sequence with order 3	45
Figure 24: Schematic illustration of the three major stages of a PBM experiment	46
Figure 25: Schematic illustration of the generation of the training input set.....	49
Figure 26: Schematic illustration of the screening algorithm.	50
Figure 27: Two different PWM representations	52
Figure 28: Generation of the tree structure based on mutual information	59
Figure 29: Illustration of a complete graph with 6 nodes and 15 edges.....	59
Figure 30: Schematic illustration of the model validation procedure.	65
Figure 31: ROC curves for 4 distinctive intensity thresholds.....	66
Figure 32: Schematic illustration of the different algorithmic steps performed to detect causal	68
Figure 33: Construction of a personal genome by vcf2diploid tool.....	69
Figure 34: Graphical illustration of the procedure to generate a p-value for TF binding	71
Figure 35: Illustration of the calculation the SNE – or SNP effect - p-value.	72
Figure 36: Schematic illustration of the principle ChIP-seq analysis including a peak detection step.....	73
Figure 37: Example FASTQ-format	73
Figure 38: Forward and reverse read density profiles.....	77
Figure 39: Graphical illustration of cases considered when merging screening data of common peaks.....	81
Figure 40: Alleleseq pipeline	83
Figure 41: Graphical illustration of the relation of positive and negative reference data sets to true positives, true negatives, false positives and false negatives.	85
Figure 42: Bar chart illustrating the distribution of AUC-mean differences over all datasets.....	88
Figure 43: Three example validation profiles for the UniPROBE mouse dataset subgroup classification	89
Figure 44: Bar chart representing the subgroups of those TF-sets showing a better validation profile with the CLTree in comparison to a PWM.....	89
Figure 45: Exemplary AUC-profiles for the UniPROBE TF Spdef.....	90
Figure 47: Barchart representing the percentage of improved validation profiles by means of the ET model	91
Figure 48: Validation profiles (AUC-curve) for Gcm1 and Gata6	91
Figure 49: Venn-Diagram showing the overlap of the 19 TFs published from Bulyk <i>et al</i> with those TFs showing better validation profiles with a tree structure based model.....	92
Figure 50: AUC-profile of E2F2.....	97
Figure 51: AUC-profile of Eomes	98
Figure 52: AUC-profile of Esrra.....	99
Figure 53: AUC-profile of Gcm1.....	100
Figure 54: AUC-profile of Myb	100
Figure 55: AUC-profiles for EGR1, IRF4, MAX, SPI1 and SRF	104
Figure 56: Relative distribution of SNPed versus not-SNPed peaks in the parental genomes	108
Figure 57: Relative distribution of SNPed versus not-SNPed and lost peaks in the subgroup of peaks only detected in one genome.	109
Figure 58: Sensitivity/Specificity plot for model comparison for EGR1, IRF4, MAX, SRF and SPI1.....	111
Figure 59: Graphical illustration of winning-parent assignment of Alleleseq.....	112
Figure 60: Column charts illustrating the relative amount of correctly and incorrectly predicted winning parent for all TFs and all models.	112
Figure 61: Bar chart illustrating the relative amount of parental BS, where both, only one or none BS are co-located with a SNP.....	113
Figure 62: Graphical illustration of the final amount of potentially causative SNPs.....	114
Figure 63: Venn Diagram illustrating that the resulting candidate list is quasi exclusively based on the ET model based rSNP detections.	115
Figure 64: Top causative SNP candidate for EGR1.	118
Figure 65: Paternal/maternal peak sequence alignment rs750239.....	118
Figure 66: TFBS landscape for EGR1 with the ET model.....	119

Figure 67: Top causative SNP candidate for SRF.	120
Figure 68: Paternal/maternal peak sequence alignment rs5570871.	121
Figure 69: TFBS landscape for SRF with the ET model	121
Figure 70: AUC profiles for EGR1 and SPI1 trained and validated on 2,000 randomly chosen ChIP-seq peaks.	130
Figure 71: Schematic illustration of biological dependencies to be checked when investigating the biological plausibility of an rSNP.	140
Figure 72: Pilot relational database scheme developed to mirror biological connections to be considered to investigate biological plausibility.	141

8.3. Mixture Model of CENTIPEDE

The hierarchical mixture model (only a superficial introduction will be given) is represented by

$$P(D_l) = P(Z_l = 1|G_l)P(D_l|Z_l = 1) + P(Z_l = 0|G_l)P(D_l|Z_l = 0)$$

with

l denoting a motif match

D_l denote for observed experimental data, wherein D_l can come from two underlying distributions that form the mixture model, representing the bound state of the TF with $Z=1$: bound and $Z=0$: unbound.

G_l denote for the prior information around the motif match

The hierarchical character of the model results from the fact, that the prior probabilities of observing a bound motif are based on another prior information, which is G_l (e.g. conservation, distance to TSS, score of the PWM). For each potential binding location l the prior probability $\pi_l = P(Z_l = 1|G_l)$, that the site is used, is calculated. This prior probability depends on various genomic information, which can describe a site l and is represented by a logistic regression model

$$\log\left(\frac{\pi_l}{1 - \pi_l}\right) = \beta_0 + \beta_1 \times PWMscore_l + \beta_2 \times Cons.Score_l + \beta_3 \times TSSProximity_l + \dots$$

with

model parameters β estimated by maximizing the likelihood function using an expectation maximization (EM) algorithm.

As experimental data CENTIPEDE can combine many different types of experiments, like DNase-seq or histone modification ChIP-seq reads.

A single type of experimental data D_l is determined for a particular motif instance l as

$$P(X_l) = \pi_l P(X_l|Z_l = 1) + (1 - \pi_l) P(X_l|Z_l = 0)$$

with

$\pi_l = P(Z_l = 1|G_l)$ and $P(Z_l = 0|G_l) = 1 - \pi_l$,

For multiple experimental data-types the model is for example represented by

$$P(X_l^{(1)}, X_l^{(2)}, X_l^{(3)}) = \pi_l P(X_l^{(1)}, X_l^{(2)}, X_l^{(3)}|Z_l = 1) + (1 - \pi_l) P(X_l^{(1)}, X_l^{(2)}, X_l^{(3)}|Z_l = 0)$$

with

$X_l^{(1)}, X_l^{(2)}, X_l^{(3)}$ denoting three different discrete-count data types (general annotation: $X_l^{(k)}$, for the k^{th} data type).


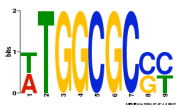



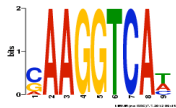




The model assumes independence between the different data types determining

$$P(X_l^{(1)}, X_l^{(2)}, X_l^{(3)}|Z_l = 1) = P(X_l^{(1)}|Z_l) P(X_l^{(2)}|Z_l) P(X_l^{(3)}|Z_l)$$

8.4. Comparison of STEME detected Motifs with MEME

In order to evaluate the goodness of STEME's motif detection, the sequence set has been analyzed by means of MEME (Bailey *et al.*, 2009).

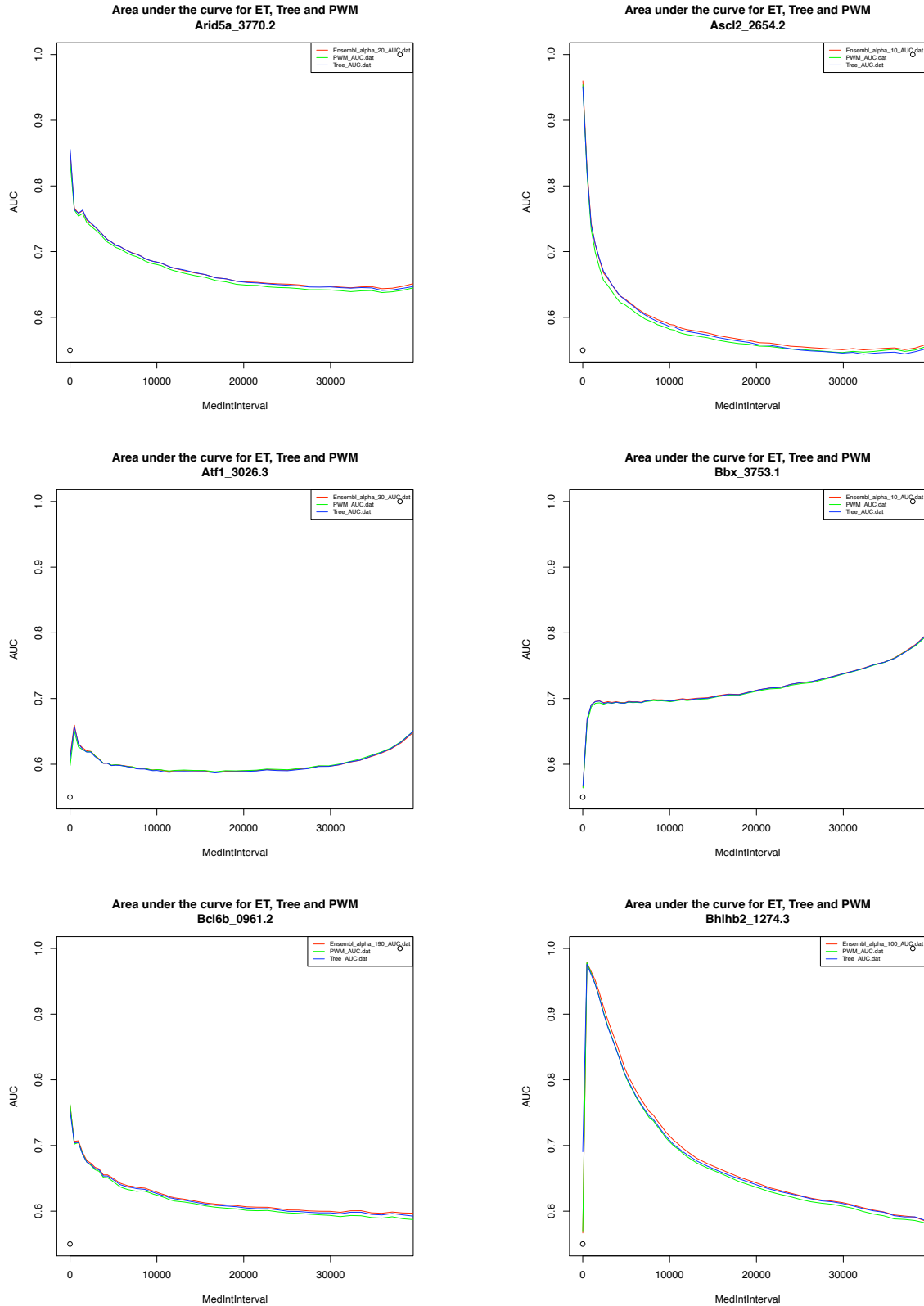
Table 27: Listing of STEME and MEME detected motif for E2F2, E2F3, Eomes, Esrra, Gcm1 and Myb.

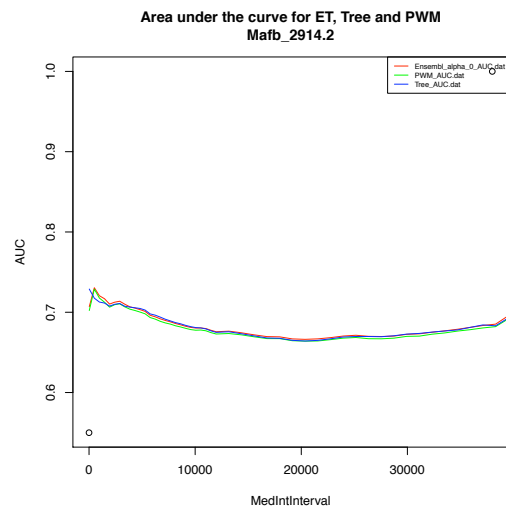
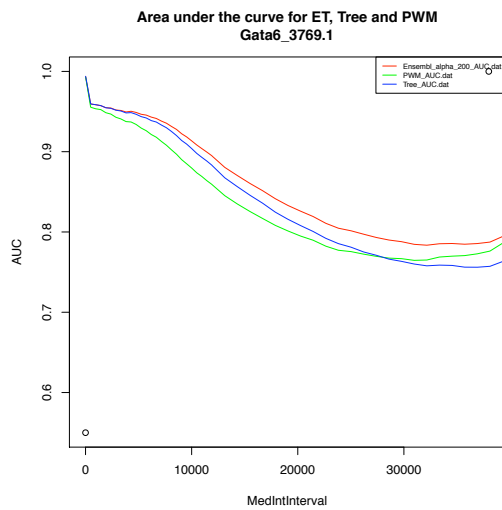
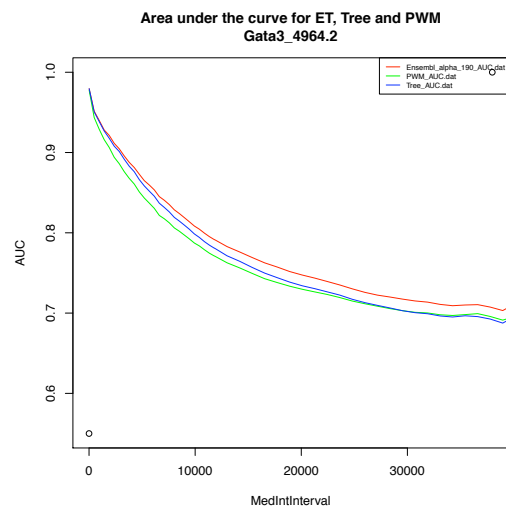
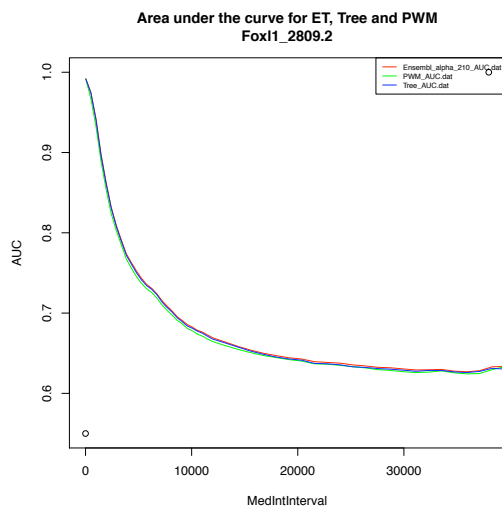
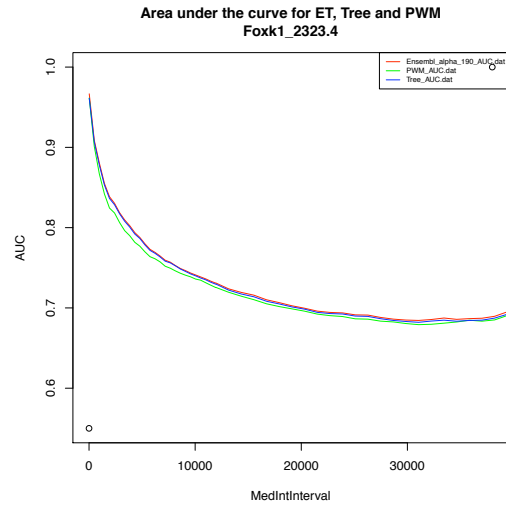
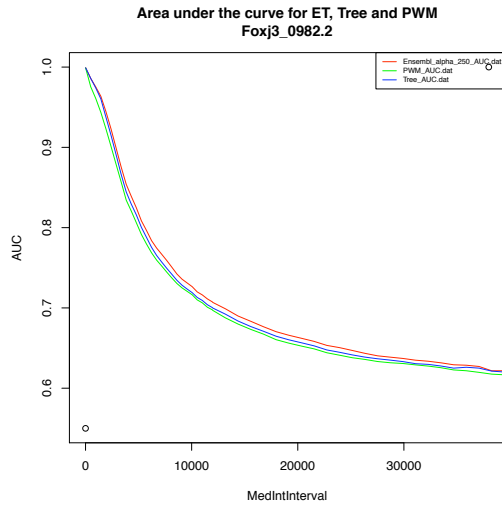
TF	STEME motif	MEME motif
E2F2		
Eomes		
Esrra		
Gcm1		
Myb		

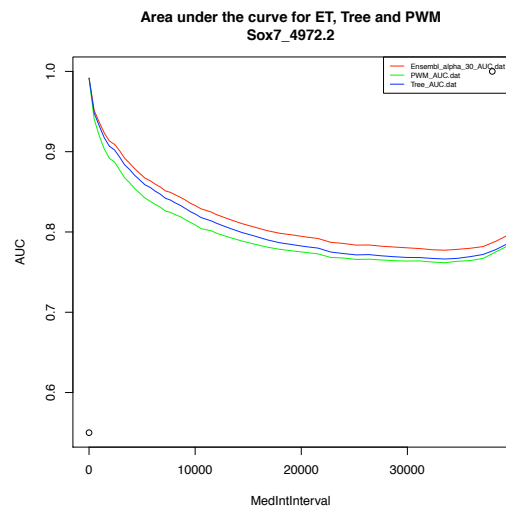
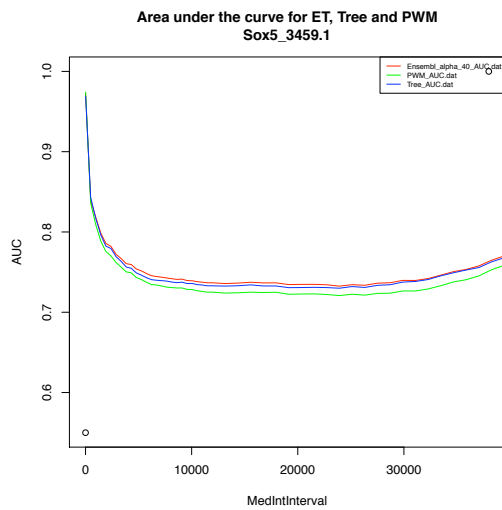
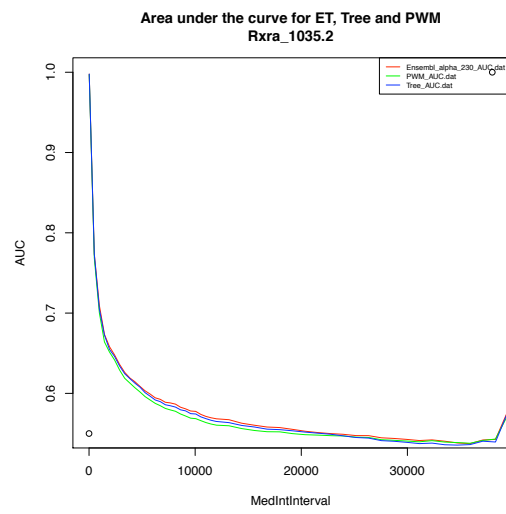
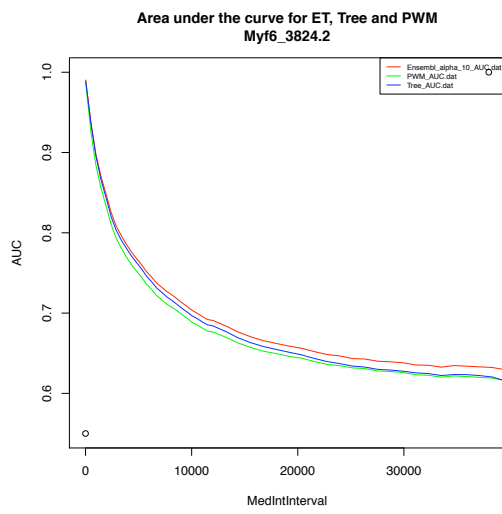
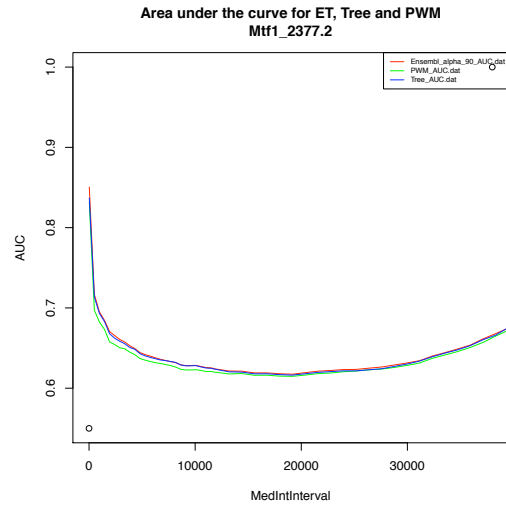
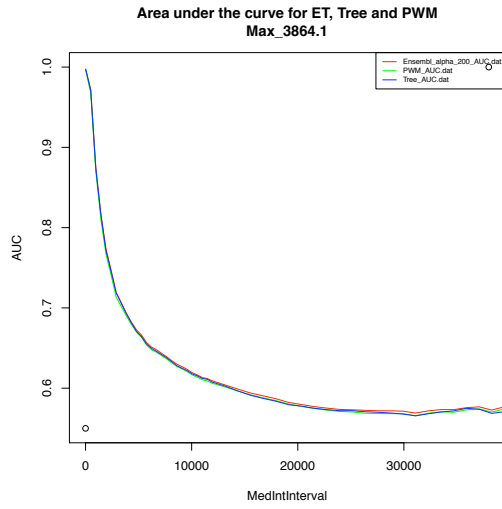
The following options have been selected for MEME: distribution of a single motif among the sequences: "Any number of repetitions", the minimum and maximum sequence with has been set, so that the motif width was fixed on 9, the number of motifs to find was set to 1. The motifs detected by STEME and MEME for the same input sequences are for those TFs listed in Table 27. Comparing the detected motifs for STEME and MEME no fundamental differences are observable. The most prominent nucleotide positions and core motifs are present in all motifs detected by both tools. In comparison to MEME, the STEME motifs show a more variable distribution of position conservation. For the MEME motifs the height, or conservation, of the different nucleotides is for a relative high amount of positions uniformly maximal. However, since the main core motif match well, the performance of STEME in comparison to MEME is considered as sufficiently well.

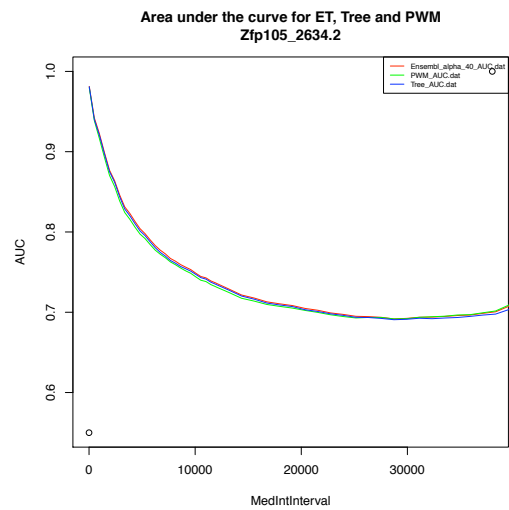
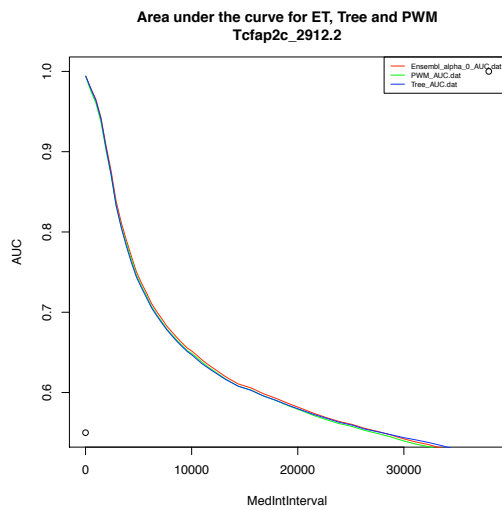
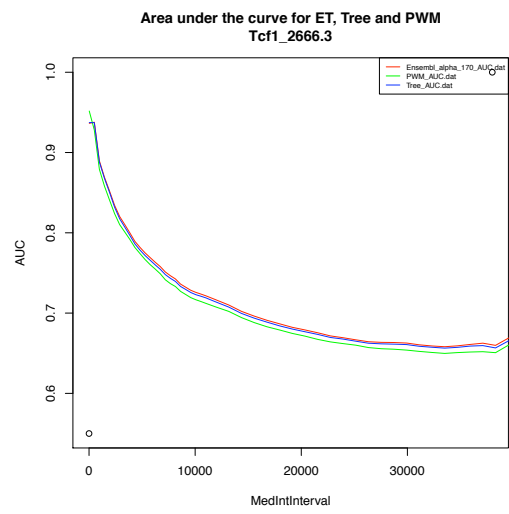
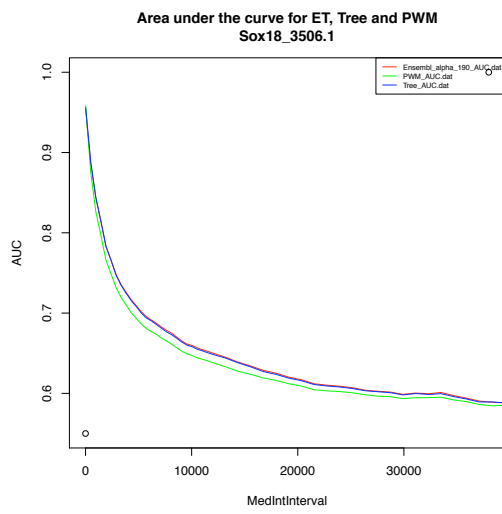
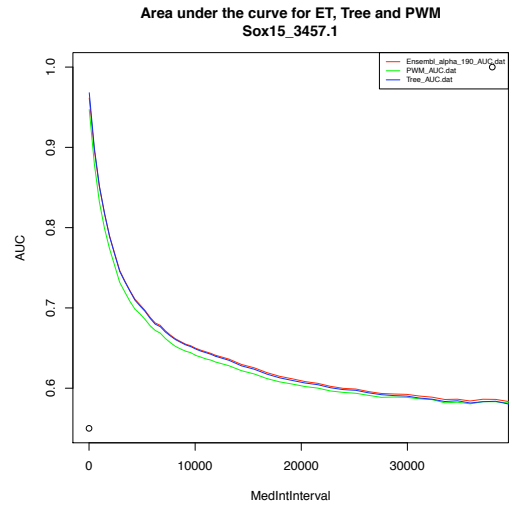
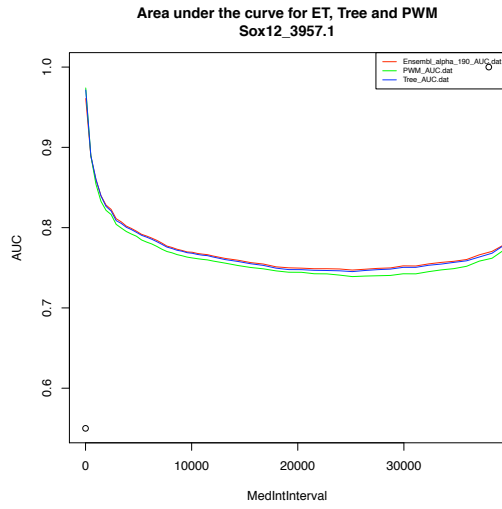
8.5. AUC profiles of UniPROBE Mus musculus set

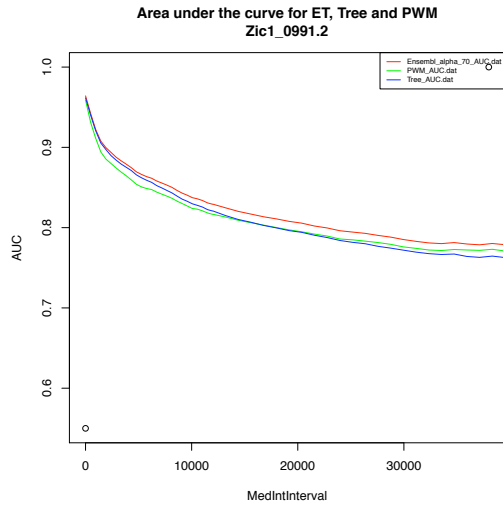
Tree better than PWM, small effect, improvement by ET model



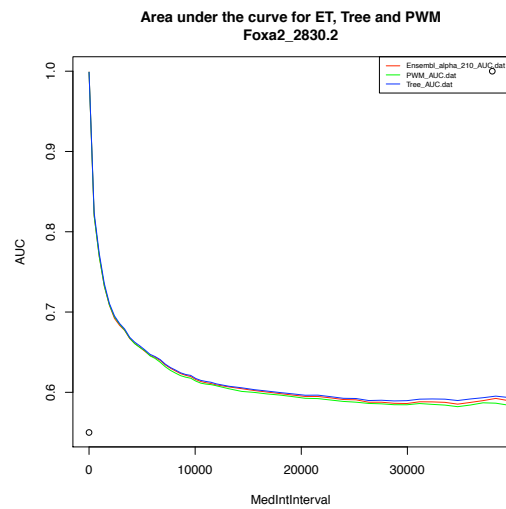
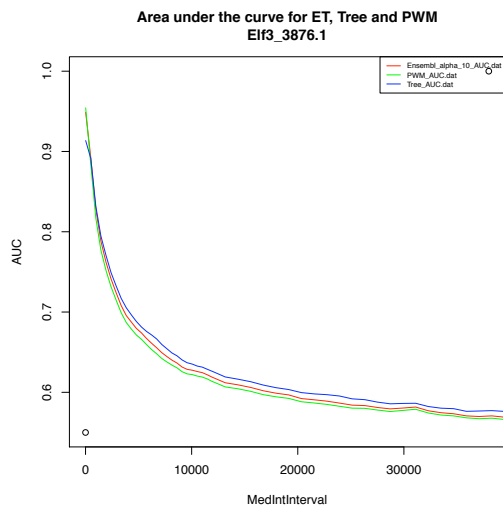
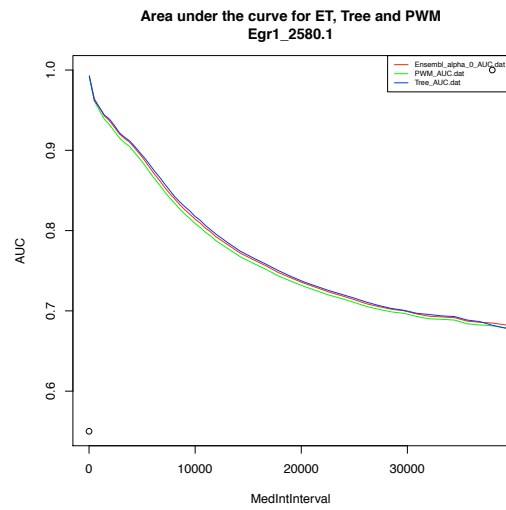
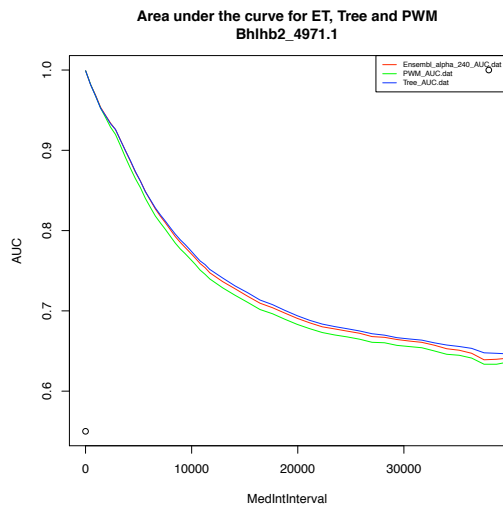


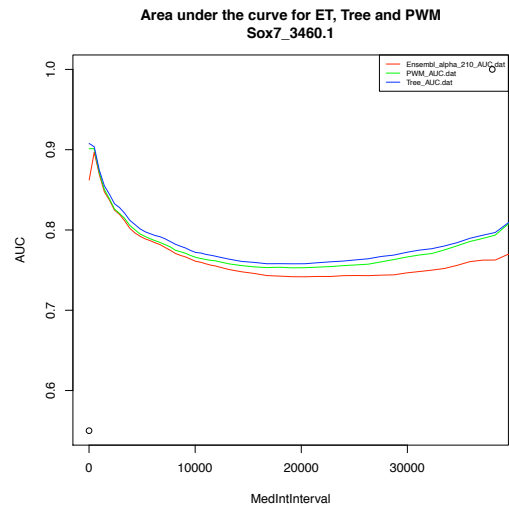
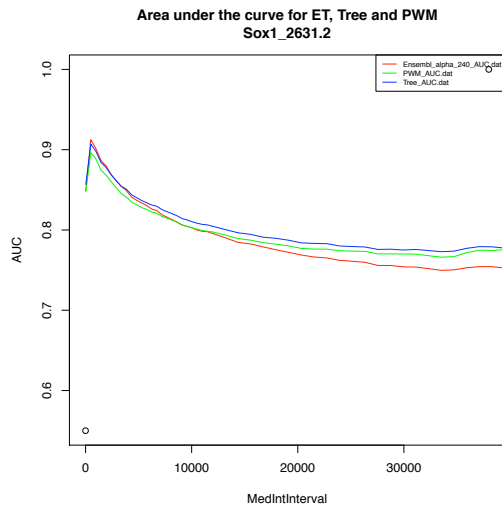
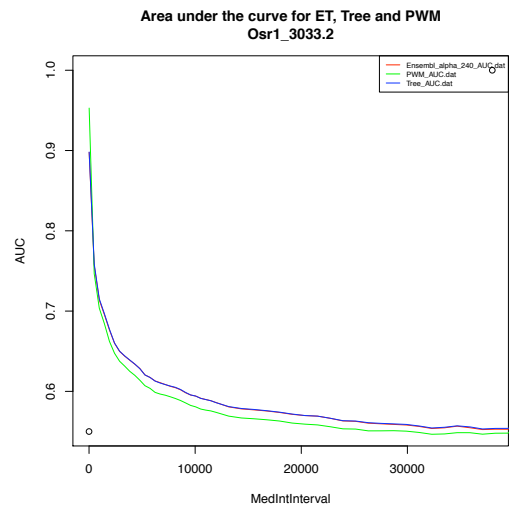
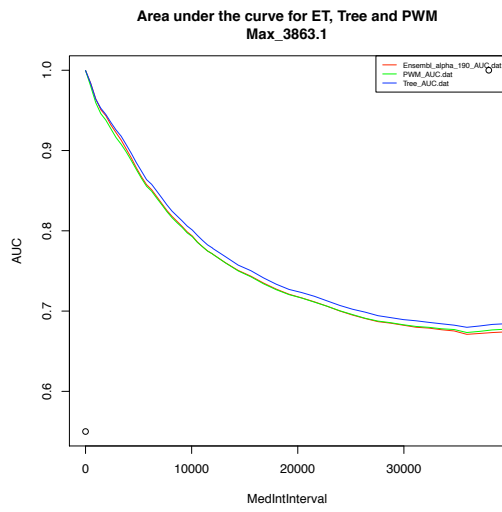
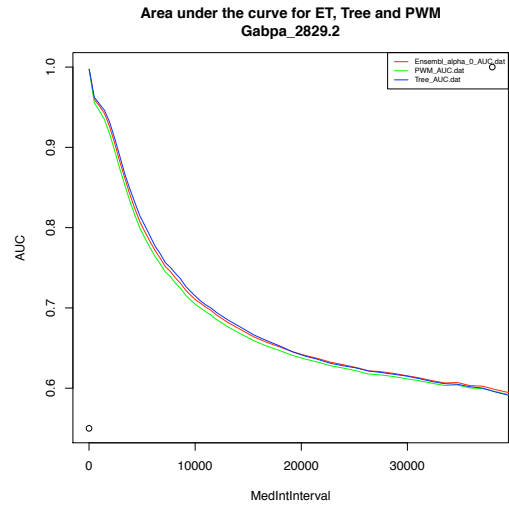
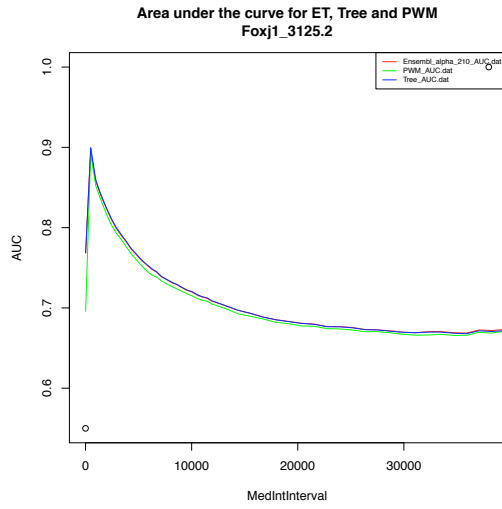


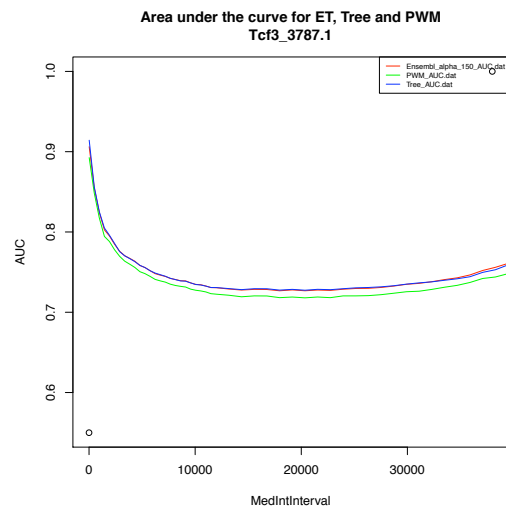
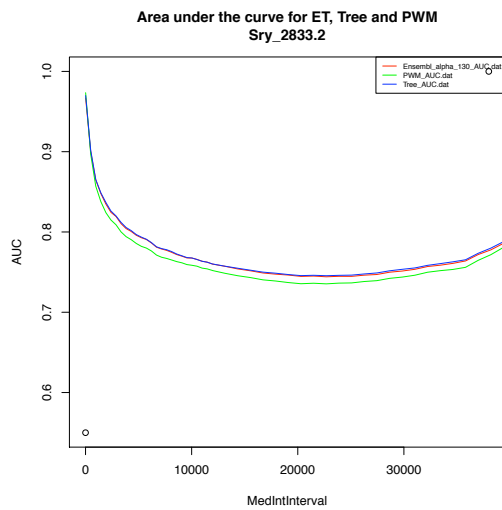
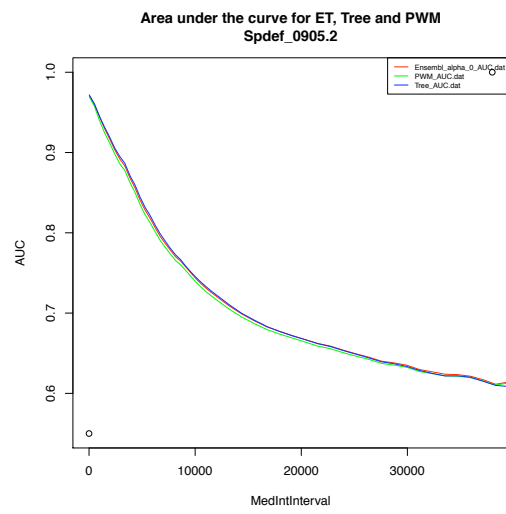
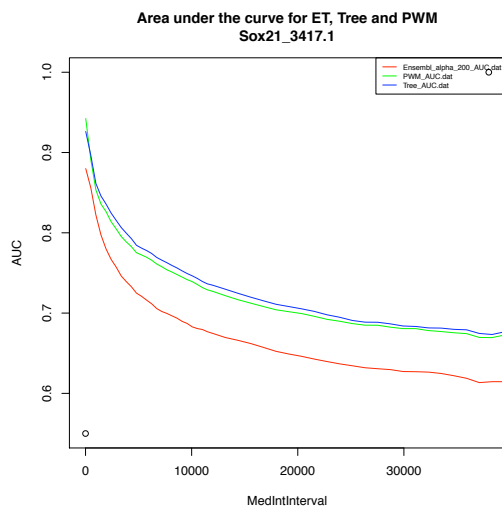
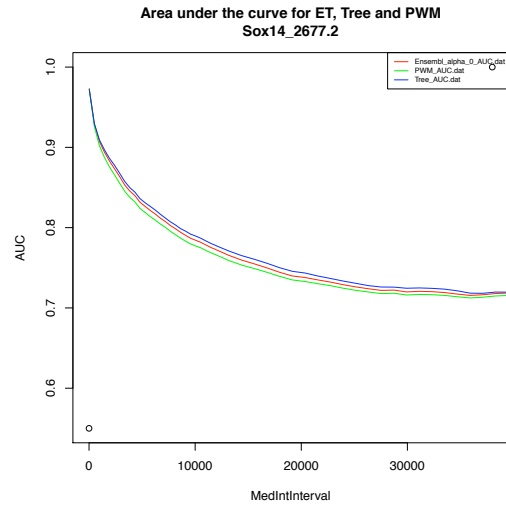
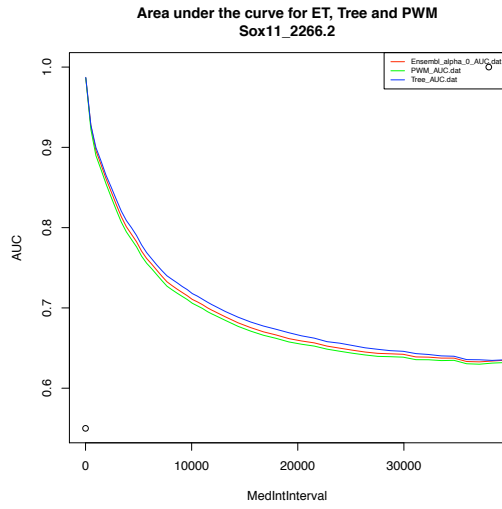


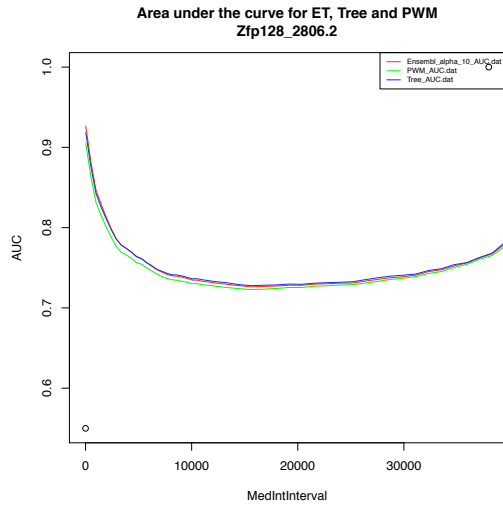


Tree better than PWM, small effect, no improvement by ET model

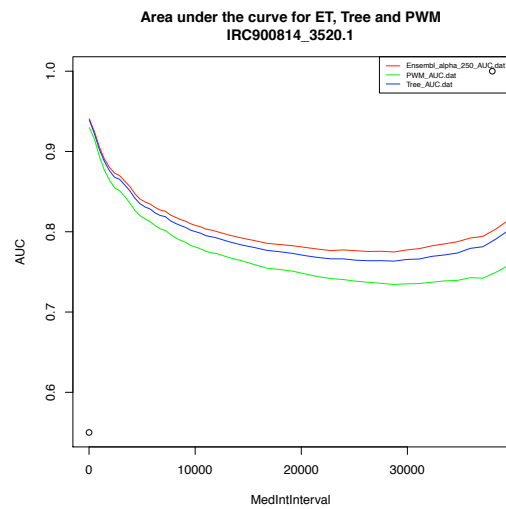
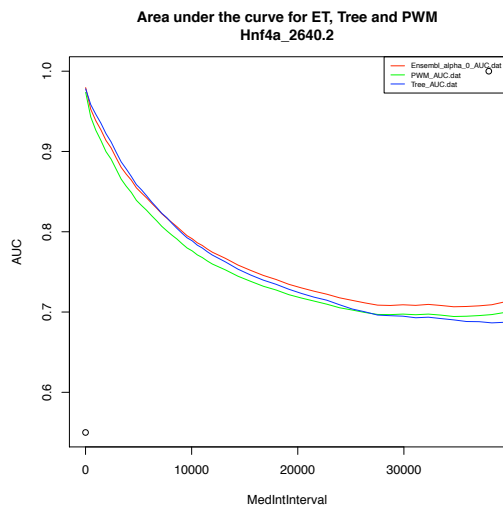
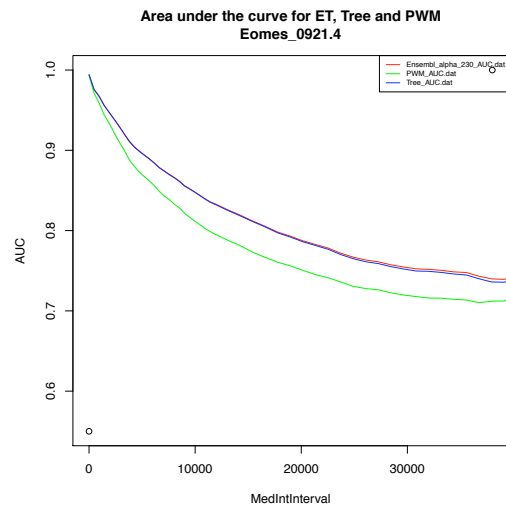
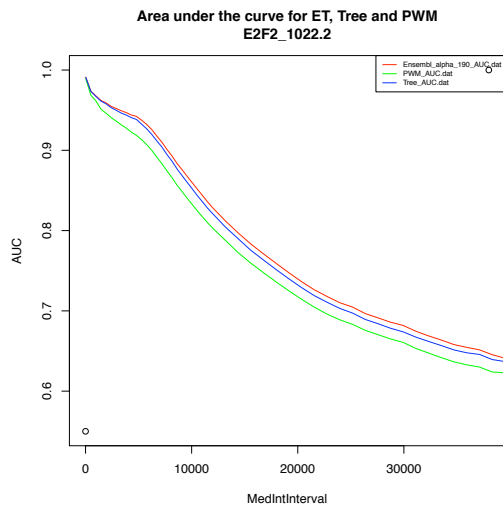


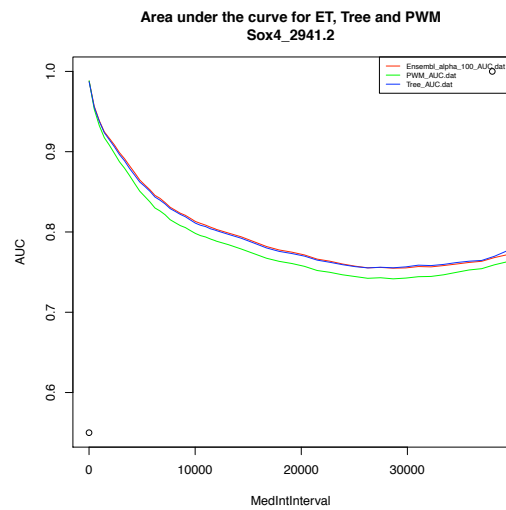
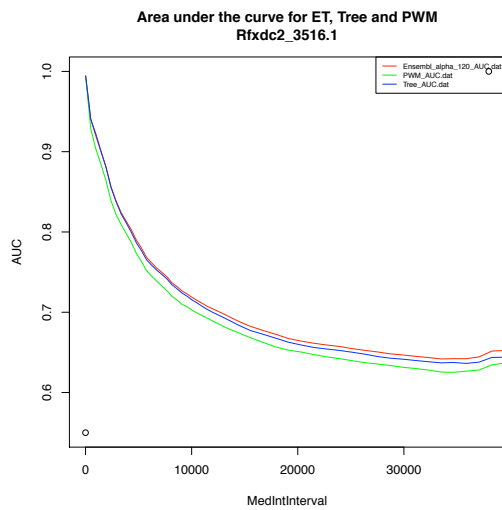
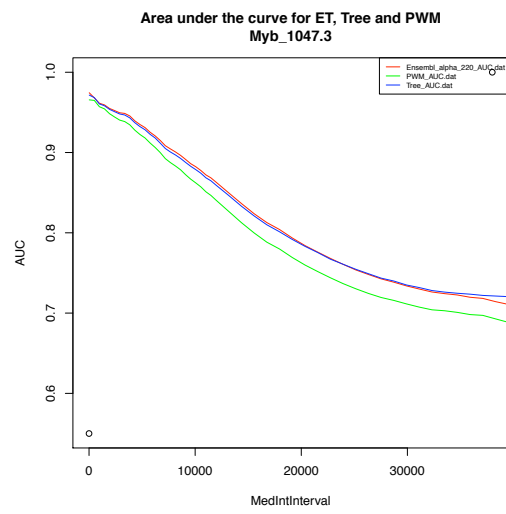
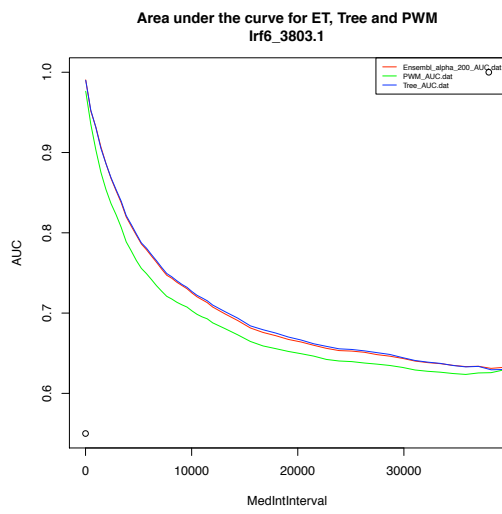
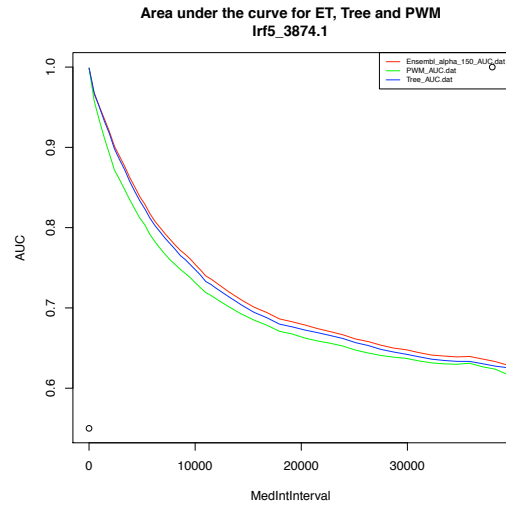
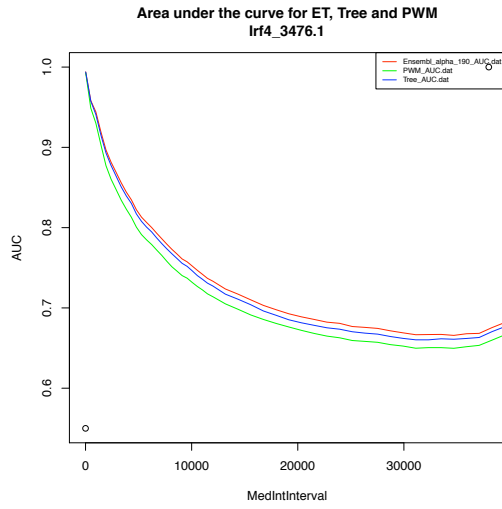


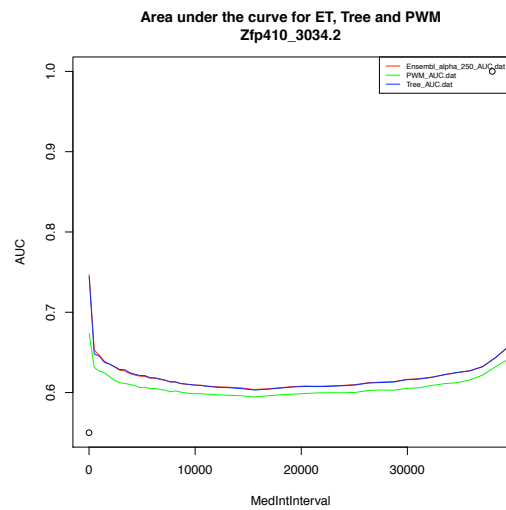
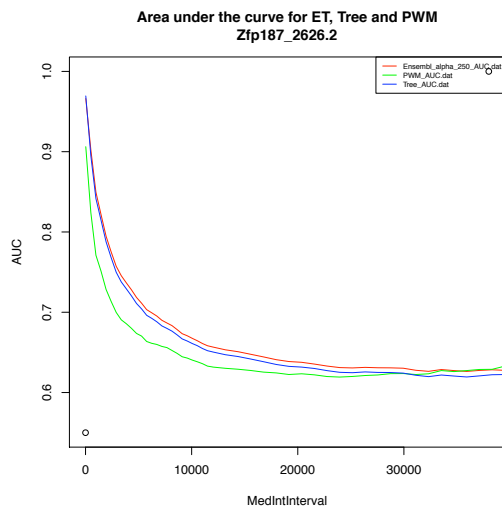
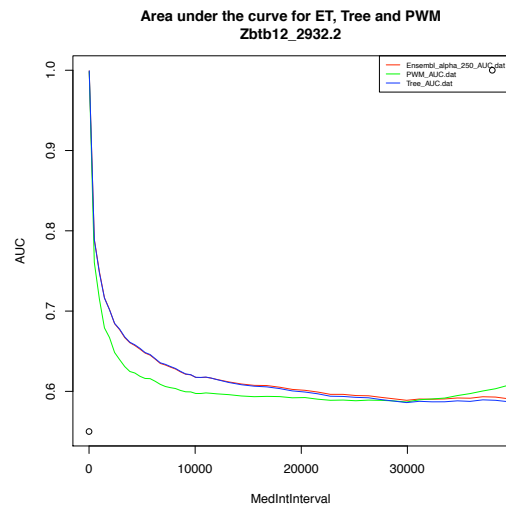
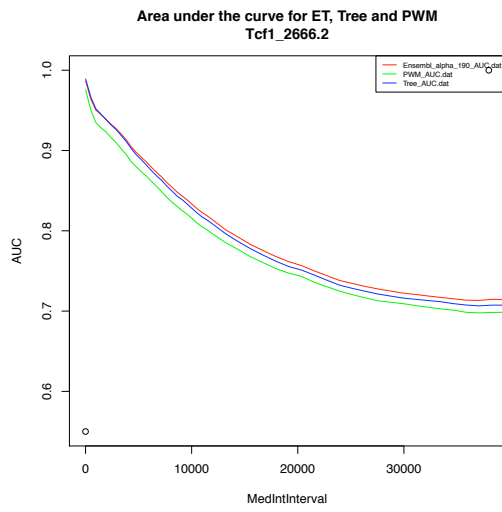
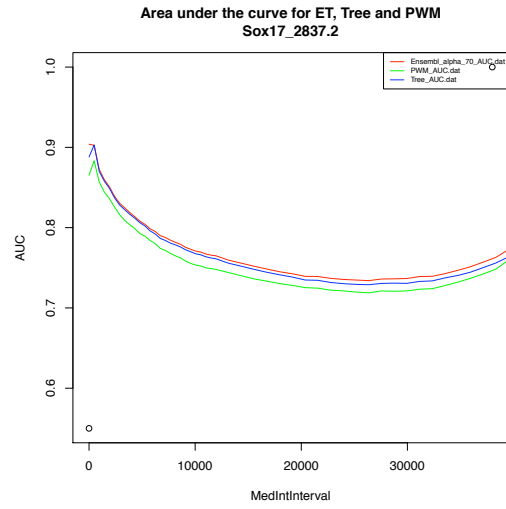
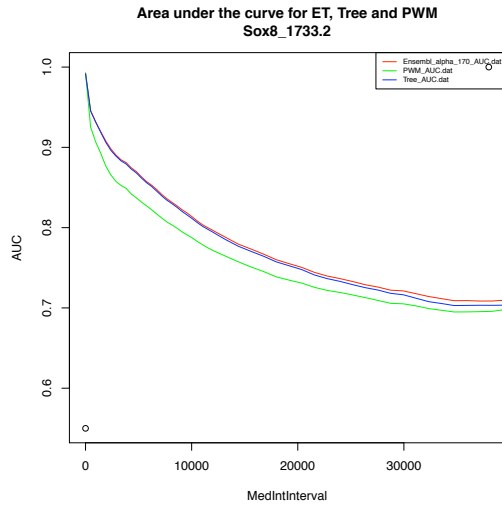


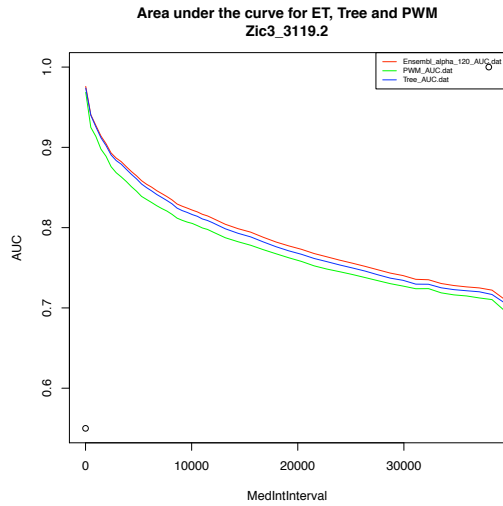


Tree better than PWM, medium effect, improvement by ET model

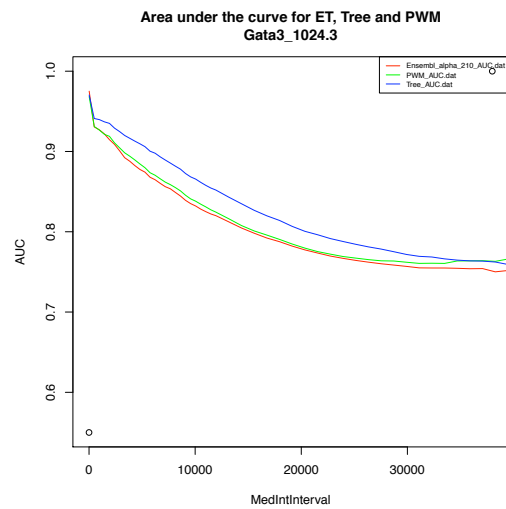
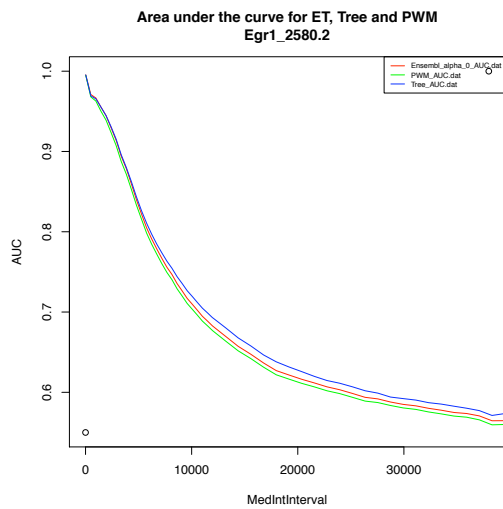
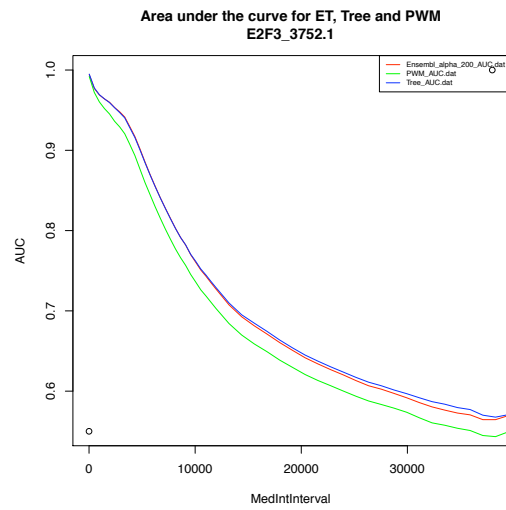
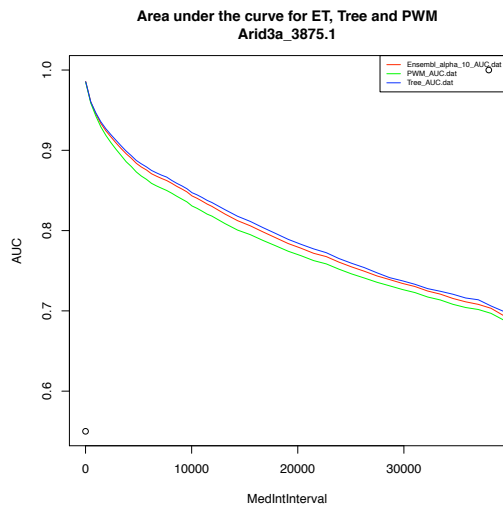


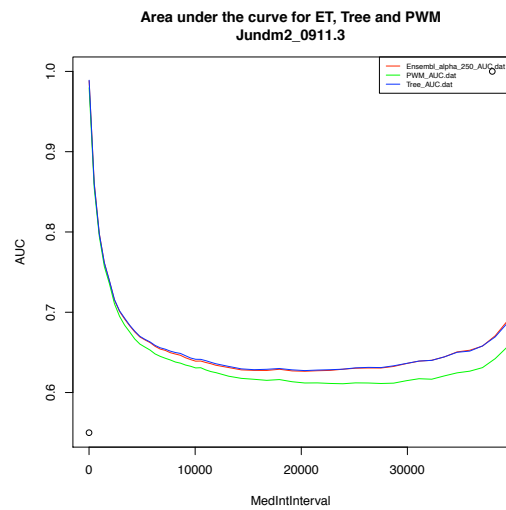
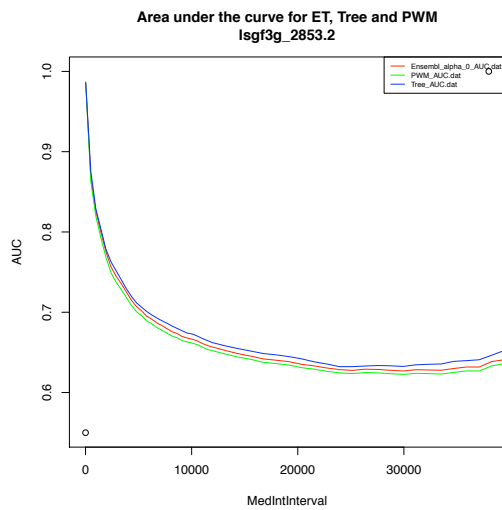
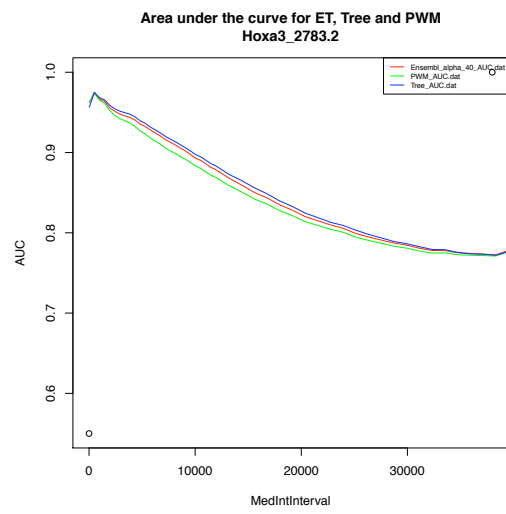
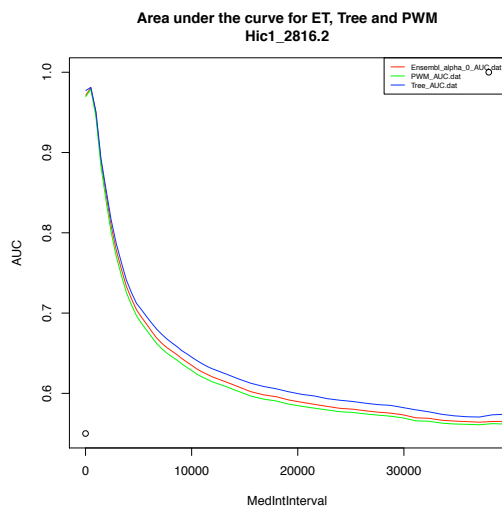
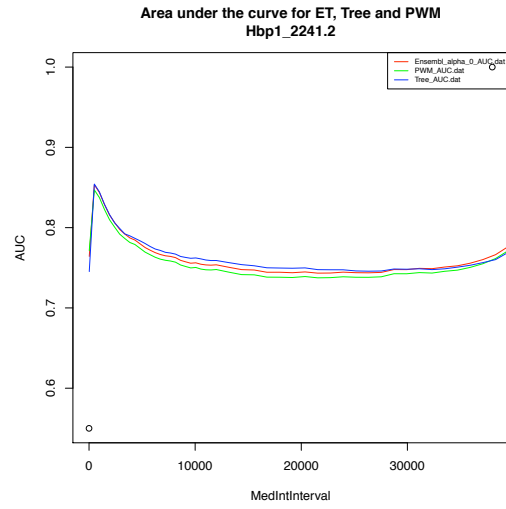
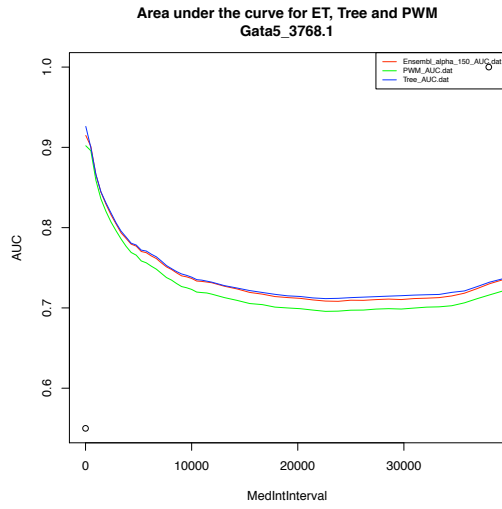


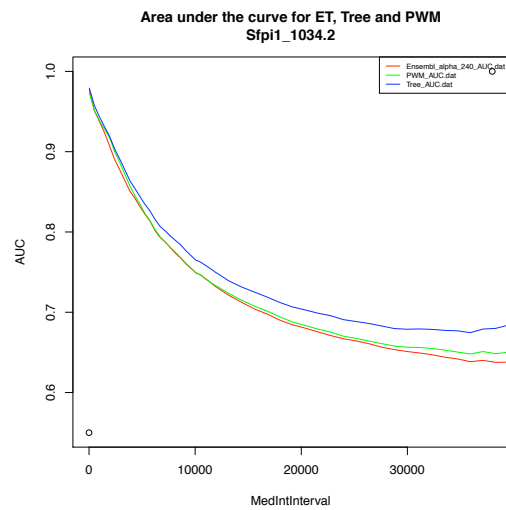
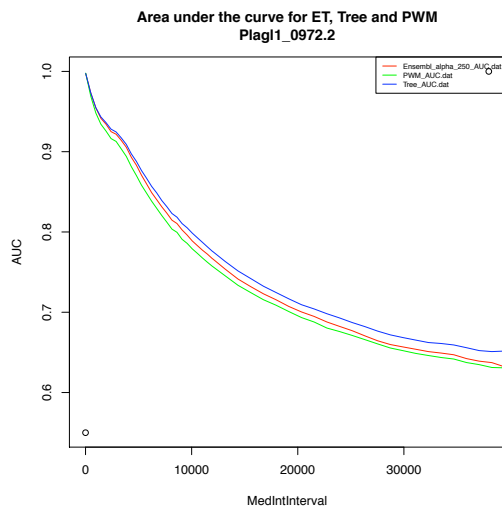
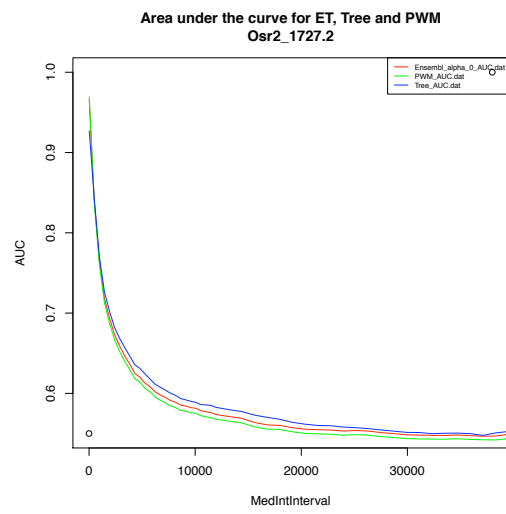
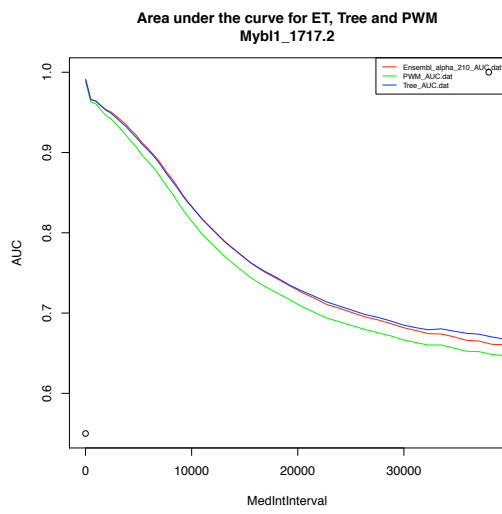
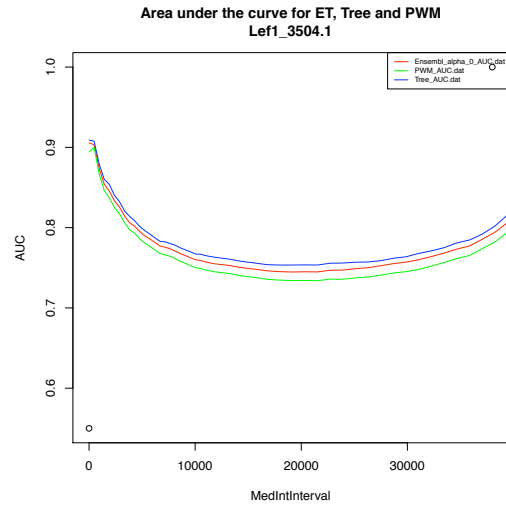
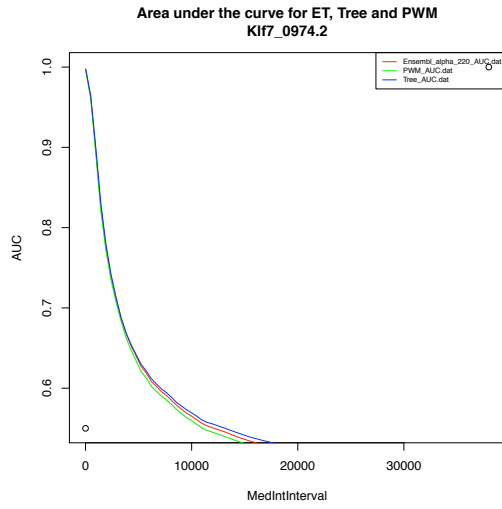


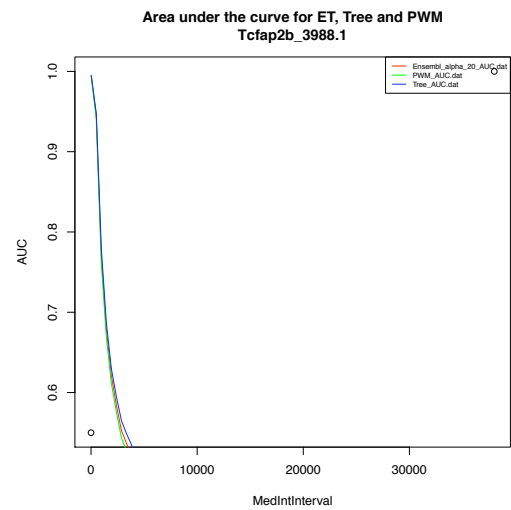
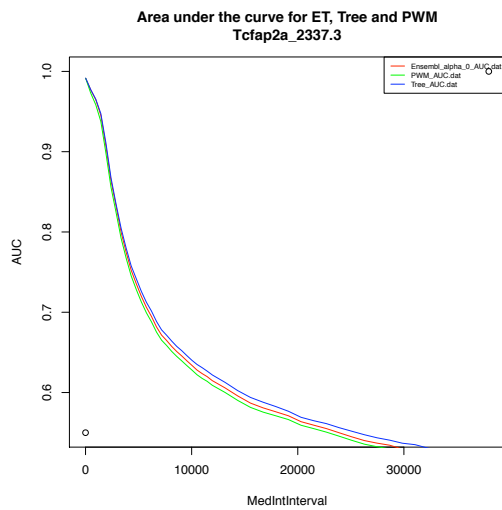
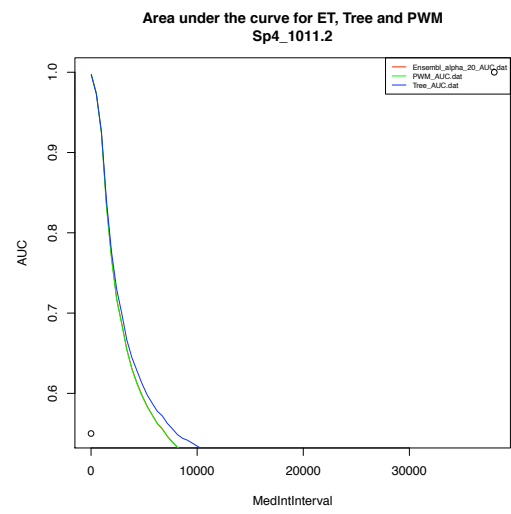
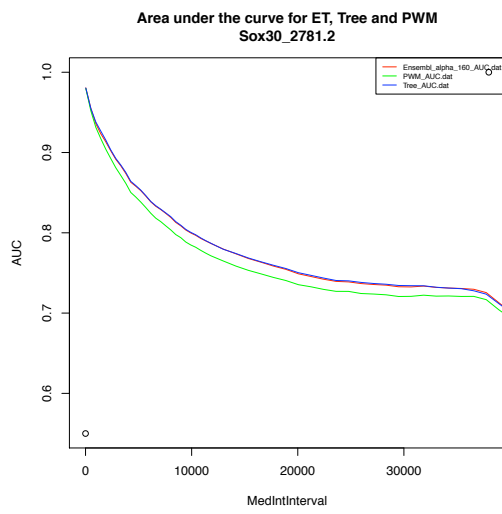
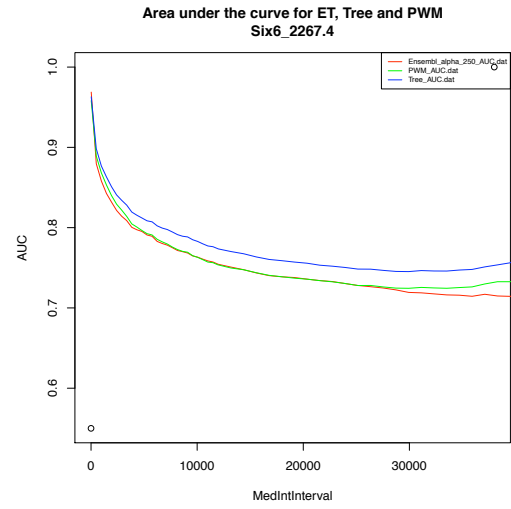
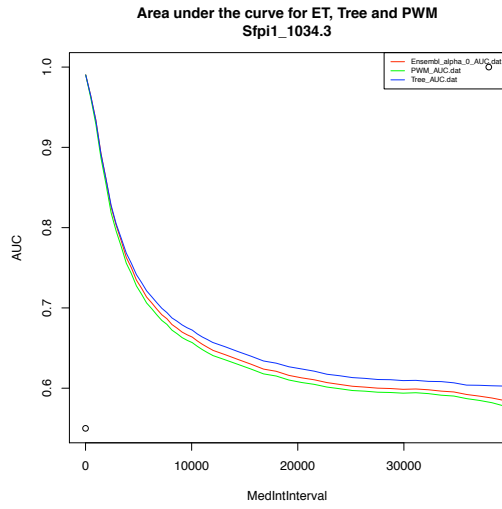


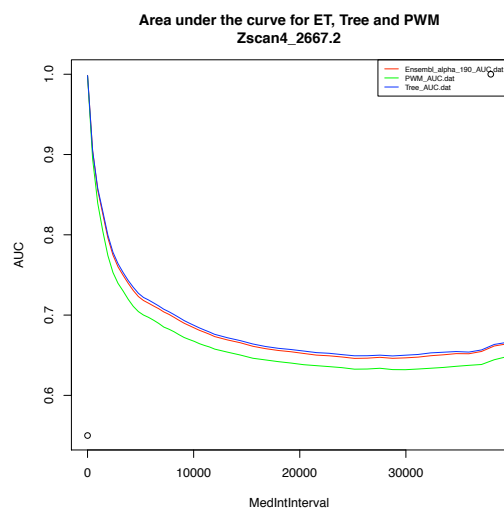
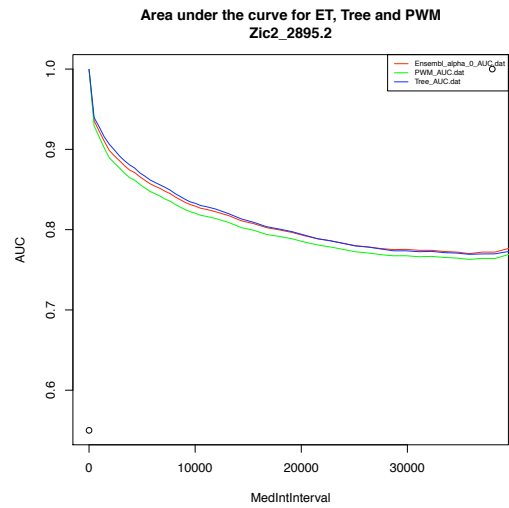
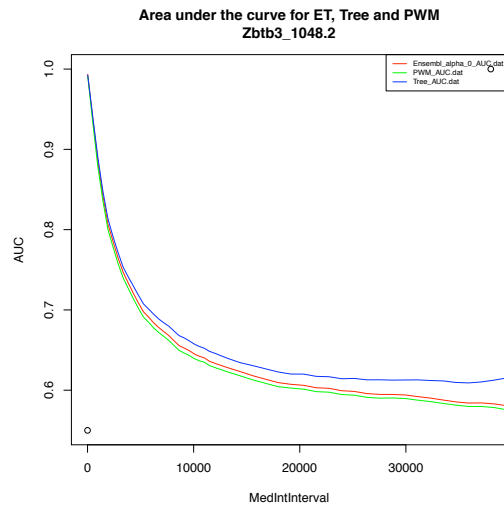
Tree better than PWM, medium effect, no improvement by ET model



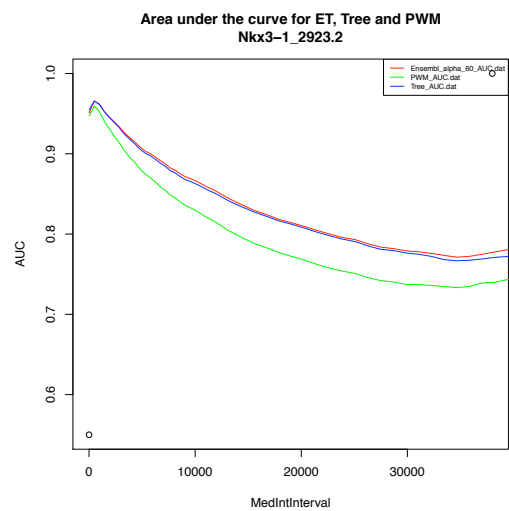
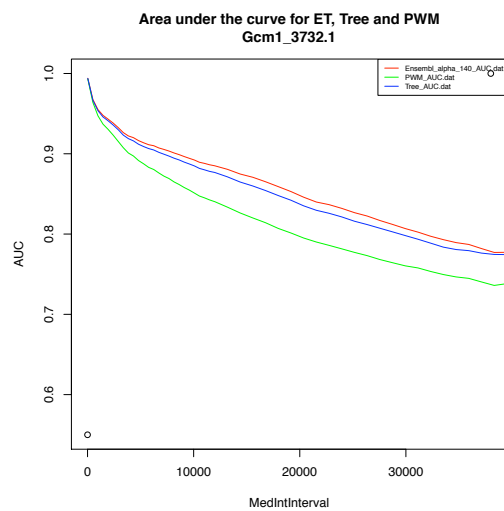


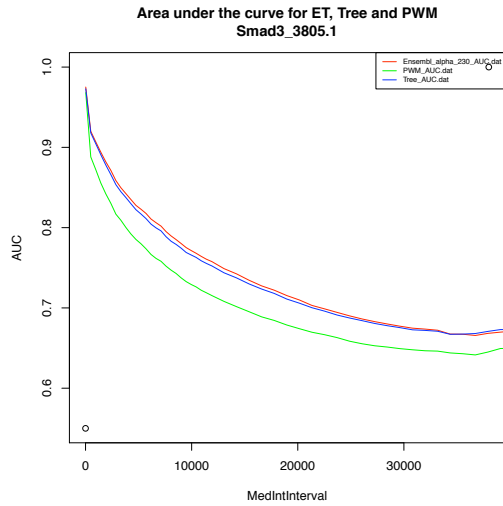




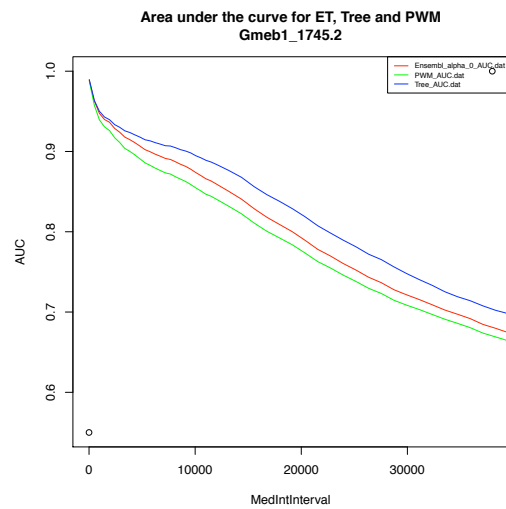
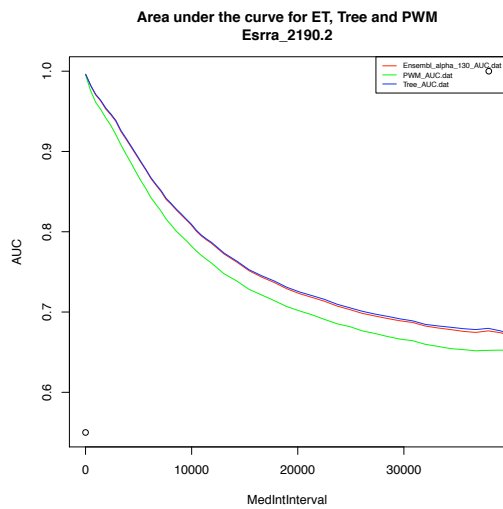
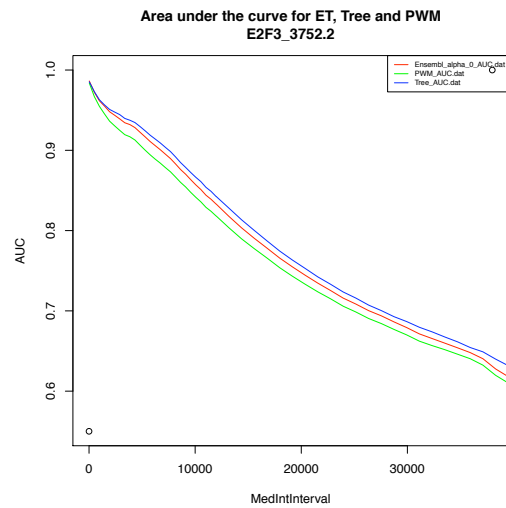
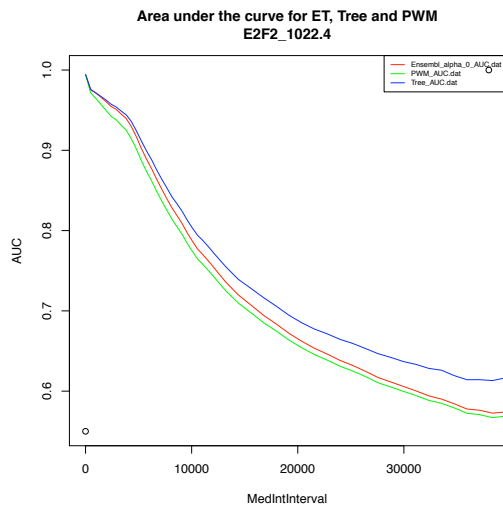


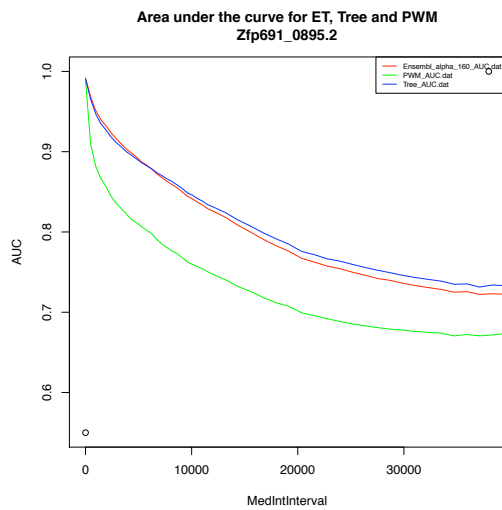
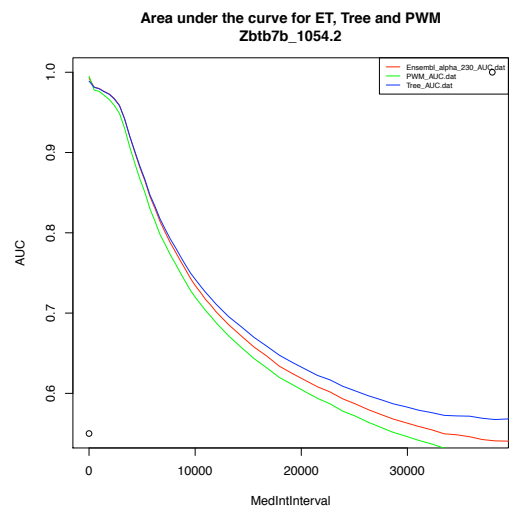
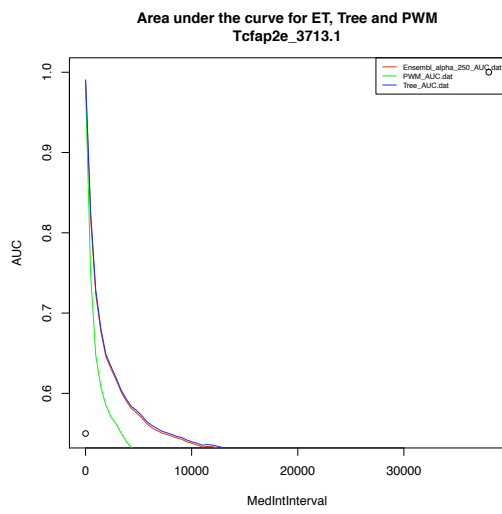
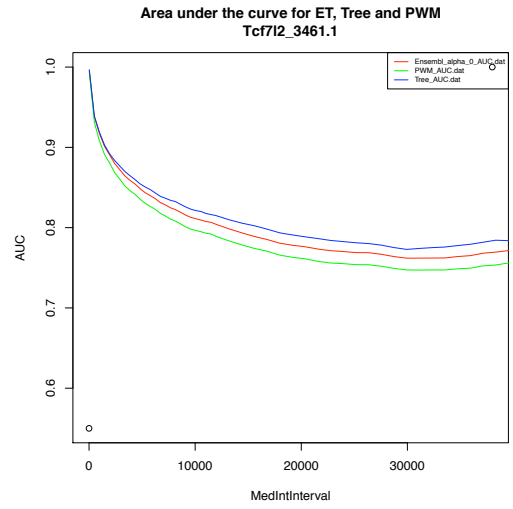
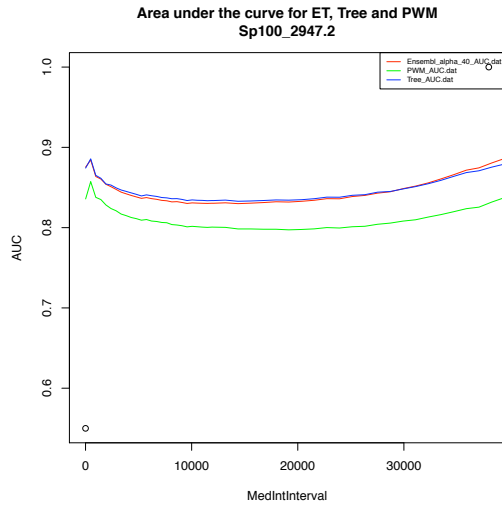
Tree better than PWM, strong effect, improvement by ET model



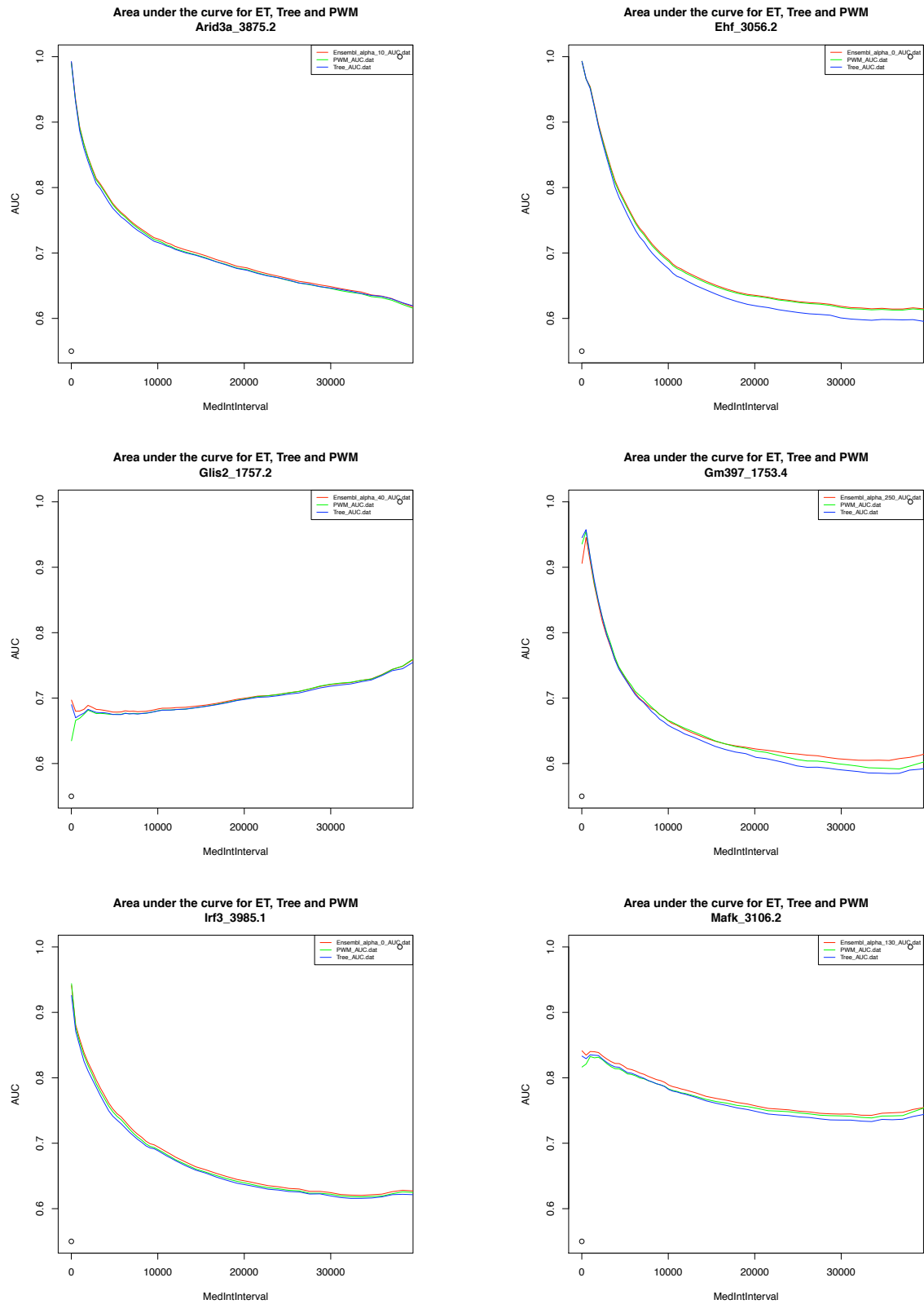


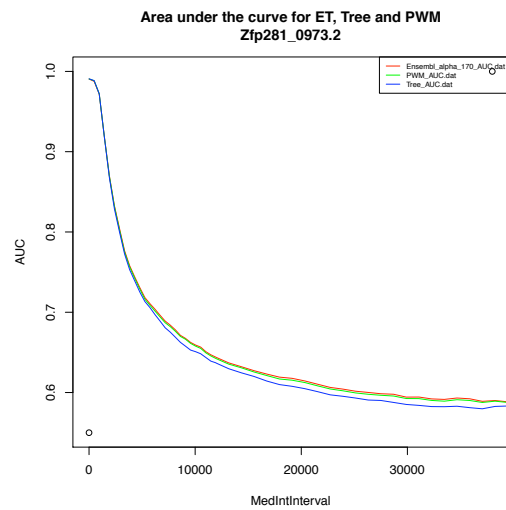
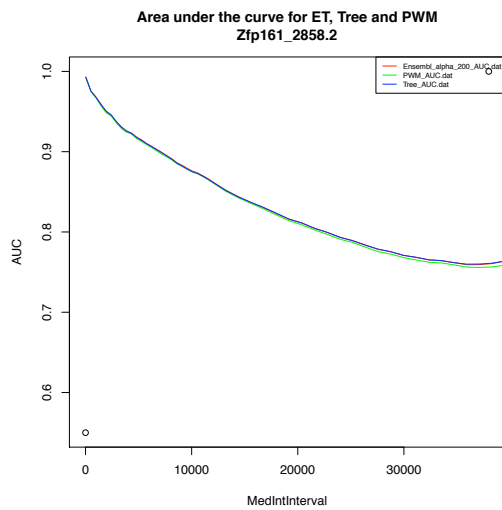
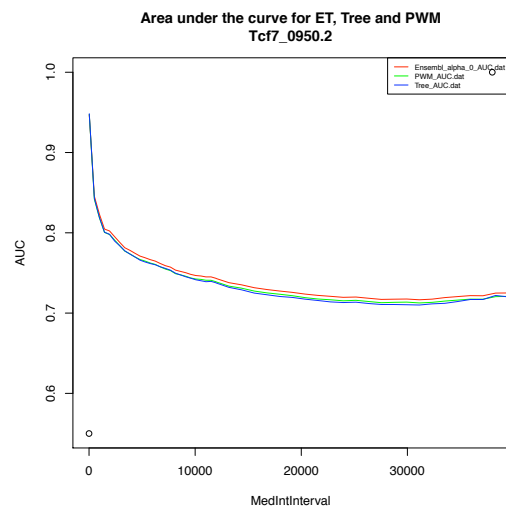
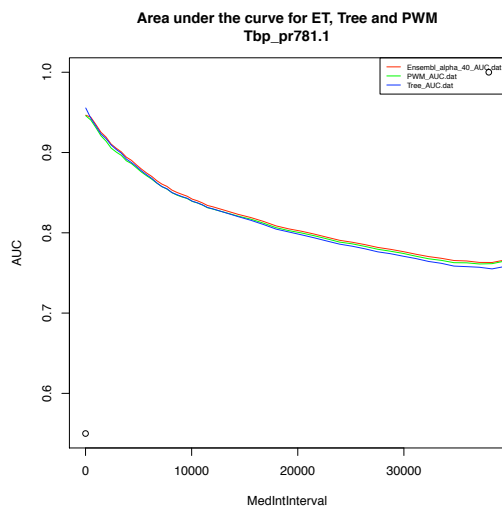
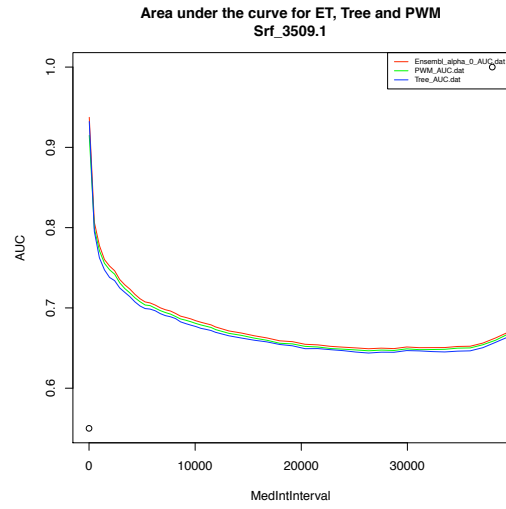
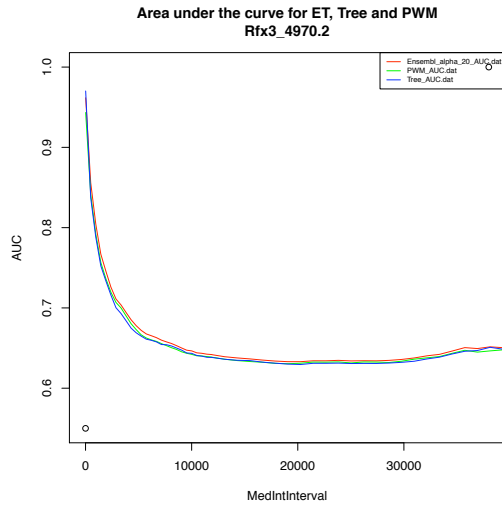
Tree better than PWM, strong effect, no improvement by ET model

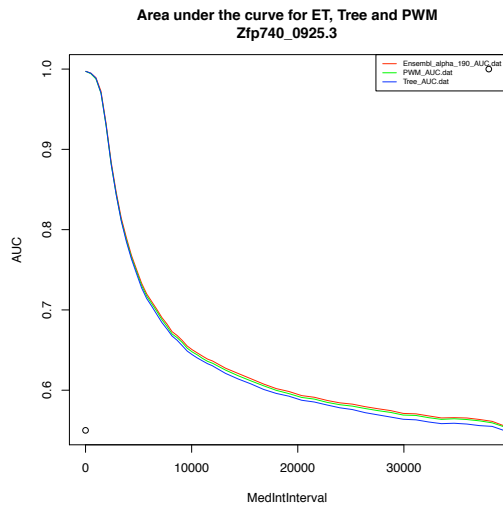




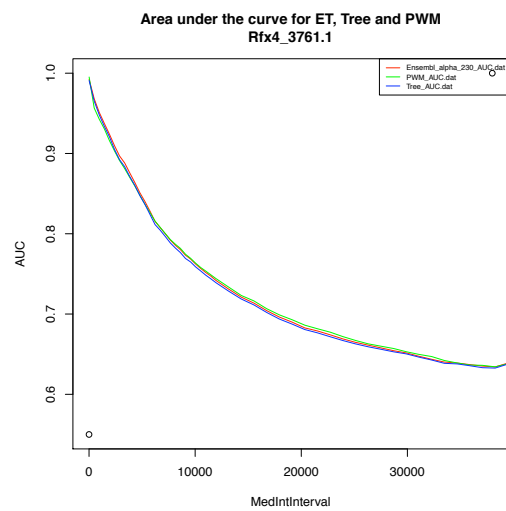
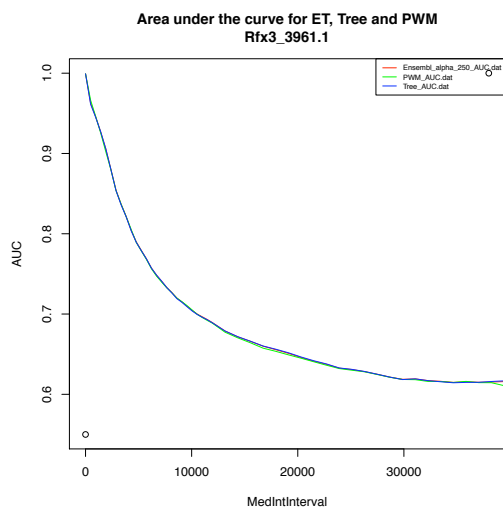
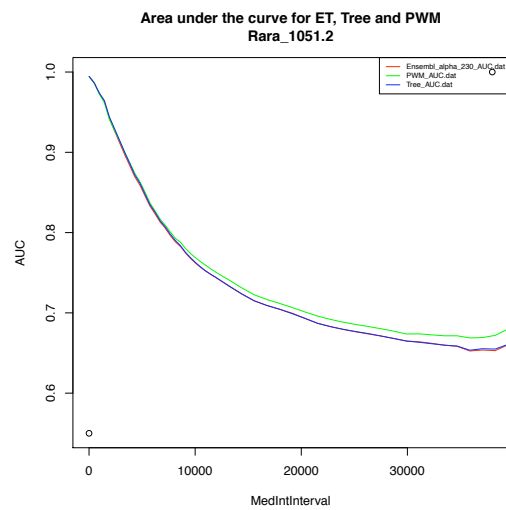
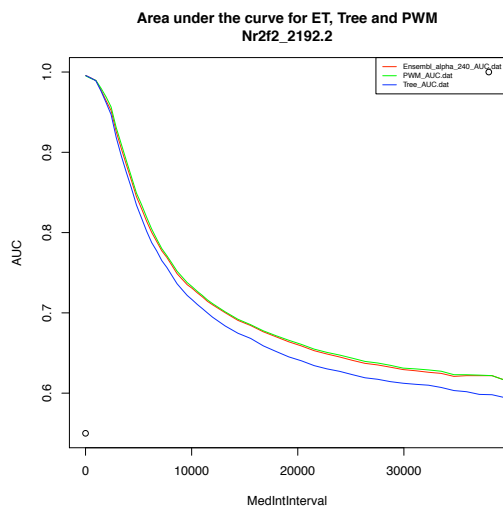
PWM better than Tree, improvement by ET model

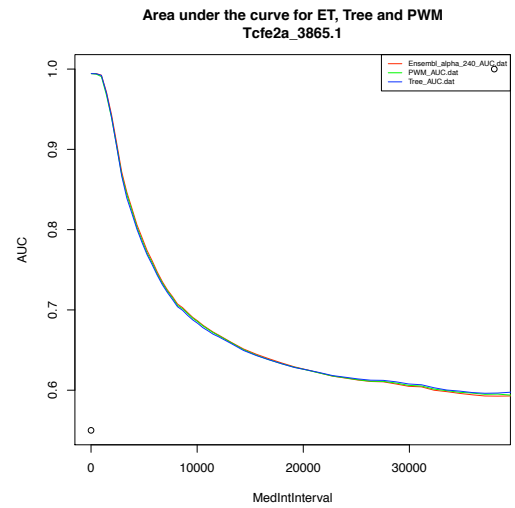
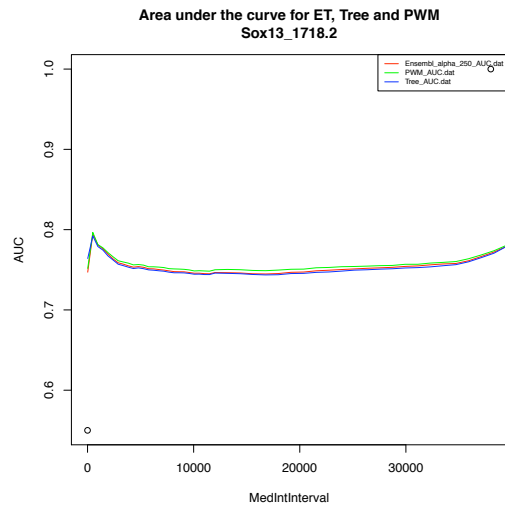






PWM better than Tree, no improvement by ET model





8.6. Numerical Instabilities in ET- α -Training

The original idea of a pilot study to improve the way to select the best α by a different objective function, was a maximum-log-likelihood approach. In this regard the ET model has been systematically tested for α values between 0 and 1,000 and it was found that the optimal α seems to be infinitively high. To recall, the role of α , specifically in combination with the exponential function, is to amplify or enhance the difference of the MI between the different parent-child relationships within a tree to reweight positional interdependencies. The pilot study revealed, that at a certain α -value (>250), dependent on the TF considered, the system collapses. How does this come? The explanation is, that at this collapsing point the enhancement of α leads to a MI difference which is infinitive, resulting in a negative value for the determinant of the Laplacian matrix Q . In general the determinant of a matrix can be negative if the matrix holds a high amount of “0”s. This is not directly the case in the Laplacian matrix, but the calculation of β by $e^{(\alpha \cdot MI)}$ provokes a similar effect.

Considering the Laplacian matrix Q at the collapsing point, with matrix cell values determined by β , being in turn determined by $e^{(\alpha \cdot MI)}$, the cell value correlating with the edge assigned with the highest MI is infinitively high, while the others are not. This situation is as extreme as if a matrix was equipped in one or two cells with a “1” and in the remaining ones with “0”. Thus, the determinant of this matrix gets negative, resulting in non-sense likelihoods for the ET model.

8.7. Number of SNPs per Peak

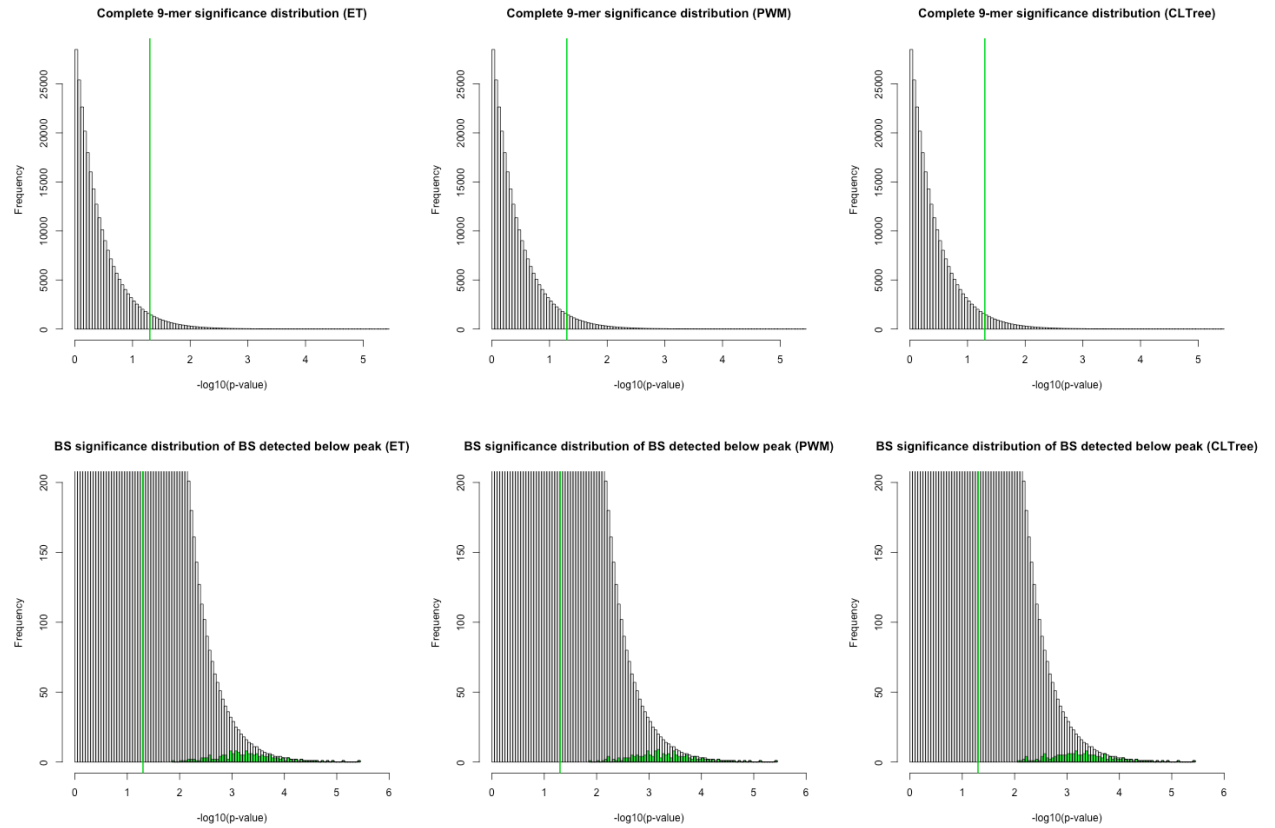
Most of the peaks are co-localized with only one SNP. However a tiny amount is highly affected by genetic variants, with up to 7 SNPs within a range of 200bp (length of a peak center) as one can take from the table below. The relative amount of peaks affected by a certain number of different SNPs is over all TF within the different groups stable. In the group of common peaks, for each TF ~80% are co-localized with one SNP, ~15% with 2 SNPs, ~2.5% with 3 SNPs, between 0.5 and 1% of peaks with 4 SNPs. More than 4 SNPs are to observe very rarely and only sporadic for different TFs. For example for SPI1 5 common peaks on the paternal genome are co-localized with 5 SNPs and only for SPI1 some peaks are detected, where both, the maternal and paternal peaks comprise 7 SNPs.

q		EGR1		IRF4		MAX		SPI1		SRF	
	#SNPs										
	s	M	P	M	P	M	P	M	P	M	P
SNPed Common Peaks	1	1,029	1,030	1,189	1,186	742	745	4,700	4,705	736	739
	2	180	184	213	217	150	148	865	858	139	136
	3	33	27	38	37	15	14	153	156	19	20
	4	11	12	7	8	7	6	39	39	4	4
	5	3	3	3	2	2	3	5	3	2	1
	6	0	0	1	1	1	1	3	5	0	0
	7	0	0	0	0	0	0	3	2	0	0
	8	0	0	0	0	0	0	0	0	0	0
Peaks detected only in the maternal genome	1	33		33		25		16		18	
	2	9		9		7		4		9	
	3	1		2		3		1		3	
	4	1		1		0		1		1	
	5	0		0		0		0		0	
	6	2		0		1		0		0	
	7	0		0		0		0		0	
	8	0		1		0		0		0	
Peaks detected only in the paternal genome	1		28		11		21		2		10
	2		11		4		4		0		5
	3		4		0		5		0		2
	4		0		1		0		1		1
	5		0		0		0		0		0
	6		0		0		0		0		0
	7		0		0		0		0		0
	8		0		0		0		0		0

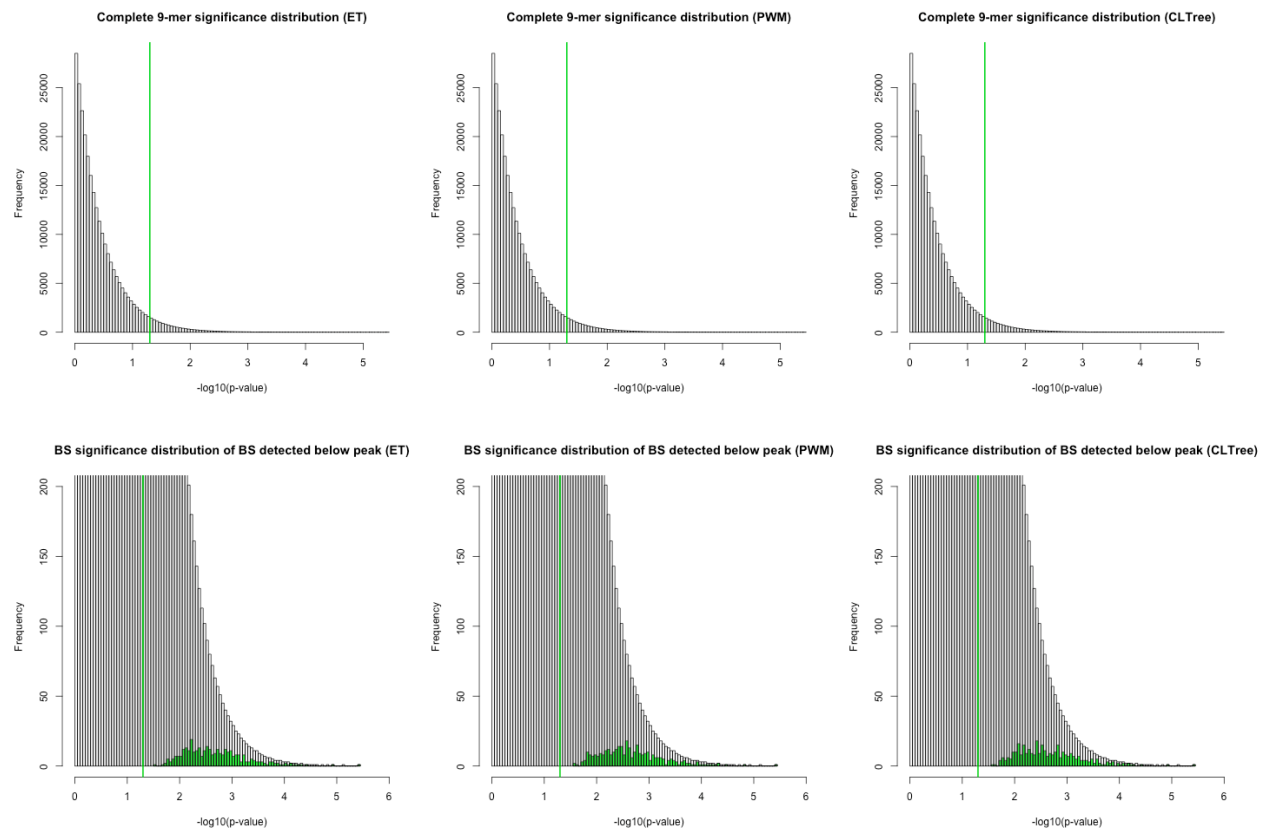
Within the group of peaks only detected in a single parent the relative amounts are slightly shifted towards a higher number of SNPs/peak. Here, ~70% of peaks comprise 1 SNP, between 20% (EGR1, only maternal) and 30% (SRF, only maternal) are co-localized with 2 SNPs, between 2% (EGR1, only maternal) and 17% (MAX, only paternal) hold 3 SNPs. Also here higher SNP numbers per peak are rare, for example for MAX a single peak is detected with 6 SNPs or for IRF4 one peak in the maternal genome with 8 SNPs.

8.8. BS Significance Distributions

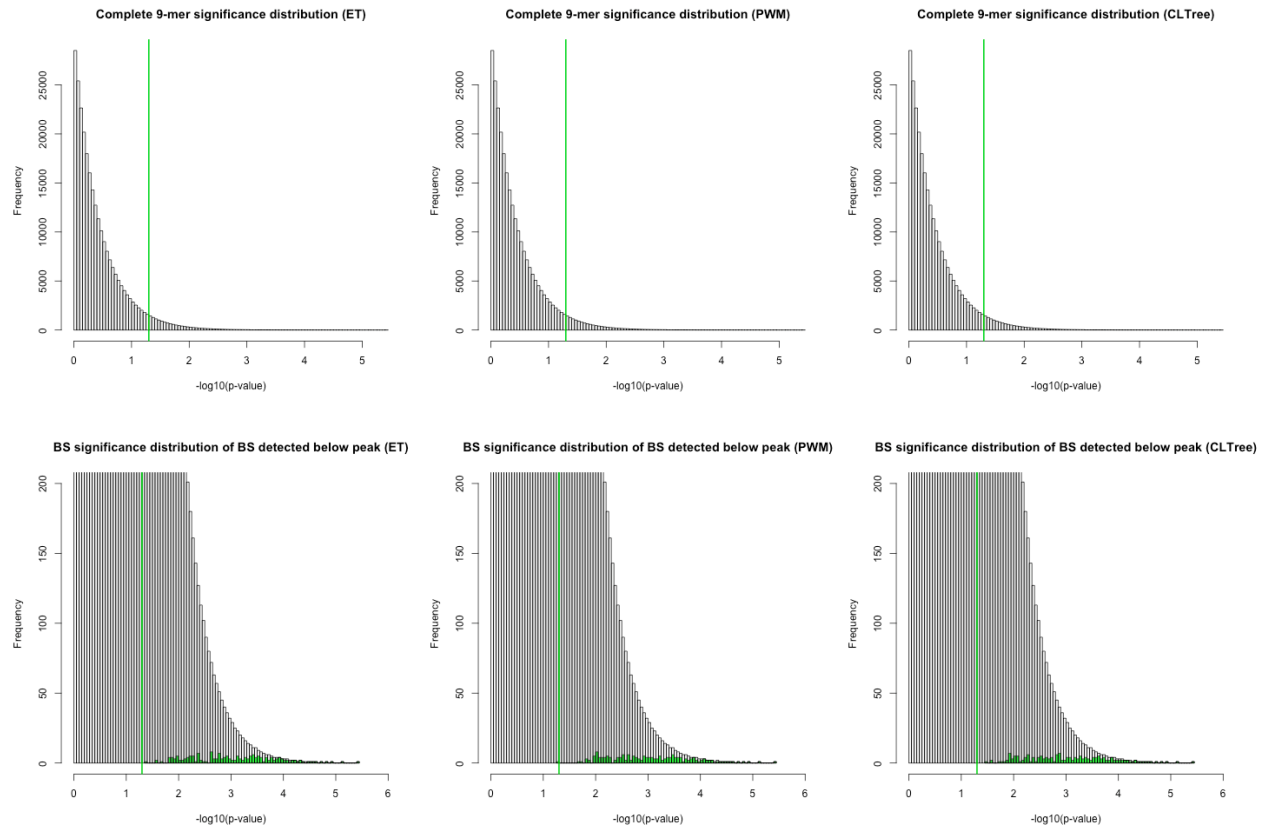
EGR1



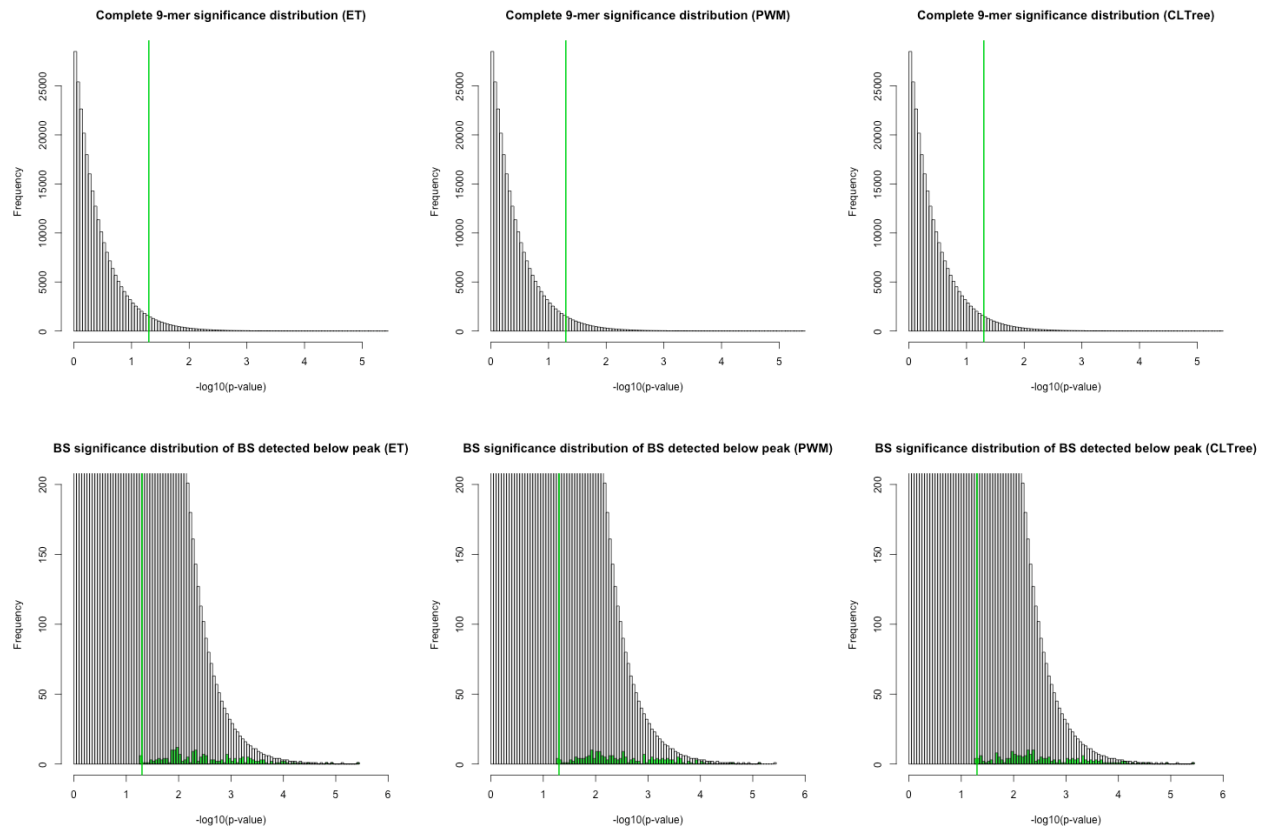
IRF4



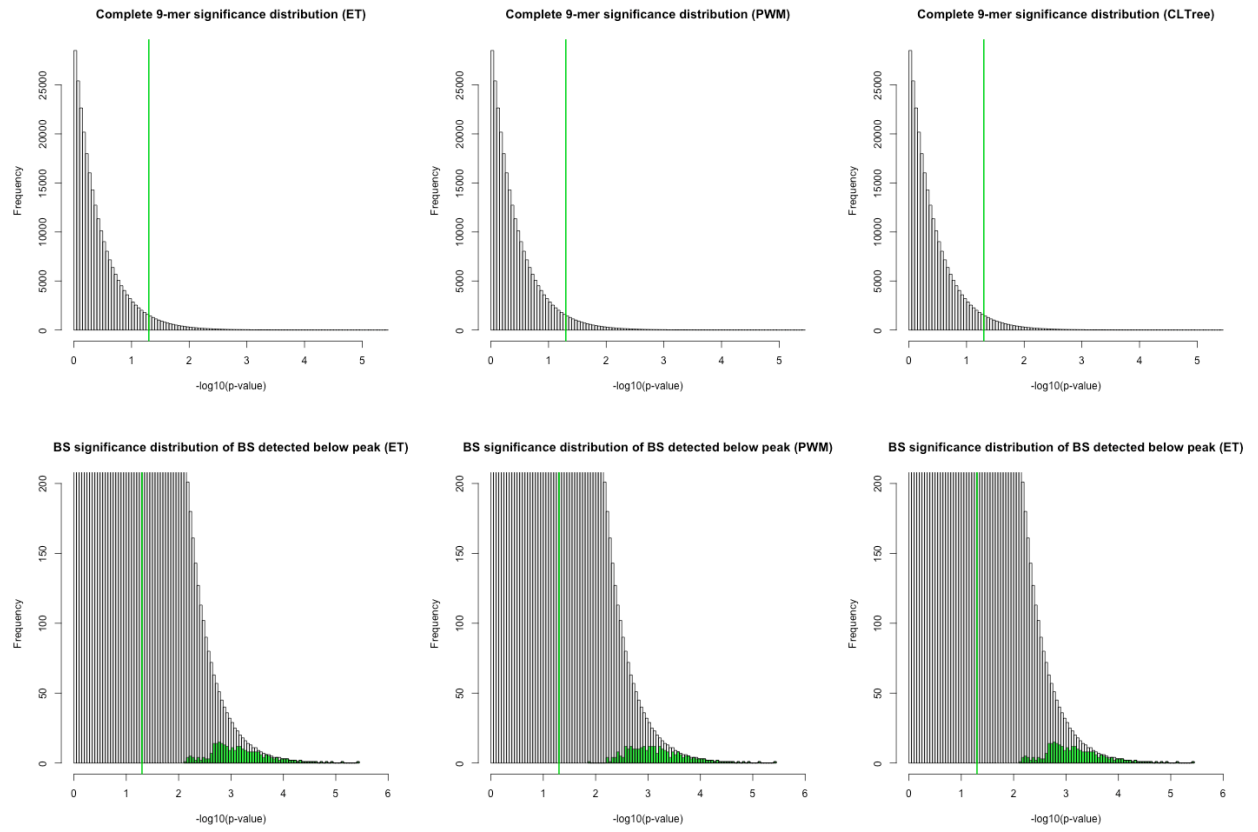
MAX



SRF

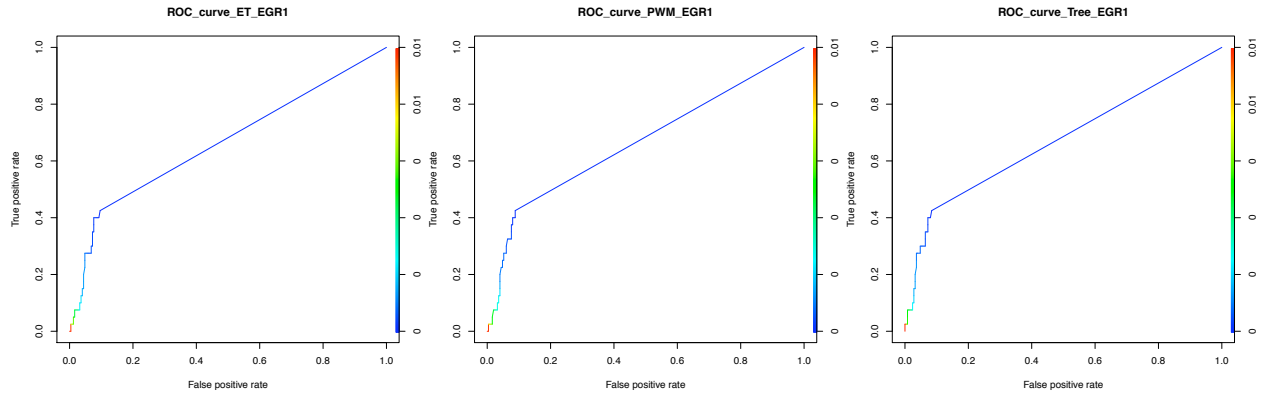


SPI1

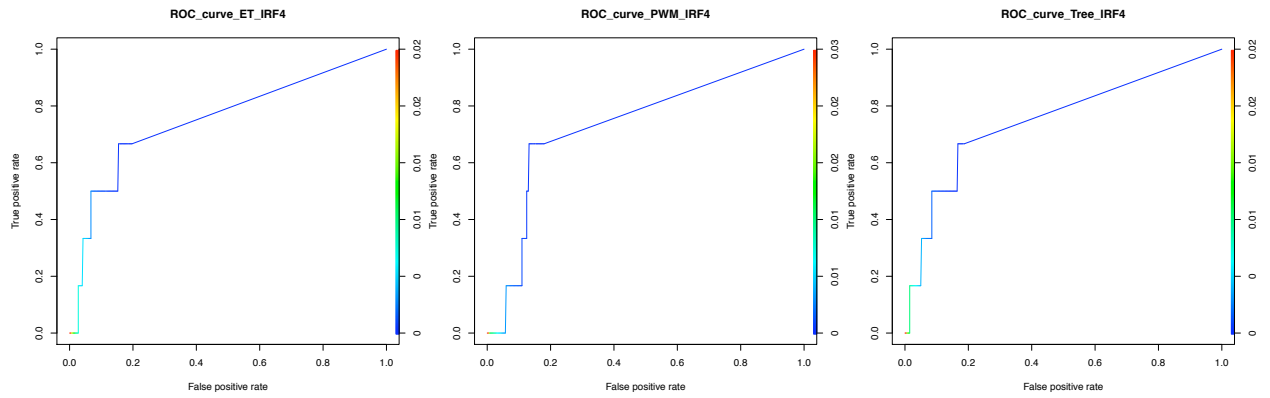


8.9. ROC curve Analysis for BS significance Difference Threshold Detection

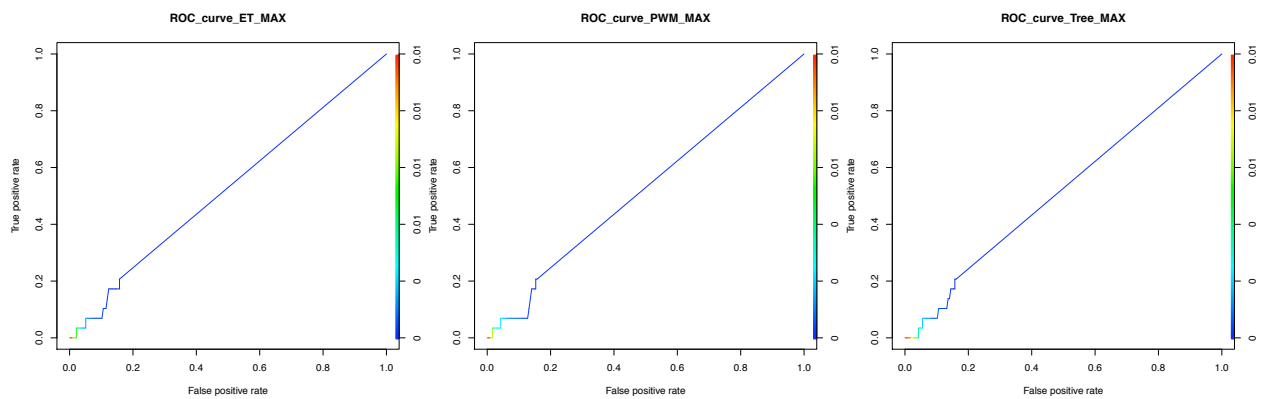
EGR1



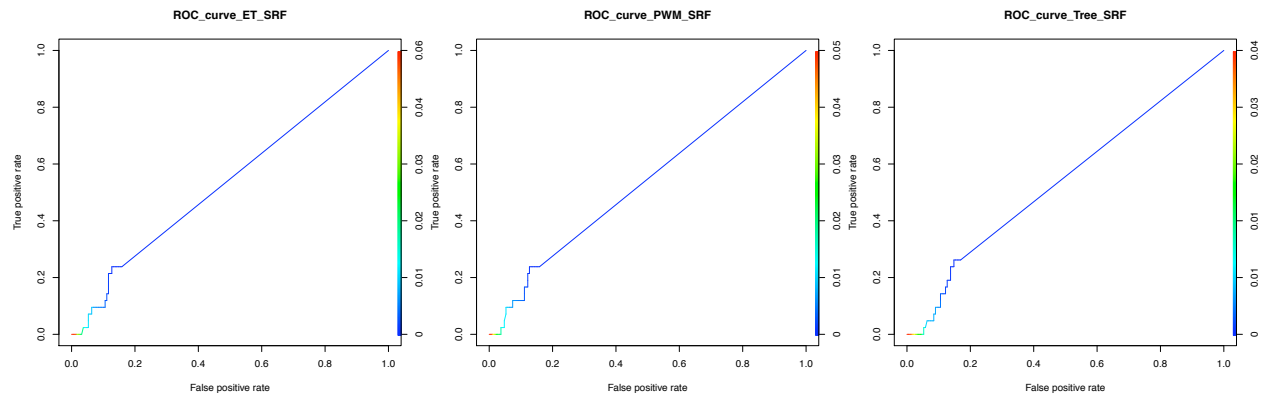
IRF4



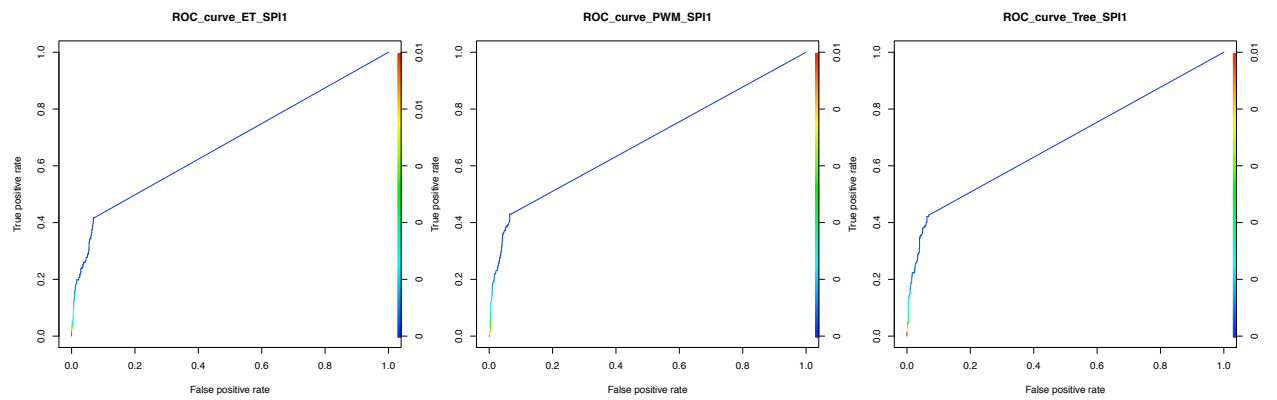
MAX



SRF



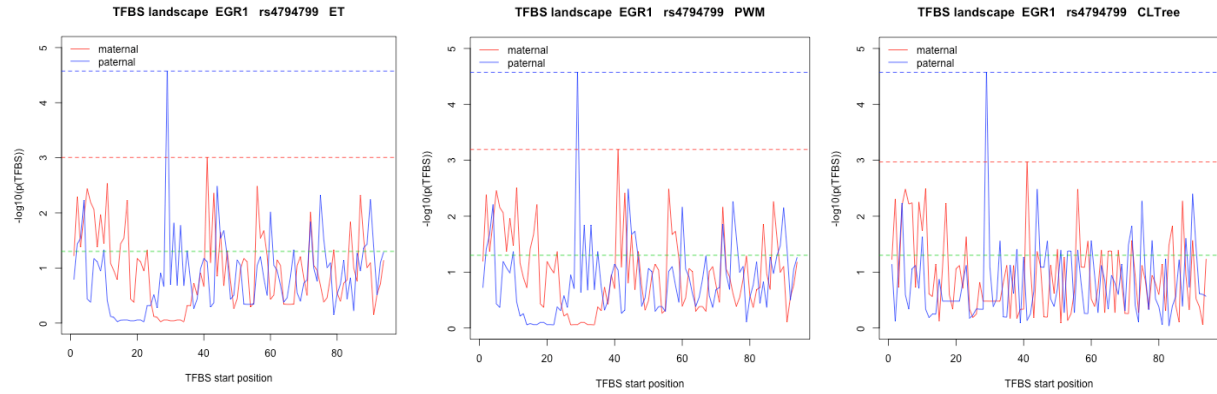
SPI1



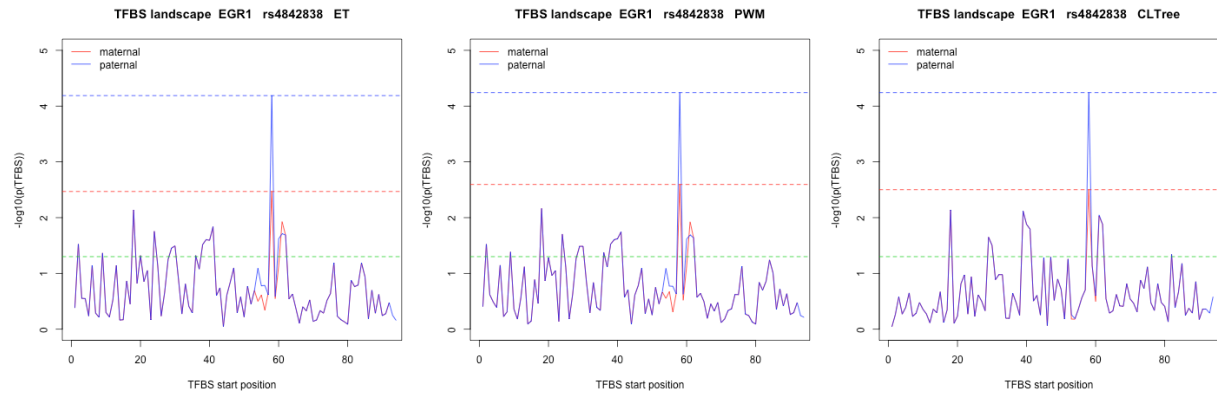
8.10. TFBS Landscapes Candidate SNP List

(sequences not trimmed, so that shifts due to different flanking sequences are visible)

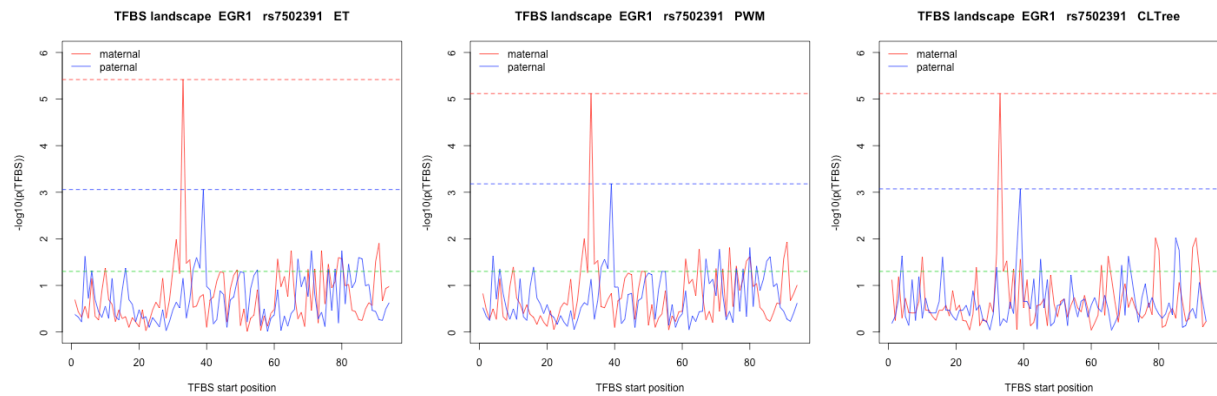
EGR1 rs4794799



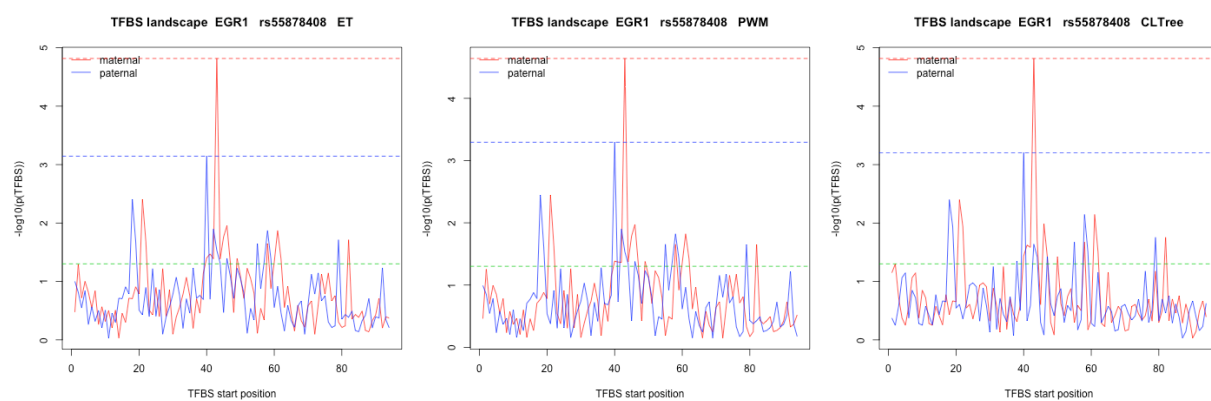
EGR1 rs4842838



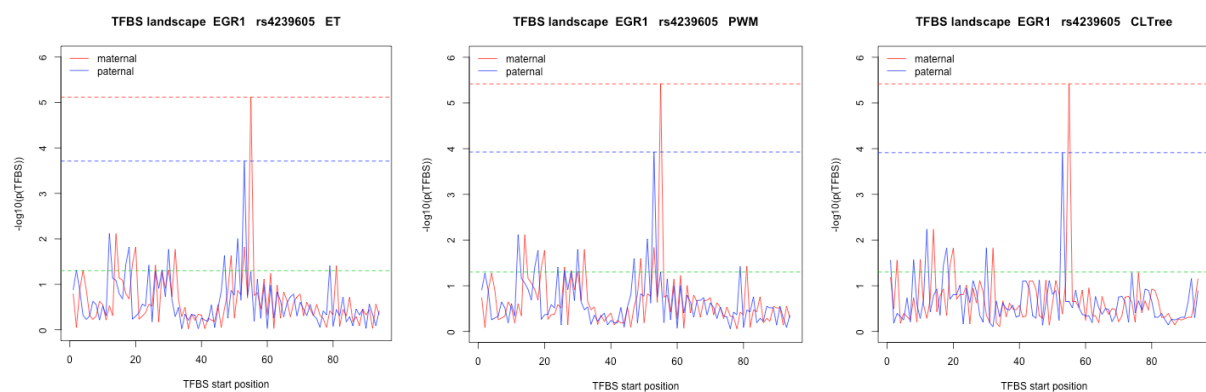
EGR1 7502391



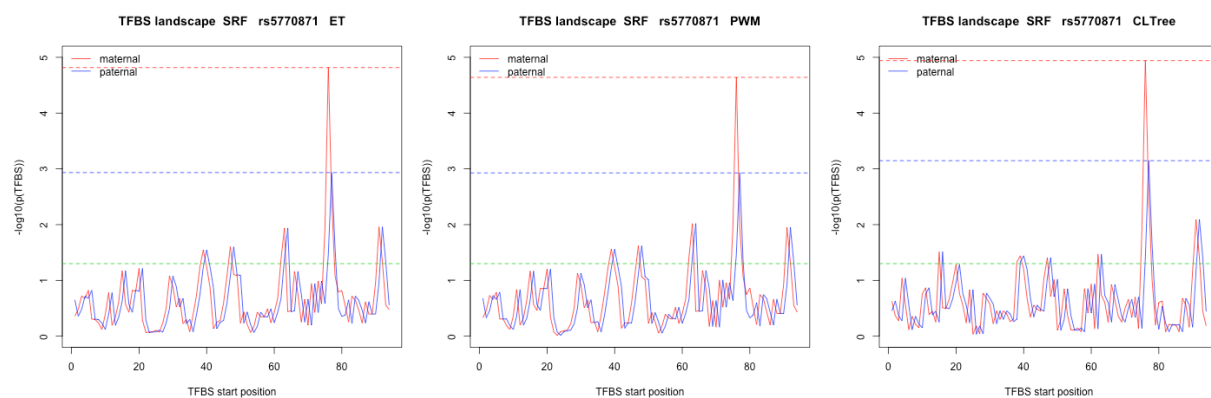
EGR1 55878408



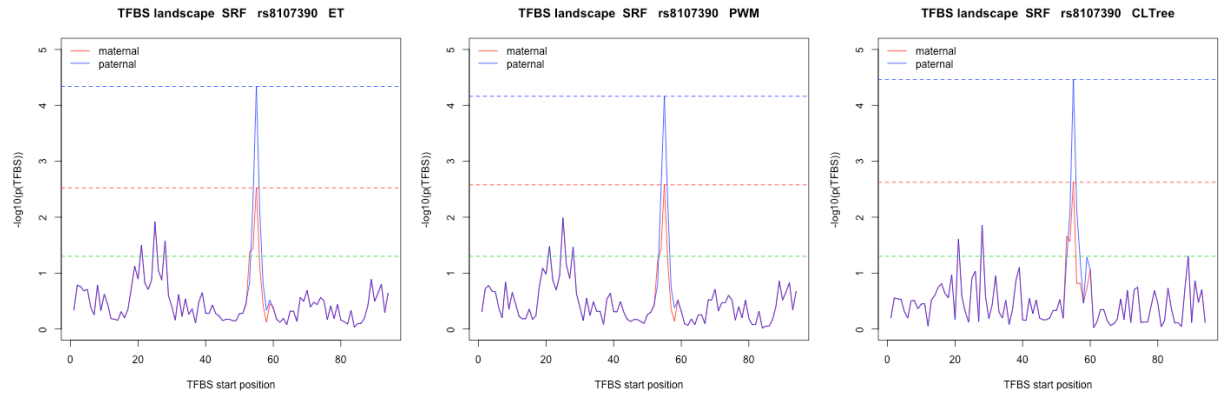
EGR1 rs4239605



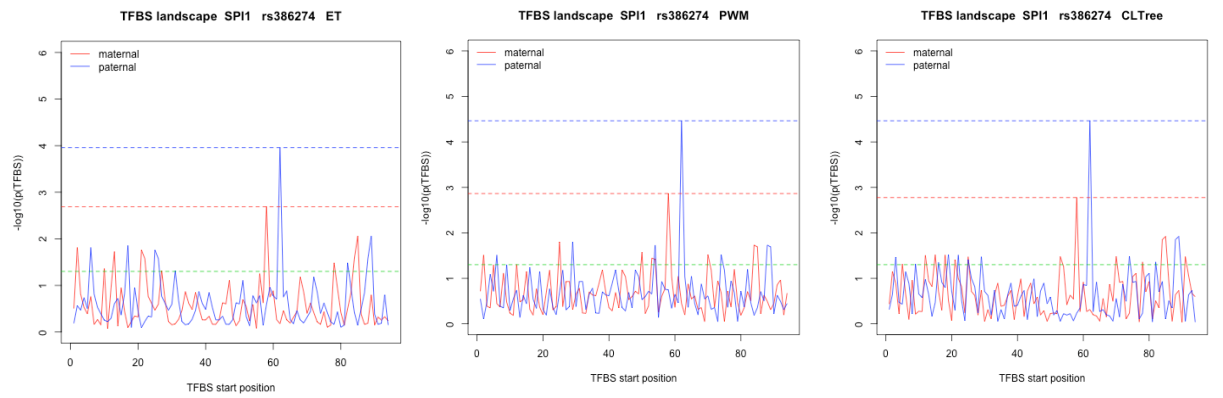
SRF rs5770871



SRF rs8107390



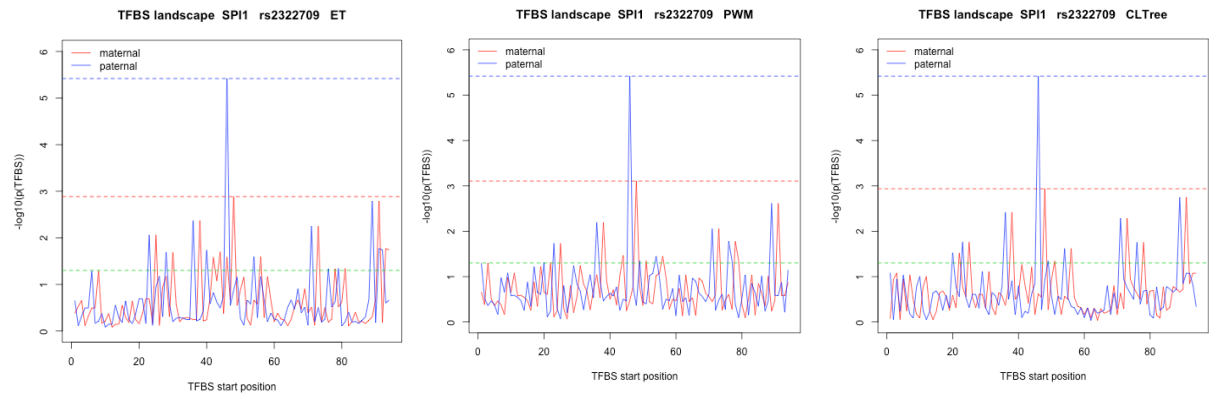
SPI1 rs386274



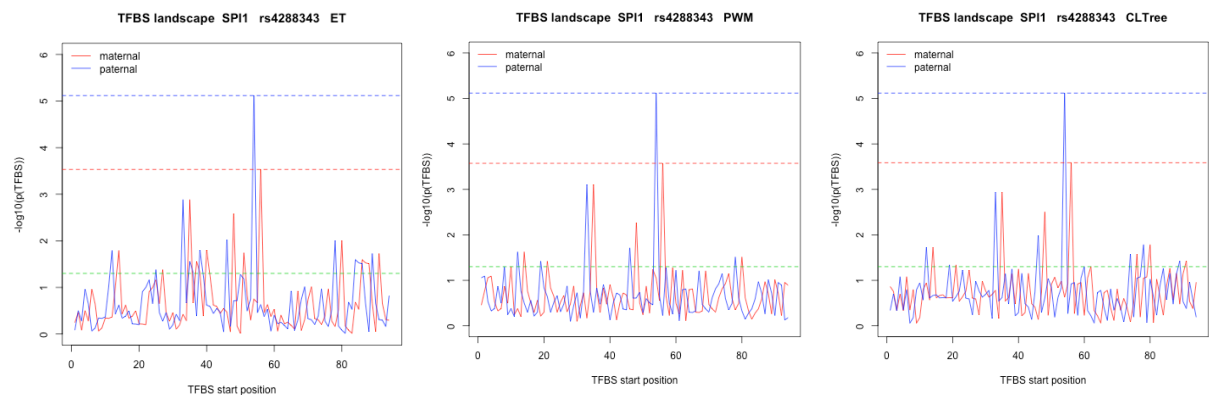
SPI1 rs2249189



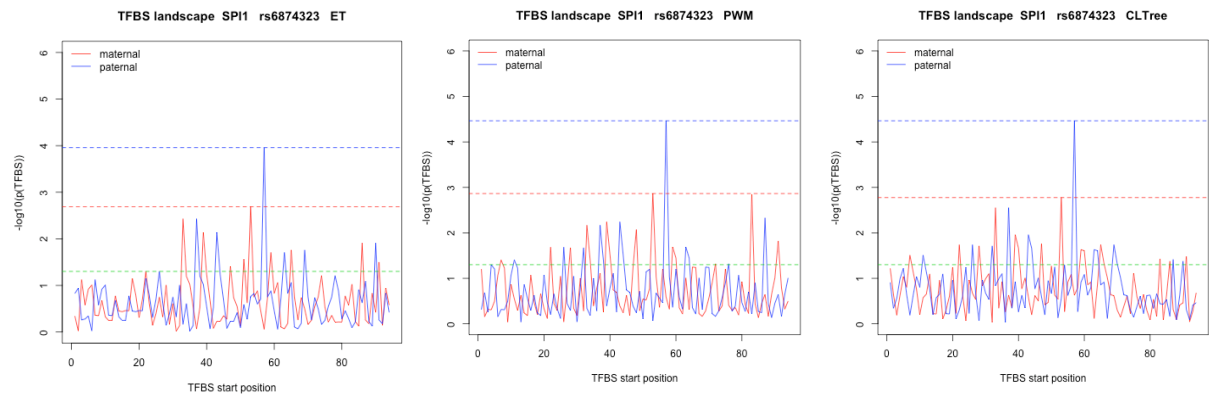
SPI1 rs2322709



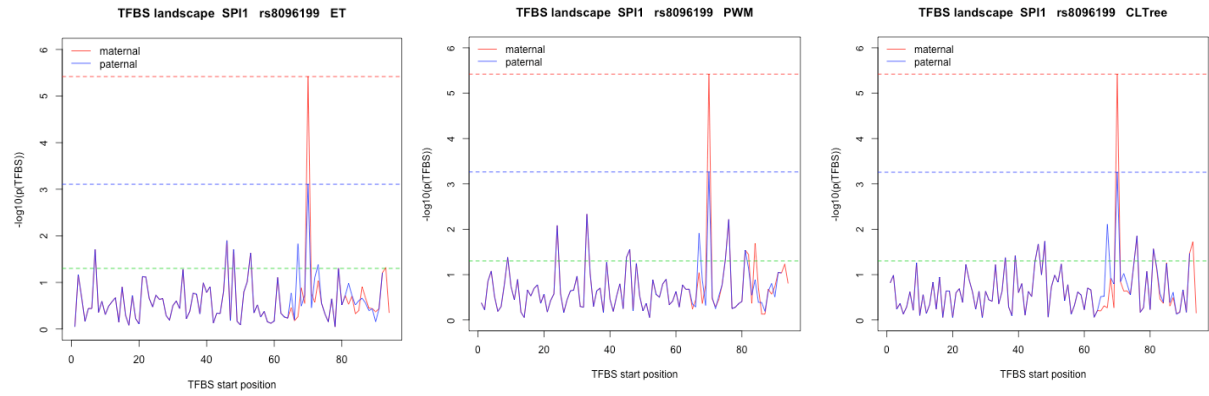
PI1 rs4288343



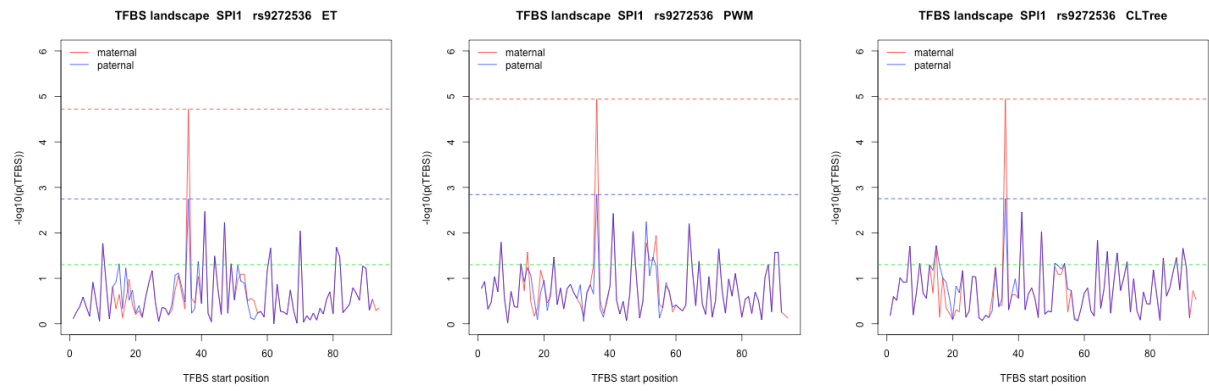
SPI1 rs6874323



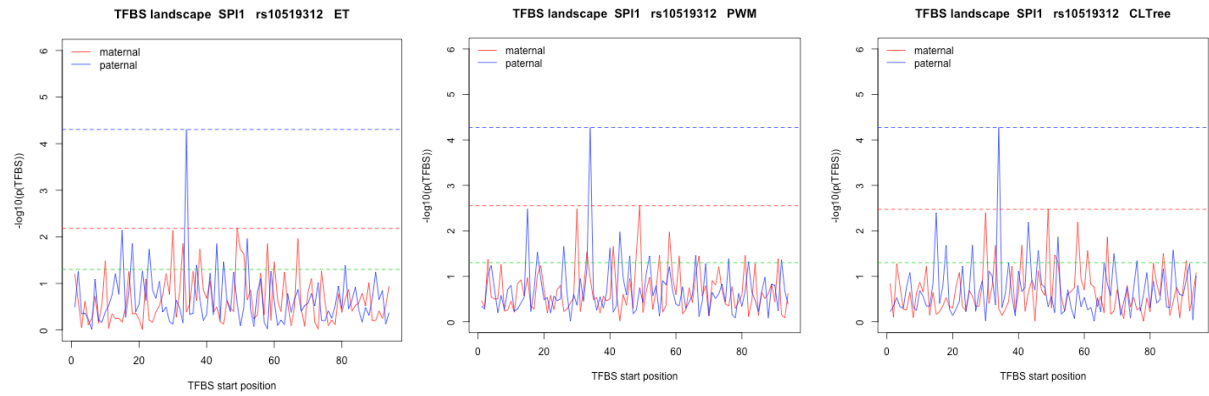
SPI1 rs8096199



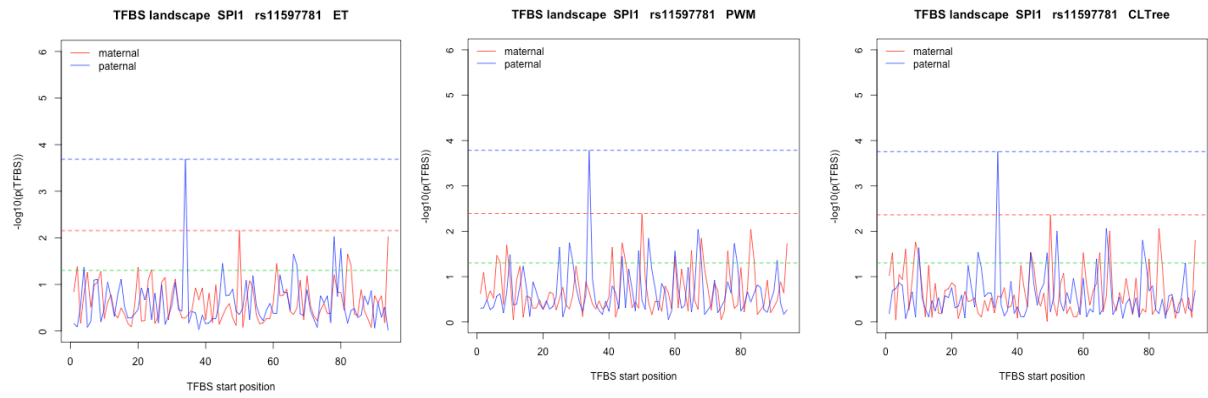
SPI1 rs9272536



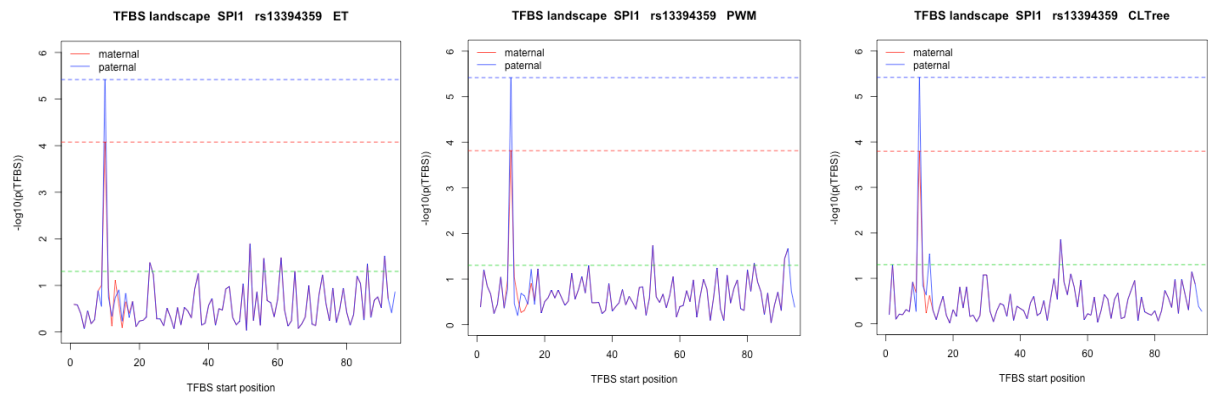
SPI1 rs10519312



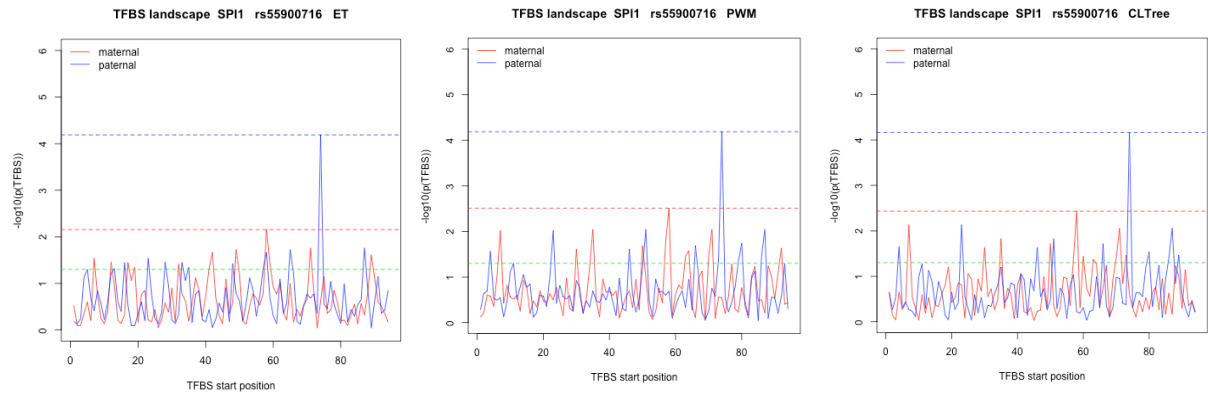
SPI1 rs11597781



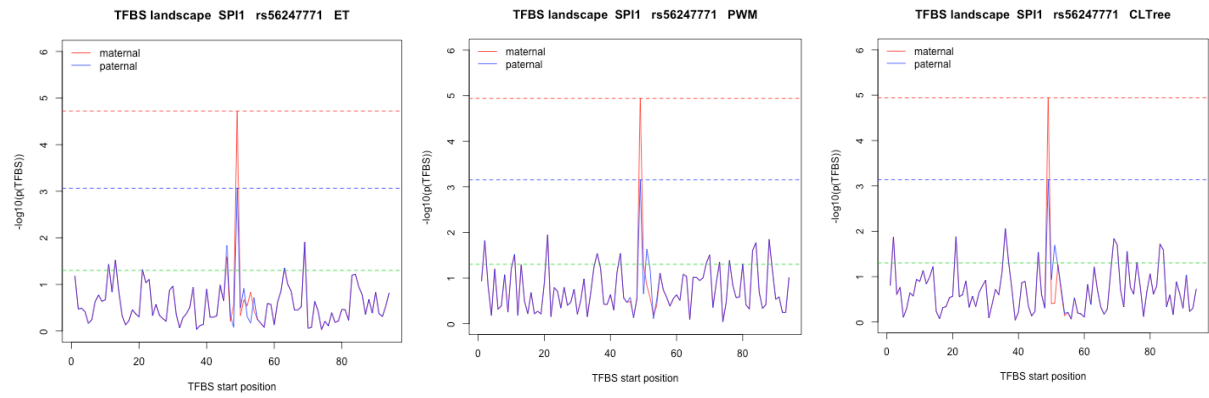
SPI1 rs13394359



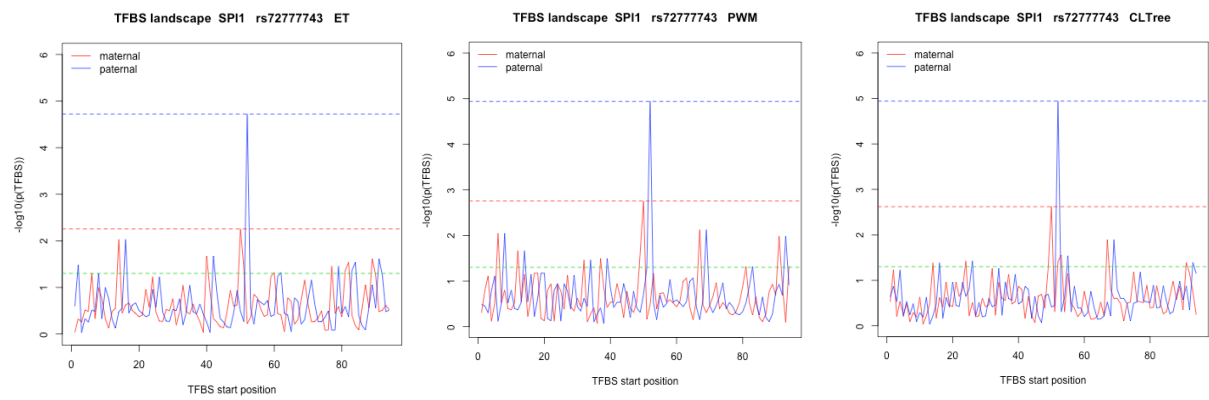
SPI1 rs55900716



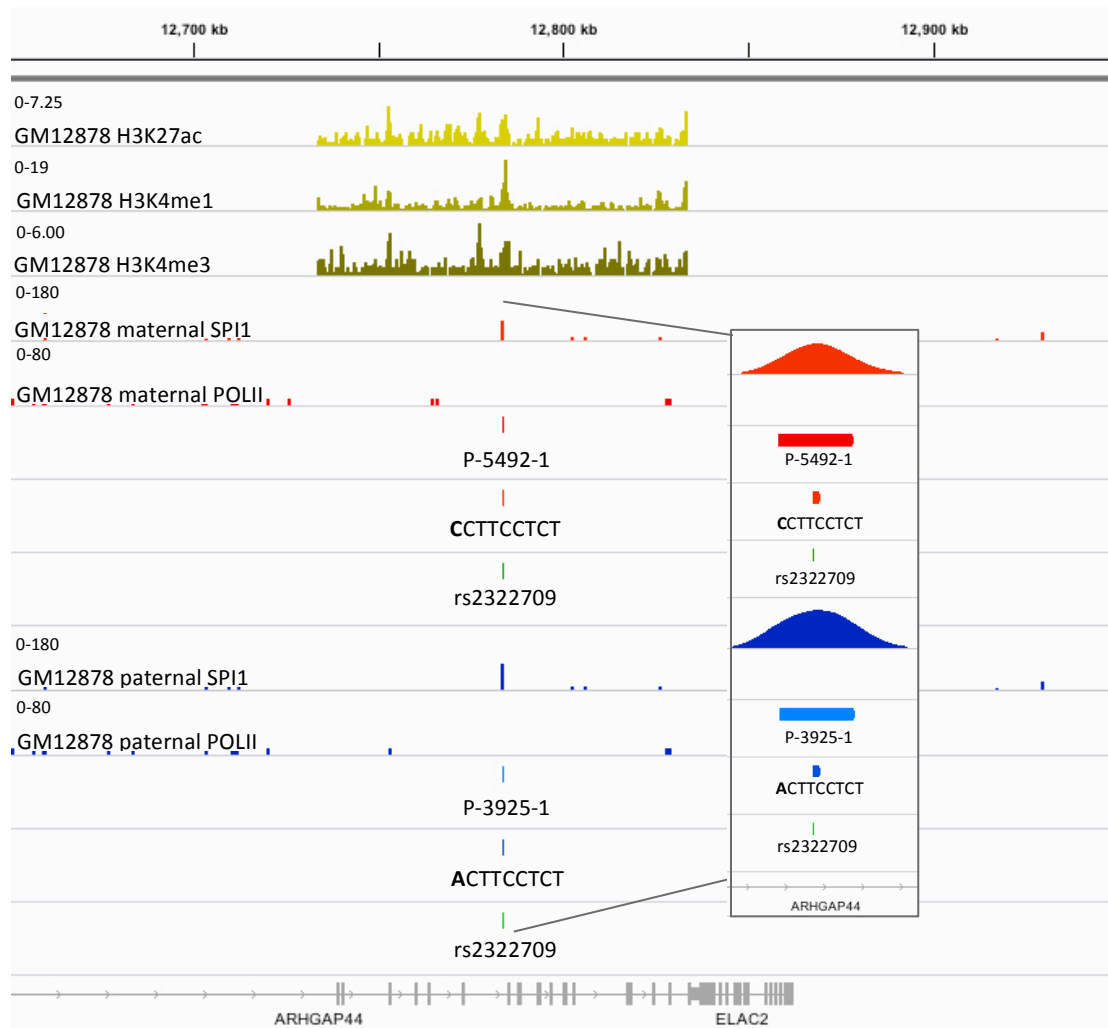
SPI1 rs56247771



SPI1 rs72777743



8.11. Top causative SNP Candidate for SPI1



CAAGCATAACATTGTCTAATCTGTACTTCAAGCTCTCTGGAGAACACCTTCCTCTTTTCACTCTGTGAGTCTGTTCTGTTGCCTTGTGTACAAGAACTCTT
AGCATAACATTGTCTAATCTGTACTTCAAGCTCTCTGGAGAACACACTTCCTCTTTTCACTCTGTGAGTCTGTTCTGTTGCCTTGTGTACAAGAACTCTTGA

↑
rs2322709

8.12. A relational Database System to Investigate Biological Plausibility

The database intends to integrate high-throughput technologies like next generation sequencing or microarray data with a special focus on gene expression regulation analysis. Furthermore, methods used to analyze those data are under high development, so that the comparability of different published datasets is not necessarily given. Accordingly, beside the location of regulatory elements and score information also the way of analysis needs to be considered for consistent data integration.

This database is not previewed to serve as a “collector” of links referring to different primary or secondary databases, but as a “collector” of the information stored behind those. Therefore a huge effort has to be made with regard to data pre-processing and consistency checks before any data set enters the database. The data pre-processing and insert functions are mainly implemented in Python, but also Perl, C, C++ and Unix shell scripts.

The selection of public data sources has been done “by hand” with following criteria:

- Data quality, structure and formats
- Possibilities for automation of data queries and provided interfaces

Furthermore, genome-wide *in silico* predicted, therefore ChIP-seq independent TFBS for all published PWMs (JASPAR, UniPROBE and TRANSFAC®) are integrated; tree-based screening results are still lacking until the training set issue will be solved. A consistency with regard to annotation is addressed by the integration of public ontologies for genes, cell types, anatomical systems, pathological stages and human diseases. The issue of database cross-reference consistency is considered in the data pre-processing as well as in the python-implemented data insertion steps. Furthermore, the database as itself, as a relational database system, provides, based on primary and secondary keys a last consistency check, before finally entering a dataset. However, an inconsistency during the final insertion leads to tremendous time losses depending on the data to be inserted, so that in most cases prevented by our extensive data pre-processing. As experimental data sources published and Inhouse generated microarray or ChIP- and FAIRE-seq experiments are considered, wherein extensions for further approaches like RNA-seq should be designed. It has been decided to not integrate any automated data processing steps for an on-the-fly analysis, since beside the data points as themselves, also the hypothesis, meaning the research question and biological intention, has to be integrated to assure comparability. Therefore the experimental design, preliminary quality assurance (e.g. antibody binding specificity for ChIP-seq experiments) and the applied analysis method have to be collected and evaluated.

8.13. Publications

Peer-reviewed publications:

Susanne E. Reinsbach, Petr V. Nazarov, Demetra Philippidou, Martina Schmitt, Anke Wienecke-Baldacchino, Arnaud Muller, Laurent Vallar, Iris Behrmann and Stephanie Kreis. "Dynamic regulation of microRNA expression following Interferon- γ -induced gene transcription", RNA Biology, 2012 July 1;9(7)

Francisco J Azuaje, Michaël Heymann, Anne-Marie Ternes, Anke Wienecke-Baldacchino, Daniel Struck, Danièle Moes and Reinhard Schneider. "Bioinformatics as a driver, not a passenger, of translational biomedical research: Perspectives from the 6th Benelux Bioinformatics Conference", J Clin Bioinforma. 2012 Mar 13;2:7.

John E, Wienecke-Baldacchino A, Liivrand M, Heinäniemi M, Carlberg C, Sinkkonen L. "Dataset integration identifies transcriptional regulation of microRNA genes by PPAR γ in differentiating mouse 3T3-L1 adipocytes", Nucleic Acids Res. 2012 Feb 7.

Pehkonen P, Welter-Stahl L, Diwo J, Ryyänen J, Wienecke-Baldacchino A, Heikkinen S, Treuter E, Steffensen KR, Carlberg C., "Genome-wide landscape of liver X receptor chromatin binding and gene regulation in human macrophages", BMC Genomics. 2012 Jan 31;13:50.

Peer-review publications in revision or submitted:

Martina J. Schmitt, Demetra Philipidou, Susanne Reinsbach, Anke Wienecke-Baldacchino, Christiane Margue, Iris Behrmann and Stephanie Kreis, "Interferon-gamma-induced activation of Signal Transducer and Activator of Transcription 1 (STAT1) up-regulates the tumor suppressing microRNA-29 family in melanoma cells", submitted to Cell Communication and Signaling in August 2012.

Merja Heinäniemi, Matti Nykter, Roger Kramer, Anke Wienecke-Baldacchino, Lasse Sinkkonen, Joseph Xu Zhou, Richard Kreisberg, Stuart A. Kauffman, Sui Huang and Ilya Shmulevich. "Gene pair signatures in cell type transcriptomes reveal lineage control", Nature Methods, in revision

Oral Presentations:

Merja Heinäniemi, Anke Wienecke-Baldacchino, "Data Integration and Mining in Context of Gene Regulation", AI Lecture Series – Data Mining Applications, 9th November 2010

Anke Wienecke-Baldacchino, "Biological Reasoning by massive Data Integration", Capita Selecta in complex Disease Analysis (CSCDA), Leuven (Belgium), 25-27 August 2010

Anke Wienecke, "The Regulatory SNP Project", LuciLinX Bioinformatics Symposium 2009, University of Luxembourg

Anke Wienecke, "Experimentally based in silico selection approach for candidate SNPs in Type 2 Diabetes", Journal Club, Institute for Genetic Epidemiology Göttingen, Germany, May 2008

Anke Wienecke, "Experimentally based in silico selection approach for candidate SNPs", Center of Excellence, University of Kuopio, Finland, April 2008

Poster (presenting/main author indicated with *):

Anke Wienecke-Baldacchino*, Merja Heinäniemi, Nikos Vlassis, "Detecting regulatory SNPs: A Bayesian approach" a) Life Science PhD Days, September 2011, University of Luxembourg
b) 1st International Systems Biomedicine Symposium, LCSB, Cine Belval, September 2011 Luxembourg

Elisabeth John*, Anke Wienecke-Baldacchino, Merja Heinäniemi, Carsten Carlberg and Lasse Sinkkonen, "Dataset integration identifies transcriptional regulation of microRNA genes by PPAR γ in mouse adipogenesis", Cell Signal-omics 2011, Luxembourg 26.-28.01.2011, New Conference Center Kirchberg

Janine Diwo*, Lynn Welter-Stahl, Sophie Losciuto, Anke Wienecke-Baldacchino, Carsten Carlberg, "LXR activation potentiates the IL-response in human macrophages", Life Science PhD Days, September 2010, University of Luxembourg

Wienecke-Baldacchino A.*, Heinäniemi M., Carlberg C., "D2K – data to knowledge – data integration for biological reasoning", 9th European Conference on Computational Biology, Ghent, Belgium, 26-29 September 2010

Lynn Welter-Stahl*, Janine Diwo, Anke Wienecke, Carsten Carlberg, "Liver X receptors modulate the differentiation of human macrophages" EMBO conference for Nuclear Receptor: from molecular mechanism to molecular medicine", 25-29 September 2009 Cavtat, Croatia

Merja Heinäniemi*, Matti Nykter, Lasse Sinkkonen, Anke Wienecke, Ilya Shmulevich, Carsten Carlberg, "Data integration across cell lineages to discover transcriptional regulatory networks", Sixth International Workshop on Computational Systems Biology, WCSB 2009, June 10-12, 2009, Århus, Denmark

Demetra Philippidou, Dirk Moser, Anke Wienecke, Iris Behrmann and Stephanie Kreis*, "Regulation of miRNA expression by cytokine-induced STATs", Keystone April, 2009 "The Biology of RNA Silencing", Victoria, Canada

Merja Heinäniemi*, Anke Wienecke*, Harald Grallert, Barbara Thorand, Thomas Illig, Heinz-Erich Wichmann, Christian Herder, Wolfgang Rathmann, Heike Bickeböller, Carsten Carlberg, "Omics resource for PPARs: application of quantitative *in silico* binding predictions in the identification of regulatory SNPs in type 2 diabetes", 4th EMBO Conference: From Functional Genomics to Systems Biology, EMBL Heidelberg, 15 - 18 November 2008

8.14. Computing Resources Used and Source Code Snippets

Hardware:

Mac OS X, 2x3 GHz Quad-Core Intel Xeon, 16GB 667 MHz DDR2 FB-DIMM, 4TB HD

Programming/Scripting Languages:

- Python (Version 2.7)
(necessary packages different from default: Numpy, Scipy, Pygraph)
- R (Version 2.11.1)
- Shell Scripting (bash)

Source Code

In the following only the main functions for the model training will be listed to get a kind of flavor. The complete code listing would go beyond the scope of this work in particular due to the special notation of python code (basically determining the code structure by indentation). All code can be requested from the author.

```

#----- calculate mutual information from single and paired marginals -----
def getMI(pos_pairs_norm, pos_single_norm):
    MI_dict = {}
    for i in pos_pairs_norm:
        MI = 0
        for j in pos_pairs_norm[i]:
            log_factor = pos_pairs_norm[i][j] / (pos_single_norm[i[0]][j[0]]*pos_single_norm[i[1]][j[1]])
            MI_temp = pos_pairs_norm[i][j] * math.log(log_factor, math.e)
            MI = MI + MI_temp
        if MI < 0:
            MI_dict[i] = 0
        else:
            MI_dict[i] = MI
    return MI_dict

#----- get the maximum spanning tree -----
def getMaxSpanTree(MI_dict, lenBS):
    '''
    {(7, 3): 0.00074755677006516422, (4, 7): 0.00069190891349291432, ...
    '''
    g = graph()
    for i in range(0, lenBS):
        g.add_node(i)
        MI_dict[(i, i)] = 0.0
    #print MI_dict
    for i in MI_dict:
        try:
            g.add_edge(i, -MI_dict[i]+10) #the algorithm does not like values between 0 and -1 !!!!
        except:
            #print sys.exc_info()
            continue
    tree = minimal_spanning_tree(g)
    return tree

#----- calculate probability score of CLTree model -----
def calculateTreePx(sequence, tree, norm_matrix_pair, norm_matrix_single):
    #print tree
    #print norm_matrix_pair
    #print norm_matrix_single
    #print sequence
    p_x = 0
    for i in tree:
        if tree[i] != None:
            p_x_factor = math.log(norm_matrix_pair[(i, tree[i])][sequence[i]+sequence[tree[i]]]/norm_matrix_single[tree[i]][sequence[tree[i]]], math.e)
        else:
            #print i, sequence[i], norm_matrix_single[i][sequence[i]]
            p_x_factor = math.log(norm_matrix_single[i][sequence[i]], math.e)
        p_x = p_x + p_x_factor
    return math.exp(p_x)

```

```

#----- calculate probability score of PWM model -----
def calculatePWMPx(seq, PWM_tf):
    #print PWM_tf
    tf_prob = 0
    for j in range(len(seq)):
        tf_prob = tf_prob + math.log(PWM_tf[j][seq[j]], math.e)
    p_sx = math.exp(tf_prob)
    return p_sx

#----- calculate Q-matrix for ET model -----
def get_Q_BETA(MI, alpha):
    '''
        0          1          2          3          4
    0  +sum(row) -MI(0,1) -MI(0,2) -MI(0,3) -MI(0,4)
    1  -MI(1,0) +sum(row) -MI(1,2) -MI(1,3) -MI(1,4)
    2  -MI(2,0) -MI(2,1) +sum(row) -MI(2,3) -MI(2,4)
    3  -MI(3,0) -MI(3,1) -MI(3,2) +sum(row) -MI(3,4)
    4  -MI(4,0) -MI(4,1) -MI(4,2) -MI(4,3) +sum(row)
    '''

    #print MI, max(max(MI.keys()))

    max_index = max(max(MI.keys()))
    mx = make_matrix(max_index+1, max_index+1)
    mx_temp = make_matrix(max_index+1, max_index+1)

    for i in MI:
        mx[i[0]][i[1]] = -math.exp(float(alpha) * MI[i])
        mx_temp[i[0]][i[1]] = math.exp(float(alpha)*MI[i])
        #mx[i[0]][i[1]] = -float(alpha) * MI[i]
        #mx_temp[i[0]][i[1]] = float(alpha)*MI[i]

    for row in range(0,len(mx_temp)):
        for column in range(0,len(mx_temp)):
            if row == column:
                diagonal = sum(mx_temp[row]) #possible, since the 0/0, 1/1...values are on zero by default
                #print mx_temp[row], sum(mx_temp[row])
                mx[row][column] = diagonal

    Q = copy.deepcopy(mx)
    print Q
    print len(Q)

    #cut out 1 row and column
    Q.pop(len(mx)-1)
    for i in Q:
        i.pop(len(mx)-1)

    detQBeta = numpy.linalg.det(Q)
    return detQBeta

```

```

#----- calculate w for ET model -----
def get_wx(seq, pos_pairs_norm, pos_single_norm, MI, alpha):

    #print pos_single_norm.keys()

    max_index = max(max(MI.keys()))
    mx = make_matrix(max_index+1, max_index+1)

    for i in range(len(seq)):
        for j in range(len(seq)):
            if i == j:
                continue
            else:
                #print i, j, seq[i], seq[j], pos_pairs_norm[(i,j)][seq[i]+seq[j]], pos_single_norm[i][seq[i]], pos_single_norm[j][seq[j]], MI[(i,j)]
                #mx_factor = (pos_pairs_norm[(i,j)][seq[i]+seq[j]]/(pos_single_norm[i][seq[i]]*pos_single_norm[j][seq[j]]))*float(alpha)*MI[(i,j)]
                mx_factor = (pos_pairs_norm[(i,j)][seq[i]+seq[j]]/(pos_single_norm[i][seq[i]]*pos_single_norm[j][seq[j]]...
                                                                    ...[seq[j]]))*math.exp(float(alpha)*MI[(i,j)])

                mx[i][j] = - mx_factor

    for row in range(0,len(mx)):
        for column in range(0,len(mx)):
            if row == column:
                diagonal = sum(mx[row]) #possible, since the 0/0, 1/1...values are on zero by default
                #print mx_temp[row], sum(mx_temp[row])
                mx[row][column] = -diagonal

    #print mx

    Q = copy.deepcopy(mx)

    #cut out 1 row and column
    Q.pop(len(mx)-1)
    for i in Q:
        i.pop(len(mx)-1)

    detQwx = numpy.linalg.det(Q)
    return detQwx

```

```

#----- Main function for ET model training and validation -----
#python Main_Steme_EnsembleTree_AUC_weighted.py Uniprobe/Myb/TreeMotifRun/motif-00/motif-sites.txt Uniprobe/Myb/Myb_1047.3_v1_deBruijn.txt
Uniprobe/Myb/Myb_1047.3_v2_deBruijn.txt Uniprobe/Myb/ 220

def __main__():
    steme = sys.argv[1]
    filename_train_uniprobe = sys.argv[2]
    filename_val = sys.argv[3] #pbm dataset array 2
    workingDirectory = sys.argv[4]
    res_filename_temp = "Screening_Array_2"
    alpha = sys.argv[5]
    res_filename = workingDirectory + "/" + res_filename_temp + "_alpha_" + alpha + ".txt"

    #os.system("mkdir -p " + workingDirectory)
    #os.system("mkdir -p " + workingDirectory + "/RocCurveData")

    ###calculate correction factor for distanz of 9-mer
    print "loading correction factor for glass slide distance..."
    CorrFactor = CallPositionCorrectionR(workingDirectory, filename_train_uniprobe)
    print "..done"

    print "reading steme data:", steme
    stemeData = readStemeOutput(steme)
    data_train = readFile(filename_train_uniprobe)

    normInputTrain = normalizeInputWeights(data_train)

    weightedSteme = generateWeightsForSteme(stemeData, normInputTrain)
    correctedSteme = correctStemeWeights(weightedSteme, CorrFactor)

    print "calculating matrixes"
    pos_pairs, norm_matrix_pair, pair_header, nt_pair = getWeightedPairCountsSTEME(correctedSteme)
    pos_single_tree, single_header, n = getSingleCountsFromPair(norm_matrix_pair, 9)

    lenBS = len(correctedSteme[0]["Seq"])
    #print "calculating MI"
    MI = getMI(norm_matrix_pair, pos_single_tree)
    detQbeta = get_Q_BETA(MI, alpha)
    Tree = getMaxSpanTree(MI, lenBS)

    # "...screening variable region of Uniprobe validation set"
    subMain_ScreenUniprobeVarSeq(filename_val, res_filename, Tree, pos_single_tree, detQbeta, MI, norm_matrix_pair, lenBS, CorrFactor, alpha)

    ##AUC calculation
    print "starting AUC processing"
    numberOfIterations = 50

```

```

data_val = readFile(filename_val)
print "number of iterations:", numberOfIterations

model = ["Ensembl"]
screenResult = res_filename

stepSize = []

start = len(data_val)-1
stop = int(start-(start/1.4))
stepsize = (start-stop)/(numberOfIterations/2)
stepSize.append((start, stop, stepsize))
start = stop
stop = 1
stepsize = (start-stop)/(numberOfIterations/2)
stepSize.append((start, stop, stepsize))

print "respective step sizes:", stepSize

for mm in model:
    for j in stepSize:
        start = j[0]
        iterator = j[2]
        stop = j[1]
        while start>= stop:
            getRocCurveData = "python getRocCurveDataEnsTreeOnly.py " +screenResult +" " +str(start) +" " +mm +" > " +workingDirectory
+ "/RocCurveData/temp_RocCurve_" +str(start) +"_" +mm +"_alpha_" +alpha +".dat"
            print getRocCurveData
            "\n*****"
            os.system(getRocCurveData)
            start = start-iterator
            print "\n*****"
            getRocCurves = "Rscript getROCurvesAlpha.R " +workingDirectory +"/RocCurveData/" +" " +mm +" " +alpha
            print getRocCurves
            print "\n*****"
            os.system(getRocCurves)
            print "\n*****"
            getAUCData = "Rscript RocPythonAlpha.R " +mm +" " +workingDirectory +"/RocCurveData/" +alpha
            print getAUCData
            print "\n*****"
            os.system(getAUCData)
            print "\n*****"
            print "make AUC-graphs..."
            getAUCGraph = "Rscript AUC_graph_EnsTreeOnly_alpha.R " +workingDirectory +"/RocCurveData/" +alpha
            print getAUCGraph
            print "\n*****"
            os.system(getAUCGraph)

print "FINISHED..."

```