Phenotype-related regulatory element and transcription factor identification via phylogeny-aware discriminative sequence motif scoring

Linking complex phenotypic changes
to the divergence of the gene-regulatory landscape

DISSERTATION

zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat)

vorgelegt an der
Technischen Universität Dresden
Fakultät Informatik

eingereicht von
BJÖRN LANGER
Dipl.-Math., Dipl.-Inf.
geboren am 31. Mai 1987 in Zittau

Tag der Einreichung: 3. November 2017

GUTACHTER:

Prof. Ivo F. Sbalzarini, Center for System Biology Dresden (CSBD)

Prof. Peter F. Stadler, Universität Leipzig

Contents

LIST OF ABBREVIATIONS 7

LIST OF FIGURES IX

LIST OF TABLES XI

ABSTRACT XIII

ACKNOWLEDGEMENTS XV

1 INTRODUCTION 1

1.1 FORWARD GENOMICS 4

2 REFORGE 7

2.1 COMPUTING TF BINDING AFFINITY TO A SEQUENCE 8
2.1.1 STUBB'S SCORING APPROACH 8
2.1.2 SCORING BIAS FROM DIFFERING LENGTHS AND GC-CONTENTS 10
2.1.3 SCORING METHOD CORRECTION 11
2.2 SCORING MULTIPLE PHYLOGENETICALLY DEPENDENT SEQUENCES 15
2.3 FROM BRANCH SCORES TO ELEMENT RANKS 16
2.4 VERIFICATION ON SYNTHETIC DATA 19
2.4.1 CRE EVOLUTION SIMULATION SETUP 19
2.4.2 REFORGE'S PROOF OF CONCEPT 23
2.4.3 METHOD OPTIMIZATION AND ASSESSMENT ON DIFFERENT DATASETS 24
2.5 VALIDATION ON BIOLOGICAL DATA 29
2.5.1 VISION-IMPAIRMENT PHENOTYPE 29
2.5.2 REFORGE IDENTIFIES PUTATIVE VISION-RELATED REGULATORY ELEMENTS 31
2.5.3 SCORING METHOD COMPARISON 33
2.5.4 COMPARISON TO FORWARD GENOMICS 37
2.6 SUMMARY & DISCUSSION 39

3 TFFORGE 41

3.1 TRANSCRIPTION FACTOR LIBRARY 42
3.1.1 PWM TO ENSEMBL ID MAPPING 43
3.1.2 MOTIF SIMILARITY AND CLUSTERING 43
3.2 TFFORGE'S PROOF OF CONCEPT 44
3.3 TECHNICAL AND METHODOLOGICAL VARIANTS 44
3.3.1 SEQUENCE SCORING METHOD 45
3.3.2 TF RANKING METHOD 48
3.4 BIOLOGICAL RELEVANT DATA VARIANTS 48
3.4.1 NUMBER OF REGULATORY ELEMENTS AND TRAIT-LOSS AGE 48
3.4.2 PLEIOTROPIC VS. SPECIFIC REGULATORY ELEMENTS 49

3.5 VALIDATION ON BIOLOGICAL DATA 50
3.6 SUMMARY & DISCUSSION 52

4 COMBINED APPLICATION 55

4.1 RUNTIME ANALYSIS 56
4.2 USE CASE - TOWARDS GENE-REGULATORY NETWORKS 58

5 SUMMARY 61

APPENDIX A.
FIGURES I

APPENDIX B.
SOFTWARE XV

APPENDIX C.
REFERENCES XVI

ERKLÄRUNGEN ZUR ERÖFFNUNG DES PROMOTIONSVERFAHRENS XIX

List of Abbreviations

AFTP ..... *ancestrally fixed transition probabilities*
ATAC ..... *assay for transposase-accessible chromatin*
bp ..... *base pairs*
ChIP ..... *chromatin immunoprecipitation*
ChIP-seq ..... *ChIP-sequencing*
CNE ..... *conserved non-coding element*
CRE ..... *cis-regulatory element*
CRX ..... *cone-rod homeobox protein*
DNA ..... *deoxyribonucleic acid*
EM ..... *expectation maximization*
evo-devo ..... *evolutionary developmental biology*
HMM ..... *hidden markov model*
HPC ..... *high-performance computing, high performance computing*
NRL ..... *neural retina leucine zipper*
NSC ..... *null score correction*
NTPC ..... *null transition probability correction*
PAX6 ..... *paired box 6*
PWM ..... *position weight matrix*
ROC ..... *receiver operating characteristic*
TF ..... *transcription factor*
TFBS ..... *transcription factor binding site*

List of Figures

Figure 1.1: Evolutionary model and assumptions behind Forward Genomics approach 5
Figure 2.1: Stubb's hidden Markov model 9
Figure 2.2: Sequence length dependency of Stubb's score 11
Figure 2.3: GC-content dependency of Stubb's score 12
Figure 2.4: Genome browser screenshot 15
Figure 2.5: Scheme of Forward Genomics' branch method 16
Figure 2.6: Fitch parsimony versus Dollo parsimony 17
Figure 2.7: PEBCRES methodology 20
Figure 2.8: Logos of transcription factors used in CRE evolution simulation 21
Figure 2.9: Phylogeny used in CRE evolution simulation 22
Figure 2.10: Synthetic CRE length histogram 24
Figure 2.11: Scoring window length influence on REforge's performance on synthetic data 25
Figure 2.12: Ranking method influence on REforge's performance on synthetic data 25
Figure 2.13: Scoring method influence on REforge's performance on synthetic data 26
Figure 2.14: Ancestral filtering influence on REforge's performance on synthetic data 27
Figure 2.15: REforge's performance on biological relevant dataset variants 29
Figure 2.16: Phylogeny underlying the vision-impairment phenotype 31
Figure 2.17: REforge's performance on real data 32
Figure 2.18: Scoring method influence on REforge's performance on real data 33
Figure 2.19: Ancestral element filtering influence on REforge's performance on real data .. 34
Figure 2.20: Scoring window length influence on REforge's performance on real data 35
Figure 2.21: Ranking method influence on REforge's performance on real data 36
Figure 2.22: Tree manipulation 37
Figure 2.23: REforge versus Forward Genomics on synthetic data 38
Figure 2.24: REforge versus Forward Genomics on biological data 39
Figure 3.1: Example of similar motifs 43
Figure 3.2: Motif ranking with TFforge 45
Figure 3.3: TFforge's discriminative power in dependence of different filters 46
Figure 3.4: Influence of the count filter on TFforge's performance on synthetic data 47
Figure 3.5: Influence of the ranking method on TFforge's performance on synthetic data ... 47
Figure 3.6: TFforge's performance in dependence of the number of CNEs 49
Figure 3.7: TFforge's performance on pleiotropic elements 50
Figure 3.8: CRX logo 51
Figure 3.9: CRX ranking score for vision-impairment species and other trait-loss species 52
Figure 4.1: Summary scheme 55
Figure 4.2: Combination of REforge and TFforge 58
Figure 4.3: Combined TFforge+REforge analysis on subterranean mammals 60
Figure A.1: Score distribution assessment method comparison ii
Figure A.2: Influence of the scoring method on REforge's performance on synthetic data.... iii
Figure A.3: Influence of the scoring method on REforge's performance on synthetic data.....iv
Figure A.4: Influence of the scoring method on REforge's performance on synthetic data..... v

Figure A.5: Influence of the scoring method on REforge's performance on synthetic data vi
Figure A.6: Influence of the scoring method on REforge's performance on real data vii
Figure A.7: Influence of ancestral element filtering on REforge's performance on real data. vii
Figure A.8: Influence of the scoring window length on REforge's performance on real data viii
Figure A.9: Influence of the ranking method on REforge's performance on real data viii
Figure A.10: Motif ranking with TFforge ix
Figure A.11: TFforge's detection sensitivity at a precision of 80% on synthetic data with long trait-loss age..... x
Figure A.12: TFforge's detection sensitivity at a precision of 80% on synthetic data with medium trait-loss age..... xi
Figure A.13: TFforge's detection sensitivity at a precision of 80% on synthetic data with short trait-loss age..... xii
Figure A.14: TFforge's detection sensitivity at a precision of 100% on synthetic data with long trait-loss age..... xiii
Figure A.15: TFforge's detection sensitivity at a precision of 100% on synthetic data with medium trait-loss age..... xiv
Figure A.16: TFforge's detection sensitivity at a precision of 100% on synthetic data with short trait-loss age..... xv

List of Tables

Table 1: score comparison 15
Table 2: Total number of inserted TFBSs into 1000 ancestral sequences 22
Table 3: ATAC-seq dataset names and their underlying data 30
Table 4: Enrichment of top-ranking CNEs in tissue-specific ATAC-seq elements..... 34
Table 5: Score correction method comparison 44
Table 6: Runtime examples of REforge/TFforge's scoring module 57

Abstract

Understanding the connection between an organism's genotype and its phenotype is a key question in evolutionary biology and genetics. It has been shown that many changes of morphological or other complex phenotypic traits result from changes in the expression pattern of key developmental genes rather than from changes in the genes itself. Such altered gene expression arises often from changes in the gene regulatory regions. That usually means the loss of important transcription factor (TF) binding sites within these regulatory regions, because the interaction between TFs and specific sites on the DNA is a key element of gene regulation.

An established approach for the genome-wide mapping of genomic regions to phenotypes is the Forward Genomics framework. This approach compares the genomic sequences of species with and without the phenotype of interest based upon two ideas. First, the initial loss of a phenotype relaxes selection on all phenotypically related genomic regions and, second, this can happen independently in multiple species. Of interest are such regions that diverged specifically in phenotype-loss species. Although this principle is general, the current implementation is only well-suited for the identification of phenotype related gene-coding regions and has a limited applicability on regulatory regions. The reason is its reliance on sequence conservation as divergence measure, which does not accurately measure functional divergence of regulatory elements.

In this thesis, I developed REforge, a novel implementation of the Forward Genomics principle that takes functional information of regulatory elements in the form of known phenotype-related TF into account. The consideration of the flexible organization of TF binding sites within a regulatory region, both in terms of strength and order, allows the abstraction from the region's sequence level to its functional level. Thus, functional divergence of regulatory regions is directly compared to phenotypical divergence, which tremendously improves performance compared to Forward Genomics, as I demonstrated on synthetic and real data.

Additionally, I developed TFforge which follows the same approach but aims at identifying the TFs relevant for the given phenotype. Given a multi-species alignment with a phenotype annotation and a set of regulatory regions, TFforge systematically searches for TFs whose changes in binding affinity between species fit the phenotype signature. The reported output is a ranking of the TFs according to their level of correspondence. I prove the concept of this approach on both biological data and artificially generated regions. TFforge can be used as a standalone analysis tool and also to generate the input set of TFs for a subsequent REforge analysis. I demonstrate that REforge in combination with TFforge is able to substantially outperform standard Forward Genomics, i.e. even without foreknowledge of relevant TFs.

Overall, the in this thesis introduced methods are examples for the power of computational tools in comparative genomics to catalyze biological insights. I did not only show a detailed description of the methods but also conducted a real data analysis as validation. REforge and TFforge have a wide applicability on endless phenotypes, both on their own in the association of TF and regulatory region to a phenotype. Moreover, particularly their combination constitutes in respect to gene regulatory network analyses a valuable tool set for evo-devo studies.

Acknowledgements

First of all, I would like to express my gratitude and sincere thank to my supervisor *Michael Hiller* for this unique opportunity to work in this great environment and on this impressive project. I very much enjoyed the possibility of a change of expertise gain of entirely new knowledge that this project offered. In this context, I also thank *Ivo Sbalzarini* and *Jochen Rink* for being my TAC members and sharing very constructive comments and helpful suggestions in all the meetings.

The staff of MPI-CBG's computer department, especially *Oscar Gonzalez* and *Juraj Matejcic* deserve a big thank you for their intense efforts to keep our computational environment always running smoothly.

A very special thank you goes to all the present and past members of the *Hillerlab* with which I have always enjoyed working very much and that gave every day a comfortable feeling.

I am very grateful for being so lucky to always get great support and motivation from my parents *Birgit und Andreas Langer* as well as my sister *Anne Langer*, whenever there is need. Furthermore, I thank all my friends in and outside of Dresden for their motivation, understanding, but also welcome distraction. A particular "thank you" to *Kathrin Seibt*, *Ulrik Günther* and *Daniel Rolle* for taking the time to read parts of my thesis and to give very constructive comments.

Lastly, simply, but most affectionately: thank you *Marija Matejcic!*

"There is a grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved."

Charles Darwin
On the origin of species (1859)

1 Introduction

The importance of computation in biology

The growing availability of data is ubiquitous, not only but in particular in biology. These amounts of data require and at the same time allow the development of methods for analysis, leading to the emergence of novel research fields and their exponential growth. Amongst them is bioinformatics, which rapidly gained importance in the mid-1990s by the advances in sequencing technology. Sequencing here refers to the determination of DNA's nucleotide[1] base pair[2] (bp) order. Following the Human Genome Project that, by 2003 officially finished sequencing a human genome reference and identified all human genes, hundreds of animals have been sequenced. This number is ever growing, resulting in projects like i5k and Genome 10K. These two projects aim for nothing less but sequencing the genomes of 5000 insects and the genome of at least one individual from each vertebrate genus and thus sequencing about 10,000 vertebrate genomes, respectively.

This increasing number of available genomic datasets allows different comparative approaches to mine the genome for meaningful differences and similarities like gene candidates, lineage-specific accelerated regions, ultra-conserved regions and losses or divergences of generally conserved regions. Therefore, they compare this variety of species on a "building plan" level.

Evolution of morphological phenotypes

As many species exist, as vast is the diversity of ***phenotypes*** (the observable traits) that exists and needs to be understood. In this context, the connection of the phenotype and the ***genotype*** (the genetic building plan) is a key question particularly for evolutionary biology and genetics. Our nowadays understanding of the emergence of this endless phenotype diversity was heavily shaped by the rapid growth of evo-devo (evolutionary developmental biology) in the late seventies. Evo-devo added the development as a third key element for the quest of understanding form to first Darwin's evolution theory and second, the molecular biological basis of heredity. Amongst the surprising insights that evo-devo brought are the concepts of deep homology and pleiotropy. Deep homology refers to the concept of the development of similar structures in extremely distant species, like fins and digits in zebrafish and mice, respectively (Nakamura, et al. 2016), being controlled by deeply conserved genetic mechanisms. The genes that are involved in such fundamental developmental processes like the establishment of morphology are usually signaling genes. Due to their fundamental role, they often act on many seemingly unrelated phenotypic traits, a concept that is called ***pleiotropy***. Consequently, function-influencing mutations in such genes cause manifold effects and are therefore mostly lethal, putting their genetic sequence under a strong ***selection***.

[1] The four nucleotides adenine, guanine, cytosine and thymine abbreviated by A, G, C and T, respectively, represent the monomeric units of the DNA.

[2] Due to the double-stranded nature of DNA in which two nucleotides form a pair, the DNA's building blocks are commonly called base pairs

Selection, the fundamental principle of Darwin's evolution theory, is the concept of differential survival and reproduction arising from heritable traits. The underlying mechanism for heredity is the passing on of genomic information. This information exists in various types like how a protein is built and when and in which tissue it is produced. These types of information are encoded in different kinds of DNA regions like genes and gene regulatory regions and just as different as their encoded information is, as different is the code itself. These differences cause the selection, from a sequence-perspective, to impose very different restrictions upon the underlying DNA regions. For example, due to the precise and strictly defined grammar of the genetic code, the sequence variability of genes is very low. Hereby, computational methods are able to identify and analyze genes very accurately. This in turn promotes deeper insights into the specific rules of how genes encode information, how this information is translated biologically into gene products, what their molecular function is and how genes have evolved in different species. Yet as stated before, the functional understanding of genes is only a part of understanding the phenotype-genotype relation but due to its defined encoding the most approachable. (Wray 2007) summarized this with "genetic code makes it easy to identify, accurately and comprehensively, mutations that alter protein sequences from DNA sequence comparisons alone (that is, non-synonymous substitutions, frameshifts, premature stop codons), whereas the same is not true of mutations that alter transcription, splicing, transcript stability and other regulatory processes".

The importance of gene regulation both on an organism's morphological development and on its evolution was another key insight from evo-devo. The correct spatiotemporal activity of the developmental genes is generally crucial for the organism's development, but at the same time changes of this spatiotemporal expression pattern allow the alteration of traits. Space, in this context, refers to the tissue and time refers to the developmental stage at which a gene is expressed. Accordingly, famous examples like loss of pelvic structures in several freshwater populations of threespine stickleback fish (Shapiro, et al. 2004), the loss of limbs in snakes (Kvon, et al. 2016; Leal and Cohn 2016) and the wing development in bats (Booker, et al. 2016) show the conservation of the protein sequence of important genes like *Pitx1, Shh* and the HoxD cluster but changes in their spatiotemporal expression patterns. Instead, these changes result from mutations in regulatory regions of the genes. This establishment of the complex spatiotemporal expression patterns of key developmental genes by many different regulatory regions is considered to be a general principle. The regulatory regions themselves are located either in direct vicinity of the gene like the promoter at the transcription start site or even up to thousands or a million bases away from the target gene. Since such distal regulatory regions are nevertheless on the same chromosome, they are also called ***cis-regulatory elements (CRE***). Often, CREs are modular in the sense that one CRE regulates the gene expression in only a subset of tissues and developmental stages. This provides a larger mutational space for innovation than the gene-coding sequence, whereas simultaneously allowing the avoidance of pleiotropic costs. The idea of a vast phenotype divergence that results from gene regulatory changes instead of minor gene changes has been proposed already in (King and Wilson 1975) and reviewed in (Wray 2007). Finally, Carroll summarized it in his 'genetic theory of morphological evolution' with two major points: *'(1) form evolves largely by altering the expression of functionally conserved proteins; and (2) such changes largely occur through mutations in the cis-regulatory regions of mosaically pleiotropic developmental regulatory genes and of target genes within the vast regulatory networks they control*' (Carroll 2008).

Transcription factors and their role for the function of a regulatory element

Because of the ambiguous "grammar" of the regulatory code, regulatory elements are rather elusive to knowledge acquisition compared to genes. This refers to questions like "Which genes does a CRE affected and by which mechanisms", "In which tissue and at which time point is the CRE active?" and especially "How can they be reliably predicted?". Generally, the function of a CRE is defined indirectly by the proteins that bind to short subsequences of the CRE (typically 5-15 bp in eukaryotes) and control the expression of the regulatory element's target gene. These proteins are called ***transcription factors*** (***TFs***). Using experimental approaches to identify potential binding sites of TFs, it has been found that such ***binding sites of TF*** (***TFBSs***) are usually degenerate. This means a TF binds not only to a single sequence but a set of sequences with more or less strong preferences of particular nucleotides at each position of the binding sequence. To represent the binding preference in a computationally approachable way, strings or strings with k mismatches are therefore inadequate, but the common representation are rather ***position weight matrices*** (***PWMs***) (Stormo 2000). A PWM for the binding sites of a particular TF describes position-wise the nucleotide preference, i.e. the frequency of each of the four nucleotides at this particular position in experimentally validated binding sequences. Their common visualization by frequency and resulting information content is called ***binding sequence logo*** (Figure 2.8 shows 5 examples).

It follows that mutations in the sequence of regulatory elements can severely influence the binding of TFs by inactivating or creating binding sites. Therefore, *minor sequence changes can majorly affect the regulatory element's function*. On the other hand, flexibility of TFBSs is not only observed with respect to their sequence but also with respect to their position, both in absolute terms in the regulatory region and relative to each other. That is, many regulatory regions are considered to act in a "billboard-like" manner with respect to the TFBS arrangement (Smith, et al. 2013). This model allows for example "binding site turnover": the gain of a redundant binding site allows the loss of a previously functional TFBS elsewhere in the CRE. Together with the little mutational constraint outside of TFBSs, it is, therefore, evident that the *function of a regulatory element can stay preserved, despite major sequence divergence*. It follows that the sequence conservation and divergence is not an appropriate indicator of CRE function. A better predictor is the presence and strength of TFBSs. A particular reason for this is the generally high conservation of the binding preferences of TFs even between distantly related species. (Liu, et al. 1997; Nitta, et al. 2015)

In summary, to understand the phenotype-genotype relationship from an evolutionary developmental perspective, we want to know which TFs are involved in a phenotypic change and via which regulatory elements they are acting. That is in a more general perspective: What does the underlying gene regulatory network look like? Comparative genomics provides here an unprecedented opportunity to target this question by uncovering the genomic changes that underlie phenotypic differences between species. Therefore, the question particularly addressed in this thesis is: ***How can we computationally identify high-quality candidates for such CREs and TFs that are connected to a given phenotype***, to circumvent the costly experimental screenings?

Several tools targeting somewhat related problems are existing. A big portion of them is covering the *de novo*-discovery of motifs, meaning the identification of motifs from sequences without prior knowledge of binding sites. This can be done by discriminating two sets of sequences like DIPS (Sinha 2006), DME (Smith, et al. 2005) or in a phylogenetic way like PhyME (Sinha, et al. 2004) and PhyloGibbs-MP (Siddharthan 2008). PhyloGibbs-MP handles the input

data also discriminatively, but differentiates between two sets of genomic regions instead of two sets of species, meaning it allows for example the comparison of conserved CREs regulating two sets of genes. SADMAMA (Keich, et al. 2008) compares two sets of sequences with respect to their number and/or quality of binding sites, according to a given PWM but does not consider any phylogenetic relatedness of the sequences.

A comparative approach that predicts the phenotypic function of CREs is "Reverse Genomics" (Marcovitz, et al. 2016), yet it does not predict CREs related to a given phenotype. A computational approach that has been developed to associate genetic regions to a given phenotype is "Forward Genomics" (Hiller, et al. 2012). This comparative genomics method detects a species-specific conservation decay and associates it with the corresponding phenotype signature. Forward Genomics is currently the best approach to answer the above stated question, at least with respect to CREs (but not TFs), nevertheless it is only applicable to CREs in a limited sense (see Chapter 1.1). Therefore, I use the basic idea of Forward Genomics as the foundation of my methods and outline Forward Genomics and in particular its underlying principle in the following.

1.1 Forward Genomics

The Forward Genomics framework links phenotypic differences between species to their underlying genomic differences by focusing on the parallel loss of a trait in independent lineages. For this, it relies on the two concepts: "use it or lose it" and "repeated evolution". The first one describes the consequences of selection on genetic information (e.g. encoding a protein or defining a regulatory element's function), underlying a phenotype. Necessary information is preserved throughout evolution by selection, unnecessary information on the other hand decays by the accumulation of mutations (neutral evolution) over time. A loss of a phenotype therefore leads to a loss of genetic information relevant for this phenotype. The second concept points out that a phenotype loss in independent lineages leads eventually to an information loss in all of these lineages. Taking this together, Forward Genomics' idea is to search genome-wide for a specific evolutionary signature that matches the species' phenotype signature: ***Which independent losses of an otherwise conserved sequence happened specifically in the trait-loss species.*** The repeated evolution here adds specificity to this search. Figure 1.1 illustrated the evolutionary model and the assumptions behind Forward Genomics.

The current implementation of Forward Genomics relies on sequence identity as a measure for species divergence (Hiller, et al. 2012; Prudent, et al. 2016). Sequence identity is defined as the branch length normalized fraction of identical bases. As the sequence identity is, due to the stringent genetic code, an adequate proxy for gene function, this method was applied successfully to several different traits in the past. One example is the "synthesis of vitamin C", which is independently lost in primates including human, guinea pigs and many bats. The identified cause is a sequence divergence in the Gulo gene, which encodes a key enzyme in the vitamin C synthesis (Hiller, et al. 2012). Another example is the independent loss of vision in the blind mole rat and the cape golden mole. Here, (Prudent, et al. 2016) used Forward Genomics to identify several genes enriched in functions related to eye development, the perception of light and the circadian rhythm.

Following the working theory "Morphological phenotypic changes largely arise by changes in gene expression" (Roscito, et al. 2018) recently associated on a large scale, via Forward

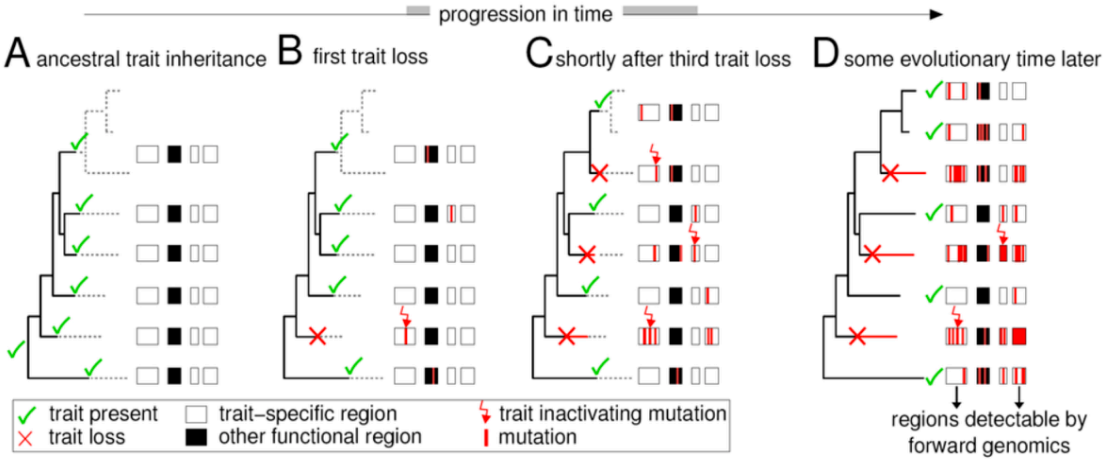


Figure 1.1: Evolutionary model and assumptions behind Forward Genomics approach (Figure from (Hiller, et al. 2012))

(A) The preservation of an ancestral trait in descendant species causes genomic regions, required for this trait, to evolve under natural selection and therefore stay preserved as well.
(B) An inactivating mutation in a trait-relevant region causes a lineage to lose the ancestral trait.
(C) The previous loss of the trait removes the selective pressure from the trait-specific regions, which can therefore evolve neutrally and accumulate random mutations. In parallel, two additional lineages lost independently the trait, due to inactivating mutations in the same or other trait-relevant regions.
(D) The trait-specific regions of the three trait-loss species continue to diverge neutrally, while neutral selection conserves the corresponding regions in the trait-preserving lineages. Forward Genomics detects such characteristic differential loss signatures to reveal functional genomic components underlying the trait.

Genomics, diverged sequences with the loss of two different phenotypes: limbs in snakes and functional eyes in subterranean mammals. The results show for the limb phenotype on one hand the maintenance of limb genes and on the other hand the snake-specific divergence of hundreds of putative limb-regulatory elements. While for the vision phenotype, as stated before, diverging genes are known, also a large set of putative vision-related CREs was identified. In this study we, furthermore, analyzed the diverged putative CREs with respect to binding sites of TFs and found a correlation between trait-loss species-specific divergence of CREs and loss of TFBSs, specific for trait-relevant TFs. We therefore concluded that the observed CRE sequence divergence is likely a result of their phenotype-related functional inactivation by the loss of phenotype-related TFBSs.

This finding consolidates the importance of TFBSs for the function of regulatory elements. In additional consideration of the previously described lack of correlation between sequence conservation and CRE function, it is conclusive that the effectiveness of the Forward Genomics idea could be substantially improved by its application onto the functional level of regulatory elements instead of the sequence level. I pursue this idea by the implementation of REforge, a regulatory element-specific version of Forward Genomics (see Chapter 2). TFforge (see Chapter 3), an adaptation of REforge, is designed to identify binding motifs and thereby TFs that are involved in phenotypic changes, which REforge uses as input. Hence, both methods can complement each other (see Chapter 4). Each method is individually tested on simulated data, validated on biological data and compared with respect to their runtime (see Chapter 4). Lastly, I give in Chapter 5 an overall summary.

2 REforge

REforge (regulatory element forward genomics) aims for the identification of CREs that are relevant for an analyzed phenotype, by incorporating knowledge of trait-relevant TFs. The expected input data consists of a set of candidate CREs, a set of trait-relevant TFs and their motifs and a genome alignment, including the species that lost the phenotype during their evolution. To achieve this aim, REforge applies Forward Genomics' basic principles (see Chapter 1.1) to TFBSs:

conceptual idea

Functional information has a tendency to get lost due to the sequence's accumulation of mutations over time (in neutral evolution) unless the information is required by the organism or implies some fitness advantage. In this case selection counteracts the process of information loss and preserves it. Consequently, mutations under negative selection will likely not spread i.e. will not be fixed in the population and remain constrained to a small fraction of a population. Since regulatory elements are functionally defined by their TFBSs, the following is assumed: During the course of evolution, following a trait loss, binding sites of trait-relevant TF in trait-relevant regulatory elements are lost at a higher rate than binding sites in other regulatory elements. In the trait-non-relevant elements, binding sites are either only lost (and gained) randomly due to turnover or represent random motif matches without functional relevance. Thus, given a set of TFs that are relevant for a specific trait, I identify such regulatory elements that lost TFBSs preferentially in trait-loss species and preserved the TFBSs in trait-preserving species.

specific idea

In particular, I first assess for every element the collective binding affinity of given set of TFs to each species' sequence (see Chapter 2.1). Second, since the species' sequences are dependent based on the species' phylogeny, I adopt Forward Genomics' branch method (Prudent, et al. 2016) to control for the phylogenetic relatedness by transforming the sequence scores to branch scores. These represent the binding affinity changes during times of independent evolution (see Chapter 2.2). In order to associate the phenotype with specific regulatory elements, I contrast the trait-loss and the trait-preserving species by comparing their corresponding branch scores (see Chapter 2.3). The result is a ranking of the regulatory elements according to their phenotypic relevance, which is imposed by their TFBS signature. The only requirement for input elements is that they lie within aligning regions of the considered species' genomes for the aforementioned application of the branch method.

To verify and optimize REforge, I developed a CRE evolution simulation (see Chapter 2.4) which allows tests on ground-truth data. Additionally, I assess REforge's performance on real data of mammalian species with vision impairment as phenotype with respect to experimentally identified vision regulatory elements (see Chapter 2.5). At last, I compare REforge with Forward Genomics both on the synthetic and real data (see Chapter 2.5.4).

2.1 Computing TF binding affinity to a sequence

In order to utilize knowledge of phenotypically relevant TFs in the analysis to abstract from the sequence level to a functional level, the assessment of the binding affinity to a sequence is an essential step. For many regulatory regions, the position of their TFBSs within the region is functionally not important, but only their presence. Consequently, the turnover of TFBSs, meaning the gain of a new binding site by chance that allows loss of a previously functional site, is common (Villar, et al. 2014). Furthermore, if the element's function is defined only by the TFBSs, it follows that every part of the regulatory element's sequence, which is not bound by transcription factors, is free to diverge without effect on the elements function. It follows that, albeit functional conservation, sequence conservation is not necessarily given and the variable position of the TFBS within the CRE together with mutations outside of TFBSs represents noise for Forward Genomics.

Additional to this freedom of arrangement of TFBSs within a regulatory element, it is known that both strong and also weak TFBSs can exhibit a biological function (Farley, et al. 2015). Therefore, we want to measure the affinity of TFs to a sequence in a way that does not imply constraints on the arrangement of TFBSs and takes varying strengths into account. A big variety of motif scoring tools like FIMO (Grant, et al. 2011), RSAT (Thomas-Chollier, et al. 2008) and MOODS (Korhonen, et al. 2009) is existing that evaluates the sequence position-wise and call motif occurrences based on a threshold. The drawback of such an approach is that binding sites that score below the threshold are not considered, which may apply to weak TFBSs. Moreover, it remains unclear how the TFBS scores of these methods can be combined into a CRE score and in this context also how TFBSs that overlap each other could be treated. A different approach to score sequences is used by Stubb (Sinha, et al. 2003; Sinha, et al. 2006) (described in detail in Chapter **Error! Reference source not found.**). Stubb utilizes a ***hidden Markov model (HMM)*** and thereby circumvents the need for a threshold by assigning a score to the entire sequence under consideration of weak, strong and overlapping binding sites. For this reason (and its possibility to consider multiple motifs simultaneously), I use Stubb to score sequences and, due to observed scoring biases (see Chapter 2.1.2), introduce different score correction methods (see Chapter 2.1.3).

2.1.1 Stubb's scoring approach

Stubb has been developed to discover CREs in genomic sequences based on clusters of binding sites, required for the elements activity. This means, Stubb compares different loci within a single genome. For REforge's purpose I make use of its scoring function but in order to compare orthologous[1] loci from the genomes of different species.

Stubb computes a score for a given sequence s essentially by comparing the two probabilities of s being generated by two different HMMs. The first HMM contains a background state and motif states, whereas the second HMM is a single state HMM with only the background state. For both HMMs the background state is represented by a PWM of length 1, which is derived from a given background sequence. The motif states are defined by the motifs' PWMs and thus the number of states depends on the number of considered

[1] Orthologous regions descended from the same ancestral region as a result of species divergence into two separate species.

motifs. It follows that one can score a sequence with respect to a single TF motif or a set of TF motifs. A scheme of the HMM is given in Figure 2.1. In each state, the Markov model samples from the state's PWM and transitions into another state i with probability t_i. Therefore, also weak binding sites contribute to the final score and the contributions of the binding sites are proportional to their strengths. The same holds for overlapping binding sites which are considered via distinct paths of the HMM.

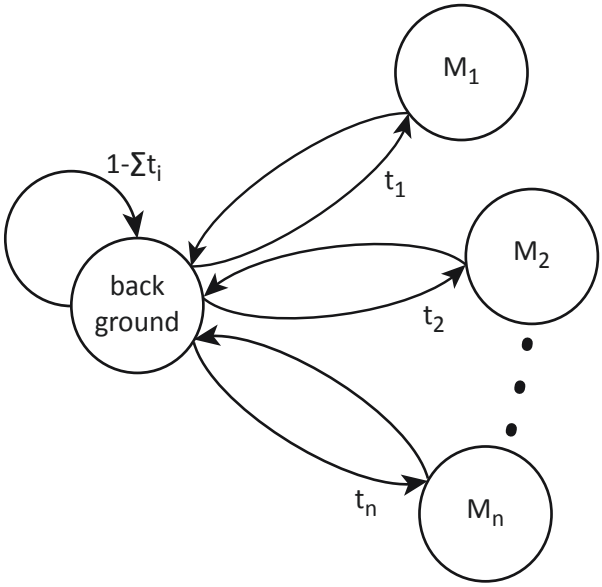


Figure 2.1: Stubb's hidden Markov model
Stubb computes the score based upon the likelihood of a sequence under this "n+1 state HMM". The background state emission derives from a background sequence and the n motif states (M_1, M_2, …, M_x) emit according to the motif PWM. Transition probabilities unequal to 1 are indicated. The transition probabilities t_i are preset or found via Expectation Maximization.

The transition probabilities can be considered as "motif weight" or frequency of motif occurrence. They can be explicitly provided by the user but since they are generally not known prior, Stubb by default conducts an Expectation Maximization (EM) of the resulting score over the transition probabilities. Considering a sequence s, let

- $\vec{t}$ be the vector of transition probabilities t_i,
- $\vec{0}$ be the zero vector
- $P(s|\vec{t})$ be the probability with which s is generated by the HMM and which the Forward algorithm[1] computes by effectively summing over all possible paths followed by the hidden states,
- and hence $P(s|\vec{0})$ the probability of the background HMM generating s

then the Stubb score can be described by

$$F(s) = \max_{\vec{t}} \log\left(\frac{P(s|\vec{t})}{P(s|\vec{0})}\right) \quad (1)$$

In order to have Stubb fitting well into REforge's HPC workflow, I minimized its file input and output by adding the possibility of

- passing the sequence itself as parameter to Stubb instead of a sequence file
- a brief output that only prints the score and the optimal transition probabilities without creating any output files.

[1] dynamic programming algorithm, computing $P(s)$ for a sequence s under a given HMM

Together with the possibility of passing the transition probabilities to Stubb via `<(echo 'file content')` instead of a file, my modification of Stubb generally reads only the PWMs and a background sequence file, which are both static and can therefore be cached by the operating system.

Additional changes introduced to Stubb's code include the minimal length of sequences and the removal of hardcoded pseudocounts. Support of sequences shorter than 200 bp was found to be necessary, since a significant portion of the analyzed sequences turn out to be shorter. Stubb's original sequence length constraint results from its usage as CRE detection tool in combination with its possibility to compute the background distribution based on the input sequence instead of a background sequence. Therefore, I generated a set of random background sequences from which REforge always chooses the one that matches the input sequence by GC-content, prior to calling Stubb. Stubb applies by default a fixed pseudocount to the given matrix in order to guarantee every motif state to sample any sequence with a probability greater zero. The removal of this internally added pseudocounts was necessary, since I use PWMs that are, first, already pseudocount-corrected and, second, and more importantly, contain frequencies instead of counts.

2.1.2 Scoring bias from differing lengths and GC-contents

I observed that the binding affinity score of a sequence is influenced by the length and the GC-content of the sequences, irrespective of changes in the binding sites. This can be problematic, since thereby insertions, deletions and mutations of the sequence outside of the actual TFBS can influence the score, although they are unlikely affecting the CRE's functionality.

The influence of the sequence length on the score was assessed with the help of random sequences of different lengths into which the consensus of a motif was inserted at a random position. With increasing sequence length, the score decreases (see Figure 2.2), which can be explained and derived from the way Stubb computes its score. With increasing sequence length, the score contribution of a single motif occurrence decreases. To avoid bias due to different sequence lengths, the sliding window option of Stubb is used and the sequence score are defined as the maximum over all window scores. This approach basically identifies the cluster of highest TFBS density, by which Stubb originally defines CREs.

It is evident that the probability of a motif occurrence also depends on the sequence's nucleotide distribution or, since TFBS can be located on both DNA strands, to the sequence's GC-content. The likelihood is maximal if the GC-contents of sequence and motif match. In this case, the score is expected to be smallest. I assessed the influence of the sequence's GC-content analogously to the scoring window length's influence previously. This means that I replaced a subsequence of randomly generated 1000 bp sequences, with predetermined GC-content, by the consensus of a motif. As shown exemplarily for the TF complex Notch1-CSL in Figure 2.3 the scores match the expectation and depend on the sequence's GC-content with a minimum around the motif's GC-content.

Rapid GC-content changes along a single phylogenetic branch are unlikely and therefore by the transition to branch scores (see Chapter 2.2) the GC-content influence remains minor, especially by adapting the background accordingly for every sequence. Nevertheless, the usage of score correction methods and AFTP, that is described in the following Chapter, can further reduce it.

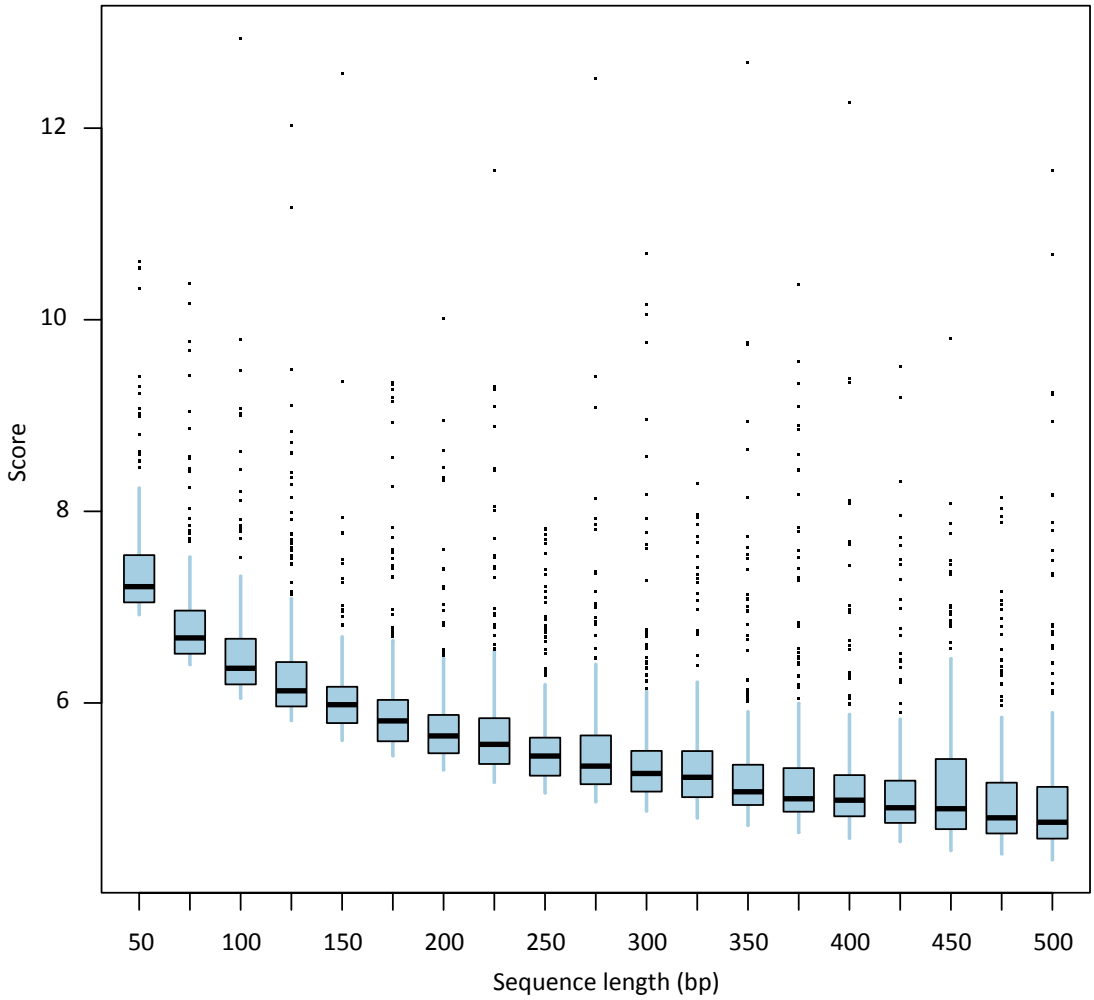


Figure 2.2: Sequence length dependency of Stubb's score
Shown are the Stubb scores of sequences in dependence on their length in nucleotides. Each box comprises the scores of 250 randomly generated sequences with uniform nucleotide distribution into which at a random position the consensus CCGTGGGAACTA of the Notch1-CSL PWM (see Figure 2.3 for sequence logo) from UniPROBE (Hume, et al. 2015) was introduced.

2.1.3 Scoring method correction

In the previous chapter, I observed consistently positive scores of random sequences (see Figure 2.3) while comparing the Stubb scores of random sequences with and without inserted TF binding motif instances. Generally, Stubb computes a score, that is greater than 0, due to the probability for random motif occurrences. Nevertheless, a random sequence generally should not contain any effective binding site, and thus should get a score of 0. I call the average score of a sequence with a specific GC-content the null score. As (Kim, et al. 2010) describes, Stubb asks whether "*the motif helps 'explaining' the data (sequence) significantly better than the background model can*". If Equation (1) is rewritten as

$$F(s) = \max_{\vec{t}} \log\left(\frac{P(s|\vec{t})}{P(s|\vec{0})}\right) = \log\left(\frac{\max_{\vec{t}} P(s|\vec{t})}{P(s|\vec{0})}\right) \qquad (2)$$

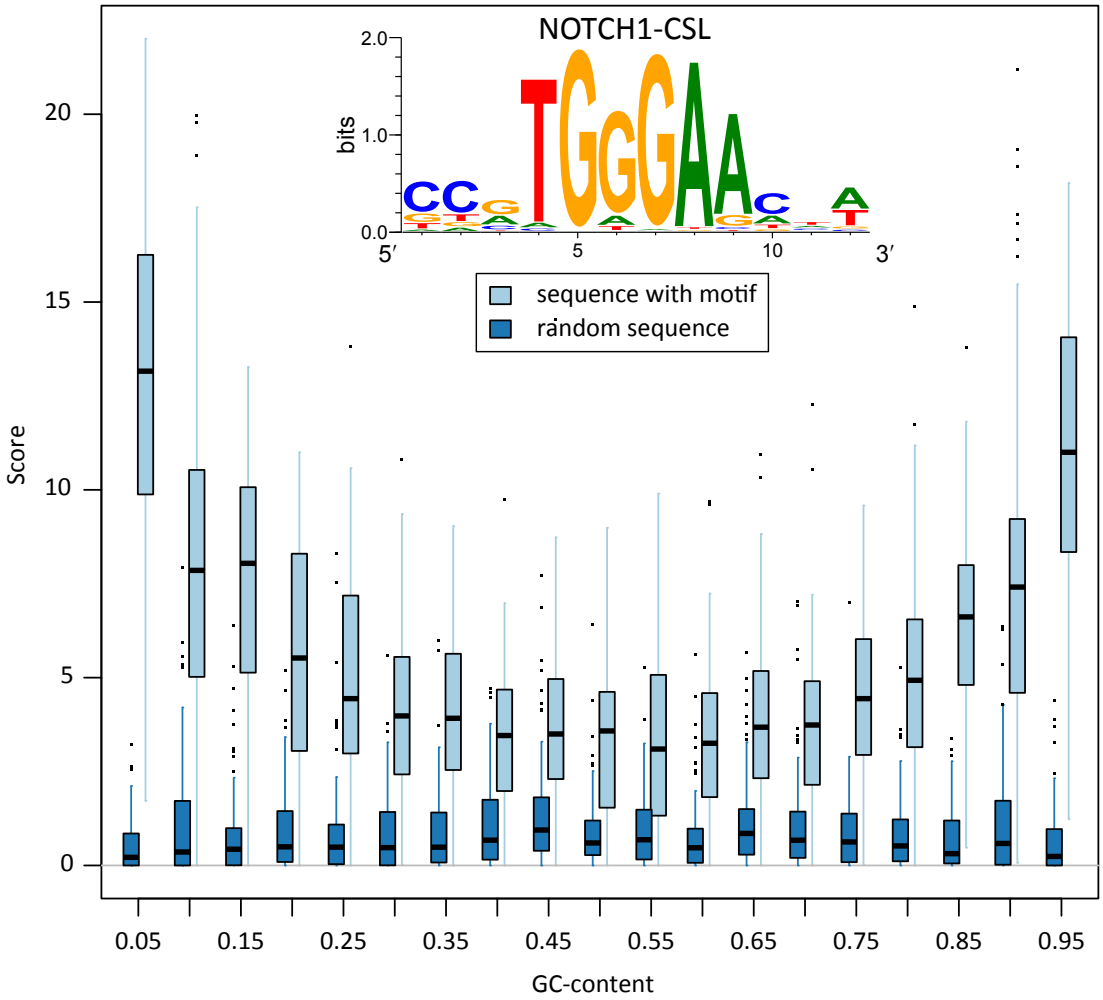


Figure 2.3: GC-content dependency of Stubb's score
Comparison of Stubb scores of random sequences without (dark) and with (light) an inserted sampled motif in dependence of the GC-content of the sequence. Each box comprises 100 1000 bp sequences that were scored with a 200 bp scoring window with respect to the Notch1-CSL PWM from UniPROBE (Hume, et al. 2015). The motif's GC-content is 53%. Its sequence logo representation is shown in the top.

it is obvious that the background score is independent from the EM step and serves solely as a normalization such that the score is positive if and only if the argument of the maximum is not the zero vector. That means, Stubbs functionality can be rephrased as the question *whether the sequence has a positive motif occurrence frequency*. The problem of positive null scores now arises from the (matched) background motif of Stubbs denominator. This normalization does capture the nucleotide distribution similarity between sequence and motif but fails to capture the probability of a motif occurrence in dependence of the nucleotide distribution. To counteract this effect, three options to filter or correct the score are possible:

1. Apply a motif and sequence GC-content specific filter onto the scores.
2. Correct by subtracting the null score, meaning the average score of shuffled sequences. Stubbs question is therefore changed to:

   Does the motif help 'explaining' the sequence better than it helps 'explaining' similar sequences in terms of length and nucleotide distribution?

   or in terms of motif occurrences:

   Does the sequence contain more or better motifs than random sequences with the same GC content and sequence length?

3. Correct by learning the motif weights, i.e. transition probabilities into the motif states from a background sequence (e.g. shuffled versions of the input sequence) and compute the given sequence's score with these motif weights as a correction term for the Stubb score. Under the new null model, the method now asks:

   "*Given that we must use a [n]-state HMM to explain/parse a sequence, are we significantly better off using a higher motif weight than the value learned from background sequences?*" (Kim, et al. 2010)

   or in terms of motif occurrences:

   Does the sequence contain more or better motifs than we would expect it to contain by our knowledge of background sequences?

Stubb with minimal score filter (MSF)

By assessing the scores of random sequences with inserted binding sites that have been sampled from the PWM, one can define thresholds for every motif in dependence of the sequence's length and GC-content. For the following applications of this score was the 10th quantile chosen as threshold. The length of the random sequences is 200, matching the scoring window length.

Stubb with Null Score Correction (NSC)

This correction method subtracts the average score of shuffled versions of the sequence from the sequence score. The shuffling preserves the sequence's nucleotide composition and length but no potential binding sites which gives rise to a "binding site less background score". Note that, due to the identical nucleotide composition is $P(s|\vec{0}) = P(r|\vec{0})$. Hence, the corrected score is

$$F_{NSC}(s) = F(s) - \operatorname*{avg}_r F(r) \tag{3}$$

$$= \log\left(\frac{\max_{\vec{t}} P(s|\vec{t})}{P(s|\vec{0})}\right) - \operatorname*{avg}_r \log\left(\frac{\max_{\vec{t}} P(r|\vec{t})}{P(r|\vec{0})}\right)$$

$$= \log\left(\frac{\max_{\vec{t}} P(s|\vec{t})}{P(s|\vec{0})} \frac{P(r|\vec{0})}{\operatorname*{avg}_r \max_{\vec{t}} P(r|\vec{t})}\right)$$

$$= \log\left(\frac{\max_{\vec{t}} P(s|\vec{t})}{\operatorname*{avg}_r \max_{\vec{t}} P(r|\vec{t})}\right) \tag{4}$$

This score correction therefore basically replaces the score normalization term by a null score correction term. Note that, due to this change, the positivity of the scores is not necessarily given anymore. For the following applications of this correction method is the average null score always computed from the scores of ten shuffled sequences.

Stubb with Null Transition Probability Correction (NTPC)

This approach follows (Kim, et al. 2010)'s idea of adapting Stubb's null model for their motif scanning tool SWAN (which unlike Stubb scores a sequence with respect to only a single TF). Before the correction step, the optimal transition probabilities of the model are learned for shuffled sequences representing a background. This means, with r denoting the shuffled sequences, one can define

$$\overrightarrow{t^r} = \operatorname*{avg}_r \operatorname*{argmax}_{\vec{t}} P(r|\vec{t})$$

Then t_i^r refers to the "optimal" transition probability into motif state i for which the score for the sequence r is maximal. Next, the sequence score is corrected by the score of the same sequence under the fixed model parameter $\overrightarrow{t^r}$, that means:

$$F_{NTPC}(s) = F(s) - F(s|\overrightarrow{t_r}) \qquad (5)$$
$$= \log\left(\frac{\max_{\vec{t}} P(s|\vec{t})}{P(s|\vec{0})}\right) - \log\left(\frac{P(s|\overrightarrow{t^r})}{P(s|\vec{0})}\right)$$
$$= \log\left(\frac{\max_{\vec{t}} P(s|\vec{t})}{P(s|\overrightarrow{t^r})}\right)$$

This method therefore computes the log likelihood ratio of the sequence's maximum energy and the sequence's expected energy by the background transition probabilities. For the following applications of this null correction method the average null transition probability is always, unless explicitly mentioned, computed from the optimal transition probabilities of 50 shuffled sequences.

Stubb with Ancestrally Fixed Transition Probabilities (AFTP)

The random appearance and disappearance of TFBSs within an element during the evolution along a phylogenetic branch introduces noise to REforge's scoring method and becomes the more likely the more motifs are considered simultaneously. To counteract this noise, the motif set can be restricted to its ancestrally present subset. One can further even fix the motif weights throughout the whole phylogeny, if one assumes that the ancestral element is optimal with respect to its function. This is equivalent to the assumption that TFBSs stay qualitatively preserved in terms of their relative occurrence frequency (under the additional assumption that it is unlikely that one TF takes over the function of another).

This idea of fixed relative motif occurrence frequencies can be facilitated by fixing the transition probabilities ancestrally while computing the species' sequence scores of a regulatory element. This means that the Stubb score of the ancestral regulatory element is computed first and the corresponding optimal transition probabilities are saved. For every descendent species Stubb can now skip its EM step and instead use the ancestrally optimal transition probabilities to compute the score. This approach can be combined with each of the correction methods described above. In case of the NTPC the transition probabilities can be ancestrally fixed in two versions: for the numerator or for both numerator and denominator.

A summary of the described combinations of score correction methods is given in Table 1. An empirical comparison of these methods is done on synthetic data in Chapter 2.4.3 and on real in Chapter 2.5.3.

Since the CRE's sequences in different species are not independent of each other but related according to the phylogeny, the set of scores of trait-loss species cannot directly be compared to the set of scores of trait-preserving species. This problem can be solved by converting the sequence scores to branch scores.

	Species specific scoring	Ancestrally fixed transition probabilities (AFTP)	
Stubb	$\log\left(\frac{\max_{\vec{t}} P(s\|\vec{t})}{P(s\|\vec{0})}\right)$	$\log\left(\frac{P(s\|\overrightarrow{t_{anc}})}{P(s\|\vec{0})}\right)$	with $\overrightarrow{t_{anc}} = \operatorname{argmax}_{\vec{t}} P(s_{anc}\|\vec{t})$
Stubb with NSC	$\log\left(\frac{\max_{\vec{t}} P(s\|\vec{t})}{\operatorname{avg}_r \max_{\vec{t}} P(r\|\vec{t})}\right)$	$\log\left(\frac{P(s\|\overrightarrow{t_{anc}})}{\operatorname{avg}_r P(r\|\overrightarrow{t_{anc}})}\right)$	
Stubb with NTPC	$\log\left(\frac{\max_{\vec{t}\,:\,\vec{t}\in R^n, t_n < t_n^r} P(s\|\vec{t})}{P(s\|\overrightarrow{t^r})}\right)$	(1) $\log\left(\frac{P(s\|\overrightarrow{t_{anc}})}{P(s\|\overrightarrow{t^r})}\right)$ (2) $\log\left(\frac{P(s\|\overrightarrow{t_{anc}})}{P(s\|\overrightarrow{t^r_{anc}})}\right)$	
	with $\overrightarrow{t^r} = \operatorname{avg}_r \operatorname{argmax}_{\vec{t}} P(r\|\vec{t})$		

Table 1: score comparison

2.2 Scoring multiple phylogenetically dependent sequences

As phylogenetically closely related species share a relatively recent common ancestor, their sequences are generally more similar than the sequences of more distant species, as for example Figure 2.4 illustrates, where human and chimp sequence are identical and highly similar to the gorilla sequence. Yet, all the three species are closely related. It follows that the resulting scores are dependent in accordance to the phylogeny as well. I tackle this problem by adapting the branch method from (Prudent, et al. 2016) (see Figure 2.5) and use the branches of the elements underlying species phylogeny as set of evolutionary independent events. For this the phylogenetic tree needs to be known. Since only sequenced species are analyzed and their alignment is expected as input, the tree can be considered as given (it could be inferred from the genomic sequence data otherwise).

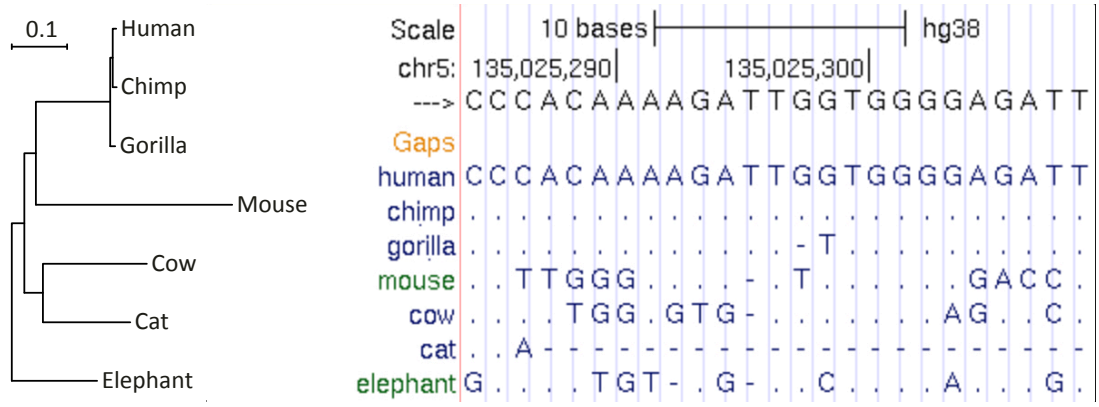


Figure 2.4: Genome browser screenshot

The alignment of a genomic sequence of 25 bp for human, chimp, gorilla, mouse, cow, cat and elephant is shown (left) together with the phylogenetic relation of the species (right). Dots indicate identical bases with respect to the human sequence, which is the reference. Letters and dashes represent mismatches and deletions, respectively.

A score for a phylogenetic branch is computed as the difference of the binding scores of its end and start node. Thus, positive branch scores represent gain or strengthening of binding sites and negative scores indicate binding site losses along the branch. If the score is approximately 0, one can differentiate two cases:

- if both start and end node of the branch were scoring positively, then TF binding sites were qualitatively preserved (not necessarily at the same position);
- if both nodes were low-scoring, then the factor has no binding affinity to the considered element.

In the latter case, a ***branch filter*** can be applied: Since such branches carry no relevant information for the trait-loss analysis, they can be considered as noise and can be filtered out. The decision between the two cases is made threshold based, with the threshold depending on the score correction method. If sequence scores are not corrected, REforge filters upon motif and sequence's GC-content dependent precomputed thresholds; if a score correction is used, all branch scores are kept for which either the start or the end node sequence score is greater zero. A test on 1000 synthetic sequences (see Chapter 2.4.1) and 567 phenotype-unrelated TF motifs showed, that the latter filter removes on average 44% of the scores.

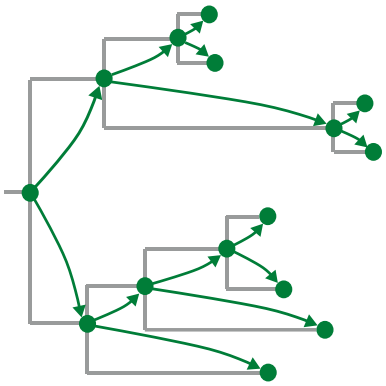


Figure 2.5: Scheme of Forward Genomics' branch method
By comparing every node to its direct ancestral node, the changes along the branches are measured.

With the similar reasoning, also a more general ***CRE filter*** can be applied to CREs whose ancestral sequence is low-scoring. A low ancestor score means that the CRE is not expected to exhibit regulatory function with respect to the phenotype. Therefore, any possible gains and losses of TFBS in descendent species are most likely also phenotype-unrelated, since the phenotype is ancestrally present. The influence of this filter on REforge is later assessed (see Chapter 2.4.3.1).

2.3 From branch scores to element ranks

The knowledge of the TF's binding preferences in terms of motifs enables us to compute for every element the above described branch scores, which quantify the change in binding affinity of the TF set to this particular element along the phylogenetic branch. In order to classify whether the binding affinity preferentially decreased during the evolution of trait-loss species, the branches are classified into "trait-loss" and "trait-preserving" branches. A gain of a trait along a branch is not considered here, since I focus on traits that are absent only in a minority of species and, thus, present in the ancestor of the considered species. Therefore, a

trait gain would be preceded by a trait loss, which would contradict Dollo's law of irreversibility[1].

In order to classify the branches, REforge infers the phenotype of every ancestral sequence for any element. For that, every CRE is treated independently because the input sequence alignments may be partly incomplete, meaning the sequence data of some species for some CREs could be unknown. In this case, the phylogenetic tree for these CREs is a subtree [2] of the considered phylogenetic tree. This states a problem particularly in combination with the ambiguity of the ancestor naming, which conventionally names any ancestral node by two of its descendent terminal species. Therefore, if any of these two species is missing, the very same ancestor is named after different species. To resolve this problem, the phylogenetic tree is pruned for every CRE to the subtree that contains only the species with known sequence information. Then, the phenotype for every ancestral species is inferred on only this CRE specific subtree.

REforge uses the Dollo parsimony method to infer the ancestral phenotypes. Dollo parsimony assumes that (parallel) gains of a phenotype are unlikely, compared to phenotype losses. Therefore, it allows only for a single trait emergence that can be subsequently lost as often as necessary (see Figure 2.6 for illustration of Dollo Parsimony vs. "standard" or Fitch parsimony).

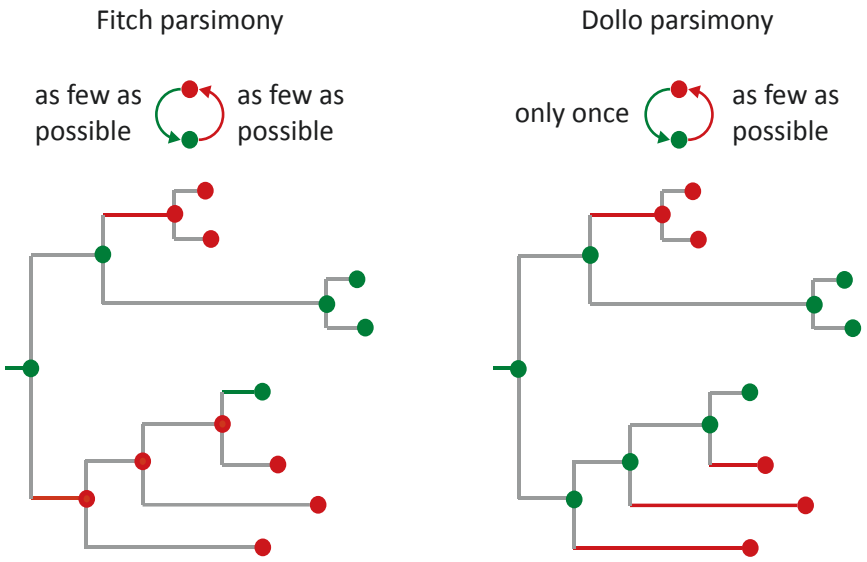


Figure 2.6: Fitch parsimony versus Dollo parsimony
While Fitch parsimony (left) infers ancestral phenotypes from the terminal species such that the number of gains (green branches) and losses (red branches) is minimal, Dollo parsimony has the additional constraint that only a single trait gain (usually at the root) is allowed.

Next, every branch is assigned to either the "trait-preserving" or "trait-loss" class according to the phenotype of its end node. The branch scores are, due to their construction independent measures; therefore, the "trait-preserving" class and the "trait-loss" class can be statistically compared in their entirety. Because of the ancestral presence of the trait, relevant

[1] Dollo stated his hypothesis first in 1893: "An organism is unable to return, even partially, to a previous stage already realized in the ranks of its ancestors"

[2] Subtree is here used in the graph theoretical sense, hence a tree forming subgraph

TFBSs are also ancestrally present and can therefore only be lost or conserved, whereas other TFBSs can be gained and lost randomly. Hence, either no qualitative difference between the two classes (thus a TFBS preservation) or a negative shift of the "trait-loss" branch score distribution (meaning TFBS loss) is expected. A statistically significant shift implies that TFBSs have been preferentially lost along these branches and, therefore, REforge concludes that the element is relevant for the given trait.

In order to assess the difference of the two branch score distributions and to finally rank the elements by their relevance, several options are possible:

- Pearson correlation coefficient: *Does a linear correlation between scores and the corresponding phenotype (0 or 1) exist?* The significance of the Pearson correlation coefficient has been used by (Prudent, et al. 2016) for assessing the difference in their sequence identity branch score distributions.
- Spearman correlation coefficient: *Does a monotonic correlation between scores and the corresponding phenotype (0 or 1) exist?* Spearman correlation, in contrast to the Pearson correlation tests for correlation of the ranks, and is therefore less affected by outliers but on the other hand does not capture the effect size. Both of these methods are usually used for continuous variables instead of binary. Analog to the Pearson correlation test, a significance can be derived from the correlation coefficient and used for the ranking. Rankings referring to the Pearson or Spearman correlation coefficient will, in the following, always be based upon the derived p-value.
- One-sided t-test: *Is the trait-loss branch score distribution compared to the trait-preserving branch score distribution negatively shifted?* The t-test compares the means under the assumption of a normal distribution of scores.
- One-sided Wilcoxon rank-sum test: *Is the trait-loss branch score distribution compared to the trait-preserving branch score distribution negatively shifted?* The Wilcoxon rank-sum compares the means of the ranks and handles therefore also non-normal distributed data. Rankings referring to the t-test or Wilcoxon rank-sum test will, in the following, always be based on the p-value of the one-sided test.
- Cohen's D: *What is the effect size based on differences between means?* Cohen's D computes the distributions mean difference and normalizes it by their variances. This is especially in case of a big number of scores underlying the distribution more meaningful than the statistical tests.
- Upper bound of one-sided t-tests confidence interval: *How big is the negative shift of the trait-loss branch distribution compared to the trait-preserving branch distribution at least?* The confidence interval describes both the confidence that the t-test expresses via the p-value and the effect size. The interval becomes smaller with increasing confidence and shifts negatively with greater effect size. Since the test is one-sided, the lower bound is $-\infty$.
- Upper bound of one-sided Wilcoxon tests confidence interval: *How big is the negative shift of the trait-loss branch distribution compared to the trait-preserving branch distribution at least?* The difference of this measure to the previous one lies only in the underlying test. Rankings referring to the confidence interval of either t-test or Wilcoxon rank-sum test will, in the following, always be based on the upper bound of the confidence interval.

The current implementation of REforge computes each of these measures. Yet, an empirical comparison of these ranking methods with respect to REforge's performance will be later conducted on synthetic and on real data (see Chapter 2.4.3 and 2.5.3).

2.4 Verification on synthetic data

In order to verify and validate the general principle of REforge and optimize the method with respect to the scoring procedure and the ranking scheme, experiments need to be performed under controlled conditions. This means knowledge of the ground truth is necessary, i.e. which TFs are involved in the phenotype loss of which species and ideally at which time point the loss happened. To this end, I start by creating a regulatory element evolution simulation. In brief, this simulation evolves a set of ancestral CREs along a phylogeny to obtain sets of CRE sequences in 20 species. A detailed description of the simulation setup is given in the following Chapter 2.4.1. Afterwards, I describe the usage of this synthetic data for a proof of REforge's concept (see Chapter 2.4.2). Furthermore, the synthetic CREs can be used to compare the different scoring and ranking methods for REforge and to explore the stability of the results upon varying input data (see Chapter 2.1.3).

2.4.1 CRE evolution simulation setup

Given a phylogeny with predefined trait-loss time points and a starting sequence that represents the ancestral version of the regulatory element, I simulate the element's evolution for every phylogenetic branch separately. For such evolutionary steps along a branch PEBCRES (Duque, et al. 2014) was adapted.

PEBCRES

PEBCRES uses a sequence and several model parameters as input and applies a discrete-time Wright-Fisher simulation with a fixed population size to the input sequence. That is, a set ("population") of sequences undergoes alternatingly a mutation and a selection step. In the beginning, the population is initialized with copies of the starting sequence. The number of iterations ("generations") is chosen proportional to the phylogenetic branch length. Each mutation step introduces with a predefined rate insertions, deletions and substitutions to every individual of the population independently. I chose the mutation rate to be 10^{-4} because this allows the simulation of an evolutionary distance of 0.1 substitutions per neutral site with 1000 iterations. The following selection step starts with the computation of the fitness of every individual with respect to its regulatory functionality. Then, a new population is generated by sampling each individual of the new generation independently from the parent generation with a probability distribution, that depends on the parental fitnesses. The fitness of an individual, therefore, defines the its survival, death and/or reproduction. PEBCRES computes the fitness by predicting an expression profile via GEMSTAT (He, et al. 2010) (see following paragraph) and comparing it to a given "ideal" expression profile. The entire functionality of PEBCRES is schematically outlined in Figure 2.7A.

In order to simulate a trait loss two options are possible, which depend on the desired function of the regulatory element. The regulatory element can be specific in its function to the regarded trait or it can be a pleiotropic element. The latter regulates gene expression in multiple different domains and thus is relevant for multiple different traits. In case of a specific element, neutral evolution upon a trait loss can be expected and this can be realized by removing the fitness influence during the selection step of the Wright–Fisher model. The

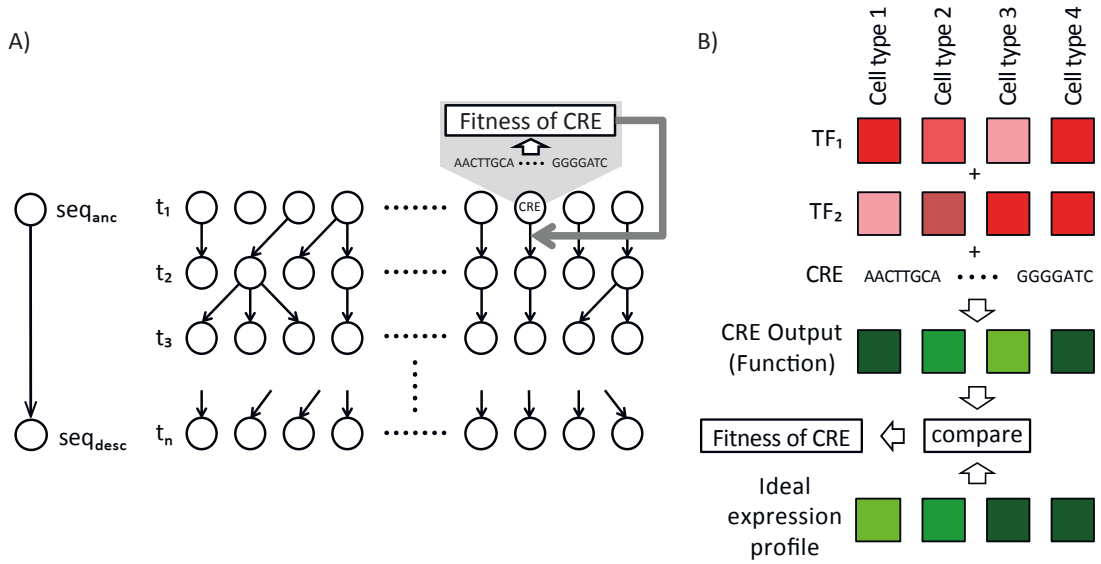


Figure 2.7: PEBCRES methodology
(adapted from (Duque, et al. 2014))
(A) Illustration of PEBCRES' underlying Wright-Fisher simulation. The first generation (t_1) of a fixed-sized population of sequences (CREs) is initialized with the sequence seq_{anc} and then evolved by multiple mutation and selection steps. The output sequence seq_{desc} arises from the final population t_n. In each generation, random mutations are introduced. Every individual of a generation is randomly sampled from the previous generation's population with a probability distribution dependent upon the fitnesses. The fitness of an individual is according to (B) determined via GEMSTAT (He, et al. 2010).
(B) Given a set of relevant TFs (red), their binding motifs and their concentration (brighter shades represent higher concentration) in different cell types, GEMSTAT predicts from the CRE's sequence CRE's output (green). This output is the CRE's regulatory function in terms of a gene expression in the corresponding cell type. A comparison of this expression profile with a given "ideal expression profile" gives rise to the CRE's fitness (similar profiles result in a high fitness)

removal of this influence basically switches off the selection pressure, that assures that fit individuals preferentially reproduce, and PEBCRES is basically only introducing mutations.

To consider pleiotropic elements, GEMSTAT's possibility of taking multiple cell types into account can be utilized. Assuming that evolution under selection is simulated with respect to two different cell types, the trait-loss case can be modeled by removal of the "lost" cell type from the fitness computation. This assures that the regulatory element is free to functionally diverge for the "trait-loss" tissue but preserves its function with respect to the second tissue.

In summary, the only PEBCRES input parameters that change throughout the simulation are the input sequence, the number of iterations (generations) and, in case of trait loss, the selection coefficient or cell type definition. It is therefore inefficient to create for every PEBCRES application a new config file, which PEBCRES uses by default as input. For this reason, I added additional parameters for the sequence file, the number of iterations, a flag for neutral evolution and a seed to the PEBCRES code. All other parameters can be defined in the config file, that is therefore valid for the entire simulation. The neutral evolution flag causes PEBCRES to ignore the selection coefficient, defined in the config file, and instead set it to 0. The seed is used to initialize the random generator, so that the simulation becomes for debugging purposes deterministic.

The output of every PEBCRES evolution is a population. As the final sequence for each node and leaf, the sequence with the median fitness out of the population is used. Since the

sequence is further evolved by subsequent PEBCRES applications, at least for every internal node of the phylogeny, I also added the possibility of initializing PEBCRES with a population instead of a single sequence.

GEMSTAT

GEMSTAT is a thermodynamics-based sequence-to-expression model. Based on the attributes of the given TFs, like their binding preferences (PWM), concentration and functional strength (as an activator or repressor), it predicts the functionality of the regulatory sequence in terms of the resulting gene expression level. This prediction is possibly done with respect to multiple cell types, that differ in their concentration levels of TFs. GEMSTAT's general methodology is outlined in Figure 2.7B.

In the following, I describe the setup of a tissue-specific CRE evolution simulation for 1000 regulatory elements that are relevant and specific to trait, which is lost in some species. The case of pleiotropic CREs is treated analogously with the difference of GEMSTAT considering two cell types. For tissue-specific CREs only a single cell type together with its TF concentrations and its "ideal" (i.e. target) expression is needed. Furthermore, the case of irrelevant CREs is treated analogously but without defining a trait loss.

I choose a set of five sufficiently different motifs (see Figure 2.8) as "target set TFs". These function as activators with equal activating strength and all five are with equal concentrations present. Since the trait shall be generally present throughout the evolution (I will consider the trait losses later on) the target expression is set to 100%. These parameters are maintained throughout the subsequent CRE evolution along the whole phylogeny.

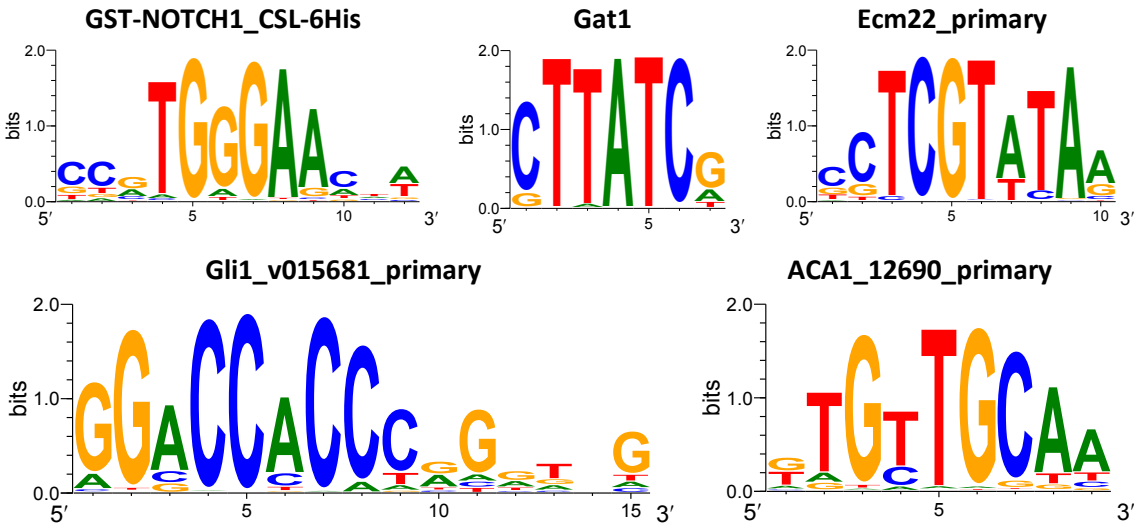


Figure 2.8: Logos of transcription factors used in CRE evolution simulation
The logos visualize the binding preferences of the TFs according to the PWMs. The x-axis indicates the sequence position. The height of each letter stack indicates the information content of this position, while the relative letter height corresponds to their frequency. Logos were created with WebLogo 3.5.0 (Crooks, et al. 2004)

The phylogeny was adopted from a mammalian phylogeny, with three species being defined as trait-loss species. The time of trait loss was chosen such that the age of trait losses corresponds to 0.09 substitutions per neutral site, which corresponds approximately to the divergence of the rat from the rat-mouse ancestor (the distance from human to human-rhesus ancestor is about 0.04 substitutions per neutral site in comparison). The phylogeny with the time points of trait losses and a scale in PEBCRES generations is depicted in Figure 2.9.

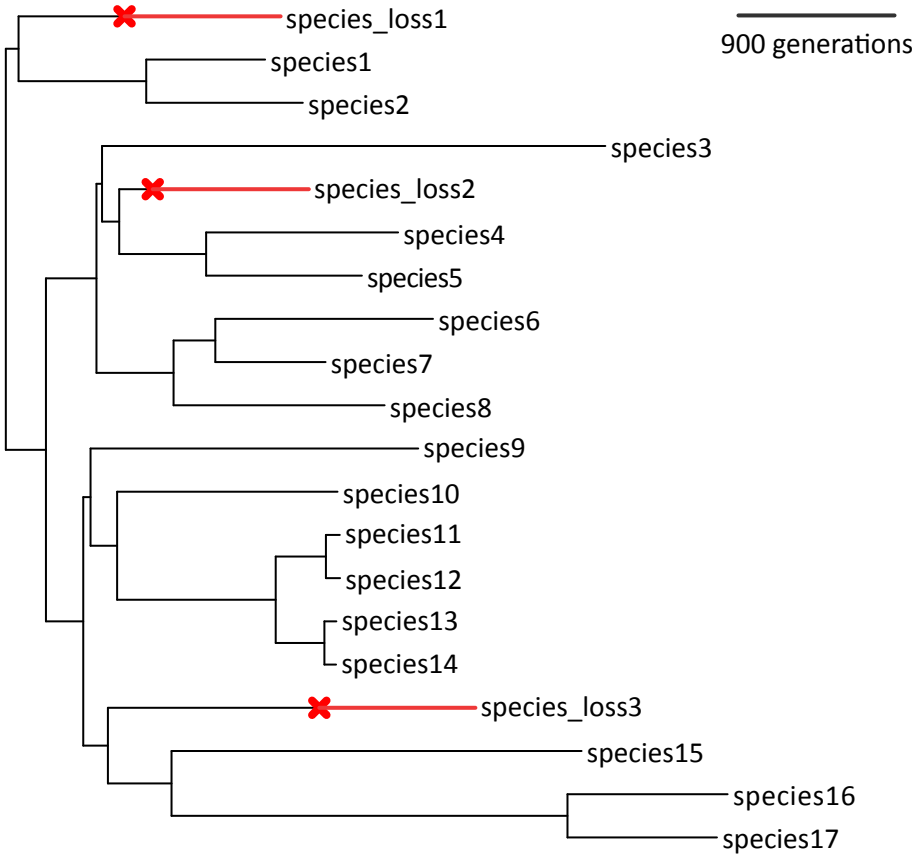


Figure 2.9: Phylogeny used in CRE evolution simulation
A phylogeny representing the relatedness of 17 trait-preserving and three trait-loss species, with branch lengths corresponding to the time of their evolution. Red crosses mark the trait loss, which is the starting point of neutral evolution (red branches) for the trait-loss species. The length of neutral evolution in this scenario is 900 generations.

As ancestral elements, I used 200 bp randomly (with uniform base distribution) chosen sequences into which consensus TFBSs were planted at five randomly, but distinctly chosen loci, such that the GEMSTAT fitness of these sequence was greater than 0.85. I decided to plant the TFBSs into the sequence instead of evolving a high-fitness sequence from a random sequence, because the latter proved to introduce bias in the usage of certain motifs. Upon testing the effect of planting motifs into a random sequence, five inserted motifs proved to be sufficient for a 200 bp long sequence to reliably achieve an element fitness > 0.85. As the overall number of included motif in all 1000 elements show in Table 2, this approach leads to a rather uniform incorporation of the motifs.

TF	ACA	Ecm	GST	Gat	Gli
Count	1014	1022	1033	967	964

Table 2: Total number of inserted TFBSs into 1000 ancestral sequences

To guarantee a maximal ancestral fitness, I refine the starting sequence by iteratively evolving it with PEBCRES for 200 iterations each and a high mutation rate of 10^{-3} until the fitness is either > 0.99 or the relative improvement $\frac{f-f_{old}}{1-f_{old}} < 0.1$.

For the successive application of PEBCRES to every phylogenetic branch, the simulation framework from (Prudent, et al. 2016) was used. In accordance to the given trait-loss scenario, it splits beforehand the trait-loss branches by introducing intermediate nodes, that mark the beginning of neutral evolution. Next, PEBCRES is applied via the framework to every node to evolve the sequence in dependence of the branch length. PEBCRES is run with a population size of 50, a substitution probability of 95%, insertion probability of 2.5% and a deletion probability of 2.5%. Furthermore, tandem repeats can be introduced with a probability of 20% if an insertion is simulated.

2.4.2 REforge's proof of concept

The CRE evolution simulation, described in the previous chapter, allows to proof the concept of REforge. With the five TF motifs shown in Figure 2.8 as the "target TF set", I generated two sets of CREs according to the CRE evolution simulation setup:

- "Foreground-Set":
  The key feature of this set of 1000 CREs is, that the sequences were evolved neutrally for an evolutionary distance of 0.09 substitutions per site on the three branches leading to the trait-loss species. The ancestral sequences contain motifs of the "target TF set", which stay functionally preserved in all the non-trait-loss species
- "Background-Set":
  This set of 10000 CREs was generated in the same manner like the foreground set, except for evolving all sequences (including the "trait-loss" species) under selection, i.e. de facto no trait loss happened. Therefore, this set represents regulatory elements that are controlled by the same TFs as the Foreground-Set but are unrelated to the analyzed phenotype. This means, that the TFs are pleiotropic in the sense of controlling at least two different phenotypes – the trait-loss phenotype and a different phenotype that creates selective pressure on this set of CREs.

In view of the fact, that the common use-case of REforge is the classification of a small set of relevant elements and a considerably bigger set of irrelevant elements, I aim for a high ***precision***, i.e. the fraction of identified elements that are relevant. For a defined target precision, I analyze REforge's performance with respect to ***sensitivity*** (recall), i.e. the fraction of relevant elements identified. An optimization for a high specificity in case of such unbalanced group sizes would be inappropriate, since in this scenario a high specificity could easily be achieved by a trivial classifier that assigns every element to the irrelevant set.

The results show, that REforge is able to distinguish a randomly chosen 200 CRE Foreground subset from a 10000 CRE background set with a sensitivity of up to 95% for a given precision of 90%. This concludes the proof of REforge's principle.

To further improve REforge's performance, I empirically compare, in the following, different variations, e.g. score correction methods, score filters and distribution assessment methods, via the generated CRE datasets.

2.4.3 Method optimization and assessment on different datasets

In order to, optimize parameters and other technical details of REforge and additionally analyze its performance under different biological conditions, more synthetic CREs were generated:

- Medium and short trait-loss age Foreground-Set variants:
  These two CRE sets are identical to the Foreground-Set, except for the trait-loss species' sequences, which have been evolved neutrally for a shorter evolutionary distance of 0.06 and 0.03 substitutions per site, respectively, instead of 0.09.
- "Background-Set 2":
  This set of 5000 CREs has been evolved fully under selection just like Background-Set 1, but with respect to a different set of TFs. Thus, these CREs represent unrelated regulatory elements (or background generally).

2.4.3.1 Technical and methodological variants

Scoring window length

I analyze the scoring window length influence, described in Chapter 2.1.2, by scoring (with NSC) the CRE Foreground- and Background-Set with different window lengths ranging from 50 and to 300 (see Figure 2.11). Small scoring windows here effectively cause a restriction to subsequences without a correcting benefit, since the sequences show only limited variation in length (see Figure 2.10). This is also the reason, why the results of the 250 bp and 300 bp scoring windows are almost identical, since the majority of sequences are shorter and full window size actually rarely used.

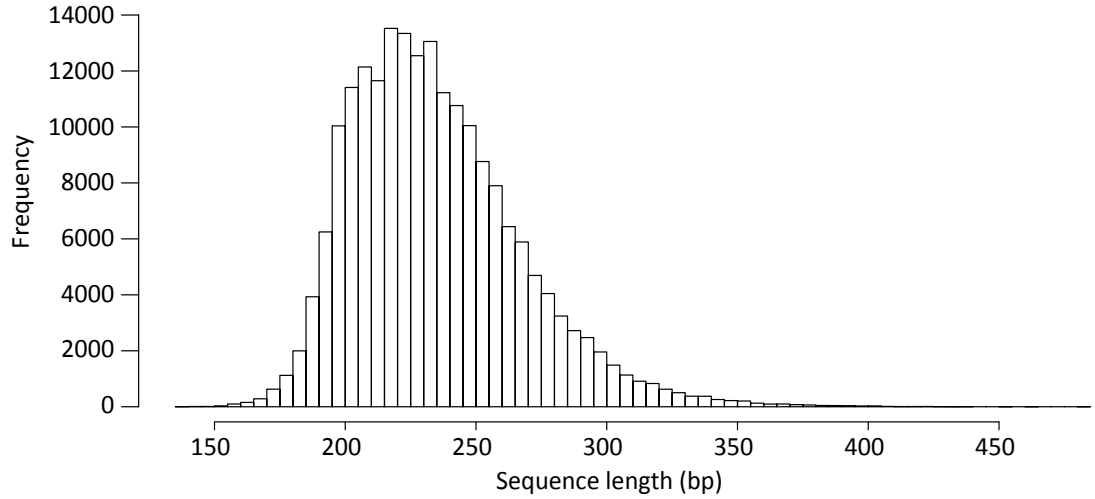


Figure 2.10: Synthetic CRE length histogram
The histogram summarizes the sequence lengths of all 10000 background CREs in all 20 species.

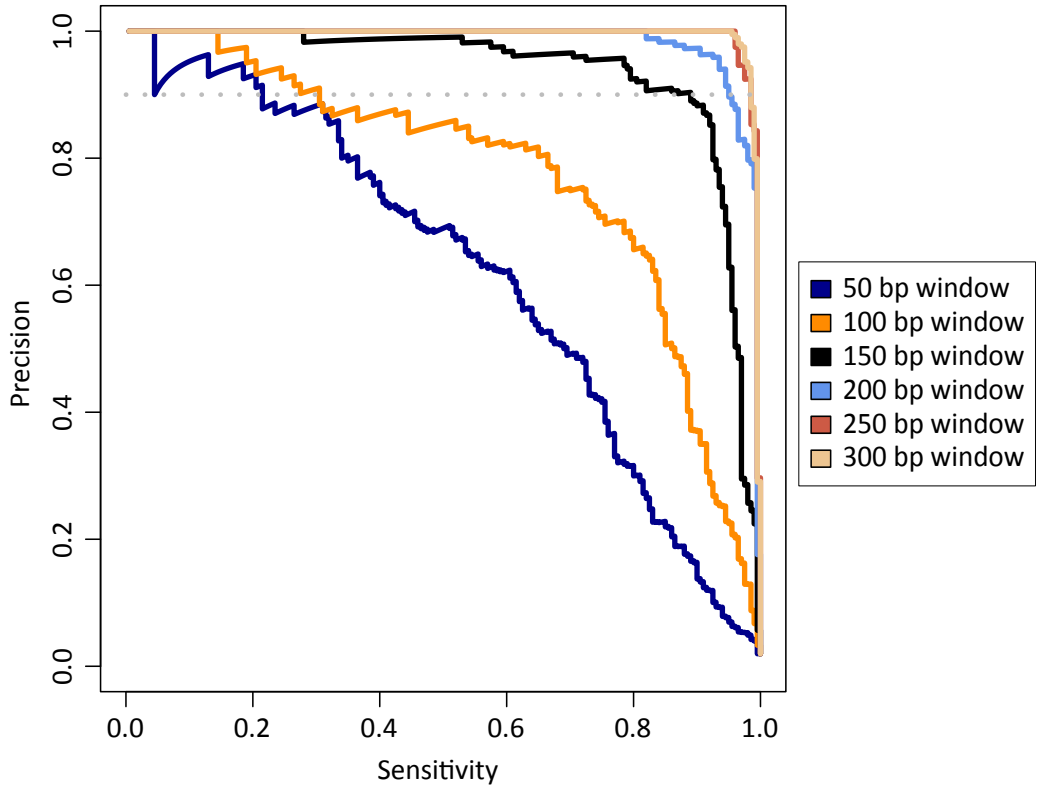


Figure 2.11: Scoring window length influence on REforge's performance on synthetic data

The precision-sensitivity curve shows the discriminative power of REforge with NSC and Cohen's D as ranking method on a dataset of 200 foreground CREs with a long trait-loss time and 10000 background CREs in dependence of the scoring window length. The grey dotted line indicates a precision of 90%.

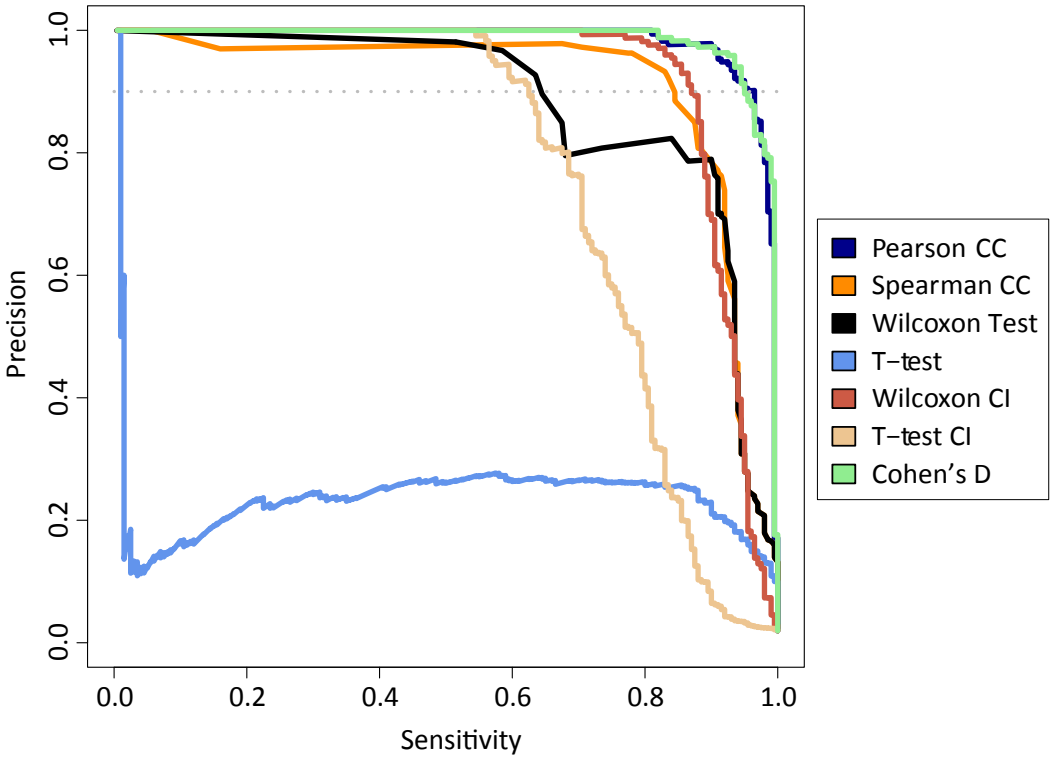


Figure 2.12: Ranking method influence on REforge's performance on synthetic data

The precision-sensitivity curve shows the discriminative power of REforge with NSC on a dataset of 200 foreground CREs with a long trait-loss time and 10000 background CREs in dependence of the ranking method. Here CC refers to the correlation coefficient and CI to the upper bound of the confidence interval. The scoring window length is fixed to 200. The grey dotted line indicates a precision of 90%.

Ranking method (branch score distribution comparison)

I compare the methods for assessing the significance of the branch score distribution difference, that result in the final ranking (see Chapter 2.3), for their effect on REforge's performance. As shown exemplarily in Figure 2.12. (see also Appendix A.I.1), the best performance is consistently achieved with either Cohen's D or the significance of a positive Pearson correlation coefficient.

Scoring method

I compare REforge's performance in dependence of the sequence scoring variations described in Chapter 2.1.3, including an ancestral fixation of motif weights. As shown exemplarily in Figure 2.13 and under various other conditions in Appendix A.I.2, Stubb with transition probability correction generally results in a lower detection sensitivity than Stubb with NSC and Stubb with MSF. The differences between each scoring method and its AFTP counterpart seem only marginal. This is not unexpected, since the simulation generates CREs with highly functional ancestral sequences. These sequences likely do not change their TFBS composition during the evolution, in the sense of lost binding sites of one TF that are compensates by binding sites of another TF, especially because the TFs were chosen such, that the motifs are not similar. This likely causes the transition probabilities to vary only little and therefore causes the ancestral fixation to not have an effect.

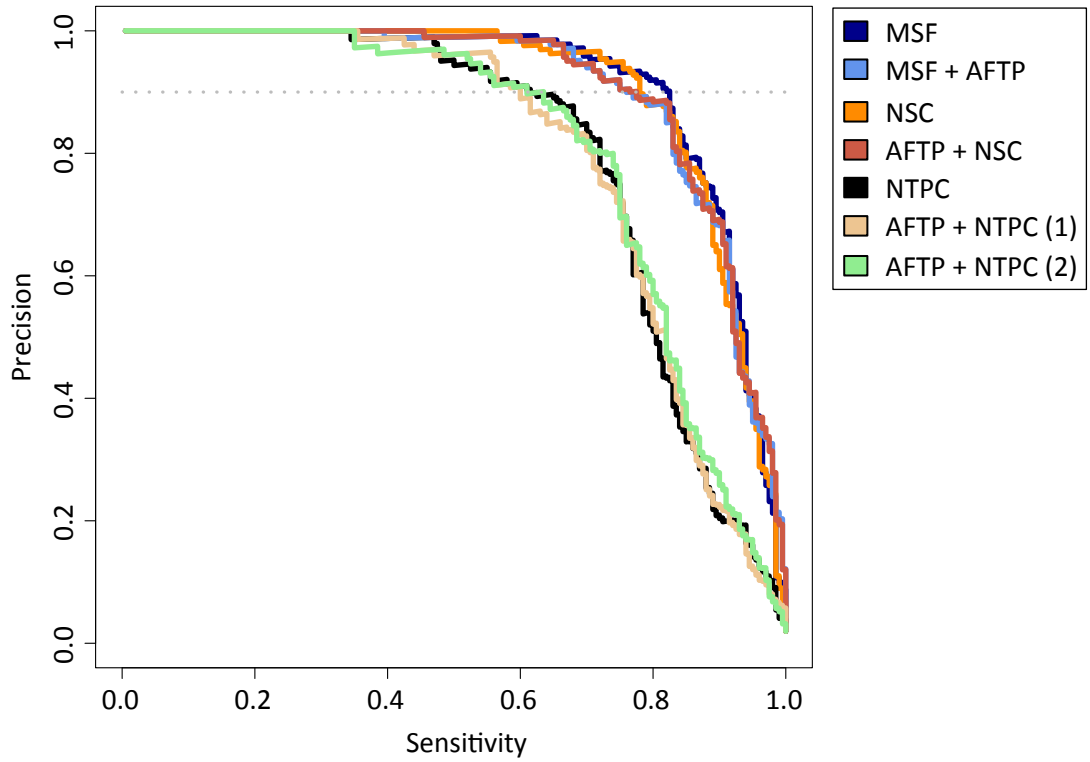


Figure 2.13: Scoring method influence on REforge's performance on synthetic data

The precision-sensitivity curve shows the discriminative power of REforge with Cohen's D based ranking in dependence of the different score corrections, described in Chapter 2.1.3. The underlying dataset consists of 200 foreground CREs with a long trait-loss time and 10000 background CREs. The scoring window length is fixed to 200. The grey dotted line indicates a precision of 90%.

Ancestral score based CRE filter

For identifying the Foreground-CREs out of the Background-Set 1 CREs, the ancestral score based CRE filter does not affect the performance, since the background CREs do not score zero by definition. The contrary is observed, if the background contains or consists of Background-Set 2 CREs. Since these CREs contain binding sites of relevant TFs only by chance, the ancestral sequences likely score zero, are therefore filtered and cannot be found anymore as false positives. This increases the sensitivity consistently slightly and for the NTPC drastically. The differences between the sensitivities at a precision of 90% when applying REforge with and without ancestral CRE filtering on different datasets is shown in Figure 2.14.

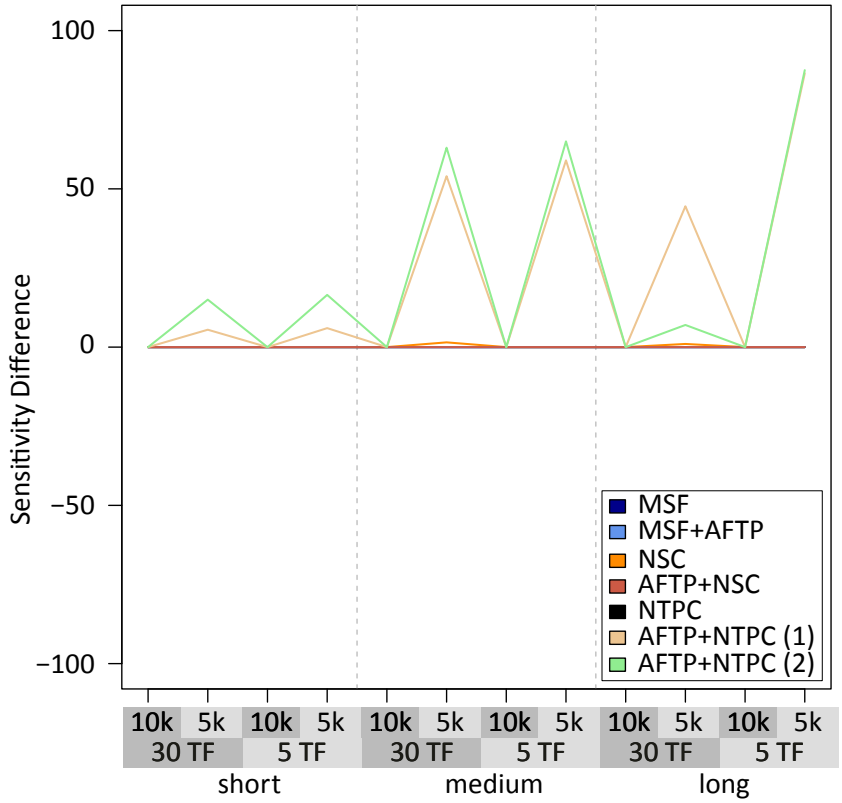


Figure 2.14: Ancestral filtering influence on REforge's performance on synthetic data
Sensitivity difference at a precision of 90% for every scoring method computed between the version with ancestral filtering and the version without ancestral filtering. The comparison is done for three different trait-loss ages (short, medium and long), 2 different TF sets (30 TF and 5 TF) and 2 different background sets (10k and 5k).

2.4.3.2 Biological relevant dataset variants

I further test the stability of the obtained performance results on different input datasets, that resemble different biological scenarios. These scenarios refer to the absence/ presence of TFBSs of the target TF set, the specificity of this TF set and the age of the trait-loss. For a comparison of all combinations, the achieved sensitivities at a precision of 90% and 99% are used. Additionally, this data stability comparison is performed with all scoring methods to identify possible advantages and drawbacks of the scoring methods with certain data types. Figure 2.15 gives the results of this comparison, which are discussed scenario-wise in the following paragraphs.

Presence of TFBSs in background elements

We want to know whether the existence of TFBSs in the "background" CREs influences REforge's performance. So far, I used CREs as background that contain binding sites for the target TFs, i.e. these CREs are controlled by the same set of TFs but exhibit function for different phenotypes in different tissues and/or developmental time points. Now, regions without such binding sites will be considered additionally. These regions could be regulatory regions controlled by different TFs, with or without relevant function for the considered phenotype or simply any other genomic region. Specifically, I compare the results obtained on a background consisting of 10000 CREs from Background-Set 1 against a background consisting of 5000 CREs each from Background-Set 1 and Background-Set 2 (see Figure 2.15). Because of the lack of considered TFBSs, the expectation is no or only a minor negative influence of such regions onto the performance due to noise. If, on the other hand, ancestral element filtering is applied, the majority of such elements are expected to be filtered, which would even cause a positive effect (due to the effectively smaller negative set). Indeed, the results match these expectations, and I observe no influence except for the effect due to ancestral filtering the background.

Specificity of the "relevant TF set"

When defining the set of relevant TFs, a trade-off between specificity and sensitivity needs to be done, since many TFs exhibit multiple functions and additionally the knowledge of a TF function is generally incomplete. If a very restrictive choice is made and only certainly phenotype-specific TFs are included, signal arising from pleiotropic TFs will clearly be missed. Therefore, I analyze the influence of a too general choice of TFs by including 25 unrelated TF into the TF set. The results show that the sensitivity is not majorly affected by including these unrelated TFs into the score computation and only a slight tendency of the methods with AFTP to gain advantage over their non-fixed counterparts is observed (see Figure 2.15). This is because AFTP effectively restricts the set of TF motifs to the ancestrally present ones, which reduces noise arising from random gains and losses of the other TF motifs.

Age of the trait losses

An older trait loss implies a longer time for neutral evolution to mutate TFBSs and should therefore increase the signal-to-noise ratio. Since REforge achieves with the initially chosen trait-loss scenario of 0.09 substitutions per neutral site already a very high performance, I reduced the age of trait loss, by using the medium and short trait-loss Foreground-Set variants. As the analysis shows (see Figure 2.15), this results in a lower sensitivity and is out of all assessed variants the main contributor to sensitivity changes. Nevertheless, REforge is able to achieve on the short trait-loss scenario still a sensitivity of around 30% at a precision of 90%.

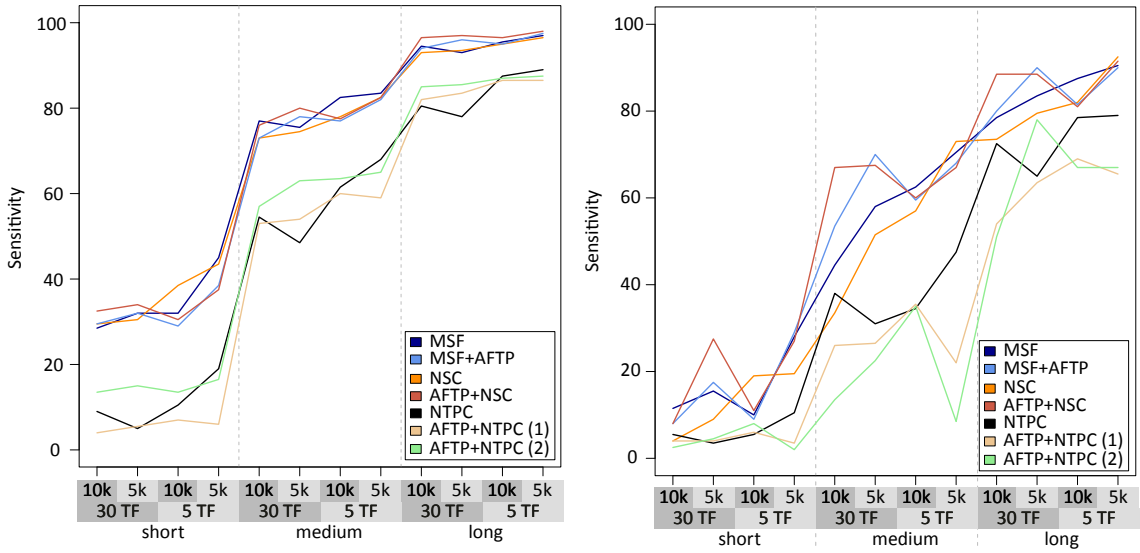


Figure 2.15: REforge's performance on biological relevant dataset variants
Compared are the different scoring methods on sets of elements that consist of 1000 foreground CNEs generated according to the short, medium or long trait-loss scenario and 10000 Background-Set CNEs. The background CNEs are represented either by the Background-Set 1, i.e. CREs with relevant TFBSs (10k) or by a combination of CREs with and without relevant TFBSs (5k). The results are measured by the achieved sensitivity at a precision of 90% (A) and 99% (B). The ancestral score based CRE filter was used for every scoring method.

2.5 Validation on biological data

I have so far given a proof of principle of the regulatory element adaptation of Forward Genomics on synthetic data. By the introduction of variants to this data, different theoretical aspects of real data were assessed and REforge's performance under these various conditions was analyzed. In the following, I will show the application of REforge on real data on the example of the vision-impairment phenotype in subterranean mammals, which is described in Chapter 2.5.1. Afterwards, I demonstrate the functionality of REforge on this data (see Chapter 2.5.2) and reverify the scoring method comparison results from synthetic data on real data (see Chapter 2.5.3). The vision-impairment phenotype is used because it has already proven itself to be well suited for Forward Genomics both for the identification of gene losses (Prudent, et al. 2016) and for the identification of diverging regulatory elements (Roscito, et al. 2018). This allows therefore also a direct comparison of REforge and Forward Genomics (see Chapter 2.5.4).

2.5.1 Vision-impairment phenotype

A remarkably complex phenotype to study is the degeneration of the visual system in subterranean mammals, ranging in the degree of degeneration from disorganized lenses as well as reduced eyes with loss of optic nerve connections and thinner retinas to completely skin-covered eyes. The blind mole-rat (*Spalax galili*) and the cape golden mole (*Chrysochloris asiatica*) show this latter most extreme phenotype and the star-nosed mole (*Condylura cristata*) and naked mole-rat (*Heterocephalus glaber*) still retained tiny eyes and optic nerves and therefore a poor visual acuity. Interestingly, as the phylogeny in Figure 2.16 shows, these

four species are not directly related, which means that these cases of vision degeneration developed independently during evolution. This predestines this phenotype loss for comparative genomics analysis like Forward Genomics on different types of genomic regions. Due to the gene divergence of highly specialized tissues, (Prudent, et al. 2016) was able to identify with Forward Genomics already many genes with functions in eye development and perception of light. However, gene loss alone cannot explain reduced eyes and retinal tissue and therefore understanding the regulatory landscape divergence is crucial. For this reason, (Roscito, et al. 2018) analyzed the vision-impairment phenotype and identified potential vision-related regulatory regions. I will summarize in the following the underlying dataset of this study because it allows a direct comparison and evaluation of REforge with Forward Genomics.

Vision-impairment dataset

Based upon a genome-wide alignment of 29 species, including mouse as reference, 25 other mammalian species and green anole lizard, frog and chicken, 491576 regions were identified that are conserved in at least 15 species, are at least 30 bp long and are not overlapping any genes. These regions are in the following referred to as ***CNE***s (***Conserved Non-coding Element***). Conservation of non-exonic regions is here used as a proxy for putative regulatory function, since many regulatory elements show higher conservation compared to non-functional regions (Bejerano, et al. 2004; Woolfe, et al. 2005). For 351279 out of these CNEs, sequence is available in all 4 vision loss species. From this set were, via Forward Genomics, 9364 CNEs identified that are significantly diverged in the four vision-impaired species.

Independent of this analysis, (Roscito, et al. 2018) determined open chromatin region in mouse eyes, limbs and midbrain at embryonic day E11.5 and separately in lens and retina at embryonic day E14.5 using ATAC-seq. Open chromatin regions are genomic regions in which the DNA is accessible, which for example allows TF binding and is therefore an indication of regulatory activity in the corresponding tissue at the corresponding developmental time point. The combination and removal of these sets from each other gives tissue-specific sets of regions with putative regulatory activity:

eyeE115.vs.limb+midbrain	eye E11.5 regions not overlapping midbrain or limb regions
retinaE145.vs.limb+midbrain	retina E14.5 regions not overlapping midbrain or limb regions
lensE145.vs.limb+midbrain	lens E14.5 regions not overlapping midbrain or limb regions
limbE115.vs.midbrain+Eye	regions specific to limb, since they do not overlap any other region
midbrainE115.vs.limb+Eye	regions specific to midbrain, since they do not overlap any other region
eyeE115E145.vs.rest	general eye regions not overlapping midbrain or limb regions
lensE145.vs.rest	regions specific to lens, since they do not overlap any other region
retinaE145.vs.rest	regions specific to retina, since they do not overlap any other region

Table 3: ATAC-seq dataset names and their underlying data

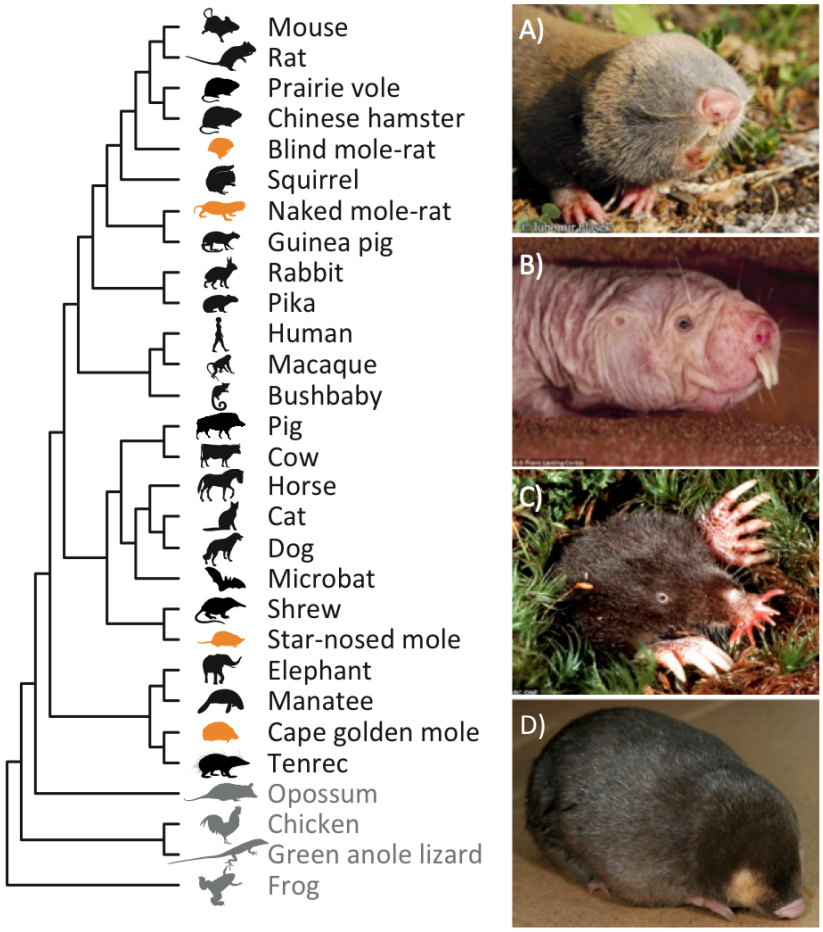


Figure 2.16: Phylogeny underlying the vision-impairment phenotype
The left panel shows the phylogeny of mammals used for the vision-impairment dataset, together with the outgroup species (grey). Pictures of the 4 vision-impairment species (orange) are shown in the right panel:
(A) blind mole-rat
(B) naked mole-rat
(C) star-nosed mole
(D) cape golden mole

These sets of tissue-specific regulatory regions can be used to assess the "quality" of the CNE set obtained by Forward Genomics. If the identified CNEs are vision-related CREs then an enrichment in the overlap with eye-related ATAC-seq elements is expected. On the other hand, no enrichment for the midbrain- and limb-specific regions should be observed. The results of this analysis are later shown in Chapter 2.5.4, where they are compared to the results of an identical analysis for REforge.

2.5.2 REforge identifies putative vision-related regulatory elements

To analyze the set of putative vision-regulatory elements, i.e. CNEs (defined in the previous Chapter 2.5.1) with respect to their functional relevance in development of the eye, a set of vision-related TFs is necessary. Following (Cvekl and Mitton 2010) the transcription factors Pax6, Prox1, Six3, Sox2, Sox11, c-Maf, Hsf4, Pitx3, Sox1, Crx, Mitf, Nrl, Pax2, Vax1, Vax2, Vsx2, Hes1, Lhx2, Rax and MafB are involved in retinal or lens development. To obtain a motif set for the vision phenotype, I overlap this TF list with the available TF motifs from three databases (see Chapter 3.1) and the corresponding cluster representatives (see Chapter 3.1.2). This results in a list of 30 motifs.

I use the same phylogeny of 4 vision-impaired species and 21 additional placental mammals as (Roscito, et al. 2018) to rank the CNEs by the relevance that the differential binding site losses of the TF motifs impose on them. In order to assess whether the obtained

top-ranking CNE set consists of likely eye-related CREs, the mouse tissue-specific ATAC-seq elements, described in the previous Chapter 2.5.1, are used. To do so, the overlap with these elements is computed and compared to the overlaps of randomly chosen CNE subsets of same size.

I find that the top-ranking CNEs generally are highly enriched in accessible regions that are specific to one vision-related tissue or the combination of all three, each with limb and midbrain as background (see Figure 2.17). In contrast, the open chromatin regions that are specific to limb or midbrain do not show an enrichment. This confirms that REforge-detected CNEs are highly enriched for genomic regions that are regulating the development of the considered trait. The retina and lens time point specific elements show only a medium enrichment in their overlap with the top-ranking CNEs, due to their very restrictive construction as regions that are neither active in another vision-related tissue nor at a different developmental time.

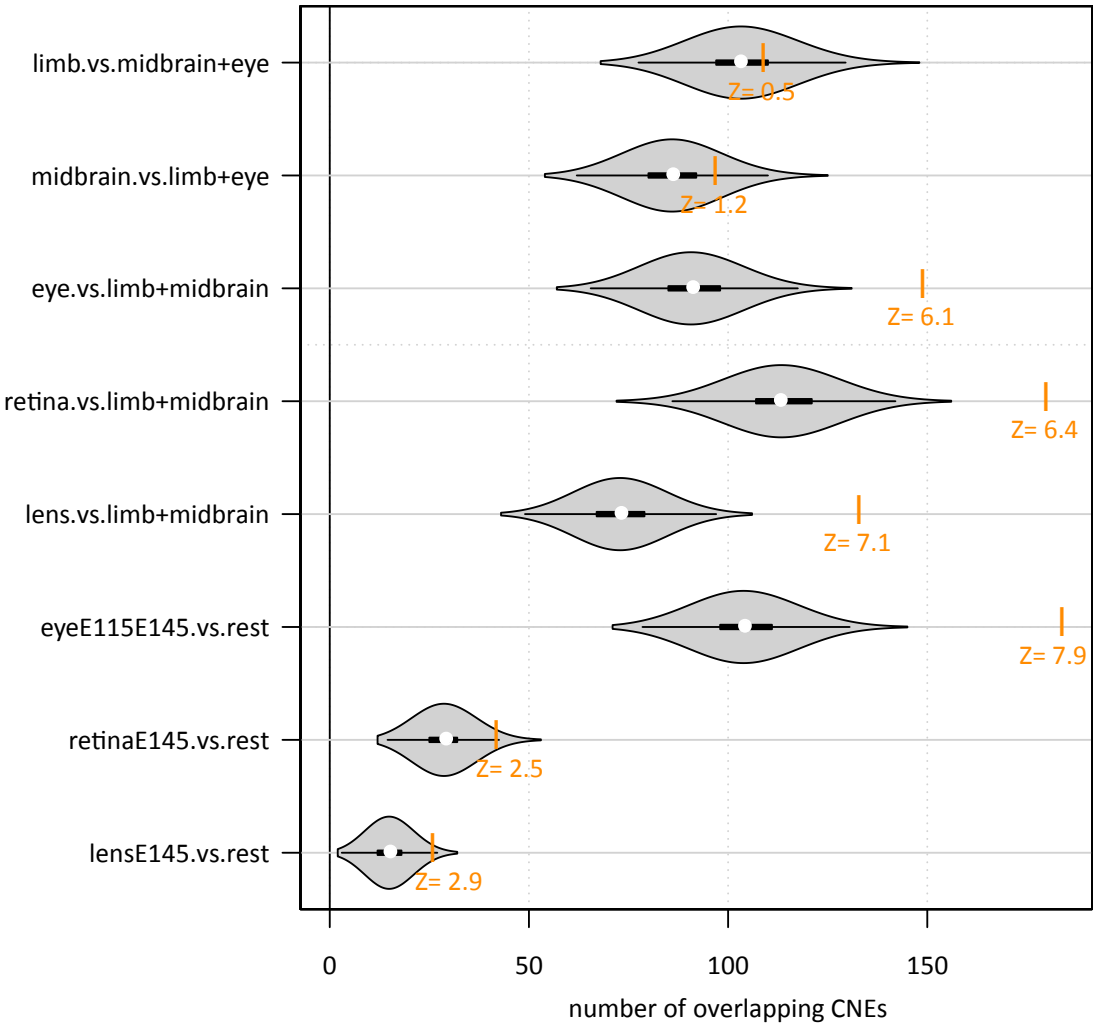


Figure 2.17: REforge's performance on real data
The overlap of different sets of tissue- and developmental time-specific ATAC-seq element with the top 9364 CNEs computed via REforge is shown (orange) in comparison to the overlap, that is expected by chance with a 9364-element subset of all 351279 CNEs (grey violin plots). The enrichment is assessed by the z-score. The underlying score were computed with a window of 200 bp, NSC, AFTP and ancestral element filtering on the ancestor of placental mammals. The element ranking is based upon Cohen's D.

2.5.3 Scoring method comparison

The next step is to reassess and confirm the results obtained from synthetic data with respect to the scoring methods and the ancestral filtering of scores, on real data. This analysis is done on a set of 100.000 CNEs to which REforge is applied with all seven different scoring methods variants, each with and without ancestral filtering. The comparison of the resulting rankings is done via an overlap of the top 2500, top 5000 and top 10000 CNEs with the ATAC-seq regions sets, analog to Chapter 2.5.2. As shown exemplarily for REforge with NSC, AFTP and ancestral element filter in Table 4, enrichments are generally independent of the chosen cutoff. In matters of the scoring method one can observe that Stubb with NSC generally outperforms Stubb with MSF and Stubb with NTPC (see Figure 2.18 for the top 5000 CNEs and Appendix A.I.3 for results on the top 2500 and 10000 CNEs). Moreover, every method shows

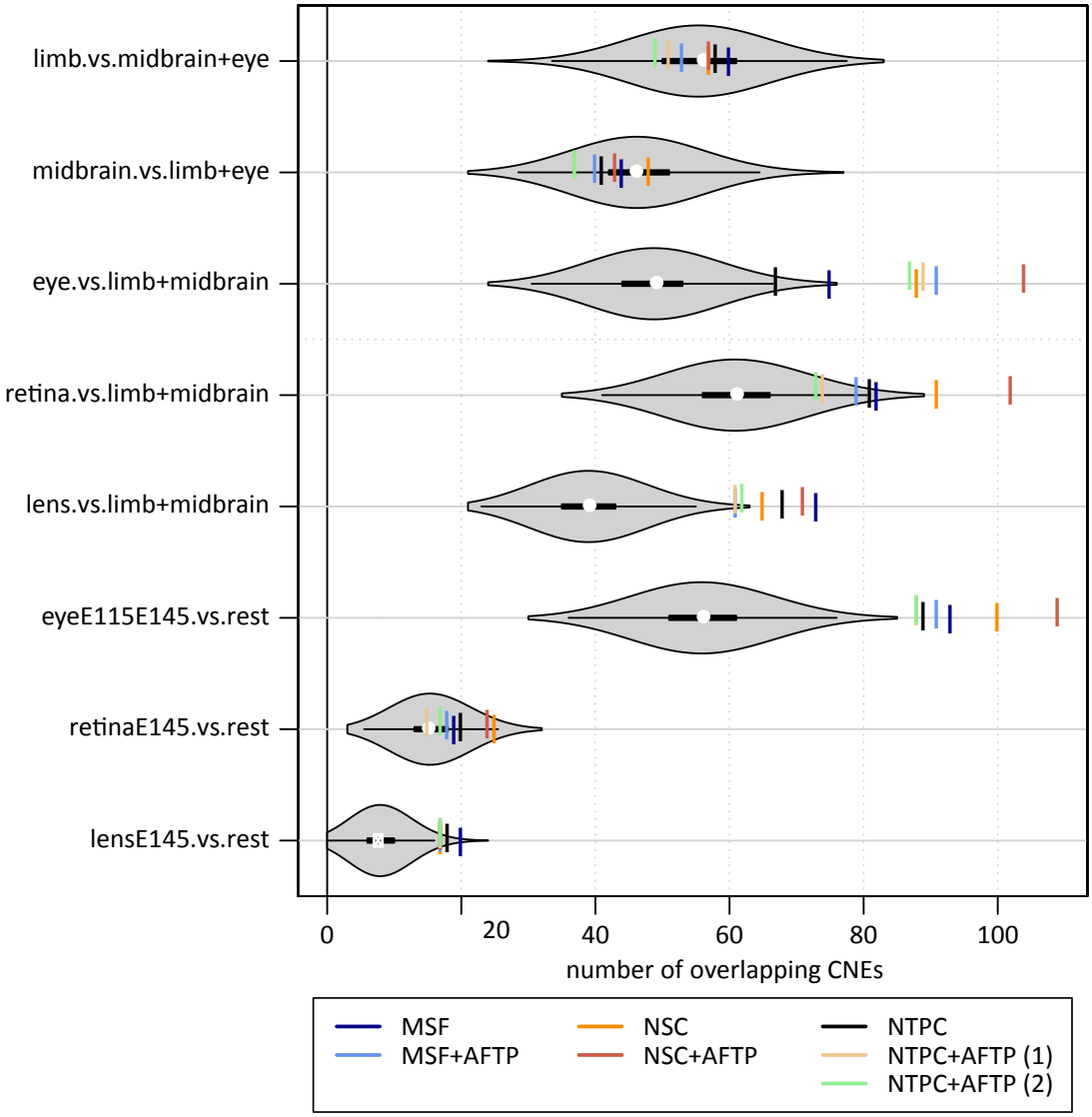


Figure 2.18: Scoring method influence on REforge's performance on real data
The effect of different scoring methods on REforge's performance is compared. The top 5000 CNEs from the resulting rankings with Cohen's D were overlapped with multiple tissue- and developmental time-specific ATAC-seq element sets. This overlap is shown in comparison to the by chance expected overlap (grey violin plots). A scoring window size of 200 bp was used for every method.

ATAC-seq dataset	*Top2500*	*Top5000*	*Top10000*
eyeE115.vs.limb+midbrain	7.23	6.24	7.43
retinaE145.vs.limb+midbrain	4.11	4.81	6.27
lensE145.vs.limb+midbrain	5.74	5.28	6.34
eyeE115E145.vs.rest	5.85	6.56	8.06
retinaE145.vs.rest	1.53	2.92	3.44
lensE145.vs.rest	4.51	4.24	3.75
limbE115.vs.rest	0.6	0.19	-0.78
midbrainE115.vs.rest	-0.43	-0.34	-1.52

Table 4: Enrichment of top-ranking CNEs in tissue-specific ATAC-seq elements
Enrichments, indicated as z-scores, of the top-ranked CNEs in overlap with various mouse tissue-specific ATAC-seq elements. The scores were computed with NSC and AFTP and CNE's were, after ancestral filtering, ranked with Cohen's D

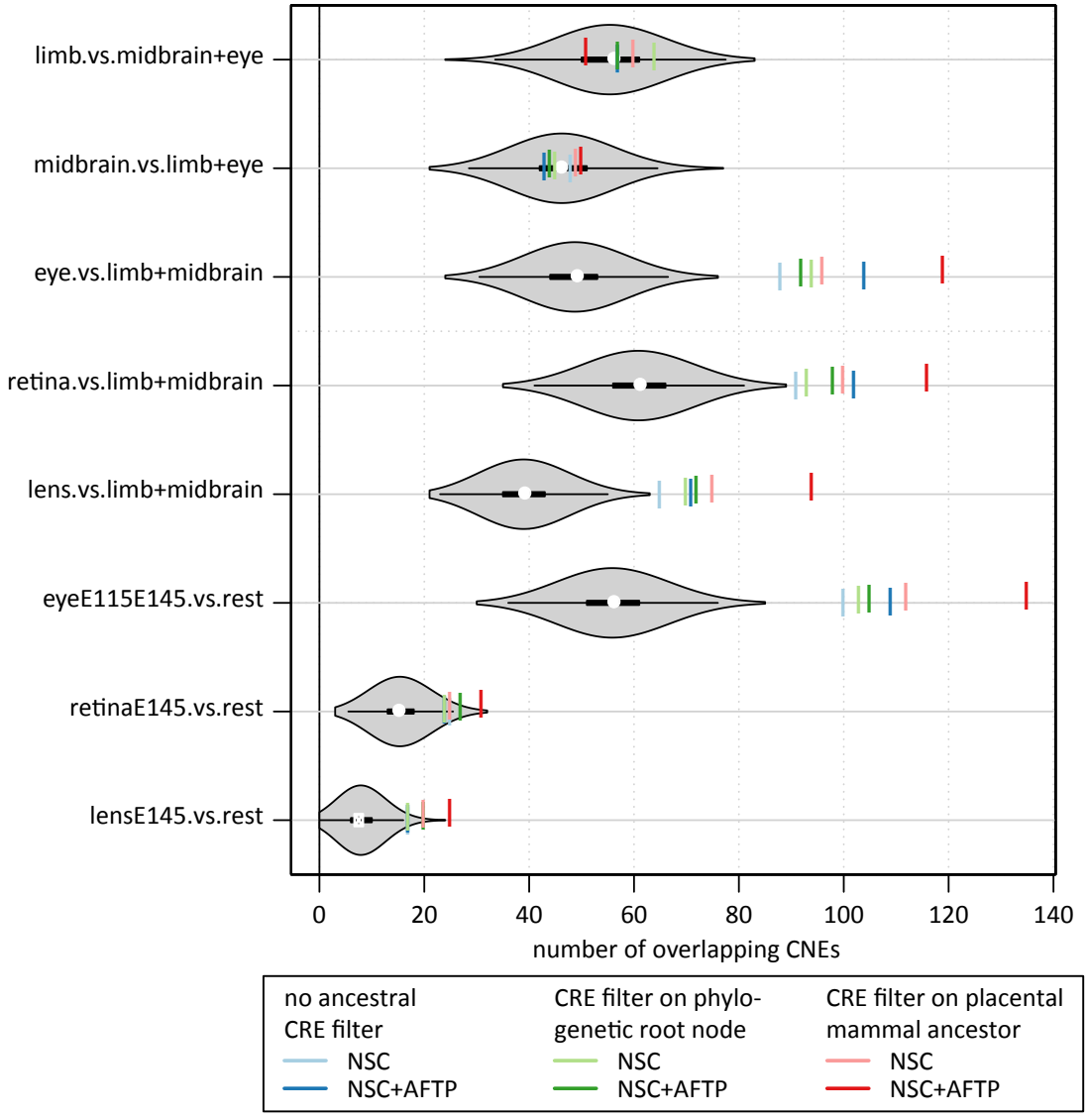


Figure 2.19: Ancestral element filtering influence on REforge's performance on real data
The effect of an ancestral element filter based on the ancestor of placental mammals (red) and the oldest possible ancestor (green) is compared to REforge's performance without an ancestral element filter (blue). Scoring was conducted with a window size of 200 bp, NSC and with (dark) or without (light) AFTP. The top 5000 CNEs from the resulting rankings with Cohen's D were overlapped with multiple tissue- and developmental time-specific ATAC-seq element sets. These overlaps are shown in comparison to their expectation (grey violin plots)

a slightly better performance if the transition probabilities are ancestrally fixed (AFTP). Fixation of the transition probabilities to the ancestrally optimal values is equivalent to the assumption that the relative TFBS occurrence frequencies remain rather constant throughout the evolution. This in particular effectively reduces the set of considered TFs to these that bind the ancestral sequence. Hereby, the influence of randomly created and destroyed TFBSs is minimized, an effect that becomes more likely with increasing TF motif set size. On the other hand, this approach misses regulatory elements in which the function of one TF has been taken over by another one, even if both TF belong to considered set of relevant TFs. My expectation, also supported by the results of the score method comparison, is that this happens only rarely.

An additional step beyond the ancestral sequence-based restriction of the TF set is the application of an element filter based upon the ancestral sequence's score. There is a clear trend for ancestral CRE filtering to improve REforge's performance, independent of the additional use of AFTP as shown for the top 5000 CNEs in Figure 2.19 and for the top 2500 and

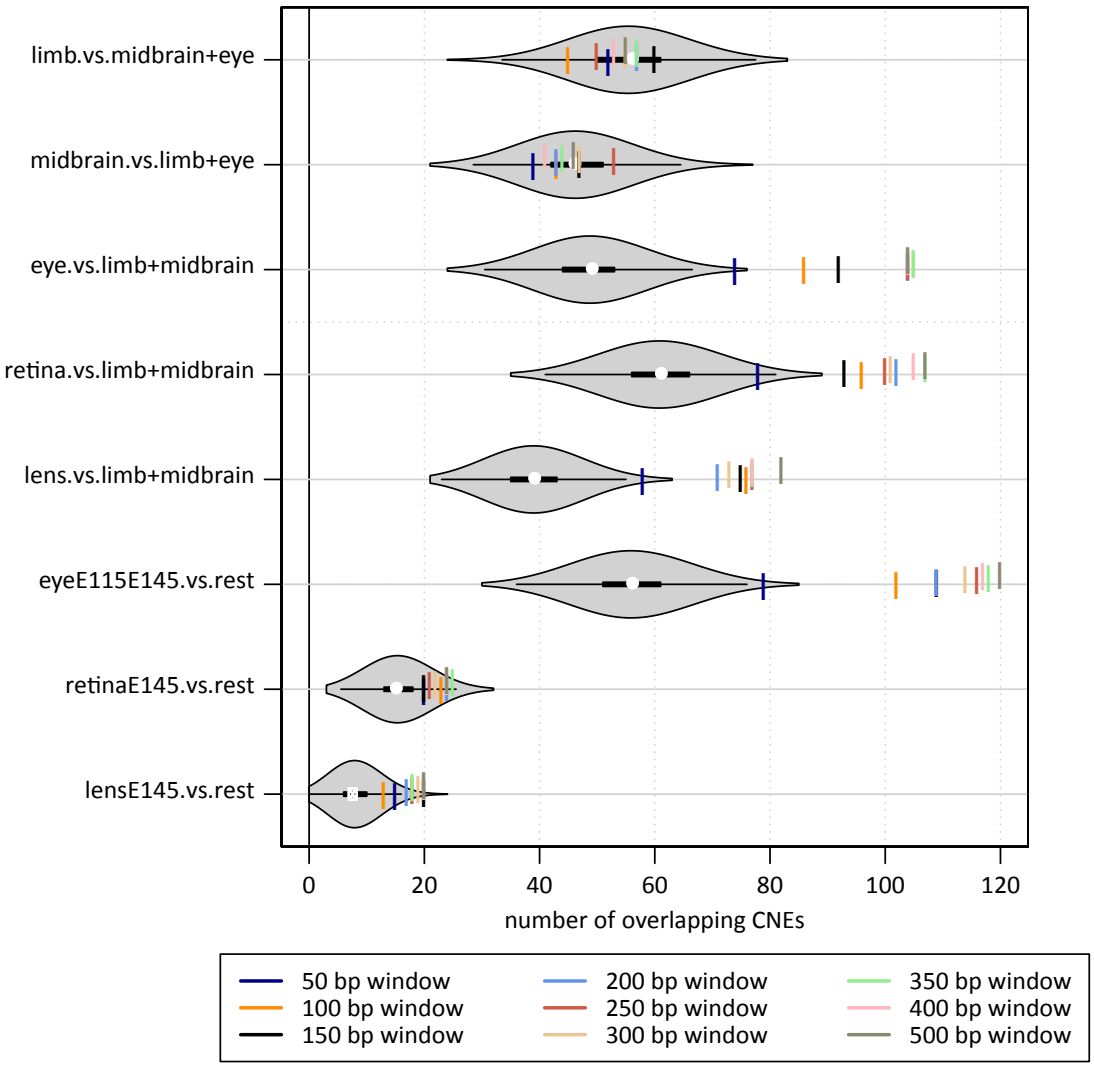


Figure 2.20: Scoring window length influence on REforge's performance on real data
The effect of different scoring windows on REforge's performance is compared. The scoring was conducted in every case with NSC and AFTP. The top 5000 CNEs from the resulting rankings with Cohen's D were overlapped with multiple tissue- and developmental time-specific ATAC-seq element sets. This overlap is shown in comparison to the by chance expected overlap (grey violin plots).

top 10000 CNEs in Appendix A.I.4. This improvement arises from minimizing the chance of false positives due to score fluctuations from randomly gained and lost TFBSs. The additional performance gain from AFTP is presented by Figure 2.19 again.

A comparison of the performance in dependence of the scoring window length (see Figure 2.20 for the top 5000 CNEs and Appendix A.I.5 for top 2500 and top 10000 CNEs) suggests, that the scoring window should be chosen such that the majority of elements are scored in their entirety. The set of 100000 CNEs, that are considered for this analysis ranges in length from 50 bp to 2883 bp (median 131, third quartile 222).

An analogous comparison of the results based upon different ranking methods (see Figure 2.21 for top 5000 CNEs and Appendix A.I.6 for the top 2500 and top 10000 CNEs) validates the Pearson correlation coefficient and Cohen's D as the best ranking methods)

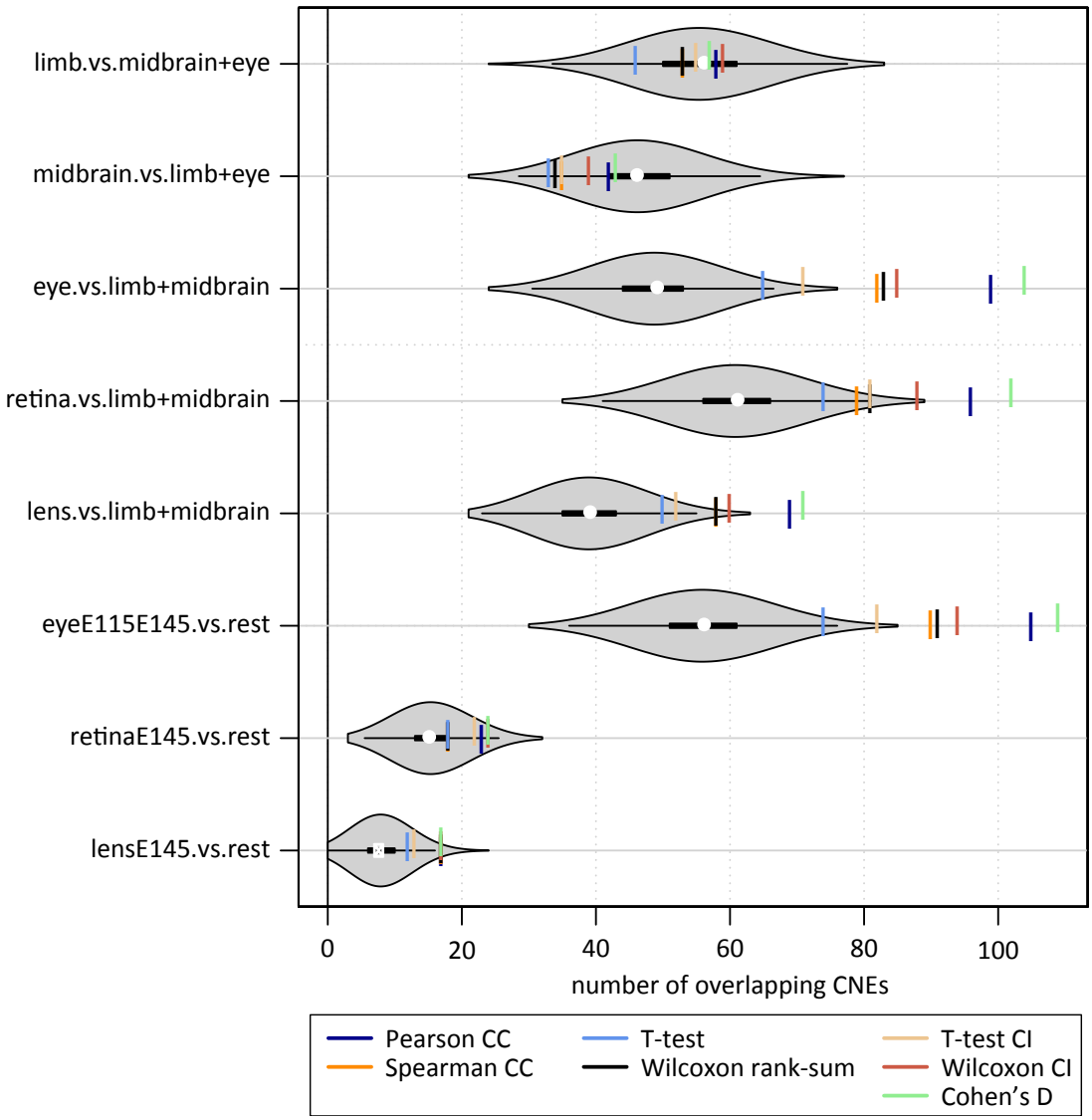


Figure 2.21: Ranking method influence on REforge's performance on real data

The effect of different ranking methods on REforge's performance is compared. The scoring was conducted with NSC and AFTP and a window size of 200. The top 5000 CNEs from the resulting rankings with each method were overlapped with multiple tissue- and developmental time-specific ATAC-seq element sets. This overlap is shown in comparison to the by chance expected overlap (grey violin plots).

2.5.4 Comparison to Forward Genomics

I will in the following compare Forward Genomics to the new REforge method via their performances on both synthetic data and the vision-impairment dataset. While REforge needs for any CRE only the sequences of every species (including the ancestral species), Forward Genomics needs these sequences in an aligning format, to be able to compute the sequence identity. Since the CRE evolution simulation provides only sequences but no alignment, I created a sequence alignment for every synthetic CRE with PRANK (Löytynoja 2014). In order to have PRANK considering the known ancestral sequences (which is an uncommon use case), I manipulated the phylogenetic tree by introducing for every ancestral species two additional internal node with connecting branches of length zero, according to Figure 2.22. Thereby the original ancestral node becomes a terminal node for which the sequence can be given to PRANK. Because of the branch lengths of 0, this sequence is "inferred" for the newly introduced internal nodes and thereby is set as ancestral sequence of the corresponding subtree.

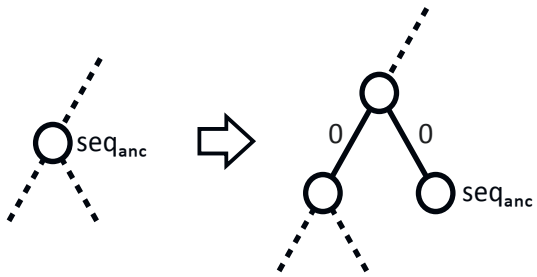


Figure 2.22: Tree manipulation
Insertion of internal nodes with connecting branches of length 0 allows incorporation of known ancestral sequences into an alignment, generated with PRANK.

The comparison of Forward Genomics and REforge is done both on the long and the short trait-loss scenario. The long trait-loss scenario is expected to be the "easier" one for both methods, since it implies a high sequence divergence along the trait-loss branches and therefore the highest signal-to-noise ratio. As the ROC curves in Figure 2.23 show, REforge greatly outperforms Forward Genomics. REforge's outstanding performance on these datasets was already shown in Chapter 2.4. The reason for Forward Genomics bad performance is the sequence divergence in trait-preserving species that is as high as for the trait-loss species. This follows from the simulation setup, which sets only a fraction (namely the TFBSs) of the sequence under constraint.

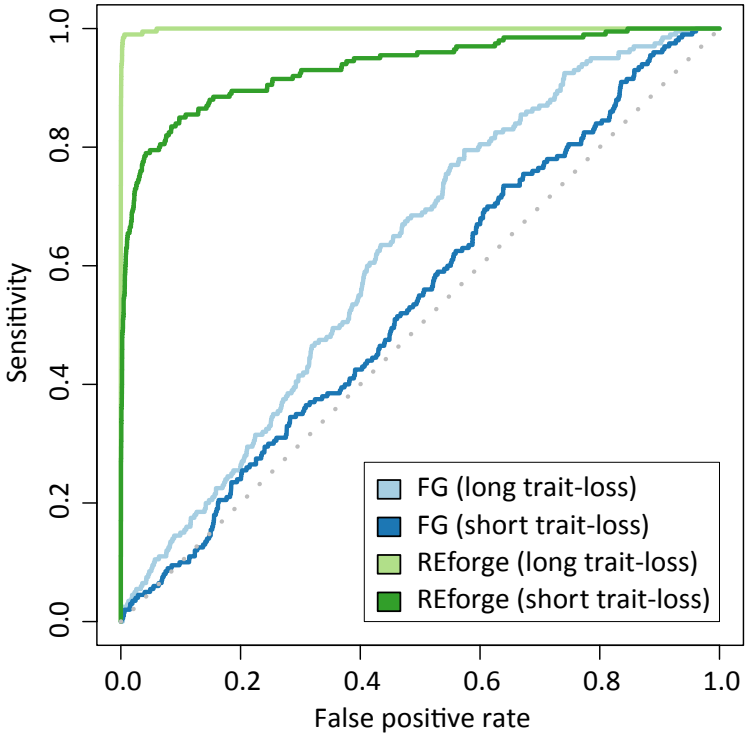


Figure 2.23: REforge versus Forward Genomics on synthetic data
ROC curve for REforge (200 bp scoring window, NSC, ranking with Cohen's D) (green) and Forward Genomics (blue) for distinguishing 200 relevant from 10000 non-relevant CREs. The set of relevant CREs evolved neutrally for a short (dark) or long (light) time

For a comparison of both methods on real data, REforge is applied to the same dataset of 351279 CNEs from which (Roscito, et al. 2018) identified 9364 diverged CNEs in the vision-impaired species. Therefore, also the top 9364 CNEs from REforge's ranking with Cohen's D, after a scoring with NSC, AFTP and ancestral element filtering on the placental mammals' ancestor, are used. For both of these CNE sets, the overlaps with the tissue-specific ATAC-seq elements sets are computed. As shown in Figure 2.24, REforge clearly outperforms Forward Genomics in terms of enrichment of the top CNE set in all vision-tissue related ATAC-seq element sets. An exception is the lens time point-specific set with respect to which both methods perform equally.

A remarkable difference is Forward Genomics' depletion in CNEs overlapping limb and midbrain-specific elements. This likely arises from sequence divergence constraints of the limb and midbrain regulatory landscape, that cause a higher sequence identity of the limb- and midbrain CREs also in the vision-impairment species and thus a low Forward Genomics ranking. Since REforge does not rely on sequence identity but only on TFBS preservation and only vision TFs and no limb or midbrain TFs were considered, such a depletion for the two tissues is neither expected nor observed.

Interestingly, the intersection of the top Forward Genomics and the top REforge CNEs is with 687 CNEs relatively small, compared to the absolute sizes of both sets, yet more than the by chance expected overlap of $\frac{9364^2}{351279} \approx 250$ CNEs. This reflects the different methodologies of both methods, which are nevertheless not independent from each other. The biological importance of this intersection remains to be explored.

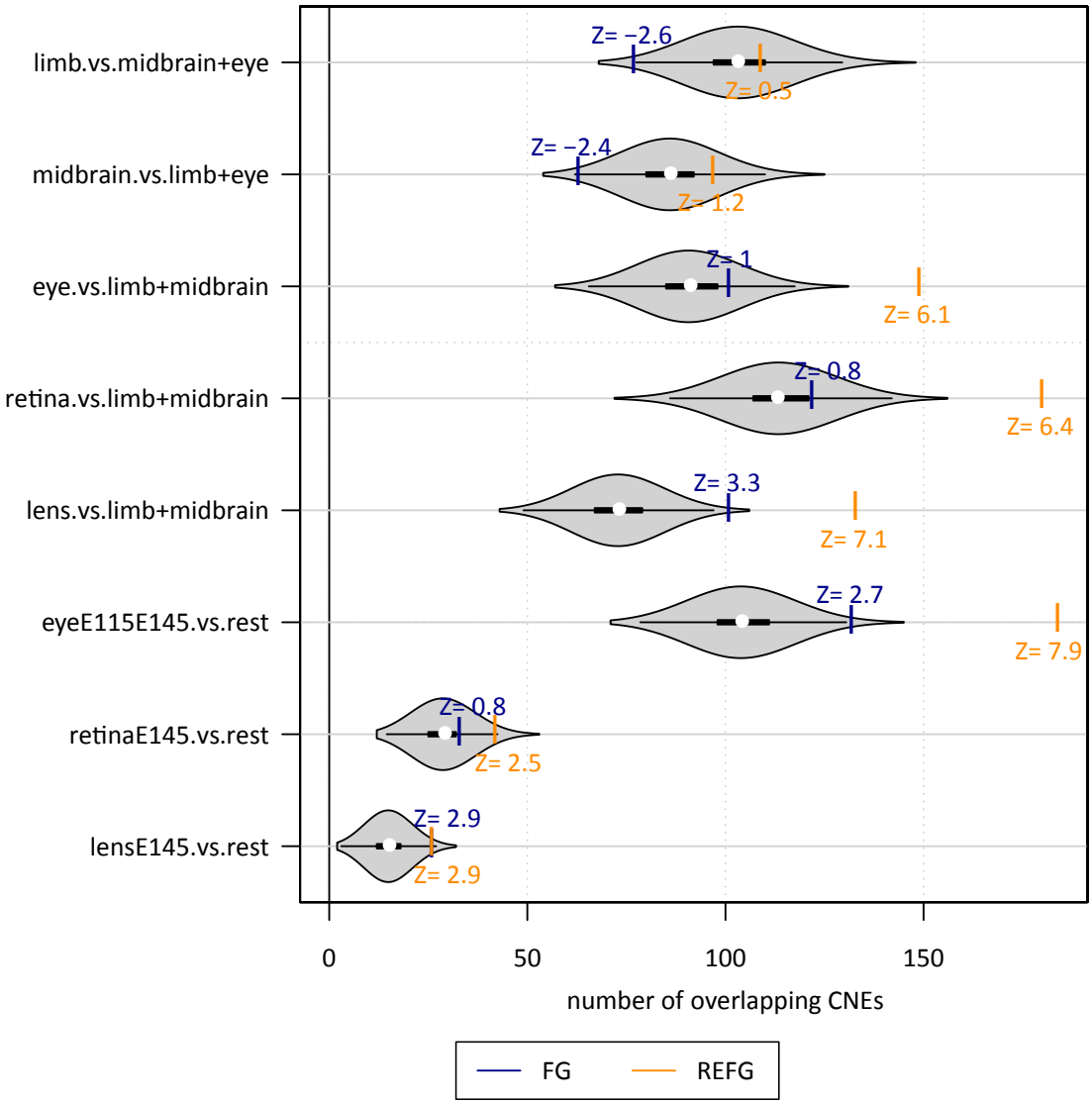


Figure 2.24: REforge versus Forward Genomics on biological data
Violin plots show the number of elements of a randomly chosen 9364-element subset of all 351279 CNEs that overlap the differentially enriched ATAC-seq elements in mouse limb, midbrain, eye (E11.5), retina (E14.5) and lens (E14.5) in comparison to each other. In contrast are the enriched overlaps of the top 9364 identified CNEs identified by FG (blue) and REforge (orange) indicated. REforge's underlying scores were computed with a window of 200 bp, NSC, AFTP and ancestral element filtering on the ancestor of placental mammals. The element ranking was based upon Cohen's D.

2.6 Summary & Discussion

I developed REforge as a novel computational approach to discover CREs that are relevant for the development of a specific phenotype by incorporating knowledge of phenotype-relevant TFs into the rationale of Forward Genomics and, therefore, by scoring the corresponding motifs in a phylogeny-aware and discriminative manner. Thus, REforge's novelty arises from the consideration of TF binding knowledge as the "grammar" that underlies the CRE's function. This means, REforge identifies relevant CREs by their TFBSs that are lost in trait-loss species and preferentially preserved in trait-preserving species. The examination of TFBSs allows the abstraction from the sequence level towards the functional

level of CREs and therefore results in a more accurate association with the analyzed phenotype. REforge is designed as a flexible algorithm that can be applied to any set of species and any binary phenotypic change, given a genome alignment of the species, their underlying phylogeny, a set of motifs belonging to phenotypically relevant TFs and a set of putative regulatory regions. In order to rank these CREs according to their putative relevance with respect to the phenotype, REforge conducts four steps:

- Assessing the binding affinity of the TF set to every species' sequence of the CRE
- Transforming sequence scores into phylogenetically independent branch scores
- Categorizing branch scores into "trait-preserving" and "trait-loss"
- Ranking the CREs by relevance with respect to the phenotype

I verified REforge and optimized its performance with respect to various variants on synthetic data, which was generated by a CRE evolution simulation that I built for this purpose. Furthermore, I validated REforge on biological data of the example of the vision-impairment phenotype in subterranean mammals. The results show, that REforge considerably outperforms Forward Genomics, the current approach for this research question, both on synthetic and real data.

Recently, another study examining the same phenotype with respect to both genes and gene regulatory regions has been published (Partha, et al. 2017). Their method follows in its essence Forward Genomics' principle too, with the difference of comparing trait-loss and trait-preserving species based upon an element-specific phylogenetic branch length estimation instead of the sequence identity. Nevertheless, this approach relies on sequence divergence and, therefore, their method will still suffer the same drawbacks as Forward Genomics with respect to analyzing regulatory regions.

Hence, by the TFBS scoring REforge facilitates the identification of many more relevant CREs with a higher precision compared to currently available tools. This allows the avoidance of time-consuming and costly high-throughput experimental screening approaches and allows to narrow down the dataset of regulatory regions to a high-quality candidate CRE set for experimental validation.

Yet, in order to apply REforge, phenotype relevant TFs and their binding preferences need to be known. Unfortunately, such knowledge is not always available or complete. An idea to approach this problem is the identification of relevant TFs via the detection of a preferential global loss of their binding sites in the trait-loss species. The rationale here is in accordance to REforge to detect motifs of relevant TFs from putative regulatory regions. Such putative regulatory regions could be for example conserved non-coding regions, genomic regions of open chromatin in the corresponding tissue or genomic regions that are bound by a known TF for which co-factors shall be identified. Another tool, implementing this perspective is described in the following Chapter.

3 TFforge

TFforge (Transcription Factor forward genomics) aims for the identification of TFs that are relevant for an analyzed phenotype, by a TF binding-based comparative genomics approach. The given data consists of a set of candidate TF motif library, a set of CREs that contains (but not necessarily consists of) trait-relevant CREs and a genome alignment, including species that lost the phenotype during their evolution. To achieve this aim, TFforge follows the basic Forward Genomics principles (see Chapter 1.1), applied on TFBS:

conceptual idea

The loss of selective pressure, due to the loss of a complex phenotypic trait, upon regulatory regions, which are specific to this trait, allows for the accumulation of mutations. Thus, trait-relevant TFBSs decay during the course of evolution in trait-loss species at a higher rate than other binding sites underlying selection, which are only randomly lost (and gained) due to turnover. Given a set of more or less trait-specific regulatory regions we want to identify such TFs, whose binding sites show a differential loss pattern in trait-loss and trait-preserving species. Therefore, a library of TF motifs is screened to identify these motifs that show a global loss of binding sites (with respect to the given regulatory elements) in the trait-loss species while showing binding sites preservation in trait-preserving species.

specific idea

The procedure to identify such differentially lost TFBSs is analogous to REforge's identification of trait-relevant regulatory elements. This means, assuming the ancestral sequences are already inferred, every sequence is scored with respect to the considered TF motif (as described in Chapter 2.1). Since TFforge analyzes all motifs separately, every sequence is scored with respect to *only one TF motif at a time* instead of all given motifs together, like REforge's sequence scoring. The next step is the computation of the phylogenetically independent branch scores (see Chapter 2.2) from these sequence scores. This gives a measure of motif gains and losses during evolution. After an inference of the ancestral phenotypes, these branch scores are categorized into scores of "trait-preserving" and "trait-loss" branches. The final evaluation of the relevance of a TF motif is based on the shift between the trait-loss and the trait-preserving branch score distribution, essentially analogous to REforge's ranking method, described in Chapter 2.3. Yet, contrary to REforge, which considers the branch scores that belong to a single element, TFforge compares the trait-loss and the trait-preserving score distribution that comprehend the branch scores obtained from *all given (putative) CREs*. If the trait-loss score distribution is statistically significantly shifted towards negative scores, it can be concluded that the analyzed TF lost binding sites preferentially along the trait-loss branches compared to the trait-preserving branches. This implies relevance of the TF for the given trait.

To summarize, the TFforge method ranks every TF motif with respect to its trait relevance, computed from the differential binding site losses. It follows, that a preferably exhaustive collection of TF motifs is needed, as this allows for an accurate TF identification. For this, a mapping from the motifs to actual TFs is additionally necessary to allow for the investigation of the biological meaning of the identified binding sites in a later stage. I start, therefore, prior to the establishment of the algorithm, with the compilation of a collection of TF motifs from

public databases (see Chapter 3.1). Next, I give a proof of principle of the method on synthetic data (see Chapter 3.2) and evaluate, analogous to the proceedings for REforge, the effects of scoring and ranking variants on the TFforge performance (see Chapter 3.3) as well as its performance on biologically relevant dataset variants (see Chapter 3.4). Finally, I validate TFforge on real data (see Chapter 3.5) and conclude with a summary the method and the findings (see Chapter 3.6).

3.1 Transcription factor library

In order to compile a TF library, I aggregated TF motifs from three widely used TF databases:

UniPROBE (Universal PBM Resource for Oligonucleotide-Binding Evaluation)

UniPROBE (Hume, et al. 2015) is database of protein-binding microarray data on protein-DNA interactions. PWMs can be downloaded from UniPROBE's website[1] which stores every PWM both in forward and reverse complement orientation. Since Stubb assesses both DNA strands, the latter ones are redundant and were skipped. To identify the TF that corresponds to a motif, UniPROBE offers links to other databases, that generally include the UniProt ID or Swiss-Prot ID. This information was obtained by crawling the TF detail pages found on the website[2], which contained 566 TF entries at the time of accession (April 2016).

TRANSFAC (TRANScription FACtor database)

The TRANSFAC database (Matys, et al. 2006) is a database for eukaryotic TFs, including their binding profiles. TRANSFAC Pro 2014.3 provides in total 5474 PWMs, half of which are motifs of vertebrate TFs. From these I used the motifs that either have a statistical base of at least 20 sequences or were 3D structure-based computed. TRANSFAC offers a variety of different external database identifiers.

JASPAR

JASPAR (Mathelier, et al. 2014) is an open-access database of TF binding profiles for TFs from species in six different taxonomic groups. I use the collection of non-redundant vertebrate TFs, which can be downloaded from the (recently updated) website[3]. At the time of accession (April 2016) this set contained 519 motifs. The website also offers a complete SQL dump, which allows the extraction of UniProt IDs for all PWMs.

All of the obtained matrices were normalized subsequently to convert all frequency matrices to probability matrices. After removing all low-quality matrices by filtering for an information content above five bits, the remaining matrices were converted into Stubb's weight matrix input file format. The next step was to convert the different obtained database identifiers to Ensembl IDs (described in the following Chapter 3.1.1). The filter of PWMs for which this step was possible left 2272 matrices from which I chose via clustering 638 PWMs

[1] http://the_brain.bwh.harvard.edu/uniprobe/downloads/All/All_PWMs.zip

[2] http://the_brain.bwh.harvard.edu/uniprobe/browse.php

[3] http://jaspar.genereg.net/download/CORE/JASPAR2018_CORE_vertebrates_redundant_pfms_jaspar.txt

(see Chapter 3.1.2). The cluster step is not necessary but reasonable to reduce the computational effort of TFforge, since very similar PWMs will result in very similar results and will therefore obtain subsequent ranks.

3.1.1 PWM to Ensembl ID Mapping

After an application of TFforge to identify a set of relevant motifs with respect to a certain phenotype, a major point of interest will be the corresponding set of TFs and their possibly already existing functional annotations. In order to allow the creation of this link in an easy manner, a mapping of the PWMs to a consistent gene database identifier is desirable. I chose the Ensembl ID as reference identifier as Ensembl is a centralized database for genomic studies with a wide range of gene annotations (Yates, et al. 2015). Since the databases offer different external database identifiers but not consistently, I acquired Ensembl IDs and additionally Swiss-Prot or UniProt IDs for every PWM if possible. The latter ones can be translated to Ensembl IDs via the *idmapping_selected.tab* table from UniProt's website[1]. PWMs for which this mapping to Ensembl IDs as a gene identifier was not possible with the described approach, due to missing data, were dismissed.

3.1.2 Motif similarity and clustering

In order to cluster redundant TF motifs and motifs of similar TFs, I computed pairwise PWM similarity scores via the motif comparison tool Tomtom (Gupta, et al. 2007) with the Euclidian distance as distance measurement (parameters: "-thresh 1", "-dist ed"). Motif clusters were then defined such that each of the pairwise similarity scores between all cluster members are below the threshold 0.0001. This was done with a greedy approach, starting with the most similar motifs. To illustrate this, the logos of the two most similar not-clustered motifs are shown in Figure 3.1. The PWM that was most similar to all other PWMs of its cluster, is used as cluster representative. By this approach the original PWM set of 2272 motifs is reduced to my final motif library consisting of the 638 clusters representatives.

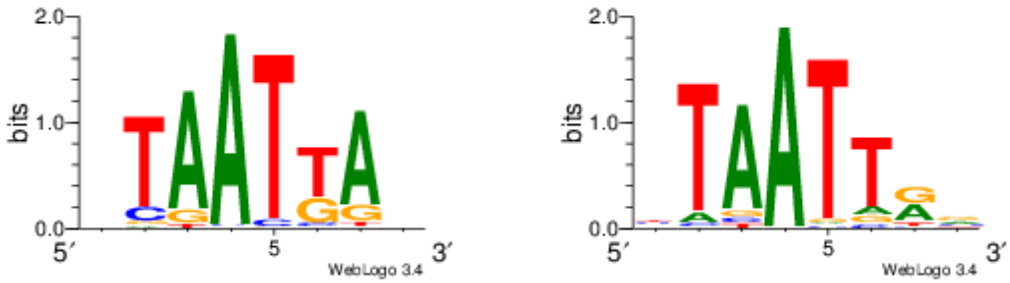


Figure 3.1: Example of similar motifs
The binding motifs of TF MEOX2 (left) and SHOX (right) are the two most similar motifs that, with a Tomtom score of 0.000100008, did not pass the clustering threshold of 0.0001.

[1] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping_selected.tab.gz

To evaluate the performance of TFforge on synthetic data, I refined the motif library. The set of relevant TFs is already given by the five TF motifs that were used in the CRE evolution simulation to create the synthetic dataset. The refined library should, therefore, represent the set of irrelevant TFs. For this purpose, I further filtered all motifs from the motif library that are similar to the "target TF set" of the CRE evolution simulation with a stricter threshold of 0.01. The obtained "background set" consists of 567 motifs.

3.2 TFforge's proof of concept

The proof of concept and a verification of TFforge is, just like for REforge, based upon synthetic data, for which the ground truth is known in terms of correct ancestral sequences and trait-involved TFs. I use the "Foreground-Set" (see Chapter 2.4.2), that consists of 1000 phenotype-specific CREs with corresponding sequences in 20 species including three species, that lost the phenotype a "long" time ago. As previously described (see Chapter 2.4.1), five TF were used for the CRE evolution simulation, which represent the "positive set", whereas the "negative set" is constituted of the 567 motifs of the library that do not show strong similarity with the "positive set". Upon applying TFforge onto the CRE set and ranking all 572 motifs via the upper bound of the t-tests confidence interval, I find that all five relevant TF motifs can be clearly distinguished from the other motifs (precision and sensitivity 100%). All relevant TF motifs have a rank score < -2.23 whereas the others rank score is > -0.35 (see Figure 3.2). The same result of a clear margin between both TF sets is obtained also with different ranking methods like Cohen's D, which is for all relevant TF motifs < -2.06 and all other motifs > -0.57 (see Appendix A.II.1). This shows, in summary, that TFforge is able to detect relevant TF motifs with a high precision by their differential binding site loss.

3.3 Technical and methodological variants

The provided proof of concept suggests a great discriminative power of TFforge. With the aim of maximizing this power, I assess the influence of TFforge's underlying sequence scoring method (see Chapter 3.3.1) and ranking method (see Chapter 3.3.2) on its performance. These analyses are conducted on the set of 1000 synthetic CREs (see Chapter 2.4.1), that are generated according to a long trait-loss scenario. Each analysis ranks the same set of TFs, consisting of five relevant and 567 irrelevant TFs, that were used for the proof of principle.

Score correction method	$avg\,\Delta TCI$	ΔTCI
MSF	-0.435	1.835
NSC	1.527	2.274
NTPC	0.450	0.911
MSF + AFTP	-0.172	0.527
NSC + AFTP	1.238	1.889
NTPC + AFTP (1)	0.461	0.942
NTPC + AFTP (2)	0.518	0.945

Table 5: Score correction method comparison The column "avg ΔTCI" summarizes the average differences in the t-test's upper confidence interval bounds, computed on 100 sets of 100 randomly chosen CREs, while the column "ΔTCI" refers to the score difference computed on all 1000 CREs. All CREs belong to the long trait-loss scenario set. The used score correction methods are described in Chapter 2.1.3.

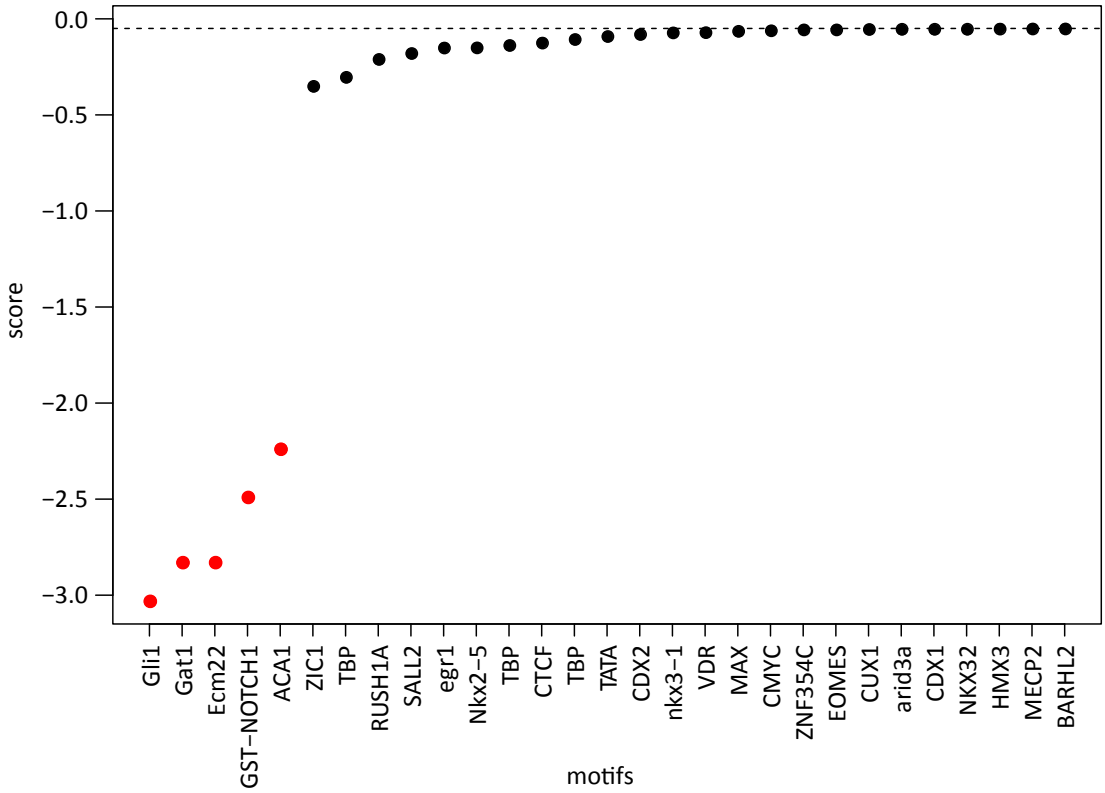


Figure 3.2: Motif ranking with TFforge
Shown are in ascending order of ranking score the TF motifs, whose score < -0.05. The ranking bases on the t-tests confidence interval. The underlying data are 1000 synthetic CREs with a long trait-loss age scored by Stubb with a 200 bp window, NSC and AFTP. Red indicates the five target TF motifs.

3.3.1 Sequence scoring method

TFforge differs from REforge with respect to the scoring setup in the sense that TFforge scores sequences always with respect to only a single motif. REforge, however, computes sequence scores with respect to a set of motifs. This possibly influences the score corrections methods (see Chapter 2.1.3) which are therefore reevaluated. In summary, the simple version of Stubb with minimal score filter (MSF) is compared to the correction methods NSC and NTPC and to all methods with ancestrally fixed transition probabilities (AFTP). This is done analogous to the proof of concept by assessing the t-test's confidence interval score margin between the "worst" true positive and the "best" false positive, i.e. the discriminative power between relevant and irrelevant TF set. Negative margins represent therefore the result of incorrect classifications by means of false positives in the top 5 ranking TFs. Table 5 summarizes these score differences computed on either all 1000 CREs or as average over 100 subsamples of size 100. Similar to the findings for REforge, a scoring with NSC with and without AFTP results in best the performance. In this analysis, NSC gives a slightly better performance without AFTP. This is expected by the insights from REforge, for which AFTP mainly effects unspecific CREs, i.e. CREs without ancestral binding sites (which are not present amongst analyzed CREs). However, REforge's benefit of the AFTP also arises from an effective reduction of the considered TF motif set for each element. Since TFforge considers from the beginning only one TF at a time, the AFTP method is not expected to bring performance improvements on other datasets, yet this remains to be tested.

Following the sequence scoring, multiple filter methods can be applied:

- *CRE filter*: "irrelevant" CREs, i.e. CREs without ancestral TFBSs are filtered
- *Branch filter*: as described in Chapter 2.2, scores of "non-informative" branches can be filtered. I apply this filter for all TFforge analyses by default to reduce the number of branch scores, considered in the following ranking.
- *TF filter*: TFs are required to retain after the previous filter steps a minimum of trait-loss and trait-preserving branch scores. If a TF's score count falls below the threshold, the TF is by default considered to be non-relevant. The idea behind this filter is that the decision of TF relevance shall be performed on a substantial number of CREs.

I analyzed the influence of the CRE and the TF filter with the aim to further improve the discriminative power of TFforge. The analysis bases upon the confidence interval bound margin between relevant and irrelevant motifs, which was also used before. The results shown in Figure 3.3 demonstrate a minor tendency to improve the discriminative power between both TF sets with the ancestral score based CRE filter. A bigger but less consistent influence shows the TF filter.

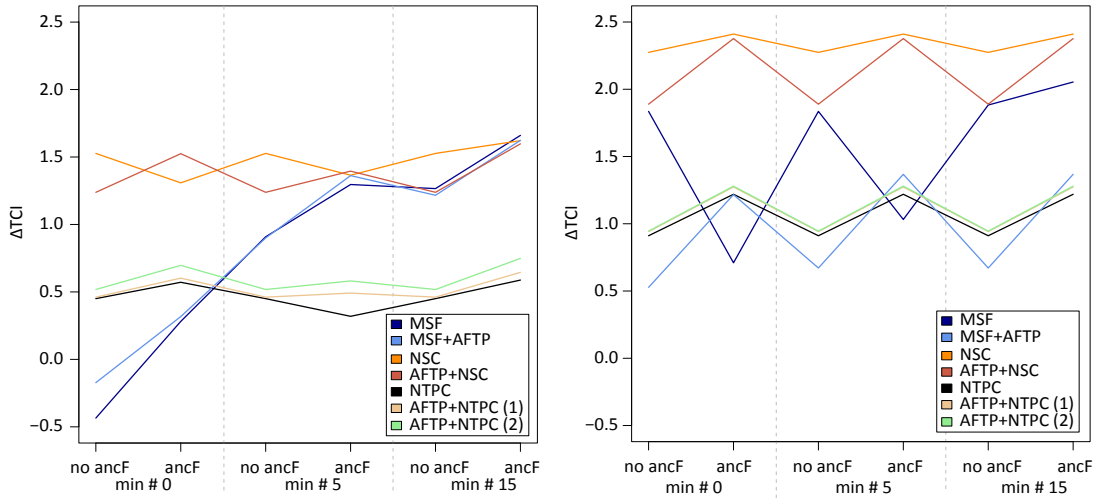


Figure 3.3: TFforge's discriminative power in dependence of different filters

The differences in the t-test's upper confidence interval bounds, computed and averaged over 100 sets of 100 randomly chosen CREs (left) and computed on all 1000 CREs (right) are shown for different filter combinations. The filters are the ancestral score-based CRE filter and the TF filter based upon the number of branch scores. All CREs belong to the long trait-loss scenario of the synthetic CREs.

Since the TF filter seems to potentially improve TFforge's performance substantially, I analyzed the gain in dependence of the number of CREs on a more fine-grained scale. Therefore, TFforge's performance in terms of sensitivity at a precision of 80% (corresponding to one irrelevant TF in the top set of predicted TFs) is assessed on different-sized subsamples of the synthetic CRE set, with either no TF filter or a requirement of at least 5 or 15 trait-loss and trait-preserving branch scores, each. As Figure 3.4 shows, a filter threshold of 5 generally increases the performance but not uniformly. Instead, the sensitivity on small sample sizes is more drastically affected than for bigger sample sizes. This trend is strengthened for a threshold of 15, albeit in case of the smallest sample size of 20 the threshold causes also true positives to be filtered. The conclusion is that the filter threshold should be chosen in dependence of the expected number of branch scores, which directly depends on the number of given elements and the phylogeny size.

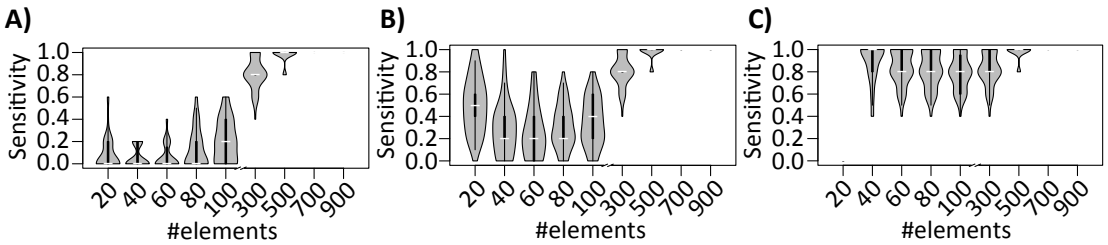


Figure 3.4: Influence of the count filter on TFforge's performance on synthetic data
(A) Sensitivity at precision of 80% in dependence of the number of considered CREs, scored with NSC, AFTP and ranked with Cohen's D. Additionally, TFs with less than 5 (B) or 15 (C) trait-loss or trait-preserving branch scores were filtered. All CREs belong the long trait-loss scenario set.

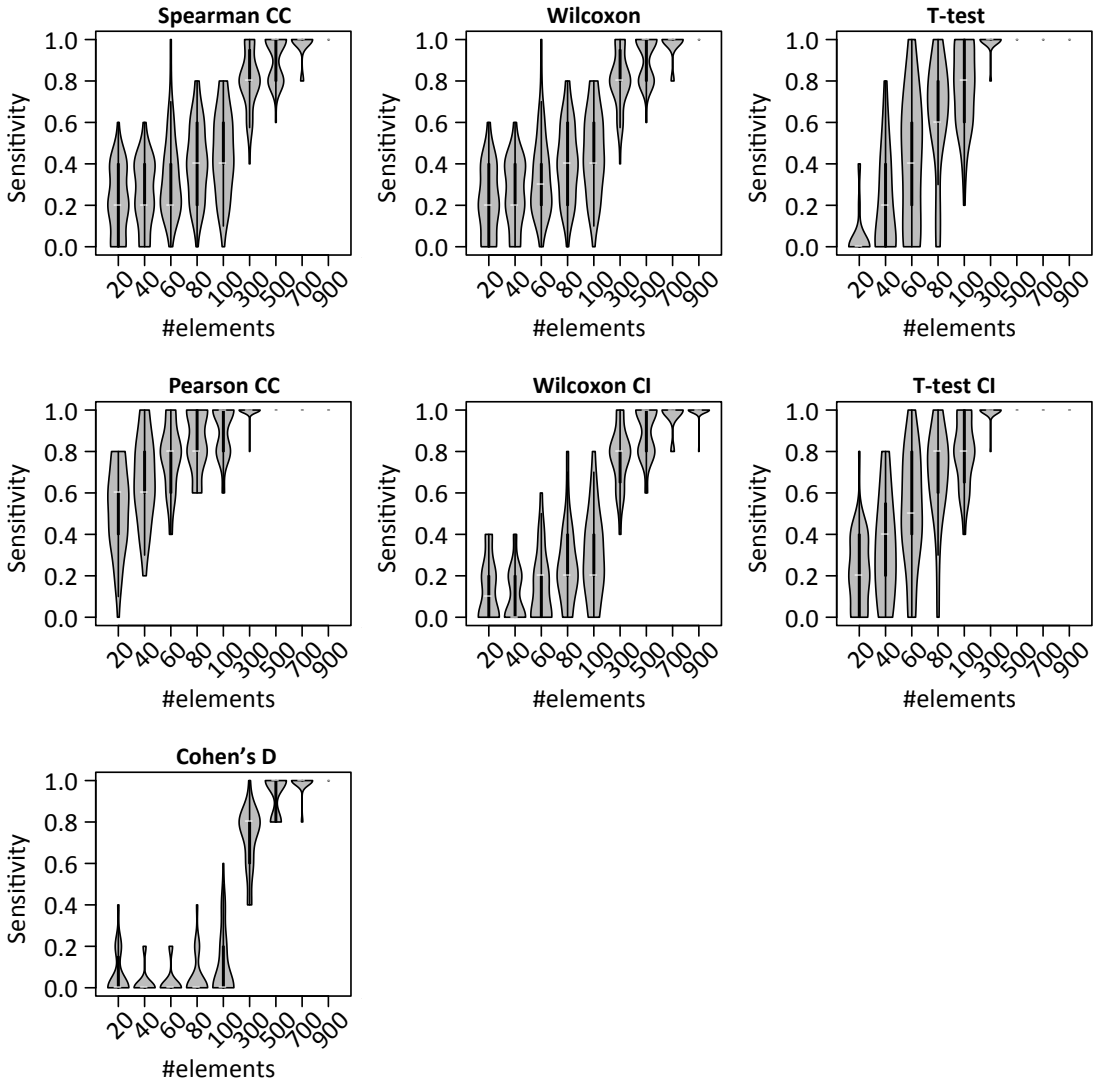


Figure 3.5: Influence of the ranking method on TFforge's performance on synthetic data
Violin plots show TFforge's detection sensitivity at a precision of 100% on synthetic data with short trait-loss age in dependence of the number of elements. The underlying ranking is indicated by the title of each plot. Abbreviated are correlation coefficient (CC), Wilcoxon rank-sum test (Wilcoxon) and the upper bound of the confidence interval (CI).

3.3.2 TF ranking method

Contrary to the REforge method, the number of scores considered during a single distribution comparison step is for TFforge not bound by the number of phylogenetic branches but by the number of phylogenetic branches times the number of elements. This additional magnitude of scores that underlie the trait-loss and trait-preserving branch score distributions, has likely an effect on the power of the statistical tests, with which both distributions are compared and the TF is finally ranked. Therefore, I reevaluate in the following the ranking methods for TFforge. In order to be able to observe performance differences, the comparison is performed on the "difficult" scenario of a "short" trait-loss and on differently sized subsets of the 1000 synthetic CREs. TFforge's performance is assessed via the sensitivity at a precision of 80% or 100%, hence the maximum number top-ranking TFs that contain one or no TF of the "negative set", respectively.

As shown for the sensitivity at precision of 100% in Figure 3.5, Pearson correlation coefficient's p-value and the t-test's upper confidence interval bound are the measures leading consistently to the best CRE ranking. The results on different trait-loss age scenarios with respect to the sensitivity at the precision of both 80% and 100% with a more fine-grained scale of numbers of elements can be found in Appendix A.II.2.

3.4 Biological relevant data variants

In the following, I test TFforge's performance under different biologically relevant variations of the input data. These variations include the number of given regulatory elements and the evolutionary time for which these elements evolved neutrally in trait-loss species (see Chapter 3.4.1). Additionally, pleiotropic elements are considered, i.e. elements that are relevant with respect to multiple phenotypes (see Chapter 3.4.2). In all scenarios, TFforge scores with NSC with AFTP and ranks the results via the Pearson correlation coefficient.

3.4.1 Number of regulatory elements and trait-loss age

For further analyses of TFforge's performance on different datasets, first the number of regarded elements is lowered. This allows the establishment of a lower bound for the elements necessary to the achieve a high detection level, i.e. 100% sensitivity at a precision of 80% or 100%. I conduct the test on subsamples of the set of 1000 synthetic CREs. As Figure 3.6 shows, the target sensitivity level at a precision of 80% or 100% is reliably reached with 40 to 400 elements or 50 to 400 elements, respectively, depending on the time of the trait-loss (0.09, 0.06 or 0.03 substitutions per site after trait-loss).

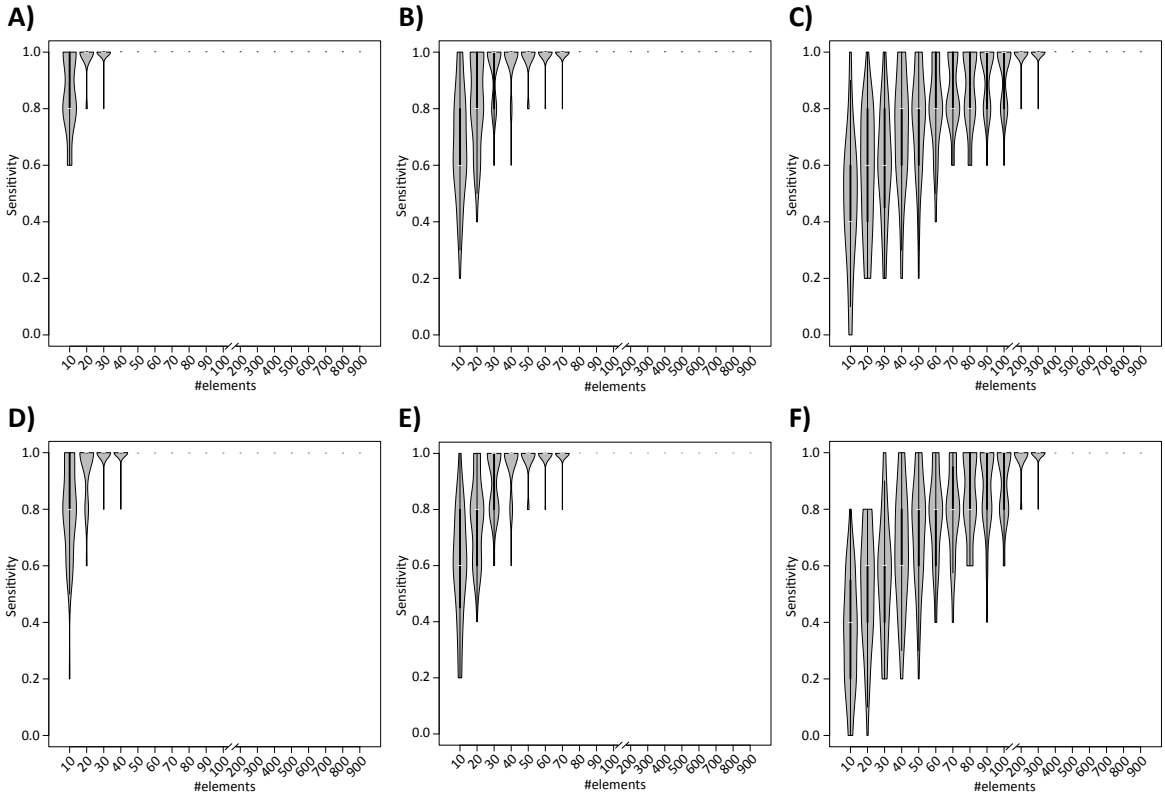


Figure 3.6: TFforge's performance in dependence of the number of CNEs
Shown is the achieved sensitivity in identifying the relevant TFs with TFforge with NSC, AFTP and ranking according to the Pearson correlation coefficient. The sensitivities are computed at a precision of 80% (A-C) or 100% (D-F). The time of the trait-loss in the underlying CNEs is 0.09 (A, D), 0.06 (B, E) and 0.03 (C, F) substitutions per neutral site.

3.4.2 Pleiotropic vs. specific regulatory elements

In correspondence to the theory that morphological changes largely arise from gene expression changes due to CRE mutations, the assumption has been so far, that CREs act modular on a specific phenotype. However, examples like the shared limb and phallus enhancers, described by (Infante, et al. 2015) reveal that this assumption is sometimes violated. Their study showed in mice and *Anolis* lizards that the CRE activity patterns in limb and genital tissues during development overlap. This shared activity is likely a reason for the also observed retention of limb CREs in snakes. HLEB (hindlimb enhancer B), for example, is a CRE with confirmed hindlimb and genitalia function, whose snake sequence retained functionality in the developing genitals while losing the ability to drive gene expression in (mouse) hindlimbs (Infante, et al. 2015).

The example of functional divergence of HLEB is of particular interest for TFforge, because it likely arises from binding sites of different TFs that evolved independent from each other in the snake's HLEB sequence. While limb TFBSs diverged, the genital-specific TFBSs likely remained functional. In order to test TFforge's ability to analyze such pleiotropic CREs and to differentiate their degenerated from their conserved TFBSs, a pleiotropic set of synthetic CREs is necessary. Such elements can be evolved, as described in Chapter 2.4.1, under selective constraint for 2 tissues. In this case, both tissues have their own tissue-specific (thus disjunct) set of TFs. During a trait loss, the selection constraint upon one tissue is removed, while

leaving the other tissue unaffected. Therefore, neutral evolution can degrade the trait-loss tissue-specific TFBSs, while the element itself stays under selection. This means, in particular, that the TFBSs specific to the other tissue stay under selection.

With this new CRE dataset, I analyze TFforge's discriminative power on pleiotropic elements. The results show the same performance dependency on the number of CREs and the time of trait-loss as in the case of trait-specific CREs. Using sets of 20 CREs, that are combined from tissue-specific and pleiotropic CREs in different ratios, I find that TFforge's performance of 60% sensitivity with precision of 100% is independent of the number of pleiotropic CREs (see Figure 3.7).

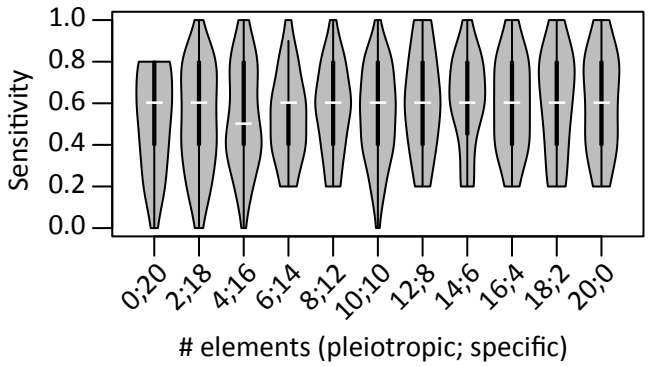


Figure 3.7: TFforge's performance on pleiotropic elements
Sensitivity at precision of 100% with Pearson correlation coefficient-ranking on sets of 20 synthetic CREs. The sets contain 20 to 0 randomly chosen pleiotropic CREs and 0 to 20 randomly chosen specific CREs. The underlying scores were computed with a 200 bp window, NSC and AFTP.

In summary, I conclude that TFforge is able to discover TF motifs that are involved in trait-loss. Its performance correlates with the number of relevant CREs and the time of binding site degradation. Whether the CRE's function is specific or shared between different tissues does not influence the results.

3.5 Validation on biological data

In the previous chapters, I gave a proof of principle for TFforge's paradigm and tested its performance under different parameter conditions and input variants. Since this analysis was solely based upon synthetic elements, I will in the following validate TFforge's function on biological data. For this, I use again the example of vision-impairment in the subterranean mammals, that was introduced in Chapter 2.5.1, yet on a different dataset.

A rather well-studied and photoreceptor-specific TF is the cone-rod homeobox protein ***CRX***, which is involved in photoreceptor cell differentiation and cone and rod function maintenance. Therefore, the following expectation is reasonable: genomic regions bound by CRX are of regulatory importance for the photoreceptor development and thus vision in general. A biochemical method for the identification of TF-DNA binding interactions globally in the genome is ***ChIP-seq***. (Corbo, et al. 2010) performed ChIP-seq experiments for CRX on adult mouse retinas. From this dataset, I use the overlapping peaks of both replicates that have a quality score of at least 45. Additionally, I restrict the analysis to the 200 bp center

regions of these peaks. Another requirement, that is sufficient to ensure the existence of ancestral sequence, is an overlap of the regions with at least 100 bp of conserved elements of the 29-way alignment. This results in 1080 genomic regions, of which the majority (773) does not overlap promoter regions. These regions are used to search for TF motifs that are associated with loss of vision.

The starting point for this search is the, in chapter 3.1 described, motif library which I extend by the CRX motif identified in (Lee, et al. 2010) from gel-shift assays (see Figure 3.8). This motif has been shown to very closely resemble the most overrepresented motif of the ChIP-seq dataset, suggesting that it will capture CRX's binding specificity changes on this dataset more accurately than the library's motifs of CRX and TFs with similar binding preferences. I will from now on refer to this motif of CRX with CRX.

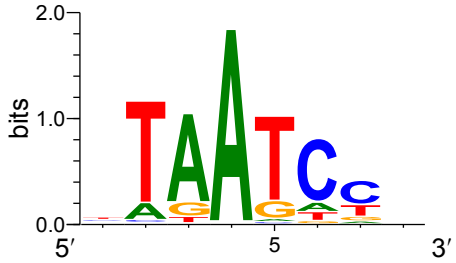


Figure 3.8: CRX logo
Shown is the sequence logo representation of the CRX's DNA-binding preference identified by (Lee, et al. 2010) via gel-shift assays and visualized with WebLogo 3.5.0 (Crooks, et al. 2004).

TFforge identifies CRX to be the most relevant out of all TF motifs. The next step is a verification, that this finding is indeed connected to the vision-impairment phenotype instead of being only a result of the mere abundance of CRX binding sites in the data. Therefore, I investigate TFforge's motif ranked in dependence of the choice of "trait-loss" species. That is, instead of the four vision-impairment species (blind mole rat, cape golden mole, star-nosed mole and naked mole-rat) four other species are defined as "trait-loss" species. I consider each of the 4809 combinations of four placental mammals (see Figure 2.16 for the phylogeny) that contain neither any of the four outgroup species, nor any of the four real vision-impaired species and also no pair of sibling species. Since all of these species sets have a generally good vision, no preferential vision-related (thus no CRX) binding site loss is expected in these species compared to the remaining species. Indeed, CRX is found to be most relevant motif only in case of "vision-impairment" as trait-loss. Furthermore, a comparison of the ranking score of CRX in the "vision-impairment" scenario with its ranking scores in the other scenarios also shows the specificity of the CRX relevance with respect to the "vision-impairment" species (see Figure 3.9). The results remain unchanged if the analysis is restricted to non-promoter regions, meaning to *cis*-regulatory regions exclusively, to avoid potential bias from possibly different TFBS composition of CREs and promoters.

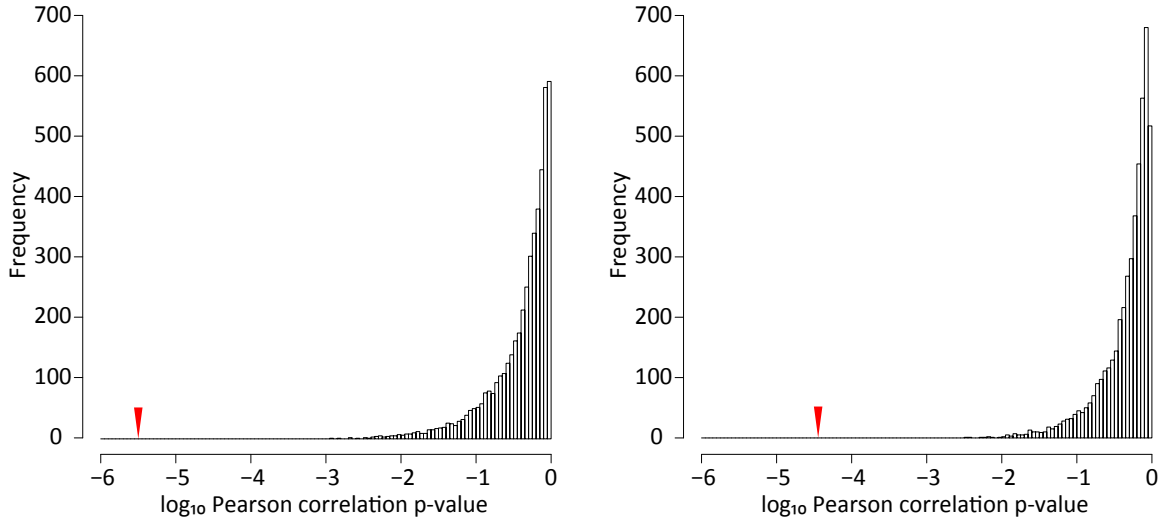


Figure 3.9: CRX ranking score for vision-impairment species and other trait-loss species

The histograms summarize CRX's $\log_{10}$ ranking score resulting from TFforge analyses on the same CRE set with 4809 differently chosen trait-loss species sets. In contrast is the $\log_{10}$ ranking score resulting from 4 vision-impairment species as trait-loss species indicated (red triangle). The ranking scores are based upon the Pearson correlation of trait-loss and trait-preserving branch score distributions. The set of underlying CREs consists of either all curated regions (A) or non-promoter regions (B).

3.6 Summary & Discussion

I developed the TFforge pipeline for the identification of the functional role of transcription factors by extending the Forward Genomics strategy to TFs and by scoring the corresponding motifs in a phylogeny-aware and discriminative manner. TFforge is designed as a flexible tool that can be applied to any set of species and any binary phenotypic change, given a genome alignment, the species' phylogenetic relationship, a set of regulatory regions and a general motif library. In order to rank the motifs of this library according to their putative relevance with respect to the phenotype, TFforge conducts four steps:

- Assessing the binding affinity of a TF to every sequence
- Transforming sequence scores into phylogenetically independent branch scores
- Categorizing branch scores into "trait-preserving" and "trait-loss"
- Ranking the motifs by relevance with respect to the phenotype

I provided a proof of concept on synthetic data and used this synthetic data, furthermore, to analyze the influence of different scoring method corrections, filters and finally the ranking method. The results from these analyses indicate that TFforge performs best with null score corrected sequence scores (NSC). A fixation of the transition probabilities (AFTP) does not provide any improvements for the synthetic dataset, although this remains to be confirmed on other (synthetic or real) datasets. The CRE filter based on ancestral sequence scores generally improves the performance as it removes noise arising from likely unrelated CREs. Thus, its effect is similar to the branch filter that filters non-informative branches. Yet, the application of a subsequent TF filter based upon the number of remaining branch scores has a more pronounced effect. However, the threshold for this filter is less straightforward due to its dependence on the number of considered CREs and species. To identify the relevant TFs based on their branch score distributions difference, the Pearson correlation coefficient gives the most consistent ranking.

A parameter that was not yet assessed is the scoring window size influence. However, since it is a parameter of the sequence scoring method, which is shared between TFforge and REforge, I expect the results of REforge to be applicable to TFforge. This would mean, that the scoring window length should be chosen such that it covers the majority of the sequences completely.

To validate the method on a real dataset, I used experimentally obtained CRX ChIP-seq regions (Corbo, et al. 2010) and ran TFforge on a motif library, which includes the CRX motif, with respect to different sets of trait-loss species. CRX is a known vision-related TF. The results show that the motif ranking heavily depends on the chosen set of trait-loss species. This means that the CRX motif in comparison to other motifs is found to be the most relevant one if and only if the previously introduced four vision-impairment species are chosen as vision-loss species. The comparison of CRX's ranking scores instead of the ranking position shows again the vision-impairment species resulting in, by far, the most extreme score for CRX.

The dataset of CRX bound regions in combination with the vision-impairment phenotype likely also allows deeper biological insights. For example, the top-ranked TF motifs contain besides CRX and several other very similar motifs also a motif of the ubiquitously expressed TF SP1, which is known to interact with CRX (Lerner, et al. 2005) and has no motif-wise similarity to CRX. The function and biological relevance of other top-ranking TFs remain to be assessed and verified.

Furthermore, the aforementioned ATAC-seq datasets of mouse retina and lens cells in combination with the vision-impairment phenotype phylogeny represents a great resource for additional validations and analyses. Since CRX plays an important role in retinal cells but not for the lens, an analysis of the retina ATAC-seq dataset is expected to rank CRX high. The contrary, however, is expected for the lens dataset. The next step would be an investigation of the top-ranking TFs in all vision-related ATAC-seq datasets with other ATAC-seq datasets, e.g. limb and/or midbrain ATAC-seq data, as control. Since vision-related TFs are unlikely to show also a differential binding site loss in limb or midbrain regulatory regions, these control datasets could help assessing the false positive rate.

The set of possible input datasets can easily be extended to further ChIP-seq datasets of known vision-related TFs like NRL or PAX6 or the promoter regions of known vision-related genes. In summary, the TFforge method has a wide applicability and can constitute a valuable tool to discover transcription factors that are potentially involved in the evolution of phenotypic changes between species.

4 Combined application

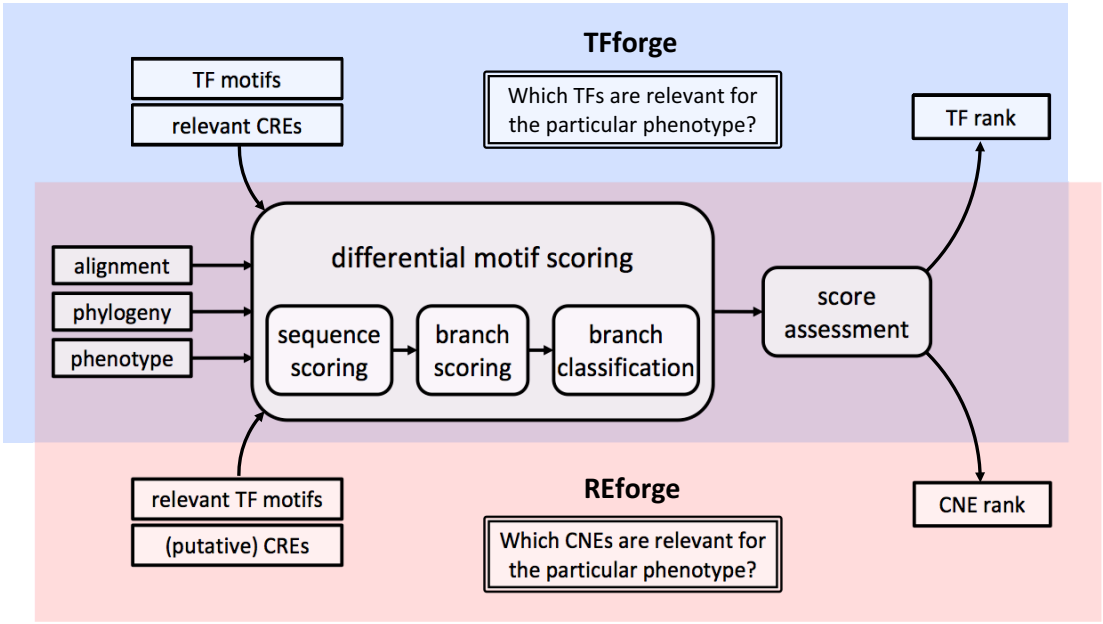


Figure 4.1: Summary scheme
TFforge (top) und REforge (bottom) share the differential motif scoring module (center), that uses sets of TFs and CREs as well as an alignment, a phylogeny and the phenotype annotation as input and provides classified branch scores upon which, via a score assessment, the final rank is computed. The differential motif scoring scores sequences, transforms the results into branch scores and classifies these according to the phenotype.

Regarding the phenotypes of organisms, not only genes encoding the proteins are important but also the regulation of their expression. This regulation is mainly dependent on regulatory sequences that contain binding sites for transcription factors. However, the interactions amongst TFs and between TFs and regulatory sequences, that finally control the gene expression form a complex network.

The developed methods REforge and TFforge both target the question of what we can learn about a gene regulatory network that underlies a specific phenotype, given a genome alignment and a set of phenotype-loss species. This question involves the identification of components of the regulatory network and an understanding of how the network changes during the evolution of species that lost that particular phenotypic trait. The difference between the methods is their focus on either the TFs or the genomic regions upon which the TFs act. In particular, TFforge helps answering the question “Which TFs are relevant for the particular phenotype”, whereas REforge helps answering “Which CNEs are relevant for the particular phenotype”. The close relatedness of both aspects raises the question of a possible combination of both methods.

To answer the questions for phenotype-relevant TFs and CREs, both methods rely on a central differential motif scoring module that assesses the changes in TF binding affinity to CRE sequences throughout the evolution of species with a special focus on the trait-loss species in comparison to the trait-loss species. This differential motif scoring, which is

constituted by the combination of sequence scoring, branch scoring and branch classification, accounts for the main computational load of the developed methods.

In the following, I discuss the runtime of the sequence scoring and ranking for both TFforge and REforge (see Chapter 4.1). Afterwards, I give an outline of the motivation and application of a combination of both methods (see Chapter 4.2).

4.1 Runtime analysis

The main computational effort of REforge and TFforge is the sequence scoring. In case of REforge, the complexity is proportional to the number of elements and the number phylogenetic nodes (thus $2n - 1$ in case of n species). For TFforge, the number of TFs is an additional factor (REforge considers all TF simultaneously). All of these operations are independent and could theoretically be parallelized. Yet, the scoring procedure of REforge and TFforge conducts a parallelization only with respect to the transcription factors and elements, not with respect to the sequences. Combining the computation of all sequence scores of an element, i.e. the scores of the entire phylogenetic tree, with respect to a TF (set) into a single task brings two advantages. At first, this allows an easy exploitation of the phylogenetic relatedness of the sequences in the sense that closely related species often show identical sequences, such that the corresponding scores can be cached. Secondly and more importantly, the scoring procedure can this way immediately convert sequence scores into branch scores. These branch scores are collected in a score output file. For example, the TFforge analysis with NSC, AFTP and ancestral filtering of a 1000 CNE set with respect to 572 TF that has been conducted in Chapter 3.3.1, took overall 266,7 min of computation time. 258 min (97%) of this were HPC time on the MPI-CBG's HPC-cluster, which offers 105 nodes with 24 cores each. I used in this case 1000 cores for the computation of the sequence scores and branch scores. Hereby, the total cluster time was 124923 min (2082 h) accumulated by 1000 jobs, each computing 572 element-TF combinations. Therefore, a single job, computing sequence and branch scores of a 20-species phylogeny took on average 13 s. For two other analyses with identical setup, the total cluster time summed up to 2129 h and 1544 h. The score computation of the same dataset without NSC in comparison took in total 93,5 min of which 82 min were HPC time (1205 h distributed over 1000 jobs).

Regarding the REforge algorithm such analyses are faster due to the - by a factor equal to the motif library size - lower computational demand. The background set of synthetic CRE in combination with the simulation phylogeny, both described in Chapter 2.4.1, needs on average a runtime of 14 min 20 s. About 1 min of this is overhead and the rest is HPC time with 500 parallelized jobs, if scores are computed with NSC. Similar results are obtained, when observing the runtimes of the scoring procedure on the 1000-element foreground set of synthetic CREs. The individual runtimes are listed in Table 6.

In the current setup, the sequence scores and branch scores are computed in a separate module, common to REforge and TFforge. This module summarizes the list of filtered branch scores per element and per motif or motif set in an output file. The subsequent step of the REforge pipeline is the branch classification and branch score distribution assessment, that results in the final ranking of e.g. all 11000 synthetic CREs (1000 Foreground-Set CREs and 10000 Background-Set CREs). It needs consistently about 8 min 20 s on a standard desktop computer of which 6 min 30s are used for reading the scorefile and categorizing the branches, and 103 s to conduct the ranking. These run times are consistent with TFforge's branch

#elements	#jobs	HPC time	HPC total time	real time
10000	500	14.0 min	46.7 h	15m11.186s
10000	500	12.0 min	50.2 h	13m10.839s
10000	500	12.0 min	54.0 h	13m12.971s
10000	500	18.0 min	57.3 h	19m11.172s
10000	500	9.0 min	41.2 h	10m10.785s
10000	500	14.0 min	57.9 h	15m11.293s
1000	100	3.0 min	2.6 h	3m24.533s
1000	100	2.0 min	2.4 h	2m24.377s
1000	100	5.0 min	6.8 h	5m26.223s
1000	100	3.0 min	2.6 h	3m24.559s
1000	100	6.0 min	7.4 h	6m25.710s
1000	100	4.0 min	4.3 h	4m24.611s
1000	100	6.0 min	7.7 h	6m25.727s
1000	100	4.0 min	4.7 h	4m24.639s
1000	100	6.0 min	7.6 h	6m25.781s
1000	100	4.0 min	4.5 h	4m24.616s

Table 6: Runtime examples of REforge/TFforge's scoring module
The real HPC time, total HPC time and the overall time including HPC preparation are listed in dependence of the number of scored elements (#elements), that are scored with respect to a 20-species phylogeny and split into #jobs processes. The scoring was conducted with NSC.

classification and ranking step. In case of the analysis of 30000 CREs (described in the following Chapter 4.2) the ranking of a TF takes 105 s each, after the data has been read and branches classified. But since 638 TFs need to ranked, this overall runtime sums up to more than 18.5 h. Therefore, this step has been parallelized with respect to the TFs. However, both the ranking runtimes of REforge and TFforge contain the computation of *all* previously described statistical tests. The computation of the Pearson correlation coefficient alone would decrease the runtime considerably.

Another possible runtime optimization exists, related to scoring module. As described, the current setup aims for the reusability of the scores for different trait-loss scenarios, like it has been done for TFforge's validation in Chapter 3.5. This, on the other hand, implies that the branches remain to be classified according to the phenotype in an additional step. While this step needs to be done only once for synthetic data, for real data the classification is done for each CRE, due to the possibly incomplete sequence alignment, as described in Chapter 2.3.

Therefore, if we abandon the possibility of reusing branch scores for different phenotypes, then the classification can be done together with the branch score computation. This prevents the need for a second traversal of the tree. Another possibility would be the enumeration of all possible names of a specific species, based upon the complete phylogenetic tree.

4.2 Use case - Towards gene-regulatory networks

Previously, I showed for both methods (see in Chapter 2.4.3.2 and 3.4.2) that uncertainty in the functional relatedness of the input does not affect the functionality drastically, since the comparison between trait-loss and trait-preserving species is the main component of the analysis. Therefore, I test how well TFforge and REforge can complement each other. The common application for this combination is the scenario, in which the user

- has a phenotype that has been lost independently in multiple species
- has a sequence alignment of trait-loss and trait-preserving species
- has a set of genomic regions, constituted for example by conserved regions from the alignment or by experimentally acquired regions of open chromatin
- wants to identify regulatory regions involved in the phenotype development
- wants to identify TFs involved in the phenotype development

In this setup, TFforge can initially identify a set of TFs with differentially lost binding sites. In a second step, this refined TF set can be used by REforge to identify relevant CREs. More iterations of this dataflow, according to the schematic Figure 4.2, starting with the refined CRE set are possible.

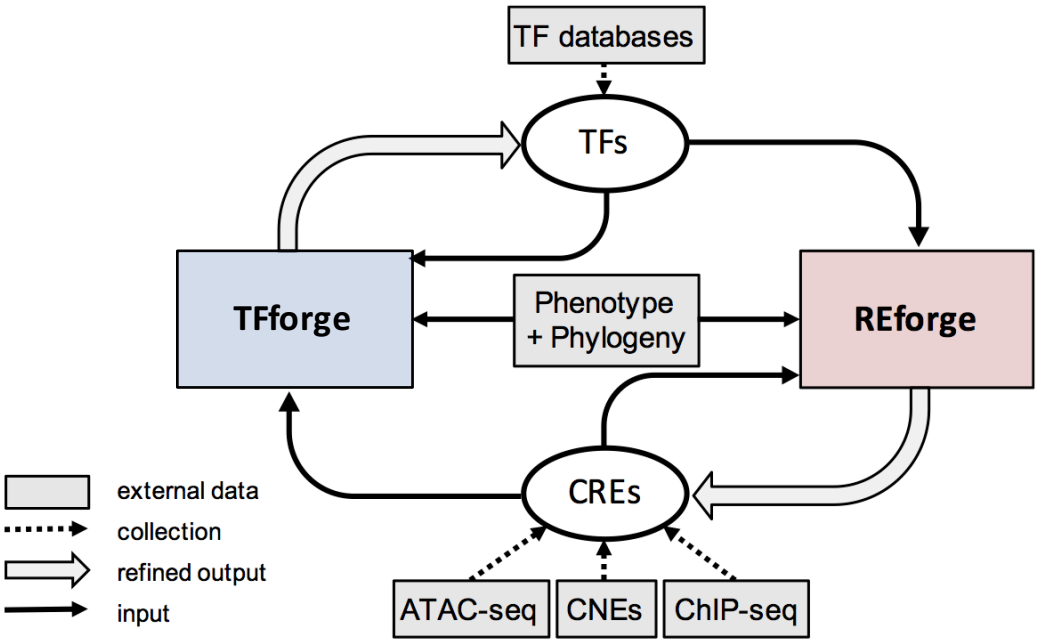


Figure 4.2: Combination of REforge and TFforge
Initial TF input data is acquired from public TF databases, whereas the initial CRE set arises from computational (e.g. CNEs) or experimental (e.g. ATAC- or ChIP-seq) analysis, from which TFforge refines the set of relevant TFs. REforge can, via these TFs, refine the set of CREs. This dataflow can possibly be iterated further.

The previously described vision-impairment phenotype together with the set of CNEs, that have been described in Chapter 2.5.1, is an example of a possible application of the combination of TFforge and REforge, if no *a priori* knowledge of relevant TFs is assumed. I evaluate in an exemplary analysis of this dataset the performance of the combination of both methods. Therefore, I use a subsample of 10000 out of the total 351279 CNEs to obtain a TF ranking from TFforge. Since I previously used 30 vision-related TFs, I now choose the top 30

TFs from this ranking as input for the REforge analysis step of the complete CNE set. To assess the quality of the results, the overlap of the top-ranking CNEs with ATAC-seq elements from different tissues is computed analogous to Chapter 2.5.2. The tissue-specific enrichments in top-ranking CNEs resulting from this combination of TFforge and REforge are visualized in Figure 4.3. Additionally, the Figure shows for comparison the same overlaps for the CNEs identified in Chapter 2.5.4 with Forward Genomics and REforge based on vision-related TF. The comparison shows that the combination of TFforge and REforge achieves a substantially higher enrichment in ATAC-seq elements for almost all vision-related tissues than Forward Genomics can achieve. This means that the developed combination of TFforge and REforge outperforms Forward Genomics even without foreknowledge of relevant TFs.

It is, however, evident that the combination of TFforge and REforge achieves a lower enrichment in ATAC-seq elements for the vision-related tissues than REforge with foreknowledge of relevant TFs does. This means that, unsurprisingly, foreknowledge is beneficial for the performance of the methods for retrieving a high precision set of putative trait-related CREs. Interestingly though, the initial TF unspecificity causes the top-ranking CNEs to be significantly enriched in overlaps with limb-specific ATAC-seq elements. This might reflect limb adaptations to the subterranean lifestyle but is up to further investigation. In this context, also the set of identified TF motifs needs to be analyzed, with respect to existing knowledge about their function. For that, the motif-protein mapping via Ensembl IDs can be used (see Chapter 3.1.1).

The identification of relevant TFs additionally to the CREs lays the foundation of a partial reconstruction of the gene regulatory network, that underlies the studied phenotype. Such a network consists of genes and other molecular regulators as well as regulatory regions that control a specific function by direct and indirect interaction. The set of TFs, which are simply the products of specific genes, represent an essential subset of the regulators of the network. Furthermore, the identified CREs represent the regulatory regions upon which this set of TFs acts. Repeated application of REforge on subsets of the relevant TF set could produce a fine-grained resolution of putative TF-CRE interactions. Additionally, target genes could be assigned to the regulated regions, for example based on their distance or by more sophisticated methods, relying on gene expression data, chromatin state information or gene expression prediction, like McEnhancer (Hafez, et al. 2017) uses. An investigation of the function of these genes can confirm known phenotypical relevance. It, furthermore, can reveal additional TFs involved in the gene regulatory network, which could be the ground for additional iterations of TFforge and REforge.

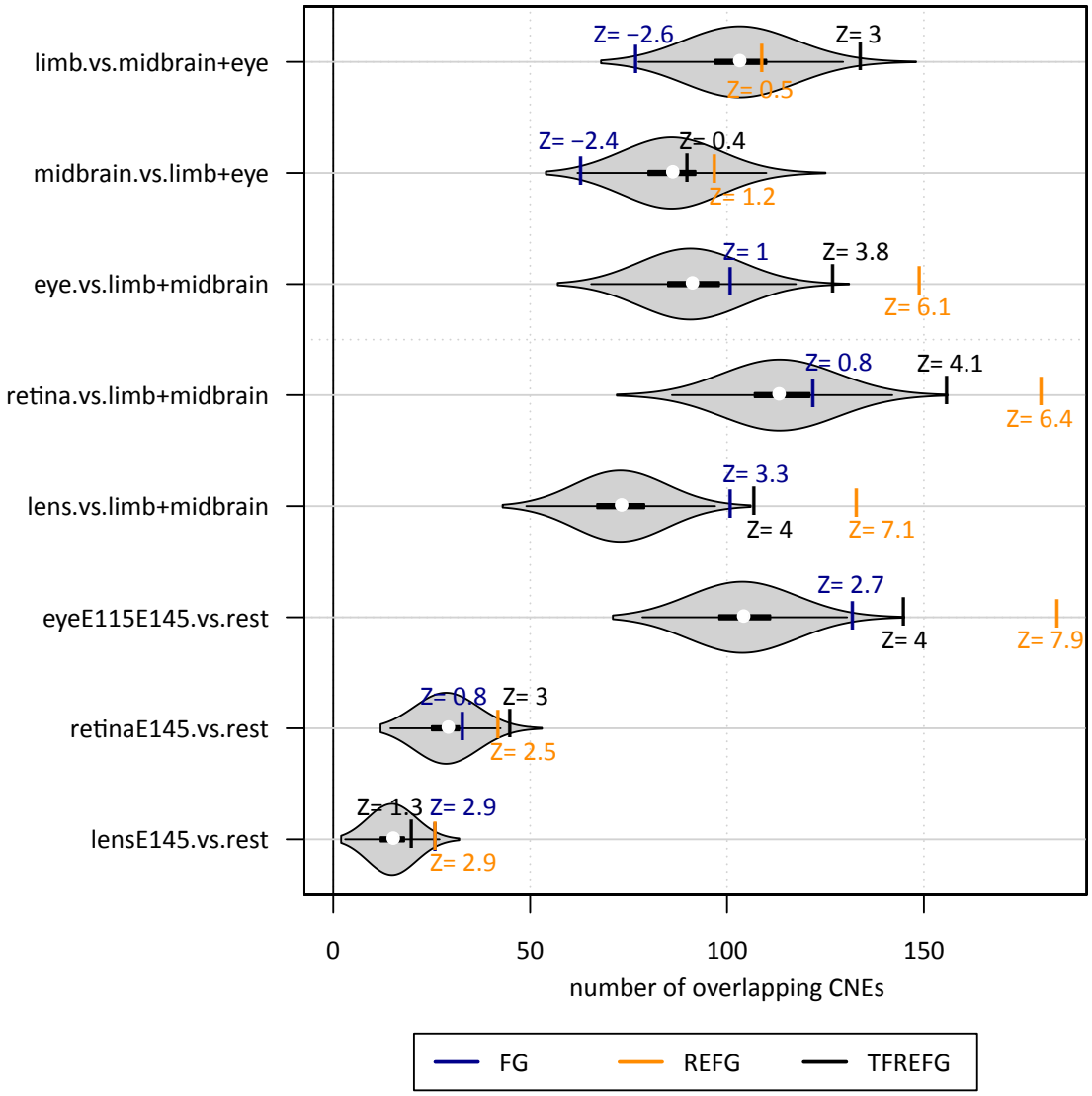


Figure 4.3: Combined TFforge+REforge analysis on subterranean mammals
A comparison of the precision of Forward Genomics (blue), REforge with given vision TFs (orange) and REforge with TFs from TFforge (black), measured by the overlap with multiple tissue- and developmental time-specific ATAC-seq element sets. In comparison is the by chance expected overlap shown (grey violin plots). Considered are the top 9364 CNEs ranked by each of the methods.

5 Summary

With the growing availability of genomic sequence data from a variety of species, the question for computational solutions to biological problems that were until now primarily experimentally answered emerges. This includes for example the extraction of knowledge about the functional importance of the genomic regions and their connection to phenotypic traits. In particular, the high complexity of gene expression regulation including for example tissue specificity, developmental activation or inactivation and interaction of different transcription factors limits the current understanding of the development of complex phenotypes. Consequently, future research will depend on the targeted identification of suitable candidate loci and candidate molecules for experimental validation analysis of their interaction.

I developed two novel methods, REforge and TFforge, that follow the idea of Forward Genomics that biological function divergence (referring to either TFs or CREs) can be associated with a matching signature of phenotype divergence among the considered species. This means, that loss patterns on the sequence level are identified, that match the observable phenotypic loss pattern. Different genomic "grammars" encoding information in different types of regions, like coding and regulatory regions, imply different selective constraints on the sequence. If we focus on regulatory elements the sequence divergence itself is a suboptimal measure for functional divergence. Here, the incorporation of TF binding knowledge as the "grammar" underlying CRE function allows the abstraction from the sequence level towards the functional level.

The first pipeline, REforge, focuses on the association between gene-regulatory regions and phenotype by exploiting knowledge of phenotype-related TFs – their existence and their binding preferences. I showed that this knowledge helps achieving a considerably higher accuracy in the CRE-phenotype mapping than Forward Genomics. As a result, many more relevant CREs can be discovered with REforge. The pipeline is designed for a general application and, after a verification on synthetic data, has been successfully tested for the phenotype of functional eyes in subterranean mammals. REforge can be utilized for any phenotype, for which species are known that have lost it and for which the loss likely happened due to or involving gene regulatory changes. The only prerequisites for the application of REforge on such a phenotype are, first, the existence of a genome alignment including the trait-loss species and related trait-preserving species and, second, knowledge about phenotype-related TFs including their binding preferences.

The second pipeline, TFforge, changes the viewpoint from a TF-centered to a CRE-centered perspective. It, thereby, allows the identification of phenotype-related motifs and consequently the indirect identification of phenotype-related TFs. This is based on published binding motifs of transcription factors available from a variety of databases. I verified and validated the method on synthetic data and the biological example of the vision-TF CRX. In summary, the tool has just like REforge a wide application field, e.g. for the identification of common co-binding TFs.

Thus, in summary, both methods search for a biological function divergence, predicted from the sequence, by measuring existence of TFBSs in regulatory regions and comparing it between different species. The minor difference here is the focus either on a single TF in

combination with all relevant regulatory regions or on a single CRE and all relevant TFs that bind to it.

The combination of both TFforge and REforge allows the parallel identification of CREs and TFs whose functional divergence patterns match the phenotypical divergence pattern of the analyzed set of species. In particular, the methods present in combination with each other an opportunity to identify regulatory elements without foreknowledge of trait-relevant TFs and therefore with the same input knowledge that Forward Genomics expects. Yet, I showed on the example of the trait-impairment phenotype that the subsequent application of TFforge and REforge is able to substantially outperform Forward Genomics.

The only assumptions that REforge and TFforge state in addition to Forward Genomics' assumptions are related to the TFs. This is, at first, the existence of a motif library, which can be generated as described in Chapter 3.1. Second, the TFs are assumed to be conserved throughout the species in the sense that their binding preferences remain unaltered. Since TFs are generally highly pleiotropic developmental genes (Liu, et al. 1997; Nitta, et al. 2015), this assumption is reasonable. The additional assumptions, inherited from Forward Genomics, contain a given species alignment, a given set of candidate CREs and the existence of trait-loss species in the aligning species. The performance gain resulting from a higher number of independent trait-loss species has been already assessed for Forward Genomics (Prudent, et al. 2016), yet it was also applied to the phenotype of limb-loss in snakes, which constitutes only two dependent losses. Although the performance of REforge and TFforge will most likely increase with a higher number of trait losses as well, this could be topic of future investigation. In this context, the performance of REforge on a single trait loss would be interesting. While Forward Genomics in such a scenario identifies species-specific divergence generally, with or without connection to the phenotype of interest, REforge offers the possibility to focus on phenotype-specific diverged CREs, due to the incorporation of TF knowledge.

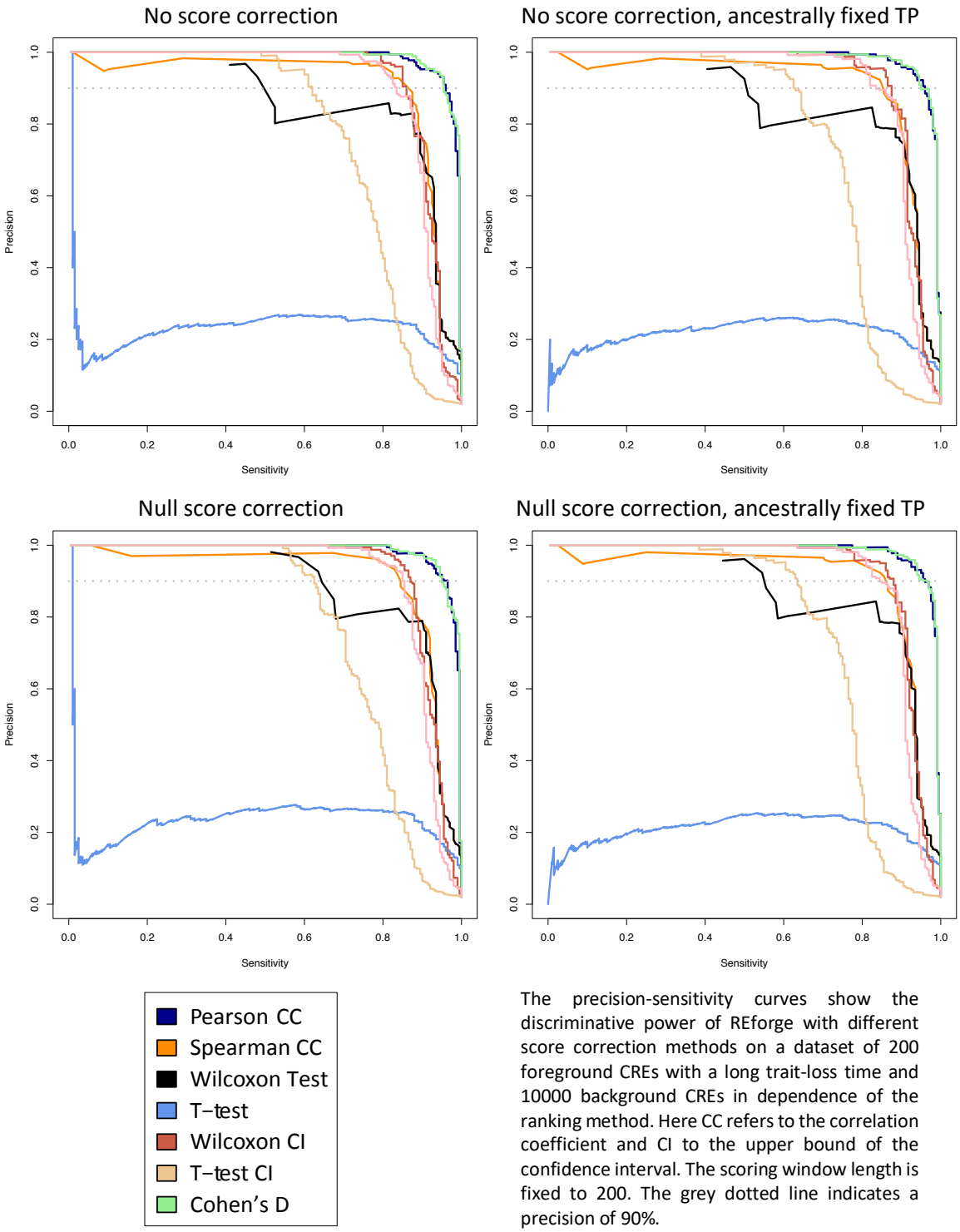
Another minor difference in the assumptions of Forward Genomics and REforge (and TFforge) is that Forward Genomics needs the complete phylogenetic tree including the branch lengths, whereas my developed methods rely on the structure of the tree exclusively. Although this is no constraint of Forward Genomics, since the branch lengths are usually inferred from the divergence sequences, the resulting branch lengths represent the divergence globally. This possibly leads to inaccuracies in Forward Genomics, if the analyzed regions underlie differing selective constraints. Due to the abstraction to the functional level, which is considered to be conserved, REforge and TFforge circumvent this issue.

In summary, both REforge and TFforge have a wide applicability on endless phenotypes on their own and, in particular, their combination in respect to gene regulatory network analysis can constitute a valuable toolset for evo-devo studies.

Appendix A. Figures

Appendix A.I. REforge

Appendix A.I.1. Score correction assessment on synthetic data



The precision-sensitivity curves show the discriminative power of REforge with different score correction methods on a dataset of 200 foreground CREs with a long trait-loss time and 10000 background CREs in dependence of the ranking method. Here CC refers to the correlation coefficient and CI to the upper bound of the confidence interval. The scoring window length is fixed to 200. The grey dotted line indicates a precision of 90%.

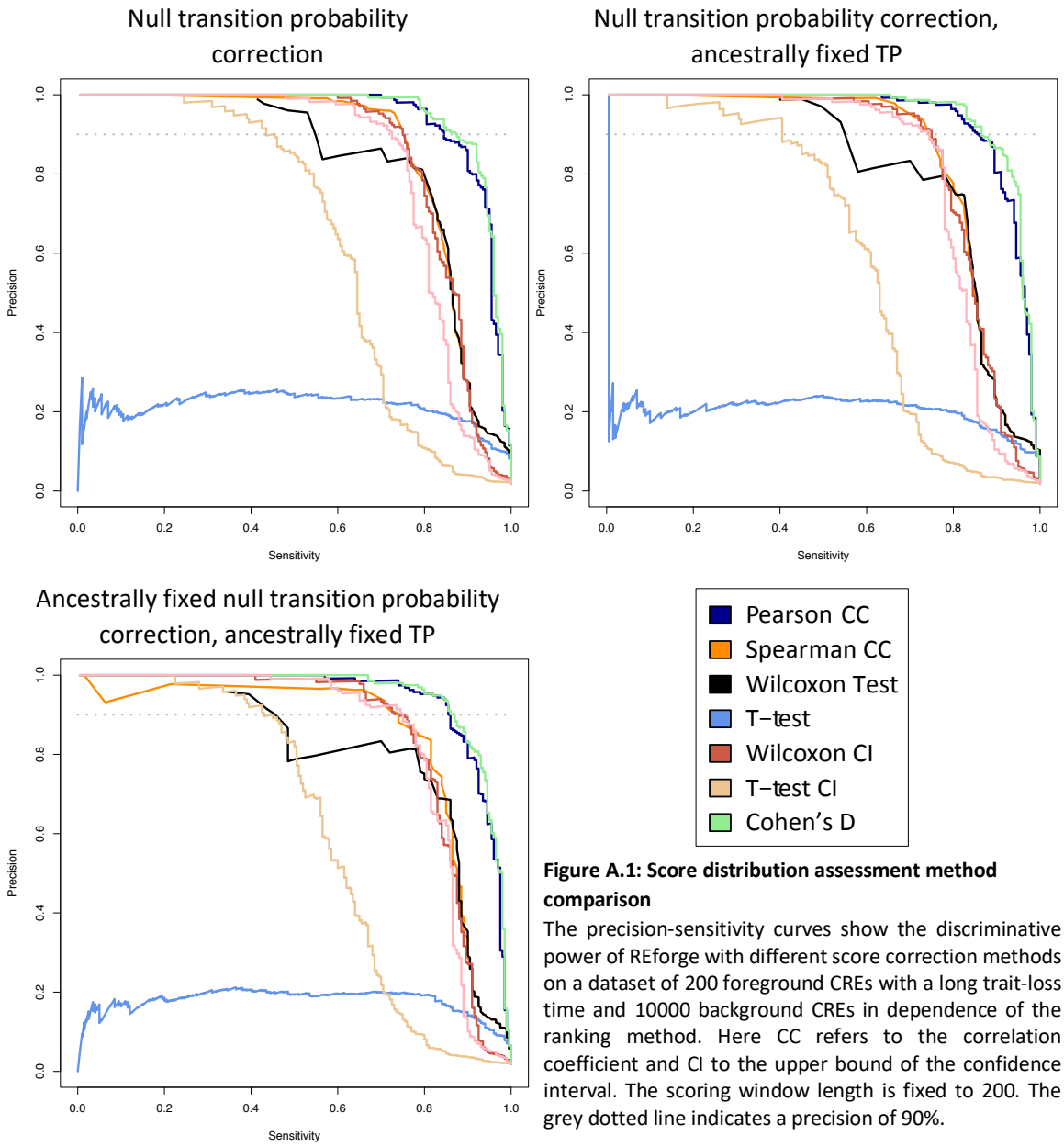


Figure A.1: Score distribution assessment method comparison

The precision-sensitivity curves show the discriminative power of REforge with different score correction methods on a dataset of 200 foreground CREs with a long trait-loss time and 10000 background CREs in dependence of the ranking method. Here CC refers to the correlation coefficient and CI to the upper bound of the confidence interval. The scoring window length is fixed to 200. The grey dotted line indicates a precision of 90%.

Background 5000+5000; 5 TF

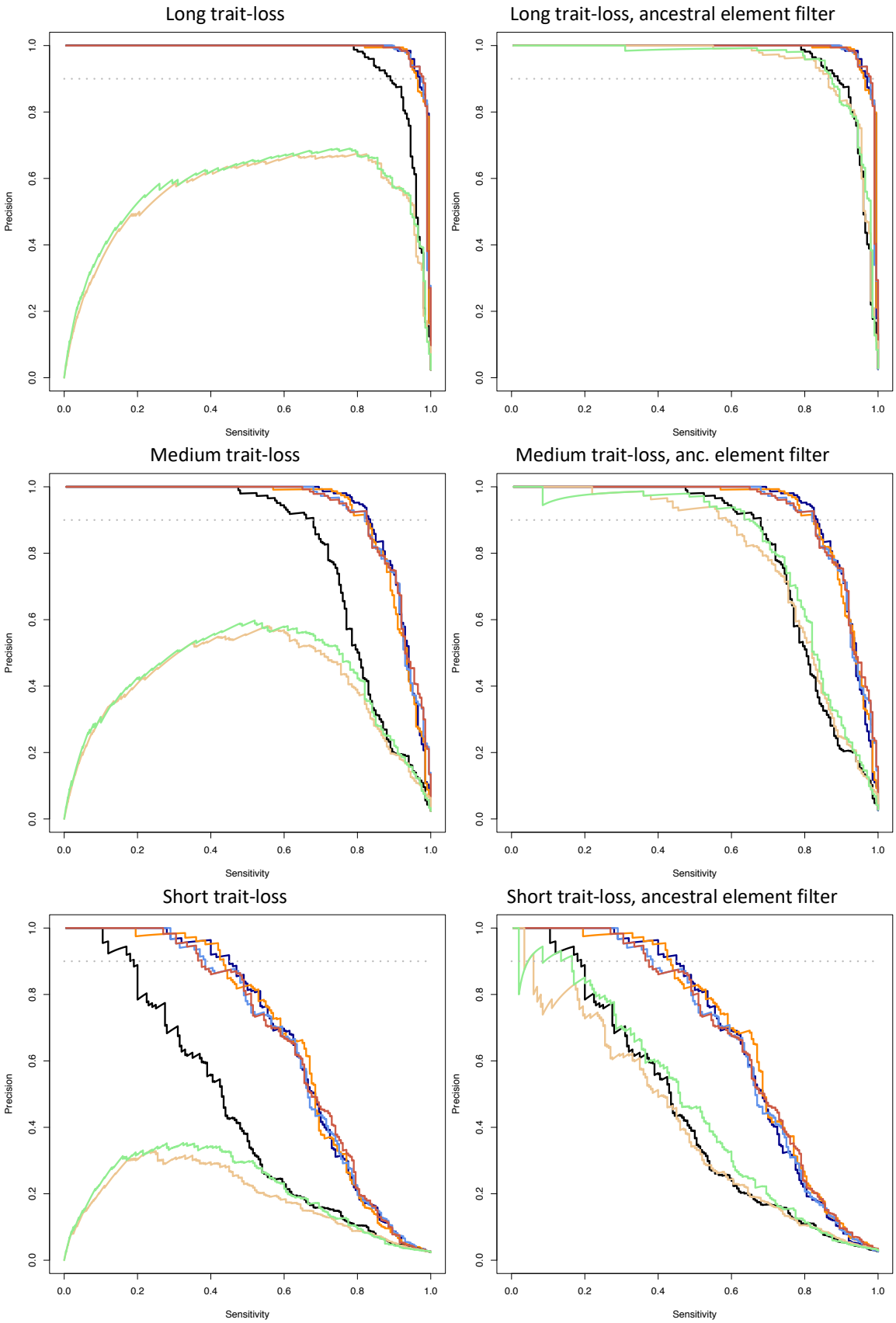


Figure A.2: Influence of the scoring method on REforge's performance on synthetic data

Precision-sensitivity curves on Foreground CNEs from 3 different trait-loss scenarios and 5000 CNEs from each background set with 5 relevant TFs, scored with (right) and without (left) ancestral element filter.

Background 5000+5000; 30 TF

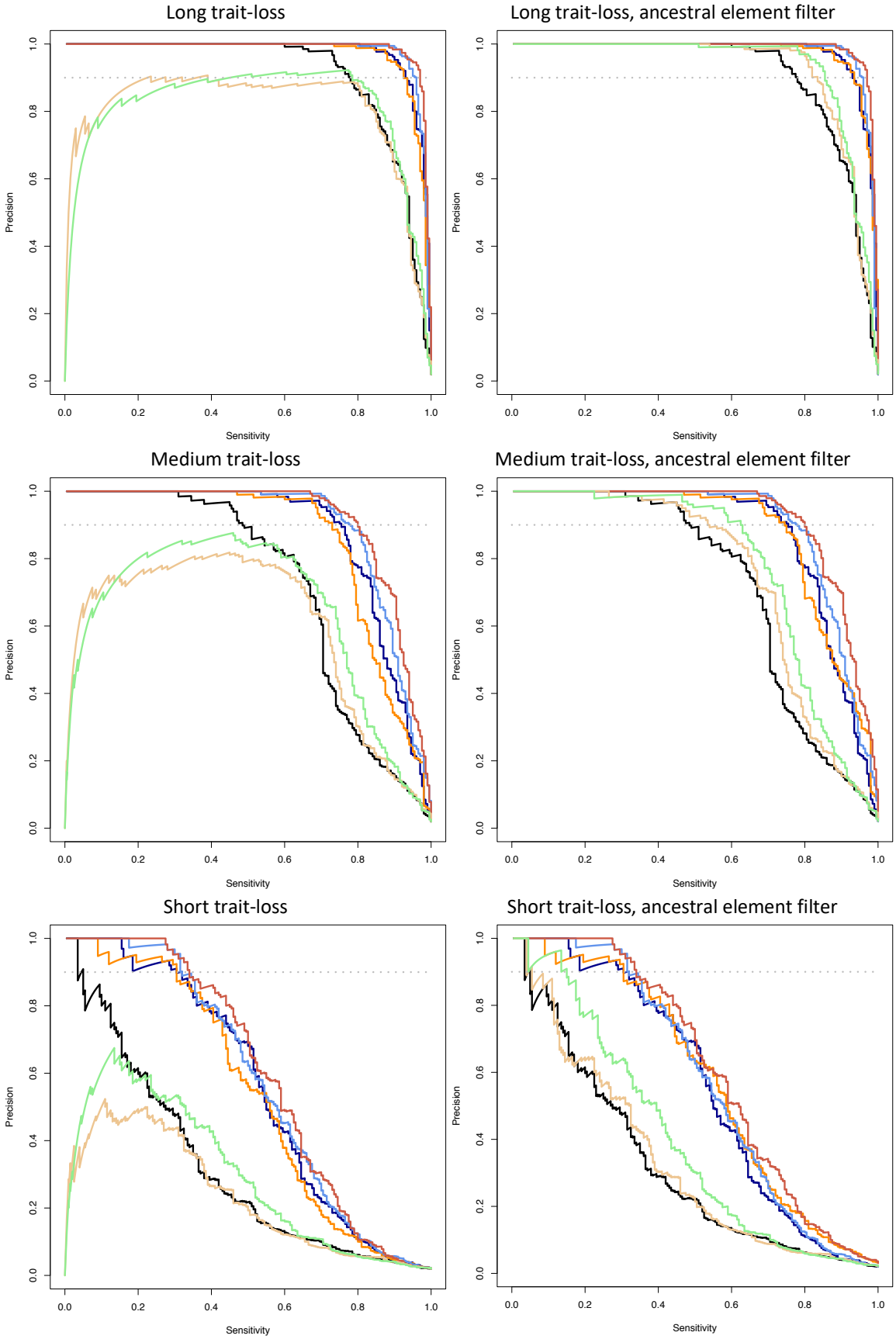


Figure A.3: Influence of the scoring method on REforge's performance on synthetic data

Precision-sensitivity curves on Foreground CNEs from 3 different trait-loss scenarios and 5000 CNEs from each background set with 5 relevant and 25 irrelevant TFs, scored with (right) and without (left) ancestral element filter.

Background 10000; 5 TF

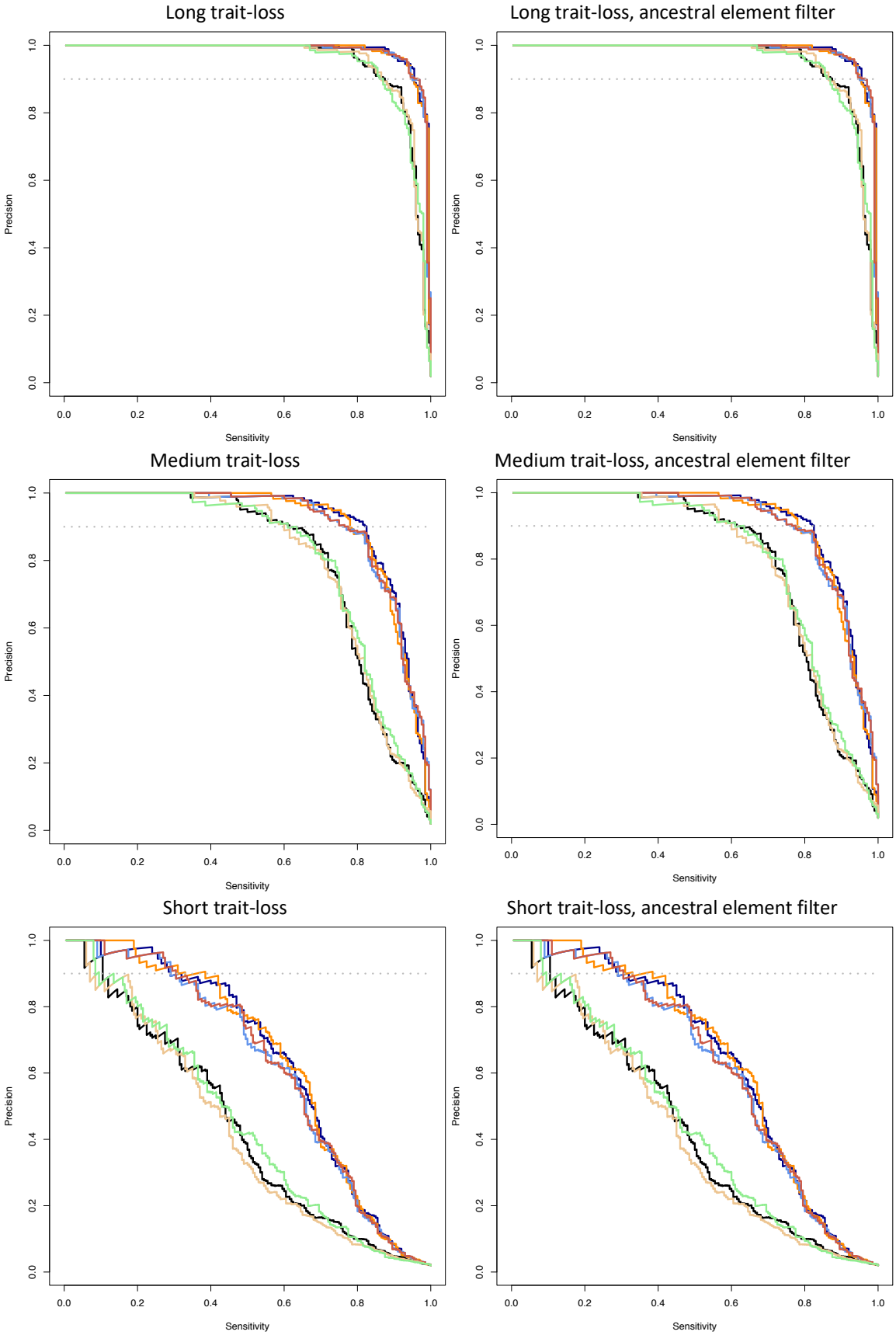


Figure A.4: Influence of the scoring method on REforge's performance on synthetic data

Precision-sensitivity curves on Foreground CNEs from 3 different trait-loss scenarios and Background-Set 1 CNEs with 5 relevant TFs, scored with (right) and without (left) ancestral element filter.

Background 10000; 30 TF

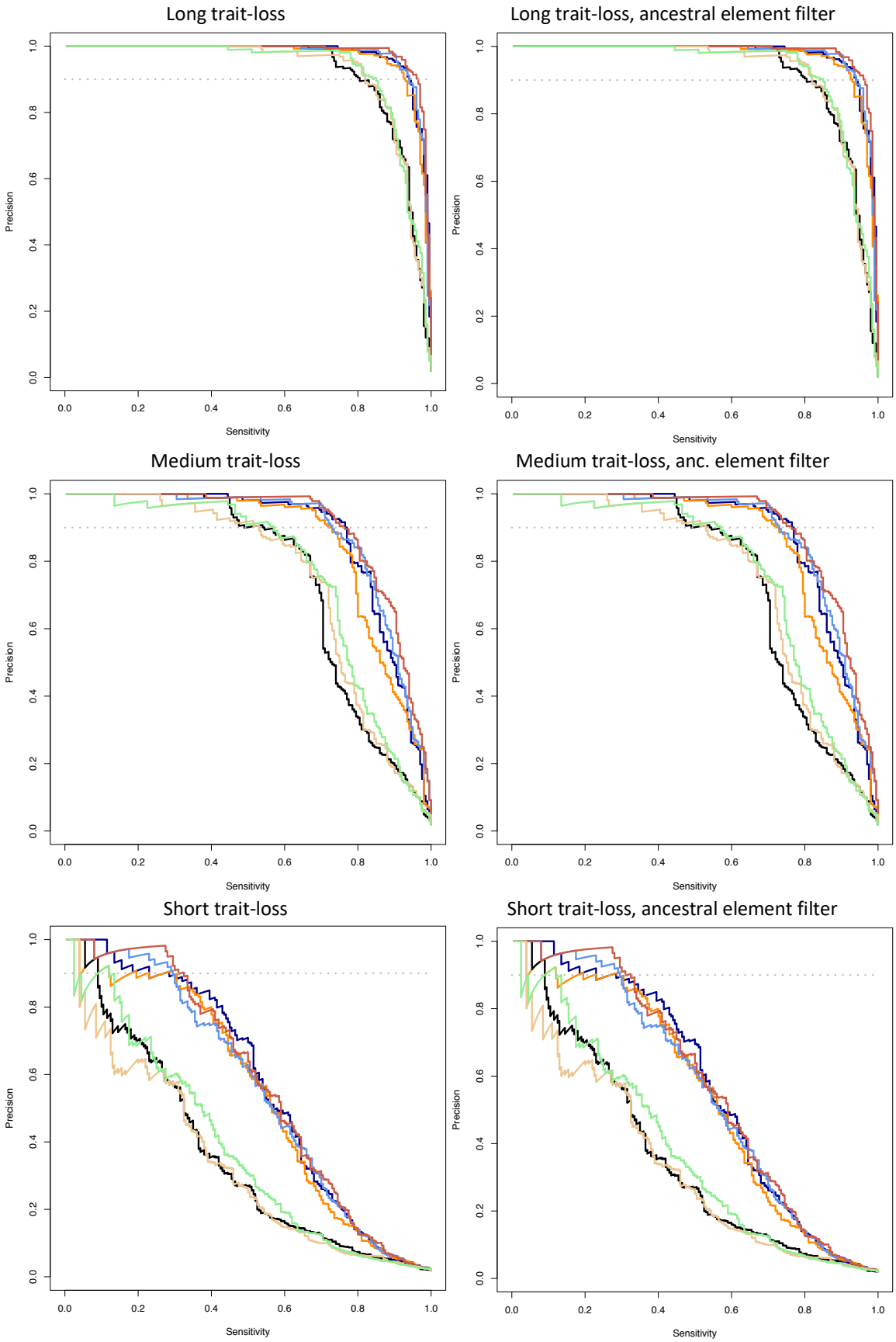


Figure A.5: Influence of the scoring method on REforge's performance on synthetic data

Precision-sensitivity curves on Foreground CNEs from 3 different trait-loss scenarios and Background-Set 1 CNEs with 5 relevant and 25 irrelevant TFs, scored with (right) and without (left) ancestral element filter.

Appendix A.I.3. Scoring method assessment on real data

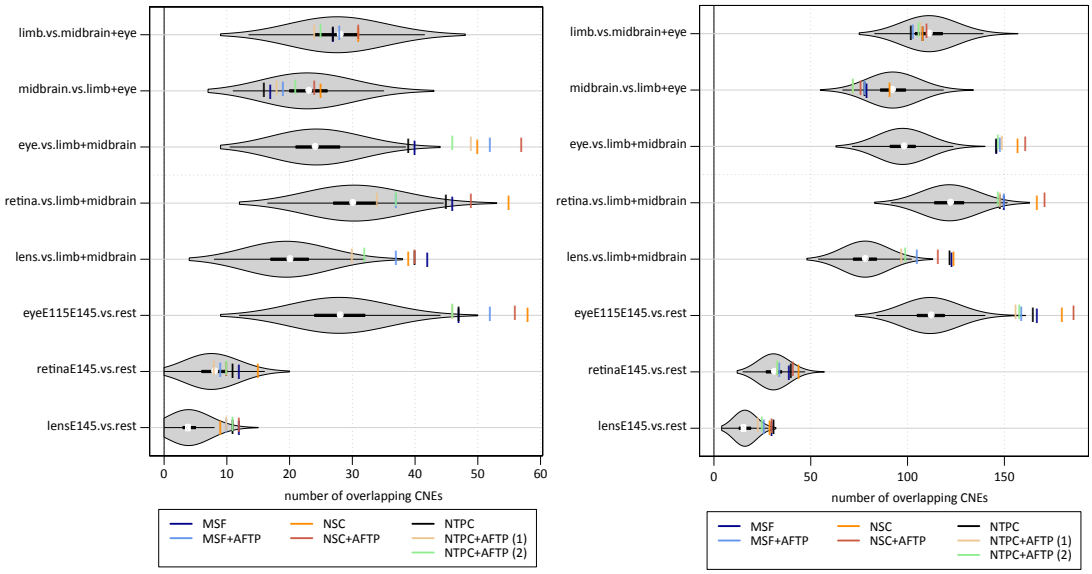


Figure A.6: Influence of the scoring method on REforge's performance on real data

The effect of different scoring methods on REforge's performance is compared. Independent of method was a window of 200 bp. The 2500 CNEs (left) and top 10000 (right) CNEs from the resulting rankings with Cohen's D were overlapped with multiple tissue- and developmental time-specific ATAC-seq element sets. This overlap is shown in comparison to the by chance expected overlap (grey violin plots).

Appendix A.I.4. Ancestral filter assessment on real data

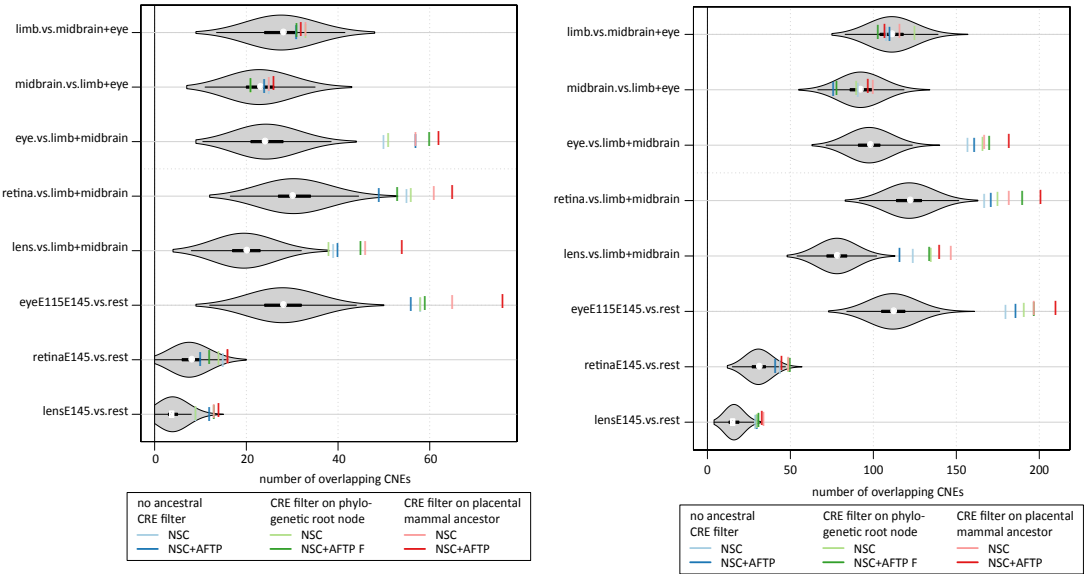


Figure A.7: Influence of ancestral element filtering on REforge's performance on real data

The effect of an ancestral element filter based on the ancestor of placental mammals (red) and the oldest possible ancestor (green) is compared to REforge's performance without an ancestral element filter (blue). Independent of the ancestral element filter was the scoring conducted with a window of 200 bp, NSC and with (dark) or without (light) AFTP. The 2500 CNEs (left) and top 10000 (right) CNEs from the resulting rankings with Cohen's D were overlapped with multiple tissue- and developmental time-specific ATAC-seq element sets. This overlap is shown in comparison to the by chance expected overlap (grey violin plots).

Appendix A.I.5. Scoring window length assessment on real data

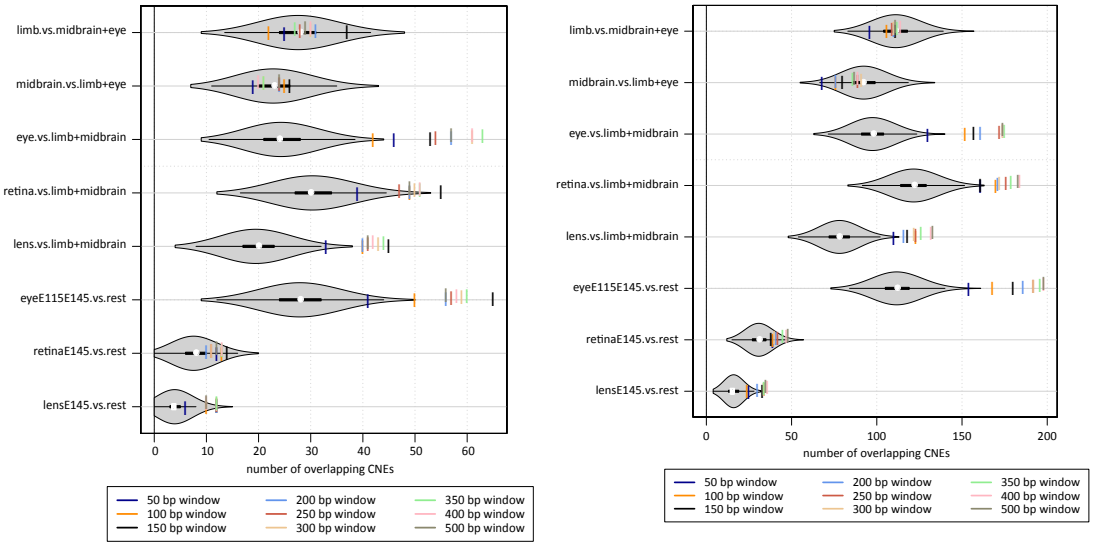


Figure A.8: Influence of the scoring window length on REforge's performance on real data

The effect of different scoring window on REforge's performance is compared. Independent of the window size was the scoring conducted with NSC and AFTP. The 2500 CNEs (left) and top 10000 (right) CNEs from the resulting rankings with Cohen's D were overlapped with multiple tissue- and developmental time-specific ATAC-seq element sets. This overlap is shown in comparison to the by chance expected overlap (grey violin plots).

Appendix A.I.6. Ranking method assessment on real data

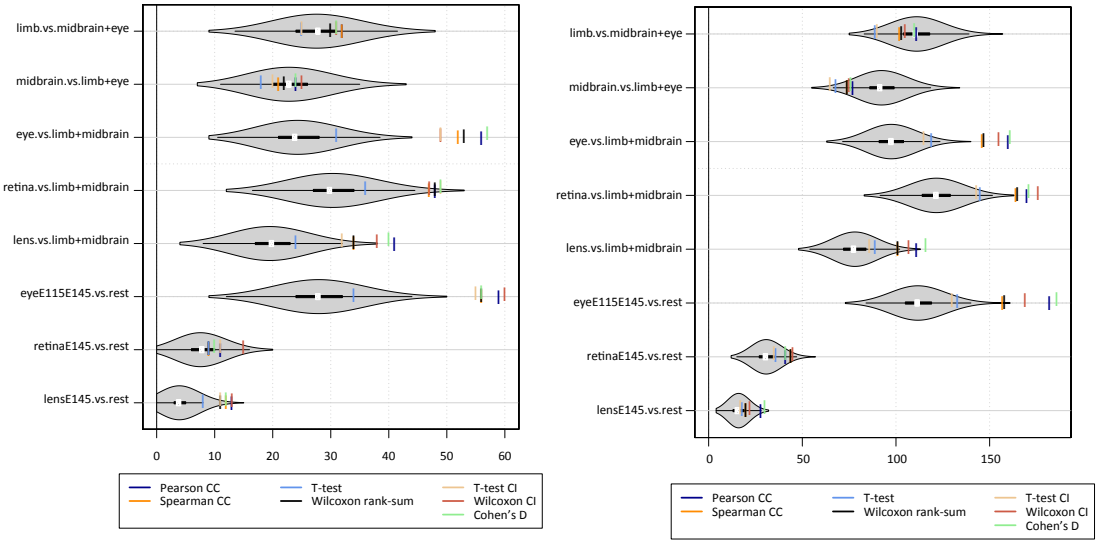


Figure A.9: Influence of the ranking method on REforge's performance on real data

The effect of different ranking window on REforge's performance is compared. The scoring conducted with NSC and AFTP and a window size of 200. The top 2500 CNEs (left) and top 10000 (right) from the resulting rankings with each method were overlapped with multiple tissue- and developmental time-specific ATAC-seq element sets. This overlap is shown in comparison to the by chance expected overlap (grey violin plots).

Appendix A.II.1. Proof of Concept

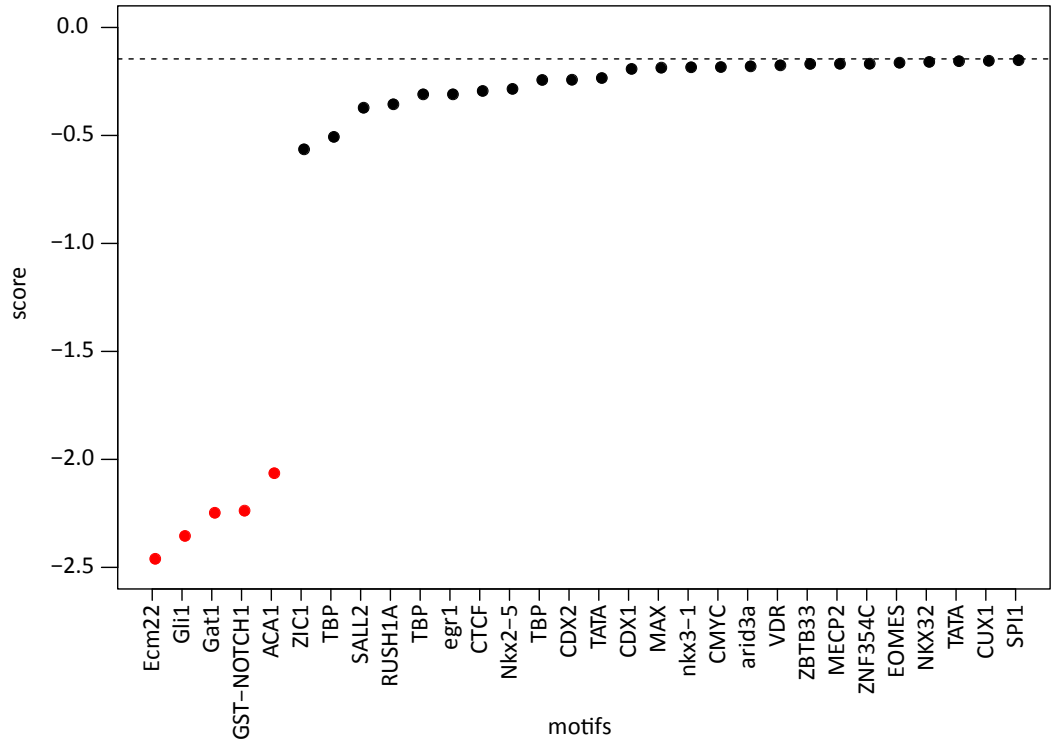


Figure A.10: Motif ranking with TFforge
Shown are in ascending order the TF motifs, whose ranking score < -0.05. The ranking bases on Cohen's D. The underlying data are 1000 synthetic CREs with a long trait-loss age scored by Stubb with a 200 bp window, NSC and AFTP. Red indicates the five target TF motifs.

Appendix A.II.2. Ranking method comparison

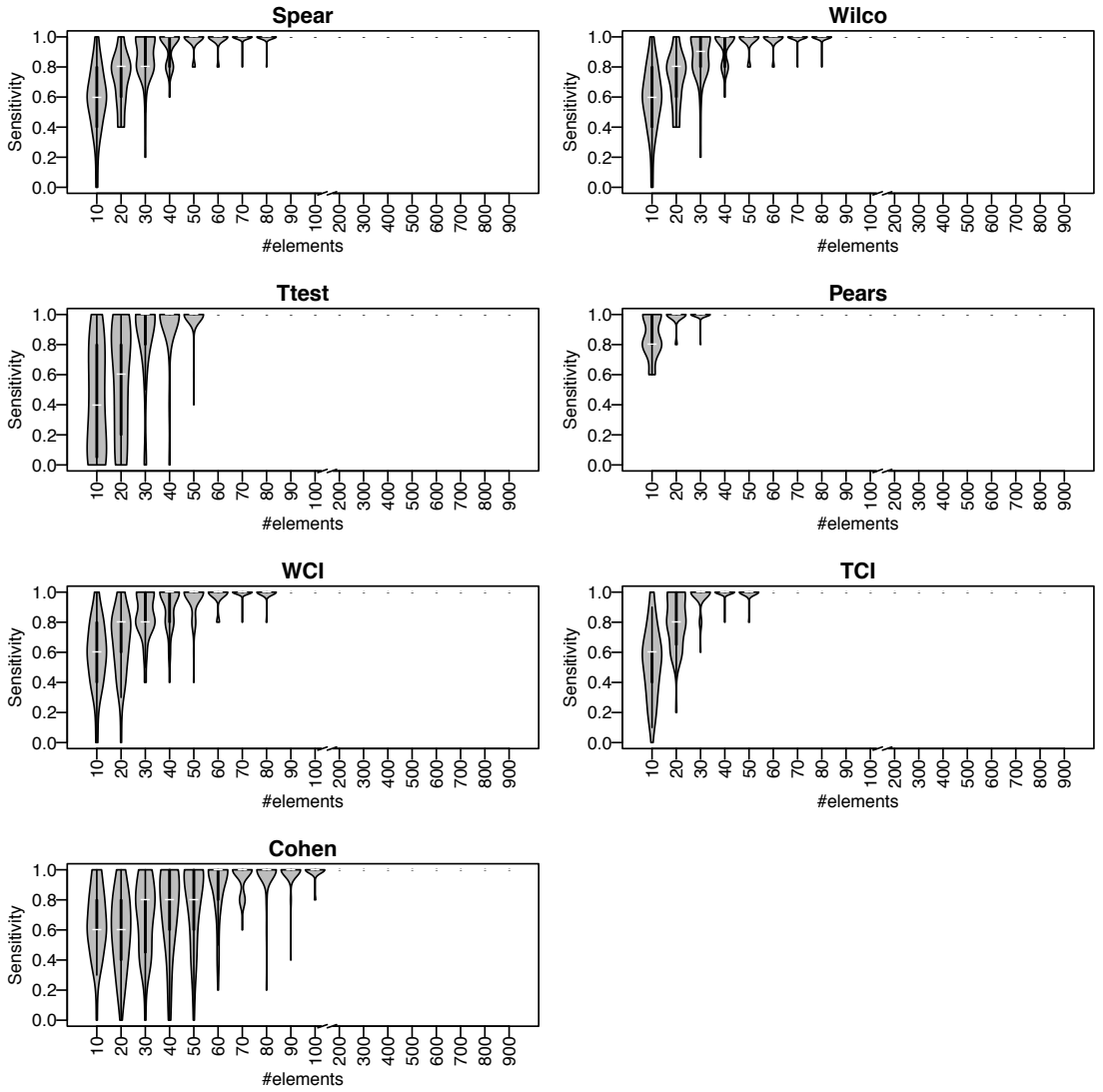


Figure A.11: TFforge's detection sensitivity at a precision of 80% on synthetic data with long trait-loss age
Violin plots show TFforge's detection sensitivity at a precision of 80% on synthetic data with long trait-loss age in dependence of the number of elements. The underlying ranking is indicated by the title of each plot. Abbreviated are correlation coefficient (CC), Wilcoxon rank-sum test (Wilcoxon) and the upper bound of the confidence interval (CI).

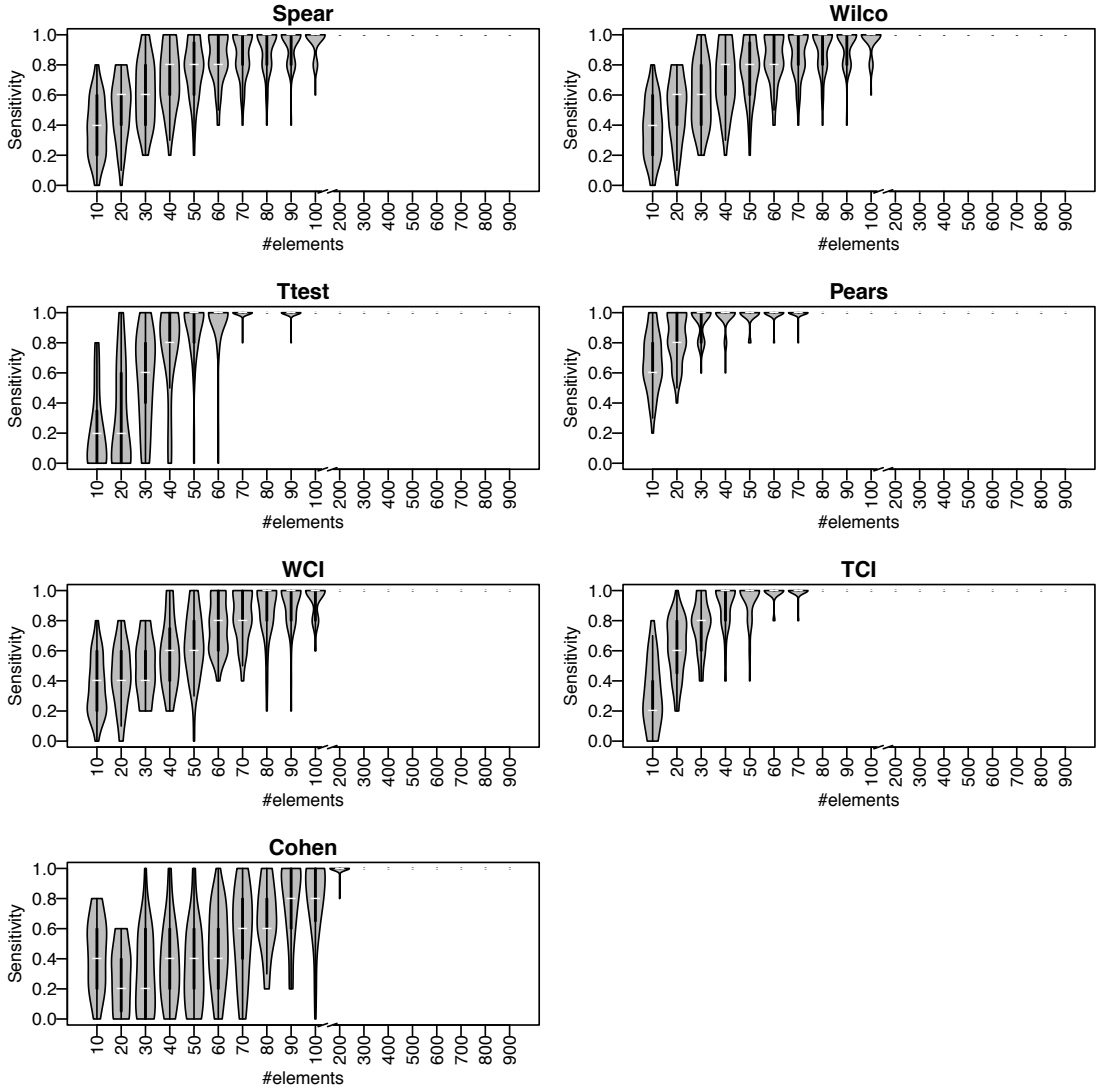


Figure A.12: TFforge's detection sensitivity at a precision of 80% on synthetic data with medium trait-loss age

Violin plots show TFforge's detection sensitivity at a precision of 80% on synthetic data with medium trait-loss age in dependence of the number of elements. The underlying ranking is indicated by the title of each plot. Abbreviated are correlation coefficient (CC), Wilcoxon rank-sum test (Wilcoxon) and the upper bound of the confidence interval (CI).

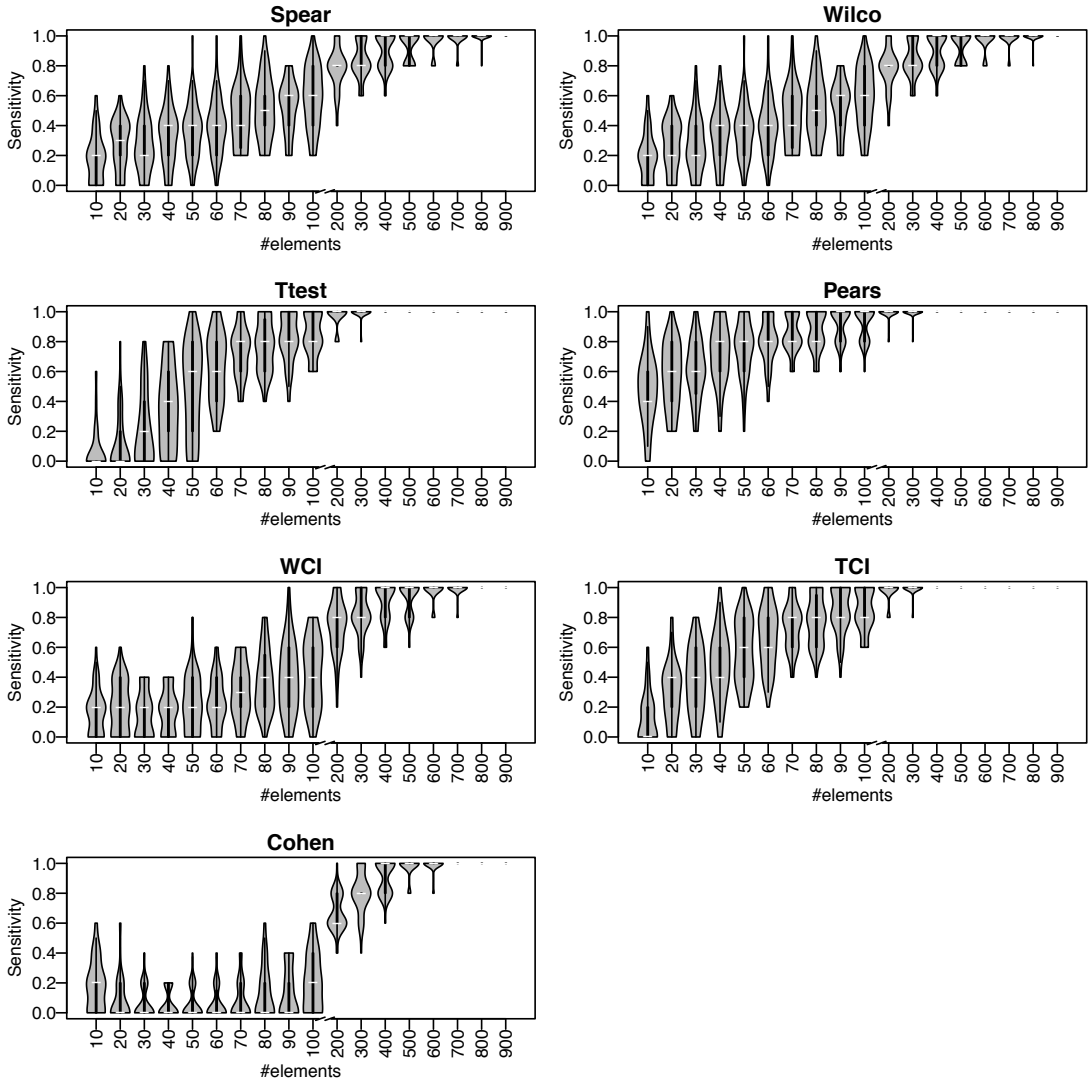


Figure A.13: TFforge's detection sensitivity at a precision of 80% on synthetic data with short trait-loss age
Violin plots show TFforge's detection sensitivity at a precision of 80% on synthetic data with short trait-loss age in dependence of the number of elements. The underlying ranking is indicated by the title of each plot. Abbreviated are correlation coefficient (CC), Wilcoxon rank-sum test (Wilcoxon) and the upper bound of the confidence interval (CI).

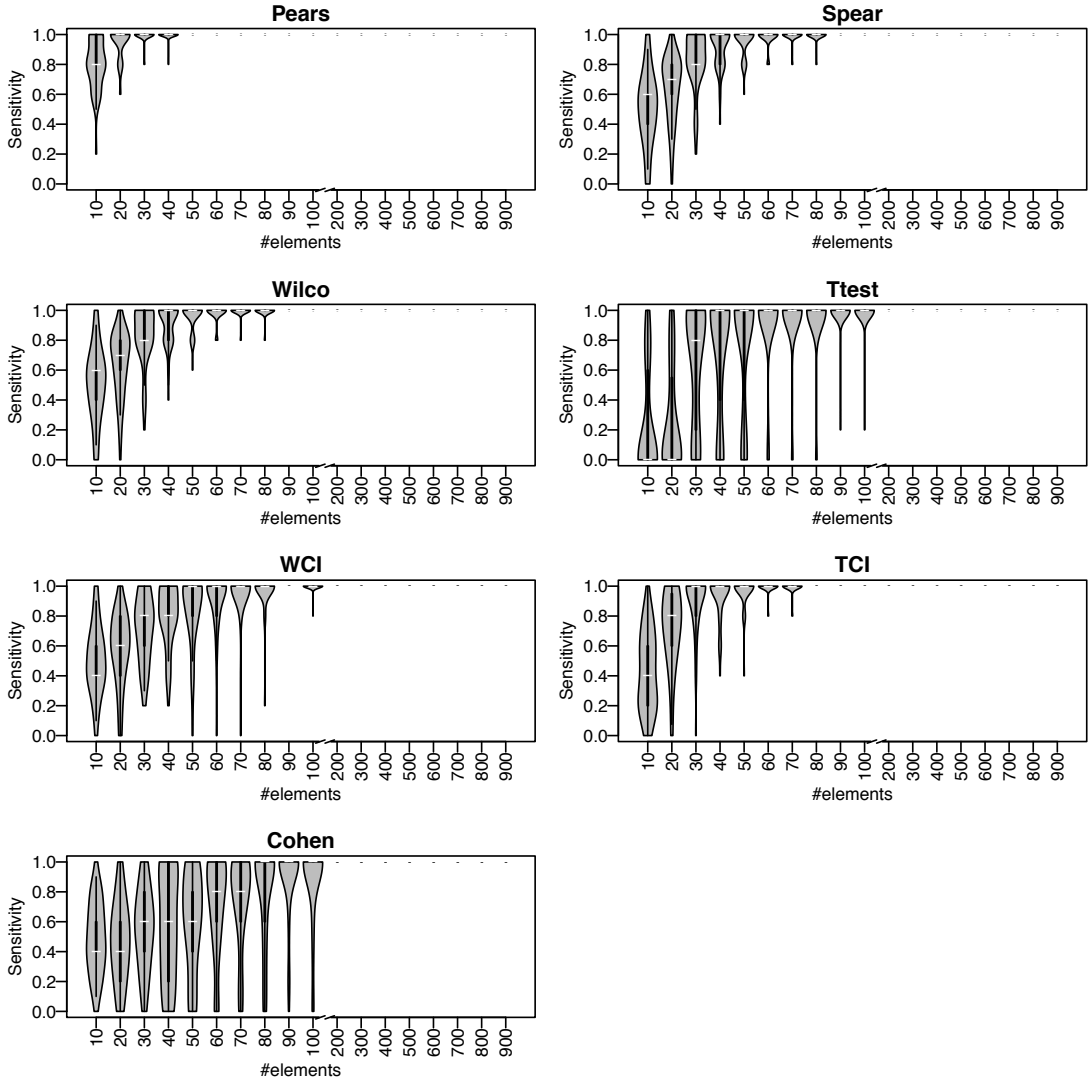


Figure A.14: TFforge's detection sensitivity at a precision of 100% on synthetic data with long trait-loss age
Violin plots show TFforge's detection sensitivity at a precision of 100% on synthetic data with long trait-loss age in dependence of the number of elements. The underlying ranking is indicated by the title of each plot. Abbreviated are correlation coefficient (CC), Wilcoxon rank-sum test (Wilcoxon) and the upper bound of the confidence interval (CI).

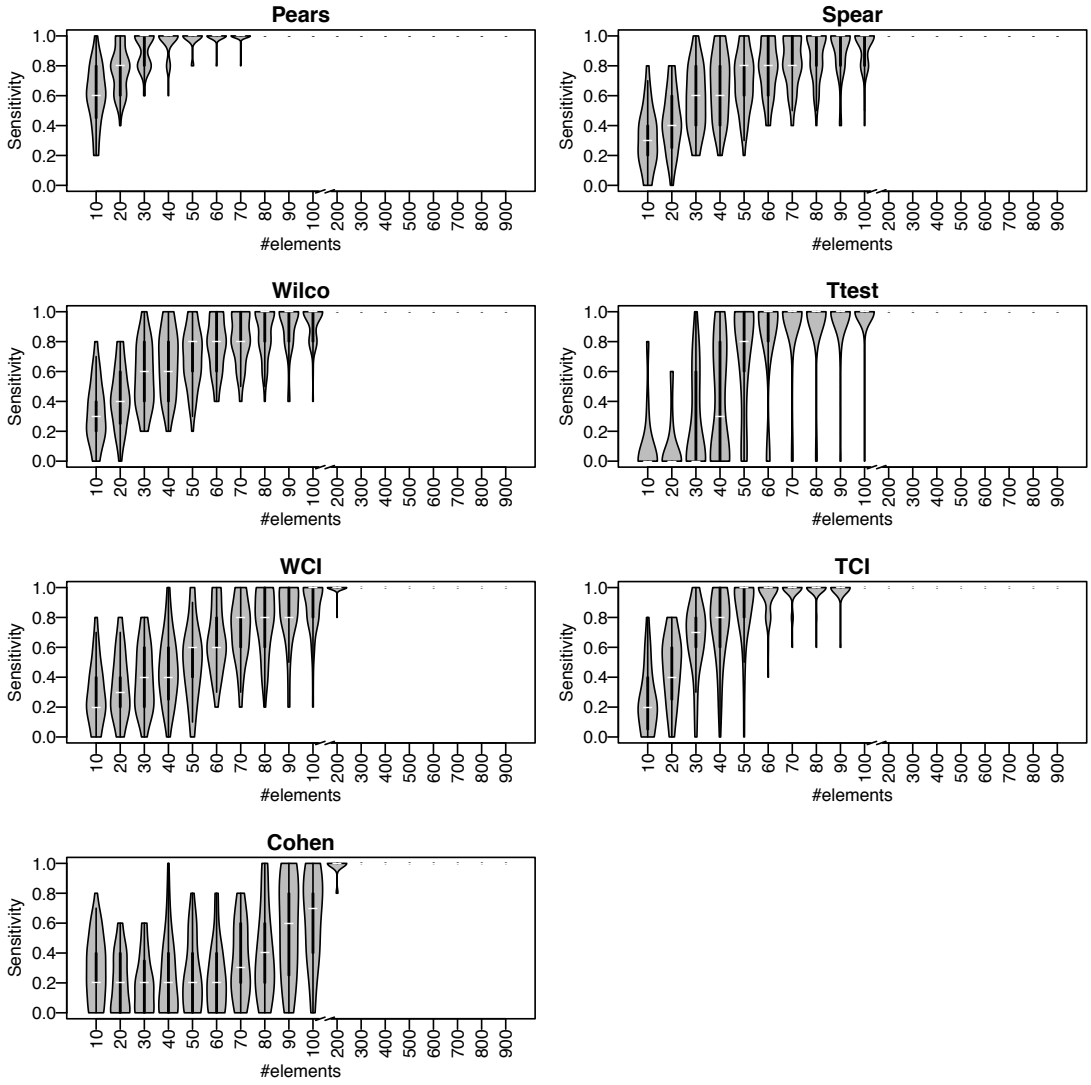


Figure A.15: TFforge's detection sensitivity at a precision of 100% on synthetic data with medium trait-loss age

Violin plots show TFforge's detection sensitivity at a precision of 100% on synthetic data with medium trait-loss age in dependence of the number of elements. The underlying ranking is indicated by the title of each plot. Abbreviated are correlation coefficient (CC), Wilcoxon rank-sum test (Wilcoxon) and the upper bound of the confidence interval (CI).

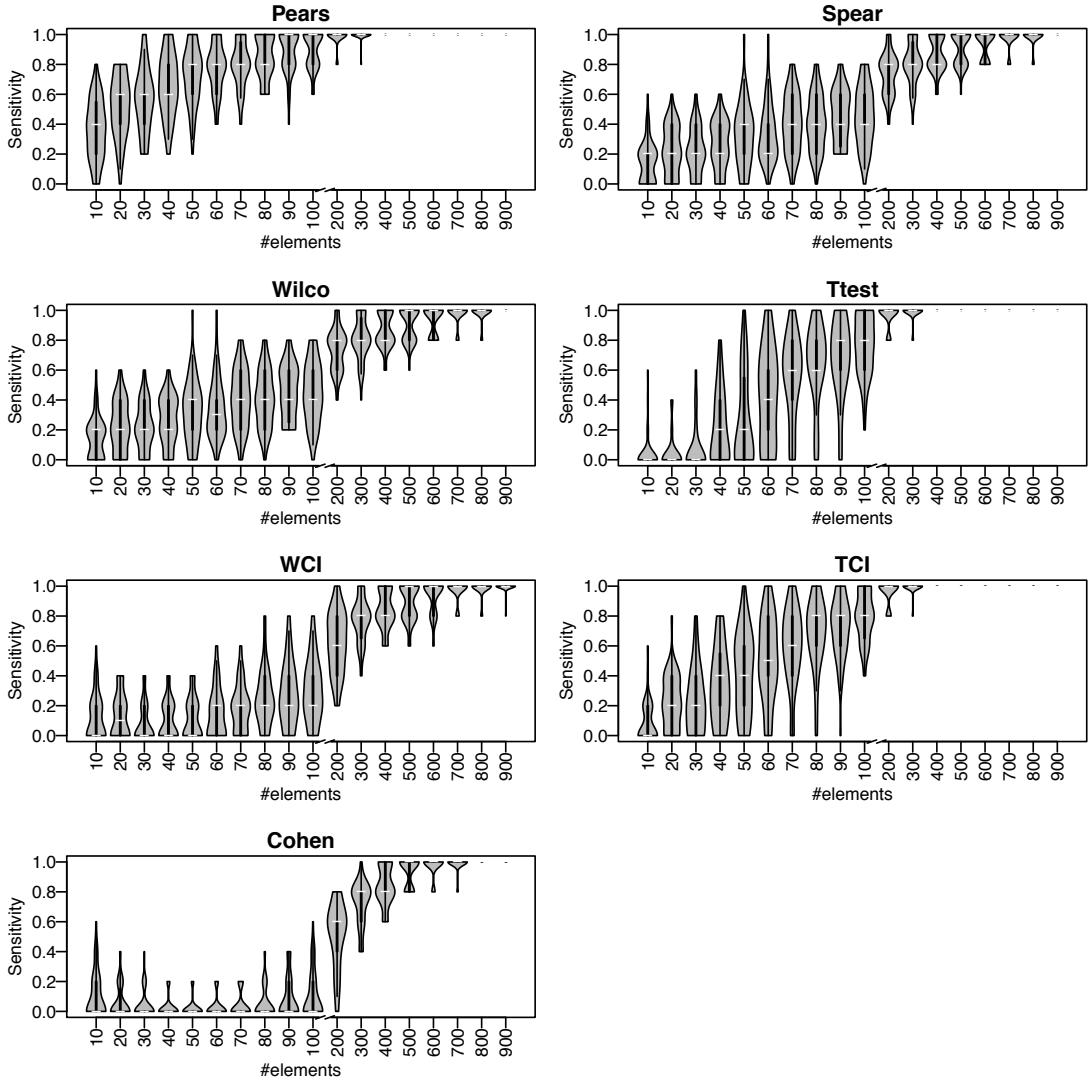


Figure A.16: TFforge's detection sensitivity at a precision of 100% on synthetic data with short trait-loss age
Violin plots show TFforge's detection sensitivity at a precision of 100% on synthetic data with short trait-loss age in dependence of the number of elements. The underlying ranking is indicated by the title of each plot. Abbreviated are correlation coefficient (CC), Wilcoxon rank-sum test (Wilcoxon) and the upper bound of the confidence interval (CI).

Appendix B. Software

The following version of software were used:

Python	3.5.1
R	3.1
PEBCRES	1.02
Tomtom	4.10.0
Stubb	2.1
PRANK	v.170427

Appendix C. References

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. Science 304:1321-1325.

Booker BM, Friedrich T, Mason MK, VanderMeer JE, Zhao J, Eckalbar WL, Logan M, Illing N, Pollard KS, Ahituv N. 2016. Bat Accelerated Regions Identify a Bat Forelimb Specific Enhancer in the HoxD Locus. PLoS Genet 12:e1005738.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134:25-36.

Corbo JC, Lawrence KA, Karlstetter M, Myers CA, Abdelaziz M, Dirkes W, Weigelt K, Seifert M, Benes V, Fritsche LG, et al. 2010. CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. Genome Research 20:1512-1525.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Research 14:1188-1190.

Cvekl A, Mitton KP. 2010. Epigenetic regulatory mechanisms in vertebrate eye development and disease. Heredity (Edinb) 105:135-151.

Duque T, Samee MA, Kazemian M, Pham HN, Brodsky MH, Sinha S. 2014. Simulations of enhancer evolution provide mechanistic insights into gene regulation. Molecular Biology and Evolution 31:184-200.

Farley EK, Olson KM, Zhang W, Brandt AJ, Rokhsar DS, Levine MS. 2015. Suboptimization of developmental enhancers. Science 350:325-328.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics (Oxford, England) 27:1017-1018.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. Genome Biol 8:R24.

Hafez D, Karabacak A, Krueger S, Hwang YC, Wang LS, Zinzen RP, Ohler U. 2017. McEnhancer: predicting gene expression via semi-supervised assignment of enhancers to target genes. Genome Biol 18:199.

He X, Samee MA, Blatti C, Sinha S. 2010. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. PLoS Computational Biology 6.

Hiller M, Schaar BT, Indjeian VB, Kingsley DM, Hagey LR, Bejerano G. 2012. A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species. Cell Rep 2:817-823.

Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML. 2015. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Research 43:D117-122.

Infante CR, Mihala AG, Park S, Wang JS, Johnson KK, Lauderdale JD, Menke DB. 2015. Shared Enhancer Activity in the Limbs and Phallus and Functional Divergence of a Limb-Genital cis-Regulatory Element in Snakes. Developmental Cell 35:107-119.

Keich U, Gao H, Garretson JS, Bhaskar A, Liachko I, Donato J, Tye BK. 2008. Computational detection of significant variation in binding affinity across two sets of sequences with application to the analysis of replication origins in yeast. BMC Bioinformatics 9:372.

Kim J, Cunningham R, James B, Wyder S, Gibson JD, Niehuis O, Zdobnov EM, Robertson HM, Robinson GE, Werren JH, et al. 2010. Functional characterization of transcription factor motifs

using cross-species comparison across large evolutionary distances. PLoS Computational Biology 6:e1000652.

King M-C, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees.

Korhonen J, Martinmaki P, Pizzi C, Rastas P, Ukkonen E. 2009. MOODS: fast search for position weight matrix matches in DNA sequences. Bioinformatics (Oxford, England) 25:3181-3182.

Kvon EZ, Kamneva OK, Melo US, Barozzi I, Osterwalder M, Mannion BJ, Tissieres V, Pickle CS, Plajzer-Frick I, Lee EA, et al. 2016. Progressive Loss of Function in a Limb Enhancer during Snake Evolution. Cell 167:633-642 e611.

Leal F, Cohn MJ. 2016. Loss and Re-emergence of Legs in Snakes by Modular Evolution of Sonic hedgehog and HOXD Enhancers. Current Biology 26:2966-2973.

Lee J, Myers CA, Williams N, Abdelaziz M, Corbo JC. 2010. Quantitative fine-tuning of photoreceptor cis-regulatory elements through affinity modulation of transcription factor binding sites. Gene Therapy 17:1390-1399.

Lerner LE, Peng GH, Gribanova YE, Chen S, Farber DB. 2005. Sp4 is expressed in retinal neurons, activates transcription of photoreceptor-specific genes, and synergizes with Crx. Journal of Biological Chemistry 280:20642-20650.

Liu XD, Liu PC, Santoro N, Thiele DJ. 1997. Conservation of a stress response: human heat shock transcription factors functionally substitute for yeast HSF. The EMBO journal 16:6466-6477.

Löytynoja A. 2014. Phylogeny-aware alignment with PRANK. Multiple sequence alignment methods:155-170.

Marcovitz A, Jia R, Bejerano G. 2016. "Reverse Genomics" Predicts Function of Human Conserved Noncoding Elements. Molecular Biology and Evolution 33:1358-1369.

Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H, et al. 2014. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Research 42:D142-147.

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, et al. 2006. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Research 34:D108-110.

Nakamura T, Gehrke AR, Lemberg J, Szymaszek J, Shubin NH. 2016. Digits and fin rays share common developmental histories. Nature 537:225-228.

Nitta KR, Jolma A, Yin Y, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EE. 2015. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. Elife 4:e04837.

Partha R, Chauhan BK, Ferreira Z, Robinson JD, Lathrop K, Nischal KK, Chikina M, Clark NL. 2017. Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. Elife 6.

Prudent X, Parra G, Schwede P, Roscito JG, Hiller M. 2016. Controlling for Phylogenetic Relatedness and Evolutionary Rates Improves the Discovery of Associations Between Species' Phenotypic and Genomic Differences. Molecular Biology and Evolution 33:2135-2150.

Roscito JG, Sameith K, Parra G, Langer BE, Petzold A, Moebius C, Bickle M, Rodrigues MT, Hiller M. 2018. in preparation. In.

Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jonsson B, Schluter D, Kingsley DM. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature 428:717-723.

Siddharthan R. 2008. PhyloGibbs-MP: module prediction and discriminative motif-finding by Gibbs sampling. PLoS Computational Biology 4:e1000156.

Sinha S. 2006. On counting position weight matrix matches in a sequence, with application to discriminative motif finding. Bioinformatics (Oxford, England) 22:e454-463.

Sinha S, Blanchette M, Tompa M. 2004. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. BMC Bioinformatics 5:170.

Sinha S, Liang Y, Siggia E. 2006. Stubb: a program for discovery and analysis of cis-regulatory modules. Nucleic Acids Research 34:W555-559.

Sinha S, van Nimwegen E, Siggia ED. 2003. Stubb: A probabilistic method to detect regulatory modules. Bioinformatics (Oxford, England) 19:i292-i301.

Smith AD, Sumazin P, Zhang MQ. 2005. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. Proceedings of the National Academy of Sciences of the United States of America 102:1560-1565.

Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, Ovcharenko I, Ahituv N. 2013. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nature Genetics 45:1021-1028.

Stormo GD. 2000. DNA binding sites: representation and discovery. Bioinformatics (Oxford, England) 16:16-23.

Thomas-Chollier M, Sand O, Turatsinze JV, Janky R, Defrance M, Vervisch E, Brohee S, van Helden J. 2008. RSAT: regulatory sequence analysis tools. Nucleic Acids Research 36:W119-127.

Villar D, Flicek P, Odom DT. 2014. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. Nature Reviews Genetics 15:221-233.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biology 3:e7.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nature Reviews Genetics 8:206-216.

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L. 2015. Ensembl 2016. Nucleic Acids Research 44:D710-D716.

Erklärungen zur Eröffnung des Promotionsverfahrens

1. Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.
2. Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts habe ich Unterstützungsleistungen von folgenden Personen erhalten: Michael Hiller.
3. Weitere Personen waren an der geistigen Herstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich nicht die Hilfe eines kommerziellen Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.
4. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt und ist auch noch nicht veröffentlicht worden.
5. Ich bestätige, dass ich die geltende Promotionsordnung der Fakultät Informatik der Technischen Universität Dresden anerkenne.

Dresden, 3. November